

# **The molecular characterisation of foveal hypoplasia**

Mohammed Adil Mohammed Derar,

Submitted in accordance with the requirements for the degree of  
Doctor of Philosophy

The University of Leeds

School of Medicine

October 2024

I confirm that the work submitted is my own work and that appropriate credit has been given where reference has been made to the work of others.

### **Contributions to chapter 3:**

The author has compiled *SLC38A8* variants identified in foveal hypoplasia from the analysis of peer reviewed articles published before May 2024, the 100,000 genomes project and from additional patients obtained through collaboration with the European Retinal Disease Consortium, updated variants with the Human Genome Variation Society nomenclature, reanalysed the curated variants using a comprehensive suite of *In silico* pathogenicity assessment tools, reclassified variants using the American College of Medical Genetics and Genomics guidelines, submitted the curated *SLC38A8* variants to Leiden Open Variation Database, performed haplotype analysis to confirm a founder variant in *SLC38A8*, generated an *SLC38A8* mutation spectrum in foveal hypoplasia, performed a comparative analysis of deleterious and benign missense variants using homology modelling and variant simulation, conducted statistical validation for the association of deleterious missense variants with an increased electrostatic potential, assessed for *SLC38A8* conservation through multiple sequence alignment in 150 orthologs, evaluated the phenotype reported in the probands with foveal hypoplasia, designed single guide RNAs to knock out *SLC38A8* for functional validation and evaluated *SLC38A8* expression in multiple tissues and cell lines.

The previous PhD candidate at the University of Leeds, Dr Emma C. Lord has identified a novel *SLC38A8* inversion in proband F1310 as part of the local unsolved foveal hypoplasia cohort. Other contributing individuals include Dr Andrew R. Webster and Dr Gavin Arno at the University College London and Prof Elfride De Baere at Ghent University, whom have identified *SLC38A8* variants in their respective cohorts as part of the collaboration with the European Retinal Disease Consortium. Dr Jing Yu who is based at the University of Oxford has developed a bioinformatic tool known as SVRare for the analysis of structural variants in patients recruited to the 100,000 genomes project. Dr Jing Yu has compiled 554060126 structural variants identified in 71408 patients with rare disease. The raw output files containing the annotated but unfiltered structural variants in *SLC38A8* were made accessible in the 100,000 genomes project research environment for further analysis. My supervisors Prof Chris Inglehearn, Dr Carmel Toomes, and Dr Sandra Bell supervised the entire project

**Contributions to chapter 4:**

The author has performed autozygosity mapping on proband F1335, conducted variation segregation analysis based on the identified homozygous *HPS5* missense variant, reanalysed and reannotated the whole genome of proband F1337, analysed the foveal hypoplasia cohort in the 100,000 genomes project using dedicated bioinformatic tools for single nucleotide polymorphisms identification and structural variant discovery, performed haplotype analysis to confirm a founder structural variant in *OCA2* that is restricted individuals of African descent, has reviewed the retinal disorders gene panel used in the 100,000 genomes project and the National Health Service Genomic Medical Centre, reanalysed and validated structural variants identified by SVRare in proband F1369 and proband F1071 and these structural variants were subsequently confirmed and recharacterised using long read sequencing. All variants were reported using the Human Genome Variation Society nomenclature, American College of Medical Genetics and Genomics classification and the likely consequences of the structural variants was determined.

The previous PhD candidate at the University of Leeds, Dr Emma C. Lord has analysed the genomic data belonging to probands F1335 and F1337 but failed to identify the genomic aetiology and thus these cases remain unsolved. The collaborating UK Inherited Retinal Disease Consortium member, Dr Jing Yu has analysed the solved case F1310 for validation purposes and two unsolved foveal hypoplasia cases F1369 and F1071 for structural variant discovery using SVRare. Dr Jing Yu performed the aggregation and annotation of structural variants in the local unsolved foveal hypoplasia cohort. My supervisors, Dr Carmel Toomes, Prof Chris Inglehearn and Dr Sandra Bell directed the project.

**Contributions to chapter 5:**

The author has performed autozygosity mapping on proband F1288 for a reanalysis of the whole exome, has analysed variants detected using *In silico* pathogenicity assessment tools, reclassified variants using the American College of Medical Genetics and Genomics guidelines, performed multiple sequence alignment for each candidate variant, conducted segregation analysis of candidate variants using Sanger sequencing, performed restriction fragment length polymorphism for confirmation of segregation pattern, verified *LAMP1* variant as the most plausible candidate gene for foveal hypoplasia in proband F1288, evaluated *LAMP1* conservation using 150 orthologs, analysed the *LAMP1*

cohort in the 100,000 genomes project, analysed *LAMP1* variant site for canonical and noncanonical splice sites, generated *LAMP1* midigene splice assays along with the corresponding isogenic control, verified the *LAMP1* midigene splice vectors using long read sequencing, designed single guide RNAs and the single stranded oligonucleotide donor template for *LAMP1* knock in, investigated *LAMP1* expression in cell lines and tissues, assessed the transfection efficiency in ARPE19 cell line, performed plasmid purification, performed restriction digest of 10 plasmids, generated DNA libraries for sequencing using Oxford Nanopore Technology, developed a novel bioinformatic pipeline to distinguish reads of different origin without the need of barcode sequences, developed a complementary *de novo* assembly pipeline, produced and contrasted read statistics from reference and *de novo* assemblies.

The previous PhD candidate at the University of Leeds, Dr Emma C. Lord has analysed the whole exome data belonging to proband F1288 and has highlighted *LAMP1* as a promising candidate gene for foveal hypoplasia. The PhD student at the University of Leeds, Dr Benjamin McClinton introduced the author to the long read sequencing technology and the standard bioinformatic pipeline for BAM file generation, Dr Christopher Watson at the University of Leeds provided review on the data and my supervisors Dr Carmel Toomes, Prof Chris Inglehearn and Dr Sandra Bell oversaw the completion of the project.

This copy has been supplied on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

The right of Mohammed Adil Mohammed Derar to be identified as Author of this work has been asserted by him in accordance with the Copyright, Designs and Patents Act 1988.

© 2023 The University of Leeds and Mohammed Adil Mohammed Derar

## Acknowledgements

I would like to begin by acknowledging the efforts of my supervisors, Prof Chris Inglehearn, Dr Sandra Bell and Dr Carmel Toomes, for their expertise and in overseeing the completion of my doctoral study. My academic journey is a collective effort owing to the contribution of my lead supervisor, Prof Chris Inglehearn, who provided constructive criticism to redirect my project and provided ample opportunities for my growth and development as a scientist. I would like thank Prof Chris Inglehearn for providing additional support and supervision during the COVID-19 pandemic. I would also like to thank Dr Carmel Toomes whose input was crucial in troubleshooting and her diverse experience allowed me to adopt various experimental techniques to produce high quality data. I am also grateful for the opportunities Dr Carmel Toomes has organised such as the research placement at Radboud University Hospital, Netherlands under the supervision of Dr Susanne Roosing, which maximised my learning experience in such a stimulating environment. The involvement of my co-supervisor, Dr Sandra Bell was highly appreciated as it facilitated handling tissue cultures and cloning techniques to successfully achieve my experimental objectives with great confidence. The regular evaluations of my progress by the three aforementioned supervisors allowed me to attain higher levels of critical thinking and sound judgement to produce works of publishable quality.

Special thanks to Dr Fiona Errington-Mais and Dr Nicolas Orsi who provided additional support to make sure that my thesis deadlines were met and ensured the completion of my doctoral study within the allocated time frame. I would also like thank the staff at the Wellcome Trust Brenner Building, St James's University Hospital including Ewa Jaworska, for maintaining high laboratory standards to ensure my wellbeing and the integrity of my experiments. Other supportive academic staff members include Dr Manir Ali whose conversations motivated me to persevere during hard times. I would also like to extend my gratitude to Dr James Poulter and Dr Christopher Watson at St James's University hospital for sharing their expertise in protein modelling and Oxford Nanopore's DNA sequencing, respectively.

It has been a pleasure to work alongside my colleagues at the Leeds Vision Research Group, Benjamin McClinton, Dong Sun, George Parpas, Esra Ermis, Danah Albuainain, Erica Harris, Bushra Ahmed, Ahlam Altowairqi, Afrah Alshammari, Ummey Haney and Katarzyna Szymanska who made my doctoral journey more rewarding with their pleasant interactions.

The emotional and personal backing from my family played a significant role in my perseverance during my academic endeavours. I do recognise the love received from my family, Adil Dirar, Nemat Magzoub, Dalya Dirar, Dr Waad Dirar, and Prof Hatim Dirar for their constant positive reinforcements. This also includes special friends whom I consider my family, Dr Wael Osman and Mohamed Ahmed.

I am also eternally grateful to the Sanctuary Scholarship team who have funded this doctoral study allowing me to undertake world leading research at the University of Leeds and The Turing scheme committee for their financial contribution towards my overseas research placement in the Netherlands.

## Abstract

Foveal hypoplasia (FH) is a phenotype observed in association with a series of rare and overlapping inherited retinal developmental diseases, which hinders diagnosis in patients. A recessive manifestation known as foveal hypoplasia 2 (FVH2), comprises of FH without pigmentary abnormalities, and is caused by pathogenic variants in *SLC38A8*.

This doctoral study began with genetic characterisation of *SLC38A8* variants in FH, achieved by analysing an FH cohort from the European Retinal Disease Consortium, 100,000 genomes project (100KGP) and cases in the published literature. This expanded the known *SLC38A8* mutation spectrum to 51 unique variants in 63 FVH2 affected probands. Haplotype analysis in an Indian, Pakistani and Bangladeshi patients harbouring the NP\_001073911.1:p.(Tyr88\*) variant confirms a founder mutation restricted to people of South Asian ancestry. Homology modelling and variant simulation suggest loss of function as the underlying disease mechanism in FVH2. Reviewing phenotypes in the 63 FVH2 affected probands confirms that the *SLC38A8* phenotype constitutes FH and chiasmal misrouting, with limited evidence to support the presence of hypopigmentation in a few cases.

In a parallel study, whole genome sequencing in local patients and the 100KGP has identified 19 unique single nucleotide variants and 11 structural variants that cause FH. Amongst these are novel SVs affecting *PAX6* and *GPR143*, which were redefined using Nanopore sequencing. The genomic diagnosis in 26 FH probands was attributed to *PAX6* (5), *GPR143* (5), *OCA2* (4), *SLC38A8* (4), *TYR* (2), *CACNA1F* (2), *HPS5* (1), *CNGA3* (1), *CNGB3* (1) and *ZNF408* (1). The diagnostic yield achieved from the analysis of patients in both the 100KGP and the local cohort with the clinical description of FH was 25% (14/56). The study also identified a founder SV in *OCA2* that occurs at a high frequency in the African population. The haplotype in two African and an Afro-Caribbean patients with NC\_000015.10: g.28017719\_28020673del in *OCA2* support the hypothesis.

Finally, autozygosity mapping and variant analysis in an unsolved FH patient highlighted *LAMP1* as a candidate FH gene. Midigene splicing constructs were generated to assess the molecular consequence of *LAMP1* variant NP\_005552.3:p.(Gly370Alafs\*14). These vectors were validated using a novel bioinformatic workflow developed for a cost effective, multiplexed and barcode free assembly of plasmids using Nanopore sequencing.



## Table of Contents

<b>Intellectual Property Rights and Publication Statements.....</b>	<b>i</b>
<b>Acknowledgements.....</b>	<b>iv</b>
<b>Abstract .....</b>	<b>vi</b>
<b>Table of Contents.....</b>	<b>viii</b>
<b>List of Figures.....</b>	<b>xviii</b>
<b>List of Tables .....</b>	<b>xxiii</b>
<b>Abbreviations.....</b>	<b>xxv</b>
<b>Chapter 1 : Introduction .....</b>	<b>1</b>
1.1 General overview .....	1
1.2 Eye anatomy and physiology .....	2
1.3 Ocular embryogenesis .....	5
1.4 Retinal architecture .....	9
1.5 Photopigments .....	12
1.6 The phototransduction cascade .....	14
1.7 The visual cycle (Retinoid cycle).....	15
1.8 The fovea .....	17
1.9 Foveal pit development .....	18
1.10 Inherited retinal disease (IRDs).....	20
1.11 Retinitis pigmentosa and allied disease .....	21
1.12 Foveal hypoplasia .....	24
1.13 Genetics of Isolated FH.....	28
1.13.1 <i>PAX6</i> .....	28
1.13.2 <i>SLC38A8</i> .....	28
1.13.3 <i>AHR</i> .....	29
1.13.4 <i>FRMD7</i> .....	29
1.14 Albinism.....	30
1.15 Genetics of Albinism .....	32
1.15.1 Oculocutaneous albinism (OCA) .....	32
1.15.2 Ocular Albinism (OA).....	32
1.15.3 OCA syndromes .....	33
1.15.3.1 Hermansky-Pudlak syndrome (HPS) .....	34

1.15.3.2 Chediak-Higashi syndrome (CHS).....	35
1.16 Albinism phenotypic overlap with <i>SLC38A8</i> isolated FH .....	36
1.17 Chiasmal misrouting.....	37
1.18 Next generation sequencing (NGS) .....	39
1.18.1 Second generation sequencing (Short read).....	40
1.18.2 NGS applications in genomic medicine .....	42
1.18.2.1 Whole exome sequencing (WES).....	42
1.18.2.2 Whole genome sequencing (WGS) .....	42
1.18.2.3 Targeted panel-based testing .....	44
1.18.3 Target enrichment .....	44
1.18.4 Limitations of short-read NGS .....	47
1.19 Third generation sequencing (long read) .....	48
1.19.1 Oxford Nanopore Technologies (ONT) platform.....	50
1.20 The UK 100,000 Genomes Project (100KGP) .....	52
1.20.1 Genomics England Clinical Interpretation Partnership (GECIP) ....	54
1.20.2 Bioinformatic pipelines and variant annotations .....	54
1.21 Ethical concerns in DNA sequencing .....	55
1.22 Aims of the project .....	56
<b>Chapter 2 : Materials and Methods .....</b>	<b>58</b>
2.1 Materials.....	58
1X Tris-ethylenediaminetetraacetic acid (TE) buffer (pH 8.0) .....	58
50X Tris acetate ethylenediaminetetraacetic acid (TAE) buffer .....	58
Luria-Bertani (LB) broth .....	58
NZY+ broth .....	58
2.2 Patients .....	58
2.3 Methods .....	59
2.3.1 Polymerase chain reaction (PCR) .....	59
2.3.1.1 Standard PCR.....	59
2.3.1.2 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR) .....	59
2.3.1.3 Long Range Polymerase Chain Reaction (LR-PCR).....	60
2.3.1.4 Whole Genome Amplification (WGA).....	60
2.3.2 Restriction enzyme digestion.....	61
2.3.3 Restriction Fragment Length Polymorphism (RFLP).....	61
2.3.4 Agarose gel electrophoresis .....	61

2.3.5 DNA purification .....	62
2.3.5.1 DNA gel extraction .....	62
2.3.5.2 Column purification .....	62
2.3.5.3 Magnetic beads clean up .....	63
2.3.6 DNA quantification: .....	63
2.3.6.1 NanoDrop™ spectrophotometer .....	63
2.3.6.2 Qubit® fluorometer .....	63
2.3.7 Ethanol precipitation .....	64
2.3.8 Sanger Sequencing .....	64
2.3.9 ONT long read sequencing.....	65
2.4 Molecular cloning .....	65
2.4.1 Gateway cloning .....	65
2.4.2 Site Directed Mutagenesis.....	66
2.4.3 Primer design for site directed mutagenesis .....	67
2.4.4 Bacterial transformation.....	67
2.4.5 Plasmid isolation and purification .....	68
2.4.5.1 Plasmid Miniprep .....	68
2.4.5.2 Plasmid Maxiprep .....	68
2.4.6 Cryopreservation of transformants .....	69
2.5 Tissue Culture .....	69
2.5.1 Cell lines and cell culture .....	69
2.5.2 Storage of cell lines .....	70
2.5.3 Cell counting.....	70
2.5.4 RNA extraction .....	71
2.5.5 RNA quality assessment .....	71
2.5.6 Transfection: Nucleofection .....	71
2.5.7 Microscopy .....	72
2.6 Bioinformatics.....	72
2.6.1 Integrated Genomics Viewer (IGV).....	72
2.6.2 4Peaks (nucleobytes).....	73
2.6.3 Ensembl genome browser.....	73
2.6.4 Genome Aggregation Database (gnomAD).....	73
2.6.5 Online Mendelian Inheritance in Man (OMIM).....	74
2.6.6 Encyclopedia of DNA elements (ENCODE) .....	74
2.6.7 The Human Protein Atlas .....	74

2.6.8 Leiden open variation database (LOVD) .....	74
2.6.9 Primer design .....	75
2.6.10 sgRNA designs for CRISPR-Cas9 mutagenesis .....	75
2.6.11 Autozygosity mapping .....	76
2.6.12 Restriction mapping .....	76
2.6.13 Variant interpretation tools.....	77
2.6.13.1 Grantham Matrix .....	77
2.6.13.2 Sorting Intolerant From Tolerant (SIFT).....	77
2.6.13.3 Polymorphism Phenotyping v2 (PolyPhen-2) .....	77
2.6.13.4 Align-Grantham Variation Grantham Deviation (GVGD).....	78
2.6.13.5 Protein Analysis Through Evolutionary Relationships (PANTHER).....	78
2.6.13.6 Combined Annotation Dependent Depletion (CADD) .....	78
2.6.13.7 MutationTaster .....	79
2.6.13.8 SpliceAI .....	79
2.6.13.9 varSEAK .....	79
2.6.13.10 Ensembl Variant Effect Predictor (VEP) .....	80
2.6.14 Protein sequence alignments .....	80
2.6.15 Computational modelling of proteins and variant simulation .....	81
2.6.15.1 Protein homology modelling.....	81
2.6.15.2 Variant Simulation.....	81
2.6.16 Statistical Package for Social Sciences (SPSS).....	82
2.7 The 100KGP workflows.....	82
2.7.1 Targeted analysis of <i>SLC38A8</i> cohort in the rare disease cohort ....	82
2.7.2 Analysis of participants with a confirmed FH diagnosis.....	83
2.7.3 Haplotype analysis using whole genome VCF files.....	84
2.7.4 Gene specific analysis of non coding and coding variants in the rare disease cohort .....	84
2.7.5 Gene specific analysis of SVs in the rare disease cohort.....	85
2.8 Bioinformatic pipelines .....	85
2.8.1 Quality assessment of BAM files using FASTQC .....	85
2.8.2 Annotation of VCF files using VEP and additional plug ins .....	86
2.8.3 Multiplexing assembly of plasmids using ONT sequencing.....	86

<b>Chapter 3 : Spectrum of genetic variation in <i>SLC38A8</i> and associated phenotype.....</b>	<b>88</b>
3.1 Introduction .....	88
3.2 Results .....	89
3.2.1 The reanalysis of novel <i>SLC38A8</i> variants obtained through collaboration .....	89
3.2.2 Bioinformatic analysis of the 100KGP datasets.....	93
3.2.2.1 Targeted analysis of the biallelic <i>SLC38A8</i> cohort .....	93
3.2.2.2 Haplotype analysis of a potential founder mutation in <i>SLC38A8</i> .....	98
3.2.2.3 Analysis of the monoallelic <i>SLC38A8</i> cohort .....	100
3.2.2.4 Probing for deep intronic variants in <i>SLC38A8</i> .....	103
3.2.2.5 Detecting SVs in <i>SLC38A8</i> .....	106
3.2.2.6 Summary of the 100KGP findings.....	108
3.2.3 Defining an <i>SLC38A8</i> mutation spectrum .....	108
3.2.3.1 Interrogating the published literature .....	108
3.2.3.2 <i>SLC38A8</i> mutation spectrum in FH .....	111
3.2.4 Computational protein modelling of <i>SLC38A8</i> and missense variants simulation .....	114
3.2.4.1 Generation of an <i>SLC38A8</i> wildtype and mutant models .....	114
3.2.4.2 Comparative analysis of pathogenic missense variants and common polymorphisms in FVH2 .....	117
3.2.4.3 Assessing <i>SLC38A8</i> amino acid conservation .....	121
3.2.4.4 Evaluating <i>SLC38A8</i> phenotype in a large cohort .....	122
3.2.5 Curating <i>SLC38A8</i> variants for a LOVD .....	124
3.2.6 CRISPR-cas9 knock out of <i>SLC38A8</i> . .....	125
3.2.6.1 sgRNA design and evaluation.....	125
3.2.6.2 Assessing tissues and cell lines for <i>SLC38A8</i> expression.....	127
3.3 Discussion .....	129
3.3.1 Overview of the results .....	129
3.3.2 Expanding on the known <i>SLC38A8</i> mutation spectrum .....	130
3.3.3 Confirming a founder <i>SLC38A8</i> mutation in South Asians.....	132
3.3.4 Insight into missense variants pathogenic mechanism in FVH2 ....	134
3.3.5 Redefining the <i>SLC38A8</i> phenotype .....	136
<b>Chapter 4 : Genomic discoveries in unsolved FH cohorts .....</b>	<b>142</b>

4.1 Introduction .....	142
4.2 Results .....	143
4.2.1 Reanalysis of local unsolved FH cases .....	143
4.2.1.1 Summary of local unsolved FH cases.....	143
4.2.1.2 Previous work performed by Toomes's group on proband F1335 .....	144
4.2.1.3 Detecting a homozygous <i>HPS5</i> missense variant using autozygosity mapping in proband F1335 .....	145
4.2.1.4 Segregation analysis of the <i>HPS5</i> missense variant .....	145
4.2.2 Identification of SVs using SVRare in local unsolved FH cases .....	150
4.2.2.1 Pilot study: application of SVRare on solved FH case F1310..	150
4.2.2.2 The reanalysis of a heterozygous <i>PAX6</i> deletion in proband F1369 and a hemizygous <i>GPR143</i> deletion in proband F1071 .....	151
4.2.2.3 Characterisation of identified SVs using ONT sequencing .....	153
4.2.2.4 Investigating the genomic features at the SV breakpoints .....	157
4.2.2.5 Reanalysis of F1377 genome .....	159
4.2.3 Bioinformatics analysis of 100KGP patients diagnosed with FH ....	164
4.2.3.1 The FH cohort in the 100KGP .....	164
4.2.3.2 Identifying pathogenic variants in the 100KGP FH cohort .....	168
4.2.3.3 Identifying SVs in the FH cohort .....	174
4.2.3.4 The discovery of a potential founder SV in albinism .....	176
4.2.3.5 Detecting SVs in additional participants with FH .....	179
4.2.3.6 Reviewing Gene panels associated with FH.....	184
4.3 Discussion .....	189
4.3.1 Overview of the analysis in the local unsolved FH cohort .....	189
4.3.2 Genomic findings in the unsolved local FH cohort .....	190
4.3.2.1 Autozygosity mapping identifies a novel <i>HPS5</i> variant in an FH case .....	190
4.3.2.2 The characterisation of a novel <i>PAX6</i> deletion in aniridia using ONT-based long read sequencing .....	192
4.3.2.3 The identification of a novel <i>GPR143</i> deletion in ocular albinism .....	195
4.3.2.4 Searching for a pathogenic variant in proband F1377 .....	196
4.3.3 100KGP diagnostic odyssey in FH .....	197
4.3.3.1 Overview of the FH cohort .....	197

4.3.3.2 The FH diagnostic yield in the 100KGP .....	198
4.3.3.3 Challenges in defining an FH cohort in the 100KGP .....	199
4.3.3.4 Interpreting pathogenic variants in <i>TYR</i> and <i>OCA2</i> that are enriched in albinism .....	200
4.3.3.5 Revealing a founder SV in <i>OCA2</i> that is of African ancestry ...	203

## **Chapter 5 : The identification of *LAMP1* as a candidate for isolated FH. 207**

5.1 Introduction .....	207
5.2 Results .....	209
5.2.1 Autozygosity mapping in proband F1288 for autosomal recessive gene discovery .....	209
5.2.2 Characterisation of candidate variants identified .....	210
5.2.2.1 <i>In silico</i> pathogenicity assessment.....	210
5.2.2.2 Multiple sequence alignment .....	211
5.2.2.3 Segregation analysis in unsolved FH family F5 .....	214
5.2.3 Confirming <i>CHRNA4</i> variant segregation with FH.....	216
5.2.4 <i>LAMP1</i> as a potential novel FH gene .....	217
5.2.5 Assessing evolutionary conservation in <i>LAMP1</i> .....	218
5.2.6 Assessing the <i>LAMP1</i> variant site for splice sites .....	220
5.2.7 Bioinformatic analysis of the 100KGP .....	222
5.2.8 Generating <i>LAMP1</i> midigene splice vectors.....	228
5.2.9 Generating an isogenic control of the <i>LAMP1</i> midigene .....	233
5.2.10 Multiplexed assembly of recombinant plasmids using barcode-free ONT sequencing.....	237
5.2.11 Generating CRISPR-Cas9 knock-in models of <i>LAMP1</i> .....	246
5.2.11.1 Designing sgRNAs for <i>LAMP1</i> knock-in and knock-out models .....	246
5.2.11.2 Investigating <i>LAMP1</i> expression in available cell lines and tissues .....	249
5.2.11.3 Assessing the transfection efficiency of ARPE19 cell line .....	252
5.3 Discussion .....	254
5.3.1 The identification of a novel pathogenic variant in <i>LAMP1</i> .....	254
5.3.2 Ascribing a potential clinical phenotype to pathogenic variants in <i>LAMP1</i> .....	258
5.3.2.1 <i>LAMP1</i> could arrest foveal development through autophagy ..	259

5.3.2.2 LAMP1 as a potential downstream component in the retinal pigmentation pathway .....	260
5.3.3 A barcode free multiplexing strategy for plasmid assembly using ONT sequencing .....	261
5.3.4 Limitations of the study .....	266
<b>Chapter 6: General discussion .....</b>	<b>268</b>
6.1 Molecular characterisation of <i>SLC38A8</i> in <i>FVH2</i> .....	268
6.2 The application of WGS in FH .....	269
6.3 The unsolved FH cases .....	270
6.4 Technologies to improve the diagnostic yield in FH .....	273
6.4.1 Long read sequencing for comprehensive variant discovery .....	273
6.4.2 Optical genome mapping (OGM) in SV identification .....	276
6.5 Diagnostic limitation of the 100KGP .....	278
6.6 Future direction .....	280
6.6.1 Insight into molecular consequences of <i>SLC38A8</i> variants in <i>FVH2</i> .....	280
6.6.1.1 Elucidating the pathogenic mechanisms of missense variants .....	280
6.6.1.2 Finding an appropriate cell line expressing <i>SLC38A8</i> .....	281
6.6.2 Maximising the translational potential of genomic findings from the 100KGP .....	282
6.6.2.1 Phasing alleles in an African founder haplotype responsible for albinism .....	282
6.6.3 Assessing <i>LAMP1</i> as a candidate gene underlying FH .....	283
6.6.3.1 Testing whether the NM_005561.4:c.1109delG variant alters <i>LAMP1</i> splicing .....	284
6.6.3.2 Genome editing <i>LAMP1</i> in ARPE19 cell line and evaluation of protein trafficking .....	284
6.6.3.3 smMIP screen of <i>LAMP1</i> in unsolved IRD cases .....	285
<b>References .....</b>	<b>286</b>
<b>Chapter 8 Appendix .....</b>	<b>318</b>
8.1 Appendix A : gene list and phenotypic keywords in filtering .....	318
8.1.1 All the known genes underlying FH .....	318
8.1.2 Curated FH genes for a gene-specific analysis using SVRare .....	318



8.1.3 HPO terms to identify a suspected FH phenotype in the 100KGP .	319
8.2 Appendix B : Scripts .....	320
8.2.1 Gene-Variant workflow v1.6 for the analysis in the 100KGP .....	320
1. Copy the workflow to own directory: .....	320
2. Specify genes .....	320
3. Specify GECIP project and dataset for analysis: .....	320
4. Define project: re_gecip_hearing_and_sight .....	321
8.2.2 100K Gene_Variant filter configuration and execution .....	321
1. Copy script to own directory: .....	321
2. Activate idppy3 .....	321
3. Run filtering script on Gene-Variant Workflow output: .....	321
8.2.3 ONT pipeline for MinION sequencing using the Flongle R9.4.1 .....	322
1. Guppy script v5 .....	322
2. Installing Miniconda 3 .....	322
3. Performing base calling using Guppy v5: .....	323
4. Concatenating FASTQ files .....	323
5. Adaptor trimming by Porechop .....	323
6. Obtain read statistics on the trimmed FASTQ file using NanoStat .....	323
7. Filter reads based on quality and length using NanoFilt .....	323
8. Obtain read statistics on the filtered FASTQ file using NanoStat .	323
9. Alignment of reads to reference sequence using Minimap2 .....	324
10. Sort and index BAM files using Samtools .....	324
8.2.4 Quality control using FASTQC .....	324
8.2.5 Genome wide annotation using VEP plugins of CADD, SpliceAI and UTRannotator .....	324
1. Script to download the Ensembl VEP .....	324
2. Script to download VEP plugins of SpliceAI, CADD and UTRannotator .....	324
3. Script for VEP annotation using GRCh37 .....	326
8.2.6 Splitting VCF by chromosome and subset by FH gene .....	327
1. Install BCFTiiks v1.21 using Conda package repository .....	327
2. Subset VCF files by chromosome .....	327
3. Unzip the compressed VCF files .....	327
3. Subset VCF files by gene of interest .....	327

8.2.7 Scripts and bioinformatic commands for ONT sequencing in plasmid assembly .....	328
1. Guppy script v5 .....	328
2. Installing Miniconda 3.....	328
3. Performing base calling using Guppy v5.....	329
4. Concatenating FASTQ files.....	329
5. Adaptor trimming by Porechop .....	329
6. Obtain read statistics on the trimmed FASTQ file using NanoStat	329
7. Filter reads based on quality and length using NanoFilt.....	329
8. Obtain read statistics on the filtered FASTQ file using NanoStat .	330
9. Alignment of reads to reference sequence using Minimap2 .....	330
11. Removing Soft clipped reads .....	330
12. Extract reads using a plasmid's unique sequence .....	330
13. Install Canu .....	330
14. Perform <i>de novo</i> assembly .....	330
8.3 Appendix C: Figures and Tables.....	331

## List of Figures

Figure 1.1 Anatomy of the eye.....	2
Figure 1.2 Embryonic developmental timeline of the vertebrate eye.....	6
Figure 1.3 Retinal lamination in humans .....	10
Figure 1.4 Photoreceptor schematic and rhodopsin structure .....	13
Figure 1.5 Phototransduction in rods.....	15
Figure 1.6 The retinoid cycle. ....	16
Figure 1.7 The fovea in primates. ....	18
Figure 1.8 Stages of foveal pit development. ....	19
Figure 1.9 Graph of mapped and identified IRD genes. ....	21
Figure 1.10 Characteristics of FH detected using in vivo imaging.....	26
Figure 1.11 Clinical phenotype of albinism. ....	31
Figure 1.12 Genes underlying albinism .....	33
Figure 1.13 The optic pathway in humans.....	37
Figure 1.14 The visual pathway in albinism and <i>SLC38A8</i> isolated FH .....	39
Figure 1.15 Illumina NGS workflow. ....	41
Figure 1.16 smMIPs target enrichment workflow.....	46
Figure 1.17 Oxford nanopore sequencing overview.....	52
Figure 2.1 Barcode free assembly of multiplexed plasmids using ONT sequencing. ....	87
Figure 3.1 FH families solved through the identification of biallelic <i>SLC38A8</i> variants. ....	91
Figure 3.2 Characterisation of SVs identified in the local FH cohort. ....	92
Figure 3.3 Biallelic <i>SLC38A8</i> variants affecting amino acid residue 308.....	97
Figure 3.4 Haplotype of participants harbouring p.(Tyr88*) in <i>SLC38A8</i> .....	99
Figure 3.5 <i>SLC38A8</i> inversion detected in the 100KGP.....	107
Figure 3.6 <i>SLC38A8</i> mutation spectrum.....	112

Figure 3.7 Distribution of pathogenic variants in SLC38A8. ....	113
Figure 3.8 Quality assessment of swiss-model SLC38A8 protein structure. ..	116
Figure 3.9 Predicted effects of missense changes and common polymorphisms on SLC38A8 .....	118
Figure 3.10 Statistical validation of missense effects on surface electrostatic potentials .....	119
Figure 3.11 Localisation of missense variants in the SLC38A8 protein.....	120
Figure 3.12 Conservation profile of SLC38A8 .....	121
Figure 3.13 Clinical manifestations of <i>SLC38A8</i> . ....	123
Figure 3.14 <i>SLC38A8</i> database in LOVD .....	124
Figure 3.15 Guide RNAs designed for <i>SLC38A8</i> knock out and sequencing primers design. ....	126
Figure 3.16 SLC38A8 gene expression profile. ....	128
Figure 4.1 ROH identified in proband F1335 .....	145
Figure 4.2 NGS discovery of <i>HPS5</i> variant in F1335 .....	146
Figure 4.3 Validation of WGA in two unsolved FH families. ....	148
Figure 4.4 Segregation analysis of <i>HPS5</i> variant in family F5.....	149
Figure 4.5 Genetic characterisation of deletions in F1369 and F1071 .....	152
Figure 4.6 Optimisation of LR-PCR on unsolved FH cases.....	154
Figure 4.7 Characterisation of <i>PAX6</i> deletion using long read sequencing....	155
Figure 4.8 Characterisation of an SV in F1301 using long read sequencing..	156
Figure 4.9 Annotation of genomic region at the breakpoints of characterised SVs in F1369 and F1071 .....	158
Figure 4.10 Sequence homology surrounding the nc_000023.11:g.9384915_9982211del breakpoints in F1071 .....	158
Figure 4.11 NGS identification of NM_001080442.3:c.2t>c, NP_001073911.1:p.(Met1?) in F1377 .....	159
Figure 4.12 Read quality control checks using F1377 BAM file .....	161

Figure 4.13 Variant statistics in the F1377 genome. ....	162
Figure 4.14 Characteristics of the 100KGP FH cohort. ....	165
Figure 4.15 Specific disease diagnoses of FH probands. ....	167
Figure 4.16 Confirmation of detected variants in 10 probands with FH.....	173
Figure 4.17 <i>OCA2</i> genotype in proband P18 with unsolved FH.....	175
Figure 4.18 Characteristics of participants with NC_000015.10:g.28017719_28020673del in <i>OCA2</i> . ....	177
Figure 4.19 Validation of SVs identified in the 100KGP. ....	183
Figure 4.20 Isolated FH genes reviewed in Retinal Disorders v5.11.....	186
Figure 4.21 Categorisation of genes in the Retinal Disorders v5.11 panel.....	187
Figure 4.22 <i>PAX6</i> protein schematic. ....	193
Figure 4.23 Protein alignment of <i>TYR</i> and <i>OCA2</i> orthologs.....	202
Figure 5.1 Pedigree of unsolved FH family F5.....	208
Figure 5.2 Autozygosity mapping in proband F1288. ....	209
Figure 5.3 Multiple sequence alignments in candidate FH genes. ....	213
Figure 5.4 Confirmation and segregation of the candidate variants. ....	215
Figure 5.5 <i>CHRNA4</i> restriction fragments pattern in family F5.....	216
Figure 5.6 Variant prioritisation strategy. ....	218
Figure 5.7 <i>LAMP1</i> conservation profile and protein model.....	219
Figure 5.8 Splice sites in region downstream intron 7 of <i>LAMP1</i> .....	221
Figure 5.9 100KGP case with potential FH.....	223
Figure 5.10 WGS analysis in 100KGP proband with <i>LAMP1</i> variant and a complex phenotype.....	224
Figure 5.11 <i>LAMP1</i> midigene schematic. ....	229
Figure 5.12 Restriction digest of <i>LAMP1</i> midigenes.....	230
Figure 5.13 Validation of <i>LAMP1</i> midigenes.....	232
Figure 5.14 Genotyping SMD <i>LAMP1</i> midigenes using RFLP.....	235

Figure 5.15 ONT sequencing of the <i>LAMP1</i> WT SDM midigene vector generated by SDM.....	236
Figure 5.16 Plasmids included in multiplexed nanopore sequencing. ....	239
Figure 5.17 Linearisation of plasmids using restriction enzymes. ....	240
Figure 5.18 Bioinformatics pipeline used to demultiplex plasmid reads. ....	241
Figure 5.19 Plasmid assembly using multiplexed nanopore sequencing. ....	243
Figure 5.20 Genome editing strategy for <i>LAMP1</i> knock in. ....	248
Figure 5.21 Tissue expression of <i>LAMP1</i> and <i>LAMP2</i> . ....	251
Figure 5.22 Fluorescent imaging of GFP transfected ARPE19 cells. ....	253
Figure 5.23 <i>LAMP1</i> protein schematic.....	257
Supplementary Figure 3.1. Aberrant mRNA sequences of 3 SVs identified in the local cohort.....	332
Supplementary Figure 3.2. Variant simulation of <i>SLC38A8</i> missense variants in FH.....	338
Supplementary Figure 3.3. Evolutionary conservation of <i>SLC38A8</i> 6 <sup>th</sup> transmembrane domain .....	340
Supplementary Figure 3.4. Further testing for <i>SLC38A8</i> expression .....	345
Supplementary Figure 4.1 Predicted aberrant <i>PAX6</i> mRNA and its translation.....	350
Supplementary Figure 4.2. <i>GPR143</i> locus with primer binding sites and deletion breakpoints. ....	351
Supplementary Figure 4.3. Primers flanking the 2.7 kb deletion in <i>OCA2</i> .....	353
Supplementary Figure 5.1 PCR amplification of variants in candidate genes	354
Supplementary Figure 5.2 Assessing segregation of candidate variants in affected sibling F1287. ....	355
Supplementary Figure 5.3 <i>LAMP1</i> midigene reference sequence.....	361
Supplementary Figure 5.4 RFLP strategy for genotyping <i>LAMP1</i> SDM clones.....	362

Supplementary Figure 5.5 MinION run report for multiplexed sequencing  
of 10 plasmids..... 364

## List of Tables

Table 1.1 Genetic heterogeneity in RP and associated disease. ....	24
Table 1.2 Genes implicated in FH and the corresponding inheritance pattern. ....	27
Table 1.3 Classification of HPS subtypes and associated phenotypes. ....	35
Table 3.1 <i>SLC38A8</i> biallelic cases with ophthalmic disorders in the 100KGP .	96
Table 3.2 <i>SLC38A8</i> monoallelic cohort in the 100KGP. ....	102
Table 3.3 Hidden variants in the <i>SLC38A8</i> monoallelic cohort. ....	105
Table 3.4 All <i>SLC38A8</i> variants identified in FVH2. ....	110
Table 3.5 Comparison of candidate <i>SLC38A8</i> models. ....	115
Table 4.1 Local cohort of unsolved FH .....	144
Table 4.2 Amino acid conservation at <i>HPS5</i> variant site. ....	147
Table 4.3 Additional variants identified in proband F1377 .....	163
Table 4.4 Solved FH cases in the 100KGP .....	166
Table 4.5 <i>In-silico</i> pathogenicity assessment of the 100KGP FH cohort. ....	171
Table 4.6 Haplotypes of three african descendants heterozygous for NC_000015.10:g.28017719_28020673del. ....	178
Table 4.7 Pathogenic SVs in FH-associated genes identified in the 100KGP	181
Table 4.8 Gene list in the Retinal Disorders v5.11 panel. ....	188
Table 5.1 <i>In-silico</i> pathogenicity assessment of candidate variants. ....	212
Table 5.2 MTMR14 corrected sequence alignment. ....	214
Table 5.3 100KGP rare disease cohort with <i>LAMP1</i> biallelic variants. ....	227
Table 5.4 Variants detected in <i>LAMP1</i> midgenes. ....	233
Table 5.5 Statistical summary of multiplexed sequencing of 10 plasmids. ....	244
Table 5.6 Comparative analysis of <i>de novo</i> and reference sequence assemblies. ....	245
Supplementary Table 3.1 Entire <i>SLC38A8</i> biallelic cohort in the 100KGP. ....	333



Supplementary Table 3.2 Entire <i>SLC38A8</i> monoallelic cohort in the 100KGP.....	334
Supplementary Table 3.3. <i>SLC38A8</i> monoallelic cohort in the 100KGP analysed for exonic and intronic hits.....	336
Supplementary Table 3.4. Concordance in ACMG classification of <i>SLC38A8</i> variants .....	337
Supplementary Table 3.5. Phenotypes of probands with FVH2 in 63 families.....	343
Supplementary Table 3.6. Predicted off target effects of <i>SLC38A8</i> guide RNAs .....	344
Supplementary Table 4.1. <i>HPS5</i> primers design.....	346
Supplementary Table 4.2 <i>PAX6</i> and <i>GPR143</i> primers used in LR-PCR.....	346
Supplementary Table 4.3. VEP annotation of <i>SLC38A8</i> genomic sequence.	349
Supplementary Table 4.4 Haplotype analysis in three families of African descent harbouring NC_000015.10:g.28017719_28020673del .....	352
Supplementary Table 5.1 Off target effects predicted using <i>LAMP1</i> sgRNA1	356
Supplementary Table 5.2 Potential off target using <i>LAMP1</i> sgRNA2 .....	358
Supplementary Table 5.3 Primer sets for <i>LAMP1</i> , <i>LAMP2</i> and <i>TP53</i> .....	358
Supplementary Table 5.4 Plasmids unique features .....	363

## Abbreviations

100KGP	100,000 genomes project
5mC	5-methylcytosine
ABCA4	ATP-binding cassette sub-family A, member 4
ACMG	American college of medical genetics
AP	Adaptor protein complexes
ASD	Anterior segment dysgenesis
ASIC	Application-specific integrated circuit
$\beta$ -ME	Beta-mercaptoethanol
BAM	Binary alignment map
BBS	Bardet-Biedl syndrome
BCVA	Best corrected visual acuity
BLOC	Biogenesis of lysosome-related organelles
BSA	Bovine serum albumin
CADD	Combined annotation dependent depletion
Cas9	Clustered regularly interspaced palindromic repeats - associated protein
CDS	Coding sequence
cGMP	Cyclic guanosine monophosphate
CHS	Chediak-Higashi syndrome
CNG	Cyclic nucleotide gated
COBALT	Constraint-based multiple alignment tool
CRE	Cis-regulatory elements
CRISPR	Clustered regularly interspaced palindromic repeats
dA	Adenosine monophosphate
dELS	Distal enhancer like signatures
dH <sub>2</sub> O	Deionized water
DMEM	Dulbecco's modified eagle medium
DMSO	Dimethyl sulfoxide

DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide triphosphate
DTT	Dithiothreitol
EDTA	Ethylenediaminetetraacetic acid
ENCODE	Encyclopaedia of DNA elements
ER	Endoplasmic reticulum
ERAD	Endoplasmic reticulum associated degradation
ERDC	European retinal disease consortium
ESE	Exonic splicing enhancer
ExAC	Exome aggregation consortium
FAZ	Foveal avascular zone
FBS	Foetal bovine serum
FERM	Band 4.1, ezrin, radixin and moesin domain
FFPE	Formalin fixed paraffin embedded
FH	Foveal hypoplasia
FHONDA	Foveal hypoplasia, optic nerve decussation defect and anterior segment dysgenesis
FVH2	Foveal hypoplasia 2
GABA	$\gamma$ -aminobutyric acid
GC	Guanine and cytosine
GCL	Ganglion cell layer
GDP	Guanosine diphosphate
GECIP	Genomics England Clinical Interpretation Partnership
GEL	Genomics England
GFP	Green fluorescent protein
gnomAD	Genome aggregation database
GQX	genotyping quality
GTP	Guanosine triphosphate
GVGD	Grantham variation grantham deviation
GWAS	Genome wide association studies

GYGQ	SNP rs187887338, p.(Ser192Tyr), rs147546939 and p.(Arg402Gln).
HCl	Hydrochloric acid
HDR	Homology directed repair
HFL	Henle's fibre layer
HiFi	High fidelity
HPO	Human phenotype ontology
HLH	Hemophagocytic lymphohistiocytosis
HPC	High performance computing
HPS	Hermansky-Pudlak syndrome
IGV	Integrated genomics viewer
ILM	Inner limiting membrane
INDEL	Insertions or deletion
INL	Inner nuclear layer
IPL	Inner plexiform layer
ipRGC	Intrinsically photosensitive retinal ganglion cell
IRBP	Interphotoreceptor retinoid binding protein
IRD	Inherited retinal disease
IRE1	Inositol requiring kinase 1
IVA	Interactive variant analysis
L-DOPA	L-3,4-Dihydroxyphenylalanine
LB	Luria-Bertani
LCA	Leber congenital amaurosis
LDDT	Local distance difference test
LINE	Long interspersed nuclear elements
LOF	Loss of function
LOVD	Leiden open variation database
LR-PCR	Long range polymerase chain reaction
LRAT	Lecithin retinol acyltransferase
LRO	lysosome-related organelles

LTR	Long terminal repeat
MAF	Minor allele frequency
MgCl <sub>2</sub>	Magnesium Chloride
MHC	Major histocompatibility complex locus
mRNA	Messenger RNA
mtDNA	Mitochondrial DNA
NaCl	Sodium Chloride
NCBI	National centre for biotechnology information
NCKX	Na <sup>+</sup> /Ca <sup>2+</sup> -K <sup>+</sup> exchanger
NEB	New England Biolabs
NFL	Nerve fibre layer
NGRL	National genomics research library
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NHS	National Health Service
NMD	Nonsense mediated decay
NUMT	Nuclear mitochondrial sequences
OA	Ocular Albinism
OCA	Oculocutaneous albinism
OCT	Optical coherence tomography
OCTA	Optical coherence tomography angiography
OGM	Optical genome mapping
OLM	Outer limiting membrane
OMIM	Online Mendelian inheritance in man
ONL	Outer nuclear layer
ONT	Oxford nanopore technologies
OPL	Outer plexiform layer
ORF	Open reading frame
p-value	Probability value
PacBio	Pacific biosciences

PAGE	Polyacrylamide gel electrophoresis
PAM	Protospacer adjacent motif
PANTHER	Protein analysis through evolutionary relationships
PBS	Phosphate buffer saline
PCR	Polymerase chain reaction
PDE	Phosphodiesterase
PE	Phosphatidylethanolamine
PEDF	Pigment epithelium-derived factor
pELS	Proximal enhancer like signatures
PLS	Promoter like signatures
PolyPhen-2	Polymorphism phenotyping v2
POS�	Photoreceptor outer segment layer
PTC	Premature stop codon
RCSB-PDB	Research collaboratory for structural bioinformatics - protein data bank
RDH5	Retinol dehydrogenase 5
RDH8	Retinol dehydrogenase 8
RFLP	Restriction fragment length polymorphism
RIN <sup>e</sup>	RNA integrity number equivalent
RMSD	Root mean square deviation
RNA	Ribonucleic acid
ROH	Regions of homozygosity
RP	Retinitis pigmentosa
RPE	Retinal pigmented epithelium
RT-PCR	Reverse transcriptase Polymerase chain reaction
SD-OCT	spectral domain optical coherence tomography
SDM	Site Directed Mutagenesis
sgRNA	Single guide RNA
SIFT	Sorting intolerant from tolerant
SINE	Short interspersed nuclear elements

smMIP	Single molecule molecular inversion probe
SMRT	Single molecule real time sequencing technology
SNAT	Sodium coupled neutral amino acid transporter
SNV	Single nucleotide variant
SOC	Super optimal catabolite
SPSS	Statistical Package for Social Sciences
ssODN	Single stranded oligodeoxynucleotide
SV	Structural variant
TADs	Topologically associated domains
TAE	Tris acetate ethylenediaminetetraacetic acid
TE	Tris-ethylenediaminetetraacetic acid
TID	Transillumination defect
T <sub>m</sub>	Melting temperature
TM-score	Template modelling score
UKIRDC	UK inherited retinal disease consortium
UMI	Unique molecular index
UPR	Unfolded protein response
USH3	Usher syndrome type 3
UTR	Untranslated region
VAF	Variant allele frequency
VCF	Variant calling format
Ensembl VEP	Ensembl-Variant effect predictor
VEP	Visual evoked potentials
VUS	Variant of uncertain significance
WES	Whole exome sequencing
WG	Weeks of gestation
WGA	Whole genome amplification
WGS	Whole genome sequencing
$\lambda_{max}$	Wavelength of maximum absorption

## Chapter 1 : Introduction

### 1.1 General overview

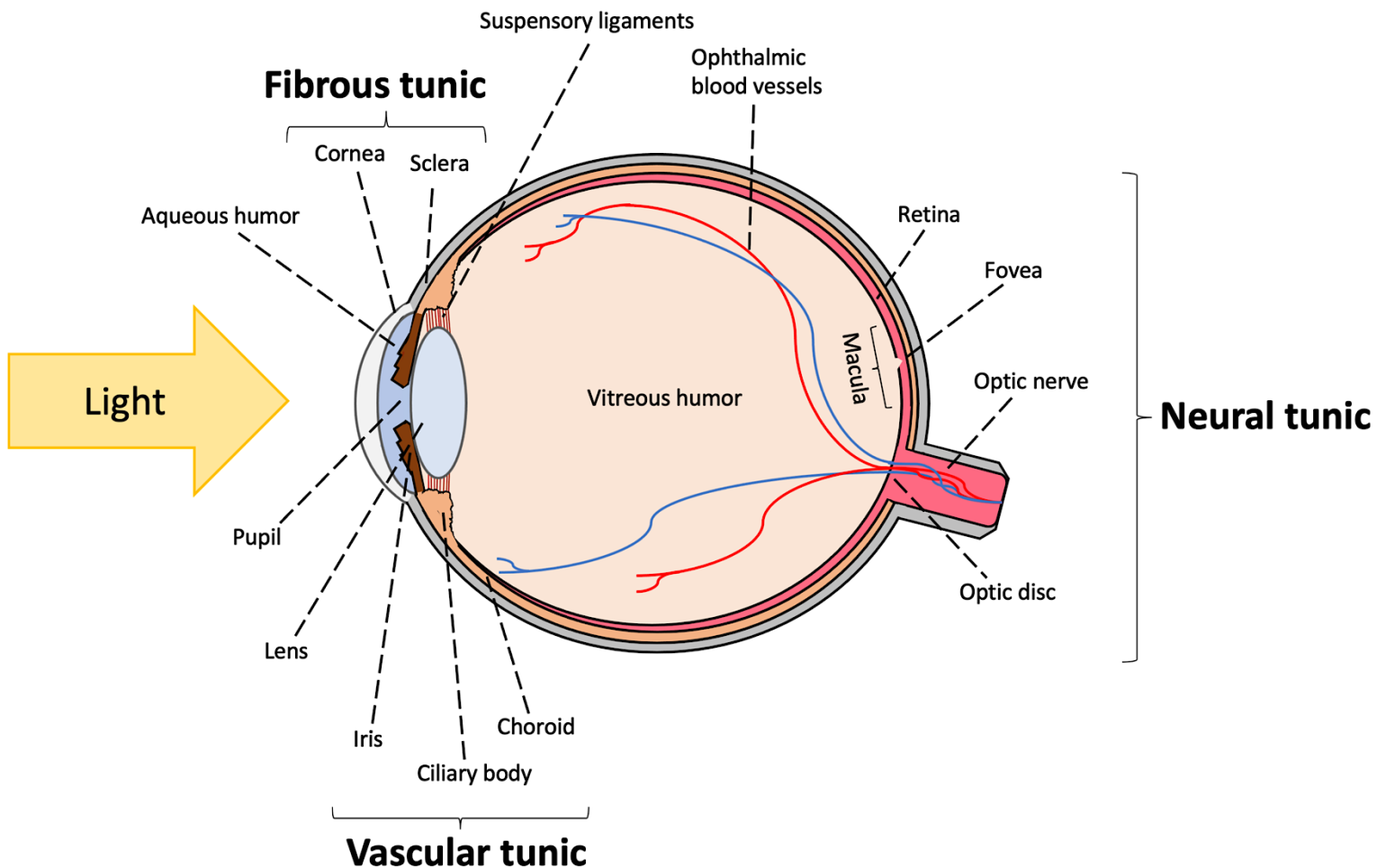
Inherited retinal disease (IRD) has an incidence between 1 in 2000-4000, making it the leading cause of vision loss within the working age demographic in the UK (Pontikos et al., 2020). The resulting annual economic burden of IRDs is estimated to be £523.3 million in the United Kingdom (Galvin et al., 2020). This group of Mendelian disorders pose a clinical challenge due to the profound genetic heterogeneity, as exemplified in foveal hypoplasia (FH), a developmental condition which can be inherited in an autosomal recessive, autosomal dominant or X-linked manner and is known to be caused by pathogenic variants in 52 genes to date (RetNet, <https://web.sph.uth.edu/RetNet/>, as read on 19.12.23). FH can occur as an isolated entity or as a component of syndromic disease, meaning FH can also be part of the phenotypic spectrum in other IRDs (Schiff et al., 2021). The diagnostic challenge imposed by IRDs such as FH requires genetic testing for better disease characterisation, especially when phenotypic overlaps exist (Britten-Jones et al., 2023). The progressive decrease in costs associated with genomic testing and the advances in molecular research have increased identification of the underlying genomic aetiology in IRDs (Ellingford et al., 2016).

This doctoral study focuses on the genetic characterisation of FH. One aim is to provide further insights into the mutation spectrum, genotype-phenotype correlation and pathogenic mechanisms in *SLC38A8*-related isolated FH. Another aim is to increase diagnostic yield through analysing next generation sequencing data in cohorts of unsolved FH obtained locally or derived from the 100,000 genomes project, and thus improve variant interpretation in FH. Concurrent with the above analyses, a third aim strives to characterise variants in *LAMP1* as a potential novel cause of isolated FH through the use of functional assays and clustered regularly interspaced palindromic repeats (CRISPR) and CRISPR-associated protein (Cas9) mediated cellular models.



## 1.2 Eye anatomy and physiology

The human eye is a highly specialised sensory organ forming the part of the central nervous system responsible for visual perception. It is composed of three layers, the fibrous tunic, vascular tunic and the neural tunic (Artal, 2016) (Figure 1.1). The fibrous tunic is the outermost layer consisting of the sclera and cornea. Tough and fibrous layers of collagen form the main structural component of both tissues, functioning to protect against mechanical trauma (Boote et al., 2020). The sclera is white and opaque due to the irregular organisation and density of mainly type I collagen fibrils (Sridhar, 2018, Fullwood et al., 2011). This opacity prevents off axial light from interfering with ocular optics and ensures light only enters the eye through the pupil. The sclera is also lined by a thin mucous membrane, known as bulbar conjunctiva, that has a protective role acting as a structural barrier and contributes to the tear film through mucin production (Prajna and Vijayalakshmi, 2017).



**Figure 1.1 Anatomy of the eye.** Sagittal section of an adult human eye annotated with different ocular layers and the associated structures.

Unlike the sclera, the cornea is a transparent membrane encasing the eye in front of the iris. The arrangement of type I collagen fibres in the cornea is uniform and in parallel, which confers tensile strength and transparency (Sridhar, 2018). The cornea's distinct curvature provides a convex surface to refract light rays at their point of contact (Sridhar, 2018). The refractive index of the cornea is estimated to be 1.335 -1.432, and this accounts for the majority of the eye's refractive power (Patel and Tutchenko, 2019).

The vascular tunic, also known as the uvea, is composed of the iris, ciliary body, and choroid (Yanoff and Sassani, 2020). The iris is a contractile circular ring with a middle aperture (pupil) located at the anterior of the vascular tunic (Davis-Silberman and Ashery-Padan, 2008). Iris pigmentation provides opacity to restrict light entry to the pupil, preventing stray light from impeding vision (Kruijt et al., 2011). The density of melanocytes and composition of melanin in the stroma determines iris colour. Iris sphincter and dilator pupillae act antagonistically to regulate the amount of light entering the eye by adjusting the pupil diameter (Marumo and Nakano, 2021). The sphincter pupillae are smooth muscles that encircle the pupil, and upon parasympathetic stimulation they constrict (miosis) whereas dilator pupillae are arranged radially and have sympathetic innervations to mediate pupil dilation (mydriasis) (Nakano et al., 2021).

Behind the iris is a structure known as the lens, which is a transparent biconvex disc. It lies directly in the path of light that enters the pupil (Yanoff and Sassani, 2020). The lens functions to refract light further, focusing it onto the retina for imaging. The refractive index of the human crystalline lens is approximately  $1.408 \pm 0.005$ , which allows for fine tuning of light rays (Uhlhorn et al., 2008). A basement membrane known as the lens capsule engulfs the lens to maintain its structural integrity and to regulate diffusion of metabolic substrates (Danysh et al., 2010). The lens is predominantly composed of a single cell type, fibre cells, which are arranged in a concentric array and are tightly packed at the centre. These highly specialised cells lack cytoplasmic organelles to prevent obstruction of light (Hejtmancik and Shiels, 2015).

The ciliary body is another component of the anterior segment and is also part of the vascular tunic. It is made up of the ciliary muscles and ciliary processes (Yanoff and Sassani, 2020). The ciliary muscles have a unique architecture composed of smooth muscle fibres oriented longitudinally, concentrically, and radially, that act antagonistically to regulate the refractive power of the lens to focus on objects at variable distances (Knaus et al., 2021). During accommodation, the ciliary smooth muscles constrict, causing the suspensory ligaments to loosen and consequently the lens thickens and becomes more convex to support near vision. The biomechanical role of ciliary muscles in altering lens morphology during accommodation allows for the maintenance of the focal point on the retina for a clear image (Martin et al., 2005).

The surface of the ciliary body forms folds with small projections known as ciliary processes that govern diffusion between blood and aqueous humour (Delamere, 2005). These protrusions consist of an epithelium bilayer. The inner non-pigmented layer develops from the neural retina and is thought to synthesize the macromolecules that make up the vitreous humour, a transparent gel that occupies a large cavity at the posterior of the eye known as the vitreous chamber (Bishop et al., 2002). The outer pigmented layer is continuous with the retinal pigmented epithelium and is responsible for secreting aqueous humour (Delamere, 2005). This is a plasma like fluid that maintains intraocular pressure and nourishes avascular ocular structures. This transparent fluid fills two cavities, one between the cornea and iris, known as the anterior chamber, and the other between the iris and the suspensory ligaments, termed the posterior chamber (Murthy et al., 2015).

The choroid is an intermediate layer of connective tissue and blood vessels enveloping the eye forming the posterior section of the vascular tunic (Nickla and Wallman, 2010). The choroid consists of five layers, the suprachoroid, lamina fusca, stroma, Haller's and Sattler's vasculature, and choriocapillaris (Yanoff and Sassani, 2020). The choriocapillaris is a sublayer consisting of a dense network of capillaries branching off the arterioles in Sattler's layer, which supply the outer retina with oxygen and nourishment (Nickla and Wallman, 2010). The Bruch's membrane lies between the choriocapillaris and the retinal pigmented epithelium

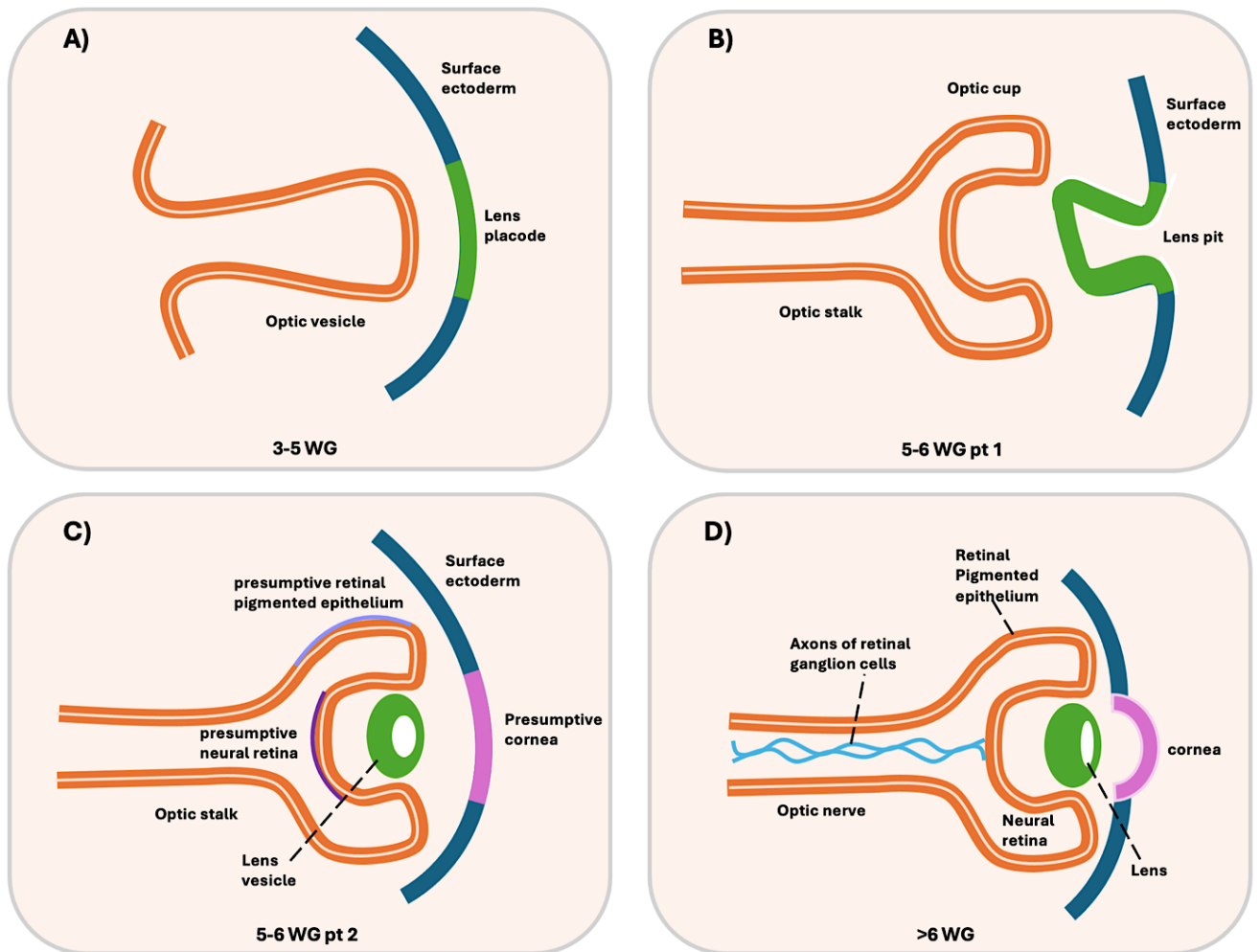
to regulate diffusion of biomolecules and oxygen, meeting the high metabolic demands of the retina (Booij et al., 2010). Numerous melanocytes are present within the choroidal stroma and suprachoroid. This pigmentation allows for the absorption of stray light to limit internal reflections which would otherwise degrade the retinal image (Nickla and Wallman, 2010).

The neural tunic is the innermost layer, consisting of the retina and optic nerve. The retina is a component of the central nervous system. It is the sensory tissue housing photoreceptor cells responsible for phototransduction, the process by which the focused photons reaching the retina are converted to electrical signals. The retina is continuous with the optic nerve, which transmits the electrical signals along the visual pathway to the visual cortex in the brain, resulting in visual perception. The point where the optic nerve leaves the eye is termed the optic disc and is a blind spot in the retina due to the absence of photoreceptors in this region.

### **1.3 Ocular embryogenesis**

Ocular development begins after the embryo (gastrula) forms three germinal layers comprising the endoderm at the innermost, the mesoderm as the intermediate layer and the ectoderm at the outermost surface (Miesfeld and Brown, 2019). All three layers will eventually differentiate into the different cells that constitute the human body. The ectoderm gives rise to the nervous system, epidermis and migratory multipotent progenitors known as the neural crest cells (Kiecker et al., 2016).

Early embryonic growth of the human eye is initiated and orchestrated by transcription factors, most notably *PAX6* for temporal and spatial gene expression (Ashery-Padan and Gruss, 2001). Subsequently, a series of complex morphological changes give rise to unique structures that compartmentalise derivatives of the ectoderm, the neuroectoderm and surface ectoderm, physically separating the development of the anterior and posterior eye structures (Figure 1.2) (Miesfeld and Brown, 2019).



**Figure 1.2 Embryonic developmental timeline of the vertebrate eye.** A) The optic vesicle protrudes, coming in close proximity to the surface ectoderm which induces lens placode formation. B) Simultaneous invagination of the optic vesicle and lens placode results in the formation of the optic cup and lens pit, respectively. C) The lens pit pinches off from the surface ectoderm forming a lens vesicle. The surface ectoderm overlying the lens vesicle is rebuilt and serves as the future cornea. D) The outer layer of the optic cup develops into the retinal pigmented epithelium while the inner layer forms the neural retina. The surface ectoderm along with periocular mesenchymal deposits forms the cornea. Retinal ganglion cells in the neural retina have their axons elongated, passing through the optic stalk. At this stage the dense optic stalk becomes the optic nerve.

At three weeks of gestation (WG), cells of the neuroectoderm in the diencephalon evaginate bilaterally, forming structures known as the optic sulcus or groove. These protrude towards the surface ectoderm, passing through the mesenchyme, which is derived from neural crest cells and mesoderm (Forrester et al., 2016). Each elongated optic sulcus expands to develop an optic vesicle on the distal end and becomes narrower towards the lumen of the forebrain to form the optic stalk, which later develops into the optic nerve (Graw, 2010).

At five WG, each optic vesicle contacts the surface ectoderm, inducing a localised cellular expansion restricted to the region overlying the optic vesicle by adherence to the extracellular matrix. This thickened section is known as the lens placode (Huang et al., 2011). The lens placode gets depressed, forming a lens pit that progressively deepens and forms an enclosed cavity termed the lens vesicle that isolates lens development (Miesfeld and Brown, 2019). Further lens development at six WG, consists of elongation of the primary and secondary lens fibres and concentric arrangement in the lumen, followed by loss of nuclei and mitochondria through differentiation at a later stage (Graw, 2010).

The lens vesicle eventually detaches from the surface ectoderm and the renewed surface ectoderm seal (presumptive cornea) above the lens vesicle becomes the corneal epithelium (Miesfeld and Brown, 2019). During the fifth WG, waves of infiltrating periorbital mesenchymal cells originating from the neural crest of the dorsal neural tube fill the empty space between the lens vesicle and the corneal epithelium, giving rise to the corneal endothelium (Graw, 2010). Collagen deposits arranged in lamellae along with extracellular matrix proteins and periorbital mesenchymal cells form the corneal stroma (Cvekl and Tamm, 2004). Later, mesenchymal cells differentiate into quiescent keratocytes which function in wound repair and produce water soluble proteins known as crystallins. These underlie corneal transparency by homogenising the refractive index in the stroma to promote light to pass through without scattering (Jester et al., 1999).

Concurrent with the lens placode self-folding, there is an inward folding of the optic vesicle along with the neuroectoderm to form a bilayered structure known

as the optic cup (Casey et al., 2021). The optic cup experiences asymmetrical growth around the circumference which results in a groove known as the choroid fissure, sometimes also known as the optic fissure. The hyaloid artery passes through the fissure and extends into the optic stalk, reaching the posterior pole of the lens, where it acts as a temporary embryonic vascular system to support ocular development (Graw, 2010). The optic fissure fuses during the sixth WG, sealing the hyaloid artery at the centre of optic stalk (Forrester et al., 2016). Periocular mesenchyme encases the optic cup and condenses to form the vascular choroid on its inner side and the sclera at its outermost surface (Forrester et al., 2016).

The outer layer of the optic cup give rise to the retinal pigmented epithelium while the inner layer forms the neural retina. These layers are separated by an intraretinal space (Forrester et al., 2016). Multipotent retinal progenitor cells first begin to differentiate into retinal ganglion cells, followed by amacrine cells, cone photoreceptors and horizontal cells, and at a later stage into bipolar cells and rod photoreceptors (Miesfeld and Brown, 2019). During the seventh WG, the axons of the retinal ganglion cells converge through the optic stalk, reaching the diencephalon and forming a connection between the brain and the prospective eye referred to as the optic nerve (Graw, 2010). By 30 WG, myelination of the optic nerve occurs, ensuring a more rapid signal transmission along the axons of the retinal ganglion cells. The margin of the optic cup at this stage is known as the optic cup lip.

During the 12<sup>th</sup> and 13<sup>th</sup> WG the neuroepithelial cells at the optic cup lip proliferate, extending it (Graw, 2010). The vascular supply introduced into the growing optic cup neuroectoderm splits into the blood vessels of the iridopupillary membrane and capsulopupillaris to provide nourishment to the developing iris (Forrester et al., 2016). The iris smooth muscles differentiate from the neuroectoderm cells, with the sphincter pupillae developing first at 13-14 WG followed by the dilator pupillae at 25 WG (Forrester et al., 2016).

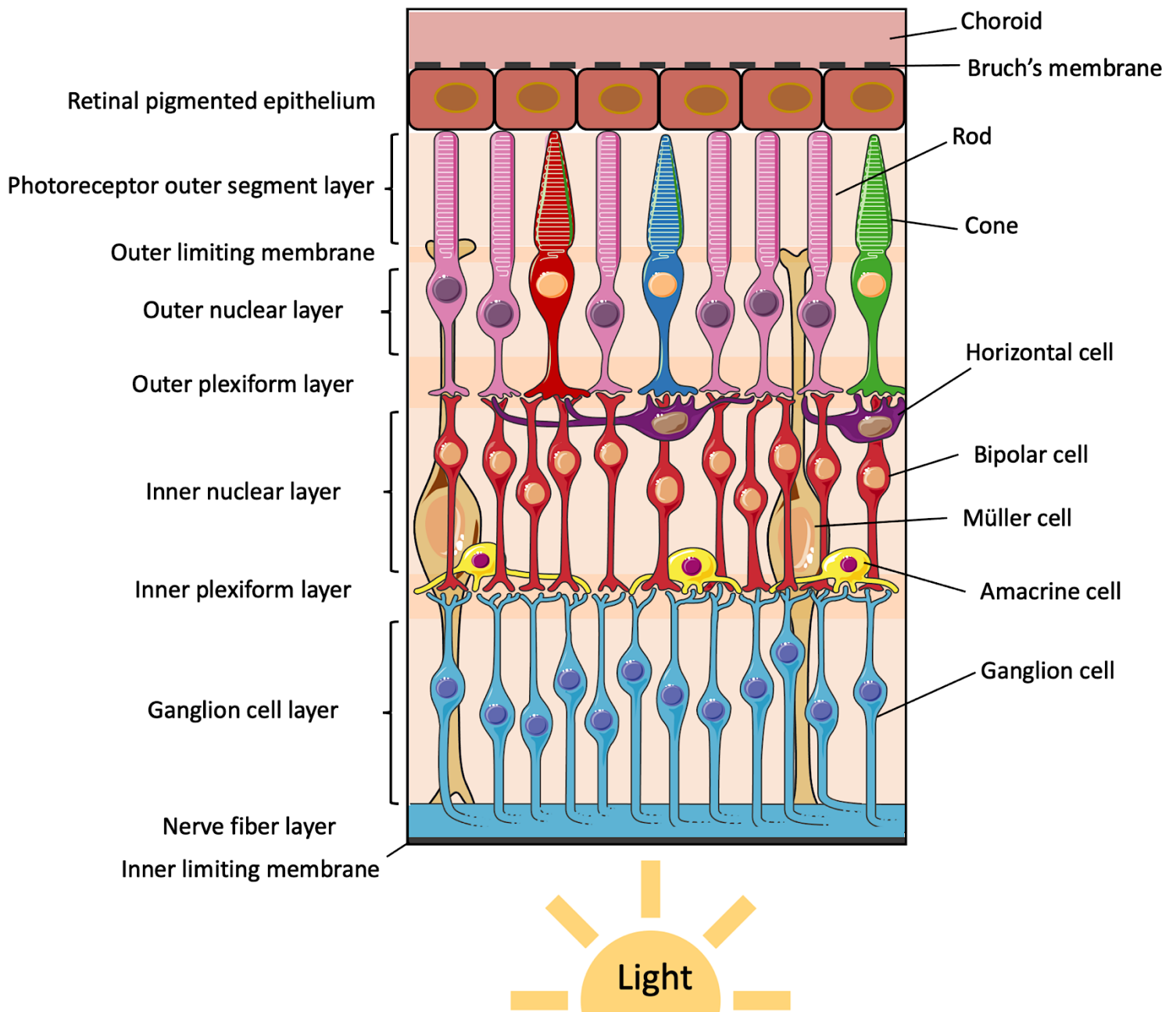
## 1.4 Retinal architecture

The vertebrate retina is inverted with respect to incoming light, meaning that light rays need to pass through different retinal layers to reach the photoreceptor cells (Kroger and Biehlmaier, 2009, Bringmann et al., 2018). This is generally considered optically disadvantageous. However, this unique retinal architecture provides efficient space management to accommodate complex neural networks (Kroger and Biehlmaier, 2009).

The human retina is a highly laminated structure composed of ten layers visible on microscopy, which can be grouped into the layers of the outer and inner retina (Figure 1.3). The outer retina includes the photoreceptors and the retinal pigmented epithelium (RPE). The RPE is a monolayer of pigmented epithelial cells that forms the blood-retinal barrier, regulating diffusion of nutrients and waste products across the choroid via tight junctions (Kocaoglu et al., 2016). The RPE is also involved in recycling the vitamin A derivatives of the visual cycle which are responsible for light capture by photoreceptors (see section 1.6) (Sahu and Maeda, 2018). Amongst its other functions is to promote photoreceptor's health through phagocytosing parts of the outer segments in a controlled manner (Kocaoglu et al., 2016).

Cone and rod outer segments delineate the photoreceptor outer segment layer (POSL), while their inner segments form a tight junction with Müller cells at the outer limiting membrane (OLM) and their cell bodies mark the outer nuclear layer (ONL) (Kobat and Turgut, 2020, Zhao et al., 2021, Bringmann and Wiedemann, 2022). The outer plexiform layer (OPL) contains synaptic terminals of photoreceptors and dendrites of interneurons (bipolar and horizontal cells) (Bringmann and Wiedemann, 2022).





**Figure 1.3. Retinal lamination in humans.** Illustration of the neural architecture of the retina. The retinal layers (left) and the various cellular components including additional anatomical structures (right) are labelled. Light focused on the retina passes through the inner limiting membrane and the structures of the inner retina to reach the photoreceptor layer where it is captured. Cone variants are depicted in different colours corresponding to the different wavelength detected. Diagram generated with modified resources from Servier Medical Art (<https://smart.servier.com>). Servier Medical Art is licensed under CC BY 4.0.

Cones are concentrated at the centre of the retina (the fovea) and are responsible for high acuity vision under brightly illuminated conditions (photopic vision) (Lamb, 2016). There are three subtypes of cones with different absorption spectra due to evolutionary variations in photopigment proteins known as opsins. These cone variants are S-cones (short wave-length) which are sensitive to blue light, M-cones (medium wavelength) for green light and L-cones (long wave-length) for red light (Kawamura et al., 2012). The combined sensory input from the different cones enables the recognition of a range of wavelengths corresponding to the different colours.

In contrast, rods have peak density at the macula, are absent from the fovea, and are present but decline in density towards the retinal periphery. Rods are highly sensitive photoreceptors that generate achromatic images under dim light to enable scotopic vision (Lamb, 2016). The high sensitivity of rods is attributed to their neural wiring whereby multiple rods (~1000) connect to a single ganglion cell, reducing the threshold for activation down to a single photon (Bringmann and Wiedemann, 2022).

The inner retina constitutes complex neuronal circuits involved in relaying and modulating visual information from the photoreceptors (Masland, 2012). It starts at the inner nuclear layer (INL) which contains the nuclei of various interneurons (bipolar, horizontal and amacrine cells) and neuron supporting cells (Müller cells) (Hoon et al., 2014, Masland, 2012). These intermediate neurons relay sensory information from the hyperpolarised photoreceptors to the ganglion cells at the inner plexiform layer (IPL) (Hoon et al., 2014). Horizontal and amacrine cells within the inner nuclear layer (INL) respectively regulate the synaptic transmissions of photoreceptors and ganglion cells via  $\gamma$ -aminobutyric acid (GABA) mediated inhibitory feedback to enhance contrast and image contour (Kaneko and Tachibana, 1987, Bringmann and Wiedemann, 2022). Retinal ganglion cells in the ganglion cell layer (GCL) receive inputs from bipolar and amacrine cells and transmit the encoded visual signal along their axons (Bringmann and Wiedemann, 2022). These axonal projections are confined to the nerve fibre layer (NFL) and they converge into the optic nerve to relay the signal to the brain for higher processing (Hoon et al., 2014). At the interface

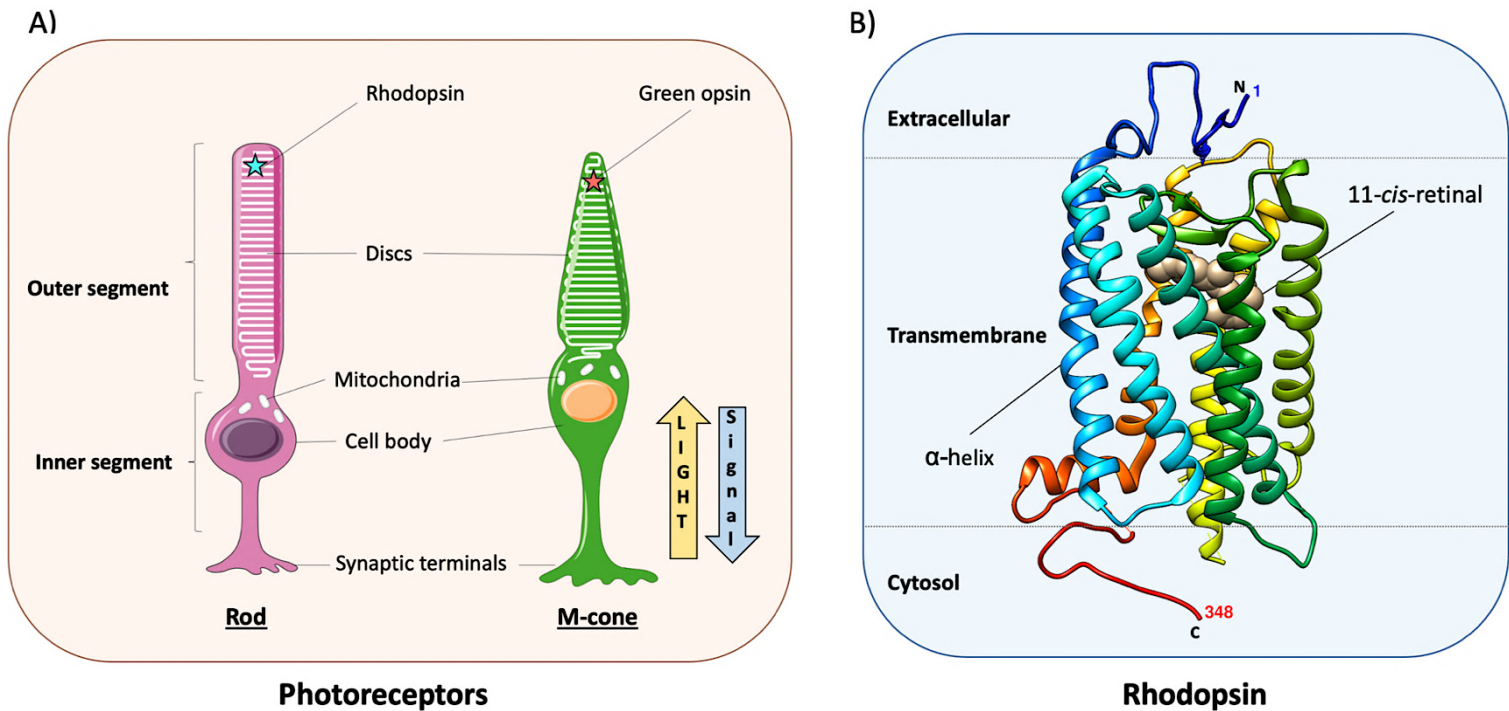
between the retina and the vitreous humour is a barrier made up of Müller cell terminal footplates and astrocytes. This is known as the inner limiting membrane (ILM) (Fine and Zimmerman, 1962).

## 1.5 Photopigments

Conversion of light stimulus into electrical potentials is dependent on the functional properties of photopigments. These are protein complexes constituting a G-protein coupled receptor (the opsin protein) covalently linked by a Schiff base to a light reactive molecule (the chromophore 11-*cis*-Retinal, a vitamin A analogue) (Figure 1.4B) (Foster, 2017). Photopigments are present at high levels in the outer segments of photoreceptors. The opsins are transmembrane proteins and they are present in the vertical stack of membranous discs found in the photoreceptor outer segments (Figure 1.4A). Both the length of the outer segment and photopigment density dictate the light absorbing capacity of a photoreceptor (Jacobs, 2021). The outer segment discs are continuously shed at the distal end, adjacent to the RPE, and renewed at the proximal end, at a rate required to maintain functional integrity while eliminating accumulated photo-oxidative damage. This diurnal shedding process is facilitated by the RPE, which phagocytoses discs at the apex of the outer segment at varying rates which are regulated by the circadian rhythm (Kocaoglu et al., 2016).

Opsins are embedded in the cell membrane and have a structure consisting of seven transmembrane  $\alpha$ -helical domains (Figure 4B) (Hargrave et al., 1983). Different opsins are unique to each photoreceptor, with rhodopsin in rods, green, blue and red opsins in cone variants, and melanopsin in the intrinsically photosensitive retinal ganglion cells (ipRGCs) (Fu and Yau, 2007, Provencio et al., 2000). Opsins modulate the absorption spectra of the chromophore ligand (11-*cis*-retinal) for vision under varying conditions of illumination (Terakita, 2005). The wavelength of maximum absorption ( $\lambda_{max}$ ) for rhodopsin is 500 nm, and for the cone photopigments are 420 nm for blue opsin, 530 nm for green opsin and 560 nm for red opsin (Merbs and Nathans, 1992, Kawamura et al., 2012). The rhodopsin is encoded by the *RHO* gene on 3q22.1 and melanopsin by *OPN4* on 10q23.2. The genes encoding opsins of S-cone, M-cone and L-cone are

*OPN1SW* (7q32.1), *OPN1MW* (Xq28) and *OPN1LW* (Xq28) respectively (Terakita, 2005). Genomic insults to these genes are associated with photoreceptor degeneration as seen in IRDs such as retinitis pigmentosa (OMIM: 180380), or with colour blindness (OMIM: 613522, 300821 and 300822) in the case of cone opsin deficiency.

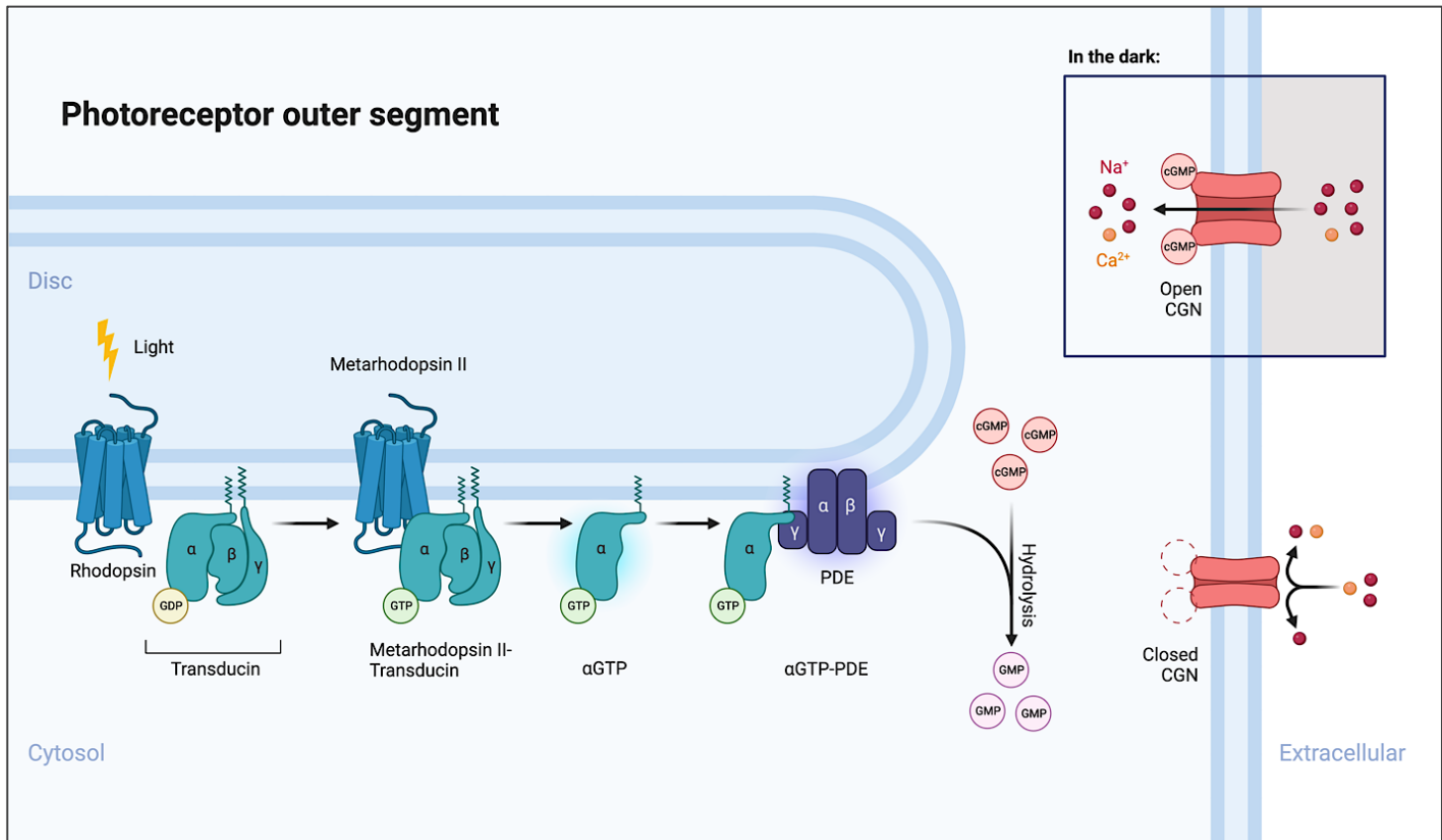


**Figure 1.4 Photoreceptor schematic and rhodopsin structure.** A) The dominant features of both cone and rod gross anatomy is the outer segment, which is composed of membranous discs that are densely packed with photopigments containing opsins (star). Arrows depict the direction of light towards the outer segment and the electric potentials travelling in the opposite direction approaching the synaptic terminals. The Photoreceptor structure image was designed in part using Servier Medical Art (<https://smart.servier.com>). Servier Medical Art is licensed under CC BY 4.0. B) Protein model of bovine rhodopsin in the inactive state (dark-adapted), displaying the seven helical transmembrane domains and the chromophore ligand (11-*cis*-retinal) attached to residue 296. Protein model retrieved from the RCSB Protein Data Bank (ID:1HZX) and visualised on UCSF Chimera.

## 1.6 The phototransduction cascade

Photons reaching the photoreceptor outer segment induce a conformational change of the chromophore 11-*cis*-retinal, which initiates a biochemical signalling cascade that amplifies signal strength, resulting in the generation of electric potentials (Figure 1.5) (Fu and Yau, 2007). The following describes the phototransduction cascade in rods. The process is essentially the same in cones but is mediated by a different opsin, which activates transducins and phosphodiesterases composed of cone-specific subunits, resulting in differences in the activating wavelength of light and chromophore regeneration (Srinivasan et al., 2018, Srinivasan et al., 2014).

In rods, the activation of rhodopsin is contingent on the photoisomerization of 11-*cis*-retinal to all-*trans*-retinal, which in turn initiates a series of thermal reactions resulting in metarhodopsin II (Okano et al., 1992). This activated form of rhodopsin interacts with a heterotrimeric G protein known as transducin, causing guanosine diphosphate (GDP) to be displaced by guanosine triphosphate (GTP) (Teller et al., 2001). The  $\alpha$  subunit of transducin binds to GTP ( $\alpha$ GTP) and dissociates from the  $\beta$  and  $\gamma$  subunits. It is now free to stimulate phosphodiesterase (PDE). The  $\alpha$ GTP has an affinity for PDE which, when bound, eliminates the inhibition imposed by PDE  $\gamma$  subunits. The liberated PDE has its catalytic activity amplified, increasing the hydrolysis of cyclic guanosine monophosphate (cGMP) (Fu and Yau, 2007). The reduction of cytosolic cGMP causes the cyclic nucleotide gated (CNG) channels to close, thereby restricting  $\text{Na}^+$  and  $\text{Ca}^{2+}$  entry, while  $\text{Ca}^{2+}$  extrusion continues through the  $\text{Na}^+/\text{Ca}^{2+}\text{-K}^+$  exchanger (NCKX). This ultimately results in photoreceptor hyperpolarisation. (Lamb and Pugh, 2006, Barret et al., 2022). The resultant reduction in membrane potential inhibits glutamate secretion at the post-synaptic terminals (Barret et al., 2022, Agosto et al., 2021). Glutamate is a neurotransmitter with different effects on bipolar cells. Its reduction alleviates the inhibition of ON-centre bipolar cells so that they can depolarise (Agosto et al., 2021). The electric signal generated encodes visual information that is modulated and processed by the neural retina prior to reaching the visual cortex where it is further processed (Kaneko and Tachibana, 1987, Bringmann and Wiedemann, 2022).

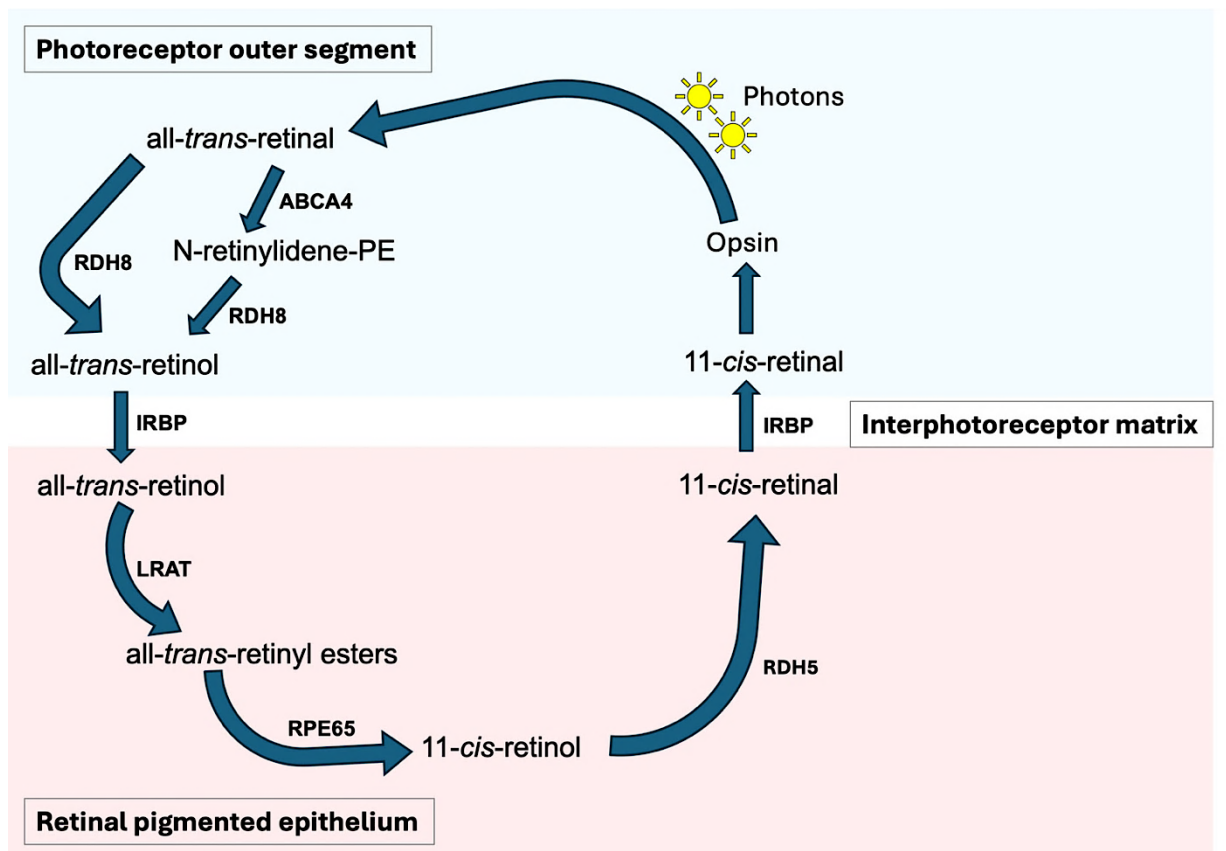


**Figure 1.5 Phototransduction in rods.** Light activation of rhodopsin leads to an active configuration known as metarhodopsin II. Metarhodopsin II binds to transducin causing GDP to be replaced by GTP. The alpha subunit of transducin dissociates from the complex and is free to interact with PDE, eliminating the inhibition imposed by the gamma subunit of the enzyme. Increased PDE hydrolysis of cGMP causes the closure of CNG channels, thereby preventing entry of  $\text{Na}^+$  and  $\text{Ca}^{2+}$  from the extracellular space into the cytosol. The membrane potential gets reduced until reaching the threshold for hyperpolarisation. Diagram created with BioRender.com.

### 1.7 The visual cycle (Retinoid cycle)

The retinoid cycle is a process of sequential enzymatic reactions that maintains photoreceptor activity through recycling a series of vitamin A derivatives and by-products of the phototransduction cascade to ensure a constant supply of the light reactive molecule, 11-*cis*-retinal (Figure 1.6) (Wright et al., 2015). Upon opsin activation, the product of photoisomerization, the all-*trans*-retinal,

dissociates from opsin and either enters the cytoplasm directly or binds to phosphatidylethanolamine (PE) at the disc membrane, forming a complex known as N-retinylidene-PE, that gets shuttled into the cytoplasm via ATP-binding cassette sub-family A, member 4 (ABCA4) (Quazi et al., 2012). In the cytoplasm, the all-*trans*-retinal is reduced into all-*trans*-retinol (vitamin A) by retinol dehydrogenase 8 (RDH8) (Maeda et al., 2005). The all-*trans*-retinol gets transported out of the photoreceptor outer segment into the RPE by interphotoreceptor retinoid binding protein (IRBP) where it gets esterified by lecithin retinol acyltransferase (LRAT) into all-*trans*-retinyl esters (Jin et al., 2009). These esters undergo concurrent hydrolysis and isomerization into 11-*cis*-retinol by retinal pigment epithelium 65 (RPE65), followed by oxidation to 11-*cis*-retinal is by retinol dehydrogenase 5 (RDH5) (Jang et al., 2001, Moiseyev et al., 2005). Finally, the newly regenerated 11-*cis*-retinal is delivered back to the photoreceptor outer segments by IRBP to replenish the visual chromophore (Jin et al., 2009).



**Figure 1.6 The retinoid cycle.** Key enzymatic reactions taking place converting all-*trans*-retinal into 11-*cis*-retinal, regenerating the chromophore to maintain phototransduction by opsins. IRBP shuttles molecules between the photoreceptor and RPE through the interphotoreceptor matrix.

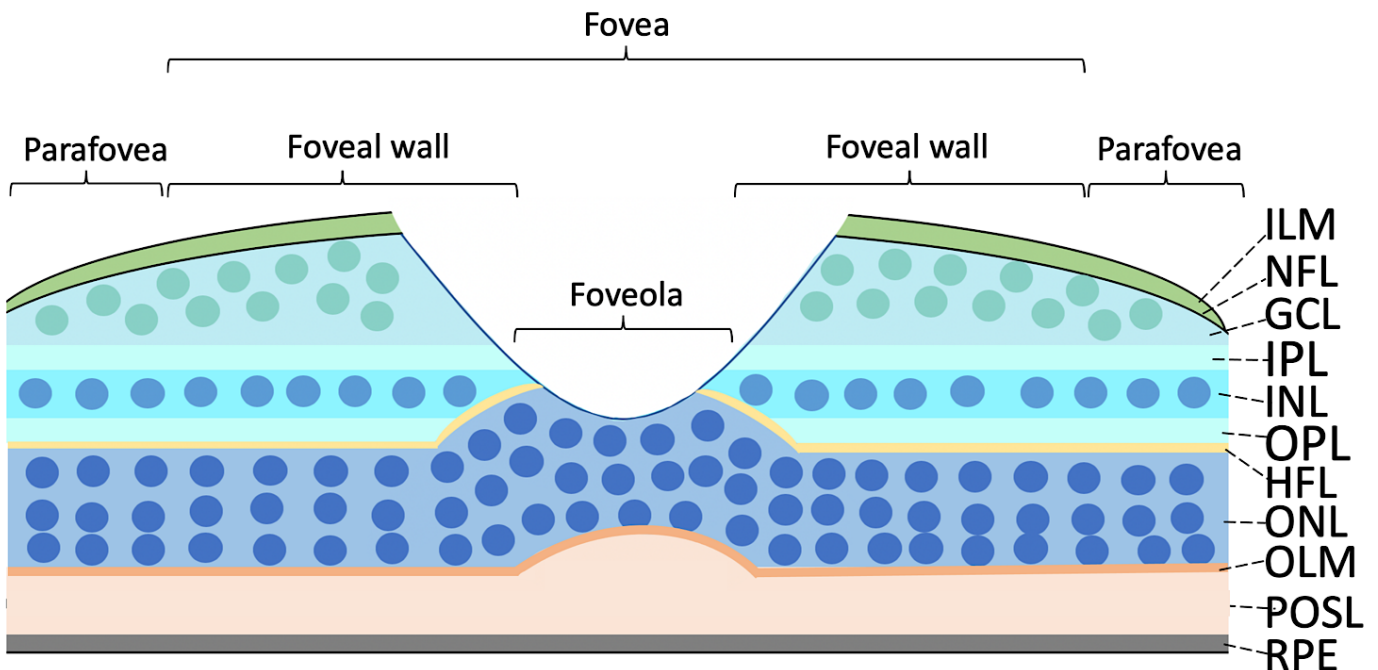
## 1.8 The fovea

At the centre of the retina lies the macula lutea (macula), which is a region responsible for high acuity vision. Within the macula is a depression that is temporal to the optic axis, known as the fovea centralis (fovea) (Provis et al., 2013). The concaviclivate (bowl shaped) fovea is a highly specialised structure that emerged in diurnal primates to compensate for the optical disadvantage of an inverted retina (Bringmann et al., 2018, Slonaker, 1897). This structure is densely packed with only cone photoreceptors (199,000-324,000 cones/mm<sup>2</sup>) (Curcio et al., 1990) and has no overlying retinal layers in the path of the light, to ensure maximum exposure and minimal light scattering (Figure 1.7). The cellular density at the fovea ensures that each cone photoreceptor is connected to an individual retinal macroglial cell (Müller cell) (Reichenbach and Bringmann, 2010). In the retina, these elongated and radial Müller cells strategically span nearly the entire retina from the outer limiting membrane to the inner limiting membrane (Reichenbach and Bringmann, 2010). Müller cells have unique optical properties, allowing them to guide light through the neural retina to reach the photoreceptors with negligible scatter or loss of intensity (Franze et al., 2007). However at the fovea, Müller cell morphology and organisation hinders their wave-guiding properties, so the fovea relies on cones to guide light from their inner segments to their outer segments (Provis et al., 2013).

The fovea is not just devoid of rods, it also lacks capillaries, creating the foveal avascular zone (FAZ). The optical advantage of having no vasculature at the fovea is to prevent visual distortions resulting from the retinal blood vessels being superimposed on images formed by the lens (Provis et al., 2013). The FAZ is defined prior to foveal pit formation by anti-angiogenic factors like pigment epithelium-derived factor (PEDF), which is expressed in retinal ganglion cells in the region of the developing fovea (Kozulin et al., 2010). Axonal guidance factors encoded by *NTNG1* (Netrin G1) and *EPHA6* (EphA6) also contribute to FAZ formation by moving axons away from the incipient foveal region at different stages during development (Provis et al., 2013). The diameter of the FAZ is influenced by ethnicity, gender, retinal thickness and premature birth (Pang et al., 2023, O'Shea et al., 2022, Mintz-Hittner et al., 1999). The development of the fovea is dependent on the presence of the FAZ, as it defines the region for the



incipient fovea by making it more susceptible to orthogonal force generated by intraocular pressure (Springer and Hendrickson, 2004).

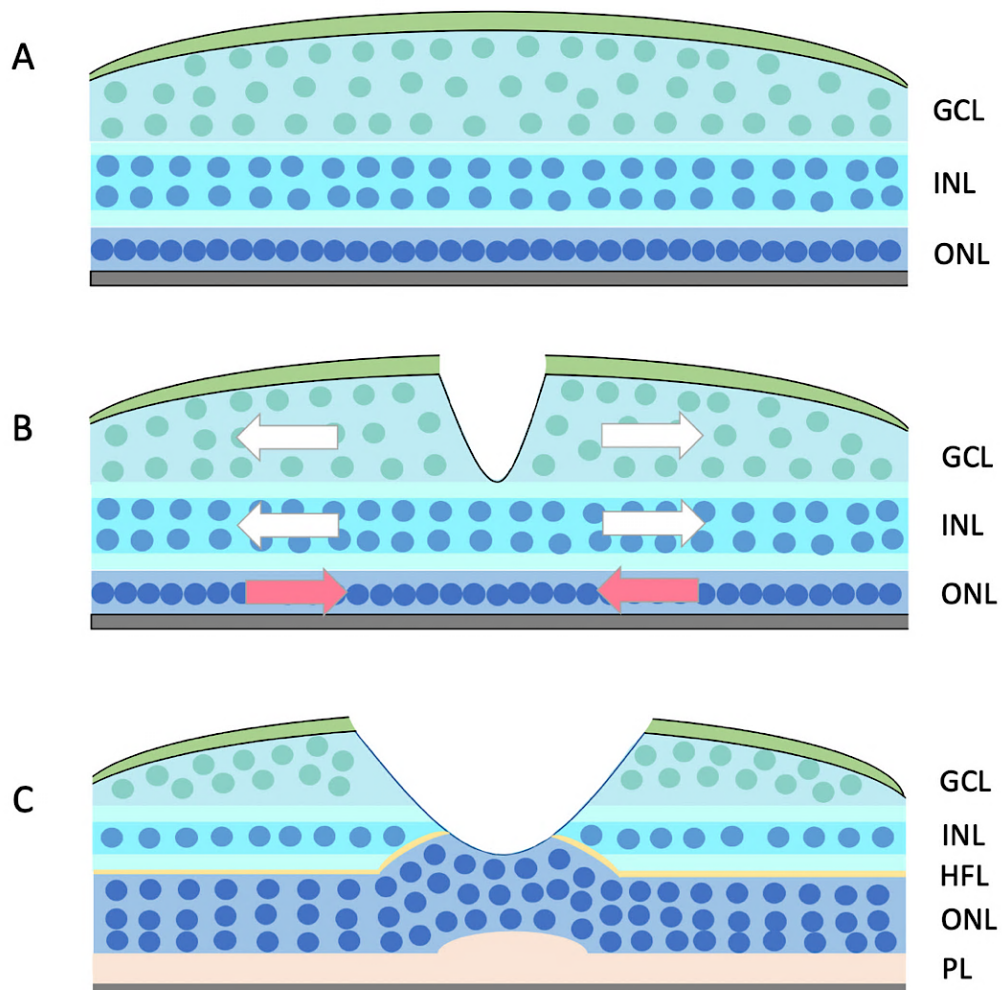


**Figure 1.7 The fovea in primates.** Graphical representation of an intact foveal pit in a healthy eye. Nuclei of neuronal cells are visible in their respective retinal layers. Note the variation in thickness of different layers. The fovea encompasses the foveola and foveal wall. The parafovea is the macular region surrounding the fovea. ILM: inner limiting membrane; NFL: nerve fibre layer; GCL: ganglion cell layer; IPL: inner plexiform layer; OPL: outer plexiform layer; HFL: Henle's fibre layer; ONL: outer nuclear layer; OLM: outer limiting membrane; POSL; photoreceptor outer segment layer; RPE: retinal pigmented epithelium.

### 1.9 Foveal pit development

The prenatal development of the fovea is initiated at 23 WG, while its maturation lasts up to 45 weeks post birth (Hendrickson and Yuodelis, 1984). The development of the fovea centralis is dependent on the bidirectional migration of neurons in the INL, ONL and GCL (Figure 1.8). The cells in the GCL and INL are displaced away from the centre, resulting in the formation of a depression (Kondo, 2018, Hendrickson and Yuodelis, 1984). Cones undergo morphological changes whereby they decrease in width and their outer segment lengthens. This is known as cone specialisation (Hendrickson and Yuodelis, 1984). The cone

photoreceptor cells in the ONL are displaced inwards, causing them to be densely packed for maximum spatial acuity (Thomas et al., 2011, Hendrickson and Yuodelis, 1984). This is followed by the axonal elongation of both cones and rods, that bend to form a cylindrical densely packed structure referred to as Henle's fibre layer (HFL) where the synaptic terminals occur at the outer plexiform layer (Lujan et al., 2011).

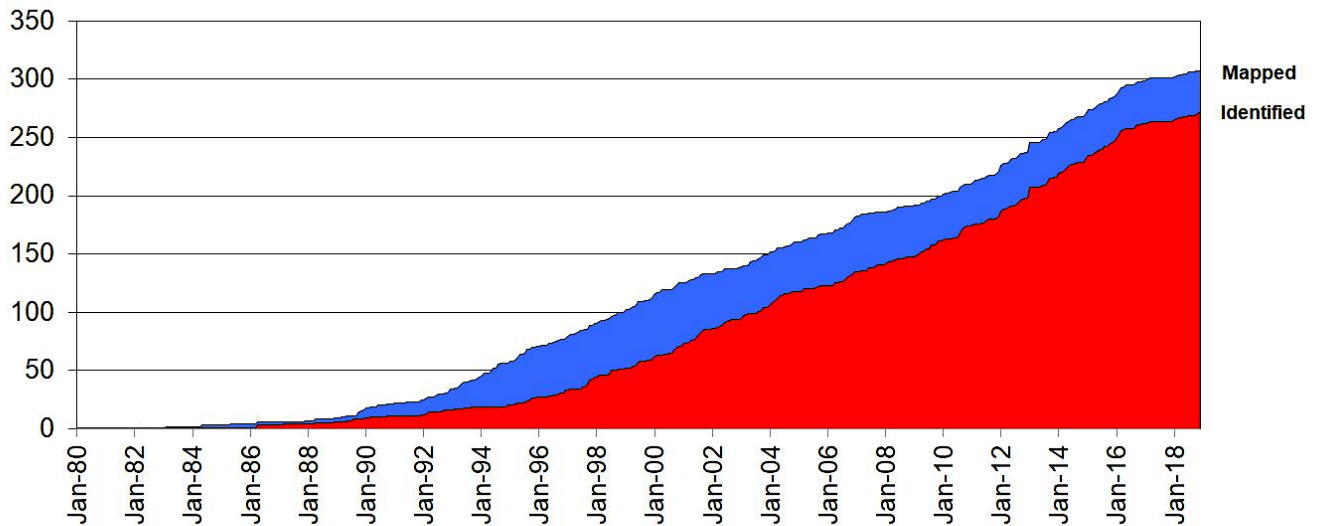


**Figure 1.8 Stages of foveal pit development.** A) The GCL is intact until late in the second trimester. B) At 23 WG, the cells in the INL and GCL are displaced away from the centre (centrifugally) forming a hollowed-out pit, while the cones in the ONL get displaced towards the centre (centripetally). C) The final stage of foveal development occurs postnatally. The foveal pit volume increases and the cones in the ONL become tightly packed and protrude, becoming more prominent. Note that the ONL widens while the INL and GCL get compressed. The outer segments of cones in the POSL also lengthen, forming an obstruction that elevates foveolar cones. The formation of the HFL marks the end of foveal pit development.

### **1.10 Inherited retinal disease (IRDs)**

The advent of the ophthalmoscope during the 19<sup>th</sup> century enabled the examination of the fundus and consequently the characterisation of hereditary retinal disease (Francis, 2006). IRDs is a broad term describing a heterogeneous group of visually debilitating diseases affecting the retina. IRDs are caused by genetic changes that affect genes essential for retinal function or structure (Duncan et al., 2018). To date, there are 281 genes implicated, and these predominantly exhibit autosomal recessive (195), autosomal dominant (66), X-linked (13) or mitochondrial inheritance (7) (RetNet, <https://web.sph.uth.edu/RetNet/>, as read on 19.12.23). The financial burden of IRDs is estimated to cost the United Kingdom £523.3 million every year (Galvin et al., 2020). However, this value does not account for less common IRDs that can collectively increase the health cost significantly due to prolonged diagnostic journey (Lam et al., 2021).

Diagnosis of IRD has proven difficult due to the profound genetic heterogeneity. For instance, retinitis pigmentosa can be caused by variants in >100 genes, while variants in some genes can cause multiple IRDs, making it challenging to identify the genetic aetiology. There is often an overlap in the clinical manifestations of different IRDs, which hinders differential diagnosis based on phenotyping (Lam et al., 2021). The clinical symptoms of an IRD can be further complicated by variable penetrance or by syndromic manifestations (Ellingford et al., 2016). Despite these challenges, advances in molecular research have aided the diagnosis of IRDs through the increased characterisation of the underlying genes (Figure 1.9).



**Figure 1.9 Graph of mapped and identified IRD genes.** 317 disease loci have been progressively mapped and 281 causative genes identified over four decades, as documented by the RetNet website (<https://sph.uth.edu/RetNet/>, viewed 19.12.23).

### 1.11 Retinitis pigmentosa and allied disease

The classification of IRDs is largely dependent on the retinal region affected, the pattern of photoreceptor degeneration or disease progression. The major IRD subgroups can be defined as macular dystrophy, chorioretinal dystrophy, cone dystrophy, cone-rod dystrophy, rod-cone dystrophy, cone dysfunction syndromes, rod dysfunction syndromes and Leber congenital amaurosis / early onset severe retinal dystrophy (Georgiou et al., 2021). However, the limitation of such a classification system is that many IRD don't fit into these narrow categories. Furthermore, some IRDs are actually syndromic ciliopathies with systemic defects.

The most prevalent IRD phenotypes in the UK include retinitis pigmentosa (RP) and Usher syndrome respectively (Galvin et al., 2020). RP falls under the rod-cone dystrophy subgroup, which denotes a progressive degeneration of rods followed by cones. RP manifests as a reduction in peripheral vision (tunnel vision) and nyctalopia (night blindness), with subsequent impairment of central vision leading to complete blindness (Cremers et al., 2018). The fundus of RP patients

contains pigment deposits (bone spicules), the optic discs are pale (optic disc pallor) and retinal blood vessels are attenuated (thin) (Hamel, 2006). The frequency of RP in the general population is around 1:3000, making it the most common IRD (Francis, 2006). Non-syndromic RP can be caused by variants in 95 different genes and is inherited in autosomal recessive (61), autosomal dominant (24), both autosomal recessive and dominant (7) and X-linked inheritance patterns (3) (Table 1.1). There have also been reports of non Mendelian inheritance in RP owing to digenic inheritance in *ROM1* and *RDS* (Kajiwara et al., 1994), mitochondrial heteroplasmy in *MT-ATP6* (Duno et al., 2013) and uniparental isodisomy in *RPE65* and *MERTK* (Thompson et al., 2002).

Syndromic forms of RP can present as Usher syndrome, with an estimated frequency of 1 in 6000 (Kimberling et al., 2010). The symptoms of Usher syndrome are RP, hearing loss and balance complications (vestibular dysfunction) (Ebermann et al., 2009). The visual and auditory defects in Usher syndrome are caused by biallelic mutations in 16 genes (Table 1.1). This syndrome is classified into 3 categories based on severity, with Usher syndrome Type 1 being the most severe with an earlier onset of retinal degeneration (Ferrari et al., 2011). Usher syndrome type 2 patients suffer more moderate hearing loss that is correctable with hearing aids, have less severe RP and lack vestibular dysfunction (Ebermann et al., 2009). Usher syndrome type 3 (USH3) is a rare and more variable phenotype that is more common in specific demographics like Finnish and Ashkenazi Jewish populations due to founder effects (Pakarinen et al., 1995, Ness et al., 2003). Linkage studies on Finnish patients showed a common haplotype containing c.300T>G, p.(Tyr100\*) in an uncharacterised gene on 3q21-q25 which is now characterised as *CLRN1* (Joensuu et al., 2001). The other founder mutation was a homozygous mutation NM\_174878.3:c.143T>G, NP\_777367.1:p.(Asn48Lys) in *CLRN1* that was found to be causative for USH3 in 30 American families of Ashkenazi descent (Ness et al., 2003).

Bardet-Biedl syndrome (BBS) comprises RP, postaxial polydactyly, hypogonadism, renal disease, obesity and developmental disorders (cognitive impairment) (Elawad et al., 2022). This is an autosomal recessive condition arising from dysfunction in primary cilia that results in systemic manifestations

(Tsang et al., 2018). To date, there are 24 genes implicated in BBS (Table 1.1). The incidence of BBS in Europeans is approximately 1 in 160,000, but it is much higher in the Bedouins of Kuwait (1 in 13,500), where the consanguinity rate is remarkably high (Frag and Teebi, 1989).

Leber congenital amaurosis (LCA) refers to a group of congenital retinal dystrophies with severe reduction in visual acuity (~20/400) or complete blindness (Tsang and Sharma, 2018a). The clinical symptoms include severe visual loss, involuntarily eye movement (nystagmus), tendency to rub the eyes (oculodigital reflex), hyperopia, deep set eyes, lens clouding (cataracts) and cone shaped protruding corneas (keratoconus) (Tsang and Sharma, 2018a). There are 25 genes underlying LCA and these are involved in various retinal functions, such as photoreceptor morphogenesis, phototransduction, the retinoid cycle, ciliary transport processes and guanine synthesis (Kumaran et al., 2017). Certain genotypes can modify the clinical manifestation of LCA to include extra-ocular anomalies. Patients with mutations in *CEP290* can display LCA with intellectual disability and ataxia similar to Joubert syndrome (Perrault et al., 2007). A recurrent deep intronic mutation NM\_025114.4:c.2991+1655A>G, NP\_079390.3:p.(Cys998\*) resulting in a hypomorphic allele was detected in 21% of a European cohort with LCA (n=77) (den Hollander et al., 2006). This common mutation is currently a therapeutic target in for antisense oligonucleotide therapy known as sepiofarsen to block the insertion of a cryptic exon in the messenger RNA (mRNA) and has shown promising results (Russell et al., 2022).

Phenotype	Genes	Inheritance
Retinitis pigmentosa	<i>ABCA4, ADGRA3, AGBL5, AHR, ARHGEF18, ARL6, ARL2BP, BBS1, BBS2, C8orf37, CC2D2A, CERKL, CLCC1, CLRN1, CNGA1, CNGB1, CRB1, CWC27, CYP4V2, DHDDS, DHX38, EMC1, ENSA, EYS, FAM161A, HGSNAT, IDH3B, IFT140, IFT172, IMPG2, KIAA1549, KIZ, LRAT, MAK, MERTK, MVK, NEK2, NEUROD1, PCARE, PDE6A, PDE6B, PDE6G, POMGNT1, PRCD, PROM1, PROS1, RAX2, RBP3, REEP6, RGR, RLBP1, RP1L1, SAMD11, SLC7A14, SPATA7, TRNT1, TTC8, TULP1, USH2A, ZNF408, ZNF513, ADIPOR1, ARL3, CA4, CRX, FSCN2, GUC A1B, HK1, IMPDH1, IMPG1, KIF3B, KLHL7, PRPF3, PRPF4, PRPF6, PRPF8, PRPF31, PRPH2, RDH12, ROM1, RP9, SEMA4A, SNRNP200, SPP2, TOPO RS, NR2E3, NRL, BEST1, RP1, RPE65, SAG, RHO, OFD1, RP2 and RPGR</i>	Autosomal recessive, dominant and X-linked
Usher Syndrome	<i>ABHD12, ADGRV1, ARSG, CDH23, CEP250, CEP78, CIB2, CLRN1, ESPN, HARS, MYO7A, PCDH15, USH1C, USH1G, USH2A and WHRN</i>	Autosomal recessive
Bardet-Biedl syndrome	<i>ADIPOR1, ARL6, BBIP1, BBS1, BBS2, BBS4, BBS5, BBS7, BBS9, BBS10, BBS12, C8orf37, CEP19, CEP290, IFT172, IFT27, INPP5E, LZTFL1, MKKS, MKS1, NPHP1, SDCCAG8, TRIM32 and TTC8</i>	Autosomal recessive
Leber congenital amaurosis	<i>IMPDH1, OTX2, CRX, AIPL1, CABP4, CCT2, CEP290, CLUAP1, CRB1, DTHD1, GDF6, GUCY2D, IFT140, IQCB1, KCNJ13, LCA5, LRAT, NMNAT1, PRPH2, RD3, RDH12, RPE65, RPGRIP1, SPATA7 and TULP1</i>	Autosomal recessive and dominant

**Table 1.1 Genetic heterogeneity in RP and associated disease.** Complete list of identified genes underlying four IRDs obtained from Retnet (<https://sph.uth.edu/RetNet/>, viewed 19.12.23). Colour code implies the inheritance pattern, with genes implicated in autosomal recessive diseases depicted in black, autosomal dominant in blue, X-linked in red and genes where different variants can cause either recessive or dominant disease are highlighted in green.

## 1.12 Foveal hypoplasia

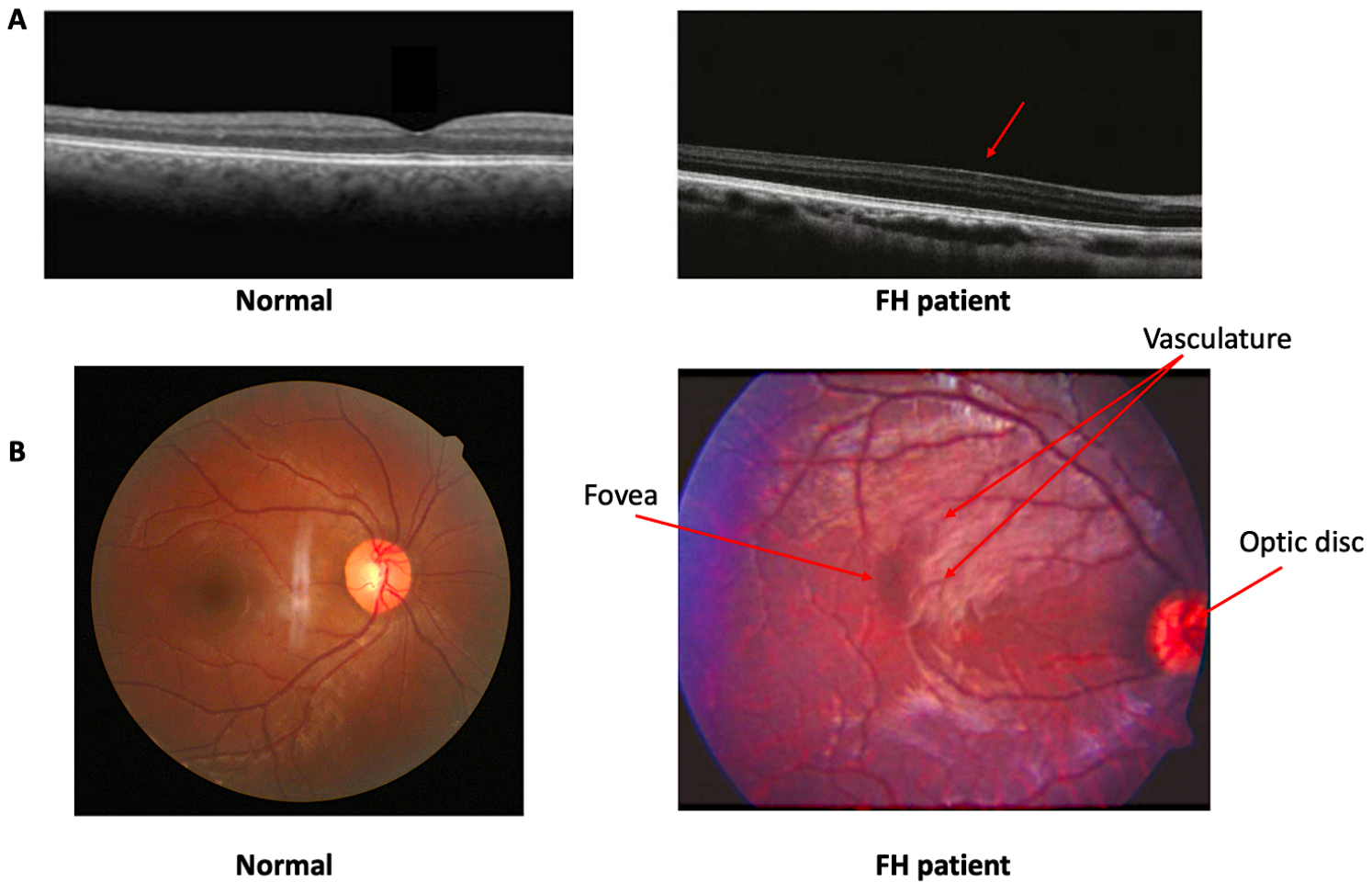
Foveal hypoplasia (FH) is a congenital defect resulting in the complete or partial arrest of foveal pit development. This manifests as the continuity of all retinal layers, including vascularization, at the foveal region. Interruptions to different foveal developmental stages, in the chronological order of centrifugal displacement of neurons, cone photoreceptor specialization and centripetal migration of cone photoreceptors, dictates the severity of FH (Hendrickson and Yuodelis, 1984). Morphological changes relating to foveal development can be visualised by optical coherence tomography (OCT) and are used to grade FH

(Grade 1-4). A higher grade corresponds to a more severe FH and ultimately a worse visual acuity (Thomas et al., 2011, Kruijt et al., 2018). Both grade 1 and 2 FH display photoreceptor outer segment lengthening (cone specialisation) and ONL widening (centripetal migration of cones), but grade 1 has a shallow foveal pit while grade 2 lacks a foveal pit (centrifugal displacement of neurons). Grade 3 and 4 exhibit lack of photoreceptor outer segment lengthening and absence of foveal pit, but ONL widening is present in grade 3 and absent in grade 4 (Thomas et al., 2011). However, there are reports of patients with grade 3 and 4 FH displaying good best corrected visual acuity (BCVA: 20/25 - 20/60). This appears to contradict this grading system, which limits the prognostic utility of using retinal morphology in predicting vision (Kirchner et al., 2019).

Patients with FH typically exhibit absence of or a shallower foveal pit, reduced visual acuity, absence of or a poorly defined foveal avascular zone, lack of foveal reflex and nystagmus (Ehrenberg et al., 2021). Nystagmus is the term used to describe involuntarily eye oscillations due to oculo-motor dysfunction and it is frequently found in FH as well as other early onset IRDs (Choi et al., 2018). Nystagmus makes it difficult to image the fovea using regular OCT due to the constant ocular movement. The FH hallmarks can be detected using an ophthalmoscope, spectral domain OCT (SD-OCT), fundus autofluorescence or OCT angiography (OCTA) to probe the fovea (Figure 1.10) (Kondo, 2018, Gabai et al., 2015, Toral et al., 2017). Alternative imaging techniques like infrared reflectance and ultra-wide field fundoscopy can reveal concentric macular rings, a sign exclusive to foveal hypoplasia (Cornish et al., 2014, Ramtohl et al., 2020). These circular patterns are the result of the unique arrangement of photoreceptor axons in the Henle's fibre layer of FH patients (Ramtohl et al., 2020).

FH is most commonly reported in association with other congenital oculopathies encompassing aniridia, Stickler syndrome, Aland Island disease, optic nerve hypoplasia, achromatopsia, retinopathy of prematurity, incontinentia pigmenti, familial exudative vitreoretinopathy, nanophthalmos, albinism and albinism associated syndromes. However it can also manifest as an isolated entity, though this is less common (Table 1.2) (Schiff et al., 2021).





**Figure 1.10 Characteristics of FH detected using *In vivo* imaging.** A) OCT scan exposing the underlying retinal layers reveals absence of the foveal pit in an FH patient. B) Fundoscopic image of the posterior surface (fundus) of the eye displays an undefined foveal avascular zone in an FH patient. Patient images modified with the permission of Al-Arimi et al. (2013) and Poulter et al. (2013). Normal images obtained from the Leeds Vision Research Group.

Phenotype	Genes & loci	Inheritance
Isolated foveal Hypoplasia	<i>PAX6</i> , <i>SLC38A8</i> , <i>AHR</i> and <i>FRMD7</i>	Autosomal dominant, recessive and X-linked
Ocular albinism	<i>GPR143</i>	X-linked
Oculocutaneous albinism	<i>TYR</i> , <i>TYRP1</i> , <i>OCA2</i> , <i>SLC45A2</i> , <i>OCA5</i> , <i>LRMDA</i> , <i>DCT</i> and <i>SLC24A5</i>	Autosomal recessive
Hermansky-Pudlak syndrome	<i>HPS1</i> , <i>HPS3</i> , <i>HPS4</i> , <i>HPS5</i> , <i>HPS6</i> , <i>AP3B1</i> , <i>DTNBP1</i> , <i>BLOC1S3</i> , <i>BLOC1S5</i> , <i>PLDN</i> and <i>AP3D1</i>	Autosomal recessive
Chediak-Higashi syndrome	<i>LYST</i>	Autosomal recessive
Aniridia	<i>PAX6</i>	Autosomal dominant
Stickler syndrome	<i>COL2A1</i> and <i>COL11A1</i>	Autosomal dominant
Aland Island disease	<i>CACNA1F</i>	X-linked
Incontinentia pigmenti	<i>IKBKG</i>	X-linked
Achromatopsia	<i>CNGB3</i> , <i>CNGA3</i> , <i>GNAT2</i> , <i>ATF6</i> , <i>PDE6C</i> and <i>PDE6H</i>	Autosomal recessive
Optic nerve hypoplasia	<i>HESX1</i> , <i>SOX2</i> , <i>SOX3</i> , <i>VAX1</i> , <i>ATOH7</i> , <i>PROKR2</i> , <i>PAX6</i> , <i>OTX2</i> and <i>NR2F1</i>	Autosomal recessive and dominant
Nanophthalmos	<i>MFRP</i> , <i>PRSS56</i> , <i>CRB1</i> , <i>TMEM98</i> and <i>BEST1</i>	Autosomal recessive and dominant

**Table 1.2 Genes implicated in FH and the corresponding inheritance pattern.** A list of FH associated disorders and the observed inheritance patterns are listed. Genes associated with autosomal recessive disease are highlighted in black, autosomal dominant in blue, X-linked in red and those associated with both autosomal recessive and dominant inheritance in green. *OCA5* refers to a locus mapped to 4q24 for which no causative gene has been identified.

## 1.13 Genetics of Isolated FH

Four genes have been implicated in isolated FH to date, *PAX6* (Azuma et al., 1996), *SLC38A8* (Poulter et al., 2013), *AHR* (Mayer et al., 2019) and *FRMD7* (Thomas et al., 2014a).

### 1.13.1 *PAX6*

*PAX6* encodes a transcription factor responsible for upregulating genes involved in ocular mesenchymal differentiation (Torkashvand et al., 2018). Monoallelic *PAX6* variants are typically associated with aniridia, nystagmus and anterior segment abnormalities, with FH being part of this clinical spectrum (Thomas et al., 2014b). However, heterozygous missense variants in *PAX6* have also been reported to cause isolated FH that is autosomal dominant (MIM:607108) (Azuma et al., 1996). Missense mutations in *PAX6* account for 70% of the non-aniridia phenotypes which includes isolated FH (Tzoulaki et al., 2005). These variants are confined to exons 5, 6 and 9 which encode the paired domain (exon 5 and 6) and the homeobox domain (exon 9) required for deoxyribonucleic acid (DNA) binding activity (Lima Cunha et al., 2019, Tzoulaki et al., 2005). It is thought that the highly pathogenic loss of function (LOF) variants in *PAX6* usually result in aniridia, but more subtle genetic insults result in a milder phenotype of isolated FH.

### 1.13.2 *SLC38A8*

*SLC38A8* encodes a sodium dependent amino acid transporter that has a higher affinity for L-glutamine, L-alanine, L-histidine, L-aspartate and L-arginine (Hagglund et al., 2015). This transporter belongs to the sodium coupled neutral amino acid transporter (SNAT) family and is thought to regulate the glutamate/glutamine cycle in the CNS (Hagglund et al., 2015). Protein localisation studies conducted on human biopsies demonstrated *SLC38A8* localisation to neuronal cells and a subset of glial cells in the brain, and to the neural retina of the eye (Poulter et al., 2013). In contrast, the findings in murine biopsies suggest that *SLC38A8* expression is limited to the eye (Perez et al., 2014).

Biallelic pathogenic variants in *SLC38A8* (MIM:615585) were found to cause autosomal recessive isolated FH and chiasmal misrouting, with or without anterior segment dysgenesis, in the absence of any pigmentary abnormality, in 7 families (Poulter et al., 2013). This was a paradigm shift, since FH with chiasmal misrouting is one of the diagnostic criteria indicative of albinism (Al-Araimi et al., 2013). Additional families with pathogenic *SLC38A8* variants were reported with the same phenotype and without any pigmentation defects (Schiff et al., 2021, Kuht et al., 2020, Kruijt et al., 2022). However, chiasmal misrouting was not evaluated in some cases (Perez et al., 2014, Campbell et al., 2019, Ehrenberg et al., 2021, Weiner et al., 2020, Toral et al., 2017, Lasseaux et al., 2018, Hayashi et al., 2021, Kruijt et al., 2022).

### **1.13.3 AHR**

The *AHR* gene encodes a ligand activated transcription factor that upregulates xenobiotic metabolizing enzymes, a class of catalytic proteins involved in detoxification (Mayer et al., 2019). Splicing mutations in this gene are responsible for autosomal recessive RP (RP85; MIM:618345) (Zhou et al., 2018). Recent works by Mayer and colleagues (Mayer et al. 2019) identified a homozygous nonsense variant as the cause of autosomal recessive infantile nystagmus with isolated FH in a single family. This premature stop codon is predicted to induce nonsense mediated decay (NMD) resulting in null alleles. Based on this loss of function concept, it suggests that variants interfering with the activity of the Q rich domain required for transcriptional activation would also manifest isolated FH.

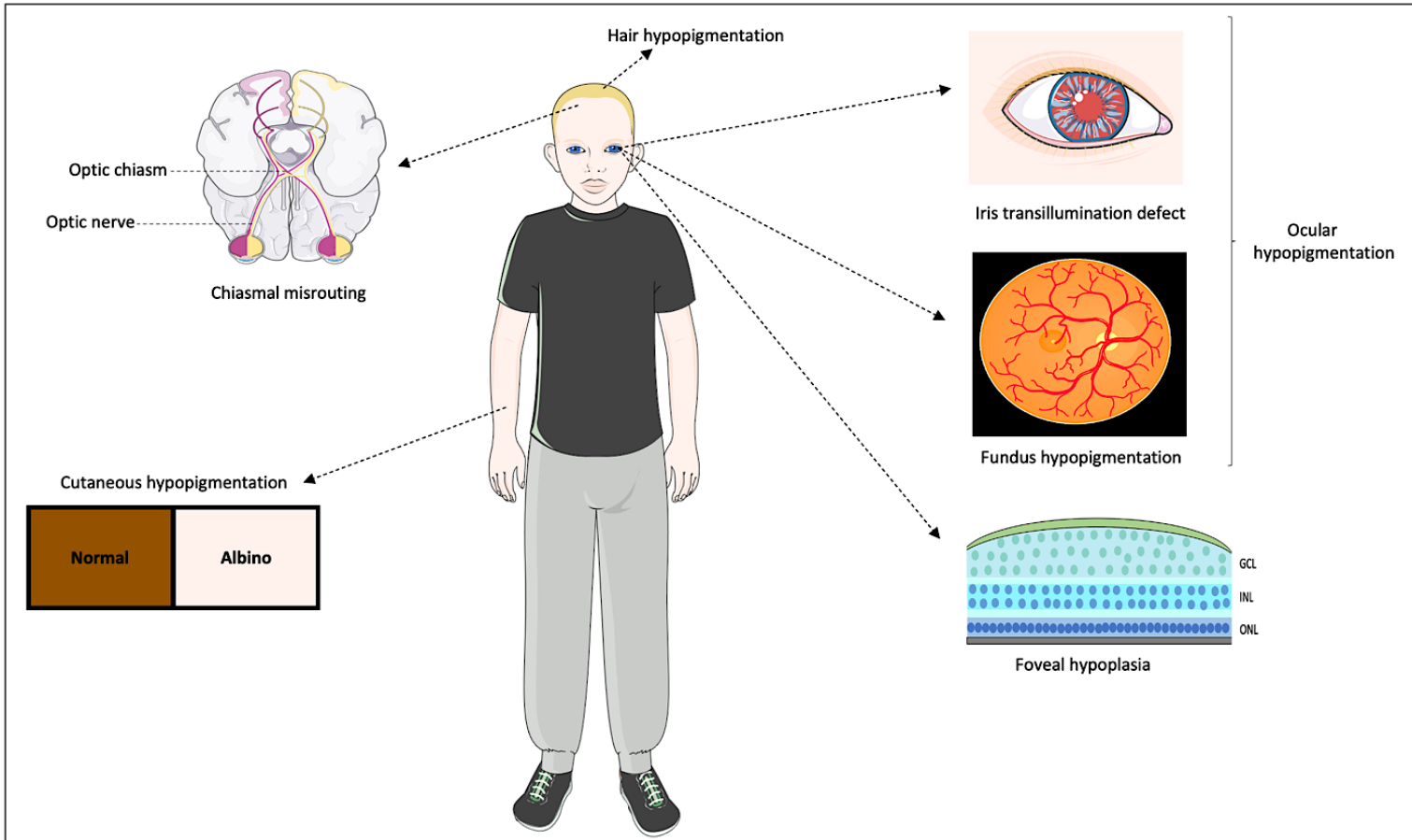
### **1.13.4 FRMD7**

Mutations in *FRMD7* are known to cause infantile nystagmus with an X-linked inheritance pattern (Tarpey et al., 2006). Recently, with the aid of SD-OCT, retinal examination of 45 subjects with *FRMD7* mutations revealed a shallower foveal pit (Grade 1) in 12 patients, in contrast to the total absence of the foveal pit observed in typical FH cases (Thomas et al., 2014a). This phenotype was attributed to splice site and nonsense variants that result in null alleles or missense changes that disrupt the band 4.1, ezrin, radixin and moesin domain (FERM) or the FERM-adjacent domain of the *FRMD7* protein (Thomas et al.,

2014a). The function of FRMD7 is yet to be elucidated, but the expression profile using reverse transcriptase polymerase chain reaction (RT-PCR) in human embryo brain biopsies (37 days post ovulation) hinted at an oculo-motor developmental role (Tarpey et al., 2006).

### **1.14 Albinism**

Albinism describes a group of hereditary disorders impairing melanin homeostasis, which result in systemic or ocular hypopigmentation (Montoliu et al., 2014). In oculocutaneous albinism, the skin, hair and eyes are affected, while in ocular albinism the pigmentation defect is limited to the eyes (Tsang and Sharma, 2018b). Other syndromic forms of albinism include Chediak-Higashi syndrome and Hermansky-Pudlak syndrome, which are accompanied by other pathological anomalies such as immunodeficiency and bleeding diathesis (Huizing et al., 2020). All forms of albinism typically exhibit ophthalmic features of reduced visual acuity, nystagmus, fundus hypopigmentation, FH, chiasmal misrouting and an iris transillumination defect (TID) (Kruijt et al., 2018) (Figure 1.11). Iris TID is a pigmentation defect due to low melanin content in the Iris stroma, which allows light to pass through the translucent iris to be reflected back, appearing as red illumination. Chiasmal misrouting is a term describing aberrant optic nerve projections that hinder stereoscopic vision (see section 1.17). The phenotypic expression of albinism clinical features remains variable in terms of severity as reflected by the different grading of TID, FH and fundus hypopigmentation (Kruijt et al., 2018, Zhong et al., 2021).



**Figure 1.11 Clinical phenotype of albinism.** African child exhibiting the classical manifestations of oculocutaneous albinism. Box displays normal cutaneous pigmentation in contrast to an albino's hypopigmented skin. Ocular defects comprise absence of foveal pit, pale fundus with vascular incursion at the fovea, light coloured irides that are translucent and optic nerve chiasmal misrouting. Diagram generated using modified graphical models obtained from Servier Medical Art (<https://smart.servier.com>). Servier Medical Art is licensed under CC BY 4.0.

## 1.15 Genetics of Albinism

### 1.15.1 Oculocutaneous albinism (OCA)

There is considerable genetic heterogeneity in albinism. OCA can be classified into eight subtypes (OCA1-OCA8) based on the gene affected, with 7 genes and 1 locus implicated (Figure 1.12). OCA exhibits an autosomal recessive inheritance pattern due to biallelic LOF of enzymes (*TYR*, *TYRP1* and *DCT*), membrane bound transporters (*OCA2*, *SLC45A2*, *SLC24A5*) and a structural protein (*LRMDA*) involved in the melanogenesis pathway (Chan et al., 2021, Gronskov et al., 2013, Tomita et al., 1989, Rinchik et al., 1993, Wei et al., 2013, Newton et al., 2001, Pennamen et al., 2021, Boissy et al., 1996). The prevalence of different OCA subtypes varies globally due to the presence of founder mutations in local populations. For example, OCA due to *OCA2* pathogenic variants are more frequent in individuals of African descent (1:15000) than in European populations (1:30000) (Okoro, 1975). A known founder haplotype on chromosome 11q14 known as GYGQ constitute SNP rs187887338 (G), p.(Ser192Tyr) (Y), rs147546939 (G) and p.(Arg402Gln) (Q). This haplotype is exclusive to Europeans and harbours two *TYR* variants in cis, NM\_000372.5:p.[(Ser192Tyr);(Arg402Gln)]. These mutations can be missed in genetic screening due to the high carrier frequency and the milder phenotype (Gronskov et al., 2019). This complex allele is considered pathogenic based on evidence from seven unrelated OCA families reported to have the NM\_000372.5:p.[(Ser192Tyr);(Arg402Gln)] variant in trans with a pathogenic *TYR* variant (Campbell et al., 2019).

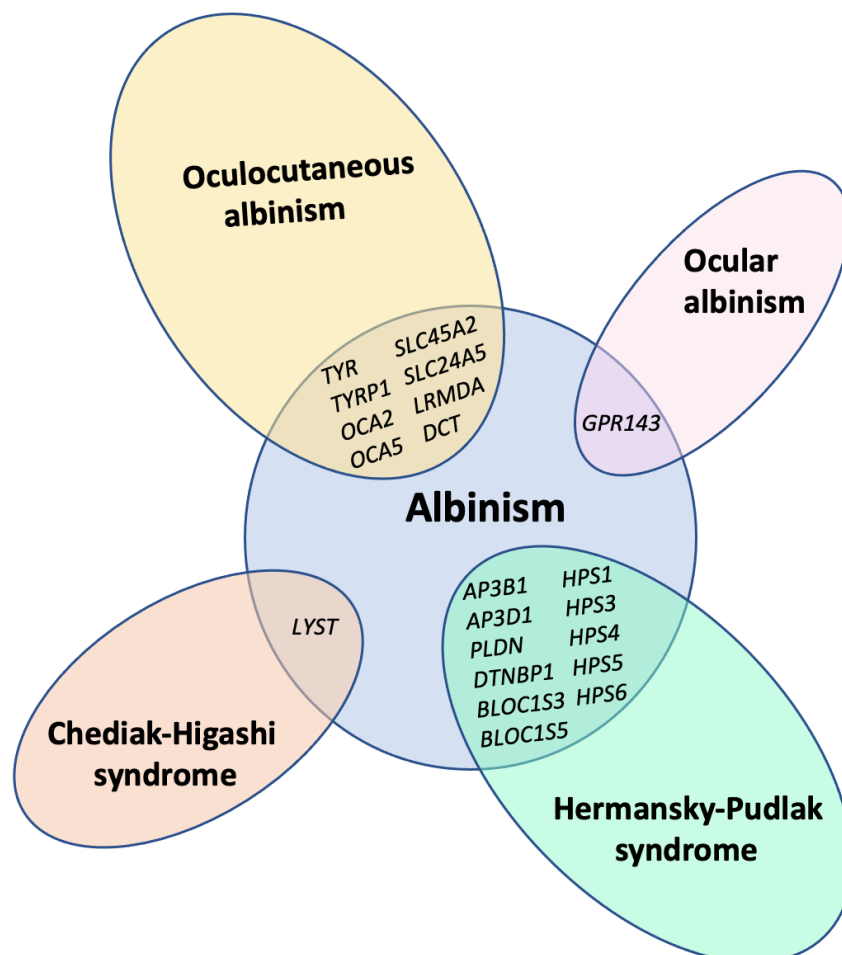
### 1.15.2 Ocular Albinism (OA)

OA is a monogenic disease that follows an X-linked pattern of inheritance and is caused by pathogenic variants in *GPR143* (Bassi et al., 1995). The gene encodes a G protein coupled receptor expressed in melanosomes and the RPE, where it acts as a receptor for L-3,4-Dihydroxyphenylalanine (L-DOPA), a melanin precursor in the melanogenesis pathway (Lopez et al., 2008). Female carriers present with an intermediate phenotype of mild fundus hypopigmentation attributed to X-inactivation, causing distinctive mosaic patches of hypopigmentation (Mao et al., 2021, Zhong et al., 2021). *GNAI3* has also been

identified to be a promising candidate for OA. It shares the same signalling pathway as *GPR143* (Young et al., 2016), and complete loss of function of *GNAI3* in mice resulted in macromelanosomes and reduced melanosomal density in RPE, both of which are observed in OA in humans (Young et al., 2011).

### 1.15.3 OCA syndromes

Hermansky-Pudlak and Chediak-Higashi syndromes are autosomal recessive OCA syndromes, with the former being caused by variants in 11 genes and the latter being specifically due to variants in the *LYST* gene (Figure 1.12). Both syndromes share a common aetiology of defective intracellular trafficking of transport and storage vesicles that leads to hypopigmentation with bleeding susceptibility in all patients.



**Figure 1.12 Genes underlying albinism.** Illustration of causative genes underlying non-syndromic and syndromic forms of albinism. OCA5 refers to a genetic locus of 4q24.



### 1.15.3.1 Hermansky-Pudlak syndrome (HPS)

HPS is caused by the loss of function of genes involved in lysosomal trafficking and biogenesis of lysosome-related organelles (LRO) (Dell'Angelica et al., 1999, Falcón-Pérez et al., 2002). LROs are subcellular mobile compartments that transport various cargoes in the endolysosomal system (Dell'Angelica et al., 2000). LROs such as melanosomes in melanocytes, dense granules in platelets and lytic granules in cytotoxic T cells, when defective, give rise to the hallmarks of HPS (Falcón-Pérez et al., 2002). HPS is characterised by OCA and an increased susceptibility to bleeding (bleeding diathesis). Additional anomalies of immunodeficiency (HPS2 and HPS10), granulomatous colitis (HPS1, HPS4 and HPS6) or pulmonary fibrosis (HPS1, HPS2 and HPS4) can manifest but are subtype specific (Hermansky and Pudlak, 1959, Li et al., 2022).

To date 11 genes have been identified in HPS, corresponding to the different subtypes (Table 1.3). HPS genes encode subunits of the Biogenesis of Lysosome-related Organelles Complexes (BLOCs) and adaptor protein complexes (AP) (Nazarian et al., 2003, Di Pietro et al., 2004). Three BLOCs have been implicated in HPS, BLOC-1, BLOC-2 and BLOC-3 (Nazarian et al., 2003, Di Pietro et al., 2004, Falcón-Pérez et al., 2002). Both BLOC-1 and BLOC-2 orchestrate cargo delivery to LROs by defining endosomal tubules that act as routes to direct recycling endosomes to targets (Dennis et al., 2015, Delevoye et al., 2016). In contrast, BLOC-3 regulates the localisation of lysosomes and late endosomes (Nazarian et al., 2003, Dennis et al., 2015, Delevoye et al., 2016).

The subunits of the adaptor protein AP-3, encoded by *AP3B1* and *AP3D1*, have been involved in HPS. The protein product of *AP3B1* is the  $\beta$ 3A subunit while *AP3D1* encodes the  $\delta$  subunit (Badolato and Parolini, 2007). AP-3 is involved in selective sorting of cargo in the secretory and endocytic transport pathways. For example, AP-3 mediates sorting of tyrosinase from early endosomes into melanosomes through the recognition of a dileucine sorting motif (EEKQPLL) required for the correct localisation of tyrosinase in melanocytes (Theos et al., 2005). Unlike adaptor protein AP-1, dysfunction in ubiquitously expressed AP-3

is compatible with life, affecting only cell specific sorting events to manifest HPS (Zizioli et al., 1999).

HPS subtype	Gene	Complexes	Phenotypic severity	OMIM
HPS1	<i>HPS1</i>	BLOC-3	Severe phenotype, granulomatous colitis and pulmonary fibrosis	<a href="#">203300</a>
HPS2	<i>AP3B1</i>	AP-3	Immunodeficiency and pulmonary fibrosis	<a href="#">608233</a>
HPS3	<i>HPS3</i>	BLOC-2	Mild phenotype	<a href="#">614072</a>
HPS4	<i>HPS4</i>	BLOC-3	Severe phenotype, granulomatous colitis and pulmonary fibrosis	<a href="#">614073</a>
HPS5	<i>HPS5</i>	BLOC-2	Mild phenotype	<a href="#">614074</a>
HPS6	<i>HPS6</i>	BLOC-2	Mild phenotype and granulomatous colitis	<a href="#">614075</a>
HPS7	<i>DTNBP1</i>	BLOC-1	Classic HPS	<a href="#">614076</a>
HPS8	<i>BLOC1S3</i>	BLOC-1	Classic HPS	<a href="#">614077</a>
HPS9	<i>PLDN</i>	BLOC-1	Classic HPS	<a href="#">614171</a>
HPS10	<i>AP3D1</i>	AP-3	Immunodeficiency	<a href="#">617050</a>
HPS11	<i>BLOC1S5</i>	BLOC-1	Mild phenotype	<a href="#">619172</a>

**Table 1.3 Classification of HPS subtypes and associated phenotypes.** All forms of HPS exhibit OCA and bleeding diathesis. Different subtypes exhibit varying severity and are associated with additional pathologies. Immunodeficiency constitutes neutropenia or hemophagocytic lymphohistiocytosis (HLH) in some cases. The disease identifier in OMIM database is provided for reference.

### 1.15.3.2 Chediak-Higashi syndrome (CHS)

CHS is caused by biallelic pathogenic variants in the *LYST* gene, which encodes a lysosome trafficking regulator protein. These lead to the accumulation of enlarged lysosomes and LROs within leukocytes, platelets, and melanocytes (Zelickson et al., 1967, Gil-Krzewska et al., 2016). The exact pathomechanism remains poorly understood. Histological studies suggest that the aberrantly enlarged melanosomes (macromelanosomes) within melanocytes are formed by either the fusion or the unregulated growth of premelanosomes. These

macromelanosomes exceed the transport capacity of the trafficking machinery for delivery to target cells, and this, coupled with the reduced availability of normal melanosomes, results in hypopigmentation (Zelickson et al., 1967).

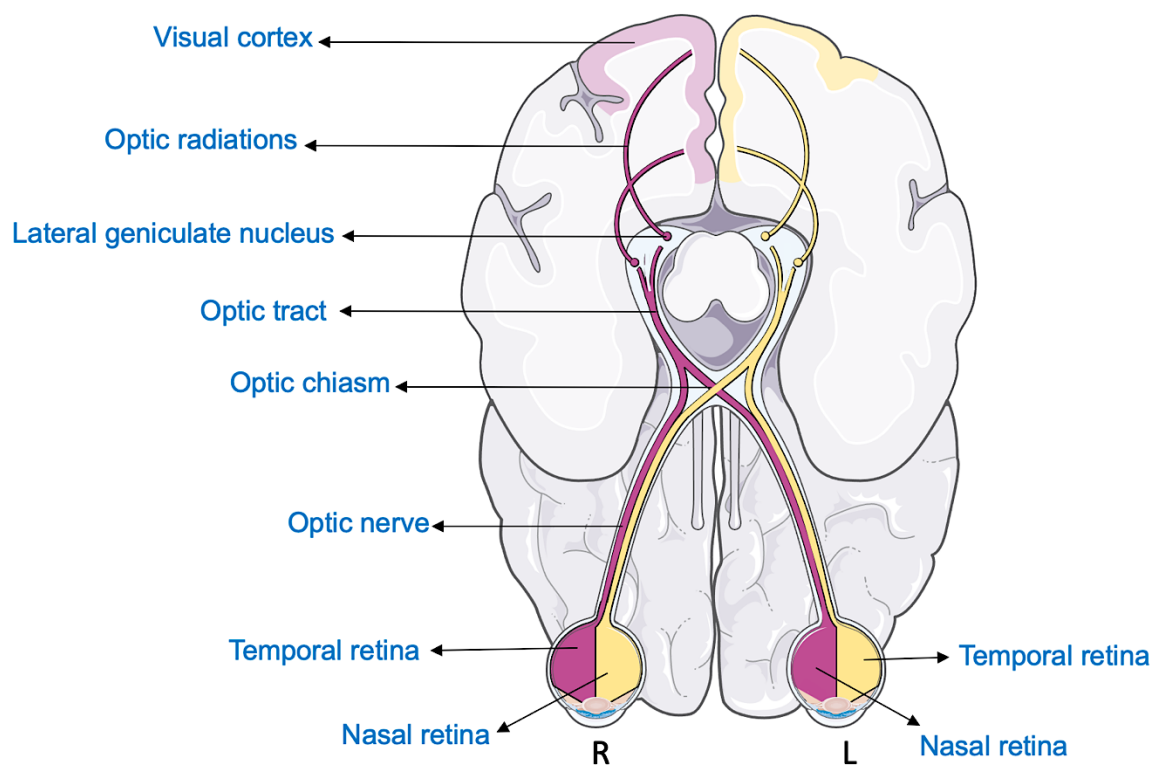
Pathogenic variants affecting the ARM/HEAT or BEACH domains of the LYST protein impair the morphology, transport, and exocytosis of lytic granules (secretory lysosomes), which underlies the defective cytotoxicity of natural killer cells and T lymphocytes in CHS (Gil-Krzewska et al., 2016). Peripheral blood smears of CHS patients also show distinctive and enlarged lytic granules (azurophilic granules) present within neutrophils. This morphological change serves as a pathognomonic feature that aids in differential diagnosis (Antunes et al., 2013, Dame et al., 2019). In terms of genotype-phenotype correlation, most patients will develop albinism, immune dysfunction, bleeding tendency and HLH. Once HLH develops, the condition is said to be in the CHS accelerated phase (Bharti et al., 2013). HLH is an immunological syndrome resulting in multiorgan dysfunction and mortality in patients within the first decade of life (Maaloul et al., 2016). The urgency for a genomic diagnosis in albinism is important to support early medical interventions and prolong life.

### **1.16 Albinism phenotypic overlap with *SLC38A8* isolated FH**

The difficulty in detecting systemic or ocular hypopigmentation in Northern European descendants prompted the use of FH and chiasmal misrouting as diagnostic criteria for albinism (Kruijt et al., 2018, Campbell et al., 2019). This means that a significant proportion of *SLC38A8*-associated FH cases are misdiagnosed as albinism based on phenotypic examination, even though for the most part they have no obvious pigmentation defects. Reports of iris TID in five families with *SLC38A8*-related isolated FH further blurs the phenotype, suggesting the possibility of mild pigmentary disturbance and making *SLC38A8*-associated FH indistinguishable from OA in some cases (Kuht et al., 2020, Lasseaux et al., 2018, Schiff et al., 2021). The validity of these reports is discussed in section 3.3.5. Nevertheless, to aid in differential diagnosis, *SLC38A8* must be investigated in albinism cases (Dumitrescu et al., 2021).

## 1.17 Chiasmatic misrouting

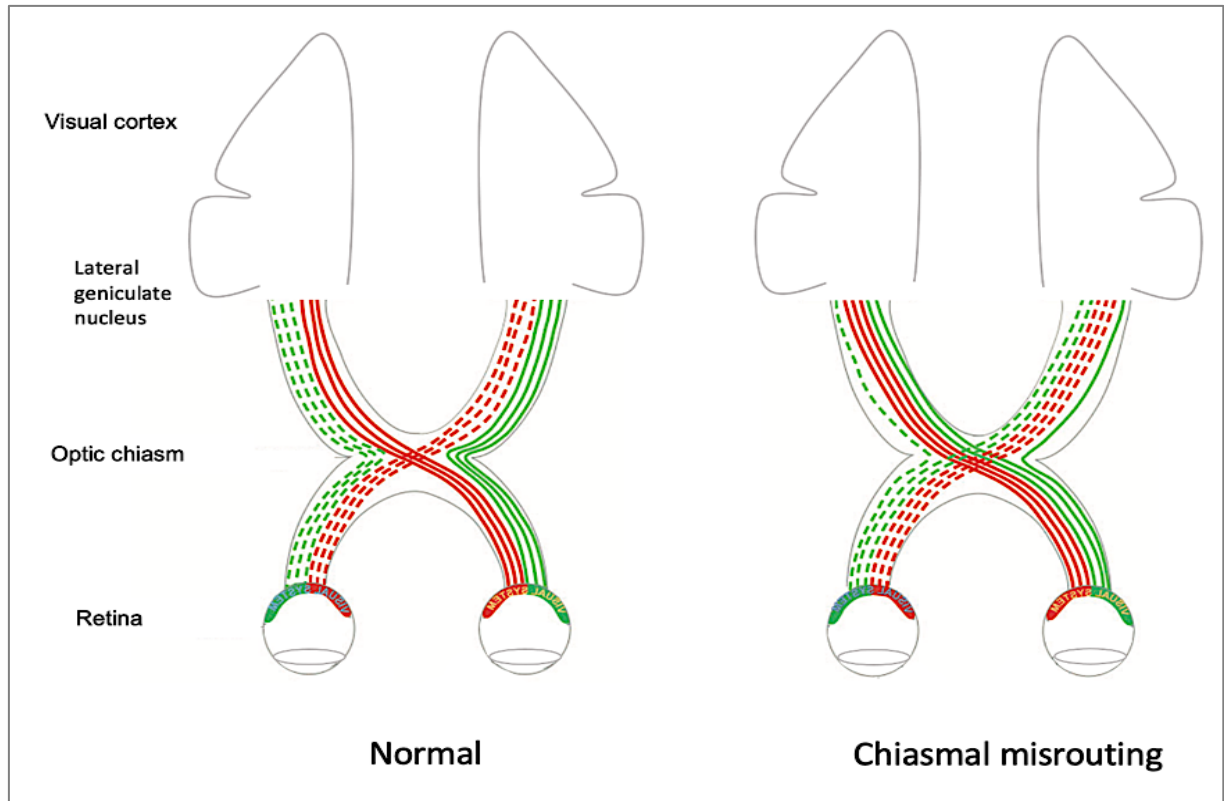
In healthy individuals, the temporal hemiretina (half of the retina close to the temples) of one eye projects its optic nerve fibres to the ipsilateral (same side) visual cortex of the occipital lobe, while the nasal hemiretina (retinal half close to the nose) projections cross to the contralateral (opposite) occipital cortex (Figure 1.13). This allows each cerebral hemisphere to receive visual inputs from both eyes, which forms the basis of binocular vision (stereoscopic vision) (Hoffmann and Dumoulin, 2015). The junction where contralateral projections cross each other to reach the opposite cerebral hemisphere is termed the optic chiasm (Ather et al., 2019). These retinal projections are almost equally split to each hemisphere, 53% being contralateral projections and 47% are considered ipsilateral (Kupfer et al., 1967).



**Figure 1.13 The optic pathway in humans.** Visual representation of the optic pathway responsible for binocular vision. The key anatomical structures are labelled and the optic nerve projections reaching the visual cortex are colour coded. Left visual cortex is in violet while the right visual cortex is in yellow. R: right eye; L: left eye. Diagram was designed using graphical elements from Servier Medical Art (<https://smart.servier.com>). Servier Medical Art is licensed under CC BY 4.0.

Chiasmal misrouting, also known as an optic decussation defect, refers to a skewing of this split in optic nerve projections. This results in partial representation of ipsilateral nerve fibres at the occipital cortex and consequently hinders stereoscopic vision (Neveu et al., 2003) (Figure 1.14). This is due an increase in nerve fibres aberrantly crossing from the temporal hemiretina to the contralateral visual cortex, instead of reaching the ipsilateral occipital cortex (Guillery et al., 1975). This neuronal maldevelopment was initially considered to be a sign exclusive to albinism and was thought to be due to the potential effects of melanin precursor (L-DOPA) on axonal decussation (Ather et al., 2019, Ilia and Jeffery, 1999). However, the identification of the *SLC38A8* phenotype of FH and chiasmal misrouting in the absence of pigmentation defects created a paradigm shift that challenges this notion (Poulter et al., 2013, Kruijt et al., 2022). The routing of optic nerve projections is evaluated using visual evoked potentials (VEP) (Apkarian et al., 1983). Some of the common VEP modalities include flash VEP, which uses flashing lights to measures amplitude, and pattern VEP, which displays a checkerboard like pattern to gauge latency (Carroll et al., 1980, Neveu et al., 2003).

There have been reports of normal optic chiasm in two infants (< 3 years) affected by *SLC38A8*-associated FH (Campbell et al., 2019, Schiff et al., 2021). In one study, normal axonal projections were initially reported during infancy, but subsequent referral at an older age (14 years) showed evidence of chiasmal misrouting (Campbell et al., 2019). The phenotypic discrepancy in these two infant probands is likely to be attributed to the difficulty in obtaining readings from a constantly moving child or can be related to the sensitivity of the test not being effective in infants. The use of VEP testing for chiasmal misrouting must therefore consider a sensible age cut off to obviate unnecessary subsequent molecular investigations and referrals. Investigating age-related changes in the visual pathway showed that the interhemispheric asymmetry in albino patients remained stable past the age of 18 years (Neveu et al., 2003). There were no significant differences in the ipsilateral and contralateral visual cortex for both latency and amplitude of flash VEP. This suggests a reliable age threshold of 18 years or above to be considered for retesting of VEP subjects whom had no evidence of chiasmal misrouting at a young age.



**Figure 1.14 The visual pathway in albinism and *SLC38A8* isolated FH.** Axonal projections in a normal optic chiasm and in chiasmal misrouting. In the normal image, the axonal projections of the nasal retina cross normally to the contralateral visual cortex with no defects. In chiasmal misrouting, there is an aberrant branching of the optic nerve from the temporal retina, such that some axonal projections reach the contralateral instead of the ipsilateral visual cortex. The outcome is a reduced representation of the temporal retina in the respective ipsilateral visual cortex. Image reproduced and adapted from (Erskine and Herrera, 2014) under the licence CC BY 4.0.

### 1.18 Next generation sequencing (NGS)

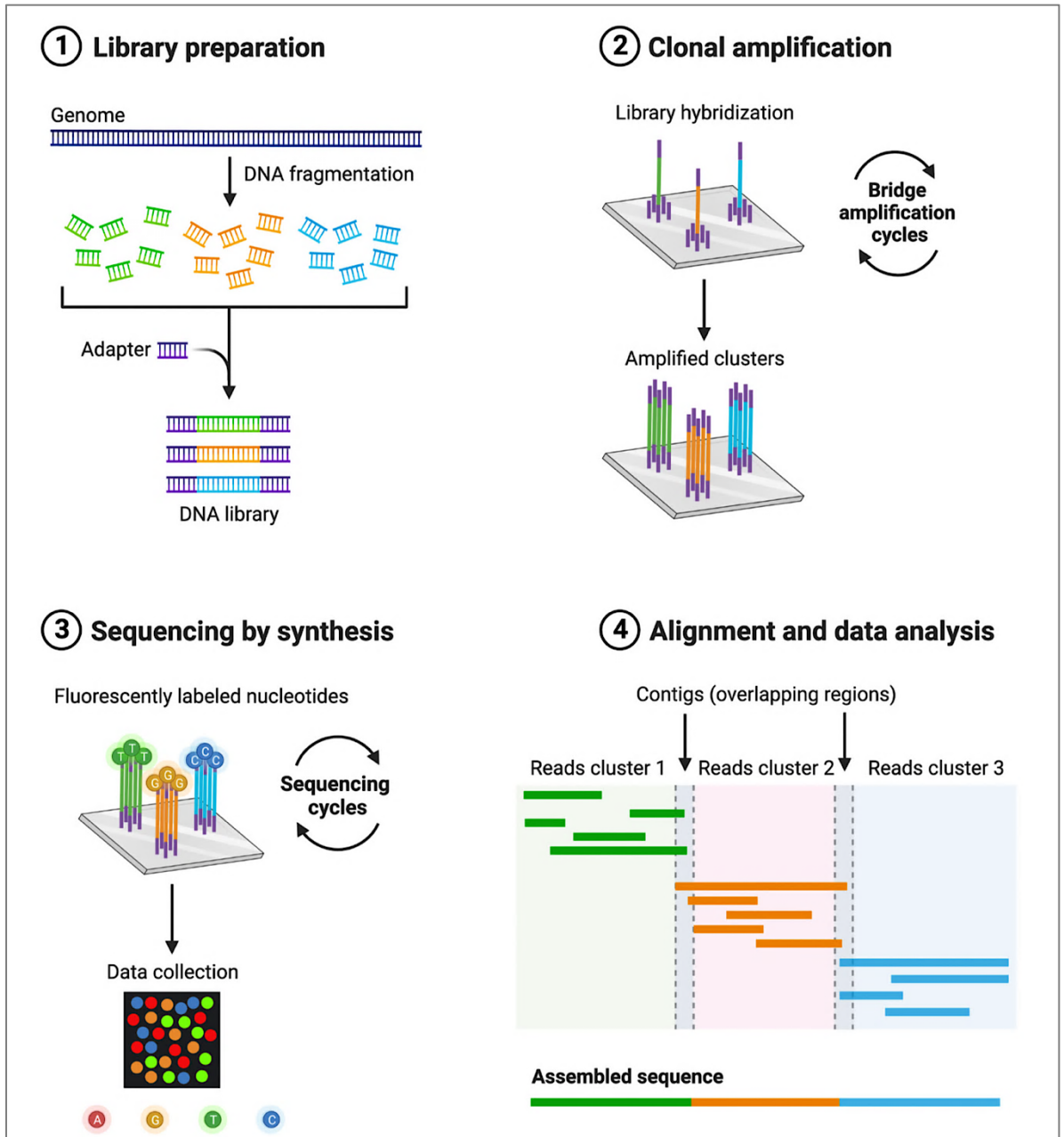
NGS refers to a technology that allows for the massively parallel sequencing of millions of individual DNA molecules in a rapid and cost-effective manner (Myllykangas et al., 2012). This is a significant advancement from first generation sequencing, which is limited to sequencing a single DNA template at a time (Sanger et al., 1977). NGS is an umbrella term encompassing the successors of Sanger sequencing, the second and third generation sequencing technologies

that are classified based on read length (Slatko et al., 2018). High throughput sequencing has revolutionised clinical investigations in rare inherited disorders, providing a diagnostic yield of 52-74% in IRDs (Britten-Jones et al., 2023, Carss et al., 2017). This relatively high diagnostic rate is attributed to the increase in the number of genes being interrogated for germline mutations and the improved resolution of large genomic aberrations (Britten-Jones et al., 2023, Ellingford et al., 2016).

### **1.18.1 Second generation sequencing (Short read)**

The hallmarks of second generation sequencing are multiple clonally amplified short DNA fragments ~300 bp that are sequenced in both directions (paired end reads) to generate multiple representations of all target loci (coverage) (Slatko et al., 2018). The works presented in this thesis relied on the Illumina NGS platform to sequence DNA extracted from patients with the clinical presentation of FH. Illumina benchtop instruments like the MiSeq, HiSeq and NextSeq have become the predominant NGS technology owing to their data quality and accuracy, with a 0.1-0.5% error rate (Stoler and Nekrutenko, 2021, Slatko et al., 2018). The sequencing methodology underlying Illumina technology is known as sequencing by synthesis (Figure 1.15).

The concept is to immobilise the template DNA onto a solid support (flow cell) through hybridisation with complementary oligonucleotide probes, followed by clonal amplification resulting in discrete clusters of identical DNA copies (Bentley et al., 2008). The single stranded DNA is used as a substrate for the incorporation of proprietary fluorescently labelled nucleotides that permit the addition of one nucleotide at a time. A laser is used to excite the fluorophore after each addition of dNTP. The emitted wavelength and signal intensity is captured to determine the base call of each position in the growing chain (Bentley et al., 2008, Myllykangas et al., 2012). This base-by-base sequencing method is performed simultaneously across millions of clusters on the flow cell, generating a significant read output with improved accuracy (Myllykangas et al., 2012).



**Figure 1.15 Illumina WGS workflow.** 1) Template DNA is fragmented into the appropriate length of 350-500 bp. The ends of DNA strands are repaired, the 3' end dA tailed and adaptor sequences are ligated. 2) Denatured template DNA is anchored to complementary sequences on the flow cell and is subjected to massively parallel isothermal amplification. 3) DNA polymerase incorporates fluorescently labelled dNTPs with a reversible terminator on the forward strand that blocks the addition of the next base. Imaging of the incorporated base takes place in cycles, recording a characteristic signal unique to each nucleotide. The process is repeated for the reverse strand using a sequencing primer to generate bidirectional data. 4) Base calling identifies the sequence of each amplicon (read) and these short reads are collated to assemble a full length sequence that is aligned against a reference genome to identify genomic discrepancies. Diagram created with BioRender.com.



## **1.18.2 NGS applications in genomic medicine**

There are many applications for NGS analysis, including sequencing an entire genome (whole genome sequencing), restricting investigation to only the known protein coding regions and splice junctions, amounting to around 1% of the genome (whole exome sequencing) or screening a selected set of disease-specific genes (targeted gene panel).

### **1.18.2.1 Whole exome sequencing (WES)**

The exome and canonical splice sites harbour approximately 85% of disease-causing variants in Mendelian disease (Choi et al., 2009). According to the recommendations of the American College of Medical Genetics (ACMG), this NGS strategy should be considered for clinical indications whereby the disease is uncharacterised, there is a profound genetic heterogeneity or in cases where prior genetic screening in known disease associated genes were unsatisfactory (Directors, 2012). WES provides a cheaper alternative with more manageable data in comparison to whole genome sequencing. However, the exome capture does not achieve complete coverage (~95% coverage) of targets due to PCR sensitivity to guanine and cytosine (GC) content which causes variants to be missed in GC rich regions due to reduced coverage (Meienberg et al., 2015, Choi et al., 2009). The rate of false negatives in WES can reach 0.42% as opposed to whole genome sequencing in exonic variant detection (Meienberg et al., 2016). The limited structural variant (SV) resolution offered by WES is contingent on the SV breakpoints being in the exons and the identification of copy number variants (CNV) is exacerbated by variability in coverage due to varying GC content across the exome which makes it difficult to interpret reduction or increase in reads (Du et al., 2016). The selective enrichment of the entire exome requires additional steps in library preparation that is discussed in 1.18.3.

### **1.18.2.2 Whole genome sequencing (WGS)**

The unbiased WGS strategy was initially clinically effective as a second line molecular test to detect elusive mutations in exons, introns or GC rich regions and improves the resolution of large structural aberrations (Carss et al., 2017)

with a 29% uplift in IRDs diagnostic solve rates (Ellingford et al., 2016). The reduction in WGS associated cost is making the technology more accessible as a first line test to increase the range of different variant classes detected (Abbott et al., 2022). The trade-off for WGS benefits by comparison with WES or a targeted approach is a significant informatics burden relating to storage and variant interpretation. Analysing the exonic, intronic and intergenic segments of the genome captures a significantly larger amount of variants by about several orders of magnitude, this increases the burden for *In silico* analysis (Hocking et al., 2023, Abbott et al., 2022). The interpretation of many noncoding variants according to ACMG guidelines pose a diagnostic challenge since without supplementary functional assays, their characterisation remains obscure (Richards et al., 2015).

WGS uses raw genomic DNA as template without the need for prior amplification. This helps to overcome PCR bias due to intrinsic differences in GC content or amplicon length that affects the melting temperature ( $T_m$ ) and consequently WGS results in a more uniform coverage (Ding et al., 2023). Using raw genomic DNA reduces obstacles to amplification due to extreme base composition such as being highly GC-rich in particular (Aird et al., 2011). The drawback of PCR free libraries is the requirement for larger amounts of input DNA that is of high quality, which can limit application in clinical practice, especially for formalin fixed paraffin embedded (FFPE) biopsies (Robbe et al., 2018).

Recently two major UK genomic research initiatives using WGS have been established to accelerate the translation of genomic findings into the clinical setting. One is the 100,000 genomes project detailed in section 1.20, while the other is the National Genomics Research Library (NGRL) (Caulfield et al., 2020). WGS has shown valuable research potential in maximising scientific discoveries of novel genes and in facilitating genome wide association studies (GWAS) (Carss et al., 2017).

### **1.18.2.3 Targeted panel-based testing**

Diagnostic workflows frequently use targeted NGS analysis as a first line test to interrogate specific genes using a customisable panel which minimises incidental findings and restricts variant discovery to genes of interest (Hocking et al., 2023). This targeted NGS approach offers a higher depth of coverage for a more accurate variant identification, but the spectrum of detectable mutations are limited to those within the coding sequence and splice sites of select genes. The SV detection is limited by the necessity of both breakpoints of the SV being within the gene panel (Wang et al., 2018). As well as screening in Mendelian disease, the high coverage (300X) achieved by gene panels also makes them suitable for use in cancer diagnostics to detect somatic variants at much lower frequencies and for monitoring residual disease post chemotherapy (López-Oreja et al., 2023). Focusing the diagnostic test to a narrow set of genetic targets reduces the sequencing expenditure and computational demands for analysis, resulting in a quicker turnaround time.

The application of virtual gene panels after WES or WGS has become a routine standard of care for many clinical indications, including IRDs, as it yields more manageable data focused on variants in genes relevant to the patient's phenotype while also reserving the exome or genome for future reanalysis (Bernardis et al., 2016, Hocking et al., 2023). This screening approach is constrained by the gene panel, meaning that frequent reviews are required to incorporate newly discovered genes associated with the phenotype (Jespersgaard et al., 2019).

### **1.18.3 Target enrichment**

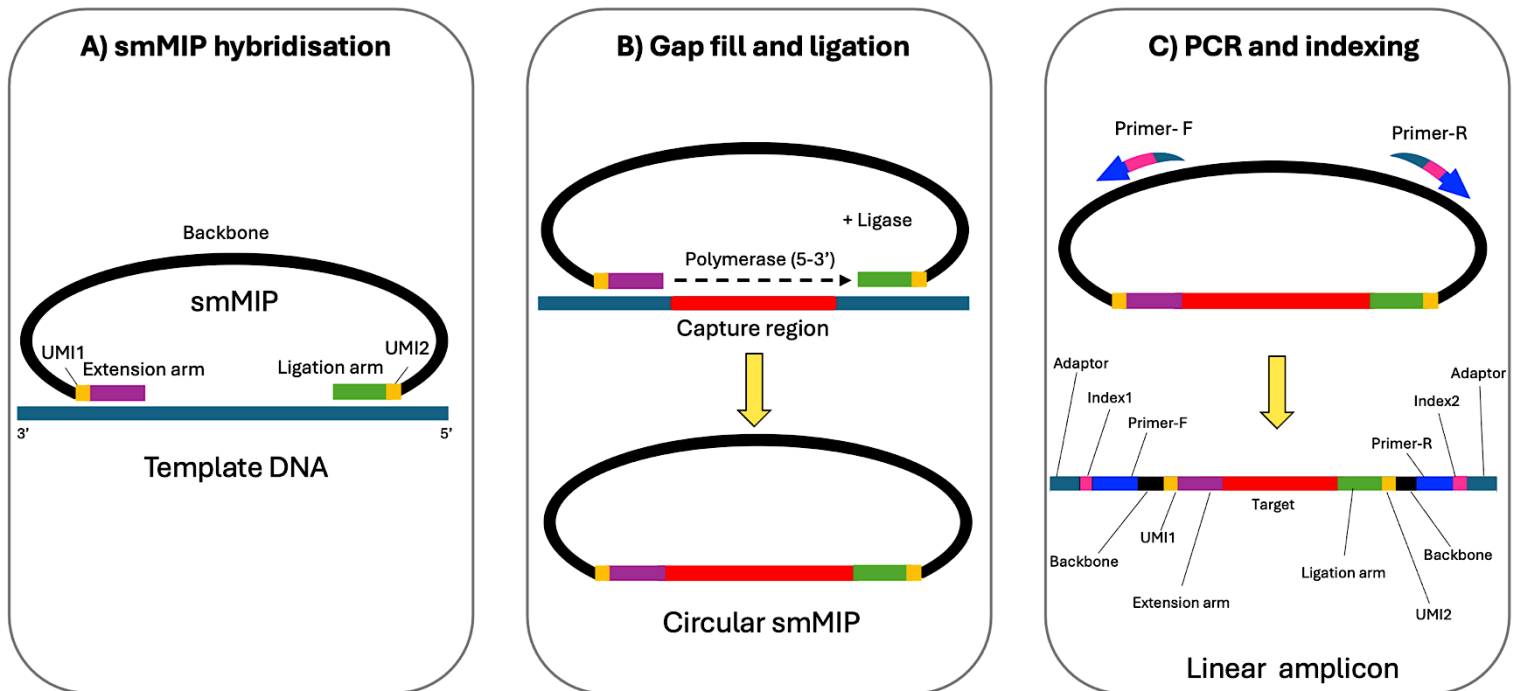
Selective isolation of target genes in panel-based testing, or the entire exome in WES, is a prerequisite in the library preparation for these NGS applications. This can be achieved through different techniques but most notably by, PCR-based amplification to define genes of interest using a cocktail of PCR primers or through hybridisation techniques to capture the DNA corresponding to target genes or entire exomes (Samorodnitsky et al., 2015). The purpose of enrichment

strategies is to maximise on target sequences for better coverage and sequencing efficiency with more manageable data interpretation .

Hybridisation enrichment is the mainstream method owing to its superior performance in terms of size and uniformity of target region captured (~70 Mb), with more accurate variant calling (variant identification), in contrast to the PCR-based capture methods. Hybridisation capture uses single stranded DNA or RNA probes that are either immobilised on microarrays or suspended in solution. Solution-based hybridisation is currently the preferred approach as it improves on the limitations of microarrays in exon capture that is restricted to ~250 bp, offers scalability and is cheaper. This strategy exploits biotinylated oligonucleotide probes that are complementary to the target sequences, followed by the physical isolation of regions of interest using streptavidin-coated magnetic beads (Gnirke et al., 2009).

Other target enrichment techniques include single molecule molecular inversion probe (smMIP) which entails the use of two oligonucleotide probes known as the extension and ligation arms that are connected by a common linker of backbone sequence (Hiatt et al., 2013) (Figure 1.16). The target enrichment involves the hybridisation of the ligation and extension probe arms to a complementary target sequence that flanks a 225 bp region of interest (Panneman et al., 2023). Numerous smMIPs can be combined in a single reaction to target different loci and without the interference with each other (Mc Clinton et al., 2023). Upon PCR amplification, the extension probe arm on the 5' acts as a primer allowing the polymerase to extend the sequence along the target region followed by ligation to the ligation arm at the 3', resulting in a circular probe containing the captured target (Hiatt et al., 2013). A subsequent exonuclease treatment is performed to eliminate unhybridized smMIPs and free template DNA followed by PCR amplification using universal primers that contain indexes and adaptor sequences compatible with the desired sequencing instrument. These primers bind to the common backbone sequence on the circular smMIPs resulting in linearised amplicons that are uniquely tagged and contain the region of interest (Hiatt et al., 2013). The indexes are unique identifier sequences of for the identification of each pooled sample (Hitti-Malin et al., 2022). The smMIPs used also have unique

molecular tags for sequencing to group reads generated from a single smMIP to eliminate PCR artefacts and increase the accuracy of variant calling (Mc Clinton et al., 2023).



**Figure 1.16 smMIPs target enrichment workflow.** A) oligonucleotide probes flank the target region of 225 bp with the 16-20 bp extension arm on the 5' and a 20-24 ligation arm on the 3'. B) DNA polymerase extends the sequence from the extension arm in the 5-3' direction. DNA ligase forms phosphodiester bonds in the backbone of the newly synthesised DNA which seals the gap and consequently forms a circular smMIP. C) Exonuclease III digestion is performed to remove unincorporated smMIPs and excess DNA template. Subsequent target amplification is achieved using PCR and 56-58 bp universal primers that anneal to the 28 bp backbone sequence of the circular smMIP. The captured sequence is PCR amplified resulting in a linear amplicon with 8-12 bp indexes for sample identification and adaptor sequences to bind to the flow cell of the sequencing instrument. The PCR product will also contain 12 bp unique molecular index (UMI) which is unique to each smMIP and is used to group reads generated by a single smMIP. The amplicons are purified using magnetic beads that selectively isolate DNA which is eluted at a later stage once the contaminants are removed. The purified amplicons are quantified and pooled accordingly into single library to be sequenced.

The large multiplexing capacity of smMIPs can offer a low cost of £0.22 per sample per gene and can achieve an average coverage of 431 X for 1712 exons in 372 probands (Hitti-Malin et al., 2022). The advantage of this PCR based enrichment is the requirement of low amount of template DNA with 10-25 ng being adequate and the high depth of coverage is sufficient for multiplexing at a large scale and provides high sensitivity for the detection of variants that occur at low frequency in a sample (Hiatt et al., 2013). The target enrichment workflow is simple, rapid and the assay can be customised to add or remove individual smMIPs without the need to redesign and order the entire probes in the panel as in the case with microarrays and other hybridisation capture methods. These qualities make the high throughput target capture technique of smMIPs applicable to analysing large cohorts especially for laboratories of low budget (Panneman et al., 2023).

#### **1.18.4 Limitations of short-read NGS**

One of the challenges of short-read NGS lies in sequence assembly of highly repetitive genomic regions (Ebbert et al., 2019). For example, failure to assemble sequences covering the highly repetitive *RPGR* ORF15, which accounts for a significant proportion of RP cases, leads to a reduced diagnostic outcome in NGS analysis of IRD patients (Chiang et al., 2018, Bernardis et al., 2016). The bioinformatic dilemma of unambiguous mapping also extends to telomeres and centromeres and these are highly repetitive. Single nucleotide variants, short insertions-deletions (INDELs) and structural variants with breakpoints within these regions may evade detection (Chang and Larracunte, 2019).

Another challenge experienced by amplicon based sequencing or PCR based enrichment techniques for targeted analysis using NGS, is variable PCR performance. The target amplification can be inhibited by purine rich sequences due to their higher denaturing temperature. Genomic regions containing GC-rich sequences have a low amplification efficiency that hampers variant discovery due to a low coverage (Carss et al., 2017). Other intractable genomic features include homopolymers, pseudogenes, highly homologous genes and segmental duplications. Variants residing within these “deadzones” are frequently not

detected and may contribute to missing pathogenic variants in patients (Alkan et al., 2011, Samorodnitsky et al., 2015).

The current human reference genomes of GRCh37 and GRCh38 represents mainly the euchromatin which was easier to sequence for the Human Genome Project (Nurk et al., 2022). The limitation of the human reference genome is that it shows obscurity at intractable regions that are highly repetitive and polymorphic which coincides with the heterochromatin and this genomic gap may cause to the pathogenic variants within these regions to evade detection using short read NGS. To overcome this issue, long read sequencing with the aid of other sequencing, epigenetic and cytogenomic technologies offers a gapless reference sequence with complete representation of 22 autosomes and an X chromosome (100%) encompassing both euchromatin (92%) and heterochromatin regions (8%) (Nurk et al., 2022). The newly assembled reference genome (T2T-CHM13) improved upon GRCh38 by 238 Mb but it lacks allelic diversity and the Y chromosome (Nurk et al., 2022).

### **1.19 Third generation sequencing (long read)**

The inherent limitations relating to short-read sequencing and assembly are addressed by newer long-read sequencing technologies known as third generation sequencing. These are single molecule sequencing technologies that can process DNA fragments of up to 10 kb in length or more in real time (Logsdon et al., 2020). The longer reads are more likely to span entire repeats or homopolymer tracts, overcoming many of the limitations of short-read sequencing to give better coverage that is uniform across the genome (Watson et al., 2021).

Furthermore, long reads have greater overlap than short reads, allowing for more efficient *de novo* assembly to reconstruct a representative consensus sequence of greater accuracy without the need for a supplied reference genome. This is particularly useful in evolutionary biology to assemble a novel genome from multiple shorter reads belonging to a newly identified species (Lok et al., 2017).

A limiting factor for third generation sequencing is the preservation of high molecular weight DNA, which is susceptible to fragmentation during library preparation (Hu et al., 2021).

The adoption of long-read sequencing as a second line test has significantly improved the diagnostic turnaround time in comparison to short-read sequencing strategies. The ability of long-read sequencing to analyse data as it is being sequenced in real time allows for quicker clinical decisions to be made once a plausible variant is identified, without the need to probe for further variants (Watson et al., 2021). Another difficulty sometimes resolved through long-read sequencing is variant phasing to establish the haplotype without the need for parental testing (McClinton et al., 2023b). This is particularly useful in reclassifying VUS in autosomal recessive disorders if the variant resides in trans (Gupta et al., 2023). Elucidating the allelic orientation of variants is especially important to determine pathogenicity in recessive disorders that are prone to complex alleles. For example, in Stargardt disease, a well characterised IRD, the variants NP\_000341.2:p.[Leu541Pro; Arg1443His] in cis and NP\_000341.2:p.(Asp915Asn) in trans in the *ABCA4* gene are known to be sufficient for biallelic pathogenicity (Salles et al., 2017). The association of these variants in a different chromosomal allelic configuration would greatly reduce penetrance.

The methylation profile revealed using long read sequencing is a significant improvement to short read whole genome bisulphite sequencing as it does not involve treatments that degrades the DNA (Sigurpalsdottir et al., 2024). PCR free long read sequencing can provide more accurate identification of epigenetic modifications such as the 5-methylcytosine (5mC) at CpG sites to infer methylation and transcription regulation (Sigurpalsdottir et al., 2024). Monitoring the epigenetic landscape at a target locus of medical relevance can aid in the diagnosis of imprinting disorders such as Prader–Willi syndrome and Angelman syndrome. Both of these syndromes are due to aberrant methylation of a CpG island near the promoter of *SNRPN* on chromosome 15q11.2 (Yamada et al., 2023)



Currently, there are two long-read sequencing technologies with varying performance in terms of error rates and coverage. These are the sequencing platforms offered by the commercial companies of Pacific Biosciences (PacBio) and Oxford Nanopore Technologies (ONT). The work presented in this thesis used the latter for its simplified workflow and because it is more affordable at relatively small scale.

### **1.19.1 Oxford Nanopore Technologies (ONT) platform**

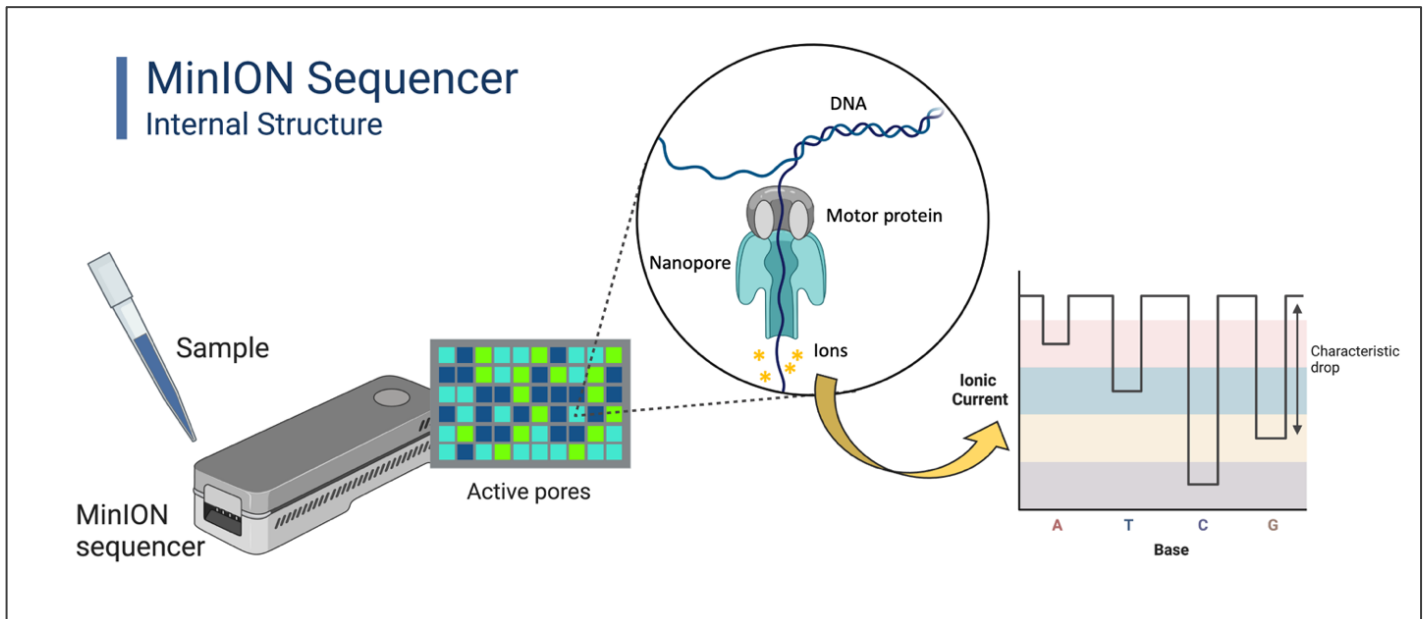
Different ONT products of varying throughput capacity are available for different experimental demands. These are, in order of increasing sequencing volume, the Flongle, MinION, GridION and PromethION. The lower-end product, the Flongle (R9.4.1), was selected for work presented in this thesis due to its cost efficiency of £72 per flow cell, its adequate capacity for small-scale sequencing assays and its rapid performance. The Flongle R9.4.1 achieves a sequencing rate of 240 bases per second and has demonstrated an accuracy of 95.8% corresponding to a quality score of Q13.5 using the Guppy base calling algorithm (version 6) (Ni et al., 2023). This highly portable flow cell adaptor provides a theoretical sequencing output of 2.8 GB, making it ideal for field work and low budget projects (Hu et al., 2021).

The sequencing methodology of ONT involves a motor protein, the phi29 DNA polymerase, that unwinds DNA and propels the single stranded template through an  $\alpha$ -hemolysin protein pore embedded in an electro-resistant membrane. In the presence of applied current, the incoming nucleotides cause a characteristic disruption that enables the identification of each base (Figure 1.17).

Structural variants are often hard to characterise using second generation sequencing platforms due to the low resolution of short-read sequencing. The ONT platform is able to resolve SVs at nucleotide resolution to facilitate SV characterisation for accurate reporting and submission to population databases (McClinton et al., 2023a). Nanopore sequencing is also capable of adaptive sampling, whereby reads can be selectively sequenced based on desired targets

in real time. This dramatically increases the coverage of the target sequence without the need for complex enrichment methods (Martin et al., 2022). Another major advantage of the ONT platform is the generation of ultra-long reads (>100 kb) for reliable *de novo* assembly of genomes, and in targeted sequencing of large targets of medical relevance such as the major histocompatibility complex (MHC) locus. ONT sequencing was able to sequence the entire MHC locus in single ultra-long reads of 4 Mb with 30X coverage (Jain et al., 2018) .

The simplified workflow of ONT eliminates barriers to entry by providing equipment at low or even no cost, with profit made primarily from the sale of consumables. This gives it a competitive advantage over the PacBio platform, which requires significant laboratory expenditure for the purchase of an expensive sequencing instrument (Watson et al., 2021). The ONT workflow begins by removing contaminants using magnetic beads that selectively bind DNA for isolation. Amplicons of different genes from different patients can be pooled after purification for multiplex sequencing, without the need for indexes (McClinton et al., 2023a). The subsequent library preparation entails a DNA repair stage, followed by adaptor ligation to facilitate strand capture and recruitment of processive enzymes at the 5' end (Yahya et al., 2023). Sequencing can be performed for longer durations to increase read output. The drawback of ONT sequencing is the high error rate ~13%, which hinders clinical applications for single nucleotide variant (SNV) discovery (Dohm et al., 2020). However, this can be alleviated by obtaining sufficient read-depth, and there is also significant progress in the development of bioinformatics tools for base call correction (Dohm et al., 2020). The clinical utility of third generation is predicated on the reduction in false positive and false negative calls to justify implementation in clinical and diagnostic services.



**Figure 1.17 Oxford Nanopore sequencing overview.** Samples are loaded onto a flow cell that has sufficient active pores, each containing a viable proprietary protein complex tethered to a polymer membrane. The motor protein (phi29 DNA polymerase) on the 5' end unwinds the double stranded DNA as it pulls it through the nanopore (*Staphylococcus aureus*  $\alpha$ -hemolysin pore). This rate of translocation can be regulated by an internal charge. Each nanopore is connected to a sensory array known as an Application-Specific Integrated Circuit (ASIC) that monitors changes in current. Characteristic disruptions unique to each nucleotide are recorded as an ionic trace that is translated for base calling (Lu et al., 2016). Diagram generated using resources from BioRender.com

## 1.20 The UK 100,000 Genomes Project (100KGP)

The reduction in sequencing costs of massively parallel sequencing increased the feasibility of population sequencing initiatives to accelerate the translation of genomic findings (Myllykangas et al., 2012). Building upon the success of the 1000 genomes project, the National Health Service (NHS) in the United Kingdom partnered with Illumina under the management of Genomics England (GEL) to deliver the largest national sequencing project at the time (Dheensa et al., 2018, Consortium et al., 2015). The project sequenced 100,000 genomes from patients with rare diseases or cancer, supporting the delivery of an integrated genomic

medicine service for the NHS with equitable access to treatments (Trotman et al., 2022, Dheensa et al., 2018). Initially, the project performed two pilot studies between 2014-2015 to assess the feasibility of sample procurement, genome sequencing, data handling and scientific analysis (Peplow, 2016). The clinical merit of WGS was apparent in these trials through its ability to detect a larger repertoire of variants and to identify novel genes underlying disease (Carss et al., 2017).

The 100KGP only targeted patients who had proved negative in prior genetic screening. This project provided patients with the opportunity to benefit from a potential diagnosis through WGS which, at the time, was not routinely available through the NHS, (Dheensa et al., 2018). The insights obtained from interrogating the genome can also redefine a diagnosis into more specific subtypes as exemplified in HPS1-HPS11 subtypes that are based on the affected gene, manage therapeutic options and inform a more accurate prognosis (Trotman et al., 2022). The agnostic approach of WGS demonstrated a 31-33% increase in diagnostic yield in a mix of patients with rare diseases who did or did not previously undergo a genomic test (Investigators et al., 2021). Receiving a molecular diagnosis puts an end to what is often a prolonged diagnostic odyssey for patients, providing them with closure. Delineating the genotype underlying their condition in these patients allows for recruitment to clinical trials in hopes of receiving a stratified medical therapy (Dheensa et al., 2018).

The GEL main programme was launched in 2015-2018 and eligible patients were recruited from across the UK to participate in the 100KGP, based on medical criteria of having a rare inherited disorder or cancer, and lacking a molecular diagnosis at the time of recruitment (Best et al., 2022b). DNA samples from consenting patients were processed at 13 regional genomic medicine centres, all of which conformed to medical laboratory accreditation (ISO15189) for a standardised process with minimal inter-laboratory discrepancies (Turnbull et al., 2018). Large scale WGS was performed at the Wellcome Genome Campus at the Sanger Centre in Hinxton, Cambridgeshire, using Illumina's flagship HiSeq X instrument for high throughput and deep sequencing with streamlined workflows (Peplow, 2016). Quality control parameters for germline samples required a 95%

genome alignment at 15X coverage for reads with a mapping quality >10. In an attempt to limit computational bias, a unified bioinformatic pipeline, Illumina's North Star pipeline (version 2.6.53.23) was applied to all samples for variant discovery and CellBase (version 90/GRCh38) for variant annotation (The 100,000 genomes project protocol v3, accessed 13/11/23). These bioinformatic pipelines are constantly evolving to improve the genomic investigations and reduce the error rate.

### **1.20.1 Genomics England Clinical Interpretation Partnership (GECIP)**

Genomic and clinical data belonging to patients was de-identified and stored in a secure virtual research environment referred to as the GEL embassy, hosted by Invuka. Currently the main programme data release v18 holds 106,263 genomes belonging to 90,178 participants (accessed 04/01/24). Access to the data is restricted to a consortium of approved researchers and clinicians based on expertise in particular disease or function specific domains (Turnbull et al., 2018). A plethora of open-source software suites and command line tools are available to GECIP members within the GEL virtual research environment, to maximise productivity and accelerate scientific innovation. The vision behind the formation of the GECIPs was to enhance the clinical interpretation of genomic data for diagnostic and therapeutic purposes, with convincing results fed back to clinicians.

### **1.20.2 Bioinformatic pipelines and variant annotations**

Genomic data contained within the research embassy is presented in datasets of either rare diseases (73,512 genomes) or cancer (32,751 genomes), with frequent updates in data with each dataset release (<https://re-docs.genomicsengland.co.uk/release18/>). The default analysis pipeline of the rare disease data follows a variant prioritisation framework known as tiering. The tiering workflow annotates variants found in virtual gene panels which have been defined by the scientific community on PanelApp for a targeted analysis of the genome that is relevant to the patient's phenotype (Investigators et al., 2021). The PanelApp is a crowd sourcing gene panel application for the curation of genes by expert reviews from GECIP members (Martin et al., 2019). The tiering

annotation provides supplementary evidence of variant segregation, compliance with the expected mode of inheritance, allele frequency, protein alterations and association with virtual gene panels applicable to the patient's phenotype, to collectively classify variants (Best et al., 2022a). Variants are assigned to three categories (Tier 1-3) (Investigators et al., 2021). Tier 1 variants are considered the most pathogenic category, and it includes protein truncating variants or *de novo* rare variants that segregates with disease, follows the relevant mode of inheritance for the gene and affect a known gene part of a relevant gene panel (Investigators et al., 2021). Tier 2 is similar, but the distinction is that it includes other types of variants such as missense changes that fit the above criteria. Tier 3 variants are plausible variants in genes that are aren't included in gene panels specific to the patient's phenotype, while untiered variants are those considered to be polymorphisms. Only Tier 1 and 2 are clinically assessed by NHS genomic medicine centres for diagnostic purposes.

A second annotation pipeline exclusive to the rare disease cohort is the Exomiser variant prioritisation framework, which retains coding variants with a minor allele frequency (MAF)  $<0.1\%$  in publicly accessible population databases (Smedley et al., 2015). Exomiser uses a scoring system that factors compatibility with the mode of inheritance, population frequencies, pathogenicity predictions and phenotypic homology to known clinical manifestations associated with the gene and orthologues or model organism (Smedley et al., 2015, Robinson et al., 2014). The Exomiser score is generated for each variant and are ordered in a hierarchy ranking the most plausible variants in a participant (Robinson et al., 2014) .

## **1.21 Ethical concerns in DNA sequencing**

The drawback of the untargeted screening by whole genome or exome approaches is that it could reveal genetic changes of medical relevance in genes unrelated to the patient's clinical inquiry. These incidental findings raise ethical concerns surrounding the disclosure of such medically actionable information (Green et al., 2013). The current ACMG reporting guidelines advise geneticists to disclose only secondary findings of pathogenic variants in select rare disease or cancers that are monogenic, of high penetrance and where treatments or

preventative measures are available (Green et al., 2013). Withholding the disclosure of genomic findings in genes outside of the ACMG list would prevent earlier medical intervention in asymptomatic individuals at risk of late onset conditions like age related macular degeneration. To preserve patient autonomy and the fiduciary duty of clinicians, patients are required to consent as to whether they would like to be informed about incidental findings prior to research or diagnostic testing.

Trio testing of probands and their parents is becoming more routinely used in clinical practice to elucidate variant segregation with disease and assess *de novo* status, both of which aid in variant interpretation (Wright et al., 2019). Investigating the inheritance of variants using whole genomes or exomes in this manner can reveal non-paternity, which complicates genetic counselling of the parents. The moral dilemma exists as to whether the counsellor should disclose misattributed parentage and how to communicate its implications for reoccurrence risk in future pregnancies (Cunningham et al., 2024). The increase in NGS use in the absence of an adequate reporting framework is likely to increase this burden on clinicians and genetic counsellors.

## **1.22 Aims of the project**

The work presented in this thesis aims to use modern molecular genetics and cell biology techniques to provide insight into FH pathogenesis. Three separate approaches are employed, as follows.

Firstly, this project aims to characterise the full spectrum of variants in *SLC38A8* that cause FH and uses *In-silico* protein modelling to elucidate the underlying pathogenic mechanisms of missense variants. This data should help clinical geneticists to distinguish likely pathogenic variants from background population variation, improving the diagnostic rate for *SLC38A8*-related FH, and may lead to new insights into disease mechanism and genotype-phenotype correlation.

Secondly, this project aims to increase the diagnostic yield in FH more broadly through detailed analyses of short-read NGS data from cohorts of unsolved FH from the 100KGP and locally recruited cases, with the use of long read sequencing for SV validation.

Lastly, this project aims to perform functional analyses of a novel variant in a potential candidate gene involved in isolated FH, *LAMP1*. This entails the use of *In vitro* splice assays to determine the molecular consequence of the *LAMP1* variant and applying CRISPR-Cas9 gene editing in appropriate cell lines to assess the cellular consequences of the variant.



## Chapter 2 : Materials and Methods

### 2.1 Materials

#### **1X Tris-ethylenediaminetetraacetic acid (TE) buffer (pH 8.0)**

10 mM Tris-HCl (Hydrochloric acid) (pH 8.0)

1 mM EDTA (ethylenediaminetetraacetic acid)

#### **50X Tris acetate ethylenediaminetetraacetic acid (TAE) buffer**

2 M Tris base

0.97 M Glacial acetic acid

50 mM EDTA (pH 8.0)

#### **Luria-Bertani (LB) broth**

1% [w/v] peptone, 0.5% [w/v] yeast extract and 0.5% [w/v] Sodium Chloride (NaCl). Ampicillin or kanamycin were added to achieve a final concentration of 100 µg/ml or 50 µg/ml respectively. LB-agar plates were prepared using 1.5% [w/v] agar granules.

#### **NZY+ broth**

0.98% [w/v] peptone, 0.5% [w/v] yeast extract, 0.094% [w/v] Magnesium Chloride (MgCl<sub>2</sub>), 0.5% [w/v] NaCl and 0.019% [w/v] Citric Acid (anhydrous).

### 2.2 Patients

Ethical approval for this project was granted by the Leeds East Teaching Hospitals NHS Trust Research Ethics Committee (Project number 17/YH/0032), and by other local Research Ethics Committees in other centres. Approval for research on the 100,000 Genomes consortium dataset was granted by the East of England - Cambridge South Research Ethics Committee (project number 14/EE/1112). All patients gave informed written consent, in accordance with the

principles of the Declaration of Helsinki. Affected individuals and their family members were recruited through national and international collaborations with ophthalmologists and clinical geneticists. All patients recruited to the study have a phenotype consistent with foveal hypoplasia. Genomic DNA was obtained from peripheral blood and DNA was extracted by the NHS North East and Yorkshire Genomics laboratory hub using standard protocols.

## **2.3 Methods**

### **2.3.1 Polymerase chain reaction (PCR)**

#### **2.3.1.1 Standard PCR**

Standard PCR utilised 25 µl reactions comprised of 1X reaction buffer (Invitrogen), 200 µM of deoxynucleotide triphosphate (dNTP) mix (Thermo Fisher Scientific), 0.4 µM of each forward and reverse primer (Sigma-Aldrich), 1 mM MgCl<sub>2</sub> (Invitrogen), 25-50 ng template DNA, 1 unit (U) of Taq polymerase (Invitrogen) and nuclease free deionized water (dH<sub>2</sub>O). PCR cycling parameters entailed an initial denaturation step of 95°C for 3 minutes, followed by 35 cycles of 94°C for 30 seconds, 55-65°C for 60 seconds, 72°C for 30 seconds with a final extension stage of 72°C for 10 minutes. PCR was performed on a PTC-200 (MJ Research) or Veriti™ Dx 96 well (Applied Biosystems) thermocycler. Optimisation of PCR conditions utilised PCR<sub>x</sub> enhancing buffer (Invitrogen), Dimethyl sulfoxide (DMSO) (2-6%) (Invitrogen) and manipulation of magnesium concentrations (1 mM - 2 mM) for standard troubleshooting of low amplifications products. Alternatively, Hotshot Diamond Master Mix (Clontar Life Sciences) was used according to the manufacturer's instructions. Briefly this consisted of 1X hotshot Master mix (400 µM dNTP and 6 mM MgCl<sub>2</sub>), Primers at 0.25 µM, 10-20 ng of template DNA and made up to 10µl using dH<sub>2</sub>O

#### **2.3.1.2 Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)**

Ribonucleic acid (RNA) was converted to cDNA using the SuperScript™ II reverse transcriptase kit (Invitrogen). A 12 µl reaction was made up consisting of 50-250 ng of random primers, 500 ng of mRNA or 1-5 µg of total ribonucleic acid

(RNA), a final concentration of dNTP mix (Invitrogen) at 0.5 mM and nuclease free water. This was incubated at 65°C for 5 minutes and then placed on ice. 1X First strand buffer (250 mM Tris-HCl, pH 8.3, 375 mM KCl, 15 mM MgCl<sub>2</sub>; Invitrogen) and 1 mM dithiothreitol (DTT; Invitrogen) was added, followed by an incubation at 25°C for 2 minutes. Optionally, 40 U RNaseOUT (Invitrogen) was added to inhibit ribonucleases. The last step required the addition of 200 U of SuperScript™ II reverse transcriptase (Invitrogen) with sequential incubation at 25°C for 10 minutes, 42°C for 50 minutes and 70°C for 15 minutes.

### **2.3.1.3 Long Range Polymerase Chain Reaction (LR-PCR)**

Amplification of large PCR targets was achieved using the SequalPrep™ Long PCR Kit with dNTPs (Invitrogen). Reactions were carried out in a 20 µl volume comprised of 1-100 ng template DNA, 1X SequalPrep™ reaction buffer (Invitrogen), 0.5-1X SequalPrep™ enhancer A or B (Invitrogen), 1.8 U SequalPrep™ long polymerase (Invitrogen), 0.5 µM primers (Sigma-Aldrich), DMSO (Invitrogen) at 2% of total volume and DNase free water. Thermocycler parameters were an initial denaturation at 94°C for 2 minutes, then 10 cycles of 94°C for 10 seconds, 50-65°C for 30 seconds and 68°C for 1 minute/kb, followed by 30 cycles of 94°C for 10 seconds, 50-65°C for 30 seconds and 68°C for 1 minute/kb (+20 seconds per cycle). A final extension was selected at 72°C for 5 minutes.

### **2.3.1.4 Whole Genome Amplification (WGA)**

Genomic DNA was amplified using the GenomiPhi™ v2 DNA Amplification Kit (Cytiva; supplied by Sigma-Aldrich) to replenish patient's DNA stock when necessary. Reactions were carried out according to the manufacturer's instructions. Single reactions consisted of 1 µl (10 ng) of DNA and 9 µl of sample buffer. The amplification reaction was denatured by heating to 95°C for 3 minutes and was subsequently cooled on ice. 9 µl of reaction buffer and 1 µl of enzyme mix was added to the chilled solution and the samples were incubated at 30°C for 90 minutes for isothermal amplification. Heat inactivation of Phi29 DNA polymerase relied on heating at 65°C for 10 minutes. Serial dilutions of amplification products were conducted at 1/10 and 1/50 to achieve DNA

concentrations compatible with PCR reactions. The WGA products were stored at -20°C.

### **2.3.2 Restriction enzyme digestion**

Restriction enzymes were purchased from Invitrogen, New England Biolabs (NEB) or Promega. All the reactions were prepared according to the manufacturer's specifications except that for NEB enzymes, the reactions were performed in a smaller volume of 25 µl. Each reaction consisted of 1 µg of DNA, 5-20 U of restriction enzyme, the required buffer at the appropriate concentration of 1X and DNase free water. Reactions were incubated at 37°C for 180 minutes then heat inactivated at 65°C or 80°C for 20 minutes. Restriction fragments were predicted using SnapGene Viewer (Dotmatics) (section 2.6.12) then visualised by agarose gel electrophoresis (section 2.3.4).

### **2.3.3 Restriction Fragment Length Polymorphism (RFLP)**

An alternative genotyping strategy was to perform RFLP analysis to exploit variations in the DNA sequence at restriction sites in the target DNA. PCR amplicons or plasmids were digested using the appropriate restriction enzymes (Promega) in a 20 µl reaction consisting of 1X buffer, 2 µg of acetylated bovine serum albumin (BSA), 1 µg of DNA, 5-20 U of restriction enzyme and DNase free water. Incubations were performed at 37°C for 3 hours followed by a 20 minute heat inactivation at 65°C or 80°C depending on the restriction enzyme. The restriction sites were identified using computational tools outlined in section 2.6.12. Digested products were analysed on a 2% agarose gel by electrophoresis as described below.

### **2.3.4 Agarose gel electrophoresis**

DNA was fractionated and visualised by agarose gel electrophoresis, with size based discrimination achieved using a 0.7-3% [w/v] agarose gel. The gels were prepared using UltraPure™ agarose (Invitrogen) and 1X TAE buffer according to the required concentration. Gel staining was performed using 3 µl of Midori

Green Advanced DNA Stain (Geneflow) per 100 ml of agarose gel. Electrophoresis was conducted at a voltage of 150 V and DNA fragments size was estimated by comparison with the DNA size standard of EasyLadder 1 (Bioline) or a 1 kb Plus DNA Ladder (Invitrogen). Agarose gels were imaged using Gel Doc XR+ (Bio-Rad) and banding patterns were visualised using Image Lab v6.0.1 (Bio-Rad).

### **2.3.5 DNA purification**

#### **2.3.5.1 DNA gel extraction**

DNA to be purified from agarose gels were excised under ultraviolet light to isolate thin slices (<0.4 g) containing the target discrete DNA bands. A QIAquick Gel Extraction Kit (Qiagen) was then used to extract the DNA from agarose. 3 volumes of buffer QG was added to 1 gel volume followed by an incubation period at 50°C for 10 minutes. Vortexing was performed after the addition of buffers for a thorough mix. 1 gel volume of isopropanol was then added to the dissolved gel mixture prior to transfer onto a spin column. The column was then centrifuged at 17,900 x g for 1 minute and the flowthrough was discarded. Sequential addition of buffers followed by centrifugation was performed. 500µl of buffer QG was added and the filtrate was discarded following a spin at 17,900 x g for 1 minute. 750 µl of wash buffer PE is added and an additional centrifugation step is performed to remove any residual buffer. DNA was isolated from the column using a 10 µl volume of buffer EB (10 mM Tris-Cl, pH 8.5).

#### **2.3.5.2 Column purification**

Silica membrane-based purification of PCR products was performed using the MinElute PCR Purification Kit (Qiagen) microcentrifuge protocol. Five times the volume of buffer PB was added for each volume of PCR reaction. The microcentrifuge was calibrated for 17,900 x g for each spin to be performed. The DNA and buffer PB mixture was transferred to a spin column in a 2 ml collection tube and was subjected to a spin to bind the DNA. 750 µl of PE wash buffer was then added to the column followed by two rounds of centrifugation, with the eluted liquid discarded in each case. The DNA was eluted by adding 10 µl of EB to the

centre of the column followed by a 1 minute incubation and a spin (10 mM Tris-Cl, pH 8.5). The purified DNA eluate was stored at -18°C.

### **2.3.5.3 Magnetic beads clean up**

Large DNA amplicons generated by LR-PCR (section 2.3.1.3) were purified using magnetic beads that selectively bind to DNA. Equal volume of AMPure XP reagent (Beckman Coulter) was added to samples and the tubes were placed on a hula mixer and incubated for 5 minutes at room temperature. Samples were placed on a magnetic rack to pellet the DNA and the supernatant was discarded. The pellet was washed using 200 µl of 70% ethanol twice before decanting the supernatant. A brief spin of 1 minute at 17,900 x g was used to remove residual ethanol and the pellet was allowed to air dry for 2 minutes on the magnetic rack. The samples were removed from the magnetic rack and the pellet was resuspended in the desired volume of DNase free water. The tubes were placed back on the magnetic rack and the eluate was extracted and stored.

## **2.3.6 DNA quantification:**

### **2.3.6.1 NanoDrop™ spectrophotometer**

DNA concentration was measured using a NanoDrop™ 8000 or NanoDrop™ 1000 spectrophotometer (Thermo Fisher Scientific). DNase free water was used to initiate the instrument and clean the pedestals prior to sample loading. Software ND-8000 v2.1 or ND-1000 v3.5 (Thermo Fisher Scientific) was accessed on a computer for calibration using the appropriate blanks of DNase free water, EB buffer for Miniprep isolates or TE buffer for Maxiprep isolates. 2 µl of sample DNA was pipetted on to the sample pedestal and measured in ng/µl.

### **2.3.6.2 Qubit® fluorometer**

Quantification of DNA also used a Qubit v1.01 fluorometer. Samples were prepared according to Qubit™ dsDNA BR Assay Kit (Invitrogen) instructions. Master mix was made using a 1:200 dilution of BR reagent in BR buffer for each sample. Individual 200µl reactions consisted of 1 µl of DNA or 10 µl of supplied

standards. The fluorometer Instrument was calibrated using the supplied standards 1 and 2 generating a linear regression curve marking the detectable range of concentrations. Samples DNA concentration was determined after factoring in the 200x dilution factor.

### **2.3.7 Ethanol precipitation**

To precipitate DNA from solution, 1/10<sup>th</sup> volume of 3 M sodium acetate (pH 5.2) was added and mixed by flicking the tube. Subsequent addition of 2 ½ volume of ice cold 100% ethanol to DNA sample followed by a 60 minute incubation at -20°C to precipitate DNA. The mixture was centrifuged at 15,000 x g for 30 minutes at 4°C forming a pellet of DNA. The supernatant was discarded, and the pellet was washed with 0.5 ml of 70% ethanol. This was spun at 15,000 x g for 20 minutes at room temperature and the supernatant was decanted. Once air dried, the pellet was redissolved in the appropriate volume of TE buffer to achieve the desired concentration.

### **2.3.8 Sanger Sequencing**

PCR products were prepared using ExoSAP-IT™ reagent (Thermo Fisher Scientific) in a ratio of 5 µl of amplicon: 2 µl of reagent to enzymatically remove contaminants. The mix was incubated at 37°C for 15 minutes followed by 80°C for 15 minutes. Sequencing was carried out in 10 µl reactions consisting of 1.5 µl of BigDye™ Terminator v3.1 sequencing buffer (Applied biosystems), 0.75 µl of BigDye™ Terminator 3.1 (Applied biosystems), 1 µl of 1.6 µM primer (Sigma-Aldrich), 1 µl of ExoSAP-IT treated DNA and 5.75 µl of DNase free water. Thermocycle sequencing was carried out with an initial denaturation step at 96°C for 60 seconds followed by 25 cycles of 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 240 seconds. The temperature ramp rate selected was 1°C/second. After sequencing, DNA products were precipitated by the addition of 5 µl of 125 mM EDTA and 60 µl of 100% ethanol and mixed by inverting the tube. The solution was centrifuged at 3061 x g for 30 minutes at 20°C and the supernatant was decanted. In the next step, DNA was cleaned using 60 µl of 70% ethanol followed by a 15 minutes spin at 805 x g at 4°C. Once the supernatant was discarded, then DNA pellet was redissolved in 10 µl Hi-Di™ Formamide (Thermo

Fisher Scientific) and subjected to a 5 second pulse spin. DNA sequencing was performed by loading 10 µl on an ABI3130xl Genetic Analyser (Applied Biosystems) and DNA traces were visualised on 4Peaks (Nucleobytes).

### **2.3.9 ONT long read sequencing**

Long DNA fragments were sequenced using ONT sequencing for full representation of the sequence in a single read. Plasmids were linearised by restriction endonuclease digestion (section 2.3.2). Contaminants were removed from DNA using Magnetic bead clean up (section 2.3.5.3). Nucleic acid quantification was carried out by Qubit® fluorometry (Invitrogen) (section 2.3.6.2). When sequencing multiple DNA fragments in the same reaction, template DNAs were pooled into a single reaction mix prior to DNA library preparation with approximately equal molarity, to a total of 100 fmol (~500 ng) in 23.5 µl. DNA library preparation used the ONT Ligation Sequencing Kit (SQK-LSK110) for DNA end repair, deoxy-adenosine monophosphate tailing and adapter ligation as per manufacturer's instruction. The DNA library mixture (30 µl) was loaded onto a Flongle flow cell (R9.4.1) with >50 pores. Sequencing parameters were defined using minKNOW v21.11.7 (ONT) for a 24 hour sequencing duration at -180V. All ONT sequencing was performed on a MinMk1B (ONT) instrument.

## **2.4 Molecular cloning**

### **2.4.1 Gateway cloning**

Midigene splice assays were generated using Gateway® BP Clonase™ II Enzyme Mix and LR Clonase™ II Enzyme Mix. The strategy involves Lambda based recombination events between the insert and intermediate vectors. To generate the insert, primers were designed with additional sequence to generate amplicons flanked with the attB recombination sites. LR-PCR (section 2.3.1.3) was performed to generate large amplicons which were then purified using the QIAquick Gel Extraction Kit method (Qiagen) (section 2.3.5.1). Entry clones were generated using 10 µl reactions consisting of 15-150 ng of PCR product and 150 ng of donor vector (pDONR™221) made up to 10 µl with TE buffer (pH 8.0). 2 µl of BP Clonase™ II enzyme mix was added to the reaction, followed by a 3 hour



incubation at 25°C. The reaction was terminated by addition of 2 µg of Proteinase K solution followed by incubation at 37°C for 10 minutes. Entry clones were transformed into One Shot® OmniMAX 2 T1 Phage-Resistant Cells using heatshock transformation (section 2.4.4) and transformants were spread onto LB-agar plates with kanamycin at a final concentration of 50 µg/ml then incubated overnight at 37°C. Entry clones were picked as single colonies, then plasmid DNA was purified using the QIAprep Spin Miniprep Kit (Qiagen) method described in section 2.4.5.1.

Midigene splice vectors were generated using destination vector pCI-NEO-RHO (vector map in Figure 5.14). This vector was a gift from Dr Erwin Van Wyk (Radboud University Nijmegen) and is a gateway-adapted vector containing the RHO exons 3, 4, and 5 with the recombination sequences located between RHO exons 3 and 4 and between exons 4 and 5 (Sangermano et al., 2016). The splice vectors were generated in a 10 µl reaction consisting of 150 ng of entry clone and 150 ng of pCI-NEO-RHO made up to 10 µl with TE buffer (pH 8.0). 2 µl of LR Clonase™ II enzyme mix was added, then the mixture was incubated for 3 hours at 25°C. 2 µg of Proteinase K solution was added with subsequent incubation at 37°C for 10 minutes. Midigene vectors were then transformed into One Shot® OmniMAX 2 T1 Phage-Resistant Cells using heatshock transformation (section 2.4.4). Cells were plated onto LB-agar petri dishes containing ampicillin at a final concentration of 100 µg/ml, followed by an overnight incubation at 37°C. Plasmid maps of all the vectors utilised in Gateway cloning are provided in (Figure 5.14).

#### **2.4.2 Site Directed Mutagenesis**

Introducing point mutations in Gateway cloning constructs was achieved using the QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies) to generate isogenic controls. Mutagenic primers were designed via the web link <https://www.agilent.com/store/primerDesignProgram.jsp> according to the criteria outlined in section 2.4.3 Primers were manufactured and purified by polyacrylamide gel electrophoresis (PAGE) through Sigma-Aldrich. Reactions were carried out in 50 µl consisting of 1X reaction buffer, 10-50 ng of plasmid DNA, 125 ng of each mutagenic primer, 1 µl of dNTP mix, 3 µl of QuikSolution

reagent and 2.5 U PfuUltra High Fidelity DNA polymerase and made up to the correct volume with DNase free water. Amplification was achieved with an initial denaturation step at 95°C for 1 minute, followed by 18 cycles consisting of 95°C for 50 seconds, 60°C for 50 seconds and 68°C for 1 minute/kb of plasmid, then a final extension step at 68°C for 7 minutes. Restriction digestion of the methylated parental plasmid (unmutated) was achieved by adding 10 units of DpnI and incubating at 37°C for 2 hours. DpnI treated DNA was subsequently transformed into XL10-Gold ultracompetent cells (Agilent Technologies) (section 2.4.4). Transformed cells were plated on LB-agar using the desired antibiotics for each construct and incubated overnight at 37°C.

### **2.4.3 Primer design for site directed mutagenesis**

Mutagenic primers were designed to target midgene splicing constructs to generate matching controls for variants being investigated. These oligonucleotides were designed to be complementary to the target sequence in the insert but including the desired sequence change. Primers were designed using the QuikChange Primer Design Program (<https://www.agilent.com/store/primerDesignProgram.jsp>). Design criteria were set to restrict primers to 25-45 bp in length, with GC content >40%,  $T_m$  of  $\geq 78^\circ\text{C}$  and to have a terminal guanine or cytosine base. Selected primers were manufactured by Sigma-Aldrich and were purified using PAGE.

### **2.4.4 Bacterial transformation**

DNA was transformed into XL10-Gold ultracompetent cells (Agilent Technologies) or One Shot<sup>®</sup> OmniMAX<sup>™</sup> 2 T1 Phage-Resistant Cells (Thermo Fisher Scientific) using heat shock transformation. To increase the transformation efficiency of XL10-Gold ultracompetent cells (Agilent Technologies), 2  $\mu\text{l}$  of beta-mercaptoethanol ( $\beta$ -ME) was added to thawed cells on ice and these were then incubated for 10 minutes. The target DNA was then added and further incubation on ice for 30 minutes. Heat shock was performed using a 30 second exposure to heat in a 42°C water bath, followed immediately by incubation on ice for 2 minutes. 0.5 ml of preheated NZY+ broth was added to transformed XL10-Gold

ultracompetent cells or 250 µl of super optimal catabolite (SOC) medium (2% w/v bacto-tryptone, 0.5% w/v bacto-yeast extract, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl<sub>2</sub>, 20 mM glucose) was added to transformed One Shot<sup>®</sup> OmniMAX<sup>™</sup> 2 T1 Phage-Resistant Cells and incubated for 1 hour at 225 rpm and 37°C in a bacterial shaking incubator. This cellular suspension was then spread on LB-agar plates with the appropriate antibiotics of either kanamycin or ampicillin at a final concentration of 50 µg/ml and 100 µg/ml, respectively. The LB-agar plates were incubated overnight at 37°C.

## **2.4.5 Plasmid isolation and purification**

### **2.4.5.1 Plasmid Miniprep**

Plasmid DNA was purified from bacterial culture using the QIAprep Spin Miniprep Kit (Qiagen). 1 ml of bacterial culture was pelleted at 17,900 x g for 3 minutes, supernatant containing the growth media was removed and the pellet was resuspended in 250 µl of buffer P1. Cells were then subjected to alkaline lysis for 4 minutes using 250 µl of buffer P2. Neutralisation buffer N3 was added to the mixture using a 350 µl volume and a spin was performed at 17,900 x g for 10 minutes to pellet the cell debris. The supernatant was transferred to a silica membrane-based spin column which was centrifuged at 17,900 x g for 1 minute to capture the DNA. The column was washed with 0.5 ml of buffer PB to remove endonucleases, then with 0.75 ml of buffer PE to eliminate salts. A spin at 17,900 x g for 1 minute was performed after the addition of each buffer. A final spin was conducted to eliminate any residual buffer. Finally, DNA was eluted using 40 µl EB and quantified using by Nanodrop<sup>™</sup> spectrophotometer (section 2.3.6.1).

### **2.4.5.2 Plasmid Maxiprep**

When plasmid DNA for mammalian transfection were required, the plasmids were purified using an EndoFree<sup>®</sup> Plasmid Maxi Kit (Qiagen). 250 ml of bacterial culture was pelleted at 6000 x g for 15 minutes at 4°C. The supernatant was discarded and the pellet was resuspended in 10 ml of buffer P1. 10 ml of lysis buffer P2 was then added and mixed by inverting the tube, followed by a 5 minute incubation at room temperature. 10 ml of neutralisation buffer P3 was added to

the mixture to prevent further lysis. The cell debris was pelleted by centrifugation at 6000 x g for 30 seconds at room temperature. The lysate was transferred to a QIAfilter Cartridge and 2.5 ml of buffer ER was added then the mix was incubated on ice for 30 minutes. Filtered lysate was added to an activated QIAGEN-tip for DNA capture. The column was washed twice with 30 ml of buffer QC then the DNA was eluted using 15 ml of buffer QN. Ethanol precipitation was performed by adding 10.5 ml of isopropanol to the eluted DNA. After mixing, the solution was centrifuged at 15,000 x g for 30 minutes at 4°C and the supernatant was discarded. The pellet was washed using 5 ml of 70% ethanol and centrifuged at 15,000 x g for 10 minutes at 4°C. The supernatant was decanted, and the pellet was allowed to air dry for 10 minutes to eliminate residual ethanol. The pellet was redissolved in 400 µl of TE buffer and quantified using by Nanodrop™ spectrophotometer (section 2.3.6.1).

#### **2.4.6 Cryopreservation of transformants**

Glycerol stocks were prepared for long-term storage of bacterial cells containing desired plasmids. 1 ml of a bacterial subculture was transferred to a cryovial along with an equal volume of 50% glycerol diluted in deionised water. Once mixed thoroughly, the glycerol stocks were stored at -80°C. To recover bacteria, a sterile loop was used to scrap the glycerol surface, and this was subsequently streaked onto a LB-agar plate with the appropriate antibiotics.

### **2.5 Tissue Culture**

#### **2.5.1 Cell lines and cell culture**

ARPE-19 cells (ATCC) are spontaneously immortalized, epithelial cells derived from normal human retinal pigmented epithelia. These were cultured in Dulbecco's Modified Eagle Medium (DMEM) F12 Nutrient Mixture (Gibco) supplemented with 10% foetal bovine serum (FBS), 1% Glutamax 100X (Gibco) and 0.4% penicillin G-streptomycin. Cells were split upon reaching 80% confluency using 0.5% trypsin-EDTA (Gibco) and phosphate buffer saline (PBS; Gibco) for wash. Cells were split into 1/5 or 1/10 in a T25 flask and incubated at 37°C with 5% CO<sub>2</sub> in an MCO-20AIC cell culture incubator (Sanyo). Cell culturing

tasks were performed in NuAire Labgard 437 ES Class II Biosafety Cabinets while adhering to aseptic techniques.

Routine passaging entailed removal of the growth media and a wash using 2 ml of PBS performed twice. 1.5 ml of trypsin was then added to cover the entire 25 cm<sup>3</sup> surface of the container and the flask was incubated at 37°C for 5 minutes. Light microscopy at 10X was used to confirm cells being detached. 5 ml of culture media was added to inactivate the trypsin and the suspended cells were transferred to a Corning centrifuge tube (Thermo Fisher Scientific) and centrifuged at 100 x g for 5 minutes at room temperature. The supernatant was discarded and the cell pellet was then resuspended in an equal starting volume of fresh growth media. Depending on the split, cells were seeded accordingly into a new T25 flask with the appropriate volume of growth media.

### **2.5.2 Storage of cell lines**

Cells were cultured in a T75 flask and trypsinised upon reaching 80% confluency. Pelleted cells were resuspended in a solution comprising 90% FCS and 10% DMSO to maintain cellular integrity when frozen. The cell-solution mix was aliquoted into 1.5 ml cryovials and placed in a Mr. Frosty (Sigma-Aldrich) freezing container that is insulated with 100% isopropanol for cryopreservation. Cells contained in the Mr Frosty were frozen in a -80°C U725 freezer (New Brunswick Scientific) prior to transfer to liquid nitrogen for long term storage.

### **2.5.3 Cell counting**

Automated cell counting utilised a Countess 3 automated cell counter (Thermo Fisher Scientific) to estimate cell density and the proportion of live cells. Cells grown to an 80% confluency in a T75 flask were trypsinised and resuspended in 5 ml of fresh growth media. Slides were prepared using 10 µl of the cell suspension mixed with an equal volume of 0.4% Trypan Blue dye. After 30 seconds incubation at room temperature, a single slide was inserted into the Countess 3 for analysis.

#### **2.5.4 RNA extraction**

Total RNA extraction from cells was performed using the RNeasy Plus Minikit (Qiagen) according to the manufacturer's instructions. The number of live cells (not exceeding  $1 \times 10^7$ ) was confirmed using the Countess 3 (section 2.5.3). Cells were pelleted using a centrifuge calibrated at  $8000 \times g$  for 15 minutes at  $4^\circ\text{C}$ . Cell lysis was performed using  $600 \mu\text{l}$  of buffer RLT plus followed by vortexing for 30 seconds. The lysate was transferred to a gDNA Eliminator spin column and centrifuged at  $8000 \times g$  for 30 seconds at room temperature to remove the cell debris. The filtrate was retained and  $600 \mu\text{l}$  of 70% ethanol was added and mixed via pipetting. RNA capture required transferring the filtrate to an RNeasy spin column and centrifuging at  $8000 \times g$  for 15 seconds at room temperature. The next steps consisted of a sequential series of buffer additions (buffer RW1 followed by RPE) with subsequent centrifugation to discard the flowthrough between each buffer. RNA was eluted using  $30\text{-}50 \mu\text{l}$  of nuclease free water and a final spin at  $8000 \times g$  for 1 minute at room temperature. RNA concentration was determined using the NanoDrop™ spectrophotometer (section 2.3.6.1) and the extracted RNA was immediately stored at  $-80^\circ\text{C}$ .

#### **2.5.5 RNA quality assessment**

RNA quality was determined using the TapeStation 2200 (Agilent technologies) automated electrophoresis system.  $2 \mu\text{l}$  of RNA extract was diluted in  $1 \mu\text{l}$  of High Sensitivity RNA ScreenTape Sample Buffer. The solution was vortexed and centrifuged at  $805 \times g$  for 1 minute. High sensitivity RNA screen tape (Agilent technologies) was loaded into the instrument for reliable separation of total RNA. The 2200 TapeStation analysis software v A. 01.05 was utilised to introduce a virtual ladder in the analysis. RNA integrity number equivalent ( $\text{RIN}^e$ ) was used as a measure of RNA quality.

#### **2.5.6 Transfection: Nucleofection**

Transgenic ARPE-19 cell lines were generated using a Nucleofector Kit V (Lonza/Amaxa) and Nucleofector Device v3 (Lonza/Amaxa) according to the manufacturer's instructions. DMEM culture media used was supplemented with

HAM F12 (2.5 mM glutamine, 1.5 g/L NaHCO<sub>3</sub> (90%) and FBS (10%). Centrifuge parameters were adjusted to 90 x g for 10 minutes to harvest the required number of cells after trypsinisation. 1x10<sup>6</sup> live cells were resuspended in 100 µl of Nucleofector Solution. 2 µg of plasmid DNA was transfected into the 100 µl cell solution using Nucleofector program NX-001. 500 µl of growth media was added to the transformed cells and the cells were subsequently cultured in a 6 well plate at a final volume of 1.5 ml per well using conditioned media containing ARPE-19 growth factors and signal peptides to aid in cell recovery post transfection. The conditioned media was obtained from an ARPE-19 culture after 24 hours incubation in fresh growth media of DMEM F12 (Gibco) with 10% FBS, 1% Glutamax 100X (Gibco) and 0.4% penicillin G-streptomycin. The cells were incubated at 37°C and 5% CO<sub>2</sub> for 24 hours.

### **2.5.7 Microscopy**

Cellular morphology and confluency were observed in culture flasks using a CKX41 light microscope (Olympus) at different magnifications of 4X, 10X, 20X, 40X and 100X. Fluorescently labelled cells expressing green fluorescent protein (GFP) were visualised under EVOS FL (Thermo Fisher Scientific) fluorescent microscope at various magnification stages and using overlay imaging.

## **2.6 Bioinformatics**

All the genomic coordinates presented in this thesis have been according to GRCh38 unless otherwise then the genomic build would be stated.

### **2.6.1 Integrated Genomics Viewer (IGV)**

The IGV desktop application v2.8.13 was used to render and interact with BAM and VCF files for comparative purposes to the reference genome (Robinson et al., 2011). With this high-performance tool a user can examine the sequences of individual reads with indications for different variant types. Both NGS short-read and ONT long-read sequencing BAM files were inspected in IGV to validate variants, and to detect SVs not called by the standard bioinformatic pipeline.

Plasmid sequence alignments were also inspected in IGV, using custom references containing both the insert and plasmid's bacterial backbone.

### **2.6.2 4Peaks (nucleobytes)**

This is a software used to visualise the electropherogram traces of Sanger sequencing. Innovative options are also included to edit the nucleotide trace for correction purposes and to reverse the sequence to display the traces on the opposite DNA strand. The supported file formats are not limited to ab1, SCF, ZTR, EXP, BIO, and CTF.

### **2.6.3 Ensembl genome browser**

The Ensembl genome browser (<https://www.ensembl.org/index.html>) was used to navigate *Homo sapiens* genome build GRCh37 and GRCh38 (Cunningham et al., 2022). This is an open access platform providing a visual representation of the genome annotated with protein coding genes, pseudogenes and noncoding RNA genes. Additional genetic information was inferred from target loci through selected tracks for gene regulation, corresponding chromosomal band and gnomAD (section 2.6.4) (Karczewski et al., 2020) reported variants. Exploring target genes provided access to nucleic acid and protein sequence for downstream analysis.

### **2.6.4 Genome Aggregation Database (gnomAD)**

The gnomAD database (v4.1.0) contains variant information drawn from 76215 genomes and 730947 exomes datasets. These are derived from control groups and from studies of late-onset diseases, and they specifically exclude severe paediatric diseased individuals, making this resource relevant for the study of early onset inherited retinal diseases. Information such as depth of coverage is displayed in gnomAD along with the entire list of variants reported in a particular gene. Allele frequencies of different variants were extracted from gnomAD v4.1.0 (GRCh38) (<https://gnomad.broadinstitute.org/>) based on the matching ethnicity for the proband (Chen et al., 2022) and gnomAD v2.1.1 was used in filtering variants from the raw VCF file and in variant interpretation.



### **2.6.5 Online Mendelian Inheritance in Man (OMIM)**

OMIM (<https://www.omim.org/>) is a reliable source of genetic and phenotypic data of medically relevant genes of both Mendelian and non Mendelian inheritance (Amberger et al., 2015). A brief overview of the disorder is provided along with comprehensive information on the causative gene in terms of function, expression and genetic mapping with additional insights from cloning studies and animal models.

### **2.6.6 Encyclopedia of DNA elements (ENCODE)**

The ENCODE project (<https://www.encodeproject.org/>) is a repository of known functional DNA elements in the human and mouse genomes (Consortium et al., 2020). The project relies on epigenomic signatures of H3K4me3, H3K27ac, CTCF and DNase to characterise candidate cis-regulatory elements (CRE). These CREs include promoter like signatures (PLS), proximal enhancer like signatures (pELS), distal enhancer like signatures (dELS), DNase-H3K4me3 and CTCF-only. In attempts to characterise intronic variants detected in our cohort, the ENCODE database was accessed to identify overlapping cis regulatory elements.

### **2.6.7 The Human Protein Atlas**

This database (<https://www.proteinatlas.org/>) provides the expression profiles of specific mRNAs and proteins in different organisational structures of tissues, cells and subcellular structures (Ponten et al., 2008). This initiative endeavours to characterise the temporal (time dependent) and spatial (tissue specific) expression of all human proteins through collating data from proteomics, transcriptomics, and immunohistochemistry studies. Both the global (entire body) mRNA expression pattern of target and protein localisation at subcellular levels were accessed to characterise novel candidate genes underlying FH.

### **2.6.8 Leiden open variation database (LOVD)**

LOVD (<https://www.lovd.nl/>) v3.0 is an open access database for genome sequence variants that are curated and maintained by experts. The database

currently holds 911342 and these are reported in compliance with Human Genome Variation Society (HGVS) nomenclature and may include the ACMG classification, variant segregation and available references from the published literature. The clinical data is also displayed which includes the diagnosis and symptoms associated with the variants.

### **2.6.9 Primer design**

Based on the genomic coordinates of targets, primers were designed using the Primer3Web (<http://primer3.ut.ee/>) and the ExonPrimer (<https://ihg.helmholtz-munich.de/ihg/ExonPrimer.html>) tools (Untergasser et al., 2012). The design parameters were set to select primers of 18-30 bp in length, with 40-60% GC content,  $T_m$  ranging from 55-65°C and within 5°C of each other, and avoiding runs of 4 or more identical bases or dinucleotide repeats. Oligonucleotides were subjected to physiochemical assessment using PrimerBLAST (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/>) (Ye et al., 2012) and SNPCheck V3 (<https://genetools.org/SNPCheck/snpcheck.htm>) for design validation in terms of self-complementarity and presence of polymorphisms in the primer binding site. Selected primers were ordered from Sigma-Aldrich and prepared into aliquots of 100 µM and 10 µM concentration using DNase free water.

### **2.6.10 sgRNA designs for CRISPR-Cas9 mutagenesis**

Gene editing by CRISPR-Cas9 mediated SDM required the use of a single guide RNA (sgRNA) that is homologous to the target sequence and this sgRNA guides the Cas9 endonuclease to selectively cleave the DNA at the target site. sgRNAs for use with conventional Cas9 were designed using the Invitrogen™ TrueDesign™ Genome Editor (<https://apps.thermofisher.com/apps/genome-editing-portal/#/select/experiment>). For gene correction or knock-in experiments, silent blocking mutations were introduced at the protospacer adjacent motif (PAM) in the single stranded oligodeoxynucleotide (ssODN) donor template. sgRNAs were designed to satisfy the following conditions; a 20 bp oligonucleotide complementary to the target, with a PAM sequence (NGG) on the 3' side of the non-target DNA strand, the edit site ~6 bp from cut site, the ssODN template and

sgRNA targeting different strands and using an asymmetrical donor template (Richardson et al., 2016). Other recommended modifications included phosphorothioate on the terminal 3 bases of the 3' and 5' end (Sanjurjo-Soriano et al., 2020). sgRNAs designed were assessed for targeting efficiency using IDT CRISPR Cas9 Design Checker ([https://eu.idtdna.com/site/order/designtool/index/CRISPR\\_SEQUENCE](https://eu.idtdna.com/site/order/designtool/index/CRISPR_SEQUENCE)) and Invitrogen™ TrueDesign™ Genome Editor (<https://apps.thermofisher.com/apps/genome-editing-portal/#/select/experiment>).

The edit site, PAM and the region downstream of a mutation were screened for potential splice sites and exonic splicing enhancers (ESEs) to prevent edits that may disrupt splice consensus sequences. This was facilitated using ESE finder 3.0 (<https://esefinder.ahc.umn.edu/cgi-bin/tools/ESE3/esefinder.cgi>) (Cartegni et al., 2003) and NN splice v0.9 ([https://fruitfly.org/seq\\_tools/splice.html](https://fruitfly.org/seq_tools/splice.html)) (Reese et al., 1997).

### 2.6.11 Autozygosity mapping

The identification of novel candidate genes in pedigrees with autosomal recessive inheritance and consanguineous relationships used autozygosity mapping (Carr et al., 2013). This strategy narrows the disease locus containing the causative gene to regions that are identical by descent due to a common ancestor. A single bioinformatic tool, AutoMap (<https://automap.iob.ch/>) (Quinodoz et al., 2021) was utilised to identify regions of autozygosity in VCFs from both WES and WGS.

### 2.6.12 Restriction mapping

When planning restriction digestion of amplicons or plasmid (section 2.3.2), the NEBcutter v3.0.17 website (<https://nc3.neb.com/NEBcutter/>) was used to create a graphical representation of the substrate DNA annotated with recognition sites for different restriction enzymes. Agarose gel simulations could then be produced, revealing the expected fragment sizes and the predicted band

migration pattern. SnapGene viewer v6.2.1 (Dotmatics) was used to generate comprehensive plasmid schematics annotated with different plasmid features for consideration when selecting the appropriate restriction enzymes.

### **2.6.13 Variant interpretation tools**

The following is a list of the tools and software packages used to assess variants, filter them and prioritise any with the potential to disrupt gene, mRNA or protein function to imply pathogenicity. All variants were reported in accordance to the HGVS nomenclature and have been classified according ACMG guidelines.

#### **2.6.13.1 Grantham Matrix**

This is scoring system for missense variants that is based on evolutionary conservation of amino acids to predict the severity of amino acid substitutions (Grantham, 1974). A higher score corresponds to a more biochemically dissimilar amino acid change that is not conserved under protein evolution and is considered deleterious. Grantham score classification comprises conservative substitution (0-50), moderately conservative (51-100), moderately radical (101-150) and radical (151<).

#### **2.6.13.2 Sorting Intolerant From Tolerant (SIFT)**

SIFT (<https://sift.bii.a-star.edu.sg/>) predicts the effects of nonsynonymous amino acid substitutions on protein function based on evolutionary conservation (Vaser et al., 2016). Multiple sequence alignments were used to generate probability scores at each position of the queried peptide sequence. A SIFT score of <0.05 suggest that the variant was deleterious, while variants with scores  $\geq 0.05$  were considered to be likely tolerated.

#### **2.6.13.3 Polymorphism Phenotyping v2 (PolyPhen-2)**

Polyphen-2 (<http://genetics.bwh.harvard.edu/pph2/>) predicts amino acid substitution effects based on sequence homology and structural considerations (Adzhubei et al., 2010). The algorithm uses a probabilistic model known as Naïve

Bayes classifier to predict pathogenicity with varying degrees of confidence. For the diagnosis of Mendelian disease, HumVar score was used for better differentiation between disease causing missense variants and common single nucleotide polymorphisms. Polyphen-2 uses a score range between (1) to imply probably pathogenic and (0) for benign.

#### **2.6.13.4 Align-Grantham Variation Grantham Deviation (GVGD)**

Align-GVGD ([http://agvgd.hci.utah.edu/agvgd\\_input.php](http://agvgd.hci.utah.edu/agvgd_input.php)) relies on Grantham matrix and multiple sequence alignments to provide a verdict on potential pathogenicity of amino acid substitutions. Align-GVGD categorises variants into groups in which class C65 is the most pathogenic and class C0 is the least pathogenic. (Tavtigian et al., 2006).

#### **2.6.13.5 Protein Analysis Through Evolutionary Relationships (PANTHER)**

This webserver interrogates phylogenetic trees of protein families, orthologs and paralogs (<http://www.pantherdb.org/tools/csnpScoreForm.jsp>), and uses the evolutionary preservation of amino acids as an indicator for the likely functional impact of a genetic variant that alters them (Thomas et al., 2022). This approach is known as specific evolutionary preservation and is considered superior to simple amino acid conservation as it assesses ancestral changes in the lineage specific to the protein of interest (Tang and Thomas, 2016). The preservation time was interpreted as, >450 million years (my) for probably damaging, <450 my but >200 my for possibly damaging and <200 my for probably benign.

#### **2.6.13.6 Combined Annotation Dependent Depletion (CADD)**

The CADD (<https://cadd.gs.washington.edu/snv>) is a score which incorporates data on pathogenicity of SNVs and INDELS from many different sources using a machine learning model to simulate variants at each position of the reference genome (Rentzsch et al., 2019). CADD annotations utilise various stand-alone tools to explore amino acid conservation, physiochemical alterations, epigenetic modifications, and gene regulatory elements to consolidate a Phred like score for interpretation. The scaled CADD score threshold was adjusted to 15 which

indicates the variant being in the 5% most pathogenic changes in the human genome.

#### **2.6.13.7 MutationTaster**

Gathering further data on the functional consequences of a variant was facilitated by MutationTaster (<https://www.mutationtaster.org/>) (Schwarz et al., 2014). The overall pathogenicity prediction with this webserver was provided by a machine learning algorithm known as Bayes classifier and it uses allele frequencies and disturbances to regulatory features, splice sites and amino acid conservation to make judgements on whether the variant is classified as disease causing or a polymorphism.

#### **2.6.13.8 SpliceAI**

SpliceAI (<https://spliceailookup.broadinstitute.org/>) uses deep learning to predict whether synonymous changes, nonsynonymous substitutions or INDELs are likely to affect splice acceptor, donor, enhancer or inhibitor sites in a way that alters splicing (Jaganathan et al., 2019). The webserver enables the identification of splicing consequences across large genomic distances from the variant site through neural networks model and was used to assess the effects of mainly deep intronic variants in splicing. SpliceAI generates a score estimating the probability of a splicing event to occur and a conservative threshold of 0.5 was selected to imply interference with a splicing event.

#### **2.6.13.9 varSEAK**

Noncoding variants were also evaluated for their potential to mediate aberrant splicing events using the VarSEAK online webserver (<https://varseak.bio/index.php>). This was used to determine whether candidate variants disrupted canonical splice sites or activated a potential cryptic splice site. VarSEAK generates a score of -100 to 100 to determine whether the splice site is functional. Delta score calculates the difference in the score for the splice site on the wildtype and variant sequence. The scores are used to generate a classification of class 1-5 with class 1 denoting no splicing effect while class 5

confidently predicts the occurrence of a splicing event. The webserver also integrates MaxEntScan scores, a complementary splice site prediction tool that uses maximum entropy principle to predict the likelihood of a splicing event to occur (Yeo and Burge, 2004).

#### **2.6.13.10 Ensembl Variant Effect Predictor (VEP)**

Ensembl VEP is a comprehensive suite that allows the user to characterise variants based on phenotypic associations (ClinVar), allele frequencies (gnomAD and 1000 Genomes), protein domains affected (InterProScan), overlapping regulatory elements (ENCODE) and *In-silico* pathogenicity prediction packages (SIFT, PolyPhen-2, CADD and Splice AI). Prioritising variants using the Ensembl VEP ([https://www.ensembl.org/Homo\\_sapiens/Tools/VEP](https://www.ensembl.org/Homo_sapiens/Tools/VEP)) tool provided additional auxiliary information for consideration in determining which variants were most likely to cause disease (McLaren et al., 2016). VEP also provides access to UTRannotator, a plug in that annotates variants in non-coding regulatory regions that can influence gene expression. This program identifies small variants (1-5 bp) in the 5' untranslated region (UTR) that disrupt or create uORFs (upstreamORFs) though altering or creating new start and stop codons (Zhang et al., 2021).

#### **2.6.14 Protein sequence alignments**

Assessing evolutionary conservation of amino acids was made possible using Homologene (<https://www.ncbi.nlm.nih.gov/homologene>). The webserver is available on National Centre for Biotechnology Information (NCBI), which provides sequence alignments against a default set of orthologs. Custom alignments utilised Needle (EMBOSS) ([https://www.ebi.ac.uk/Tools/psa/emboss\\_needle/](https://www.ebi.ac.uk/Tools/psa/emboss_needle/)) for pairwise and global sequence alignment using Needleman-Wunsch algorithm (Needleman and Wunsch, 1970). For more than 2 sequence alignments, Clustal Omega (<https://www.ebi.ac.uk/Tools/msa/clustalo/>) was employed (Sievers et al., 2011).

## 2.6.15 Computational modelling of proteins and variant simulation

### 2.6.15.1 Protein homology modelling

Protein homology modelling was carried out using default parameters on Phyre2 (<http://www.sbg.bio.ic.ac.uk/phyre2>) (Kelley et al., 2015), I-TASSER (<https://seq2fun.dcmf.med.umich.edu//I-TASSER/>) (Yang and Zhang, 2015) and SWISS-MODEL (<https://swissmodel.expasy.org/>) to generate three candidate SLC38A8 models (Waterhouse et al., 2018). These web servers rely on alignment of the queried peptide sequence to a homologous pre-existing template, followed by annotating the likely side chains conformations (rotamers). RCSB-Protein Data Bank pairwise structure alignment (<https://www.rcsb.org/alignment>) was performed on all three models using Rigid body alignment for a sensitive structural comparisons amongst closely related models (Berman et al., 2000). The best candidate was selected for further comparison against an existing reference SLC38A8 protein model (A6NNN8) on AlphaFold protein structure database (<https://alphafold.ebi.ac.uk/>) using SWISS-MODEL structure assessment tool (<https://swissmodel.expasy.org/assess>) and comparison tool (<https://swissmodel.expasy.org/comparison/>). Quality assessment was performed using ProTSAV (<http://www.scfbio-iitd.res.in/software/newProsav/modules.jsp>), a webserver which employs various stand-alone tools for quality control and consolidates the result into a single score of higher accuracy and sensitivity (Singh et al., 2016). A representative model was selected based on ProTSAV score of  $\leq 5$  Å RMSD (Root Mean Square Deviation) to imply a suitable model for studying the effects of mutations. Molecular imaging was performed using UCSF Chimera v1.16 to render and manipulate 3D protein models (Pettersen et al., 2004).

### 2.6.15.2 Variant Simulation

Predicting the effects of missense variants required computational simulations on a generated protein 3D model to explore biochemical, and structural consequences. *In silico* mutagenesis of target residues was achieved using UCSF Chimera prior to computational analysis. Changes in physiochemical parameters including hydrophobicity mapping and electrostatic potentials were evaluated via UCSF Chimera. The effects of variants on protein stability, residue



exposure and solvent accessibility were predicted by Site Directed Mutator (<http://www-cryst.bioc.cam.ac.uk/~sdm/sdm.php>) (Worth et al., 2011). Gauging structural changes on the protein model was facilitated by Missense3D (<http://missense3d.bc.ic.ac.uk/~missense3d/>) (Ittisoponpisan et al., 2019). Establishing amino acid conservation utilised the Constraint-based Multiple Alignment Tool (COBALT) ([https://www.ncbi.nlm.nih.gov/tools/cobalt/re\\_cobalt.cgi](https://www.ncbi.nlm.nih.gov/tools/cobalt/re_cobalt.cgi)) for multiple sequence alignment of SLC38A8 orthologs and the ConSurf server ([https://consurf.tau.ac.il/consurf\\_index.php](https://consurf.tau.ac.il/consurf_index.php)) to integrate conservation scores into the protein models (Ashkenazy et al., 2010).

### **2.6.16 Statistical Package for Social Sciences (SPSS)**

Statistical validation of biological data was carried out using the SPSS v27 software package to perform statistical tests comprising the T-test, Chi-squared and Fisher's exact test to investigate the significance of observed associations between two variables. A probability value (p-value) of  $\leq 0.05$  is interpreted as being statistically significant. This software suite provides a variety of tests to handle different variables that are measurable (scale), categorical (nominal) and ordered (ordinal) and can provide descriptive statistics to evaluate data distributions.

## **2.7 The 100KGP workflows**

All of the following analysis has been performed using software tools or scripts within the GEL research environment (<https://re.extge.co.uk/ovd/>).

### **2.7.1 Targeted analysis of SLC38A8 cohort in the rare disease cohort**

The Main Programme v11 dataset available in the Labkey application v23.7 was selected for a gene specific analysis to detect either biallelic or monoallelic variants in SLC38A8. Participants whom already received a molecular diagnosis were discarded through inspection of Labkey GMC exit questionnaire. Proband were filtered by the phenotype and were retained if they are reported with Human Phenotype Ontology (HPO) terms suggesting retinal defects. These are the

most common HPO terms reported in the Data Discovery application v6.0 for the specific disease categories assigned to relevant IRDs, which are developmental macular and foveal dystrophy, inherited macular dystrophy, LCA or early onset retinal dystrophy and rod cone dystrophy. The list of the HPO terms used are available in Appendix A section 8.1.3. Only the monoallelic *SLC38A8* cohort was further refined to exclude participants with a benign synonymous or splice region variant ( $\geq 3$  bp) using CADD, SpliceAI and Varseak to inform of any predicted consequences on splicing.

All *SLC38A8* variants (Biallelic and monoallelic) detected in the remaining probands were annotated and filtered at  $\leq 0.1\%$  for heterozygous and homozygous variants, or  $\leq 2\%$  for compound heterozygous variants by the Exomiser workflow. These variants were subjected to *In silico* pathogenicity assessment and allele frequency assessment using gnomAD and the 100KGP database. The 100KGP annotations by Exomiser score and Tiering classification provided supplementary evidence in variant interpretation. Variant interpretation of monoallelic variants was facilitated by alternative scores that are independent of the mode of inheritance. The Variant score assessed the degree of pathogenicity for the variants identified and the Gene-Pheno score evaluates the participants phenotype for similarity to *SLC38A8* related FH. All variants were inspected on the BAM file to eliminate artefactual calls and each variant analysed was reported using the HGVS nomenclature and classified by the ACMG criteria.

### **2.7.2 Analysis of participants with a confirmed FH diagnosis**

The main programme v18 dataset was filtered based on the HPO terms for “foveal hypoplasia, hypoplasia of the fovea and aplasia of the fovea”. Each participant was inspected using the gmc exit questionnaire on Labkey to eliminate participants that have already received a genomic diagnosis. Variants were filtered at  $\leq 0.1\%$ , or  $\leq 2\%$  for compound heterozygous variants in gnomAD and the 100KGP through the Exomiser workflow. The variants detected were filtered for 52 FH genes based on the mode of inheritance (Section 8.1.1). Biallelic variants were retained if detected in 39 autosomal recessive FH genes. Monoallelic variants were filtered for 11 autosomal dominant or 39 autosomal

recessive FH genes. Variants on the X chromosome were restricted to 5 X linked FH genes. The variants were characterised using computational tools to inform pathogenicity and using population frequencies to deduce whether the variant is rare. The variant score served as supplementary evidence that confirms the pathogenicity predictions and aids in interpretation of population frequencies. All the variants were categorised based on the ACMG classification.

### **2.7.3 Haplotype analysis using whole genome VCF files**

Haplotype analysis was performed on participants suspected with a founder variant. VCF files belonging to each participant was accessed using the `genome_file_paths_and_types` option on LabKey v23.7 (accessed on 17/02/2021) and the genomic coordinates were adjusted to GRCh38 to facilitate comparative analysis of haplotypes. The locus surrounding the target variant in all three patients was examined on IGV browser v2.15.4. The analysis excluded variant calls that had been filtered out due to low genotyping quality (GQX). This is a Phred scaled confidence metric for genotype assignment based on variant site annotation. Only variants of high confidence were retained and each variant was subsequently investigated using the participants BAM file to eliminate artefactual variants. Allele frequencies of the relevant ethnicity to the participants were derived from gnomAD. The alleles presented in the haplotype were encrypted to comply with the publication policy of GEL.

### **2.7.4 Gene specific analysis of non coding and coding variants in the rare disease cohort**

The Gene-Variant Workflow v1.6 provided in the GEL research environment was utilised to aggregate *SLC38A8* variants including intronic variants from each participant's VCF file (Appendix B: section 8.2.1). Connecting to the GEL high performance computing (HPC) cluster was achieved through a unique GECIP address (`mderar@corp.gel.ac@phpgridzlogn00N.int.corp.gel.ac`) to borrow the computational power required to analyse the entire rare disease cohort of 73512 genomes. The parameters requested are 2 processing units (CPU cores) on the same node (server) with 10 Gb of memory (RAM) each. The 100K Gene Variant Filter ([https://github.com/sunaynabest/filter\\_100K\\_gene\\_variant\\_workflow](https://github.com/sunaynabest/filter_100K_gene_variant_workflow)) was

applied to eliminate variants with a MAF >0.002 (0.2%) in gnomAD and the 100KGP. Intronic variants were evaluated by CADD, SpliceAI and varSEAK to determine impact on splicing and using UTRannotator to detect variants in the 5' untranslated region (UTR) that could disrupt the open reading frame (ORF). The ENCODE project database (section 2.6.6) was used to assess if the intronic and the intergenic variants overlap CREs.

### **2.7.5 Gene specific analysis of SVs in the rare disease cohort**

The already generated output files containing annotated SV's by SVRare (Yu et al., 2022) for FH genes in 71408 participants were made accessible for a gene specific analysis in the 100KGP research environment ([/re\\_gecip/shared\\_allGeCIPs/JingYu-SV-query/protein\\_coding\\_genes\\_2021-11-17](/re_gecip/shared_allGeCIPs/JingYu-SV-query/protein_coding_genes_2021-11-17)). SVs identified in participants whom have already received a genomic diagnosis through the 100KGP were eliminated. SVs were also discarded for being too common if reported with an allele count  $\geq 60$  unless the FH gene follows an X linked inheritance then allele count is adjusted to  $\geq 120$  since female carriers of pathogenic SVs would contribute to a higher allele count. The remaining SVs were analysed to discard false positive calls and to determine whether, either alone or in combination with existing variants, they could explain the participant's phenotype. Candidate SVs were manually inspected using participant's BAM file on IGV browser v2.15.4 for validation of SV calls and to correct the genomic coordinates that are often ambiguously defined by the SV caller.

## **2.8 Bioinformatic pipelines**

### **2.8.1 Quality assessment of BAM files using FASTQC**

The FASTQC v0.12.1 was downloaded using the following URL <https://www.bioinformatics.babraham.ac.uk/projects/download.html#fastqc>. This tool performed quality control checks on the supplied BAM file using a single command (Appendix B: section 8.2.4). FASTQC provided a statistical overview of the total reads analysed and a graphical representation for multiple quality control metrics. This includes the GC content, sequence length distribution,

average quality score for all reads, quality score at each position in the sequence, nucleotide distribution, percentage of undetermined bases (N content), overrepresented sequences and proportion of adapter sequences. The single command used specified the path to the output directory and the path to the input file

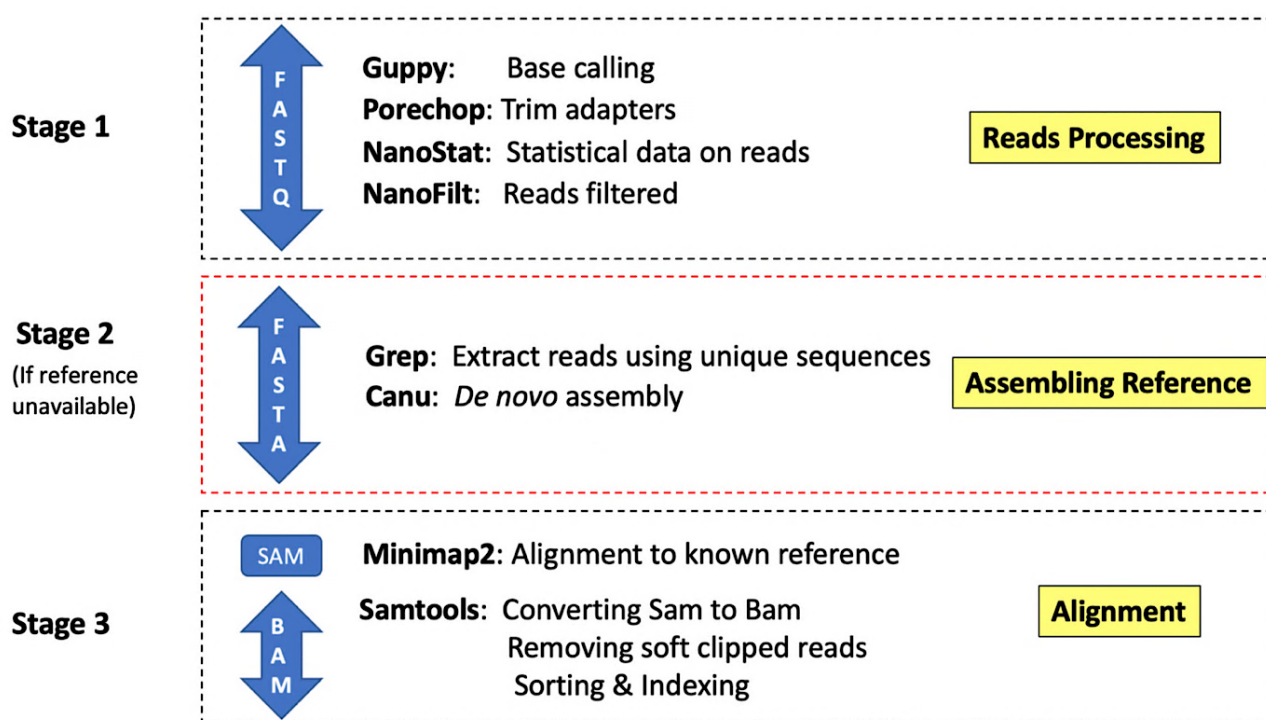
### **2.8.2 Annotation of VCF files using VEP and additional plug ins**

Connection to the host HPC at the University of Leeds was established using a unique address (ummamd@arc4.leeds.ac.uk). The reannotation of target VCF file was achieved using Ensembl VEP and additional plug ins for CADD, SpliceAI and UTRannotator to characterise exonic, intronic and intergenic variants detected. A bioinformatic environment was created to download Ensembl VEP toolkit for GRCh37 using a script (Appendix B: section 8.2.5). A second script was used which specified a runtime of up to 48 hours and a 1 gigabyte RAM to download the CADD, SpliceAI and UTRannotator plug ins via the web address provided. Once these tools have been downloaded, a third script was executed to perform the annotation of variants called and this script detailed the path to the input file and VEP plug ins, input file format, the selected species and the corresponding genomic assembly and the title for the output VCF file and the overall variant statistic file to be produced. The genomic VCF was split by chromosome using bcftools v1.21 for a more efficient data processing by the local desktop computer. The subsequent stage entailed creating subset of the genomic VCF to extract only variants in the exon and introns of 52 FH genes using the commands detailed in Appendix B: section 8.2.6.

### **2.8.3 Multiplexing assembly of plasmids using ONT sequencing**

The University of Leeds HPC systems was accessed using the address (ummamd@arc4.leeds.ac.uk) to support the analysis described in this section. The bioinformatic pipeline used involves three critical stages in isolating a single plasmid's reads (Figure 2.1). Stage 1 is a quality control stage that filters reads based on quality (Q10) and size discrimination to eliminate reads from fragmented templates or those formed by unwanted ligation during library preparation. Stage 2 is an optional stage reserved for *de novo* assembly, which

entails isolating reads based on unique sequences to differentiate between contaminating reads belonging to other plasmids. The FASTQ file containing all the reads is filtered to extract only reads containing the unique target sequences (Appendix C: Supplementary Table 5.4). Finally, stage 3 discards soft clipped reads i.e., reads that partially align to the reference sequence to retain only reads of high confidence. This proposed workflow supports two assembly methods, one being the assembly reference which is more rapid and uses a known reference in alignment, the other being a *de novo* assembly if the reference sequence is unavailable. This method generates a contiguous sequence based on consensus from reads of the same origin. The bioinformatic script is available on Appendix B: section 8.2.7).



**Figure 2.1. Barcode free assembly of multiplexed plasmids using ONT sequencing.** The workflow is divided into three sections with the first involving read processing of the FASTQ file. The second stage highlighted in red is optional and involves a *de novo* assembly to generate a consensus sequence in FASTA format that represents the plasmid. The last stage involves the alignment of reads to the supplied reference and the generation of the BAM file. The different file formats at each stage is contained within blue arrows or rectangles.

## Chapter 3 : Spectrum of genetic variation in *SLC38A8* and associated phenotype

### 3.1 Introduction

In 2013 Al-Araimi and colleagues described a recessive disorder comprising FH, an optic nerve decussation defect (chiasmal misrouting) and anterior segment dysgenesis (ASD), which they named FHONDA syndrome (Al-Araimi et al., 2013). Genome wide linkage analysis performed on FHONDA affected families mapped the disorder to a 6 Mb locus on 16q23.3-24.1. Subsequent autozygosity mapping and Sanger sequencing in seven affected families identified pathogenic variants in *SLC38A8* as the genetic aetiology (Poulter et al., 2013). This condition has subsequently been renamed as foveal hypoplasia 2 (FVH2), after ocular findings in additional families prompted the revision of the phenotype to constitute FH and chiasmal misrouting, with ASD being a variable phenotype (Ehrenberg et al., 2021, Schiff et al., 2021, Kuht et al., 2020, Weiner et al., 2020, Lasseaux et al., 2018, Perez et al., 2014, Poulter et al., 2013, Toral et al., 2017). Further reports of an iris TID in five unrelated European families with biallelic *SLC38A8* variants introduces a significant phenotypic overlap with albinism (Kuht et al., 2020, Lasseaux et al., 2018, Schiff et al., 2021).

The lack of a consensus surrounding the *SLC38A8* phenotype, coupled with the diagnostic challenges in detecting hypopigmentation or chiasmal misrouting, means that a proportion of patients with *SLC38A8* biallelic variants are misdiagnosed as albinism or an another FH related IRD, suggesting that the condition is underdiagnosed (Campbell et al., 2019, Lasseaux et al., 2018, van Genderen et al., 2006, Jackson et al., 2020). This is a common problem for rare disorders, which are often neglected in clinical pathways, especially if the disorder has only recently been characterised. To address this diagnostic issue, unbiased genomic screening provided by WGS has become more routinely used in genomic investigations. This circumvents the issue of phenotypic complexity and provides actionable information to increase recruitment for clinical trials and inform reproductive choices. The use of WGS has been endorsed and encouraged by the NHS ([www.england.nhs.uk/genomics/nhs-genomic-med-service/](http://www.england.nhs.uk/genomics/nhs-genomic-med-service/)) and the UK government

([www.gov.uk/government/publications/genome-uk-the-future-of-healthcare/genome-uk-the-future-of-healthcare](http://www.gov.uk/government/publications/genome-uk-the-future-of-healthcare/genome-uk-the-future-of-healthcare)), with the aim of revolutionising patient care in the UK through genomic medicine. FH patients have also been recruited to the 100KGP with their genomes sequenced in a diagnostic setting and made available to researchers. This provided a large cohort for analysis, with considerable translational potential. This study undertakes bioinformatic analysis of genomic data from both the 100KGP and local patients suspected of having FVH2, and reviews the published literature on the *SLC38A8* variants and phenotypes, to characterise the spectrum of pathogenic variants in this gene that underlie FVH2.

## 3.2 Results

### 3.2.1 The reanalysis of novel *SLC38A8* variants obtained through collaboration

Being part of the European Retinal Disease Consortium (ERDC) facilitated collaborative efforts to identify causative variants in families with FH for which a genomic diagnosis had not yet been obtained. NGS performed by local colleagues and members of the ERDC on four families with FH and without apparent hypopigmentation has identified four novel and one known *SLC38A8* variant. All five variants were included in this study for the genetic characterisation of *SLC38A8* and have been reanalysed using various *Insilico* pathogenicity tools and were reclassified as pathogenic or likely pathogenic according to ACMG criteria (Figure 3.1).

A novel inversion, NC\_000016.10:g.84015931\_84027116inv was detected in proband F1310, who is from a family (F1) of Middle Eastern origin (Figure 3.1). This balanced SV was detected in a homozygous state and co-segregated with FH and nystagmus in the proband's family (Lord et al., 2017). Evaluation of chiasmal misrouting was not performed in this case. Initially, Sanger sequencing did not detect mutations in the coding segments of *SLC38A8* or at the exon-intron boundaries, but WGS revealed an 11.1 kb inversion spanning exons 7-9 which is predicted to alter the reading frame (Figure 3.2A). Further analysis showed that

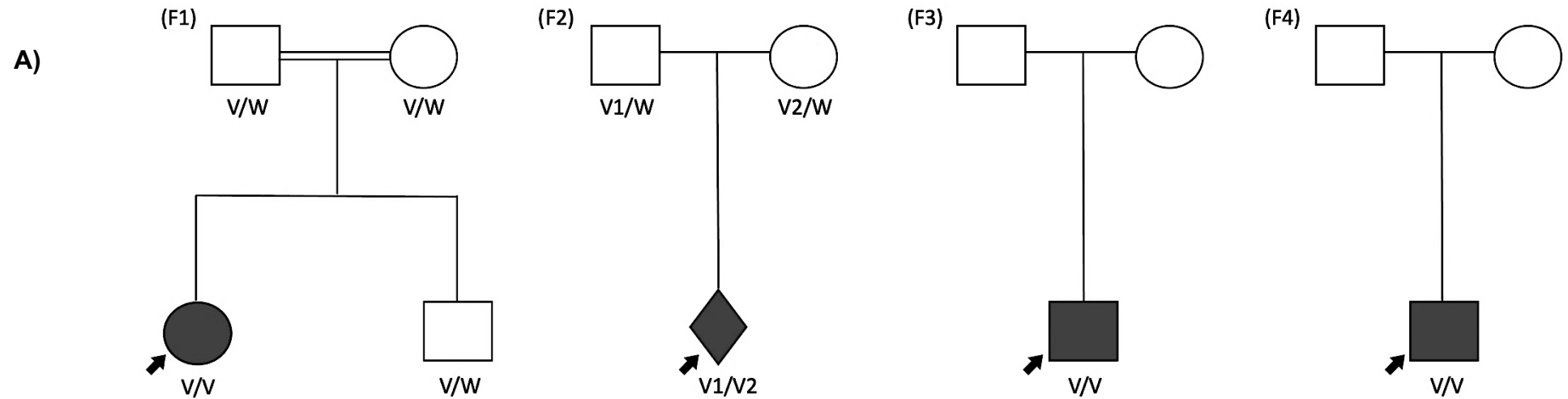


the breakpoints of this pathogenic inversion overlaps short interspersed nuclear elements (SINE) at chr16:84015711-84015978 and chr16:84027071-84027357 (Figure 3.2B).

In family 2 (F2), the proband ERD01 harboured compound heterozygous deletions of 2.7 kb and 12.5 kb detected in trans (Figure 3.1). The proband presented with FH but was not assessed for the presence of chiasmal misrouting. WGS confirmed the inheritance of NC\_000016.10:g.84006453\_84019002del (V1) from the father and NC\_000016.10:g.84034763\_84037497del (V2) variant from the mother. The NC\_000016.10:g.84034763\_84037497del variant eliminates exon 3 to induce a frameshift (Figure 3.2A). The NC\_000016.10:g.84006453\_84019002del abolishes exon 8-11 without altering the reading frame but gains additional incorrect residues from translating intergenic sequences that are present downstream the deletion. The sequence of the aberrant mRNA transcripts produced by these SVs is provided in (Appendix C: Supplementary Figure 3.1). The breakpoints in both the pathogenic SVs overlapped at least one SINE or a long terminal repeat (LTR) (Figure 3.2B).

The proband ERD02 in family 3 (F3) is a 5 year old male who presented with FH, congenital nystagmus and myopia. Assessment for chiasmal misrouting was not performed in this patient. The proband underwent WES which identified a homozygous frameshift variant NM\_001080442.3:c.669delC; NP\_001073911.1:p.(Thr224Profs\*44). This likely pathogenic variant is very rare as it is absent from both gnomAD and the 100KGP.

Molecular screening in proband F1368 of family 4 (F4) who exhibited FH, nystagmus and ASD has led to the discovery of a homozygous missense variant, NM\_001080442.3:c.848A>C; NP\_001073911.1:p.(Asp283Ala) (Figure 3.1). Similarly, the presence of chiasmal misrouting was not evaluated in this patient. This variant has been previously reported in gnomAD, with a higher prevalence in Ashkenazi Jewish populations. The allele frequency in that ethnic group is 232/29608, with four homozygotes reported. The variant has been classified as likely pathogenic by ACMG.



**B)**

ID	Source	Ethnicity	Phenotype	Genotype	Variant	gnomAD	100KGP	Mutation taster	Ensemble VEP	CADD	SIFT	Polyphen-2	Align GVGD	ACMG
F1 (F1310)	Local Cohort	Middle Eastern	Foveal hypoplasia and nystagmus	Homozygous	<i>g.84015973_84027074inv</i>	-	2/142816	-	-	-	-	-	-	Pathogenic
F2 (ERD01)	ERDC	-	Foveal hypoplasia	Compound heterozygote	<i>g.84006453_84019002del</i> <i>g.84034763_84037497del</i>	-	-	-	-	-	-	-	-	Pathogenic Pathogenic
F3 (ERD02)	ERDC	-	Foveal hypoplasia, nystagmus and myopia	Homozygous	<i>c.669delC</i> <i>p.(Thr224Profs*44)</i>	-	-	Disease causing	High impact	-	-	-	-	Likely pathogenic
F4 (F1368)	Local Cohort	European	Foveal hypoplasia, nystagmus and anterior segment dysgenesis	Homozygous	<i>c.848A&gt;C</i> <i>p.(Asp283Ala)</i>	34/1180002	4/124172	Disease causing	Moderate impact	27.1	Damaging	Probably damaging	Class C65	Likely pathogenic

**Figure 3.1. FH families solved through the identification of biallelic *SLC38A8* variants.** A) Pedigree of families with a suspected FVH2. B) Clinical findings and pathogenicity assessment of *SLC38A8*.

A)

**5'3' Frame 2**  
 QILGTHGPDHLLSS-VWRGLTLTHTR-GSFS--LTHPWSSNLASPTPLAPRNATEPPGVSPSPIPNSQRLSGTLKVGFGAMEGQTPG  
 SRGLPEKPHPATAAATLSSMGAVFILMKSALGAGLLNFPWAFSKAGGVVPAFLVELVSLVFLISGLVILGYAAAVSGQATYQGVVRGLC  
 GPAIGKLCCEACFLNLLMISVAFRLVIGDQLEKLCDSLLSGTTPAPQPWYADQRFTLPLLSVLVILPLSAPREIAFQKYTSILGTLAAC  
 YLALVITVQYYLWPQGLVRESHPSLSPASWTSVFSVFPPTICFGFQCHEAAVSIYCSMRKRSLSHWALVSVLSLLACCLLYSLTGVYGFLL  
 TFGTEVSADVLMSPGNDMVIIVARVLFVAVSIVTVYPIVFLGRSVMQDFWRRSCLGGWGPSALADPSGLWVRMPLTILWVTVTLAMAL  
 FMPDLSEIVSIIIGGISFFIFIFPGLCLICAMGVEPIGPRVKCCLEVVWGVVSVLVGTFIFGQSTAAAVWEMF-WAASAGQEGALRGLTL  
 RGCCMQPDRCHFFSS-RCWRD-

Wildtype SLC38A8

**5'3' Frame 2**  
 QILGTHGPDHLLSS-VWRGLTLTHTR-GSFS--LTHPWSSNLASPTPLAPRNATEPPGVSPSPIPNSQRLSGTLKVGFGAMEGQTPG  
 SRGLPEKPHPATAAATLSSMGAVFILMKSALGAGLLNFPWAFSKAGGVVPAFLVELVSLVFLISGLVILGYAAAVSGQATYQGVVRGLC  
 GPAIGKLCCEACFLNLLMISVAFRLVIGDQLEKLCDSLLSGTTPAPQPWYADQRFTLPLLSVLVILPLSAPREIAFQKYTSILGTLAAC  
 YLALVITVQYYLWPQGLVRESHPSLSPASWTSVFSVFPPTICFGFQVCASSVQWVSSI-DQESSAAWRSGEWSLCSAPSLSGRARRQRS  
 GRCSDGQLVPRKGPSSG-PYVAAVCSQETDAISFPHKDAGET

NC\_000016.10: g.84015931\_84027116inv

**5'3' Frame 2**  
 QILGTHGPDHLLSS-VWRGLTLTHTR-GSFS--LTHPWSSNLASPTPLAPRNATEPPGVSPSPIPNSQRLSGTLKVGFGAMEGQTPG  
 SRGLPEKPHPATAAATLSSMGAVFILMKSALGAGLLNFPWAFSKAGGVVPAFLVELCVTPSCLAPRPPRSRGTQTSASPCPCSPCWSSC  
 PCLPRGRSPSRNTQAS-ALWLPVTWPWSSPCSTTSGPRASCVSPILH-ALPPGPLCSVSPSPSASGFSVTKLPSPTAACANGASPTGP  
 WCLCCPCWPAASSIH-RGFMAS-LLGQKFLITS-CPTQAMIWSSSLWPGSEFLLSPS-LSTPSCSSWGGQ-CRTSGGGAAGWGGPAPWPT  
 PQCGSGCR-PSCGSP-RSPWRCLCLTSARSSASSEASVPSSSSSSVCASSVQWVSSI-DQESSAAWRSGEWSLCSAPSLSGRARRQ  
 RSGRCSDGQLVPRKGPSSG-PYVAAVCSQETDAISFPHKDAGET

NC\_000016.10: g.84034763\_84037497del

**5'3' Frame 2**  
 QILGTHGPDHLLSS-VWRGLTLTHTR-GSFS--LTHPWSSNLASPTPLAPRNATEPPGVSPSPIPNSQRLSGTLKVGFGAMEGQTPG  
 SRGLPEKPHPATAAATLSSMGAVFILMKSALGAGLLNFPWAFSKAGGVVPAFLVELVSLVFLISGLVILGYAAAVSGQATYQGVVRGLC  
 GPAIGKLCCEACFLNLLMISVAFRLVIGDQLEKLCDSLLSGTTPAPQPWYADQRFTLPLLSVLVILPLSAPREIAFQKYTSILGTLAAC  
 YLALVITVQYYLWPQGLVRESHPSLSPASWTSVFSVFPPTICFGFQCHEAAVSIYCSMRKRSLSHWALVSVLSLLACCLLYSLTVSFPSSG  
 KMEQGSVQSRSLVRIKDDGSRVADGIRHAQWRPKQCSFIPRRPRS-LARGT-VSWEPRG

NC\_000016.10: g.84006453\_84019002del

B)

ID	Structural variant	Breakpoint 1			Breakpoint 2		
		Repeat element	Locus	Size (bp)	Repeat element	Locus	Size (bp)
F1310	84015931_84027116inv	SINE (AluSq2)	16:84015711-84015978	268	SINE (AluSg)	16:84027071-84027357	287
ERD01	g.84006453_84019002del	SINE (MIRb)	16:84006313-84006540	228	LTR (MER34D)	16:84018616-84019084	469
	g.84034763_84037497del	-	-	-	SINE (AluSx1)	16:84037479-84037784	306

**Figure 3.2 Characterisation of SVs identified in the local FH cohort.** The predicted translation of *SLC38A8* coding sequence. The reading frame is highlighted in pink with the start site, methionine (M) is in red and the termination signal marked by “-”.

### **3.2.2 Bioinformatic analysis of the 100KGP datasets**

The launch of the 100KGP provided access to genomes from a large cohort of rare disease patients (71682 participants) who lacked a molecular diagnosis at the time of recruitment, including many with ophthalmic manifestations. This facilitated the identification of *SLC38A8* variants causing FVH2, with the aim of further characterising the spectrum of *SLC38A8* variants underlying FH and to increase the diagnostic yield in the 100KGP for patients with FH. The rare disease cohort was selected for analysis due to its relevance to the FVH2 phenotype, which is of very low prevalence and is caused by germline variants of high penetrance. In this study, the clinical and genomic data of participants belonging to the Main Programme v11 dataset was investigated (accessed 28/12/21). The *SLC38A8* cohort within this dataset comprised 875 variant entries belonging to probands and their relatives. These variants were filtered at 0.1% MAF and annotated by the Exomiser variant prioritisation framework and the variants were also supplemented by the Tiering classification for evidence of variant segregation.

#### **3.2.2.1 Targeted analysis of the biallelic *SLC38A8* cohort**

The 100KGP dataset was initially screened for biallelic variants in *SLC38A8*, consistent with the established autosomal recessive inheritance of *SLC38A8*-related FH (425 variant entries). In this analysis probands were excluded unless they had been submitted with HPO terms indicative of retinal defects. The phenotypic criteria used were deliberately vague to compensate for the inconsistencies in the reporting of phenotypes in the 100KGP, possibly due to reporting errors at the recruiting NHS genomic medicine centre (Best et al., 2022a). HPO term selection was based on the 100KGP specific disease categories assigned to relevant inherited retinal diseases, which are developmental macular and foveal dystrophy, inherited macular dystrophy, Leber congenital amaurosis or early onset retinal dystrophy and rod cone dystrophy. The most common HPO terms reported for these four ophthalmic disease categories were selected as the retention criteria used for filtering of biallelic *SLC38A8* cases (Appendix A: section 8.1.3).

To eliminate redundancy in the analysis, participants were discarded if a diagnosis involving other genes was delivered through the 100KGP. This stringent filtering criteria resulted in 19 probands with biallelic *SLC38A8* variants and varying ocular manifestations. The *SLC38A8* variants in the 19 probands were subjected to *In silico* pathogenicity assessment using a set of diverse and open access tools (Appendix C: Supplementary Table 3.1). Five probands were initially classified as confirmed FVH2 cases following the detection of two deleterious *SLC38A8* variants in each participant (Table 3.1). In each case, the pathogenicity for these variants was supported by a predicted pathogenic consequence and a low allele frequency of  $\leq 0.00033$  for the relevant specific population in gnomAD and  $\leq 0.0001$  for the global population of the 100KG. With exception being NM\_001080442.3:c.922A>G; NP\_001073911.1:p.(Thr308Ala), which had an allele frequency of 0.0006 in gnomAD and 0.0005 in the 100KGP. The *SLC38A8* variants in the five convincing FVH2 cases displayed an Exomiser score  $\geq 0.8$ , suggesting that this may be a reliable threshold to identify *SLC38A8* variants contributing to participants' phenotype.

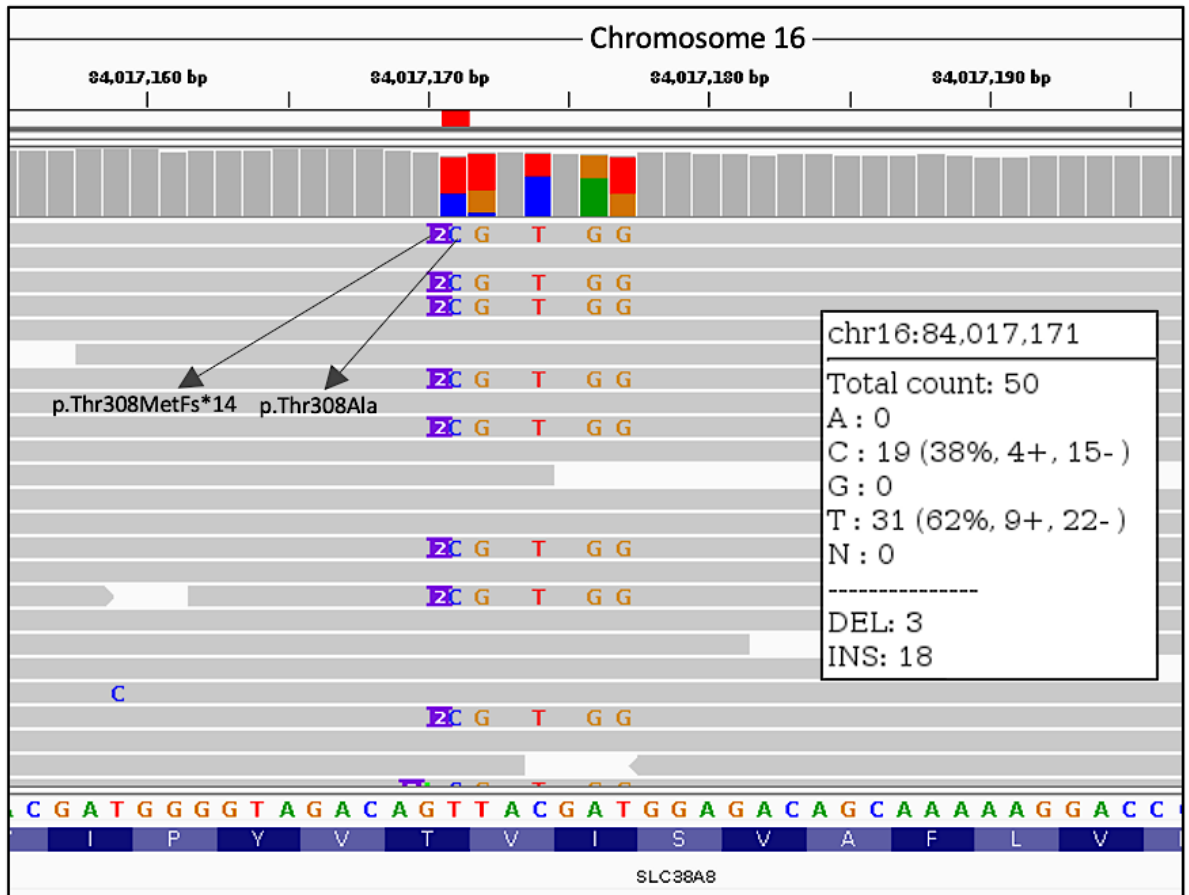
One proband, case 1 (C1) was identified with two *SLC38A8* variants affecting the same amino acid residue, a pathogenic frameshift variant, NP\_001073911.1:p.(Thr308Metfs\*14) and a missense, NP\_001073911.1:p.(Thr308Ala) that is classified as a VUS. This seemed a highly unlikely finding for biallelic variants. Further investigation of the patient's BAM file revealed that both variants were in cis which suggest that one variant is actually an artefact, and this case was therefore excluded from the analysis as being monoallelic (Figure 3.3).

The remaining four participants with *SLC38A8* variants included three individuals (C2-C4) homozygous for a nonsense variant NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) and one proband (C5) who is compound heterozygous for a nonsense variant NM\_001080442.3:c.435G>A; NP\_001073911.1:p.(Trp145\*) and a splice donor variant NM\_001080442.3:c.632+1G>A, detected in trans. The nonsense variants NP\_001073911.1:p.(Tyr88\*) and NP\_001073911.1:p.(Trp145\*) induce a premature stop codon in exons 3 and 4 respectively, with the mutated mRNA

transcripts predicted by MutationTaster to undergo NMD. The splice donor variant abolishes a canonical splice recognition sequence (GT) downstream of exon 5, and is predicted by spliceAI to disrupt splicing. Variants NP\_001073911.1:p.(Tyr88\*), NP\_001073911.1:p.(Trp145\*) and NM\_001080442.3:c.632+1G>A were all classified as pathogenic having fulfilled the ACMG criteria. At the time of the discovery, all three variants in participants II-V were novel, but they were later published as causative for FVH2 (Schiff et al., 2021).

ID	Ethnicity	Phenotype	Genotype	Variants	gnomAD	100KGP	Exomiser	Tier	CADD	MutationTaster	SIFT	PolyPhen-2	varSEAK	SpliceAI	ACMG
C1	European	Visual loss	Compound heterozygous	c.922_923insTG p.(Thr308Metfs*14)	0	-		3	-	Disease causing	-	-	-	-	Likely Pathogenic
				c.922A>G p.(Thr308Ala)	689/1179964 (1)	61/126698	0.0465	3	28	Disease causing	Deleterious	Probably damaging	-	-	VUS
C2	South Asian	Non-progressive visual loss, abnormal macular morphology, central scotoma, abnormal fundus morphology and visual impairment	Homozygous	c.264C>G p.(Tyr88*)	30/91074	16/126698	0.912	1	55	Disease causing	-	-	-	-	Pathogenic
C3	South Asian	Progressive visual loss, central scotoma, visual impairment, reduced visual acuity and retinal dystrophy	Homozygous	c.264C>G p.(Tyr88*)	30/91074	16/126698	0.882	1	55	Disease causing	-	-	-	-	Pathogenic
C4	South Asian	Congenital nystagmus	Homozygous	c.264C>G p.(Tyr88*)	30/91074	16/126698	0.840	1	55	Disease causing	-	-	-	-	Pathogenic
C5	European	Visual impairment and nystagmus	Compound heterozygous	c.435G>A p.(Trp145*)	1/349594	3/126698		3	48	Disease causing	-	-	-	-	Pathogenic
				c.632+1G>A	2/1111810	2/126698	0.997	3	32	-	-	-	Class 5	DL (0.95)	Pathogenic

**Table 3.1. *SLC38A8* biallelic cases with ophthalmic disorders in the 100KGP.** Participants with biallelic *SLC38A8* variants retained after *In silico* pathogenicity assessment. All variants were reported using the MANE transcript for *SLC38A8* (NM\_001080442.3). Allele frequencies for the relevant ethnicity were obtained from gnomAD v4.0. The number in brackets indicates the homozygous count. 100KGP allele frequencies were extracted from interactive variant analysis (IVA) v.2.2.3 available in the GEL research environment. Splice variants were interrogated using CADD score, varSEAK and SpliceAI.



**Figure 3.3. Biallelic *SLC38A8* variants affecting amino acid residue 308.** BAM file of participant C1 harbouring both NP\_001073911.1:p.(Thr308Metfs\*14) and NP\_001073911.1:p.(Thr308Ala) reveals the presence of both variants only on the same reads. Note that at chr16:84017171, the alternative allele C is identified in 38% of reads while the insertion CA allele is in 36%.

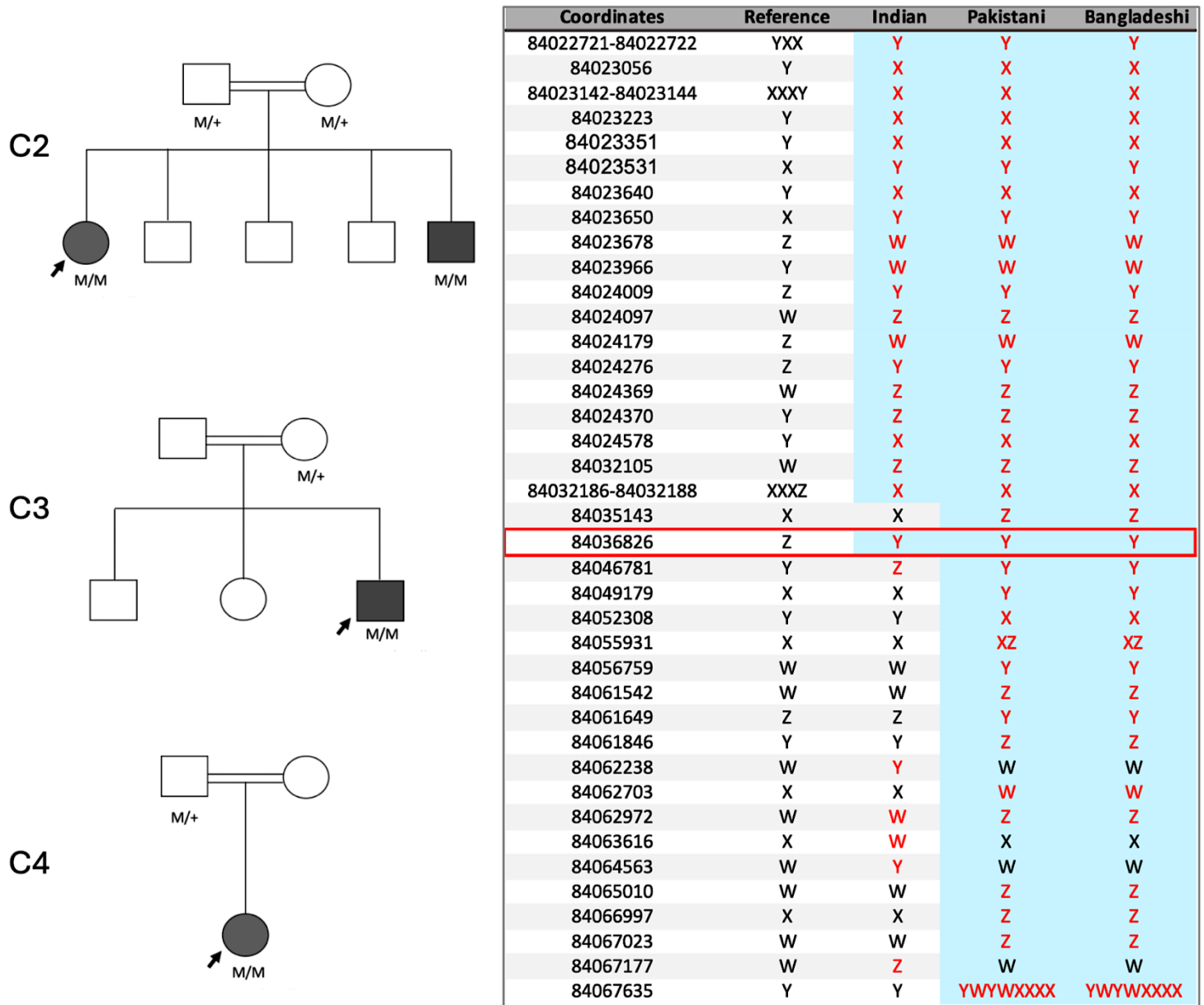


### 3.2.2.2 Haplotype analysis of a potential founder mutation in *SLC38A8*

The investigation detailed above led to the discovery of three South Asian patients (C2-C4) homozygous for a nonsense pathogenic variant, NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*), all reported to have ocular phenotypes compatible with FH in the 100KGP (Table 3.1). Despite having the same genotype, these three patients had been reported to have distinct ocular manifestations. The finding of apparently unrelated Indian (C2), Bangladeshi (C3) and Pakistani (C4) patients harbouring NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) warranted an investigation into the possibility of a founder mutation restricted to patients of South Asian ancestry.

Variant segregation analysis was performed through analysis of the genomes of available family members whom were recruited alongside the proband to the 100KGP (Figure 3.4). The pathogenic variant NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) was detected in a heterozygous state in all asymptomatic family members and was detected as homozygous only in affected individuals.

The locus surrounding the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) variant in all three patients was examined on IGV browser v2.15.4. Analysis of the sequence in a 44.9 kb region surrounding the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) has revealed identical haplotypes encompassing the variant in patients of Bangladeshi (C3) and Pakistani (C4) origin. A partially overlapping haplotype was also found in the patient of Indian ancestry (C2). The homology block examined across all patients spans a 9.5 kb region from chr16:84022721-84032188. A single discordant allele at chr16:84035143 is observed in the Indian patient which is upstream of the shared nonsense variant NP\_001073911.1:p.(Tyr88\*). In addition, a large discordant region of 20.9 kb (chr16:84046781-84067635) is downstream of the shared NP\_001073911.1:p.(Tyr88\*) variant.



**Figure 3.4. Haplotype of participants harbouring p.(Tyr88\*) in SLC38A8.**

Alleles were encrypted in compliance with GEL publication policy. Reference allele is depicted in black while alternate alleles are in red. Identical alleles between participants is highlighted in a light blue. Homology between two patients of Bangladeshi (C3) and Pakistani (C4) origin and apparent discrepancy in alleles with a participant of Indian ancestry (C2). Red box highlights the shared *SLC38A8* variant detected in all 3 patients of South Asian descent.

### 3.2.2.3 Analysis of the monoallelic *SLC38A8* cohort

The conventional rare disease annotation pipelines of Exomiser and Tiering are effective but are restricted to SNVs and INDELS in the coding segments and at canonical splice sites of the genome. These pipelines do not capture deep intronic variants or SVs, which causes a proportion of pathogenic variants in *SLC38A8* to remain undetected. This restricted analytical approach of the 100KGP can contribute to the underdiagnosis of FVH2. In particular, compound heterozygous probands in whom one allele has been missed will be discounted in filtering for biallelic cases. Based on this rationale, the Main Programme v11 dataset was revisited for a selective analysis to capture participants with deleterious monoallelic *SLC38A8* variants. These were then subjected to further screening for potentially pathogenic second alleles involving non-coding variants or SVs.

In total, 450 monoallelic *SLC38A8* variant entries were identified in the 100KGP participants. The participants were filtered according to the established HPO terms derived from the 4 specific disease groups compatible with an IRD (Appendix A: section 8.1.3). This filtering criteria resulted in 49 *SLC38A8* variants in probands with phenotypes consistent with FH. Probands whom received a molecular diagnosis unrelated to *SLC38A8* via the 100KGP were eliminated from the analysis. Subsequent selection excluded synonymous and splice region variants that are  $\geq 3$  bp away from a canonical splice site, unless they were predicted by spliceAI to alter splicing of the MANE transcript. Variants identified in the remaining 29 participants underwent computational pathogenicity assessment in combination with allele frequency evaluation and GEL annotation for characterisation (Appendix C: Supplementary Table 3.2). Variant interpretation was facilitated by alternative scores since the default Exomiser score and Tiering classification algorithms takes account of biallelic inheritance in *SLC38A8* and therefore the low score generated will reflect the lack of a second variant. Instead, the variant score was applied to assess pathogenicity irrespective of the mode of inheritance, while the Gene-Pheno score assessed phenotypic homology to the known *SLC38A8* phenotype of FVH2.

14/29 participants had variants with an allele frequency below 0.001 and these were predicted to be deleterious based on the collective evidence from different computational tools (Table 3.2). However, the ACMG classification for 13 monoallelic *SLC38A8* variants was VUS and only a single variant, NM\_001080442.3:c.389-2A>G was categorised as likely pathogenic. The genomes belonging to these 14 probands (C6-C19) were retained to further bioinformatic analyses to identify a potential second pathogenic allele in *SLC38A8*.

ID	Ethnicity	Phenotype	Genotype	gnomAD	100KGP	Variant score	Gene-Pheno	CADD	Grantham	SIFT	PolyPhen-2	Align-GVGD	varSEAK	SpliceAI	ACMG
C6	European	Nystagmus	c.723C>A p.(Ser241Arg)	10/1110398	2/126698	0.9999	0.1710	23.4	110	Deleterious	Probably damaging	Class C65	-	-	VUS
C7	European	Nystagmus and gaze evoked nystagmus	c.389-2A>G	4/1103970	1/126698	0.9996	0.1889	33	-	-	-	-	Class 5	AG:0.43 AL:0.97	Likely Pathogenic
C8	European	Progressive visual loss	c.776T>C p.(Leu259Pro)	16/1180028	1/16638	0.9982	0.2567	28.9	98	Deleterious	Probably damaging	Class C65	-	-	VUS
C9	African	Abnormality of saccadic eye movement	c.895C>T p.(Arg29Trp)	3/74974	1/126698	0.9973	0.2075	26.1	101	Deleterious	Probably damaging	Class C65	-	-	VUS
C10	European	rod-cone dystrophy, visual impairment and sensorineural hearing impairment	c.1289C>T p.(Ala430Val)	39/1613746	4/126698	0.9945	0.2743	24.9	64	Deleterious	Possibly damaging	Class C55	-	-	VUS
C11	European	Progressive visual loss, macular degeneration, central scotoma and visual impairment	c.792C>G p.(Ile264Met)	24/1179996	1/126698	0.9918	0.3173	26.1	10	Deleterious	Probably damaging	Class C0	-	-	VUS
C12	African	rod-cone dystrophy and visual impairment	c.109G>A p.(Ala37Thr)	0/75034	7/126698	0.9906	0.2904	20.7	58	Deleterious	Benign	Class C55	-	-	VUS
C13	European	Macular dystrophy and retinal dystrophy	c.127C>G p.(Leu43Val)	113/117994	16/126698	0.9963	0.3487	27.7	32	Deleterious	Probably damaging	Class C25	-	-	VUS
C14	European	Progressive visual loss, optic neuropathy and visual impairment	c.127C>G p.(Leu43Val)	113/117994	16/126698	0.9963	0.2926	27.7	32	Deleterious	Probably damaging	Class C25	-	-	VUS
C15	European	Iris coloboma and visual impairment	c.127C>G p.(Leu43Val)	113/117994	16/126698	0.9963	0.2517	27.7	32	Deleterious	Probably damaging	Class C25	-	-	VUS
C16	East Asian	Rod-cone dystrophy	c.131A>G p.(Asn44Ser)	2/39666	1/126698	0.9963	0.2930	22.7	46	Tolerated	Possibly damaging	Class C45	-	-	VUS
C17	South Asian	visual impairment	c.787C>G p.(Leu263Val)	4/91092	2/126698	0.9948	0.3011	23.2	32	Tolerated	Possibly damaging	Class C25	-	-	VUS
C18	European	Visual impairment	c.1205C>T p.(Pro402Leu)	780/1179970 (2)	71/126698	0.9624	0.3474	23.8	98	Tolerated	Possibly damaging	Class C0	-	-	VUS
C19	African	progressive visual loss	c.487C>T p.(Pro163Ser)	47/75052 (1)	17/126698	0.9904	0.2904	24.8	74	Deleterious	Probably damaging	Class C65	-	-	VUS

**Table 3.2. SLC38A8 monoallelic cohort in the 100KGP.** Clinical and genomic data relating to participants with potentially deleterious *SLC38A8* variants. All variants were reported using the MANE transcript (NM\_001080442.3). gnomAD allele frequencies are displayed for the relevant ethnicity, with homozygotes reported in brackets. The 100KGP population frequencies obtained by IVA v2.2.3 are not refined by ethnicity.

#### 3.2.2.4 Probing for deep intronic variants in *SLC38A8*

19 participants possessing a single predicted deleterious variant and an ocular phenotype indicative of FH in the 100KGP were reanalysed for the purpose of detecting a second pathogenic allele in the noncoding genomic region. This cohort comprised the 14 participants with potentially deleterious monoallelic variants (Table 3.2) and 5 additional participants discarded from the targeted analysis of the biallelic cohort due to having a single deleterious variant and a likely benign variant (Appendix C: Supplementary Table 3.1).

To detect hidden pathogenic variants in the noncoding and coding regions of the genome, a bioinformatics tool native to the GEL research environment known as Gene-Variant Workflow v1.6 was used and changes to the script were made to target *SLC38A8* (Appendix B: section 8.2.1). This pipeline is exclusive to the rare disease cohort and was used to extract all the *SLC38A8* variants from each participant's VCF. The output for *SLC38A8* (44,497 variant entries) was split into GRCh37 (11,348) and GRCh38 (33,149). The large number of *SLC38A8* variants detected was refined to exclude variants with a MAF >0.002 (0.2%) in gnomAD and the 100KGP using a script known as 100K Gene Variant Filter (Appendix B: section 8.2.2). Intronic variants in 19 select participants were extracted and analysed by online pathogenicity assessment tools to determine their potential to alter splicing (Appendix C: Supplementary Table 3.3). Additionally, the noncoding variants were investigated to using the ENCODE database to determine whether they overlap CREs and might therefore influence transcription. Another annotation provided by UTRannotator was used to infer whether intronic variants in the 5' UTR could disrupt the reading frame. The analysis in 12/19 participants did not yield any significant results.

Genomic findings in 7/19 participants showed rare intronic variants but none of these altered splicing nor affected the 5'UTR (Table 3.3). However, three participants (C11, C19 and C20) had SNVs in a shared CRE which is a distal enhancer like signature (ENCODE ID:EH38E3194808). This CRE is 207 bp in length spanning chr16:84025275-84025482. Without functional assays the pathogenicity of these variants cannot be ascertained.

Interestingly, the analysis also uncovered hidden exonic and potentially deleterious variants not called by Exomiser in participants C21 and C22, as their allele frequency is greater than 0.001 (0.1%). The two variants are NM\_001080442.3:c.440C>G; NP\_001073911.1:p.(Ala147Gly) and NM\_001080442.3:c.743C>G; NP\_001073911.1:p.(Ser248Cys), which are predicted by *In silico* tools to be tolerated and deleterious respectively. The allele frequency for both variants is <0.004 in the disease enriched population of the 100KGP, but the general population data of gnomAD suggests that these may in fact be polymorphisms. The NM\_001080442.3:c.440C>G; NP\_001073911.1:p.(Ala147Gly) variant was reported in 469 homozygotes and NM\_001080442.3:c.743C>G; NP\_001073911.1:p.(Ser248Cys) in 505 homozygotes. The ACMG classification for both variants is a VUS. It was therefore concluded that no plausible hidden second variants were detected in the 19 participants included in this targeted analysis.

ID	Variants	CADD	Splice AI	VarSEAK	ENCODE	UTRannotator	SIFT	PolyPhen-2
C1	<b>c.922A&gt;G; p.(Thr308Ala)</b>	28	-	-	-	-	Deleterious	Probably damaging
	c.632+24G>A	0.024	DG: 0.01	Class 1	DNase-H3K4me3 (EH38E3194817)	none	-	-
C6	<b>c.723C&gt;A; p.(Ser241Arg)</b>	23.4	-	-	-	-	Deleterious	Probably damaging
	c.805+54A>G	2.092	0	Class 1	none	none	-	-
	c.1162+941dupG	1.146	0	Class 1	none	none	-	-
	c.691-494G>T	1.410	AG: 0.02	Class 1	none	none	-	-
	c.389-473G>T	1.71	0	Class 1	none	none	-	-
	c.190-911_190-907dupTTTT	1.207	0	Class 1	none	none	-	-
	c.190-1931C>T	0.64	0	Class 1	pELS (EH38E3194821)	none	-	-
c.1162+940_1162+941dupGG	-	-	-	none	none	-	-	
C11	<b>c.792C&gt;G; p.(Ile264Met)</b>	26.1	-	-	-	-	Deleterious	Probably damaging
	c.691-2585C>T	3.279	0	Class 1	dELS (EH38E3194808)	none	-	-
C19	<b>c.487C&gt;T; p.(Pro163Ser)</b>	24.8	-	-	-	-	Deleterious	Probably damaging
	c.1163-615C>T	1.24	0	Class 1	none	none	-	-
	c.389-224G>C	0.816	DG: 0.01	Class 1	none	none	-	-
	c.633-1156G>T	1.909	DG: 0.03	Class 1	none	none	-	-
	c.691-363T>C	1.563	DG: 0.01	Class 1	none	none	-	-
	c.806-2725C>T	1.223	DG: 0.01	Class 1	none	none	-	-
	c.691-2425T>C	1.254	0	Class 1	dELS (EH38E3194808)	none	-	-
C20	<b>c.487C&gt;T; p.(Pro163Ser)</b>	24.8	-	-	-	-	Deleterious	Probably damaging
	c.189+1028delG	0.267	0	Class 1	pLS (EH38E3194823)	none	-	-
	c.690+2355A>G	0.813	0	Class 1	none	none	-	-
	c.691-2425T>C	1.254	0	Class 1	dELS (EH38E3194808)	none	-	-
	c.691-363T>C	1.563	DG: 0.01	Class 1	none	none	-	-
	c.633-1156G>T	1.909	DG: 0.03	Class 1	none	none	-	-
	c.389-224G>C	0.816	DG: 0.01	Class 1	none	none	-	-
C21	<b>c.189G&gt;A; p.(Leu63=)</b>	22	DG: 0.28 DL: 0.01	Class 5	-	-	-	-
	c.471C>T; p.(Ser157=)	0.179	0	Class 1	none	none	-	-
	c.440C>G; p.(Ala147Gly)	13.15	0	-	none	none	Deleterious	Benign
	c.690+863G>C	0.369	0	Class 1	none	none	-	-
	c.806-369A>G	0.045	0	Class 1	none	none	-	-
	c.805+2428G>A	2.927	0	Class 1	none	none	-	-
	c.691-363T>C	1.563	DG: 0.01	Class 1	none	none	-	-
C22	<b>c.189G&gt;A; p.(Leu63=)</b>	22	DG: 0.28 DL: 0.01	Class 5	-	-	-	-
	c.743C>G; p.(Ser248Cys)	22.7	0	Class 1	none	none	Deleterious	Possibly damaging
	c.690+665G>A	1.735	0	Class 1	dELS (EH38E3194812)	none	-	-
	c.531-554A>G	2.116	0	Class 1	DNase only (EH38E3194819)	none	-	-
	c.690+683A>G	1.526	0	Class 1	dELS (EH38E3194812)	none	-	-
	c.690+3207A>C	2.296	0	Class 1	none	none	-	-
	c.690+804G>A	0.209	0	Class 1	dELS (EH38E3194812)	none	-	-
C23	c.806-3C>G	23.8	AL: 0.89	Class 5	none	none	-	-
	c.806-443C>T	3.016	0	Class 1	none	none	-	-

**Table 3.3. Hidden variants in the SLC38A8 monoallelic cohort.** Genomic findings in participants with a single potentially pathogenic SLC38A8 variant and HPO terms suggestive of an IRD. Variants in bold are the identified pathogenic SNVs called by Exomiser, with additional variants detected by Gene-Variant Workflow v1.6 being listed below. ENCODE ID is provided for each CRE identified. pELS-proximal enhancer like signature; dELS-distal enhancer like signature; pLS-promoter like signature.



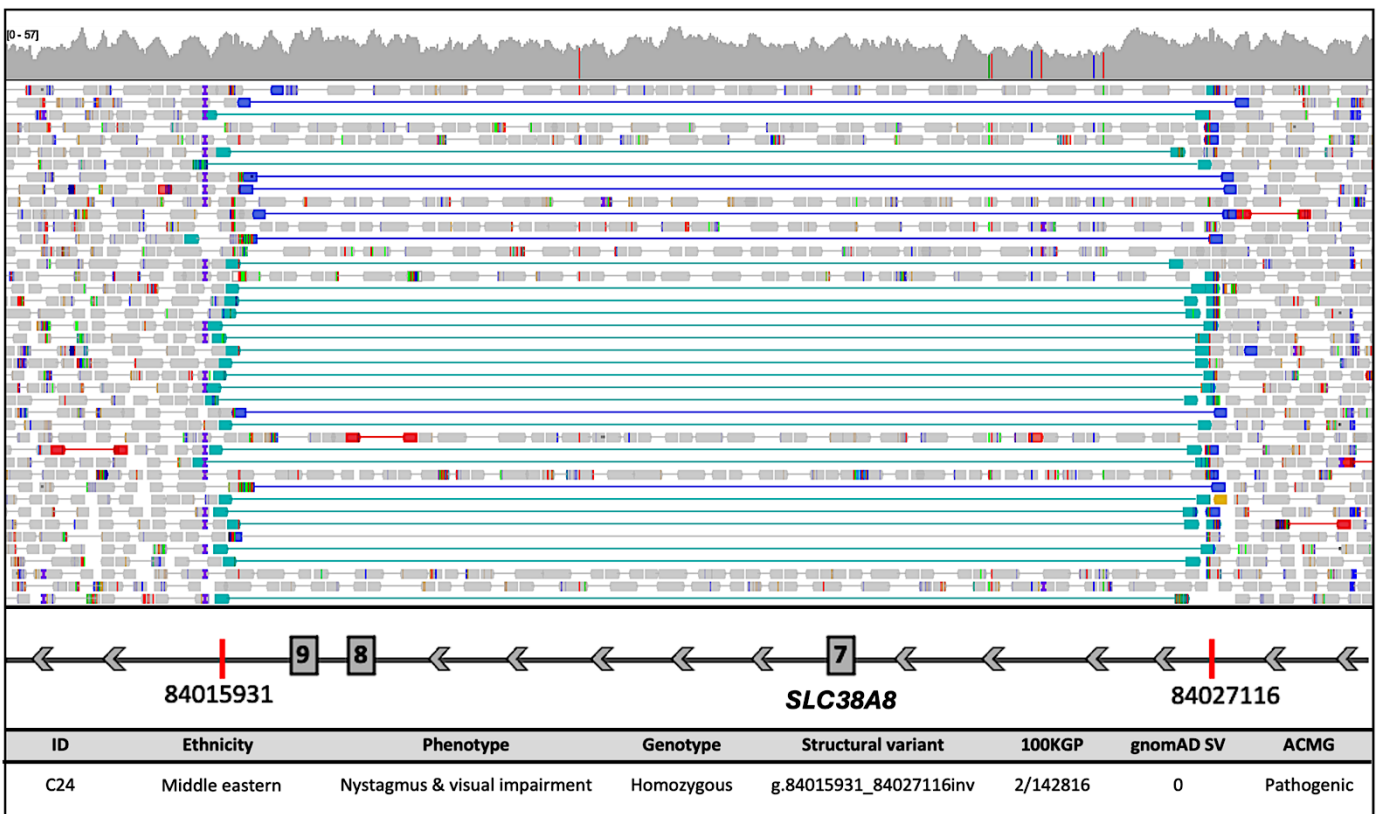
### 3.2.2.5 Detecting SVs in *SLC38A8*

SVs are often undetected unless dedicated tools are used to evaluate changes in coverage indicative of copy number variation (CNV) or to detect discordant read pairs (aberrant alignment) consistent with inversion or translocation events. The 100KGP provides genomic files (VCF) of SVs called individually by the bioinformatics tools MANTA for SV discovery and CANVAS for CNV identification. However, the data is difficult to interpret without annotation. To overcome this, collaborating researcher Dr Jing Yu developed a structural variant discovery tool known as SVRare for the analysis of all the protein coding genes in the 100KGP (Yu et al., 2022). SVRare annotated and aggregated findings for 554,060,126 SVs (CNVs and inversions) called by CANVAS and MANTA in 71,408 participants belonging to the rare disease cohort. These SVs are characterised by SVRare based on their ability to disrupt the coding sequence, the underlying genes affected and the allele frequencies in the 100KGP. At the time of the analysis, the script was not made available to GECIP members. However, the unfiltered output files for *SLC38A8* were shared for a gene specific analysis. These output files for SVRare were available in the research environment.

SVRare has identified 49 SVs in *SLC38A8* belonging to 663 participants. These SVs were analysed to discard false positive calls and to determine whether, either alone or in combination with existing variants, they could be the cause of FH in participants. Two SVs were discarded for having an allele count >60. Of the remaining 47 SVs, 21 disrupted the coding sequence (CDS) of *SLC38A8* and these were taken forward for further analysis. Participants with a heterozygous SV in the *SLC38A8* CDS were assessed for second hits using Gene-Variant Workflow v1.6 for pathogenic SNVs and using SVRare to screen for compound heterozygous SV. Candidate SVs were manually inspected using participant's BAM file for validation of variant calls.

One participant (C24) was homozygous for the inversion NC\_000016.10:g.84015931\_84027116inv (Figure 3.5). This was reported in only a single participant in the 100KGP, with a phenotype of infantile nystagmus.

Interestingly, this SV was also discovered in our local cohort in a proband (F1310) of the same ethnicity (Figure 3.1), and through subsequent discussion with the clinician involved, it was later confirmed that this is the same case (discussed in section 6.5). The SV was verified upon inspection of the participant's genomic file (BAM). This 11,185 bp inversion encompasses exon 7, 8 and 9 of *SLC38A8*, accounting for 33% of the entire gene. No further potentially pathogenic *SLC38A8* allele combinations including SVs were detected in this analysis.



**Figure 3.5. *SLC38A8* Inversion detected in the 100KGP.** Diagram displays paired read alignments in the BAM file of a participant with infantile nystagmus. Primers in turquoise (forward) and blue (reverse) are primer pairs of the same orientation due to the inverted sequence. The *SLC38A8* gene schematic underneath the sequence alignment shows the inversion breakpoints and the exons affected. Below the genomic track is the tabulated clinical and genomic findings in the 100KGP for this patient along with the population data.

### 3.2.2.6 Summary of the 100KGP findings

Analysis of the *SLC38A8* gene in the 100KGP dataset has delivered a genomic diagnosis in four participants with ocular conditions consistent with a diagnosis of isolated FH, the condition known as FVH2. These four participants harbour pathogenic variants in *SLC38A8*, comprising nonsense variants NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) and NM\_001080442.3:c.435G>A; NP\_001073911.1:p.(Trp145\*), a splice site variant NM\_001080442.3:c.632+1G>A and a large inversion NC\_000016.10:g.84015931\_84027116inv. Three of these variants have since been published by others (Schiff et al., 2021) though all were novel when first identified in this study. The *SLC38A8* inversion detected in the 100KGP has been previously identified in our local cohort. Analysis of the monoallelic cohort and subsequent non-coding genome investigations for a second hit did not detect further biallelic pathogenic variants.

### 3.2.3 Defining an *SLC38A8* mutation spectrum

#### 3.2.3.1 Interrogating the published literature

In order to facilitate variant interpretation, the following work endeavoured to capture all *SLC38A8* variants implicated in FVH2 according to the published literature, 100KGP and in the local solved cohort. A better understanding of the spectrum of known pathogenic *SLC38A8* variants will aid in variant interpretation using ACMG classification and can provide additional insights into the pathogenic mechanism underlying FVH2 in these patients. Published *SLC38A8* variants were curated from 11 peer-reviewed publications and 2 conference abstracts and were reanalysed using *In silico* tools and in light of frequency data from the online databases of gnomAD and the 100KGP. Clinical details of probands were recorded for systemic evaluation of the *SLC38A8* phenotype, along with the proband's ethnicity and other supplementary data.

To date, 47 unique *SLC38A8* variants have been reported in the published literature in 59 FVH2 affected families (Table 3.4). All 47 variants are analysed and listed according to HGVS nomenclature except for two uncharacterised deletions. The deletions with undefined breakpoints are,

NM\_001080442.3:c.(?\_1)\_(189+1\_190-1)del for exon 1 deletion, NM\_001080442.3:c.(805+1\_806-1)\_(953+1\_954-1)del for exon 7 deletion, two entries for NM\_001080442.3:c.(805+1\_806-1)\_(1162+1\_1163-1)del for exon 7-8 deletion and NM\_001080442.3:c.(189+1\_190-1)\_(690+1\_691-1)del for exon 2-5 deletion.

All of these 47 variants have been reclassified according to ACMG guidelines for a standardised interpretation. This led to the revision of 9 published pathogenic or likely pathogenic variants as VUSs despite the pathogenicity assessment tools predicting a deleterious effect. These variants are NM\_001080442.3:c.101T>G; NP\_001073911.1:p.(Met34Arg), NM\_001080442.3:c.692G>A; NP\_001073911.1:p.(Cys231Tyr), NM\_001080442.3:c.1078\_1104del; NP\_001073911.1:p.(Ala360\_Leu368del), NM\_001080442.3:c.855G>C; NP\_001073911.1:p.(Leu285Phe), NM\_001080442.3:c.1126G>A; NP\_001073911.1:p.(Gly376Arg), NM\_001080442.3:c.922A>G; NP\_001073911.1:p.(Thr308Ala), NM\_001080442.3:c.527C>G; NP\_001073911.1:p.(Thr176Arg), NM\_001080442.3:c.923C>G; NP\_001073911.1:p.(Thr308Ser) and NM\_001080442.3:c.845\_847delCTG; NP\_001073911.1:p.(Ala282del). Moreover, a VUS, NM\_001080442.3:c.1214+5G>C has been reclassified as likely pathogenic in this analysis based on additional supporting evidence of co-segregation with disease and being in trans with a pathogenic variant. The concordance between the published interpretation (Schiff et al., 2021, Kuht et al., 2020, Lasseaux et al., 2018) and the ACMG classification for the *SLC38A8* variants in FVH2 was 33% (10/30) (Appendix C: Supplementary Table 3.4).

The addition of 4 novel variants from our solved local cohort (Figure 3.1) to the list of 47 published *SLC38A8* variants expands the mutation spectrum in FVH2 to 51 variants in 63 families (Table 3.4). According to ACMG classification this definitive list of *SLC38A8* variants comprised of 17 VUS, 20 likely pathogenic and 13 pathogenic variants. A single uninformative variant, a “large deletion” was not classified by ACMG. This large cohort of variants in patients with a confirmed genomic diagnosis of FVH2 provides a reliable dataset for downstream analysis to investigate the genotype-phenotype correlation in *SLC38A8*.

Cohort	Variant	Variant (p.annotation)	gnomAD	100KGP	Reference	ACMG
Local cohort	NC_000016.10:g.84015973_84027074inv	NP_001073911.1:p.?	0	1/71408	(Lord, 2018)	Pathogenic
Local cohort	NC_000016.10:g.84034763_84037497del	NP_001073911.1:p.?	0	0	This thesis	Pathogenic
Local cohort	NC_000016.10:g.84006453_84019002del	NP_001073911.1:p.?	0	0	This thesis	Pathogenic
Local cohort	NM_001080442.3:c.669delC	NP_001073911.1:p.(Thr224Profs*44)	1/1614052	0	This thesis	Likely pathogenic
Literature and local cohort	NM_001080442.3:c.848A>C	NP_001073911.1:p.(Asp283Ala)	291/1613852 (4)	4/126712	(Toral et al., 2017, Lasseaux et al., 2018, Ehrenberg et al., 2021, Schiff et al., 2021, Kruijt et al., 2022)	Likely pathogenic
Literature and 100KGP	NM_001080442.3:c.264C>G	NP_001073911.1:p.(Tyr88*)	32/1613888	16/126712	(Chaudhuri et al., 2018, Schiff et al., 2021)	Pathogenic
Literature and 100KGP	NM_001080442.3:c.435G>A	NP_001073911.1:p.(Trp145*)	1/1614032	3/126712	(Schiff et al., 2021)	Pathogenic
Literature and 100KGP	NM_001080442.3:c.632+1G>A	NP_001073911.1:p.?	3/1613894	2/126712	(Schiff et al., 2021)	Pathogenic
Literature	NM_001080442.3:c.922A>G	NP_001073911.1:p.(Thr308Ala)	800/1613912 (3)	61/126712	(Lasseaux et al., 2018)	VUS
Literature	NM_001080442.3:c.534C>G	NP_001073911.1:p.(Ile178Met)	18/1614064	4/126712	(Campbell et al., 2019)	Likely pathogenic
Literature	NM_001080442.3:c.1002delG	NP_001073911.1:p.(Ser336Alafs*15)	5/1613296	0	(Poulter et al., 2013, Kruijt et al., 2022)	Pathogenic
Literature	NM_001080442.3:c.101T>G	NP_001073911.1:p.(Met34Arg)	0	0	(Poulter et al., 2013, Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.1029delG	NP_001073911.1:p.(Leu344Cysfs*7)	7/1613788	0	(Poulter et al., 2013)	Likely pathogenic
Literature	NM_001080442.3:c.1078_1104del	NP_001073911.1:p.(Ala360_leu368del)	0	0	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.1126G>A	NP_001073911.1:p.(Gly376Arg)	18/1613954	1/126712	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.1214+5G>C	NP_001073911.1:p.?	0	0	(Kuht et al., 2020)	Likely pathogenic
Literature	NM_001080442.3:c.1234G>A	NP_001073911.1:p.(Gly412Arg)	4/1613868	2/126712	(Poulter et al., 2013, Kruijt et al., 2022)	Likely pathogenic
Literature	NM_001080442.3:c.1256G>T	NP_001073911.1:p.(Gly419Val)	0	0	(Kruijt et al., 2022)	VUS
Literature	NM_001080442.3:c.160G>T	NP_001073911.1:p.(Gly54*)	13/1609190	0	(Ehrenberg et al., 2021, Kruijt et al., 2022)	Pathogenic
Literature	NM_001080442.3:c.260C>T	NP_001073911.1:p.(Thr87Ile)	2/1613980	0	(Kruijt et al., 2022)	VUS
Literature	NM_001080442.3:c.269G>T	NP_001073911.1:p.(Gly90Val)	33/1613958	1/126712	(Ehrenberg et al., 2021)	VUS
Literature	NM_001080442.3:c.388+5G>A	NP_001073911.1:p.?	47/1614000 (1)	1/126712	(Ehrenberg et al., 2021, Kruijt et al., 2022)	Likely pathogenic
Literature	NM_001080442.3:c.490_491delICT	NP_001073911.1:p.(Leu164Valfs*41)	2/1614070	0	(Weiner et al., 2020)	Pathogenic
Literature	NM_001080442.3:c.527C>G	NP_001073911.1:p.(Thr176Arg)	4/1614012	0	(Lasseaux et al., 2018)	VUS
Literature	NM_001080442.3:c.558C>A	NP_001073911.1:p.(Tyr186*)	2/1614222	0	(Kuht et al., 2020)	Likely pathogenic
Literature	NM_001080442.3:c.598C>T	NP_001073911.1:p.(Gln200*)	14/1614068	0	(Poulter et al., 2013, Kruijt et al., 2022)	Pathogenic
Literature	NM_001080442.3:c.6_9delIGGGA	NP_001073911.1:p.(Glu2Aspfs*32)	0	0	(Lasseaux et al., 2018)	Likely pathogenic
Literature	NM_001080442.3:c.632+2T>G	NP_001073911.1:p.?	20/1613580	1/126712	(Kuht et al., 2020)	Likely pathogenic
Literature	NM_001080442.3:c.644G>T ^	NP_001073911.1:p.(Trp215Leu) ^	12/1614084	0	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.682G>A ^	NP_001073911.1:p.(Gly228Arg) ^	46/1613894	7/126712	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.676T>C	NP_001073911.1:p.(Cys226Arg)	0	0	(Ehrenberg et al., 2021)	VUS
Literature	NM_001080442.3:c.692G>A	NP_001073911.1:p.(Cys231Tyr)	1/1589962	0	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.695A>G	NP_001073911.1:p.(His232Arg)	3/1601784	0	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.697G>A	NP_001073911.1:p.(Glu233Lys)	49/1602028	0	(Poulter et al., 2013, Lasseaux et al., 2018, Ehrenberg et al., 2021, Kruijt et al., 2022)	Likely pathogenic
Literature	NM_001080442.3:c.698A>G	NP_001073911.1:p.(Glu233Gly)	2/1603304	0	(Schiff et al., 2021)	Likely pathogenic
Literature	NM_001080442.3:c.707T>A	NP_001073911.1:p.(Val236Asp)	0	0	(Poulter et al., 2013)	Likely pathogenic
Literature	NM_001080442.3:c.800T>G	NP_001073911.1:p.(Leu267Arg)	0	0	(Kruijt et al., 2022)	VUS
Literature	NM_001080442.3:c.845_847delCTG	NP_001073911.1:p.(Ala282del)	2/1613970	0	(Poulter et al., 2013, Kruijt et al., 2022)	VUS
Literature	NM_001080442.3:c.855G>C	NP_001073911.1:p.(Leu285Phe)	0	0	(Kuht et al., 2020)	VUS
Literature	NM_001080442.3:c.923C>G	NP_001073911.1:p.(Thr308Ser)	4/1614016	0	(Lasseaux et al., 2018, Schiff et al., 2021)	VUS
Literature	NM_001080442.3:c.954-1G>C	NP_001073911.1:p.?	0	0	(Kuht et al., 2020)	Pathogenic
Literature	NM_001080442.3:c.95T>G	NP_001073911.1:p.(Ile32Ser)	7/1613980	0	(Perez et al., 2014, Weiner et al., 2020, Ehrenberg et al., 2021, Kruijt et al., 2022)	Likely pathogenic
Literature	NM_001080442.3:c.964C>T	NP_001073911.1:p.(Gln322*)	37/1612362	2/126712	(Kuht et al., 2020)	Pathogenic
Literature	NM_001080442.3:c.995dupG	NP_001073911.1:p.(Trp333Metfs*35)	30/1613022	0	(Kuht et al., 2020)	Pathogenic
Literature	NM_001080442.3:c.(805 + 1_806-1)_(1162 + 1_1163-1)del	NP_001073911.1:p.?	0	0	(Kruijt et al., 2022)	Likely pathogenic
Literature	NM_001080442.3:c.(?_1)_(189+1_190-1)del	NP_001073911.1:p.?	-	-	(Kuht et al., 2020)	Likely pathogenic
Literature	NM_001080442.3:c.(805+1_806-1)_(953+1_954-1)del	NP_001073911.1:p.?	-	-	(Lasseaux et al., 2018)	Likely pathogenic
Literature	NM_001080442.3:c.(805+1_806-1)_(1162+1_1163-1)del	NP_001073911.1:p.?	-	-	(Schiff et al., 2021)	Likely pathogenic
Literature	NM_001080442.3:c.(189+1_190-1)_(690+1_691-1)del	NP_001073911.1:p.?	-	-	(Campbell et al., 2019)	Likely pathogenic
Literature	Large deletion	NP_001073911.1:p.?	-	-	(Poulter et al., 2013, Kruijt et al., 2022)	-
Literature	Whole gene deletion	NP_001073911.1:p.?	-	-	(Ehrenberg et al., 2021, Kruijt et al., 2022)	Likely pathogenic

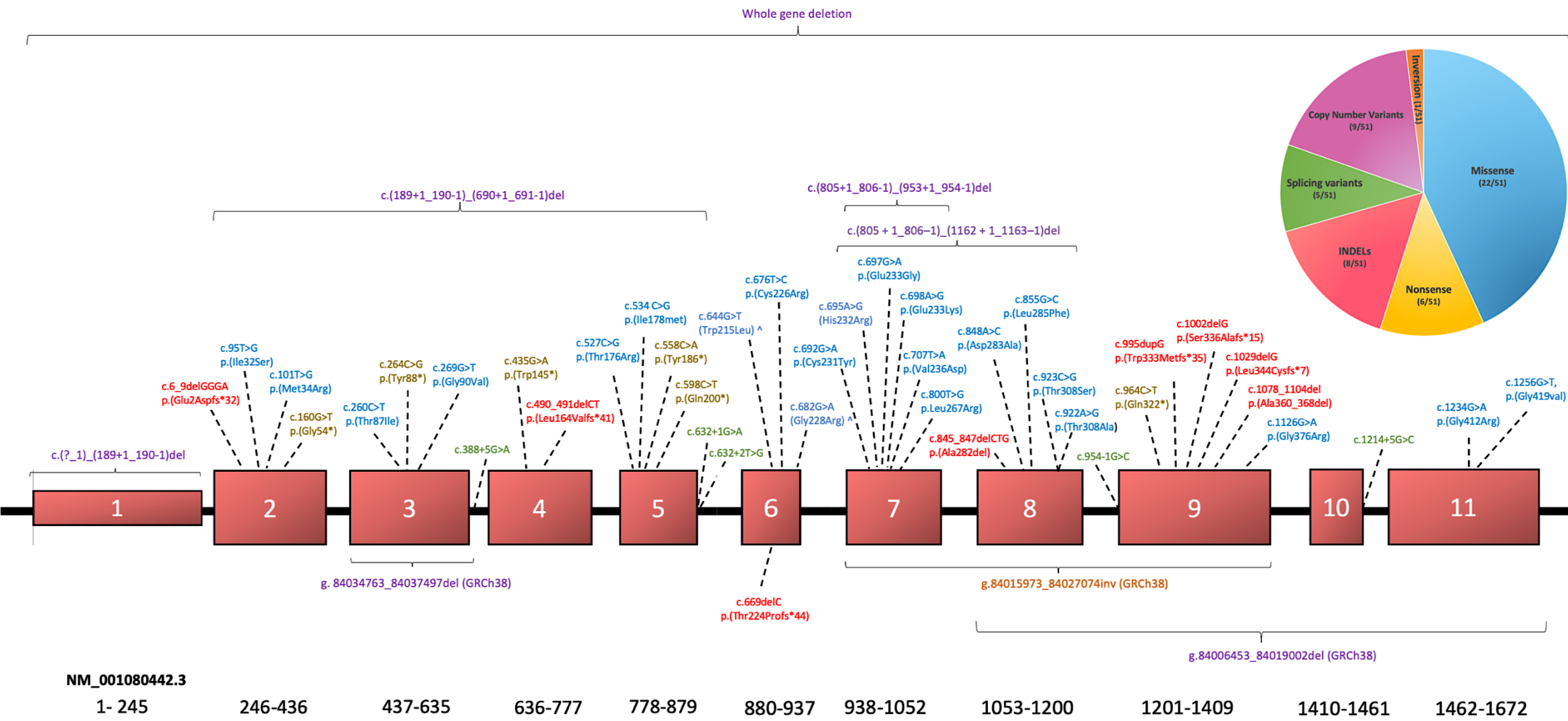
**Table 3.4. All SLC38A8 variants identified in FVH2.** 51 unique variants curated from the local cohort (4 variants), 100KGP and published literature (47 variants). Unknown consequences on the protein is represented by p.?. The “^” denotes variants being in cis. The homozygote count is displayed in brackets.

### 3.2.3.2 *SLC38A8* mutation spectrum in FH

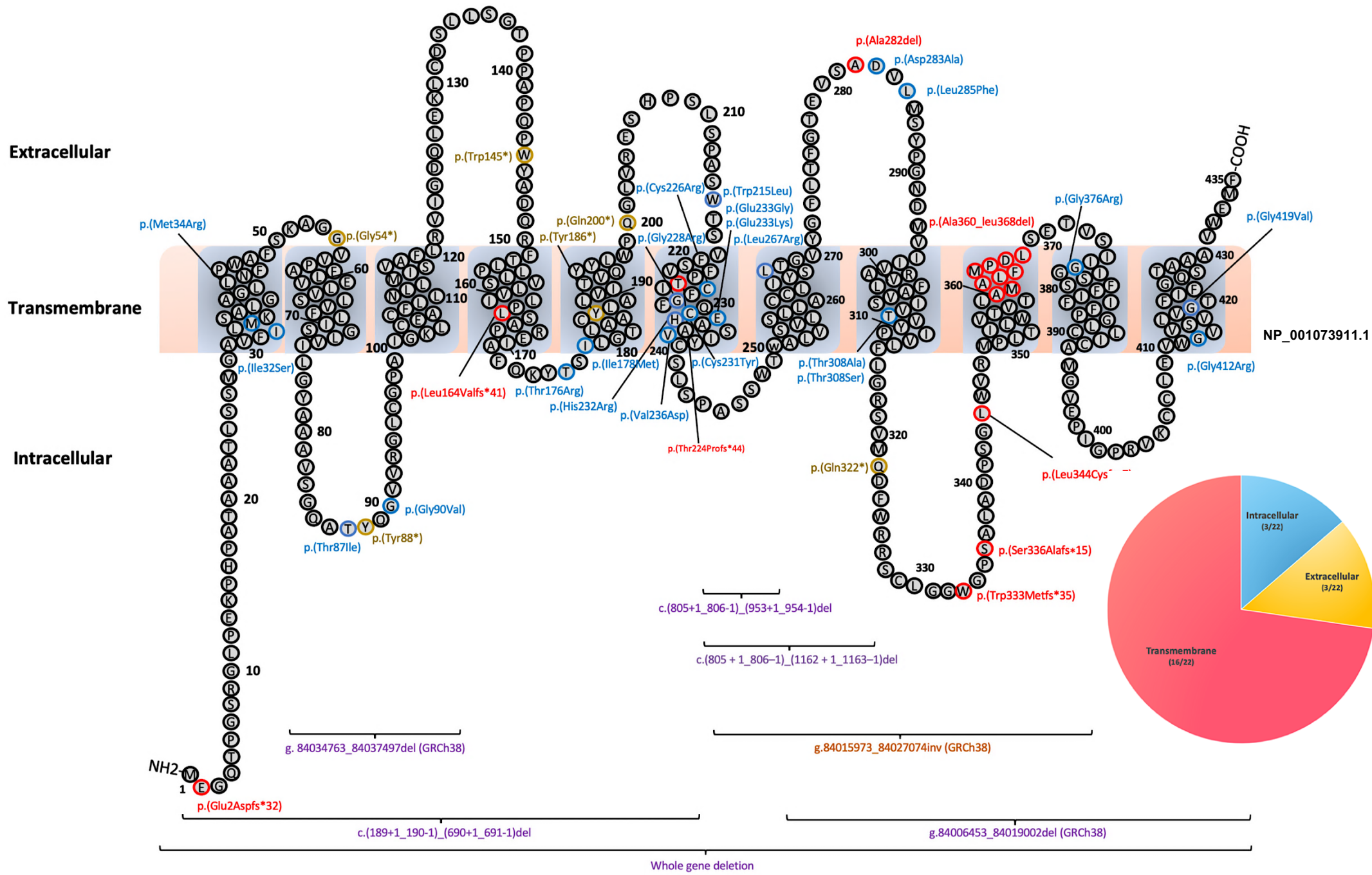
In total, 51 *SLC38A8* variants causing FVH2 were reported in the literature, our local cohort and the 100KGP (12/11/2023) (Table 3.4). These constitute 22 missense, 6 nonsense, 8 INDELS, 5 splicing and 10 SVs and CNVs. Missense variants are the largest category of *SLC38A8* variants responsible for FVH2 (43%), while high impact variants (nonsense, frameshift INDELS, splice donor or acceptor variants and SVs) collectively account for 49%. The remaining 8% of variants in FVH2 are attributed to in-frame deletions and splice region variants (>2 bp from canonical splice sites). All 51 variants are distributed throughout the gene, targeting all 11 exons of *SLC38A8*, with no indication of a mutational hotspot at the DNA level (Figure 3.6). Only one variant has been reported in the non-coding exon 1, that being a complete deletion of the exon, which is likely to impact regulation of gene expression.

*SLC38A8* is a member of the sodium-coupled neutral amino acid transporter family and the topology of the reference *SLC38A8* protein was predicted according to the MANE transcript (NP\_001073911.1), using PROTTER (<https://wlab.ethz.ch/protter/start/>). Mapping the deleterious variants onto the affected amino acid residues provided insight into protein domains disrupted (Figure 3.7). The protein structure of *SLC38A8* is of 435 amino acids and is oriented with the N-terminus within the intracellular cytosol and the C-terminus in the extracellular space. The protein is mainly composed of 11 transmembrane domains spanning 233 amino acids which make up the pore channel. The transmembrane domain is the largest component of *SLC38A8*, and it harbours 16 missense variants and 1 inframe INDEL.

High impact variants like nonsense, frameshift and splice variants are likely to abolish the mRNA transcript through NMD resulting in no *SLC38A8* generated. Based on this rationale, the functional analysis in *SLC38A8* focused on missense changes. Missense variants are scattered across the intracellular and extracellular domains, but tend to cluster within the transmembrane domains (72.7%). A high density of missense variants (7) is detected in the 6<sup>th</sup> transmembrane domain and this equates to 32% (7/22) of missense variants in 5.3% (23/435 aa) of the protein.



**Figure 3.6. SLC38A8 mutation spectrum.** Gene Schematic capturing to date 51 damaging SLC38A8 variants implicated in FH. These are extracted from the reported literature (47) and our local cohort (4). 7/9 copy number variations have no defined breakpoints, and one “large deletion” cannot be mapped due to insufficient information from the publication. Two undefined CNVs affect exon 7-8, c.(805+1\_806-1)\_(1162+1\_1163-1)del. Variants extracted from the literature are positioned above the gene while novel variants reported from our local cohort are labelled below. Variants are colour coded according to their molecular consequences, frameshifts (red), missense (blue), nonsense (golden yellow), splicing variants (green), copy number variations (purple) and inversions (orange). The reference sequence (NM\_001080442.3) is presented as a scale at the bottom of the figure. ^ denotes variants in cis.



**Figure 3.7. Distribution of pathogenic variants in SLC38A8.** The protein is composed of 11 transmembrane domains, 5 intracellular and 5 extracellular loops. The pie chart shows the localisation of missense variants at different protein domains. The reference protein sequence (NP\_001073911.1) is illustrated using circles to represent each residue, spanning 1-435 amino acids. Splice variants and a “large deletion” are excluded.

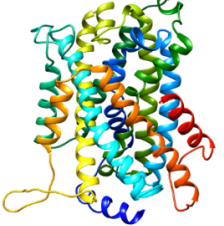

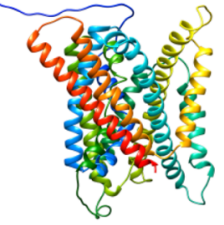
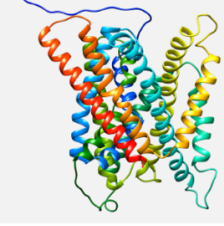


### 3.2.4 Computational protein modelling of SLC38A8 and missense variants simulation

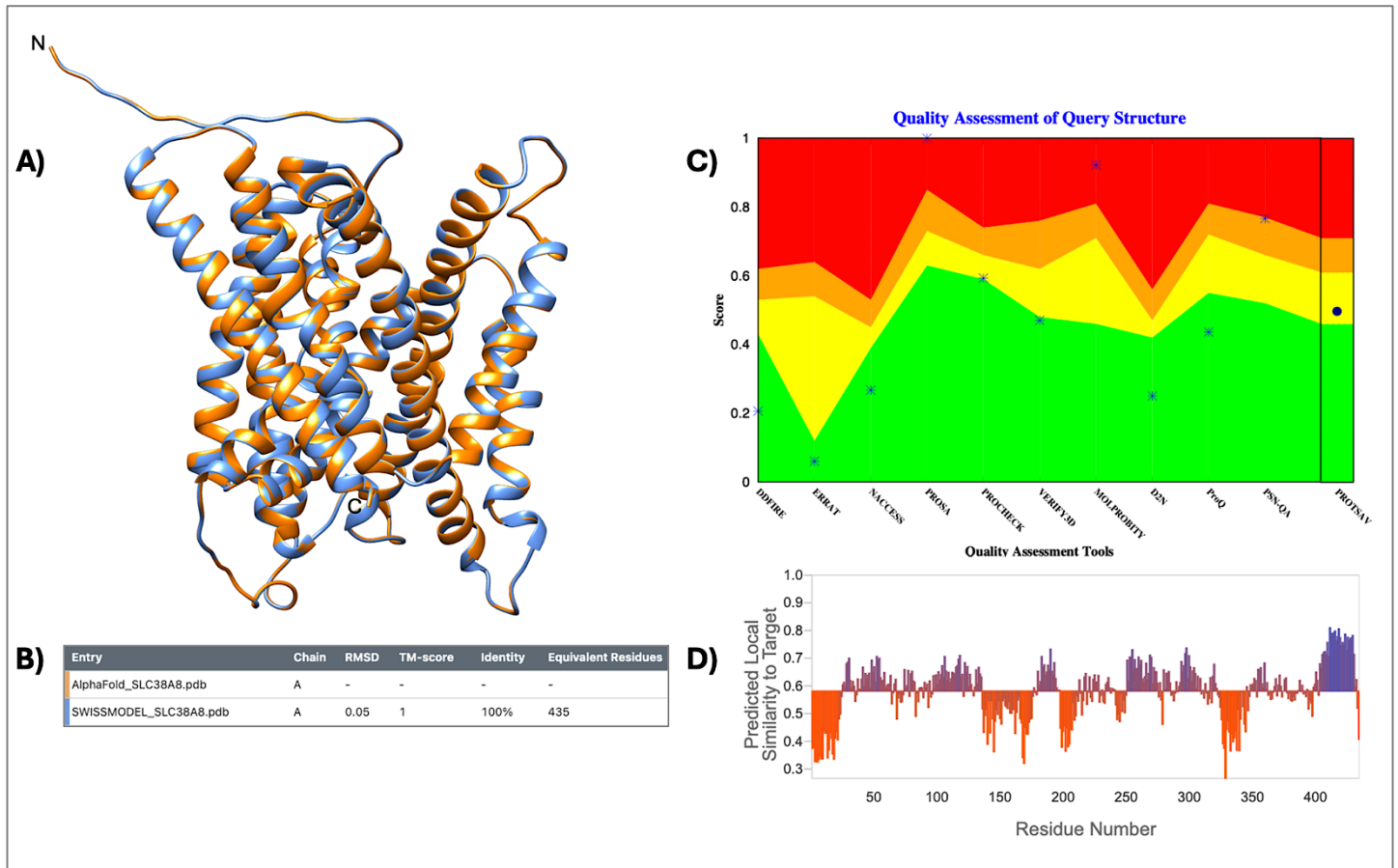
#### 3.2.4.1 Generation of an SLC38A8 wildtype and mutant models

*In-silico* analysis based on SLC38A8 modelling was used to determine the possible consequences of missense variants. The lack of an existing SLC38A8 protein model on RCSB-Protein Data Bank (RCSB-PDB) necessitated a study to generate a reliable SLC38A8 model using the amino acid sequence of the MANE transcript (NP\_001073911.1). Candidate protein models were generated by homology modelling using Phyre2, I-TASSER and SWISS-MODEL as described in section 2.6.15.1. The candidate protein models were validated via structural alignment against a pre-existing AlphaFold2 model (A6NNN8.1.A) to determine the most homologous structure (Table 3.5). The SWISS-MODEL generated the most reliable SLC38A8 structure covering all 435 amino acids and yields 100% homology with AlphaFold2 (Figure 3.8).

The SWISS-MODEL outperformed the reference AlphaFold2 model and the other candidate models, achieving the lowest MolProbity score (1.03) for a more accurate predicted protein (Table 3.5). Quality assessment using ProtSAV generated a score of 2-5 Å, qualifying the model for interrogating variants and in evaluating functional interactions. The SWISS-MODEL also achieved the highest QMEANDisCo quality score of 0.58 to imply a good quality structure. The SWISS-MODEL SLC38A8 structure was selected to represent the wildtype SLC38A8 protein and was subsequently manipulated on UCSF Chimera to mutate 22 target residues corresponding to the reported *SLC38A8* missense changes causative for FVH2.

Models	Length (aa)	Alignment to reference model	ProtSAV	MolProbity	QMEANDisCo	3D Structure
Phyre2	418	IDDT: 0.29 TM-Score: 0.74 RMSD: 7.58	0-2 Å	MolProbity: score 3.21 Clash score: 151.55 Ramachandran favored: 89.3 % Ramachandran outlier: 2.68% Rotamer outliers: 0.29% Bad angles: 188/4424 Bad bonds: 39/3240	0.39	
I-TASSER	435	IDDT: 0.51 Tm-score: 0.80 RMSD: 8.35	0-2 Å	MolProbity: 2.97 Clash score: 7.05 Ramachandran favored: 72.7 % Ramachandran outlier: 12.3% Rotamer outliers: 10.1% Bad angles: 67/4294 Bad bonds: 3/3272	0.52	
SWISS-MODEL	435	IDDT: 1 TM-Score: 1 RMSD: 0.05	2-5 Å	MolProbity: 1.03 Clash score: 0.30 Ramachandran favored: 94 .5% Ramachandran outlier: 0.92% Rotamer outliers: 1.12% Bad angles: 21/4615 Bad bonds: 0/3376	0.58	
AlphaFold2	435	IDDT: 1 TM-Score: 1 RMSD: 0	2-5 Å	MolProbity: 1.29 Clash score: 0.90 Ramachandran favored: 94 % Ramachandran outlier: 1.62% Rotamer outliers: 1.40% Bad angles: 12/4615 Bad bonds: 0/3376	0.58	

**Table 3.5. Comparison of candidate SLC38A8 models.** SWISS-MODEL generated the most reliable SLC38A8 structure based on quality assessment. Homology to reference AlphaFold2 model is inferred using local distance difference test (IDDT), template modelling score (TM-score) and root mean square deviation (RMSD). Quality assessment of SLC38A8 models used ProtSAV, MolProbity and QMEANDisCo. Angstrom (Å) is a measure of length representing  $10^{-10}$  metre.



**Figure 3.8. Quality Assessment of SWISS-MODEL SLC38A8 protein structure.** A) Superimposed protein structures of SWISS-MODEL and reference AlphaFold2 model. B) Table displaying a 100% identity between the superimposed structures with an RMSD of 0.05 and TM-score of 1, indicating a perfect match. C) ProtSAV plot with individual scores for 10 stand-alone tools and a consolidated score of 2-5 Å for ProtSav. The ProtSAV scores are depicted as, 0–2 Å in green region, 2–5 Å in yellow, 5–8 Å in orange and >8 Å in red. D) QMEANDisCo quality score plot for each amino acid residue is displayed with a middle line depicting the average quality score of the entire structure at 0.58.

### 3.2.4.2 Comparative analysis of pathogenic missense variants and common polymorphisms in FVH2

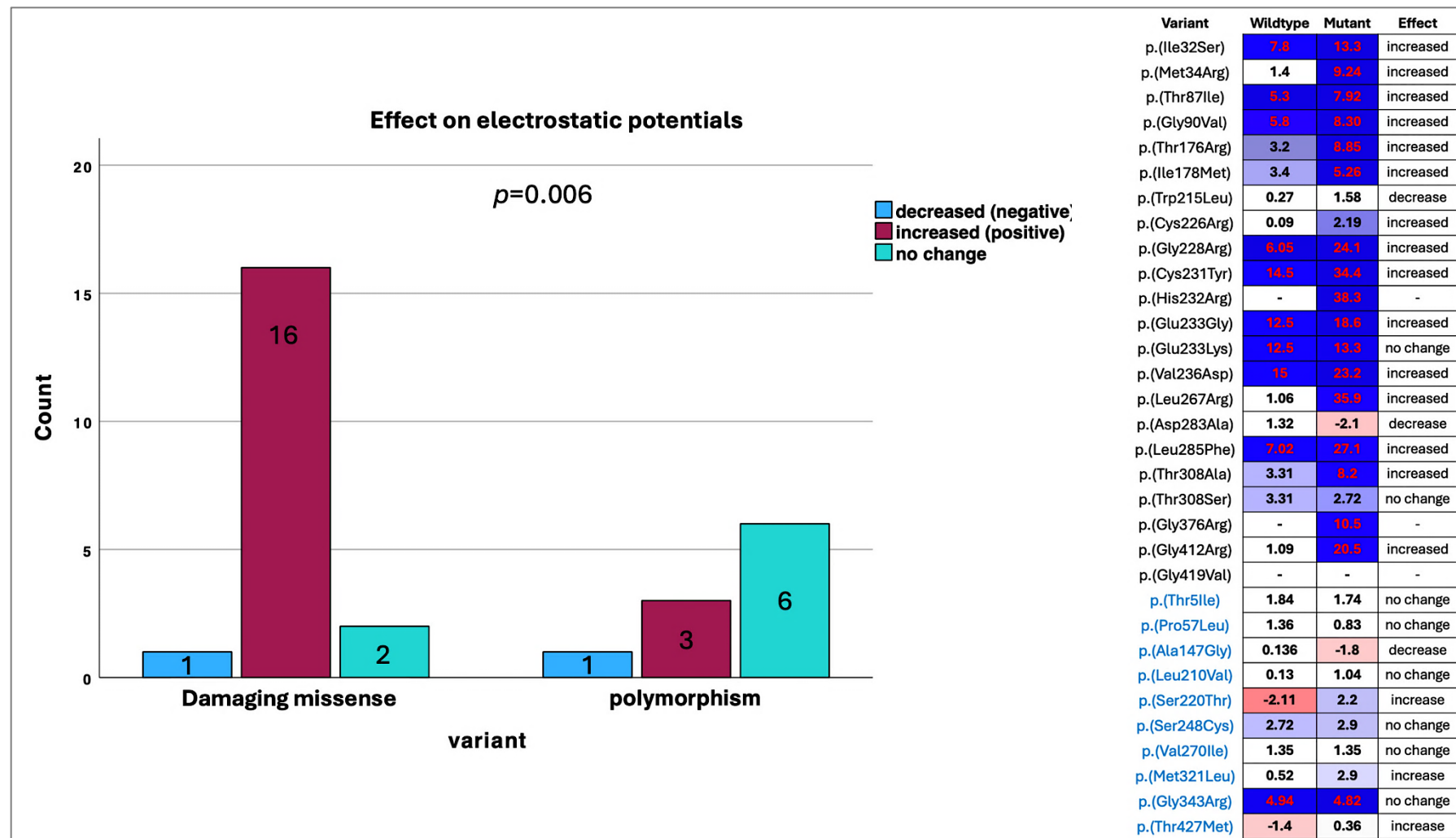
To determine why certain *SLC38A8* missense variants cause FVH2, variant simulations were performed using the SWISS-MODEL generated *SLC38A8* structure. The *In-silico* experiment contrasted mutant protein structures generated by the 22 missense variants causative for FVH2 against those with the 10 most common and benign *SLC38A8* amino acid substitutions found in the gnomAD database. The parameters selected in the analysis includes amino acid conservation, protein stability, hydrophobicity profile, electrostatic potentials, residue exposure levels and solvent accessibility (Figure 3.9). The raw data for each metric is also provided in the Appendix C: Supplementary Figure 3.2. These analyses showed that the deleterious missense variants usually affect conserved residues (86%), have a higher incidence of destabilizing effects (55%) and structural damage (50%), cause severe changes in hydrophobicity (68%) and affect residues that are electropositive (68%) resulting in an increased electrostatic potential (84%). In comparison, the polymorphisms affect residues of varying evolutionary conservation with no preference (50% neutral, 30% conserved and 20% variable) are less likely to destabilize the protein (20%) or cause structural damage (10%), are less prone to cause profound changes in hydrophobicity (30%) or electrostatic potentials (40%). The association between pathogenic missense variants and an increase in the electrostatic potentials was statistically validated using the Fisher Exact test (Figure 3.10). The *p* value obtained for this small cohort was 0.006 which allows us to reject the null hypothesis. The results therefore prove that there is a significant correlation whereby missense variants are more likely to increase the electrostatic potentials.

Localisation of the deleterious missense variants and polymorphisms on the *SLC38A8* 3D structure was also explored to elucidate which regions of the folded protein were more likely to be affected by each category of variants (Figure 3.11). The data showed that the pathogenic variants aggregated around the pore (7/22), inside the pore (5/22) or within deep structures (6/22). On the other hand, the common polymorphisms did not show any consistent localisation bias but were limited to the surface and away from the pore (10/10).

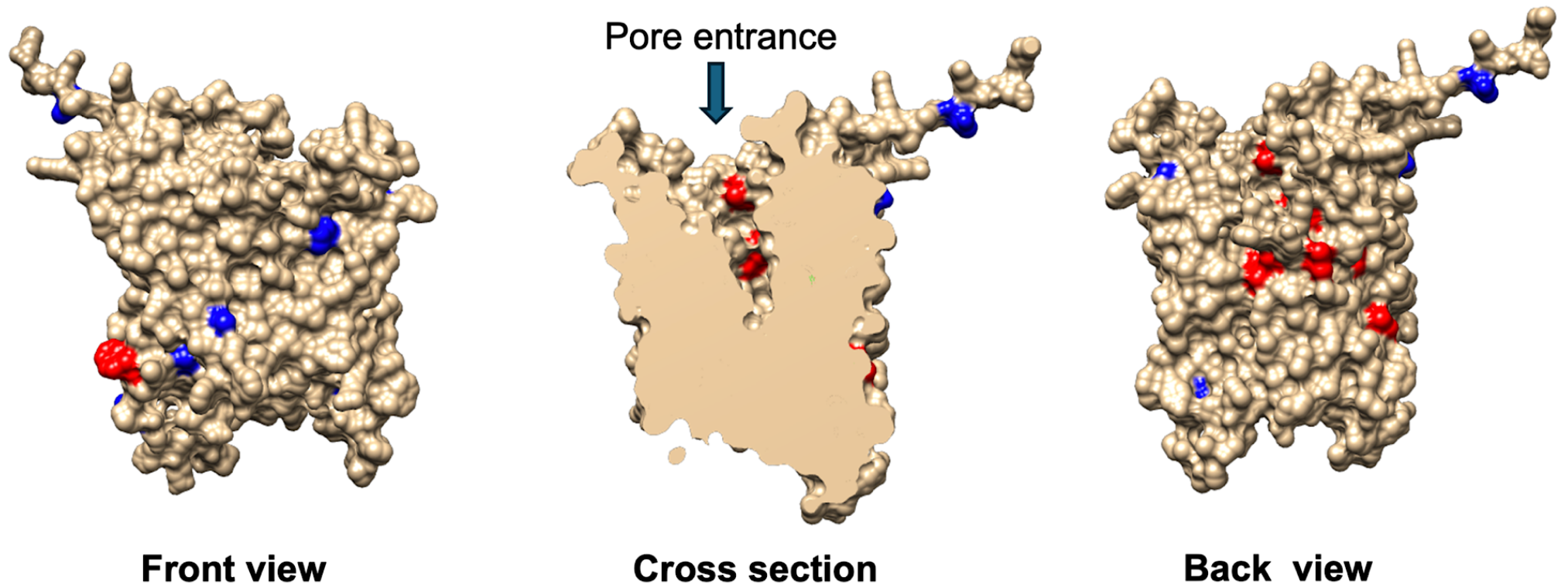
Missense	Conservation	Protein Stability (kcal/mol)	Residue Exposure (Å)	Solvent Accessibility (%)	Hydrophobicity	Electrostatics potential kcal/(mol*e)	Structural Impact
p.(Ile32Ser)	9	-2.313	-0.7	+5.1	-5.3	+5.5	Cavity expansion
p.(Met34Arg)	8	0.077	0	+9.9	-6.4	7.84	
p.(Thr87Ile)	8	0.041	+0.2	-10.6	+5.2	-2.62	
p.(Gly90Val)	7	1.549	-1.1	+6.7	+4.60	-2.52	Buried residue replaced
p.(Thr176Arg)	8	1.040	+2	+3.2	-5.20	+5.65	Cavity expansion
p.(Ile178Met)	4	-0.152	-0.1	-2.2	-2.60	+1.86	
p.(Trp215Leu)	4	-0.052	+0.4	+2.5	+4.7	1.31	
p.(Cys226Arg)	7	-0.200	-0.3	0	-7	+2.1	Buried residue replaced
p.(Gly228Arg)	6	1.517	+135.7	-1.1	-4.10	+12.22	Buried residue replaced
p.(Cys231Tyr)	9	-0.376	+0.2	-1.1	-3.80	+12.3	
p.(His232Arg)	9	-0.955	-0.3	0	-1.30	-	
p.(Glu233Gly)	9	0.701	-1.5	+16.4	+3.10	+6.1	Cavity expansion
p.(Glu233Lys)	9	0.126	-3	+7.8	-0.40	+6.8	
p.(Val236Asp)	8	-1.917	+1.2	0	-7.70	+6.2	Buried residue replaced
p.(Leu267Arg)	8	-0.223	-1	+9.4	-8.3	+34.84	
p.(Asp283Ala)	9	0.500	+1.2	-0.4	+5.3	+3.42	Buried residue replaced
p.(Leu285Phe)	9	-0.614	-1401.6	+0.6	-1	+20.02	
p.(Thr308Ala)	9	-0.044	-0.6	-4.2	+2.5	+4.89	
p.(Thr308Ser)	9	-0.499	-0.3	-3.7	-0.10	-0.59	
p.(Gly376Arg)	9	0.104	+902.8	0	-4.10	-	Buried residue replaced
p.(Gly412Arg)	7	-0.165	+0.2	0	-4.10	+19.41	Buried residue replaced
p.(Gly419Val)	9	0.848	+0.3	0	+4.60	-	Buried residue replaced
p.(Thr5Ile)	3	0.231	0	+0.2	+3.80	-0.10	
p.(Pro57Leu)	5	1.990	-0.1	+11.7	+5.4	-0.43	
p.(Ala147Gly)	4	-0.636	0	+1.8	-2.20	-1.936	
p.(Leu210Val)	2	-0.188	+0.1	+2.3	+0.40	+0.91	
p.(Ser220Thr)	8	0.746	0	-5.9	+0.10	+4.31	
p.(Ser248Cys)	6	0.060	-0.1	+0.4	+3.30	0.18	
p.(Val270Ile)	5	0.265	+0.4	+3.8	+0.30	0	
p.(Met321Leu)	5	0.046	0	-0.8	+1.90	+2.38	
p.(Gly343Arg)	8	0.519	-0.8	-12.2	-4.10	-5.12	Residue at bend curvature replaced
p.(Thr427Met)	8	0.438	0	+5.2	+2.60	+1.04	



**Figure 3.9. Predicted effects of missense changes and common polymorphisms on SLC38A8.** Mosaic illustration summarises the results of *in silico* assessment of evolutionary conservation, biochemical and structural alterations. The left hand column lists the deleterious missense variants in black font and 10 common polymorphisms extracted from gnomAD in blue. Protein stability effects are inferred using a positive  $\Delta G$  value to imply a stabilising effect and a negative  $\Delta G$  for a destabilising effect. Changes in residue exposure levels greater than 0.1 Å were recorded as significant. Structural damage refers to any of cavity expansion, replacement of a buried residue with an exposed residue or substitutions of charged and hydrophilic residues with uncharged and hydrophobic amino acids. Values presented for the conservation, residue exposure levels, solvent accessibility, hydrophobicity and electrostatic potentials show the changes in measurements between wildtype and mutant amino acid. Conservation profile for each amino acid residue on SLC38A8 follows a scale of 1-9, with cut offs at 1-3 corresponding to variable (turquoise), 4-6 for neither conserved or variable (white) and 7-9 for conserved (maroon). Hydrophobicity scores are according to the Kyte-Doolittle scale with thresholds of -4.50 for hydrophilic (blue), 0 for neutral (white) and 4.50 for hydrophobic (orange). Electrostatic potentials uses Adaptive Poisson-Boltzmann Solver (APBS) thresholds of -5 kcal/(mol\*e) for negative (red), 5 kcal/(mol\*e) for positive (blue) and 0 kcal/(mol\*e) for neutral (white). Colour keys are grouped under the relevant columns with darker colours corresponding to a greater effect on the protein and vice versa.



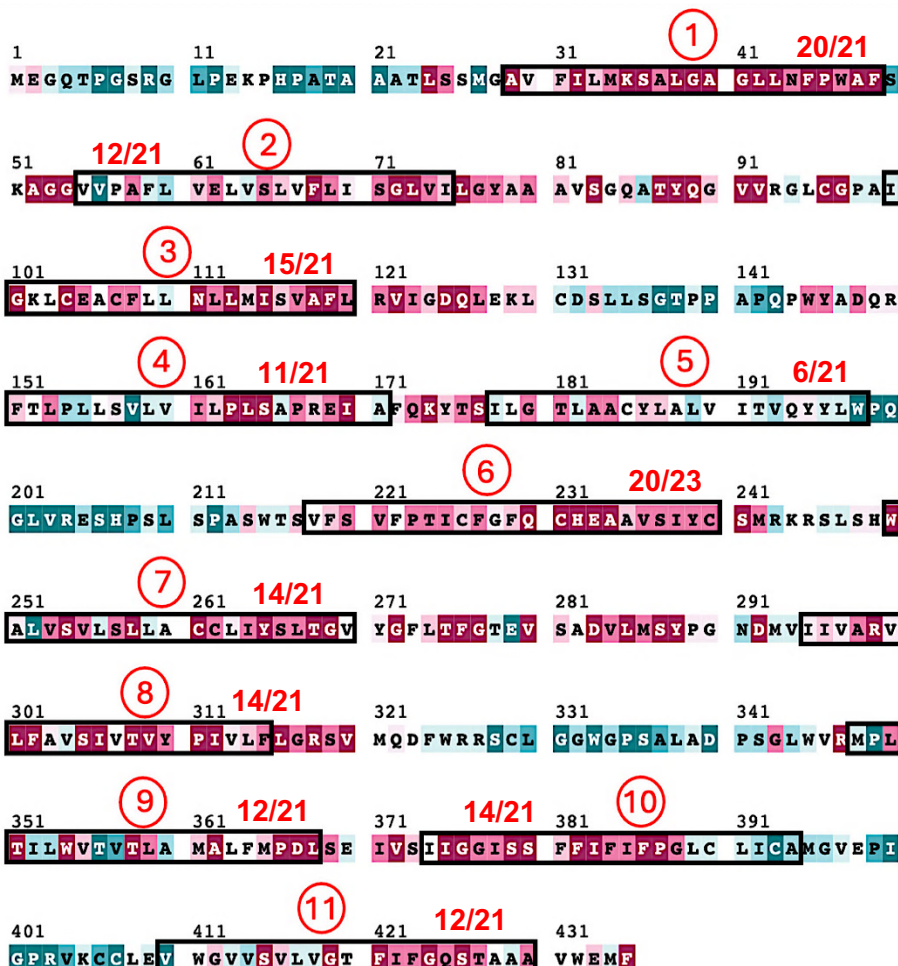
**Figure 3.10. Statistical validation of missense effect on surface electrostatic potentials.** Computational assessment of the effect of 22 deleterious *SLC38A8* missense variants compared with 10 most common polymorphisms extracted from gnomAD. The deleterious missense variants are associated with increased electrostatic potentials while polymorphisms have variable effects. Fisher Exact test generated a probability value of 0.006. Table presents the raw electrostatic measurements obtained by APBS. Significant change is interpreted as  $\geq 1$  kcal/(mol\*e). The values for NP\_001073911.1:p.(Gly419Arg), NP\_001073911.1:p.(Gly376Arg) and NP\_001073911.1:p.(His232Arg) were not available due to a limitation using UCSF Chimera that did not compute the electrostatic potentials at these residues.



**Figure 3.11. Localisation of missense variants in SLC38A8 protein.** Protein model annotated with 22 deleterious missense changes and 10 polymorphisms in different orientations including front, back and cross-sectional view. Pathogenic missense variants are highlighted in red while the polymorphisms are marked in blue. The deleterious missense variants aggregate within the pore or around the opening (82%) while a minority are scattered at the surface (18%). Polymorphisms are positioned away from the pore and are scattered on the surface (100%)

### 3.2.4.3 Assessing SLC38A8 amino acid conservation

To further characterise SLC38A8 and increase our understanding of the transmembrane domains, the study investigated the degree of amino acid conservation across the entire protein (435 aa). Multiple sequence alignments were performed against orthologs in 150 different species for an accurate evolutionary assessment of amino acids (Appendix C: Supplementary Figure 3.3). The species in the phylogenetic tree ranges from mammals through to fish. Each residue was scored 1-9 to generate a reliable conservation profile. SLC38A8 is composed of 244 conserved amino acid residues and 67% (150/244) of these are within the eleven transmembrane domains (Figure 3.12). The data shows that the transmembrane domains show variable degrees of conservation. The first and sixth transmembrane domains are highly conserved, with >87% of their residues are conserved, while the fifth transmembrane domain has more variable residues (<30% conserved). The remainder of the transmembrane domains have intermediate conservations with 52-71% of their residues being conserved. The missense variants in FVH2 tend to aggregate within the sixth transmembrane domain that is highly conserved.



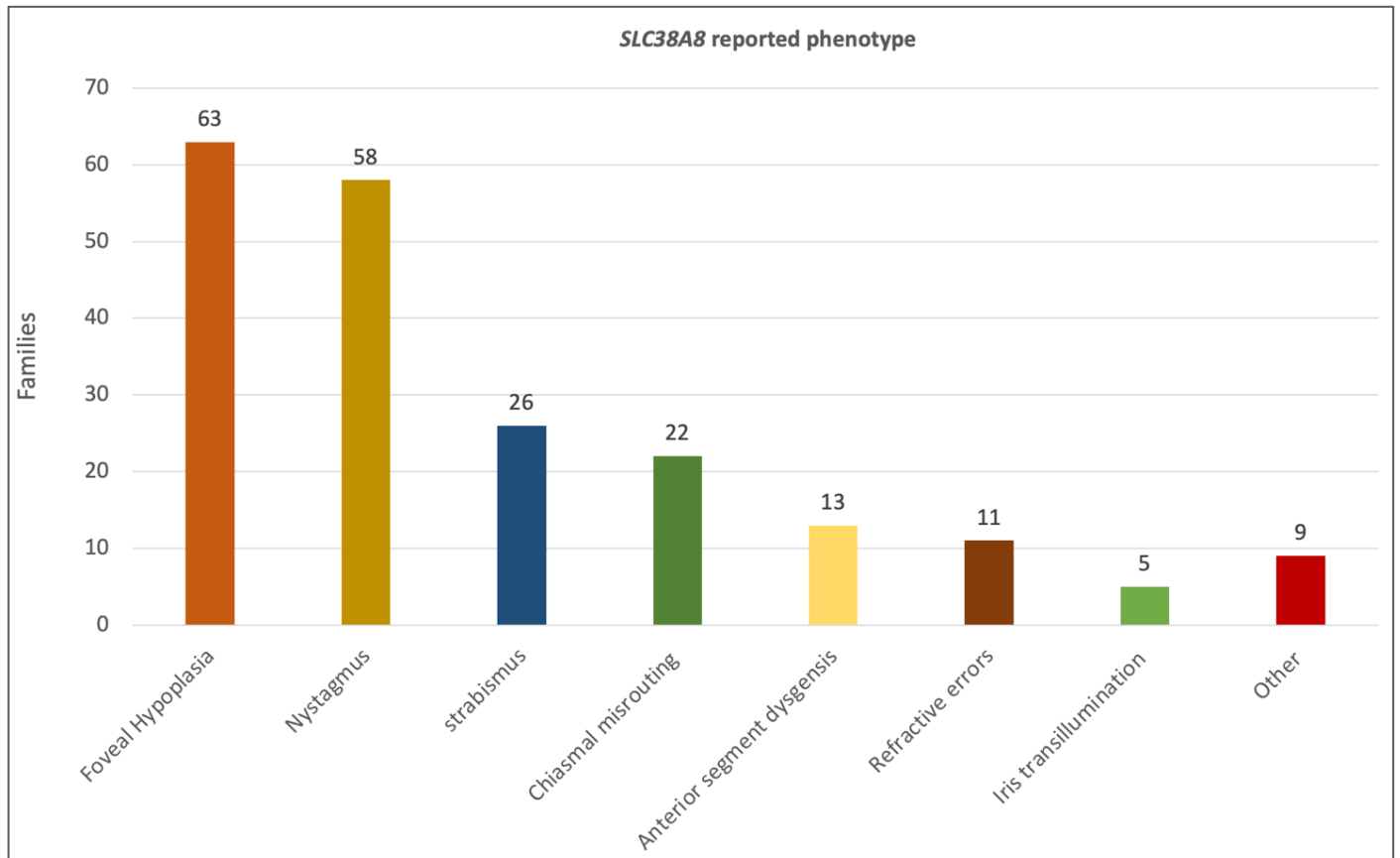
**Figure 3.12. Conservation profile of SLC38A8.** A) Residues are colour coded based on evolutionary conservation scores, with turquoise for variable (1-3), white for neither conserved nor variable (4-6) and maroon for conserved (7-9). Black boxes highlight the transmembrane domains of SLC38A8 and each are labelled by a number corresponding to their order on the 2D protein structure. The conserved residue count for each transmembrane domain is also displayed. All 435 amino acids have been annotated with the residue number according to the reference protein sequence NP\_001073911.1.



#### **3.2.4.4 Evaluating SLC38A8 phenotype in a large cohort**

The variability in phenotypic reports of families affected by *SLC38A8* related FH includes overlap with other FH encompassing disorders. Clinical manifestations were re-examined in 63 affected families from the published literature, local cohort and the 100KGP to try to improve the definition of the *SLC38A8* phenotype (Appendix C: Supplementary Table 3.5). All affected individuals in these families have been clinically evaluated for FH using OCT, but VEP was not performed in some studies (Perez et al., 2014, Ehrenberg et al., 2021, Toral et al., 2017, Lasseaux et al., 2018).

The established *SLC38A8* phenotype consists of FH together with chiasmal misrouting in the absence of hypopigmentation, with ASD as an additional variable phenotype. Reviewing the phenotypes reported in the entire cohort of families with FVH2 showed consistent reports of FH (100%) accompanied mostly with oculomotor defects comprising of nystagmus (92%) and strabismus (41%) (Figure 3.13). Less frequent reports include chiasmal misrouting (35%), ASD (21%) and refractive errors (17%). However, some reports of refractive errors comprising astigmatism, myopia and hyperopia lack grading in terms of severity, which hinders interpretation. Contradicting the established orthodoxy that the *SLC38A8* phenotype does not involve any pigmentation defects, is the reports of five families with iris TID in association with FH and chiasmal misrouting (Schiff et al., 2021, Kuht et al., 2020, Lasseaux et al., 2018).



**Figure 3.13. Clinical manifestations of *SLC38A8*.** Phenotypes observed across the entire cohort of FVH2 affected families (63) derived from the published literature (59) and the local cohort (4). The number of probands/families reported with the phenotypic trait is displayed in brackets above the bar. Anterior segment dysgenesis includes posterior embryotoxon (10), Axenfeld anomaly (1), blue dot cataract (1) and retinochoroidal coloboma (1). Refractive errors refer to myopia (5), hypermetropia (8) and astigmatism (8). Different refractive errors are sometimes reported in different affected individuals of the same family. Other clinical findings documented include developmental delay (3), Kartagener syndrome (1), microphthalmia (1), macrocephaly (1), infantile hypotonia and generalized joint hypermobility (1), goniodysgenesis (1) and hypospadias, patent foramen ovale and high muscle tone (1).

### 3.2.5 Curating *SLC38A8* variants for a LOVD

To aid clinicians in variant interpretation of *SLC38A8* genetic changes and to maximise the translational potential of the work described here, all variants confirmed as causative for FVH2 in the published literature, 100KGP and the local cohort were published by the author in an open access variation database (<https://databases.lovd.nl/shared/genes/SLC38A8>). Authorisation for gene specific curation was granted to provide full access to edit and manipulate data in LOVD v3.0. The database was updated with all deleterious *SLC38A8* variants known and/or published to date in FVH2 (25/02/2024), in addition to those of no clinical consequences that are reported in the literature (Figure 3.14). These variants were reported using the MANE transcript (NM\_001080442.3) in accordance with HGVS nomenclature, and given an ACMG classification, including the corresponding evidence to support each category. 56 new variant entries in 42 individuals were submitted using phenotypic data in HPO format. Supplementary data relating to the family's pedigree, ethnicity, gender and consanguinity status were supplied where available to increase the quality of the entries. This comprehensive database currently holds 151 variants with 88 unique variants in 154 individuals. The high quality of data submitted improved the reliability rating of the *SLC38A8* LOVD database to the maximum score (4 stars), which exceeds that of public variant archives like ClinVar (3 stars)

#### The *SLC38A8* gene homepage

General information	
Gene symbol	SLC38A8
Gene name	solute carrier family 38, member 8
Chromosome	16
Chromosomal band	q23.3
Imprinted	Unknown
Genomic reference	<a href="#">NC_000016.9</a>
Transcript reference	<a href="#">NM_001080442.1</a> , <a href="#">NM_001080442.3</a>
Exon/intron information	<a href="#">NM_001080442.1 exon/intron table</a>
Associated with diseases	<a href="#">FVH2</a>
Citation reference(s)	-
Refseq URL	<a href="#">Genomic reference sequence</a>
Curators (1)	<b>Mohammed A.M Derar</b>
Total number of public variants reported	151
Unique public DNA variants reported	87
Individuals with public variants	154
Hidden variants	-
Download all this gene's data	<a href="#">Download all data</a>
Notes	Establishment of this gene variant database (LSDB) was performed by Johan den Dunnen, supported by <a href="#">Global Varome</a> .
Date created	May 03, 2013
Date last updated	December 21, 2023
Version	<b>SLC38A8:231221</b>

#### User account #04246 (Mohammed A.M Derar, Leeds, UK)

User ID	04246
ORCID ID	<a href="#">0000-0003-0293-909X</a>
Name	Mohammed A.M Derar
Institute	University of Leeds
Department	Leeds Institute of Medical Research
Telephone	07729264022
Address	Leeds Institute of Medical Research Wellcome Trust Brenner Building (Level 8) St James's University Hospital Beckett Street Leeds LS9 7TF
City	Leeds
Country	United Kingdom (Great Britain)
Email address	<a href="mailto:ummamd@leeds.ac.uk">ummamd@leeds.ac.uk</a>
API token	<a href="#">(Show / More information)</a>
API token expiration	-
Curator for 1 gene	<a href="#">SLC38A8</a>
Collaborator for 0 genes	-
Data owner for 182 data entries	<a href="#">42 individuals</a> , <a href="#">42 screenings</a> , <a href="#">56 variants</a> , 42 phenotypes
Has created 182 data entries	<a href="#">42 individuals</a> , <a href="#">42 screenings</a> , <a href="#">56 variants</a> , 42 phenotypes
Shares access with 0 users	-
Default license	<a href="#">(Change)</a>
User level	Curator
Created by	<a href="#">Mohammed A.M Derar</a>
Date created	2022-01-19 12:24:41 +01:00 (CET)

Database name	Curators	Rating	Unique variants	Software	Last updated
Global Varome shared LOVD <a href="https://databases.lovd.nl/shared/genes/SLC38A8">https://databases.lovd.nl/shared/genes/SLC38A8</a>	Mohammed A.M Derar University of Leeds	★★★★	88	LOVD 3.X	2023-12-21
ClinVar at NCBI <a href="https://www.ncbi.nlm.nih.gov/clinvar/?term=SLC38A8[gene]">https://www.ncbi.nlm.nih.gov/clinvar/?term=SLC38A8[gene]</a>	Admin BRAINN	★★★★	250	ClinVar	2024-01-07
BIPMed SNP Array - HG19 <i>Not curated, does not accept submissions</i> <a href="http://bjpmed.igm.unicamp.br/snparray_hg19/genes/SLC38A8">http://bjpmed.igm.unicamp.br/snparray_hg19/genes/SLC38A8</a>	Admin BRAINN	★★★☆☆	-	LOVD 3.X	2018-10-04

**Figure 3.14. *SLC38A8* database in LOVD.** *SLC38A8* Gene curator status was granted to the author on the 19/01/2022. Images from LOVD were obtained on 11/02/24.

### **3.2.6 CRISPR-cas9 knock out of *SLC38A8*.**

Further insight into the pathogenic mechanisms of variants underlying FVH2 can be obtained using cellular models. For this purpose, genome editing of available cell lines was investigated as a way to generate viable mutant cell lines for downstream *In vitro* functional analysis.

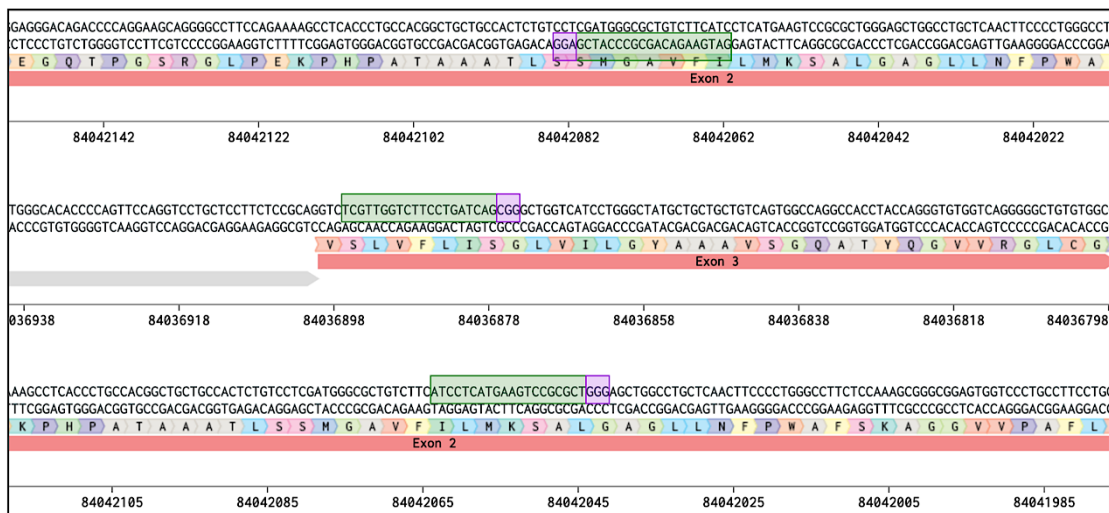
#### **3.2.6.1 sgRNA design and evaluation**

CRISPR-Cas9 mediated mutagenesis was evaluated as a method for generating knockout cellular models to investigate the pathogenic mechanisms underlying FVH2. The guide RNAs were designed to direct the Cas9 endonuclease to cleave *SLC38A8* at exon 2 or 3 causing gene disruption (Figure 3.15). This approach capitalises on the predominant error prone DNA repair mechanism of non-homologous end joining (NHEJ) to introduce frameshift variants that are likely to knock out the gene. Three sgRNAs were designed and are predicted to have variable targeting performance when assessed by IDT and Invitrogen computational tools (section 2.6.10). SgRNA1 and sgRNA2 have a risk in that the target sequences are prone to known SNPs which may reduce the targeting efficiency. The rs151146991 SNP has a low global MAF of 0.0000316 (51/1613966) but the rs151146991 has an allele frequency of 0.338 (542965/1607100).

For validation purposes using Sanger sequencing, primers were designed to flank the edit site to confirm the introduction of random INDELS by NHEJ. Potential off target effects arising from the use of the designed sgRNAs were anticipated and are catalogued (Appendix C: Supplementary Table 3.6). The genomic coordinates of these off-target edits will be used in designing primers for screening purposes.

A)

## Target locus



B)

## CRISPR guides

ID	Sequence	PAM	Strand	Location (GRCh38)	Score (%)	On target score	Off target score	off-targets	SNP risk
sgRNA1	GATGAAGACAGCGCCATCG	AGG	+	16:84042061	74.70	37	78	12	rs151146991
sgRNA2	TCGTTGGTCTTCTGATCAG	CGG	-	16:84036897	57.99	54	70	14	rs1317524
sgRNA3	ATCCTCATGAAGTCCGCGCT	GGG	-	16:84042064	57.31	24	89	3	-

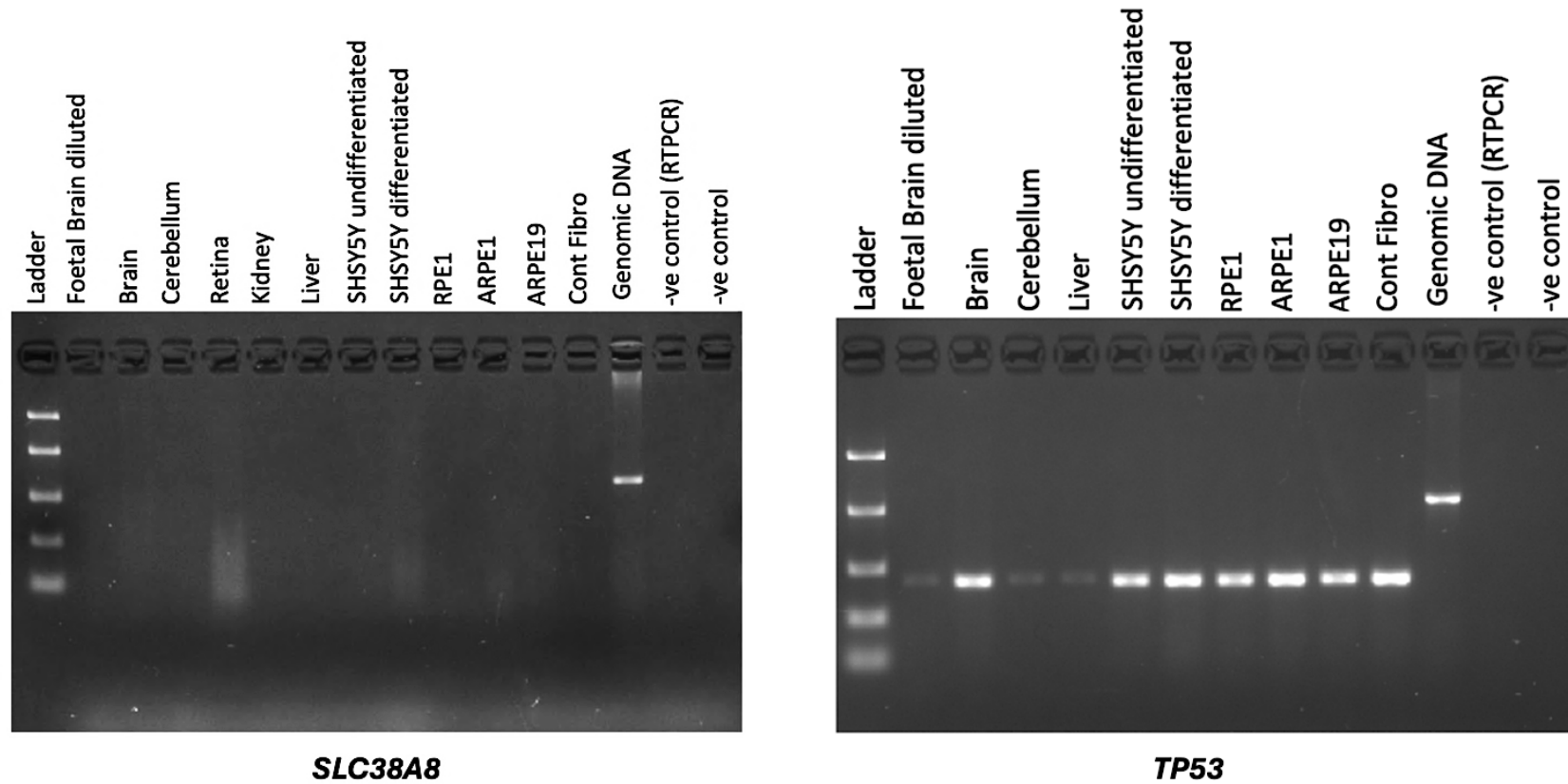
## Sequencing primers

Forward Primer	Sequence	Tm (°C)	Reverse Primer	Sequence	Tm (°C)	Amplicon (bp)
1F-1	ATGCAGCACCCCTTTGGAAAG	59.0	1R-1	TTACCAGTCCACCAGGAAG	59.0	353
1F-2	GTGGGTAGGGAATGGGAGAC	59.2	1R-2	CAGAAAAGCCAAAGCCACCT	59.0	456
1F-3	CTTAGCCATGGAGGGACAG	59.0	1R-3	GCTTACAGGACACGCAACT	59.2	301
2F-1	TTCCAGGTTACAGCCTAG	59.0	2R-1	TGAGCAGGTTGAGGAGGAAG	59.0	403
2F-2	TAGACTGCCAACCCATGTT	58.9	2R-2	CGTCTGCTCAACTGGAAAC	59.1	421
2F-3	GCACTGCAGCTCTGATTCA	58.8	2R-3	CTCACACAGCTTCCAATGG	58.8	331
3F-1	ATGCAGCACCCCTTTGGAAAG	59.0	3R-1	TTACCAGTCCACCAGGAAG	59.0	353
3F-2	GACTAGAAGTCGTGGGCAGA	58.8	3R-2	CAGAAAAGCCAAAGCCACCT	59.0	429
3F-3	CTTAGCCATGGAGGGACAG	59.0	3R-3	GCTTACAGGACACGCAACT	59.2	301

**Figure 3.15. Guide RNAs designed for *SLC38A8* knock out and sequencing primers design.** A) Three sgRNAs targeting exon 2 (sgRNA1 and 3) and exon 3 (sgRNA2) of the MANE transcript of *SLC38A8* (ENST00000299709.8). sgRNA target sequences are highlighted in green and the PAM recognition sequence for *S.pyogenes* Cas9 is highlighted in purple. The amino acid corresponding to each codon in the exon is highlighted in unique colours. The genomic coordinates are displayed on the bottom track. B) sgRNA design parameters are tabulated, including the PAM sequence and the genomic location of edit sites. SNPs that may occur in the sgRNA target sequence are reported using dbSNP identifiers. The score is the predicted guide RNA performance using Invitrogen™ TrueDesign™ Genome Editor algorithm. The higher On target or Off target score implies a more favourable outcome with reduced unintended genomic edits. Three primer pairs were designed for each sgRNA to sequence the target site for confirmation of mutants with the desired edit.

### 3.2.6.2 Assessing tissues and cell lines for *SLC38A8* expression

Identification of an appropriate cell line for CRISPR mediated genome editing initially required the assessment of *SLC38A8* expression. Cells expressing *SLC38A8* could potentially serve as a model in which to observe the cellular consequences of *SLC38A8* knockout. RT-PCR was performed on a range of different tissues comprising foetal brain, brain, cerebellum, liver and kidney, together with several commercial cell lines including SH-SY5Y, of neuronal origin, and RPE1, ARPE1 and ARPE19, all of retinal origin. The cDNA generated was used as a template for PCR reactions to detect the expression of a housekeeping gene (*TP53*) to determine RNA integrity. *TP53* was ubiquitously expressed with uniform intensity across the different samples, except for three samples which appeared to have lower RNA quality. These three samples were from foetal brain, cerebellum and liver (Figure 3.16). Genomic DNA was selected as a positive control to confirm primer performance, with a larger band expected due to inclusion of one or more introns. Two negative controls were used to rule out contamination in the RT-PCR and subsequent PCR reaction. The negative control for the RT-PCR did not include the reverse transcriptase enzyme required for cDNA generation while the PCR negative control contained dH<sub>2</sub>O instead of sample cDNA. The *SLC38A8* PCR experiment was replicated using 3 different primer sets targeting the UTR or exon 7 and 8 (Appendix C: Figure 3.4). The results showed no *SLC38A8* expression in any of the cell lines tested, causing the proposed functional study to be halted.



**Figure 3.16. *SLC38A8* gene expression profile.** *SLC38A8* did not generate any amplicons for the 12 samples tested. *TP53* was expressed equally in 10 samples, with 3 samples demonstrating lower signal intensity, suggesting compromised RNA integrity. The predicted cDNA amplicon for *TP53* is 408 bp and for *SLC38A8* is 188 bp. No contamination was detected in the negative control and the genomic DNA control generated the expected band size of 1057 bp for *TP53* and 600 bp for *SLC38A8*. The 1 kb ladder consists of bands of 100, 250, 500, 1000 and 2000 bp. 10 $\mu$ l of PCR product was loaded on a 2% agarose gel. The amplicons are displayed for primer set 3 which targets exons 7 and 8 of *SLC38A8*.

### 3.3 Discussion

#### 3.3.1 Overview of the results

This study began with analysis of probands with four novel variants and one previously reported variant, recruited through collaboration with the ERDC (Figure 3.1). Next the coding and noncoding regions of *SLC38A8* were investigated in the participants of the 100KGP, resulting in new diagnoses for three further probands (Table 3.1). This relatively small number of cases identified within two large IRD cohorts highlights the very low prevalence of FVH2, consistent with its being a rare subtype of an already group of disease (IRD). The 100KGP analysis uncovered four SNVs of which two nonsense and one splice region variant were subsequently published elsewhere (Schiff et al., 2021). Haplotype analysis was performed to validate the hypothesis of a founder variant in three 100KGP participants of South Asian descent who share a mutual nonsense variant NP\_001073911.1:p.(Tyr88\*) (Figure 3.4). Additional analysis in the 100KGP have also uncovered a large *SLC38A8* inversion that was found to be a redundant case submitted to the 100KGP despite already being solved in the local cohort through the ERDC collaboration (Figure 3.5).

Next, the *SLC38A8* literature to date was reviewed and the published variants were redefined using HGVS nomenclature and characterised according to ACMG classification. In total, a catalogue of 51 *SLC38A8* variants were obtained from the published literature (47) and ERDC collaboration (4). These 51 *SLC38A8* variants were detected in probands from 63 FVH2 affected families which served as a reliable cohort for investigations to further characterise and understand the spectrum of variants underlying FVH2 (Table 3.4). The phenotype in the 63 families afflicted by FVH2, drawn from published cases, those recruited to 100KGP or included in the local cohort, was analysed in an attempt to refine *SLC38A8* phenotype (Figure 3.13). The definitive list of *SLC38A8* variants along with auxiliary information related to the pedigree and phenotype was made available on an LOVD database to support variant interpretation by clinicians and other laboratories (Figure 3.14).



The *SLC38A8* mutation spectrum in FH was reviewed, providing novel insights into domains of potential significance as hinted by missense variant aggregation in the highly conserved 6<sup>th</sup> transmembrane domain (Figure 3.7). Protein modelling was performed to generate a novel *SLC38A8* protein model, followed by missense simulations to elucidate the likely loss of function mechanisms of these variants in FVH2 (Table 3.5 and Figure 3.9). Revealing the 3D protein structure of *SLC38A8* showed that deleterious missense variants tend to localise at deeper structures within or around the pore of the transmembrane protein (Figure 3.11). Statistical validation of pathogenic missense variants' association with increased in electrostatic potentials ( $p=0.006$ ) suggested changes in electrostatic potentials as a possible loss of function mechanism for missense variants underlying FH (Figure 3.10).

CRISPR-cas9 mutagenesis was considered to investigate the effects of *SLC38A8* loss of function at the cellular level. However, no cell lines or tissues tested were found to express *SLC38A8* causing this functional study to be postponed until a suitable cell line is available (Figure 3.16).

### 3.3.2 Expanding on the known *SLC38A8* mutation spectrum

In this study five variants responsible for FVH2 in four unrelated families were reanalysed (Figure 3.1). These are comprised of four novel *SLC38A8* variants, these are the NC\_000016.10:g.84015973-84027074inv, NC\_000016.10:g.84034763\_84037497del, NC\_000016.10:g.84006453\_84019002del, NM\_001080442.3:c.669delC; NP\_001073911.1:p.(Thr224Profs\*44) and the previously reported missense variant NM\_001080442.3:c.848A>C; NP\_001073911.1:p.(Asp283Ala). The report of these four novel variants expands the *SLC38A8* mutation spectrum in FVH2 to 51 unique variants.

This study has characterised an *SLC38A8* inversion as a pathogenic variant responsible for FH and nystagmus. Inversions exert their pathogenicity through interruption or rearrangement of the coding sequence or regulatory elements if the breakpoints occur at or near a gene (Venables, 2007). The homozygous

NC\_000016.10:g.84015973\_84027074inv was detected in a proband from a consanguineous family (F1) (Figure 3.1) and the same proband was also encountered in the 100KGP (Figure 3.5). The inversion encompasses exon 7-9 and is predicted to disrupt the ORF to induce a frameshift (Figure 3.2A). In the event of a PTC being introduced, NMD is likely to degrade the mutant transcript to result in a null allele, though this remains to be proven experimentally. Empirical evidence from RT-PCR is required, converting the *SLC38A8* mRNA transcript into cDNA for analysis, to ascertain whether this results in NMD or aberrant splicing events.

In family F2, the proband ERD01 was a compound heterozygote for large deletions NC\_000016.10:g.84034763\_84037497del and NC\_000016.10:g.84006453\_84019002del. The NC\_000016.10:g.84034763\_84037497del variant deletes exon 3 of the MANE transcript which alters the ORF resulting in a frameshift. This frameshift is predicted to introduce a PTC at amino acid position 113 which is likely to lead to NMD eliminating the aberrant mRNA transcript (Figure 3.2A). The NC\_000016.10:g.84006453\_84019002del variant deletes exons 8-11 without altering the reading frame and may incorporate additional intergenic sequences downstream the deletion into the aberrant mRNA transcript (Appendix C: Supplementary Figure 3.1). This would result in a truncated protein of 321 residues, with 268 correct residues but gaining additional 53 incorrect amino acids (Figure 3.2A). Though this prediction requires empirical evidence to ascertain the exact molecular consequence of the deletion.

A homozygous frameshift variant NM\_001080442.3:c.669delC; NP\_001073911.1:p.(Thr224Profs\*44) was also identified in proband ERD02 in family F3 (Figure 3.1). The introduced PTC in exon 7 is likely to cause the aberrant transcript to be targeted by NMD, resulting in a null allele. The analysis also detected a frequent variant, NM\_001080442.3:c.848A>C; NP\_001073911.1:p.(Asp283Ala) in a European proband (F1368) in family F4, from the USA. The variant was predicted as pathogenic using pathogenicity assessment tools but it was reported in four homozygotes in gnomAD. This variant is almost exclusively found in people of Ashkenazi Jewish ethnicity and

has an allele frequency of 0.007836 (232/29608) in that population. A founder effect is the likely explanation for the high allele frequency of this variant. The NM\_001080442.3:c.848A>C; NP\_001073911.1:p.(Asp283Ala) variant has been previously reported by Toral et al. (2017), Sciff et al. (2021) and Ehrenberg et al. (2021), causing FH in a total of nine unrelated Ashkenazi Jewish families. Data from the *SLC38A8* modelling showed that the mutation affects a conserved residue, substituting a buried residue that is charged (aspartic acid) with an uncharged residue (alanine) (Figure 3.9). The variant is also predicted to disrupt hydrogen bonds formed by the aspartic acid to inflict further damage to the protein structure.

### 3.3.3 Confirming a founder *SLC38A8* mutation in South Asians

The decrease in allelic diversity due to endogamy within specific communities gives rise to enriched haplotypes of medical relevance within subpopulations (Perez et al., 2014). Founder mutations such as the *SLC38A8* NM\_001080442.3:c.848A>C; NP\_001073911.1:p.(Asp283Ala) variant described above, contribute to the ethnic variation in disease prevalence and predisposition due to reduction in the gene pool within ethnic groups. There has also been a report of a second potential founder mutation in *SLC38A8*, described by Schiff et al. (2021), who suggested a similar hypothesis for the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) variant, given that it emerged in three unrelated and consanguineous Indian, Pakistani and Bangladeshi patients. This study encountered the same three patients in the 100KGP cohort and investigated the hypothesis that this variant is a founder mutation (Table 3.1). Haplotype analysis in a 44.9 kb locus spanning the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) variant in patients of Pakistani and Bangladeshi origin supports the hypothesis of a common ancestor, as indicated by identical homozygous genotypes spanning the entire 44.9 kb region from chr16:84022721-84067635 (Figure 3.4). The Indian patient shared the same haplotype proximal to the nonsense variant for 9.5 kb region (chr16: 84022721-84032188) but revealed a discordant haplotype on the distal side, which may imply a more distant shared ancestry and a crossover within this small region between the Indian and the Pakistani and Bangladeshi cases.

The gnomAD database reported this variant as almost exclusive to the South Asian population and the 100KGP database reported this variant only in South Asian individuals, with a global allele frequency of 0.0000198 (32/1613888) and 0.000329 (30/91074) for the South Asian demographic in gnomAD, with no homozygotes reported (accessed 30/03/2024). The 100KGP total allele frequency is 0.0000129 (16/124176). The gnomAD database provides evidence based on the general population, who are not affected by a severe early onset inherited disease. In contrast, the 100KGP population is enriched with rare disease patients, including FH patients. The evidence from these contrasting population databases provides further confirmation that the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) is indeed a founder mutation in South Asians. The clinical utility of this revelation requires testing on a larger cohort of South Asian patients with FH for the presence of the NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*) to accurately determine what proportion of South Asian FH cases this variant accounts for. Depending on the proportion of FH cases attributed to NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*), the genomic testing approach of FH in South Asians could be highly targeted. Sanger sequencing can be performed to target this founder mutation as a first line molecular test that is rapid and doesn't incur incidental findings potentially associated with NGS. Furthermore, NGS may not be widely available in developing South Asian countries, thus Sanger sequencing may be a more accessible technology.

Other founder mutations in *SLC38A8* were reported in reproductively isolated populations or confined to a smaller geographical region. These include the NM\_001080442.3:c.95T>G; NP\_001073911.1:p.(Ile32Ser) variant in Indian and Karaite Jews and the NM\_001080442.3:c.995dupG; NP\_001073911.1:p.(Trp333Metfs\*35) variant in Koreans (Weiner et al., 2020, Ehrenberg et al., 2021, Perez et al., 2014, Toral et al., 2017, Schiff et al., 2021, Kuht et al., 2020, Kruijt et al., 2022, Hayashi et al., 2021). It may also be the case that the NM\_001080442.3:c.101T>G; NP\_001073911.1:p.(Met34Arg) variant, detected in two independent Turkish families afflicted by FH, is a founder mutation in patients of Turkish ancestry (Poulter et al., 2013, Kuht et al., 2020).

However, based on NM\_001080442.3:c.101T>G; NP\_001073911.1:p.(Met34Arg) being absent from population databases, additional reports of the variant in Turkish families and subsequent haplotype analysis is required to ascertain whether a founder effect could explain this observation.

### **3.3.4 Insight into missense variants pathogenic mechanism in FVH2**

SLC38A8 is a sodium-dependent transporter and consists of 11 transmembrane domains and 5 extracellular and 5 intracellular loops (Figure 3.7). The transmembrane domains account for around 54% of protein residues and these regions are conserved especially in the first and sixth and transmembrane domains, indicating functional significance (Figure 3.12). Multiple studies were conducted on SLC38A2 (SNAT2), a closely related protein with similar characteristics to SLC38A8 in terms of sodium dependency and ability to transport N-methylated amino acids (Hagglund et al., 2015). Mutations in SNAT2 conserved residues of Asn<sup>82</sup> and Tyr<sup>337</sup>, in the respective first and seventh transmembrane, resulted in a decreased affinity for Na<sup>+</sup> and consequently a reduced uptake of alanine when compared to the wildtype protein (Zhang et al., 2008). Further studies on SNAT2 showed that a conserved residue in the eighth transmembrane domain (Thr<sup>384</sup>) is required for Na<sup>+</sup> binding (Zhang et al., 2009). The analogous residue on SLC38A8 protein is Thr<sup>308</sup> in transmembrane 8. The evidence from both studies suggests that transmembrane regions influence interactions with Na<sup>+</sup> required to mediate conformational changes to bind alanine. The eleven transmembrane domains collectively form the pore of SLC38A8 that is conserved, and its function is sensitive to amino acid substitutions imposed by missense mutations (Zhang et al., 2008).

The mutation spectrum showed that the damaging missense variants in *SLC38A8* disproportionately affect the transmembrane domains (16/22) (Figure 3.7) and through homology modelling these variants are predicted to alter the physiochemical properties of the protein via various mechanisms, suggesting they alter normal protein structure and might therefore affect protein function and localisation. 18/22 missense variants have severe changes in the hydrophobicity

of amino acids that are  $\geq 2.5$  on the Kyte-Doolittle scale (Figure 3.9). These dramatic changes alter 9/10 neutral residues into strongly hydrophobic or hydrophilic residues. Furthermore, all the affected hydrophobic residues (8/22) are reduced by the missense variants into 4 strongly hydrophilic, 2 neutral and 2 mildly hydrophobic residues (Appendix C: Supplementary Figure 3.2). Changes in hydrophobicity within the transmembrane region could alter the protein conformation and impede interactions with the substrates (Partridge et al., 2004).

Furthermore 16/19 missense variants implicated in FH increased the surface electrostatic potentials of the channel pore (Figure 3.10). The co-transport of amino acids by SLC38A8 is dependent on inward currents, abiding by the stoichiometry of 1 neutral amino acid for each Na<sup>+</sup>, which may be disrupted by alterations to electrostatic potentials (Mackenzie et al., 2003). For example, the increase in the local electrostatic potential within the ribosomal tunnel due to mutant positively charged residues can alter translation rates, arrest peptide elongation, and hamper secondary structure formation, leading to misfolding events (Lu and Deutsch, 2008). Misfolded or unfolded proteins aggregation in the endoplasmic reticulum (ER) can induce unfolded protein response (UPR) to activate inositol requiring kinase 1 (IRE1) signalling pathway. This will ultimately activate ER associated degradation (ERAD) for the ubiquitination and degradation of these aberrant proteins by cytosolic proteosomes (Haeri and Knox, 2012).

The evidence presented for both possible pathogenic mechanisms suggests that missense variants in *SLC38A8*, like the many nonsense, frameshift and splice site variants, all act by causing a potential loss of function. This is substantiated by the observations in FVH2 whereby patients with mutations that result in null alleles (whole gene deletion, nonsense and frameshift variants) have the same phenotype as those with missense variants. This reinforces loss of function as the likely the pathogenic mechanism underlying *SLC38A8* related FH.

Homology modelling has been used extensively to study the SLC38 family, displaying the translational potential and reliability of this computational approach

(Zhang et al., 2009, Schioth et al., 2013). Our modelling data is also in concordance with other studies using *SLC38A8* protein modelling for variant characterisation (Toral et al., 2017). Their data showed that NP\_001073911.1:p.(Asp283Ala) also increased the electrostatic potentials as a mechanism to destabilise the protein. However, more efforts are required in functionally validating *SLC38A8* missense variants using protein uptake assays and other experiments outlined in section 6.6.1 to further characterise the protein, revealing functional domains and improving our understanding of the molecular mechanisms underlying FH. The outcome of functionally characterising *SLC38A8* domains could also be translated to *SLC38A7* considering the close evolutionary relationship and sequence identity of 42% between *SLC38A8* and *SLC38A7* (SNAT7) (Hagglund et al., 2015). *SLC38A7* shares a similar expression profile in the CNS as *SLC38A8* and is thought to have analogous functions in the glutamine-glutamate cycle by channelling L-glutamine in neurones (Hagglund et al., 2011). The SNAT7 is associated with tumour growth thus the further characterisation of the protein will aid therapeutic interventions to halt cancer progression (Verdon et al., 2017).

### **3.3.5 Redefining the *SLC38A8* phenotype**

FVH2 can encounter phenotypic overlap with other FH related diseases, especially with albinism. The *SLC38A8*-related phenotype needs to be redefined to aid in differential diagnosis based on phenotyping. The current consensus on the *SLC38A8* phenotype is that the patients exhibit FH, with chiasmal misrouting, but with the exclusion of pigmentation defects. However, there are some inconsistencies in the clinical manifestations derived from published reports of *SLC38A8* related FH (Appendix C: Supplementary Table 3.5). In the cohort of 63 families affected by FVH2, from both the published literature or families identified through the ERDC, FH was present in all cases, chiasmal misrouting was detected in 35% and ASD in 21% of families (Figure 3.13). The lack of reporting of chiasmal misrouting is likely to be due to inequitable access to VEP testing, technical challenges in conducting VEP tests on infants and inconsistent VEP results obtained during childhood. It has been shown that chiasmal misrouting becomes more prominent with age, which causes misleading results if VEP is conducted in young patients (Campbell et al., 2019). Campbell and colleagues

found apparently normal VEP responses during infancy in patients who, in subsequent referrals at an older age (14 years), showed evidence of chiasmal misrouting. The application of VEP testing therefore requires an age cut off for correct phenotyping to avoid inappropriate molecular investigations in genes unrelated to the patient's true phenotype. Furthermore, the presence or absence of chiasmal misrouting will not distinguish between albinism and *SLC38A8*-associated FH since this phenomenon is also commonly observed in albinism patients (Montoliu and Kelsh, 2014).

The variable feature that is presumably secondary to FVH2 and may or may not present, is ASD (21%). Posterior embryotoxon which accounts for 77% of ASD in FVH2 affected families and manifesting in 16% of all FVH2 cases, has been suggested to be a distinctive feature of *SLC38A8*, prompting targeted analysis of this *SLC38A8* if the patient exhibits FH with posterior embryotoxon in the absence of iris TID (Ehrenberg et al., 2021). Some authors speculated that ASD is a consequence of more deleterious variants (Toral et al., 2017). This is not true as patients of the same genotype may not express ASD, for example, patients homozygous for *SLC38A8* NM\_001080442.3:c.95T>G; NP\_001073911.1:p.(Ile32Ser) were reported in 9 families but ASD was reported in only two of these cases (Weiner et al., 2020, Ehrenberg et al., 2021, Perez et al., 2014, Kruijt et al., 2022). Similarly, in three families harbouring homozygous *SLC38A8* NM\_001080442.3:c.264C>G; NP\_001073911.1:p.(Tyr88\*), only one was described as exhibiting ASD (Schiff et al., 2021). Moreover, homozygous NM\_001080442.3:c.698A>G; NP\_001073911.1:p.(Glu233Lys) was reported in a proband of an Indian family with coloboma but not in a Turkish-Iranian family which was also homozygous for the variant (Poulter et al., 2013, Ehrenberg et al., 2021). This implies that the presence other factors such as modifier genes, the combined effect of multiple polymorphisms in *SLC38A8*, or environmental factors could dictate ASD expression. Kuht et al. (2020) and Kruijt et al. (2022) argued that ASD could be a phenocopy due to the frequency of posterior embryotoxon across all age groups in the UK general population being 6.8% (Rennie et al., 2005). The prevalence of posterior embryotoxon across the age group of 40-64 was significantly higher in a study (14.7%) with a higher frequency in women at this age group (p=0.023) (Hashemi et al., 2015). The prevalence of



posterior embryotoxon reported in the literature ranges from 6-14.7% which is comparable to the prevalence of 16% reported in families with FVH2 (16%), this may suggest that this ASD manifestation is an incidental finding in the general population.

Nystagmus almost always accompanies FH as it is secondary to reduced visual acuity in early life. In FVH2, 92% of cases in the 63 families described here were reported to have nystagmus. A single study accounts for the discrepancy in nystagmus reports (Poulter et al., 2013). It is likely that the study did not disclose the full clinical spectrum of the patients since subsequent phenotyping of their patients (4/7) in another study revealed nystagmus in all 4 patients tested (Kruijt et al., 2022). Investigating nystagmus characteristics in FH patients with *SLC38A8* mutations revealed consistent conjugate and horizontal nystagmus (4 pendular and 7 Jerk nystagmus) across all 11 patients in one study (Kuht et al., 2020). These findings are in line with another study reporting 8 patients with FVH2 displaying horizontal nystagmus (7 pendular and 1 Jerk nystagmus) and one with rotary nystagmus (Schiff et al., 2021). These horizontal and conjugate ocular oscillations are also present in albinism and idiopathic infantile nystagmus which implies a common pathophysiology (Kumar et al., 2011). The prevalence of nystagmus waveforms differs slightly for each disease. Albinism patients (n=52) had a higher incidence (67%) of jerk oscillations while in idiopathic infantile nystagmus the patients (n=51) displayed a similar frequency (42.7%) for both pendular and jerk movements (Kumar et al., 2011). On the other hand, subjects with FVH2 (n=19) had an increased tendency for pendular nystagmus (57.9%) (Schiff et al., 2021, Kuht et al., 2020).

Characterizing FH through the recommended structural grading system helps to reveal subtle distinctive features capable of aiding differential diagnosis (Thomas et al., 2011). The FH grading was not reported in all 63 families affected by FVH2 but was available for some probands in the published literature. Revealing the retinal architecture via OCT examination in 34 probands with *SLC38A8* mutations showed consistent classification of grade 3 or 4 FH (Ehrenberg et al., 2021, Kuht et al., 2020, Schiff et al., 2021, Kruijt et al., 2022). These severe forms of FH are due to early interruption of foveal development whereby cone specialisation and

centrifugal migration of neurones in the ganglion cell layer and inner nuclear layer are interrupted (Thomas et al., 2011). Quantifying the retinal changes in a cohort of 10 patients with *SLC38A8* mutations revealed thicker retinal nuclear fibre, ganglion cell, inner plexiform, inner nuclear and outer plexiform layers, but a thinner cone photoreceptor outer segment layer, indicative of disruption of foveal developmental (Kuht et al., 2020). Despite the consistent grading of FH in patients with *SLC38A8* mutations, these categories of grades 3 and 4 cannot serve as indicators for FVH2 since some albinism patients also display high grade FH, though the grading is more diverse (grades 1-4) (Thomas et al., 2011, Kruijt et al., 2022). In a cohort of 10 subjects with *PAX6* related FH, the FH grades were limited to grade 1-3, which implies a less severe FH in comparison to *SLC38A8*-related FH (Thomas et al., 2011). However, the visual acuity is more severely reduced in *PAX6*-associated FH because of the additional anterior segment and neurological anomalies comprising of cataracts, glaucoma, keratopathy, and optic nerve hypoplasia (Hingorani et al., 2009).

Perhaps the most distinctive feature of the *SLC38A8*-associated phenotype by contrast with ocular albinism is the lack of hypopigmentation. However, estimating pigmentation levels poses a diagnostic challenge, especially in fair skinned populations (Campbell et al., 2019). Cutaneous and ocular pigmentation are variable and can become more profound with age (Campbell et al., 2019). In these populations, given the difficulty in defining a skin pigmentation, the diagnostic signs for albinism are therefore FH in conjunction with chiasmal misrouting (Montoliu and Kelsh, 2014). This causes a significant proportion of patients with FVH2 to be misdiagnosed as albinism. Infants who exhibit FH and have a lighter complexion than their older relatives would be suspected of albinism since chiasmal misrouting cannot be reliably evaluated during infancy. Progressive pigmentation changes would make distinguishing between FH and albinism in infants very challenging (Campbell et al., 2019, Schiff et al., 2021). Without molecular testing, it seems likely that a significant proportion of FVH2 cases would be misclassified as albinism.

The phenotypic overlap between albinism and FVH2 is further exacerbated by reports of iris TID in 5 unrelated families harbouring *SLC38A8* variants (Schiff et

al., 2021, Kuht et al., 2020, Lasseaux et al., 2018). Iris TID are caused by the loss of pigmentation in the iris stroma, a common sign of albinism (Sjodell et al., 1996). Cases of FVH2 with iris TID exhibit a phenotype indistinguishable from ocular albinism, which would inevitably lead to a false diagnoses. Since iris TID has only been reported in 8% of FVH2 cases and these were reported only in families of Caucasian descent (4 European and 1 Ashkenazi Jewish), it seems plausible that this finding may be unrelated to hypopigmentation but is rather a result of light pigmented irides in fair populations (Campbell et al., 2019). Without additional phenotypic data on the grading of Iris TID or slit lamp examination results from additional non affected family members for comparison, it is not possible to determine whether these iris hypopigmentation reports are reliable. Moreover, evidence from a medaka fish model of an *SLC38A8* ortholog knock out generated by morpholino did not display hypopigmentation defects (Poulter et al., 2013). In addition, a mouse knockout model of the FVH2 phenotype generated by CRISPR-Cas9 mutagenesis also did not show any signs of systemic or ocular hypopigmentation (Guardia et al., 2023). The melanin content measured in the retina, iris and RPE in these mice did not show significant discrepancy to the wildtype mouse (C57BL/6J). Although hypopigmented patches were observed in the choroid, the melanin content of the iris was similar to wildtype.

A diagnostic workflow has been proposed by Kruijet et al. (2018) to discriminate between albinism and *SLC38A8* phenotype based on the phenotypic expression of three major criteria comprising FH (>grade 2), chiasmal misrouting and absence of ocular hypopigmentation, or two of these three major criteria supplemented with two minor criteria of nystagmus, absence of skin or hair hypopigmentation or absence of FH grade 1. The utility of this diagnostic criteria is controversial, given the redundancy of FH grading as both a major and minor criterion, the lack of definitive estimates for hypopigmentation, the unequitable access to VEP testing for chiasmal misrouting detection and variability of VEP results, and the variation in pigmentation with age (Dumitrescu et al., 2021). The proposed diagnostic workflow may therefore serve as medium to narrow down the suspected FH to albinism or FVH2, but further distinction would require molecular analysis.

In summary, the hallmarks of the *SLC38A8* phenotype comprise FH and chiasmal misrouting without hypopigmentation. Establishing a defined cohort suspected of harbouring *SLC38A8* variants is challenging due to the phenotypic overlap with ocular albinism and the technical limitations surrounding diagnostic tests like VEP. To capture genuine FVH2 cases, *SLC38A8* needs to be incorporated in FH, nystagmus and albinism gene panels to compensate for the lack of chiasmal misrouting reports and the difficulty in defining pigmentation status.

## Chapter 4 : Genomic discoveries in unsolved FH cohorts

### 4.1 Introduction

NGS has revolutionised genomic investigations in IRDs through the increased number of genes for interrogation and the ability to detect different mutational classes including SVs. This represents a significant improvement over conventional Sanger sequencing which was previously the gold standard in clinical diagnosis of inherited disease. Sanger sequencing is restricted to a single genomic target of ~500 bp and has limited scope for the identification of SV's, which reduces its diagnostic potential especially in conditions with profound genetic heterogeneity such as FH, which is caused by variants in 52 currently known genes (Table 1.2). The FH encompassing disorders also display substantial phenotypic overlap, as exemplified by FVH2 and albinism (section 3.3.5). This hinders differential diagnosis based on phenotyping to narrow down the suspected genes for Sanger sequencing. Collectively, these factors resulted in a significant proportion of FH patients not receiving a genetic diagnosis due to the limitations of Sanger sequencing.

Patients with unsolved FH have been referred to Dr Carmel Toomes's research laboratory (University of Leeds), primarily for *SLC38A8* screening, but where that proved negative, as part of this study these cases were subject to a wider genomic investigation by NGS. The progressive decrease in WGS and WES associated costs made short read NGS technology accessible to many establishments including the Leeds Institute of Medical Research. The initial genetic screen provided at Dr Toomes laboratory was WES, for an affordable and more manageable genomic analysis focused on protein coding regions including the intronic boundaries. In the event that WES did not identify any clinically significant results, the unsolved cases underwent a more comprehensive analysis provided by WGS to cover both the coding (exonic) and non-coding (intronic and intergenic) segments of the genome.

The study described in this chapter aimed to increase the diagnostic yield in FH by utilising updated computational tools including a relatively new SV discovery

tool (SVRare) to reanalyse and reinterpret genomic data available for the unsolved FH families obtained locally or through collaborative links with members of the ERDC. Expertise in IRDs was also applied along with bioinformatics analyses, to detect the genetic aetiology in patients recruited to the 100KGP with a phenotype of, or genotype consistent with, FH.

## **4.2 Results**

### **4.2.1 Reanalysis of local unsolved FH cases**

#### **4.2.1.1 Summary of local unsolved FH cases**

Through collaboration with fellow researchers at the ERDC and with clinicians at local hospitals, six patients were identified with FH and without obvious pigmentary defects (Table 4.1). The phenotypic descriptions available for these patients are often incomplete. For example, a single proband (F1071) exhibited ASD but the exact manifestation was not disclosed. Only two patients were assessed for chiasmal misrouting and neither nystagmus nor hypopigmentation status were consistently reported. All six patients have a confirmed clinical diagnosis of FH but previous analyses in various laboratories did not detect causative mutations in known FH genes.

Genomic DNA from these six patients was analysed by WES, with library preparation and sequencing being performed by the University of Leeds NGS facility at St James's University Hospital. The raw whole exome data was obtained for bioinformatic analysis and checked for variants in *SLC38A8* and other candidate FH genes. The WES data was previously analysed by Dr Emma Lord as part of her PhD project at the time but no genomic diagnosis was obtained in the six patients (Lord, 2018). Five patients subsequently underwent further molecular testing using WGS performed by Edinburgh Genomics at the University of Edinburgh. The WES and WGS output files of BAM and VCF belonging to the unsolved six patients were revisited for an updated analysis using contemporary bioinformatic assessment tools, with updated population databases and ACMG classification.

Case ID	Phenotype	Consanguinity	Ethnicity	Gender	NGS
F1335	FH	Yes	South Asian	F	WGS
F1369	FH	No	European	F	WGS
F1071	FH, chiasmal misrouting and ASD	No	South Asian	M	WGS
F1377	FH, nystagmus and chiasmal misrouting	No	European	M	WGS
F1287 *	FH and no iris TID	Yes	South Asian	F	WGS
F1288 *	FH and no iris TID	Yes	South Asian	F	WES

**Table 4.1. Local cohort of unsolved FH.** Six patients displaying FH but without a molecular diagnosis. These cases represent an ethnically diverse cohort of European and South Asian descent. Three individuals of South Asian ancestry are the progeny of consanguineous relationships. Two patients from the same family are marked by “\*”.

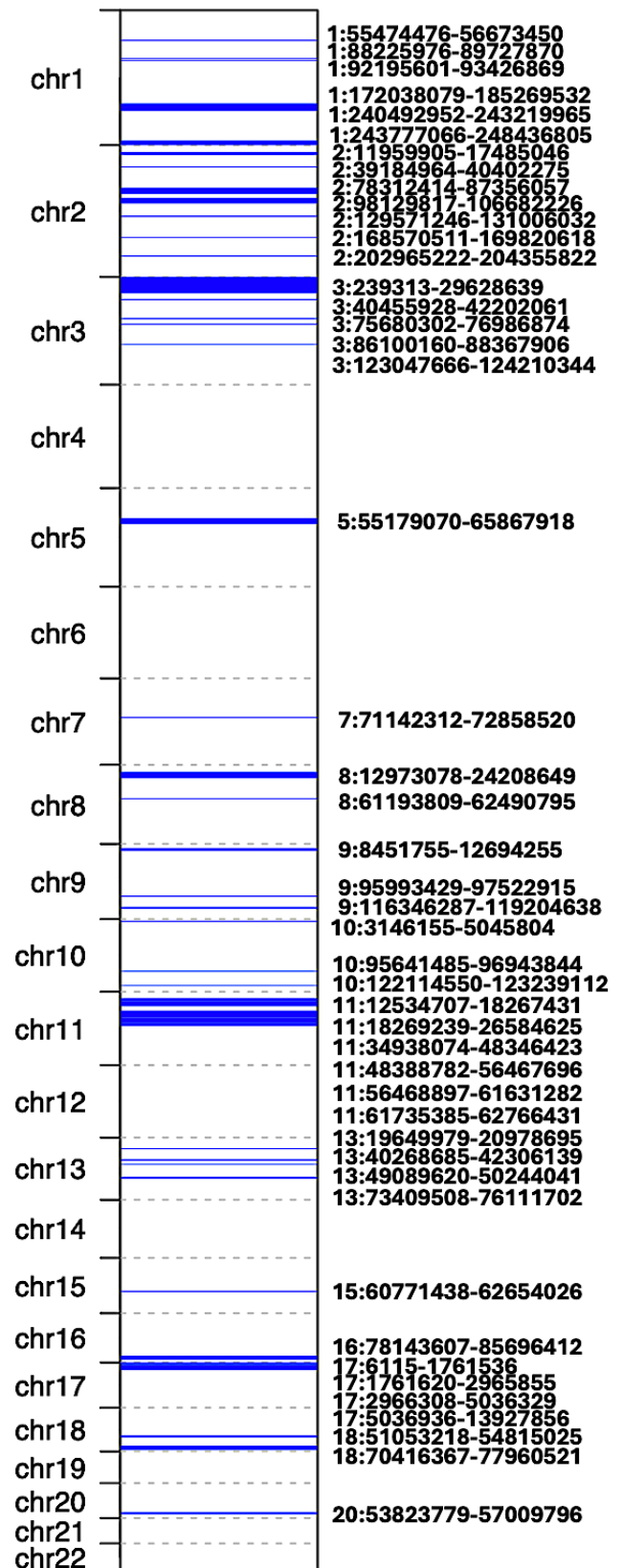
#### 4.2.1.2 Previous work performed by Toomes’s group on proband F1335

A single patient (F1335) of South Asian descent was described as having only FH with no additional phenotypic features reported. This proband was analysed using WES by Dr Emma Lord and using ExomeDepth for CNV discovery which revealed an apparent novel homozygous duplication in exon 6 of *SLC38A8*. The preliminary coordinates assigned to the SV was chr16:84050746-84056494 (GRCh37) (Lord, 2018). WGS was then carried out to deduce the breakpoints of the SV, but instead this showed that the *SLC38A8* duplication was an artefact. Therefore no pathogenic mutations were identified in *SLC38A8* (Lord, 2018), and the case was not examined further.

#### 4.2.1.3 Detecting a homozygous *HPS5* missense variant using autozygosity mapping in proband F1335

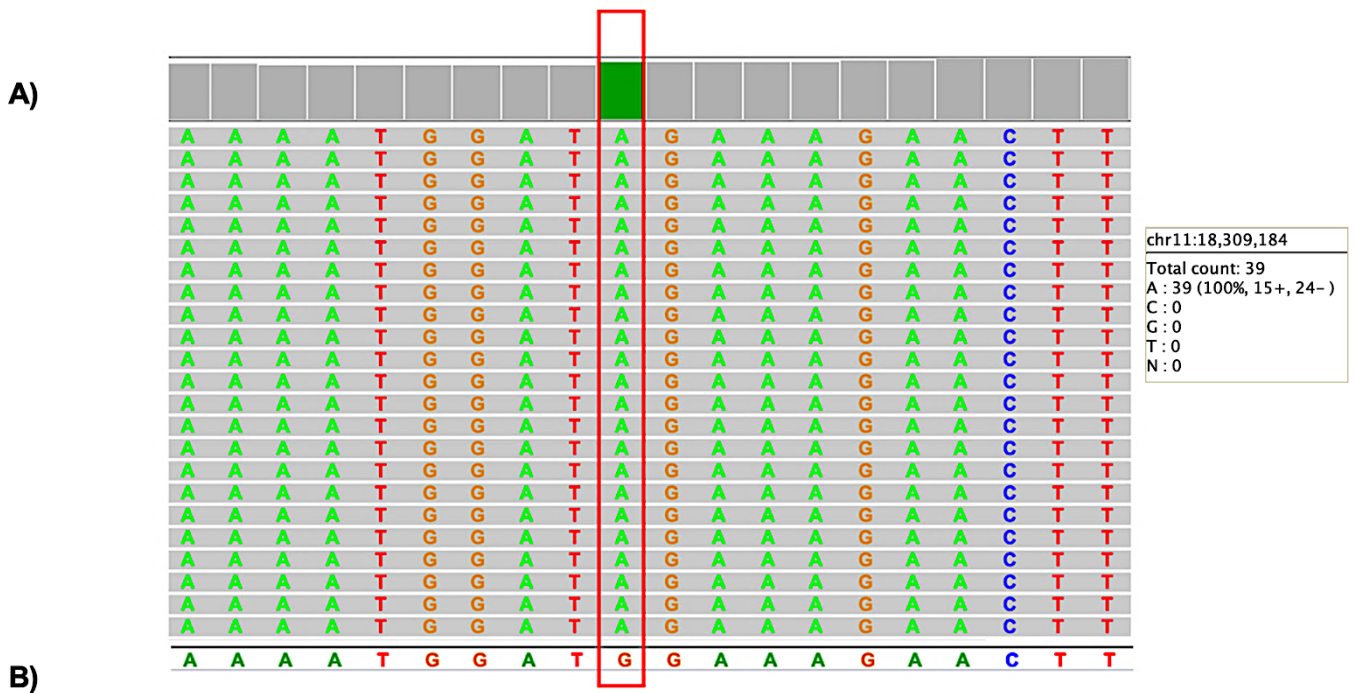
The presence of a consanguineous relationship between the parents of the proband, along with the apparent autosomal recessive inheritance in the pedigree, suggested autozygosity mapping could help to delineate the genetic aetiology in the proband. The exome of proband F1335 was annotated with autozygous regions also known as regions of homozygosity (ROH) using AutoMap (Figure 4.1). This analysis revealed multiple ROH, the 10 largest of which mapped according to GRCh37 on chr3:239313-29628639 (29.4 Mb), chr11:34938074-48346423 (13.4 Mb), chr1:172038079-185269532 (13.2 Mb), chr8:12973078-24208649 (11.24 Mb), chr5:55179070-65867918 (10.69 Mb), chr2:78312414-87356057 (9.0 Mb), chr17:5036936-13927856 (8.9 Mb), chr2:98129817-106682226 (8.6 Mb), chr11:18269239-26584625 (8.3 Mb) and chr11:48388782-56467696 (8.1 Mb).

**Figure 4.1. ROH identified in proband F1335.** Ideogram of autosomes 1-22 mapped with ROH. The ROH identified by AutoMap v1.0 are highlighted in blue and are portrayed based on the relative size of the ROH. The genomic coordinates provided are in GRCh37.





The already generated and available VCF was inspected, guided by the autozygosity mapping data, to restrict variant interpretation to variants residing within ROH. This selective approach led to the discovery of a homozygous missense variant in an ROH spanning chr11:18269239-26584625 (GRCh37). The missense variant was detected in *HPS5*, NM\_181507.2:c.2615C>T, NP\_852608.1p.(Pro872Leu) at chr11:18309184 (GRCh37). The variant allele was detected in all 39 reads at the locus and is consistent with a homozygous genotype (Figure 4.2). This *HPS5* variant is considered rare in gnomAD, with an allele frequency of 0.0000439, and was absent from the 100KGP. Various pathogenicity assessment tools unanimously predicted a deleterious outcome on protein function and the ACMG classification for this variant was VUS. The amino acid conservation profile of *HPS5* showed that the residue 872 is conserved in multiple species including *D. rerio* (zebra fish) and *X. tropicalis* (western clawed frog) but not in *C. elegans* (round worm) (Table 4.2).



Variant	Transcript	gnomAD	100KGP	CADD	Grantham	SIFT	PolyPhen-2	Align GVGD	ACMG
c.2615C>T (p.Pro872Leu)	ENST00000349215.8	4/91072	0	28.1	98	deleterious	Probably damaging	Class C65	VUS

**Figure 4.2. NGS discovery of *HPS5* variant in F1335.** A) A homozygous variant was detected in *HPS5* as depicted by the presence of the variant in all the reads at the locus. The reads in the BAM file were aligned to genome assembly GRCh37 and the depth of coverage for the variant allele is 39 X. B) Summarising the results of pathogenicity assessment tools and displaying the allele frequency for the South Asian demographic in gnomAD.

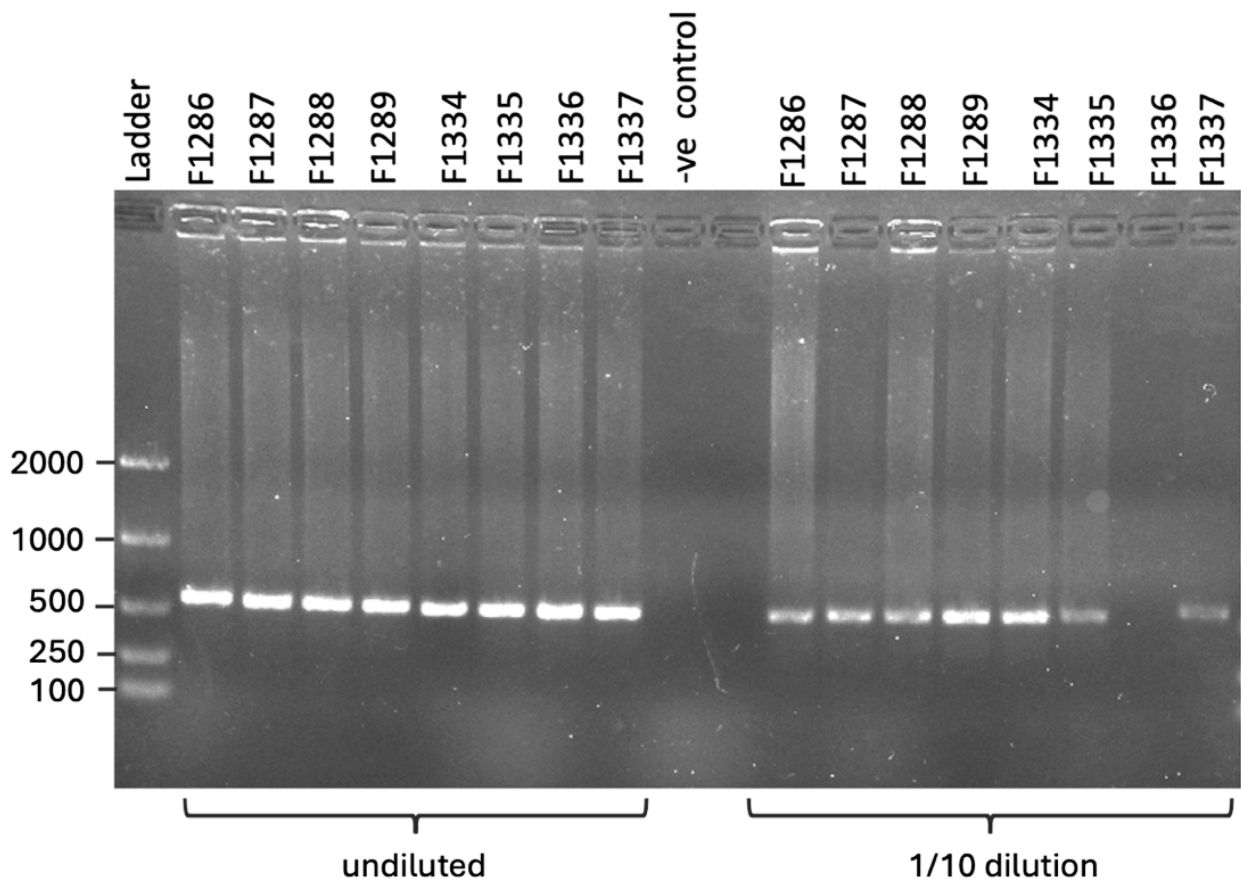
species	gene	aa	alignment
<i>H.sapiens</i>	<a href="#">ENSG00000110756.18</a>	872	ESALRSLIKFFPSILPSDIIQLCH
<i>P.troglodytes</i>	<a href="#">ENSPTRG00000003419</a>	872	ESALRSLIKFY <b>P</b> SILPSDIIQLC
<i>F.catus</i>	<a href="#">ENSFCAG00000011158</a>	831	ESALRSLIKFY <b>P</b> SILPSDIIQLC
<i>M.musculus</i>	<a href="#">ENSMUSG00000014418</a>	869	ESALRACIKFY <b>P</b> SISPSDIAQLC
<i>G.gallus</i>	<a href="#">ENSGALG00000006266</a>	883	ETALRPLVKFY <b>P</b> SIVPSDIKQ
<i>T.rubripes</i>	<a href="#">ENSTRUG00000008835</a>	854	EAAVRSLSQFY <b>P</b> TVTPADVAAMS
<i>D.rerio</i>	<a href="#">ENSDARG00000071062</a>	856	EVAVRAYPQFY <b>P</b> TILPSDIMAM
<i>X.tropicalis</i>	<a href="#">ENSXETG00000016685</a>	840	ETAIRSLTRFY <b>P</b> SIVPLDVMQLC
<i>C.elegans</i>	<a href="#">W09G3.6</a>	287	QIPRAPPISF-RSAFPTDSAQFC

**Table 4.2. Amino acid conservation at HPS5 variant site.** Multiple sequence alignment of HPS5 orthologs ranging from humans to nematodes. The Ensembl reference ID for each ortholog is displayed and the adjacent column displays the position of the first amino acid residue of the corresponding protein sequence, *Homo sapiens*: humans; *Pan troglodytes*: Chimpanzee; *Felis catus*: domestic cat; *Mus musculus*: house mouse; *Gallus gallus*: chicken; *Takifugu rubripe*: japanese pufferfish; *Danio rerio*: zebrafish; *Xenopus tropicalis*: western clawed frog; *Caenorhabditis elegans*: roundworm.

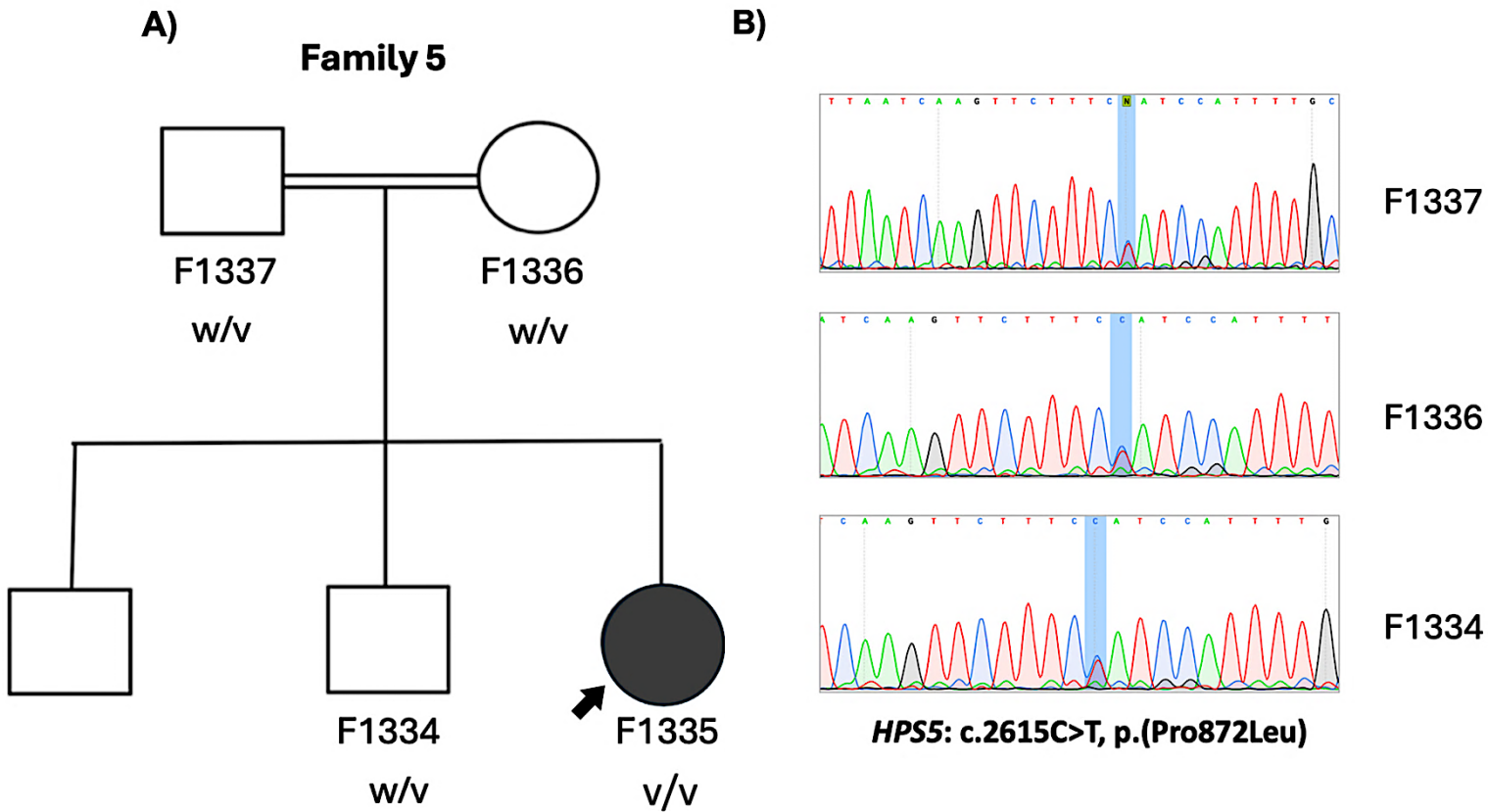
#### 4.2.1.4 Segregation analysis of the HPS5 missense variant

The homozygous variant detected in *HPS5* may suggest a potential genomic diagnosis of an albinism syndrome known as HPS (section 1.15.3.1), which would explain FH in proband F1335 (Zhang et al., 2003). DNA samples from four available members of family 5 (F5) along with another unrelated family (F6) were subjected to whole genome amplification (WGA) in order to increase the available DNA stock to support subsequent genetic analysis. The amplicons were further diluted in deionised water at a dilution factor of 1/10 and 1/50. The undiluted and diluted amplicons were tested by PCR targeting the *HPS5* gene to confirm successful genome amplification and for determination of a reliable DNA working concentration (Figure 4.3). Primers (TTCATAAATTGAATATTTCCGCC and GAGAAACAATGCCCTGTCAAA) were designed to target exon 18 of *HPS5*, with minimal self-complementarity at the 3' (Appendix C: Supplementary Table 4.1). The undiluted and 1/10 DNA dilution resulted in successful PCR amplification. However, the 1/50 dilution had a very low DNA concentration that did not generate any products (data not shown).

Sanger sequencing was performed on the *HPS5* amplicons for the members of family F5 to confirm the NGS findings, ensuring that the detected *HPS5* variant is not an artefact, and to increase confidence in this clinically actionable result (Figure 4.4). Each individual in family F5 was genotyped for the presence or absence of NM\_181507.2:c.2615C>T; NP\_852608.1:p.(Pro872Leu) in *HPS5*. The *HPS5* variant was present in a heterozygous state in all the asymptomatic individuals tested (F1336, F1337 and F1334).



**Figure 4.3. Validation of WGA in two unsolved FH families.** PCR products of *HPS5* amplification in members of two families (F5 and F6). Family F5 is composed of F1334, F1335, F1336 and F1337 while another family (F6) constitutes F1286, F1287, F1288 and F1287. All samples tested have generated the expected amplicon of 600 bp except for one sample. All samples were loaded on a 1% agarose gel and were run at 95 V for 40 minutes.



**Figure 4.4. Segregation analysis of *HPS5* variant in family F5.** Sanger sequencing data on three family members. A) Pedigree summarising the *HPS5* genotype in family F5 with “v” symbolising the variant allele and “w” for wildtype. B) Electropherograms generated for three family members, with the variant site highlighted in blue. Heterozygous genotype is apparent by the two peaks of equal intensity at position c.2615 representing both the wildtype in blue (cytosine) and variant allele in red (thymine). Note that the peak amplitude for heterozygous alleles are much lower in comparison to homozygous alleles. DNA traces visualised on 4Peaks (Nucleobytes) version 1.5.

## 4.2.2 Identification of SVs using SVRare in local unsolved FH cases

### 4.2.2.1 Pilot study: application of SVRare on solved FH case F1310

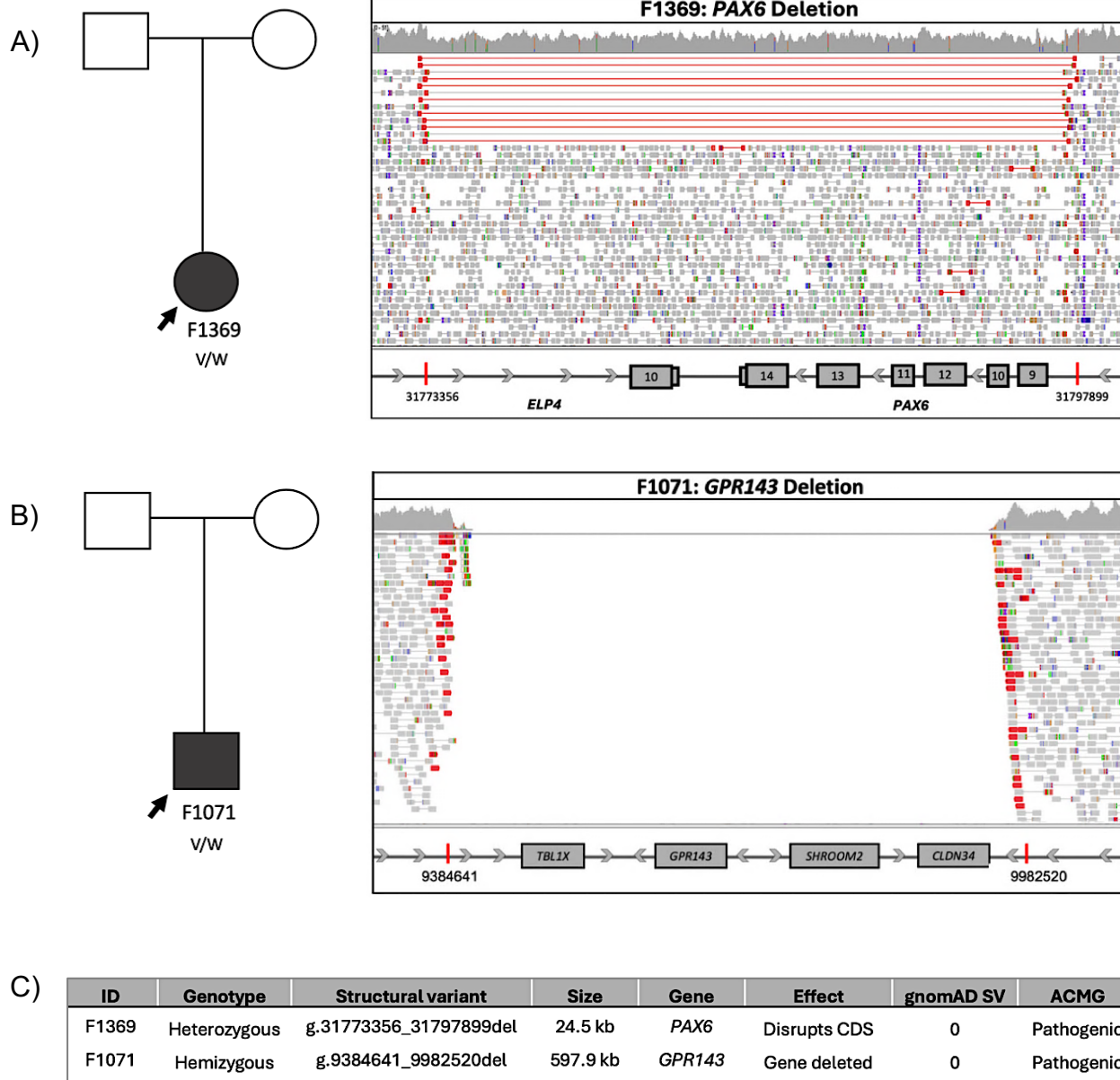
Proband F1310 presented with FH and nystagmus but chiasmal misrouting was not evaluated using VEP. This individual is from a Middle Eastern (Iranian) family of consanguineous marriage. Previous work by Dr Emma Lord using WGS in F1310 revealed an *SLC38A8* inversion that spans chr16:84015931-84027116 (Figure 3.1). The 11.2 kb inverted sequence includes exon 7, 8 and 9 which disrupts the CDS. Segregation analysis confirmed the co-segregation of NC\_000016.10:g.84015931\_84027116inv with the FH phenotype in the pedigree.

Proband F1310 served as a control to test the efficiency of SVRare, developed by collaborating UK Inherited Retinal Disease Consortium (IRDC) member Jing Yu for SV identification. Dr Yu analysed the Leeds unsolved FH cases for which SV calls by MANTA and CANVAS were available, then used the SVRare pipeline to compile and annotate all *SLC38A8* SVs in F1310. Bioinformatic analysis by Dr Yu has narrowed down the genetic aetiology in proband F1310 to a single SV, the known inversion in *SLC38A8*, NC\_000016.10:g.84015931\_84027116inv (Figure 3.1). The BAM file for proband F1310 was revisited and manually inspected to rule out artefacts and to confirm the coordinates of the SV detected. The NC\_000016.10:g.84015931\_84027116inv was reanalysed based on allele frequency in gnomAD SV and the 100KGP. The SV was also assessed for the predicted effect and compliance with the mode of inheritance related to the phenotype (Figure 3.2)

#### **4.2.2.2 The reanalysis of a heterozygous *PAX6* deletion in proband F1369 and a hemizygous *GPR143* deletion in proband F1071**

The successful pilot study of SVRare in proband F1310 displayed the potential benefit of SVRare application in the local cohort of unsolved FH patients. The VCF files of probands F1369 and F1071 generated by WGS were also reanalysed by Dr Yu using SVRare. Proband F1369 represents a sporadic case of FH in a proband of European descent. The analysis of the annotated SVs in proband F1369 has uncovered, among other CNVs, a heterozygous deletion of 24.5 kb on chromosome 11. The detected SV was confirmed by the author through reassessing the BAM file and the preliminary breakpoints assigned to this deletion were chr11:31773356-31797899 (Figure 4.5). This CNV encompasses exon 9 -14 of *PAX6* and exon 10 of *ELP4*, and has been ACMG classified as pathogenic. *PAX6* is a well characterised gene underlying autosomal dominant FH manifestations and monoallelic LOF variants are a likely cause of aniridia as opposed to isolated FH (Azuma et al., 1996, Tzoulaki et al., 2005). This convincing *PAX6* deletion was missed in previous investigations by Dr Emma Lord using ExomeDepth, BreakDancer and DELLY2 for SV discovery (Lord, 2018).

The South Asian proband F1071 presented with FH, chiasmal misrouting and anterior segment abnormalities. No hypopigmentation was reported and an iris TID was not observed in this patient. Initial analysis of WES and WGS data by Dr Emma Lord did not identify the genetic aetiology for FH in this proband (Lord, 2018). The proband's genome was therefore reanalysed by MANTA and CANVAS and the SV calls were merged and annotated using the SVRare tool by Dr Jing Yu. This uncovered multiple SVs including a CNV encompassing *GPR143*. The male proband was hemizygous for a large deletion of 597.9 kb on the X chromosome that was classified as pathogenic by ACMG criteria. The proband's BAM file was revisited by the author which confirmed the deletion to be located at chrX:9384641-9982520 and that it deletes four entire genes constituting *TBL1X*, *GPR143*, *SHROOM2* and *CLDN34*. Pathogenic *GPR143* variants are a known cause of X-linked ocular albinism (Bassi et al., 1995), making this a strong candidate pathogenic variant in this proband.



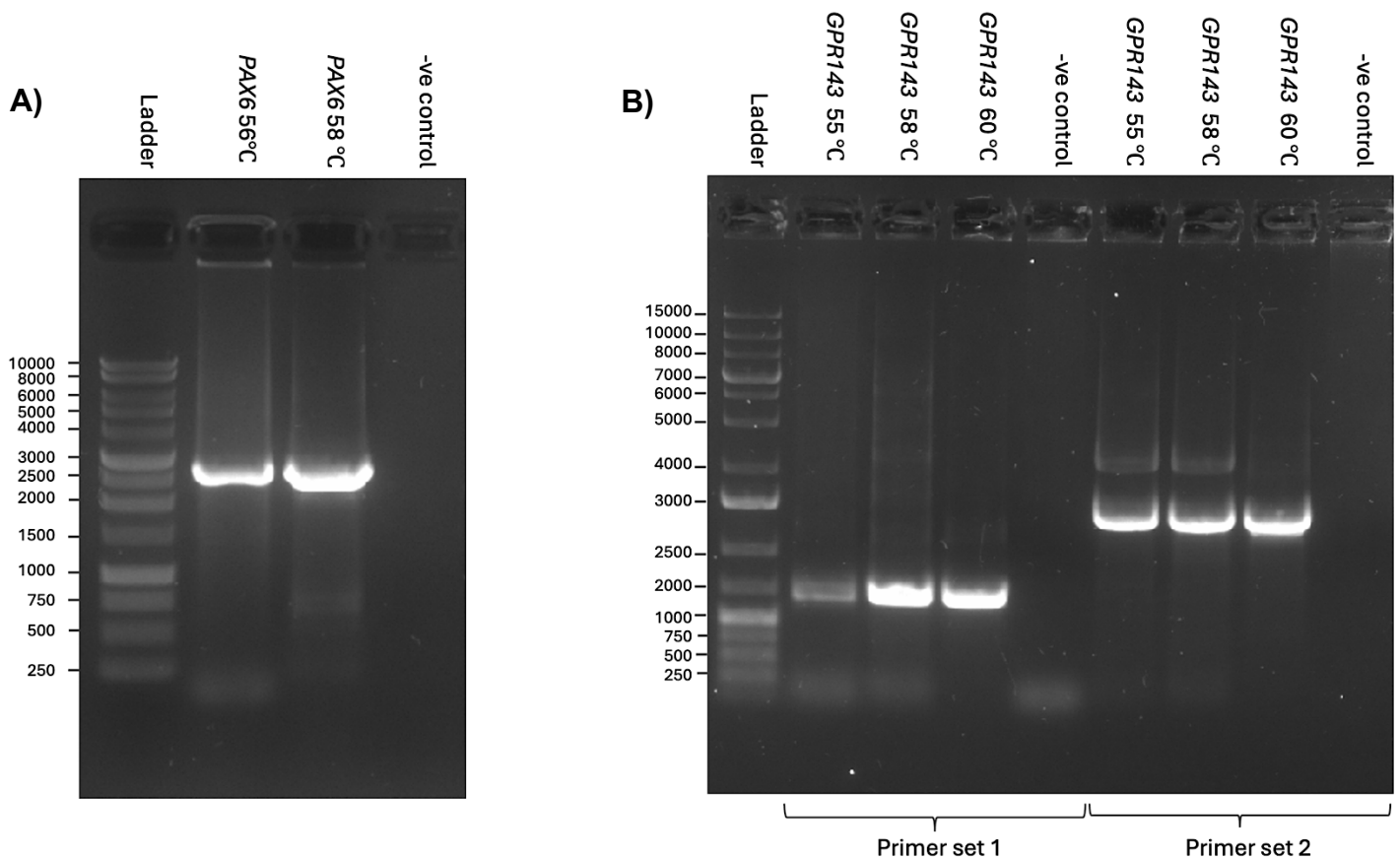
**Figure 4.5. Genetic characterisation of deletions in F1369 and F1071.** A) Representation of a family with sporadic or autosomal recessive inheritance of FH in individual F1369. BAM file displaying discordant primer pairs with deletions represented by a red marker to indicate an alignment insert larger than the expected size. The coverage track displays the Sashimi plot for reads along the entire window, with a visible reduction in coverage in the deleted region (chr11:31773356-31797899) indicating a heterozygous deletion. The deletion encompassing *PAX6* was detected in a region with a read count of 51. The genetic map at the bottom depicts exons as grey rectangles with the arrows matching the orientation of the sequence and the vertical red lines at the breakpoints. B) Pedigree of unaffected parents and a single affected male offspring, F1071. BAM file displaying no reads in the deleted region of chrX:9384641-9982520, indicating a hemizygous deletion. The coverage at the region preceding the deletion is 37 X. The schematic of the deleted region displays four deleted genes as rectangular grey boxes with vertical red lines defining the deduced breakpoints and arrows for the orientation of the sequence. C) Genetic characterisation of the *PAX6* deletion (NC\_000001.11:g.31773356\_31797899del) and *GPR143* deletion (NC\_000023.11:g.9384641\_9982520del) is tabulated along with the phenotype and additional information such as the size of the deleted region and details of the proband's clinical manifestation, gender and ethnic demographic.

#### 4.2.2.3 Characterisation of identified SVs using ONT sequencing

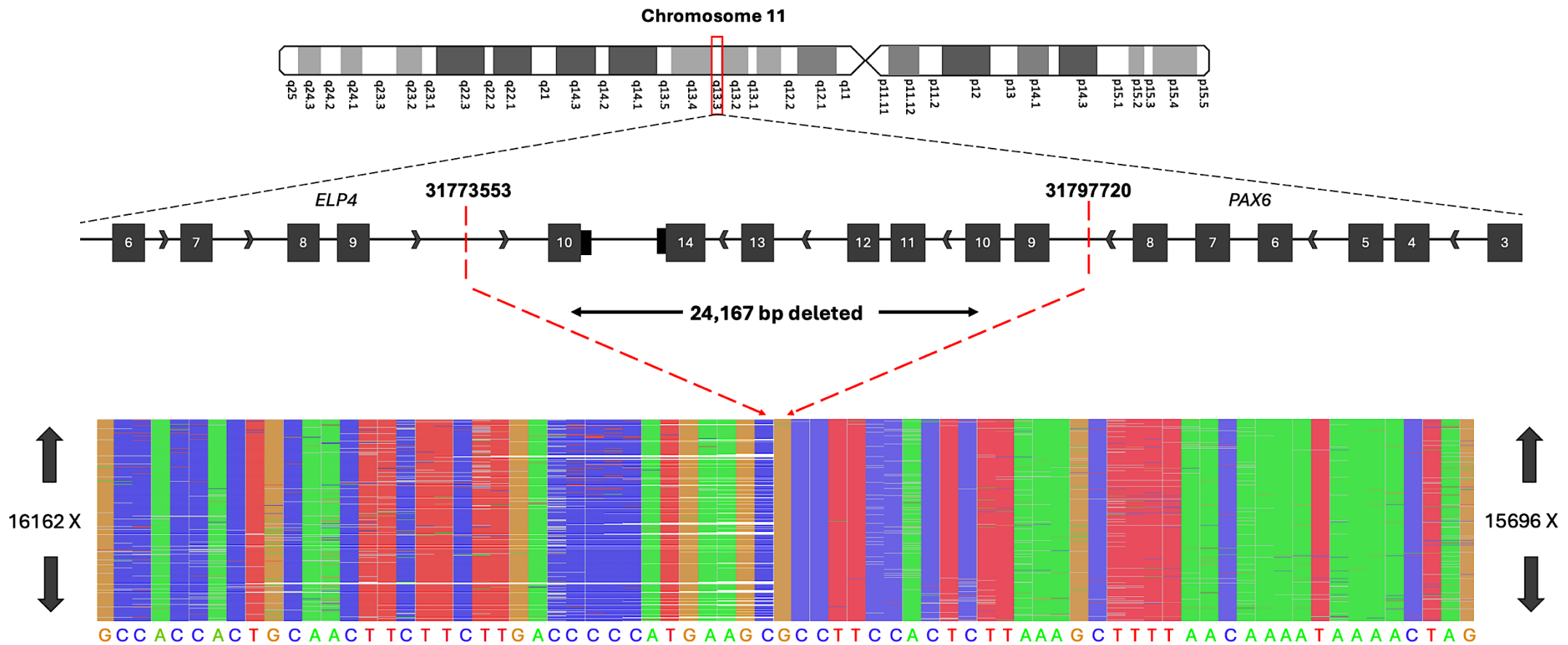
The SVs detected in the local unsolved FH cohort required a secondary test for confirmation of the NGS findings. Long read ONT sequencing using a MinION instrument was performed to validate and characterise the identified SVs in probands F1369 and F1071 (Figure 4.5). LR-PCR was conducted as described in section 2.3.1.3 to generate large amplicons (up to 20 kb) using primers flanking the chr11:31771367-31798086 interval in F1369 and spanning the chrX:9384591-9982856 and chrX:9384225-9983186 region in F1071 (Appendix C: Supplementary Table 4.2). Given the size of the deletion in each case, these primers would not be expected to generate any product for the wildtype alleles. The expected amplicon with the deletion allele in *PAX6* is 2607 bp, and the amplicons with deletion encompassing *GPR143* are 1640 bp and 2336 bp for the two primer sets used. The amplicon for the wildtype allele of *PAX6* is 26719 bp and for *GPR143* these are 598961 bp and 598961 bp, all of which would fail to amplify with SequalPrep<sup>™</sup> long-PCR polymerase that is limited to 20 kb products. As a result, the PCR thermocycler parameters used for the extension step (1 minute/kb) was adjusted to amplify just the variant allele. This introduced a PCR bias, making the variant alleles appear falsely as homozygous (Figure 4.6).

The amplified products of the variant alleles in probands F1369 and F1071 were subjected to ONT sequencing using Flongle flow cells for accurate determination of the deletion breakpoints. These PCR products were pooled with other plasmids or distinct amplicons for multiplexed sequencing. The bioinformatic script for the generation of BAM file is provided in Appendix B: section 8.2.3. The higher coverage and greater mapping accuracy of long reads refined the breakpoints for the deletion in F1369 to chr11:31773553-31797720 (Figure 4.7). The region upstream of the deletion has a coverage of 16162 X and downstream the deletion is 15696 X. As for the large deletion in chromosome X of F1071, the deletion is mapped to chrX:9384915-9982211 which eliminates a region of 597.3 kb (Figure 4.8). The read count surrounding the deleted region is 49135 X.

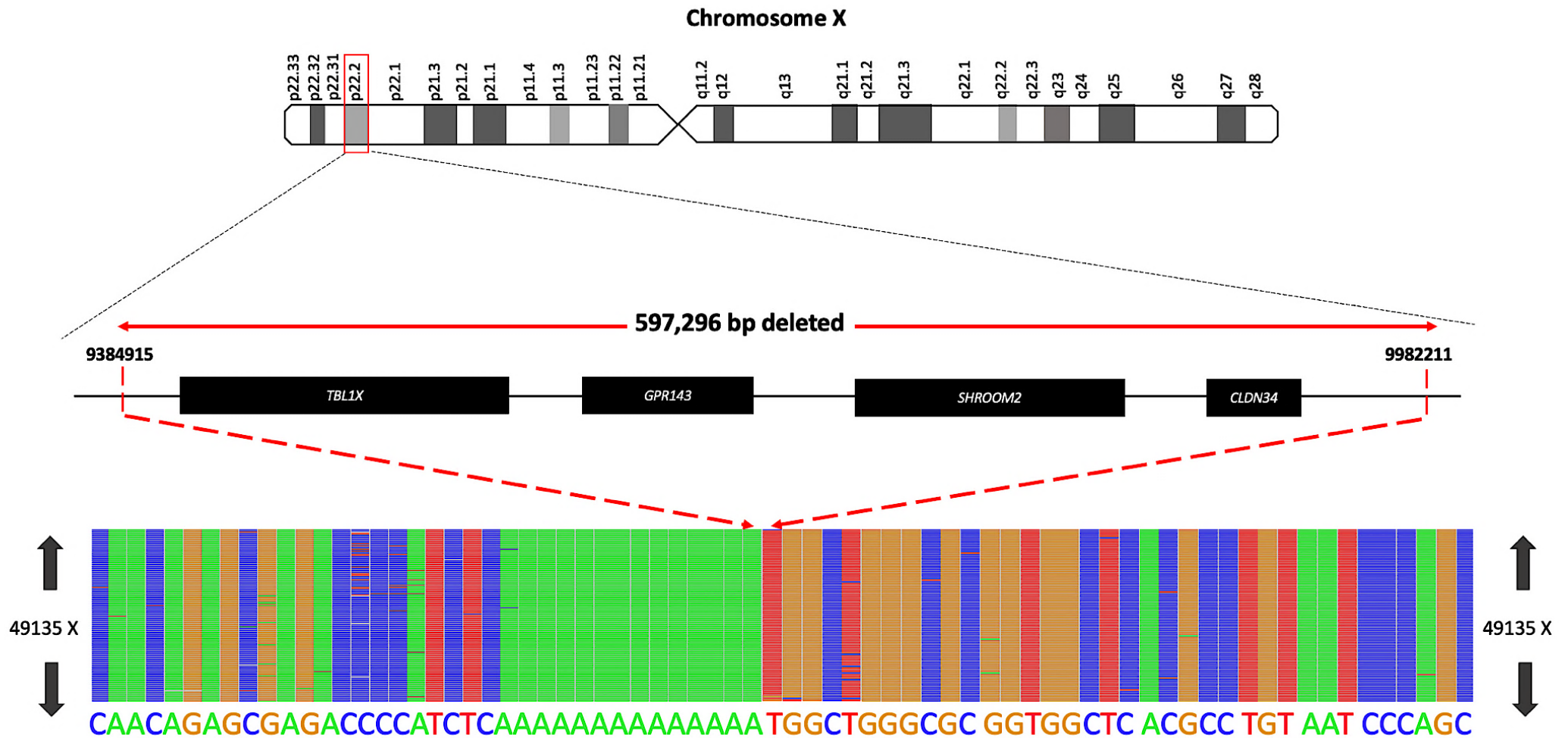




**Figure 4.6. Optimisation of LR-PCR on unsolved FH cases.** A) Amplification of *PAX6* in F1369 results in a visible discrete band of 2607 bp for the deleted allele and the wildtype amplicon of 26719 bp was not amplified. B) The PCR product for the variant allele in *GPR143* is 1640 bp with primer set 1 and 2336 bp for primer set 2. The amplicons for the wildtype alleles would be 598265 bp for primer set 1 and 598961 for primer set 2. Both wildtype amplicons massively exceed the capacity for PCR amplification by LR-PCR, so the wildtype allele did not generate any amplicons. Nonspecific amplification products of approximately 4000 bp produced by primer set 2 were eliminated at 60°C.



**Figure 4.7. Characterisation of *PAX6* deletion using long read sequencing.** Schematic of a locus on chromosome 11q13.3 that is affected by a large deletion spanning chr11:31773553-31797720 in proband F1369. The G-banding pattern of chromosome 11 is in accordance with the International System for Human Cytogenomic Nomenclature (ISCN). The exons are depicted in black rectangles and the arrows indicate the orientation of the gene. Red arrows points at the terminal nucleotide at the breakpoint. The depth of coverage at each breakpoint is provided on either side. The deletion is present in the middle of the read with part of the sequence mapping upstream the deletion and another part mapping downstream the deletion.

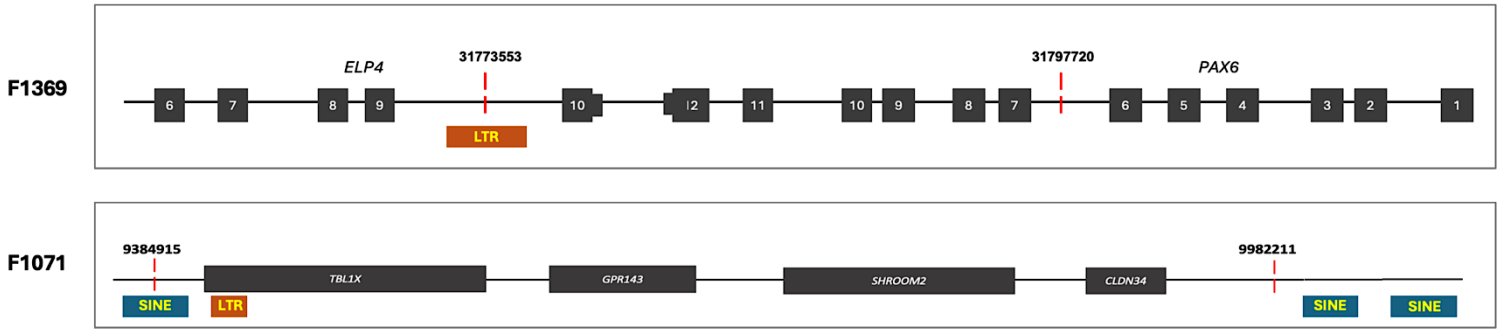


**Figure 4.8. Characterisation of an SV in F1301 using long read sequencing.** Chromosome X ideogram with emphasis on Xp22.2 which harbours a large deletion at chrX:9384915-9982211. Genes within the deleted region are represented by black rectangular boxes. The nucleotide sequence at the break point is marked by red arrows and the corresponding read count at each breakpoint is displayed. Sequencing data belongs to amplicons of *GPR143* primer set 2.

#### 4.2.2.4 Investigating the genomic features at the SV breakpoints

The genomic landscape at the defined breakpoints for the deletions detected in *PAX6* of F1369 and *GPR143* of F1071 was investigated to look for features of the sequence that might have given rise to these CNVs. The 50 bp region downstream and upstream of each breakpoint for these CNVs was explored for evidence of interspersed repeats, including SINEs, LTRs, long interspersed nuclear elements (LINEs) and low complexity regions. The loci investigated were chr11:31773503-31773603 and chr11:31797670-31797770 for F1369 and chrX:9384865-9384965 and chrX:9982161-9982261 for F1071. The data from RepeatMasker track on UCSC genome browser (<https://genome-euro.ucsc.edu/index.html>) shows that the SV in *PAX6* has an overlapping LTR at only one breakpoint at chr11:31773553 while the deletion encompassing *GPR143* has SINEs at or nearby both breakpoints at chrX:9384915 and chrX:9982211 (Figure 4.9). In total four repeat elements comprising three SINES of 312, 303 and 276 bp, and a single LTR of 162 bp are in the vicinity of NC\_000023.11:g.9384915\_9982211del in *GPR143* of F1071.

Investigating microhomology surrounding the SV breakpoints using Basic Local Alignment Search Tool (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>) did not reveal any sequence homology between *PAX6* breakpoints (chr11:31773553-31797720) that is indicative of non allelic homologous recombination. However, in a region 200 bp upstream (chrX:9384715-9384915) and 200 downstream (chrX:9982211-9982411) the SV breakpoints at chrX:9384915-9982211, there is a 50% sequence homology with 95% identical sequence between the two distant loci (Figure 4.10).

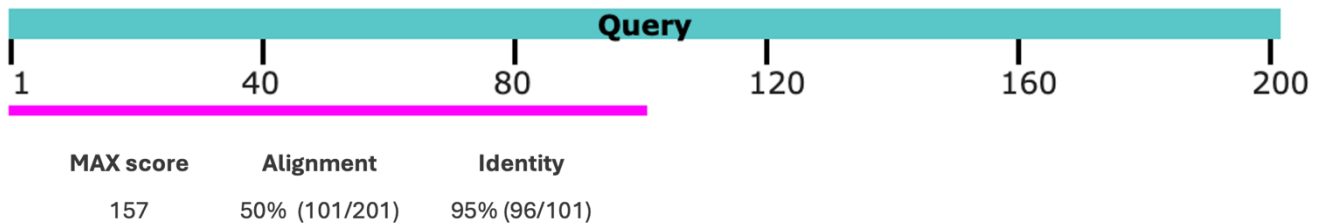


ID	Structural variant	Breakpoint 1			Breakpoint 2		
		Repeat element	Locus	size	Repeat element	Locus	size
F1369	g.31773553_31797720del	LTR (LTR67B)	11:31773266-31773866	526 bp	-	-	-
F1071	g.9384915_9982211del	SINE (AluY)	X:9384618-9384929	312 bp	SINE (AluJb)	X:9981909-9982184	276 bp
		LTR (MLT1J)	X:9384930-9385091	162 bp	SINE (AluY)	X:9982213-9982515	303 Bp

**Figure 4.9. Annotation of genomic region at the breakpoints of characterised SVs in F1369 and F1071.** Exons are numbered and are depicted in black squares while genes are represented as black rectangles. The breakpoint of the SVs is marked by a red line. The detected LTRs and SINEs are symbolised as orange and blue rectangles, respectively and are scaled based on their relative size.

```

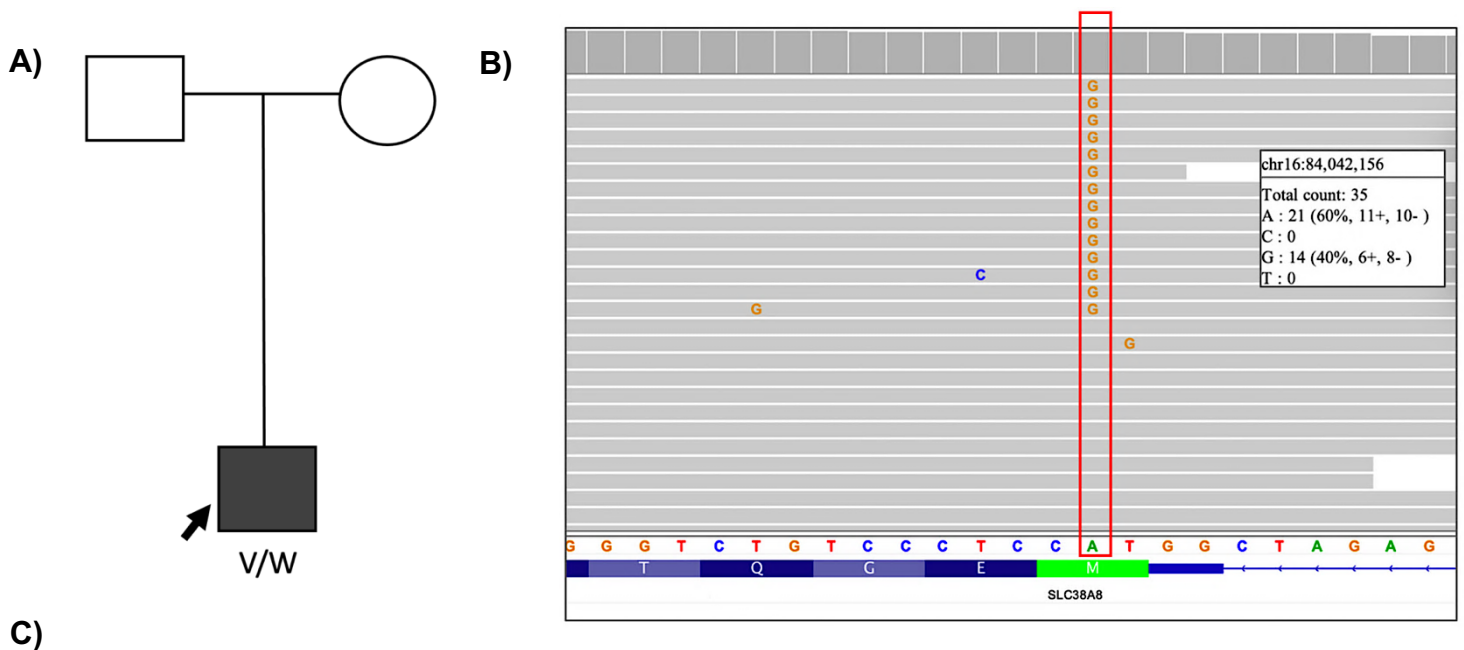
Query 1 TGGTGAAACCCCGTCTCTACTAAAATACaaaaaaaaTAAGCCGGGCGTGGTGGCGGGCGC 60
Sbjct 100 TGGTGAAACCCCGTCTCTACTAGAAAATACAAAAAA-TAAGCTGGGCGTGGTGGCGGGCAC 158
Query 61 CTGTAGTCCCAGCTACTCGGGAGGCTGAGGCAGGAGAATGG 101
Sbjct 159 CTGTAGTCCCAGCTACTCAGGAGGCTGAGGCAGGAGAATGG 199
    
```



**Figure 4.10. Sequence homology surrounding the NC\_000023.11:g.9384915\_9982211del breakpoints in F1071.** A) Alignment of a 101 bp sequence between chrX:9384715-9384915 and chrX:9982211-9982411. B) Visual representation of sequence alignment with the queried sequenced represented in teal and the aligned sequences in magenta. Maximum score (MAX score) is the highest quality score for the alignment achieved.

#### 4.2.2.5 Reanalysis of F1377 genome

A single proband from the unsolved FH cohort (F1377) presented with FH, chiasmal misrouting and without pigmentation defects (Figure 4.11A). WGS outsourced to Edinburgh genomics (University of Edinburgh) and subsequent bioinformatic analysis by Dr Emma Lord detected a heterozygous *SLC38A8* variant NM\_001080442.3:c.2T>C, NP\_001073911.1:p.(Met1?) that is predicted to affect the start codon (Lord, 2018). Validation of this NGS finding was performed using Sanger sequencing, which confirmed NM\_001080442.3:c.2T>C, NP\_001073911.1:p.(Met1?) the presence of this variant (Lord, 2018). This SNV was reanalysed by the author, and various pathogenicity assessment tools unanimously predicted a deleterious effect. However, the variant was classified as a VUS by ACMG (Figure 4.11B and C). The proband F1377 remains without a genomic diagnosis as no additional plausible variants were detected in *SLC38A8* or in other FH associated genes.



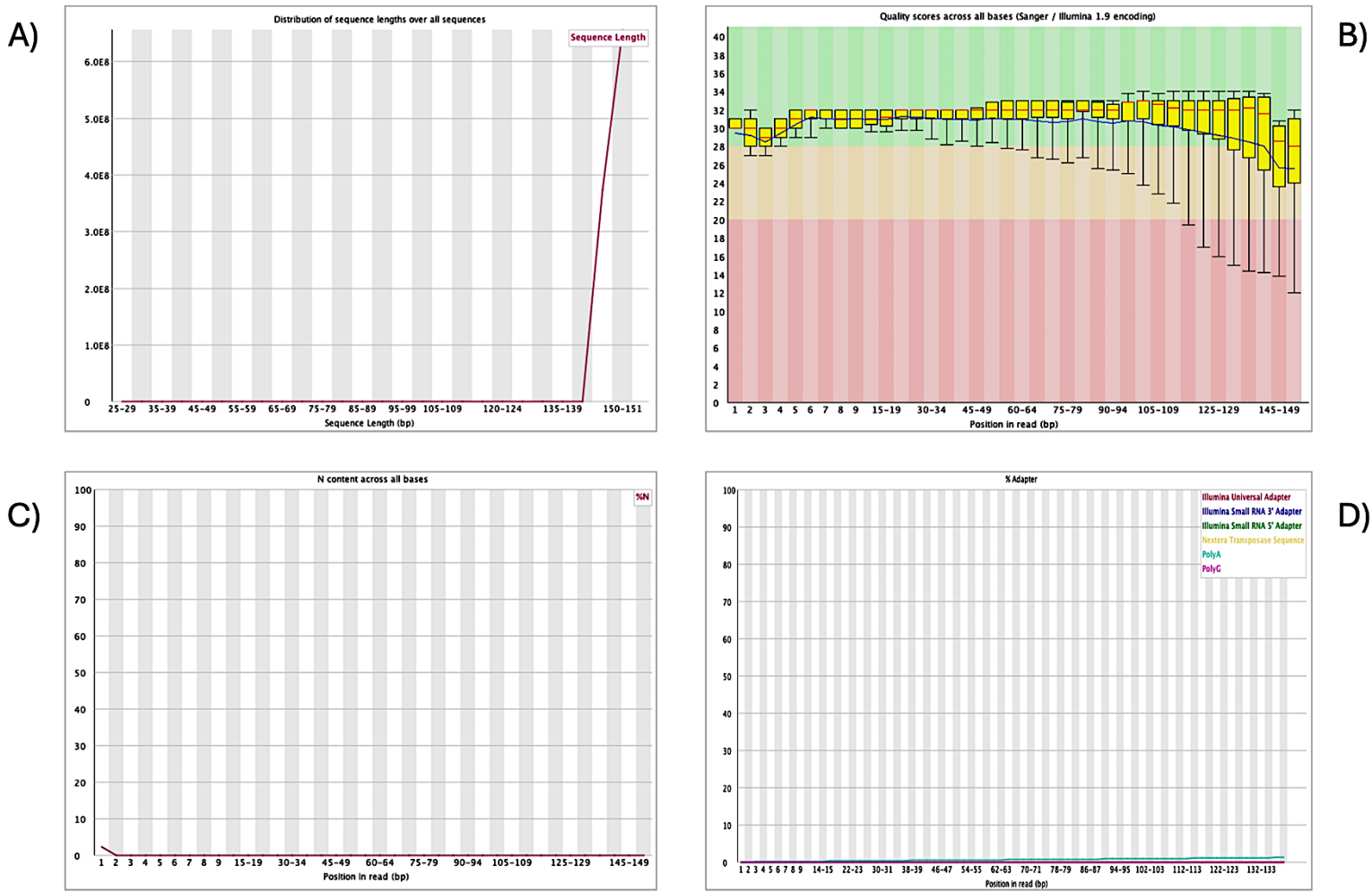
**Figure 4.11. NGS identification of NM\_001080442.3:c.2T>C, NP\_001073911.1:p.(Met1?) in F1377.** A) Pedigree displaying unaffected parents with an affected offspring. B) Variant NM\_001080442.3:c.2T>C, NP\_001073911.1:p.(Met1?) is detected in 60% (21/35) of the reads at chr16:84042156 which coincides with exon 1. C) Pathogenicity assessment using four different computational tools and the allele frequencies are derived from two population databases.

variant	Transcript	Gene	gnomAD	100KGP	Ensemble VEP	SIFT	Polyphen-2	CADD	ACMG
c.2T>C, p.(Met1?)	ENST00000299709.8	SLC38A8	8/1178240	3/126704	High Impact	Deleterious	Probably damaging	19.56	VUS

A reinvestigation of the WGS data was performed for this project, to look for a second potentially pathogenic variant in *SLC38A8* that might have been missed in previous analyses. The read alignment in the BAM file was evaluated for quality control using FASTQC (Appendix B: section 8.2.4). The BAM file contained 1033870317 reads with lengths between 139-151 bp and the average Phred quality score being  $\geq Q28$  for each base in the reads but at the terminal region (145-149 bp) it decreases to Q26 (Figure 4.12A and B). The base calling at each position in the reads has negligible or no ambiguous base calls assigned "N" (Figure 4.12C). No contaminating adapter sequences were detected in the reads except for an insignificant percentage of poly adenosine signal (Figure 4.12D). The satisfactory quality control results of the BAM file provided confidence in using the already generated VCF for proband F1377.

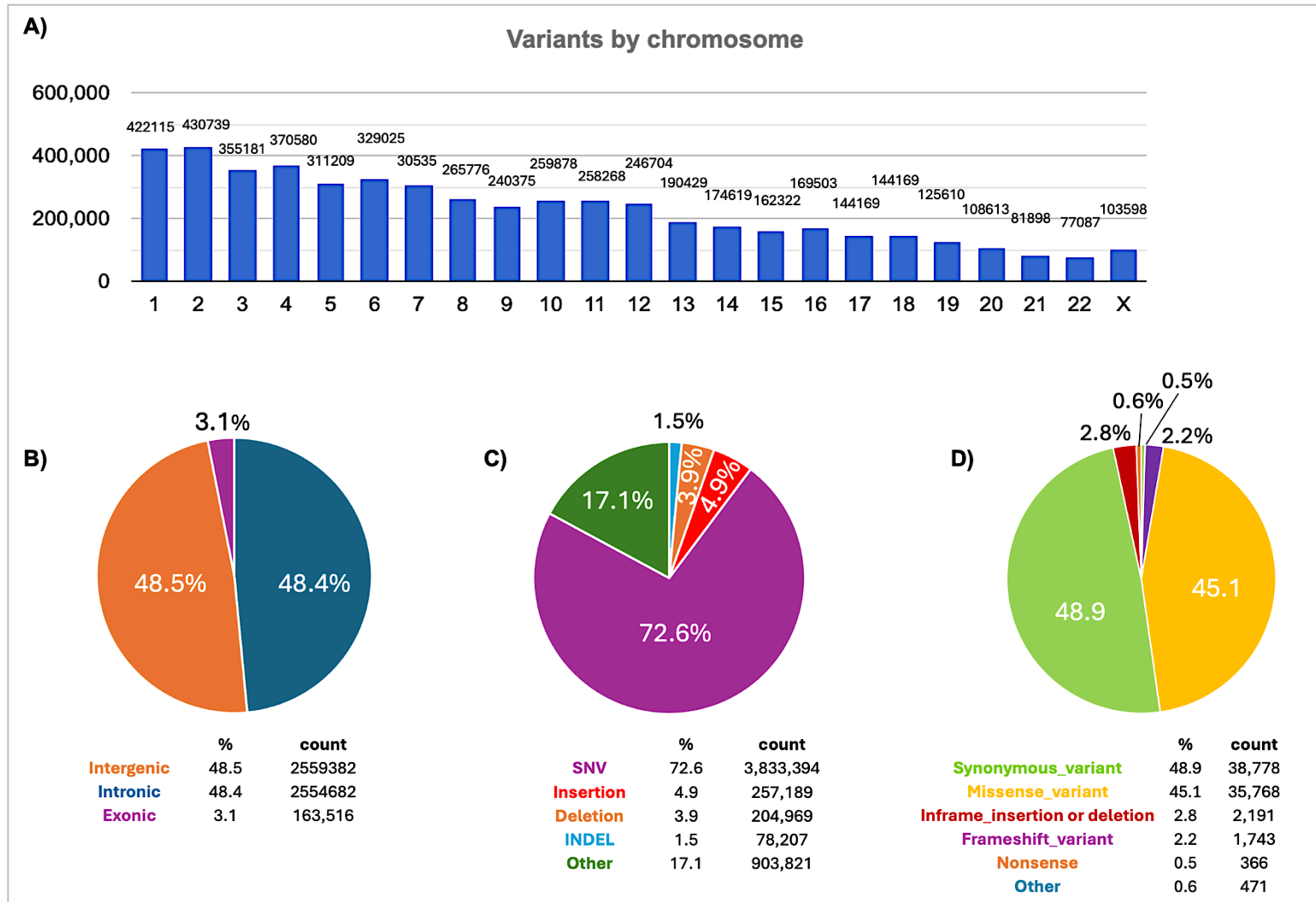
The available VCF generated by Emma Lord contained variant calls in GRCh37 that were filtered to exclude variants with MAF  $>0.01\%$  in dbSNP or Exome Aggregation Consortium (ExAC). This VCF was reannotated using Ensembl VEP, CADD, SpliceAI and UTRannotator, to expand the previous genomic investigation to include noncoding variants that interfere with mRNA splicing or variants at the 5' UTR that could potentially disrupt the ORF. Three scripts were used to download the Ensembl VEP and the VEP plugins for CADD, SpliceAI and UTRannotator, followed by whole genome annotation based on GRCh37 to match the coordinates in the VCF file (Appendix B: section 8.2.5). The whole genome except for the Y chromosome was characterised, with 5277580 variants identified (Figure 4.13A). The majority of variants detected were in intergenic (48.4%) and intronic (48.3%) regions of the genome (Figure 4.13B). The most frequent genetic alterations were SNVs (72.6%) (Figure 4.13C). Within the CDS the most common molecular consequences are synonymous variants (48.9%) and missense changes (45.1%) (Figure 4.13D). The whole genomic annotation file was split by chromosome to generate smaller files that require less computational power to analyse (Appendix B: section 8.2.6). A locus at chr16:84036349-84084427 (GRCh37) which encompasses *SLC38A8* and regions upstream and downstream the gene was interrogated. This locus was filtered for variants with a CADD score  $\geq 15$  or a SpliceAI delta score  $\geq 0.5$  for any of splice donor or splice acceptor gain or loss (Appendix C: Supplementary table

4.3). Only the previously detected NM\_001080442.3:c.2T>C; NP\_001073911.1:p.(Met1?) was identified.



**Figure 4.12. Read quality control checks using F1377 BAM file.** A) Read length distribution. B) Quality score distribution across the entire reads. The box plot shows the lower quartile (25<sup>th</sup> percentile) up to upper quartile (75<sup>th</sup> percentile) in yellow with a red horizontal line depicting the median (50<sup>th</sup> percentile). The upper whisker is the 90<sup>th</sup> percentile and the bottom whisker is the 10<sup>th</sup> percentile. Blue line is the mean quality score for each position in the read. Region in green indicates a high quality score, yellow is for moderate and red is a low quality score. C) Percentage of undetermined bases at each position across all reads. D) Proportion of reads with contaminating adapter sequences.





**Figure 4.13. Variant statistics in F1377 genome.** A) Variant count on chromosomes 1-22 and the X chromosome. A total of 5,277,580 variants were detected. B) Variant classification by genomic region. Variants within coding segments of the gene (exonic), within the non coding segments of a gene (intronic) or between genes (intergenic). C) Variant types. SNV relates to a single nucleotide change. Deletion or insertions refers to an individual deletion or insertion event affecting  $\geq 2$  nucleotides. INDEL refers to both an insertion and deletion event of  $\geq 2$  nucleotides. Others includes SVs and repeat elements D) Molecular consequences for the CDS. A total of 7937 variants were identified that are predicted to affect the CDS. The category “Other” encompasses variants affecting the start or termination codon. The text is colour coded to correspond to the region of interest on the pie chart.

Investigating additional 51 FH genes in proband F1377 revealed two variants in *TYR* that were predicted as deleterious but had a high allele frequency in the European demographic of gnomAD (Table 4.3). These two variants were ACMG classified as benign and VUS. Another intronic variant was detected in *ATF6* had contradictory data with varSEAK predicting an aberrant splicing event affecting a splice acceptor site but SpliceAI does not support this. The investigation in F1377 did not reveal pathogenic variants compatible with the mode of inheritance for the respective FH gene analysed. This proband was retained for further analysis in novel FH gene discovery.

Variant	Gene	Genotype	Transcript	gnomAD	CADD	SIFT	PolyPhen-2	SpliceAI	varSEAK	ACMG
c.575C>A p.(Ser192Tyr)	<i>TYR</i>	Compound heterozygous	NM_000372.5	432912/1179856 (80073)	23.8	deleterious	Probably damaging	0	-	Benign
c.1205G>A p.(Arg402Gln)	<i>TYR</i>			346238/1177910 (51519)	27.2	Deleterious	Probably damaging	0	-	VUS
c.1434-3T>A	<i>ATF6</i>	Homozygous	NM_007348.4	755110/1135416 (251495)	16.3	-	-	0	Class 5	VUS

**Table 4.3 Additional variants identified in proband F1377.** All variants listed had a CADD score exceeding the threshold of a score 15. The allele frequencies obtained from gnomAD are based on the European demographic and the homozygous count is presented in brackets. ACMG classification was performed on Franklin (<https://franklin.genoox.com/clinical-db/home>).

### **4.2.3 Bioinformatics analysis of 100KGP patients diagnosed with FH**

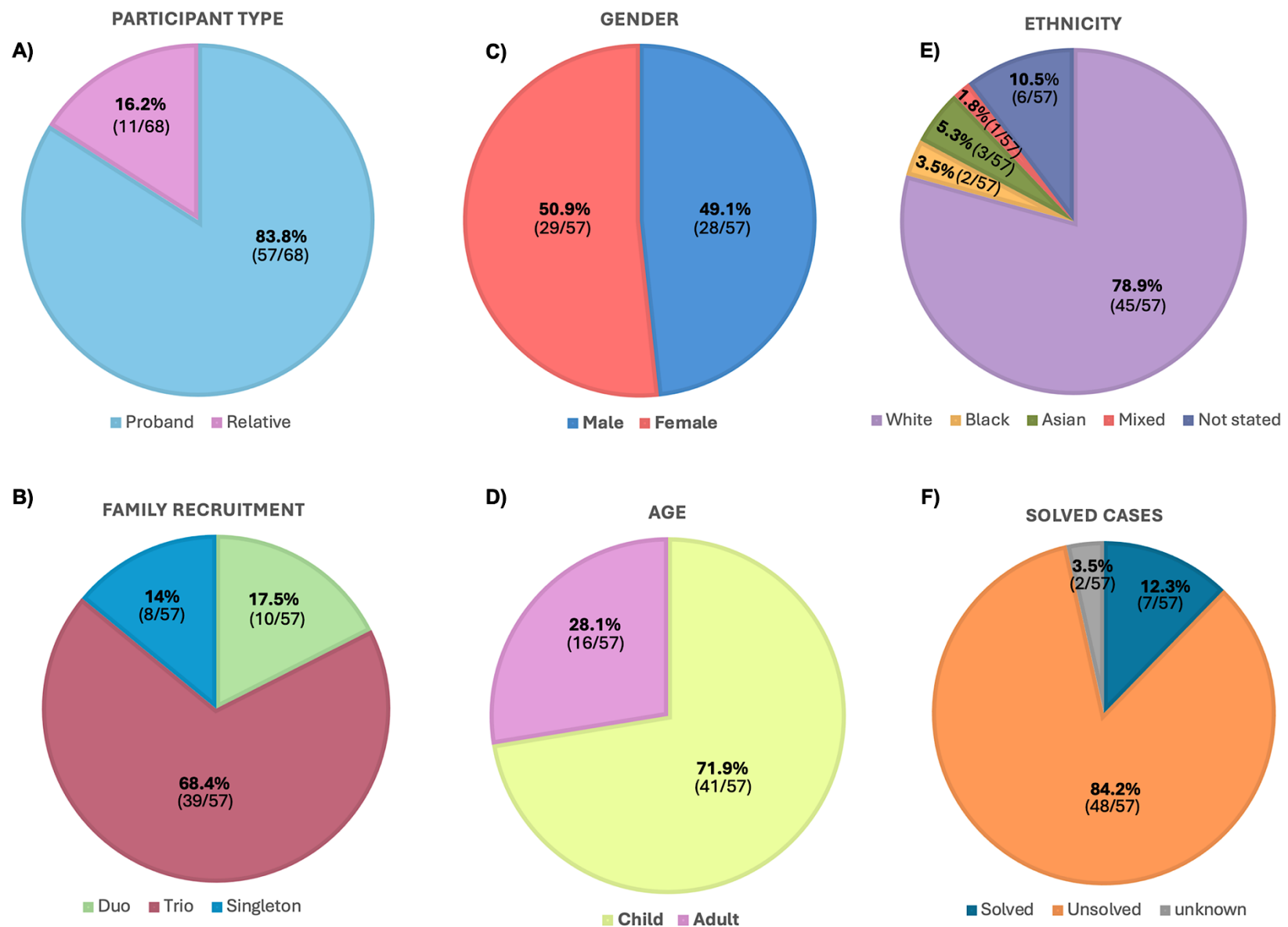
Previous 100KGP analysis was targeted towards a particular gene of interest, *SLC38A8*, to capture novel variants causative for FH (Section 3.2.2). Using the genotype to infer the phenotype in genes of high penetrance is referred to as reverse phenotyping. In contrast, the analytical strategy described in this section focusses primarily on the analysis of genotypes in a defined cohort of individuals who share a common FH phenotype.

#### **4.2.3.1 The FH cohort in the 100KGP**

The 100KGP recruited 58 probands and 11 relatives with a confirmed FH manifestation identified on Participant Explorer v5.6.0 using the HPO terms hypoplasia of the fovea, foveal hypoplasia (HPO: 0007750) and aplasia of the fovea (HPO:0008060) (Figure 4.14A). A single duplicated entry was discarded, which brings the total number of probands to 57. Of these, 39 probands were recruited alongside both parents (Trio), 10 were recruited with a single parent (Duo) (Figure 4.14B) and the rest includes eight probands who enrolled in the 100KGP without their parents (singleton).

In terms of the social demographics, this cohort displays an almost equal split for both genders, consisting of 29 females and 28 males (Figure 4.14C). With regard to age, 41 probands are children and only 16 were recruited as adults (Figure 4.14D), reflecting the fact that FH is generally a congenital condition. The ethnic breakdown of this cohort shows Europeans as the predominant ethnic group, with 45 individuals, followed by three South Asians and two individuals of African descent (Figure 4.14E). One individual was of mixed ancestry and six were classed as being of other ethnicities.

A genomic diagnosis was already obtained through the 100KGP in seven FH cases, which were considered solved before the analysis described here (Figure 4.14F). Two probands were said to be of unknown status and 48 probands were classed as unsolved.



**Figure 4.14. Characteristics of the 100KGP FH cohort.** A) A total of 68 FH participants were recruited, these being 57 probands and 11 affected relatives. B) “Trio” relates to a proband for whom both parents were recruited. “Duo” represents a proband and a single parent. “Singleton” refers to a proband enlisted without the parents. C) The biological sex of the probands determined at birth. D) Child is defined by <18 years and adult as ≥18 years. E) Ethnic demographics of the probands, with white assigned to European descent and black for African descent. F) Proportion of probands with a genomic diagnosis.

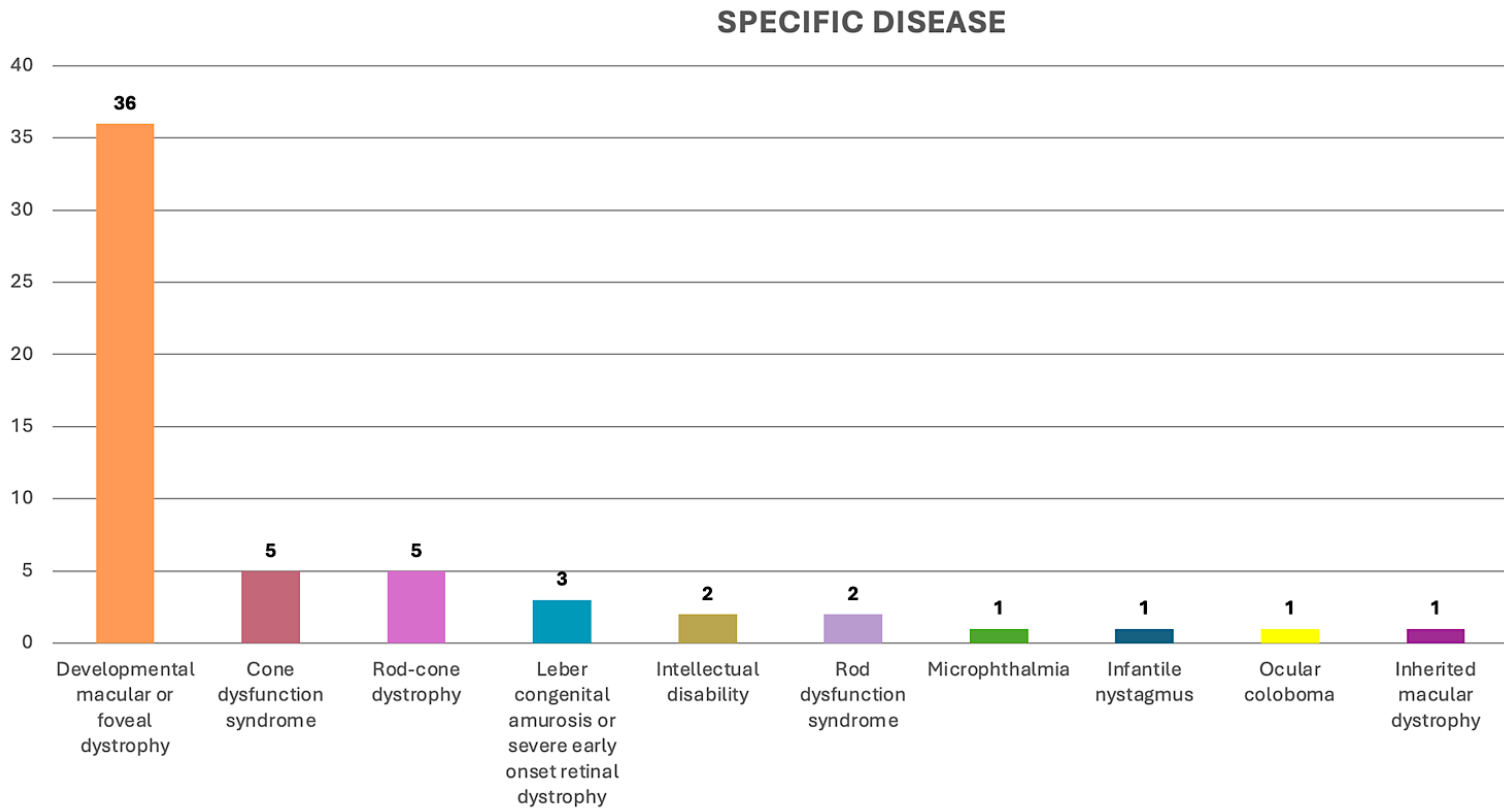
The genomic diagnosis in the seven already solved cases (P1-7) involves four pathogenic, three likely pathogenic variants and two VUS in *TYR*, *OCA2*, *USH2A*, *CNGB3*, *SRD5A3* and *GJA1* (Table 4.4). Pathogenic variants in *SRD5A3* and *GJA1* are not known to cause FH and these genes are not characterised as part of the 872 IRD genes (<https://retnet.org/symbols>). Both genes are not usually considered IRD genes, but the *SRD5A3* phenotype can overlap IRD (Taylor et al., 2017), and *GJA1* is a cause of microphthalmia as part of oculodentodigital dysplasia (Kumar et al., 2020). Given the well documented issues with HPO term usage during patient recruitment to the 100KGP (Best et al., 2022a), this may explain why these cases have been assigned the HPO term FH and have been included in this cohort.

ID	Phenotype	Genotype	Gene	Variants	Exomiser	TIER	ACMG	OMIM
P1	Nonprogressive visual loss, hypoplasia of the fovea and reduced visual acuity	Homozygous	<i>TYR</i>	c.1255G>A, p.(Gly419Arg)	0.976	2	Pathogenic	Oculocutaneous albinism
P2	Visual impairment, hypoplasia of the fovea and reduced visual acuity	Homozygous	<i>OCA2</i>	c.1255G>A, p.(Val419Ile)	0.989	3	Likely Pathogenic	Oculocutaneous albinism
P3	Visual impairment and hypoplasia of the fovea	Compound heterozygous	<i>OCA2</i>	c.1863dupA, p.(Ile622Hisfs*4)	0.988	1	Pathogenic	Oculocutaneous albinism
				c.1255G>A, p.(Val419Ile)		2	Likely Pathogenic	
P4	Progressive visual loss and hypoplasia of the fovea	Compound heterozygous	<i>USH2A</i>	c.4714C>T, p.(Leu1572Phe)	0.965	2	VUS	Usher syndrome Retinitis pigmentosa
				c.1679delC, p.(Pro560Leufs*31)		1	Pathogenic	
P5	Abnormal light adapted electroretinogram and hypoplasia of the fovea	Compound heterozygous	<i>CNGB3</i>	c.1148delC, p.(Thr383Ilefs*13)	0.974	1	Pathogenic	Achromatopsia
				c.889T>C, p.(Ser297Pro)		2	VUS	
P6	Cone-rod dystrophy and hypoplasia of the fovea	Homozygous	<i>SRD5A3</i>	c.57G>A, p.(Trp19*)	0.865	3	Pathogenic	Congenital disorder of glycosylation Kahrizi syndrome
P7	Hypoplasia of the fovea	Homozygous	<i>GJA1</i>	c.638T>A, p.(Met213Lys)	0.951	3	Likely Pathogenic	Craniometaphyseal dysplasia Oculodentodigital dysplasia Syndactyly

**Table 4.4. Solved FH cases in the 100KGP.** Seven probands with a genomic diagnosis in four known IRD genes and two genes not considered to be IRD genes. The Exomiser scores for these variants exceed 0.8. Kahrizi syndrome comprises intellectual disability, coloboma, cataracts, spine curvature (kyphosis) and coarse facial features.

The specific disease assigned to the 57 probands reported with FH includes 11 different diagnoses (Figure 4.15). The most common specific disease is developmental macular or foveal dystrophy, which accounts for 63% of the cases. Other specific diseases include a range of IRD subgroups comprising cone dysfunction syndrome (8.8%), rod-cone dystrophy (8.8%), rod dysfunction syndrome (3.4%), LCA or early onset severe retinal dystrophy (5.3%) and

macular dystrophy (1.8%). Less common specific diseases include microphthalmia (1.8%), infantile nystagmus (1.8%), intellectual disability, (3.5%) and ophthalmological features like ocular coloboma (1.8%).



**Figure 4.15. Specific disease diagnoses of FH probands.** Bar chart displaying the frequency of 10 specific diseases assigned to 57 probands reported with FH. The height of each bar is relative to the proband count that is displayed at the top. Data obtained through the Participant Explorer v5.6.0.

#### 4.2.3.2 Identifying pathogenic variants in the 100KGP FH cohort

The 100KGP latest dataset, main programme v18 was filtered to retain only participants with the HPO terms for “foveal hypoplasia, hypoplasia of the fovea or aplasia of the fovea”. HPO terms were used as the filtering criteria instead of a specific disease since patients with a confirmed FH diagnosis can be assigned to variable and sometimes incorrect specific disease (Figure 4.15).

This filtering resulted in the established FH cohort of 57 probands, with 3899 variants annotated and filtered to exclude variants with a MAF in gnomAD of  $\geq 0.1\%$ , or  $\geq 2\%$  for biallelic variants using Exomiser. The probands were further refined by excluding the seven participants (P1-7) that have already received a genomic diagnosis through the 100KGP. For a comprehensive yet manageable assessment, the remaining 50 probands were assessed for variants in 52 FH-implicated genes identified from the available literature (Appendix A: section 8.1.1).

This analysis was achieved through partitioning 3157 rare variants detected in 57 FH probands into different groups based on the mode of inheritance. A sub-group of 1273 biallelic variants was filtered for autosomal recessive FH genes. A monoallelic sub-group of 1702 heterozygous variants was filtered for autosomal dominant FH genes. Another group of 168 variant entries on the X chromosome were filtered for variants in X-linked FH genes. The remaining 14 variants in mitochondrial genes were not analysed since none of the known FH genes listed are in the mitochondrial genome. This analysis resulted in 14 probands (P8-21) with potentially deleterious genotypes in known FH-associated genes. These 14 probands carried between them a total of 21 variants that were predicted to alter the MANE mRNA transcript.

All 21 variants were subjected to *In-silico* pathogenicity assessment combined with allele frequency scrutiny (Table 4.5). Four participants (P18-21) harboured at least one variant that was predicted to be benign in nature and these were therefore eliminated from the analysis. The analysis led to the detection of 11 pathogenic and likely pathogenic variants and 3 VUSs consistent with the relevant mode of inheritance for the FH gene in 10 probands (P8-17). These 14

variants all display an Exomiser score  $\geq 0.85$  except for the variant in proband P16. The Tiering data was not always available for each variant as a supporting evidence of variant segregation with FH. The computational pathogenicity prediction of these 14 variants points to them being deleterious. The allele frequency for the 14 variants is  $\leq 0.00591$  (0.59%) in gnomAD and  $\leq 0.00648$  (0.65%) in the 100KGP. There is a high homozygote count for variants in *TYR*, *OCA2* and *CNGB3* despite them being deleterious. All 14 variants in these 10 probands were further validated and confirmed by investigating the BAM file in IGV for genuine variant calls (Figure 4.16).

This analysis led to a potential genomic diagnosis of autosomal recessive oculocutaneous albinism in three patients (P19, P10 and P11), autosomal recessive achromatopsia in two patients (P8 and P12) and the remaining diagnosis are attributed to autosomal dominant aniridia, optic nerve hypoplasia or isolated FH (P13 and P15), X-linked Aland Island disease (P14), autosomal dominant Stickler syndrome (P17) and a case of autosomal dominant familial exudative vitreoretinopathy (P16). A clinical collaboration request has been sent to the responsible clinician in each case to inform them on the genomic diagnosis for confirmation.

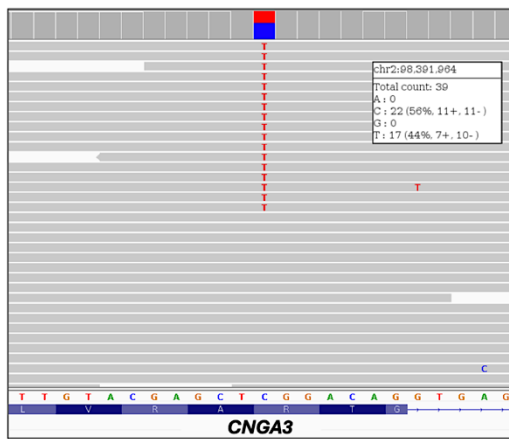


ID	Phenotype	Genotype	Gene	Variants	gnomAD	100KGP	Exomiser	Tier	CADD	SIFT	PolyPhen-2	SpliceAI	varSEAK	ACMG
P8	Cone-rod dystrophy, visual impairment and hypoplasia of the fovea	Compound heterozygous	CNGA3	c.667C>T p.(Arg223Trp)	212/1613534	17/126698	0.982	2	25.4	Deleterious	Probably damaging	-	-	Pathogenic
			CNGA3	c.1279C>T p.(Arg427Cys)	827/1613950 (3)	68/126698		2	27.5	Deleterious	Probably damaging	-	-	Likely pathogenic
P9	Macular dystrophy, reduced visual acuity and hypoplasia of the fovea	Compound heterozygous	TYR	c.1467dupT p.(Ala490Cysfs*20)	372/1613874	44/126698	0.933	-	23.8	-	-	-	-	Pathogenic
			TYR	c.1217C>T p.(Pro406Leu)	7323/1611690 (23)	605/126698		-	25.6	Deleterious	Probably damaging	-	-	Likely pathogenic
P10	Nonprogressive visual loss, visual impairment, reduced visual acuity and hypoplasia of the fovea	Compound heterozygous	OCA2	c.1025A>G p.(Tyr342Cys)	619/1613560	44/126698	0.971	-	25.4	Deleterious	Probably damaging	-	-	Likely pathogenic
			OCA2	c.1255G>A p.(Val443Ile)	9545/1613948 (39)	821/126698		-	25.8	Deleterious	Probably damaging	-	-	Likely pathogenic
P11	Visual impairment, reduced visual acuity and hypoplasia of the fovea	Compound heterozygous	TYR	c.1118C>A p.(Thr373Lys)	1423/1613434 (1)	143/126698	0.986	-	23	Deleterious	Benign	-	-	Pathogenic
			TYR	c.1217C>T p.(Pro406Leu)	7323/1611690 (23)	605/126698		-	25.6	Deleterious	Probably damaging	-	-	Likely pathogenic
P12	Progressive visual loss, retinal dystrophy, visual impairment and hypoplasia of the fovea.	Homozygous	CNGB3	c.1148delC p.(Thr383Ilefs*13)	3651/1606146 (7)	359/126698	0.987	-	33	-	-	-	-	Pathogenic
P13	Nonprogressive visual loss, visual impairment, reduced visual acuity and hypoplasia of the fovea	Heterozygous	PAX6	c.399+1G>A	0	1/126698	0.98	1	34	-	-	DL:0.99 DG:0.52	Class 5	Pathogenic
P14	Nonprogressive visual loss, visual impairment, reduced visual acuity and hypoplasia of the fovea	X-linked	CACNA1F	c.875G>A p.(Cys292Tyr)	0	3/126698	0.98	2	23.3	Deleterious	Probably damaging	-	-	VUS
P15	Nonprogressive visual loss, visual impairment, reduced visual acuity and hypoplasia of the fovea	Heterozygous	PAX6	c.817T>C p.(Ser273Pro)	0	3/126698	0.986	3	29.3	Deleterious	Probably damaging	-	-	Likely pathogenic
P16	Abnormality of retinal pigmentation, visual impairment, retinal dystrophy and hypoplasia of the fovea	Compound Heterozygous	ZNF408	c.1599C>A p.(Phe533Leu)	12/1612078	1/126698	0.477	-	23.8	Deleterious	Possibly damaging	-	-	VUS

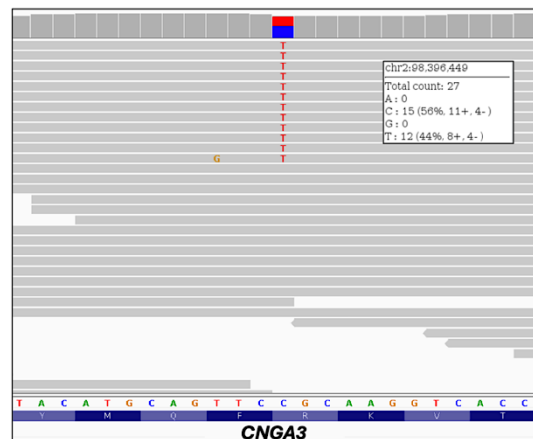
P17	Visual impairment, macular dystrophy and hypoplasia of the fovea	Heterozygous	<i>COL11A1</i>	c.206G>T p.(Gly69Val)	3/1613430	1/126698	0.850	3	18.7	Tolerated	Possibly damaging	-	-	VUS
P18	Visual impairment and hypoplasia of the fovea	Compound Heterozygous	<i>OCA2</i>	c.2323G>C p.(Gly775Arg)	6/1613792	3/126698	0.9151	3	25.6	Deleterious	Probably damaging	-	-	Pathogenic
			<i>OCA2</i>	c.365C>T p.(Thr122Ile)	714/1614254 (2)	32/126698		3	7.39	Tolerated	Benign	-	-	Likely benign
		X-linked	<i>IKBKG</i>	c.7A>C p.(Arg3=)	687/1082804 (4)	54/126698	0.0847	-	8.612	-	-	0	Class 1	Benign
P19	Nonprogressive visual loss, visual impairment, reduced visual acuity and hypoplasia of the fovea	Compound heterozygous	<i>SLC38A8</i>	c.273G>A p.(Val91=)	21267/1614068 (175)	1702/126698	0.0064	-	6.812	-	-	0	Class 1	Benign
			<i>SLC38A8</i>	c.1074G>A p.(Thr358=)	5803/1613952 (12)	427/126698		-	0.019	-	-	AG:0.01	Class 1	Benign
P20	Abnormal fundus morphology and hypoplasia of the fovea	X-linked	<i>GPR143</i>	c.17C>A p.(Arg6Leu)	0	44/126698	0.388	-	9.158	Tolerated	Benign	-	-	VUS
P21	Hypoplasia of the fovea	Heterozygous	<i>BEST1</i>	c.880A>G p.(Arg294Gly)	0	1/126698	0.926	-	1.881	Tolerated	Benign	0	Class 1	VUS

**Table 4.5. *In-silico* pathogenicity assessment of the 100KGP FH cohort.** 10/14 probands with an FH compatible phenotype have been identified with deleterious genotypes in *CNGA3*, *CNGB3*, *TYR*, *OCA2*, *PAX6*, *COL11A1*, *ZNF408* and *CACNA1F*. Amongst these, are three novel mutations responsible for FH, NM\_001368894.2:c.399+1G>A and NM\_001368894.2:c.817T>C, NP\_001355823.1:p.(Ser273Pro) in *PAX6* and NM\_001256789.3:c.875G>A in *CACNA1F*, found in probands P13, P15 and P14, respectively. Probands P14, P18 and P20, who carry variants in X-linked FH-associated genes, are males. The allele frequencies obtained from gnomAD and the 100KGP population data are not refined by ethnicity thus the overall frequency is displayed. In probands where the ethnicity is not disclosed (P11 and P12), the total allele count is used in gnomAD. Splicing variants are interpreted using CADD score, SpliceAI and varSEAK.

P8

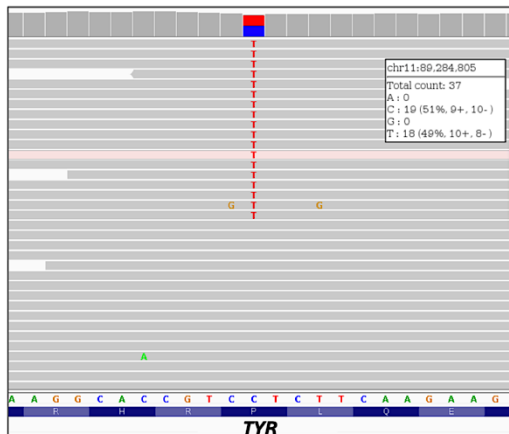


NM\_001298.3:c.667C>T  
NP\_001289.1:p.(Arg223Trp)

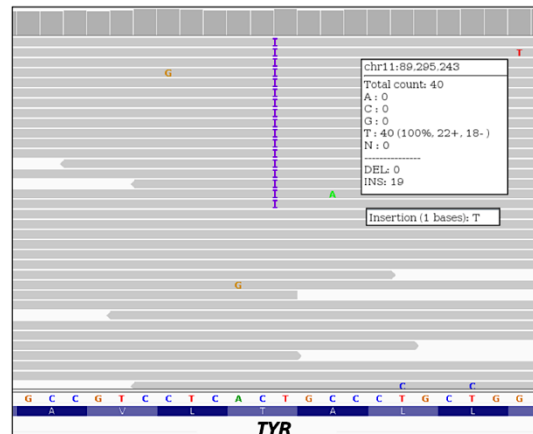


NM\_001298.3:c.1279C>T  
NP\_001289.1:p.(Arg427Cys)

P9

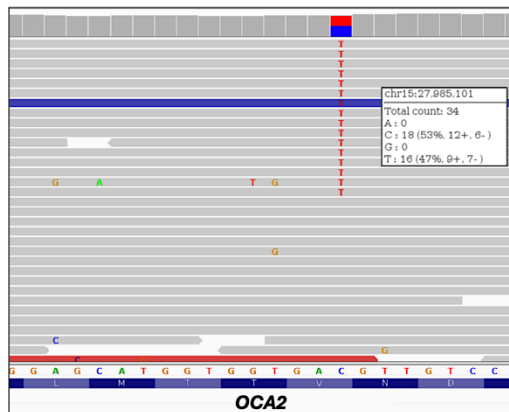


NM\_000372.5:c.1217C>T  
NP\_000363.1:p.(Pro406Leu)

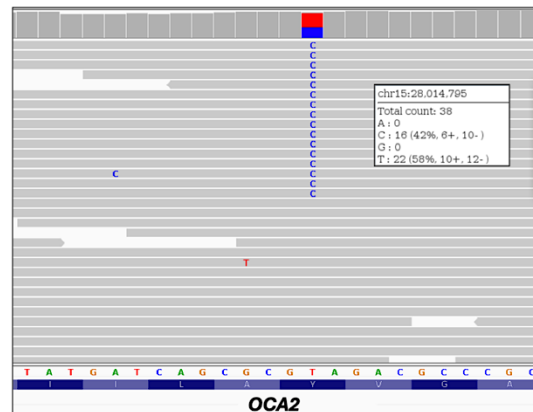


NM\_000372.5:c.1467dupT  
NP\_000363.1:p.(Ala490Cysfs\*20)

P10

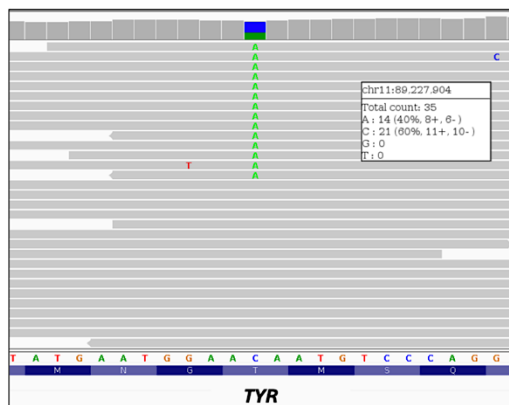


NM\_000275.3:c.1255G>A  
NP\_000266.2:p.(Val443Ile)

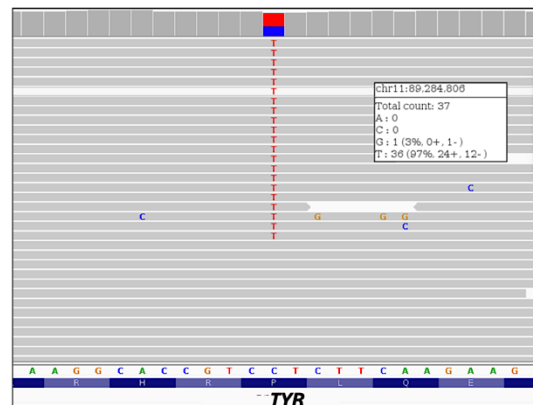


NM\_000275.3:c.1025A>G  
NP\_000266.2:p.(Tyr342Cys)

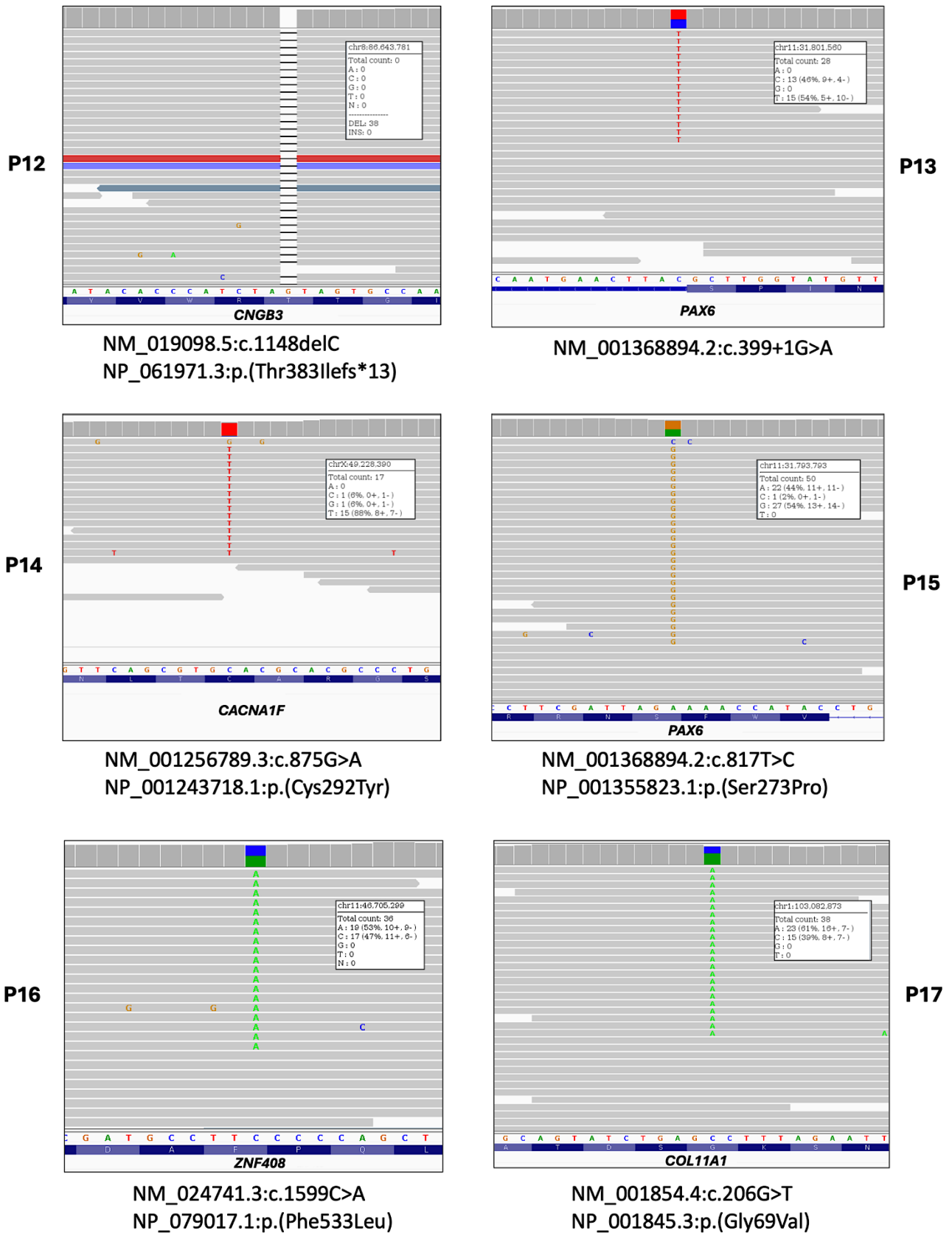
P11



NM\_000372.5:c.1118C>A  
NP\_000363.1:p.(Thr373Lys)



NM\_000372.5:c.1217C>T  
NP\_000363.1:p.(Pro406Leu)

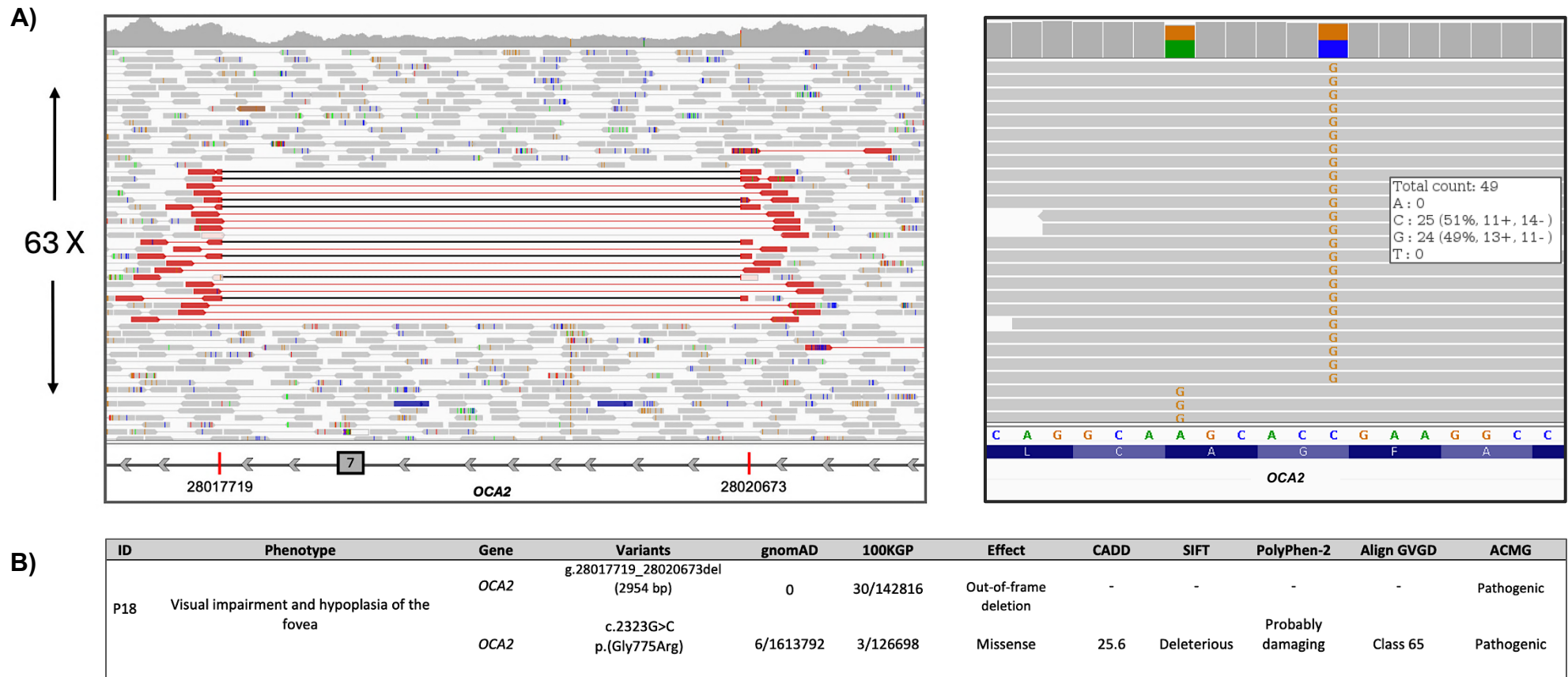


**Figure 4.16. Confirmation of detected variants in 10 probands with FH.** BAM alignment visualised on IGV v2.15.4 for ten potentially solved FH cases identified in the 100KGP. A small window displays the genomic coordinates for the variant according to GRCh38. The allele count at the variant site is also displayed.

#### 4.2.3.3 Identifying SVs in the FH cohort

The output files containing annotated SV calls by SVRare were examined for a gene specific analysis in the remaining 40 probands with unsolved FH. In total 3294 SVs in 52 FH-associated genes were compiled for 71408 participants with rare disease. 1450 SVs were retained based on allele count  $\leq 60$  in the 100KGP and the CDS being disrupted. Only SVs detected in the 40 probands with unsolved FH were analysed and using the BAM file for inspection to eliminate artefactual calls and correct the genomic coordinates that are often ambiguously defined by the SV caller. Confirmed SVs were assessed either alone or in combination with other variants in the CDS, depending on the relevant mode of inheritance for the gene involved.

This analysis uncovered a deletion in *OCA2* (NC\_000015.10: g.28017719\_28020673del) as the second pathogenic variant in proband (P18) who was also previously found to carry a single pathogenic missense variant NM\_000275.3:c.2323G>C, NP\_000266.2:p.(Gly775Arg) in *OCA2* (Table 4.5). This CNV deletes 2955 bp sequence containing exon 7 of *OCA2* and is predicted to disrupt the ORF (Figure 4.17A). On examination, this SV was reported to also be present in 29 additional participants in the rare disease cohort of the 100KGP (Figure 4.17B). 28 of these did not harbour a second pathogenic allele in *OCA2*. No pathogenic SVs or a combination of pathogenic SVs and SNVs were detected in FH genes in the remaining 39 participants. These 39 probands with unsolved FH were therefore retained for SNV discovery in the noncoding segment of their genome. Communicating the genomic diagnosis in P18 to the clinicians was achieved through the clinical collaboration request.

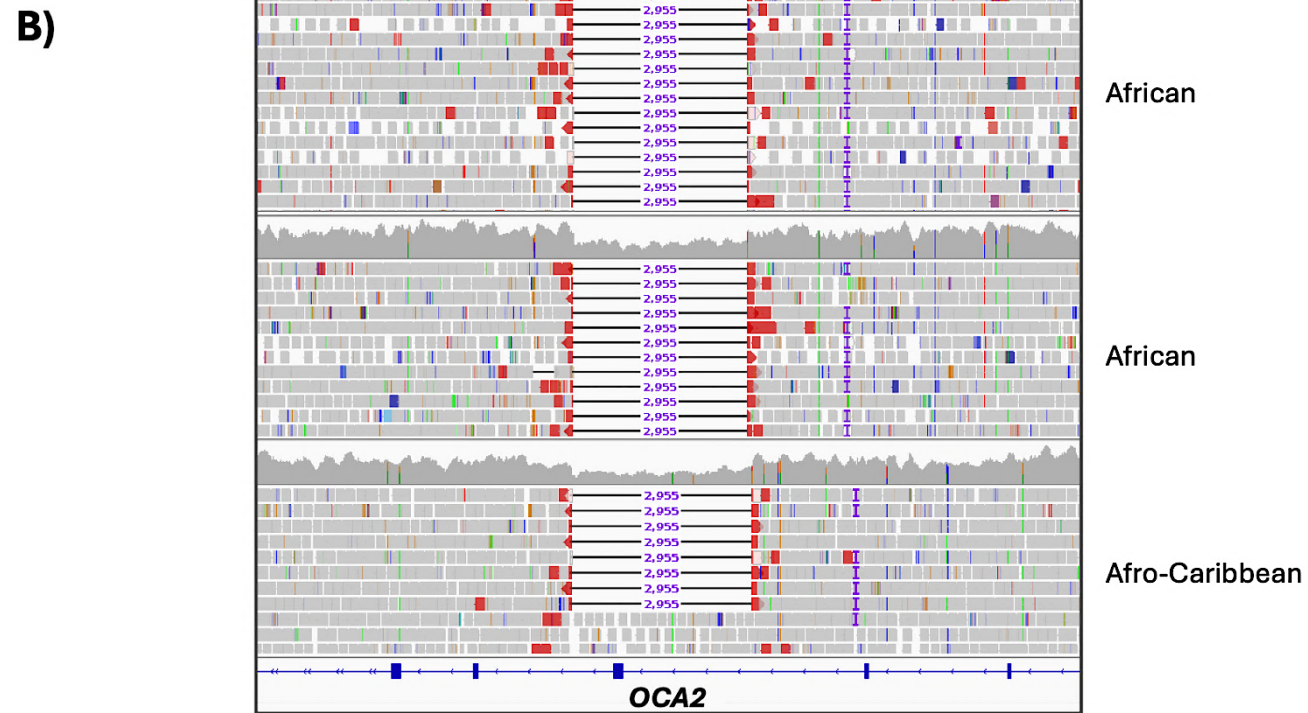
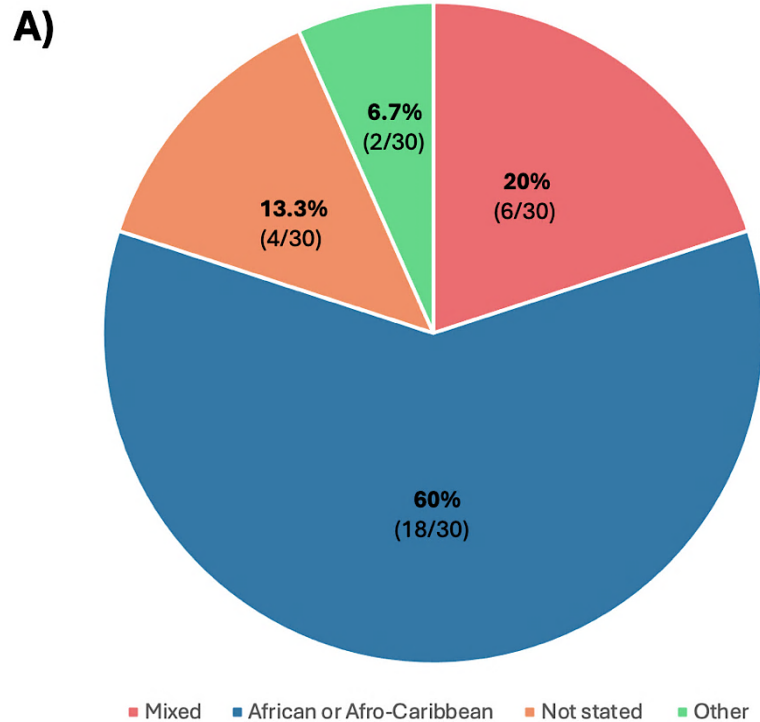


**Figure 4.17. *OCA2* genotype in proband P18 with unsolved FH.** A) heterozygous deletion of 2954 bp in *OCA2* was identified in proband P18. The deduced breakpoints are chr15:28017719-28020673. Reduction in read coverage at the deleted region is apparent in the Sashimi plot, with almost half the expected read depth across the deleted region. Gene schematic is shown at the bottom track with arrows depicting the sequence orientation and the breakpoints being marked by a red line. The window to the left displays the variant call at chr15:27851397. The variant was detected in 49% of reads (24/49). B) Table presents the *OCA2* genotype in the proband along with the variants interpretation. Both gnomAD and the 100KGP column display the overall allele frequency.

#### **4.2.3.4 The discovery of a potential founder SV in albinism**

The NC\_000015.10:g.28017719\_28020673del identified in OCA2 was found to be present in a heterozygous state in 30/71408 individuals in the 100KGP. The ethnic breakdown of these participants include African or Afro-Caribbean (60%), mixed African or mixed Afro-Caribbean (20%), other ethnicities (6.7%) and undisclosed ethnicity (13.3%) (Figure 4.18A). The data regarding the 30 participants ancestry, age, gender and hospital of recruitment was accessed through Labkey v23.7. The overwhelming African/Afro-Caribbean demographic (80%) strongly suggest a potential founder effect that is present amongst African descendants. The NC\_000015.10:g.28017719\_28020673del carrier status was confirmed in all 30 participants through the manual inspection of the locus in IGV using the BAM file which revealed the deletion as heterozygous.

Haplotype analysis was performed to test the hypothesis that a founder effect resulted in the enrichment of NC\_000015.10:g.28017719\_28020673del in African descendants. Three participants of diverse yet related backgrounds of African and Afro-Caribbean were selected as representatives (Figure 4.18B). Using both the proband's and their parent's genomes, the genotypes were extracted in a 45208 bp region (chr15:27979889-28025097) centred on the OCA2 exon 7 deletion (Appendix C: Supplementary Table 4.4). Variants included in the region have quality score  $\geq 79$  for confidence and have been confirmed in the participants' BAM files. The haplotype was constructed with the alleles phased based on the parental origin in the three probands (Table 4.6). The data shows all three participants have genotypes consistent with a shared haplotype in a 44807 bp region from chr15:28024696-27979889 which encompasses the NC\_000015.10:g.28017719\_28020673del. A single discrepant allele is detected in an African proband and is present downstream the homology block at chr15:28024900 .



**Figure 4.18. Characteristics of participants with NC\_000015.10:g.28017719\_28020673del in OCA2.** A) Ethnic demographic of NC\_000015.10:g.28017719\_28020673del carriers. Other ethnicities comprise a South Asian and a European. Mixed ethnicity denotes a South Asian mixed with African ethnicity or a European mixed with either African or Afro-Caribbean. B) BAM file for three participants with reads aligned at chr15:28008134-28030259. The NC\_000015.10:g.28017719\_28020673del is present in all three individuals. The coverage obtained for each of undefined African, East African and Afro-Caribbean participants is 65 X, 109 X and 68 X respectively.



GRCh38	Reference	African	African	Afro-Caribbean	African AF	European AF	dbSNP
15:27979889	Z	W	W	W	0.9719	0.9607	rs11074315
15:27980004	Z	Y	Y	Y	0.04376	0.00001476	rs59967422
15:27992851	Y	Z	Z	Z	0.9982	0.9887	rs4499192
15:27994398	X	Y	Y	Y	0.1626	0.00319	rs57376758
15:27999839	Z	Y	Y	Y	0.128	0.0003386	rs12101660
15:27999998	X	Y	Y	Y	0.9999	0.9979	rs4640132
15:28014907	Z	Y	Y	Y	0.1373	0.04837	rs1800401
15:28017719-28020673	-	Del	Del	Del	0.002543	0	-
15:28020856	Z	Z	Z	Y/Z	0.08265	0.0004705	rs77490758
15:28021089	W	W	W	X/W	0.3672	0.4944	rs746861
15:28021139	Y	Y	Y	Z/Y	0.08184	0.000147	rs74531804
15:28021881	Z	Y	Y	Y	0.3221	0.1683	rs2122005
15:28022387	Z	ZWZZWZYZY	ZWZZWZYZY	ZWZZWZYZY	0.8557	0.2215	rs11283428
15:28023490	Z	Z	X/Z	Z	0.3323	0.003131	rs7174197
15:28023862	Y	X	X	X	0.8548	0.2197	rs3794606
15:28024696	X	X	X/W	X	0.3327	0.00274	rs28546555
15:28024900	X	Y	X	X	0.00001332	0	-
15:28025097	Z	Y	Y	Y	0.187	0.0004116	rs73375883

**Table 4.6. Haplotypes of three African descendants heterozygous for NC\_000015.10:g.28017719\_28020673del.** The alleles are encrypted to maintain anonymity of the participants. Alleles were documented spanning a chr15:27979889-28025097 locus. The genomic coordinates are listed in ascending order. Alleles in orange signify the maternal allele and blue for the paternal allele. Light blue shading denotes a shared allele in all three probands. Unphased alleles whereby the parental alleles cannot be inferred, are depicted in black and the “/” separates the allele pair. The red box marks the deleted region of chr15:28017719-28020673. Allele frequency (AF) is displayed for the alternate alleles in the African/African American and European demographic in gnomAD.

#### 4.2.3.5 Detecting SVs in additional participants with FH

Many 100KGP participants with IRDs are reported with vague phenotypes or with partial information on the clinical manifestations. A proportion of these participants are likely to have FH but are not reported with the HPO terms of foveal hypoplasia, hypoplasia of the fovea or aplasia of the fovea during recruitment to the 100KGP. To compensate for the uninformative phenotypic reporting, a gene specific analysis was performed on all 100KGP participants, irrespective of any phenotypic criteria. This reverse phenotyping or reverse genetics approach was conducted to analyse all the SVs in 27 FH-associated genes using SVRare. The select 27 genes are responsible for isolated FH, aniridia, albinism, incontinentia pigmenti, Aland island disease and Stickler syndrome (Appendix A: section 8.1.2).

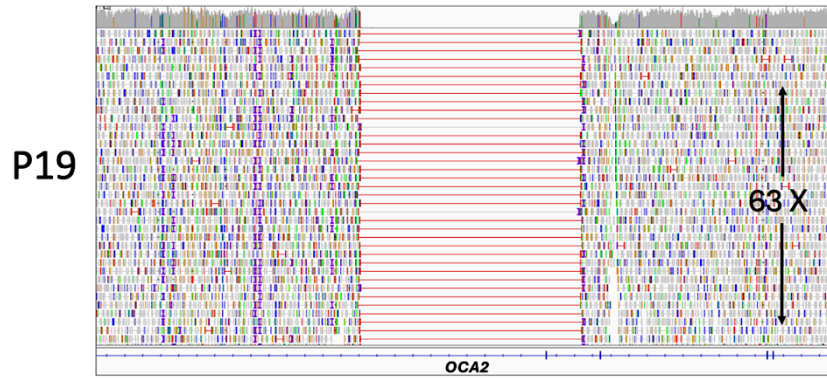
In total, 2207 unfiltered SVs were compiled by SVRare in these 27 genes for 71408 participants with rare disease and these were subsequently filtered to exclude those with an allele count  $\geq 60$  in the 100KGP and to retain only SVs that disrupted the CDS. This filtering strategy has yielded 969 potential SVs. Each SV was inspected using the BAM file for validation and the probands were screened for additional variants in the same gene depending on the mode of inheritance of FH associated with the gene. This led to the detection of nine deleterious SVs that disrupt the CDS in nine participants who have not yet been given a molecular diagnosis (Table 4.7). The SVs identified include deletions in *PAX6*, *OCA2*, *GPR143* and *CACNA1F* with the coverage ranging from 31-78 X (Figure 4.19), meaning all nine SVs were detected at a sufficient read depth to provide confidence in the SVs called. The reported HPO terms in these patients are compatible with FH phenotypes which further provides reassurance in the genomic finding. The genomic diagnoses were reported to the recruiting clinicians, and these constitute ocular albinism (P21-P24), oculocutaneous albinism (P19-P20), aniridia (P25 and P27) and Aland Island disease (P26).

Amongst the SVs listed, the deletion in *OCA2* (NC\_000015.10: g.28017719\_28020673del) was reported with the highest allele frequency of 0.01416765053 in gnomAD for the African ethnicity and also has the highest

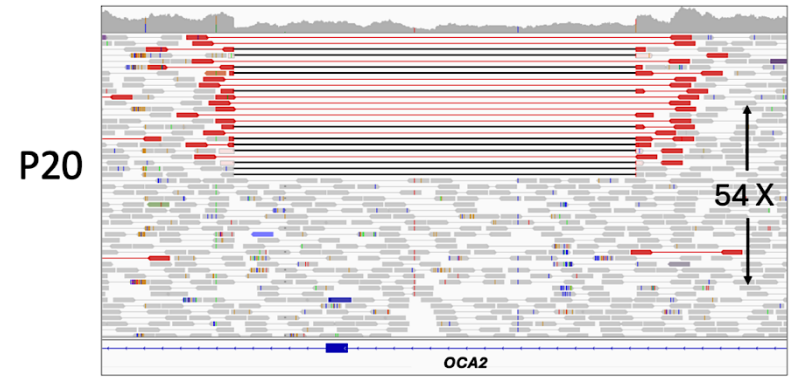
allele count of 30 in the 100KGP. This *OCA2* deletion was encountered previously in another proband (P18) (Figure 4.17). A recurrent deletion in *GPR143* (NC\_000023.11:g.9476905\_9954651del) was also identified in two participants of European descent whom are males, and this has an allele count of 14 in the 100KGP but was not reported in gnomAD. Upon inspection of the BAM files for the remaining 12 participants with this deletion, it transpired that these were false positive SV calls.

ID	Phenotype	Gene	Genotype	Class	SV	Effect	100KGP	gnomAD SV	Gender	ACMG
P19	Retinal dystrophy, central scotoma, visual impairment, nonprogressive visual loss and reduced visual acuity	<i>OCA2</i>	Homozygous	Deletion	<i>g.27823041_27849013del</i> (25972 bp)	Out-of-frame deletion of exon 23	2/71408	0	F	Likely pathogenic
P20	Nonprogressive visual loss, central scotoma, visual impairment, macular atrophy and reduced visual acuity	<i>OCA2</i>	Compound heterozygous	Deletion	<i>g.28017719_28020673del</i> (2954 bp)	Out-of-frame deletion of exon 7	30/71408	0	F	Pathogenic
				Missense	<i>c.2228C&gt;T, p.(Pro743Leu)</i>	-	-	287/1613590		Pathogenic
P21	Nonprogressive visual loss, retinal dystrophy, visual impairment, central scotoma and macular atrophy	<i>GPR143</i>	Hemizygous	Deletion	<i>g.9476905_9954651del</i> (477746 bp)	Gene deletion	14/71408	0	M	Pathogenic
P22	Nonprogressive visual loss, retinal dystrophy, reduced visual acuity, central scotoma and visual impairment	<i>GPR143</i>	Hemizygous	Deletion	<i>g.9476905_9954651del</i> (477746 bp)	Gene deletion	14/71408	0	M	Pathogenic
P23	Visual impairment, retinal dystrophy, reduced visual acuity, central scotoma and nonprogressive visual loss	<i>GPR143</i>	Hemizygous	Deletion	<i>g.9743119_10024473del</i> (281354 bp)	Deletion of exon 1-6. Effect cannot be determined	1/71408	0	M	Likely pathogenic
P24	Visual impairment, retinal dystrophy, central scotoma and nonprogressive visual loss	<i>GPR143</i>	Hemizygous	Deletion	<i>g.9760253_9761254del</i> (1001 bp)	Out-of-frame deletion of exon 2	1/71408	0	M	Likely pathogenic
P25	Aniridia	<i>PAX6</i>	Heterozygous	Deletion	<i>g.31393557_31812922del</i> (419365 bp)	Gene deletion	1/71408	0	M	Pathogenic
P26	Visual impairment, abnormal light adapted electroretinogram, nonprogressive visual loss and reduced visual acuity	<i>CACNA1F</i>	Hemizygous	Deletion	<i>g.49224054_49226365del</i> (2311 bp)	In-frame deletion of exon 12-14	2/71408	0	M	VUS
P27	Aniridia	<i>PAX6</i>	Heterozygous	Deletion	<i>g.28833790_37084393del</i> (8250603 bp)	Gene deletion	1/71408	0	F	Pathogenic

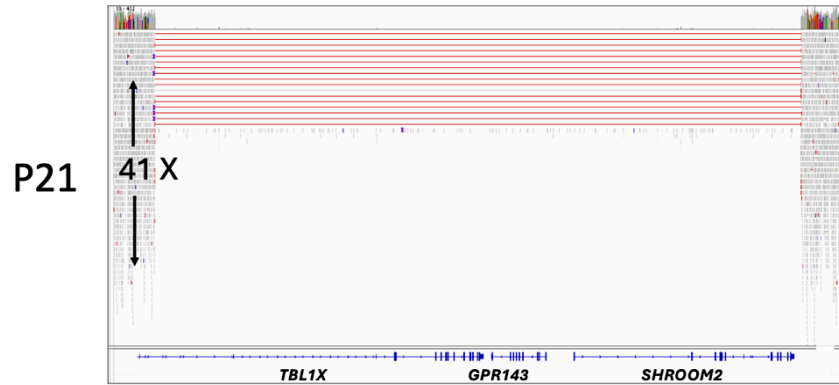
**Table 4.7. Pathogenic SVs in FH-associated genes identified in the 100KGP.** Table summarising the clinical and genetic findings of 9 probands potentially solved through the identification of deleterious SVs in FH genes. *GPR143* and *CACNA1F* variants follow an X-linked inheritance pattern while *PAX6* variants show autosomal dominant inheritance. The size of the SVs detected range from 1001-8250603 bp. M: male; F: female. The 100KGP column displays the participant count for each SV listed.



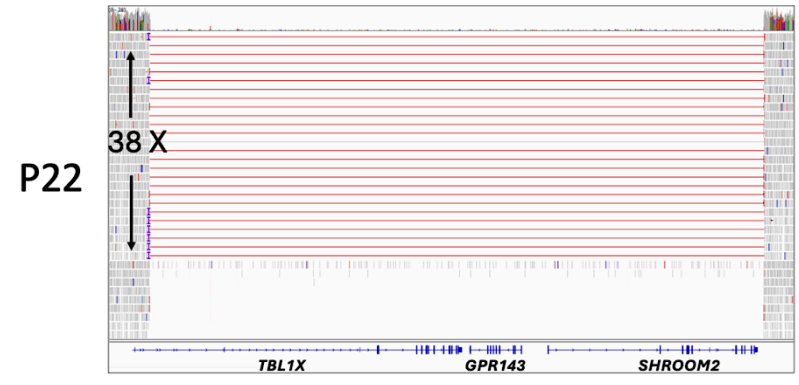
NC\_000015.10:g.27823041\_27849013del



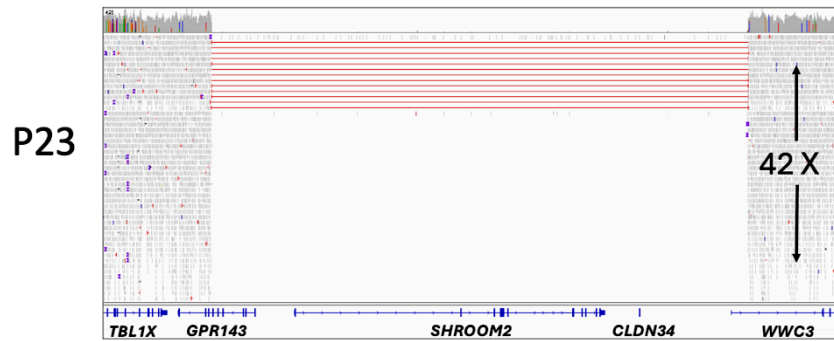
NC\_000015.10:g.28017719\_28020673del



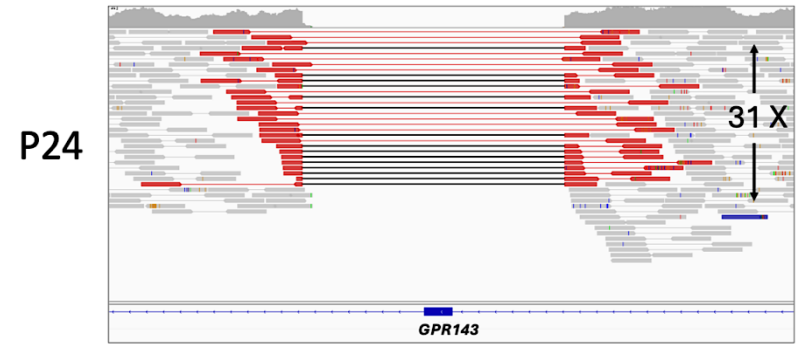
NC\_000023.11:g.9476905\_9954651del



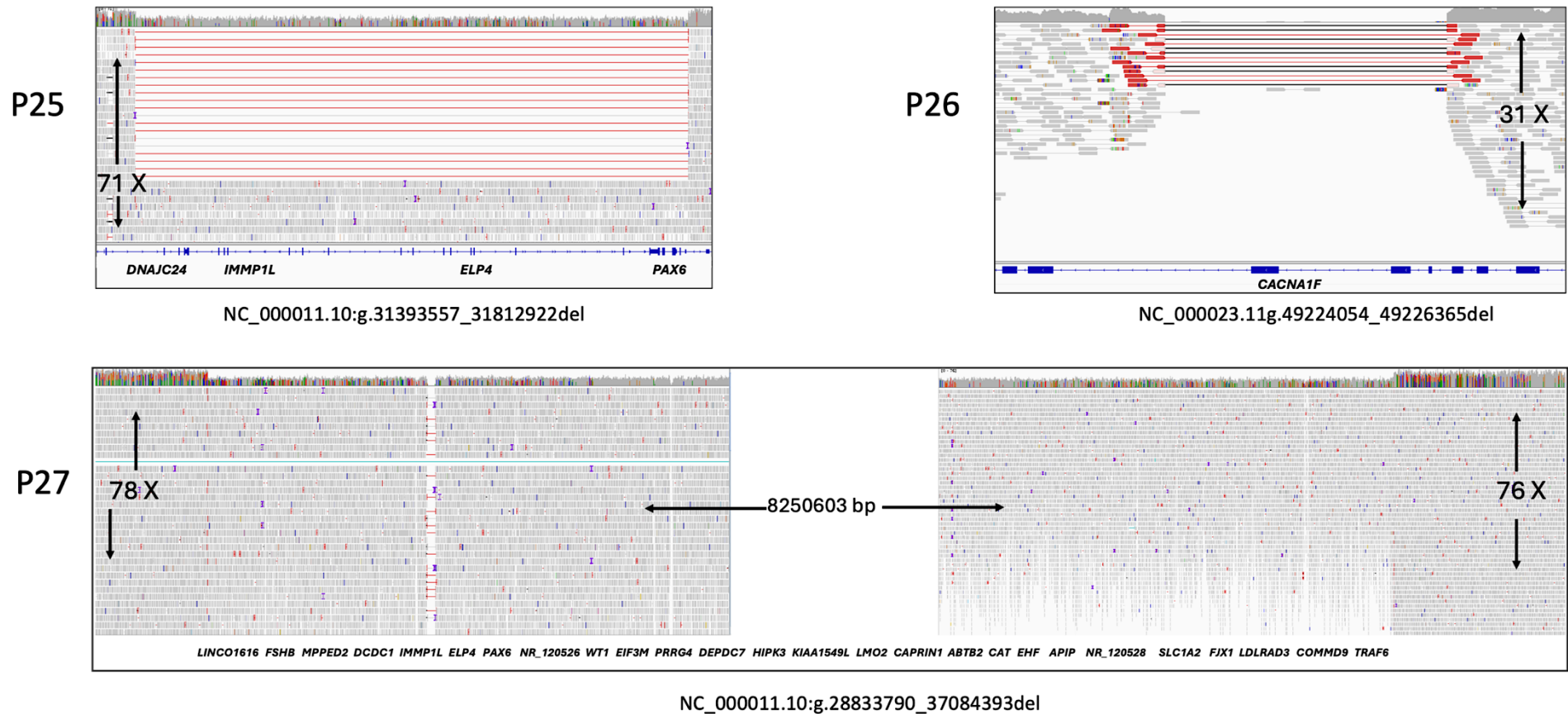
NC\_000023.11:g.9476905\_9954651del



NC\_000023.11:g.9743119\_10024473del



NC\_000023.11:g.9760253\_9761254del



**Figure 4.19. Validation of SVs identified in the 100KGP.** BAM file displaying the SVs at the target loci. Sashimi plot represents the coverage along the genomic region inspected. The total read count is provided and the gene schematic is present at the bottom. For heterozygous SVs, the read depth in the Sashimi plot drops to around half the expected reads (P20 and P25-P27), while for homozygous or hemizygous deletions it drops to zero (P19, and P21- P24). Gene schematic is present at the bottom of the IGV diagram. Red markers signify a larger alignment insert size than the expected size as an indicator for a deletion.

#### 4.2.3.6 Reviewing Gene panels associated with FH

The virtual gene panels on PanelApp are curated and reviewed by a team of approved researchers and clinical specialists as part of the GECIP consortium. An application for PanelApp reviewer status was granted under the affiliation of University of Leeds. This provided the opportunity to manipulate the Retinal Disorders gene panel in the 100KGP to provide a more effective variant interpretation in FH and ultimately increase the diagnostic yield. The eligibility criteria for the application of the Retinal Disorders gene panel in 100KGP participants is based on a suspicion of posterior segment abnormalities, cone dysfunction syndrome, developmental macular or foveal dystrophy, inherited macular dystrophy, Leber congenital amaurosis or early-onset severe retinal dystrophy, rod dysfunction syndrome, rod-cone dystrophy, familial exudative vitreoretinopathy, Sorsby retinal dystrophy or Doyme retinal dystrophy.

The Retinal Disorders virtual gene panel v2.7 contained 398 genes but it did not contain all the genes known to be involved in isolated FH. Two such genes are *AHR* and *FRMD7*, which were not considered as IRD causing genes. These two genes were classified as of moderate evidence (Amber), meaning that additional evidence from at least one further unrelated family is required, together with a consensus from other reviewers, for pathogenic variants in these genes to be classified as clinically actionable. As part of this study, both *AHR* and *FRMD7* were reviewed in the Retinal Disorders gene panel v5.11 and were supported with the required evidence from the published literature for  $\geq 3$  unrelated FH families reported with pathogenic variants in *AHR* or *FRMD7*, in order to substantiate the gene association with IRDs (Figure 4.20).

Another isolated FH gene in the retinal disorders panel v5.11 that was also reviewed by the author is *PAX6*. It was characterised in the previous list as being of low confidence in IRD (Red). This is in part due to the varying clinical manifestations associated with *PAX6* variants, which caused the gene to be incorporated into other panels such as optic nerve hypoplasia and aniridia but not the IRD panel. Additional evidence from the published literature was provided to support the reclassification of this gene as involved in IRDs, including FH.

The only isolated FH gene in the Retinal Disorders gene panel v5.11 that was categorised as a diagnostically actionable gene (Green) was *SLC38A8*. Information on this gene was reviewed and updated with the latest developments from the published literature and additional information from this thesis. The submitted reviews for all four genes, *AHR*, *FRMD7*, *PAX6* and *SLC38A8*, are currently pending review by other gene reviewers.

Despite the additions of isolated FH genes, the latest Retinal Disorders panel v5.11 that is currently used in analysing eligible participants still does not cover all the genes known to be associated with FH. This version of the gene panel contained 423 genes, with 261 being “green”, 56 “amber” and 104 are classified as “red” (Figure 4.21) (visited on 26/07/24). There are 18 missing genes implicated in FH, these comprise *DCT*, *LYST*, *HPS1*, *HPS3*, *HPS4*, *HPS5*, *HPS6*, *AP3B1*, *DTNBP1*, *BLOC1S3*, *BLOC1S5*, *PLDN*, *AP3D1*, *HESX1*, *PROKR2*, *SOX3*, *PRSS56* and *TMEM98*. Pathogenic variants in these genes are responsible for oculocutaneous albinism, CHS, HPS, optic nerve hypoplasia and nanophthalmos. Some genes that are missing are part of other gene panels that are very specific to a particular phenotype, such as optic nerve hypoplasia (Table 4.8). The albinism related gene *DCT* was added to the retinal disorders gene panel and is pending review from other experts to be assigned a classification.



Green	<b>SLC38A8</b>	6 reviews ✓ You reviewed  4 green	BIALLELIC, autosomal or pseudoautosomal	<b>Sources</b> <ul style="list-style-type: none"> <li>Expert Review Green</li> <li>Literature</li> <li>NHS GMS</li> </ul> <b>Phenotypes</b> <ul style="list-style-type: none"> <li>Foveal hypoplasia 2, with or without optic nerve misrouting and/or anterior segment dysgenesis OMIM:609218</li> <li>foveal hypoplasia - optic nerve decussation defect - anterior segment dysgenesis syndrome MONDO:0012216</li> <li>chiasmal misrouting</li> </ul> <b>Tags</b>
Amber	<b>AHR</b>	4 reviews ✓ You reviewed  1 green	BIALLELIC, autosomal or pseudoautosomal	<b>Sources</b> <ul style="list-style-type: none"> <li>Expert Review Amber</li> <li>NHS GMS</li> <li>RetNet</li> </ul> <b>Phenotypes</b> <ul style="list-style-type: none"> <li>?Retinitis pigmentosa 85, OMIM:618345</li> <li>Retinal dystrophy</li> </ul> <b>Tags</b> watchlist
Amber	<b>FRMD7</b>	2 reviews ✓ You reviewed  1 green	X-LINKED: hemizygous mutation in males, biallelic mutations in females	<b>Sources</b> <ul style="list-style-type: none"> <li>Expert Review Amber</li> <li>Literature</li> </ul> <b>Phenotypes</b> <ul style="list-style-type: none"> <li>Nystagmus 1, congenital, X-linked, OMIM:310700</li> <li>Nystagmus, infantile periodic alternating, X-linked, OMIM:310700</li> <li>foveal hypoplasia, MONDO:0044203</li> </ul> <b>Tags</b>
Red	<b>PAX6</b>	4 reviews ✓ You reviewed  2 green 2 red	MONOALLELIC, autosomal or pseudoautosomal, imprinted status unknown	<b>Sources</b> <ul style="list-style-type: none"> <li>Eligibility statement prior genetic testing</li> <li>Expert Review Red</li> <li>NHS GMS</li> </ul> <b>Phenotypes</b> <ul style="list-style-type: none"> <li>Foveal Hypoplasia and Presenile Cataract Syndrome</li> <li>Developmental macular and foveal dystrophy (foveal hypoplasia in the context of aniridia)</li> </ul> <b>Tags</b>

### Mohammed Derar (University of Leeds)

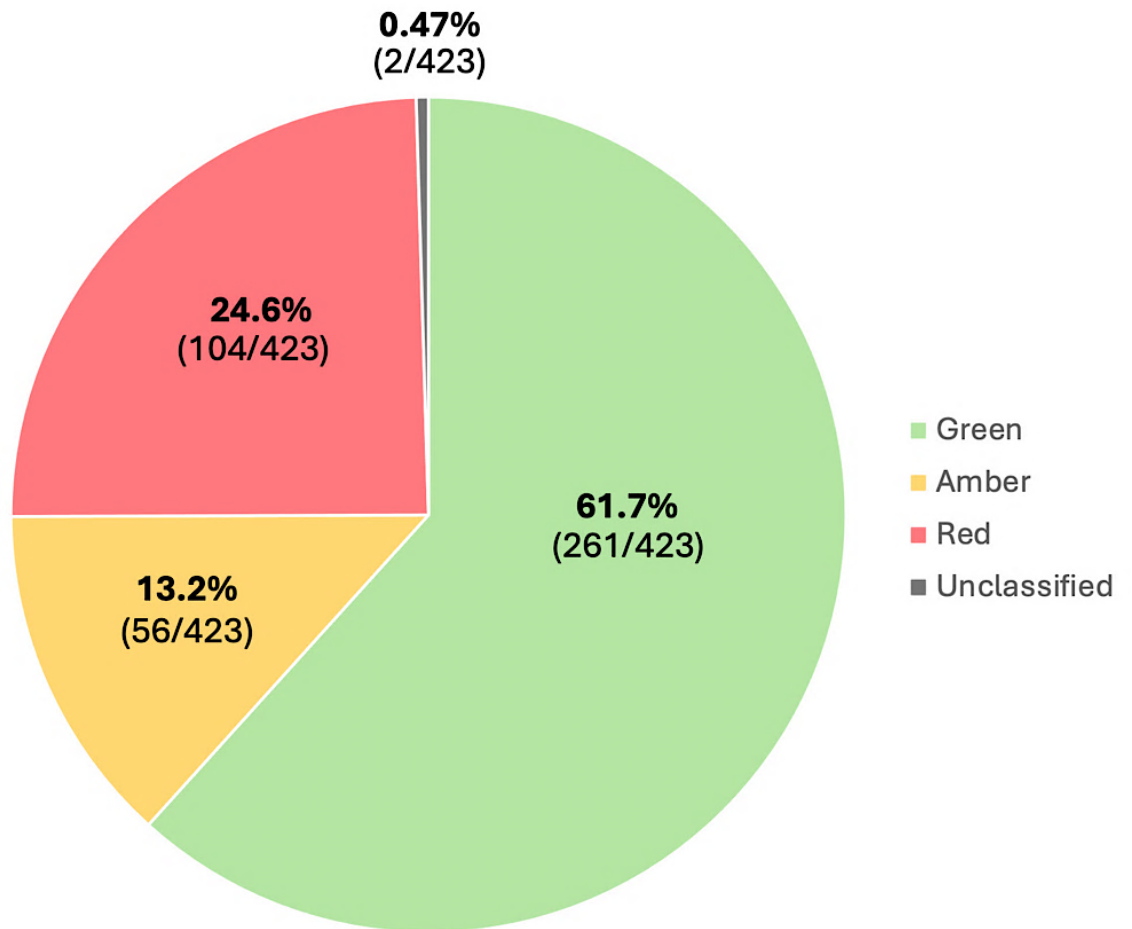
Group: GeCIP domain

Workplace: Research lab

Authorization: Token 0ed0424d9f28cbcd99513f4cf7aea7bc7f17c14d



**Figure 4.20 Isolated FH genes reviewed in retinal disorders v5.11.** Four genes implicated in FH were reviewed in the panel using data from the published literature. Access to gene panels was facilitated using PanelApp <https://panelapp.genomicsengland.co.uk>. The gene reviewer authorisation token for the specific user is displayed.



**Figure 4.21. Categorisation of genes in the retinal disorders v5.11 panel.**

Only genes categorised as green (61.7%) are diagnostically reportable and clinically actionable. Genes classified as amber (13.2%) have moderate evidence to support gene-disease association but are not to be used in a diagnostic test. These genes have evidence for disease in less than three unrelated family members. The genes listed as red have insufficient evidence, making them the least reliable, and are not included in variant interpretation. 0.4% of the gene panel remains unclassified. (PanelApp visited on 02/07/24).

ABCA4	CACNA2D4	DHDDS	KCNV2	PAX2	RGS9	TSPAN12	GDF6	SPP2	FREM2	PODNL1
ABCC6	CAPN5	DRAM2	KIAA1549	PCDH15	RHO	TTC8	IFT81	SP TLC1	FSCN2	POMZP3
ABHD12	CC2D2A	EFEMP1	KIF11	PCYT1A	RIMS2	TLL5	JAG1	TTC21B	FUT5	PRTFDC1
ACBD5	CDH23	ELOVL4	KIZ	PDE6A	RLBP1	TUB	KIF3B	TTPA	GNPTAB	RB1
ACO2	CDH3	ERCC6	KLHL7	PDE6B	RNU4ATAC	TUBB4B	LIG3	UBAP1L	GP1BA	RGS9BP
ADAM9	CDHR1	ERCC8	LAMA1	PDE6C	ROM1	TUBGCP4	LRR32	VWA8	GRIP1	RIMS1
ADAMTS18	CEP164	EYS	LAMP2	PDE6G	RP1	TUBGCP6	MAPKAPK3	ADGRA3	HARS	SLC24A5
ADGRV1	CEP250	FAM161A	LCA5	PDSS1	RP1L1	TULP1	MCOLN1	AMN	HKDC1	SLC45A2
AFG3L2	CEP290	FAM57B	LRAT	PEX1	RP2	UNC119	MIR204	AP3B2	HMCN1	SLC7A14
AGBL5	CEP78	FLVCR1	LRIT3	PEX2	RP9	USH1C	GDF6	ARMS2	HTRA1	SMOC1
AHI1	CERKL	FZD4	LRP2	PEX6	RPE65	USH1G	IFT81	ATP13A2	INVS	SOX2
AIPL1	CFH	GNAT1	LRP5	PEX7	RPGR	USH2A	JAG1	ATXN7	IRX5	SPTLC2
AIRE	CHM	GNAT2	LZTFL1	PHYH	RPGRIP1	USP45	KIF3B	B3GLCT	IRX6	STRA6
ALDH3A2	CLN3	GNB3	MAK	PLA2G5	RPGRIP1L	VCAN	LIG3	BBIP1	ITIH2	TCTN1
ALMS1	CLN5	GNPTG	MED12	PLK4	RS1	VPS13B	LRR32	BCOR	ITM2B	TCTN2
ALPK1	CLN6	GPR143	MERTK	PNPLA6	SAG	WDPCP	MAPKAPK3	BMP4	KCTD7	TCTN3
AMACR	CLN8	GPR179	MFRP	POC1B	SCAPER	WDR19	MCOLN1	C2	KIF7	TEAD1
ARHGEF18	CLRN1	GRK1	MFS08	POMGNT1	SDCCAG8	WHRN	MIR204	C3	LRMDA	TEX28
ARL13B	CNGA1	GRM6	MKKS	POMT1	SGSH	ZFYVE26	MPDZ	C5orf42	LRP1	TMEM126A
ARL2BP	CNGA3	GRN	MKS1	PPT1	SLC24A1	ZNF408	MT-ATP6	CA4	MFN2	TMEM67
ARL3	CNGB1	GUCA1A	MMACHC	PRCD	SLC38A8	ZNF423	MT-TH	CCZ1B	MT-ND1	TRIM32
ARL6	CNGB3	GUCA1B	MSTO1	PRDM13	SLC6A6	ADIPOR1	MT-TL1	CEP41	MT-ND4	TYR
ARSG	CNNM4	GUCY2D	MTTP	PROM1	SNRNP200	AHR	MT-TP	CFB	MT-ND6	TYRP1
ATF6	COL11A1	HCCS	MYO7A	PRPF3	SPATA7	ASRGL1	MT-TS2	CIB2	MYOC	VAX1
ATOH7	COL18A1	HGSNAT	NDP	PRPF31	SRD5A3	C12orf65	MVK	COL11A2	NAALADL1	VSX2
BBS1	COL2A1	HK1	NEUROD1	PRPF4	SSBP1	CCT2	NBAS	COQ4	NEK2	WASF3
BBS10	COL4A1	HMX1	NMNAT1	PRPF6	STN1	CEP19	OPN1SW	CROCC	NR2F1	WFS1
BBS12	COL9A1	IDH3A	NPHP1	PRPF8	TIMM8A	CFAP20	PDE6H	CTSF	NUMB	WT1
BBS2	COL9A2	IDH3B	NPHP3	PRPH2	TIMP3	CLCC1	PGK1	CUBN	OCA2	ZNF513
BBS4	COL9A3	IFT140	NPHP4	PRPS1	TINF2	CLUAP1	POC5	CYP1B1	OPA1	ZPR1
BBS5	COQ2	IFT172	NR2E3	RAB28	TMEM216	COQ5	POMGNT2	CYP27A1	OPA3	EVR3
BBS7	CRB1	IFT27	NRL	RAX2	TMEM218	CTNND1	PYGM	DTHD1	OR2M7	ATXN7_CAG
BBS9	CRX	IFT74	NYX	RBP3	TMEM231	CYP2R1	RDH11	EMC1	PAK2	
BEST1	CSPP1	IKBK G	OAT	RBP4	TMEM237	DHX38	RTN4IP1	FAM71A	PAX6	
C1QTNF5	CTC1	IMPDH1	OFD1	RCBTB1	TOPORS	DMD	SAMD11	FBLN5	PDAP1	
C21orf2	CTNNA1	IMPG1	OPN1LW	RD3	TPP1	DYNC2H1	SAMD7	FOXC1	PDZD7	
C2orf71	CTNNB1	IMPG2	OPN1MW	RDH12	TRAF3IP1	ELOVL1	SEMA4A	FOX E3	PITPNM3	
C8orf37	CTSD	INPP5E	OTX2	RDH5	TREX1	ESPN	SLC25A46	FOX I2	PITX2	
CABP4	CWC27	IQCB1	P3H2	REEP6	TRNT1	EXOSC2	SLC37A3	FRAS1	PITX3	
CACNA1F	CYP4V2	KCNJ13	PANK2	RGR	TRPM1	FRMD7	SPG7	FREM1	PLD4	

**Table 4.8. Gene list in the retinal disorders v5.11 panel.** All 423 genes included in the gene panel are displayed and highlighted in colours according to their classification. Green highlights genes that are diagnostic grade and currently used in variant interpretation. Amber or red denotes genes that require additional evidence to support their clinical association in IRD. Neither amber nor red genes are utilised in variant interpretation using the retinal disorders v5.11 panel. Genes that are pending a classification are highlighted in black.

## 4.3 Discussion

### 4.3.1 Overview of the analysis in the local unsolved FH cohort

The genomic investigations in the local cohort of unsolved FH has delivered a genomic diagnosis in 3/5 probands. In proband F1335, a VUS variant NM\_181507.2:c.2615C>T; NP\_852608.1p.(Pro872Leu) was detected in *HPS5* and was predicted to have a deleterious effect and it affects residues conserved in multiple species including *Danio rerio* (zebrafish) and *Xenopus tropicalis* (western clawed frog) but not in *Caenorhabditis elegans* (roundworm). This missense variant was missed in previous targeted analysis which focused only on *SLC38A8*. Though variant segregation in the family, particularly in the proband needs be reformed using Sanger sequencing for confirmation of the NGS finding in F1335.

Access to a novel SV calling tool known as SVRare allowed for the analysis of SVs called in two patients from the local unsolved FH cohort, proband F1369 and F1071. Applying domain specific knowledge of FH for the interpretation of SVs detected enabled the identification of two pathogenic CNVs in F1369 and F1071 that were missed on previous WGS analysis by the Carmel Toome's group. The CNVs were validated with their breakpoints being characterised using long read ONT sequencing which provided a better resolution of SVs in comparison to short read NGS. The CNVs, NC\_000011.10:g.31773553\_31797720del encompasses *PAX6* and *ELP4* in F1369 and the NC\_000023.11:g.9384915\_9982211del eliminates *GPR143*, *TBL1X*, *SHROOM2* and *CLDN34* in F1071. Analysis of the genomic sequences surrounding the defined breakpoints of both SVs showed that NC\_000023.11:g.9384915\_9982211del in particular had significant repeat elements of three SINEs and an LTR in close proximity. Moreover, only the g.9384915\_9982211del showed significant sequence homology of 50% for a 201 bp sequence close to both breakpoints at chrX:9384715-9384915 and chrX:9982211-9982411.

In another unsolved FH proband (F1377), the genomic investigation entailed a reanalysis of the whole genome to detect pathogenic variants in *SLC38A8* and other FH genes that were missed in previous WGS analysis prior to the start of

this doctoral study. The reannotation of the whole genome was conducted using VEP plugins of CADD, SpliceAI and UTRannotator to characterise all variants including those residing within the UTR intronic and intergenic regions of the genome. However, the analysis did not detect significant variants that could explain the FH phenotype in the proband.

### **4.3.2 Genomic findings in the unsolved local FH cohort**

#### **4.3.2.1 Autozygosity mapping identifies a novel *HPS5* variant in an FH case**

Autozygosity mapping coupled with NGS is a powerful diagnostic strategy for identifying the genetic aetiology in autosomal recessive conditions when there is suspected consanguinity (Maria et al., 2015). The underlying principle is that pathogenic alleles, despite being individually rare in the population, can have a higher chance of co-occurring in the offspring of closely related individuals. This is due to the presence of large genomic regions known as ROH that are identical by descent from a common ancestor (Wang et al., 2009). Pathogenic variants within these ROH will be homozygous and will occur in linkage disequilibrium with other SNPs in the ROH. This causes them to be co-inherited during meiotic recombination (Saqib et al., 2015).

Autozygosity mapping was performed in proband F1335 based on the autosomal recessive inheritance pattern of FH and the presence of consanguineous parents (Figure 4.1). This investigative approach led to the identification of a rare missense variant NM\_181507.2:c.2615C>T; NP\_852608.1p.(Pro872Leu) in *HPS5* that was classified as a VUS. This *HPS5* variant call is of high confidence considering the variant allele was detected in all reads at 39 X (Figure 4.2A). Genotyping the proband's family (F5) using Sanger sequencing for confirmation purposes showed that the *HPS5* variant was present as a heterozygous variant in all asymptomatic family members. The *HPS5* missense variant NM\_181507.2:c.2615C>T; NP\_852608.1p.(Pro872Leu) is absent from clinically significant variant and population databases like the 100KGP, ClinVar and LOVD, implying it is a very rare allele. Nevertheless, the variant was present in the gnomAD database in four South Asian carriers at an allele frequency of

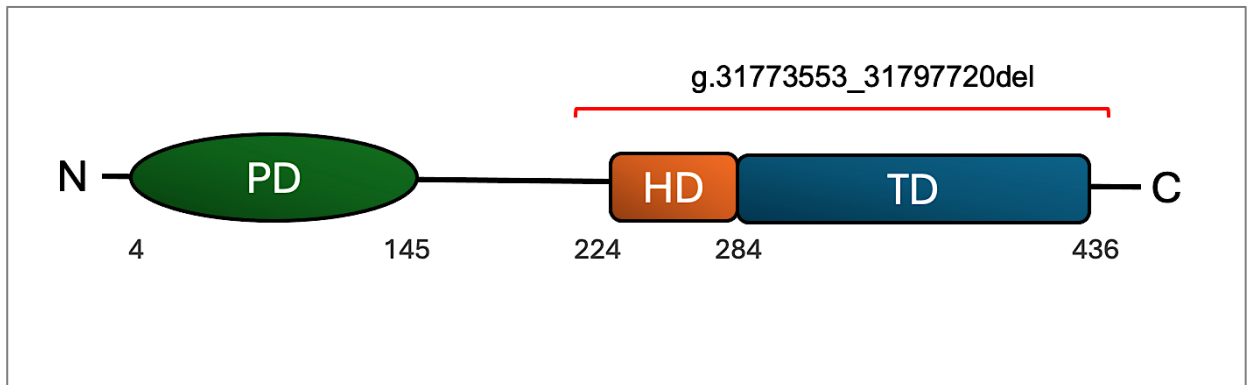
0.0000439 (4/91072) (Figure 4.2B). To the best knowledge of the author, there are no reports of this variant in the published literature, which suggests that NM\_181507.2:c.2615C>T; NP\_852608.1p.(Pro872Leu) represents a novel *HPS5* variant causing HPS type 5, though it will be necessary to resequence the *HPS5* gene in the proband in order to confirm this finding. This rare variant substitutes a conserved proline at residue 872 for leucine. Proline and leucine are both nonpolar and hydrophobic, but the replacement of proline at this residue would be expected to have severe consequences on protein folding and assembly (Alderson et al., 2018). The spatial arrangement of molecules around the peptidyl bond of proline have a dihedral angle ( $\omega$ ) of  $0^\circ$  which significantly alters the secondary structure of the nascent protein (Gurung et al., 2023). The actual protein domain affected by the missense variant is yet to be elucidated. Currently, the only characterised domain in *HPS5* is the WD40 domain located at the *N*-terminus, which is required for interactions with BLOC-2 subunits of HPS3 and HPS6 (Daly et al., 2013). Therefore, the likely pathogenic mechanism for this variant cannot be determined at this time.

The identification of homozygous blocks using SNP microarrays has in the past been used to narrow down genomic targets for Sanger sequencing in consanguineous or endogamous families with a recessive disorder (Ghofrani et al., 2017). In the era of massively parallel sequencing, autozygosity mapping can still be useful for variant identification in known genes but it is also applicable in novel gene discovery, by mapping the disease locus to an ROH shared amongst affected members of the same family (Carr et al., 2013). In hindsight, instead of using autozygosity mapping with NGS data, analysis of the coding segments in albinism genes using a gene panel or WES alone would have been more appropriate and would have led to a more rapid identification of the homozygous variant in *HPS5* in proband F1335. The decision not to perform an initial assessment of albinism genes may have occurred because the phenotypic report of the proband details only FH (Table 4.1). The lack of reporting of classic symptoms of HPS, including bleeding diathesis and oculocutaneous albinism, may imply a milder phenotype which is more challenging to detect. This is a diagnostic dilemma that particularly affect patients with *HPS3*, *HPS5* and *HPS6* variants, who present milder forms of HPS that often gets misdiagnosed as ocular albinism (Michaud et al., 2017).

#### 4.3.2.2 The characterisation of a novel *PAX6* deletion in aniridia using ONT-based long read sequencing

The clinical utility of large variant callers such as SVRare was demonstrated in our local unsolved FH cohort, where it first confirmed the presence of a known *SLC38A8* inversion in a proband F1310, then identified a pathogenic 24.5 kb deletion in chromosome 11 of proband F1369 (Figure 4.5A). The genomic coordinates assigned to the SV using short-read WGS analysis were chr11:31773356-31797899, but this was an estimate since short reads offer low resolution. Accurate determination of the exact breakpoints was achieved using long read sequencing, which refined the genomic coordinates to chr11:31773553-31797720 (Figure 4.7). There is a significant discrepancy of 376 bp in the genomic coordinates defined by short and long-read sequencing, which further highlights the limitation of short-read NGS in SV discovery.

This heterozygous SV, NC\_000011.10:g.31773553\_31797720del deletes exons 9-14 of *PAX6* and exon 10 of *ELP4* but the exact molecular consequence cannot be determined without empirical evidence. The aberrant *PAX6* transcript may be subjected to NMD and no protein would be produced. Another possible outcome is that the transcript could include additional sequence that is immediately downstream of the breakpoint in *ELPF4* intron 10. In the case of this transcript escaping NMD then the protein will lose 248 amino acids (residue 188-436) encoded by exon 9-14 and will incorporate 70 incorrect amino acids to result in a truncated and aberrant protein of 258 amino acids (Appendix C: Supplementary Figure 4.1). The protein effectively loses 57% of its residues (248/436) including an evolutionarily conserved DNA binding motif known as homeodomain and a transactivating domain (Figure 4.22). The homeodomain (224-284 aa) that is deleted is required for the activation of target genes during ocular development. The transactivating domain is a proline, serine and threonine rich region that is involved in the regulation of transcription and the DNA binding affinity of *PAX6* (Mishra et al., 2002, Lima Cunha et al., 2019).



**Figure 4.22. PAX6 protein schematic.** Molecular consequences of NC\_000011.10:g.31773553\_31797720del on PAX6 protein. PD: paired domain, HD: homeobox domain and TD: transactivating domain. The amino acid residues are annotated based on NP\_001355823.1.

The confirmation of the NGS finding by a secondary test (ONT sequencing) allowed the characterisation of the *PAX6* deletion as NC\_000011.10:g.31773553\_31797720del. This deletion is likely to have a relatively severe impact on *PAX6* function, which would be expected to result in the more severe phenotype of aniridia, as opposed to isolated FH which is mainly attributed to missense variants (70%) (Lima Cunha et al., 2019). Aniridia is a malformation of the iris, which would be hard to miss on clinical examination. Nevertheless, the lack of iris hypoplasia or coloboma reports in proband F1369 may suggest a non-aniridia phenotype such as isolated FH and optic nerve hypoplasia.

This NC\_000011.10:g.31773553\_31797720del SV was not reported in publicly available variant archives. The population database of gnomAD CNV and gnomAD SV were inspected but they contain very limited data. The gnomAD CNV contains a single *PAX6* entry for a deletion of 43413 bp which spans chr11:31763314-31806727. Similarly, the larger gnomAD SV database has reports of three deletions in *PAX6* that are  $\geq 1$  kb (visited 12/07/24). The gnomAD SV v4 contains 1199117 SVs in total, in comparison to gnomAD v4 which holds 786500648 small variants (<https://gnomad.broadinstitute.org/stats>, Visited 04/08/24). The paucity in both benign and pathogenic *PAX6* SVs reported in



these population databases is in part due to SVs in general being a more complex genomic variation that occurs less frequently than SNVs or INDELS that are formed by mismatch wobble pairing and replication slippage respectively. Moreover, SVs including other pathogenic SNVs in *PAX6* are not retained through natural selection, as shown by the LOEUF score of 0.23 which indicates negative selection against LOF variants in the *PAX6* MANE transcript (ENST00000640368.2). These observations may explain why *PAX6* pathogenic variants, and particularly SVs, are relatively rare. The likely phenotype of the LOF variants is aniridia, a severe condition associated with ocular malformations including iris hypoplasia and coloboma. These abnormal and conspicuous traits are likely to reduce reproductive fitness in affected individuals. Another contributing factor could be the difficulty in detecting SVs using current diagnostic technologies like WES and Sanger sequencing which can contribute to the scarcity of SVs reported in population databases.

In spite of this apparent paucity of *PAX6* SVs in public databases, there is a study in the literature which used WGS in a previously analysed and unsolved aniridia cohort and identified 12 such variants that had been missed (Hall et al., 2024). Amongst these SVs are a g.31777022\_31796016del and g.31780738\_31796994del, both of which also encompass exons 9-14 of *PAX6* and exon 10 of *ELP4*. The NC\_000011.10:g.31780738\_31796994del from that study has a breakpoint in intron 8 of *PAX6* that is 726 bp away from the SV detected in proband F1369, NC\_000011.10:g.31773553\_31797720del. This warranted further investigation into the locus chr11:31796016-31797720 as an apparent hotspot for SVs, perhaps arising from genomic instability. This 1733 bp region was found to be rich in repeat sequences (~12%) that could predispose to replication errors that would generate SVs.

The chr11:31796016-31797720 region also contains four high epigenetic signatures for H3K4me3 and High H3K27ac that are indicative of a distal enhancer like CRE in the ENCODE database. These CREs may influence *PAX6* expression and these are at chr11:31796177-31796498 (EH38E2951364), chr11:31796830-31797024 (EH38E2951365), chr11:31797621-31797863 (EH38E2951367) and chr11:31797285-31797545 (EH38E2951366). The

experimental validation of this finding is beyond the scope of this thesis but could provide the basis for future work in *PAX6* regulation.

#### **4.3.2.3 The identification of a novel *GPR143* deletion in ocular albinism**

WGS provides uninterrupted sequence, which facilitates the identification of SVs, especially those with breakpoints in intronic or intergenic regions. However, short read NGS may not provide adequate resolution to accurately determine the breakpoints of an SV. Such information is required for accurate characterisation of SVs according to the HGVS nomenclature. Abiding by this standardised reporting guideline ensures a single variant definition is used for submission in population and variant databases.

WGS performed on proband F1071 revealed a 597.9 kb deletion on chromosome X:9384641-9982520 (Figure 4.5B). The male proband loses four entire genes, *TBL1X*, *GPR143*, *SHROOM2* and *CLDN34*. This clinically actionable result was confirmed using ONT long-read sequencing, which redefined the genomic coordinates of the SV as chrX:9384915-9982211 (Figure 4.8). These breakpoints are inadequately resolved as evident by the slight variability in read length as the primers used in PCR target amplification bind to a region rich in repeats (Appendix C: Supplementary Figure 4.2). These primers partially anneal at different positions, resulting in diverse amplicons that vary slightly in length. This generates inconsistency in reads assembled, making it difficult to define the exact genomic breakpoints of the deletion at single nucleotide resolution. To further address this issue, a primer should be redesigned to anneal to a unique sequence at chrX:9382915-9984211 that is downstream the repeat locus on chromosome X.

Amongst the four genes eliminated by NC\_000023.11:g.9384915\_9982211del, only *GPR143* is known to be of medical relevance. The phenotype manifesting from a hemizygous *GPR143* null allele is ocular albinism, which consists of FH, chiasmal misrouting, iris TID, fundus hypopigmentation and nystagmus. Nonetheless, the proband was described as having FH without any additional symptoms of ocular hypopigmentation. This could be attributed to the variability

in the phenotypic expression of hypopigmentation and chiasmal misrouting with age that has been documented in albinism (Campbell et al., 2019). Moreover, affected males sharing the same pathogenic *GPR143* genotype of hemizygous NM\_000275.3:c.623C>A; NP\_000266.2:p.(Ala208Glu) were reported to vary in ocular manifestations of iris TID, fundus hypopigmentation and nystagmus (Jung et al., 2018). The albinism in F1071 may present a mild form of iris TID that was missed in this proband, while the absence of chiasmal misrouting data may be due to inaccessibility to VEP testing or the technical challenges of performing VEP on infants. Nevertheless, the finding of NC\_000023.11:g.9384915\_9982211del in F1071 suggests that the initial diagnosis of isolated FH needs revisiting by the responsible clinician as the proband is likely to be affected by ocular albinism.

#### **4.3.2.4 Searching for a pathogenic variant in proband F1377**

The identification of a monoallelic variant in a recessive condition is of no clinical utility unless intermediate phenotypes exist in carriers of a single pathogenic allele (Kuht et al., 2023). Proband F1377 was diagnosed with FH and chiasmal misrouting in the absence of any obvious pigmentation defect. This specific phenotype narrowed the genomic investigation and led to the discovery of a single plausible variant in *SLC38A8* NM\_001080442.3:c.2T>C; NP\_001073911.1:p.(Met1?) prior to the beginning of this doctoral study. An investigation to look for a possible missing variant in this patient was carried out as part of this project. It relied on the reassessment of genomic files generated previously. The already existing BAM file was inspected for quality control, which confirmed that the DNA library preparation and sequencing performance were adequate (Figure 4.12). The reads generated are all within the range of 139-151 bp with no intermediate reads that could imply poor library preparation (Figure 4.12A). The mean quality along the entire length of the read is  $\geq$ Q28 which translates to an error rate of 1 in 633 bp (Figure 4.12B). The high accuracy of base calling enabled the determination of every base with no ambiguous bases assigned (Figure 4.12C). The adaptor trimming process was also successful as there is no significant evidence of adapter sequence contamination (Figure 4.12D).

The previously generated VCF was annotated with dedicated tools to elucidate deep intronic variants that could affect the splice site or 5' untranslated region. The reanalysis did not detect any additional significant variants in *SLC38A8* nor in any of the FH genes. This could be because the pathogenic variant or variants are in a region of low complexity that is rich in repeats. The ambiguous mapping of repetitive or highly homologous reads would result in low representation of a target region and any variant discovered would be filtered out if it does not exceed the minimum coverage threshold (Yahya et al., 2023). Further testing in the future should consider genome sequencing using long read sequencing technology. This approach would generate large individual reads that span entire intractable regions to overcome the limitations of the ambiguous mapping of repetitive sequences that is experienced with short read NGS. The longer reads would also facilitate discovery of SVs with breakpoints in low complexity regions. However, another plausible explanation is that the NM\_001080442.3:c.2T>C; NP\_001073911.1:p.(Met1?) variant is a coincidental finding and that the pathogenic variants responsible for FH reside in a novel FH gene that is uncharacterised. This may have caused variants in this gene to be missed in the targeted analysis of known FH and albinism genes.

### **4.3.3 100KGP diagnostic odyssey in FH**

#### **4.3.3.1 Overview of the FH cohort**

The generation of an FH cohort from the 100KGP dataset for this project was based on select HPO terms of hypoplasia of the fovea, foveal hypoplasia and aplasia of the fovea for the confirmed clinical diagnosis of FH (Figure 4.14). This cohort of probands with unsolved FH were further stratified based on particular demographics. For example, the majority of this FH cohort were children, which is expected for a set of inherited disorders with ocular symptoms like nystagmus and poor visual acuity being apparent during infancy and early childhood. The benefit of working on a cohort that is defined by a confirmed FH diagnosis is that it can be utilised for the purpose of novel FH gene discovery if no plausible mutation has been identified in the participants.

#### 4.3.3.2 The FH diagnostic yield in the 100KGP

This FH cohort had 47 unsolved cases and 3 probands of unknown status, who were grouped into a cohort of 50 probands with unsolved FH. The initial analysis was performed by analysing small rare variants annotated by Exomiser in 52 known FH genes. This has led to the identification of rare and deleterious variants that are compatible with the mode of inheritance for the FH gene in 10 participants (Table 4.5). The genomic investigation in the remaining cohort of 40 unsolved FH cases, was conducted using SVRare to detect variants  $\geq 50$  bp which are not called or assessed by either the Exomiser or Tiering pipelines. The remaining 40 probands with unsolved FH were analysed for the purpose of SV discovery in all 52 genes known to cause FH and this analysis identified a pathogenic deletion in *OCA2* in a proband who was already known to have a second pathogenic *OCA2* SNV (Figure 4.17). The diagnostic yield delivered using WGS in this FH cohort was therefore 22% (11/50). This is comparable to another recent study using WGS in unsolved IRDs, in which a molecular diagnosis was obtained in 13% (34/271) of the cases (Liu et al., 2024). However, that IRD cohort (n=271) was much larger (5.4X) and the number of ophthalmic genes analysed (n=792) was also significantly higher (15X) and covered 83% of the 52 known FH genes. The diagnostic analysis for FH in the 100KGP has therefore been efficient and provides an illustration of the translational potential of genomic discoveries in the larger cohort (73512 genomes) and the viable collaborative links available with the NHS genomic medicine centres.

In addition, participants with potential FH that were not reported with the HPO terms of hypoplasia of the fovea, foveal hypoplasia or aplasia of the fovea were analysed using the newly available SVRare data in a reverse phenotyping approach. This is based on establishing a pathogenic genotype to infer FH based on the high penetrance in FH genes. This additional analysis focused on identifying SVs in 27 genes implicated in isolated FH and albinism, and this endeavour led to identification of pathogenic SVs responsible for FH in nine probands (Table 4.7).

The combined 100KGP analysis in FH has yielded a genomic diagnosis in 20 probands. The analysis has also identified 12 novel variants comprising 3 SNVs and 9 SVs. The genomic diagnosis entails oculocutaneous albinism in six probands (P9, P10, P11 and P18-P20), ocular albinism in four probands (P21-P24), achromatopsia (P8 and P12), aniridia, optic nerve hypoplasia or isolated FH (P13, P15, P25 and P27), X-linked Aland Island disease (P14 and P26), Stickler syndrome (P17) and a case familial exudative vitreoretinopathy (P16). By far albinism represents the majority of the solved FH cases (50%) followed by a *PAX6*- related phenotype (20%).

#### **4.3.3.3 Challenges in defining an FH cohort in the 100KGP**

The defined FH cohort which is based on the exact phenotype report of FH has its merits but the caveat with such a selective cohort is that many IRD cases with FH are not reported with the exact clinical manifestation of FH in the 100KGP. This is attributed to the non standardised diagnostic testing performed at clinics which frequently lack OCT and VEP examination. As a result, this definitive FH cohort represents a proportion of all the cases that have an FH phenotype of high confidence.

The attempts to capture probands with FH who are not reported with HPO terms for FH could rely on less stringent phenotypic criteria indicative of an IRD instead. However, the HPO terms used in the phenotypic reports can be vague or redundant, and their usage during at clinics during recruitment has proved inconsistent. As a result, interpreting that phenotypic data for each participant in the 100KGP for categorisation based on a suspected FH-related phenotype can be challenging and is exacerbated by the phenotypic overlap of different IRDs with FH. The limited and often unreliable clinical information available from patients recruited to the 100KGP can hinder genomic investigations leading to prolonged analysis (Best et al 2022).

The alternative strategy of using the genotype to infer phenotype is complicated by the genetic and phenotypic heterogeneity in FH-related disease. To date, 52 genes and a locus (*OCA5*) can be involved in this disorder. Many of the variants

in these genes will be benign or incompatible with mode of inheritance for the disease, and some of the patients with pathogenic variants in these genes will have other related IRDs. Therefore, capturing a cohort of participants with variants in FH-related genes cannot serve as an FH cohort since not all probands will have FH.

Increasing the diagnostic yield in FH would therefore require a genomic analysis in a cohort with a confirmed phenotype of FH as indicated by the exact HPO term of FH reported (definitive FH cohort) and another gene specific cohort based on participants with variants in FH genes (potential FH cohort) that is irrespective of the phenotypes reported. The investigative approach in this chapter has partly satisfied this criteria however the assessment of deep intronic variants in the remaining unsolved 39 participants of the defined FH cohort was not performed due to the very large number of intronic variants that require interpretation and time constraint. Similarly, analysis of SVs in all 52 FH genes as part of the potential FH cohort was also not performed but was rather performed on 27 FH genes were.

#### **4.3.3.4 Interpreting pathogenic variants in *TYR* and *OCA2* that are enriched in albinism**

Interrogating the 100KGP datasets identified 14 deleterious SNVs in ten probands recruited with a confirmed phenotype of FH but without a genomic diagnosis (Table 4.5). These variants identified were reported in *CACNA1F*, *CNGA3*, *CNGB3*, *COLL1A1*, *OCA2*, *PAX6*, *TYR* and *ZNF408*. Of these variants, NM\_001368894.2:c.399+1G>A and NP\_001355823.1:p.(Ser273Pro) in *PAX6* and NP\_001243718.1:p.(Cys292Tyr) in *CACNA1F* were not previously reported in gnomAD or LOVD. The submission of these two variants to publicly available variant databases and their reporting in the literature will expand the mutation spectrum in Aland Island disease and aniridia.

Other pathogenic variants detected in the 100KGP participants include NM\_000372.5:c.1118C>A; NP\_000363.1:p.(Thr373Lys) and NM\_000372.5:c.1217C>T; NP\_000363.1:p.(Pro406Leu) in *TYR* and NM\_000275.3:c.1255G>A; NP\_000266.2:p.(Val443Ile) in *OCA2*, all of which are deleterious but have a homozygous count reported in gnomAD (Table 4.5). These likely pathogenic variants affect conserved residues (Figure 4.23). The Thr373 in *TYR* displays moderate conservation whereby the residue is conserved in mammals such as *Bos taurus* (Cattle) but not in birds like *Gallus gallus* (Chicken). The Pro406 in *TYR* is evolutionary conserved in *Dani rerio* (Zebra Fish) but not in *Xeno tropicalis* (Tropical clawed frog). The Val443 in *OCA2* is fully conserved in *Dani rerio* (Zebra Fish) and in six other evolutionarily diverse species including *Xeno tropicalis*. The high allele count for these pathogenic variants is exclusive to Europeans and it indicates a high carrier rate in the general population (Table 4.5). Interestingly, these *TYR* and *OCA2* variants are also enriched in albinism patients which further reinforces the genomic diagnosis delivered in probands P10 and P11.

Both NP\_000363.1:p.(Thr373Lys) and NP\_000363.1:p.(Pro406Leu) in *TYR* were detected in 31 patients and the *OCA2* variant NP\_000266.2:p.(Val443Ile) was in 52 patients from a cohort of 991 individuals suspected with albinism (Lasseaux et al., 2018). The high allele counts for these variants may be driven by positive selection for depigmented traits, as is thought to be the case for the *TYR* polymorphism NM\_000372.5:c.575C>A; NP\_000363.1:p.(Ser192Tyr) which is associated with lightly pigmented skin, hair and eye colour (Wilde et al., 2014). Evidence for natural selection with regards to depigmented and conspicuous features was traced in European populations using genetic markers dating back ~5000 years (Wilde et al., 2014). It has been demonstrated that the NP\_000363.1:p.(Ser192Tyr) in *TYR* is hypomorphic and when in cis with NP\_000363.1:p.(Arg402Gln) is sufficient for pathogenicity (Loftus et al., 2023). This complex allele NP\_000363.1:p.[Ser192Tyr; Arg402Gln] has been reported in trans with NP\_000363.1:p.(Thr373Lys) and NP\_000363.1:p.(Pro406Leu) in 13 individuals diagnosed with oculocutaneous albinism which further confirms the pathogenicity of these variants (Loftus et al., 2023). The caveat with this study is that albinism was diagnosed based on the observed pigmentation, which is



subjective, and that the albinism hallmark of FH and chiasmal misrouting were not consistently reported, which reduces reliability of their cohort.

				p.(Thr373Lys)																																																											
<b>TYR</b>	<i>Homo sapiens</i>	316	ADVEFCLSLTQYESGSM	KAA	NFSFRNTLEGFASPLTGIADASQSSMHNALHIYMG	TMSQVQGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Pan troglodytes</i>	316	ADVEFCLSLTQYESGSM	KAA	NFSFRNTLEGFASPLTGIADASQSSMHNALHIYMG	TMSQVQGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Mus musculus</i>	316	ADVEFCLSLTQYESGSM	DRTAN	FNSFRNTLEGFASPLTGIADPSQSSMHNALHIFMNG	TMSQVQGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Canis lupus familiaris</i>	316	ADVEFCLSLTQYESD	SMDKAA	NFSFRNTLEGFASPLTGIADASQSSMHNALHIYMG	TMSQVPGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Bos taurus</i>	316	ADVEFCLSLTQYESGSM	KAA	NFSFRNTLEGFADPVTGIADASQSSMHNALHIYMG	TMSQVPGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Gallus gallus</i>	316	SEVEFCLTLTQYESGSM	KMAN	YSFRNTLEGFADPHTAISNISQSLHNLHIYMG	TMSQVQGSANDPIFL	LHHA	FVDS	395																																																						
	<i>Xenopus tropicalis</i>	320	AVELCLSLTNYETEP	MDRSAN	FNSFRNTLEGFADPRTGIANRSQSNMHNLSLHVFLNG	SMSVQGSANDPVFL	LHHA	FVDS	399																																																						
	<i>Danio rerio</i>	318	ADVSVLRLTDYETGQ	MRRANL	SFRNALEGFANPETGLAVTGRSLMHNSLHVFMNG	SMSVQGSANDPIFI	IHHAF	IDS	397																																																						
				p.(Pro406Leu)																																																											
<b>TYR</b>	<i>Homo sapiens</i>	396	IFEQWLR	RRH	RFLQ	EVY	PEAN	APIGH	NRES	YMV	PFIP	LYR	NGD	FFISS	KDL	GDY	SYL	QDSD	PD	SFQ	DIK	SYLE	EQAS	RIW	475																																						
	<i>Pan troglodytes</i>	396	IFEQWLR	RRH	RFLQ	EVY	PEAN	APIGH	NRES	YMV	PFIP	LYR	NGD	FFISS	KDL	GDY	SYL	QDSD	PD	SFQ	DIK	SYLE	EQAS	RIW	475																																						
	<i>Mus musculus</i>	396	IFEQWLR	RRH	RFL	LEV	Y	PEAN	APIGH	NRS	Y	MV	PFIP	LYR	NGD	FFITS	KDL	GDY	SYL	QESD	PGF	RY	NI	E	PYLE	EQAS	RIW	475																																			
	<i>Canis lupus familiaris</i>	396	IFEQWLR	RRH	RFL	REV	Y	PEAN	APIGH	NRES	YMV	PFIP	LYR	NGD	FFISS	RDL	GDY	SYL	QESER	D	I	FQD	YIK	PYLE	EQAS	RIW	475																																				
	<i>Bos taurus</i>	396	IFEQWLR	RKY	H	RFL	QD	V	Y	PEAN	APIGH	NRES	YMV	PFIP	LYR	NGD	FFISS	KDL	GDY	SYL	QDSE	P	I	FQD	YIK	PYLE	EQAQ	RIW	475																																		
	<i>Gallus gallus</i>	396	IFERWLR	RRH	R	F	L	E	V	Y	PEAN	APIGH	NREN	Y	MV	PFIP	LYR	NGE	FFISS	R	E	L	G	D	Y	E	Y	L	Q	E	PAL	G	S	FQD	F	L	I	P	Y	L	K	Q	A	H	Q	I	W	475															
	<i>Xenopus tropicalis</i>	400	IFEQWLR	RRH	G	A	S	V	D	I	Y	PEAN	APIGH	N	R	G	Y	M	V	PFIP	LYR	N	G	E	F	F	A	A	S	R	D	L	G	D	Y	D	L	A	E	S	-	G	S	I	E	D	F	L	L	P	Y	L	E	Q	A	R	Q	I	W	477			
	<i>Danio rerio</i>	398	IFEQWLR	RRH	O	P	R	T	H	Y	T	A	N	A	P	I	G	H	N	D	G	Y	M	V	PFIP	LYR	N	G	D	Y	F	L	S	T	K	A	L	G	Y	E	A	L	Q	D	P	G	R	F	V	Q	E	F	L	T	P	Y	L	E	Q	A	Q	I	W
				p.(Val443Ile)																																																											
<b>OCA2</b>	<i>Homo sapiens</i>	374	PSLTHVVEWIDFETL	LALLFG	MMIL	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	A	Y	R	L	S	R	G	R	V	W	A	M	I	I	M	L	C	L	I	A	A	V	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	453				
	<i>Pan troglodytes</i>	382	PSLTHVVEWIDFETL	LALLFG	MMIL	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	A	Y	R	L	S	R	G	R	V	W	A	M	I	I	M	L	C	L	I	A	A	V	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	461				
	<i>Mus musculus</i>	369	PSLTHVVEWIDFETL	LALLFG	MMIL	V	A	V	F	S	E	T	G	F	F	D	Y	CAV	K	A	Y	Q	L	S	R	G	R	V	W	A	M	I	F	M	L	C	L	M	A	A	I	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	448				
	<i>Canis lupus familiaris</i>	380	PSLTHVVEWIDFETL	LALLFG	MMIL	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	T	Y	R	L	S	R	G	R	V	W	A	M	I	I	M	L	C	L	I	A	A	V	L	S	A	F	L	D	N	V	T	L	L	L	F	T	P	V	T	459				
	<i>Bos taurus</i>	381	PSLTHVVEWIDFETL	LALLFG	MMIL	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	V	Y	Q	L	S	R	G	R	V	W	T	M	I	F	M	L	C	L	V	A	A	V	L	S	A	F	L	D	N	V	T	V	L	L	F	T	P	V	T	460				
	<i>Gallus gallus</i>	394	PSMVKVEWIDYETL	LALLFG	MMV	L	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	A	Y	R	F	S	R	G	V	W	A	M	I	T	L	L	C	L	I	A	A	I	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	473				
	<i>Xenopus tropicalis</i>	393	PSLVKVEWIDYETL	LALLFG	MMIL	V	A	V	F	S	D	T	G	F	F	D	Y	CAV	K	A	Y	Q	L	S	R	G	R	I	W	P	M	I	I	L	C	L	I	A	A	I	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	472					
	<i>Danio rerio</i>	369	PSLMTVVEWIDYETL	LALLFG	MMIL	V	A	I	F	S	E	T	G	F	F	D	Y	CAV	K	A	Y	Q	L	S	R	G	R	V	W	P	M	I	I	L	C	L	I	A	A	I	L	S	A	F	L	D	N	V	T	M	L	L	F	T	P	V	T	448					

**Figure 4.23. Protein alignment of TYR and OCA2 orthologs.** Assessment of evolutionary conservation in TYR and OCA2 for three missense variants. *Homo sapiens*: humans; *Pan troglodytes*: *Mus musculus*: house mouse; *Canis lupus familiaris*: Dog; *Bos taurus*: cattle; *Gallus gallus*: chicken; *Xenopus tropicalis*: western clawed frog; *Danio rerio*: zebrafish.

Taken together, these three variants of NP\_000363.1:p.(Thr373Lys) and NP\_000363.1:p.(Pro406Leu) in TYR and the OCA2 variant NP\_000266.2:p.(Val443Ile) are pathogenic yet are enriched in the general population, possibly due to natural selection in Europeans dating back ~5000 years. Another explanation could be that these presumably pathogenic variants are in linkage disequilibrium with other nearby benign variants that are retained through natural selection and eventually result in ancestral haplotypes that are

common within the population. However, the confirmation of the hypothesis would require further validation, which may shed insight into the disease burden associated with recurring variants that are computationally predicted as pathogenic. The complexity associated with different genes due to evolutionary constraints further highlights the requirement for gene specific variant interpretation that accounts for the higher allele frequencies of pathogenic variants in albinism genes like *TYR* and the penetrance associated with complex alleles, both criteria cannot be adequately fulfilled by the generic ACMG classification.

#### **4.3.3.5 Revealing a founder SV in *OCA2* that is of African ancestry**

This study identified two individuals in the 100KGP with the NC\_000015.10:g.28017719\_28020673del SV in *OCA2*, each of whom also carry a second pathogenic allele in *OCA2* (Figure 4.17 and Table 4.7). This variant deletes exon 7, resulting in a frameshift that leads to a PTC in exon 8. The aberrant transcript is predicted to be eliminated by NMD. Both participants were of African or Afro-Caribbean ancestry and have a confirmed or suspected phenotype of FH but without any indication of the hypopigmentation associated with oculocutaneous albinism. The absence of such a conspicuous feature in dark skin toned individuals may reflect the relatively poor phenotypic data available through the 100KGP, which may only disclose part of the full clinical manifestation. The selective and non-uniform reporting of phenotypes in the 100KGP has been covered in section 6.5 and is also discussed elsewhere (Best et al., 2022b).

Participants recruited to the 100KGP are predominantly of Caucasian descent, with only 3.1% (2767/90178) being of African ancestry (Participant explorer v5.6.0 visited on 12/07/24). However, across the entire 100KGP rare disease cohort, regardless of the diagnosis, the NC\_000015.10:g.28017719\_28020673del SV was identified in a total of 24 participants of confirmed African, Afro-Caribbean or mixed heritage, as well as in four individuals with no stated ethnicity and two of other ethnicities (Figure 18A). The enrichment of this SV in people of African ethnicity has significant

implications for diagnostic genetic screening in oculocutaneous albinism. The analysis of the haplotype performed on three probands supports NC\_000015.10:g.28017719\_28020673del being a founder SV in *OCA2* that is restricted to people of African descent (Table 4.6). The data suggest the presence of a common haplotype extending over 44807 bp from 15:28024696-27979889 in three individuals of African and Afro Caribbean heritage. This is consistent with a very old founder variant that was initially present in a common progenitor. Confirmation of the haplotype in the three probands would require long read sequencing to distinguish between paternal and maternal alleles and to confirm the extent of the shared haplotype in these individuals.

Another founder mutation has been reported in *OCA2*, a 2.7 kb deletion that also eliminates exon 7 and was reported in unrelated individuals of African ancestry (Durham-Pierre et al., 1994). A study screening 146 South African patients with oculocutaneous albinism showed that the 2.7 kb SV was present in 78% (114/146) of the patients (Stevens et al., 1995). This recurring 2.7 kb deletion was identified using PCR, agarose gel electrophoresis and southern blotting. All these molecular techniques offer a low resolution and cannot accurately determine the size at the base pair level. As a result, the exact coordinates of the SV and reports of the variant according to HGVS nomenclature are not available. Interestingly, the primers that flank exon 7 of *OCA2* for the “2.7 kb” deletion also span the NC\_000015.10:g.28017719\_28020673del SV identified in the 100KGP (Appendix C: Supplementary Figure 4.3). It is therefore very likely that the uncharacterised 2.7 kb deletion reported in that previous study is in fact the NC\_000015.10:g.28017719\_28020673del (2.95 kb) described in this thesis using NGS, given the close proximity of the two SVs and similar size (250 bp discrepancy). If the “2.7 kb deletion” is confirmed to be the NC\_000015.10:g.28017719\_28020673del, diagnostic services operating in regions with a significant African demographic could adopt Sanger sequencing instead of NGS for a targeted assessment of the NC\_000015.10:g.28017719\_28020673del in individuals of African descent presenting with oculocutaneous albinism. Sanger sequencing is a simpler and cheaper technique that does not involve the informatics burden associated with

NGS, which makes it a more accessible technology in regions of low economic status.

Oculocutaneous albinism due to *OCA2* haploinsufficiency is more common in Africans and has variable incidence rates depending on the geographical region. In Namibia for example it is 1:1755 (Lund and Roberts, 2018). The higher prevalence of albinism in Africans could be a consequence of reoccurring variants like the NC\_000015.10:g.28017719\_28020673del SV due to founder effect or endogamy in tribal communities. Another contributing factor may be that the depigmentation is more evident in dark skin toned individuals which facilitates the diagnosis of albinism.

The NC\_000015.10:g.28017719\_28020673del SV in *OCA2* has a total allele frequency of 0.0002 (30/142,816) in the 100KGP, with no homozygotes detected (Figure 4.17). This exact SV was not reported in gnomAD, but another, NC\_000015.10:g.28017719-28020677, has one identical breakpoint while the other is 4 bp longer. Again this variant is most certainly the g.28017719\_28020673del as it is also enriched in the African demographic, with an allele frequency of 0.00254 (86/33816) and 10 homozygotes reported (gnomAD visited 12/07/24). The slight discrepancy in size with this suspected variant might be related to the variable mapping of SVs using short read sequencing or the algorithmic variation between different genomic aligners used for BAM file generation. The reporting of SVs is very sensitive to changes in the identified breakpoints and slight variation due to technological limitations in DNA sequencing may result in redundant submissions of SVs that ultimately hinders the utility of population databases in filtering variants and in variant interpretation.

A consensus in the defined genomic coordinates of an SV is required and this can be achieved using long read sequencing (McClinton et al., 2023a). However, this technology is not yet widely adopted, and the majority of SVs are identified using short read NGS. To circumvent the issues related to redundant reporting of individual SVs in population databases, SVs that are very similar in terms of genomic coordinates and size could be grouped to accommodate for slight discrepancies in the breakpoints of these large variants. This grouping is

reminiscent to dbSNP, whereby a variant ID is generated for the different possible alleles at a particular locus. This would avoid the confusion in SVs reported as described above, whereby the same SV, NC\_000015.10:g.28017719\_28020673del was most likely reported as NC\_000015.10:g.28017719-28020677 by another laboratory due to technological discrepancies.

## Chapter 5 : The identification of *LAMP1* as a candidate for isolated FH

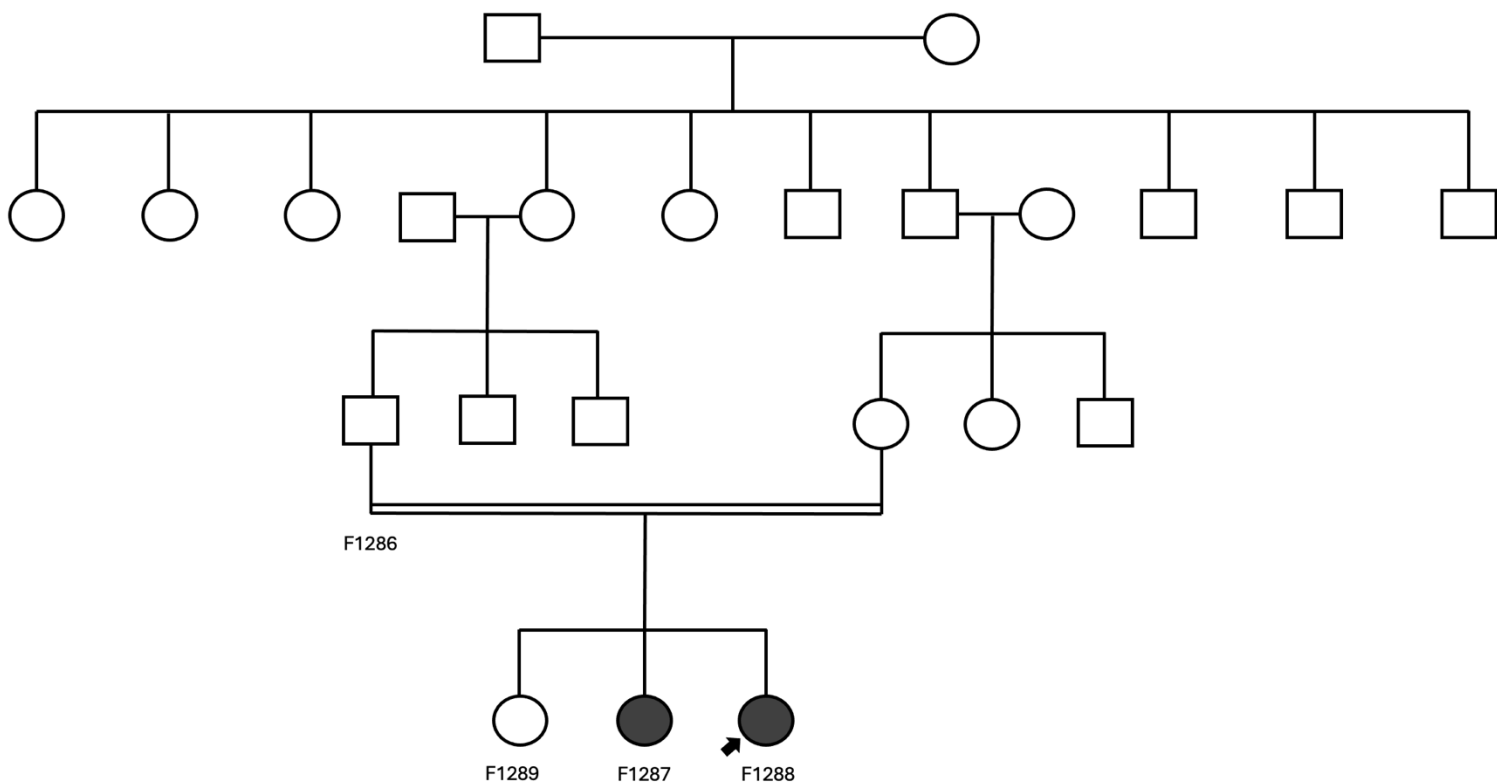
### 5.1 Introduction

In the remaining local unsolved FH cohort, family five (F5), is a South Asian family with two affected individuals and a consanguineous family history (Figure 5.1). The female proband (F1288) and her affected sibling (F1287) both display isolated FH, in the absence of an iris TID. Clinical data on nystagmus and chiasmal misrouting were not available. Previous work by Dr Emma Lord using WES failed to detect pathogenic small variants and CNVs in genes known to be involved in albinism or FH, including *SLC38A8* (Lord, 2018). The lack of plausible variants detected by WES in known FH genes in the affected proband could imply the presence of a pathogenic deep intronic variant that alters splicing or affects regulatory domains such as CREs or non-coding variants at the 3' or 5' UTR that had evaded detection. Moreover, copy number neutral SVs like inversions and translocations would most likely be missed by WES in F1288. Another explanation is that the unsolved FH in the family could be caused by variants in a gene not previously implicated in isolated FH. The existing VCF generated by Emma Lord was reanalysed for the purpose of novel FH gene discovery. This bioinformatics procedure was carried out by Dr Lord shortly before she completed her PhD studies, which led to the identification of a LOF variant in *LAMP1* as a potential novel gene causing isolated FH. However, no further analysis was carried out at that time (Lord, 2018).

The Lysosomal Associated Membrane Protein 1 (*LAMP1*; *CD107a*) gene encodes a highly N-glycosylated type I transmembrane protein that is localised to the membranes of endosomes, lysosomes and LROs such as melanosomes. *LAMP1* is one of the most abundant lysosomal membrane proteins, making it useful as a biomarker for lysosomes and LROs (Cheng et al., 2018). The putative role of *LAMP1* is to maintain lysosomal integrity through forming a glycocalyx on the luminal side of the membrane, which resists the hydrolytic activity of the enzymes encapsulated (Lu, Zhu et al. 2016). To date, no clinical phenotype has been linked to pathogenic germline variants in *LAMP1*. However, increased *LAMP1* expression on the surface of tumour cells is associated with metastatic

potential and is a marker for poor prognosis in diffuse large B-cell lymphoma (Dang, Zhou et al. 2018). A murine model deficient in *LAMP1* showed that *LAMP1* knockout is compatible with life but revealed no symptoms reminiscent of an obvious disease phenotype. However, these mice exhibited subtle abnormalities in the brain, comprising altered distribution of a lysosomal aspartyl protease known as cathepsin-D in the deep lamina VI and superficial laminae II, and detectable astrogliosis in the neocortex (Andrejewski et al., 1999).

The study described in this chapter will explore the hypothesis that loss of function in *LAMP1* is the cause of isolated FH in the unsolved family F5. The aim is that this hypothesis will be further tested by *In vitro* experiments initiated here but not completed due to insufficient time, using cellular models for the functional analysis of the *LAMP1* variant.



**Figure 5.1 Pedigree of unsolved FH family F5.** Two unaffected first cousins are the parents of two affected and one unaffected female offspring. The mode of FH inheritance displayed is likely to be autosomal recessive.

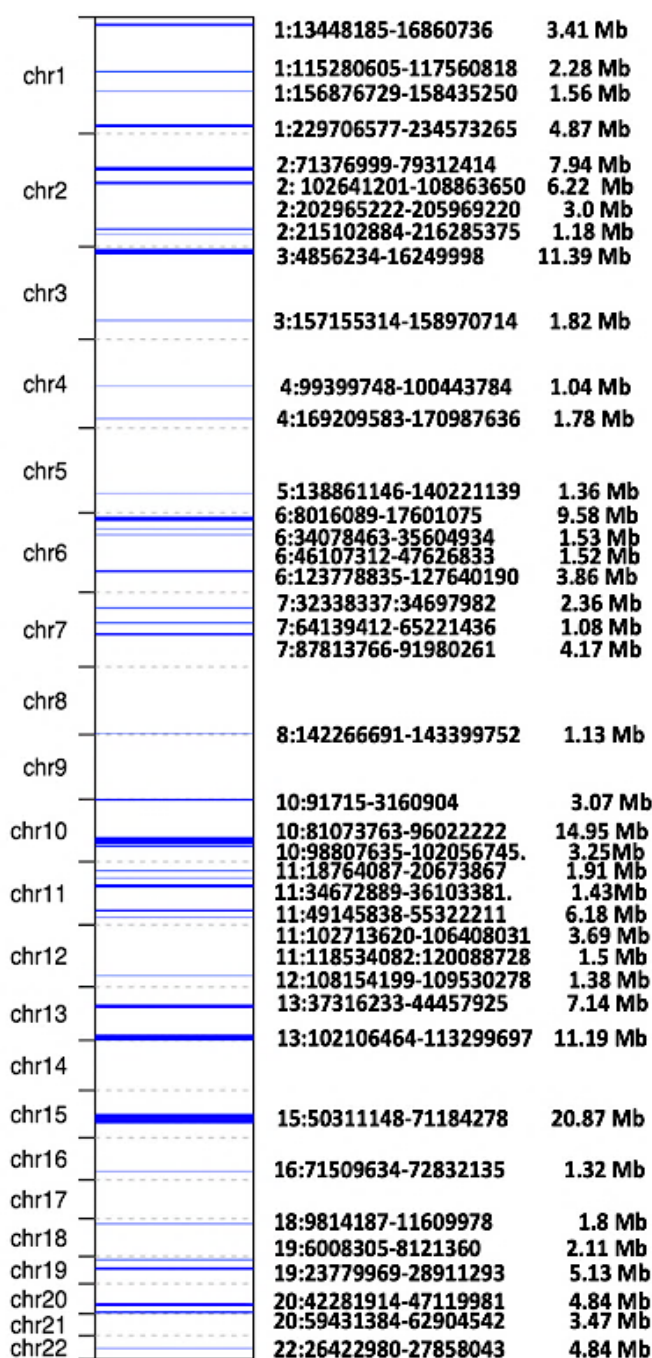
## 5.2 Results

### 5.2.1 Autozygosity mapping in proband F1288 for autosomal recessive gene discovery

The apparent autosomal recessive inheritance of FH and the presence of consanguinity in family F5 suggests that the causative variant is likely to reside in a genomic region shared by both affected siblings that is identical by descent. These autosomal regions will appear as continuous segments of homozygous genotypes known as ROH.

A reanalysis of proband F1288 WES genomic data was performed. Autozygosity mapping was conducted to identify homozygous regions in a single affected proband (F1288) to narrow down candidate variants to those residing in ROH defined by AutoMap (Figure 5.2). Only ROH that exceed 1 Mb in size and have at least 88% homozygous genotypes were retained in the analysis by default settings, to eliminate results that are artefactual or of less confidence.

**Figure 5.2 Autozygosity mapping in proband F1288.** ROH identified across 22 autosomes are highlighted in blue and are annotated with the corresponding chromosomal region (GRCh37) and the relative size in mega bases (Mb). The 40 ROH displayed are scaled based on their relative size. The dotted grey line separates each chromosome on the vertically arranged genome schematic.





The analysis detected 40 ROH in 18/22 autosomes, with none on chromosomes 9, 14, 17 and 21. The largest ROH are chr15:50311148-71184278 (20.87 Mb), chr10:81073763-96022222 (14.95 Mb), chr3:34856234-16249998 (11.39) and chr13:102106464-113299697 (11.19 Mb). The genome sequence of the other affected sibling (F1287) was inaccessible at the time the analysis was performed. Once the genomic data of F1287 was retrieved, it exceeded the computational capacity of AutoMap Webserver (VCF  $\leq$  1 GB) to identify ROH and was therefore not analysed.

## 5.2.2 Characterisation of candidate variants identified

### 5.2.2.1 *In silico* pathogenicity assessment

Filtering variants located within the genomic coordinates of the identified ROH in proband F1288 revealed rare variants in six candidate genes, *LAMP1* NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14), *CHRNA4* NM\_000744.7:c.919G>A; NP\_000735.1:p.(Gly307Ser), *HELZ2* NM\_001037335.2:c.7060G>C; NP\_001032412.2:p.(Val2354Leu), *LOXL4* NM\_032211.7:c.1334G>A; NP\_115587.6:p.(Arg445His), *TADA3* NM\_006354.5:c.737C>A; NP\_006345.1:p.(Ala246Asp), and *MTMR14* NM\_001077525.3:c.1244G>A; NP\_001070993.1:p.(Ser415Asn). Amongst these six genes, only *CHRNA4* has an associated clinical phenotype, autosomal dominant nocturnal frontal lobe epilepsy (MIM: 118504). In addition, *MTMR14* acts as a modifier of autosomal dominant centronuclear myopathy (160150). All six variants had a MAF  $\leq$  0.015% in gnomAD v4. Variant interpretation employed a variety of webserver-based prediction tools to assess the pathogenicity of each variant (Table 5.1). The *LAMP1* variant resulted in a PTC in the last exon of the MANE transcript (ENST00000332556.5), which is likely to escape NMD resulting in a truncated protein. The *CHRNA4* variant was consistently predicted to be deleterious with all the prediction tools utilised. Analysis of variants in *LOXL4*, *TADA3* and *MTMR14* generated conflicting results with no consensus as to whether the variants are pathogenic or benign. On the other hand, the prediction tools suggest that the *HELZ2* variant is tolerated. According to ACMG classification carried out using the Franklin webserver

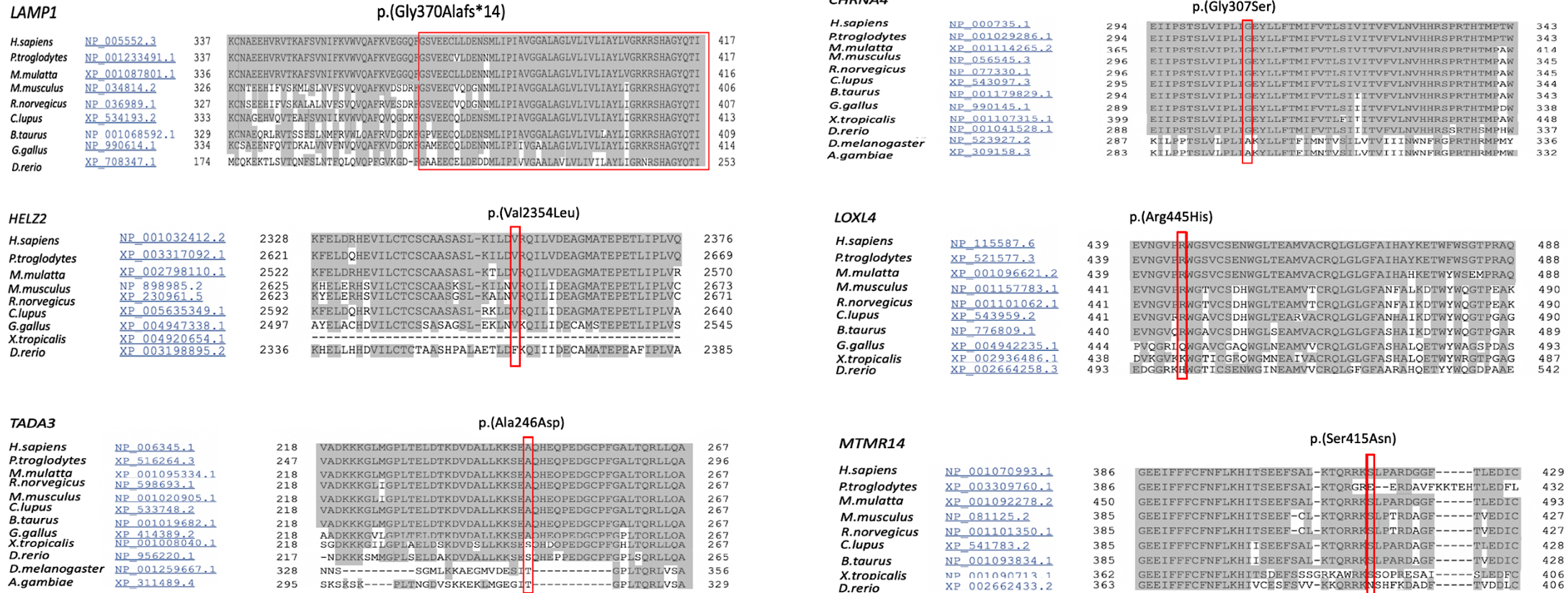
(<https://franklin.genoox.com/clinical-db/home>), all six variants identified in F1288 were considered VUS.

### 5.2.2.2 Multiple sequence alignment

The detection of six variants of uncertain significance required supplementary data to aid in variant interpretation and to validate the computational pathogenicity predictions. The additional evidence was derived from evolutionary conservation of protein sequence provided by multiple sequence alignments using *H.sapiens* (Human) peptide sequence and intermediate species in the evolutionary timeline through to *D.rerio* (Zebra fish) (Figure 5.3). The results of MSA demonstrated that the Gly370 residue of LAMP1 is fully conserved from human to *D.rerio* and that the C-terminal region of the protein beyond this frameshift variant (370-417 aa) is also highly conserved, with the exception of the residues 376, 381 and 405. Amino acid position Gly307 in CHRNA4 also remains highly conserved through to *D.rerio*. HELZ2 and TADA3 variants target amino acids at Val2354 and Ala246 respectively, which are conserved through to *G.gallus* (chicken) suggesting moderate conservation. By contrast, Arg445His (LOXL4) and Ser415Asn (MTRM14) were naturally present in the *D.rerio* peptide sequence, which suggests that these variants are probably tolerated amino acid changes. The MTMR14 amino acid sequence alignment revealed multiple deviations in *P.troglodytes* (chimpanzee), which is atypical. To investigate this, protein sequences of *H.sapiens* and *P.troglodytes* were realigned using the MANE protein transcript (XP\_016795885.2), confirming a sequence misalignment by the NCBI HomoloGene webserver (<https://www.ncbi.nlm.nih.gov/homologene>) due to using an obsolete protein transcript (XP\_003309760.1) (Table 5.2). The Arg445 residue on LOXL4 is conserved in *B.taurus* (cattle) while the Ser415 on MTMR14 is highly conserved from Human to *X.tropicalis* (Frog).

Variant	Transcript	gnomAD	ROH	CADD	Grantham	SIFT	Polyphen-2	PANTHER	Align GVGD	MutationTaster	ACMG
<b>LAMP1</b> c.1109delG p.(Gly370Alafs*14)	NM_005561.4	0	13:102106464- 113299697	23	N/A	N/A	N/A	N/A	N/A	Disease causing	VUS
<b>CHRNA4</b> c.919G>A p.(Gly307Ser)	NM_000744.6	1/86258	20:59431384- 62904542	29.8	56	Damaging	Probably damaging	Probably damaging	Class C55	Disease causing	VUS
<b>HELZ2</b> c.7060G>C p.(Val2354Leu)	NM_001037335.2	13/86258	20:59431384- 62904542	15.32	32	Tolerated	Benign	N/A	Class C25	Polymorphism	VUS
<b>LOXL4</b> c.1334G>A p.(Arg445His)	NM_032211.7	0	10:98807635- 102056745	23.1	29	Tolerated	Possibly damaging	Possibly damaging	Class C25	Disease causing	VUS
<b>TADA3</b> c.737C>A p.(Ala246Asp)	NM_006354.3	0	3:34856234- 16249998	23.8	126	Tolerated	Benign	Possibly damaging	Class C65	Disease causing	VUS
<b>MTMR14</b> c.1244G>A p.(Ser415Asn)	NM_001077525.3	0	3:34856234- 16249998	22.7	46	Tolerated	Benign	Possibly damaging	Class C45	Disease causing	VUS

**Table 5.1 *In silico* pathogenicity assessment of candidate variants.** Transcript refers to NCBI refseq ID of the MANE mRNA transcript for the gene. The ROH column details the chromosome and the ROH coordinates are according to GRCh37. The gnomAD column displays the variant allele count and the total allele count for the South Asian population including the homozygous count. The remaining columns display the results for each of the seven pathogenicity tools used to analyse the variants.



**Figure 5.3 Multiple sequence alignment in candidate FH genes.** Variant site is highlighted by a red box and the dark grey regions indicate conservation. The corresponding protein reference identifier is provided for each ortholog and the peptide sequence is numbered from the first to the last amino acid residue. *Homo sapiens*: humans; *Pan troglodytes*: chimpanzee; *Macaca mulatta*: rhesus macaque; *Canis lupus*: grey wolf; *Bos taurus*: bovine; *Mus musculus*: house mouse; *Rattus norvegicus*: brown rat; *Gallus gallus*: chicken; *Danio rerio*: zebrafish; *Xenopus tropicalis*: western clawed frog; *Drosophila melanogaster*: common fruit fly; *Anopheles gambiae*: mosquito.

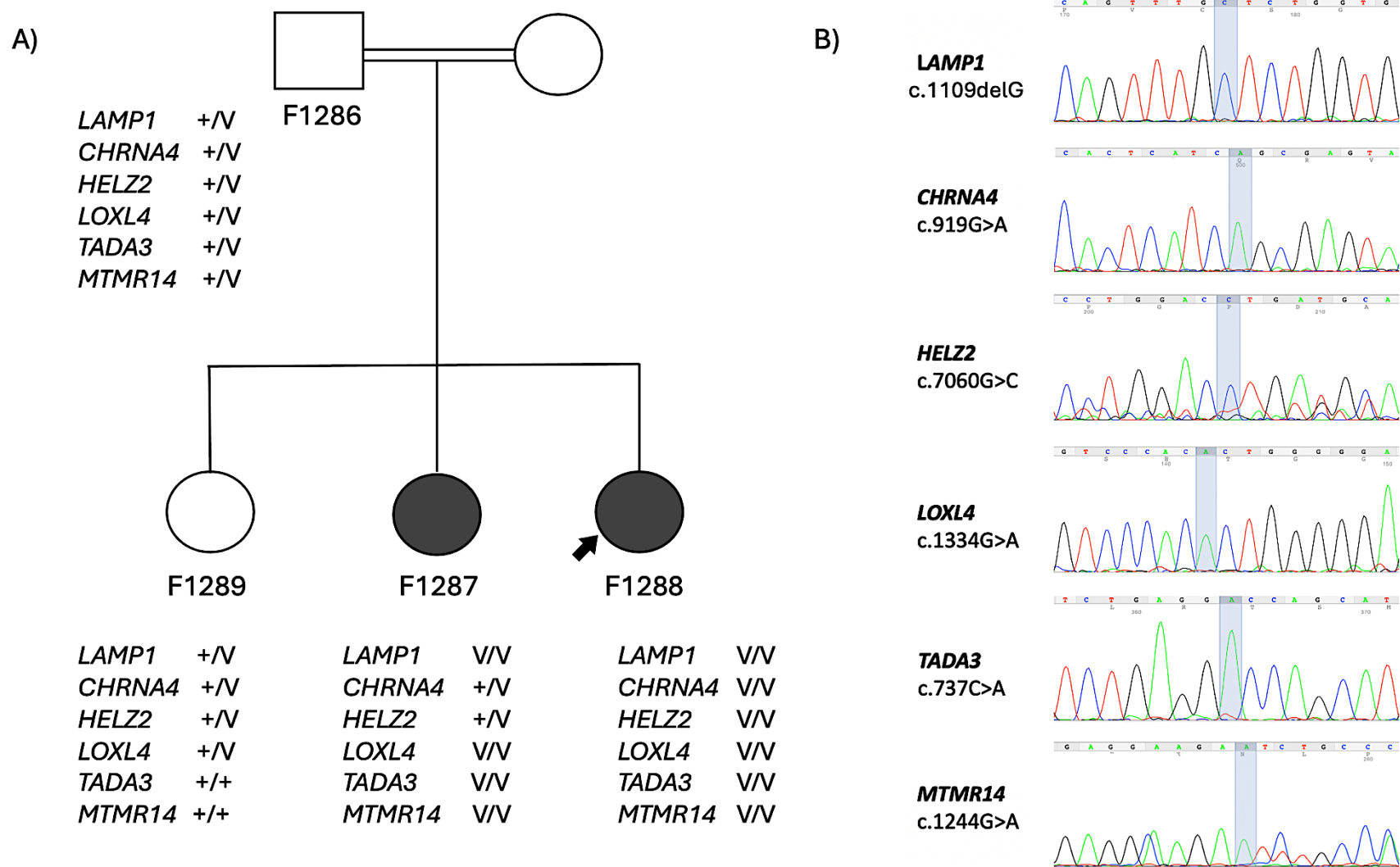
p.(Ser415Asn)

<i>Homo sapiens</i>	<a href="#">NP_001070993.1</a>	401	TSEEFSA LKTQRRKSL	PARDGGFTLEDICMLRRKDRGSTTSLGSDFSLVMESSPGATGSFTYEAVELVPAGAPTQAAWRK	480
<i>Pan troglodytes</i>	<a href="#">XP_016795885.2</a>	401	TSEEFSA LKTQRRKSL	PARDGGFTLEDICMLRRKDRGSTTSLGSDFSLVMESSPGATGSFTYEAVELVPAGAPTQAAWRK	480

**Table 5.2 *MTMR14* corrected sequence alignment.** *MTMR14* peptide sequence realignment showing complete homology between *H.sapiens* and *P.troglodytes* at a region encompassing the variant site.

### 5.2.2.3 Segregation analysis in unsolved FH family F5

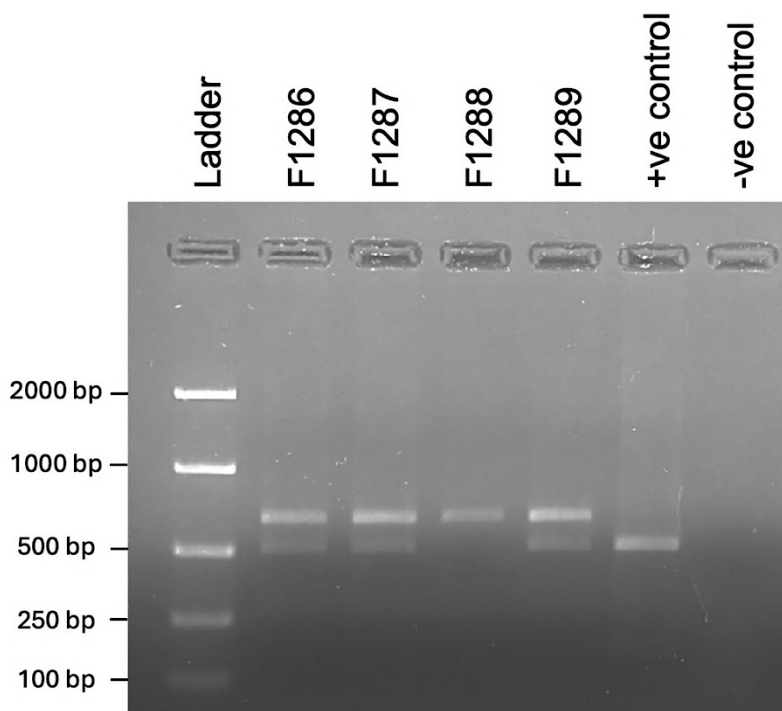
In order to confidently eliminate candidate variants and provide additional supporting evidence for accurate ACMG classification, all six variants detected by WES were validated using Sanger sequencing in the affected proband (F1288) and other family members to exclude those that did not segregate with FH. The four members of family F5 with DNA available for analysis are assigned unique laboratory IDs and were genotyped to assess for the co-segregation of candidate variants in *LAMP1* NM\_005561.4:c.1109delG, *CHRNA4* NM\_000744.7:c.919G>A, *HELZ2* NM\_001037335.2:c.7060G>C, *LOXL4* NM\_032211.7:c.1334G>A, *TADA3* NM\_006354.5:c.737C>A and *MTMR14* NM\_001077525.3:c.1244G>A with the FH phenotype (Figure 5.4). However, there was no DNA available for the unaffected mother, who was therefore excluded from the segregation analysis, though her genotype can sometimes be inferred based on the results of the other family members. The locus for each of the six variants was flanked by the appropriate primers for PCR amplification, with only samples that generate discrete and strong bands being used for Sanger sequencing (Appendix C: Supplementary Figure 5.1). All DNA samples were sequenced in both forward and reverse orientations for an increased confidence in the data generated. The analysis of the electropherograms in the context of the pedigree showed that the candidate variants in *LAMP1*, *LOXL4*, *TADA3* and *MTMR14* all segregated with the disease, while the *CHRNA4* and *HELZ2* variants were found to be heterozygous in the affected sibling (F1287) and thus were excluded. The Sanger sequencing electropherogram of the affected sibling F1287 is also available in the Appendix C: Supplementary Figure 5.2.



**Figure 5.4 Confirmation and segregation of the candidate variants.** A) Pedigree summarising the genotype in relation to the six candidate variants for each family member. B) DNA electropherogram of each candidate variant detected in the affected proband (F1288). The affected nucleotide in the electropherogram trace is highlighted in light blue. “+” indicates the wildtype allele and “V” signifies the variant allele.

### 5.2.3 Confirming *CHRNA4* variant segregation with FH

The finding that *CHRNA4* variant NM\_000744.7:c.919G>A did not segregate with the disease in the affected sibling F1287 has significant implications for variant interpretation. An additional genotyping strategy of RFLP was used to rule out any sample mix up and to confirm the findings of Sanger sequencing. PCR amplification of exon 5 in *CHRNA4* was performed and the amplicons (641 bp) were subjected to an endonuclease digestion using BslI. The palindrome recognition site (5'-CCNNNNNNNGG-3') of the enzyme overlaps the variant site so that the different alleles result in a distinctive pattern on agarose gel electrophoresis (Figure 5.5). The affected proband (F1288) was found to be homozygous for the *CHRNA4* variant NM\_000744.7:c.919G>A as expected. The remaining three family members comprised an asymptomatic parent (F1286), unaffected older sibling (F1289) and an affected sibling (F1287) were found to be heterozygous for *CHRNA4* variant NM\_000744.7:c.919G>A. The results obtained with RFLP were concordant with Sanger sequencing analysis. Furthermore, the *HELZ2* variant, which lies in the same ROH as the *CHRNA4* variant on chromosome chr20:59431384-62904542, shows the same pattern of segregation, further confirming that the *CHRNA4* variant is heterozygous in the affected sibling F1287. As a result, the *CHRNA4* variant was eliminated from further consideration.

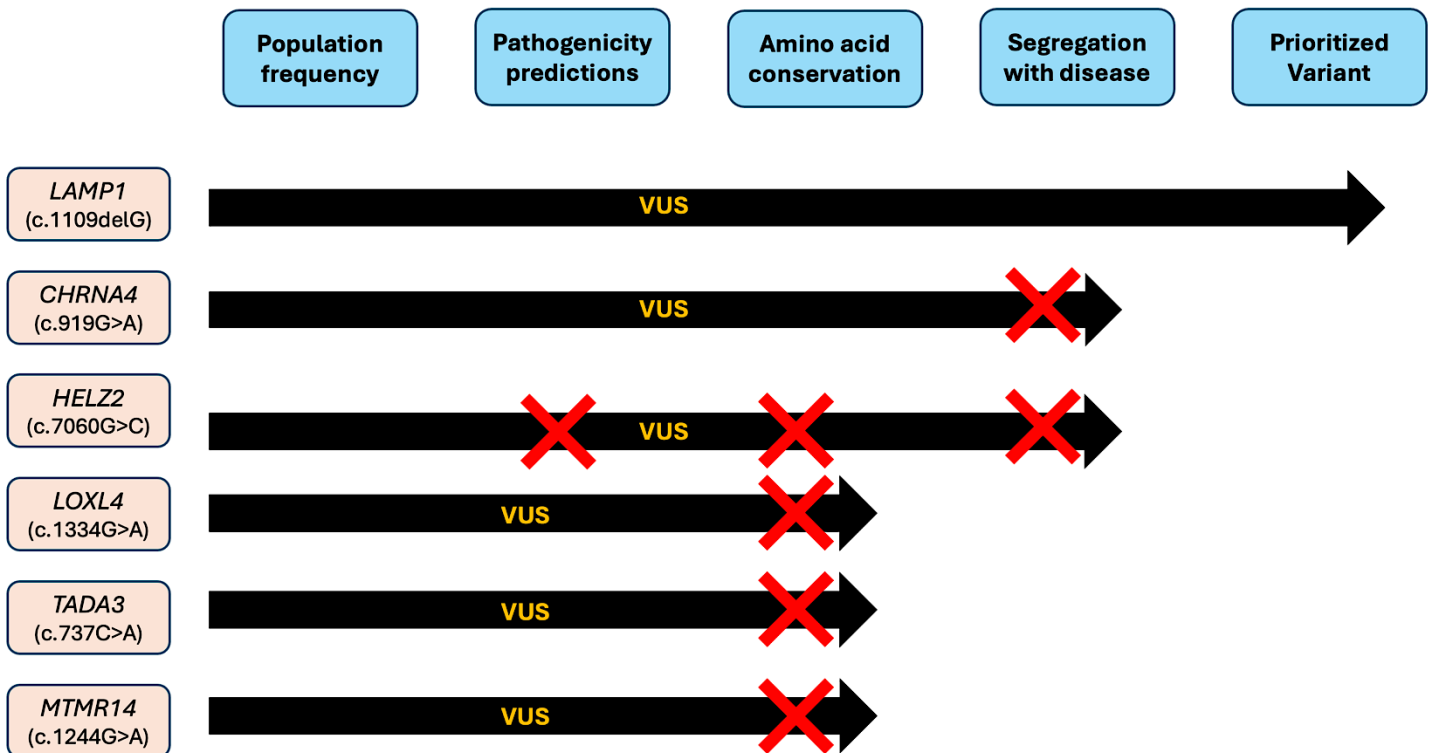


**Figure 5.5 *CHRNA4* restriction fragments pattern in family F5.** The wildtype allele is expected to be cleaved into two fragments of 493 and 148 bp while the mutant product of 641 bp will not be cleaved. The positive control obtained from a healthy unrelated individual is homozygous for the wildtype sequence. Note that the 148 bp smaller band was not resolved using the 2% agarose gel and was masked by the loading buffer (dark shadow on bottom part of the gel).

#### 5.2.4 *LAMP1* as a potential novel FH gene

The collective evidence based on computational variant characterisation, evolutionary sequence alignment and segregation analysis did not support five rare variants in *CHRNA4*, *HELZ2*, *LOXL4*, *TADA3* and *MTMR14* as the cause of FH in family F5 (Figure 5.6). The variant in *HELZ2* NM\_001037335.2:c.7060G>C; NP\_001032412.2:p.(Val2354Leu) was discarded from the analysis based on the predicted benign nature, moderate conservation and the variant not co-segregating with the FH phenotype in the family. The variants in *LOXL4* NM\_032211.7:c.1334G>A; NP\_115587.6:p.(Arg445His) and *MTMR14* NM\_001077525.3:c.1244G>A; NP\_001070993.1:p.(Ser415Asn) were deprioritised based on MSA, since these missense variants affect residues that are less evolutionarily conserved, and indeed the aberrant amino acid substitutions were found in orthologs in other species. Moreover, the pathogenicity assessment of these two variants generated conflicting data. The variant found in *TADA3* NM\_006354.5:c.737C>A; NP\_006345.1:p.(Ala246Asp) was not conserved and gave inconsistent results with the pathogenicity prediction tools. The variant in *CHRNA4* NM\_000744.7:c.919G>A, NP\_000735.1:p.(Gly307Ser) was predicted to be deleterious but did not segregate with disease in affected sibling F1287 of family F5. The remaining variant with the most favourable evidence for pathogenicity is the *LAMP1* variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) which is considered rare, highly deleterious based on altering the reading frame and co-segregates with the disease in all affected members of family F5.



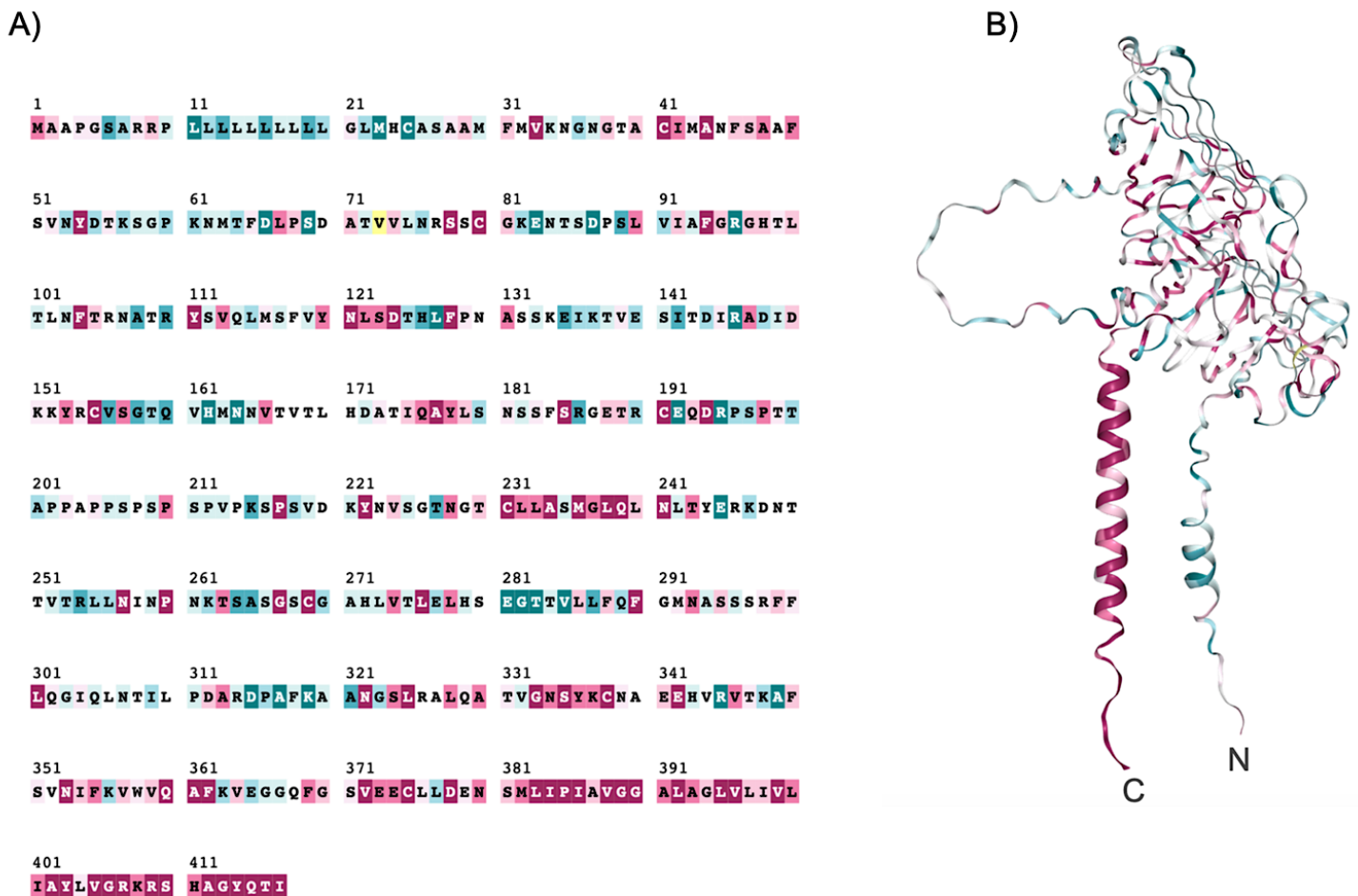


**Figure 5.6 Variant prioritisation strategy.** The variant prioritisation pipeline is depicted as a black arrow advancing through different analysis checkpoints. A red cross symbolises the elimination or deprioritisation of variants due to the corresponding evidence not supporting the variant being causative for FH in family F5.

### 5.2.5 Assessing evolutionary conservation in *LAMP1*

To further investigate the significance of the frameshift variant in *LAMP1*, the overall conservation profile of the entire protein was assessed (Figure 5.7). This analysis aimed to ascertain the extent to which either the entire *LAMP1* protein or particular domains are essential for protein function and thus have been preserved through evolution. Protein sequence alignment by the ConSurf webserver used the *Homo sapiens* *LAMP1* reference sequence (NP\_005552.3) for alignment against 150 orthologs in diverse species. The results of the alignments are depicted as a score of 1-9 for each position on the peptide sequence. This analysis includes diverse species from mammals through to fish for the accurate estimation of amino acid preservation. Interestingly, it was

evident that the residues preceding the affected region (1-369) have varying conservation estimates, but the region downstream, which is predicted to be lost due to the NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) variant, is enriched with conserved residues. This region of 48 residues is composed of 77% (37/48) of highly conserved residues identified by ConSurf.



**Figure 5.7 LAMP1 conservation profile and protein model.** A) The conservation scoring system follows a colour scheme of 1 (turquoise) to 9 (Maroon), with 9 being highly conserved and 1 being variable. The midpoint for neither conserved nor variable is a score of 5 (white). A yellow residue implies a less confident score. B) A protein model of a LAMP1 monomer annotated by the corresponding conservation colour scores generated by the ConSurf webserver (<https://consurf.tau.ac.il/>).

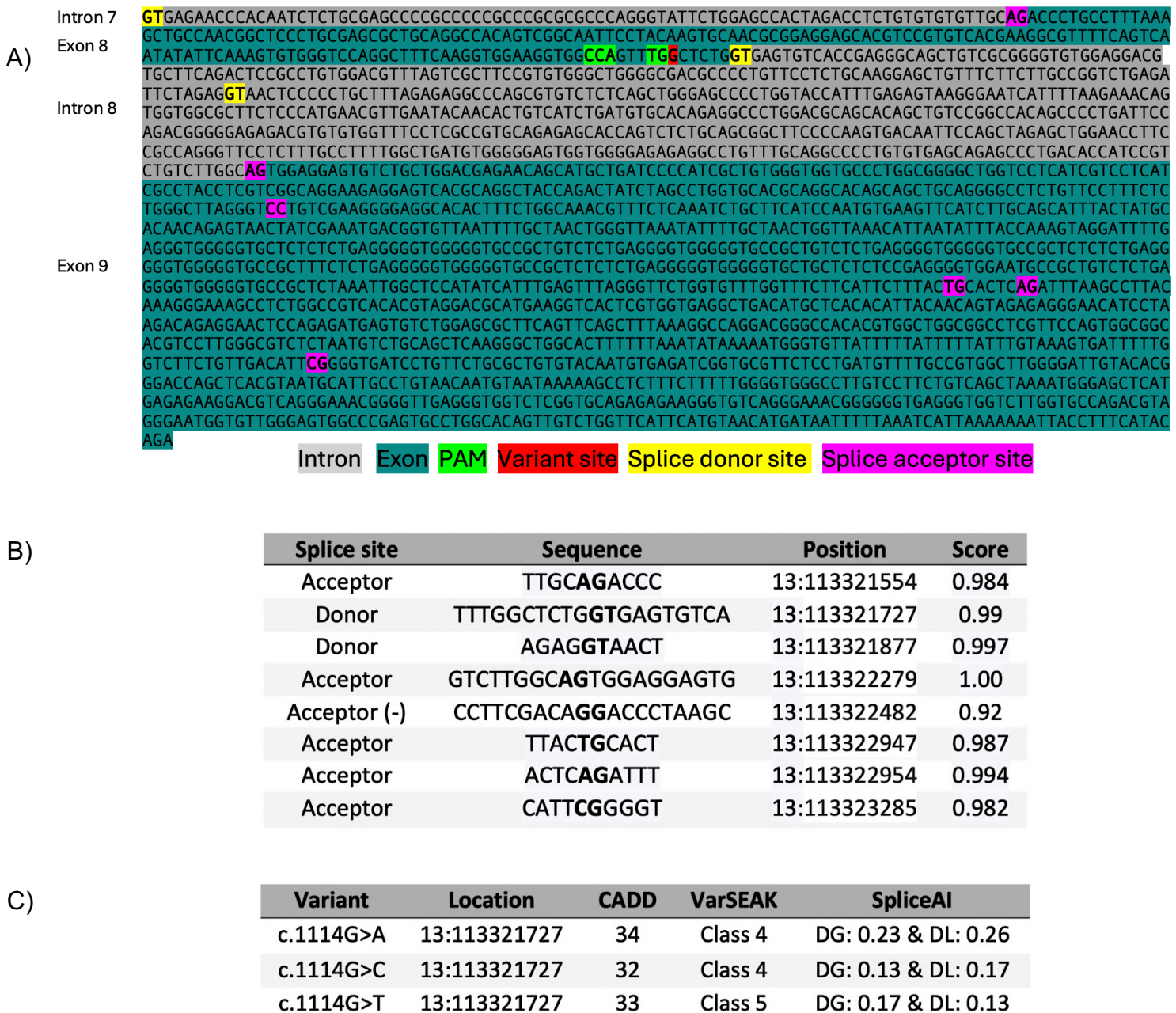
### 5.2.6 Assessing the *LAMP1* variant site for splice sites

In order to determine whether the *LAMP1* NM\_005561.4:c.1109delG induced frameshift may impact splicing, the locus at the variant site and downstream were assessed to determine whether they overlap a canonical or non-canonical splice sites. Canonical splice sites consist of highly conserved dinucleotide sequences immediately adjacent to the exons, together with less tightly conserved consensus sequences extending tens of bp into the introns. These sequences are recognized by the spliceosome and direct the assembly of mRNAs by discarding the introns. The canonical splice site sequence for a splice donor site is “GT” at the 5’ of an intron, while the splice acceptor has an invariant “AG” on the 3’ of an intron. The so-called branch point consists of a consensus adenosine upstream of the splice acceptor site, connected by a polypyrimidine tract. Non-canonical splice sites are those in which the dinucleotide sequence of the splice site is different from that of canonical introns, giving rise to intron boundaries for example with GG, TG or CG dinucleotides at the splice acceptor site.

A 2.2 kb region encompassing intron 7 to exon 9 which also includes the variant site in exon 8 was annotated for potential splice donor and acceptor sites via Spliceator (<https://www.lbgi.fr/spliceator/>) and Netgene2 (<https://services.healthtech.dtu.dk/services/NetGene2-2.42/>) (Figure 5.8A and B). Only predicted splice sites with a score  $\geq 0.9$  were considered of high confidence and genuine. The frameshift induced by c.1109delG, NP\_005552.3:p.(Gly370Alafs\*14) affects the terminal base (guanine) of exon 8 that is 5 bp away from the variant site and is at the boundary with a canonical splice donor site (GT) on intron 7.

The potential for the penultimate base at exon 8 to interfere with a splicing event was assessed using nucleotide substitutions and splice variant analysis tools. The guanine at chr13:113321727 was predicted by varSEAK to have deleterious consequences on splicing if altered by any nucleotide substitution (Figure 5.8C). On the other hand, spliceAI generated a contradictory prediction that implies a loss or gain of a splice donor or acceptor site is less likely to occur. The conflicting evidence makes it unclear whether the altered recognition of the terminal guanine

on exon 8, due to NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) may also result in a potential splicing defect.

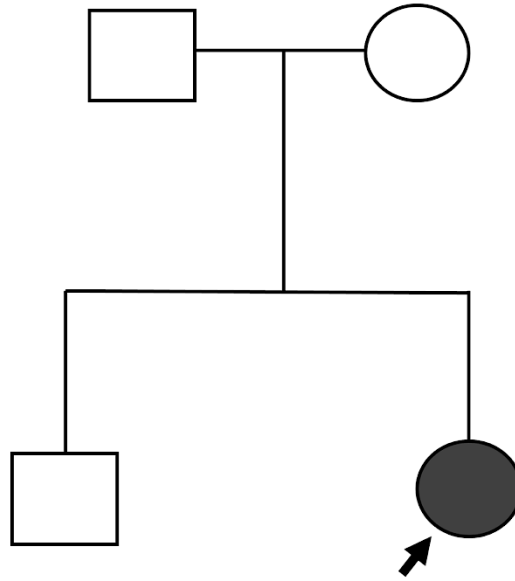


**Figure 5.8 Splice sites in region downstream intron 7 of *LAMP1*.** A) Genomic sequence of *LAMP1* annotated with splice donor and splice acceptor sites according to Spliceator and Netgene2. B) The confidence metric for each splice site prediction is provided in the score column and these range from 0 to 1, with 1 being a predicted splice site of high confidence. C) Gauging the splicing effect of variants at the penultimate nucleotide of exon 8. The pathogenicity assessment tools predict a deleterious splicing effect for the substitution of G at chr13:113321727 by any other nucleotide.

### 5.2.7 Bioinformatic analysis of the 100KGP

One way to test the hypothesis that biallelic *LAMP1* variants can cause FH is to look for additional families with biallelic *LAMP1* pathogenic variants and with an IRD phenotype. The 100KGP provides access to a large genomic dataset derived from 72937 patients with rare diseases, including some diagnosed with FH. The main-programme\_v18 dataset was therefore filtered to capture participants reported with the clinical description of FH using the HPO term of foveal hypoplasia or hypoplasia of the fovea (HP:0007750) and aplasia of the fovea (HP:0011503). The resulting cohort was 68 participants with 6764 variant entries, of which 12 received a genomic diagnosis through the 100KGP. The remaining 57 unsolved patients were screened for biallelic or monoallelic *LAMP1* variants. However, the analysis did not detect any participant with FH and biallelic variants in *LAMP1*.

Subsequent analysis involved a reduced stringency in phenotypic filtering, based on much broader criteria indicative of an IRD instead of just FH. This approach accounts for the potential misdiagnoses in phenotypic reports of FH by non-specialist clinicians and to allow for non-standardised testing at clinics (see section 6.5). The HPO terms selected were based on the most reported HPO keywords in the specific retinal disease category of the 100KGP (Appendix A: section 1.8.3). These are developmental macular and foveal dystrophy, inherited macular dystrophy, Leber congenital amaurosis or early onset retinal dystrophy and rod cone dystrophy. The 72937 participants with rare disease were filtered for HPO terms indicative of an IRD and screened for biallelic variants in *LAMP1*. This strategy identified a single participant possessing a homozygous variant NM\_005561.4:c.299C>T; NP\_005552.3:p.(Arg77Cys) in *LAMP1*, who is reported to have a complex phenotype of developmental delay including nystagmus (Figure 5.9), a common sign associated with FH. The *LAMP1* variant in this proband was predicted to be deleterious by various assessment tools and by the Exomiser variant score. This variant was classified using ACMG as a VUS based on the current supporting evidence of being rare and the predicted pathogenic nature by *In-silico* tools. This led to a clinical collaboration request being made to inquire about the full clinical manifestations of the patient and to ask whether the clinician could confirm that the patient does or does not exhibit FH.

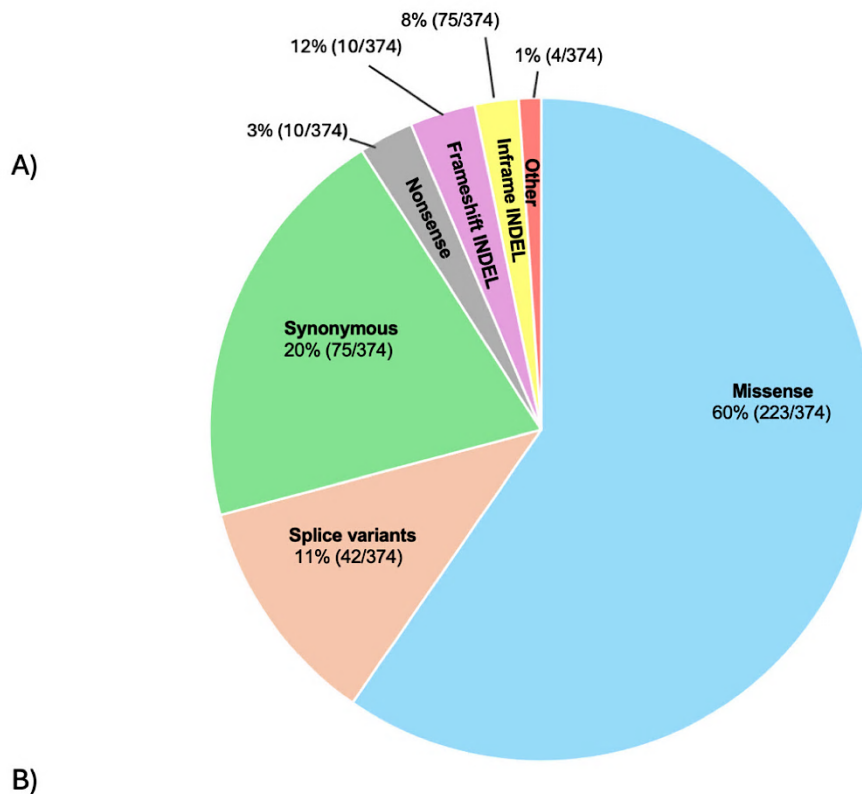


ID	Phenotype	Genotype	Variant	gnomAD	100KGP	Variant score	CADD	SIFT	Polyphen-2	Align GVD	ACMG
I	Nystagmus, ataxia and cerebral atrophy	Homozygous	c.229C>T p.(Arg77Cys)	462/1613994 (3)	2/16636	0.947	18.49	Deleterious	Benign	Class C65	VUS

**Figure 5.9 100KGP *LAMP1* case with potential FH.** Clinical and genomic findings in participant with a homozygous and potentially pathogenic variant in *LAMP1*. Patient displays a wide range neurological symptoms including nystagmus. The variant score relates to pathogenicity and a low allele frequency in gnomAD and the 100KGP.

This participant also harbours other 373 rare variants of which 223 are missense, 42 are splice variants, 75 are synonymous, 10 are nonsense, 12 frameshift variants, 8 are in-frame insertions or deletions and 4 are classified as “others” which includes stop and start loss. One of the missense variants identified, is a homozygous NM\_005561.4:c.1814A>G; NP\_005552.3:p.(Asn605Ser) in *ATXN2*. Variants in *ATXN2* are a known cause of dominantly inherited amyotrophic lateral sclerosis, Parkinson’s disease and spinocerebellar ataxia (Figure 5.10). The variant is predicted to be deleterious by most of the computational tools utilised and has an allele frequency of  $\leq 0.00018$  in both

gnomAD and the 100KGP. This variant in *ATXN2* may therefore explain at least part of the clinical spectrum seen in this patient.



Gene	Genotype	Variant	gnomAD	100KGP	CADD	SIFT	Polyphen-2	Align GVGD	ACMG
<i>ATXN2</i>	Homozygous	c.1814A>G p.(Asn605Ser)	151/1179678	23/126712	18.31	Deleterious	Benign	Class C45	VUS

**Figure 5.10 WGS analysis in a 100KGP proband with *LAMP1* variant and a complex phenotype.** A) Proportion of different classes of variants detected via Exomiser. Other denotes stop and start loss while the splice variants comprise splice donor, acceptor and region variant. B) Pathogenicity assessment of a rare homozygous missense variant in *ATXN2*.

To further investigate the hypothesis that *LAMP1* deficiency has a clinical impact, the rare disease dataset of 72937 genomes was filtered for participants with biallelic variants in *LAMP1*. This analytical approach aims to examine and capture cases harbouring two potentially pathogenic variants irrespective of the reported phenotype. In such cases it is plausible that more severe neurological symptoms might cause ocular manifestations such as FH to be overlooked and not reported.

Alternatively, the presence of likely pathological biallelic *LAMP1* variants in probands with phenotypes unrelated to FH might be considered evidence against the hypothesis that biallelic pathogenic *LAMP1* variants cause FH.

The variant score was used instead of the Exomiser score as its scoring is independent of the participant's phenotype and can be used in the analysis of a potentially new phenotypic association. This analysis led to the detection of 11 probands with varying phenotypes, none of which resemble an IRD. One of these was the proband with a complex phenotype represented above (Figure 5.9). The remaining ten probands had their variants interpreted using an array of pathogenicity assessment tools (Table 5.3). None of these participants proved to have biallelic deleterious variants in *LAMP1*. The ACMG classification for 13 of these variants was VUS and five variants were characterised as likely benign. The 100KGP data therefore did not provide any further evidence for or against the hypothesis that *LAMP1* variants cause a recessively inherited FH or IRD.



ID	Phenotype	Genotype	Variant	gnomAD	100KGP	Variant score	CADD	SIFT	Polyphen-2	Align GVGD	SpliceAI	ACMG
I	Microcephaly	Homozygous	c.163G>A p.(Asp55Asn)	164/1613892	25/126698	0.4196	11.77	Tolerated	Benign	Class C15	-	VUS
II	Progressive spastic quadriplegia and seizures	Homozygous	c.1149G>A p.(Leu383=)	2/1613704	3/16638	0	10.35	-	-	-	0	VUS
III	Progressive external ophthalmoplegia and ptosis	Compound heterozygous	c.74G>A p.(Cys25Tyr)	359/1613912	38/126698	0.2503	13.38	Tolerated	Benign	Class C65	-	VUS
			c.751A>C p.(Thr251Pro)	59/1613916	6/126698	0.2351	21.6	Deleterious	Benign	Class C35	-	VUS
IV	Breast carcinoma	Compound heterozygous	c.314C>T p.(Thr105Met)	143/1614016	7/126698	0.9831	21.8	Tolerated	Possibly damaging	Class C65	-	VUS
			c.374C>A p.(Thr125Lys)	336/1613360 (4)	29/126698	0.7267	14.6	Tolerated	Benign	Class C65	-	Likely benign
V	Parkinsonism, neurological speech impairment, gait ataxia, spasticity and dystonia	Compound heterozygous	c.314C>T p.(Thr105Met)	143/1614016	7/126698	0.9831	21.8	Tolerated	Possibly damaging	Class C65	-	VUS
			c.374C>A p.(Thr125Lys)	336/1613360 (4)	29/126698	0.7267	14.6	Tolerated	Benign	Class C65	-	Likely benign
VI	Pes planus and joint hypermobility	Compound heterozygous	c.644A>G p.(Lys215Arg)	172/1613894 (2)	29/126698	0.5114	8.792	Tolerated	Benign	Class C25	-	Likely benign
			c.893G>A p.(Arg298Gln)	73/1613718	6/126698	0.7239	12.93	Tolerated	Benign	Class C35	-	VUS
VII	Splenomegaly, hepatomegaly and osteopenia	Compound heterozygous	c.357T>C p.(Val119=)	1/1613840	2/126698	0	2.477	-	-	-	0	VUS
			c.628C>T p.(pro210Ser)	132/1613916	6/126698	0.8148	18.06	Tolerated	Probably damaging	Class C65	-	VUS
VIII	Stage 5 chronic kidney disease	Compound heterozygous	c.586C>G p.(pro196Ala)	9/1613130	1/16638	0.6740	3.375	Tolerated	Benign	Class C25	-	VUS
			c.588T>A p.(Pro196=)	10/1613398	1/16638	0	2.166	-	-	-	0	VUS
IX	Limb muscle weakness	Compound heterozygous	c.24G>A p.(Arg8=)	1/1197054	2/126698	0	13.81	-	-	-	0	VUS
			c.422C>G p.(Ser141Cys)	1/1600920	2/126698	0.9960	23.5	Deleterious	Probably damaging	Class C65	-	VUS

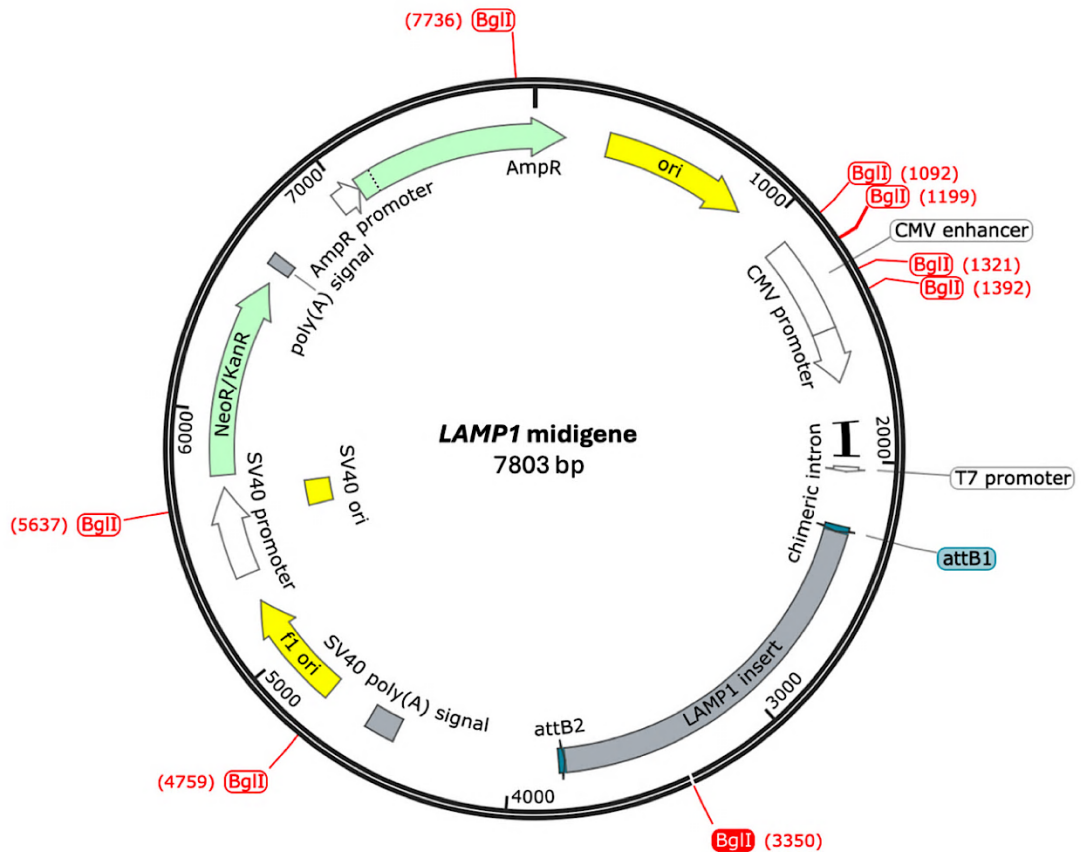
X	Global developmental delay and polymicrogyria	Compound heterozygous	c.750+11_750+49del	2602/1613274 (24)	336/126712	0.2660	25	-	-	-	-	Likely benign
			c.987G>A p.(Gln329=)	835/1614202 (1)	88/126712	0	3.149	-	-	-	0	Likely benign

**Table 5.3 100KGP rare disease cohort with *LAMP1* biallelic variants.** Ten probands with biallelic *LAMP1* variants underwent *In-silico* pathogenicity assessment. All variants were reported in HGVS format using the MANE transcript ENST00000332556.5.

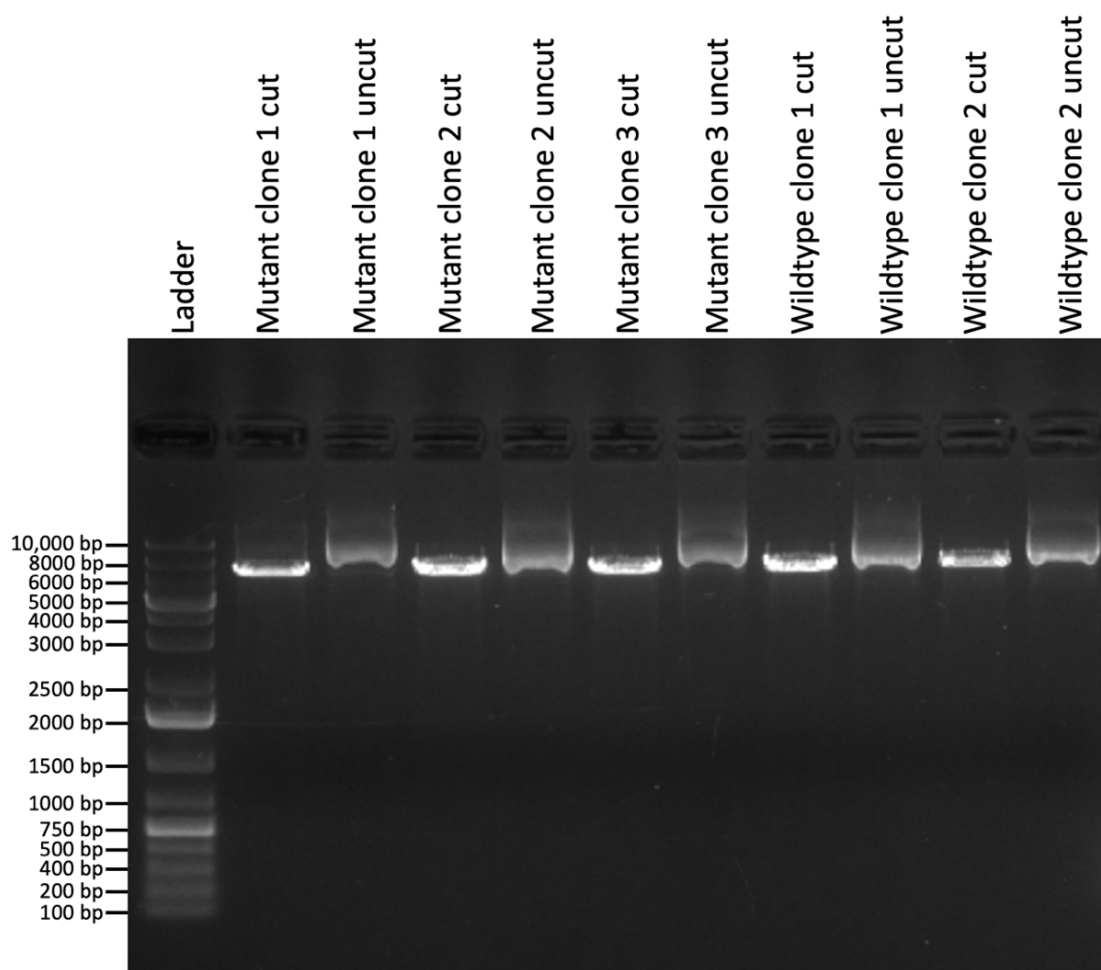
### 5.2.8 Generating *LAMP1* midigene splice vectors

The terminal base at the exon 8 junction with the canonical splice donor site (GT) at intron 8 of *LAMP1* may be affected by the frameshift variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14), which may lead to a potential defect of splicing rather than a protein truncation. The lack of availability of RNA from affected family members for RT-PCR meant that an alternative strategy was required to assess any effect on splicing caused by this *LAMP1* variant. An *In vitro* splice assay was therefore designed using *LAMP1* constructs (midigenes) for the wild-type sequence and the NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14), variant to evaluate the potential impact on splicing in HEK293T cells for a high expression of vectors containing the SV40 origin of replication. These cells are widely available and are routinely maintained in culture in the host institute, making them accessible for use.

A genomic region of 1498 bp spanning introns 6-8 of *LAMP1* was amplified from the affected proband (F1288) and an unaffected healthy control who is homozygous for the wildtype allele. These *LAMP1* amplicons were cloned into intermediate vectors of pDONR221 and pCI-neo-RHO using Gateway Cloning™ to generate *LAMP1* midigenes carrying the genomic region with the desired variant or the wildtype alleles (section 2.4.1). The generated *LAMP1* midigenes contain multiple promoters to support transgene expression in prokaryotic cells for cloning purposes and in eukaryotic cells for functional analysis (Figure 5.11). The vectors constructed for the *LAMP1* splicing assay were cloned into competent *E.coli* cells, isolated and linearised using restriction enzymes with the digested products visualised on an agarose gel (Figure 5.12). The linearised midigenes were of the expected size of 7803 bp for the wildtype and 7802 bp for the mutant *LAMP1* midigene. The undigested midigenes are circular, which caused them to adopt secondary conformations that impeded migration through the agarose gel, resulting in a smear, and the false appearance of an increased size.



**Figure 5.11 *LAMP1* midigene schematic.** Vector map of the midigene plasmid with the *LAMP1* insert, annotated with different features such as antibiotic resistance genes (ampicillin, kanamycin and neomycin) for selective isolation of clones, promoters (SV40: simian virus 40, CMV: cytomegalovirus and bacteriophage T7 promoter) and origin of plasmid replication (SV40 and bacteriophage f1) for plasmid DNA replication and transgene expression. The arrows depict the direction of transcription and the labels on the outside of the plasmid, maps the restriction sites for restriction enzymes that cleaves DNA once allowing for linearisation of the circular plasmid. Recombination sites (attb1 and attb2) flanking the *LAMP1* insert were used in Gateway cloning to induce recombination events with intermediate plasmids for the transfer of the *LAMP1* insert.



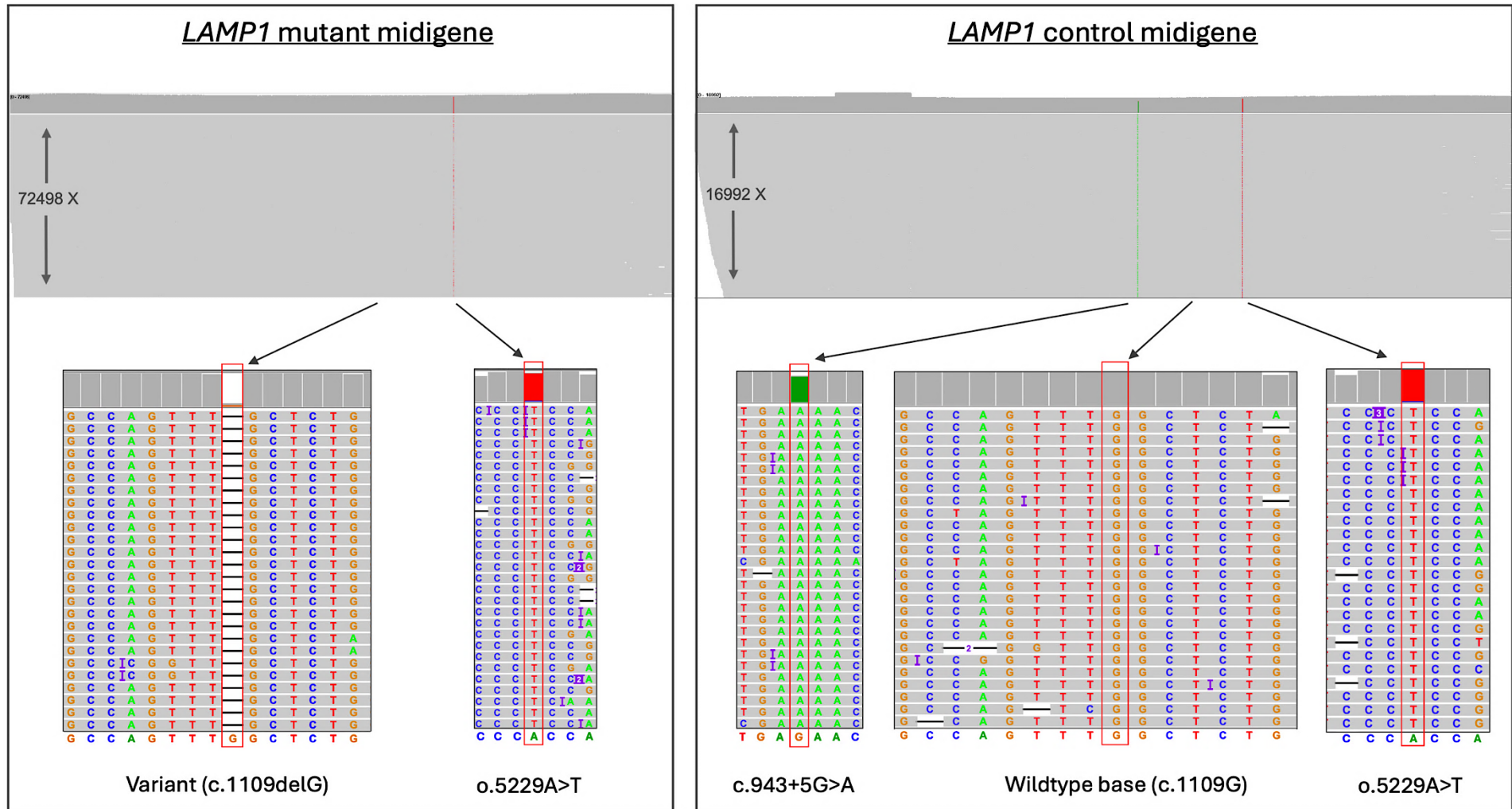
**Figure 5.12. Restriction digestion of *LAMP1*+ midigenes.** *LAMP1* midigenes linearised using BamHI (Promega) and visualised on a 0.7% agarose gel. *LAMP1* midigenes were extracted from five different clones, three being transformants with the mutant *LAMP1* midigene and two with the wildtype *LAMP1* midigene. The digested midigene that is linearised is 7803 bp (7802 bp for the variant).

Long read sequencing was used in preference to conventional Sanger sequencing for the verification of the *LAMP1* wild-type and variant midigenes. This technology supports multiplexing and provides a rapid check of sufficient quality to confirm that the desired constructs have been generated, with no errors introduced by PCR amplification or during the cloning process. The alignment of reads from the midigenes in BAM file generation required the construction of a custom reference sequence, since the sequences of this recombinant *LAMP1* plasmid will only align the *LAMP1* insert to the human reference genome, leaving

the plasmid backbone with prokaryotic components unaligned (Appendix C: Supplementary Figure 5.3).

Performing ONT sequencing generated long reads of 7803 bp (wildtype) and 7802 bp (variant) for a complete sequence representation of each *LAMP1* midigene, with a mean read quality of Q 11.6-11.8. The variant *LAMP1* midigene was loaded on a single Flongle flow cell for uniplex sequencing and it achieved 72,498 X coverage (Figure 5.13). The wildtype *LAMP1* midigene was pooled with other unrelated amplicons for a multiplexed sequencing experiment in which more stringent bioinformatics quality controls were applied in filtering reads, leading to a lower coverage of 16,992 X. The identity of both midigenes was confirmed by the detection of the variant and wildtype alleles at the variant site in 89-97% of the reads. The high error rate in ONT sequencing generated many INDELS and SNV that are present in some but not all the reads (Dohm et al., 2020). The expected identical reads generated from sequencing *LAMP1* midigenes isolated from single and discrete clones made it possible to distinguish artefactual variant calls. Artefacts were filtered out using a variant frequency threshold of greater than or equal to 0.8 (80%) in all the reads and only those that pass this are considered genuine and were retained.

This analysis led to the discovery of a variant c.5229A>T in the plasmid backbone of both the mutant and control *LAMP1* midigenes. This variant does not overlap any functional feature of the midigene such as the simian virus 40 promoter or bacteriophage f1 origin of replication (Table 5.4). In addition, another variant was identified in the *LAMP1* insert in 95% of the reads, that was exclusive to the control *LAMP1* midigene. This intronic variant, NM\_005561.4:c.943+5G>A, is rare as it is absent from the gnomAD database and has been predicted to have deleterious consequences on splicing. Sanger sequencing was not performed in the control genomic DNA nor the *LAMP1* midigene to confirm whether the variants have been introduced through the cloning process or whether the control individual carries the homozygous *LAMP1* variant NM\_005561.4:c.943+5G>A.



**Figure 5.13 Validation of *LAMP1* midigenes.** ONT sequencing of *LAMP1* mutant and wildtype midigenes. BAM file shows the long reads aligned across the entire *LAMP1* midigene sequence, with a coverage of 72498X for the mutant midigene and 16992 X for the wildtype midigene. Small window displays the nucleotide at the variant site and confirms the identity of each midigene. The corresponding reference sequence at each position is displayed at the bottom of the window. Variation to the reference genome is highlighted by colours corresponding to the variant nucleotide. A: green, G: orange, C: blue and T: red.

Variant	Read Count	Coordinate	Feature	gnomAD	CADD	SpliceAI	VarseaK	UTRannotator
c.943+5G>A	9595/10108 (W)	4237	<i>LAMP1</i> intron 7	0	18.37	0.26	Class 5	none
o.5229A>T	12065/12396 (W) 58764/67926 (M)	5229	backbone	-	-	-	-	-

**Table 5.4 Variants detected in *LAMP1* midigenes.** Details of two variants, comprise o.5229A>T variant in the plasmid backbone and the *LAMP1* intronic variant NM\_005561.4:c.943+5G>A. The read count for each variant in the wildtype (W) or mutant (M) midigene is presented.

### 5.2.9 Generating an isogenic control of the *LAMP1* midigene

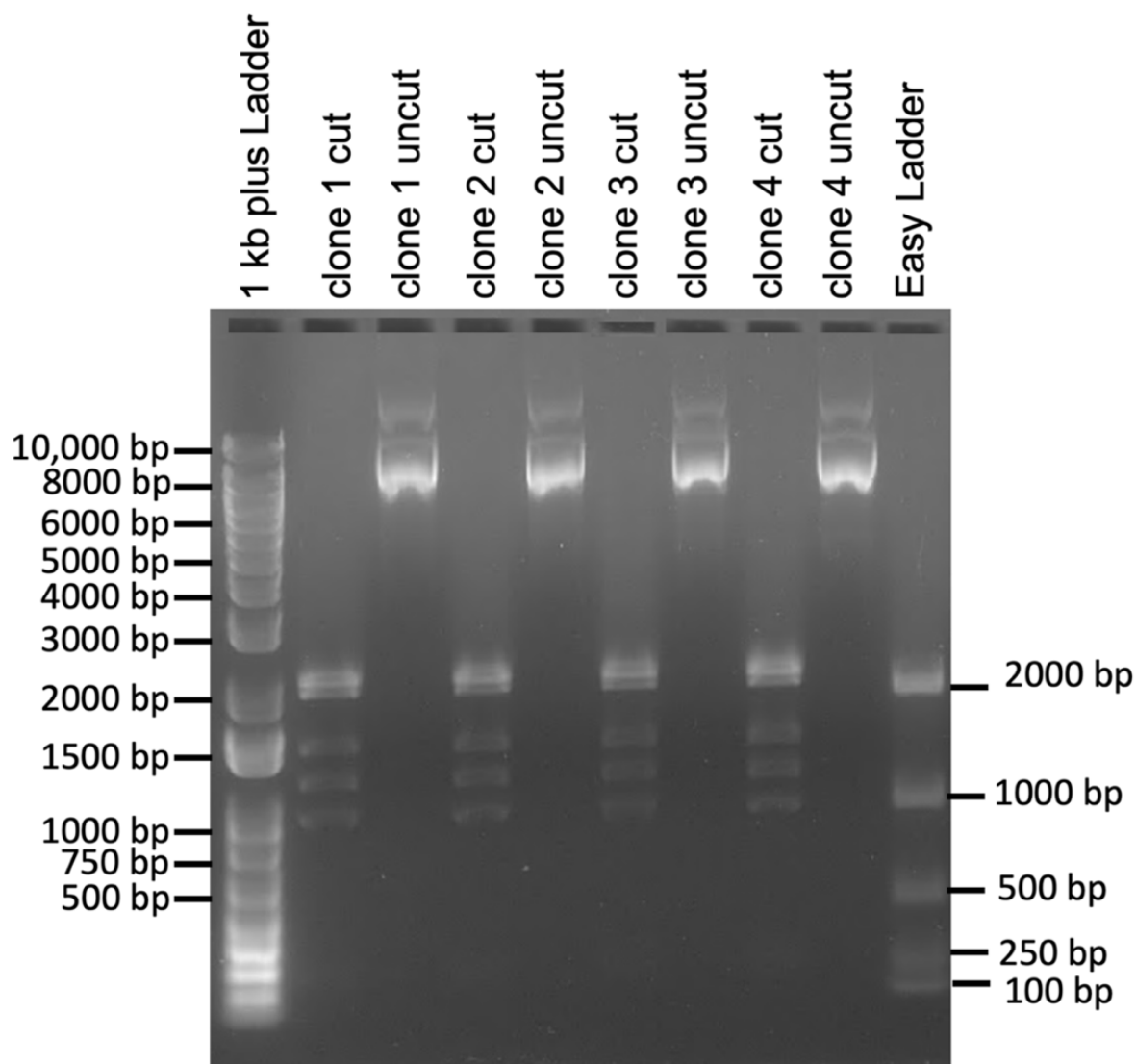
The control *LAMP1* midigene vector carries the wildtype base (guanine) instead of the variant (NM\_005561.4:c.1109delG), but it also carries an additional *LAMP1* variant which may influence splicing, NM\_005561.4:c.943+5G>A. The *LAMP1* wildtype midigene is therefore not entirely suitable as a control for the *In vitro* splice assay. This prompted a revision of the strategy to generate an isogenic control midigene. Instead of using the wildtype *LAMP1* construct for comparison against the *LAMP1* variant construct. Oligonucleotide based SDM was performed to revert the c.1109delG variant in the *LAMP1* variant midigene back to the wildtype base.

SDM was performed on the *LAMP1* variant midigene as per the protocol described in section 2.4.1. DNA was harvested from the transformed *E.coli* colonies using the miniprep method (Qiagen) and was genotyped using RFLP to confirm the presence of the wildtype base at position 1109 of the CDS. The strategy of the RFLP detection is based on the BglII recognition sequence (5'-GCCNNNNNGGC-3') overlapping the *LAMP1* variant site (Appendix C: Supplementary Figure 5.4). The presence of the NM\_005561.4:c.1109delG variant will abolish the restriction site, resulting in a distinctive banding pattern based on the genotype (Figure 5.14). Four clones carrying the midigene WT SDM vector were digested by BglII and have generated the expected banding pattern for the wildtype sequence on

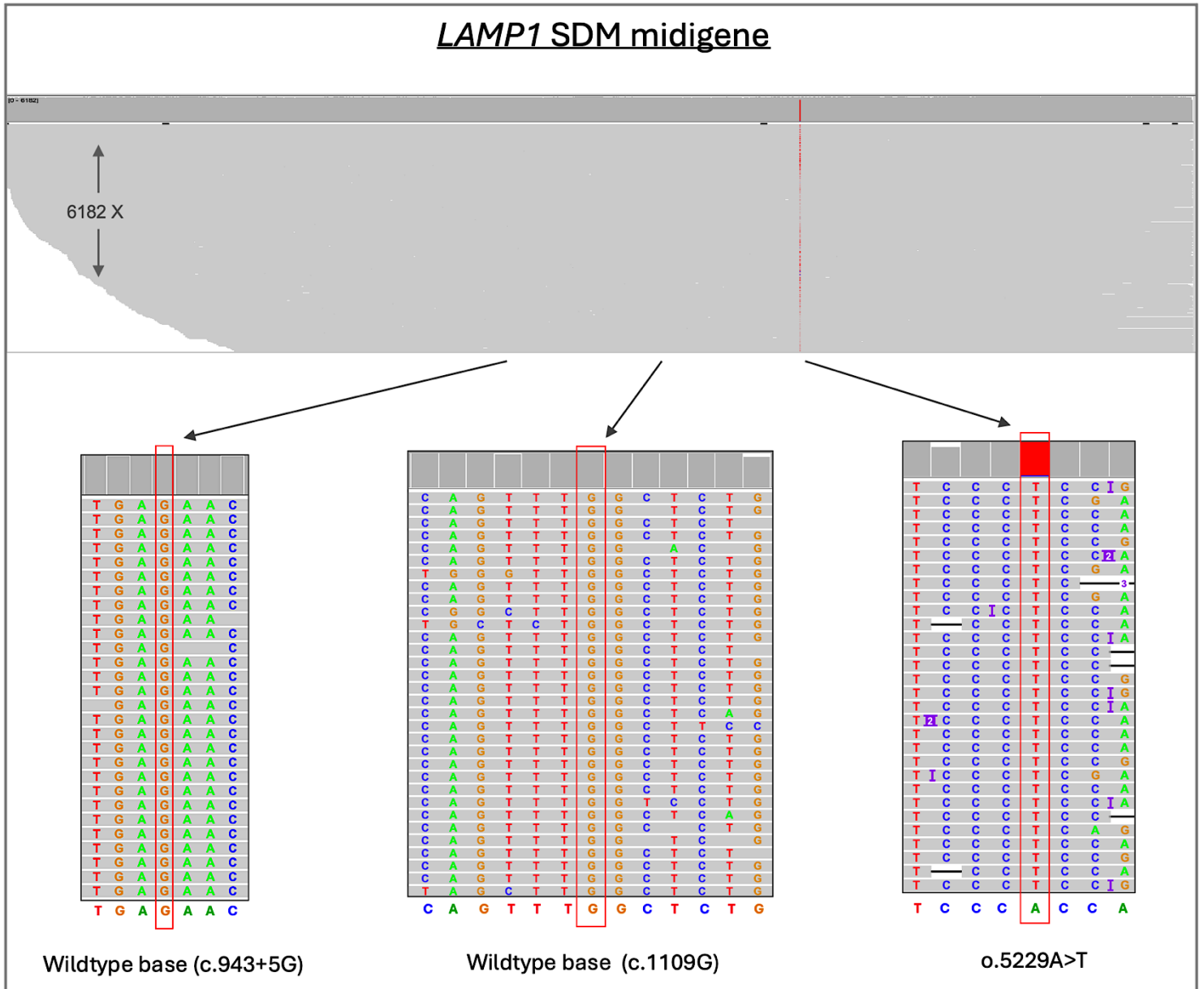


agarose gel electrophoresis, with the 2099, 1958, 1409, 1158 and 878 bp fragments visible. The remaining 122, 107 and 71 bp bands are not visible due to the low molecular weight DNA bands incorporating less of the Midori green dye in the DNA backbone or the greater diffusing capacity for the smaller products, making them harder to visualise. In contrast the smaller bands (100-200 bp) of the ladder are visible on the 0.7% agarose gel albeit at a low resolution.

The RFLP results were subsequently verified using ONT sequencing to confirm incorporation of the wildtype base guanine at the variant site, and to ensure no unintended edits were present in the newly generated control *LAMP1* WT SDM midigene vector (Figure 5.15). A single clone (clone 1) was sequenced, and the coverage obtained from the long reads was 6182 X with a mean quality of Q12. The wildtype base (guanine) was successfully introduced at the variant site and was present in 97% of the reads. As an additional quality control measure to check for sample mix up during sequencing library preparation, the variants o.5229A>T and NM\_005561.4:c.943+5G>A were used as identifiers to confirm the parental origin of the *LAMP1* WT SDM midigene. The presence of only the o.5229A>T is exclusive to the parental mutant midigene that carries the NM\_005561.4:c.1109delG in *LAMP1* (Table 5.4). The NM\_005561.4:c.943+5G>A was absent from all the reads and the o.5229A>T plasmid variant was detected in 95% of the reads of the *LAMP1* SDM midigene. This experiment therefore confirmed the successful generation of an isogenic control *LAMP1* SDM midigene for use in the *In vitro* splicing assay. However, due to time constraints it was not possible to transfect the control and variant midigene vectors into HEK293T cells to determine whether the *LAMP1* variant identified in family F5 resulted in a splicing defect rather than a frameshift that causes a protein truncation.



**Figure 5.14 Genotyping SDM *LAMP1* midigenes using RFLP.** A 0.7% agarose gel image displaying the digested and undigested SDM *LAMP1* midigenes obtained from four different clones. Digestion of the SDM midigenes by BglI resulted in five visible bands (2099, 1958, 1409, 1159 and 878 bp) and a very faint band of low resolution around the 100 bp sizing ladder. A dark shadow appears over the middle of the gel which is the migrating loading dye, and this masks the fluorescent signal of the bands. The undigested plasmid forms supercoiled structures impeding its migration through the gel, resulting in two separate and irregular bands.



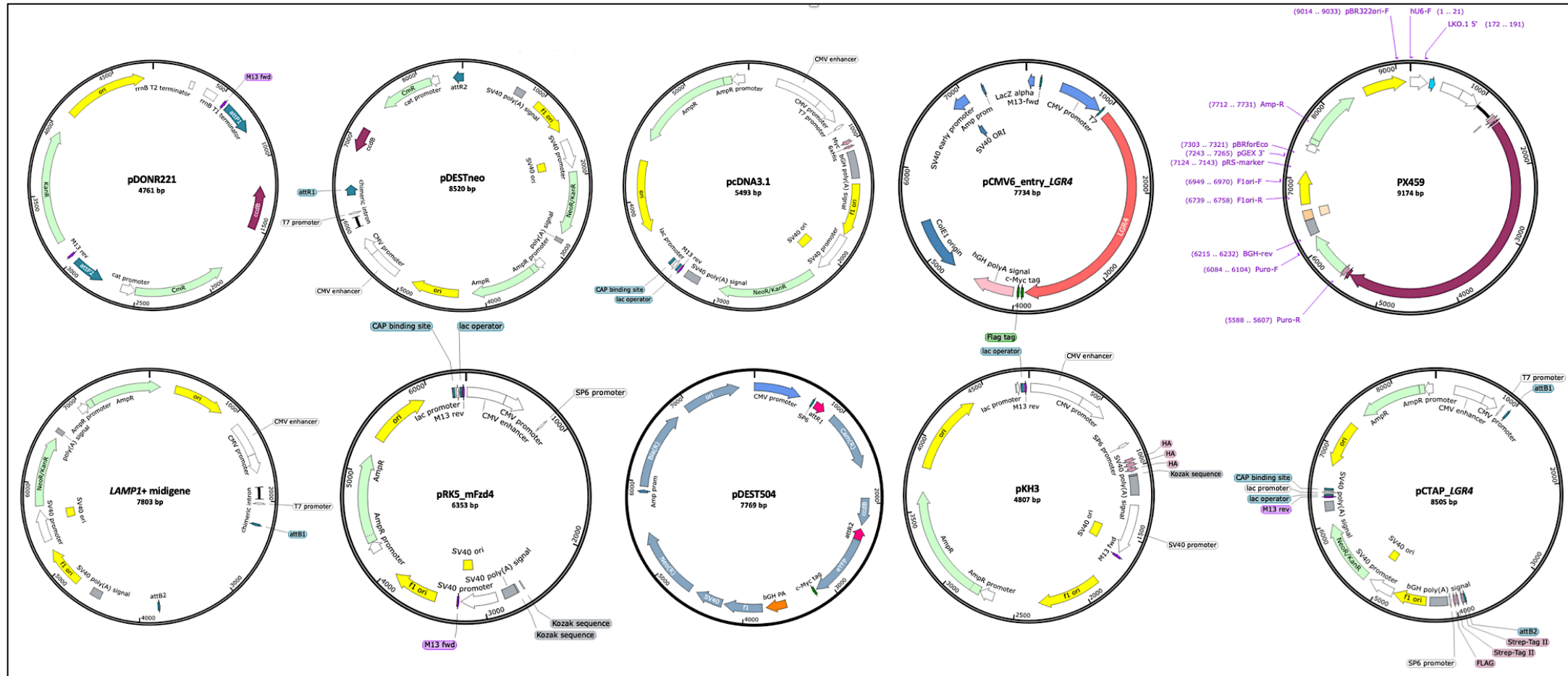
**Figure 5.15** ONT sequencing of the *LAMP1* WT SDM midigene vector generated by SDM. Verification of a single SDM product obtained from clone 1 (Figure 5.12). The depth of coverage for the long reads spanning the entire *LAMP1* midigene sequence is 6182 X. The guanine wildtype base is present at the variant site in 5741/5895 of the reads at the position coinciding with NM\_005561.4:c.1109 in the *LAMP1* cloned sequence. The variant used as a genetic marker, o.5229A>T, was detected with read count of 5262/5549 and the NM\_005561.4:c.943+5G>A *LAMP1* variant was not detected.

### 5.2.10 Multiplexed assembly of recombinant plasmids using barcode-free ONT sequencing

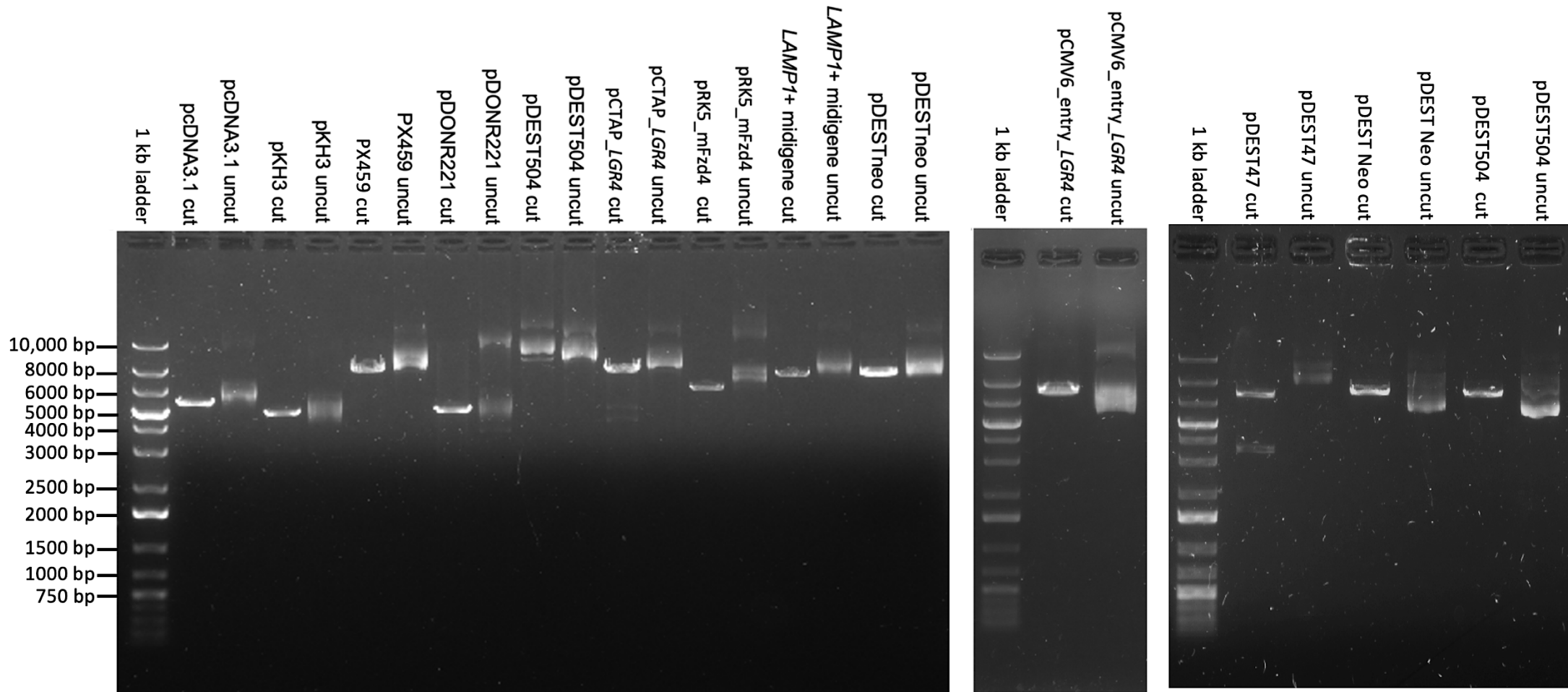
The successful application of long read sequencing in validating *LAMP1* midigenes and the associated SDM construct has demonstrated the feasibility of this technique in the co-sequencing of additional vectors, given that they are dissimilar in sequence. The ONT sequencing of *LAMP1* midigene constructs (Figure 5.13 and 5.15) and other PCR amplicons to characterise structural variants (section 4.2.2.3) were of low scale. These sequencing trials provided the bioinformatic foundation for multiplexing using more samples, for up to 10 targets for sequencing.

Using ONT based sequencing rather than conventional Sanger sequencing in plasmid assembly offered scalability, cost efficiency, a less laborious laboratory protocol and a more comprehensive analysis of the entire plasmid instead of traditional sequencing of just the insert. To test the efficacy of this potentially novel technical innovation, 10 plasmids were selected for multiplexed sequencing using the Flongle R9.4.1 flow cell. These plasmids were the *LAMP1* wildtype midigene (*LAMP1+* midigene), Gateway cloning vectors (pDONR221 and pDESTneo), Cas9 vector (PX459), other expression vectors (pDEST47, pcDNA 3.1, pRK5\_mFzd4 and pKH3), and additional plasmids derived from colleagues working on other projects (pCMV6\_entry\_LGR4, pCTAP\_LGR4) (Figure 5.16). These 10 plasmids were linearised using selected restriction enzymes, XbaI, EcoRI, BamHI and XhoI that cleaves the target circular DNA once (Figure 5.17). Each digested plasmid was quantified using Qubit fluorometry and the readings were converted to moles to support pooling in equimolar ratios. 15 femtomole (fmol) of each plasmid was pooled for equal representation of reads belonging to the 10 different plasmids sequenced. The read alignment in the BAM file for the different recombinant plasmids used reference vector sequences obtained from online repositories for the commercially available plasmids, or a custom reference genome was constructed as in the case of the *LAMP1* midigene sequence (Appendix C: Supplementary Figure 5.3).

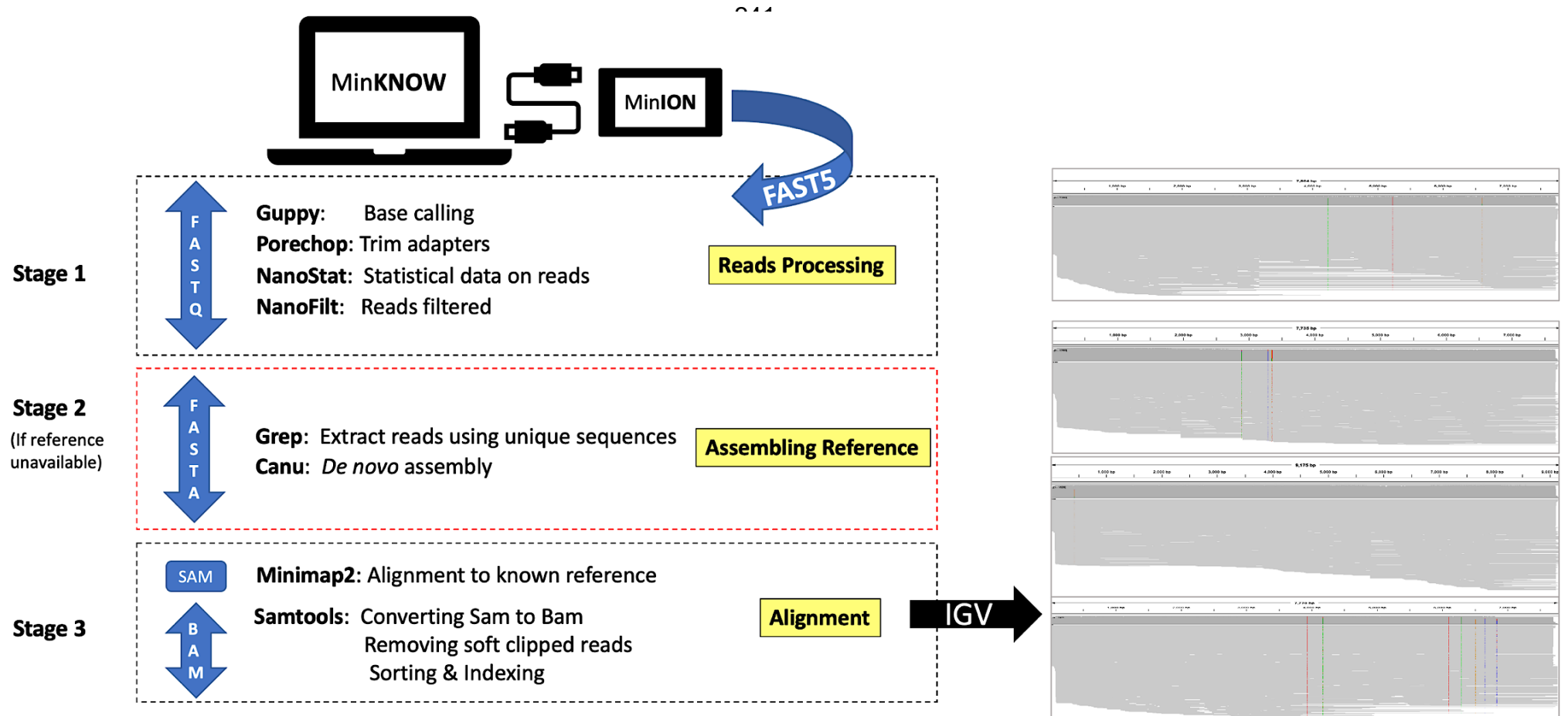
A novel and robust bioinformatic pipeline was devised to discriminate between reads of different plasmid origin, despite the significant sequence homology between the plasmids used and the lack of indexes to tag the different libraries (Figure 5.18). This pipeline used two plasmid assembly methods. A reference assembly used a known plasmid sequence for a reference guided assembly of plasmid sequence. The alternative method used was a *de novo* assembly to construct a plasmid sequence with minimal prior knowledge of the plasmid sequence. The script containing the list of commands for each assembly method to discriminate between reads of different plasmid origin is provided in the Appendix B: section 8.2.7.



**Figure 5.16 Plasmids included in multiplexed nanopore sequencing.** 10 plasmids annotated with different features including antibiotic resistance genes, promoters and selection markers. All plasmid maps were generated using SnapGene viewer v6.2.1. PX459: pSpCas9(BB)-2A-Puro V2.0 and pDESTneo: pCI-neo-RHO-ex3-5).



**Figure 5.17. Linearisation of plasmids using restriction enzymes.** Plasmids were linearised using restriction endonucleases specific to each plasmid that cleave once. pDEST504 required additional attempts to linearise the plasmid (3<sup>rd</sup> gel image). Only 10 plasmids were selected for ONT multiplexed sequencing, these are: pcDNA3.1: 5493 bp, PKH3: 4807 bp, PX459: 9174 bp, pDONR221: 4761 bp, pcTAP\_LG4R: 8505 bp, pRK5\_mFzd4: 6353 bp, *LAMP1* + midigene: 7803 bp, pDESTneo: 8520, pCMV6\_entry\_LG4R: 7734 bp and pDEST504: 7769 bp. The restriction enzymes used were XbaI (pcDNA 3.1, pKH3 and PX459), EcoRI (pDONR221 and pCTAP\_LGR4), BamHI (pcMV6\_entry LG4R, pRK5\_mFzd4 and *LAMP1* + midigene) and XhoI (pDESTneo and pDEST504).

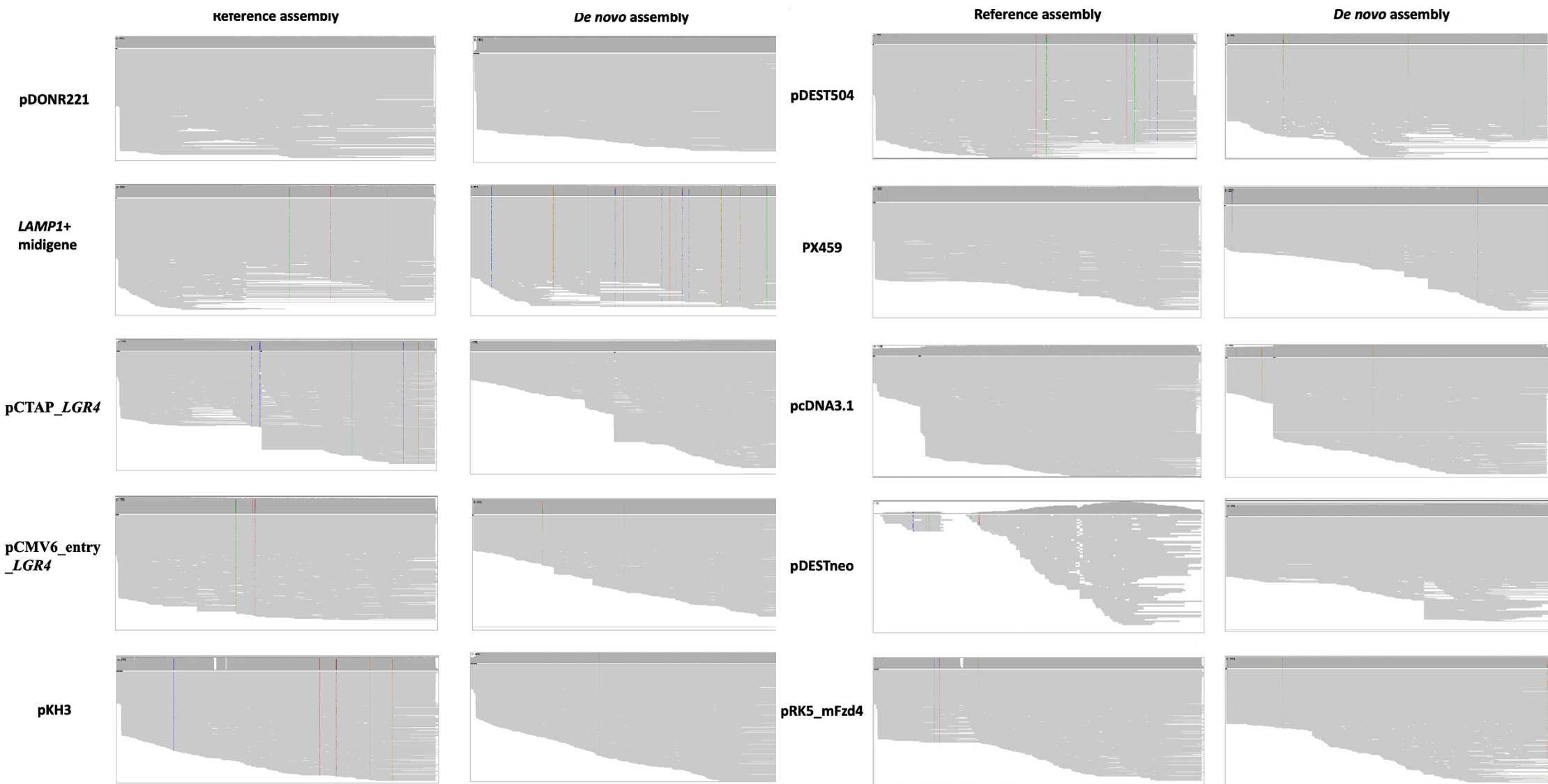


**Figure 5.18. Bioinformatics pipeline used to demultiplex plasmid reads.** Fast5 output generated by the MinION instrument undergoes a series of bioinformatics analyses for processing with different intermediate file formats to discriminate between reads of different origin. Different file formats corresponding to each computational stage are depicted within blue arrows or boxes. The computational workflow is divided into 3 stages comprising base calling and read processing to retain high quality long reads (Stage 1), *de novo* assembly of reference sequence if it is unavailable, which relies on consensus from reads that are demultiplexed (Stage 2), and finally mapping against the reference genome to identify genomic discrepancies in the alignment (stage 3). The mapped reads in the BAM file are visualised using IGV and the window displays demultiplexed and aligned reads spanning the entire length of 4 different plasmids.



Using the reference assembly, the BAM files generated contained read alignments of high coverage (~3709) for each plasmid (Figure 5.19). A significant proportion of these reads spanned the entire length of the plasmid to provide full sequence representation of the vector. To validate the multiplexed sequencing, the mean read quality, mean read length and average identity of different reads within the BAM file were interrogated (Table 5.5). The sequencing metrics for the 10 multiplexed plasmids is a mean quality score exceeding Q12 and an average identity of 93.03%. To test the workflow limits in plasmid discrimination, a single plasmid (pRK5\_mFzd4) of high sequence similarity to another (pKH3) was used as an internal control in this multiplexing experiment of 10 plasmids. The pRK5\_mFzd4 and pKH3 plasmids are 96% homologous and these were deliberately sequenced to test the sensitivity of the workflow. The pRK5\_mFzd4 plasmid has been previously sequenced on a Flongle R9.4.1 in a multiplexing trial of 5 plasmids (Data not displayed) and the sequence of both pRK5\_mFzd4 and pKH3 have been assembled correctly. However, in this experiment of multiplexed sequencing of 10 plasmids, both the pRK5\_mFzd4 and pKH3 plasmids have not been assembled correctly as they contain artefactual deletions that are consistent in 99% (3665/3707) of reads.

An outlier plasmid, pDESTneo, had the lowest coverage of 72 X and the mean read length was shorter than the expected size by 6713 bp. This suggested the possibility of an incorrect reference sequence used in the read alignment which did not resemble the actual plasmid. *De novo* assembly was performed using Canu to assemble the plasmid sequence without the need of a reference sequence. *De novo* assembly used a minimal overlap length of 100 bp. The resultant pDESTneo assembly was of 7937 bp and this was used as the representative reference for read alignment, which improved the overall sequencing metrics. The mean read length had increased to 7099.9 bp and the coverage increased to 988 X (Table 5.5). The size of the *de novo* assembled pDESTneo reference sequence of 7937 bp, matches the size of the linearised pDESTneo product, which migrates close to the 8000 bp marker on a 0.7% agarose gel (Figure 5.17).



**Figure 5.19. Plasmid assembly using multiplexed nanopore sequencing.** Long read alignments for each plasmid using different assembly methods of reference based and *de novo* assembly. Allele fraction threshold was set to 0.8 to only display variants consistent in 80% of the reads. Mismatches detected are displayed in colours corresponding to the adenine (green), thymine (red), guanine (orange) and cytosine (blue). INDELS were hidden by setting the labelling threshold for INDEL markers to 1000 bp.

Plasmid	Reference assembly						<i>De novo</i> assembly					
	Size (bp)	Coverage	Read length N50 (bp)	Mean read length (bp)	Mean read quality (Q)	Average Identity (%)	Size (bp)	Coverage	Read length N50 (bp)	Mean read length (bp)	Mean read quality (Q)	Average Identity (%)
PX459	9174	1626	9016	8116.8	12.0	93.1	9186	2207	9079	8445.6	12.2	93.1
pDONR221	4761	1863	4654	4512	12.6	94.0	4761	1561	4653	4504.1	12.5	93.8
pcDNA3.1	5493	1105	5357	4421.6	12.2	93.3	5504	836	5429	4495.4	12.3	93.4
pCMV6_entry_ <i>LGR4</i>	7734	1759	7589	6508.5	12.2	93.3	7742	2598	7668	7015.3	12.3	93.4
pCTAP_ <i>LGR4</i>	8505	1700	8341	6144	12.3	93.3	8503	2755	8420	6964.3	12.4	93.6
pDEST504	7770	727	7633	6480.5	12.4	93.4	7783	995	7703	6949.8	12.5	93.5
<i>LAMP1+</i> midigene	7803	1366	7664	7044	12.1	93.1	7806	971	7733	7044.5	12.3	93.2
pKH3	4807	3709	4655	4384.2	12.2	91.9	4761	4290	4722	4501.8	12.3	93.2
pRK5_mFzd4	6353	2011	6164	5466.3	12.1	92.5	6321	1791	6236	5677.3	12.3	93.4
pDESTneo	8520	72	2479	1807.1	12.3	92.4	7937	988	7887	7099.9	12.3	93.5

**Table 5.5 Statistical summary of multiplexed sequencing of 10 plasmids.** Read statistics of nanopore sequencing using known reference sequences and *de novo* assembly are tabulated. The reference assembly generated a coverage >1000 reads for each plasmid except for pDEST504 (727) and pDESTneo (72). The coverage for six plasmids was increased via *de novo* assembly, these being PX459 (+589 X), pCMV6\_entry\_*LGR4* (+839 X), pCTAP\_*LGR4* (+1055 X), pKH3 (+581 X), pDEST504 (+268X) and pDESTneo (916 X). A mean read quality  $\geq$ Q 12 and an average identity  $\geq$ 91% was obtained for all plasmids using both assembly methods. Sequencing metrics were extracted from NanoStat while coverage data was obtained from IGV's Sashimi plot. Mean read quality corresponds to Phred scores whereby a Q 10 cut off implies a 10% error rate in base calling. Identity refers to the percentage of identical nucleotides relative to the reference sequence used. Read length N50 denotes the shortest contig required to represent 50% of the sequence. Known reference sequences were obtained from manufacturers or submitters with different degrees of confidence.

In order to test the robustness of *de novo* assembly, a comparative analysis was performed of both reference assembly and *de novo* assembly for each plasmid (Table 5.6). For reliability purposes, the reference assembly was assigned as the control. The DNA sequence of both assemblies were aligned by global pairwise alignment using the Needleman-Wunsch algorithm. The alignment scoring system used gap penalties of 10, 0.5, 10 and 0.5 for gap open, gap extend, end gap open and end gap extend respectively. The gap is a corrective measure to resolve variations in the aligned sequences due to additional or missing bases in the pairwise alignment. The *de novo* assemblies for 9/10 plasmids were either shorter or longer than the reference, with only pDONR221 achieving the exact plasmid reference size of 4761 bp (zero difference in size). The percentage of homology between the reference and the queried *de novo* assemblies for all plasmids except pDEST neo, was  $\geq 97\%$  and the gaps in the sequence alignment amounts to  $\sim 2.5\%$ . 3/10 *de novo* assemblies were shorter in length by  $\sim 46$  bp and achieved  $>97\%$  identity. As expected, the results for pDESTneo alignment had a much larger discrepancy (-583 bp), with only 84.3% identity and 12.7% gap in alignment.

Plasmid	Size difference	Identity	Gaps
PX459	+12 bp	9103/9243 ( <b>98.5%</b> )	126/9243 ( <b>1.4%</b> )
pDONR221	0 bp	4720/4799 ( <b>98.4%</b> )	76/4799 ( <b>1.6%</b> )
pcDNA3.1	+11 bp	5473/5520 ( <b>99.1%</b> )	43/5520 ( <b>0.8%</b> )
pCMV6_entry_LGR4	+8 bp	7687/7766 ( <b>99.0%</b> )	56/7766 ( <b>0.7%</b> )
pCTAP_LGR4	-2 bp	8448/8532 ( <b>99.0%</b> )	56/8532 ( <b>0.7%</b> )
pDEST504	+13 bp	7723/7814 ( <b>98.8%</b> )	76/7814 ( <b>1.0%</b> )
LAMP1+ midigene	+3 bp	7758/7826 ( <b>99.1%</b> )	43/7826 ( <b>0.5%</b> )
pKH3	-46 bp	4716/4842 ( <b>97.4%</b> )	116/4842 ( <b>2.4%</b> )
pRK5_mFzd4	-32 bp	6243/6418 ( <b>97.3%</b> )	162/6418 ( <b>2.5%</b> )
pDESTneo	-583 bp	7408/8785 ( <b>84.3%</b> )	1113/8785 ( <b>12.7%</b> )

**Table 5.6 Comparative analysis of *de novo* and reference sequence assemblies.** Size difference refers to the difference in length of the assembled sequence between the reference and *de novo* assembly. Identity refers to the percentage of identical nucleotide matches in the alignment (identical nucleotides/total nucleotides). The gap percentage is based on the gap count divided by the total nucleotides for the plasmid. Note that the increased length of aligned sequences is due to the gaps inserted by EMBOSS Needle for an improved alignment.

### 5.2.11 Generating CRISPR-Cas9 knock-in models of *LAMP1*

To further test the effect of *LAMP1* variants on function, a study was carried out using CRISPR-cas9 gene editing to induce homology directed repair (HDR) to generate a cellular model harbouring the *LAMP1* variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14). In parallel, a second experiment was proposed to use NHEJ to create other INDEL variants at or near the same site. The aim was that these cellular models would be used in downstream experiments carrying out functional analysis to determine whether this frameshift variant and/or other similar variants affect protein function, lysosomal integrity or localisation, providing insights into intracellular trafficking of the aberrant *LAMP1* protein.

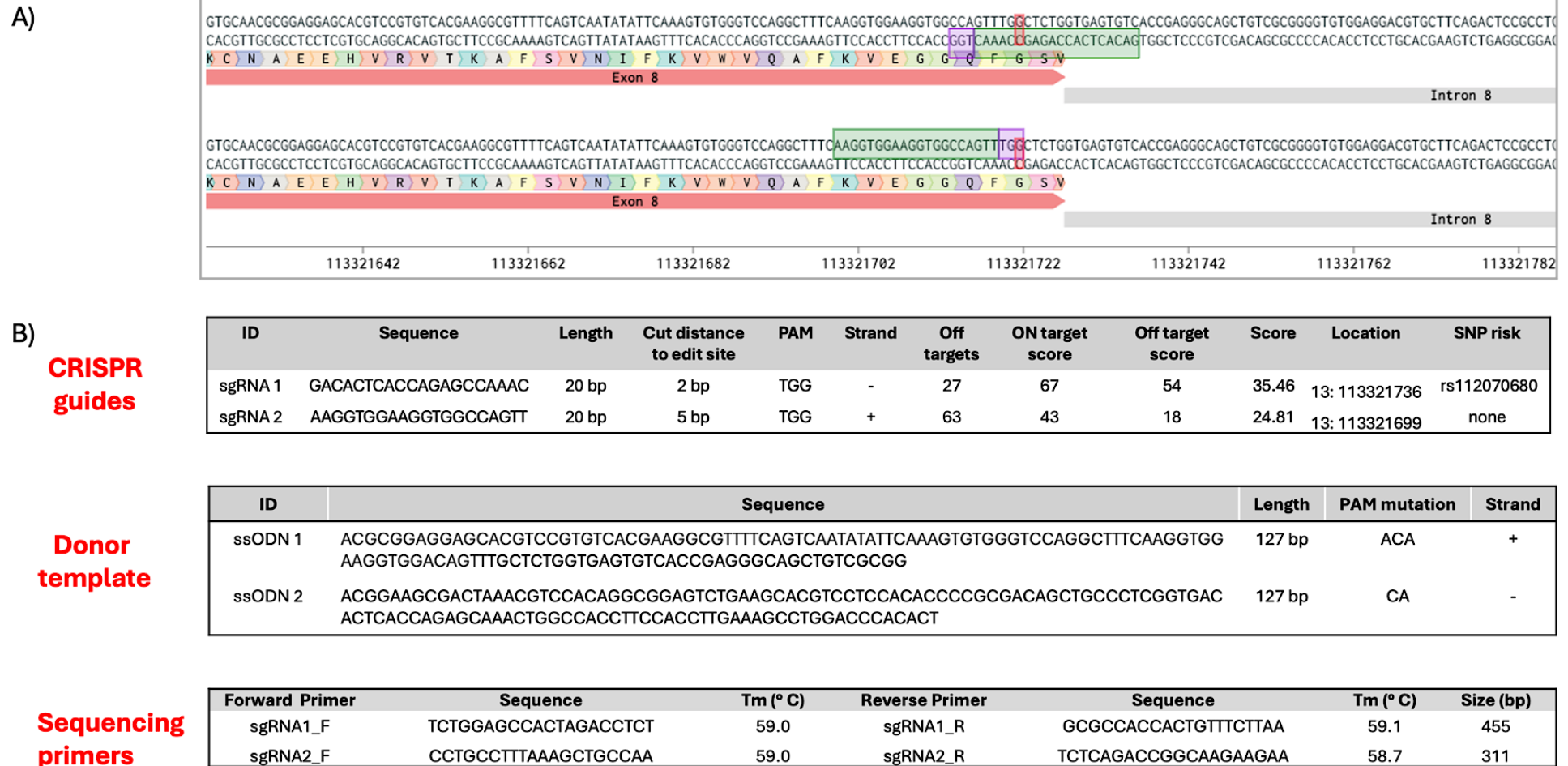
#### 5.2.11.1 Designing sgRNAs for *LAMP1* knock-in and knock-out models

Generating cellular models of the *LAMP1* variant found in family F5 required CRISPR-Cas9 mediated genome edits to recapitulate the exact NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) variant in appropriate cell lines. Guide RNAs were designed to target exon 8 at position chr13:113321721 to delete the guanine nucleotide, resulting in a frameshift and a subsequent PTC in exon 9. The aim was to use these guide RNAs to facilitate HDR in the presence of a donor repair template that is homologous to the target locus and carries the required genomic edit. To increase targeting efficiency, the sgRNAs and donor templates were designed to be of different orientation (+/-) and the donor template was asymmetrical and restricted to 127 bp (Richardson et al., 2016, Sanjurjo-Soriano et al., 2020). An asymmetrical donor template consists of 91 bp proximal and 36 bp distal to the PAM (Richardson et al., 2016).

Only two sgRNAs had the most appropriate design due to the limited PAM availability around the 3' end of exon 8 (Figure 5.20). Both are predicted to have multiple off target effects resulting from guide RNAs annealing to partially complementary sequences. These unwanted targets are documented in Appendix C: Supplementary Table 5.1 and 5.2. The donor templates were designed to have the PAM mutated so that when the desired edit is achieved, the Cas9 will not continue to cleave the DNA perpetually.

In addition, it was planned to use the same sgRNAs in the absence of a donor template to induce error prone NHEJ. This should introduce INDELS at the variant site to mimic the effect of NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) by resulting in a similar protein truncation. Primers were also designed to flank the edit sites for PCR amplification followed by Sanger sequencing for confirmation of intended genomic edit

In order to avoid any unintended effects arising from genome editing of the *LAMP1* gene, both the PAM site used by the sgRNAs and the location of the cleavage induced by the Cas9 (5 bp from the 5' of PAM) were assessed for splice sites using data from the previous analysis (Figure 5.8A and 5.8B). Neither the PAM nor the cut sites of the sgRNAs overlapped any predicted splice sites.



**Figure 5.20 Genome editing strategy for *LAMP1* knock in.** A) Visual representation of sgRNAs designed for HDR mediated mutagenesis in exon 8 of *LAMP1*. The sgRNAs are highlighted in green while the PAM is highlighted in purple and the edit site in red. The corresponding amino acid of each codon along the exon is presented at the bottom of the sequence using a one letter amino acid identifier. B) Primer design parameters of the sgRNAs to guide the Cas9, ssODNs for HDR and the sequencing primers for verification by Sanger sequencing.

### 5.2.11.2 Investigating *LAMP1* expression in available cell lines and tissues

To select the most appropriate cell line in which to conduct the proposed CRISPR-cas9 gene editing experiment, *LAMP1* expression was assessed in various tissues and cell lines. RT-PCR was performed to generate cDNA from RPE derived cell lines (ARPE19, ARPE1 and RPE1), a neuroblastoma cell line (SH-SY5Y), a fibroblast cell line created locally (Cont fibro) and RNAs derived from human brain, cerebellum, kidney and liver (Figure 5.21).

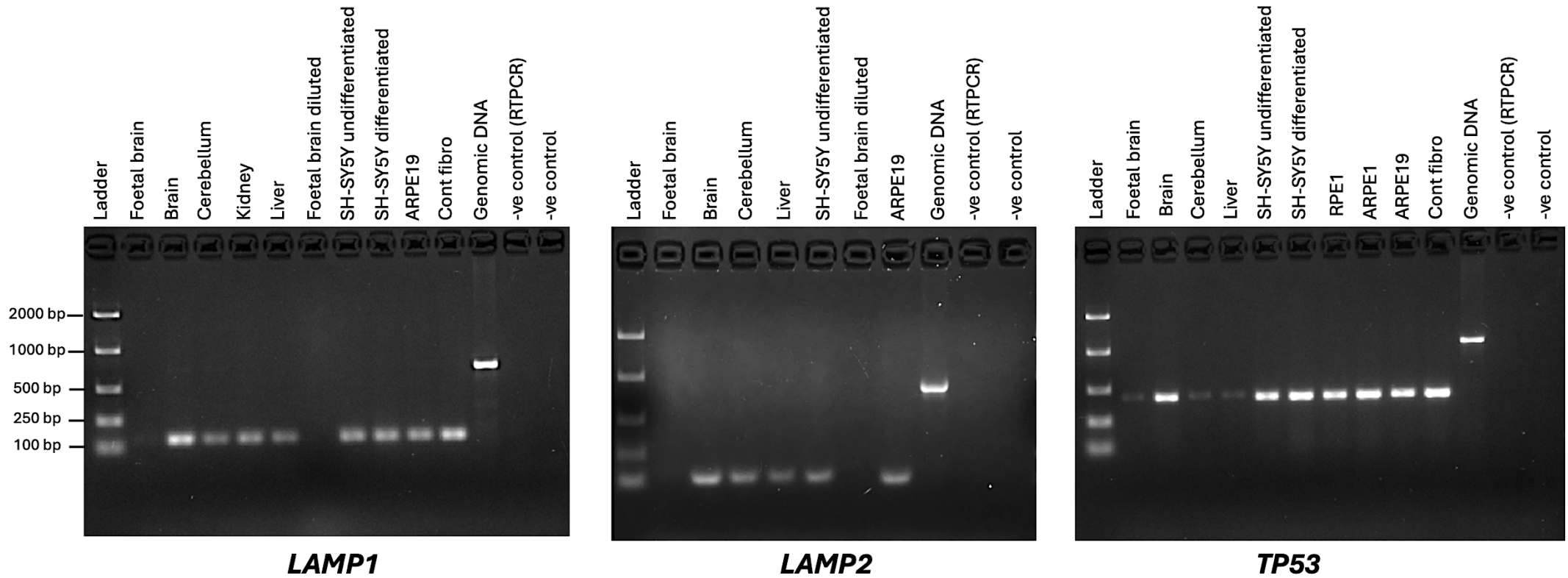
The primers were designed to have the primer binding sites on separate exons so that the genomic product will include the intron and can be distinguished from the cDNA product based on the larger size. The primer pair were also assessed for compatibility, ensuring that both primers have a  $T_m$  within a similar range for a synchronised primer performance (Appendix C: Supplementary table 5.3). To prevent non-specific target amplification such as in the case of highly homologous genes like *LAMP1* and *LAMP2*, the primer sequence was screened against the refseq mRNA and the refseq genome of *Homo sapiens* to confirm primer specificity (<https://www.ncbi.nlm.nih.gov/tools/primer-blast/index.cgi>). As an additional precaution, the primer binding site was assessed for the presence of polymorphisms that could hinder primer annealing (<https://genetools.org/SNPCheck/snpcheck.htm>).

*LAMP2* expression profile was also assessed as this is a paralog of *LAMP1* that is known to be upregulated to compensate for *LAMP1* deficiency in mice (Andrejewski et al., 1999). Monitoring *LAMP2* expression would later serve as an indirect test to confirm loss of *LAMP1* expression resulting from a gene knock out. This experiment showed that both LAMP genes (*LAMP1* and *LAMP2*) share the same expression profile for all the cell lines and tissues tested. The *LAMP* genes are ubiquitously expressed except in foetal brain. However, this may be because the cDNA is of lower concentration. In contrast, the *LAMP1* and *LAMP2* expression is the highest in brain RNA in comparison to other samples tested, which may be because the brain RNA used was a recent aliquot of better overall quality with minimal degradation.



The negative control for the PCR, which lacks template DNA did not show signs of DNA contamination in the PCR master mix. A second negative control derived from the RT-PCR experiment was included to rule out cDNA contamination in the RNA sample preparation. This RT-PCR control did not contain reverse transcriptase required to generate cDNA. The lack of any amplification in this tube has confirmed the absence of any contaminating cDNA.

The housekeeping gene *TP53* was selected as a positive control for its ubiquitous expression. *TP53* displayed consistent gene expression in all samples tested except in three tissues of foetal brain, brain and liver, where signal was lower. The lower signal intensity for these tissues may reflect a low cDNA yield, possibly due to compromised RNA integrity. DNA samples derived from kidney, RPE1 and ARPE1 cell lines were not sufficient for gene expression assessment of all three genes of *TP53*, *LAMP1* and *LAMP2*. Nevertheless, the other remaining RPE cell line was tested for expression of all three target genes. The ARPE19 cell line displayed uniform expression of *LAMP1*, *LAMP2* and *TP53* and was consequently selected for genome editing and downstream functional evaluation.



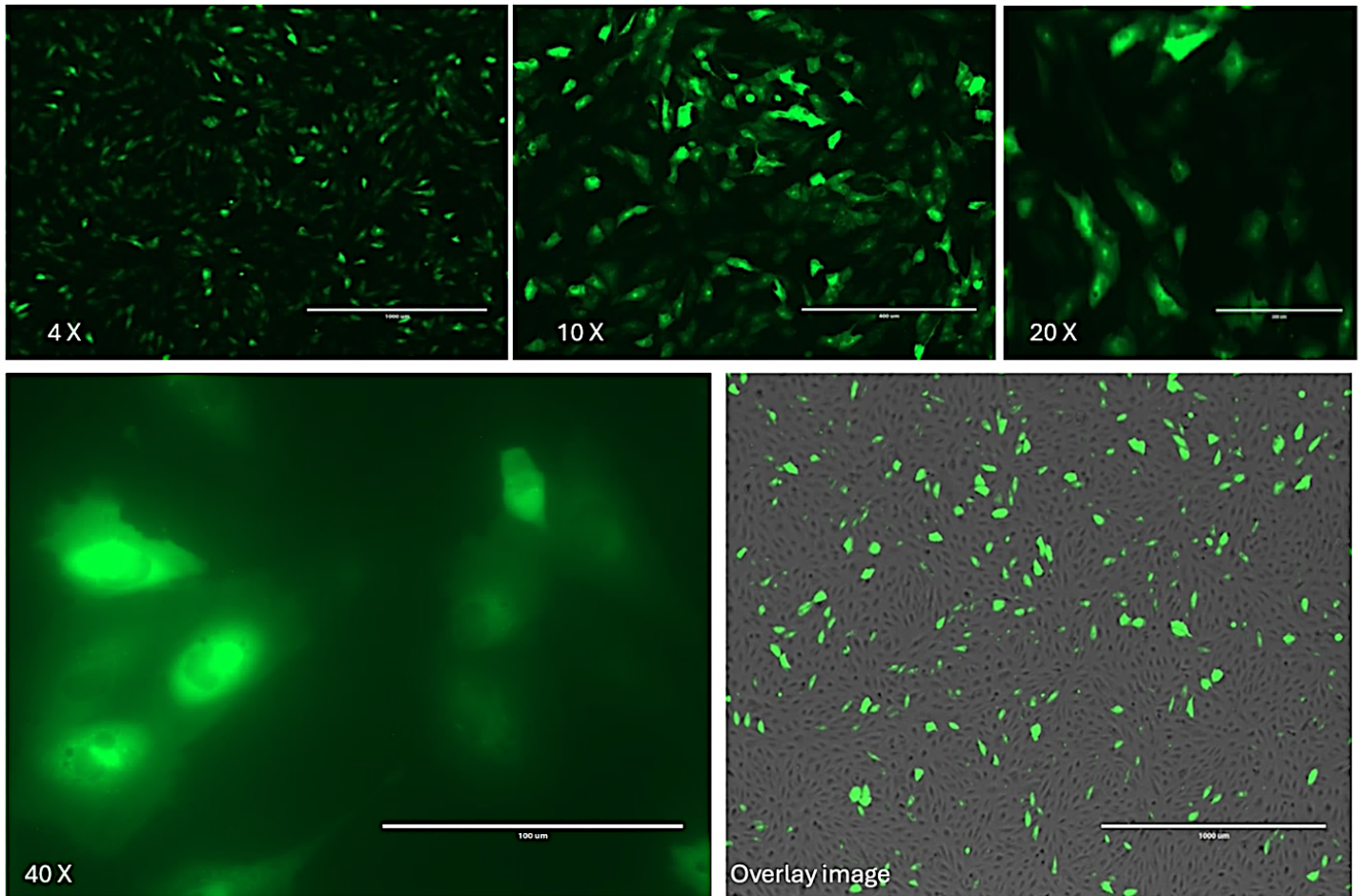
**Figure 5.21 Tissue expression of *LAMP1* and *LAMP2*.** The *TP53* control shows ubiquitous but non-uniform expression, with very low levels in foetal brain, cerebellum and liver, possibly due to RNA degradation. The *LAMP1* and *LAMP2* genes show expression in all samples tested at similar intensities, except in foetal brain. No contamination was detected in the negative control for both the RT-PCR used in cDNA generation and the subsequent PCR for evaluating tissue expression. The cDNA amplicons for *LAMP1* (exon 8), *LAMP2* (exon 7 - 8) and *TP53* are 128 bp, 108 bp, 408 bp respectively. The genomic DNA generated the expected banding pattern of 682 bp for *LAMP1*, 812 bp for *LAMP2* and 1057 bp for *TP53*. Image obtained on a 2% agarose gel.

### 5.2.11.3 Assessing the transfection efficiency of ARPE19 cell line

After selecting ARPE19 cells as an appropriate cell line that expresses *LAMP1* and *LAMP2* mRNA, the ARPE19 cells were subjected to nucleofection using 2 µg of green fluorescence protein (GFP) control vector to determine the efficiency of foreign genetic material uptake (section 2.5.6). Nucleofection is an electroporation-based transfection protocol that uses an electric pulse and proprietary chemical additives to enhance exogenous nucleic acid entry into the nucleus while maintaining the cell's viability. The GFP protein used as a control is excited by 488 nm wavelength light and the emitted wavelength of 509 nm is detected by microscopy.

ARPE19 cells were visualised post-transfection using fluorescence microscopy at various magnifications (4X - 40X) for a wider field of view or a more detailed imaging (Figure 5.22). The morphological appearance of ARPE19 cells resemble a rice grain but post-nucleofection, the ARPE19 cells expressing GFP appear distorted and enlarged, which may indicate cellular stress imposed by the transfection process or cellular toxicity arising from the amount of GFP present. However, the green fluorescence enhances the cellular morphology of ARPE19 cells which could explain the irregular cellular appearance as a more detailed morphology.

Overlay imaging was also performed to contrast the number of live cells transfected with GFP to those that did not take up the GFP plasmid. The transfection process yielded a significant number of live cells (~8%) that express GFP. Approximately 259 GFP expressing cells were detected at 4 X magnification with 3215 ARPE19 cells not displaying any fluorescence (Figure 20 overlay image). This indicates that the ARPE19 cells are suitable for CRISPR-cas9 mediated mutagenesis as they should readily take up the Cas9 plasmid to yield a high number of transformants that produce the Cas9 endonuclease. This will in turn increase the targetting efficiency and chances to achieve the desired *LAMP1* knock-in in the presence of the appropriate sgRNAs and template ssODN.



**Figure 5.22 Fluorescent imaging of GFP transfected ARPE19 cells.** A confluent culture of ARPE19 cells expressing GFP following nucleofection. The overlay image was captured at 4 X with the background cells being monochromatic while the GFP expressing cells are fluorescent green. White line depicts scale in micrometres. At 4 X it is 1000 µM, at 10 X it is 200 µM, at 20 X it is 200 µM and at 40 X it is 100 µM. Cellular imaging was performed on an EVOS FL (Thermo Fisher Scientific) fluorescent microscope.

## 5.3 Discussion

### 5.3.1 The identification of a novel pathogenic variant in *LAMP1*

This chapter describes a genomic investigation using whole exome sequencing and autozygosity mapping to delineate the genetic aetiology of FH in family F5, as well as the initiation of a functional study to investigate the cellular consequences of the most likely candidate variant. The use of variant interpretation restricted to variants in ROH, coupled with segregation analysis within the family, has identified a novel *LAMP1* variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) as a possible cause of FH in this family (Table 5.1). This rare variant was absent from the gnomAD population database, which excludes individuals with severe paediatric disease. GnomAD v4 is a large database that currently holds 76215 genomes and 730947 exomes, yet only 53 LOF variants in *LAMP1* are reported. The paucity of *LAMP1* LOF variants in gnomAD suggests that frameshift variants like p.(Gly370Alafs\*14) are not retained in the general population but are lost through natural selection. These variants are unlikely to be , as also evident by the LOEUF (loss-of-function observed / expected upper bound fraction) score of 0.338, indicating negative selection for LOF in the MANE transcript (NM\_005561.4) of *LAMP1*. Screening of the entire GEL rare disease dataset (n=72937) for further biallelic potentially pathogenic *LAMP1* variants revealed a single participant displaying a complex phenotype which includes nystagmus. However, this patient also carries a homozygous VUS missense variant in the *ATXN2* gene, and expresses neuromotor symptoms including ataxia, which may suggest the nystagmus might be a secondary clinical manifestation caused by the *ATXN2* variant. Therefore, the investigation of the 100KGP did not identify any further convincing biallelic cases, either in FH or IRD cases, which would have provided further support for *LAMP1* variants as causing FH, or in other clinical phenotypes which might have casted doubt on the proposed hypothesis.

This variant could have several possible effects on the mRNA and the encoded protein. The PTC in the mRNA could be recognised as abnormal which would lead to the transcript being subject to NMD, and therefore no protein would be produced. Alternatively, the position of the PTC in the last exon (exon 9) may

enable the aberrant transcript to escape NMD, meaning it would be translated, leading to the production of a truncated protein. A further possibility is that the altered recognition of the last base at the penultimate exon (exon 8) could affect the adjacent splice donor site on intron 8 leading to an as yet unknown splicing defect.

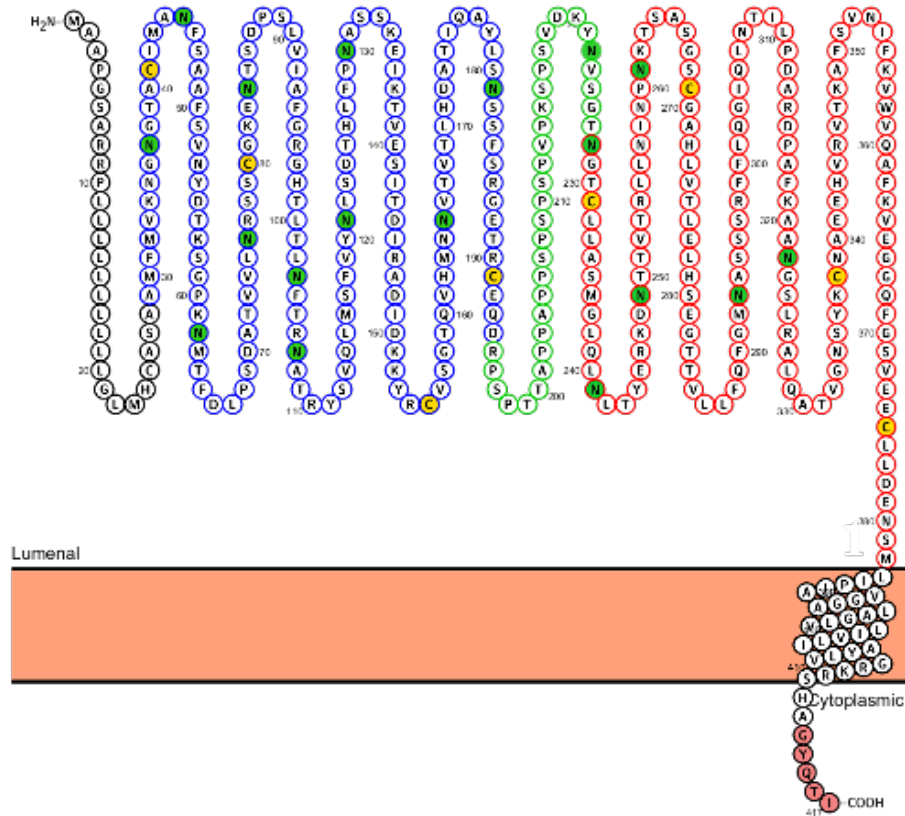
The *LAMP1* protein consists of a luminal domain (1-382 aa) that has two subdomains (the N and C domains) separated by a linker region, a transmembrane domain (383-410 aa) and a short cytoplasmic tail (411-417 aa) (Figure 5.23). In the event of the mRNA with NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) escaping NMD and a truncated protein being produced, the induced frameshift would eliminate protein residues 370-417, which show high evolutionary conservation in 150 orthologs using ConSurf analysis. The loss of 48 amino acids would result in a shorter protein of 384 residues. This deleted region coincides with the transmembrane domain and the short cytoplasmic tail protein that are almost entirely conserved through evolution with 77% (37/48) of the residues being graded as highly conserved (Figure 5.7). The evolutionary conservation of the C-terminal region suggests a biologically significant structural or functional motif that is crucial for protein activity. An example of such important motif that is lost through the NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) is the tyrosine-based sorting motif (YQTI) on the C-terminus that is required for *LAMP1* delivery via the trans Golgi network to lysosomes (Honing et al., 1996). The cytoplasmic tail sequences of YQTI are sensitive to change, especially the threonine, which when mutated can lower the binding affinity to AP1 and AP2 protein complexes by 11 fold (Obermuller et al., 2002). Both AP1 and AP2 proteins are involved in the endosomal sorting of *LAMP1* into clathrin coated vesicles for delivery from the trans-golgi network to late endosomes and lysosomes. It therefore seems likely that loss of these 48 C-terminal residues will affect protein function.

Other important signalling sequences at the C-terminus that would also be lost include the Gly413 residue. The mouse equivalent of this *LAMP1* residue is Gly378, which is needed for *LAMP1* sorting into the trans-golgi network for direct delivery to endosomes and lysosomes (Rohrer et al., 1996). This murine residue

precedes the YQTI sorting signal and when mutated, p.(Gly378Asp) results in severely reduced interactions with the  $\mu$ 3A subunit of AP3 by 7 fold and results in about half the expected *LAMP1* abundance in lysosomes (Yamaguchi et al., 2022). The NM\_005561.4:c.1109delG variant would be expected to eliminate Gly413 and the YQTI peptide signal, and might therefore be expected to cause pathology relating to *LAMP1* mislocalisation. Even if a small proportion of aberrant *LAMP1* reaches the intended lysosomal or cell surface destination, the NM\_005561.4:c.1109delG variant may also abolish the transmembrane domain, so that the truncated protein will be unable to tether to the plasma membrane and may accumulate in the cytoplasm. The exact intracellular consequences of the NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) variant are yet to be elucidated. One possible experimental approach that could determine the possible mislocalisation of *LAMP1* effect is through *In vitro* immunostaining of CRISPR-cas9 mediated knock in cell lines using anti-*LAMP1* antibodies to monitor protein trafficking and localisation.

Another possibility is that a potential splicing defect due to disruption of the nearby canonical splice site or cryptic splice site activation may result in intron retention or exon skipping. Given that the NM\_005561.4:c.1109delG variant occurs at the penultimate exon, the aberrant RNA could incorporate intron 8, exclude exon 8 or 9 or include exon 8, 9 and intron 8. Experimental evidence is required in the form of RT-PCR if patient samples are available, or using a *LAMP1* midgene splice assay, to ascertain the exact aberrant splicing event. Nevertheless, these events would most likely render the encoded protein similarly dysfunctional. The missplicing would shorten or extend the *LAMP1* transcript based on the combination of exons and introns spliced in and these may alter the position of particular domains or signal peptides. Changes in length between the sequences of the transmembrane domain and YQTI trafficking signal resulted in reduced delivery of *LAMP1* to lysosomes and redirected ~40% of the protein to the cell surface (Rohrer et al., 1996). The caveat of the study by Rohrer et al (1996) is that the changes in amino acid position are linear but were interpreted on the 3D protein structure as direct alteration to the distance between the residues. This does not take into account that the protein structure is a complex assembly based on folding to generate secondary structures and intermolecular

bonds so that the perceived distant residues may come in close proximity when the protein folds. For example the C-luminal subdomain of *LAMP1* protein adopts a conformation based on planar and bent  $\beta$  sheets known as  $\beta$ -prism fold that is maintained by four disulfide bonds at eight conserved cysteine residues of the luminal protein segment (Terasawa et al., 2016).



**Figure 5.23 *LAMP1* protein schematic.** Protein consists of several domains, with the N luminal domain (blue) at 29-194 aa, C luminal domain (red) at 228-382 and a hinge region (light green) at 195-227. The peptide sequence is also annotated with sites for post-translational modifications (green circle), disulfide bonds (yellow circle) and intracellular trafficking signal (pink circle). The luminal domain has four cysteine residues, making up a total of nine cysteine residues, and the entire protein has 18 N-glycosylation sites. The diagram does not depict the structure of *LAMP1* but rather annotates the peptide sequence with the known protein domains of significance. Protein diagram generated using UniProt *LAMP1* protein (P11279) on Protter (Omasits et al., 2014).



Determining whether the *LAMP1* variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) causes aberrant splicing or undergoes NMD requires empirical evidence obtained through RT-PCR or midigene splice assays. The work presented in this thesis has generated and validated *LAMP1* midigenes carrying the variant (Figure 5.13) and has also created an isogenic control midigene with the wildtype sequence (Figure 5.15). Unfortunately, the splicing study was not completed due to time constraints.

Further work presented in this thesis has also generated the foundation for a CRISPR-Cas9 study to generate cellular models with the *LAMP1* variant NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14). The works include the designed CRISPR cas9 guides, donor templates and sequencing primers (Figure 5.20), identifying an appropriate retinal cell line that expresses *LAMP1* and *LAMP2* (Figure 5.21) and can be readily transfected (Figure 5.22) and generating a control cell line transfected with only the Cas9 plasmid (data not presented). Though the experiment has not been completed, the work described here can be build upon to further characterise the *LAMP1* protein and shed light into the endo-lysosomal trafficking of LAMP proteins.

### **5.3.2 Ascribing a potential clinical phenotype to pathogenic variants in *LAMP1***

The retinal pigmentation pathway has been a subject of focus in IRDs due to the established link between impaired melanin biosynthesis and FH, as demonstrated in all forms of albinism (Bakker et al., 2022). The exact mechanism by which the melanogenesis pathway affects foveal development remains unknown. Of particular relevance to the potential FH phenotype due to *LAMP1* pathogenic variants in proband F1288 are the syndromic forms of albinism (HPS and CHS) which both arise from impaired endo-lysosomal trafficking of key enzymes to lysosomes or LROs. The resulting clinical phenotype of systemic hypopigmentation and the additional cell type specific anomalies are due to the defective LROs.

Nonetheless, the proband in family F5 was not reported to exhibit hypopigmentation. If systemic or ocular hypopigmentation were present in the proband, who is of South Asian descent, it would have been obvious and detected by the clinician. Moreover, there are no reports of additional clinical features or more serious and life threatening complications as experienced by CHS or HPS patients. Based on these observations, if the hypothesis that biallelic and pathogenic variants in *LAMP1* causes FH is correct, it would suggest that 1) *LAMP1* is acting on a pathway independent to the retinal pigmentation pathway in albinism; or 2) *LAMP1* is a downstream key player in the pathway, similar to *SLC38A8*, causing FH without pigmentary abnormalities or perhaps with subtle hypopigmentation.

#### **5.3.2.1 LAMP1 could arrest foveal development through autophagy**

LAMP proteins are the major constituents of membrane protein in lysosomes and lysosome derived organelles, and they are considered to be responsible for maintaining the integrity of these membrane bound vesicles. However, in double knockout mice that are deficient in both LAMP1 and LAMP2, the lysosomes in the fibroblast cells showed normal hydrolytic activity (Eskelinen et al., 2004). These mouse embryos survived up to 16.5 days and they did not show ocular manifestations but rather accumulation of large autophagic vacuoles in skeletal and cardiac muscle and impaired cholesterol metabolism, reminiscent of Danon disease in human patients, which is caused by *LAMP2* deficiency (Eskelinen et al., 2004). Danon disease is an X-linked disorder characterised by cardiomyopathy, skeletal myopathy, intellectual disability and accumulation of autophagic and glycogen vacuoles visible on histological staining (Danon et al., 1981).

Alternatively, LAMP1 may have functions beyond the putative structural role of protecting lysosomes from degradation by the encased hydrolytic enzymes which dominate the literature. LAMP1 has also been implicated in autophagy, a cellular catabolic process that eliminates obsolete cellular components and recycles the broken down products to sustain homeostasis. LAMP1, along with LAMP2 and 14 lysosomal membrane proteins, mediate fusion of autophagosomes with

lysosomes for degradation of the encapsulated cargo (Xu et al., 2023). Autophagosomes are vesicles that sequester cellular constituents destined to be degraded. The process of autophagy supports ocular development in many ways, including the formation of the organelle free zone of the developing lens (Gheyas et al., 2022) and the regression of the hyaloid blood vessel during retinal vasculature maturation (Kim et al., 2010). Moreover, evidence from studies using inhibitors of autophagy such as wortmannin and studies of murine knockout models of autophagy regulator *Ambra1* showed a reduction in neuronal markers of NeuroD and  $\beta$ -III-Tubulin, indicative of hampered neurogenesis and reduced neuronal differentiation (Vazquez et al., 2012). It has also been suggested that autophagy is potentially involved in cone differentiation through the mTOR signalling pathway (Boya et al., 2016). Considering the detected higher expression of *LAMP1* in brain (Figure 5.21), it is therefore plausible to propose that *LAMP1* disruption of autophagy during embryonic retinal neurogenesis may result in aplasia of the fovea.

### **5.3.2.2 LAMP1 as a potential downstream component in the retinal pigmentation pathway**

It remains a plausible hypothesis that the NM\_005561.4:c.1109delG; NP\_005552.3:p.(Gly370Alafs\*14) variant renders the *LAMP1* gene and/or protein dysfunctional, and that the effect of this extends to melanosomes, a melanin producing membrane bound organelle that share lysosomal membrane proteins (LAMP1 and LAMP2) and some lysosomal degradative enzymes like cathepsin B (Orlow, 1995, Diment et al., 1995). Melanin production is governed by strict intracellular conditions relating to pH homeostasis in melanosomes that is optimal at pH 6.8 and is severely reduced at pH <5.5. Changes in pH would alter the activity of tyrosinase in melanogenesis, skew the relative proportions of eumelanin (brown-black) and pheomelanin (yellow-red) and affect melanosome maturation in melanocytes (Ancans et al., 2001). It has also been demonstrated that the cation channel TPC2 on the surface of melanosomes acts as a negative regulator in melanin biosynthesis through expelling Na<sup>+</sup>. The resulting increase in the membrane potential enhances V-ATPase proton pump activity, which reduces intramelanosomal pH (Bellono et al., 2016). Similarly, LAMP1 protein was recently found to contribute to an acidic luminal environment by inhibiting the

lysosomal cation channel (TMEM175) to reduce proton efflux (Zhang et al., 2023). It is therefore possible that loss of function of LAMP1 may disrupt the pH of melanosomes and that this in turn may hamper melanin production, reducing it to sub basal levels (below normal). Subtle changes in melanin content may not be tolerated in the RPE during the early embryogenesis, leading to arrested foveal development.

### **5.3.3 A barcode free multiplexing strategy for plasmid assembly using ONT sequencing**

Traditional plasmid assembly uses Sanger sequencing which generates a single read of one orientation (forward or reverse) for an amplicon up to 500 bp. This capillary based sequencing offers the highest accuracy, with an error rate of 0.001 %, but at the expense of uniplex (one sample) sequencing (Huang et al., 1992). Based on this accuracy, Sanger sequencing has become the gold standard in many applications. In clinical diagnostics, it is used to probe regions that are hypermutable, such as the IGHV locus in chronic lymphocytic leukaemia, which is difficult to map using NGS. It is also used in validating NGS findings to rule out artefactual variant calls (Crossley et al., 2020, Davi et al., 2020). Other applications of Sanger sequencing include decoding whole plasmids using primer walking, whereby sequential sequencing is performed using overlapping reads to progressively build up the sequence of a large DNA construct (>1 kb) (Benes et al., 1997). Such assembly methods are labour intensive, have long turnaround times and are not suitable for large scale analysis, especially if whole plasmid sequencing is required instead of targeted sequencing of only the insert. The other alternative method for recombinant plasmid assembly is NGS, which does offer scalability with high accuracy in base calling with an error rate of 0.1-0.5% (Slatko et al., 2018). The main drawback of this approach is that it is not cost effective for a low samples size and it requires complex bioinformatic pipelines for analysis, leading to longer turnaround times.

The application of long read sequencing technology in plasmid sequencing serves as an intermediate technology, providing multiplexing capability at an affordable cost for small scale use and offering sufficient sequencing accuracy in

plasmid validation. This study demonstrated the feasibility of using Nanopore sequencing on a Flongle R9.4.1 flow cell to sequence 10 different plasmids simultaneously, without the use of short DNA identifiers known as barcode sequences (Figure 5.19). This greatly reduced the cost and decreased the turnaround time associated with sequencing to 1-2 days, which is equivalent to that of Sanger sequencing but without the need for designing and ordering multiple primers. The multiplexing assay generated 137,850 reads containing 866.14 Mb, over a sequencing duration of 24 hours (Appendix C: Supplementary Figure 5.5).

The mean read quality obtained from the multiplexed sequencing was Q12.2, which translates to an incorrect base error rate of 63 bases per 1000 bases sequenced and a base calling accuracy of 93.7% (Table 5.5). This is significantly lower than the quality standard of Illumina short read NGS (Miseq reagent cartridge V3), which yields  $\geq 70\%$  of reads at Q30 (1 in 1000 error rate) and is equivalent to a base calling accuracy of 99.9%. Furthermore, the high proportion of artefactual INDELS and SNVs generated by Nanopore sequencing makes it difficult to visualise BAM files on IGV, meaning that careful consideration is required to rule out artefactual variant calls.

However, using the reference assembly for alignment of reads to an existing plasmid sequence, the average coverage obtained for 9/10 plasmids (excluding pDESTneo) sequenced was 1763 X, providing enough reads to make a reliable judgement on excluding artefactual variant calls despite the high error rate. Based on the clonal origin of the plasmids sequenced, reads belonging to a single plasmid would be expected to be identical. Genuine INDELS or SNVs should therefore be represented in all reads. Allowing for the artefactual SNVs to be filtered based on an allele frequency of 0.8 or greater to eliminate erroneous variant calls, while INDELS can be interrogated based on changes to the Sashimi plot for a collective representation of coverage, instead of interpreting the scattered insertion and deletion markers assigned for INDELS.

Recent improvements in ONT sequencing technology with the introduction of the Flongle flow cell R10.4.1, together with recent advances in bioinformatic tools like Dorado toolkit, (<https://github.com/nanoporetech/dorado>), provide more accurate base calling and variant identification. These updates were not available at the time the multiplexed ONT sequencing study was conducted (section 5.2.10). Nevertheless, the novel ONT workflow introduced in this thesis has proved the concept of barcode free and multiplexed assembly of plasmids using ONT technology (Figure 5.18). Recent and future developments in Nanopore sequencing will further refine the sequencing of the pooled plasmid libraries and reduce the need for further analysis to eliminate artefactual variant calls.

The *de novo* assemblies for all plasmids except pDESTneo, had sequences with up to 2.5% discrepancy in size in comparison to the known reference sequence (Table 5.6). These discrepancies are based on artificially inserted gaps for the correction of misaligned sequences due to missing nucleotides. These gaps are distributed throughout the plasmid genome but are overrepresented at the start and end of the aligned sequences. Nonetheless, the high degree ( $\geq 97\%$ ) of identity between the *de novo* assemblies and the reference sequences for all plasmids except pDESTneo implies a partial plasmid assembled. The *de novo* assemblies of the 9/10 plasmids had minor discrepancies in comparison to the known reference sequence, which may hinder utility. Nevertheless, assemblies can be polished using tools such as those of the Nanopolish suite for base call correction (Simpson et al., 2017). The potential to perform *de novo* assembly using partial knowledge of a known reference sequence for the construction of plasmid sequence, makes the proposed ONT workflow more versatile and applicable for sequencing plasmids of unknown or uncertain origin (Figure 5.18). The application of *de novo* assembly described in the workflow is contingent on a minimum knowledge of unique sequences present in the plasmid (Appendix C: Supplementary Table 5.4). These could relate to distinctive features such as different antibiotic resistance genes, a unique insert or cloning related sequences. This partial *de novo* assembly serves as a secondary and a corrective assembly to rescue malalignments experienced with the reference assembly. All plasmids sequenced in this study using this partial *de novo* assembly achieved almost whole plasmid sequence representation in a single

read at an average coverage of 1899 X with  $\geq 93\%$  identity (Table 5.5). In cases where a complete *de novo* assembly is required, it may be necessary to use DNA barcodes in library preparation, which can then be demultiplexed and plasmids can be assembled individually without reliance on a known reference sequence.

The success of this *de novo* methodology has implications beyond the verification of cloning products in synthetic biology and, coupled with the portability of the MinION sequencer makes the application of the ONT workflow described in this thesis very useful in field study of virology. Whole plasmid sequencing using short read NGS has its merits in the identification of new strains of virus in phylogenetic analysis and in identifying contaminating pathogens to trace outbreaks in epidemiology studies (Silva et al., 2021). The application of ONT technology using *de novo* assembly in whole plasmid sequencing has offered better characterisation of viruses (Taylor et al., 2019). The long reads and the error rate achieved of 0.1% has resolved low complexity regions (repetitive sequences) for more accurate classification of viral strains (serotyping) and in detecting SNVs responsible for high virulence factors like antibiotic resistance (Taylor et al., 2019). The drawback of this reference agnostic method is that it demands high coverage and high quality reads to generate a reliable consensus read representing the target plasmid (Gallegos et al., 2020).

Benchmarking the sequencing performance data from this thesis against other similar studies has showed great potential in the ONT-workflow described in this thesis. A recent study using commercial barcoding kits for multiplexing and a MinION flow cell R9.4.1 instead of a Flongle flow cell R9.4.1, had achieved an unexpected lower coverage of  $\leq 7500$  X for plasmids of similar size that are of  $\leq 10$  kb (Johnson et al., 2023). The quality score was much higher for these long reads (Q20-21), almost double the average Q12 achieved in the multiplexing experiment described in this thesis (Johnson et al., 2023). Similarly, in a second study utilising a MinION flow cell R10.3 in plasmid assembly, the average reads generated from a 72 hour run were 3.86 million which is consistent with the significantly higher read output facilitated by the MinION flow cell and the longer run duration (Brown et al., 2023). The strategy outlined in their study uses pseudo paired reads to improve the base calling to an accuracy exceeding 99.9%. This

high accuracy satisfies the clinical research standards for variant identification, but it does not offer multiplexing capability as each plasmid is loaded on an individual MinION flow cell for sequencing.

Another study presenting barcode free and multiplexed ONT sequencing of plasmids also relied on identifying unique sequences based on a supplied reference. This study claimed a sensitivity down to 2 bp to discriminate between different plasmids (Uematsu and Baskin, 2023). Considering the high error rate of ONT sequencing using Guppy in base calling, it is difficult to reliably isolate reads based on a 2 bp identifier, especially if the identifier sequence overlaps homopolymer repeats, and if the proposed workflow did manage to filter out the target reads then the coverage is likely to be very low for any meaningful interpretation. This study demonstrated a coverage ~1000 X for sequencing a maximum of six plasmids. Both parameters are lower than the values obtained using the ONT sequencing workflow proposed in this thesis, which provided an average read depth of 1763 X for reference assembly and the multiplexing of 9 samples (excluding pDESTneo). Furthermore, that study pooled samples by standardised concentration instead of using an equimolar pooling which accounts for differences in plasmid size as was done in this study. This may explain the inconsistency in reads generated, whereby a large range in coverage 100-1000 X was obtained for the small sample size of six plasmids. A mutual quality control step in their workflow and the proposed ONT workflow in this thesis is the use of pairwise alignments for the initial assessment of compatibility of different plasmids to be sequenced in a single run. However, their study uses a more efficient and innovative approach utilizing the Smith–Waterman algorithm to generate a phylogenetic comparison with automated judgements on sequence homology. The ONT strategy proposed in this thesis is thought to be novel as it allows barcode free multiplexing, uses the lowest end ONT product known as the Flongle R9.4.1 for cost efficiency, and supports reference based and partial *de novo* assembly in plasmid construction.



### 5.3.4 Limitations of the study

The consanguineous relationship in family F5 and the autosomal recessive inheritance of FH, prompted the use of autozygosity mapping to delineate the disease loci (Figure 5.1) (Carr et al., 2012). The caveat of such an approach is that the affected family members might possess compound heterozygous pathogenic variants that are responsible for FH and these variants won't be detected as they are not confined to ROH. Furthermore, the genomic data of the proband was generated through WES which restricts the detection of novel variants to the exome and the exon-intron boundary. As a result, variants residing in non-coding regions of the genome will be missed and this class of variants is thought to account for approximately 15% of Mendelian disorders (Keogh and Chinnery, 2013). These variants include deep intronic variants that could alter splicing, and variants affecting distal regulatory elements that could interfere with the transcription of a target gene. Although it is less likely, it may also be valuable to screen the mtDNA for pathogenic mutations, whereby heteroplasmy could generate variable phenotypes in progeny. Of note, due to technical limitations only a single affected proband underwent autozygosity mapping. It would have been better if both affected siblings F1287 and F1288 were analysed to identify candidate variants that are shared by both siblings and are present in mutual ROH, but this approach was restricted by the inability to access the genomic file during COVID19 lockdown. Another weakness of the study is that CNVs within the identified ROH in proband F1288 were not analysed. This might lead to pathogenic variants causative for the isolated FH being missed in the family.

The hypothesis that *LAMP1* variants have a pathological effect is based on the identification and co-segregation with FH in the family. However, the determination of the variant's molecular consequence on the mRNA transcript and protein function is based on preliminary data and theoretical concepts (section 5.3.2). Further empirical evidence at the cellular level is required to ascertain the likely splicing consequences of the NM\_005561.4:c.1109delG variant and to formulate sound hypotheses on the possible pathogenic mechanisms underlying FH in the family. Functional studies are required to determine whether the *LAMP1* variant interferes with melanin biosynthesis, lysosomal trafficking or in hampering autophagy response in cells. To this end,

this study has generated *LAMP1* midigenes for use in *In vitro* splicing assays (Figure 5.13 and 5.15) and for testing CRISPR guides (Figure 5.20) in preparation for such studies, as well as identifying a suitable cell line for gene editing experiments targeting *LAMP1* (Figure 5.21). Additional supporting evidence could be obtained by identifying other unrelated families affected with FH and having pathogenic *LAMP1* variants, but none were found within the limits of this study. The recruitment of additional family members of family F5 for segregation analysis would also increase the confidence in proposing *LAMP1* as the potential cause of FH in the family. Future work would continue these experiments to provide further insight into the role of *LAMP1* in the mammalian retina.

The barcode free assembly of plasmid sequences using the workflow described in this thesis is dependent on the complete or partial knowledge of the sequence of the target plasmid (Figure 5.18). The reference based mapping may serve its purpose in facilitating assembly of recombinant plasmids as the sequence of the cloned plasmids are usually known. One limitation of such biased assembly is that filtering for a particular subset of reads that align to the reference sequence will overlook potential contamination in the sample. Another drawback of this multiplexed and barcode free sequencing workflow is the sample compatibility for pooling. Plasmids must be no more than 80% similar in sequence to be demultiplexed and assembled successfully using the ONT approach outlined in this thesis. This requirement of at least 20% discrepancy in sequence is particularly problematic in validating a recombinant plasmid that is cloned and isolated from different transformed colonies. The very high degree of sequence homology, as these plasmids would be almost identical, would make it difficult to distinguish the origin of different reads without index sequences (DNA barcodes).

## Chapter 6 : General discussion

### 6.1 Molecular characterisation of *SLC38A8* in *FVH2*

The identification of *SLC38A8* as a gene underlying FH at the Toomes laboratory has gained attraction within the field which prompted the screening of many IRD patients of no pigmentation defects for *SLC38A8* variants. The published literature, the 100KGP and collaboration with ERDC has provided 51 unique *SLC38A8* variants in a cohort of 63 solved *FVH2* cases for analysis (Table 3.4 and Appendix C: supplementary Table 3.5). A mutation spectrum was constructed revealing the distribution of *SLC38A8* variants on the gene schematic and protein topology. These 51 variants affected all the 11 exons of *SLC38A8* with no indication of a mutational hotspot at the DNA level (Figure 3.6). However, the missense variants had an increased tendency to cluster within the transmembrane domains of the protein (16/22), specifically at the 6<sup>th</sup> transmembrane domain (7/22) (Figure 3.7) The 6<sup>th</sup> transmembrane has 87% of its amino acids evolutionary conserved across 150 diverse species which implies a region critical for protein function (Figure 3.12). *In-silico* based *SLC38A8* protein modelling and missense variant simulation showed that the deleterious variants overwhelmingly affect conserved residues, causing severe changes in hydrophobicity and an increased the electrostatic potentials (Figure 3.9). Though the validity of these predictions are yet to be confirmed using functional studies. The consistent *FVH2* manifestation arising from both missense and LOF variants in *SLC38A8* imply a loss of function as the likely mechanism of disease in *FVH2*.

All the 51 biallelic and deleterious variants in *SLC38A8* displayed consistent phenotypes with notable discrepancies in chiasmal misrouting due to technical limitations in VEP and posterior embryotoxon that is an incidental finding in the general population (section 3.3.5). The overall evidence from interpreting the phenotype in 63 *FVH2* and the 51 deleterious variants suggest that there is a reliable genotype-phenotype correlation. The *FVH2* represents an isolated entity of FH with no significant evidence for hypopigmentation. However, the *SLC38A8* gene should be incorporated into albinism and retinal disorders gene panel to compensate for the diagnostic challenges in detecting chiasmal misrouting and

hypopigmentation which artificially generates a phenotypic overlap between *SLC38A8*, albinism and other FH disorders.

## 6.2 The application of WGS in FH

The advent of short-read NGS and the subsequent reduction in associated costs allowed more IRDs patients to undergo WGS and WES for a more comprehensive genomic analysis. The diagnostic utility of WGS has revolutionised clinical investigations into IRDs, since it enables the assessment of all the protein coding genes including the noncoding regions of the genome, when limited phenotype information is available. The application of this paired end short-read sequencing technology to patients with FH has enabled the detection of a large repertoire of variants, including 19 unique SNVs and 11 unique SVs comprising deletions and inversions, that are causative for FH in both the local unsolved FH cohort (F1335, F1369 and F1071) and the entire 100KGP analysis (C2-C5, C24,P8-P27) (Table 3.1, 4.5 and 4.7, and Figure 3.5 and 4.5). The resulting genomic diagnosis in with FH was attributed to *PAX6* (5), *GPR143* (5), *OCA2* (4), *SLC38A8* (4), *TYR* (2), *CACNA1F* (2), *HPS5* (1), *CNGA3* (1), *CNGB3* (1) and *ZNF408* (1), The genomic diagnosis in the 26 probands was communicated back to the responsible clinician using the clinical collaboration request in the 100KGP or through personal communication.

The diagnostic yield achieved from the analysis of patients in both the 100KGP and local cohort with the clinical report of FH was 25% (14/56), which is significantly lower than the reported diagnostic rate of 57.4% in IRDs using WGS (Weisschuh et al., 2024). The reduced diagnostic outcome in this doctoral study is primarily due to the 100KGP participants being a more difficult cohort to analyse since the standard genetic testing at the point of care did not detect any pathogenic variants. This cohort is therefore enriched with more elusive variants such as SVs and deep intronic variants that affect splicing or gene regulation, or variants in novel FH genes yet to be identified. Furthermore, the 100KGP analysis conducted on the FH cohort was incomplete considering that intronic variants were not assessed and this may have contributed to the reduced diagnosis.

To further improve the diagnostic yield in the 100KGP for FH, the retinal disorders v5.11 gene panel within the 100KGP was reviewed using expertise in the domain of FH (section 4.2.3.6). This panel was revised to include *PAX6*, *AHR*, *FRMD7* and *DCT* for clinical utility and the request is pending review by 33 experts assigned to this panel. The application of this virtual gene panel can benefit future analyses within the 100KGP research environment by providing supportive evidence in variant interpretation known as the Tiering classification (section 1.20.2). The advantages of using the revised gene panel would also extend to NHS genomic medicine centres as part of the national genomic test directory v7 for the clinical indication of retinal disorders (R32) (<https://www.england.nhs.uk/publication/national-genomic-test-directories/>). This would expand the genomic analysis in IRDs to 264 relevant genes and potentially including the four genes reviewed, to reduce the volume of variants requiring interpretation and ultimately reduce the turnaround time at such clinics with overwhelming workload.

### **6.3 The unsolved FH cases**

The inability to deliver a genomic diagnosis following WGS analysis in proband F1337 (section 4.2.2.5) and in other 39 probands from the 100KGP (section 4.2.3.3) could be attributed to one of a number of possible factors. Intractable genomic regions due to highly repetitive sequences can prove difficult to analyse, especially if the read length is smaller than the repeat region. These regions cannot be mapped accurately and often appear distorted or have a lower coverage. Variant calls within these regions may not pass the quality thresholds, causing them to be filtered out or not called in the first place. In addition, it can be difficult to design effective capture probes targeting these repetitive regions due to the formation of secondary structures at target site (Yahya et al., 2023). A prime example of this is *RPGR* ORF15, a hotspot for variants causing X-linked RP. This terminal exon is purine rich and contains ~999 bp of low complexity sequence (highly repetitive). Reduction in coverage is usually experienced at this region, which causes underlying *RPGR* mutations to be underreported (Chiang et al., 2018).

Another potential confounding factor in the analysis of WGS is the presence of VUSs. There is an increasing need to standardise variant interpretation to eliminate inter-laboratory reporting bias, but current guidelines can undermine the perceived pathogenicity of variants detected. The introduction of ACMG guidelines has raised the level of evidence required to support a pathogenic classification and in turn it increased the volume of variants classed as VUS. Evidence based on a variant's segregation with disease, allelic data, functional evidence, *de novo* status or phenotypic specificity are collated for each variant to determine pathogenicity (Richards et al., 2015). However, this information is not always available for each variant. The analysis performed as part of this doctoral study in the FH cohort has identified 17 SNVs that are all rare and predicted to be deleterious, yet four of these SNVs were classified as VUS (Table 3.1, 4.5, 4.7 and Figure 4.2). These are NP\_852608.1:p.(Pro872Leu) in *HPS5*, NP\_001243718.1:p.(Cys292Tyr) in *CACNA1F*, NP\_079017.1:p.(Phe533Leu) in *ZNF408* and NP\_001845.3:p.(Gly69Val) in *COL11A1*. Similarly, 17 *SLC38A8* variants that were also predicted to be deleterious were reported as VUS in solved FVH2 cases (Table 3.4), which further highlights that the ACMG classification does not always consistently reflect variants' pathogenicity and this may mislead genomic investigations, causing them to probe for further variants.

To aid in the interpretation of variants in *SLC38A8*, a comprehensive list of all the deleterious *SLC38A8* variants to date that causes FVH2 have been compiled from reported in the literature and the local cohort (Table 3.4) and made available on LOVD along with the corresponding ACMG classification (Figure 3.14). Diagnostic teams working in rare disease, are involved in genome wide variant interpretations which covers genes beyond their domain of specialism and makes it harder to appreciate the clinical consequences of VUS in these genes (Fokkema et al., 2021). Both fellow researchers and healthcare staff can benefit from accessing the up to date *SLC38A8* LOVD database (<https://databases.lovd.nl/shared/genes/SLC38A8>) to reduce interpretation bias and to support a genetic diagnosis in the event of identifying a known VUS. The publicly accessible *SLC38A8* database within LOVD contains 87 unique variants including those that are not pathogenic, that were identified in 151 individuals (visited 20/09/24). These *SLC38A8* variants are also integrated into both the

UCSC and Ensembl genome browsers (section 2.6.3) for a locus driven analysis and for convenience. The genomic and phenotypic data of the database are of high quality as the variants submitted are based on HGVS nomenclature and ACMG classification while the phenotype is detailed using HPO terms that are informative.

VUSs pose a diagnostic challenge, as the data is only clinically actionable where the degree of certainty for pathogenicity is high. In certain well characterised IRDs such as albinism, a known pathogenic allele exist in *TYR* which is composed of a combination of two non pathogenic variants that are if detected individually are not sufficient for pathogenicity unless both are reported in cis. For example, the complex allele NP\_000363.1:p.[Ser192Tyr; Arg402Gln] likely detected in F1377 (Table 4.3) constitutes a benign variant and a VUS in cis, was reported in trans with either of the pathogenic *TYR* variants identified in P9 and P11, NP\_000363.1:p.(Pro406Leu) and NP\_000363.1:p.(Thr373Lys) (Table 4.5) as the cause for oculocutaneous albinism in 13 probands (Loftus et al., 2023). This highlights the limitation of ACMG classification in variant interpretation which is more applicable to well characterised Mendelian diseases whereby additional information is available on the significance of allelic configurations of VUS, the phenotype is well defined, the disease mechanism is known, and functional data are available.

Intronic variants are often challenging to characterise due to lack of computational tools and parameters to gauge pathogenicity. The metrics currently used in intronic variant interpretation are allele frequencies, evolutionary conservation and the potential to alter splicing. However, these may not be applicable to the characterisation of intronic SVs and most computational tools including SpliceAI, CADD and varSEAK cannot process SVs to determine the likely splicing consequences. There are also several other neglected functional components of the non-coding genome that can potentially alter transcription, such as CREs. Analysing three different intronic *SLC38A8* variants in probands C11, C19 and C20 revealed that the rare variants affect a common distal enhancer-like signature at chr16:84025275-84025482 (Table 3.3). The CRE, EH38E3194808 altered by them is of 207 bp. Without functional assays the pathogenicity of

NM\_001080442.3:c.691-2585C>T and NM\_001080442.3:c.691-2425T>C cannot be ascertained. This creates a dilemma in diagnostic services whereby the identification of novel intronic variants are likely to be of no clinical utility since functional analysis including midgene splicing assays are usually performed at research institutes instead.

Other neglected regulatory domains include topologically associated domains (TADs) which potentially govern CREs (Madani Tonekaboni et al., 2019). TADs are regulatory regions that define chromatin structural boundaries on the chromosome to bring distant enhancers or repressors in contact with target promoters. Disruption to the 3D nuclear organisation provided by a TAD can disrupt gene expression, as was evident at the RP17 locus, where 8 intronic SVs with overlapping breakpoint regions were shown to cause ectopic promoter and enhancer interactions that resulted in aberrant gene activation in the retina (de Bruijn et al., 2020).

Another contributing factor to the unsolved FH cases is the presence of novel genes underlying FH that are yet to be characterised. The variants residing in these genes will evade detection since the targeted 100KGP analysis was restricted to 52 known FH genes. The unsolved FH cohort should undergo further genomic testing using third generation sequencing and next generation cytogenetic technologies outlined below (section 6.4) if accessible to rule out the known FH genes prior to attempts in novel FH gene discovery.

## **6.4 Technologies to improve the diagnostic yield in FH**

### **6.4.1 Long read sequencing for comprehensive variant discovery**

The diagnostic strategies described in this thesis failed to identify the genetic aetiology in 78% (39/50) of the 100KGP FH cohort. Patients with unsolved FH following WGS analysis in both the 100KGP and local cohort should be recruited for a secondary test using third generation sequencing, also known as long read sequencing. Longer reads provide a superior read assembly capable of generating a more contiguous genome that provides a better resolution of



variants in intractable genomic regions (Nurk et al., 2022). This will aid in the identification of variants in low complexity regions (tandem repeats, homopolymers, pseudogenes, segmental duplications and homologous genes) and identify SVs that are challenging to detect in short-read sequencing (Van Schil et al., 2018, Ebbert et al., 2019). An additional benefit of using long read sequencing in variant identification is that individual reads are long enough to encompass multiple SNVs to establish the phase of variants. This would make the genomic data more informative and would maximise the translational potential of the genomic findings as the sequencing data allows for the construction of haplotypes for analysis.

There are two long read sequencing technologies that are currently being widely used, ONT sequencing (section 1.19.1) and PacBio. Unlike ONT sequencing, the high fidelity (HiFi) sequencing technology provided by PacBio can generate long reads of up to 25 kb with an accuracy of 99.9% to support the detection of SNVs, INDELs including SVs for use in a clinical setting (Lin et al., 2024). HiFi sequencing is a single molecule real time sequencing technology (SMRT) that uses nucleotide specific fluorescent signals to decode the genome (Korlach et al., 2010). The technology involves the ligation of hairpin adaptors to both ends of the double stranded DNA to generate a pool of circularised DNA molecules known as a SMRT bell library (Travers et al., 2010). Under circular consensus sequencing provided by the Sequel® instrument, the DNA polymerase performs multiple rounds of sequencing on each SMRT bell library to generate sub-reads that are collated to produce a representative HiFi read with a Phred score of at least Q20 (Kucuk et al., 2023, Wenger et al., 2019).

HiFi whole genome sequencing offers a comprehensive diagnostic test to detect small variants, short tandem repeat expansions and SVs without the need for PCR amplification which can introduce PCR bias and artefacts (Lin et al., 2024, Ding et al., 2023). The PacBio sequencing platform can interrogate highly homologous loci of medical relevance such as the *OPN1LW* and *OPN1MW* gene cluster associated with X-linked colour vision deficiencies of protanopia (MIM:303900) and deuteranopia (MIM:303800), respectively. Variants residing in

this gene cluster can go undetected using short read NGS due to read alignment and mapping difficulties to the reference genome which contributes to the underdiagnosis of colour blindness (Haer-Wigman et al., 2022). However, the HiFi reads are not constrained by these limitations as they are long enough to encompass unique sequences that allow for the accurate alignment of *OPN1LW* and *OPN1MW* reads to the correct locus on the reference (Haer-Wigman et al., 2022). The application of HiFi sequencing is also useful in assessing mitochondrial DNA (mtDNA) for conditions like Leber hereditary optic neuropathy, since mtDNA is highly homologous to sequences in the nuclear DNA known as nuclear mitochondrial sequences (NUMT) (Lin et al., 2024). NUMTs are usually larger than the read length of short read NGS which causes a read alignment dilemma, making it difficult to differentiate nuclear DNA and mtDNA reads, which can generate misleading results (Cihlar et al., 2020). The HiFi long reads can overcome this predicament as they can span the entire mtDNA of 16.6 kb for a complete characterisation of the mitochondrial genome (Lin et al., 2024).

Comparative analysis of HiFi genome sequencing versus short read WGS in *de novo* variant detection has showed more accurate variant calls with less false positives for long read sequencing (Kucuk et al., 2023). Long read sequencing was also able to detect an average of 239176 SNVs and 418020 INDELs that were not called per trio analysis by short read sequencing. Approximately 44% (105737/239176) of these SNVs and 22% (92961/418020) of these INDELs were in regions of no coverage with WGS (Kucuk et al., 2023). Despite these benefits, the biggest barrier to entry remains the sequencing costs, which are 3-6 times more expensive than short read NGS for 30X coverage per genome (Kucuk et al., 2023). This financial evaluation does not take into account the instrument purchase and the associated costs to set up the infrastructure for HiFi sequencing, which are expensive. Due to this relatively high cost associated with acquiring a PacBio instrument in comparison to ONT technology, the Leeds Institute of Medical Research and the 100KGP can first opt for Nanopore sequencing for SV discovery in the unsolved FH cohort and if required, then SNV and SV discovery using HiFi sequencing can be outsourced.

#### 6.4.2 Optical genome mapping (OGM) in SV identification

In the event that long read sequencing did not deliver a genomic diagnosis in the FH cohort then OGM can be considered to detect very large SVs that are undetected by current DNA sequencing technologies. OGM developed by Bionano is a unique cytogenetic technology that provides a genome wide resolution of SVs ranging in size from 500 bp up to full chromosome length representation. This technology is effective in detecting insertions, deletions, inversions and chromosomal aberrations that are missed by current short read and long read sequencing technologies. OGM outperforms long read sequencing in SV characterisation to resolve complex genomic rearrangements involving more than one SV (Sund et al., 2024). The technology was able to recognise interchromosomal translocation of 2q inserted into a locus on chromosome 4 with multiple inversions and translocation events that were missed by long read sequencing (Sund et al., 2024). OGM uses ultra-long DNA molecules to overcome the size constraints of long read sequencing in large SV detection while also providing allelic phasing information (Neveling et al., 2021). This allows for the identification of all known SV classes, and these comprise deletions (>700 bp), insertions (>500 bp), CNVs (>500 kb), interspersed, tandem or segmental duplications (>30 kb), pericentric or paracentric inversions (>30 kb) and intrachromosomal or interchromosomal translocations (>70 kb) (<https://bionano.com/saphyr-systems>). This is a significant improvement over karyotyping, which is limited to the identification of only large chromosomal rearrangements (5-10 Mb) due to the poor resolution, microarrays, which cannot identify balanced translocations and inversions, and fluorescence in situ hybridisation (FISH), which uses highly targeted probes for a specific variant and therefore overlooks other SVs not accounted for (Mantere et al., 2021, Sund et al., 2024).

The underlying principle of OGM entails the use ultra-high molecular weight DNA of >250 kb as a substrate for enzyme DLE-1 to add fluorescent labels to a recurrent 6 bp motif (CTTAAG) that occurs on average every 5 kb in the genome (Sund et al., 2024). The Saphyr™ imaging system visualises the labelling patterns on the linear DNA molecules and uses positional information to generate a consensus genomic map using *de novo* assembly (Mantere et al., 2021). The

label patterns in the sample are compared to a reference genome annotated with the labelled motifs to detect changes in the pattern, spacing or quantity of the labels that are indicative of a structural variation or CNV (Dremsek et al., 2021, Neveling et al., 2021).

The additional benefits provided by OGM include a high sensitivity to detect low level somatic variants at 5% variant allele frequency (VAF) and the ability to determine mosaicism and loss of heterozygosity, which makes it a very attractive technology in haemato-oncology diagnostics (Neveling et al., 2021). The application of OGM in germline variant identification such as in rare inherited disease like FH, can maximise the diagnostic yield by revealing SVs missed by previous genetic and cytogenetic analyses. For example, OGM performed on a proband with unsolved Usher syndrome has revealed a 173 Mb inversion with breakpoints in *USH2A* that was overlooked and discarded as a false positive in previous WGS analysis (de Bruijn et al., 2023). Moreover, this whole genome imaging technology, when applied to a cohort with neurodevelopmental disorders, identified five pathogenic and likely pathogenic SVs and two SVs classified as VUSs that were previously missed by WES (Schrauwen et al., 2024). The breakpoints of these SVs were in noncoding regions of the genome and if WGS was considered, only the smaller SVs may have been detected.

The weakness of OGM includes poor detection of balanced SVs with breakpoints in heterochromatin regions coinciding with the centromere and telomeres or the p-arm of acrocentric chromosomes (chromosome 13, 14, 15, 21 and 22) (Mantere et al., 2021, Neveling et al., 2021). These regions are not adequately represented in the current reference genome (GRCh38) due to being very large (> 1 Mb) and rich in repetitive sequences (Nurk et al., 2022). These heterochromatin regions represent a blind spot in medical genomics as they are not effectively covered by DNA sequencing nor OGM. Balanced SVs with breakpoints at the centromere like pericentric inversions or chromosomal rearrangements such as Robertsonian translocations will most likely evade detection by OGM (Dremsek et al., 2021). Furthermore, the internal control database used to scrutinise allele frequencies for very large SVs is of a low sample size of 279 genomes (Schrauwen et al., 2024). The SVs within the detectable range of OGM are also infrequently reported in existing population databases like gnomAD SV and DGV. The limited number

of SVs in the available database may mislead SV interpretation due to skewed frequencies that are not representative of the general population (Dremsek et al., 2021). Since OGM is a relatively new technology, over time more users can submit their findings to publicly available SV databases including the internal control database. This will increase the knowledge base for this technology, leading to more accurate SV filtering based on allele fractions.

## **6.5 Diagnostic limitation of the 100KGP**

The application of WGS in the 100KGP has been effective in improving the diagnostic rate in rare diseases through the broad spectrum of variants detected and the identification of novel disease gene associations, owing to the optimised laboratory and analytical workflows for large scale analysis (Daniel et al., 2022). The 100KGP emphasis on quality assurance concentrated mainly on the genomic data and the bioinformatic analysis pipelines. However, the clinical aspect of the project required improvement (Best et al., 2022a). The lack of standardised diagnostic tests being performed and the inconsistent reporting of clinical manifestations at recruitment resulted in reports of patient's phenotypes being often incomplete or incorrect. This was evident in the analysis of the biallelic *SLC38A8* cohort, where 3 patients (C2-C4) with the same genotype of homozygous NM\_001080442.3:c.264C>G, NP\_001073911.1:p.(Tyr88\*) displayed varying ocular phenotypes with no reports of FH, chiasmal misrouting or hypopigmentation status (Table 3.1). Furthermore, two probands of African (P18 and P20) descent with a genomic diagnosis of oculocutaneous albinism were not reported with what must presumably be a very conspicuous phenotype of albinism (Table 4.7 and Figure 4.17). This underreporting of disease hallmarks may be falsely interpreted as absence of key diagnostic features, leading to misdiagnosis or underdiagnosis. Furthermore, the clinical descriptions are frequently reported using HPO terms that are often vague or redundant. Uninformative clinical data may prolong the genomic investigations due to an increased number of genes analysed to compensate for the less specific phenotype, which increases the burden of variant interpretation and prolongs the turnaround times for results.

Accurate phenotyping has significant impact on the direction of the genetic analysis. The variant annotation pipelines of the 100KGP are contingent on phenotypic data to identify plausible variants underlying the participant's symptoms. The Tiering workflow applies gene panels that are disease specific, meaning that erroneous phenotypic descriptions will result in incorrect gene panels being screened, thus causing pathogenic variants to be missed (Best et al., 2022a). Furthermore, the GEL variant prioritisation framework of Exomiser relies on the reported phenotype to generate a final score, reflecting a given variant's likely pathogenicity (variant score) in light of the phenotypic match to the known phenotype associated with the gene in question (gene-pheno score)(Robinson et al., 2014). The utility of Exomiser in variant interpretation and disease gene discovery is therefore hindered when inadequate phenotyping data is present, since this leads to a lower gene-pheno score being assigned to a pathogenic variant.

In order to maximise the translational potential of genomic discoveries in the 100KGP, viable communication networks with clinicians are required, both for confirmation and publication of results and to ensure the results inform patient care. The response rate of communication for the 100KGP is relatively low. This is probably for several reasons, including clinicians retiring or having obsolete contact information (Best et al., 2022a). Similarly, updates on participants' diagnostic progress are often not disclosed to the 100KGP. In some cases, delivering a genomic diagnosis to clinicians reveals that the patient had recently received a diagnosis outside of the 100KGP. For example, this was the case for proband C24 identified as homozygous for an *SLC38A8* inversion, NC\_000016.10:g.84015973-84027074inv in the 100KGP (Figure 3.5), who was recruited to the 100KGP despite being a solved FVH2 in our local cohort (Figure 3.1). The disconnect between the clinical domain and the research entity of the 100KGP hinders collaborative efforts and reduces the value of genomic findings.

In hindsight clinical investigations of FH in the 100KGP should opt for a reverse phenotyping approach to analyse a gene specific cohort irrespective of the phenotypes reported. This would circumvent the limitations of targeted analysis guided by the phenotype, which is prone to missing variants due screening the

wrong gene panels as a result of inconsistencies in phenotypic reporting. This genotype to phenotype analytic approach was applied in SV discovery in 27 FH genes and has identified 8 pathogenic SVs and 1 SV that is a VUS (Table 4.7). The analysis showcased the diagnostic potential of this analytical approach, which would have further increased the diagnostic yield in FH if SNV discovery was also performed in this manner.

## 6.6 Future direction

### 6.6.1 Insight into molecular consequences of *SLC38A8* variants in *FVH2*

Further insight into the effects of disease-causing missense variants on the protein utilised computational 3D protein models to gauge biochemical and structural alterations (section 3.2.4). A second experiment was also performed which aimed to reveal the cellular implications of *SLC38A8* loss of function to further the understanding of disease mechanisms in *SLC38A8* (section 3.2.6).

#### 6.6.1.1 Elucidating the pathogenic mechanisms of missense variants

The *SLC38A8* *In-silico* protein modelling described in section 3.2.4.2 provided computational evidence suggesting that the protein models containing pathogenic missense variants tend to have greater electrostatic potentials at the affected residues ( $p=0.006$ ) (Figure 3.10) than models containing common polymorphic variants. Functional evidence would help to further validate this hypothesis. The generation of mutant proteins with either the deleterious or benign missense variants require a genome editing of *SLC38A8* in an appropriate cell line (section 6.6.1.2). Gauging changes in electrostatic potentials in mutant proteins with the 22 documented disease-causing missense variants and 10 benign missense variants could be carried out by research groups with expertise in nuclear magnetic resonance spectroscopy (Song et al., 2021). A follow up experiment using a protein uptake assay could be performed as described in a previous study (Hagglund et al., 2015) to quantify the amount of L-glutamine transported by the wild type *SLC38A8* protein and multiple mutant proteins each carrying a single pathogenic or benign variant. Additionally, the binding of the co-

factor, Na<sup>+</sup> should be assessed in parallel. A significant reduction in the transport of substrates across the ion channel by the SLC38A8 harbouring pathogenic missense variants would provide evidence for a loss of function as the model for disease in *SLC38A8* related FH. The outcome of this study may also elucidate the underlying pathogenic mechanism of missense variants in SLC38A8 and the insight could be relevant to transmembrane proteins beyond the SLC38 protein family, depending on the degree of homology in protein structures.

Currently, there are only two reports of functional studies involving SLC38A8 in the literature. One used antisense oligonucleotides to knock down the ortholog of *SLC38A8* in medaka fish (Poulter et al., 2013) and the other overexpressed *SLC38A8* in frog oocytes to measure protein function (Hagglund et al., 2015). With this proposed third study, additional supporting evidence for the functional data criteria of ACMG would be made available to aid in variant classification of *SLC38A8* variants.

#### **6.6.1.2 Finding an appropriate cell line expressing *SLC38A8***

In this project an attempt was made to model FVH2 in cells using CRISPR-Cas9 gene editing to knock out SLC38A8. However, this experiment was halted because appropriate cells expressing SLC38A8 were not detected. Assessing the expression levels of *SLC38A8* using RT-PCR in 12 RNA samples from various cell lines including RPE1, ARPE1 and ARPE19 and tissues not limited to the brain, cerebellum and retina did not yield any detectable levels of *SLC38A8* in all samples tested (Figure 3.16 and supplementary Figure 3.4). Based on the published literature, the protein is present in the adult brain and neuronal retina (Poulter et al., 2013). However, data from the Human Protein Atlas database (section 2.6.7) contradicts this. The tissue expression profile of *SLC38A8* shows high expression in the brain, mainly the amygdala and cerebral cortex, followed by the ovary, and with no detectable levels in the retina. The inability to detect *SLC38A8* expression in this thesis means that, if this line of research is to continue, it may be necessary to carry out further testing of additional tissues and cell lines of neuronal or retinal origin. As a last resort, HEK293T cells can be



transfected with an expression vector carrying the *SLC38A8* gene to artificially generate the desired cell line.

### **6.6.2 Maximising the translational potential of genomic findings from the 100KGP**

The analytic endeavours using NGS has utilised the bioinformatic tools SVRare to annotate and facilitate the identification of plausible SVs in select genes responsible for an FH encompassing disorders. In total, there has been 9 unique SVs with 8 characterised as pathogenic or likely pathogenic and one being a VUS (Table 4.7 and Figure 3.5). The SVs were predicted to disrupt the following genes, *GPR143* (3), *PAX6* (2), *OCA2* (2), *SLC38A8* (1) and *CACNA1F* (1). Amongst these SV was NC\_000015.10: g.28017719\_28020673del that was detected in 22 100KGP participants of a confirmed African descent (Figure 4.18).

#### **6.6.2.1 Phasing alleles in an African founder haplotype responsible for albinism**

One of the most significant findings obtained in the course of this project was the observation in the 100KGP data of a founder SV (NC\_000015.10: g.28017719\_28020673del) in *OCA2* that is a very common pathogenic allele found in people of African origin. This finding of is almost certainly not novel but represents a confirmation and characterisation of a much older report from pre-NGS era publications of a poorly defined 2.7 kb deletion in *OCA2* found at relatively high frequency in Africans. The identification of a founder SV in *OCA2* that is responsible for oculocutaneous albinism in African descendants prompted an analysis of inferred haplotypes using short read NGS data (Table 4.6). However, the scope of this assessment was limited to three probands whom WGS data was also available for both parents. Using short read sequencing data (reads are ~300 bp) of the available parents in the 100KGP, the paternal and maternal alleles can be phased to generate an informative and contiguous haplotype in the proband. Future work in this regard entails performing third generation sequencing to confirm the allelic phase by generating long reads exceeding 10 kb, that are more likely to encompass multiple distant alleles in a single read to deduce the haplotype and would also confirm detected SV

(NC\_000015.10: g.28017719\_28020673del) in the proband (Gupta et al., 2023). Based on the expertise of Leeds Institute of Medical Research in Nanopore sequencing, the ONT platform could be selected, using the Flongle flow cell as it is inexpensive and can generate adequate coverage exceeding 250000 X for amplicons of 5 kb size (McClinton et al., 2023a).

The proof of concept for using long reads in allele phasing was obtained through accurate haplotyping across the *USH2A* locus of 800.6 kb using ONT sequencing on genomic DNA and using the adaptive sampling modality (section 1.19.1) for target enrichment, which yielded a 12 fold increase in coverage (Nakamichi et al., 2023). The use of LR-PCR to generate amplicons for Nanopore sequencing was reported to generate chimeric reads of (12-42%) that impeded haplotype interpretation of the *ABCA4* locus (McClinton et al., 2023b). This issue could be addressed with the use of adaptive sampling on genomic DNA through specifying *OCA2* genomic coordinates and the bioinformatic outlined in the two studies (Nakamichi et al., 2023, Gupta et al., 2023). In preparation for this genomic analysis, a clinical collaboration request has been made through the 100KGP in hopes to reach out to the clinicians dealing with the 30 participants harbouring the NC\_000015.10:g.28017719\_28020673del in *OCA2* for collaborative purposes (Figure 4.18A).

### **6.6.3 Assessing *LAMP1* as a candidate gene underlying FH**

Previous work by the Toomes group suggested a homozygous variant NM\_005561.4:c.1109delG in *LAMP1* as a possible cause of FH in a consanguineous family (F5) (Lord, 2018). *LAMP1* along with its paralog, *LAMP2*, is a highly N-glycosylated type I transmembrane protein that is present on the membranes of lysosomes (Andrejewski et al., 1999). However, *LAMP2* variants are associated with an X-linked dominant clinical phenotype of Danon disease, while *LAMP1* has no known disease associations. The work presented in this thesis included a reanalysis of WES data from proband F1288 using autozygosity mapping and subsequent variant filtering, interpretation and segregation analysis of candidate variants to confirm the identification of the *LAMP1* variant as the most plausible candidate (section 5.2.1-5.2.5). Two functional experiments were

designed to further explore this hypothesis, detailed below, but were not completed during the doctoral study due to insufficient laboratory time.

### **6.6.3.1 Testing whether the NM\_005561.4:c.1109delG variant alters *LAMP1* splicing**

To test whether this variant causes a defect in *LAMP1* splicing due to the close proximity of a canonical splice donor site that is 5 bp away (Figure 5.8), midgenes with the wildtype control sequence and *LAMP1* variant NP\_005552.3:c.1109delG were generated and validated (Figure 5.13 and 5.15). The next steps in this experiment would be to separately transfect HEK293T cells with the *LAMP1* mutant and control *LAMP1* midgenes. HEK293T is an immortalised cell line derived from human embryonic kidney T cells that expresses SV40 large T antigen to enhance replication and expression of the vectors containing the SV40 origin of replication, such as the *LAMP1* midgene constructs (Figure 5.11) (Corradi et al., 2023). The future experimental plan would be to isolate the total RNA followed by RT-PCR using primers that flank intron 6 and 8 of *LAMP1* (Reurink et al., 2022). Analysis based on the size of the amplicons with subsequent Sanger sequencing of the PCR products would reveal the exact splicing consequence if present (Sangermano et al., 2018). This experiment will be carried out by the Toomes' group at a later date.

### **6.6.3.2 Genome editing *LAMP1* in ARPE19 cell line and evaluation of protein trafficking**

Another way to test the biological significance and the likely consequences of NM\_005561.4:c.1109delG, NP\_005552.3:p.(Gly370Alafs\*14) would be to generate cellular models to provide further evidence to prove or refute the hypothesis that the variant in *LAMP1* causes FH. Prior to commencing the CRISPR-Cas9 experiment, it was intended that a pilot study would be performed to test the targeting efficiency of the designed sgRNAs on the *LAMP1* midgene splice assays (Figure 5.20). This would be followed by Sanger sequencing of the edited midgene constructs to sequence the *LAMP1* insert for confirmation of the required edits. If the results had proved satisfactory, then ARPE19 cells would then have been targeted to generate knock in and knock out models as described

in section 5.2.11). The mutant cell lines would be genotyped to confirm the presence of a cell line(s) carrying homozygous or biallelic *LAMP1* variant allele combinations, either those replicating the precise human variant or with a likely knockout allele in the same region of the gene. The edited cell lines would also be assessed for potential off target effects by the sgRNAs used, through screening selected genomic loci that are susceptible to undesired editing (Supplementary Table 5.1 and 5.2). The generation of the desired mutated cell line, if successful, will allow for *In vitro* functional evaluation of a potentially aberrant LAMP1 protein. Monitoring the protein trafficking or localisation of LAMP1 would provide further evidence to test the hypothesis that this variant causes FH. The observation of aberrant protein trafficking or mislocalisation would suggest LAMP1 protein dysfunction leading to a defect of lysosomal function. Future work in this regard will entail immunostaining using fluorescently labelled anti-LAMP1 antibodies and fluorescence microscopy to image the stained cells at different time points. Transcriptome analysis using RNA sequencing will also be considered to compare gene expression profile between the *LAMP1* knock out cells and wildtype. It is intended that this work will be continued by another member of the Toomes group.

#### **6.6.3.3 smMIP screen of *LAMP1* in unsolved IRD cases**

Additional endeavours to be performed by other members of the Toomes group is to design smMIPS (section 1.18.3) for *LAMP1* target capture for WES analysis in a European cohort of 4000 patients including 130 local cases from Leeds with unsolved IRD. This is part of a wider genomic investigation by the ERDC to screen a joint unsolved IRD cohort for variants in 100 candidate FH genes.

## References

- ABBOTT, M., MCKENZIE, L., MORAN, B. V. G., HEIDENREICH, S., HERNANDEZ, R., HOCKING-MENNIE, L., CLARK, C., GOMES, J., LAMPE, A., BATY, D., MCGOWAN, R., MIEDZYBRODZKA, Z. & RYAN, M. 2022. Continuing the sequence? Towards an economic evaluation of whole genome sequencing for the diagnosis of rare diseases in Scotland. *J Community Genet*, 13, 487-501.
- ADZHUBEI, I. A., SCHMIDT, S., PESHKIN, L., RAMENSKY, V. E., GERASIMOVA, A., BORK, P., KONDRASHOV, A. S. & SUNYAEV, S. R. 2010. A method and server for predicting damaging missense mutations. *Nat Methods*, 7, 248-9.
- AGOSTO, M. A., ADEOSUN, A. A. R., KUMAR, N. & WENSEL, T. G. 2021. The mGluR6 ligand-binding domain, but not the C-terminal domain, is required for synaptic localization in retinal ON-bipolar cells. *J Biol Chem*, 297, 101418.
- AIRD, D., ROSS, M. G., CHEN, W. S., DANIELSSON, M., FENNEL, T., RUSS, C., JAFFE, D. B., NUSBAUM, C. & GNIRKE, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol*, 12, R18.
- AL-ARAIMI, M., PAL, B., POULTER, J. A., VAN GENDEREN, M. M., CARR, I., CUDRNAK, T., BROWN, L., SHERIDAN, E., MOHAMED, M. D., BRADBURY, J., ALI, M., INGLEHEARN, C. F. & TOOMES, C. 2013. A new recessively inherited disorder composed of foveal hypoplasia, optic nerve decussation defects and anterior segment dysgenesis maps to chromosome 16q23.3-24.1. *Mol Vis*, 19, 2165-72.
- ALDERSON, T. R., LEE, J. H., CHARLIER, C., YING, J. & BAX, A. 2018. Propensity for cis-Proline Formation in Unfolded Proteins. *ChemBiochem*, 19, 37-42.
- ALKAN, C., SAJJADIAN, S. & EICHLER, E. E. 2011. Limitations of next-generation genome sequence assembly. *Nat Methods*, 8, 61-5.
- AMBERGER, J. S., BOCCHINI, C. A., SCHIETTECATTE, F., SCOTT, A. F. & HAMOSH, A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res*, 43, D789-98.
- ANCANS, J., TOBIN, D. J., HOOGDUIJN, M. J., SMIT, N. P., WAKAMATSU, K. & THODY, A. J. 2001. Melanosomal pH controls rate of melanogenesis, eumelanin/phaeomelanin ratio and melanosome maturation in melanocytes and melanoma cells. *Exp Cell Res*, 268, 26-35.
- ANDREJEWSKI, N., PUNNONEN, E. L., GUHDE, G., TANAKA, Y., LULLMANN-RAUCH, R., HARTMANN, D., VON FIGURA, K. & SAFTIG, P. 1999. Normal lysosomal morphology and function in LAMP-1-deficient mice. *J Biol Chem*, 274, 12692-701.
- ANTUNES, H., PEREIRA, A. & CUNHA, I. 2013. Chediak-Higashi syndrome: pathognomonic feature. *Lancet*, 382, 1514.
- APKARIAN, P., REITS, D., SPEKREIJSE, H. & VAN DORP, D. 1983. A decisive electrophysiological test for human albinism. *Electroencephalogr Clin Neurophysiol*, 55, 513-31.
- ARTAL, P. 2016. The Eye as an Optical Instrument. In: AL-AMRI, M. D., EL-GOMATI, M. & ZUBAIRY, M. S. (eds.) *Optics in Our Time*. Cham: Springer International Publishing.

- ASHERY-PADAN, R. & GRUSS, P. 2001. Pax6 lights-up the way for eye development. *Curr Opin Cell Biol*, 13, 706-14.
- ASHKENAZY, H., EREZ, E., MARTZ, E., PUPKO, T. & BEN-TAL, N. 2010. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*, 38, W529-33.
- ATHER, S., PROUDLOCK, F. A., WELTON, T., MORGAN, P. S., SHETH, V., GOTTLÖB, I. & DINEEN, R. A. 2019. Aberrant visual pathway development in albinism: From retina to cortex. *Hum Brain Mapp*, 40, 777-788.
- AZUMA, N., NISHINA, S., YANAGISAWA, H., OKUYAMA, T. & YAMADA, M. 1996. PAX6 missense mutation in isolated foveal hypoplasia. *Nat Genet*, 13, 141-2.
- BADOLATO, R. & PAROLINI, S. 2007. Novel insights from adaptor protein 3 complex deficiency. *J Allergy Clin Immunol*, 120, 735-41; quiz 742-3.
- BAKKER, R., WAGSTAFF, E. L., KRUIJT, C. C., EMRI, E., VAN KARNEBEEK, C. D. M., HOFFMANN, M. B., BROOKS, B. P., BOON, C. J. F., MONTOLIU, L., VAN GENDEREN, M. M. & BERGEN, A. A. 2022. The retinal pigmentation pathway in human albinism: Not so black and white. *Prog Retin Eye Res*, 91, 101091.
- BARRET, D. C. A., KAUPP, U. B. & MARINO, J. 2022. The structure of cyclic nucleotide-gated channels in rod and cone photoreceptors. *Trends Neurosci*, 45, 763-776.
- BASSI, M. T., SCHIAFFINO, M. V., RENIERI, A., DE NIGRIS, F., GALLI, L., BRUTTINI, M., GEBBIA, M., BERGEN, A. A., LEWIS, R. A. & BALLABIO, A. 1995. Cloning of the gene for ocular albinism type 1 from the distal short arm of the X chromosome. *Nat Genet*, 10, 13-9.
- BELLONO, N. W., ESCOBAR, I. E. & OANCEA, E. 2016. A melanosomal two-pore sodium channel regulates pigmentation. *Sci Rep*, 6, 26570.
- BENES, V., KILGER, C., VOSS, H., PAABO, S. & ANSORGE, W. 1997. Direct primer walking on P1 plasmid DNA. *Biotechniques*, 23, 98-100.
- BENTLEY, D. R., BALASUBRAMANIAN, S., SWERDLOW, H. P., SMITH, G. P., MILTON, J., BROWN, C. G., HALL, K. P., EVERS, D. J., BARNES, C. L., BIGNELL, H. R., BOUTELL, J. M., BRYANT, J., CARTER, R. J., KEIRA CHEETHAM, R., COX, A. J., ELLIS, D. J., FLATBUSH, M. R., GORMLEY, N. A., HUMPHRAY, S. J., IRVING, L. J., KARBELASHVILI, M. S., KIRK, S. M., LI, H., LIU, X., MAISINGER, K. S., MURRAY, L. J., OBRADOVIC, B., OST, T., PARKINSON, M. L., PRATT, M. R., RASOLONJATOVO, I. M., REED, M. T., RIGATTI, R., RODIGHIERO, C., ROSS, M. T., SABOT, A., SANKAR, S. V., SCALLY, A., SCHROTH, G. P., SMITH, M. E., SMITH, V. P., SPIRIDOU, A., TORRANCE, P. E., TZONEV, S. S., VERMAAS, E. H., WALTER, K., WU, X., ZHANG, L., ALAM, M. D., ANASTASI, C., ANIEBO, I. C., BAILEY, D. M., BANCARZ, I. R., BANERJEE, S., BARBOUR, S. G., BAYBAYAN, P. A., BENOIT, V. A., BENSON, K. F., BEVIS, C., BLACK, P. J., BOODHUN, A., BRENNAN, J. S., BRIDGHAM, J. A., BROWN, R. C., BROWN, A. A., BUERMANN, D. H., BUNDU, A. A., BURROWS, J. C., CARTER, N. P., CASTILLO, N., CHIARA, E. C. M., CHANG, S., NEIL COOLEY, R., CRAKE, N. R., DADA, O. O., DIAKOU MAKOS, K. D., DOMINGUEZ-FERNANDEZ, B., EARNSHAW, D. J., EGBUJOR, U. C., ELMORE, D. W., ETCHIN, S. S., EWAN, M. R., FEDURCO, M., FRASER, L. J., FUENTES FAJARDO, K. V., SCOTT FUREY, W., GEORGE, D., GIETZEN, K. J., GODDARD, C. P., GOLDA, G. S., GRANIERI, P. A., GREEN, D. E., GUSTAFSON, D. L., HANSEN,

- N. F., HARNISH, K., HAUDENSCHILD, C. D., HEYER, N. I., HIMMS, M. M., HO, J. T., HORGAN, A. M., et al. 2008. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456, 53-9.
- BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Res*, 28, 235-42.
- BERNARDIS, I., CHIESI, L., TENEDINI, E., ARTUSO, L., PERCESEPE, A., ARTUSI, V., SIMONE, M. L., MANFREDINI, R., CAMPARINI, M., RINALDI, C., CIARDELLA, A., GRAZIANO, C., BALDUCCI, N., TRANCHINA, A., CAVALLINI, G. M., PIETRANGELO, A., MARIGO, V. & TAGLIAFICO, E. 2016. Unravelling the Complexity of Inherited Retinal Dystrophies Molecular Testing: Added Value of Targeted Next-Generation Sequencing. *Biomed Res Int*, 2016, 6341870.
- BEST, S., INGLEHEARN, C. F., WATSON, C. M., TOOMES, C., WHEWAY, G. & JOHNSON, C. A. 2022a. Unlocking the potential of the UK 100,000 Genomes Project-lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort. *Am J Med Genet C Semin Med Genet*, 190, 5-8.
- BEST, S., YU, J., LORD, J., ROCHE, M., WATSON, C. M., BEVERS, R. P. J., STUCKEY, A., MADHUSUDHAN, S., JEWELL, R., SISODIYA, S. M., LIN, S., TURNER, S., ROBINSON, H., LESLIE, J. S., BAPLE, E., GENOMICS ENGLAND RESEARCH, C., TOOMES, C., INGLEHEARN, C., WHEWAY, G. & JOHNSON, C. A. 2022b. Uncovering the burden of hidden ciliopathies in the 100 000 Genomes Project: a reverse phenotyping approach. *J Med Genet*, 59, 1151-1164.
- BHARTI, S., BHATIA, P., BANSAL, D. & VARMA, N. 2013. The accelerated phase of chediak-higashi syndrome: the importance of hematological evaluation. *Turk J Haematol*, 30, 85-7.
- BISHOP, P. N., TAKANOSU, M., LE GOFF, M. & MAYNE, R. 2002. The role of the posterior ciliary body in the biosynthesis of vitreous humour. *Eye (Lond)*, 16, 454-60.
- BOISSY, R. E., ZHAO, H., OETTING, W. S., AUSTIN, L. M., WILDENBERG, S. C., BOISSY, Y. L., ZHAO, Y., STURM, R. A., HEARING, V. J., KING, R. A. & NORDLUND, J. J. 1996. Mutation in and lack of expression of tyrosinase-related protein-1 (TRP-1) in melanocytes from an individual with brown oculocutaneous albinism: a new subtype of albinism classified as "OCA3". *Am J Hum Genet*, 58, 1145-56.
- BOOIJ, J. C., BAAS, D. C., BEISEKEEVA, J., GORGELS, T. G. & BERGEN, A. A. 2010. The dynamic nature of Bruch's membrane. *Prog Retin Eye Res*, 29, 1-18.
- BOOTE, C., SIGAL, I. A., GRITZ, R., HUA, Y., NGUYEN, T. D. & GIRARD, M. J. A. 2020. Scleral structure and biomechanics. *Prog Retin Eye Res*, 74, 100773.
- BOYA, P., ESTEBAN-MARTINEZ, L., SERRANO-PUEBLA, A., GOMEZ-SINTES, R. & VILLAREJO-ZORI, B. 2016. Autophagy in the eye: Development, degeneration, and aging. *Prog Retin Eye Res*, 55, 206-245.
- BRINGMANN, A., SYRBE, S., GORNER, K., KACZA, J., FRANCKE, M., WIEDEMANN, P. & REICHENBACH, A. 2018. The primate fovea: Structure, function and development. *Prog Retin Eye Res*, 66, 49-84.
- BRINGMANN, A. & WIEDEMANN, P. 2022. Chapter 2 - Basic structure of the retina. In: BRINGMANN, A. & WIEDEMANN, P. (eds.) *The Fovea*. Academic Press.

- BRITTEN-JONES, A. C., GOCUK, S. A., GOH, K. L., HUQ, A., EDWARDS, T. L. & AYTON, L. N. 2023. The Diagnostic Yield of Next Generation Sequencing in Inherited Retinal Diseases: A Systematic Review and Meta-analysis. *Am J Ophthalmol*, 249, 57-73.
- BROWN, S. D., DREOLINI, L., WILSON, J. F., BALASUNDARAM, M. & HOLT, R. A. 2023. Complete sequence verification of plasmid DNA using the Oxford Nanopore Technologies' MinION device. *BMC Bioinformatics*, 24, 116.
- CAMPBELL, P., ELLINGFORD, J. M., PARRY, N. R. A., FLETCHER, T., RAMSDEN, S. C., GALE, T., HALL, G., SMITH, K., KASPERAVICIUTE, D., THOMAS, E., LLOYD, I. C., DOUZGOU, S., CLAYTON-SMITH, J., BISWAS, S., ASHWORTH, J. L., BLACK, G. C. M. & SERGOUNIOTIS, P. I. 2019. Clinical and genetic variability in children with partial albinism. *Sci Rep*, 9, 16576.
- CARR, I. M., BHASKAR, S., O'SULLIVAN, J., ALDAHMEH, M. A., SHAMSELDIN, H. E., MARKHAM, A. F., BONTHRON, D. T., BLACK, G. & ALKURAYA, F. S. 2013. Autozygosity mapping with exome sequence data. *Hum Mutat*, 34, 50-6.
- CARROLL, W. M., JAY, B. S., MCDONALD, W. I. & HALLIDAY, A. M. 1980. Pattern evoked potentials in human albinism. Evidence of two different topographical asymmetries reflecting abnormal retino-cortical projections. *J Neurol Sci*, 48, 265-86.
- CARSS, K. J., ARNO, G., ERWOOD, M., STEPHENS, J., SANCHIS-JUAN, A., HULL, S., MEGY, K., GROZEVA, D., DEWHURST, E., MALKA, S., PLAGNOL, V., PENKETT, C., STIRRUPS, K., RIZZO, R., WRIGHT, G., JOSIFOVA, D., BITNER-GLINDZICZ, M., SCOTT, R. H., CLEMENT, E., ALLEN, L., ARMSTRONG, R., BRADY, A. F., CARMICHAEL, J., CHITRE, M., HENDERSON, R. H. H., HURST, J., MACLAREN, R. E., MURPHY, E., PATERSON, J., ROSSER, E., THOMPSON, D. A., WAKELING, E., OUWEHAND, W. H., MICHAELIDES, M., MOORE, A. T., CONSORTIUM, N. I.-B. R. D., WEBSTER, A. R. & RAYMOND, F. L. 2017. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am J Hum Genet*, 100, 75-90.
- CARTEGNI, L., WANG, J., ZHU, Z., ZHANG, M. Q. & KRAINER, A. R. 2003. ESEfinder: A web resource to identify exonic splicing enhancers. *Nucleic Acids Res*, 31, 3568-71.
- CASEY, M. A., LUSK, S. & KWAN, K. M. 2021. Build me up optic cup: Intrinsic and extrinsic mechanisms of vertebrate eye morphogenesis. *Dev Biol*, 476, 128-136.
- CAULFIELD, M., DAVIES, J., DENNYS, M., ELBAHY, L., FOWLER, T., HILL, S., HUBBARD, T., JOSTINS, L., MALTBY, N., MAHON-PEARSON, J., MCVEAN, G., NEVIN-RIDLEY, K., PARKER, M., PARRY, V., RENDON, A., RILEY, L., TURNBULL, C. & WOODS, K. 2020. National Genomic Research Library. figshare.
- CHAN, H. W., SCHIFF, E. R., TAILOR, V. K., MALKA, S., NEVEU, M. M., THEODOROU, M. & MOOSAJEE, M. 2021. Prospective Study of the Phenotypic and Mutational Spectrum of Ocular Albinism and Oculocutaneous Albinism. *Genes (Basel)*, 12.
- CHANG, C. H. & LARRACUENTE, A. M. 2019. Heterochromatin-Enriched Assemblies Reveal the Sequence and Organization of the *Drosophila melanogaster* Y Chromosome. *Genetics*, 211, 333-348.



- CHEN, S., FRANCIOLI, L. C., GOODRICH, J. K., COLLINS, R. L., KANAI, M., WANG, Q., ALFÖLDI, J., WATTS, N. A., VITTAL, C., GAUTHIER, L. D., POTERBA, T., WILSON, M. W., TARASOVA, Y., PHU, W., YOHANNES, M. T., KOENIG, Z., FARJOUN, Y., BANKS, E., DONNELLY, S., GABRIEL, S., GUPTA, N., FERRIERA, S., TOLONEN, C., NOVOD, S., BERGELSON, L., ROAZEN, D., RUANO-RUBIO, V., COVARRUBIAS, M., LLANWARNE, C., PETRILLO, N., WADE, G., JEANDET, T., MUNSHI, R., TIBBETTS, K., CONSORTIUM, G. P., O'DONNELL-LURIA, A., SOLOMONSON, M., SEED, C., MARTIN, A. R., TALKOWSKI, M. E., REHM, H. L., DALY, M. J., TIAO, G., NEALE, B. M., MACARTHUR, D. G. & KARCZEWSKI, K. J. 2022. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. *bioRxiv*, 2022.03.20.485034.
- CHENG, X. T., XIE, Y. X., ZHOU, B., HUANG, N., FARFEL-BECKER, T. & SHENG, Z. H. 2018. Characterization of LAMP1-labeled nondegradative lysosomal and endocytic compartments in neurons. *J Cell Biol*, 217, 3127-3139.
- CHIANG, J. P. W., LAMEY, T. M., WANG, N. K., DUAN, J., ZHOU, W., MCLAREN, T. L., THOMPSON, J. A., RUDDLE, J. & DE ROACH, J. N. 2018. Development of High-Throughput Clinical Testing of RPGR ORF15 Using a Large Inherited Retinal Dystrophy Cohort. *Invest Ophthalmol Vis Sci*, 59, 4434-4440.
- CHOI, J. H., JUNG, J. H., OH, E. H., SHIN, J. H., KIM, H. S., SEO, J. H., CHOI, S. Y., KIM, M. J., CHOI, H. Y., LEE, C. & CHOI, K. D. 2018. Genotype and Phenotype Spectrum of FRMD7-Associated Infantile Nystagmus Syndrome. *Invest Ophthalmol Vis Sci*, 59, 3181-3188.
- CHOI, M., SCHOLL, U. I., JI, W., LIU, T., TIKHONOVA, I. R., ZUMBO, P., NAYIR, A., BAKKALOGLU, A., OZEN, S., SANJAD, S., NELSON-WILLIAMS, C., FARHI, A., MANE, S. & LIFTON, R. P. 2009. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci U S A*, 106, 19096-101.
- CIHLAR, J. C., STROBL, C., LAGACE, R., MUENZLER, M., PARSON, W. & BUDOWLE, B. 2020. Distinguishing mitochondrial DNA and NUMT sequences amplified with the precision ID mtDNA whole genome panel. *Mitochondrion*, 55, 122-133.
- CONSORTIUM, E. P., MOORE, J. E., PURCARO, M. J., PRATT, H. E., EPSTEIN, C. B., SHORESH, N., ADRIAN, J., KAWLI, T., DAVIS, C. A., DOBIN, A., KAUL, R., HALOW, J., VAN NOSTRAND, E. L., FREESE, P., GORKIN, D. U., SHEN, Y., HE, Y., MACKIEWICZ, M., PAULI-BEHN, F., WILLIAMS, B. A., MORTAZAVI, A., KELLER, C. A., ZHANG, X. O., ELHAJJAJY, S. I., HUEY, J., DICKEL, D. E., SNETKOVA, V., WEI, X., WANG, X., RIVERA-MULIA, J. C., ROZOWSKY, J., ZHANG, J., CHHETRI, S. B., ZHANG, J., VICTORSEN, A., WHITE, K. P., VISEL, A., YEO, G. W., BURGE, C. B., LECUYER, E., GILBERT, D. M., DEKKER, J., RINN, J., MENDENHALL, E. M., ECKER, J. R., KELLIS, M., KLEIN, R. J., NOBLE, W. S., KUNDAJE, A., GUIGO, R., FARNHAM, P. J., CHERRY, J. M., MYERS, R. M., REN, B., GRAVELEY, B. R., GERSTEIN, M. B., PENNACCHIO, L. A., SNYDER, M. P., BERNSTEIN, B. E., WOLD, B., HARDISON, R. C., GINGERAS, T. R., STAMATOYANNOPOULOS, J. A. & WENG, Z. 2020. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*, 583, 699-710.

- CONSORTIUM, U. K., WALTER, K., MIN, J. L., HUANG, J., CROOKS, L., MEMARI, Y., MCCARTHY, S., PERRY, J. R., XU, C., FUTEMA, M., LAWSON, D., IOTCHKOVA, V., SCHIFFELS, S., HENDRICKS, A. E., DANECEK, P., LI, R., FLOYD, J., WAIN, L. V., BARROSO, I., HUMPHRIES, S. E., HURLES, M. E., ZEGGINI, E., BARRETT, J. C., PLAGNOL, V., RICHARDS, J. B., GREENWOOD, C. M., TIMPSON, N. J., DURBIN, R. & SORANZO, N. 2015. The UK10K project identifies rare variants in health and disease. *Nature*, 526, 82-90.
- CORNISH, K. S., REDDY, A. R. & MCBAIN, V. A. 2014. Concentric macular rings sign in patients with foveal hypoplasia. *JAMA Ophthalmol*, 132, 1084-8.
- CORRADI, Z., KHAN, M., HITTI-MALIN, R., MISHRA, K., WHELAN, L., CORNELIS, S. S., GROUP, A. B.-S., HOYNG, C. B., KAMPJARVI, K., KLAVER, C. C. W., LSKOVA, P., STOHR, H., WEBER, B. H. F., BANFI, S., FARRAR, G. J., SHARON, D., ZERNANT, J., ALLIKMETS, R., DHAENENS, C. M. & CREMERS, F. P. M. 2023. Targeted sequencing and in vitro splice assays shed light on ABCA4-associated retinopathies missing heritability. *HGG Adv*, 4, 100237.
- CREMERS, F. P. M., BOON, C. J. F., BUJAKOWSKA, K. & ZEITZ, C. 2018. Special Issue Introduction: Inherited Retinal Disease: Novel Candidate Genes, Genotype-Phenotype Correlations, and Inheritance Models. *Genes (Basel)*, 9.
- CROSSLEY, B. M., BAI, J., GLASER, A., MAES, R., PORTER, E., KILLIAN, M. L., CLEMENT, T. & TOOHEY-KURTH, K. 2020. Guidelines for Sanger sequencing and molecular assay monitoring. *J Vet Diagn Invest*, 32, 767-775.
- CUNNINGHAM, E., HAYS, S., WAINSTEIN, T., ZIERHUT, H., VIRANI, A. & TRYON, R. 2024. Exploring genetic counselors' experiences with non-paternity in clinical settings. *J Genet Couns*.
- CUNNINGHAM, F., ALLEN, J. E., ALLEN, J., ALVAREZ-JARRETA, J., AMODE, M. R., ARMEAN, I. M., AUSTINE-ORIMOLOYE, O., AZOV, A. G., BARNES, I., BENNETT, R., BERRY, A., BHAI, J., BIGNELL, A., BILLIS, K., BODDU, S., BROOKS, L., CHARKHCHI, M., CUMMINS, C., DA RIN FIORETTO, L., DAVIDSON, C., DODIYA, K., DONALDSON, S., EL HOUDAIGUI, B., EL NABOULSI, T., FATIMA, R., GIRON, C. G., GENEZ, T., MARTINEZ, J. G., GUIJARRO-CLARKE, C., GYMER, A., HARDY, M., HOLLIS, Z., HOURLIER, T., HUNT, T., JUETTEMANN, T., KAIKALA, V., KAY, M., LAVIDAS, I., LE, T., LEMOS, D., MARUGAN, J. C., MOHANAN, S., MUSHTAQ, A., NAVEN, M., OGEH, D. N., PARKER, A., PARTON, A., PERRY, M., PILIZOTA, I., PROSOVETSKAIA, I., SAKTHIVEL, M. P., SALAM, A. I. A., SCHMITT, B. M., SCHUILENBURG, H., SHEPPARD, D., PEREZ-SILVA, J. G., STARK, W., STEED, E., SUTINEN, K., SUKUMARAN, R., SUMATHIPALA, D., SUNER, M. M., SZPAK, M., THORMANN, A., TRICOMI, F. F., URBINA-GOMEZ, D., VEIDENBERG, A., WALSH, T. A., WALTZ, B., WILLHOFT, N., WINTERBOTTOM, A., WASS, E., CHAKIACHVILI, M., FLINT, B., FRANKISH, A., GIORGETTI, S., HAGGERTY, L., HUNT, S. E., GR, I. I., LOVELAND, J. E., MARTIN, F. J., MOORE, B., MUDGE, J. M., MUFFATO, M., PERRY, E., RUFFIER, M., TATE, J., THYBERT, D., TREVANION, S. J., DYER, S., HARRISON, P. W., HOWE, K. L., YATES, A. D., ZERBINO, D. R. & FLICEK, P. 2022. Ensembl 2022. *Nucleic Acids Res*, 50, D988-D995.
- CURCIO, C. A., SLOAN, K. R., KALINA, R. E. & HENDRICKSON, A. E. 1990. Human photoreceptor topography. *J Comp Neurol*, 292, 497-523.

- CVEKL, A. & TAMM, E. R. 2004. Anterior eye development and ocular mesenchyme: new insights from mouse models and human diseases. *Bioessays*, 26, 374-86.
- DALY, C. M., WILLER, J., GREGG, R. & GROSS, J. M. 2013. snow white, a zebrafish model of Hermansky-Pudlak Syndrome type 5. *Genetics*, 195, 481-94.
- DAME, J. A., PHILLIPS, L.-A., DE VILLIERS, N., PILLAY, K., HLELA, C. & ELEY, B. 2019. A novel LYST mutation causing Chédiak Higashi syndrome in a South African child. *Pediatric Hematology Oncology Journal*, 4, 44-46.
- DANIEL, G., GENOMICS ENGLAND RESEARCH, C., DANIELA, P., KAREN, F., EGE, S., MOHAMMED, A.-O., ARNAUD, P. J. G., KHUSHNOODA, R., ITARU, Y., NELE, B., CHANTAL, T., BRUCE, D. G., PAUL, B., VERITY, H., JULIE, H., TOMOKI, K., SAHAR, M., MITSUO, M., TAKAKO, O., HELEN, S., KHALID, T., CLAIRE, L. S. T., FAIQA, I., SAIMA, R., TAKAYUKI, M., PIA, O., BART, L., HIROKO, M., ZUBAIR, M. A., GRAEME, M. B., KATHLEEN, F., ANDREW, M. & ERNEST, T. 2022. Genetic association analysis of 269 rare diseases reveals novel aetiologies. *medRxiv*, 2022.06.10.22276270.
- DANON, M. J., OH, S. J., DIMAURO, S., MANALIGOD, J. R., EASTWOOD, A., NAIDU, S. & SCHLISELFELD, L. H. 1981. Lysosomal glycogen storage disease with normal acid maltase. *Neurology*, 31, 51-7.
- DANYSH, B. P., PATEL, T. P., CZYMMEK, K. J., EDWARDS, D. A., WANG, L., PANDE, J. & DUNCAN, M. K. 2010. Characterizing molecular diffusion in the lens capsule. *Matrix Biol*, 29, 228-36.
- DAVI, F., LANGERAK, A. W., DE SEPTENVILLE, A. L., KOLIJN, P. M., HENGEVELD, P. J., CHATZIDIMITRIOU, A., BONFIGLIO, S., SUTTON, L. A., ROSENQUIST, R., GHIA, P., STAMATOPOULOS, K., ERIC, T. E. R. I. O. C. L. L. & THE EUROCLONALITY, N. G. S. W. G. 2020. Immunoglobulin gene analysis in chronic lymphocytic leukemia in the era of next generation sequencing. *Leukemia*, 34, 2545-2551.
- DAVIS-SILBERMAN, N. & ASHERY-PADAN, R. 2008. Iris development in vertebrates; genetic and molecular considerations. *Brain Res*, 1192, 17-28.
- DE BRUIJN, S. E., FIORENTINO, A., OTTAVIANI, D., FANUCCHI, S., MELO, U. S., CORRAL-SERRANO, J. C., MULDER, T., GEORGIU, M., RIVOLTA, C., PONTIKOS, N., ARNO, G., ROBERTS, L., GREENBERG, J., ALBERT, S., GILISSEN, C., ABEN, M., REBELLO, G., MEAD, S., RAYMOND, F. L., COROMINAS, J., SMITH, C. E. L., KREMER, H., DOWNES, S., BLACK, G. C., WEBSTER, A. R., INGLEHEARN, C. F., VAN DEN BORN, L. I., KOENEKOOP, R. K., MICHAELIDES, M., RAMESAR, R. S., HOYNG, C. B., MUNDLOS, S., MHLANGA, M. M., CREMERS, F. P. M., CHEETHAM, M. E., ROOSING, S. & HARDCASTLE, A. J. 2020. Structural Variants Create New Topological-Associated Domains and Ectopic Retinal Enhancer-Gene Contact in Dominant Retinitis Pigmentosa. *Am J Hum Genet*, 107, 802-814.
- DE BRUIJN, S. E., RODENBURG, K., COROMINAS, J., BEN-YOSEF, T., REURINK, J., KREMER, H., WHELAN, L., PLOMP, A. S., BERGER, W., FARRAR, G. J., FERENC KOVACS, A., FAJARDY, I., HITTI-MALIN, R. J., WEISSCHUH, N., WEENER, M. E., SHARON, D., PENNING, R. J. E., HAER-WIGMAN, L., HOYNG, C. B., NELEN, M. R., VISSERS, L., VAN DEN BORN, L. I., GILISSEN, C., CREMERS, F. P. M., HOISCHEN, A., NEVELING, K. & ROOSING, S. 2023. Optical genome mapping and

- revisiting short-read genome sequencing data reveal previously overlooked structural variants disrupting retinal disease-associated genes. *Genet Med*, 25, 100345.
- DELAMERE, N. A. 2005. Ciliary Body and Ciliary Epithelium. *Adv Organ Biol*, 10, 127-148.
- DELEVOYE, C., HEILIGENSTEIN, X., RIPOLL, L., GILLES-MARSENS, F., DENNIS, M. K., LINARES, R. A., DERMAN, L., GOKHALE, A., MOREL, E., FAUNDEZ, V., MARKS, M. S. & RAPOSO, G. 2016. BLOC-1 Brings Together the Actin and Microtubule Cytoskeletons to Generate Recycling Endosomes. *Curr Biol*, 26, 1-13.
- DELL'ANGELICA, E. C., MULLINS, C., CAPLAN, S. & BONIFACINO, J. S. 2000. Lysosome-related organelles. *The FASEB Journal*, 14, 1265-1278.
- DELL'ANGELICA, E. C., SHOTELERSUK, V., AGUILAR, R. C., GAHL, W. A. & BONIFACINO, J. S. 1999. Altered trafficking of lysosomal proteins in Hermansky-Pudlak syndrome due to mutations in the beta 3A subunit of the AP-3 adaptor. *Mol Cell*, 3, 11-21.
- DEN HOLLANDER, A. I., KOENEKOOP, R. K., YZER, S., LOPEZ, I., ARENDS, M. L., VOESENEK, K. E., ZONNEVELD, M. N., STROM, T. M., MEITINGER, T., BRUNNER, H. G., HOYNG, C. B., VAN DEN BORN, L. I., ROHRSCHEIDER, K. & CREMERS, F. P. 2006. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet*, 79, 556-61.
- DENNIS, M. K., MANTEGAZZA, A. R., SNIR, O. L., TENZA, D., ACOSTA-RUIZ, A., DELEVOYE, C., ZORGER, R., SITARAM, A., DE JESUS-ROJAS, W., RAVICHANDRAN, K., RUX, J., SVIDERSKAYA, E. V., BENNETT, D. C., RAPOSO, G., MARKS, M. S. & SETTY, S. R. 2015. BLOC-2 targets recycling endosomal tubules to melanosomes for cargo delivery. *J Cell Biol*, 209, 563-77.
- DHEENSA, S., SAMUEL, G., LUCASSEN, A. M. & FARSIDES, B. 2018. Towards a national genomics medicine service: the challenges facing clinical-research hybrid practices and the case of the 100 000 genomes project. *J Med Ethics*, 44, 397-403.
- DI PIETRO, S. M., FALCÓN-PÉREZ, J. M. & DELL'ANGELICA, E. C. 2004. Characterization of BLOC-2, a Complex Containing the Hermansky-Pudlak Syndrome Proteins HPS3, HPS5 and HPS6. *Traffic*, 5, 276-283.
- DIMENT, S., EIDELMAN, M., RODRIGUEZ, G. M. & ORLOW, S. J. 1995. Lysosomal hydrolases are present in melanosomes and are elevated in melanizing cells. *J Biol Chem*, 270, 4213-5.
- DING, Y., OWEN, M., LE, J., BATALOV, S., CHAU, K., KWON, Y. H., VAN DER KRAAN, L., BEZARES-ORIN, Z., ZHU, Z., VEERARAGHAVAN, N., NAHAS, S., BAINBRIDGE, M., GLEESON, J., BAER, R. J., BANDOLI, G., CHAMBERS, C. & KINGSMORE, S. F. 2023. Scalable, high quality, whole genome sequencing from archived, newborn, dried blood spots. *NPJ Genom Med*, 8, 5.
- DIRECTORS, A. B. O. 2012. Points to consider in the clinical application of genomic sequencing. *Genet Med*, 14, 759-61.
- DOHM, J. C., PETERS, P., STRALIS-PAVESE, N. & HIMMELBAUER, H. 2020. Benchmarking of long-read correction methods. *NAR Genom Bioinform*, 2, lqaa037.
- DREMSEK, P., SCHWARZ, T., WEIL, B., MALASHKA, A., LACCONE, F. & NEESEN, J. 2021. Optical Genome Mapping in Routine Human Genetic Diagnostics-Its Advantages and Limitations. *Genes (Basel)*, 12.

- DU, C., PUSEY, B. N., ADAMS, C. J., LAU, C. C., BONE, W. P., GAHL, W. A., MARKELLO, T. C. & ADAMS, D. R. 2016. Explorations to improve the completeness of exome sequencing. *BMC Med Genomics*, 9, 56.
- DUMITRESCU, A. V., PFEIFER, W. L. & DRACK, A. V. 2021. Clinical phenocopies of albinism. *J AAPOS*, 25, 220 e1-220 e8.
- DUNCAN, J. L., PIERCE, E. A., LASTER, A. M., DAIGER, S. P., BIRCH, D. G., ASH, J. D., IANNACCONE, A., FLANNERY, J. G., SAHEL, J. A., ZACK, D. J., ZARBIN, M. A. & AND THE FOUNDATION FIGHTING BLINDNESS SCIENTIFIC ADVISORY, B. 2018. Inherited Retinal Degenerations: Current Landscape and Knowledge Gaps. *Transl Vis Sci Technol*, 7, 6.
- DUNO, M., WIBRAND, F., BAGGESEN, K., ROSENBERG, T., KJAER, N. & FREDERIKSEN, A. L. 2013. A novel mitochondrial mutation m.8989G>C associated with neuropathy, ataxia, retinitis pigmentosa - the NARP syndrome. *Gene*, 515, 372-5.
- DURHAM-PIERRE, D., GARDNER, J. M., NAKATSU, Y., KING, R. A., FRANCKE, U., CHING, A., AQUARON, R., DEL MARMOL, V. & BRILLIANT, M. H. 1994. African origin of an intragenic deletion of the human P gene in tyrosinase positive oculocutaneous albinism. *Nature Genetics*, 7, 176-179.
- EBBERT, M. T. W., JENSEN, T. D., JANSEN-WEST, K., SENS, J. P., REDDY, J. S., RIDGE, P. G., KAUWE, J. S. K., BELZIL, V., PREGENT, L., CARRASQUILLO, M. M., KEENE, D., LARSON, E., CRANE, P., ASMANN, Y. W., ERTEKIN-TANER, N., YOUNKIN, S. G., ROSS, O. A., RADEMAKERS, R., PETRUCELLI, L. & FRYER, J. D. 2019. Systematic analysis of dark and camouflaged genes reveals disease-relevant genes hiding in plain sight. *Genome Biol*, 20, 97.
- EBERMANN, I., KOENEKOOP, R. K., LOPEZ, I., BOU-KHZAM, L., PIGEON, R. & BOLZ, H. J. 2009. An USH2A founder mutation is the major cause of Usher syndrome type 2 in Canadians of French origin and confirms common roots of Quebecois and Acadians. *Eur J Hum Genet*, 17, 80-4.
- EHRENBERG, M., BAGDONITE-BEJARANO, L., FULTON, A. B., ORENSTEIN, N. & YAHALOM, C. 2021. Genetic causes of nystagmus, foveal hypoplasia and subnormal visual acuity- other than albinism. *Ophthalmic Genet*, 1-9.
- ELAWAD, O., DAFALLAH, M. A., AHMED, M. M. M., ALBASHIR, A. A. D., ABDALLA, S. M. A., YOUSIF, H. H. M., DAW ELBAIT, A. A. E., MOHAMMED, M. E., ALI, H. I. H., AHMED, M. M. M., MOHAMMED, N. F. N., OSMAN, F. H. M., MOHAMMED, M. A. Y. & ABU SHAMA, E. A. E. 2022. Bardet-Biedl syndrome: a case series. *J Med Case Rep*, 16, 169.
- ELLINGFORD, J. M., BARTON, S., BHASKAR, S., WILLIAMS, S. G., SERGOUNIOTIS, P. I., O'SULLIVAN, J., LAMB, J. A., PERVEEN, R., HALL, G., NEWMAN, W. G., BISHOP, P. N., ROBERTS, S. A., LEACH, R., TEARLE, R., BAYLISS, S., RAMSDEN, S. C., NEMETH, A. H. & BLACK, G. C. 2016. Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology*, 123, 1143-50.
- ERSKINE, L. & HERRERA, E. 2014. Connecting the retina to the brain. *ASN Neuro*, 6.
- ESKELINEN, E. L., SCHMIDT, C. K., NEU, S., WILLENBORG, M., FUERTES, G., SALVADOR, N., TANAKA, Y., LULLMANN-RAUCH, R., HARTMANN, D., HEEREN, J., VON FIGURA, K., KNECHT, E. & SAFTIG, P. 2004. Disturbed cholesterol traffic but normal proteolytic function in LAMP-1/LAMP-2 double-deficient fibroblasts. *Mol Biol Cell*, 15, 3132-45.

- FALCÓN-PÉREZ, J. M., STARCEVIC, M., GAUTAM, R. & DELL'ANGELICA, E. C. 2002. BLOC-1, a novel complex containing the pallidin and muted proteins involved in the biogenesis of melanosomes and platelet-dense granules. *J Biol Chem*, 277, 28191-9.
- FARAG, T. I. & TEEBI, A. S. 1989. High incidence of Bardet Biedl syndrome among the Bedouin. *Clin Genet*, 36, 463-4.
- FERRARI, S., DI IORIO, E., BARBARO, V., PONZIN, D., SORRENTINO, F. S. & PARMEGGIANI, F. 2011. Retinitis pigmentosa: genes and disease mechanisms. *Curr Genomics*, 12, 238-49.
- FINE, B. S. & ZIMMERMAN, L. E. 1962. Muller's cells and the "middle limiting membrane" of the human retina. An electron microscopic study. *Invest Ophthalmol*, 1, 304-26.
- FOKKEMA, I., KROON, M., LOPEZ HERNANDEZ, J. A., ASSCHEMAN, D., LUGTENBURG, I., HOOGENBOOM, J. & DEN DUNNEN, J. T. 2021. The LOVD3 platform: efficient genome-wide sharing of genetic variants. *Eur J Hum Genet*, 29, 1796-1803.
- FORRESTER, J. V., DICK, A. D., MCMENAMIN, P. G., ROBERTS, F. & PEARLMAN, E. 2016. Chapter 2 - Embryology and early development of the eye and adnexa. In: FORRESTER, J. V., DICK, A. D., MCMENAMIN, P. G., ROBERTS, F. & PEARLMAN, E. (eds.) *The Eye (Fourth Edition)*. W.B. Saunders.
- FOSTER, D. H. 2017. Chromatic Function of the Cones☆. *Reference Module in Neuroscience and Biobehavioral Psychology*. Elsevier.
- FRANCIS, P. J. 2006. Genetics of inherited retinal disease. *J R Soc Med*, 99, 189-91.
- FRANZE, K., GROSCHE, J., SKATCHKOV, S. N., SCHINKINGER, S., FOJA, C., SCHILD, D., UCKERMANN, O., TRAVIS, K., REICHENBACH, A. & GUCK, J. 2007. Muller cells are living optical fibers in the vertebrate retina. *Proc Natl Acad Sci U S A*, 104, 8287-92.
- FU, Y. & YAU, K. W. 2007. Phototransduction in mouse rods and cones. *Pflugers Arch*, 454, 805-19.
- FULLWOOD, N. J., MARTIN, F. L., BENTLEY, A. J., LEE, J. P. & LEE, S. J. 2011. Imaging sclera with hard X-ray microscopy. *Micron*, 42, 506-11.
- GABAI, A., VERITTI, D. & LANZETTA, P. 2015. Fundus autofluorescence applications in retinal imaging. *Indian J Ophthalmol*, 63, 406-15.
- GALLEGOS, J. E., ROGERS, M. F., CIALEK, C. A. & PECCOUD, J. 2020. Rapid, robust plasmid verification by de novo assembly of short sequencing reads. *Nucleic Acids Res*, 48, e106.
- GALVIN, O., CHI, G., BRADY, L., HIPPERT, C., DEL VALLE RUBIDO, M., DALY, A. & MICHAELIDES, M. 2020. The Impact of Inherited Retinal Diseases in the Republic of Ireland (ROI) and the United Kingdom (UK) from a Cost-of-Illness Perspective. *Clin Ophthalmol*, 14, 707-719.
- GEORGIU, M., FUJINAMI, K. & MICHAELIDES, M. 2021. Inherited retinal diseases: Therapeutics, clinical trials and end points-A review. *Clin Exp Ophthalmol*, 49, 270-288.
- GHEYAS, R., ORTEGA-ALVAREZ, R., CHAUSS, D., KANTOROW, M. & MENKO, A. S. 2022. Suppression of PI3K signaling is linked to autophagy activation and the spatiotemporal induction of the lens organelle free zone. *Exp Cell Res*, 412, 113043.
- GHOFRANI, M., YAHYAEI, M., BRUNNER, H. G., CREMERS, F. P., MOVASAT, M., IMRAN KHAN, M. & KERAMATIPOUR, M. 2017. Homozygosity Mapping and Targeted Sanger Sequencing Identifies Three Novel CRB1

- (Crumbs homologue 1) Mutations in Iranian Retinal Degeneration Families. *Iran Biomed J*, 21, 294-302.
- GIL-KRZEWSKA, A., WOOD, S. M., MURAKAMI, Y., NGUYEN, V., CHIANG, S. C. C., CULLINANE, A. R., PERUZZI, G., GAHL, W. A., COLIGAN, J. E., INTRONE, W. J., BRYCESON, Y. T. & KRZEWSKI, K. 2016. Chediak-Higashi syndrome: Lysosomal trafficking regulator domains regulate exocytosis of lytic granules but not cytokine secretion by natural killer cells. *J Allergy Clin Immunol*, 137, 1165-1177.
- GNIRKE, A., MELNIKOV, A., MAGUIRE, J., ROGOV, P., LEPROUST, E. M., BROCKMAN, W., FENNEL, T., GIANNOUKOS, G., FISHER, S., RUSS, C., GABRIEL, S., JAFFE, D. B., LANDER, E. S. & NUSBAUM, C. 2009. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat Biotechnol*, 27, 182-9.
- GRANTHAM, R. 1974. Amino acid difference formula to help explain protein evolution. *Science*, 185, 862-4.
- GRAW, J. 2010. Chapter Ten - Eye Development. *In: KOOPMAN, P. (ed.) Current Topics in Developmental Biology*. Academic Press.
- GREEN, R. C., BERG, J. S., GRODY, W. W., KALIA, S. S., KORF, B. R., MARTIN, C. L., MCGUIRE, A. L., NUSSBAUM, R. L., O'DANIEL, J. M., ORMOND, K. E., REHM, H. L., WATSON, M. S., WILLIAMS, M. S., BIESECKER, L. G., AMERICAN COLLEGE OF MEDICAL, G. & GENOMICS 2013. ACMG recommendations for reporting of incidental findings in clinical exome and genome sequencing. *Genet Med*, 15, 565-74.
- GRONSKOV, K., DOOLEY, C. M., OSTERGAARD, E., KELSH, R. N., HANSEN, L., LEVESQUE, M. P., VILHELMSSEN, K., MOLLGARD, K., STEMPEL, D. L. & ROSENBERG, T. 2013. Mutations in c10orf11, a melanocyte-differentiation gene, cause autosomal-recessive albinism. *Am J Hum Genet*, 92, 415-21.
- GRONSKOV, K., JESPERGAARD, C., BRUUN, G. H., HARRIS, P., BRONDUM-NIELSEN, K., ANDRESEN, B. S. & ROSENBERG, T. 2019. A pathogenic haplotype, common in Europeans, causes autosomal recessive albinism and uncovers missing heritability in OCA1. *Sci Rep*, 9, 645.
- GUARDIA, A., FERNANDEZ, A., SERUGGIA, D., CHOTARD, V., SANCHEZ-CASTILLO, C., KUTSYR, O., SANCHEZ-SAEZ, X., ZURITA, E., CANTERO, M., REBSAM, A., CUENCA, N. & MONTOLIU, L. 2023. A Slc38a8 Mouse Model of FHONDA Syndrome Faithfully Recapitulates the Visual Deficits of Albinism Without Pigmentation Defects. *Invest Ophthalmol Vis Sci*, 64, 32.
- GUILLERY, R. W., OKORO, A. N. & WITKOP, C. J., JR. 1975. Abnormal visual pathways in the brain of a human albino. *Brain Res*, 96, 373-7.
- GUPTA, P., NAKAMICHI, K., BONNELL, A. C., YANAGIHARA, R., RADULOVICH, N., HISAMA, F. M., CHAO, J. R. & MUSTAFI, D. 2023. Familial co-segregation and the emerging role of long-read sequencing to re-classify variants of uncertain significance in inherited retinal diseases. *NPJ Genom Med*, 8, 20.
- GURUNG, D., DANIELSON, J. A., TASNIM, A., ZHANG, J. T., ZOU, Y. & LIU, J. Y. 2023. Proline Isomerization: From the Chemistry and Biology to Therapeutic Opportunities. *Biology (Basel)*, 12.
- HAER-WIGMAN, L., DEN OUDEN, A., VAN GENDEREN, M. M., KROES, H. Y., VERHEIJ, J., SMAIHODZIC, D., HOEKSTRA, A. S., VIJZELAAR, R.,

- BLOM, J., DERKS, R., TJON-PON-FONG, M., YNTEMA, H. G., NELEN, M. R., VISSERS, L., LUGTENBERG, D. & NEVELING, K. 2022. Diagnostic analysis of the highly complex OPN1LW/OPN1MW gene cluster using long-read sequencing and MLPA. *NPJ Genom Med*, 7, 65.
- HAERI, M. & KNOX, B. E. 2012. Endoplasmic Reticulum Stress and Unfolded Protein Response Pathways: Potential for Treating Age-related Retinal Degeneration. *J Ophthalmic Vis Res*, 7, 45-59.
- HAGGLUND, M. G., SREEDHARAN, S., NILSSON, V. C., SHAIK, J. H., ALMKVIST, I. M., BACKLIN, S., WRANGE, O. & FREDRIKSSON, R. 2011. Identification of SLC38A7 (SNAT7) protein as a glutamine transporter expressed in neurons. *J Biol Chem*, 286, 20500-11.
- HAGGLUND, M. G. A., HELLSTEN, S. V., BAGCHI, S., PHILIPPOT, G., LOFQVIST, E., NILSSON, V. C. O., ALMKVIST, I., KARLSSON, E., SREEDHARAN, S., TAFRESHIHA, A. & FREDRIKSSON, R. 2015. Transport of L-glutamine, L-alanine, L-arginine and L-histidine by the neuron-specific Slc38a8 (SNAT8) in CNS. *J Mol Biol*, 427, 1495-1512.
- HALL, H. N., PARRY, D., HALACHEV, M., WILLIAMSON, K. A., DONNELLY, K., CAMPOS PARADA, J., BHATIA, S., JOSEPH, J., HOLDEN, S., PRESCOTT, T. E., BITOUN, P., KIRK, E. P., NEWBURY-ECOB, R., LACHLAN, K., BERNAR, J., VAN HEYNINGEN, V., FITZPATRICK, D. R. & MEYNERT, A. 2024. Short-read whole genome sequencing identifies causative variants in most individuals with previously unexplained aniridia. *J Med Genet*, 61, 250-261.
- HAMEL, C. 2006. Retinitis pigmentosa. *Orphanet J Rare Dis*, 1, 40.
- HARGRAVE, P. A., MCDOWELL, J. H., CURTIS, D. R., WANG, J. K., JUSZCZAK, E., FONG, S. L., RAO, J. K. & ARGOS, P. 1983. The structure of bovine rhodopsin. *Biophys Struct Mech*, 9, 235-44.
- HASHEMI, H., KHABAZKHOOB, M., EMAMIAN, M. H., SHARIATI, M., YEKTA, A. & FOTOUHI, A. 2015. The frequency of occurrence of certain corneal conditions by age and sex in Iranian adults. *Cont Lens Anterior Eye*, 38, 451-5.
- HAYASHI, T., KONDO, H., MATSUSHITA, I., MIZOBUCHI, K., BABA, A., IIDA, K., KUBO, H. & NAKANO, T. 2021. Homozygous single nucleotide duplication of SLC38A8 in autosomal recessive foveal hypoplasia: The first Japanese case report. *Doc Ophthalmol*, 143, 323-330.
- HEJTMANCIK, J. F. & SHIELS, A. 2015. Overview of the Lens. *Prog Mol Biol Transl Sci*, 134, 119-27.
- HENDRICKSON, A. E. & YUODELIS, C. 1984. The morphological development of the human fovea. *Ophthalmology*, 91, 603-12.
- HERMANSKY, F. & PUDLAK, P. 1959. Albinism associated with hemorrhagic diathesis and unusual pigmented reticular cells in the bone marrow: report of two cases with histochemical studies. *Blood*, 14, 162-9.
- HIATT, J. B., PRITCHARD, C. C., SALIPANTE, S. J., O'ROAK, B. J. & SHENDURE, J. 2013. Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res*, 23, 843-54.
- HINGORANI, M., WILLIAMSON, K. A., MOORE, A. T. & VAN HEYNINGEN, V. 2009. Detailed ophthalmologic evaluation of 43 individuals with PAX6 mutations. *Invest Ophthalmol Vis Sci*, 50, 2581-90.
- HITTI-MALIN, R. J., DHAENENS, C. M., PANNEMAN, D. M., CORRADI, Z., KHAN, M., DEN HOLLANDER, A. I., FARRAR, G. J., GILISSEN, C., HOISCHEN, A., VAN DE VORST, M., BULTS, F., BOONEN, E. G. M.,



- SAUNDERS, P., GROUP, M. D. S., ROOSING, S. & CREMERS, F. P. M. 2022. Using single molecule Molecular Inversion Probes as a cost-effective, high-throughput sequencing approach to target all genes and loci associated with macular diseases. *Hum Mutat*, 43, 2234-2250.
- HOCKING, L. J., ANDREWS, C., ARMSTRONG, C., ANSARI, M., BATY, D., BERG, J., BRADLEY, T., CLARK, C., DIAMOND, A., DOHERTY, J., LAMPE, A., MCGOWAN, R., MOORE, D. J., O'SULLIVAN, D., PURVIS, A., SANTOYO-LOPEZ, J., WESTWOOD, P., ABBOTT, M., WILLIAMS, N., SCOTTISH GENOMES, P., AITMAN, T. J. & MIEDZYBRODZKA, Z. 2023. Genome sequencing with gene panel-based analysis for rare inherited conditions in a publicly funded healthcare system: implications for future testing. *Eur J Hum Genet*, 31, 231-238.
- HOFFMANN, M. B. & DUMOULIN, S. O. 2015. Congenital visual pathway abnormalities: a window onto cortical stability and plasticity. *Trends Neurosci*, 38, 55-65.
- HONING, S., GRIFFITH, J., GEUZE, H. J. & HUNZIKER, W. 1996. The tyrosine-based lysosomal targeting signal in lamp-1 mediates sorting into Golgi-derived clathrin-coated vesicles. *EMBO J*, 15, 5230-9.
- HOON, M., OKAWA, H., DELLA SANTINA, L. & WONG, R. O. 2014. Functional architecture of the retina: development and disease. *Prog Retin Eye Res*, 42, 44-84.
- HU, T., CHITNIS, N., MONOS, D. & DINH, A. 2021. Next-generation sequencing technologies: An overview. *Hum Immunol*, 82, 801-811.
- HUANG, J., RAJAGOPAL, R., LIU, Y., DATTILO, L. K., SHAHAM, O., ASHERY-PADAN, R. & BEEBE, D. C. 2011. The mechanism of lens placode formation: a case of matrix-mediated morphogenesis. *Dev Biol*, 355, 32-42.
- HUANG, X. C., QUESADA, M. A. & MATHIES, R. A. 1992. DNA sequencing using capillary array electrophoresis. *Anal Chem*, 64, 2149-54.
- HUIZING, M., MALICDAN, M. C. V., WANG, J. A., PRI-CHEN, H., HESS, R. A., FISCHER, R., O'BRIEN, K. J., MERIDETH, M. A., GAHL, W. A. & GOCHUICO, B. R. 2020. Hermansky-Pudlak syndrome: Mutation update. *Hum Mutat*, 41, 543-580.
- ILIA, M. & JEFFERY, G. 1999. Retinal mitosis is regulated by dopa, a melanin precursor that may influence the time at which cells exit the cell cycle: Analysis of patterns of cell production in pigmented and albino retinæ. *Journal of Comparative Neurology*, 405, 394-405.
- INVESTIGATORS, G. P. P., SMEDLEY, D., SMITH, K. R., MARTIN, A., THOMAS, E. A., MCDONAGH, E. M., CIPRIANI, V., ELLINGFORD, J. M., ARNO, G., TUCCI, A., VANDROVCOVA, J., CHAN, G., WILLIAMS, H. J., RATNAIKE, T., WEI, W., STIRRUPS, K., IBANEZ, K., MOUTSIANAS, L., WIELSCHER, M., NEED, A., BARNES, M. R., VESTITO, L., BUCHANAN, J., WORDSWORTH, S., ASHFORD, S., REHMSTROM, K., LI, E., FULLER, G., TWISS, P., SPASIC-BOSKOVIC, O., HALSALL, S., FLOTO, R. A., POOLE, K., WAGNER, A., MEHTA, S. G., GURNELL, M., BURROWS, N., JAMES, R., PENKETT, C., DEWHURST, E., GRAF, S., MAPETA, R., KASANICKI, M., HAWORTH, A., SAVAGE, H., BABCOCK, M., REESE, M. G., BALE, M., BAPLE, E., BOUSTRED, C., BRITAIN, H., DE BURCA, A., BLEDA, M., DEVEREAU, A., HALAI, D., HARALDSDOTTIR, E., HYDER, Z., KASPERAVICIUTE, D., PATCH, C., POLYCHRONOPOULOS, D., MATCHAN, A., SULTANA, R., RYTEN, M., TAVARES, A. L. T., TREGIDGO, C., TURNBULL, C., WELLAND, M.,

- WOOD, S., SNOW, C., WILLIAMS, E., LEIGH, S., FOULGER, R. E., DAUGHERTY, L. C., NIBLOCK, O., LEONG, I. U. S., WRIGHT, C. F., DAVIES, J., CRICHTON, C., WELCH, J., WOODS, K., ABULHOUL, L., AURORA, P., BOCKENHAUER, D., BROOMFIELD, A., CLEARY, M. A., LAM, T., DATTANI, M., FOOTITT, E., GANESAN, V., GRUNEWALD, S., COMPEYROT-LACASSAGNE, S., MUNTONI, F., PILKINGTON, C., QUINLIVAN, R., THAPAR, N., WALLIS, C., WEDDERBURN, L. R., WORTH, A., BUESER, T., COMPTON, C., et al. 2021. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med*, 385, 1868-1880.
- ITTISOPONPISAN, S., ISLAM, S. A., KHANNA, T., ALHUZIMI, E., DAVID, A. & STERNBERG, M. J. E. 2019. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? *J Mol Biol*, 431, 2197-2212.
- JACKSON, D., MALKA, S., HARDING, P., PALMA, J., DUNBAR, H. & MOOSAJEE, M. 2020. Molecular diagnostic challenges for non-retinal developmental eye disorders in the United Kingdom. *Am J Med Genet C Semin Med Genet*, 184, 578-589.
- JACOBS, G. H. 2021. G-proteins | Color Vision. In: JEZ, J. (ed.) *Encyclopedia of Biological Chemistry III (Third Edition)*. Oxford: Elsevier.
- JAGANATHAN, K., KYRIAZOPOULOU PANAGIOTOPOULOU, S., MCRAE, J. F., DARBANDI, S. F., KNOWLES, D., LI, Y. I., KOSMICKI, J. A., ARBELAEZ, J., CUI, W., SCHWARTZ, G. B., CHOW, E. D., KANTERAKIS, E., GAO, H., KIA, A., BATZOGLOU, S., SANDERS, S. J. & FARH, K. K. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176, 535-548 e24.
- JAIN, M., KOREN, S., MIGA, K. H., QUICK, J., RAND, A. C., SASANI, T. A., TYSON, J. R., BEGGS, A. D., DILTHEY, A. T., FIDDES, I. T., MALLA, S., MARRIOTT, H., NIETO, T., O'GRADY, J., OLSEN, H. E., PEDERSEN, B. S., RHIE, A., RICHARDSON, H., QUINLAN, A. R., SNUTCH, T. P., TEE, L., PATEN, B., PHILLIPPY, A. M., SIMPSON, J. T., LOMAN, N. J. & LOOSE, M. 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol*, 36, 338-345.
- JANG, G. F., VAN HOOSER, J. P., KUKSA, V., MCBEE, J. K., HE, Y. G., JANSSEN, J. J., DRIESSEN, C. A. & PALCZEWSKI, K. 2001. Characterization of a dehydrogenase activity responsible for oxidation of 11-cis-retinol in the retinal pigment epithelium of mice with a disrupted RDH5 gene. A model for the human hereditary disease fundus albipunctatus. *J Biol Chem*, 276, 32456-65.
- JESPERSGAARD, C., FANG, M., BERTELSEN, M., DANG, X., JENSEN, H., CHEN, Y., BECH, N., DAI, L., ROSENBERG, T., ZHANG, J., MOLLER, L. B., TUMER, Z., BRONDUM-NIELSEN, K. & GRONSKOV, K. 2019. Molecular genetic analysis using targeted NGS analysis of 677 individuals with retinal dystrophy. *Sci Rep*, 9, 1219.
- JESTER, J. V., MOLLER-PEDERSEN, T., HUANG, J., SAX, C. M., KAYS, W. T., CAVANGH, H. D., PETROLL, W. M. & PIATIGORSKY, J. 1999. The cellular basis of corneal transparency: evidence for 'corneal crystallins'. *J Cell Sci*, 112 ( Pt 5), 613-22.
- JIN, M., LI, S., NUSINOWITZ, S., LLOYD, M., HU, J., RADU, R. A., BOK, D. & TRAVIS, G. H. 2009. The role of interphotoreceptor retinoid-binding protein on the translocation of visual retinoids and function of cone photoreceptors. *J Neurosci*, 29, 1486-95.

- JOENSUU, T., HAMALAINEN, R., YUAN, B., JOHNSON, C., TEGELBERG, S., GASPARINI, P., ZELANTE, L., PIRVOLA, U., PAKARINEN, L., LEHESJOKI, A. E., DE LA CHAPELLE, A. & SANKILA, E. M. 2001. Mutations in a novel gene with transmembrane domains underlie Usher syndrome type 3. *Am J Hum Genet*, 69, 673-84.
- JOHNSON, J., SOEHNLEN, M. & BLANKENSHIP, H. M. 2023. Long read genome assemblers struggle with small plasmids. *Microb Genom*, 9.
- JUNG, J. H., OH, E. H., SHIN, J. H., KIM, H. S., CHOI, S. Y., CHOI, K. D., LEE, C. & CHOI, J. H. 2018. Identification of a novel GPR143 mutation in X-linked ocular albinism with marked intrafamilial phenotypic variability. *J Genet*, 97, 1479-1484.
- KAJIWARA, K., BERSON, E. L. & DRYJA, T. P. 1994. Digenic retinitis pigmentosa due to mutations at the unlinked peripherin/RDS and ROM1 loci. *Science*, 264, 1604-8.
- KANEKO, A. & TACHIBANA, M. 1987. GABA mediates the negative feedback from amacrine to bipolar cells. *Neurosci Res Suppl*, 6, S239-51.
- KARCZEWSKI, K. J., FRANCIOLI, L. C., TIAO, G., CUMMINGS, B. B., ALFOLDI, J., WANG, Q., COLLINS, R. L., LARICCHIA, K. M., GANNA, A., BIRNBAUM, D. P., GAUTHIER, L. D., BRAND, H., SOLOMONSON, M., WATTS, N. A., RHODES, D., SINGER-BERK, M., ENGLAND, E. M., SEABY, E. G., KOSMICKI, J. A., WALTERS, R. K., TASHMAN, K., FARJOUN, Y., BANKS, E., POTERBA, T., WANG, A., SEED, C., WHIFFIN, N., CHONG, J. X., SAMOCHA, K. E., PIERCE-HOFFMAN, E., ZAPPALA, Z., O'DONNELL-LURIA, A. H., MINIKEL, E. V., WEISBURD, B., LEK, M., WARE, J. S., VITTAL, C., ARMEAN, I. M., BERGELSON, L., CIBULSKIS, K., CONNOLLY, K. M., COVARRUBIAS, M., DONNELLY, S., FERRIERA, S., GABRIEL, S., GENTRY, J., GUPTA, N., JEANDET, T., KAPLAN, D., LLANWARNE, C., MUNSHI, R., NOVOD, S., PETRILLO, N., ROAZEN, D., RUANO-RUBIO, V., SALTZMAN, A., SCHLEICHER, M., SOTO, J., TIBBETTS, K., TOLONEN, C., WADE, G., TALKOWSKI, M. E., GENOME AGGREGATION DATABASE, C., NEALE, B. M., DALY, M. J. & MACARTHUR, D. G. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581, 434-443.
- KAWAMURA, S., HIRAMATSU, C., MELIN, A. D., SCHAFFNER, C. M., AURELI, F. & FEDIGAN, L. M. 2012. Polymorphic Color Vision in Primates: Evolutionary Considerations. In: HIRAI, H., IMAI, H. & GO, Y. (eds.) *Post-Genome Biology of Primates*. Tokyo: Springer Tokyo.
- KELLEY, L. A., MEZULIS, S., YATES, C. M., WASS, M. N. & STERNBERG, M. J. 2015. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat Protoc*, 10, 845-58.
- KIECKER, C., BATES, T. & BELL, E. 2016. Molecular specification of germ layers in vertebrate embryos. *Cell Mol Life Sci*, 73, 923-47.
- KIM, J. H., KIM, J. H., YU, Y. S., MUN, J. Y. & KIM, K. W. 2010. Autophagy-induced regression of hyaloid vessels in early ocular development. *Autophagy*, 6, 922-8.
- KIMBERLING, W. J., HILDEBRAND, M. S., SHEARER, A. E., JENSEN, M. L., HALDER, J. A., TRZUPEK, K., COHN, E. S., WELEBER, R. G., STONE, E. M. & SMITH, R. J. 2010. Frequency of Usher syndrome in two pediatric populations: Implications for genetic screening of deaf and hard of hearing children. *Genet Med*, 12, 512-6.

- KIRCHNER, I. D., WALDMAN, C. W. & SUNNESS, J. S. 2019. A Series of Five Patients with Foveal Hypoplasia Demonstrating Good Visual Acuity. *Retin Cases Brief Rep*, 13, 376-380.
- KNAUS, K. R., HIPSLEY, A. & BLEMKER, S. S. 2021. The action of ciliary muscle contraction on accommodation of the lens explored with a 3D model. *Biomech Model Mechanobiol*, 20, 879-894.
- KOBAT, S. G. & TURGUT, B. 2020. Importance of Muller Cells. *Beyoglu Eye J*, 5, 59-63.
- KOCAOGLU, O. P., LIU, Z., ZHANG, F., KUROKAWA, K., JONNAL, R. S. & MILLER, D. T. 2016. Photoreceptor disc shedding in the living human eye. *Biomed Opt Express*, 7, 4554-4568.
- KONDO, H. 2018. Foveal hypoplasia and optical coherence tomographic imaging. *Taiwan J Ophthalmol*, 8, 181-188.
- KORLACH, J., BJORNSON, K. P., CHAUDHURI, B. P., CICERO, R. L., FLUSBERG, B. A., GRAY, J. J., HOLDEN, D., SAXENA, R., WEGENER, J. & TURNER, S. W. 2010. Real-time DNA sequencing from single polymerase molecules. *Methods Enzymol*, 472, 431-55.
- KOZULIN, P., NATOLI, R., BUMSTED O'BRIEN, K. M., MADIGAN, M. C. & PROVIS, J. M. 2010. The cellular expression of antiangiogenic factors in fetal primate macula. *Invest Ophthalmol Vis Sci*, 51, 4298-306.
- KROGER, R. H. & BIEHLMAIER, O. 2009. Space-saving advantage of an inverted retina. *Vision Res*, 49, 2318-21.
- KRUIJT, B., FRANSSSEN, L., PRICK, L. J., VAN VLIET, J. M. & VAN DEN BERG, T. J. 2011. Ocular straylight in albinism. *Optom Vis Sci*, 88, E585-92.
- KRUIJT, C. C., DE WIT, G. C., BERGEN, A. A., FLORIJN, R. J., SCHALIJ-DELFOOS, N. E. & VAN GENDEREN, M. M. 2018. The Phenotypic Spectrum of Albinism. *Ophthalmology*, 125, 1953-1960.
- KRUIJT, C. C., GRADSTEIN, L., BERGEN, A. A., FLORIJN, R. J., ARVEILER, B., LASSEAUX, E., ZANLONGHI, X., BAGDONAITE-BEJARANO, L., FULTON, A. B., YAHALOM, C., BLUMENFELD, A., PEREZ, Y., BIRK, O. S., DE WIT, G. C., SCHALIJ-DELFOOS, N. E. & VAN GENDEREN, M. M. 2022. The Phenotypic and Mutational Spectrum of the FHONDA Syndrome and Oculocutaneous Albinism: Similarities and Differences. *Invest Ophthalmol Vis Sci*, 63, 19.
- KUCUK, E., VAN DER SANDEN, B., O'GORMAN, L., KWINT, M., DERKS, R., WENGER, A. M., LAMBERT, C., CHAKRABORTY, S., BAYBAYAN, P., ROWELL, W. J., BRUNNER, H. G., VISSERS, L., HOISCHEN, A. & GILISSEN, C. 2023. Comprehensive de novo mutation discovery with HiFi long-read sequencing. *Genome Med*, 15, 34.
- KUHT, H. J., HAN, J., MACONACHIE, G. D. E., PARK, S. E., LEE, S. T., MCLEAN, R., SHETH, V., HISAUND, M., DAWAR, B., SYLVIUS, N., MAHMOOD, U., PROUDLOCK, F. A., GOTTLÖB, I., LIM, H. T. & THOMAS, M. G. 2020. SLC38A8 mutations result in arrested retinal development with loss of cone photoreceptor specialization. *Hum Mol Genet*, 29, 2989-3002.
- KUHT, H. J., THOMAS, M. G., MCLEAN, R. J., SHETH, V., PROUDLOCK, F. A. & GOTTLÖB, I. 2023. Abnormal foveal morphology in carriers of oculocutaneous albinism. *Br J Ophthalmol*, 107, 1202-1208.
- KUMAR, A., GOTTLÖB, I., MCLEAN, R. J., THOMAS, S., THOMAS, M. G. & PROUDLOCK, F. A. 2011. Clinical and oculomotor characteristics of albinism compared to FRMD7 associated infantile nystagmus. *Invest Ophthalmol Vis Sci*, 52, 2306-13.

- KUMAR, V., COUSER, N. L. & PANDYA, A. 2020. Oculodentodigital Dysplasia: A Case Report and Major Review of the Eye and Ocular Adnexa Features of 295 Reported Cases. *Case Rep Ophthalmol Med*, 2020, 6535974.
- KUMARAN, N., MOORE, A. T., WELEBER, R. G. & MICHAELIDES, M. 2017. Leber congenital amaurosis/early-onset severe retinal dystrophy: clinical features, molecular genetics and therapeutic interventions. *Br J Ophthalmol*, 101, 1147-1154.
- KUPFER, C., CHUMBLEY, L. & DOWNER, J. C. 1967. Quantitative histology of optic nerve, optic tract and lateral geniculate nucleus of man. *J Anat*, 101, 393-401.
- LAM, B. L., LEROY, B. P., BLACK, G., ONG, T., YOON, D. & TRZUPEK, K. 2021. Genetic testing and diagnosis of inherited retinal diseases. *Orphanet J Rare Dis*, 16, 514.
- LAMB, T. D. 2016. Why rods and cones? *Eye (Lond)*, 30, 179-85.
- LAMB, T. D. & PUGH, E. N., JR. 2006. Phototransduction, dark adaptation, and rhodopsin regeneration the proctor lecture. *Invest Ophthalmol Vis Sci*, 47, 5137-52.
- LASSEAUX, E., PLAISANT, C., MICHAUD, V., PENNAMEN, P., TRIMOUILLE, A., GASTON, L., MONFERME, S., LACOMBE, D., ROORYCK, C., MORICE-PICARD, F. & ARVEILER, B. 2018. Molecular characterization of a series of 990 index patients with albinism. *Pigment Cell Melanoma Res*, 31, 466-474.
- LI, W., HAO, C. J., HAO, Z. H., MA, J., WANG, Q. C., YUAN, Y. F., GONG, J. J., CHEN, Y. Y., YU, J. Y. & WEI, A. H. 2022. New insights into the pathogenesis of Hermansky-Pudlak syndrome. *Pigment Cell Melanoma Res*, 35, 290-302.
- LIMA CUNHA, D., ARNO, G., CORTON, M. & MOOSAJEE, M. 2019. The Spectrum of PAX6 Mutations and Genotype-Phenotype Correlations in the Eye. *Genes (Basel)*, 10.
- LIN, Y., WANG, J., XU, R., XU, Z., WANG, Y., PAN, S., ZHANG, Y., TAO, Q., ZHAO, Y., YAN, C., CAO, Z. & JI, K. 2024. HiFi long-read amplicon sequencing for full-spectrum variants of human mtDNA. *BMC Genomics*, 25, 538.
- LIU, X., HU, F., ZHANG, D., LI, Z., HE, J., ZHANG, S., WANG, Z., ZHAO, Y., WU, J., LIU, C., LI, C., LI, X. & WU, J. 2024. Whole genome sequencing enables new genetic diagnosis for inherited retinal diseases by identifying pathogenic variants. *NPJ Genom Med*, 9, 6.
- LOFTUS, S. K., GILLIS, M. F., LUNDH, L., BAXTER, L. L., WEDEL, J. C., WATKINS-CHOW, D. E., DONOVAN, F. X., PROGRAM, N. C. S., SERGEEV, Y. V., OETTING, W. S., PAVAN, W. J. & ADAMS, D. R. 2023. Haplotype-based analysis resolves missing heritability in oculocutaneous albinism type 1B. *Am J Hum Genet*, 110, 1123-1137.
- LOGSDON, G. A., VOLLGER, M. R. & EICHLER, E. E. 2020. Long-read human genome sequencing and its applications. *Nat Rev Genet*, 21, 597-614.
- LOK, S., PATON, T. A., WANG, Z., KAUR, G., WALKER, S., YUEN, R. K., SUNG, W. W., WHITNEY, J., BUCHANAN, J. A., TROST, B., SINGH, N., APRESTO, B., CHEN, N., COOLE, M., DAWSON, T. J., HO, K., HU, Z., PULLENAYEGUM, S., SAMLER, K., SHIPSTONE, A., TSOI, F., WANG, T., PEREIRA, S. L., ROSTAMI, P., RYAN, C. A., TONG, A. H., NG, K., SUNDARAVADANAM, Y., SIMPSON, J. T., LIM, B. K., ENGSTROM, M. D., DUTTON, C. J., KERR, K. C., FRANKE, M., RAPLEY, W., WINTLE, R. F. & SCHERER, S. W. 2017. De Novo Genome and Transcriptome

- Assembly of the Canadian Beaver (*Castor canadensis*). *G3 (Bethesda)*, 7, 755-773.
- LOPEZ, V. M., DECATUR, C. L., STAMER, W. D., LYNCH, R. M. & MCKAY, B. S. 2008. L-DOPA is an endogenous ligand for OA1. *PLoS Biol*, 6, e236.
- LÓPEZ-OREJA, I., LÓPEZ-GUERRA, M., CORREA, J., MOZAS, P., MUNTAÑOLA, A., MUÑOZ, L., SALGADO, A. C., RUIZ-GASPÀ, S., COSTA, D., BEÀ, S., JARES, P., CAMPO, E., COLOMER, D. & NADEU, F. 2023. All-CLL: A Capture-based Next-generation Sequencing Panel for the Molecular Characterization of Chronic Lymphocytic Leukemia. *Hemasphere*, 7, e962.
- LORD, E. C. 2018. *The molecular genetics of foveal hypoplasia*. . PhD Thesis, University of Leeds.
- LORD, E. C., POULTER, J. A., WEBSTER, A. R., SERGOUNIOTIS, P., KHAN, K. N., BENKE, P. J., FRIEDMAN, L., ALI, M., INGLEHEARN, C. F. & TOOMES, C. 2017. Mutations in SLC38A8 and FOXD1 in patients with nystagmus and foveal hypoplasia. *Investigative Ophthalmology & Visual Science*, 58, 2786-2786.
- LU, H., GIORDANO, F. & NING, Z. 2016. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics Proteomics Bioinformatics*, 14, 265-279.
- LU, J. & DEUTSCH, C. 2008. Electrostatics in the ribosomal tunnel modulate chain elongation rates. *J Mol Biol*, 384, 73-86.
- LUJAN, B. J., ROORDA, A., KNIGHTON, R. W. & CARROLL, J. 2011. Revealing Henle's fiber layer using spectral domain optical coherence tomography. *Invest Ophthalmol Vis Sci*, 52, 1486-92.
- LUND, P. M. & ROBERTS, M. 2018. Chapter 4 - Prevalence and Population Genetics of Albinism: Surveys in Zimbabwe, Namibia, and Tanzania. *In: KROMBERG, J. & MANGA, P. (eds.) Albinism in Africa*. Academic Press.
- MAALOUL, I., TALMOUDI, J., CHABCHOUB, I., AYADI, L., KAMOUN, T. H., BOUDAWARA, T., KALLEL, C. H. & HACHICHA, M. 2016. Chediak-Higashi syndrome presenting in accelerated phase: A case report and literature review. *Hematology/Oncology and Stem Cell Therapy*, 9, 71-75.
- MACKENZIE, B., SCHAFER, M. K., ERICKSON, J. D., HEDIGER, M. A., WEIHE, E. & VAROQUI, H. 2003. Functional properties and cellular distribution of the system A glutamine transporter SNAT1 support specialized roles in central neurons. *J Biol Chem*, 278, 23720-30.
- MADANI TONEKABONI, S. A., MAZROOEI, P., KOFIA, V., HAIBE-KAINS, B. & LUPIEN, M. 2019. Identifying clusters of cis-regulatory elements underpinning TAD structures and lineage-specific regulatory networks. *Genome Res*, 29, 1733-1743.
- MAEDA, A., MAEDA, T., IMANISHI, Y., KUKSA, V., ALEKSEEV, A., BRONSON, J. D., ZHANG, H., ZHU, L., SUN, W., SAPERSTEIN, D. A., RIEKE, F., BAEHR, W. & PALCZEWSKI, K. 2005. Role of photoreceptor-specific retinol dehydrogenase in the retinoid cycle in vivo. *J Biol Chem*, 280, 18822-32.
- MANTERE, T., NEVELING, K., PEBREL-RICHARD, C., BENOIST, M., VAN DER ZANDE, G., KATER-BAATS, E., BAATOUT, I., VAN BEEK, R., YAMMINE, T., OORSPRONG, M., HSOUMI, F., OLDE-WEGHUIS, D., MAJDALI, W., VERMEULEN, S., PAUPER, M., LEBBAR, A., STEVENS-KROEF, M., SANLAVILLE, D., DUPONT, J. M., SMEETS, D., HOISCHEN, A., SCHLUTH-BOLARD, C. & EL KHATTABI, L. 2021. Optical genome

- mapping enables constitutional chromosomal aberration detection. *Am J Hum Genet*, 108, 1409-1422.
- MAO, X., CHEN, M., YU, Y., LIU, Q., YUAN, S. & FAN, W. 2021. Identification of a novel GPR143 mutation in a large Chinese family with isolated foveal hypoplasia. *BMC Ophthalmol*, 21, 156.
- MARIA, M., AJMAL, M., AZAM, M., WAHEED, N. K., SIDDIQUI, S. N., MUSTAFA, B., AYUB, H., ALI, L., AHMAD, S., MICHEAL, S., HUSSAIN, A., SHAH, S. T., ALI, S. H., AHMED, W., KHAN, Y. M., DEN HOLLANDER, A. I., HAER-WIGMAN, L., COLLIN, R. W., KHAN, M. I., QAMAR, R. & CREMERS, F. P. 2015. Homozygosity mapping and targeted sanger sequencing reveal genetic defects underlying inherited retinal disease in families from pakistan. *PLoS One*, 10, e0119806.
- MARTIN, A. R., WILLIAMS, E., FOULGER, R. E., LEIGH, S., DAUGHERTY, L. C., NIBLOCK, O., LEONG, I. U. S., SMITH, K. R., GERASIMENKO, O., HARALDSDOTTIR, E., THOMAS, E., SCOTT, R. H., BAPLE, E., TUCCI, A., BRITAIN, H., DE BURCA, A., IBANEZ, K., KASPERAVICIUTE, D., SMEDLEY, D., CAULFIELD, M., RENDON, A. & MCDONAGH, E. M. 2019. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*, 51, 1560-1565.
- MARTIN, H., GUTHOFF, R., TERWEE, T. & SCHMITZ, K. P. 2005. Comparison of the accommodation theories of Coleman and of Helmholtz by finite element simulations. *Vision Res*, 45, 2910-5.
- MARTIN, S., HEAVENS, D., LAN, Y., HORSFIELD, S., CLARK, M. D. & LEGGETT, R. M. 2022. Nanopore adaptive sampling: a tool for enrichment of low abundance species in metagenomic samples. *Genome Biol*, 23, 11.
- MARUMO, C. & NAKANO, T. 2021. Early phase of pupil dilation is mediated by the peripheral parasympathetic pathway. *J Neurophysiol*, 126, 2130-2137.
- MASLAND, R. H. 2012. The neuronal organization of the retina. *Neuron*, 76, 266-80.
- MAYER, A. K., MAHAJNAH, M., THOMAS, M. G., COHEN, Y., HABIB, A., SCHULZE, M., MACONACHIE, G. D. E., ALMOALLEM, B., DE BAERE, E., LORENZ, B., TRABOULSI, E. I., KOHL, S., AZEM, A., BAUER, P., GOTTLÖB, I., SHARKIA, R. & WISSINGER, B. 2019. Homozygous stop mutation in AHR causes autosomal recessive foveal hypoplasia and infantile nystagmus. *Brain*, 142, 1528-1534.
- MC CLINTON, B., CORRADI, Z., MCKIBBIN, M., PANNEMAN, D. M., ROOSING, S., BOONEN, E. G. M., ALI, M., WATSON, C. M., STEEL, D. H., CREMERS, F. P. M., INGLEHEARN, C. F., HITTI-MALIN, R. J. & TOOMES, C. 2023. Effective smMIPs-Based Sequencing of Maculopathy-Associated Genes in Stargardt Disease Cases and Allied Maculopathies from the UK. *Genes (Basel)*, 14.
- MCCLINTON, B., CRINNION, L. A., MCKIBBIN, M., MUKHERJEE, R., POULTER, J. A., SMITH, C. E. L., ALI, M., WATSON, C. M., INGLEHEARN, C. F. & TOOMES, C. 2023a. Targeted nanopore sequencing enables complete characterisation of structural deletions initially identified using exon-based short-read sequencing strategies. *Mol Genet Genomic Med*, 11, e2164.
- MCCLINTON, B., WATSON, C. M., CRINNION, L. A., MCKIBBIN, M., ALI, M., INGLEHEARN, C. F. & TOOMES, C. 2023b. Haplotyping Using Long-Range PCR and Nanopore Sequencing to Phase Variants: Lessons Learned From the ABCA4 Locus. *Lab Invest*, 103, 100160.

- MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol*, 17, 122.
- MEIENBERG, J., BRUGGMANN, R., OEXLE, K. & MATYAS, G. 2016. Clinical sequencing: is WGS the better WES? *Hum Genet*, 135, 359-62.
- MEIENBERG, J., ZERJAVIC, K., KELLER, I., OKONIEWSKI, M., PATRIGNANI, A., LUDIN, K., XU, Z., STEINMANN, B., CARREL, T., RÖTHLISBERGER, B., SCHLAPBACH, R., BRUGGMANN, R. & MATYAS, G. 2015. New insights into the performance of human whole-exome capture platforms. *Nucleic Acids Res*, 43, e76.
- MERBS, S. L. & NATHANS, J. 1992. Absorption spectra of human cone pigments. *Nature*, 356, 433-5.
- MICHAUD, V., LASSEAUX, E., PLAISANT, C., VERLOES, A., PERDOMO-TRUJILLO, Y., HAMEL, C., ELCIOGLU, N. H., LEROY, B., KAPLAN, J., JOUK, P. S., LACOMBE, D., FERGELOT, P., MORICE-PICARD, F. & ARVEILER, B. 2017. Clinico-molecular analysis of eleven patients with Hermansky-Pudlak type 5 syndrome, a mild form of HPS. *Pigment Cell Melanoma Res*, 30, 563-570.
- MIESFELD, J. B. & BROWN, N. L. 2019. Chapter Ten - Eye organogenesis: A hierarchical view of ocular development. In: WELLIK, D. M. (ed.) *Current Topics in Developmental Biology*. Academic Press.
- MINTZ-HITTNER, H. A., KNIGHT-NANAN, D. M., SATRIANO, D. R. & KRETZER, F. L. 1999. A small foveal avascular zone may be an historic mark of prematurity. *Ophthalmology*, 106, 1409-13.
- MISHRA, R., GORLOV, I. P., CHAO, L. Y., SINGH, S. & SAUNDERS, G. F. 2002. PAX6, paired domain influences sequence recognition by the homeodomain. *J Biol Chem*, 277, 49488-94.
- MOISEYEV, G., CHEN, Y., TAKAHASHI, Y., WU, B. X. & MA, J. X. 2005. RPE65 is the isomerohydrolase in the retinoid visual cycle. *Proc Natl Acad Sci U S A*, 102, 12413-8.
- MONTOLIU, L., GRONSKOV, K., WEI, A. H., MARTINEZ-GARCIA, M., FERNANDEZ, A., ARVEILER, B., MORICE-PICARD, F., RIAZUDDIN, S., SUZUKI, T., AHMED, Z. M., ROSENBERG, T. & LI, W. 2014. Increasing the complexity: new genes and new types of albinism. *Pigment Cell Melanoma Res*, 27, 11-8.
- MONTOLIU, L. & KELSH, R. N. 2014. Do you have to be albino to be albino? *Pigment Cell & Melanoma Research*, 27, 325-326.
- MURTHY, K. R., RAJAGOPALAN, P., PINTO, S. M., ADVANI, J., MURTHY, P. R., GOEL, R., SUBBANNAYYA, Y., BALAKRISHNAN, L., DASH, M., ANIL, A. K., MANDA, S. S., NIRUJOGI, R. S., KELKAR, D. S., SATHE, G. J., DEY, G., CHATTERJEE, A., GOWDA, H., CHAKRAVARTI, S., SHANKAR, S., SAHASRABUDDHE, N. A., NAIR, B., SOMANI, B. L., PRASAD, T. S. & PANDEY, A. 2015. Proteomics of human aqueous humor. *OMICS*, 19, 283-93.
- MYLLYKANGAS, S., BUENROSTRO, J. & JI, H. P. 2012. Overview of Sequencing Technology Platforms. In: RODRÍGUEZ-EZPELETA, N., HACKENBERG, M. & ARANSAY, A. M. (eds.) *Bioinformatics for High Throughput Sequencing*. New York, NY: Springer New York.
- NAKAMICHI, K., VAN GELDER, R. N., CHAO, J. R. & MUSTAFI, D. 2023. Targeted adaptive long-read sequencing for discovery of complex phased variants in inherited retinal disease patients. *Sci Rep*, 13, 8535.



- NAKANO, T., ICHIKI, A. & FUJIKADO, T. 2021. Pupil constriction via the parasympathetic pathway precedes perceptual switch of ambiguous stimuli. *Int J Psychophysiol*, 167, 15-21.
- NAZARIAN, R., FALCÓN-PÉREZ, J. M. & DELL'ANGELICA, E. C. 2003. Biogenesis of lysosome-related organelles complex 3 (BLOC-3): a complex containing the Hermansky-Pudlak syndrome (HPS) proteins HPS1 and HPS4. *Proc Natl Acad Sci U S A*, 100, 8770-5.
- NEEDLEMAN, S. B. & WUNSCH, C. D. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, 48, 443-53.
- NESS, S. L., BEN-YOSEF, T., BAR-LEV, A., MADEO, A. C., BREWER, C. C., AVRAHAM, K. B., KORNREICH, R., DESNICK, R. J., WILLNER, J. P., FRIEDMAN, T. B. & GRIFFITH, A. J. 2003. Genetic homogeneity and phenotypic variability among Ashkenazi Jews with Usher syndrome type III. *J Med Genet*, 40, 767-72.
- NEVELING, K., MANTERE, T., VERMEULEN, S., OORSPRONG, M., VAN BEEK, R., KATER-BAATS, E., PAUPER, M., VAN DER ZANDE, G., SMEETS, D., WEGHUIS, D. O., STEVENS-KROEF, M. & HOISCHEN, A. 2021. Next-generation cytogenetics: Comprehensive assessment of 52 hematological malignancy genomes by optical genome mapping. *Am J Hum Genet*, 108, 1423-1435.
- NEVEU, M. M., JEFFERY, G., BURTON, L. C., SLOPER, J. J. & HOLDER, G. E. 2003. Age-related changes in the dynamics of human albino visual pathways. *Eur J Neurosci*, 18, 1939-49.
- NEWTON, J. M., COHEN-BARAK, O., HAGIWARA, N., GARDNER, J. M., DAVISSON, M. T., KING, R. A. & BRILLIANT, M. H. 2001. Mutations in the human orthologue of the mouse underwhite gene (*uw*) underlie a new form of oculocutaneous albinism, *OCA4*. *Am J Hum Genet*, 69, 981-8.
- NI, Y., LIU, X., SIMENEH, Z. M., YANG, M. & LI, R. 2023. Benchmarking of Nanopore R10.4 and R9.4.1 flow cells in single-cell whole-genome amplification and whole-genome shotgun sequencing. *Comput Struct Biotechnol J*, 21, 2352-2364.
- NICKLA, D. L. & WALLMAN, J. 2010. The multifunctional choroid. *Prog Retin Eye Res*, 29, 144-68.
- NURK, S., KOREN, S., RHIE, A., RAUTIAINEN, M., BZIKADZE, A. V., MIKHEENKO, A., VOLLGER, M. R., ALTEMOSE, N., URALSKY, L., GERSHMAN, A., AGANEZOV, S., HOYT, S. J., DIEKHANS, M., LOGSDON, G. A., ALONGE, M., ANTONARAKIS, S. E., BORCHERS, M., BOUFFARD, G. G., BROOKS, S. Y., CALDAS, G. V., CHEN, N. C., CHENG, H., CHIN, C. S., CHOW, W., DE LIMA, L. G., DISHUCK, P. C., DURBIN, R., DVORKINA, T., FIDDES, I. T., FORMENTI, G., FULTON, R. S., FUNGTAMMASAN, A., GARRISON, E., GRADY, P. G. S., GRAVES-LINDSAY, T. A., HALL, I. M., HANSEN, N. F., HARTLEY, G. A., HAUKNES, M., HOWE, K., HUNKAPILLER, M. W., JAIN, C., JAIN, M., JARVIS, E. D., KERPEDJIEV, P., KIRSCHE, M., KOLMOGOROV, M., KORLACH, J., KREMITZKI, M., LI, H., MADURO, V. V., MARSCHALL, T., MCCARTNEY, A. M., MCDANIEL, J., MILLER, D. E., MULLIKIN, J. C., MYERS, E. W., OLSON, N. D., PATEN, B., PELUSO, P., PEVZNER, P. A., PORUBSKY, D., POTAPOVA, T., ROGAEV, E. I., ROSENFELD, J. A., SALZBERG, S. L., SCHNEIDER, V. A., SEDLAZECK, F. J., SHAFIN, K., SHEW, C. J., SHUMATE, A., SIMS, Y., SMIT, A. F. A., SOTO, D. C., SOVIC, I., STORER, J. M., STREETS, A., SULLIVAN, B. A., THIBAUD-

- NISSEN, F., TORRANCE, J., WAGNER, J., WALENZ, B. P., WENGER, A., WOOD, J. M. D., XIAO, C., YAN, S. M., YOUNG, A. C., ZARATE, S., SURTI, U., MCCOY, R. C., DENNIS, M. Y., ALEXANDROV, I. A., GERTON, J. L., O'NEILL, R. J., TIMP, W., ZOOK, J. M., SCHATZ, M. C., EICHLER, E. E., MIGA, K. H. & PHILLIPPY, A. M. 2022. The complete sequence of a human genome. *Science*, 376, 44-53.
- O'SHEA, S. M., O'DWYER, V. M. & SCANLON, G. 2022. Normative data on the foveal avascular zone in a young healthy Irish population using optical coherence tomography angiography. *Eur J Ophthalmol*, 32, 2824-2832.
- OBERMULLER, S., KIECKE, C., VON FIGURA, K. & HONING, S. 2002. The tyrosine motifs of Lamp 1 and LAP determine their direct and indirect targeting to lysosomes. *J Cell Sci*, 115, 185-94.
- OKANO, T., FUKADA, Y., SHICHIDA, Y. & YOSHIZAWA, T. 1992. Photosensitivities of iodopsin and rhodopsins. *Photochem Photobiol*, 56, 995-1001.
- OKORO, A. N. 1975. Albinism in Nigeria. A clinical and social study. *Br J Dermatol*, 92, 485-92.
- OMASITS, U., AHRENS, C. H., MULLER, S. & WOLLSCHIED, B. 2014. Protter: interactive protein feature visualization and integration with experimental proteomic data. *Bioinformatics*, 30, 884-6.
- ORLOW, S. J. 1995. Melanosomes are specialized members of the lysosomal lineage of organelles. *J Invest Dermatol*, 105, 3-7.
- PAKARINEN, L., KARJALAINEN, S., SIMOLA, K. O., LAIPPALA, P. & KAITALO, H. 1995. Usher's syndrome type 3 in Finland. *Laryngoscope*, 105, 613-7.
- PANG, Y., ZHANG, G., ZHANG, H., SHE, J., ZHANG, X., LI, H. & ZHANG, G. 2023. Foveal avascular zone in normal human eyes by optical coherence tomography angiography. *Photodiagnosis Photodyn Ther*, 42, 103303.
- PANNEMAN, D. M., HITTI-MALIN, R. J., HOLTES, L. K., DE BRUIJN, S. E., REURINK, J., BOONEN, E. G. M., KHAN, M. I., ALI, M., ANDREASSON, S., DE BAERE, E., BANFI, S., BAUWENS, M., BEN-YOSEF, T., BOCQUET, B., DE BRUYNE, M., DE LA CERDA, B., COPPIETERS, F., FARINELLI, P., GUIGNARD, T., INGLEHEARN, C. F., KARALI, M., KJELLSTROM, U., KOENEKOOP, R., DE KONING, B., LEROY, B. P., MCKIBBIN, M., MEUNIER, I., NIKOPOULOS, K., NISHIGUCHI, K. M., POULTER, J. A., RIVOLTA, C., RODRIGUEZ DE LA RUA, E., SAUNDERS, P., SIMONELLI, F., TATOUR, Y., TESTA, F., THIADENS, A., TOOMES, C., TRACEWSKA, A. M., TRAN, H. V., USHIDA, H., VACLAVIK, V., VERHOEVEN, V. J. M., VAN DE VORST, M., GILISSEN, C., HOISCHEN, A., CREMERS, F. P. M. & ROOSING, S. 2023. Cost-effective sequence analysis of 113 genes in 1,192 probands with retinitis pigmentosa and Leber congenital amaurosis. *Front Cell Dev Biol*, 11, 1112270.
- PARTRIDGE, A. W., THERIEN, A. G. & DEBER, C. M. 2004. Missense mutations in transmembrane domains of proteins: phenotypic propensity of polar residues for human disease. *Proteins*, 54, 648-56.
- PATEL, S. & TUTCHENKO, L. 2019. The refractive index of the human cornea: A review. *Cont Lens Anterior Eye*, 42, 575-580.
- PENNAMEN, P., TINGAUD-SEQUEIRA, A., GAZOVA, I., KEIGHREN, M., MCKIE, L., MARLIN, S., GHERBI HALEM, S., KAPLAN, J., DELEVOYE, C., LACOMBE, D., PLAISANT, C., MICHAUD, V., LASSEAU, E., JAVERZAT, S., JACKSON, I. & ARVEILER, B. 2021. Dopachrome

- tautomerase variants in patients with oculocutaneous albinism. *Genet Med*, 23, 479-487.
- PEPLOW, M. 2016. The 100,000 Genomes Project. *BMJ*, 353, i1757.
- PEREZ, Y., GRADSTEIN, L., FLUSSER, H., MARKUS, B., COHEN, I., LANGER, Y., MARCUS, M., LIFSHITZ, T., KADIR, R. & BIRK, O. S. 2014. Isolated foveal hypoplasia with secondary nystagmus and low vision is associated with a homozygous SLC38A8 mutation. *Eur J Hum Genet*, 22, 703-6.
- PERRAULT, I., DELPHIN, N., HANEIN, S., GERBER, S., DUFIER, J. L., ROCHE, O., DEFOORT-DHELLEMMES, S., DOLLFUS, H., FAZZI, E., MUNNICH, A., KAPLAN, J. & ROZET, J. M. 2007. Spectrum of NPHP6/CEP290 mutations in Leber congenital amaurosis and delineation of the associated phenotype. *Hum Mutat*, 28, 416.
- PETTERSEN, E. F., GODDARD, T. D., HUANG, C. C., COUCH, G. S., GREENBLATT, D. M., MENG, E. C. & FERRIN, T. E. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem*, 25, 1605-12.
- PONTEN, F., JIRSTROM, K. & UHLEN, M. 2008. The Human Protein Atlas--a tool for pathology. *J Pathol*, 216, 387-93.
- PONTIKOS, N., ARNO, G., JURKUTE, N., SCHIFF, E., BA-ABBAD, R., MALKA, S., GIMENEZ, A., GEORGIU, M., WRIGHT, G., ARMENGOL, M., KNIGHT, H., KATZ, M., MOOSAJEE, M., YU-WAI-MAN, P., MOORE, A. T., MICHAELIDES, M., WEBSTER, A. R. & MAHROO, O. A. 2020. Genetic Basis of Inherited Retinal Disease in a Molecularly Characterized Cohort of More Than 3000 Families from the United Kingdom. *Ophthalmology*, 127, 1384-1394.
- POULTER, J. A., AL-ARAIMI, M., CONTE, I., VAN GENDEREN, M. M., SHERIDAN, E., CARR, I. M., PARRY, D. A., SHIRES, M., CARRELLA, S., BRADBURY, J., KHAN, K., LAKEMAN, P., SERGOUNIOTIS, P. I., WEBSTER, A. R., MOORE, A. T., PAL, B., MOHAMED, M. D., VENKATARAMANA, A., RAMPRASAD, V., SHETTY, R., SAKTIVEL, M., KUMARAMANICKAVEL, G., TAN, A., MACKAY, D. A., HEWITT, A. W., BANFI, S., ALI, M., INGLEHEARN, C. F. & TOOMES, C. 2013. Recessive mutations in SLC38A8 cause foveal hypoplasia and optic nerve misrouting without albinism. *Am J Hum Genet*, 93, 1143-50.
- PRAJNA, V. & VIJAYALAKSHMI, P. 2017. Chapter 31 - Conjunctiva and subconjunctival tissue. In: LAMBERT, S. R. & LYONS, C. J. (eds.) *Taylor and Hoyt's Pediatric Ophthalmology and Strabismus (Fifth Edition)*. London: Elsevier.
- PROVENCIO, I., RODRIGUEZ, I. R., JIANG, G., HAYES, W. P., MOREIRA, E. F. & ROLLAG, M. D. 2000. A novel human opsin in the inner retina. *J Neurosci*, 20, 600-5.
- PROVIS, J. M., DUBIS, A. M., MADDESS, T. & CARROLL, J. 2013. Adaptation of the central retina for high acuity vision: cones, the fovea and the avascular zone. *Prog Retin Eye Res*, 35, 63-81.
- QUAZI, F., LENEVICH, S. & MOLDAY, R. S. 2012. ABCA4 is an N-retinylidene-phosphatidylethanolamine and phosphatidylethanolamine importer. *Nat Commun*, 3, 925.
- RAMTOHUL, P., COMET, A. & DENIS, D. 2020. Multimodal Imaging Correlation of the Concentric Macular Rings Sign in Foveal Hypoplasia: A Distinctive Henle Fiber Layer Geometry. *Ophthalmol Retina*, 4, 946-953.
- REESE, M. G., EECKMAN, F. H., KULP, D. & HAUSSLER, D. 1997. Improved splice site detection in Genie. *J Comput Biol*, 4, 311-23.

- REICHENBACH, A. & BRINGMANN, A. 2010. *Müller Cells in the Healthy and Diseased Retina*.
- RENNIE, C. A., CHOWDHURY, S., KHAN, J., RAJAN, F., JORDAN, K., LAMB, R. J. & VIVIAN, A. J. 2005. The prevalence and associated features of posterior embryotoxon in the general ophthalmic clinic. *Eye (Lond)*, 19, 396-9.
- RENTZSCH, P., WITTEN, D., COOPER, G. M., SHENDURE, J. & KIRCHER, M. 2019. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res*, 47, D886-D894.
- REURINK, J., OOSTRIK, J., ABEN, M., RAMOS, M. G., VAN BERKEL, E., OLDAK, M., VAN WIJK, E., KREMER, H., ROOSING, S. & CREMERS, F. P. M. 2022. Minigene-Based Splice Assays Reveal the Effect of Non-Canonical Splice Site Variants in USH2A. *Int J Mol Sci*, 23.
- RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K., REHM, H. L. & COMMITTEE, A. L. Q. A. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 17, 405-24.
- RICHARDSON, C. D., RAY, G. J., DEWITT, M. A., CURIE, G. L. & CORN, J. E. 2016. Enhancing homology-directed genome editing by catalytically active and inactive CRISPR-Cas9 using asymmetric donor DNA. *Nat Biotechnol*, 34, 339-44.
- RINCHIK, E. M., BULTMAN, S. J., HORSTHEMKE, B., LEE, S. T., STRUNK, K. M., SPRITZ, R. A., AVIDANO, K. M., JONG, M. T. & NICHOLLS, R. D. 1993. A gene for the mouse pink-eyed dilution locus and for human type II oculocutaneous albinism. *Nature*, 361, 72-6.
- ROBBE, P., POPITSCH, N., KNIGHT, S. J. L., ANTONIOU, P., BECQ, J., HE, M., KANAPIN, A., SAMSONOVA, A., VAVOULIS, D. V., ROSS, M. T., KINGSBURY, Z., CABES, M., RAMOS, S. D. C., PAGE, S., DREAU, H., RIDOUT, K., JONES, L. J., TUFF-LACEY, A., HENDERSON, S., MASON, J., BUFFA, F. M., VERRILL, C., MALDONADO-PEREZ, D., ROXANIS, I., COLLANTES, E., BROWNING, L., DHAR, S., DAMATO, S., DAVIES, S., CAULFIELD, M., BENTLEY, D. R., TAYLOR, J. C., TURNBULL, C., SCHUH, A. & PROJECT, G. 2018. Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project. *Genet Med*, 20, 1196-1205.
- ROBINSON, J. T., THORVALDSDOTTIR, H., WINCKLER, W., GUTTMAN, M., LANDER, E. S., GETZ, G. & MESIROV, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol*, 29, 24-6.
- ROBINSON, P. N., KÖHLER, S., OELLRICH, A., WANG, K., MUNGALL, C. J., LEWIS, S. E., WASHINGTON, N., BAUER, S., SEELOW, D., KRAWITZ, P., GILISSEN, C., HAENDEL, M. & SMEDLEY, D. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*, 24, 340-8.
- ROHRER, J., SCHWEIZER, A., RUSSELL, D. & KORNFELD, S. 1996. The targeting of Lamp1 to lysosomes is dependent on the spacing of its cytoplasmic tail tyrosine sorting motif relative to the membrane. *J Cell Biol*, 132, 565-76.
- RUSSELL, S. R., DRACK, A. V., CIDECIYAN, A. V., JACOBSON, S. G., LEROY, B. P., VAN CAUWENBERGH, C., HO, A. C., DUMITRESCU, A. V., HAN,

- I. C., MARTIN, M., PFEIFER, W. L., SOHN, E. H., WALSHIRE, J., GARAFALO, A. V., KRISHNAN, A. K., POWERS, C. A., SUMAROKA, A., ROMAN, A. J., VANHONSEBROUCK, E., JONES, E., NERINCKX, F., DE ZAEYTIJD, J., COLLIN, R. W. J., HOYNG, C., ADAMSON, P., CHEETHAM, M. E., SCHWARTZ, M. R., DEN HOLLANDER, W., ASMUS, F., PLATENBURG, G., RODMAN, D. & GIRACH, A. 2022. Intravitreal antisense oligonucleotide sepfarsen in Leber congenital amaurosis type 10: a phase 1b/2 trial. *Nat Med*, 28, 1014-1021.
- SAHU, B. & MAEDA, A. 2018. RPE Visual Cycle and Biochemical Phenotypes of Mutant Mouse Models. *Methods Mol Biol*, 1753, 89-102.
- SALLES, M. V., MOTTA, F. L., DIAS DA SILVA, E., VARELA, P., COSTA, K. A., FILIPPELLI-SILVA, R., MARTIN, R. P., CHIANG, J. P., PESQUERO, J. B. & SALLUM, J. M. F. 2017. Novel Complex ABCA4 Alleles in Brazilian Patients With Stargardt Disease: Genotype-Phenotype Correlation. *Invest Ophthalmol Vis Sci*, 58, 5723-5730.
- SAMORODNITSKY, E., JEWELL, B. M., HAGOPIAN, R., MIYA, J., WING, M. R., LYON, E., DAMODARAN, S., BHATT, D., REESER, J. W., DATTA, J. & ROYCHOWDHURY, S. 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat*, 36, 903-14.
- SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A*, 74, 5463-7.
- SANGERMANO, R., BAX, N. M., BAUWENS, M., VAN DEN BORN, L. I., DE BAERE, E., GARANTO, A., COLLIN, R. W., GOERCHARN-RAMLAL, A. S., DEN ENGELSMAN-VAN DIJK, A. H., ROHRSCHEIDER, K., HOYNG, C. B., CREMERS, F. P. & ALBERT, S. 2016. Photoreceptor Progenitor mRNA Analysis Reveals Exon Skipping Resulting from the ABCA4 c.5461-10T-->C Mutation in Stargardt Disease. *Ophthalmology*, 123, 1375-85.
- SANGERMANO, R., KHAN, M., CORNELIS, S. S., RICHELLE, V., ALBERT, S., GARANTO, A., ELMELIK, D., QAMAR, R., LUGTENBERG, D., VAN DEN BORN, L. I., COLLIN, R. W. J. & CREMERS, F. P. M. 2018. ABCA4 midigenes reveal the full splice spectrum of all reported noncanonical splice site variants in Stargardt disease. *Genome Res*, 28, 100-110.
- SANJURJO-SORIANO, C., ERKILIC, N., BAUX, D., MAMAEVA, D., HAMEL, C. P., MEUNIER, I., ROUX, A. F. & KALATZIS, V. 2020. Genome Editing in Patient iPSCs Corrects the Most Prevalent USH2A Mutations and Reveals Intriguing Mutant mRNA Expression Profiles. *Mol Ther Methods Clin Dev*, 17, 156-173.
- SAQIB, M. A., NIKOPOULOS, K., ULLAH, E., SHER KHAN, F., IQBAL, J., BIBI, R., JARRAL, A., SAJID, S., NISHIGUCHI, K. M., VENTURINI, G., ANSAR, M. & RIVOLTA, C. 2015. Homozygosity mapping reveals novel and known mutations in Pakistani families with inherited retinal dystrophies. *Sci Rep*, 5, 9965.
- SCHIFF, E. R., TAILOR, V. K., CHAN, H. W., THEODOROU, M., WEBSTER, A. R. & MOOSAJEE, M. 2021. Novel Biallelic Variants and Phenotypic Features in Patients with SLC38A8-Related Foveal Hypoplasia. *Int J Mol Sci*, 22.
- SCHIOTH, H. B., ROSHANBIN, S., HAGGLUND, M. G. & FREDRIKSSON, R. 2013. Evolutionary origin of amino acid transporter families SLC32, SLC36 and SLC38 and physiological, pathological and therapeutic aspects. *Mol Aspects Med*, 34, 571-85.

- SCHRAUWEN, I., RAJENDRAN, Y., ACHARYA, A., OHMAN, S., ARVIO, M., PAETAU, R., SIREN, A., AVELA, K., GRANVIK, J., LEAL, S. M., MAATTA, T., KOKKONEN, H. & JARVELA, I. 2024. Optical genome mapping unveils hidden structural variants in neurodevelopmental disorders. *Sci Rep*, 14, 11239.
- SCHWARZ, J. M., COOPER, D. N., SCHUELKE, M. & SEELow, D. 2014. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*, 11, 361-2.
- SIEVERS, F., WILM, A., DINEEN, D., GIBSON, T. J., KARPLUS, K., LI, W., LOPEZ, R., MCWILLIAM, H., REMMERT, M., SODING, J., THOMPSON, J. D. & HIGGINS, D. G. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol*, 7, 539.
- SIGURPALSDOTTIR, B. D., STEFANSSON, O. A., HOLLEY, G., BEYTER, D., ZINK, F., HARDARSON, M., SVERRISSON, S., KRISTINSDOTTIR, N., MAGNUSDOTTIR, D. N., MAGNUSSON, O., GUDBJARTSSON, D. F., HALLDORSSON, B. V. & STEFANSSON, K. 2024. A comparison of methods for detecting DNA methylation from long-read sequencing of human genomes. *Genome Biol*, 25, 69.
- SILVA, A. J., YANG, Z., WOLFE, J., HIRNEISEN, K. A., RUELLE, S. B., TORRES, A., WILLIAMS-HILL, D., KULKA, M. & HELLBERG, R. S. 2021. Application of whole-genome sequencing for norovirus outbreak tracking and surveillance efforts in Orange County, CA. *Food Microbiol*, 98, 103796.
- SIMPSON, J. T., WORKMAN, R. E., ZUZARTE, P. C., DAVID, M., DURSI, L. J. & TIMP, W. 2017. Detecting DNA cytosine methylation using nanopore sequencing. *Nat Methods*, 14, 407-410.
- SINGH, A., KAUSHIK, R., MISHRA, A., SHANKER, A. & JAYARAM, B. 2016. ProTSAV: A protein tertiary structure analysis and validation server. *Biochim Biophys Acta*, 1864, 11-9.
- SJODELL, L., SJOSTROM, A. & ABRAHAMSSON, M. 1996. Transillumination of iris and subnormal visual acuity--ocular albinism? *Br J Ophthalmol*, 80, 617-23.
- SLATKO, B. E., GARDNER, A. F. & AUSUBEL, F. M. 2018. Overview of Next-Generation Sequencing Technologies. *Curr Protoc Mol Biol*, 122, e59.
- SLONAKER, J. R. 1897. A comparative study of the area of acute vision in vertebrates. *Journal of Morphology*, 13, 445-502.
- SMEDLEY, D., JACOBSEN, J. O., JAGER, M., KOHLER, S., HOLTGREWE, M., SCHUBACH, M., SIRAGUSA, E., ZEMOJTEL, T., BUSKE, O. J., WASHINGTON, N. L., BONE, W. P., HAENDEL, M. A. & ROBINSON, P. N. 2015. Next-generation diagnostics and disease-gene discovery with the Exomiser. *Nat Protoc*, 10, 2004-15.
- SONG, X., WANG, M., CHEN, X., ZHANG, X., YANG, Y., LIU, Z. & YAO, L. 2021. Quantifying Protein Electrostatic Interactions in Cells by Nuclear Magnetic Resonance Spectroscopy. *J Am Chem Soc*, 143, 19606-19613.
- SPRINGER, A. D. & HENDRICKSON, A. E. 2004. Development of the primate area of high acuity. 1. Use of finite element analysis models to identify mechanical variables affecting pit formation. *Vis Neurosci*, 21, 53-62.
- SRIDHAR, M. S. 2018. Anatomy of cornea and ocular surface. *Indian J Ophthalmol*, 66, 190-194.
- SRINIVASAN, S., FERNANDEZ-SAMPEDRO, M. A., MORILLO, M., RAMON, E., JIMENEZ-ROSES, M., CORDOMI, A. & GARRIGA, P. 2018. Human Blue

- Cone Opsin Regeneration Involves Secondary Retinal Binding with Analog Specificity. *Biophys J*, 114, 1285-1294.
- SRINIVASAN, S., RAMON, E., CORDOMI, A. & GARRIGA, P. 2014. Binding specificity of retinal analogs to photoactivated visual pigments suggest mechanism for fine-tuning GPCR-ligand interactions. *Chem Biol*, 21, 369-78.
- STEVENS, G., VAN BEUKERING, J., JENKINS, T. & RAMSAY, M. 1995. An intragenic deletion of the P gene is the common mutation causing tyrosinase-positive oculocutaneous albinism in southern African Negroids. *Am J Hum Genet*, 56, 586-91.
- STOLER, N. & NEKRUTENKO, A. 2021. Sequencing error profiles of Illumina sequencing instruments. *NAR Genom Bioinform*, 3, lqab019.
- SUND, K. L., LIU, J., LEE, J., GARBE, J., ABDELHAMED, Z., MAAG, C., HALLINAN, B., WU, S. W., SPERRY, E., DESHPANDE, A., STOTTMANN, R., SMOLAREK, T. A., DYER, L. M. & HESTAND, M. S. 2024. Long-read sequencing and optical genome mapping identify causative gene disruptions in noncoding sequence in two patients with neurologic disease and known chromosome abnormalities. *Am J Med Genet A*, e63818.
- TANG, H. & THOMAS, P. D. 2016. PANTHER-PSEP: predicting disease-causing genetic variants using position-specific evolutionary preservation. *Bioinformatics*, 32, 2230-2.
- TARPEY, P., THOMAS, S., SARVANANTHAN, N., MALLYA, U., LISGO, S., TALBOT, C. J., ROBERTS, E. O., AWAN, M., SURENDRAN, M., MCLEAN, R. J., REINECKE, R. D., LANGMANN, A., LINDNER, S., KOCH, M., JAIN, S., WOODRUFF, G., GALE, R. P., BASTAWROUS, A., DEGG, C., DROUTSAS, K., ASPROUDIS, I., ZUBCOV, A. A., PIEH, C., VEAL, C. D., MACHADO, R. D., BACKHOUSE, O. C., BAUMBER, L., CONSTANTINESCU, C. S., BRODSKY, M. C., HUNTER, D. G., HERTLE, R. W., READ, R. J., EDKINS, S., O'MEARA, S., PARKER, A., STEVENS, C., TEAGUE, J., WOOSTER, R., FUTREAL, P. A., TREMBATH, R. C., STRATTON, M. R., RAYMOND, F. L. & GOTTLÖB, I. 2006. Mutations in FRMD7, a newly identified member of the FERM family, cause X-linked idiopathic congenital nystagmus. *Nat Genet*, 38, 1242-4.
- TAVTIGIAN, S. V., DEFFENBAUGH, A. M., YIN, L., JUDKINS, T., SCHOLL, T., SAMOLLO, P. B., DE SILVA, D., ZHARKIKH, A. & THOMAS, A. 2006. Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. *J Med Genet*, 43, 295-305.
- TAYLOR, R. L., ARNO, G., POULTER, J. A., KHAN, K. N., MORARJI, J., HULL, S., PONTIKOS, N., RUEDA MARTIN, A., SMITH, K. R., ALI, M., TOOMES, C., MCKIBBIN, M., CLAYTON-SMITH, J., GRUNEWALD, S., MICHAELIDES, M., MOORE, A. T., HARDCASTLE, A. J., INGLEHEARN, C. F., WEBSTER, A. R., BLACK, G. C., CONSORTIUM, U. K. I. R. D. & THE, G. P. 2017. Association of Steroid 5alpha-Reductase Type 3 Congenital Disorder of Glycosylation With Early-Onset Retinal Dystrophy. *JAMA Ophthalmol*, 135, 339-347.
- TAYLOR, T. L., VOLKENING, J. D., DEJESUS, E., SIMMONS, M., DIMITROV, K. M., TILLMAN, G. E., SUAREZ, D. L. & AFONSO, C. L. 2019. Rapid, multiplexed, whole genome and plasmid sequencing of foodborne pathogens using long-read nanopore technology. *Sci Rep*, 9, 16350.

- TELLER, D. C., OKADA, T., BEHNKE, C. A., PALCZEWSKI, K. & STENKAMP, R. E. 2001. Advances in determination of a high-resolution three-dimensional structure of rhodopsin, a model of G-protein-coupled receptors (GPCRs). *Biochemistry*, 40, 7761-72.
- TERAKITA, A. 2005. The opsins. *Genome Biol*, 6, 213.
- TERASAWA, K., TOMABECHI, Y., IKEDA, M., EHARA, H., KUKIMOTO-NIINO, M., WAKIYAMA, M., PODYMA-INOUE, K. A., RAJAPAKSHE, A. R., WATABE, T., SHIROUZU, M. & HARA-YOKOYAMA, M. 2016. Lysosome-associated membrane proteins-1 and -2 (LAMP-1 and LAMP-2) assemble via distinct modes. *Biochem Biophys Res Commun*, 479, 489-495.
- THEOS, A. C., TENZA, D., MARTINA, J. A., HURBAIN, I., PEDEN, A. A., SVIDERSKAYA, E. V., STEWART, A., ROBINSON, M. S., BENNETT, D. C., CUTLER, D. F., BONIFACINO, J. S., MARKS, M. S. & RAPOSO, G. 2005. Functions of adaptor protein (AP)-3 and AP-1 in tyrosinase sorting from endosomes to melanosomes. *Mol Biol Cell*, 16, 5356-72.
- THOMAS, M. G., CROSIER, M., LINDSAY, S., KUMAR, A., ARAKI, M., LEROY, B. P., MCLEAN, R. J., SHETH, V., MACONACHIE, G., THOMAS, S., MOORE, A. T. & GOTTLOB, I. 2014a. Abnormal retinal development associated with FRMD7 mutations. *Hum Mol Genet*, 23, 4086-93.
- THOMAS, M. G., KUMAR, A., MOHAMMAD, S., PROUDLOCK, F. A., ENGLE, E. C., ANDREWS, C., CHAN, W. M., THOMAS, S. & GOTTLOB, I. 2011. Structural grading of foveal hypoplasia using spectral-domain optical coherence tomography a predictor of visual acuity? *Ophthalmology*, 118, 1653-60.
- THOMAS, P. D., EBERT, D., MURUGANUJAN, A., MUSHAYAHAMA, T., ALBOU, L. P. & MI, H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci*, 31, 8-22.
- THOMAS, S., THOMAS, M. G., ANDREWS, C., CHAN, W. M., PROUDLOCK, F. A., MCLEAN, R. J., PRADEEP, A., ENGLE, E. C. & GOTTLOB, I. 2014b. Autosomal-dominant nystagmus, foveal hypoplasia and presenile cataract associated with a novel PAX6 mutation. *Eur J Hum Genet*, 22, 344-9.
- THOMPSON, D. A., MCHENRY, C. L., LI, Y., RICHARDS, J. E., OTHMAN, M. I., SCHWINGER, E., VOLLRATH, D., JACOBSON, S. G. & GAL, A. 2002. Retinal dystrophy due to paternal isodisomy for chromosome 1 or chromosome 2, with homoallelism for mutations in RPE65 or MERTK, respectively. *Am J Hum Genet*, 70, 224-9.
- TOMITA, Y., TAKEDA, A., OKINAGA, S., TAGAMI, H. & SHIBAHARA, S. 1989. Human oculocutaneous albinism caused by single base insertion in the tyrosinase gene. *Biochem Biophys Res Commun*, 164, 990-6.
- TORAL, M. A., VELEZ, G., BOUDREAU, K., SCHAEFER, K. A., XU, Y., SAFFRA, N., BASSUK, A. G., TSANG, S. H. & MAHAJAN, V. B. 2017. Structural modeling of a novel SLC38A8 mutation that causes foveal hypoplasia. *Mol Genet Genomic Med*, 5, 202-209.
- TORKASHVAND, A., MOHEBBI, M. & HASHEMI, H. 2018. A novel PAX6 nonsense mutation identified in an Iranian family with various eye anomalies. *J Curr Ophthalmol*, 30, 234-238.
- TRAVERS, K. J., CHIN, C. S., RANK, D. R., EID, J. S. & TURNER, S. W. 2010. A flexible and efficient template format for circular consensus sequencing and SNP detection. *Nucleic Acids Res*, 38, e159.
- TROTMAN, J., ARMSTRONG, R., FIRTH, H., TRAYERS, C., WATKINS, J., ALLINSON, K., JACQUES, T. S., NICHOLSON, J. C., BURKE, G. A. A., GENOMICS ENGLAND RESEARCH, C., BEHJATI, S., MURRAY, M. J.,



- HOOK, C. E. & TARPEY, P. 2022. The NHS England 100,000 Genomes Project: feasibility and utility of centralised genome sequencing for children with cancer. *Br J Cancer*, 127, 137-144.
- TSANG, S. H., AYCINENA, A. R. P. & SHARMA, T. 2018. Ciliopathy: Bardet-Biedl Syndrome. In: TSANG, S. H. & SHARMA, T. (eds.) *Atlas of Inherited Retinal Diseases*. Cham: Springer International Publishing.
- TSANG, S. H. & SHARMA, T. 2018a. Leber Congenital Amaurosis. In: TSANG, S. H. & SHARMA, T. (eds.) *Atlas of Inherited Retinal Diseases*. Cham: Springer International Publishing.
- TSANG, S. H. & SHARMA, T. 2018b. X-linked Ocular Albinism. *Adv Exp Med Biol*, 1085, 49-52.
- TURNBULL, C., SCOTT, R. H., THOMAS, E., JONES, L., MURUGAESU, N., PRETTY, F. B., HALAI, D., BAPLE, E., CRAIG, C., HAMBLIN, A., HENDERSON, S., PATCH, C., O'NEILL, A., DEVEREAU, A., SMITH, K., MARTIN, A. R., SOSINSKY, A., MCDONAGH, E. M., SULTANA, R., MUELLER, M., SMEDLEY, D., TOMS, A., DINH, L., FOWLER, T., BALE, M., HUBBARD, T., RENDON, A., HILL, S., CAULFIELD, M. J. & GENOMES, P. 2018. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, 361, k1687.
- TZOULAKI, I., WHITE, I. M. & HANSON, I. M. 2005. PAX6 mutations: genotype-phenotype correlations. *BMC Genet*, 6, 27.
- UEMATSU, M. & BASKIN, J. M. 2023. Barcode-free multiplex plasmid sequencing using Bayesian analysis and nanopore sequencing. *bioRxiv*.
- UHLHORN, S. R., BORJA, D., MANNS, F. & PAREL, J. M. 2008. Refractive index measurement of the isolated crystalline lens using optical coherence tomography. *Vision Res*, 48, 2732-8.
- UNTERGASSER, A., CUTCUTACHE, I., KORESSAAR, T., YE, J., FAIRCLOTH, B. C., REMM, M. & ROZEN, S. G. 2012. Primer3--new capabilities and interfaces. *Nucleic Acids Res*, 40, e115.
- VAN GENDEREN, M. M., RIEMSLAG, F. C., SCHUIL, J., HOEBEN, F. P., STILMA, J. S. & MEIRE, F. M. 2006. Chiasmal misrouting and foveal hypoplasia without albinism. *Br J Ophthalmol*, 90, 1098-102.
- VAN SCHIL, K., NAESSENS, S., VAN DE SOMPELE, S., CARRON, M., ASLANIDIS, A., VAN CAUWENBERGH, C., KATHRIN MAYER, A., VAN HEETVELDE, M., BAUWENS, M., VERDIN, H., COPPIETERS, F., GREENBERG, M. E., YANG, M. G., KARLSTETTER, M., LANGMANN, T., DE PRETER, K., KOHL, S., CHERRY, T. J., LEROY, B. P., GROUP, C. N. V. S. & DE BAERE, E. 2018. Mapping the genomic landscape of inherited retinal disease genes prioritizes genes prone to coding and noncoding copy-number variations. *Genet Med*, 20, 202-213.
- VASER, R., ADUSUMALLI, S., LENG, S. N., SIKIC, M. & NG, P. C. 2016. SIFT missense predictions for genomes. *Nat Protoc*, 11, 1-9.
- VAZQUEZ, P., ARROBA, A. I., CECCONI, F., DE LA ROSA, E. J., BOYA, P. & DE PABLO, F. 2012. Atg5 and Ambra1 differentially modulate neurogenesis in neural stem cells. *Autophagy*, 8, 187-99.
- VENABLES, J. P. 2007. Downstream intronic splicing enhancers. *FEBS Lett*, 581, 4127-31.
- VERDON, Q., BOONEN, M., RIBES, C., JADOT, M., GASNIER, B. & SAGNE, C. 2017. SNAT7 is the primary lysosomal glutamine exporter required for extracellular protein-dependent growth of cancer cells. *Proc Natl Acad Sci U S A*, 114, E3602-E3611.

- WANG, L., ZHANG, J., CHEN, N., WANG, L., ZHANG, F., MA, Z., LI, G. & YANG, L. 2018. Application of Whole Exome and Targeted Panel Sequencing in the Clinical Molecular Diagnosis of 319 Chinese Families with Inherited Retinal Dystrophy and Comparison Study. *Genes (Basel)*, 9.
- WANG, S., HAYNES, C., BARANY, F. & OTT, J. 2009. Genome-wide autozygosity mapping in human populations. *Genet Epidemiol*, 33, 172-80.
- WATERHOUSE, A., BERTONI, M., BIENERT, S., STUDER, G., TAURIELLO, G., GUMIENNY, R., HEER, F. T., DE BEER, T. A. P., REMPFER, C., BORDOLI, L., LEPORE, R. & SCHWEDE, T. 2018. SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Res*, 46, W296-W303.
- WATSON, C. M., CRINNION, L. A., SIMMONDS, J., CAMM, N., ADLARD, J. & BONTHRON, D. T. 2021. Long-read nanopore sequencing enables accurate confirmation of a recurrent PMS2 insertion-deletion variant located in a region of complex genomic architecture. *Cancer Genet*, 256-257, 122-126.
- WEI, A. H., ZANG, D. J., ZHANG, Z., LIU, X. Z., HE, X., YANG, L., WANG, Y., ZHOU, Z. Y., ZHANG, M. R., DAI, L. L., YANG, X. M. & LI, W. 2013. Exome sequencing identifies SLC24A5 as a candidate gene for nonsyndromic oculocutaneous albinism. *J Invest Dermatol*, 133, 1834-40.
- WEINER, C., HECHT, I., ROTENSTREICH, Y., GUTTMAN, S., OR, L., MORAD, Y., SHAPIRA, G., SHOMRON, N. & PRAS, E. 2020. The pathogenicity of SLC38A8 in five families with foveal hypoplasia and congenital nystagmus. *Exp Eye Res*, 193, 107958.
- WEISSCHUH, N., MAZZOLA, P., ZULEGER, T., SCHAEFERHOFF, K., KUHLEWEIN, L., KORTUM, F., WITT, D., LIEBMANN, A., FALB, R., POHL, L., REITH, M., STUHN, L. G., BERTRAND, M., MULLER, A., CASADEI, N., KELEMEN, O., KELBSCH, C., KERNSTOCK, C., RICHTER, P., SADLER, F., DEMIDOV, G., SCHUTZ, L., ADMARD, J., STURM, M., GRASSHOFF, U., TONAGEL, F., HEINRICH, T., NASSER, F., WISSINGER, B., OSSOWSKI, S., KOHL, S., RIESS, O., STINGL, K. & HAACK, T. B. 2024. Diagnostic genome sequencing improves diagnostic yield: a prospective single-centre study in 1000 patients with inherited eye diseases. *J Med Genet*, 61, 186-195.
- WENGER, A. M., PELUSO, P., ROWELL, W. J., CHANG, P. C., HALL, R. J., CONCEPCION, G. T., EBLER, J., FUNGTAMMASAN, A., KOLESNIKOV, A., OLSON, N. D., TOPFER, A., ALONGE, M., MAHMOUD, M., QIAN, Y., CHIN, C. S., PHILLIPPY, A. M., SCHATZ, M. C., MYERS, G., DEPRISTO, M. A., RUAN, J., MARSCHALL, T., SEDLAZECK, F. J., ZOOK, J. M., LI, H., KOREN, S., CARROLL, A., RANK, D. R. & HUNKAPILLER, M. W. 2019. Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol*, 37, 1155-1162.
- WILDE, S., TIMPSON, A., KIRSANOW, K., KAISER, E., KAYSER, M., UNTERLANDER, M., HOLLFELDER, N., POTEKHINA, I. D., SCHIER, W., THOMAS, M. G. & BURGER, J. 2014. Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc Natl Acad Sci U S A*, 111, 4832-7.
- WORTH, C. L., PREISSNER, R. & BLUNDELL, T. L. 2011. SDM--a server for predicting effects of mutations on protein stability and malfunction. *Nucleic Acids Res*, 39, W215-22.

- WRIGHT, C. B., REDMOND, T. M. & NICKERSON, J. M. 2015. Chapter Twenty-Five - A History of the Classical Visual Cycle. *In: HEJTMANCIK, J. F. & NICKERSON, J. M. (eds.) Progress in Molecular Biology and Translational Science*. Academic Press.
- WRIGHT, C. F., PARKER, M. & LUCASSEN, A. M. 2019. When genomic medicine reveals misattributed genetic relationships—the debate about disclosure revisited. *Genetics in Medicine*, 21, 97-101.
- XU, J., GU, J., PEI, W., ZHANG, Y., WANG, L. & GAO, J. 2023. The role of lysosomal membrane proteins in autophagy and related diseases. *FEBS J.*
- YAHYA, S., WATSON, C. M., CARR, I., MCKIBBIN, M., CRINNION, L. A., TAYLOR, M., BONIN, H., FLETCHER, T., EL-ASRAG, M. E., ALI, M., TOOMES, C. & INGLEHEARN, C. F. 2023. Long-Read Nanopore Sequencing of RPGR ORF15 is Enhanced Following DNase I Treatment of MinION Flow Cells. *Mol Diagn Ther*, 27, 525-535.
- YAMADA, M., OKUNO, H., OKAMOTO, N., SUZUKI, H., MIYA, F., TAKENOUCI, T. & KOSAKI, K. 2023. Diagnosis of Prader-Willi syndrome and Angelman syndrome by targeted nanopore long-read sequencing. *Eur J Med Genet*, 66, 104690.
- YAMAGUCHI, F., SAKANE, H., MORISHITA, Y., HATA, T. & AKASAKI, K. 2022. Importance of Glycine Preceding Pivotal Tyrosine in the Lysosome-Targeting Signal GYQTI of Lysosome-Associated Membrane Protein-1 (LAMP-1). *BPB Reports*, 5, 99-104.
- YANG, J. & ZHANG, Y. 2015. I-TASSER server: new development for protein structure and function predictions. *Nucleic Acids Res*, 43, W174-81.
- YANOFF, M. & SASSANI, J. W. 2020. 9 - Uvea. *In: YANOFF, M. & SASSANI, J. W. (eds.) Ocular Pathology (Eighth Edition)*. London: Elsevier.
- YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics*, 13, 134.
- YEO, G. & BURGE, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11, 377-94.
- YOUNG, A., DANDEKAR, U., PAN, C., SADER, A., ZHENG, J. J., LEWIS, R. A. & FARBER, D. B. 2016. GNAI3: Another Candidate Gene to Screen in Persons with Ocular Albinism. *PLoS One*, 11, e0162273.
- YOUNG, A., JIANG, M., WANG, Y., AHMEDLI, N. B., RAMIREZ, J., REESE, B. E., BIRNBAUMER, L. & FARBER, D. B. 2011. Specific interaction of Gai3 with the Oa1 G-protein coupled receptor controls the size and density of melanosomes in retinal pigment epithelium. *PLoS One*, 6, e24376.
- YU, J., SZABO, A., PAGNAMENTA, A. T., SHALABY, A., GIACOPUZZI, E., TAYLOR, J., SHEARS, D., PONTIKOS, N., WRIGHT, G., MICHAELIDES, M., HALFORD, S. & DOWNES, S. 2022. SVRare: discovering disease-causing structural variants in the 100K Genomes Project. *medRxiv*, 2021.10.15.21265069.
- ZELICKSON, A. S., WINDHORST, D. B., WHITE, J. G. & GOOD, R. A. 1967. The Chediak-Higashi syndrome: formation of giant melanosomes and the basis of hypopigmentation. *J Invest Dermatol*, 49, 575-81.
- ZHANG, J., ZENG, W., HAN, Y., LEE, W. R., LIOU, J. & JIANG, Y. 2023. Lysosomal LAMP proteins regulate lysosomal pH by direct inhibition of the TMEM175 channel. *Mol Cell*, 83, 2524-2539 e7.

- ZHANG, Q., ZHAO, B., LI, W., OISO, N., NOVAK, E. K., RUSINIAK, M. E., GAUTAM, R., CHINTALA, S., O'BRIEN, E. P., ZHANG, Y., ROE, B. A., ELLIOTT, R. W., EICHER, E. M., LIANG, P., KRATZ, C., LEGIUS, E., SPRITZ, R. A., O'SULLIVAN, T. N., COPELAND, N. G., JENKINS, N. A. & SWANK, R. T. 2003. Ru2 and Ru encode mouse orthologs of the genes mutated in human Hermansky-Pudlak syndrome types 5 and 6. *Nat Genet*, 33, 145-53.
- ZHANG, X., WAKELING, M., WARE, J. & WHIFFIN, N. 2021. Annotating high-impact 5'untranslated region variants with the UTRannotator. *Bioinformatics*, 37, 1171-1173.
- ZHANG, Z., ALBERS, T., FIUMERA, H. L., GAMEIRO, A. & GREWER, C. 2009. A conserved Na(+) binding site of the sodium-coupled neutral amino acid transporter 2 (SNAT2). *J Biol Chem*, 284, 25314-23.
- ZHANG, Z., GAMEIRO, A. & GREWER, C. 2008. Highly conserved asparagine 82 controls the interaction of Na<sup>+</sup> with the sodium-coupled neutral amino acid transporter SNAT2. *J Biol Chem*, 283, 12284-92.
- ZHAO, Y., ZHAO, J., GU, Y., CHEN, B., GUO, J., XIE, J., YAN, Q., MA, Y., WU, Y., ZHANG, J., LU, Q. & LIU, J. 2021. Outer Retinal Layer Thickness Changes in White Matter Hyperintensity and Parkinson's Disease. *Front Neurosci*, 15, 741651.
- ZHONG, J., YOU, B., XU, K., ZHANG, X., XIE, Y. & LI, Y. 2021. GPR143 genotypic and ocular phenotypic characterisation in a Chinese cohort with ocular albinism. *Ophthalmic Genet*, 42, 717-724.
- ZHOU, Y., LI, S., HUANG, L., YANG, Y., ZHANG, L., YANG, M., LIU, W., RAMASAMY, K., JIANG, Z., SUNDARESAN, P., ZHU, X. & YANG, Z. 2018. A splicing mutation in aryl hydrocarbon receptor associated with retinitis pigmentosa. *Hum Mol Genet*, 27, 2587.
- ZIZIOLI, D., MEYER, C., GUHDE, G., SAFTIG, P., VON FIGURA, K. & SCHU, P. 1999. Early embryonic death of mice deficient in gamma-adaptin. *J Biol Chem*, 274, 5385-90.

## Chapter 8 Appendix

### 8.1 Appendix A : gene list and phenotypic keywords in 100KGP filtering

#### 8.1.1 All the known genes underlying FH

<i>AHR</i>	<i>CRB1</i>	<i>IKBKG</i>	<i>PROKR2</i>
<i>AP3B1</i>	<i>DCT</i>	<i>LRMDA</i>	<i>PRSS56</i>
<i>AP3D1</i>	<i>DTNBP1</i>	<i>LRP5</i>	<i>SLC24A5</i>
<i>ATF6</i>	<i>FRMD7</i>	<i>LYST</i>	<i>SLC38A8</i>
<i>ATOH7</i>	<i>FZD4</i>	<i>MFRP</i>	<i>SLC45A2</i>
<i>BEST1</i>	<i>GNAT2</i>	<i>NDP</i>	<i>SOX2</i>
<i>BLOC1S3</i>	<i>GPR143</i>	<i>NR2F1</i>	<i>SOX3</i>
<i>BLOC1S5</i>	<i>HESX1</i>	<i>OCA2</i>	<i>TMEM98</i>
<i>CACNA1F</i>	<i>HPS1</i>	<i>OTX2</i>	<i>TSPAN12</i>
<i>CNGA3</i>	<i>HPS3</i>	<i>PAX6</i>	<i>TYR</i>
<i>CNGB3</i>	<i>HPS4</i>	<i>PDE6C</i>	<i>TYRP1</i>
<i>COL11A1</i>	<i>HPS5</i>	<i>PDE6H</i>	<i>VAX1</i>
<i>COL2A1</i>	<i>HPS6</i>	<i>PLDN</i>	<i>ZNF408</i>

#### 8.1.2 Curated FH genes for a gene-specific analysis using SVRare

<i>AHR</i>	<i>COL2A1</i>	<i>HPS4</i>	<i>PAX6</i>
<i>AP3B1</i>	<i>DCT</i>	<i>HPS5</i>	<i>PLDN</i>
<i>AP3D1</i>	<i>DTNBP1</i>	<i>HPS6</i>	<i>SLC24A5</i>
<i>BLOC1S3</i>	<i>FRMD7</i>	<i>IKBKG</i>	<i>SLC45A2</i>
<i>BLOC1S5</i>	<i>GPR143</i>	<i>LRMDA</i>	<i>TYR</i>
<i>CACNA1F</i>	<i>HPS1</i>	<i>LYST</i>	<i>TYRP1</i>
<i>COL11A1</i>	<i>HPS3</i>	<i>OCA2</i>	

### 8.1.3 HPO terms to identify a suspected FH phenotype in the 100KGP

Abnormal colour vision	Congenital nystagmus	Nystagmus
Abnormal fundus	Constriction of the peripheral visual field	Optic decussation defect
Abnormal light adapted electroretinogram	Foveal hypoplasia	Progressive vision loss
Abnormal macular morphology, Abnormality of retinal pigmentation	Foveal plana	Reduced visual acuity
Central scotoma	Hypoplasia of the fovea	Retinal dystrophy,
Chiasmal misrouting	Involuntarily eye movements	Rod-cone dystrophy
Colour vision defect	Involuntarily eye oscillations	Saccadic eye movements
Cone-rod dystrophy	Macular dystrophy	Visual impairment
	Nonprogressive vision loss,	Visual loss

## 8.2 Appendix B : Scripts

### 8.2.1 Gene-Variant workflow v1.6 for the analysis in the 100KGP

#### 1. Copy the workflow to own directory:

```
$cp -R /gel_data_resources/workflows/BRS_tools_geneVariantWorkflow/1.6
/home/mderar/My_workflows
```

#### 2. Specify genes:

```
$nano gene_list.txt
SLC38A8
```

#### 3. Specify GECIP project and dataset for analysis:

```
$nano variant_workflow_inputs.json
```

```
{
  "genetic_report.project": "re_gecip_hearing_and_sight",
  "genetic_report.data_version": "main-programme_v11_2020-12-17",
  "genetic_report.gene_input": "gene_list.txt",
  "genetic_report.vep_config_file": "vep_config_file.txt",
  "genetic_report.vep_extra_configs": {
    "GRCh37": [
      "--custom
'/public_data_resources/clinvar/20200604/clinvar/vcf_GRCh37/clinvar_2020060
2.vcf.gz,ClinVar,vcf,exact,0,CLNDN,CLNDNINCL,CLNDISDB,CLNDISDBINCL,
CLNHGVS,CLNREVSTAT,CLNSIG,CLNSIGCONF,CLNSIGINCL,CLNVC,CLN
VCSO,CLNVI'"
    ],
    "GRCh38": [
      "--custom
'/public_data_resources/clinvar/20200604/clinvar/vcf_GRCh38/clinvar_2020060
2.vcf.gz,ClinVar,vcf,exact,0,CLNDN,CLNDNINCL,CLNDISDB,CLNDISDBINCL,
CLNHGVS,CLNREVSTAT,CLNSIG,CLNSIGCONF,CLNSIGINCL,CLNVC,CLN
VCSO,CLNVI'"
    ]
  },
  "genetic_report.bcftools": "module load bio/BCFtools/1.10.2-GCC-8.3.0",
  "genetic_report.R": "module load lang/R/3.6.2-foss-2019b",
  "genetic_report.vep": "module load bio/VEP/98.0-foss-2019a-Perl-5.28.1",
  "genetic_report.vep_cache":
  "/resources/data/vep.caches/from.pegasus/98/.vep",
  "genetic_report.vep_cache_version": "98",
  "genetic_report.GRCH37_and_GRCH38": "GRCH38_GRCH37_files.txt"
```

#### 4. Define project: re\_gecip\_hearing\_and\_sight

```
$nano submit_workflow.txt.sh

#BSUB -q inter
#BSUB -P re_gecip_hearing_and_sight
#BSUB -o vw_%J.stdout
#BSUB -e vw_%J.stderr
#BSUB -cwd .
#BSUB -n 2
#BSUB -R "rusage[mem=10000] span[hosts=1]"
#BSUB -M 20000

module purge
module load bio/cromwell/51

java -Dconfig.file=cromwell.conf \
-jar $CROMWELL_JAR \
run variant_workflow.wdl \
-i variant_workflow_inputs.json \
-o variant_workflow_options.json \
-m metadata_lastrun.json
```

#### 5. Submitting workflow:

```
$bsub < submit_workflow.sh
```

### 8.2.2 100K Gene\_Variant filter configuration and execution

#### 1. Copy script to own directory:

```
$cp re_gecip/GW_SB/GeneVariantWorkflow/filter_gene_variant_workflow-py
/home/mderar/My_workflows
```

#### 2. Activate idppy3

```
$• /resources/conda/miniconda3/etc/profile.d/conda.sh
$conda activate idppy3
```

#### 3. Run filtering script on Gene-Variant Workflow output:

```
$python filter_gene_variant_workflow-py
finaloutput/data/chr16_SLC38A8_GRCh38_annotated_variants.tsv
```



## 8.2.3 ONT pipeline for MinION sequencing using the Flongle R9.4.1

### 1. Guppy script v5

```
#!/bin/bash
```

```
# Use current environment variables and modules
```

```
#$ -V
```

```
# Set current working directory
```

```
#$ -cwd
```

```
# Request 3 hours of runtime
```

```
#$ -l h_rt=2:00:00
```

```
# Request Nvidia node
```

```
#$ -l coproc_v100=1
```

```
# Email at the beginning and end of the job
```

```
#$ -m be
```

```
module load singularity
```

```
module load cuda/10.1.168
```

```
# Run the following command
```

```
singularity run --nv --bind /nobackup:/nobackup /nobackup/containers/guppy-
```

```
gpu-5.0.16.simg guppy_basecaller -x "cuda:0" -i
```

```
/nobackup/ummamd/5plasmids/fast5 -s /nobackup/ummamd/5plasmids/fastq --
```

```
flow cell FLO-MIN106 --kit SQK-LSK110
```

### 2. Installing Miniconda 3

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
```

```
-O ~/miniconda3/miniconda.sh
```

```
bash Miniconda3-py39_4.12.0-Linux-x86_64.sh
```

```
conda create --name ont_tools
```

```
conda activate ont_tools
```

### 3. Performing base calling using Guppy v5:

```
mkdir project_name
mkdir fast5 fastq
cp -r /nobackup/mderar/20220630_run149/plasmid_assembly/fast5 .
nano guppy_v5.2.sh
qsub guppy_v5.2.sh
```

### 4. concatenating FASTQ files:

```
cd /nobackup/mderar/project_name/guppy_output
cat fail/* > ./fail.fastq
cat pass/* > ./pass.fastq
gzip pass.fastq
gzip fail.fastq
cat *fastq* > filename.fastq.gz
gunzip filename.fastq.gz
```

### 5. Adaptor trimming by Porechop

```
porechop --discard_middle -i filename.fastq -o filename.adapt.trim.fastq >>
filenameadapt.trim.log.std.out >> filename.adapt.trim.log.std.err
```

### 6. Obtain read statistics on the trimmed FASTQ file using NanoStat

```
NanoStat --fastq filename.adapt.trim.fastq>
filename.adapt.trim.fastq.unfiltered.NanoStat
```

### 7. Filter reads based on quality and length using NanoFilt

```
NanoFilt filename.adapt.trim.fastq -q 10 --headcrop 75 --length 750 --maxlength
9000 | gzip > filename.concat.adapt.trim.filt.fastq.gz
```

### 8. Obtain read statistics on the filtered FASTQ file using NanoStat

```
NanoStat --fastq filename.concat.adapt.trim.filt.fastq.gz >
filename.concat.adapt.trim.filt.fastq.filtered.NanoStat
```

## 9. Alignment of reads to reference sequence using Minimap2

```
minimap2 -ax map-ont /nobackup/mderar/project_name/references/reference.fa
filename.concat.adapt.trim.filt.fastq.gz>> filename.minimap.std.err >>
filename.concat.sam
```

## 10. Sort and Index BAM file using Samtools

```
samtools view -bh filename.concat.sam > filename.concat.bam
samtools sort filename.concat.bam -o filename.concat.sorted.bam
samtools index filename.concat.sorted.bam
```

### 8.2.4 Quality control using FASTQC

```
fastqc -o . filename.bam
```

### 8.2.5 Genome wide annotation using VEP plugins of CADD, SpliceAI and UTRannotator

#### 1. Script to download the Ensembl VEP

```
curl -O ftp://ftp.ensembl.org/pub/release-
111/variation/indexed_vep_cache/homo_sapiens_vep_111_GRCh37.tar.gz && \
\
    mkdir -p homo_sapiens_vep_111_GRCh37 && \
    mv homo_sapiens_vep_111_GRCh37.tar.gz
homo_sapiens_vep_111_GRCh37/ && \
(
    cd homo_sapiens_vep_111_GRCh37 && \
    tar xzf homo_sapiens_vep_111_GRCh37.tar.gz
)
```

#### 2. Script to download VEP plugins of SpliceAI, CADD and UTRannotator

```
#$ -l h_rt=48:00:00
```

```
#$ -l h_vmem=1G
```

```
#$ -m be
```

```
#$ -M ummamd@leeds.ac.uk
```

```
#$ -cwd -V
```

```
curl -O
```

```
"https://download.molgeniscloud.org/downloads/vip/resources/GRCh37/spliceai_scores.masked.in"
```

```
curl -O
```

```
"https://download.molgeniscloud.org/downloads/vip/resources/GRCh37/spliceai_scores.masked.in"
```

```
curl -O
```

```
"https://download.molgeniscloud.org/downloads/vip/resources/GRCh37/spliceai_scores.masked.sn"
```

```
curl -O
```

```
"https://download.molgeniscloud.org/downloads/vip/resources/GRCh37/spliceai_scores.masked.sn"
```

```
curl -O
```

```
"https://kircherlab.bihealth.org/download/CADD/v1.7/GRCh37/whole_genome_SNVs.tsv.gz"
```

```
curl -O
```

```
"https://kircherlab.bihealth.org/download/CADD/v1.7/GRCh37/whole_genome_SNVs.tsv.gz.tbi"
```

```
curl -O
```

```
"https://kircherlab.bihealth.org/download/CADD/v1.7/GRCh37/gnomad.genomes-exomes.r4.0.indel"
```

```
curl -O
```

```
"https://kircherlab.bihealth.org/download/CADD/v1.7/GRCh37/gnomad.genomes-exomes.r4.0.indel"
```

```
curl -O  
https://download.molgeniscloud.org/downloads/vip/resources/GRCh37/uORF_5  
UTR_PUBLIC.txt
```

### 3. Script for VEP annotation using GRCh37

```
#$ -l h_rt=48:00:00
```

```
#$ -l h_vmem=10G
```

```
#$ -pe smp 2
```

```
#$ -m be
```

```
#$ -M ummamd@leeds.ac.uk
```

```
#$ -cwd -V
```

```
# amend directory names to ummamd
```

```
module load anaconda/2019.10
```

```
source activate /nobackup/ummamd/.conda/envs/vep
```

```
vep \
```

```
--cache \
```

```
--database \
```

```
--offline \
```

```
--dir /nobackup/ummamd/VEP_annotation/homo_sapiens_vep_111_GRCh37 \
```

```
--assembly GRCh37 \
```

```
--force_overwrite \
```

```
--species homo_sapiens \
```

```
--input_file LDS_4050.vcf.gz \
```

```
--vcf \
```

```

--plugin
UTRAnnotator,/nobackup/ummamd/VEP_annotation/uORF_5UTR_PUBLIC.txt \
--plugin
CADD,/nobackup/ummamd/VEP_annotation/whole_genome_SNVs.tsv.gz,/nobackup/ummamd/VEP_annotation/gnomad.genomes-exomes.r4.0.indel.tsv.gz \
--plugin
SpliceAI,indel=/nobackup/ummamd/VEP_annotation/spliceai_scores.masked.indel.hg19.vcf.gz,snv=/nobackup/ummamd/VEP_annotation/spliceai_scores.masked.snv.hg19.vcf.gz \
--dir_plugins /nobackup/ummamd/.vep/Plugins \
--fasta /nobackup/ummamd/ummamd/VEP_annotation/reference \
--stats_file LDS_4050.hg19.cadd.spliceai.UTRAnnotator.summary.html \
--output_file LDS_4050.hg19.cadd.spliceai.UTRAnnotator.vcf

```

## 8.2.6 Splitting VCF by chromosome and subset by FH gene

### 1. Install BCFtools v1.21 using Conda package repository

```
conda install bioconda::bcftools
```

### 2. Subset the VCF file by chromosome

```
bcftools view -r(Chromosome_number) Input_file.vcf.gz -Oz -o output_file.vcf.gz
```

### 3. Unzip the compressed VCF files

```
Gunzip *.vcf.gz
```

### 4. Subset VCF file by gene of interest

```
grep "Gene_symbol " Input_file.vcf> Output_file.vcf
```

## 8.2.7 Scripts and bioinformatic commands for ONT sequencing in plasmid assembly

### 1. Guppy script v5

```
#!/bin/bash

# Use current environment variables and modules
#$ -V

# Set current working directory
#$ -cwd

# Request 3 hours of runtime
#$ -l h_rt=2:00:00

# Request Nvidia node
#$ -l coproc_v100=1

# Email at the beginning and end of the job
#$ -m be

module load singularity
module load cuda/10.1.168

# Run the following command
singularity run --nv --bind /nobackup:/nobackup /nobackup/containers/guppy-
gpy-5.0.16.simg guppy_basecaller -x "cuda:0" -i
/nobackup/ummamd/5plasmids/fast5 -s /nobackup/ummamd/5plasmids/fastq --
flow cell FLO-MIN106 --kit SQK-LSK110
```

### 2. Installing Miniconda 3

```
wget https://repo.anaconda.com/miniconda/Miniconda3-latest-Linux-x86_64.sh
-O ~/miniconda3/miniconda.sh
bash Miniconda3-py39_4.12.0-Linux-x86_64.sh
```

```
conda create --name ont_tools
conda activate ont_tools
```

### 3. Performing base calling using Guppy v5:

```
mkdir project_name
mkdir fast5 fastq
cp -r /nobackup/mderar/20220630_run149/plasmid_assembly/fast5 .
nano guppy_v5.2.sh
qsub guppy_v5.2.sh
```

### 4. Concatenating FASTQ files:

```
cd /nobackup/mderar/project_name/guppy_output
cat fail/* > ./fail.fastq
cat pass/* > ./pass.fastq
gzip pass.fastq
gzip fail.fastq
cat *fastq* > filename.fastq.gz
gunzip filename.fastq.gz
```

### 5. Adaptor trimming by Porechop

```
porechop --discard_middle -i filename.fastq -o filename.adapt.trim.fastq >>
filenameadapt.trim.log.std.out >> filename.adapt.trim.log.std.err
```

### 6. Obtain read statistics on the trimmed FASTQ file using NanoStat

```
NanoStat --fastq filename.adapt.trim.fastq>
filename.adapt.trim.fastq.unfiltered.NanoStat
```

### 7. Filter reads based on quality and length using NanoFilt

```
NanoFilt filename.adapt.trim.fastq -q 10 --headcrop 75 --length 750 --maxlength
9000 | gzip > filename.concat.adapt.trim.filt.fastq.gz
```



**8. Obtain read statistics on the filtered FASTQ file using NanoStat**

```
NanoStat --fastq filename.concat.adapt.trim.filt.fastq.gz >
filename.concat.adapt.trim.filt.fastq.filtered.NanoStat
```

**9. Alignment of reads to reference sequence using Minimap2**

```
minimap2 -ax map-ont /nobackup/mderar/project_name/references/reference.fa
filename.concat.adapt.trim.filt.fastq.gz>> filename.minimap.std.err >>
filename.concat.sam
```

**10. Sort and Index BAM file using Samtools**

```
samtools view -bh filename.concat.sam > filename.concat.bam
samtools sort filename.concat.bam -o filename.concat.sorted.bam
samtools index filename.concat.sorted.bam
```

**11. Removing Soft clipped reads**

```
samtools sort filename_removedsoftclipping.bam -o
filename_removedsoftclipping.sorted.bam
samtools index filename_removedsoftclipping.sorted.bam
```

**12. Extract reads using a plasmid's unique sequence**

```
Grep -A1 -B1 "unique_sequence" filename.concat.adapt.trim.filt.fastq >
filename.fa
```

**13. Install Canu**

```
conda install -c conda-forge -c bioconda -c defaults canu
```

**14. Perform *de novo* assembly**

```
canu -p File_suffix -d Folder_name genomeSize=8.5k -nanopore -trimmed
filename.fa useGrid=false contigFilter="2 0 1.0 0.5 0" corOutCoverage=100
minReadLength=7000 minOverlapLength=100 2> filename.log
```

### 8.3 Appendix C: Figures and Tables

> NC\_000016.10:g.84015931\_84027116inv aberrant mrna: exon 7-9 inverted

GCAAATTCCTGGGACCCATGGCCCTGACCACTTGTGCTGTCCTCTTAGGTGTGGAGGGGCC  
 TGACCCTGACACATACTCGCTAGGGCTCCTTTAGTTAATGACTCACTCACCCCTGGCCAAGCT  
 CGAATCTTGCTTCTCCACCCCTCTGGCCCCAAGAAATGCAACTGAGCCACCTGGAGTAAGC  
 CCGAGTCCCATCCCCAACAGCCAAAGACTGTCCGGCACCTTAAAAGTTGGATTTGGTGCCAT  
 GGAGGGACAGACCCAGGAAGCAGGGGCCTTCCAGAAAAGCCTCACCCCTGCCACGGCTGC  
 TGCCACTCTGTCTCGATGGGCGCTGTCTTCATCCTCATGAAGTCCGCGCTGGGAGCTGGC  
 CTGCTCAACTTCCCCTGGGCCTTCTCAAAGCGGGCGGAGTGGTCCCTGCCTTCTGGTGG  
 AGCTGTCTCGTTGGTCTTCTGATCAGCGGGCTGGTCATCCTGGGCTATGCTGCTGCTGC  
 AGTGGCCAGGCCACCTACCAGGGTGTGGTCAGGGGGCTGTGTGGCCCTGCCATTGGGAAG  
 CTGTGTGAGGCCTGCTTCTCCTCAACCTGCTCATGATCTCCGTGGCCTTCTCAGGGTGT  
 CGGGGACCAGCTGGAGAAGCTGTGTGACTCCCTCCTGTCTGGCACCCCGCCCGCCCGCA  
 GCCGTGGTACGCAGACCAGCGCTTACCCTGCCCTGCTCTCCGTGCTGGTTCATCCTGCC  
 CTGTCTGCCCGCGGGAGATCGCCTTCCAGAAATACACAAGCATCCTAGGCACTCTGGCTGC  
 CTGTTACCTGGCCCTGGTCATCACCGTGCAGTACTACCTCTGGCCCCAGGGCCTCGTGCCT  
 GAGTCCCATCCTTCACTGAGCCCTGCCTCCTGGACCTCTGTGTTCACTGTCTTCCCACCAT  
 CTGCTTCGGGTTTCAGGTTTGTGCCTCATCTGTGCAATGGGTGTGAGCCTATAGGACCAAG  
 AGTCAA GTGCTGCCTGGAGGTCTGGGAGTGGTCTCTGTGCTGGTCGGCACCTTCATCTTTG  
 GGCAGAGCACGGCGGCAGCGGTCTGGGAGATGTTCTGATGGGCAGCTAGTGCCGGGCAGG  
 AAGGGGCCCTCCGGGGGCTGACCCTACGTGGCTGCTGTATGCAGCCAGGAGACCGATGCC  
 ATTTCTTTTCTCATAAAGATGCTGGAGAGACTGA

> NC\_000016.10:g.84034763\_84037497del aberrant mRNA: exon 3 deleted

GCAAATTCCTGGGACCCATGGCCCTGACCACTTGTGCTGTCCTCTTAGGTGTGGAGGGGCCCTGACCCTGA  
 CACATACTCGCTAGGGCTCCTTTAGTTAATGACTCACTCACCCCTGGCCAAGCTCGAATCTTGCTTCTCCA  
 CCCCTCTGGCCCCAAGAAATGCAACTGAGCCACCTGGAGTAAGCCCGAGTCCCATCCCCAACAGCCAAA  
 GACTGTCCGGCACCTTAAAAGTTGGATTTGGTGCCATGGAGGGACAGACCCAGGAAGCAGGGGGCCTTCC  
 AGAAAAGCCTCACCCCTGCCACGGCTGCTGCCACTCTGTCTCGATGGGCGCTGTCTTCATCCTCATGAAGT  
 CCGCGCTGGGAGCTGGCCTGCTCAACTTCCCCTGGGCCTTCTCAAAGCGGGCGGAGTGGTCCCTGCC  
 TTCTGGTGGAGCTGTGTGACTCCCTCCTGTCTGGCACCCCGCCCGCCCGCAGCCGTGGTACGCAG  
 ACCAGCGCTTACCCTGCCCTGCTCTCCGTGCTGGTTCATCCTGCCCTGTCTGCCCGCGGGAGATCG  
 CTTCCAGAAATACACAAGCATCCTAGGCACTCTGGCTGCCTGTACCTGGCCCTGGTTCATCACCGTGCAG  
 TACTACCTCTGGCCCCAGGGCCTCGTGCCTGAGTCCCATCCTTCACTGAGCCCTGCCTCCTGGACCTCTG  
 TGTTCACTGTCTTCCCACCATCTGCTTCGGGTTTCAGTGTACGAAGCTGCCGTCTCCATCTACTGCAGCAT  
 GCGCAAACGGAGCCTCTCCACTGGGCCCTGGTGTCTGTGCTGTCTTGGCTGGCCTGCTGCCTCATCTATT  
 CACTGACGGGGGTTTATGGCTTCTGACTTTTGGGACAGAAGTTTCTGCTGACGTCTTGATGTCCTACCCAG  
 GCAATGATATGGTCATCATTGTGGCCCGGGTCTTTTGTGCTCTCCATCGTAACTGTCTACCCCATCGTGT  
 CTTCCTGGGGAGTGTGAGTATGCAGGACTTCTGGAGGAGGAGCTGCTTGGGGGGATGGGGGCCAGCG  
 CCCTGGCCGACCCCTCAGGGCTGTGGGTCCGGATGCCGCTGACCATCCTGTGGGTACCCTGACGCTC  
 GCCATGGCGCTGTTTATGCCTGACCTCAGCGAGATCGTCAGCATCATCGGAGGCATCAGTTCCTTCTCATC  
 TTCATCTTCCCAGGTTTGTGCCTCATCTGTGCAATGGGTGTGAGCCTATAGGACCAAGAGTCAA GTGCTGC  
 CTGGAGGTCTGGGAGTGGTCTCTGTGCTGGTCGGCACCTTCATCTTGGGCAGAGCACGGCGGCAGCG  
 GTCTGGGAGATGTTCTGATGGGCAGCTAGTGCCGGGCAGGAAGGGGCCCTCCGGGGGCTGACCCTACGT  
 GGCTGCTGTATGCAGCCAGGAGACCGATGCCATTTCTTTTCTCATAAAGATGCTGGAGAGACTGA

> NC\_000016.10:g.84006453\_84019002del aberrant mRNA: exon 8-11 deleted

GCAAATTCTTGGGACCCATGGCCCTGACCACTTGTTGCTGTCCTCTTAGGTGTGGAGGGGCC  
 TGACCCTGACACATACTCGCTAGGGCTCCTTTAGTTAATGACTCACTCACCCCTGGCCAAGCT  
 CGAATCTTGCTTCTCCCACCCCTCTGGCCCCAAGAAATGCAACTGAGCCACCTGGAGTAAGC  
 CCGAGTCCCATCCCCAACAGCCAAAGACTGTCCGGCACCTTAAAAGTTGGATTTGGTGCCAT  
 GGAGGGACAGACCCAGGAAGCAGGGGCCTTCCAGAAAAGCCTCACCCCTGCCACGGCTGC  
 TGCCACTCTGTCCTCGATGGGCGCTGTCTTCATCCTCATGAAGTCCGCGCTGGGAGCTGGC  
 CTGCTCAACTTCCCCTGGGCCTTCTCAAAGCGGGCGGAGTGGTCCCTGCCTTCTGGTGG  
 AGCTG GTCTCGTTGGTCTTCTGATCAGCGGGCTGGTCATCCTGGGCTATGCTGCTGCTGTC  
 AGTGGCCAGGCCACCTACCAGGGTGTGGTCAGGGGGCTGTGTGGCCCTGCCATTGGGAAG  
 CTGTGTGAGGCCTGCTTCTCCTCAACCTGCTCATGATCTCCGTGGCCTTCTCAGGGTAT  
 CGGGGACCAGCTGGAGAAGC TGTGTGACTCCCTCCTGTCTGGCACCCCGCCCGCCCGCA  
 GCCGTGGTACGCAGACCAGCGCTTACCCTGCCCTGCTCTCCGTGCTGGTATCCTGCC  
 CTGTCTGCCCGCGGGAGATCGCCTTCCAGAAATACACAAG CATCCTAGGCACTCTGGCTGC  
 CTGTTACCTGGCCCTGGTCATCACCGTGCAGTACTACCTCTGGCCCCAGGGCCTCGTGCCT  
 GAGTCCCATCCTTCACTGAGCCCTGCCTCCTGGACCTCTGTGTTTCAAGTGTCTTCCCACCAT  
 CTGCTTCGGGTTTCAGTGTACGAAGCTGCCGTCTCCATCTACTGCAGCATGCGCAAACGGA  
 GCCTCTCCCACTGGGCCCTGGTGTCTGTGCTGTCTTGTGGCCTGCTGCCTCATCTATTCA  
 CTGACGGTCAGTTCCCTAGCGGCAAATGGAGCAGGGTTACTCTGTACAAAGCAGGTCTC  
 TTGTGAGGATTAAGACGATGATGGATCCAGAGTAGCTGATGGGATACGTCACGCACAGT  
 GGAGACCCAAGCAATGTAGCTTTATTCCCAGAAGACCAAGGTCGTGATTGGCAAGGGGTA  
 CCTGAGTTTCTGGGAGCCAAGGGGA

Exon 1 Exon 2 exon 3 exon 4 exon 5 exon 6 exon 7 exon 8 Exon 9 exon 10  
 exon 11 Intergenic sequence

**Supplementary Figure 3.1. Aberrant mRNA sequences of three SVs identified in the local cohort.** Each exon is colour coded to annotate the MANE mRNA transcript NM\_001080442.3. Exon 7-9 that is inverted by NC\_000016.10:g.84015931\_84027116inv has been removed from the mRNA transcript since these nucleotides are in the reverse orientation. The mRNA transcript of the NC\_000016.10:g.84006453\_84019002del has exon 8-11 deleted and has intergenic sequences that are downstream the deletion being potentially transcribed.

Ethnicity	Phenotype	Genotype	Variants	gnomAD	100KGP	Exomiser	Tier	CADD	Grantham	SIFT	PolyPhen-2	varSEAK	SpliceAI
South Asian	optic atrophy and visual impairment	Compound heterozygous	c.190-8C>T	786/90120 (7)	1180/126690		0	0.407	-	-	-	Class 1	0
			c.1027G>C; p.(Gly343Arg)	1350/91086 (30)	320/126698	0.148	0	15.06	125	Tolerated	Possibly damaging	-	-
South Asian	Reduced visual acuity	Homozygous	c.100A>G; p.(Met34Val)	26/91072	11/126698	0.0331	2	9.931	21	Tolerated	Benign	-	-
European	Visual impairment, vestibular dysfunction and hearing impairment	Compound heterozygous	c.617C>G; p.(Ser206Cys)	245/1180008	21/126698		3	20.3	112	Deleterious	Benign	-	-
			c.273G>A; p.Val91=	17744/1180008 (138)	1702/126698	0.0288	0	6.812	-	-	-	Class 1	0
African	Visual impairment	Compound heterozygous	c.961A>T; p.(Met321Leu)	863/75018 (10)	78/126698		0	14.13	15	Tolerated	Benign	-	-
			c.960G>A; p.Val320=	837/75016 (10)	59/126698	0.0109	0	5.203	-	-	-	Class 1	AL (0.01) DG (0.01)
European	Abnormal involuntarily eye movements	Compound heterozygous	c.190-8C>T	11648/1177714 (72)	1180/126690		0	0.407	-	-	-	Class 1	0
			c.803C>T; p.(Thr268Met)	88/1179602	10/126698	0.00270	3	33	81	Deleterious	Probably damaging	-	-
European	Abnormal saccadic eye movements, ataxia, dysarthria and hyperreflexia	Compound heterozygous	c.961A>T; p.(Met321Leu)	9/1179652	78/126698		0	14.13	15	Tolerated	Benign	-	-
			c.960G>A; p.Val320=	9/1179638	59/126698	0.000566	0	5.203	-	-	-	Class 1	AL (0.01) DG (0.01)
African	Progressive visual loss, constriction of peripheral visual field, and retinal dystrophy	Compound heterozygous	c.1305C>G; p.(Phe435Leu)	530/75028 (1)	60/126698		0	3.811	22	Tolerated	Benign	-	-
			c.189G>A; p.Leu63=	750/74154 (5)	58/126698	0.124	0	22	-	-	-	Class 5	DG (0.28) DL (0.01)
African	Progressive visual loss, visual impairment, reduced visual acuity and retinal dystrophy	Compound heterozygous	c.189G>A; p.Leu63=	750/74154 (5)	58/126698		2	22	-	-	-	Class 5	DG (0.28) DL (0.01)
			c.628C>G; p.(Leu210Val)	758/75068 (7)	51/126698	0.766	2	4.698	32	Tolerated	Benign	-	-
South Asian	Progressive visual loss, central scotoma, visual impairment, Macular dystrophy and reduced visual acuity	Compound heterozygous	c.794A>G; p.(Tyr265Cys)	156/91060 (1)	23/126698		2	31	194	Deleterious	Probably damaging	-	-
			c.805+6G>A	180/91028 (1)	30/126698	0.968	2	4.451	-	-	-	Class 1	DG (0.01)
European	Visual impairment and microphthalmia	Homozygous	c.159C>T; p.Gly53=	222/63680 (1)	300/126698	0.031	0	0.063	-	-	-	Class 1	0
European	Gaze evoked nystagmus and abnormality of eye movement	Compound heterozygous	c.1074G>A; p.Ala235=	5199/1180032 (9)	427/126698		0	0.019	-	-	-	Class 1	0
			c.273G>A; p.Val91=	17744/1180008 (138)	1702/126698	0.0024	0	6.812	-	-	-	Class 1	0
European	Visual impairment, reduced visual acuity and hypoplasia of the fovea	Compound heterozygous	c.1074G>A; p.Ala235=	5199/1180032 (9)	427/126698		0	0.019	-	-	-	Class 1	0
			c.273G>A; p.Val91=	17744/1180008 (138)	1702/126698	0.0064	0	6.182	-	-	-	Class 1	0
European	Ophthalmoparesis and abnormal saccadic eye movements	Compound heterozygous	c.159C>T; p.Gly53=	222/63680 (1)	300/126698		0	0.063	-	-	-	Class 1	0
			c.273G>A; p.Val91=	17744/1180008 (138)	1702/126698	0.00041	0	6.812	-	-	-	Class 1	0
South Asian	Abnormality of saccadic eye movements	Compound heterozygous	c.273G>A; p.Val91=	857/91078 (16)	1702/126698		0	6.812	-	-	-	Class 1	0
			c.705C>T; p.Thr358=	220/89662 (1)	45/126698	0.0011	0	0.041	-	-	-	Class 1	AG (0.04) AL (0.02)
European	Visual loss	Compound heterozygous	c.922_923insTG p.(Thr308Metfs*14)	0	-		3	-	-	-	-	-	-
			c.922A>G; p.(Thr308Ala)	689/1179964 (1)	61/126698	0.0465	3	28	58	Deleterious	Probably damaging	-	-
South Asian	Non-progressive visual loss, abnormal macular morphology, central scotoma and visual impairment	Homozygous	c.264C>G; p.(Tyr88*)	30/91074	16/126698	0.912	1	55	-	-	-	-	-
South Asian	Progressive visual loss, central scotoma, visual impairment, reduced visual acuity and retinal dystrophy	Homozygous	c.264C>G; p.(Tyr88*)	30/91074	16/126698	0.882	1	55	-	-	-	-	-
South Asian	Congenital nystagmus	Homozygous	c.264C>G; p.(Tyr88*)	30/91074	16/126698	0.840	1	55	-	-	-	-	-
European	Visual impairment and nystagmus	Compound heterozygous	c.435G>A; p.(Trp145*)	1/349594	3/126698		3	48	-	-	-	-	-
			c.632+1G>A	2/1111810	2/126698	0.997	3	32	-	-	-	Class 5	DL (0.95)

**Supplementary Table 3.1 Entire SLC38A8 biallelic cohort in the 100KGP.** 19 probands with biallelic SLC38A8 variants characterised based on *In silico* pathogenicity tools, allele frequencies and GEL annotations.

Ethnicity	Phenotype	Genotype	gnomAD	100KGP	Variant score	Gene-Pheno	CADD	Grantham	SIFT	PolyPhen-2	Align-GVGD	varSEAK	SpliceAI
European	Nystagmus	c.723C>A; p.(Ser241Arg)	10/1110398	2/126698	0.9999	0.1710	23.4	110	Deleterious	Probably damaging	Class C65	-	-
European	Nystagmus and gaze evoked nystagmus	c.389-2A>G	4/1103970	1/126698	0.9996	0.1889	33	-	-	-	-	Class 5	AG: 0.43 AL: 0.97
European	Progressive visual loss	c.776T>C; p.(Leu259Pro)	16/1180028	1/16638	0.9982	0.2567	28.9	98	Deleterious	Probably damaging	Class C65	-	-
African	Abnormality of saccadic eye movement	c.895C>T; p.(Arg299Trp)	3/74974	1/126698	0.9973	0.2075	26.1	101	Deleterious	Probably damaging	Class C65	-	-
European	Macular dystrophy and retinal dystrophy	c.127C>G; p.(Leu43Val)	113/1179774	16/126698	0.9963	0.3487	27.7	32	Deleterious	Probably damaging	Class C25	-	-
European	Progressive visual loss, optic neuropathy and visual impairment	c.127C>G; p.(Leu43Val)	113/1179774	16/126698	0.9963	0.2926	27.7	32	Deleterious	Probably damaging	Class C25	-	-
European	Iris coloboma and visual impairment	c.127C>G; p.(Leu43Val)	113/1179774	16/126698	0.9963	0.2517	27.7	32	Deleterious	Probably damaging	Class C25	-	-
East Asian	Rod-cone dystrophy	c.131A>G; p.(Asn44Ser)	2/39666	1/126698	0.9963	0.2930	22.7	46	Tolerated	Possibly damaging	Class C45	-	-
South Asian	Visual impairment	c.787C>G; p.(Leu263Val)	4/91092	2/126698	0.9948	0.3011	23.2	32	Tolerated	Possibly damaging	Class C25	-	-
European	Rod-cone dystrophy, visual impairment and sensorineural hearing impairment	c.1289C>T; p.(Ala430Val)	39/1613746	4/126698	0.9945	0.2743	24.9	64	Deleterious	Possibly damaging	Class C55	-	-
European	Progressive visual loss, macular degeneration, central scotoma and visual impairment	c.792C>G; p.(Ile264Met)	24/1179996	1/126698	0.9918	0.3173	26.1	10	Deleterious	Probably damaging	Class C0	-	-
African	Rod-cone dystrophy and visual impairment	c.109G>A; p.(Ala37Thr)	0/75034	7/126698	0.9906	0.2904	20.7	58	Deleterious	Benign	Class C55	-	-
African	Progressive visual loss	c.487C>T; p.(Pro163Ser)	47/75052 (1)	17/126698	0.9904	0.2904	24.8	74	Deleterious	Probably damaging	Class C65	-	-
European	Progressive visual loss, visual impairment, macular dystrophy and reduced visual acuity	c.806-3C>G	5/1180022	1/126698	0.798	0.381	23.8	-	-	-	-	Class 5	AL: 0.89
African	progressive visual loss, retinal dystrophy and visual impairment	c.487C>T; p.(Pro163Ser)	47/75052 (1)	17/126698	0.9904	0.2904	24.8	74	Deleterious	Probably damaging	Class C65	-	-
European	Visual impairment	c.1205C>T; p.(Pro402Leu)	780/1179970 (2)	71/126698	0.9624	0.3474	23.8	98	Tolerated	Possibly damaging	Class C0	-	-
African	Visual impairment	c.534C>G; p.(Ile178Met)	0/33476	4/126698	0.9432	0.2673	24.8	10	Deleterious	Probably damaging	Class C0	-	-
European	Gaze evoked horizontal nystagmus	c.617C>G; p.(Ser206Cys)	245/1180008	21/126698	0.9321	0.2003	20.3	112	Tolerated	Benign	Class C65	-	-
European	Vestibular dysfunction, cone-rod dystrophy and visual impairment	c.617C>G; p.(Ser206Cys)	245/1180008	21/126698	0.9321	0.2641	20.3	112	Tolerated	Benign	Class C65	-	-
European	Progressive visual loss, constriction of peripheral visual field, retinal dystrophy and visual impairment	c.1063G>T; p.(Val355Phe)	5/1112000	1/126698	0.8800	0.2668	20	50	Tolerated	Possibly damaging	Class C45	-	-
South Asian	Retinal dystrophy	c.878T>C; p.(Met293Thr)	48/91088 (2)	5/126698	0.8255	0.3368	0.291	81	Tolerated	Benign	Class C65	-	-
European	Cataract, sensorineural hearing impairment, visual impairment, and abnormal retinal morphology	c.421G>A; p.(Ala141Thr)	17/1178548	1/126698	0.7013	0.3047	4.617	58	Tolerated	Benign	Class C55	-	-
European	Nystagmus	c.88G>A; p.(Val30Ile)	48/1180030	6/126698	0.5572	0.2210	7.411	29	Tolerated	Benign	Class C25	-	-
European	Blurred vision and abnormal saccadic eye movements	c.1205C>T; p.(Pro402Leu)	780/1179970 (2)	71/126698	0.4693	0.1460	23.8	98	Tolerated	Possibly damaging	Class C65	-	-
European	Abnormality of colour vision and visual impairment	c.1205C>T; p.(Pro402Leu)	780/1179970 (2)	71/126698	0.4693	0.3538	23.8	98	Tolerated	Possibly damaging	Class C65	-	-
African	Progressive visual loss, central scotoma, visual impairment and macular dystrophy.	c.1046C>T; p.(Pro349Leu)	41/74942	19/126698	0.3847	0.0030	3.576	98	Tolerated	Benign	Class C65	-	-
Other	Nystagmus and abnormality of eye movement	c.179T>C; p.(Leu60Pro)	15/1592252	6/126698	0.2977	0.2175	24.3	98	Deleterious	Probably damaging	Class C65	-	-
European	Visual impairment and sensorineural hearing impairment	c.88G>A; p.(Val30Ile)	48/1180030	6/126698	0.2113	0.0528	7.411	29	Tolerated	Benign	Class C25	-	-
European	Abnormal saccadic eye movements, nystagmus and dysmetric saccades	c.596C>T; p.(Pro199Leu)	16/1180040	2/126698	0.2049	0.3092	23.7	98	Deleterious	Benign	Class C65	-	-

**Supplementary Table 3.2 Entire SLC38A8 monoallelic cohort in the 100KGP. 29 probands with monoallelic SLC38A8 variants underwent pathogenicity assessment.**

ID	Variants	Ensemble VEP	CADD	Splice AI	VarSEAK	ENCODE	UTRannotator	SIFT	PolyPhen-2
I	<b>c.723C&gt;A;</b> <b>p.(Ser241Arg)</b>	Moderate Impact	23.4	-	-	-	-	Deleterious	Probably damaging
	c.805+54A>G	Modifier	2.092	0	Class 1	none	none	-	-
	c.1162+941dupG	Modifier	1.146	0	Class 1	none	none	-	-
	c.691-494G>T	Modifier	1.410	AG: 0.02	Class 1	none	none	-	-
	c.389-473G>T	Modifier	1.71	0	Class 1	none	none	-	-
	c.190-911_190-907dupTTTTT	Modifier	1.207	0	Class 1	none	none	-	-
	c.190-1931C>T	Modifier	0.64	0	Class 1	pELS (EH38E3194821)	none	-	-
c.1162+940_1162+941dupGG	Modifier	-	-	-	none	none	-	-	
II	<b>c.389-2A&gt;G</b>	High Impact	33	AG: 0.43 AL: 0.97	Class 5	-	-	-	-
	c.189+2501C>G	Modifier	0.766	0	Class 2	none	none	-	-
	c.691-1832C>T	Modifier	0.079	0	Class 1	none	none	-	-
	c.530+34delG	Modifier	0.068	0	Class 1	none	none	-	-
III	<b>c.776T&gt;C;</b> <b>p.(Leu259Pro)</b>	Moderate Impact	28.9	-	-	-	-	Deleterious	Probably damaging
	c.1214+433C>T	Modifier	2.963	AG: 0.01	Class 1	none	none	-	-
	c.691-451C>G	Modifier	0.683	0	Class 1	none	none	-	-
IV	<b>c.895C&gt;T;</b> <b>p.(Arg299Trp)</b>	Moderate Impact	26.1	-	-	-	-	Deleterious	Probably damaging
	c.1215-944G>A	Modifier	0.025	0	Class 1	none	none	-	-
	c.1162+750T>G	Modifier	0.524	0	Class 1	none	none	-	-
	c.806-937delT	Modifier	0.279	0	Class 1	none	none	-	-
	c.6991-1926C>T	Modifier	0.668	0	Class 1	none	none	-	-
	c.690+3269G>T	Modifier	2.708	0	Class 1	none	none	-	-
V	<b>c.1289C&gt;T;</b> <b>p.(Ala430Val)</b>	Moderate Impact	24.9	-	-	-	-	Deleterious	Possibly damaging
	c.691-605C>G	Modifier	1.503	0	Class 1	none	none	-	-
	c.388+1445C>T	Modifier	3.02	0	Class 2	none	none	-	-
	c.190-242C>A	Modifier	0.942	0	Class 2	none	none	-	-
VI	<b>c.792C&gt;G;</b> <b>p.(Ile264Met)</b>	Moderate Impact	26.1	-	-	-	-	Deleterious	Probably damaging
	c.691-2585C>T	Modifier	3.279	0	Class 1	dELS (EH38E3194808)	none	-	-
VII	<b>c.109G&gt;A;</b> <b>p.(Ala37Thr)</b>	Moderate Impact	20.7	-	-	-	none	Deleterious	Benign
	c.690+2877C>T	Modifier	1.106	0	Class 1	none	none	-	-
	c.691-52C>T	Modifier	0.687	AG: 0.2	Class 1	none	none	-	-
VIII	<b>c.487C&gt;T;</b> <b>p.(Pro163Ser)</b>	Moderate Impact	24.8	-	-	-	-	Deleterious	Probably damaging
	c.1163-615C>T	Modifier	1.24	0	Class 1	none	none	-	-
	c.389-224G>C	Modifier	0.816	DG: 0.01	Class 1	none	none	-	-
	c.633-1156G>T	Modifier	1.909	DG: 0.03	Class 1	none	none	-	-
	c.691-363T>C	Modifier	1.563	DG: 0.01	Class 1	none	none	-	-
	c.806-2725C>T	Modifier	1.223	DG: 0.01	Class 1	none	none	-	-
	c.691-2425T>C	Modifier	1.254	0	Class 1	dELS (EH38E3194808)	none	-	-
IX	<b>c.487C&gt;T;</b> <b>p.(Pro163Ser)</b>	Moderate Impact	24.8	-	-	-	-	Deleterious	Probably damaging
	c.189+1028delG	Modifier	0.267	0	Class 1	pLS (EH38E3194823)	none	-	-
	c.690+2355A>G	Modifier	0.813	0	Class 1	none	none	-	-
	c.691-2425T>C	Modifier	1.254	0	Class 1	dELS (EH38E3194808)	none	-	-
	c.691-363T>C	Modifier	1.563	DG: 0.01	Class 1	none	none	-	-
	c.633-1156G>T	Modifier	1.909	DG: 0.03	Class 1	none	none	-	-
	c.389-224G>C	Modifier	0.816	DG: 0.01	Class 1	none	none	-	-
X	<b>c.179T&gt;C;</b> <b>p.(Leu60Pro)</b>	Moderate Impact	24.3	-	-	-	-	Deleterious	Probably damaging
	c.691-1926C>T	Modifier	0.668	0	Class 1	none	none	-	-
XI	<b>c.596C&gt;T; p.(Pro199Leu)</b>	Moderate Impact	23.7	-	-	-	none	-	-
	c.530+291C>G	Modifier	0.098	0	Class 1	none	none	-	-
	c.388+502C>T	Modifier	10.07	0	Class 1	none	none	-	-
XII	<b>c.803C&gt;T;</b> <b>p.(Thr268Met)</b>	Moderate Impact	33	-	-	-	-	Deleterious	Probably damaging
	c.1163-273G>C	Modifier	0.756	0	Class 1	none	none	-	-
XIII	<b>c.189G&gt;A;</b> <b>p.(Leu63=)</b>	Low Impact	22	DG: 0.28 DL: 0.01	Class 5	-	-	-	-
	c.471C>T; p.(Ser157=)	Low Impact	0.179	0	Class 1	none	none	-	-
	c.440C>G; p.(Ala147Gly)	Moderate Impact	13.15	0	-	none	none	Deleterious	Benign
	c.690+863G>C	Modifier	0.369	0	Class 1	none	none	-	-
	c.806-369A>G	Modifier	0.045	0	Class 1	none	none	-	-
	c.805+2428G>A	Modifier	2.927	0	Class 1	none	none	-	-
	c.691-363T>C	Modifier	1.563	DG: 0.01	Class 1	none	none	-	-

	<b>c.189G&gt;A; p.(Leu63=)</b>	Low Impact	22	DG: 0.28 DL: 0.01	Class 5	-	-	-	-
	c.743C>G; p.(Ser248Cys)	Moderate Impact	22.7	0	-	none	none	Deleterious	Possibly damaging
	c.690+665G>A	Modifier	1.735	0	Class 1	dELS (EH38E3194812)	none	-	-
XIV	c.531-554A>G	Modifier	2.116	0	Class 1	DNase only (EH38E3194819)	none	-	-
	c.690+683A>G	Modifier	1.526	0	Class 1	dELS (EH38E3194812)	none	-	-
	c.690+3207A>C	Modifier	2.296	0	Class 1	none	none	-	-
	c.690+804G>A	Modifier	0.209	0	Class 1	dELS (EH38E3194812)	none	-	-
	<b>c.794A&gt;G; p.(Tyr265Cys)</b>	Moderate Impact	31	-	-	-	-	Deleterious	Probably damaging
	c.1214+1397C>T	Modifier	0.216	0	Class 1	none	none	-	-
	c.1163-383_1163-382insC	Modifier	1.443	0	Class 1	none	none	-	-
XV	c.1163-1605_1163-1601dupCCTCC	Modifier	1.221	0	Class 1	none	none	-	-
	c.691-794G>A	Modifier	0.918	0	Class 1	none	none	-	-
	c.690+2129T>G	Modifier	1.048	0	Class 1	none	none	-	-
	c.189+335C>T	Modifier	0.785	AG: 0.01	Class 1	none	none	-	-
XVI	<b>c.922A&gt;G; p.(Thr308Ala)</b>	Moderate Impact	28	-	-	-	-	Deleterious	Probably damaging
	c.632+24G>A	Modifier	0.024	DG: 0.01	Class 1	DNase-H3K4me3 (EH38E3194817)	none	-	-
XVII	c.806-3C>G	Low Impact	23.8	AL: 0.89	Class 5	none	none	-	-
	c.806-443C>T	modifier	3.016	0	Class 1	none	none	-	-
XVIII	<b>c.127C&gt;G; p.(Leu43Val)</b>	-	Moderate Impact	27.7	-	-	-	-	Deleterious
	n/a	-	-	-	-	-	-	-	-
XIX	<b>c.127C&gt;G; p.(Leu43Val)</b>	-	Moderate Impact	27.7	-	-	-	-	Deleterious

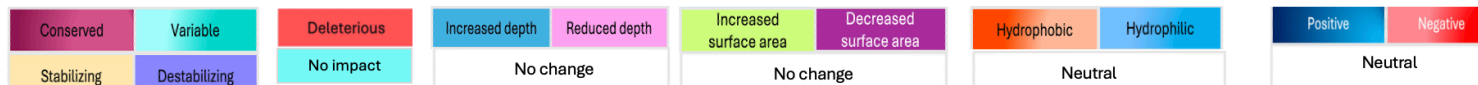
**Supplementary Table 3.3. SLC38A8 monoallelic cohort in the 100KGP analysed for exonic and intronic hits.** Genomic findings in 19 probands with a single pathogenic allele are displayed. Intronic variants analysed for splicing defects using CADD, varSEAK and SpliceAI and are assessed whether they overlap CREs or 5' UTR.

Cohort	Variant	Variant (p.annotation)	Re: ACMG	ACMG published	Reference
Literature	NM_001080442.3:c.923C>G	NP_001073911.1:p.(Thr308Ser)	VUS	Pathogenic	(Lasseaux et al., 2018)
Literature	NM_001080442.3:c.1078_1104del	NP_001073911.1:p.(Ala360_leu368del)	VUS	Likely pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.1126G>A	NP_001073911.1:p.(Gly376Arg)	VUS	Likely pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.692G>A	NP_001073911.1:p.(Cys231Tyr)	VUS	Likely pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.855G>C	NP_001073911.1:p.(Leu285Phe)	VUS	Likely pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.922A>G	NP_001073911.1:p.(Thr308Ala)	VUS	Pathogenic	(Lasseaux et al., 2018)
Literature	NM_001080442.3:c.527C>G	NP_001073911.1:p.(Thr176Arg)	VUS	Pathogenic	(Lasseaux et al., 2018)
Literature	NM_001080442.3:c.845_847delCTG	NP_001073911.1:p.(Ala282del)	VUS	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.1214+5G>C	NP_001073911.1:p.?	Likely pathogenic	VUS	(Kuht et al., 2020)
Literature	NM_001080442.3:c.558C>A	NP_001073911.1:p.(Tyr186*)	Likely pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.632+2T>G	NP_001073911.1:p.?	Likely pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	Exon 1 deletion	NP_001073911.1:p.?	Likely pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.6_9delGGGA	NP_001073911.1:p.(Glu2Aspfs*32)	Likely pathogenic	Pathogenic	(Lasseaux et al., 2018)
Literature	Exon 7 deletion	NP_001073911.1:p.?	Likely pathogenic	Pathogenic	(Lasseaux et al., 2018)
Literature	Exon 7 -8 deletion	NP_001073911.1:p.?	Likely pathogenic	Pathogenic	(Lasseaux et al., 2018)
Literature	NM_001080442.3:c.697G>A	NP_001073911.1:p.(Glu233Lys)	Likely pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.95T>G	NP_001073911.1:p.(Ile32Ser)	Likely pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.707T>A	NP_001073911.1:p.(Val236Asp)	Likely pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.1234G>A	NP_001073911.1:p.(Gly412Arg)	Likely pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.101T>G	NP_001073911.1:p.(Met34Arg)	VUS	Likely pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.644G>T ^	NP_001073911.1:p.(Trp215Leu) ^	VUS	VUS	(Kuht et al., 2020)
Literature	NM_001080442.3:c.682G>A ^	NP_001073911.1:p.(Gly228Arg) ^	VUS	VUS	(Kuht et al., 2020)
Literature	NM_001080442.3:c.695A>G	NP_001073911.1:p.(His232Arg)	VUS	VUS	(Kuht et al., 2020)
Literature	NM_001080442.3:c.954-1G>C	NP_001073911.1:p.?	Pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.964C>T	NP_001073911.1:p.(Gln322*)	Pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.995dupG	NP_001073911.1:p.(Trp333Metfs*35)	Pathogenic	Pathogenic	(Kuht et al., 2020)
Literature	NM_001080442.3:c.598C>T	NP_001073911.1:p.(Gln200*)	Pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.490_491delCT	NP_001073911.1:p.(Leu164Valfs*41)	Pathogenic	Pathogenic	(Schiff et al., 2021)
Literature	NM_001080442.3:c.1002delG	NP_001073911.1:p.(Ser336Alafs*15)	Pathogenic	Pathogenic	(Schiff et al., 2021)
Literature and local cohort	NM_001080442.3:c.848A>C	NP_001073911.1:p.(Asp283Ala)	Likely pathogenic	Likely pathogenic	(Schiff et al., 2021)

**Supplementary Table 3.4. Concordance in ACMG classification of SLC38A8 variants.** 30 variants implicated in FVH2 are listed and the corresponding reference for their ACMG classification is provided. Re: ACMG denotes the reclassification of variants using Franklin Geenox webserver (<https://franklin.genox.com/clinical-db/home>).



Missense	Conservation	Protein Stability (kcal/mol)	Structural Impact	Wildtype Residue Exposure (Å)	Mutant Residue Exposure (Å)	Wildtype Solvent Accessibility (%)	Mutant Solvent Accessibility (%)	Wildtype Hydrophobicity	Mutant Hydrophobicity	Wildtype Electrostatics potential kcal/(mol*e)	Mutant Electrostatics potential kcal/(mol*e)
p.(Ile32Ser)	9	-2.313	Cavity expansion	5.3	4.6	12.5	17.6	4.50	-0.80	7.8	13.3
p.(Met34Arg)	8	0.077		5	5	9.5	19.4	1.90	-4.50	1.4	9.24
p.(Thr87Ile)	8	0.041		3.6	3.8	60.3	49.7	-0.70	4.50	5.3	7.92
p.(Gly90Val)	7	1.549	Buried residue replaced	4.9	3.8	18.3	25	-0.40	4.20	5.8	8.30
p.(Thr176Arg)	8	1.040	Cavity expansion	2.5	4.5	18.4	21.6	-0.70	-4.50	3.2	8.85
p.(Ile178Met)	4	-0.152		3.4	3.3	63.4	61.2	4.50	1.90	3.4	5.26
p.(Trp215Leu)	4	-0.052		2.8	3.2	90.6	93.1	-0.90	3.80	0.27	1.58
p.(Cys226Arg)	7	-0.200	Buried residue replaced	10.2	9.9	0	0	2.50	-4.50	0.09	2.19
p.(Gly228Arg)	6	1.517	Buried residue replaced	12.4	148.1	1.7	0.6	-0.40	-4.50	6.05	24.1
p.(Cys231Tyr)	9	-0.376		10.2	10.4	1.1	0	2.50	-1.30	14.5	34.4
p.(His232Arg)	9	-0.955		11.2	10.9	0	0	-3.20	-4.50	-	38.3
p.(Glu233Gly)	9	0.701	Cavity expansion	8.1	6.6	10.9	27.3	-3.50	-0.40	12.5	18.6
p.(Glu233Lys)	9	0.126		8.1	5.1	10.9	18.7	-3.50	-3.90	12.5	13.3
p.(Val236Asp)	8	-1.917	Buried residue replaced	7.3	8.5	0	0	4.20	-3.50	15	23.2
p.(Leu267Arg)	8	-0.223		3.6	2.6	54.3	63.7	3.80	-4.50	1.06	26.9
p.(Asp283Ala)	9	0.500	Buried residue replaced	4.7	5.9	2.6	2.2	-3.50	1.80	1.32	-2.1
p.(Leu285Phe)	9	-0.614		10.6	-1391	0.8	1.4	3.80	2.80	7.02	27.1
p.(Thr308Ala)	9	-0.044		4.5	3.9	21.3	17.1	-0.70	1.80	3.31	8.2
p.(Thr308Ser)	9	-0.499		4.5	4.2	21.3	17.6	-0.70	-0.80	3.31	2.72
p.(Gly376Arg)	9	0.104	Buried residue replaced	5.9	908.7	0	0	-0.40	-4.50	-	10.5
p.(Gly412Arg)	7	-0.165	Buried residue replaced	6.1	6.3	0	0	-0.40	-4.50	1.09	20.5
p.(Gly419Val)	9	0.848	Buried residue replaced	6.5	6.8	0	0	-0.40	4.20	-	-
p.(Thr51Ile)	3	0.231		3.23	3.23	104.8	105	0.70	4.50	1.84	1.74
p.(Pro57Leu)	5	1.990		3.9	3.8	23	34.7	-1.60	3.80	1.36	0.83
p.(Ala147Gly)	4	-0.636		5.6	5.6	1.4	3.2	1.80	-0.40	0.136	-1.8
p.(Leu210Val)	2	-0.188		3.33	3.25	79.6	81.9	3.80	4.20	0.13	1.04
p.(Ser220Thr)	8	0.746		4.0	4.0	30.6	24.7	-0.80	-0.70	-2.11	2.2
p.(Ser248Cys)	6	0.060		3.2	3.1	94.4	94.8	-0.80	2.50	2.72	2.9
p.(Val270Ile)	5	0.265		3.64	3.95	24.8	28.6	4.20	4.50	1.35	1.35
p.(Met321Leu)	5	0.046		3.9	3.9	39.7	38.9	1.90	3.80	0.52	2.9
p.(Gly343Arg)	8	0.519	Residue at bend curvature replaced	4.3	3.5	68.1	55.9	-0.40	-4.50	4.34	4.82
p.(Thr427Met)	8	0.438		3.5	3.5	61.4	66.6	-0.70	1.9	-1.4	0.36



**Supplementary Figure 3.2. Variant simulation of SLC38A8 missense variants in FH.** The raw data for each parameter analysed is displayed. Protein stability values are for delta G ( $\Delta G$ ). Residue exposure units are in angstrom ( $\text{\AA}$ ) which implies  $10^{-10}$  metre. Hydrophobicity values are based on Kyte-Doolittle scale. Electrostatic potentials values are based on Adaptive Poisson-Boltzmann Solver (APBS) values.

1	ur NF_001073911.1 SLC38A8 Homo Sapiens	PASWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRKRSLSHWA	VSVLSLLACC
2	ur F6YH19 243_676 1.4e-250 Equus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
3	ur AOA2K5FWU4 1_405 1.9e-247 Cebus imitator	PASWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRKRSLSHWA	VSVLSLLGCC
4	ur AOA7J8CB17 1_431 5.5e-244 Melospiza molossus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
5	ur UPI0004D08FFC 1_431 1.2e-241 Galeoscoptes variegatus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
6	ur AOA8C6QL98 1_431 6.4e-239 Nannosorex galili	PASWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
7	ur AOA28Z6ZS8 1_434 1.6e-237 Artiodactyla	PASWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
8	ur GLTG21 1_434 2.6e-236 Oryzctolagus cuniculus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLGCC
9	ur AOA2Y9K2E0 17_451 5.1e-235 Caniformia	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
10	ur AOA6I9ID97 4_428 1.8e-234 Camelidae	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLGCC
11	ur Q5HZH7 1_431 1.5e-232 Muroidea	PSPWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
12	ur AOA2Y9I9X7 1_432 7.3e-232 Phocidae	PASWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
13	ur G3SXF2 1_430 3.6e-229 Elephantidae	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLGCC
14	ur AOA2Y9DN80 1_430 3.6e-227 Trichechus manatus latirostris	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
15	ur AOA1S3A2N8 1_434 3.6e-226 Erinaceus europaeus	P-SWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLGCC
16	ur AOA6G1B919 1_439 7e-222 Hyaeonidae	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
17	ur AOA835Z599 71_483 4.8e-220 Bovidae	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
18	ur AOAA40LEZ8 46_442 5.1e-217 Eptesicus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
19	ur UPI0003P0C232 187_613 5.7e-210 Elephantulus edwardii	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
20	ur UPI001E1D98E9 25_450 2.1e-204 Echinops telfairi	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
21	ur UPI001CC3EC8B 176_599 6.5e-196 Dromiciops gliroides	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
22	ur UPI0021104759 11_427 5.1e-187 Suncus etruscus	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
23	ur UPI00292CC69F 29_454 1.3e-178 Gekkota	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
24	ur AOA1U7RRL7 21_436 3.9e-178 Crocodylia	ASSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
25	ur AOAA35PDN6 1067_1492 6.2e-168 Hyloidea	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
26	ur AOA6P9CAV4 28_451 5.1e-173 Pantherophis guttatus	ASSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
27	ur UPI001C4A7820 31_445 5.4e-172 Sceloporus undulatus	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
28	ur AOA6J0I1E1 20_436 2.6e-170 Passeriformes	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
29	ur UPI00235A79B2 1_318 2.7e-169 Artibeus jamaicensis	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
30	ur UPI001ABEZF1B 10_436 7.2e-168 Hyloidea	ATSLASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
31	ur UPI001292D27D 101_530 1.8e-167 Lonchura striata domestica	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
32	ur AOA7L0BUC8 47_463 6.2e-167 Neognathae	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
33	ur AOA7L3RMI8 48_474 1.4e-165 Neognathae	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
34	ur AOA7K5HSM2 28_451 3.5e-165 Crocophaga sulcirostris	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
35	ur AOA852BM70 26_442 3.5e-164 Pitiformes	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
36	ur AOA7K6WV63 19_446 1.9e-163 Steatornis caripensis	VSSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
37	ur AOA8C5MR59 14_426 2.8e-163 Leptobranchium leishanense	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
38	ur AOA7K7SFW3 57_476 2e-162 Passeriformes	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
39	ur AOA9D3MK72 10_438 2.6e-162 Elapomorphia	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
40	ur AOA7K8PPL8 45_470 1.6e-161 Neognathae	ASSWASVF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
41	ur AOA8J4KYN3 46_473 3.7e-161 Spheniscidae	PSSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
42	ur AOA7L4HB75 52_470 2.7e-160 Podargus strigoides	ASSWASVF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
43	ur AOA852KFL7 17_430 2e-159 Urocolia indicus	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
44	ur UPI0011CEDEC 12_401 3.8e-159 Strigops habroptila	ASSWASVF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
45	ur AOA7K8SIK8 54_469 2.7e-158 Neognathae	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
46	ur AOA7L2C7Q6 43_466 1.4e-157 Alaudala cheleensis	PSSWTSVF	-----	SVFF	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
47	ur AOA3P8NQ34 23_438 4.6e-157 Pseudocrenilabrinae	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
48	ur UPI001CF0C0690 19_439 1.1e-156 Protopterus annectens	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
49	ur AOA7L0E284 1_417 2.5e-156 Trogon melanurus	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
50	ur AOA3P8V830 23_434 4.7e-156 Cynoglossus semilaevis	-SWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
51	ur AOA484D1G1 49_467 1.3e-155 Pycniidae	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
52	ur AOA7K5A101 13_426 7.2e-155 Centropus	VSSWASVF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
53	ur AOA3B3RE50 11_434 1.9e-154 Mormyridae	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
54	ur AOA6J2UTJ6 10_437 3.7e-154 Chanos_chanos	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
55	ur AOA060XHM9 23_439 1.7e-153 Protacanthopterygii	NDSCASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
56	ur AOA8C9RRB7 2_428 4.2e-153 Sclerophages formosus	SGSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
57	ur AOA437DB13 61_482 6.3e-153 Oryzias	KNSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
58	ur AOA6P7M3M7 23_438 2.3e-152 Beta splendens	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
59	ur AOA5N5MX14 14_437 3.7e-152 Siluroidei	LDASWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
60	ur UPI001C04C307 20_441 5.2e-152 Melanotaenia boesemani	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
61	ur U3IBV2 2_415 2.6e-151 Anas	ASSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
62	ur AOA9D3NRE2 14_437 1.1e-150 Siluroidei	LDASWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
63	ur UPI000FFD61FB 8_376 2.3e-150 Empidonax traillii	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
64	ur H3CV87 23_438 4.9e-150 Tetraodon nigroviridis	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
65	ur AOA3B4E689 10_429 9.5e-150 Serrasalimidae	ISTVWSMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
66	ur AOA3Q3N375 23_436 3.9e-149 Mastacembelus armatus	VSLWTSMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
67	ur AOA9Q0DYM6 23_443 9.6e-149 Murraenolepis oranienensis	ISTWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
68	ur AOA6P8TD11 50_466 3.7e-148 Notothenioidei	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
69	ur UPI001F081FBE 10_434 1.9e-147 Hypomesus transpacificus	VSSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
70	ur AOA6P7MQT2 21_443 4.3e-147 Beta splendens	IGSWASTF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
71	ur UPI0024C37C2C 50_420 8e-147 Hylla sarda	ASSLASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
72	ur AOA8C5TX84 26_424 1.5e-146 Malurus cyaneus samueli	CWGWGLSP	-----	ALSP	-	L	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
73	ur AOA674IIX0 7_385 4.9e-146 Trappena carolina tringuis	PADKRCALF	-----	SRV	-	-	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
74	ur AOA3P8W666 23_443 9.5e-146 Cynoglossus semilaevis	ISTVSSMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
75	ur AOA2F0AUA7 19_345 1.5e-145 Eschrichtius robustus	ATKKVPLT	-----	VVFL	-	R	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
76	ur AOA3B4FRA8 20_441 2e-145 Pseudocrenilabrinae	MNSWAAMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
77	ur AOA8T0B2N7 5_433 6.2e-145 Silurus meridionalis	IGFWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
78	ur AOA663DPR6 16_405 1.6e-144 Aquila chrysaetos chrysaetos	CGKSLGVGVVPAWGP	-----	STIP	-	P	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
79	ur AOA8C5EU07 22_410 2.8e-144 Blenniiformes	IGSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
80	ur AOA670YGY4 10_384 1.2e-143 Pseudonaja textilis	ARD	-----	-----	-	VPCFV	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
81	ur AOA673AAM9 18_438 1.9e-142 Sphaeramia orbicularis	AGSWASTF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
82	ur UPI0025AE890F 17_412 1.5e-142 Svonnathinae	VSSWASTF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
83	ur AOA3B4V744 31_433 5.8e-141 Seriola dumerili	-SWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
84	ur AOA850URQ4 4_365 2.5e-140 Chloropsis hardwickii	-----	-----	STIP	-	PQHPPLAL	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
85	ur UPI00234C901E 13_431 6.6e-140 Clarias gariepinus	IGFWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
86	ur AOA665UQQ1 18_441 9.7e-140 Echeneis naucrates	VGSWASMF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
87	ur AOA7K7FPG0 2_380 9.9e-139 Neognathae	ASSWASIF	-----	SVIP	-	T	-	ICFGF	-----	QCHEAAVSIYCSMRNRQLSHWA	VSVLSLLACC
88	ur UPI0015CF8C57 10_436 4.4e-138 Electrophorus electricus	ISSWTPEVF	-----	SVIP	-						



Ethnicity	MOI	Genotype	Phenotypes	Families	Source
European (British)	Compound heterozygous	c.534C>G, p.(Ile178Met)	Foveal hypoplasia, nystagmus, fundus hypopigmentation, Posterior embryotoxon and chiasmal misrouting	1	(Campbell et al., 2019)
		Deletion of exon 2-5		4	(Weiner et al., 2020)
Middle Eastern (Karaite Jewish)	Homozygous	c.95T>G, p.(Ile32Ser)	Foveal hypoplasia, nystagmus, myopia, hypermetropia and astigmatism. 6 individuals reported with strabismus. Developmental delay in a single family	3	(Perez et al., 2014)
South Asian (Indian Jewish)			Infantile nystagmus, foveal hypoplasia and astigmatism. One had posterior embryotoxon. Another with infantile hypotonia and generalized joint hypermobility	2	(Ehrenberg et al., 2021)
			Foveal hypoplasia and Nystagmus	1	(Kruijt et al., 2022)
South Asian (Indian Jewish)	Compound heterozygous	c.490_491delCT, p.(Leu164Valfs*41) c.95T>G, p.(Ile32Ser)	Foveal hypoplasia and congenital nystagmus	1	(Weiner et al., 2020)
European (Dutch)*	Compound heterozygous	c.1234G>A, p.(Gly412Arg)	Foveal hypoplasia, nystagmus, chiasmal misrouting, posterior embryotoxon and Kartagener syndrome	1	(Poulter et al., 2013)
		Large deletion		1	(Kruijt et al., 2022)
South Asian (Pakistani)	Homozygous	c.1029delG, p.(Leu344Cysfs*7)	Foveal hypoplasia	1	(Poulter et al., 2013)
Mediterranean (Turkish)	Homozygous	c.101T>G, p.(Met34Arg)	Foveal hypoplasia and chiasmal misrouting	1	(Poulter et al., 2013)
			Foveal hypoplasia, nystagmus, strabismus and chiasmal misrouting	1	(Kuht et al., 2020)
European (Dutch)*	Compound heterozygous	c.598C>T, p.(Gln200*) c.845_847delCTG, p.(Ala282del)	Foveal hypoplasia, nystagmus and chiasmal misrouting. One individual has posterior embryotoxon	1	(Poulter et al., 2013)
				1	(Kruijt et al., 2022)
South Asian (Afghan)*	Homozygous	c.1002delG, p.(Ser336Alafs*15)	Foveal hypoplasia, nystagmus, chiasmal misrouting and posterior embryotoxon	1	(Poulter et al., 2013)
South Asian (Pakistani)	Homozygous	c.707T>A, p.(Val236Asp)	Foveal Hypoplasia, chiasmal misrouting, posterior embryotoxon and <del>Anterior segment dysgenesis</del>	1	(Poulter et al., 2013)
			Foveal hypoplasia, nystagmus, astigmatism and moderate hypermetropia	1	(Toral et al., 2017)
Ashkenazi Jewish	Homozygous	c.848A>C, p.(Asp283Ala)	Infantile nystagmus, foveal hypoplasia and strabismus .	5	(Ehrenberg et al., 2021)
			Foveal hypoplasia, nystagmus, chiasmal misrouting, strabismus and iris transillumination	1	(Schiff et al., 2021)
			Foveal hypoplasia and nystagmus	2	(Kruijt et al., 2022)
			Foveal hypoplasia, nystagmus and anterior segment dysgenesis	1	(This thesis)
East Asian (Korean)	Compound heterozygous	c.692G>A, p.(Cys231Tyr)	Foveal hypoplasia, nystagmus, strabismus and chiasmal misrouting	1	(Kuht et al., 2020)
East Asian (Korean)	Compound heterozygous	c.964C>T, p.(Gln322*)	Foveal hypoplasia, nystagmus and chiasmal misrouting	1	(Kuht et al., 2020)
		c.558C>A, p.(Tyr186*)		1	(Kuht et al., 2020)
East Asian (Korean)	Compound heterozygous	c.1078_1104del, p.(Ala360_Leu368del)	Foveal hypoplasia, concentric macular rings, nystagmus and chiasmal misrouting	1	(Kuht et al., 2020)
East Asian (Korean)	Compound heterozygous	c.855G>C, p.(Leu285Phe) c.995dupG, p.(Trp333Metfs*35)	Foveal hypoplasia, concentric macular rings, nystagmus and chiasmal misrouting	1	(Kuht et al., 2020)
East Asian (Korean)	Compound heterozygous	C.954-1G>C c.995dupG, p.(Trp333Metfs*35)	Foveal hypoplasia, concentric macular rings, nystagmus, strabismus and chiasmal misrouting	1	(Kuht et al., 2020)
European	Homozygous	c.632+2T>G	Foveal hypoplasia, nystagmus, strabismus, chiasmal misrouting and iris transillumination	1	(Kuht et al., 2020)
European (British)	Compound heterozygous	Exon 1 deletion	Foveal hypoplasia, nystagmus, strabismus, chiasmal misrouting and iris transillumination	1	(Kuht et al., 2020)
		c.1126G>A, p.(Gly376Arg)		1	(Kuht et al., 2020)
East Asian (Korean)	Compound heterozygous	c.1214+5G>C c.995dupG, p.(Trp333Metfs*35)	Foveal hypoplasia, nystagmus and chiasmal misrouting	1	(Kuht et al., 2020)

East Asian (Korean)	Compound heterozygote	c.644G>T p.(Trp215Leu) ^	Foveal hypoplasia and nystagmus	1	(Kuht et al., 2020)
		c.682G>A p.(Gly228Arg) ^			
European	Compound heterozygous	c.922A>G, p.(Thr308Ala)	Foveal hypoplasia, nystagmus and iris transillumination	1	(Lasseaux et al., 2018)
		Exon 7 deletion			
-	Compound heterozygous	c.527C>G, p.(Thr176Arg)	Foveal hypoplasia, nystagmus, myopia, hypermetropia and astigmatism	1	(Lasseaux et al., 2018)
		c.848A>C, p.(Asp283Ala)			
European (French)*	Compound heterozygous	c.697G>A, p.(Glu233Lys)	Foveal hypoplasia, nystagmus, astigmatism and hypermetropia	1	(Lasseaux et al., 2018)
		c.(805 + 1_806-1)_ (1162 + 1_1163-1)del; p.(?)			
European	Compound heterozygous	c.6_9delGGGA, p.(Glu2Aspfs*32)	Foveal hypoplasia, nystagmus, strabismus, hypermetropia and Iris transillumination	1	(Lasseaux et al., 2018)
		c.923C>G, p.(Thr308Ser)			
Ashkenazi Jewish	Compound heterozygous	c.848A>C, p.(Asp283Ala)	Infantile nystagmus and foveal hypoplasia	1	(Ehrenberg et al., 2021)
		whole gene deletion			
European (German, Italian and Russian)	Compound Heterozygous	c.848A>C, p.(Asp283Ala)	Infantile nystagmus, foveal hypoplasia and strabismus	1	(Ehrenberg et al., 2021)
		c.676T>C, p.(Cys226Arg)			
Ashkenazi Jewish	Compound heterozygous	c.848A>C, p.(Asp283Ala)	Infantile nystagmus, foveal hypoplasia and strabismus (second proband in an affected family)	1	(Ehrenberg et al., 2021)
		c.269G>T, p.(Gly90Val)			
Mixed (French, Irish, Canadian and Puerto Rican)	Compound heterozygous	c.160G>T, p.(Gly54*)	Infantile nystagmus, foveal hypoplasia and macrocephaly	1	(Ehrenberg et al., 2021)
		c.388+5G>A			
Mixed (Turkish-Algerian and Iranian)	Homozygous	c.697G>A, p.(Glu233Lys)	Infantile nystagmus, foveal hypoplasia, strabismus, hypospadias, cognitive and motor delay, high muscle tone and patent foramen ovale	1	(Ehrenberg et al., 2021)
			Foveal hypoplasia, bilateral microphthalmia, retinochoroidal coloboma and strabismus	1	(Poulter et al., 2013)
South Asians (Indian, Pakistani and Bangladeshi)	Homozygous	c.264C>G, p.(Tyr88*)	Foveal hypoplasia, nystagmus and strabismus. one proband with posterior embryotoxon and chiasmal misrouting. 2 probands with concentric macular rings	3	(Schiff et al., 2021)
			Foveal hypoplasia, low vision, nystagmus and strabismus	1	(Chaudhuri et al., 2018)
European (British)	Compound heterozygous	c.435G>A, p.(Trp145*)	Foveal hypoplasia, nystagmus, strabismus and chiasmal misrouting	1	(Schiff et al., 2021)
		c.632+1G>A			
South Asian (Sri Lankan)	Homozygous	c.698A>G, p.(Glu233Gly)	Foveal hypoplasia, nystagmus, strabismus, posterior embryotoxon, concentric macular rings, bilateral peripheral iris adhesion to cornea bilateral blue dot cataract and chiasmal misrouting	1	(Schiff et al., 2021)
European (Spanish and British)	Compound heterozygous	c.923C>G, p.(Thr308Ser)	Foveal hypoplasia, nystagmus, strabismus, bilateral shallow anterior chamber, concentric macular rings and chiasmal misrouting	1	(Schiff et al., 2021)
		Deletion of exon 7-8			
European (Dutch)	Compound heterozygous	c.260C>T, p.(Thr87Ile)	Foveal hypoplasia, nystagmus and chiasmal misrouting	1	(Kruijt et al., 2022)
		c.800T>G, p.(Leu267Arg)			
European (Dutch)	Homozygous	c.598C>T, p(Gln200*)	Foveal hypoplasia, nystagmus and chiasmal misrouting	1	(Kruijt et al., 2022)
European (Swedish, Italian, Irish and British)	Compound heterozygous	c.1256G>T, p.(Gly419val)	Foveal hypoplasia and nystagmus	1	(Kruijt et al., 2022)
		c.(805 + 1_806-1)_ (1162 + 1_1163-1)del; p.(?)			

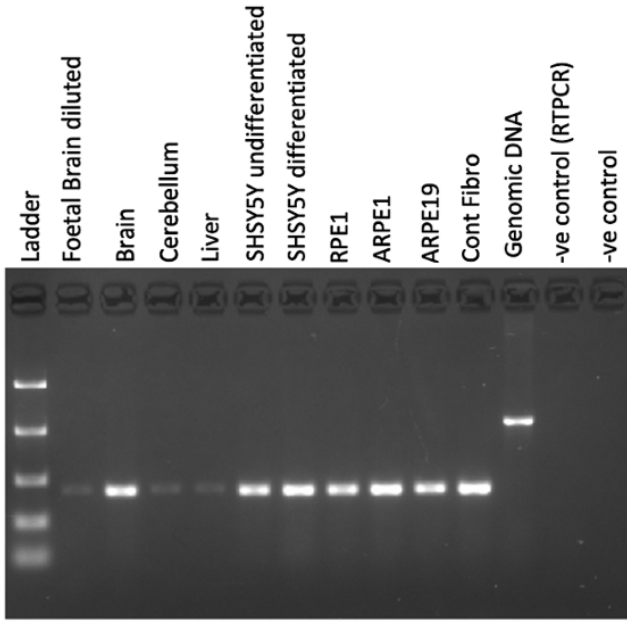
East Asian (Japanese)	Homozygous	c.995dupG (p.Trp333Metfs*35)	Foveal hypoplasia, congenital nystagmus, posterior embryotoxon, hypermetropia and goniodysgenesis.	1	(Hayashi et al., 2021)
Middle eastern	Homozygous	g.84015973_84027074inv (GRCh38)	Foveal hypoplasia and nystagmus	1	(This thesis)
-	Compound heterozygous	g.84034763_84037497del (GRCh38)	Foveal hypoplasia	1	(This thesis)
-		g.84006453_84019002del (GRCh38)			(This thesis)
-	Homozygous	c.669delC p.(Thr224Profs*44)	Foveal hypoplasia, nystagmus and myopia	1	(This thesis)

**Supplementary Table 3.5. Phenotypes of probands with FVH2 in 63 families.**

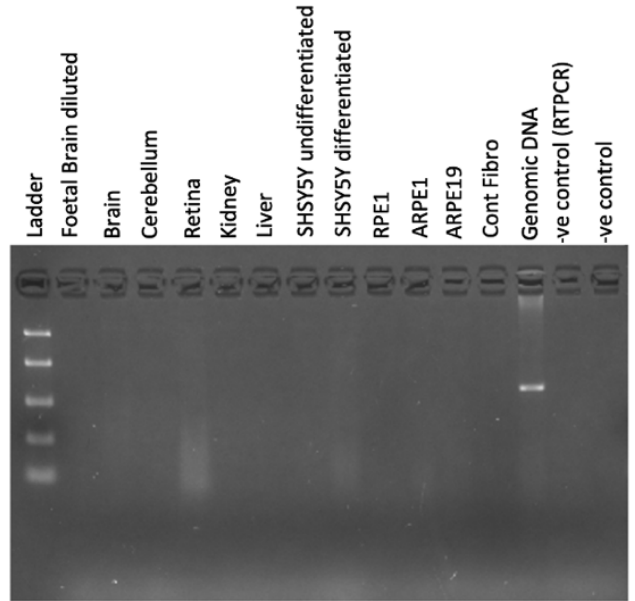
The clinical manifestation of each proband is tabulated along with the ethnicity, genotype and reference. \* denotes the same proband being reported in different studies.

Guide RNA	Off target recognition site	PAM	Genomic Location	Annotations
sgRNA1	GAGGAAGACAGGGCCCATCC	AGG	12:12878797	intergenic
sgRNA1	GATGAAGCCAGTGCCCATGG	AGG	1:5002451	intergenic
sgRNA1	GAGAAAGACAGCGCCCATGG	GAG	4:131658767	intergenic
sgRNA1	GAGGAAGACAGCCCCATCT	GAG	8:10546054	intron: PRSS55
sgRNA1	GAGGAAGGCAGCTCCCATCG	CAG	19:22627422	intergenic
sgRNA1	GATGAAGAAAGGGCCCATCT	GAG	8:136370085	intergenic
sgRNA1	GATGAAGACAGAGCCCACCA	CAG	5:148678541	intergenic
sgRNA1	GATGAAGACAGGGCCCTTCT	CAG	6:125717865	exon: LINC02523
sgRNA1	GATGAAGACAGTGCCAGCC	TAG	13:24467197	intron: PARP4
sgRNA1	GATGAAGCCAGCGCCACCC	AAG	5:171603609	intergenic
sgRNA1	GATGAAGCCGGCGCCCATCC	CAG	15:40070412	intergenic
sgRNA1	GCTGAAGACAGTGCCCATCG	CAG	2:127090263	intron: BIN1
sgRNA2	TCCTTGGTCTTCCTAAGCAG	GGG	1:168098646	intron: GPR161
sgRNA2	TCGATTGTCTTTCTGATCAG	GGG	14:22168701	intron: TRAV30
sgRNA2	TCGTTGGCCTTGGTGATCAG	AGG	13:28462759	intron: FLT1
sgRNA2	TTTTTGGTCTTCCTAATCAG	GGG	17:68581006	intron: FAM20A
sgRNA2	TCGTTGGTGTTTCATGCTCAG	TGA	18:24192211	intron: OSBPL1A
sgRNA2	TCTTTGGTCTACCTGCTCAG	AGA	12:53783171	intergenic
sgRNA2	TGGTTAGTCTTCCTTATCAG	GGA	10:70782252	intron: TBATA
sgRNA2	TTGTTGGACTTCCTGATCGG	TGA	2:109634238	intergenic
sgRNA2	TAGTTGGTTTTCTGATAAG	AAG	5:158773694	intron: EBF1
sgRNA2	TCATTGCTCTTCCTGGTCAG	GAG	10:9228305	intron: LOC101928272
sgRNA2	TCGTGGGTCTTCCTGAATAG	AAG	7:71275897	intron: GALNT17
sgRNA2	TCGTTGGTCTTCAAGATGAG	CAG	14:77931601	exon: ADCK1
sgRNA2	TGGTTGGTTTTCTGATCTG	CAG	8:134390506	intergenic
sgRNA2	TGGTTTATCTTCCTGATCAG	GAG	7:45349273	intergenic
sgRNA3	AACCTCATGAAGTAAGCGCT	GGG	6:154933482	intergenic
sgRNA3	ATCCTCATGGAGCCCACGCT	AGG	20:46993646	intron: EYA2
sgRNA3	ATCCTCATGAAGTCCTGCA	GGA	2:26026885	intergenic

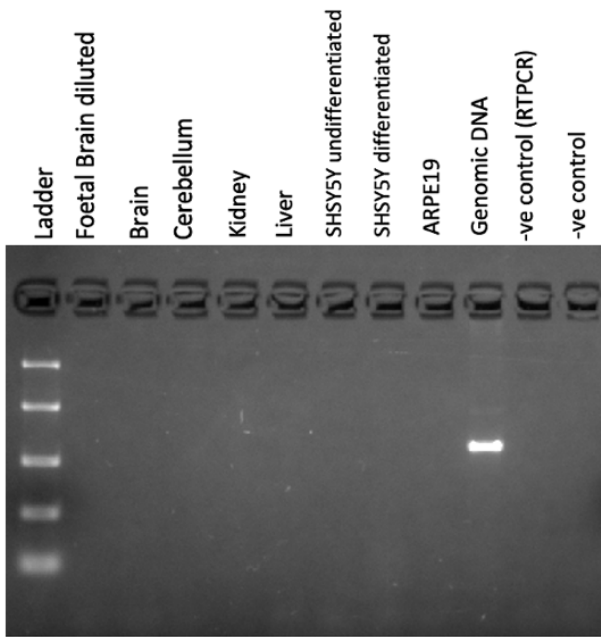
**Supplementary Table 3.6. Predicted off target effects of *SLC38A8* guide RNAs.** sgRNA1, sgRNA2 and sgRNA 3 have 12, 14 and 3 potentially unwanted genomic edits, respectively.



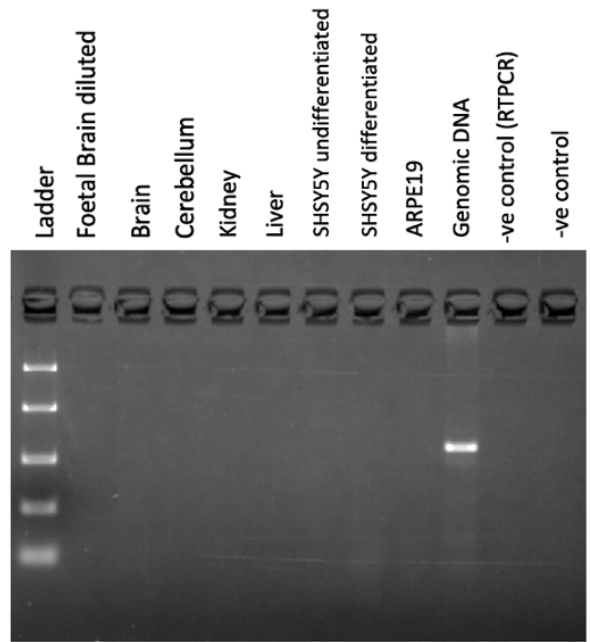
**TP53**



**SLC38A8 (exon 7 & 8)**



**SLC38A8 (exon 7 & 8)**



**SLC38A8 (UTR primers)**

**Supplementary Figure 3.4. Further testing for *SLC38A8* expression. No expression of *SLC38A8* detected in all cell lines and tissues tested using three different primer sets.**



Sequence (5'->3')	Length	Tm (°C)	GC (%)	Self complementarity	Self 3' complementarity	Product size
TTCATAAATTGAATATTTCCGCC	23	53.8	30.43	6	1	
GAGAAACAATGCCCTGTCAAA	21	56.99	42.86	4	1	600 bp

**Supplementary Table 4.1. *HPS5* primer design.** A single primer set flanking exon 18 of *HPS5*.

Primer	Sequence	Length (bp)	GC (%)	Tm (°C)	Self-complementarity	Self 3' complementarity	Wildtype product	Mutant product
<i>PAX6</i> _1F	TCTTGCTTGTTTTCCGCCTC	20	50	59.1	2	0	26719 bp	2607 bp
<i>PAX6</i> _2R	TGCATGACTTTGTAAGGATGTGG	23	43.5	59.2	4	0		
<i>GPR143</i> _1F	CCACCGAGTAGAAAGAGCCT	20	55	58.8	3	3	598265 bp	1640 bp
<i>GPR143</i> _1F	GATTACATGCGCCCAGGAAT	20	50	58.4	6	6		
<i>GPR143</i> _2F	TGTCAGCAGGTCCAATGTCT	20	50	58.9	3	0	598961 bp	2336 bp
<i>GPR143</i> _2F	GCTAGGTGAAGTGGATGGGT	20	55	59	4	0		

**Supplementary Table 4.2 *PAX6* and *GPR143* primers used in LR-PCR.** A single primer set used in the amplification of *PAX6* and 2 primer sets for *GPR143*. The expected product size for the wildtype allele and for the allele harbouring the relevant deletion is displayed.

CHROM	COORDINATE	dbSNP	REF	ALT	COV	FILTER	CADD_PHRED	DS_AG	DS_AL	DS_DG	DS_DL	GENE	TRANSCRIPT_ID
16	84036349	rs3880	G	A	45712	PASS	4.349	0	0	0	0	NECAB2	ENST00000305202
16	84036357	rs8045237	G	A	10257.9	PASS	0.218	0	0	0	0	NECAB2	ENST00000305202
16	84036596	rs10552328	AAGGT	A	31161.5	PASS	1.452	0	0	0	0	NECAB2	ENST00000305202
16	84036608	rs10552329	TGACA	T	31763.9	PASS	1.255	0	0	0	0	NECAB2	ENST00000305202
16	84036654	rs11348799	T	T	24875.6	PASS	2.101	0	0	0	0	NECAB2	ENST00000305202
16	84036681	rs8045339	A	G	22478.5	PASS	0.59	0	0	0	0	NECAB2	ENST00000305202
16	84037062	rs4343273	G	C	26246	PASS	0.433	0	0	0	0	NECAB2	ENST00000305202
16	84037115	rs4644852	G	C	24696.7	PASS	0.848	0	0	0	0	NECAB2	ENST00000305202
16	84037124	rs4347611	T	A	30480	PASS	1.334	0	0	0	0	NECAB2	ENST00000305202
16	84037242	rs4447415	C	G	21419.7	PASS	0.163	0	0	0	0	NECAB2	ENST00000305202
16	84037293	rs4274423	G	C	15745	PASS	0.047	0	0	0	0	NECAB2	ENST00000305202
16	84037315	rs5818478	C	AAAAA,CAAAA	18861.5	PASS	0.995,0.968,0.943,0.917,0.893,1.106	0	0	0	0	NECAB2	ENST00000305202
16	84037461	rs6076950	A	C	23493.5	PASS	3.092	0	0	0	0	NECAB2	ENST00000305202
16	84037496	rs60320303	G	A	23444	PASS	0.345	0	0	0.02	0	NECAB2	ENST00000305202
16	84037541	rs59177635	T	G	9856.89	PASS	2.294	0	0	0	0	NECAB2	ENST00000305202
16	84037860	rs9944299	A	G	23892.5	PASS	1.178	0	0	0	0	NECAB2	ENST00000305202
16	84038178	rs8047263	T	C	23941.7	PASS	1.58	0	0	0	0	NECAB2	ENST00000305202
16	84038368	rs8047640	T	A	26152	PASS	0.553	0.01	0	0	0	NECAB2	ENST00000305202
16	84038509	rs8047833	T	C	37885.5	PASS	0.01	0	0	0	0	SLC38A8	ENST00000299709
16	84038518	rs6563989	G	A	44101.1	PASS	0.388	0	0	0	0	SLC38A8	ENST00000299709
16	84038703	rs7187556	A	G	31852.6	PASS	0.041	0	0	0	0	SLC38A8	ENST00000299709
16	84038749	rs1876963	A	G	31130.6	PASS	0.198	0	0	0	0	SLC38A8	ENST00000299709
16	84039014	rs66972156	C	G	11914.9	PASS	3.783	0	0	0	0	SLC38A8	ENST00000299709
16	84039901	rs9889066	G	A	20658.6	PASS	1.281	0	0	0	0	SLC38A8	ENST00000299709
16	84040043	rs9889063	C	T	23548.9	PASS	0.551	0	0	0	0	SLC38A8	ENST00000299709
16	84040096	rs9934217	A	C	32112.3	PASS	0.736	0	0	0	0	SLC38A8	ENST00000299709
16	84040478	rs11360242	TAA	T,TA,TA	47126.5	PASS	0.456,0.469,0.494	0	0	0	0	SLC38A8	ENST00000299709
16	84040651	.	CAAAAA	AAA,C,CAAAA	34760	cheINDEL99	0.534,0.547,0.556,0.52,0.002,0.6	0	0	0	0.02	SLC38A8	ENST00000299709
16	84040653	.	A	*,C	788.94	PASS	.,0.808	0	0	0	0	SLC38A8	ENST00000299709
16	84040752	rs56979643	G	A	10299.3	PASS	0.132	0	0	0	0	SLC38A8	ENST00000299709
16	84040789	rs112042694	A	C	8192.55	PASS	2.736	0	0	0	0	SLC38A8	ENST00000299709
16	84040950	rs151029891	AAC	A	20575.4	PASS	0.736	0.01	0	0	0	SLC38A8	ENST00000299709
16	84040951	rs72253426	AC	*,A	245.92	PASS	.,0.74	0	0	0	0	SLC38A8	ENST00000299709
16	84040952	rs74818493	C	*,A	2168.62	PASS	.,0.311	0	0	0	0	SLC38A8	ENST00000299709
16	84040955	.	CAAAAA	:AAAAAA,AAA	2824.84	PASS	.,0.944,0.032	0.01	0	0	0	SLC38A8	ENST00000299709
16	84040957	rs201712234	A	*,AAC	20658.3	PASS	.,0.918	0	0	0	0	SLC38A8	ENST00000299709
16	84041217	rs75397367	T	C	8659.54	PASS	0.88	0.08	0	0	0	SLC38A8	ENST00000299709
16	84041258	rs8063697	C	T	30612.4	PASS	0.442	0	0	0	0	SLC38A8	ENST00000299709
16	84041658	rs17724230	T	G	45216	PASS	2.479	0	0	0	0	SLC38A8	ENST00000299709
16	84041761	rs12448870	G	A	11305.2	PASS	1.973	0	0	0.01	0	SLC38A8	ENST00000299709
16	84041766	rs11185262	C	T	8093.4	PASS	2.692	0	0	0	0	SLC38A8	ENST00000299709
16	84041779	rs12448803	A	G	9271.68	PASS	0.553	0	0	0	0	SLC38A8	ENST00000299709
16	84041967	.	.	:TTT,CITTTTT	3703.76	PASS	1.362,1.398,1.326,0.932	0	0	0	0	SLC38A8	ENST00000299709
16	84041970	.	CTTTTT	CITTT,CITTTT	24201	ExcessHet	.,0.999,0.867,0.892,0.822,0.843	0	0	0	0	SLC38A8	ENST00000299709
16	84041991	rs367553600	T	G,TG	5494.08	PASS	0.043,0.024	0	0	0	0	SLC38A8	ENST00000299709
16	84042082	rs2175247	G	A	15113.4	PASS	1.179	0	0	0	0	SLC38A8	ENST00000299709
16	84042270	rs6563990	C	T	27768.7	PASS	0.069	0	0	0	0	SLC38A8	ENST00000299709
16	84042316	rs6563991	A	G	76288.8	PASS	1.055	0	0	0	0	SLC38A8	ENST00000299709
16	84042848	rs8049856	C	G	7747.41	PASS	0.119	0	0	0	0	SLC38A8	ENST00000299709
16	84042866	rs7192437	T	A	26532.5	PASS	3.217	0	0	0	0	SLC38A8	ENST00000299709
16	84043129	rs6563992	T	C	60429.1	PASS	0.531	0	0	0	0	SLC38A8	ENST00000299709
16	84043380	rs7196330	A	G	23140.5	PASS	0.855	0	0	0	0	SLC38A8	ENST00000299709
16	84043672	.	AT	TTTT,ATTTTT	17088.7	cheINDEL99	2.914,2.851,2.788,3.108,3.043,2.796	0	0	0	0	SLC38A8	ENST00000299709
16	84043751	rs111856518	C	G	4433.48	PASS	0.181	0	0	0	0	SLC38A8	ENST00000299709
16	84043976	rs9673442	G	C,A	79665.2	PASS	0.68,1.017	0	0	0	0	SLC38A8	ENST00000299709
16	84044010	rs113614475	G	A	5310.42	PASS	3.941	0	0	0	0	SLC38A8	ENST00000299709
16	84044022	rs112046380	G	C	10600.7	PASS	3.108	0	0	0	0	SLC38A8	ENST00000299709
16	84044032	rs112622822	C	T	10139.4	PASS	1.302	0	0	0	0	SLC38A8	ENST00000299709
16	84044131	rs9673448	G	C	10420.9	PASS	0.364	0	0	0	0	SLC38A8	ENST00000299709
16	84044383	rs68025616	C	T	22810.1	PASS	0.709	0	0	0	0	SLC38A8	ENST00000299709
16	84044566	rs1533077	C	A	20692.2	PASS	0.035	0	0	0	0	SLC38A8	ENST00000299709
16	84044700	rs1533078	C	G	20389.2	PASS	1.126	0	0	0	0	SLC38A8	ENST00000299709
16	84044737	rs111696687	A	G	5511	PASS	0.003	0	0	0.02	0	SLC38A8	ENST00000299709
16	84045242	rs76191844	G	A	6105.4	PASS	0.299	0	0	0	0	SLC38A8	ENST00000299709
16	84045361	rs113190222	C	T	5740.41	PASS	0.389	0	0	0	0	SLC38A8	ENST00000299709
16	84045445	rs11862888	C	T	29607.6	PASS	0.039	0	0	0	0	SLC38A8	ENST00000299709
16	84045455	rs11862917	C	T	15361	PASS	1.778	0	0	0	0	SLC38A8	ENST00000299709
16	84045463	rs11862976	G	T	81307.6	PASS	1.272	0	0	0	0	SLC38A8	ENST00000299709
16	84045480	rs111875323	C	G	5253.41	PASS	1.294	0	0	0	0	SLC38A8	ENST00000299709
16	84045519	rs11859043	T	A	24552.4	PASS	3.716	0.01	0	0	0	SLC38A8	ENST00000299709
16	84045761	rs60876328	T	A	14348.3	PASS	1.842	0	0	0	0	SLC38A8	ENST00000299709
16	84045940	rs17724308	C	G	6755.77	PASS	1.695	0.08	0	0	0	SLC38A8	ENST00000299709
16	84046105	rs9931086	A	C	26451.7	PASS	3.712	0	0	0	0	SLC38A8	ENST00000299709
16	84046402	rs11864124	G	C	35425.6	PASS	0.72	0	0	0	0	SLC38A8	ENST00000299709
16	84046412	rs11149615	T	C	37004.3	PASS	0.042	0	0	0.01	0	SLC38A8	ENST00000299709
16	84046432	rs11864037	A	G	26589.1	PASS	1.516	0	0	0	0	SLC38A8	ENST00000299709
16	84046499	rs11864095	C	T	27972.1	PASS	0.313	0	0	0	0	SLC38A8	ENST00000299709
16	84046715	rs11864146	A	G	9048.12	PASS	3.004	0	0	0	0	SLC38A8	ENST00000299709
16	84046762	rs11864211	C	T	14224.1	PASS	3.319	0	0	0	0	SLC38A8	ENST00000299709
16	84046769	rs13336030	G	A	14246.9	PASS	0.452	0	0	0	0	SLC38A8	ENST00000299709
16	84046853	rs13335951	A	G	16885.9	PASS	0.51	0	0	0	0	SLC38A8	ENST00000299709
16	84046906	rs13336065	G	A	9668.1	PASS	0.994	0	0	0	0	SLC38A8	ENST00000299709
16	84046922	rs79936773	T	A	4635.82	PASS	0.305	0	0	0	0	SLC38A8	ENST00000299709
16	84047027	.	GTGTTT	TT,GTTTTTTT	5535.74	PASS	0.171,0.707,0.132,0.121,0.144	0	0	0	0	SLC38A8	ENST00000299709
16	84047029	.	GTTTTT	TT,GTTTTTTT	5459.31	PASS	0.095,0.098,0.092,.,0.073,0.033	0	0	0	0	SLC38A8	ENST00000299709
16	84047031	.	.	*,TGTG	653.71	PASS	.,0.765	0	0	0	0	SLC38A8	ENST00000299709
16	84047040	.	.	GTTG,G,TGT	2516.59	PASS	0.63,1.121,0.647	0	0	0	0	SLC38A8	ENST00000299709
16	84047167	rs58383994	C	A	12731.9	PASS	1.017	0	0	0	0	SLC38A8	ENST00000299709

16	84047182	rs7199105	C	A	13677.7	PASS	0.575	0	0	0.01	0	SLC38A8	ENST00000299709
16	84047193	rs7198939	A	G	26363.8	PASS	1.211	0	0	0	0	SLC38A8	ENST00000299709
16	84047334		CITTTT	CITTTT,CITTT	19618.5	ExcessHet	0.796,2.141,1.991,1.936,.	0	0	0	0	SLC38A8	ENST00000299709
16	84047505	rs7201097	T	C	21987.9	PASS	0.96	0	0	0	0	SLC38A8	ENST00000299709
16	84047512	rs7199727	C	A	17994.6	PASS	0.619	0	0	0	0	SLC38A8	ENST00000299709
16	84047531	rs7200051	G	A	6491.36	PASS	1.005	0	0	0	0	SLC38A8	ENST00000299709
16	84047623	rs72799212	C	G	6570.88	PASS	0.707	0	0	0	0	SLC38A8	ENST00000299709
16	84047973	rs7200988	G	A	6970.07	PASS	2.577	0	0	0	0	SLC38A8	ENST00000299709
16	84048014	rs7202119	T	C	19049.4	PASS	4.483	0	0	0	0	SLC38A8	ENST00000299709
16	84048081	rs73240258	G	C	4133.36	PASS	0.105	0	0	0	0	SLC38A8	ENST00000299709
16	84048286	rs77228145	G	T	5845.4	PASS	1.215	0	0	0	0	SLC38A8	ENST00000299709
16	84048322	rs7194792	G	A	33849.4	PASS	1.508	0	0	0	0	SLC38A8	ENST00000299709
16	84048540	rs6563993	A	G	6496.45	PASS	9.227	0	0	0.01	0	SLC38A8	ENST00000299709
16	84049182	rs56160820	C	CG	131774	PASS	0.291	0	0	0	0	SLC38A8	ENST00000299709
16	84049862	rs9927180	A	G	69539.4	PASS	2.092	0	0	0	0	SLC38A8	ENST00000299709
16	84049909	rs9938112	T	G	56280.8	PASS	0.005	0	0	0	0	SLC38A8	ENST00000299709
16	84051348	rs7190846	A	T	115208	PASS	2.062	0	0	0	0	SLC38A8	ENST00000299709
16	84051358	rs7192538	T	G	110456	PASS	2.766	0	0	0	0	SLC38A8	ENST00000299709
16	84051486	rs7191206	A	G	73185.4	PASS	3.927	0	0	0	0	SLC38A8	ENST00000299709
16	84051877	rs12598583	G	A,C	17011.6	PASS	6.563,5.703	0	0	0	0	SLC38A8	ENST00000299709
16	84051930	rs2326103	T	C	55635.4	PASS	6.31	0	0	0	0	SLC38A8	ENST00000299709
16	84052141	rs9923716	T	C	20349.6	PASS	10.44	0	0	0	0	SLC38A8	ENST00000299709
16	84052171	rs11864782	C	G	18892.6	PASS	7.765	0	0	0	0	SLC38A8	ENST00000299709
16	84052466	rs8055931	A	T	73828.9	PASS	10.21	0	0	0	0	SLC38A8	ENST00000299709
16	84052470	rs8055932	A	G	73664.3	PASS	9.627	0	0	0	0	SLC38A8	ENST00000299709
16	84052602	rs2175246	T	C	69347	PASS	2.074	0	0	0	0	SLC38A8	ENST00000299709
16	84052662	rs2136660	A	T	69004	PASS	2.624	0	0	0	0	SLC38A8	ENST00000299709
16	84052690	rs9926164	T	*,A	16539.6	PASS	.,2.169	0	0	0	0	SLC38A8	ENST00000299709
16	84052764	rs9935978	G	C	74236.8	PASS	0.664	0	0	0	0	SLC38A8	ENST00000299709
16	84052833	rs2136659	C	G	45932.9	PASS	1.129	0	0	0	0	SLC38A8	ENST00000299709
16	84052929	rs9926354	T	G	21186.2	PASS	3.168	0	0	0	0	SLC38A8	ENST00000299709
16	84053035	rs7200174	T	C	28370.9	PASS	0.313	0	0	0	0	SLC38A8	ENST00000299709
16	84053171	rs7205065	T	A	100111	PASS	1.788	0	0	0	0	SLC38A8	ENST00000299709
16	84053172	rs7203751	C	G	98257.4	PASS	0.788	0	0	0	0	SLC38A8	ENST00000299709
16	84053259	rs7203786	A	G	73212.7	PASS	2.638	0	0	0.03	0	SLC38A8	ENST00000299709
16	84053277	rs9938437	G	T	21453.8	PASS	0.759	0	0	0	0	SLC38A8	ENST00000299709
16	84053301	rs9938227	A	G	21807.8	PASS	1.859	0.03	0	0	0	SLC38A8	ENST00000299709
16	84053417	rs12325554	G	C	11066.2	PASS	1.722	0	0	0	0	SLC38A8	ENST00000299709
16	84053746		GTTT	GT,G,GTTTT,G	23666.7	ExcessHet	5.739,5.681,5.612,6.432,6.376	0	0	0	0	SLC38A8	ENST00000299709
16	84053967	rs2326151	C	T	45220.7	PASS	0.917	0	0	0	0	SLC38A8	ENST00000299709
16	84053973	rs9938837	A	G	24289.7	PASS	1.504	0	0	0	0	SLC38A8	ENST00000299709
16	84054140	rs2326152	C	G	71974.7	PASS	0.887	0	0	0	0	SLC38A8	ENST00000299709
16	84054262	rs80336158	G	A	15770.1	PASS	2.153	0	0	0	0	SLC38A8	ENST00000299709
16	84054830	rs60682377	G	A	17829.5	PASS	0.212	0	0	0	0	SLC38A8	ENST00000299709
16	84054991	rs58108683	G	A	17109.5	PASS	4.562	0	0	0	0	SLC38A8	ENST00000299709
16	84055273	rs58499828	C	T	28775.7	PASS	3.92	0	0	0	0	SLC38A8	ENST00000299709
16	84055282	rs56738423	G	C	28685.7	PASS	0.918	0	0	0	0	SLC38A8	ENST00000299709
16	84055515	rs28533213	A	C	20045.7	PASS	0.204	0	0	0	0	SLC38A8	ENST00000299709
16	84055673	rs57640244	G	A	15713.9	PASS	0.312	0	0	0	0	SLC38A8	ENST00000299709
16	84055769	rs11860273	G	C	81992.5	PASS	0.859	0	0	0	0	SLC38A8	ENST00000299709
16	84055860	rs9924163	C	T	91294.4	PASS	0.825	0	0	0	0	SLC38A8	ENST00000299709
16	84055898	rs2875853	A	G	75630.4	PASS	1.932	0	0	0	0	SLC38A8	ENST00000299709
16	84056325	rs5818479	CTT	C	52071.6	PASS	0.417	0	0	0	0	SLC38A8	ENST00000299709
16	84056661	rs11861366	C	T	22938.9	PASS	1.859	0	0	0	0	SLC38A8	ENST00000299709
16	84056746	rs78001770	TTC	T	18266.8	PASS	1.099	0	0	0	0	SLC38A8	ENST00000299709
16	84056747	rs11865720	T	C,*	25157.3	PASS	2.843,.	0	0	0.01	0	SLC38A8	ENST00000299709
16	84056794		AC	A	703.92	PASS	3.201	0	0	0	0	SLC38A8	ENST00000299709
16	84056828	rs73247332	C	T	12777	PASS	1.503	0	0	0.02	0	SLC38A8	ENST00000299709
16	84057736	rs9929736	A	C	43532.6	PASS	0.12	0	0	0	0	SLC38A8	ENST00000299709
16	84057742	rs9929958	G	C	42382.1	PASS	1.356	0	0	0	0	SLC38A8	ENST00000299709
16	84057749	rs71404128	T	C	36126.7	PASS	3.595	0	0	0	0	SLC38A8	ENST00000299709
16	84057881	rs12933901	G	C	38240.3	PASS	4.319	0	0	0	0	SLC38A8	ENST00000299709
16	84057974	rs12933666	A	G	67284.2	PASS	0.104	0	0	0	0	SLC38A8	ENST00000299709
16	84057975	rs12933796	C	G	67284.2	PASS	0.496	0	0	0	0	SLC38A8	ENST00000299709
16	84058183	rs12928777	C	T	91084.4	PASS	0.732	0	0	0	0	SLC38A8	ENST00000299709
16	84058227	rs72799224	A	C	6565.72	PASS	1.119	0	0	0	0	SLC38A8	ENST00000299709
16	84058430	rs146244220	G	A	6845.59	PASS	0.809	0	0	0	0	SLC38A8	ENST00000299709
16	84058499	rs150832828	T	C	6735.93	PASS	3.17	0	0	0	0	SLC38A8	ENST00000299709
16	84058538	rs77603919	G	A	6833.45	PASS	0.088	0	0	0	0	SLC38A8	ENST00000299709
16	84058641	rs12930439	T	C	73795.4	PASS	0.618	0	0	0.01	0	SLC38A8	ENST00000299709
16	84058754	rs12929637	C	T	70185.1	PASS	0.005	0	0	0	0	SLC38A8	ENST00000299709
16	84058763	rs141444908	A	ATGC	93482	PASS	0.899	0	0	0	0	SLC38A8	ENST00000299709
16	84059077	rs72799227	C	T	6805.45	PASS	2.366	0	0	0	0	SLC38A8	ENST00000299709
16	84059137	rs72799228	G	A	2836.12	PASS	1.107	0	0	0	0	SLC38A8	ENST00000299709
16	84059342	rs11862961	T	C	52050.3	PASS	5.917	4	0	0	0	SLC38A8	ENST00000299709
16	84059455	rs11866872	G	A	50762.9	PASS	0.945	0	0	0	0	SLC38A8	ENST00000299709
16	84059486	rs78382568	T	A	12710.8	PASS	0.491	0	0	0	0	SLC38A8	ENST00000299709
16	84059594	rs11866810	A	G	15175.6	PASS	2.884	0	0	0.01	0	SLC38A8	ENST00000299709
16	84059804	rs77675267	T	A	7765.44	PASS	0.188	0	0	0	0	SLC38A8	ENST00000299709
16	84059833	rs111629379	G	GTGTTT	17661.6	PASS	1.058	0	0	0	0	SLC38A8	ENST00000299709
16	84059859	rs11649522	G	C	41215.1	PASS	0.684	0	0	0	0	SLC38A8	ENST00000299709
16	84060942	rs138796615	A	AAT	29682.1	PASS	0.715	0	0	0	0	SLC38A8	ENST00000299709
16	84060995	rs28726675	C	A,*	25155.6	PASS	1.114,.	0	0	0	0	SLC38A8	ENST00000299709
16	84061050	rs12922255	C	T	40372.3	PASS	0.567	0	0	0	0	SLC38A8	ENST00000299709
16	84061650		AIT	T,ATTT,A,ATTT	38870	ExcessHet	0.99,0.993,0.923,0.967	0	0	0	0	SLC38A8	ENST00000299709
16	84062175		GA	A,G,GAAA,G	34563.7	PASS	0.37,0.346,0.36,0.35	0	0	0	0	SLC38A8	ENST00000299709
16	84063320	rs143276186	A	AAGGAAAG,G	67855.9	PASS	8.619,8.837	0	0	0	0	SLC38A8	ENST00000299709
16	84063715	rs12919122	G	A	47012.8	PASS	0.529	0	0	0	0	SLC38A8	ENST00000299709
16	84064817	rs2010830	A	G	97217.2	PASS	0.547	0	0	0	0	SLC38A8	ENST00000299709
16	84065710	rs907043	A	G	133685	PASS	0.114	0.03	0	0	0	SLC38A8	ENST00000299709
16	84065790	rs140016057	TTTG	*,T	103901	PASS	.,0.077	0	0	0	0	SLC38A8	ENST00000299709

16	84065790	rs140016057	TTTG	*,T	103901	PASS	..0.077	0	0	0	0	SLC38A8	ENST00000299709
16	84065933	rs1806488	A	C	26850.8	PASS	0.92	0	0	0	0	SLC38A8	ENST00000299709
16	84067422	rs4782878	A	G	112393	PASS	7.821	0	0	0	0	SLC38A8	ENST00000299709
16	84067622	rs9922874	A	G	108858	PASS	3.405	0	0	0	0	SLC38A8	ENST00000299709
16	84068625	rs11149616	C	A	31059.9	PASS	6.633	4	0	0	0	SLC38A8	ENST00000299709
16	84069222	rs12922815	T	A	59927.6	PASS	3.59	0	0	0	0	SLC38A8	ENST00000299709
16	84069660	rs9933841	C	G	102074	PASS	0.406	0	0	0	0	SLC38A8	ENST00000299709
16	84069768	rs11860327	T	C	109271	PASS	4.945	0	0	0.01	0	SLC38A8	ENST00000299709
16	84071061	rs11149617	C	G	71375.8	PASS	1.366	0	0	0	0	SLC38A8	ENST00000299709
16	84071375	rs36155963	A	AG	27208.7	PASS	0.546	0	0	0	0	SLC38A8	ENST00000299709
16	84071376	rs36197780	A	AG	28394.3	PASS	0.568	0	0	0	0	SLC38A8	ENST00000299709
16	84071915	.	CA	C,CAA	7856.97	PASS	0.749,0.793	0	0	0	0	SLC38A8	ENST00000299709
16	84073089	rs4782880	G	C	95355.6	PASS	0.342	0	0	0	0	SLC38A8	ENST00000299709
16	84073348	.	CAAAAA	CA,CAAAA,CA	16865.7	PASS	0.756,0.737,0.778,0.07,,0.699	0	0	0	0	SLC38A8	ENST00000299709
16	84073349	rs4782882	A	*,C	513.28	PASS	..1.252	4	0	0	0	SLC38A8	ENST00000299709
16	84073350	.	A	*,C	603.91	PASS	..0.147	0	0	0	0	SLC38A8	ENST00000299709
16	84073351	.	A	*,C	1404.06	PASS	..0.152	0	0	0	0	SLC38A8	ENST00000299709
16	84073413	rs148942102	G	GAGCC	22934.6	PASS	3.417	0	0	0.01	0	SLC38A8	ENST00000299709
16	84073530	rs7196363	A	G	91516.3	PASS	0.918	0	0	0	0	SLC38A8	ENST00000299709
16	84074875	rs34492560	G	A	25255.3	PASS	0.006	0	0	0	0	SLC38A8	ENST00000299709
16	84074921	rs12716746	G	C	89697.1	PASS	1.383	0	0	0	0	SLC38A8	ENST00000299709
16	84074978	rs9928085	C	T	34604.8	PASS	1.364	0.03	0	0	0	SLC38A8	ENST00000299709
16	84075138	rs9928107	A	G	91116.3	PASS	3.9	0	0	0	0	SLC38A8	ENST00000299709
16	84075546	rs144889482	C	T	1128.29	PASS	3.598	0	0	0	0	SLC38A8	ENST00000299709
16	84075563	rs1105355	C	A	64529.8	PASS	0.893	0	0	0	0	SLC38A8	ENST00000299709
16	84075761	.	A	G	311.95	PASS	17.41	0	0	0	0	SLC38A8	ENST00000299709
16	84077987	rs12928842	T	C	85031.7	PASS	5.219	0	0	0	0	SLC38A8	ENST00000299709
16	84078089	rs12927808	A	G	52195	PASS	0.79	0	0	0	0	SLC38A8	ENST00000299709
16	84078169	rs34441238	G	T	44761	PASS	1.657	0	0	0	0	SLC38A8	ENST00000299709
16	84078448	rs11640940	A	G	65940.2	PASS	6.459	0	0	0	0	SLC38A8	ENST00000299709
16	84078762	rs7195511	G	A	28883.5	PASS	1.2	0	0	0	0	SLC38A8	ENST00000299709
16	84079226	rs8058441	A	G	38184.8	PASS	2.088	0	0	0	0	SLC38A8	ENST00000299709
16	84079320	rs6563995	G	C	67940.3	PASS	1.79	0	0	0	0	SLC38A8	ENST00000299709
16	84079818	rs13332704	C	A	57424.4	PASS	3.093	0	0	0	0	SLC38A8	ENST00000299709
16	84079996	rs12445301	C	T	26069.5	PASS	0.062	0	0	0	0	SLC38A8	ENST00000299709
16	84080087	rs8061408	T	G	43350.9	PASS	5.562	0	0	0	0	SLC38A8	ENST00000299709
16	84080147	.	ATTTTTT	TTTTTT,ATTTTT	2050.95	cheINDEL99.	2.076,2.128,2.025,1.841,1.911	0.03	0	0	0	SLC38A8	ENST00000299709
16	84080386	rs4782884	C	G	59872.2	PASS	0.227	0	0	0	0	SLC38A8	ENST00000299709
16	84080502	rs62048168	C	A	15244.2	PASS	0.367	0	0	0	0	SLC38A8	ENST00000299709
16	84080524	rs11149619	G	A	44883	PASS	0.694	0	0	0	0	SLC38A8	ENST00000299709
16	84080769	rs11644047	G	A	84067	PASS	2.825	0	0	0	0	SLC38A8	ENST00000299709
16	84080893	rs11643984	A	G	30580.6	PASS	0.346	0	0	0	0	SLC38A8	ENST00000299709
16	84082228	rs11149620	T	C	29544.6	PASS	0.247	0	0	0	0	SLC38A8	ENST00000299709
16	84082650	rs8056413	G	T	68011.3	PASS	0.682	0	0	0.01	0	MBTPS1	ENST00000343411
16	84082706	rs8057497	T	G	25057	PASS	1.585	0	0	0	0	MBTPS1	ENST00000343411
16	84082784	rs8057673	T	C	68378.2	PASS	3.48	0	0	0	0	MBTPS1	ENST00000343411
16	84083300	rs907031	T	C	46145.6	PASS	0.059	0	0	0	0	MBTPS1	ENST00000343411
16	84083368	rs907030	G	A	48640.5	PASS	0.089	0	0	0	0	MBTPS1	ENST00000343411
16	84084419	.	C	T	9872.73	PASS	7.58	0	0	0	0	MBTPS1	ENST00000343411
16	84084420	rs78112308	A	G	30095.9	PASS	8.359	0	0	0	0	MBTPS1	ENST00000343411
16	84084427	rs201558841	C	G,*	7531.82	PASS	8.068,	0	0	0	0	MBTPS1	ENST00000343411

### Supplementary Table 4.3. VEP annotation of SLC38A8 genomic sequence.

CHROM: chromosome; REF: reference allele; ALT: alternative allele; COV: coverage; DS\_AG: delta score acceptor gain; DS\_AL: delta score acceptor loss; DS\_DG: delta score donor gain; DS\_DL: delta score donor loss. Filter column indicates if the variants have been retained in the analysis by satisfying the filtering criteria. Coordinates are according to GRCh37 and the transcript ID is from Ensembl.

>Aberrant *PAX6* mRNA

GCATGTTGCGGAGTGATTAGTGGGTTTGAAAAGGGAACCGTGGCTCGGCCTCATTCCCGC  
TCTGGTTCAGGCGCAGGAGGAAGTGTGTTTGGCTGGAGGATGATGACAGAGGTCAGGCTTCGC  
TAATGGGCCAGTGAGGAGCGGTGGAGGCGAGGCCGGGCGCCGGCACACACACATTAAC  
ACACTTGAGCCATACCAATCAGCATAGGAATCTGAGAATTGCTCTCACACACCAACCCAG  
CAACATCCGTGGAGAAACTCTCACCAGCAACTCCTTAAAACACCGTCATTTCAAACCATT  
GTGGTCTTCAAGCAACAACAGCAGCACAAAAAACCCCAACCAAACAAAACCTTTGACAGAA  
GCTGTGACAACCAGAAAGGATGCCTCATAAAGGGGGAAGACTTTAACTAGGGGCGCGCAG  
ATGTGTGAGGCCTTTTATTGTGAGAGTGGACAGACATCCGAGATTTAGAGCCCCATATTCGA  
GCCCGTGGAAATCCCGCGGCCCCAGCCAGAGCCAGCATGCAGA**ACAGTCACAGCGG**  
**AGTGAATCAGCTCGGTGGTGTCTTTGTCAACGGGCGGCCACTGCCGGACTCCACCCGG**  
**CAGAAGATTGTAGAGCTAGCTCACAGCGGGGCCCGGCCGTGCGACATTTCCCGAATTCT**  
**GCAGACCCATGCAGATGCAAAAGTCCAAGTGCTGGACAATCAAACGTGTCCAACGGAT**  
**GTGTGAGTAAAATTCTGGGCAGGTATTACGAGACTGGCTCCATCAGACCCAGGGCAATC**  
**GGTGGTAGTAAACCGAGAGTAGCGACTCCAGAAGTTGTAAGCAAATAGCCAGTATAAG**  
**CGGGAGTGCCCGTCCATCTTTGCTTGGGAAATCCGAGACAGATTACTGTCCGAGGGGGT**  
**CTGTACCAACGATAACATACCAAGCGTGTCAATAAACAGAGTTCTTCGCAACCTGGC**  
**TAGCGAAAAGCAACAGATGGGCGCAGACGGCATGTATGATAAACTAAGGATGTTGAACG**  
**GGCAGACCGGAAGCTGGGGCACCCGCCCTGGTTGGTATCCGGGGACTTCGGTGCCAG**  
**GGCAACCTACGCAAG****GCTTCATGGGGGTCAAGAAGAAGTTGCAGTGGTGGCAGTGGGA**  
**TTACTAACCCCAAGAGTGCAGCCCATACATGTGTTCAATTAACCTCCAGTGGTGGTC**  
**TCCCAATTCTTACCTTGGTACCTGTGCCTGACTATGGTACAGGTAGTTGCTTCCCTGCTG**  
**GCCAGTTGATGGCAATCTCCTCAGCCCTTCAATTATTTACTAA**

>Aberrant *PAX6* protein

MQNSHSGVNLGGVFNQRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQLDNQNV  
SNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSIAQYKRECPSIFAWEIRDRLLESEGVCNT  
DNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQ  
**GFMGVKKKLQVWVQWDLTPRVQPHTCVQLIPPVVVSQFFTLVPVPDYGTGSCFPAGQFDGN**  
**LLQPFQLFH**

>Wildtype *PAX6* protein

MQNSHSGVNLGGVFNQRPLPDSTRQKIVELAHSGARPCDISRILQTHADAKVQLDNQNV  
SNGCVSKILGRYYETGSIRPRAIGGSKPRVATPEVVSIAQYKRECPSIFAWEIRDRLLESEGVCNT  
DNIPSVSSINRVLRNLASEKQQMGADGMYDKLRMLNGQTGSWGTRPGWYPGTSVPGQPTQ  
DGCQQQEGGENTNSISSNGEDSDEAQMRLQLKRKLQRNRTSFTQEIEALEKEFERHTYPD  
VFARERLAAKIDLPEARIQVWFSNRRAKWRREEKLRNQRQASNTPSHIPSSSFTSVYQPIPQ  
PTTPVSSFTSGSMLGRTDALTNTYSALPPMPSFTMANNLPMQPPVPSQTSSYSCMLPTSPSVN  
GRSYDYTPPHMQTHMNSQPMGTSMTSTGLISPGVSPVQVPGSEPDMSQYWPRLQ

**Supplementary Figure 4.1 Predicted aberrant *PAX6* mRNA and its translation.** Potential effect of g.31773553\_31797720del in *PAX6*. Blue highlights the nucleotide and amino acid sequences derived from *PAX6* and green is for sequences that derive from *ELPF4*. Bold text marks the reading frame in the aberrant mRNA which gets translated into protein. Red text highlights the additional amino acids that gets incorporated to the protein. Genomic sequence was translated into protein sequence using ExPASy (<https://web.expasy.org/translate/>).

X: 9383915-9984211

aacaagaatgtctgtatgctatgggtggagaaacccactctgcagtaataaaacctgcaatthtttagctgagtcctggccaccagaatccagatcacatttccagcctcattg  
cagctatatgcatctatagtaagctctcaccactgggatggacagaaatgatgtgtcaactcagaggaaagaagcatgccctcctgtcttcatcggttcagctgtcaaat  
ccgtagactgcgggacttagacaactgacatttatttccctattctggaggctgggaggtccaaaatcaagggtgctcagcagggtcctcctgagggcctcacaacctataac  
gacagctgtctcctcattgttcttacttggtagatcgcgagagagaagcagaagcctctctggtttctccttataagggcactaaacctatcgcgaggtccacaacctataac  
ctcatcagctccaaaaggcttcatctctgacacccatccattggaggttagagttcaacatataaattgggcatgacgcaagcatgagctgttaacagctcccctttctctttc  
tccttgatcacaacgttatgattggacacggaacagcggcttagaccacaagttggacactgccaaccaagaagagccacggagtagaaagagcctggctcatggctggg  
cgggtggctcacgctgtaatccagcactttgggagggcaaggcgggctgagcaggaggtcaggagatcgagaccatccggctatcatgtgaaacccgctctactaaaa  
atacaaaaaataagccggcgtgggtggcggcgcctgtagtccagctactcgggaggtgagggcaggagaatgggggtgaacccgggagcgggagcttgagtgagccgaga  
tcgcccactgcactccagcctgggcgacagagcgagactccgtctcaaaaaaaaaaaaaaacaaaacaaaacaaaagaaaagaaaagaaaagaaaagaaagagcctgg  
ctctgctgttgcagaacactgcagcagctctgaacctccgccccaaacttaacctgagagagaatgagcttctgctggcttatgcaatttggggcctctgtatcttgagc  
caaatgaatcAATTCCTACTGTATTAGGGGAACCAAGTTCCTGACTGCCTGAGAGAGAGTCACTCGGGCAGGTGAGGAAAAAGGAT  
GAGGCTGCTGTGGAACGGACTGAAGCAGGGTTGTCTTCTGACTGCCGTGCAACCTTGACCTTGAAAGAAATTGGCCCCAC  
GCCTTCAGATGTCTCAGGTTTACAGATGTCTCAAAGCATAATGTTTTCCCTTGCAATGCAAGACTGtagacttccaagacttaactgattgtg  
acaagagaaggagcgaagtcaagttgcaatagtgatccacacatggctgtctcagctcatctgtagtggtcactgcagcacactgtcccactgtgccgaataaact  
gactgcagcaaaaaaacacttgcagctgtcggaaaccagggccactccagattattgattctcaatcactaaatcataaatgtgtgaaatctCTGATTCaggggtcagca  
aactgcagcctggagccaaacccagccactgtctgtttgtacagctcatgagctatgatttttcaattttaaatataattttttcagcagtgctcact  
ctgttcccaggctgcagctgagttgtatcacggctcactgcagactccactcctggctcaagtaactcctcctgcctcagcctctgtagtgctggaactacaggcatgac  
caccatgccagctaatgtattatttttttagagatgaggtctccctatgtgccaggctgtctcaactcctg



cttaagcctcggcctccaaagtgtgggattacaggcCTTGTGGCACCTGGCCTCAAAGAACTTTTTAATTGACACTTAATAGCAAGCAACAACAGA  
TTACAATTAGGCCAGTCTGTGGCAGGAAAAACAAGAAAGGGGAGAGGAGAGAAGAACTCAATTTCTGGAAGACAAGCAAAACAATT  
GCTTGTCTCattctgattggaacattctccattggaacaGGAGAGCAAACCAACAGTACTTGCTAATAGATTTGATGACGGGGGAGGGTAGA  
GGAAGTATATCAGATCTGTTTTCTTTTTAAGAGACAATGAAAGCAGAAGATATAAAAAGAAAAACAATTTTTgttctctgttctgtctgtaattatt  
ttcaagtttgaagttctgttttccctctgtgtggcagggcaaggtcacagaatgatttaagttgcaagcctgtcactgttaacaaactgccttgttctgcttgaagctgc  
ttgctgccctacagttttgtgctatcaaaactggccaacccctctggatgcatgataaaagtaagccctgtcttgggggctcagccttggatgtaaatcgtgggcccgt  
gagcactaataaactcctctgtccaccattgtctctactgtcccttaattcctgcaataacaaggtctcactttgtgccaagctggagtgctggtggcacagtaagctcac  
tgaccctcacttgaagggccaaagcctcctccactcagcctctggagtagctcgggctacagcgtacaccaccagctggctgatttttggtttttagagacagggttc  
cctttgtgagcaggctggctcctgaaactgtggcctaaagtatcctcctgcctcagcctccaaagtgtgggattacaggcgtgagccactgcacctggcTGGAAACTCT  
TAAAAAGCTTGGGTTggccggcgtgggtcactcactgtaatccagcactttgggagggcaagcgggagatcatgaggtcaggagatcaagaccatcctggctaa  
catggtgaaacccgtctactagaatacaaaaaataagctggcgtgggtggcggcacctgtagtcccagctactcaggaggtgagggcaggagaatggcaagaaccagg  
aggcagagcttgagtgagccgagattgcaccaccgcactccagcctgggcaacagagcgagaccatcctcaaaaaaaaaaaaaaaaaaaaaaGTTTGAGGTTAGT  
GGTCCatttttatttatttttaattgacaataaattgtacatattctggggtgcatagtgatgtcaatacatagaatgtacagcaatcagatcaggggtgagtagt  
ccgtaatcacaacattatcatttctgttgaaaaactcaatcctccttttagctataatttcaatttaattcattttcaattttattttattttatttttagagacagtgct  
cactgtcggcaggctgagtgctgtggcacaatcacaattcatagcagcctcaaattcctgggcatgtaactcctcctcagctcttgaattcagctattcaaaactattt  
atttttaactctagctcatcaacagttggtatagaacactagaattatttctccaatctagctgtcattttgtatcctttaaagaaattAAATGAAAGTTAGAGAAAACCAA  
CTCAAAGTACTTAAAAATAAATCAGGCTGTCCAAGTAGTAGTTTCAGAAACAGGTGGTTCACAGATCTGATATActtctgtaccctcccc  
catcgtcaaatctattgcaaaactgttggtttctcctcagtagctggaccatccacttagcaaaccaagcttccatcttctcagcagaacccaaatgaga  
tggtttcctcacttctactgttgcctttctgattggtccatttttaaACAAATTCTACTTTACCAAAAAGTCACTTTTCTGAAAAGTCCGGGTATGAAGGAAAAG  
GTGTTGGTTTTGCAGCCAGTGTGTTGGCATAAAGCTGTGTGGTGATCAAAAAGAGGCCTCCAGTGGAAGGAGAAATCCACCCATCCA  
GCCGCTGGACATGGCATTGCCGATCCCCCGGGCTGTTGGTGGGAACCTTGAGCTTTAGAACCTCCACAGGTAATATGGGCA  
ACGTAGGACACGGACGCTCAAGTAGTGCAGTCCGGCCAGAAAAACAGAGACTCCCCAGCAATGTTGACACTTCTCTCCCAAGAACT  
CCTCCTCGGCCCCAGCATCTCAGCCACATCAGTCCAGGATGACCGCCACAAGCCCCAGAGAACCCACAGCCCCGGCCACGCAC  
GTGAGGACACGAGCCACCTGGAGTTGGCAGCCAGGTTCTTTGGGCTGGCACACACTGGCACCTGTGTCTCGAGTCACACATGCTTC  
CAAAGGCCAGCATGACTTCTCACTCTCGAGTAACCCGAGGAGAAAGTcggccactgaggtatgacacacagggcactgaggtatgactggc  
aggcactgaggtatgcaGGAAACAGACGCGCCAGCCACAGAGGCCAAGAGCCCCACATTTGAATCAGCCAGATCGCCCTGCTTTT  
ATCCTGCTCTCCAGGCCAAGTATTCTGGGAAACTGTCTCTCGGTACAACAGGACTGGCTCGTTAGATTGACATCTTGACGGATT  
GCCGAGTTCGTGCCAGCTGGTCTGCAACCATTAGTTAGTCCCTTTGTCAACCAACCAGAATGCAGCTGCTTATAAATCGGGTACTTG  
GAACTGGGCTGCTATTGTTCAACCCTAACAACGCAGATGAATACCTACCATCAAAATAGGCCAATAGAAAACCT

**Supplementary Figure 4.2. *GPR143* locus with primer binding site and deletion breakpoints.** DNA sequence annotated with g.9384915\_9982211del and the primer binding sites for LR-PCR. Primer set 1 is in blue and primer set 2 is in yellow. Repeat sequences are highlighted in lowercase (Soft-masked).

GRCh38	Reference	African 1	Father	Mother	African2	Father	Mother	Afro-Caribbean	Father	Mother	African AF	European AF	dbSNP
15:27979889	Z/Z	W/W	W/W	W/W	W/W	W/W	W/W	W/W	W/W	W/W	0.9719	0.9607	rs11074315
15:27980004	Z/Z	Z/Y	Z/Y	Y/Z	Y/Z	Z/Y	Z/Z	Z/Y	Y/Z	Z/Z	0.04376	0.0001476	rs59967422
15:27992851	Y/Y	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	0.9882	0.9887	rs4499192
15:27994398	X/X	X/Y	X/X	X/X	Y/X	X/Y	X/X	X/Y	X/X	X/X	0.1626	0.00319	rs57376758
15:27997088	Y/Y	Y/Y	Y/Z	Y/Y	Y/YZYYZZYYZ	Y/Y	Y/YZZYYZZYYZ	Y/ZYYZZYYZ	Y/Y	Y/ZYYZZYYZ	0.1714	0.0005601	rs368785616
15:2799839	Z/Z	Z/Y	Z/Y	Z/Y	Z/Y	Y/Z	Z/Z	Z/Y	Y/Z	Z/Z	0.128	0.0003386	rs12101660
15:27999898	X/X	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	0.9999	0.9979	rs4640132
15:28000188	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Y/Z	Z/Z	Z/Z	0.3992	0.7946	rs28818582
15:28000414	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Z/Y	Y/Y	Z/Y	0.01754	0.0000294	rs149366789
15:28000665	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Z	Y/Y	Y/Y	Y/Y	0.1545	0.0003087	rs58913900
15:28002351	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Y/Z	Z/Z	Z/Z	Z/Z	0.1534	0.0003087	rs74005195
15:28002374	X/X	X/X	X/X	X/X	X/W	X/X	X/W	X/X	X/X	X/X	0.135	0.0002499	rs74005197
15:28003579	X/X	Y/X	X/X	X/X	X/X	X/X	X/X	X/X	X/X	X/X	0.1071	0.0003089	rs16950781
15:28003682	W/W	W/W	W/W	W/W	W/W	X/W	X/W	X/W	W/W	X/W	0.157	0.04946	rs57242412
15:28005010	W/W	X/W	X/X	W/X	W/W	X/W	X/W	X/W	W/W	X/W	0.309	0.09547	rs4778136
15:28011531	Z/Z	Z/YZ	Z/YZ	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	0.1091	0.0008183	rs112437786
15:28011881	X/X	X/X	X/X	X/X	X/X	X/X	X/X	W/X	X/X	W/X	0.01654	0	rs143157494
15:28013609	X/X	X/X	X/X	X/X	X/W	X/X	W/X	X/X	X/X	X/X	0.06956	0.0001617	rs74005198
15:28014717	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Z	Y/Y	Y/Y	Z/Y	0.2194	0.04995	rs4640131
15:28014907	Z/Z	Z/Y	Z/Z	Y/Z	Y/Z	Y/Z	Z/Z	Z/Y	Y/Z	Z/Z	0.1373	0.04837	rs1800401
15:28017719-28020673	-	W/Del	W/W	W/Del	Del/W	Del/W	W/W	W/Del	Del/W	W/W	0.002543	0	-
15:28019364	X/X	X/Del	X/X	Y/Del	Del/X	Del/X	X/X	Y/Del	Del/X	Y/X	0.02032	0	rs112638567
15:28019708	Y/Y	Y/Del	Y/Y	Z/Del	Del/Y	Del/Y	Y/Z	Z/Del	Del/Z	Z/Z	0.4398	0.7649	rs2871875
15:28020616	Z/Z	Z/Del	Z/Z	Z/Del	Del/Y	Del/Y	Y/Z	Z/Del	Del/Z	Z/Z	0.02916	0.0000588	rs78145457
15:28020856	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Z/Z	Y/Z	Y/Z	Z/Z	0.08285	0.0004705	rs77490758
15:28021089	W/W	W/W	W/W	W/W	W/W	X/W	X/W	X/W	X/W	X/W	0.3672	0.4944	rs746861
15:28021139	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Z/Y	Z/Y	Z/Y	0.08184	0.000147	rs74531804
15:28021881	Z/Z	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Z	Y/Y	Y/Y	Z/Z	0.3221	0.1683	rs2122005
15:28022387	Z/Z	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	ZWZZZWZYZZWZZZWZYZZ	0.8557	0.2215	rs11283428
15:28022612	Y/Y	Z/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	0.08737	0.009229	rs41534647
15:28022823	Y/Y	X/Y	Y/X	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	Y/Y	0.0134	0.0000441	rs148755931
15:28022866	W/W	W/W	W/W	W/W	W/W	W/W	X/W	W/W	W/W	X/W	0.01773	0	rs114176032
15:28023490	Z/Z	X/Z	X/X	Z/Z	X/Z	X/Z	X/Z	Z/Z	Z/Z	Z/Z	0.3323	0.003131	rs7174197
15:28023844	X/X	X/X	X/X	X/X	X/X	X/X	X/X	Y/X	X/X	Y/X	0.2137	0.1458	rs749846
15:28023862	Y/Y	X/X	X/X	X/X	X/X	X/X	X/X	X/X	X/X	X/Y	0.8548	0.2197	rs3794606
15:28024696	X/X	W/X	W/W	X/X	X/W	X/X	X/W	X/X	X/X	X/X	0.3327	0.00274	rs28546555
15:28024900	X/X	X/Y	X/X	X/Y	X/X	X/X	X/X	X/X	X/X	X/X	0.0001332	0	-
15:28025097	Z/Z	Z/Y	Z/Z	Y/Z	Y/Z	Z/Y	Z/Z	Z/Y	Y/Y	Z/Z	0.187	0.0004116	rs73375883

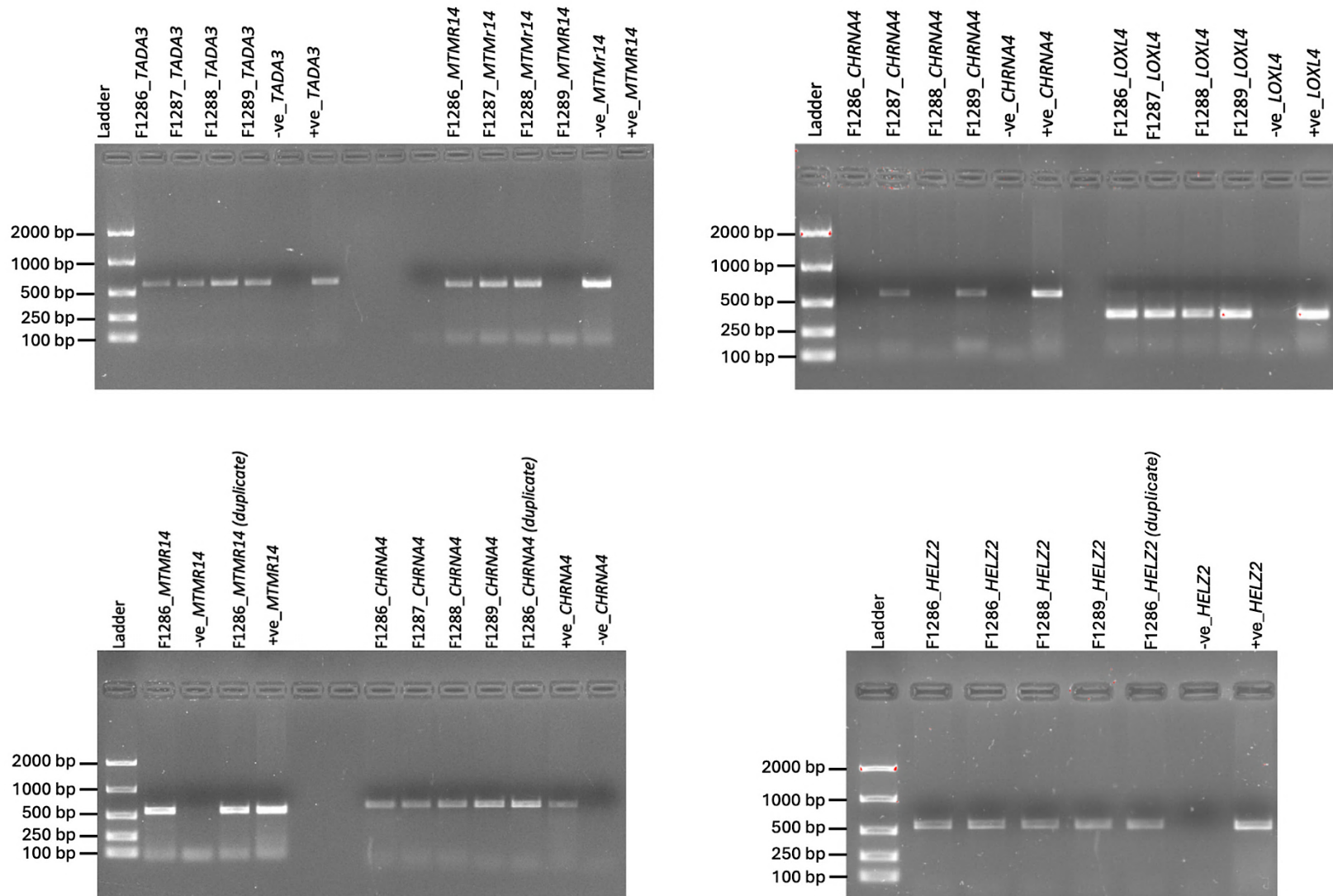
**Supplementary Table 4.4. Haplotype analysis in three families of African descent.** The alleles are presented in a 45208 bp region spanning chr15:27979889-28025097. The maternal alleles are highlighted in orange and the paternal alleles are in blue. Uninformative alleles that cannot be phased in the proband are in black. The “/” separates the alleles on the homologous chromosomes. The red box encompasses the deleted region by NC\_000015.10:g.28017719\_28020673del. The allele frequency is for the variant allele and is derived from the African and European demographic in gnomAD.

**TTTGCTCTAGGTTTCAGGCGAG**TTGTTTCTGCTGCAGCCTTATTCTGGGATCAGGAGTGGGCGGTTAACTGATGTCCGCCAGCCTTTCT  
GGGTGAAAGGGTGGGAGGGCTGTTTCTCAGCTCTCCAGCCAGGGTCCAGAGCCTATTTTGGAAATGGACATGGTCATTAAGGTGTCTGTG  
CATGAAGGCCCTGAAGCCCCGGTACTGCTAGTGCAGGCGGGCATCCGTGGAGGGTCTGGAGAGGTGGGACCTCCGATGCAGTG  
CTCTGATCAGGCAAATATTTTATGATTATTTAAAAAGGAATTAATTCAGTTAGCTAAATGGCAAGTGTTTTTTCTTTAGTCTACTTACTTG  
ATAGCCTCAGTGAAGTAAATTTAGGAACTGGCTTCTTCTGCAAGTGTCTGTAATCTGTTGCATAGTCTTGGTTTTTGTAGTCTTCTG  
GCGAGTCACATACATCTTCCAGGACACCTCTCCTAGCCTGACCTACCCAGCAGGGCACCAGCCAGCCTTCCAGAAAGGCCTGG  
CACTTGAGGCTTCTGGTGAGCCGGGGGCACACTGGTCACAGTTTGTGCTTCCCTCTGCTCAGTGGATGGAGGCACCCACAGTTTTAGA  
AGCACTCCACTTCCACTGGGGTCCAGGGCGGTGCTGATGTCAGTGTGAGTGCAGTGTGCACGTGTTTGTATTTTCTCATCACTTTAAACCGT  
GCCTGAGCTGTAAAGTATCTCCATGT**GAGAGAACC GAAGAGCTGTGTGATTTGACAGACTTTGTTTCTATTCTTGTGCCTCGACCTCTGC**  
**ATTACTTCTCCAGGATGAAGAAAGCCAAGACTTGCCTGTGTTCTGGCCCCCTCATGAAGTCCAGGACATGCCACCTCTGGGG**  
**AGGACTTTCCCGAGGGGGAAGGGGAGATAGAGCAACCGGCTGCCAGGCTGACCCAGGAACCTGAAGGGGGAGGCTGGATCCTGTC**  
**CAGAAACCGTCTCAGTGGGACTCAGGCAGGAGGGGGAGATTTCCGGGAACACTGGCACCAACTTGATCAGTCACTAGAGATCCAC**  
**GCGAGCACAGCAGCTTAGGTCACTAAGATGAGATGATAGCCTCAGCCCTCTGGTCATCCACCCAGGCCAGTGTGCCCTTTATT**  
**TAAGGCATTTGCGAGCTCCTTCTCGAGCAGCCTGAGTGCACCAGGCCCCACGCAGGACCCACAGCCTGAGTCCCTCCAGCAG**  
**GGGGCAGGGCAGCCTGCCCTTTTGGCAGGGGGTGGGGATGGGTGGGGCGGGGAGCTGTCTCGCCTGCTCCCTGTGTGGCCTTCG**  
**AATTTGGATGAAGAGGGAGAGAGAGAGCCTGGGAGAGATGAAGGCCGAGGGGGGAAGACGGGTGCTGCTCCTCCAGCGCAGTGG**  
**GTCACCAGCCAGATGAGGGGTTGATTTGAGCCTCCAGCAGGCTCCTTGGAAAGGGCAGTGTCCAGGGATTAACCACGCTTTCCG**  
**GAAACTCACGGACCCTCTTTGTGAAAGCTGAGCTTTGTGAAAGCTGCAGGCCTGAGGCCTGCCTGGGGAGGAGGGGAAGCCAGCT**  
**GGACCCTGCCCCCAAGGTGATTCCCTGAGAAAGGGCGTGCAGTATAGGCGTGAGGTAGGAGGGGCTCGGCCTGGTTGGAAAG**  
**GCCACAGCAATGCTCAGGGCTGTCTTCTGAGCAAGGAGCTCGAGGCCTCCTCCGGCCCTCTGGCCTGCCCTGTGGAGGAGGC**  
**GGGTGAAGGACACCAGGGTGTCTGCTGTGAGGGGAGGAGCTTCTGGAGAAAGCACCACCTCTCCTTTAGCCGGTGAATAGCGAAGCCA**  
**CCTTTCTTTCGGAACCTCAGTGTCTCCCGCAGCAGCCAGGCGGTGTCGGTTCGGTAGTTGCAGGGTCTATTTCAGAAGCCTGTAATTGA**  
**GCACCTGAAGTCTTTGCTCCTGAATTGTTGTCTGAAGGAGGCCTGTGAGGCTCCAGGGCACTGGGCAAATGAACAGAAGCCAGCAA**  
**CCTCTACAACCAGGTCAAGTAAAGGGCGCCGCTTGTCTCCCGCTCCTTTCTTCTCCTCCCTCCTTCTGCTCTTTGCTGCGTC**  
**CTCTTTCTCCTCTCCCTTCTCTCTCCCTCCCTTCTCCCATCTCCTATTCTCCTCCCTCTTCCCATCCCTTTCATTCTCTCC**  
**TTTTCCCATGTGATGCAATTTGATCACACCCAGCGCTGCATCCTGGGCTCTCAGAGGACATGGAGGAGACTGTGAGATTCAGAGGGGA**  
**GGAGAAAACACTGTTTCCACCCCTCCTAGCTTCTCAACCAGGGCCCTGCAGATTAGACACGAAGGACAGGTTAACAAGAAAAGGC**  
**ATACTCATGATTTTATATAAGCGTTGCTTGCAGAAAGCAGCTGAGCTTGGTGTGTTTATGCTAGGTTTATGAGAGTGGGGTCCCTGAAAG**  
**GTGTGATAGGACGAGAGCATGAGCAAGCGCAGGAAACGGGCGGGGAGGGGCCCTGGCGAGGCCTGTCCATCAGGTCCCTCTTG**  
**GCCTCCCTCCATCTTTGAGCTAAGGATGCACCCTTCCCTGGTGTGGGGGCACCTCTACAAGAGGGTCACGACGATCTGCTTCAG**  
**GGGAGTTAGGCTTGAATGAGGGCTGAGATGTGGTAATAAGAAAGGCAGGAAGATCGTGTAAATGGCCAAGGGCGTGGGGCACCTGG**  
**GGAAACGCATTTCTCACACTGTCTCAGAGGAGGGTGTGCGTCTGGCGGGCGGGGACATGGGGTTCTCCTGTCTGCCTTGTGGT**  
**CTCCTTGACACTCTCCGAGCCGTGAACCTTAGCAGCCACGTGGACTCCACGCTGCTGCAGGTGGACCTGGCAGGGGCCCTAGTGGCC**  
**AGTGGGCCGAGTCGTCCTGGGAGGGGAAGAGCACATCGTGGTGGAGCTGACCCAGGCTGACGCTTTGGGCTCCAGGTGGCGGGCC**  
**ACAGCAG**GTTATGCTGATAATTTATTTGAAAGGAATTGTGAAATCTCATTTATCTCATAATAAATTGTGTTTCTTGAATGGATTGATGG  
GAGCCACAGGGCAGGACACCATTCTAGTCAATTCATTGGCTCTTCTTATCTTTGCTTTGCTTTGAAATCTCTTTTACCATTCTCTATT  
ATGAAACTCTTGCCATTTAGTAAATCAGGCTTGAAGTCTTTATTCAAACACTGTTTGTGTTGGCCATTGAGGAAACATACATTTGAAATG  
AAAAAACTTTTTTTTTTTTTTTGGTGTGCTTGAAGTGCAGTTAAGTCCGGGCCTACGTCAGGGTTCAGAAGACACTTCCACAGCACGGT  
ACTGTGTTTGGACATCTGCTAAGACAGTGCCTACGTTCTGAGCTGCTGTATCTGTGTAGGCCTGTGCTCAGCCTTGTCTGAGCCTGT  
CAAGCCTTCAATCACATCTGCAAAGGCAGCCAGGAAG**CCGGTGGCTGTCTAGGC**AGCCCAGAGCCTAGACTCCAGAGGCAGGGCTCC  
TTCCACGCAAGCCTGACTCCCCGAGCCCTGGCCTCGGAGACTCCACTGGATTATCACGGTACTTCTGCATGGAAGTCGTAGCCCTG  
GTTGATGGATGTTCCCTGTTATAACCAAGTAGAGGGAAACCCCGCCATGCACAGGCTGTGTACCTGCAGTGGAGCAT**CTGAAGAAATGAA**  
**TGCACCCTCC**

### Supplementary Figure 4.3. Primers flanking the 2.7 kb deletion in *OCA2*.

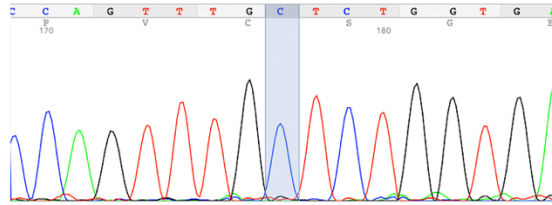
Genomic loci chr15:28017662-28021422 annotated with primer binding sites (yellow), exon 7 (blue) and the deleted sequence (red). Three primers were designed by the author and were used in genotyping the patients for the 2.7 kb deletion, are highlighted in yellow (Durham-Pierre et al., 1994). The additional primer in the deleted region is a control for confirmation of the amplicons generated by the primer set. The expected amplicon size for the wildtype allele is 3761 and 238 bp but for the deletion is 1077 bp.



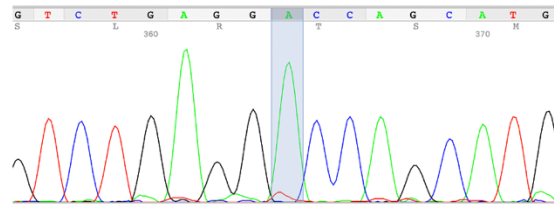


**Supplementary Figure 5.1 PCR amplification of variants in candidate genes in family F5.** All PCR products were discrete and single bands implying a homogenous amplicon. The product sizes are for 592 bp for *TADA3*, 564 bp for *MTMR14*, 641 bp for *CHRNA4*, 353 bp for *LOXL4* and 546 bp for *HELZ2*.

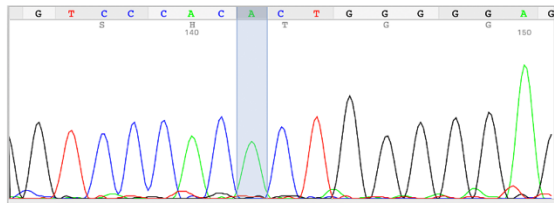
**LAMP1**  
c.1109delG



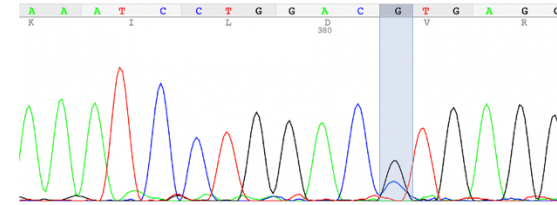
**TADA3**  
c.737C>A



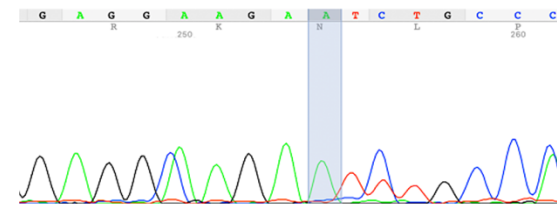
**LOXL4**  
c.1334G>A



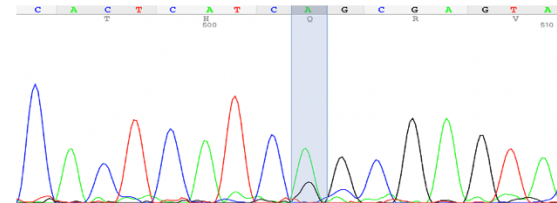
**HELZ2**  
c.706G>C



**MTMR14**  
c.1244G>A



**CHRNA4**  
c.919G>A



**Supplementary Figure 5.2 Assessing segregation of candidate variants in affected sibling F1287.** Electropherogram of six variants detected. The sibling is heterozygous for *HELZ2* and *CHRNA4* variants and homozygous for the remaining four variants in *LAMP1*, *TADA3*, *LOXL4* and *MTMR14*.

Off target recognition site	PAM	Mismatch Position	Genomic Location	Annotations
GACTCACCAGAGCCAGAA	AGG	12,18,20	chr6[917717]	intergenic
GCACTCACCAGAGCCACAC	AGG	2,12,18	chr10[117499933]	intron: EMX2OS
GACTCAGCAGAGCAAACC	AGG	9,16,19	chr20[43729504]	intergenic
CACACACACCAGAACCAAAC	CAG	1,6,14	chr21[36825450]	intron: HLCS
GACTCACCAGAACCAGAC	CAG	14,18	chr3[27018056]	intergenic
GACTTACCAGAGACAAGC	AGG	7,15,19	chr13[83268378]	intergenic
GGCACTCCCAGAGCCACAC	TGG	2,8,18	chr18[79170228]	intron: ATP9B
GACTGGCCGAGCCAAAC	CAG	7,8,11	chr6[42218018]	intergenic
GACTCATCAAAGTCAAAC	TGG	9,12,15	chr8[72552661]	intron: KCNB2
GAGACTCCCTAGAGCCAAAC	TAG	3,8,10	chr3[141069389]	intron: SPSB4
GACTCTACCAGAACCAGAC	CAG	4,14,18	chr2[41910305]	intergenic
TACTCTACCAAAGCCAAAC	AGA	1,4,12	chr2[134578325]	intron: TMEM163
GACAGTCCCAGAGCCAAGC	AAG	5,8,19	chr22[20152588]	intergenic
GAAACTAACAGAGCCTAAC	CAG	3,7,17	chr20[58963422]	intergenic
GACTCACCAGAGGCAGAC	TGG	15,18	chr12[103627215]	intron: STAB2
GACTCACCAGAGCCATTC	TGA	12,18,19	chr19[31940856]	intergenic
GACTCACCACAGACAGAC	AAG	12,15,18	chr11[123884083]	exon: TMEM225
GACCCTACCAGAACCCAAC	CAG	4,14,17	chr17[72355587]	intergenic
GACAATCAGCAGAGCCAGAC	TGA	5,9,18	chr8[142369517]	intron: TSNARE1
GACTAAACAGAGCCACAC	AGA	7,9,18	chr8[19418626]	intron: CSGALNACT1
GACTCATCAGAGCCAGGC	TGA	9,18,19	chr14[52063488]	intron: NID2
GACTCATCTGAGCCAAAT	GGA	9,11,20	chr12[108336771]	intron: CMKLR1
GACTCCCATACCCAAAC	TGA	8,12,14	chr7[33977137]	intron: BMPER
GACCCTACCAGTGCCTAAC	TAG	4,13,17	chr7[1863305]	intron: MAD1L1
GACTCTACCAGTGTCAAAC	CAG	4,13,15	chr22[17067265]	intergenic
GACTCACCAGGGGAAAC	AGA	13,15,16	chr7[17236764]	intergenic
GACTTAACCAGAGTCAAAC	AGA	4,7,15	chr5[125031363]	intergenic

**Supplementary Table 5.1 Off target effects predicted using *LAMP1* sgRNA1.**

27 off-target genomic edits which could potentially occur when using sgRNA1:

GACTCACCAGAGCCAAAC.

Off target recognition site	PAM	Mismatch Position	Genomic Location	Annotations
AAGGTGGAAGGCAACCAAGTT	CGG	12,13,14	chr12[1864710]	intron: CACNA2D4
CAGGTGGAAGGAGGCCAATT	AGG	1,12,18	chr8[22237667]	intergenic
CAGGTTGGAGGTGGCCAGTT	GGG	1,6,8	chr6[56153961]	intron: COL21A1
AAGGTTGAAGATGCCCAAGTT	AGG	6,11,14	chr17[14951592]	intergenic
AAGTTGCAATGTGGCCAGTT	TGG	4,7,10	chr11[46216717]	intergenic
AGGGTGAAGGTGGCTATTT	GGG	2,16,18	chr19[16122966]	intron: RAB8A
AAGGTGGAAGGAGGCCAACT	GGG	12,18,19	chr1[7231038]	intron: CAMTA1
TAGAGGGAAGGTGGCCAGTT	TAG	1,4,5	chr8[60703746]	intron: CHD7
AAAGTGAAAGGTGGCCAGTG	TGG	3,7,20	chr5[146850499]	intron: PPP2R2B
ATAATGGAAGGTGGCCAGTT	GAG	2,3,4	chr8[60469443]	intergenic
CAGGAGGAAGGTGGCCAGCT	GGG	1,5,19	chr6[33752286]	intergenic
CAGGTGGAAGGGGGCCAGCT	GGG	1,12,19	chrX[64432783]	intergenic
AAGGTAGAATGTGGCCAGTA	CAG	6,10,20	chr3[44997273]	intron: EXOSC7
AAGCTGGAAGTTGGCCAGTA	TGG	4,11,20	chr12[24800231]	intergenic
AAGGTGGTAGGCAGCCAGTT	AAG	8,12,13	chr18[75763650]	intergenic
AAGGTGTAAGGAGGCCAGTG	TGG	7,12,20	chr17[31663034]	intergenic
TAGTTGGAAGGTGGACAGTT	TGG	1,4,15	chr6[53009001]	intron: CLIK1
AAGGTGGCAGGTGGCCAGTG	GGG	8,20	chr2[132275747]	intergenic
AAGGTGCCAGGTGGCCAGGT	TGG	7,8,19	chr15[81094731]	intergenic
AAGGTGGAATGTAGCCAGTC	TGG	10,13,20	chr3[5988041]	intergenic
AAGGTGGATGCTGGCAAGTT	GAG	9,11,16	chr15[80903214]	intron: CEMIP
AAGTGGGAAGGTGGCCAGCT	GGG	4,5,19	chr7[28104124]	intron: JAZF1
AAGGTGGCAGGTGGCAAGTA	GAG	8,16,20	chr8[48553112]	intron: LOC101929268
AAGGTGGGAGGTGCCCAAGTG	GGG	8,14,20	chr15[76983341]	intergenic
AAGGTTGAAGATGGCCAGCT	GAG	6,11,19	chr2[181932512]	intergenic
AAGGAGAAAGGTGGCCAGCT	GAG	5,7,19	chrX[119396643]	intergenic
AAGGTGCAGGGTGGCAAGTT	GGA	7,9,16	chr11[21757808]	intergenic
AAGGTGCAGGGTGGCAAGTT	GGA	7,9,16	chr18[11432908]	intergenic
AAAGGGGAAGGTGGCCAGGT	AAG	3,5,19	chr21[25611597]	intergenic
GAGGTGGAGGGTGGCCAGTG	TAG	1,9,20	chr1[2569401]	intron: LOC100996583
AAGCTGGAAGAGGCCAGTT	AGA	4,10,12	chr16[77097681]	intergenic
AATGTGGAAGGCGGACAGTT	CAG	3,12,15	chr16[86540020]	intron: MTHFSD
AAGCTGGAAGGTGTCCAGTG	TGG	4,14,20	chr20[32588019]	intron: NOL4L-DT
AAAGTGGAAGCTGGCCAGAT	CAG	3,11,19	chr6[107863942]	intergenic
CGGGTGAAGGTGCGCCAGTT	GGA	1,2,13	chr2[20666364]	intergenic
AGGGTGAAGGTGGCAAGCT	GGA	2,16,19	chr4[102614553]	intron: NFKB1
AAGGTGGAAGGTGCCCTGGT	GGG	14,17,19	chr7[102682306]	intron: RASA4DP
AAGGTGGAAGGTGCCCTGGT	GGG	14,17,19	chr7[102583245]	intron: RASA4
AAGGTGGAAGGTGCCCTGGT	GGG	14,17,19	chr7[44029808]	intron: RASA4CP
AAGGTGGAAGGTGCCCTGGT	GGG	14,17,19	chr7[102484066]	intron: RASA4B
AAGATGGAAGGAGGCCAGAT	GGA	4,12,19	chr1[68744238]	intergenic
AAGGTGGGAGGTGGTAAGTT	CAG	8,15,16	chr4[92354879]	intron: GRID2
AAGGTGTAAGATTGCCAGTT	AGA	7,11,13	chr11[98507274]	intergenic
CAGGGGGAAGGTGGCCAGAT	GGA	1,5,19	chr16[85970542]	intergenic
GAGGGGGAAGGTGGCCAGGT	GGA	1,5,19	chr19[41089125]	intron: CYP2A13
GAGGGGGAAGGTGGCCAGGT	GGA	1,5,19	chr19[40849784]	intron: CYP2A6
AAGGTGAAAGGTGCCCAAGT	AGA	7,14,19	chr1[229912161]	intergenic
AAGGTGGAAGGGATCCAGTT	TGA	12,13,14	chrX[12765451]	intergenic
AAGGTGGAAGGGATCCAGTT	TGA	12,13,14	chrX[12670555]	intron: FRMPD4
AAGGTGGATGGTGGCTATTT	GGA	9,16,18	chr7[151617460]	intron: PRKAG2

AAGGTGGATGGTGGCTATTT	GGA	9,16,18	chr7[151617460]	intron: PRKAG2
AAGGAGGAAGGTGGTAAGTT	CAG	5,15,16	chr1[35861052]	intergenic
AAGGGGCAAGGTGGACAGTT	GGA	5,7,15	chr1[48644717]	intron: AGBL4
AAGGTGGAAGGTGCCAAAT	TGA	14,18,19	chr2[106202577]	intergenic
AAAGTGAAGGTGGGCATTT	CAG	3,15,18	chr10[355055]	intron: DIP2C
AAGGTGGAAGGTTGCCTGTA	GAG	13,17,20	chr1[163592280]	intergenic
AAGGTGGAAACTGGCCTGTT	TGA	10,11,17	chr1[220426098]	intergenic
AAGGTGGGAGGTGGTCAGCT	GAG	8,15,19	chr18[47309775]	intron: MIR4527HG
AATGTGGAAGGTGGCCAGGG	AGA	3,19,20	chr8[47365218]	intron: SPIDR
GAGGTGGAGGGTGGGCAGTT	GGA	1,9,15	chr10[14319795]	intron: FRMD4A
AAGCTGGAAGGTGGGCAGAT	AAG	4,15,19	chr14[93626275]	intron: UNC79
AAGGTGGAAGGTGGGAAGTG	GAG	15,16,20	chr4[165379671]	intron: CPE
AAGGAGCAAGGTGGGCAGTT	AGA	5,7,15	chr6[169942738]	intergenic
AAGGGGCAAGGTGGTCAGTT	GGA	5,8,15	chr10[19711967]	intron: LOC101928834

### Supplementary Table 5.2 Potential off targets using *LAMP1* sgRNA2. 63

potential off-target genomic edits predicted to occur when using sgRNA2:

AAGGTGGAAGGTGGCCAGTT.

ID	Sequence (5'→3')	Length	Tm (°C)	GC (%)	Self complementarity	Self 3' complementarity	Product size
<i>LAMP1</i> _1F	CTTTC AAGGTGGAAGGTGGC	20	59.04	55.00	5.00	2.00	Genomic: 600 bp
<i>LAMP1</i> _1R	GTAGGCGATGAGGACGATGA	20	59.05	55.00	2.00	1.00	cDNA: 128 bp
<i>LAMP2</i> _1F	CTGAAGGAAGTGAACATCAGCA	22	58.59	45.45	4.00	0.00	Genomic: 812 bp
<i>LAMP2</i> _1R	GCACATATAAGAACTTCCCAGGG	23	58.86	47.83	5.00	5.00	cDNA: 108 bp
<i>TP53</i> _1F	GTACTCCCCTGCCCTCAACA	20	61.20	60.00	4.00	0.00	Genomic: 1057 bp
<i>TP53</i> _1R	CTGGAGTCTTCCAGTGTGAT	20	56.56	50.00	7.00	2.00	cDNA: 408 bp

### Supplementary Table 5.3 Primer sets for *LAMP1*, *LAMP2* and *TP53*.

Oligonucleotide probes designed to flank exonic sequences of the target gene to generate distinctive amplicons based on the template DNA used. Genomic DNA results in a larger amplicon due to the presence of intronic sequences in the amplified product.

**>LAMP1 control midigene sequence**

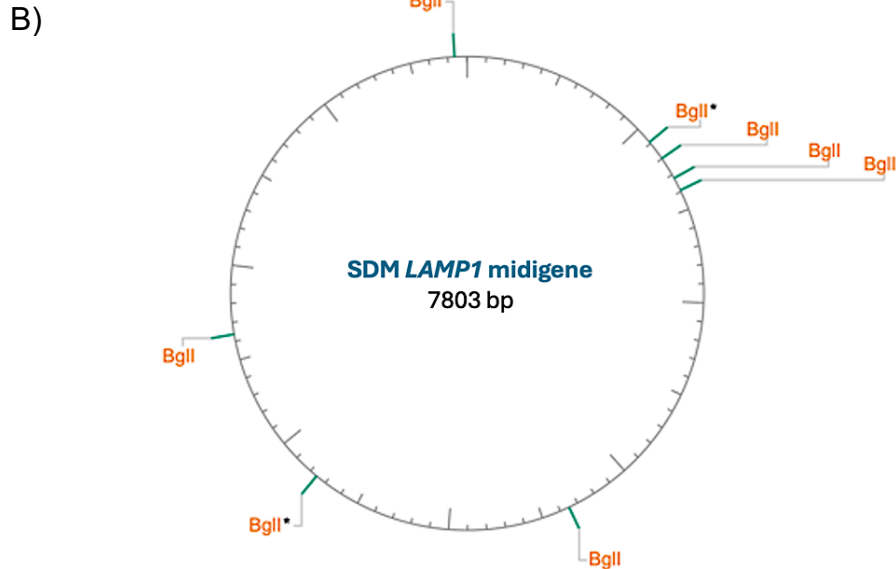
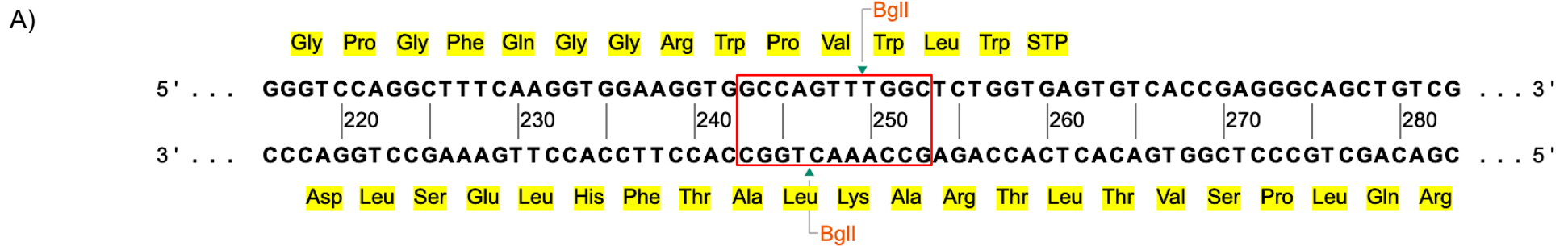
gatccgcgatggtgactctcagtacaatctgctctgatgccgatagttaagccagccccgacacccgcca  
acacccgctgacgcgccctgacgggctgtctgctcccggcatccgcttacagacaagctgacccgtctccg  
ggagctgcatgtcagaggtttcaccgtcatcccgaaacgcgagacgaaagggcctcgtgatacggc  
tattttataggtaatgcatgataataatggttcttagacgtcaggtggcacttttcggggaaatgtgcgcgaa  
cccctattgtttattttctaaatacattcaaataatgtatccgctcatgagacaataaccctgataaatgctcaata  
atattgaaaaaggaagagtatgagtattcaacatttccgtgtcgccctattccctttttgcggcattttgccttctg  
ttttgctcaccagaaaacgctggtgaaagtaaaagatgctgaagatcagttgggtgacagagtggttacatc  
gaactggatctcaacagcggtaagatccttgagagttttcgccccgaagaacgtttccaatgatgagcactttt  
aaagtctgctatgtggcggttattatcccgtattgacgccccggaagagcaactcggctcgccgatacacta  
ttctcagaatgacttggtgagtaactcaccagtcacagaaaagcatcttacggatggcatgacagtaagagaat  
tatgcagtgtgccataaccatgagtgataaactgcgccaacttacttctgacaacgatcggaggaccgaa  
ggagctaaccgctttttgcacaacatgggggatcatgtaactgccttgatcgttgggaaccggagctgaatg  
aagccataccaaacgacgagcgtgacaccacgatgcctgtagcaatggcaacaacgttgcgcaaactatta  
actggcgaactacttacttagcttcccggcaacaattaatagactggatggaggcggataaagttgcaggac  
cacttctgcgctcggccctccggctggctggtttattgctgataaatctggagccggtgagcgtgggtctcggg  
tatcattgcagcactggggccagatggtaagccctcccgtatcgtagtattctacacgacggggagtcaggca  
actatggatgaacgaaatagacagatcgtgagataggtgcctcactgattaagcattggtaactgtcagacc  
aagttactcatatatactttagattgattaaaacttcatttttaattaaaaggatctaggatgaagatccttttgataa  
tctcatgacaaaaatcccctaactgagttttcgttccactgagcgtcagacccccgtagaaaagatcaaaggat  
cttctgagatcctttttctgcgctaactgctgctgcaaaaaaaaccaccgctaccagcgggtggtttgttt  
gccggatcaagagctaccaactcttttccgaaggtaactggcttcagcagagcgcagataccaaatactgttc  
ttctagttagccgtagttaggccaccactcaagaactctgtagcaccgctacatacctcgtctgtaaatcct  
gttaccagtggtgctgccagtgccgataagtcgtgtcttaccgggttgactcaagacgatagttaccggata  
aggcgcagcggctcgggctgaacggggggtcgtgcacacagcccagcttgagcgaacgacctacaccg  
aactgagatacctacagcgtgagctatgagaaagcgcacgctcccgaaggagaaaggcggacaggt  
atccggtaagcggcagggctcggaacaggagagcgcagaggagctccaggggaaacgcctggtatc  
ttatagtcctgtcgggttcgccacctctgacttgagcgtcgatttttgatgctcgtcaggggggaggcctat  
ggaaaaacgccagcaacgccccttttacgggtcctggccttttgctggcctttgctcacatggctcgacagat  
ctcaatattggccattagccatatttcttattggttatatagcataaatcaatattggctattggccattgcatacgtt  
gtatctatatcataatatgtacatttatattggctcatgtccaatgatgcccatgttggcattgattattgactagta  
ttaatagtaatcaattacggggcattagttcatagccatataatggagttccggttacataactacggtaaatg  
gcccgcctggctgaccgccaacgacccccgccattgacgtcaataatgacgtatgttccatagtaacgcc  
aatagggactttccattgacgtcaatgggtggagtattacggtaaaactgccacttggcagttacatcaagtgtat  
catatgccaagtccgccccctattgacgtcaatgacggtaaatggcccgcctggcattatgccagttacatgac  
cttacgggactttcctacttggcagttacatctacgtattagtcacgtattaccatgggtgatcgggtttggcagta  
accaatggcggtggatagcgggttgactcaggggatttcaagtctccacccattgacgtcaatgggagttg  
tttggcaccaaaatcaacgggactttccaaaatgtcgtacaactgcgatcgcccggccggtgacgcaaat  
ggcggttagcgtgtacgggtgggaggtctataaagcagagctcgttagtgaaccgtcagatcactagaagc  
tttattgcggtagtttatcacagttaaattgtaacgcagtcagtgcttctgacacaacagttcgaacttaagctg  
cagtgactcttaaggtagccttcagaagttggtcgtgaggcactgggaggttaagatcaaggttacaaga  
caggttaaggagaccaatagaaactgggctgtcgagacagagaagactcttgcgtttctgataggcacctat  
tggtcttactgacatccactttgcctttctccacaggtgtccactcccagttcaattacagctcttaaggctagagt

acttaatacgcactactataggctagcctcgagaattccggagggtcaacaacgagctctttgtcatctacatgttc  
gtggccacttcacccatccccatgattatcatcttttctgctatgggcagctcgtcttcaccgtcaaggaggtagcg  
gccgggggggtggcgccctcacggctctgagggccagccccagcatgcatctgcggctctctccctgg  
aggagccatatacaagttgtacaaaaagcaggctagatgcagcagatgatgtggcggggctgctgtgc  
ccacagtggagtggtgcaggggagggcatgcaggccgtgcggcctctggcttcagatgctgctgcctgt  
ggttccgtgggtgcatctctgcctggaggactcctgtgcagttagcaacataagacatgaagcatataattgtc  
actgtcagctcttttaaatcgctgtgcaaatgaattaacagttcagtttctataatttagctttccaaatcatgcttc  
ctgtgctgtgatagctcctgcagctccccgtttccagagactgactctcaagaggctggaaggaccagcctgg  
gcagcacagggaggctccatttctgcaataataaaacgagttagctggcgtagtggcgcacacctgtggt  
cccagctactgggaggctgaggtgggaggatcactgagcccaggagtaaggtgcatgagccgtgatc  
actccactgcactccagcctgggtgacagagcaagactctcagagaagggtctggaaggaactaaccaaa  
atcttattccccagagatacacaacgcctttatagacaagatctttgcagttttgttctaataactagaaatgtagg  
aagtcaggttattaccaatgaccattcacgtttgatataaatgtgtgaatctactggggttaaagatcatctttct  
atgaatctgctccgtgattaaagatcattttcttttaagaatgcaagttctagccggttttctacaaggaatcca  
gttgaatacaattctcctgacgccagagggtgagaaccacaatctctgcgagccccgccccgcccgcg  
cccaggtattctggagccactagacctctgtgtgtgagaccctgcctttaaagctgccaacggctccctg  
cgagcgtgcagggcacagctggcaattcctacaagtgcaacgcggaggagcacgtccgtgtcacgaagg  
cgtttcagtcaatataattcaaagtggtgggtccaggcttcaaggtggaagggtggccagttggctctggtgagtg  
caccgagggcagctgtcgcgggggtgtgaggacgtgcttcagactccgcctgtggacgttttagctcctccgtg  
tggtctggggcgacgccccgttctctgcaaggagctgtttctctgcccgtctgagattctagaggaactcc  
ccctgcttagagaggcccagcgtgtctctcagctgggagccccgttaccattgagagtaagggaatcatttt  
aagaaacagtggtggcgcttctccatgaacgttgaatacaacactgtcatctgatgtcacagaggccctgg  
acgcagcacagctgtccggccacagcccctgattccagacgggggagagacgtgtgtggttctcgcggctg  
cagagagcaccagtctctgcagcggctcccaagtgacaattcaccagcttctgtacaagtgggatga  
gaggtacctccgaggggtaaacagttgggtaaacagctctgaaagtcagctctgccattttctagctgtatggcc  
ctgggcaagtcaattccttctctgtgctttggtttctcatccatagaaaggtagaaagggcaaacaccaact  
cttgattacaagagataattacagaacaccctggcacacagagggccacatgaaatgtcacgggtgaca  
cagccccctgtgctcagctccctggcatctctaggggtgaggagcgtctgcctagcaggtcccaccaggaag  
ctggatttgatggatggggcgctggaatcgtgaggggcagaagcaggcaaaagggtcggggcgaacctca  
ctaactgtccagttccaagcacactgtgggcagccctggccctgactcaagcctctgcctccagttccggaa  
ctgcatgtcaccaccatctgctgcggcaagaaccactgggtgacgatgaggcctctgctaccgtgtccaag  
acggagacgagccaggtggccccggcctaagacctgcttaggactctgtggccgactataggcgtctccat  
cccctacacctgtcgaccggggcgccgcttcccttagtgagggtaatgcttcgagcagacatgataagata  
cattgatgagttggacaaaccacaactagaatgcagtgaaaaaatgctttattgtgaaattgtgatgctattg  
ctttattgtaaccattataagctgcaataaacaagttaacaacaacaattgcattcattttatgtttcaggtcaggg  
ggagatgtgggaggtttttaaagcaagtaaacctctacaaaatgttgtaaaatccgataaggatcgatccgg  
gctggcgtaatagcgaagaggcccgcaccgatgccctccaacagttgctgcagcctgaatggcgaatgg  
acgcgccctgtagcggcgcaatagcgcggcggtgtggtggttacgcgcagcgtgaccgctacacttgcca  
gcgccctagcgcggctccttctgcttctcccttctctcgcacgctgcgggcttccccgtcaagctctaaa  
tcgggggctcccttaggggtccgatttagtcttacggcacctcgacccccaaaaaactgattaggggtgatggt  
cacgtagtgggcatcgccctgatagacggttttcggccttgacgttgaggtccacgttcttaatagtgactct  
tgttccaaactggaacaacactcaacctatctcggctattctttgattataagggattttgccgatttcggcctat  
tggttaaaaaatgagctgatttaacaaaaatfaacgcgaatttaacaaaatfaacgcttacaattcctgatg  
cggattttctccttacgcatctgtgcggtatttcacaccgcatacgcggatctgcgcagcaccatggcctgaaat  
aacctctgaaagaggaactgggttaggtacctctgaggcggaagaaccagctgtggaatgtgtgctcagttta

gggtgtggaaagtccccaggctccccagcaggcagaagtatgcaaagcatgcatctcaattagtcagcaac  
cagggtgtggaaagtccccaggctccccagcaggcagaagtatgcaaagcatgcatctcaattagtcagcaa  
ccatagtcgcccctaaactccgcccattcccggcccctaaactccgcccagttccgcccatttccgcccattggc  
tgactaattttttatattatgagagggccgaggccgctcggcctctgagctattccagaagtagtgaggaggctt  
tttgaggcctaggcttttgcaaaaagctgattctctgacacaacagctctcgaacttaaggctagagccacca  
tgattgaacaagatggattgcacgcaggttctccggccgctgggtggagaggctattcggctatgactgggca  
caacagacaatcggctgctctgatgccgccgtgtccggctgacagcagggggcgcccgggtctttttgcaag  
accgacctgtccgggtgccctgaatgaactgcaggacgaggcagcgcggctatcgtggctggccacgacgg  
gcttccttgccagctgtgctcagcttgcactgaagcgggaaggactggctgctattgggcaagtgcg  
ggcgaggatctcctgtcatctcacctgtcctgccgagaaagtatccatcatggctgatgcaatgcggcggct  
gcatacgcttgatccggctacctgccattcgaccaccaagcgaacatcgcatcgagcagcagcactcgt  
gatggaagccggctctgtcagatcaggatgatctggacgaagagcatcaggggctcgcgccagccgaactgtt  
cgccaggctcaaggcgcgcatgcccagcggcagggatctcgtcgtgacctatggcgtgctgcttgcgga  
atatcatgggtggaaaatggccgctttctggattcatcgactgtggccggctgggtgtggcggaccgctatcagg  
acatagcgttggtacctgctgattgctgaagagctggcggcgaatgggctgaccgcttctcgtgctttacg  
gtatcgccgctcccattcgcagcgcacgcttctatcgcttcttgacgagttcttctgagcgggactctggggt  
tcgaaatgaccgaccaagcagcgcaccaacctgccatcacgatggccgcaataaaatatctttattttcattaca  
tctgtgtgtggtttttgtgtgaatcgatagcgataag

**Supplementary Figure 5.3 *LAMP1* midigene reference sequence.** The sequence of the *LAMP1* control midigene used as a reference for read alignment is of 7803 bp.





C)

#	Ends	Coordinates	Length (bp)
1	BglI - BglI	5638-7736	2099
2	BglI - BglI	1393-3350	1958
3	BglI - BglI	3351-4759	1409
4	BglI - BglI	7737-1092	1159
5	BglI - BglI	4760-5637	878
6	BglI - BglI	1200-1321	122
7	BglI - BglI	1093-1199	107
8	BglI - BglI	1322-1392	71

**Supplementary Figure 5.4 RFLP strategy for genotyping *LAMP1* SDM clones.** A) The BglI restriction sites highlighted in red overlaps the variant site in *LAMP1*. B) Restriction map of the SDM *LAMP1* midigene annotated with BglI restriction sites. C) The BglI is expected to cleave the plasmid at eight different positions, resulting in eight discrete fragments ranging from 71 – 2099 bp.

Plasmid	Unique sequence	Component
PX459	GGCGATAGCCTGCACGAGCACATTGCCAATCTGGCCGGCAGCCCCGCCATTAAGAAGGGC	Cas9
pDONR221	TCAGAATTGGTTAATTGGTTGTAACACTGGCAGAGC	Backbone sequence
pcDNA3.1	CATCATCACCATCACCAT	6X His affinity tag
pCMV6_entry_LGR4	GATTACAAGGATGACGACGATAAG	FLAG Tag
pCTAP_LGR4	TGGAGCCACCCTCAGTTCGAGAAG	Strep tag II
pDEST504	CGCACCATCTTCTTCAAGGACGACGGCAACTACAAGACCCGCGCCGAGGTGAAGTTCGAGG	eYFP
LAMP1+ midigene	CCCGCCCCGCCCCGCGCGCCAGGGTATTCTGGAGCCACTAGACCTCTGTGTGTGTTGCAGA	LAMP1 gene
pKH3	TACCCATACGATGTTCCAGATTACGCT	Human influenza hemagglutinin tag
pRK5_mFzd4	GACCAAGATGCCCAACTTAGTGGGACACGAGCTGCAGACAGACGCCGAGCTGCAGCTGACAACTTTCACG	Backbone sequence
pDESTneo	.1TAGCTCCTGAAAATCTCGATAACTCAAAAAATACGCCCGTAGTGATCTTATTTTCATTATGGTGAAAGTTGGAACCTCTTACGTGCCGATCA	Cat promoter

**Supplementary Table 5.4 Plasmids unique features.** The sequence belonging to these unique features are only present in a single respective plasmid. These unique sequences are obtained via SnapGene Viewer v6.2.1 and were used in extracting reads belonging to a single plasmid.

## Run summary

### DATA OUTPUT



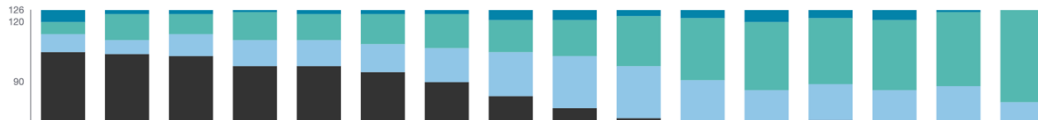
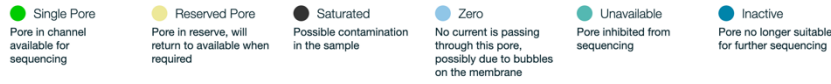
### RUN DURATION



### PORE SCAN

A Pore scan is performed at configurable time intervals to determine the current status of pores within channels on a Flow Cell. For this run a Pore scan is performed every 1.5 hrs.

#### Legend



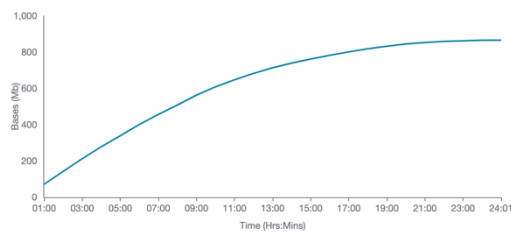
### CUMULATIVE OUTPUT

The cumulative output shows the total amount of bases or reads sequenced over time by your device.

#### Bases

##### Legend

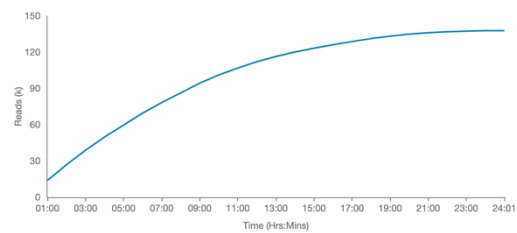
— Estimated  
Predicted total number of bases, prior to basecalling



#### Reads

##### Legend

— Total  
Total number of reads, including passed, failed and skipped.

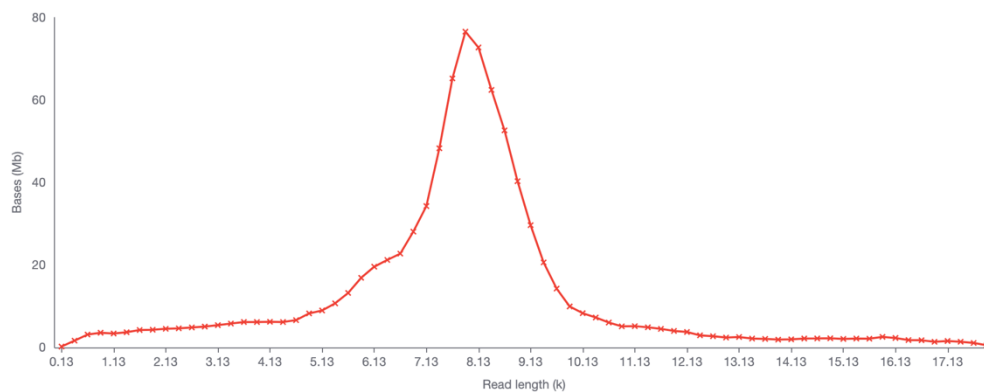


### READ LENGTHS · OUTLIERS REMOVED

The read length graph shows the total number of bases vs the read length. The longest 1% of strands are classified as outliers, and excluded to allow focus on the main body of data.

#### Legend

■ Basecalled    ×× Estimated



**Supplementary Figure 5.5 MinION run report for multiplexed sequencing of 10 plasmids.** The sequencing performance for the trial of multiplexed sequencing shows >60 pores available for sequencing which resulted in 137.85K reads generated over 24 hours.