

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]



UNIVERSITY OF SOUTHAMPTON

FACULTY OF MEDICINE

HUMAN DEVELOPMENT & HEALTH

**The identification of causal variants in nystagmus and
primary open-angle glaucoma patients through
analyses of next-generation sequencing data**

Author:

Luke O’Gorman

Supervisory Team:

Prof Sarah Ennis, Dr Jane Gibson, Ms Angela Cree,

Mr Jay Self and Prof Andrew Lotery

July 2019

Abstract

UNIVERSITY OF SOUTHAMPTON

Faculty of Medicine

Human Development & Health

Doctor of Philosophy

The identification of causal variants in nystagmus and primary open-angle glaucoma patients through analyses of next-generation sequencing data

by

Luke O’Gorman

Background

As next-generation sequencing (NGS) becomes more feasible and accessible, its application in medical genetics becomes increasingly prevalent. Ophthalmic diseases including primary open-angle glaucoma (POAG), oculocutaneous albinism and nystagmus have a strong underlying genetic basis. Understanding the genetic basis is therefore important to facilitate early diagnoses, and aid clinicians in directing targetted treatment to provide best outcome for patients. The molecular basis remains undetermined for many patients. However, high-throughput sequencing and development of bioinformatic tools to process such data has enabled a robust capacity to identify likely causal variants in disease.

Aims

By analysing targetted and whole-exome sequencing (WES) data on patients individu-

ally and across cohorts, this thesis aims to interrogate candidate genes which are known or predicted to be causal in POAG, albinism and nystagmus. From these analyses, much needed further insight into causal variants underpinning these ophthalmic diseases can be provided and, importantly, a molecular diagnosis for the patients involved in this work could be ascertained.

Outcomes

The work presented demonstrate herein the importance of *TYR* tri-allelic genotype in oculocutaneous albinism. With the realisation of its importance in molecular diagnosis of oculocutaneous albinism, it was demonstrated how the tri-allelic *TYR* genotype could be successfully incorporated into a diagnostic workflow. Furthermore, we were also able to demonstrate that by employing a similar workflow across consanguineous Pakistani ophthalmic disease patients, comparably high diagnostic yields were returned.

This thesis confirms that 3.07% of British patients with POAG had likely causal variants in the coding region of the most common causal POAG gene, myocilin (*MYOC*). Therefore, genetic testing for patients in *MYOC* coding sequences (namely exon 3) remains the most efficient plan of action when screening this gene. Other known Mendelian-like genes in POAG were able to further account for POAG in 1.17% of the UK POAG cohort. Functional assessment and larger samples sizes are required to test new hypotheses implicating molecular pathologies and other genes as Mendelian-like.

Conclusion

This thesis has expanded on the known molecular genetics underpinning ophthalmic diseases. Importantly, this work has informed clinicians and geneticists on diagnostic workflows and the variants which should be considered in order to achieve successful molecular diagnoses in patients. Crucially, for many of the cases detailed in this work, a molecular diagnosis has successfully been made which will ultimately help ensure the

best possible care for the patients involved.

Contents

Abstract	ii
Contents	xiii
List of Figures	xvi
List of Tables	xxi
Acknowledgements	xxii
Ethics Approval	xxiii
Funders	xxiv
Declaration of Authorship	xxv
Publications	xxvi
Abbreviations	xxviii
1 Background	1
1.1 Prevalence of ophthalmic diseases	1
1.2 Nystagmus	4
1.2.1 Description	4
1.2.2 Diagnosis	5
1.2.3 Genetic cause	5
1.2.4 Treatment	9
1.3 Primary open-angle glaucoma	10
1.3.1 Description	10
1.3.2 Diagnosis of POAG	12

CONTENTS

1.3.3	Genetic causes of POAG	13
1.3.4	Myocilin	14
1.3.5	Treatment	15
1.4	Methods of genetic analysis	16
1.4.1	Genetic linkage	16
1.4.2	Genome wide association studies	18
1.4.3	DNA sequencing	20
1.4.3.1	Sanger sequencing	20
1.4.3.2	Next (second) generation sequencing technology	22
1.4.3.3	Next (third) generation sequencing technology	25
1.4.3.4	NGS pipelines	26
1.4.3.5	Whole exome sequencing	28
1.4.3.6	Whole genome sequencing	29
1.4.3.7	Somatic mutations	30
1.4.3.8	Custom targeted sequencing	30
1.5	Aims	33
2	Targeted sequencing and analysis of the myocilin gene (<i>MYOC</i>) in a selected cohort of primary open-angle glaucoma patients	35
2.1	Synopsis	35
2.2	Background	35
2.3	Aim	38
2.4	Methods	39
2.4.1	Patient selection	39
2.4.2	Target selection	39
2.4.3	Quality control on kit design	40
2.4.4	Targeted sequencing	41

CONTENTS

2.4.5	Bioinformatic pipeline	41
2.4.6	Quality control of NGS data	42
2.4.7	Variant contextualisation	43
2.4.8	Prioritisation of variants	44
2.5	Results	45
2.5.1	Patient clinical traits	45
2.5.2	Quality control	47
2.5.2.1	Quality control on kit design	47
2.5.2.2	Quality control of NGS data	47
2.5.3	<i>MYOC</i> variant contextualisation	51
2.5.4	Exonic variants	53
2.5.5	Non-coding variants	57
2.6	Discussion	59
3	Next-generation sequencing analysis of 66 POAG genes across a selected cohort of severe primary open-angle glaucoma patients	65
3.1	Synopsis	65
3.2	Background	65
3.2.1	Mendelian-like genes	66
3.2.2	POAG genes identified through GWAS	69
3.2.3	Paucity of associations in POAG GWAS	70
3.2.4	Database of disease-associated variants	71
3.2.5	NGS to identify causal genes and variants	72
3.2.6	Aim	72
3.3	Methods	72
3.3.1	Patient selection	72
3.3.2	Gene selection	73

CONTENTS

3.3.3	Target gene capture and sequencing	73
3.3.4	Bioinformatic pipeline and filtering parameters	73
3.3.5	Filtering of variants	73
3.3.6	Interacting genes	75
3.3.7	Whole-gene pathogenicity	76
3.4	Results	77
3.4.1	Patient clinical traits	77
3.4.2	Gene selection for customised panel	78
3.4.3	Capture kit selection	78
3.4.4	Quality control	82
3.4.5	Filtering of variants	82
3.4.5.1	Standard filtering (Mendelian-like genes, n=7)	83
3.4.5.2	Stringent filtering of complex POAG genes (coding genes, n=56)	86
3.4.5.3	Filtering of non-coding genes (ncRNA, n=3)	88
3.4.6	Interacting proteins	88
3.4.6.1	Indirect protein-protein interactions	88
3.4.6.2	Direct protein-protein interactions	89
3.4.7	Whole gene pathogenicity scores	90
3.4.7.1	Batch effect	91
3.4.7.2	Overview of whole gene pathogenicity scores across the POAG cohort	92
3.4.7.3	Comparing POAG whole gene pathogenicity scores against a control cohort	96
3.4.8	Copy Number Variants (CNVs)	100
3.5	Discussion	102

CONTENTS

3.5.1	The seven Mendelian-like genes	102
3.5.2	Assumed complex genes	105
3.5.3	Limitations	110
3.6	Conclusion	111
4	Nystagmus whole exome analysis	113
4.1	Synopsis	113
4.2	Background	113
4.3	Aim	114
4.4	Methods	115
4.4.1	Patient selection	115
4.4.2	Sequencing and bioinformatic pipeline	116
4.4.3	Quality control	118
4.4.4	Variant prioritisation	119
4.5	Results	121
4.5.1	Quality control	121
4.5.2	Causal variant analysis	123
4.5.3	Patients with verified likely causal variants	124
4.5.4	Patients undergoing follow-up	126
4.5.5	Patients without likely causal variants	126
4.6	Conclusion	128
5	Identification of a functionally significant tri-allelic genotype in the tyrosinase gene causing hypomorphic oculocutaneous albinism (OCA1B)	129
5.1	Synopsis	129
5.2	Background	129
5.3	Aim	133

CONTENTS

5.4	Methods	133
5.4.1	Patients	133
5.4.2	Sequencing and bioinformatic pipeline	133
5.4.3	Quality control of NGS data	134
5.4.4	Identification of candidate causal variants	135
5.5	Results	135
5.5.1	Quality control	135
5.5.2	Candidate causal variant identification	138
5.5.3	Segregation analysis	140
5.6	Discussion	141
6	The utility of a sequencing panel for clinical use in patients with nys- tagmus and albinism	145
6.1	Synopsis	145
6.2	Background	145
6.2.1	Infantile nystagmus syndrome and albinism	145
6.2.2	Genetic basis	146
6.2.3	Phenotype of INS and albinism	147
6.2.4	Diagnostic gene panels	148
6.2.5	Variant interpretation	149
6.3	Aim	149
6.4	Methods	150
6.4.1	Patients	150
6.4.2	Phenotyping	150
6.4.3	Grouping by phenotype criteria	151
6.4.4	Library preparation and NGS	152
6.4.5	UKGTN gene panel	153

CONTENTS

6.4.6	Bioinformatic pipeline	155
6.4.7	Quality control	155
6.4.8	Variant prioritisation	156
6.4.9	Sanger verification	157
6.5	Results	158
6.5.1	Quality control	158
6.5.1.1	Sequence quality	158
6.5.1.2	Coverage, contamination and variant concordance . . .	158
6.5.1.3	Filtering of the patient cohort	160
6.5.2	Causal variant analysis	163
6.5.2.1	Assumed pathogenic variants	163
6.5.2.2	Assumed pathogenic and assumed likely pathogenic variants	167
6.5.2.3	PAX6 variant verification	170
6.5.2.4	TYR tri-allelic genotypic cause of albinism	175
6.5.2.5	Albinism patients with partially resolved genetic aetiology	177
6.5.2.6	Overview of diagnostic results	179
6.6	Discussion	179
6.7	Conclusion	184
7	Determining causal genetic variants for ocular disease in a consanguineous Pakistani cohort	185
7.1	Synopsis	185
7.2	Ophthalmic diseases	185
7.2.1	Cataracts	186
7.2.2	Waardenburg syndrome	186
7.2.3	Microcornea	187

CONTENTS

7.2.4	Joubert syndrome	188
7.2.5	Usher syndrome	188
7.3	Population genetics	189
7.3.1	Consanguinity	190
7.3.2	Ophthalmic disease in consanguineous pedigrees	191
7.4	Aim	193
7.5	Methods	193
7.5.1	Patient selection	193
7.5.2	Sample processing	195
7.5.3	Bioinformatic pipeline and quality control	195
7.5.4	Allocation of candidate gene lists	196
7.5.5	Variant prioritisation	198
7.6	Results	199
7.6.1	Quality control	199
7.6.1.1	Sequence	199
7.6.1.2	Coverage	199
7.6.1.3	Contamination	201
7.6.1.4	Quality control summary	203
7.6.2	Causal variant analysis	203
7.6.2.1	Runs of homozygosity	203
7.6.2.2	Albinism and nystagmus	204
7.6.2.3	<i>TYR</i> tri-allelic genotypes	206
7.6.2.4	Congenital cataract	208
7.6.2.5	Waardenburg syndrome and Usher syndrome	210
7.6.2.6	Eye malformations	212
7.6.2.7	Potential diagnostic yield	214

CONTENTS

7.7	Discussion	215
7.8	Conclusion	220
8	Conclusions	221
9	Bibliography	225
A	Appendix Supplementary Data	282
B	Appendix Supplementary Data	291
C	Appendix Supplementary Data	318
D	Appendix Supplementary Data	319
E	Appendix Supplementary Data	321
F	Appendix Supplementary Data	331

List of Figures

1.1	Causes of visual impairment and blindness in England and Wales. . .	2
1.2	Classification of childhood nystagmus.	5
1.3	Diagram of the anatomy underlying POAG and normal aqueous flow of the eye.	10
1.4	Example of a pedigree with every affected member bearing the disease causing allele.	17
1.5	Illumina sequencing process.	22
1.6	Illumina multiplexing process.	24
1.7	Next generation WES sequencing method.	28
1.8	Read depth example.	31
1.9	Illumina targeted sequencing library preparation methods.	32
2.1	Protein structure of myocilin.	37
2.2	Flow chart of the steps involved in the bioinformatic pipeline.	42
2.3	Comparison of sub-phenotypic clinical characteristics of POAG.	46
2.4	All pairwise comparisons between samples.	49
2.5	Per base analysis of variants and regions across the myocilin gene. . .	52
2.6	Human Myocilin protein structure with its domains and variants mapped.	56
3.1	Human OPTN protein structure and interacting binding interfaces. . .	67
3.2	Filtering methods used for three groups of genes.	75
3.3	Protein-protein interaction network of the 66 genes selected for targeted sequencing.	89
3.4	Crystal structure visualisation of the variant in context with the inter- acting TBK1 protein.	90

LIST OF FIGURES

3.5	Principle component analysis of GenePy scores for each sample grouped by batch number.	91
3.6	65 POAG gene GenePy scores across the POAG cohort (n=358). . . .	93
3.7	IGV image of sample GB194 for the five variant cluster in the IL1B gene.	95
3.8	IGV image of sample QG195 for variant rs771362103 in the FAM27E5 gene.	96
3.9	IGV image for POAG sample GL189 of ABO variant n.207G>A (chr9:133259827).	99
3.10	IGV image of sample QG032 for SIX6 variant, c.C182T;p.A61V. . . .	100
4.1	Pedigrees for the 9 unrelated nystagmus patients with their sample ID below.	116
4.2	Overview of the stages involved in identifying the causal variants. . . .	117
5.1	Depiction of the tri-allelic causal genotype.	132
5.2	Filter steps involved in variant prioritisation.	138
5.3	Segregation of the tri-allelic genotype is depicted with family pedigrees.	141
6.1	Patient selection for sequencing.	152
6.2	Two steps of sample omissions from the analysis.	163
6.3	PAX6 gene structure, sequencing depth and genomic context.	172
6.4	Sanger sequencing chromatograms for the six samples which are queried for the presence of c.G858T and c.C856A (produced by Thermo Fisher Scientific Sanger Quality Check application).	173
7.1	Global map of marriages between couples listed as related second degree relatives or closer.	190
7.2	Causes of visual impairment and blindness in Pakistan.	192
7.3	Box plots of heterozygosity and variant concordance between all samples.	202

LIST OF FIGURES

A.1	The Integrative Genomics Viewer (IGV) image of two samples with two variants, Q368* and T419A.	282
B.1	61 POAG genes GenePy scores from the POAG cohort (n=358) corrected by GDI and gene length.	299
E.1	IGV image of reads aligning to the tyrosinase gene.	327

List of Tables

1.1	Summary of the causes of INS, the associated genes and clinical presentations.	8
1.2	Table of the most common genes causal for POAG.	13
2.1	Myocilin gene promoter, intronic, exonic and intergenic region locations within hg38 human reference genome.	40
2.2	Demographic and clinical characteristic summaries of age, intraocular pressure (IOP), cup:disc ratio (CDR) and visual field mean deviation (VFMD) for POAG cohort (n=372).	45
2.3	Quality control test run summary provided by the concierge service.	47
2.4	Summary of quality control statistics for all four batches.	49
2.5	Quality statistics before sample omissions and after sample omissions.	51
2.6	Summary of the number of variants identified across all features of the myocilin gene.	53
2.7	Annotation of all coding sequence variants.	56
2.8	Five non-coding variants remain following initial filtering of the 160 non-coding variants.	58
3.1	A gene list was formed of POAG associated genes that could prospectively be sequenced.	79
3.2	Coverage proportion at 1X, 20X, 50X and 100X using the Illumina Design Studio BED file for NRCC design and the UCSC RefSeq (RefGene) BED file.	80
3.3	Thirty-four filtered variants across the Mendelian-like POAG genes.	84
3.4	Sixty-four filtered variants across the 56 coding assumed complex genes in POAG.	87
3.5	Five non-coding RNA variants following the application filter criteria.	88

LIST OF TABLES

3.6	GenePy scores for five IL1B variants.	94
3.7	Mann-Whitney p-values for comparisons of gene level pathogenicity scores between the POAG (n=350) and non-eye disease (n=403) cohort. . . .	97
3.8	GenePy scores for all ABO variants detected across the POAG cohort.	98
3.9	CNV gains and losses detected by CNVKit.	101
4.1	List of nine unrelated nystagmus patients and their phenotypes. . . .	116
4.2	List of 45 nystagmus genes.	120
4.3	List of 168 nystagmus with albinism or nystagmus with ataxia genes. .	120
4.4	List of 242 retinal disorder genes.	121
4.5	Coverage across the exome for 5X, 10X and 20X depth.	122
4.6	Assessment of X chromosome heterozygosity and the VerifyBamID freemix for 9 samples.	123
4.7	Pairwise calculation of to show the percentage of total variants common to both samples.	123
4.8	Potential causal variants identified following the applied filtering for each patient.	124
5.1	OCA and OA genes targeted to identify genetic causality in eighteen partial albinism patients.	130
5.2	Coverage for eight genes of interest sequenced in eighteen probands with hypomorphic albinism.	137
5.3	Assessment of heterozygosity and the VerifyBamID freemix for eighteen samples.	137
5.4	Predicted causal variants, in eighteen probands with phenotypes match- ing hypomorphic albinism.	139
6.1	Selection criteria of the four phenotype groups.	152

LIST OF TABLES

6.2	HGNC approved gene names for genes listed in the UKGTN gene panel for nystagnus and albinism with the associated OMIM inheritance pattern and phenotype.	154
6.3	Overall coverage, contamination and file sizes of 134 samples in batches 1-6.	159
6.4	Family member pairs identified from the variant similarity matrix. . .	160
6.5	Coverage at 20x depth across the 31 gene panel for the three sample duplicates and 33 samples of coverage less than 90 percent.	161
6.6	Seventeen samples omitted that were not probands, did not have horizontal waveform nystagnus or had abnormal ERG results.	162
6.7	Overall coverage, contamination and file sizes of 81 retained samples in batches 1-6.	162
6.8	Seventeen of 81 patients with assumed pathogenic variants which were determined to be likely causal genotypes.	166
6.9	Sixty-four patients investigated for likely causal genotypes as assumed likely pathogenic or as a combination of assumed pathogenic and assumed likely pathogenic.	169
6.10	Fifty-five patients which did not have likely causal genotypes with assumed pathogenic or assumed likely pathogenic variants.	176
6.11	Forty-six patients were analysed for single heterozygous assumed pathogenic or assumed likely pathogenic variants in at least one albinism gene. . .	178
6.12	Summary table summarising the number of samples identified with likely causal genotypes.	179
7.1	List of the 57 samples and their given phenotype.	194
7.2	Gene list of UKGTN cataract disorder genes (n=113).	197
7.3	Gene list of UKGTN, eye malformations 204 gene exome panel (n=204). .	197

LIST OF TABLES

7.4	Gene list of UKGTN hearing loss, syndromic and non syndromic (n=95).	198
7.5	Coverage statistics, contamination and file size quality control for the Pakistani cohort (n=57).	200
7.6	OCA and cataract sub-cohort filtered variant positions which were found within runs of homozygosity (ROH).	204
7.7	Fourty-seven albinism/nystagmus patients investigated for likely causal genotypes.	205
7.8	Two albinism/nystagmus patients (NG444 and NG470) not previously identified to have likely causal genotype were identified to have a likely causal tri-allelic genotype for albinism within TYR.	206
7.9	Five congenital cataract and one blind patient investigated for likely causal genotypes.	209
7.10	Two patients with Waardenburg syndrome and Usher syndrome and their 24 variants.	211
7.11	Two patients with microcornea and Joubert syndrome investigated for likely causal genotypes.	213
7.12	Potential diagnostic yield overview.	215
A.1	Demographic and clinical information of the ten family member pairs which were included in the study regardless of meeting patient selection criteria.	282
A.2	All variants identified across the entire myocilin gene.	283
A.3	11 cases carrying a candidate disease-causing myocilin mutation.	289
B.1	UCSC RefSeq (RefGene) hg38 coordinates and coverage of at least 1X depth for SH3PXD2B, SOD2, NTM, APOE, and TXNRD2.	291
B.2	OPTN predicted interface coordinates (Interactome INSIDER) BED file (hg38).	292

LIST OF TABLES

B.3	TBK1 predicted interface coordinates (Interactome INSIDER) BED file (hg38).	292
B.4	CDC7 predicted interface coordinates (Interactome INSIDER) BED file (hg38).	295
B.5	CDKN2A predicted interface coordinates (Interactome INSIDER) BED file (hg38).	296
B.6	Mann-Whitney test p-values between the POAG (n=350) and IBD (n=403) cohort with Bonferroni and False Discovery Rate (FDR) corrected p-values calculated for 64 single genes and six permutations of interacting genes.	300
B.7	CNVKit output for all samples and genes.	302
C.1	Whole exome sequencing sample VerifyBamID freemix results.	318
D.1	TruSight One batch 1 VerifyBamID freemix results.	319
D.2	TruSight One batch 2 VerifyBamID freemix results.	320
E.1	Primers designed for Sanger sequencing.	321
E.2	Percentage coverage at 20x depth across the 31 gene panel using the TruSight One BED file.	321
E.3	Pair-wise percentage concordance of variants between samples.	326
E.4	Extract of the sample sheet for batch 6 samples.	327
E.5	Cohort (n=81) mean percentage coverage at 20x depth across the 31 gene panel using collapsed RefSeq (curated) transcripts from UCSC database.	329
F.1	Runs of homozygosity (ROH) identified for each sample.	331

Acknowledgements

I would firstly like to thank Prof Sarah Ennis, Dr Jane Gibson, Ms Angela Cree, Mr Jay Self and Prof Andrew Lotery for giving me the opportunity to work on this PhD project under their supervision. I am very grateful to my supervisors for providing me with their guidance and support throughout. I would like to thank all of my colleagues at both the Genomics Informatics and Vision Sciences groups for their contributions and for making my experience an enjoyable one.

I thank Gift of Sight and the International Glaucoma Association (IGA) for funding this work. Thank you to the patients and families who have provided their samples and consented to this research.

Thank you to all of my loved ones who have supported me unconditionally throughout this process.

Ethics Approval

Patients were recruited following the tenets of the declaration of Helsinki, informed consent was obtained and the research was approved by the Southampton & South West Hampshire Research Ethics Committee (REC 028/04/t; REC 05/Q1702/8). All methods were carried out in accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and, if subjects were under 18, from a parent and/or legal guardian.

Funders

This studentship was supported by Gift of Sight and the International Glaucoma Association (IGA).

Declaration of Authorship

I, **Luke O’Gorman**, declare that this thesis and the work presented in it is my own and has been generated by me as the result of my own original research.

Title of thesis: **The identification of causal variants in nystagmus and primary open-angle glaucoma patients through analyses of next-generation sequencing data.**

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as detailed overleaf.

Signed:

Date:

Publications

Published

1. Mossotto E, Ashton J, **O’Gorman L**, Pengelly R, Beattie R, MacArthur B, Ennis S. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC Bioinformatics* (2019).
2. **O’Gorman L**, Cree A, Ward D, Griffiths H, Sood R, Denniston A, Self J, Ennis S, Lotery A and Gibson J. Comprehensive sequencing of the myocilin gene in a selected cohort of severe primary open-angle glaucoma patients. *Scientific Reports* (2019).
3. Arshad M, Harlalka G, Lin S, D’Atri I, Mehmood S, Shakil M, Hassan M, Chioza B, Self J, Ennis S, **O’Gorman L**, Norman C, Aman T, Ali S, Kaul H, Baple E, Crosby A, Ullah M and Shabbir M. Mutations in *TYR* and *OCA2* associated with oculocutaneous albinism in Pakistani families. *Meta Gene* (2018).
4. **O’Gorman L***, Norman C*, Gibson J, Pengelly R, Baralle D, Ratnayaka J, Griffiths H, Rose-Zerilli M, Ranger M, Bunyan D, Lee H, Page R, Newall T, Shawkat F, Mattocks C, Ward D, Ennis S, Self J. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (*TYR*) causing hypomorphic oculocutaneous albinism (OCA1B). *Scientific Reports* (2017). *Joint first authors.

Submitted

1. ***O’Gorman L***, Norman C*, Michaels L, Newall T, Crosby A, Mattocks C, Cree A, Lotery A, Baple E, Ratnayaka J, Baralle D, Lee H, Osborne D, Shawkat F, Gibson J, Ennis S, Self J.* A small gene sequencing panel realises a high diagnostic rate in patients with congenital nystagmus following basic phenotyping. *Joint first authors.

Abbreviations

1000g	1000 Genome project
1000g EUR	1000 Genome project (European population)
AC	Allele count
ACMG	American College of Medical Genetics and Genomics
AMP	Association for Molecular Pathology
AF	Allele frequency
AD	Autosomal dominant
AR	Autosomal recessive
BAM	Binary Alignment/Map files
BED	Browser extensible data
BQSR	Base Quality Score Recalibration
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation Dependent Depletion
CCT	Central corneal thickness
CDR	Cup:disc ratio
Chr	Chromosome
CN	Congenital nystagmus
CNV	Copy number variation
CSNB	Congenital stationary night blindness
Ctcf	Distal CTCF/Candidate Insulators
CTD	C-terminal domain
CtrcfO	Distal CTCF/Candidate Insulators
dbSNP	Database of single nucleotide polymorphisms
DECIPHER	DatabasE of Chromosomal Imbalance and Phenotype in Humans

ABBREVIATIONS

	using Ensembl Resources
DFP	Disease-associated polymorphisms with supporting functional evidence
DM	Disease-causing mutation
DM?	Possible pathological mutation
DNA	Deoxyribonucleic acid
EA2	Episodic ataxia type 2
Enh	Enhancer candidates
EnhF	Enhancer flanking open chromatin
EnhW	Weaker enhancer candidates
EnhWF	Weaker enhancer flanking open chromatin
ERG	Electroretinogram
ExAC	Exome Aggregation Consortium
ExAC ALL	Exome Aggregation Consortium (all populations)
ExAC NFE	Exome Aggregation Consortium (non-Finnish European population)
ESP6500	Exome Sequencing Project 6500
ESP6500 (EA)	Exome Sequencing Project 6500 (European American population)
FATHMM	The Functional Analysis through Hidden Markov Models
GATK	The Genome Analysis Toolkit
GDI	Gene damage index
GERP	Genomic evolutionary rate profiling
gnomAD	The Genome Aggregation Database
GWAS	Genome wide association studies
HET	Heterozygous
HI	Haploinsufficiency score
Hg19	Human genome 19
Hg38	Human genome 38

ABBREVIATIONS

HGMD	Human Gene Mutation Database
HOM	Homozygous
HTG	High tension glaucoma
IGV	Integrative Genomics Viewer
IIN	Idiopathic infantile nystagmus
Indels	Insertions/deletions
INS	Infantile nystagmus syndrome
IOL	Intraocular lens
IOP	Intraocular pressure
IQR	Interquartile range
Lod	Logarithm of the odds
Low	Low activity proximal to active states
MAF	Minor allele frequency
MRI	Magnetic resonance imaging
NGS	Next-generation sequencing
NRCC	Nextera Rapid Capture Custom
NTG	Normal tension glaucoma
OA	Ocular albinism
OA1	Ocular albinism type 1
OCA	Oculocutaneous albinism
OCA1A	Oculocutaneous albinism type 1A
OCA1B	Oculocutaneous albinism type 1A
OCA2	Oculocutaneous albinism type 2
OCA3	Oculocutaneous albinism type 3
OCA4	Oculocutaneous albinism type 4
OCA5	Oculocutaneous albinism type 5

ABBREVIATIONS

OCA6	Oculocutaneous albinism type 6
OCA7	Oculocutaneous albinism type 7
OCT	Optical coherence tomography
P	Pathogenic
PCA	Principal component analysis
PCR	Polymerase chain reaction
PhyloP	Phylogenetic P-values
POAG	primary open-angle glaucoma
Pos	Position
QC	Quality control
Quies	Quiescent states
Repr	Repressed regions
ReprW	Weaker repressed regions
ROH	Runs of Homozygosity
RP	Retinitis pigmentosa
SCA6	Spinocerebellar ataxia-6
SIFT	Sorting Intolerant From Tolerant
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
Tss	Transcription start sites region
TssF	Transcription start site flanking
UCSC	University of Santa Cruz
UKGTN	UK Genetic Testing Network
VCF	Variant Call Format
VEP	Visual Evoked Potential
VFMD	Visual Field Mean Deviation

ABBREVIATIONS

VUS	Variants of unknown significance
WES	Whole exome sequencing
WGS	Whole genome sequencing
WISH	Wessex Investigational Sciences Hub laboratory
WRGL	Wessex Regional Genetics Laboratory
XL	X-linked
XLD	X-linked dominant

1 Background

1.1 Prevalence of ophthalmic diseases

There are estimated to be 285 million people visually impaired and 39 million blind world-wide [1]. In England and Wales, the main causes of blindness are degeneration of the macular and posterior pole (58.6%), glaucoma (8.4%), diabetic retinopathy (6.3%), retinal disorders (5.5%), optic atrophy (4.2%), visual cortex disorders (2.3%), retinal vascular occlusions (1.8%) [2] (Figure 1.1). For less severe cases with only visual impairment, the leading causes were similarly degeneration of the macular and posterior pole (57.2%), glaucoma (7.4%), diabetic retinopathy (7.6%), retinal disorders (4.2%), optic atrophy (2.9%), visual cortex disorders (2.9%), retinal vascular occlusions (1.2%) [2]. Nystagmus is not known to be a cause for blindness, however, it has been found to account for 0.83% of visual impairment of people in the United Kingdom [2].

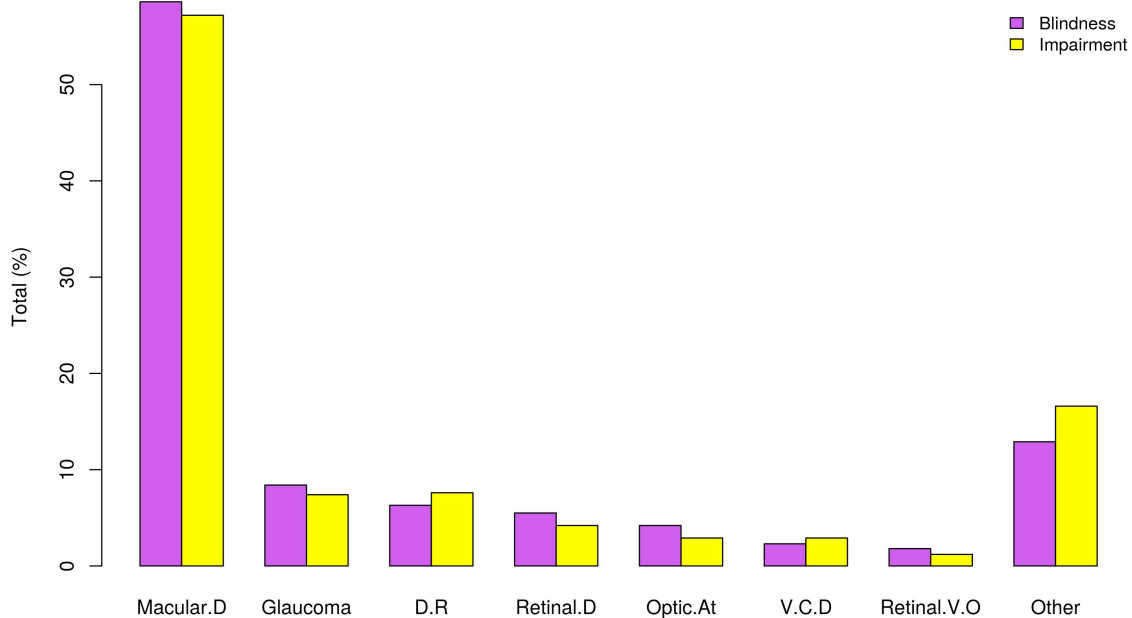


Figure 1.1: *Causes of visual impairment and blindness in England and Wales.* Totals extracted from Bunce et al, 2010 [2] for degeneration macular & posterior pole (Macular.D), glaucoma, diabetic retinopathy/maculopathy (D.R), retinal disorders (R.D), optic atrophy (Optic.At), visual cortex disorders (V.C.D), retinal vascular occlusion (Retinal V.O) and others.

Twin studies have been used to examine the heritability and the weight of genetic influence on a given trait. This is able to be achieved by comparing monozygotic twins and dizygotic twins using methods such as structural equation modelling (SEM) [3]. If monozygotic twins were to show more similarity for a given trait compared to dizygotic twins, this would suggest that the underlying genetic basis has a significant role in causing that trait. Twin studies have found that ophthalmic traits have moderate to very high heritability [4]. The higher concordance of glaucoma amongst monozygotic twins compared with dizygotic twins suggests that there is a significant genetic factor underlying glaucoma [5, 6].

Genetic analysis is able to help identify known or novel gene mutations involved in the pathogenic cause of a disease. Following the identification of novel variants or associated causal genes [7], it is then possible to understand the mechanism by which disease state phenotypes arise. Tailored treatments can be planned for surgery, pharmacotherapy, counselling and in some cases, gene therapy [8, 9] depending on the molecular mechanism underpinning their disease [10].

There have been a range of successful studies in the discovery of novel variants and genes for ophthalmic disease including nystagmus [11, 12] and POAG [13, 14] with clinical evidence to support them.

The evolution of high-throughput sequencing [15], alignment [16], variant calling [17] and variant annotation tools [18] has allowed an increased capacity for identifying causal variants and resolving genetic aetiology. The following research aims to focus on the identification of causal variants and genes for both nystagmus and primary open-angle glaucoma (POAG) patients.

1.2 Nystagmus

1.2.1 Description

Infantile nystagmus is an ophthalmic condition which has a prevalence of 24 in 10,000 in the United Kingdom [19]. It is characterised by the oscillation of one or both eyes about one or more axes which ultimately leads to visual impairment of an individual [20]. Infantile nystagmus is a clear and obvious symptom of a range of underlying conditions from metabolic disease to retinal disease to apparently isolated nystagmus. Childhood nystagmus can be categorised into congenital (CN), also known as Infantile Nystagmus Syndrome (INS), Spasmus Nutans, manifest latent or acquired (AN) (Figure 1.2) which is more frequent in adults [21].

AN can develop following drug toxicities in the body or is seen in metabolic or neurological disorders [20]. Herein focus of genetic analyses of nystagmus patients will be of INS with an aim to identify causal aetiology in cases from University Hospital Southampton.

INS is nystagmus present at birth or in infancy. It can be caused by congenital idiopathic nystagmus, albinism or retinal associated diseases (Figure 1.2). There are causes of nystagmus which are not discussed here in detail such as neurological disorders. Idiopathic and albinism are the two most common forms of INS [19].

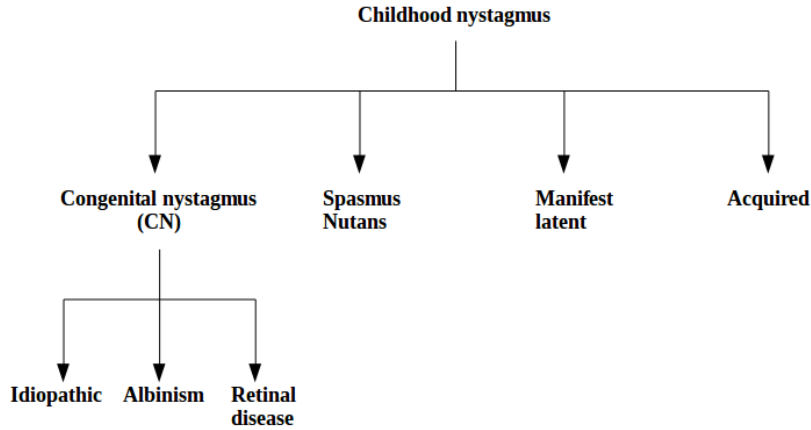


Figure 1.2: *Classification of childhood nystagmus.* Table is adapted from Hussain, 2016 [21].

1.2.2 Diagnosis

Nystagmus is an abnormal eye movement initiated by a slow phase which is followed by either a fast movement (jerk nystagmus) or a second slow movement (pendular nystagmus). Both nystagmus and the saccadic oscillations can increase visual impairment of the eye [20]. Drifting of the eyes away from the central position with an increasing speed is encountered most commonly in the horizontal plane in congenital nystagmus [20, 22]. INS diagnoses are made on the basis that slow phases show an increasing (accelerating) waveform [23].

1.2.3 Genetic cause

Idiopathic infantile nystagmus syndrome (INS)

Idiopathic infantile nystagmus syndrome (IINS), also known as congenital idiopathic nystagmus, most commonly follows an X-linked mode of inheritance with He *et al* concluding that upto 47% of causal mutations are found in the X chromosome FERM domain gene, *FRMD7* [24]. Tarpey *et al* [12] identify that upto 57% putative X-linked

pedigrees and 94% proven X-linked pedigrees of causal mutations are found in *FRMD7* [12]. Whilst the function of *FRMD7* is still relatively poorly understood, there is restricted expression of *FRMD7* found in human developing retina and brain, suggesting that *FRMD7* has a significant role in the control of eye movement [12].

INS in albinism

Nystagmus is a known clinical feature identified in patients diagnosed with ocular albinism (OA) or oculocutaneous albinism (OCA) (Table 1.1). Ocular albinism is a form of albinism presenting primarily in the eyes [25] and is inherited in X-linked pattern of inheritance. Oculocutaneous albinism is a form of albinism presenting in the eyes, skin and hair [25] and is usually inherited as a recessive disease. Additional ophthalmic features for OA and OCA include impaired visual acuity, strabismus, and photophobia. Ocular albinism type 1 (OA1) is the most common form of ocular albinism [25]. OCA1 has two sub-types in the form of OCA1A and OCA1B. OCA1A describes complete loss of tyrosinase activity and subsequently lack of melanin production throughout life [25, 26]. However, OCA1B retains partial tyrosinase activity, allowing pigment to accumulate. Likewise, OCA2, OCA3 and OCA4 show partial pigment accumulation [25, 26].

INS in Ataxia and Noonan syndrome

Episodic ataxia (EA) is a neurologic condition which involves incoordination and imbalance [27]. EA type 2 (EA2) is the most common form of EA [27] and is often associated with nystagmus [28].

Noonan syndrome (NS) is an autosomal dominant disorder which involves clinical pre-

sentations of being short in stature, facial dysmorphism and congenital heart defects [29]. In a study of 58 Noonan syndrome patients, it was found that nystagmus is present in 9% of Noonan syndrome cases [29].

CHAPTER 1

Table 1.1: **Summary of the causes of INS, the associated genes and clinical presentations.** Three causes of INS are described in the table (idiopathic, albinism and retinal disorders). For each cause of INS, the associated HUGO Gene Nomenclature Committee (HGNC) gene (or loci if gene not known) and mode of inheritance (AD: autosomal dominant, AR: autosomal recessive, XL: X-linked). For each subgroup of INS cause, the clinical features are listed. Retinal disorders associated with nystagmus can also encompass congenital cataracts, corneal opacities, developmental disorders of the optic disc and retina and retinopathy of prematurity (see Papageorgiou et al [30] for further details).

¹ There is also evidence of AD [31, 32, 33, 34] and AR pedigrees [32], however, no gene has been identified for these modes of inheritance.

² Retinal disorders of rod and cone systems are listed (achromatopsia and Leber Congenital Amaurosis (LCA)). Three achromatopsia genes were identified for achromatopsia (Michaelides et al [35]) and four genes were identified for LCA with nystagmus clinical features [36, 37].

Cause	Cause subgroup	HGNC gene name/ locus	Inheritance	Prevalence of cause subgroup (population if known)	Clinical features
Idiopathic ¹	Idiopathic	<i>FRMD7</i>	XL	1:1000 [12]	Nystagmus Head oscillations Impaired vision [12]
Albinism	OA1	<i>GPR143</i>	XL	1:60000 [38]	Nystagmus Photophobia Impaired vision Albino pupillary reflex Depigmented fundus Prominent choroidal vessels [39]
	OCA1A	<i>TYR</i>	AR	1:40000 [25]	Nystagmus Photophobia intense Severely reduced visual acuity White hair and skin Pink irides (childhood) Blue-grey irides (adult) [25]
	OCA1B	<i>TYR</i>	AR	1:40000 [25]	Nystagmus Photophobia Reduced visual acuity Some pigmentation of hair, skin and iris [25]
	OCA2	<i>OCA2</i>	AR	1:36000 (European) [25]	Nystagmus Variable hypopigmentation of the skin and hair Reduced visual acuity [25]
	OCA3	<i>TYRP1</i>	AR	1:8500 [25]	Nystagmus Rufous or brown albinism and occurring mainly in the African population Visual anomalies unclear [25]
	OCA4	<i>SLC45A2</i>	AR	1:85000 (European, Asian) [25]	Clinical features indistinguishable to OCA2 [25]
	OCA5	4q24	AR	-	Nystagmus White skin and golden hair Photophobia Foreal hypoplasia Impaired visual acuity [40]
	OCA6	<i>SLC24A5</i>	AR	-	Nystagmus White skin and light hair Transparent irides Photophobia Foreal hypoplasia Reduced visual acuity [41]
	OCA7	<i>LRMDA</i>	AR	-	Nystagmus Hypopigmentation of hair Hypopigmentation of eyes Iris transillumination [42]
Retinal disorder ²	Achromatopsia 2	<i>CNGA3</i>	AR	1:30000 [45]	Nystagmus Photophobia Reduced visual acuity Unable to discriminate between colours [43]
	Achromatopsia 3	<i>CNGB3</i>	AR		Nystagmus Photophobia Reduced visual acuity Unable to discriminate between colours [43] Develops cataract [44]
	Achromatopsia 4	<i>GNAT2</i>	AR		Nystagmus Photophobia Reduced visual acuity Unable to discriminate between colours [43]
	LCA8	<i>CRB1</i>	-	1:81000 (North American) [37]	Nystagmus Vision loss Severe retinal dysfunction [37] High to extreme hyperopia [46]
	LCA1	<i>GUCY2D</i>	AR		Nystagmus Vision loss Severe retinal dysfunction [37]
	LCA7	<i>CRX</i>	-		Nystagmus noted in the first few months of life Freund 1998 Severe vision loss Severe retinal dysfunction [37]
	LCA6	<i>RPGRIPI</i>	-		Nystagmus Vision loss Severe retinal dysfunction [37] Degeneration of both rod and cone photoreceptors [47]

1.2.4 Treatment

For most forms of treatment, the focus is on the reduction of amplitude or frequency of oscillations rather than the underlying molecular mechanism. Most success has been found with gabapentin in the treatment of acquired pendular nystagmus and baclofen in the treatment for acquired periodic alternating nystagmus [48, 22, 49]. On a molecular level, Gabapentin targets the calcium channel subunit A2D1 [50] and memantine is a non-competitive antagonist of the N-methyl-D-aspartate receptor [51]. Both drugs are effective for reducing median eye speed, however, are not ideal treatments for acquired forms of nystagmus due to side effects observed. Both gabapentin and memantine have been found to have similar side effects mostly consisting of fatigue [52, 53]. In congenital nystagmus, therapies not requiring surgery such as convergence prisms, contact lenses and afferent stimulation can be used to improve visual performance [54]. With the discovery of novel genes and variants, a better understanding of the molecular mechanism underlying nystagmus may indicate how pathways targetable with drugs can be used to target the molecular cause and not just suppress symptoms.

1.3 Primary open-angle glaucoma

1.3.1 Description

Glaucoma

Normally, the aqueous humor is able to flow out of the anterior chamber of the eye or ‘drain’ through the trabecular meshwork at the angle where the cornea and iris meet [55] (Figure 1.3). In glaucoma, the aqueous flow can become blocked which leads to an increased intra-ocular pressure (IOP) and progressively leads to blindness as a result of optic nerve damage and loss of retinal ganglion cells [55]. Glaucoma presents in multiple forms, however, primary open-angle (POAG) is the most common form [56, 57].

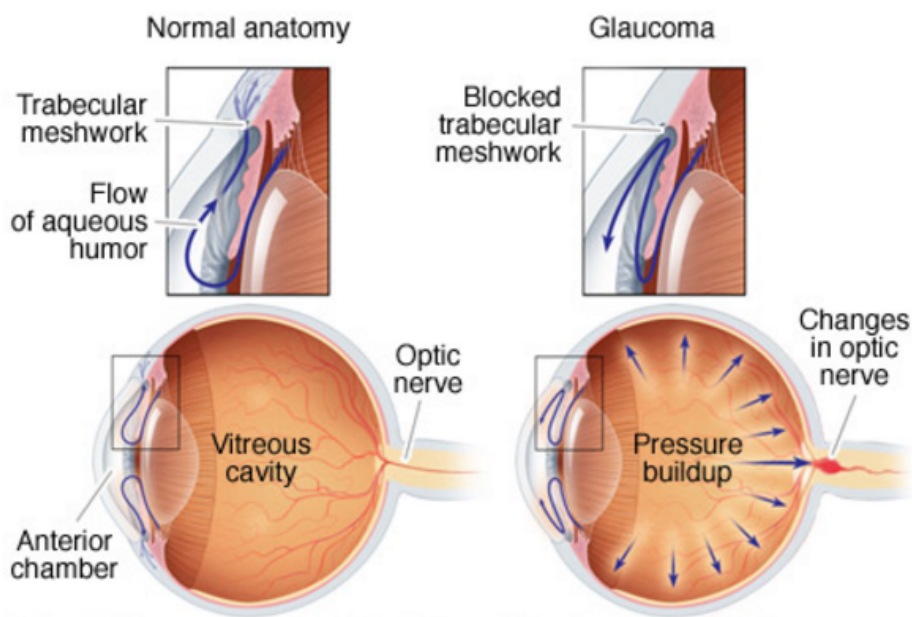


Figure 1.3: *Diagram of the anatomy underlying POAG and normal aqueous flow of the eye.* Normal eye aqueous flow, left; POAG eye aqueous flow, right. Extracted from MAYO foundation for Medical Education and Research [58].

POAG

POAG is a progressive disease and is more common in an elderly population. It affects 1-2% of Caucasians in the UK over the age of 40 years old [59, 60] and 4% over the age of 80 years old [59, 61].

POAG is characterised by an open anterior chamber angle, damage to the optic nerve and visual field loss [56, 57]. There are two subtypes of POAG, high tension glaucoma (HTG) and normal tension glaucoma (NTG). HTG is a subtype of POAG which has higher IOP (Figure 1.3) and NTG is where the IOP is not elevated [57]. It is not associated with known ocular or systemic disorders which could cause optic nerve damage or reduction of aqueous flow.

Other forms of glaucoma

Other less common forms of glaucoma include closed-angle, congenital and secondary glaucomas [56, 57]. Closed-angle glaucoma is caused by the angle between the cornea and the iris narrowing to prevent the aqueous flow to the angle, leading to ocular hypertension (elevated IOP) [62, 55]. This is diagnosed as a medical emergency and laser surgery is required [55]. Congenital glaucoma occurs from defects in the angle in children which causes impaired vision. Early treatment with surgery is likely to result in restoration of vision [55]. Secondary glaucoma is diagnosed where elevated IOP is caused by underlying diseases or causes such as pseudoexfoliation, surgery, advanced cataract, diabetes or from drugs [57]. Treatment can be given with laser or conventional surgery [55].

1.3.2 Diagnosis of POAG

It is essential to diagnose POAG as early as possible as the visual impairment which typically ensues is irreversible. POAG symptoms are multi-factorial and progressive [63]. Therefore, younger glaucoma patients are more likely to be asymptomatic and an early diagnosis is not straightforward. When determining a diagnosis or prognosis for a patient, it is important to take into account the risk factors in addition to the physical examination.

POAG risk factors

POAG is associated with several risk factors. IOP is considered the most important risk factor in developing POAG. It is also the only risk factor which can be treated. As the population age increases, the prevalence of POAG also increases [59]. POAG is also known to be more common in Africans more than other ethnic groups. Africans have also shown worse prognoses by having an earlier onset and faster progression of POAG [64, 65]. Family history is also an important risk factor for developing POAG, especially if the family member is a first degree relative [66]. Some studies have also shown that increased refractive error is also a risk factor [67], however, this remains controversial [68]. A decrease of central corneal thickness (CCT) is also believed to be a risk factor for ocular hypertension (OHT) patients developing POAG.

Physical examination

POAG is examined for diagnosis focusing on multiple sub-phenotypes. IOP can be examined with tonometry which measures the pressure in millimetres of mercury (mm Hg). Normal ranges for IOP are between 10-20 mm Hg [57, 69, 70] and HTG POAG

patients will present with an IOP higher than the normal range. Determining a normal appearing but open anterior chamber angle can be determined with slit-lamps and gonioscopy [57]. Optic disc damage can be determined with ophthalmoscopy and slit-lamp biomicroscopy [57]. ‘Cup:disc ratio’ (CDR) is normally <0.7 and a patient with POAG would typically exceed this [57]. However, the patient may or may not show these sub-phenotypes. Ultimately, it is the visual field which becomes constricted and peripheral vision lost in late disease [57, 60, 71, 72]. The visual field deviation can be measured using standard automated perimetry (SAP) which compares the visual field of a patient against a reference visual field which has been corrected for age [73].

1.3.3 Genetic causes of POAG

Collectively, monogenic causes for POAG account for approximately 5% of all POAG cases (see Table 1.2). Variants within the myocilin gene, *MYOC* (p.T293K, p.R296C and p.Q368*), were identified for 2.2% of a United Kingdom POAG cohort [75]. These *MYOC* variants possibly contributed towards a small proportion of monogenic causes of POAG. The optineurin gene, *OPTN*, has been identified as being accountable for up to 1.5% of NTG cases [74, 75, 76]. The majority of causes of POAG are believed to be complex (polygenic) genetic causes [77, 78].

Table 1.2: *Table of the most common genes causal for POAG.*

Gene	Mutation Prevalence
<i>MYOC</i>	2.2% (United Kingdom; POAG) [79]
<i>OPTN</i>	1.5% (NTG/ POAG) [74, 75, 76]
<i>WDR36</i>	1.6-17% (POAG) [80]
<i>NTF4</i>	1.7% (European; POAG) [81]
<i>TBK1</i>	1.3% (NTG) [77]

1.3.4 Myocilin

Mutations in the myocilin gene, *MYOC* (located in chromosome 1q24), were previously identified in families with autosomal dominant POAG [82, 83]. The *MYOC* gene is comprised of three exons (see Figure 2.5) and encodes a predicted 504 amino acid polypeptide. The translated polypeptide consists of a leucine zipper, an N-terminal myosin-like domain and a C-terminal olfactomedin-like domain [84]. The majority of *MYOC* mutations are found in exon 3 [79, 85] where the most common pathogenic mutations are found to have a penetrance of up to 90% in POAG cases [86].

Myocilin expression within the eye is ubiquitous [86]. The global expression of *MYOC* in a range of organs and tissue including the heart, skeletal muscle and bone marrow amongst others implies that it does not have eye specific functionality [86]. *MYOC* is believed to have a role in cytoskeletal function, and once mutated, causes an increase in intraocular pressure through obstruction of the aqueous outflow [87, 88, 83].

A range of glaucoma-causing *MYOC* mutations have been identified and collated in databases including the ‘Myocilin allele-specific glaucoma phenotype database’ [89]. Exon 1,2 and 3 have 32, 1 and 62 known glaucoma-causing variants respectively. However, the promoter does not have any known glaucoma-causing genes identified to date. Haploinsufficiency has been excluded as a disease mechanism underlying *MYOC* associated POAG [86] and under or over-expression of wild-type myocilin appears not to cause POAG [90]. However, evidence suggests that POAG can be caused by gain-of-function mutations in the *MYOC* gene [84].

1.3.5 Treatment

Most commonly, eye drops and drugs are provided to treat POAG by lowering IOP by promoting aqueous flow or through suppressing aqueous humor production [55]. Historically, non-surgical treatments have focused on lowering IOP with prostaglandin analogues, beta-blockers, alpha-agonists, and carbonic anhydrase inhibitors [91]. Drugs developed to lower the IOP have generally targeted the trabecular meshwork poorly. Therefore, combination treatments are often used for POAG patients which are tailored to fit the specific complex case of the patient [91]. Alternatively, laser trabeculoplasty can be performed to allow aqueous humor to pass through the trabecular meshwork. However, this is not seen as a viable long term option [55]. Other surgery that can be performed is filtration surgery. This creates a new passage for aqueous humor to be drained from the anterior chamber. It is perceived as the last option following earlier failures of other treatments [55].

1.4 Methods of genetic analysis

1.4.1 Genetic linkage

The method of genetic linkage has historically proven to be successful in identifying the regions of causal genes underlying disease [32]. When two loci are closer in proximity on a chromosome, the alleles tend to transmit through meiosis together. Therefore, two such loci are referred to as ‘linked’ [92]. When genes are located on the same chromosome and further apart, they can undergo the process of homologous recombination. This is a process that happens during the early phase of meiosis in which homologous chromosomes randomly exchange matching fragments [93]. Two loci that are proximal will have less recombination events between them and will, therefore, be more likely linked. To measure linkage quantitatively, we can calculate the recombination frequency [92]. Recombination frequency is calculated by the number of recombinant progeny divided by the total progeny. The recombination frequency can give an approximation of distance apart on the chromosome. A centimorgan defines two loci with an expected recombination frequency of 1% which typically translates to approximately 1 million base pairs apart. Therefore, two loci with a small recombination frequency are likely closer in proximity.

In familial studies of disease inheritance, the disease locus can be assigned to a region on a chromosome defined through genetic markers such as restriction-fragment length polymorphism (RLFP) or highly polymorphic microsatellites. If a marker maps in close proximity to a causal disease gene it is likely that all affected family members will inherit the marker along with the disease causing allele (Figure 1.4). By following the recombination events, the region containing the disease gene can be refined to hundreds or thousands of kilobases [92].

Comparison of recombination frequencies can also be used to understand the order of genes on a chromosome. Measuring recombination frequencies for closer proximity gene pairs minimises double crossovers occurring and being unaccounted for.

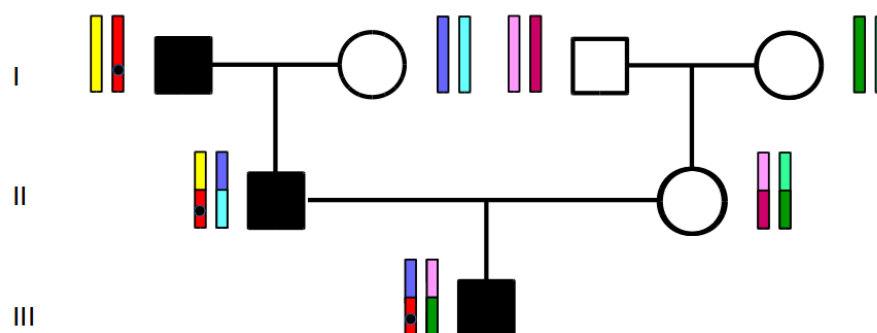


Figure 1.4: *Example of a pedigree with every affected member bearing the disease causing allele. Recombination events have reduced transmission of the disease causing haplotype (red), however, the causal mutation (black circle) is prevalent throughout.*

Another method of calculating the significance of linkage was optimised by Newton Morton in 1955. The logarithm of odds (lod) score [94] compares the null hypothesis (that the disease is not linked to the marker gene location) against the alternative hypothesis (that the disease is linked to the marker gene location). Positive lod scores will suggest that the disease and loci are linked. For Mendelian diseases, a typical lod score of 3 would suggest that it is 1000 times more likely that linkage is caused by co-segregation rather than by chance [95]. However, linkage studies fail to provide locations of causal genes when diseases are common or polygenic [96].

The first gene to be associated with primary open-angle glaucoma (POAG) was the *MYOC* gene which has been described previously. The glaucoma-causing mutations were firstly identified through linkage studies of juvenile open-angle glaucoma [83, 82].

Linkage analyses for POAG disease causing genes have also been carried out identifying the *OPTN* gene's role in POAG [97]. As for X-linked congenital idiopathic nystagmus (CIN), a linkage analysis found significant evidence for linkage to Xq26-q27, which contains *FRMD7* [12].

1.4.2 Genome wide association studies

Association studies can be performed through case-controls designed studies. This can identify association by identifying markers which are observed more often than by chance in individuals with the trait (cases) compared to the individuals without the trait (controls) [92, 98]. The completion of the Human reference genome in 2003 [99], HapMap project [100] and dbSNP [101] enabled genome-wide association studies (GWAS) to study genetic variation across the entire human genome and to identify variants associated with observable traits or disease.

The cases and control cohorts (often >1000 in number) are matched for confounding variables such as ethnicity and gender [98]. If population stratification exists, sub-populations will have differences in allelic frequencies which could lead to false positives [102]. Micro-array genotyping is applied for known SNPs which can be millions in number [103]. This is followed by imputation of haplotypes to increase the number of SNPs queried for association and therefore increase the power of the study [104]. Odds ratios and p-values are calculated for all SNPs. The non-parametric chi-squared test is used to test how likely it is that an observed distribution is due to chance by comparing case and control cohort genotypes [98]. Multiple-testing corrections such as the Bonferroni correction [105] can then be applied.

A p-value of 0.05 is considered as the nominal threshold of significance. However, currently, a p-value threshold of 5×10^{-8} is considered GWAS significant by taking into account multiple-testing [106, 107]. The association can then be further validated by replicating the GWAS within a large and independent cohort. This can verify the significant p-value detected in the initial discovery cohort and verify association [98].

GWAS do not include related individuals since there may be rare innocuous mutations appearing to have significant association with the disease. Therefore, an association study should only use unrelated individuals within a cohort of matched ethnicity. Overall, this should minimise the false-positives observed for the GWAS and detect risk variants specific to the disease [98].

The first successful GWAS identified the complement factor H (*CFH*) gene as a risk factor for age-related macular degeneration (AMD) [108, 109, 110]. The first GWAS applied to primary open-angle glaucoma was performed by Nakano *et al* on 418 POAG cases and 300 controls [111]. None of the SNPs achieved a genome-wide significance. GWAS has had limited success in the identification of genes in complex disease such as POAG even with large sample sizes [112, 113]. The first genome-wide association which was significant in POAG was the rs3213787 SNP in *SRBD1* which was detected in a Japanese cohort of NTG cases [114] which was since replicated in Caucasian [115] but not African-caribbean [116].

Within Southampton (UK), a POAG quantitative traits analysis was shown to replicate association for known POAG risk genes, however, there were no associations of novel genes with POAG risk factors through the GWAS performed [115]. The GWAS had limitations of a small sample size with a discovery case cohort of n=387 being used.

Structural variation and rare variants present in 5% or less of the population are poorly captured by GWAS as they are designed to capture more common variation [117, 118].

Consortium have formed to conduct large scale meta-analyses in POAG. The additional sample size reinforcement which have contributed to the identify replicated genes and identified a variant near the caveolin 1 (*CAV1*) and caveolin 2 (*CAV2*) genes on chromosome 7 associated with POAG [112, 119]. The largest meta-analysis to date focused on IOP of 139,555 participants from UK Biobank, EPIC-Norfolk, International Glaucoma Genetics Consortium (IGGC) and the NEIGHBORHOOD cohort [120]. This study identified 68 novel loci associations with IOP and suggested that with larger meta-analyses and a focus on precise sub-phenotypes of POAG, more success could be generated for such studies.

Rare variants across multiple genes could be a major factor in the cause of the substantial missing heritability in POAG. Therefore, next-generation sequencing (NGS) analyses for rare variants could contribute much of the remaining heritability. The rare variants which are detected through this method would not be suitable for analysis in GWAS as such studies require common SNPs. Furthermore, many of the GWAS studies indicate a substantial proportion of the significant associations in intergenic regions. Therefore, the consideration of epigenetics should not be overlooked [121].

1.4.3 DNA sequencing

1.4.3.1 Sanger sequencing

The first sequencing technology was developed in 1977 by Frederick Sanger and was known as ‘Sanger sequencing’ [122]. In this early sequencing methodology the DNA

sample is divided into four separate sequencing reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP), primers and DNA polymerase. One of four dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP) is added to each reaction, therefore, terminating the polymerase reaction. Following rounds of the reaction different size DNA fragments based upon the ddNTP that was incorporated in the polymerase reaction are generated. The radio-labeled fragments are run with electrophoresis on a polyacrylamide gel. The chain terminations closest to the primer generate the smallest DNA molecules which pass further down the gel, and chain terminations further from the primer generate larger DNA molecules which are travel do not pass as far down the gel. All four reactions are lined up and the DNA sequence can be read from the ‘ladder’ of fragment bands 5’ to 3’ from bottom to top.

The Sanger method has advanced since 1977 to become more automated using capillary electrophoresis and record dye fluorescence to output data in a chromatogram. Sanger sequencing is still used for verification of variants found in a sequence due to its high sequencing accuracy [123].

In 1995 the first bacterial genome was sequenced [124] which was followed by the sequencing and release of the first Human genome in 2003 (and the more complete version published in 2004) [125]. The human reference genome, approximately 3 billion base pairs in length, is under continuous curation of its assembly by the Genome Reference Consortium (GRC) to improve upon its previous versions. The latest release was in December 2013. Although Sanger sequencing produces long sequence reads (approximately 1,000bp), NGS is high-throughput and longer continuous sequences (contigs) can be constructed from the high output of shorter sequence reads.

1.4.3.2 Next (second) generation sequencing technology

NGS Sequencers

Next-generation sequencing (NGS), also known as massively parallel sequencing, is a technological advancement in recent years, giving rise to a more rapid understanding of the genome by providing millions of DNA sequence reads in just one run. NGS sequencers have massively multiplied the output in very little time [126, 127].

Next (2nd) Generation Sequencing includes Illumina sequencing, Ion Torrent and the now obsolete 454 sequencing [126]. Illumina currently dominates the industry and its method is illustrated in Figure 1.5 [128].

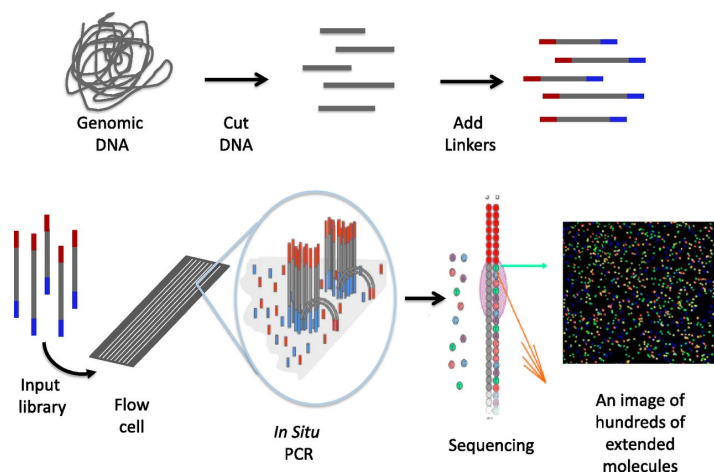


Figure 1.5: ***Illumina sequencing process.** DNA is sheared; adapters are added; DNA strands are captured on a flow cell; clusters are bridge amplified; following the annealing of primers, fluorescent reversible terminators added and the sequence is detected. Extracted from Johnsen et al [128]).*

Illumina Library preparation and cluster amplification

The genomic DNA is fragmented into desired sizes by enzymatic reactions or sonication.

Sample-specific barcodes are added to the DNA fragment to allow sample multiplexing 1.5. The DNA is then ligated to adapter sequences which are complementary to the oligonucleotides in sequencing flow cells. Libraries are purified and amplified in preparation for NGS input. Flow cells surfaces within the NGS machinery are coated with oligonucleotides in the manufacturing process. As the library is loaded into the flow cell, fragments in the library bind to the complementary oligos on the flow cell surface. The opposite end of a ligated fragment bends over to ‘bridge’ to another oligo on the surface. DNA polymerase and dTNPs are added to initiate ‘bridge amplification’ which forms a double stranded ‘bridge’ [15, 129]. The original strand is washed away leaving only the synthesised strands with oligos attached to the flow cell. The process is repeated which leads to the clustering of identical fragments [15, 129] which is essential for the fluorescent signal to be sufficient during sequencing.

Illumina sequencing-by-synthesis (SBS)

Illumina sequencing-by-synthesis (SBS) technology utilises a reversible terminator-based methods [129], as opposed to the irreversible chain termination utilised in Sanger sequencing. This method uses four fluorescently labeled nucleotides to sequence the millions of clusters in parallel. As the first base is incorporated, the emission from each cluster is recorded. The emission wavelength and intensity are used to identify the base. Since reversible terminators are used, incorporation biases are minimised [130]. This ‘sequence by synthesis’ method is repeated for the reversible strand for paired-ends.

Illumina Paired-end sequencing

Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Since the distance between each paired read is known, alignment algorithms can use this in-

formation to align the reads over repetitive regions more accurately [131]. This results in more accurate alignment of the reads.

Illumina Multiplexing

When a batch of sequences are to be sequenced, library multiplexing allows large numbers of the libraries to be pooled and sequenced simultaneously during a single sequencing run [132, 133]. Figure 1.6 illustrates the demultiplexing process of multiplexed data. Two distinct libraries are attached to unique index sequences. These index sequences are attached in library preparation. The libraries for each sample are then pooled and loaded into the same flow cell lane and are sequenced in one run. All sequences are outputted and are demultiplexed to sort the reads into different files according to their indexes. These files should correspond to respective samples [132].

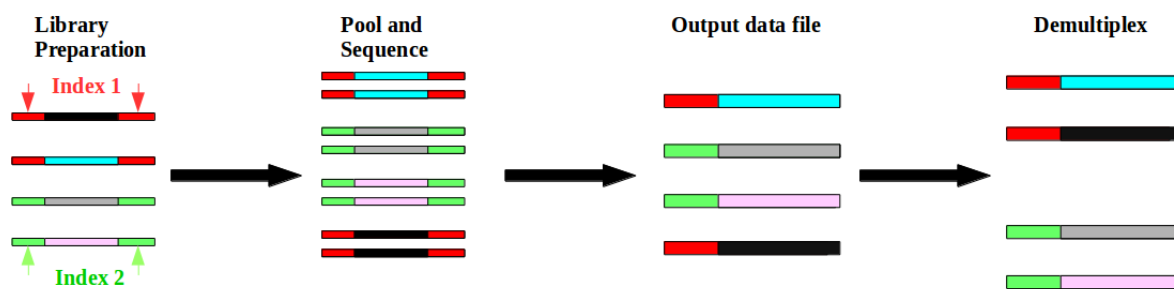


Figure 1.6: *Illumina multiplexing process.* During library preparation each sample library has unique index sequences (illustrated in red and green) attached to the DNA fragments (illustrated as black, cyan, grey and pink). Libraries are then pooled and sequenced together in a single run. The sequences are collected in a single output. Demultiplexing algorithms use the unique index sequences to sort the reads to files according to their index sequences.

Other sequencing technologies include Applied Biosystems (ABI) Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system which uses emulsion PCR amplification [15] and Complete Genomics which uses DNA Nano-Ball (DNB) arrays. However, it is Illumina MiSeq/HiSeq/NextSeq and the Ion Torrent Personal Genome Machine (PGM)

NGS platforms that have been more widely used for diagnostic purposes [134, 135].

1.4.3.3 Next (third) generation sequencing technology

Third generation sequencing includes the Pacific BioSciences RS and (SMRT) and Oxford Nanopore technology (MinION) [15]. SMRT is one of the main third generation sequencers used in the genetic industry [15] and enables single molecule real time sequencing and is capable of generating reads of 50 kb [127]. Bound to a DNA template, polymerase molecules synthesise a second strand of DNA with fluorescently labelled nucleotides and G-phosphate. This occurs in a 50nm wide well which is a zero-mode waveguide in which light cannot pass, but energy can to an extent, excite the fluorescent markers. There is a different fluorescence colour for each nucleotide which is detected as each nucleotide base is incorporated [15]. The long reads generated by SMRT are ideal for *de novo* assemblies [136], more accurate alignments and revealing long-range genomic structures [137, 127]. However, the PacBio SMRT has limited throughput and high costs ($\sim$ \\$1000 per Gb) and requires high coverage to overcome a high single pass error rate [127].

Oxford Nanopore sequencing uses a narrow pore to ensure that the native order of the nucleotides are detected and reported. Unlike PacBio SMRT, Nanopore sequencing involves a voltage potential being applied across the nanopore and the different nucleotides can be detected through a change of current through it [138]. The Oxford Nanopore technology allows the generation of ultra-long reads up to 200 kb for the Oxford Nanopore MK 1 MinION. This platform is also able to be run from a USB-based device, allowing rapid clinical responses and superior portability [127]. However,

Oxford Nanopore sequencing technology typically incurs higher error rates with the Oxford Nanopore MK 1 MinION bearing a $\sim 12\%$ error rate [127].

1.4.3.4 NGS pipelines

Alignment

NGS analysis is dependent on the accurate alignment or mapping of NGS reads to a reference genome. Mapping software enable the vast number of reads produced with NGS sequencers, to be assembled together in a meaningful way [139].

Dynamic programming alignments use an algorithmic approach termed ‘seed-and-extend’ to improve the efficiency of the mapping process [140]. Initially this technique finds substrings that match in both the reference genome and the reads to be mapped. These short matches are then extended into longer matches depending on whether gaps and mismatches are allowed.

Current alignment tools can generally be categorised into hashing and FM-indexing based methods [131]. The hashing method builds a hash table of subsequences of the genome. For each read, a k-mer is selected and for each aligning window, or seed, the alignment is extended similarly to the Needleman-Wunsch algorithm. This method is slower than the FM-indexing method, demands more memory during computation and is less sensitive at repetitive regions. However, it is better at tolerating regions of high genomic variation [131].

Other alignment tools such as Burrows-Wheeler Aligner (BWA) [141] and Bowtie [139]

use the Burrows-Wheeler Transform (BWT) method to efficiently index the genome using sorted suffixes [142]. This method enables fast, memory efficient alignment.

Effective gapped aligners such as BWA and novoalign are generally better at aligning sequences containing insertions or deletions (indels) by allowing gaps in the alignment. This is an important consideration when selecting an alignment tool as a gapped alignment is essential to the variant discovery [131].

Variant calling

Variation in an individual with respect to a reference genome can be identified using several tools including SAMtools [143] and GATK [144]. SAMtools determines the most likely genotype at the position using a Bayesian approach and assigns a Phred-scaled score for each call. The ‘GATK HaplotypeCaller’ utilises a PairHMM algorithm to give a matrix of likelihoods of haplotypes and uses a Bayesian approach to assign the most likely genotype [144].

Variant annotation tools

Variants are annotated to help identify a causal variant for a genetic disease [145]. Following calling, variants can be annotated with tools such as ANNOVAR [146], VEP [147] and SNPeff [148].

Variants are initially annotated by the chromosome, base location and quality scores. Further annotation allows the variants to be annotated against a range of public databases including variant type, context and further information on the consequence. The variants can also be annotated using minor allele frequencies such as 1000 Genomes

Project [149, 150], Exome Sequencing Project (ESP) [151] and the Exome Aggregation Consortium (ExAC) [152] and pathogenicity scores such as SIFT [153] and GERP++ [154] to supplement variant prioritisation.

1.4.3.5 Whole exome sequencing

Whole exome sequencing (WES) is an NGS method of targeting the protein coding regions of the genome - the exome (Figure 1.7). This involves the fragmentation of gDNA which then undergoes library preparation with enrichment. The captured library preparation of the exome is then sequenced and data is analysed. It is now a widely used method for causal variant and causal gene identification in diseases [155, 156, 157]. Prior to WES, molecular diagnostics had only been carried on a few loci [158]. However, the development of faster, cheaper and higher resolution NGS has enabled an alternative approach to screen all genes in one project [158].

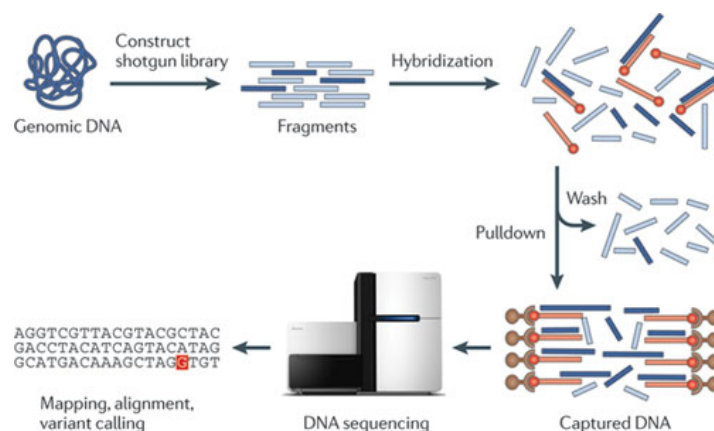


Figure 1.7: *Next generation WES sequencing method.* Genomic DNA is extracted, fragmented and ligated to short DNA sequences complementary to the oligonucleotide used in the amplification and the sequencing steps. The bound DNA fragments are ‘pulled-down’ with biotin-streptavidin-based amplification. NGS then takes place. Once the sequence is aligned to a reference genome, variants can be called [145] (extracted from Bamshad et al [159]).

WES applications in diagnostics have been able to successfully resolve genetic aetiology for 25-48% of patients [160, 161, 162]. WES has proven more successful at resolving patient aetiology than other genetic analyses utilising methods such as Sanger sequencing of single genes and chromosomal rearrangements [161]. WES analyses have also allowed for rediagnoses due to incorrect phenotyping of patients which was instigated by identifying a variants in genes causal for different diseases [156].

1.4.3.6 Whole genome sequencing

Whole genome sequencing (WGS) is the complete sequencing of all genomic regions of organisms [163]. It has the advantage of capturing all variants including non-coding variants which may be missed by methods such as whole exome sequencing (chapter 1.4.3.5) or custom targeted sequencing (chapter 1.4.3.8) which only captures a small fraction of the genome.

Reduced costs of NGS have made projects such as the 1000 Genomes project [149, 150] and 100,000 Genomes project possible. The 1000 Genomes Project (2008 to 2015) quantifies human variants of at least 1% frequency in African, American, European, South Asian and East Asian populations. The 1000 Genomes Project sequenced 2504 individuals from 26 different populations using both whole genome sequencing and exome sequencing. The 100,000 Genome Project, which has a more clinical focus, has sequenced 100,000 genomes from approximately 85,000 people and aims to aid molecular diagnoses of patients.

WGS requires extensive time consuming analysis, intensive computation power and demanding storage space. It is estimated that 85% of disease causing mutations are

found in the exome or canonical splice sites [159, 155, 156]. Therefore, targeted and whole exome sequencing (WES) offers cheaper and more direct approaches for analyses than WGS. However, further understanding regulation and the contribution of non-coding mutations to human disease could lead to an emphasis and shift to WGS analyses [164].

1.4.3.7 Somatic mutations

Somatic mutations are mutations which occur within a cell that can be passed on during cell proliferation. Somatic mutations can occur in any cells types except the germ line cells and are not passed on to an individual's offspring. Such variants may arise during development to cause mosaicism or may occur later in life. Clonal hematopoiesis of indeterminate potential (CHIP) is defined as the expansion of hematopoietic (immature cell) clones carrying somatic mutations in several genes. Individuals with CHIP have an increased risk of mortality and coronary heart disease. Jaiswal *et al* concluded that somatic mutations in hematopoietic cells contribute to development of atherosclerosis and coronary heart disease [165]. Therefore, whilst somatic mutations are not analysed in this thesis, it is important to consider that somatic mutations could influence Mendelian and non-malignant disease. There are examples of ophthalmic diseases which are believed to be caused by somatic mutations, including Coats' disease which can lead to secondary forms of glaucoma and cataracts [166].

1.4.3.8 Custom targeted sequencing

If a prospective study were to focus on a specific subset of genes and a larger cohort, utilisation of WES would not be a cost-effective method to achieve this. Rather,

targeted sequencing can offer the most efficient approach to sequence the target genes at superior depths to WES [167]. The depth is number of reads representing a base following alignment of NGS data to a reference genome (Figure 1.8). A typical WGS project would reach coverage of 30x depth per genome [168], whilst targeted sequencing can reach coverage of 500x-1000x or more overall [167]. This method of sequencing can therefore be particularly useful to detecting somatic variants present at very low allelic proportion [169].

Reference:	T G A A T A T A T C A G T C G A T C A T G T A T A C T T G C
Read 1:	T G A A T A T A T C A G T C G A T C
Read 2:	A T A T C A G T C G A T C A T G T A T
Read 3:	A G T C G A T C A T G T A T A C T T G C
Depth:	1 1 1 1 1 2 2 2 2 2 3 3 3 3 3 3 3 3 2 2 2 2 2 2 1 1 1 1 1 1

Figure 1.8: **Read depth example.** Three read sequences are shown following alignment to the reference genome. Depth is reported per base.

However, custom targeted sequencing is limited as the genes selected are subjective to the phenotypes reported for the patients studied. Therefore, if inaccurate or incomplete phenotype information is provided, important genes could be missed.

Through the use of capture kits, gene panels can be designed for targeting a subset of the genome for sequencing on NGS platforms. Current high-profile capture kits involve Agilent’s Haloplex, Illumina’s TruSeq Custom Amplicon and Nextera Rapid Capture Custom (NRCC). Illumina’s Truseq Custon Amplicon is based on an amplicon method of capture (Figure 1.9A) whereas Haloplex and NRCC uses an enrichment based method (Figure 1.9B) [169].

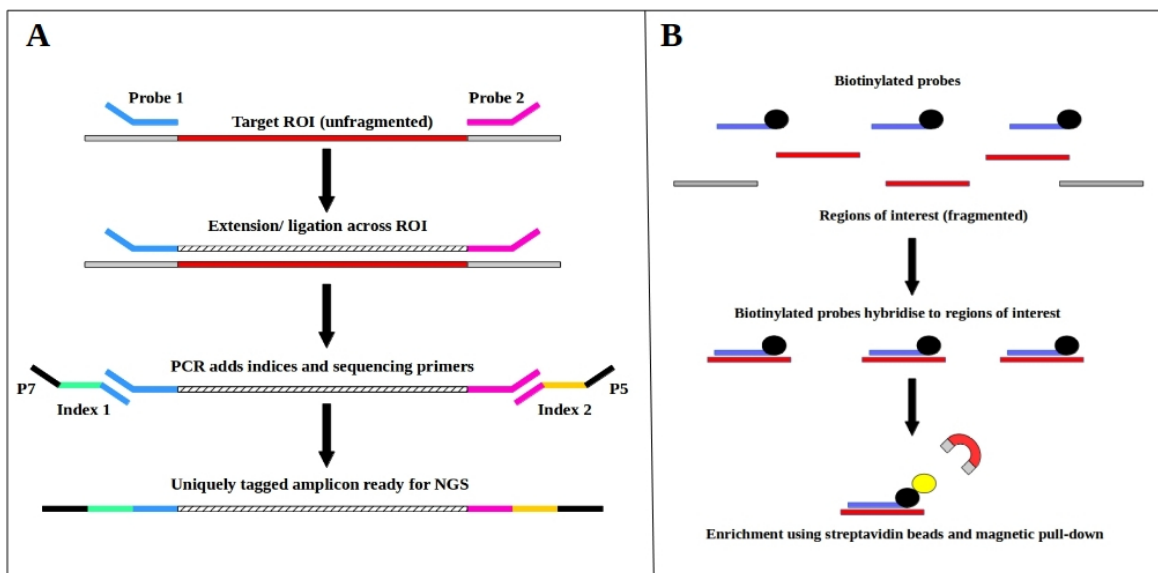


Figure 1.9: *Illumina targeted sequencing library preparation methods: amplicon (A) and enrichment (B)*. A: amplicon based sequencing amplifies regions of interest. B: enrichment works by capturing regions of interest by hybridising biotinylated probes to be magnetically pulled-down.

For TruSeq Custom Amplicon, a pair of oligo-nucleotides are designed for an amplicon target. These oligos are then hybridised to unfragmented genomic DNA which are then extended and ligated to form DNA templates of the primer flanked target region. PCR amplifies these targets and the amplicons are used as input for NGS [169].

However, NRCC uses an enrichment-based assay rather than an amplicon based-assay. Nextera-based library preparation initially involves ‘tagmentation’. In ‘tagmentation’ target DNA is fragmented and tagged by an enzyme complex of transposase and transposon ends. Following ‘tagmentation’ of the DNA fragment, PCR is performed produce an enrichment-ready fragment. The enrichment-ready fragments are denatured and biotin-labelled probes are annealed to complement the targeted region. Streptavidin beads are added, which bind to the biotinylated probes, allowing a magnetic ‘pull-down’ of the streptavidin beads to enrich the targeted regions. The enriched DNA fragments

are then eluted and a further round is completed before being used as input for NGS [169, 170]. Similarly, the predesigned capture kit, Illumina TruSight One (Illumina 5200 Illumina Way San Diego, California USA), follows this enrichment-based method. It focuses on a pre-determined subset of the exome for regions harbouring disease-causing variants of clinical relevance.

Highly curated databases can be utilised to aid the design of a custom gene panel for targeted sequencing. Such databases include the UK Genetic Testing Network (UKGTN) [171] and the Human Gene Mutation Database (HGMD) [172]. These databases contain respectively derived lists of genes associated to specific disease phenotypes.

1.5 Aims

In the introduction chapter, I have outlined a summary of the known genetic basis and the clinical characteristics of congenital nystagmus, albinism and primary open-angle glaucoma. This work aims to improve the understanding of the genetic basis of these ophthalmic diseases and help provide information to clinicians to guide molecular genetic testing in patients. Ultimately, these analyses strive to identify causal variants in patients to help contribute to accurate diagnoses.

2 Targeted sequencing and analysis of the myocilin gene (*MYOC*) in a selected cohort of primary open-angle glaucoma patients

2.1 Synopsis

This chapter interrogates the comprehensive sequencing performed on the *MYOC* gene across a cohort of homogeneously selected primary open-angle glaucoma (POAG) patients for the first time. In this work, all variant types across the intergenic, promoter, UTR, exonic coding sequences and intronic regions are described to expand on the known underlying role of *MYOC* in POAG.

Luke O’Gorman performed the design of the gene panel, the processing and interpretation of NGS data. Angela Cree was a supervisor and performed an in-depth analysis of splice region variants in *MYOC*. Helen Griffiths performed library preparation and sequencing of all samples. Prof Sarah Ennis and Dr Jane Gibson acted as main supervisors overseeing the project and provided guidance in the analysis and interpretation of the data. Prof Andrew Lotery provided supervision and clinical guidance for the project.

2.2 Background

Glaucoma

Glaucoma is a progressive ophthalmic disease and accounts for 7.9% of blindness in the UK [173]. It is characterised by a progressive loss of retinal ganglion cells, atrophy of the optic nerve and degradation of the visual field [174]. Glaucoma is presented in multiple

forms, however, primary open-angle (POAG) is the most common form [56, 57].

Primary open-angle glaucoma

Primary open-angle glaucoma (POAG) is characterised by an open anterior chamber angle, damage of the optic nerve and visual field loss [56, 57]. HTG POAG is further characterised by raised intra-ocular pressure (IOP), whereas NTG has normal IOP. POAG is known to affect at least 1% of Caucasians in the UK over the age of 40 years old [59, 60]. For more detail on glaucoma and the spectrum of its forms, please refer to Chapter 1.3.

It has historically been difficult to identify causal genes due to the complexity of the molecular genetics of POAG. It is currently known that $\sim 5\%$ of POAG is accounted for by monogenic, Mendelian-like mutations located mainly in the myocilin (*MYOC*) or optineurin (*OPTN*) genes [77]. The majority of POAG cases are assumed to be accounted for by combined effects of both genetic heterogeneity and non-genetic risk factors [77]. IOP is considered the most important risk factor in POAG, however, important risk factors in POAG include race, refractive error, central corneal thickness and a family history of POAG (for more information on the risk factors, examination and diagnosis of POAG, please refer to Chapter 1.3).

Myocilin function and expression

The myocilin gene, *MYOC*, encodes the protein myocilin which is believed to have a role in cytoskeletal development and regulation of IOP [88]. *MYOC* is expressed ubiquitously within the eye including the trabecular meshwork, and is also known as the trabecular meshwork glucocorticoid-inducible response protein (*TIGR*) [87, 88]. Mutations in *MYOC* have also been identified as the cause of hereditary juvenile-onset open-angle glaucoma (JOAG) [70, 74, 83, 175].

Myocilin expression within the trabecular meshwork and ciliary body have led to speculation that it causes an increase in IOP through obstruction of the aqueous outflow [83]. It has also been expressed at similar levels in a range of organs and tissue including the heart, skeletal muscle and bone marrow amongst others [86].

Myocilin structure

Myocilin is located on chromosome 1 and is encoded on the negative strand. *MYOC* is comprised of three exons which consist of 785bp (exon 1), 126bp (exon 2) and 604bp (exon 3) respectively. *MYOC* currently has one known RefSeq transcript (NM_000261.1) [176, 177] and encodes a 504 amino acid polypeptide (Figure 2.1) which consists of an N-terminal helix-turn-helix domain and two coil-coils. Myocilin also has a leucine zipper which can interact with other leucine zippers allowing MYOC to dimerise or form oligomers with itself [178]. The C-terminal olfactomedin-like domain [84] is part of a family of mucus proteins which are mainly found in nasal mucus [179].

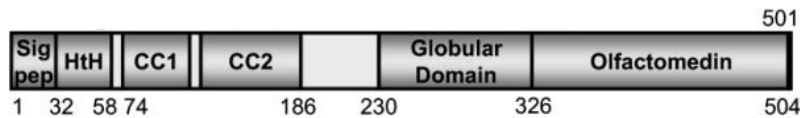


Figure 2.1: *Protein structure of myocilin (adapted from Resch et al [180]).* Secondary structures found in myocilin include a helix-turn-helix (HtH) domain, two coil-coil (CC) domains, and a C-terminal globular domain (230-504) that contains the olfactomedin-like domain (326-501), a protein family involved in many diverse functions.

Variants identified in myocilin for glaucoma

The majority of pathogenic *MYOC* mutations are found in exon 3 [79, 85] where the most prevalent pathogenic mutations are found to have a penetrance of up to 90% in POAG cases [86]. However, more recently a lower penetrance was reported for the common causal glaucoma variant, Q368*, in European populations with glaucoma

[181, 182]. Although it remains a high-risk effect size variant for advanced glaucoma, it is not necessarily always disease-causing [182]. A range of glaucoma-causing mutations have been identified and can be found in databases including the ‘Myocilin allele-specific glaucoma phenotype database’ [89]. Within this database, exons 1, 2 and 3 have 32, 1 and 62 known glaucoma-causing variants respectively. The promoter, intronic and intergenic regions do not currently harbour any known glaucoma-causing variants [89]. The database reports disease-causing mutations comprising of missense (83.7%), nonsense (5.8%), <21bp deletion (4.8%) and <21bp insertion (4.8%). The Exome Aggregation Consortium (ExAC) [152] scores the probability of *MYOC* being loss-of-function (LoF) intolerant (pLI) as 0.00. This indicates that *MYOC* is tolerant for loss-of-function. Due to a lack of glaucoma phenotype present in knockout mice for *MYOC*, haploinsufficiency has been excluded as a disease mechanism underlying *MYOC* associated POAG [183]. Under or over-expression of normal myocilin also appears not to cause POAG [90]. Shepard *et al* suggested that POAG patients with *MYOC* mutations, a gain-of-function has been suggested as the likely cause of POAG and concluded Y437H mutations in human *MYOC* induce exposure of a cryptic amino acid sequence which interacts with peroxisomal targeting signal-1 receptor (PTS1R) to elevate IOP [184].

2.3 Aim

In this study, the full region of the myocilin gene assessed in a cohort of 358 individuals with POAG selected from the UK. Through the use of next-generation sequencing (NGS), application of bioinformatic tools and strategic filtering of variants, mutations in the intergenic, promoter, UTR, exonic and intronic regions are reported for the first time.

2.4 Methods

2.4.1 Patient selection

Patients over the age of 40 were recruited across Wessex, West Midlands, Devon and Cambridgeshire regions (UK) with ethical approval provided by the Southampton and South West Hampshire Local Research Ethics Committee (05/Q1702/8). Patient data was collected for the sex, ethnicity, family history of POAG, specific diagnosis of the patient, the age at diagnosis, intra ocular pressure (IOP), cup:disc ratio (CDR), central corneal thickness (CCT) and visual field mean deviation (VFMD). Initially, 372 patients were selected for next-generation sequencing (NGS) from a glaucoma database of 1679 individuals. The filter criteria to select patients included: Caucasian ethnicity, $21 \text{ mm Hg} \leq \text{IOP} \leq 40 \text{ mm Hg}$, cup:disc ratio ≥ 0.6 and visual field mean deviation ≤ -3 . We specifically included ten first degree relatives and reduced the inclusion criteria for one relative to allow for possible studies of these interesting familial cases (see Appendix Table A.1 for clinical details of the ten family member pairs).

2.4.2 Target selection

Genes were selected using the HGMD web-interface browser and using ‘POAG’ as a search term phenotype. Additional genes which were found to be associated with POAG in Caucasians were selected. This was achieved by filtering for ‘genetics’ and ‘POAG’ in titles/abstracts of review articles in PubMed five years preceding the study (2011-2016). A gene list of 66 POAG associated genes were prioritised for inclusion on a targeted sequencing panel which included the entirety of *MYOC* gene including intronic, exonic, UTR and promoter regions (Table 2.1) and an additional 1000bp upstream of the Eukaryotic Promoter Database defined promoter coordinates (hg19

coordinates: chr1:171621818-171621877) [185, 186]. Promoters are defined as DNA structures containing a complex array of *cis*-acting regulatory elements required for accurate and efficient initiation of transcription and for controlling expression of a gene. It is also possible that multiple promoters and transcription start sites are used in gene expression [187]. However, only one known promoter is identified in the Eukaryotic Promoter Database (EPD) [185]. More detail on the capture design of the remaining 65 genes captured can be found in Chapter 3. Twenty-four SNP base pair locations defined by Pengelly *et al* (2013) [188] were included in the design to enable the checking of sample provenance. In total, this formed a cumulative target region of 297,304bp which is detailed in Chapter 3, however, this Chapter focuses exclusively on the complete *MYOC* gene region.

Table 2.1: *MYOC promoter, intronic, exonic and intergenic region locations within hg38 human reference genome.*

	Chromosome	Start	End	Length (bp)
Promoter	1	171652678	171652737	59
Prom-5'UTR	1	171652633	171652678	45
5'UTR	1	171652611	171652633	22
Exon 1	1	171652007	171652611	604
Intron 1	1	171638722	171652007	13285
Exon 2	1	171638596	171638722	126
Intron 2	1	171636709	171638596	1887
Exon 3	1	171635924	171636709	785
3'UTR	1	171635416	171635924	508

2.4.3 Quality control on kit design

The Illumina concierge service optimised the NRCC design by performing a quality control (QC) on an initial round of oligo probes that were designed in Illumina Design Studio. Following the QC, further design and synthesis of an iterative ‘round’ of oligos were performed to minimise inadequately covered regions identified from the initial rounds. Designs were tested and Picard coverage metrics (<https://broadinstitute.github.io/picard/>) were recorded with BWA-mem [141].

2.4.4 Targeted sequencing

DNA samples were enriched using the Nextera Rapid Capture Custom (NRCC) capture platform (Illumina 5200 Illumina Way San Diego, California USA). 2,262 probes were designed on hg19 through the Illumina Design Studio (<https://designstudio.illumina.com/>) with ‘dense’ (120bp) probe spacing. Three batches of 96 samples and one batch of 84 samples were run on the NextSeq Mid Output (300 cycles), providing 150bp paired-end sequence data.

2.4.5 Bioinformatic pipeline

The steps involved from processing the FastQ files of the POAG samples to shortlisting candidate causal variants are depicted in Figure 2.2. Base call intensity files were de-multiplexed for each sample and converted into FastQ files. These FastQ files were aligned against the human reference genome (hg38) using BWA-mem [141]. Variant calling was performed using GATK v3.6 [144] and ANNOVAR [146] was applied for variant annotation (Figure 2.2). Annotation was performed against a database of RefSeq transcripts [177], dbSNP v144, Exome Aggregation Consortium (ExAC) [152] and conservation-based pathogenicity scores of SIFT [153], PhyloP, PhastCons [189] and GERP++ [154]. Further annotation was performed by incorporating the ‘Myocilin allele-specific glaucoma phenotype database’ [89] (accessed December 2016). Variants were excluded if they had a read depth < 4.

Copy number variation (CNVs) were detected using CNVkit (v0.8.5) [190] for all 66 targeted genes in all samples. Mean read depths were calculated and binned (default bin size of 267bp) before normalisation using a pooled reference for each batch. Copy number calls for each sample were made using segmented log₂ ratios with the default CNVkit thresholds for gains and loss of copy [190]. CNVKit assigned copy numbers

(CN) as either CN=0 (double deletion), CN=1 (single deletion), CN=2 (wild-type), CN=3 (single gain), or CN=4 (double gain).

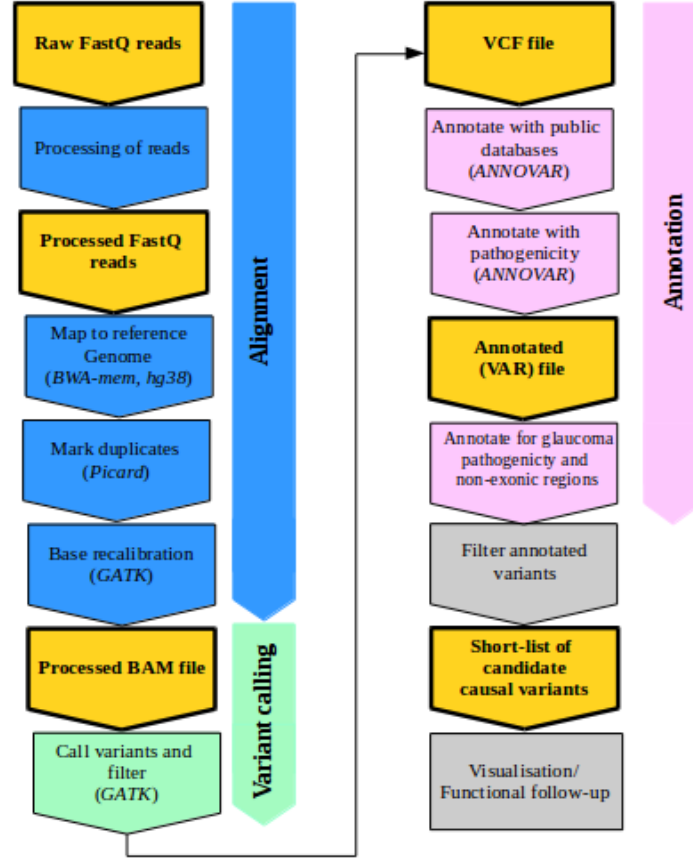


Figure 2.2: *Flow chart of the steps involved in the bioinformatic pipeline. Alignment of raw FastQ data to hg38 Human reference genome (blue), variant calling (green) and annotation with a range of public databases (pink). This is followed by prioritisation of variants for functional follow-up. Orange boxes depict key intermediate files.*

2.4.6 Quality control of NGS data

Quality control was performed on the aligned data to assess coverage. A per gene coverage analysis was conducted using the NRCC Illumina Design studio BED file and SAMtools ‘depth’ [143]. Variant sharing between the targeted NGS samples was checked for consistency with sample relationships and ethnicities. VerifyBamID [191]

software was used to determine the degree of non-reference bases and excessive heterozygosity observed across reference sites and so provide an estimate of the possibility of contamination. The resultant ‘freemix’ value was examined and a threshold value of >0.02 was applied [192].

2.4.7 Variant contextualisation

A *MYOC* graphical gene structure was constructed using the R package ‘sushi’ to represent the key metrics assessed across the gene as follows:

Per base depth was calculated using all samples’ BAM files within each batch using the SAMtools ‘depth’ option [143]. Median depth was calculated at each base across the all samples.

Mappability tracks of 20mer sequence uniqueness were obtained from UCSC. A value of 1 indicates that the sequence was completely unique and a value of 0 indicates that the sequence occurs 4 or more times in the genome [193].

In order to observe the ‘genomic state’ accross the *MYOC* gene, ENCODE annotation derived from genome-segmentation results were included [194]. This annotation included integrated ENCODE data from ChIP-seq, DNase-seq, and FAIRE-seq data using a Hidden Markov Models (HMM). Annotations included transcription start sites region (Tss), transcription start site flanking regions (TssF), strong enhancer candidates (Enh), strong enhancers flanking open chromatin (EnhF), weaker enhancer candidates (EnhW), weaker enhancers flanking open chromatin (EnhWF), distal CTCF/candidate insulators (CtcfO/Ctcf), low activity proximal to active states (Low), repressed regions (Repr), weaker repressed regions (ReprW) and quiescent states (Quies).

Repetitive regions were extracted from the UCSC table browser’s RepeatMasker database

which screens DNA sequences for interspersed repeats and low complexity DNA sequences. RepeatMasker includes short interspersed nuclear elements (SINE), long interspersed nuclear elements (LINE), long terminal repeat elements (LTR), DNA repeat elements (DNA), simple repeats (micro-satellites), low complexity repeats, satellite repeats and RNA repeats [193].

Rare variants (AF<5%) identified within the POAG cohort were plotted. Similarly, the known rare variants within 1000gEUR (AF<5%) were plotted as a control comparison. Hotspots of variation within *MYOC* were plotted by presenting all variants identified in 1000gEUR [149, 150].

2.4.8 Prioritisation of variants

Variants were considered as known glaucoma-causing variants if they were identified as a ‘Glaucoma-causing’ variant in the ‘Myocilin allele-specific glaucoma phenotype database’ or ClinVar. Coding region variants were prioritised with Combined Annotation Dependent Depletion (CADD) Phred scores greater than 15 [195]. CADD is a pathogenicity prediction score based on a machine learning algorithm which integrates multiple annotations into one metric. For improved interpretability, the scores are transformed into a Phred scores [196]. Exonic splice variants were prioritised if they exceeded a 0.5 MutPred Splice score threshold [197]. MutPred Splice is a machine learning based tool used to score coding region splice-altering variants [197].

Non-coding variants were prioritised using the Functional Analysis through Hidden Markov Models tool (FATHMM) based on its suitability for annotating non-coding variants [198]. Variants were filtered on a FATHMM score default threshold of ≥ 0.5 [198].

2.5 Results

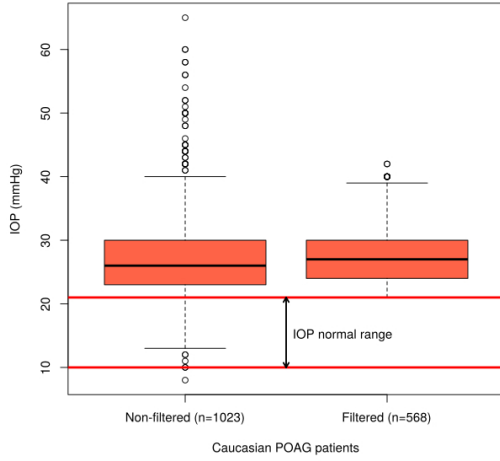
2.5.1 Patient clinical traits

Prior to filtering of the glaucoma cohort, there were 1679 patients with a diverse range of glaucoma diagnoses which included 1348 POAG, 256 NTG and 75 other diagnoses of various forms of glaucoma. Of the 1348 POAG patients, a total of 1033 patients were reported as Caucasian in the clinical database and retained. The 1033 patients were then filtered on the basis of phenotype to select those with $21 \text{ mm Hg} \leq \text{IOP} \leq 40 \text{ mm Hg}$, cup:disc ratio ≥ 0.6 and visual field mean deviation (VFMD) ≤ -3 . The number of remaining patients totalled 569. The 569 patients were further characterised on the basis of IOP, CDR, CCT and VFMD (Figure 2.3). Patients were prioritised on the basis of most severe VFMD for NGS analysis.

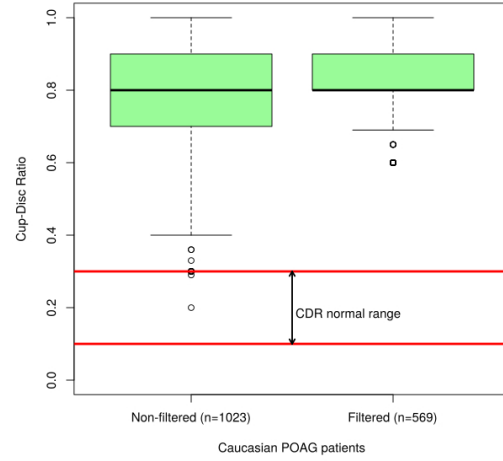
The 372 POAG patient samples which were sequenced (using three batches of 96 samples and one batch of 84 samples) had mean clinical traits of Age: 66, IOP: 27.9 mmHg, CDR: 0.82 and VFMD: -14.52 as summarised in Table 2.2.

Table 2.2: *Demographic and clinical characteristic summaries of age, intraocular pressure (IOP), cup:disc ratio (CDR) and visual field mean deviation (VFMD) for POAG cohort (n=372).*

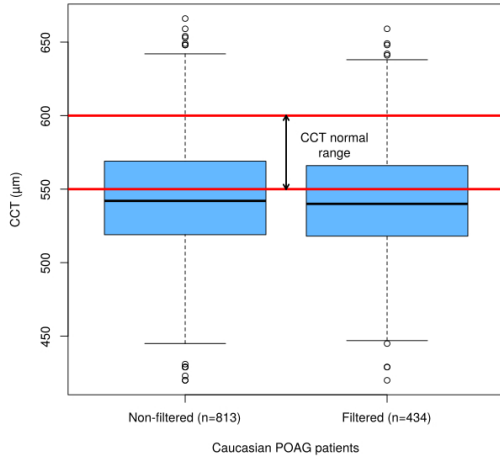
	Mean	Median	SD
Age (years)	66	66	11
IOP (mmHg)	27.9	27.0	4.7
CDR	0.82	0.80	0.09
VFMD	-14.52	-14.51	7.68



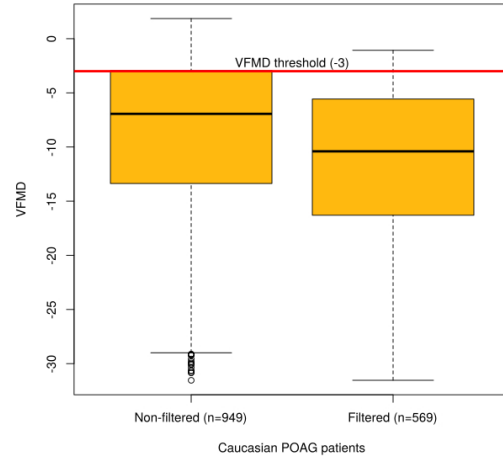
(a) *Intraocular pressure (IOP).*



(b) *Cup:disc ratio (CDR).*



(c) *Central corneal thickness (CCT).*



(d) *Visual field mean deviation (VFMD).*

Figure 2.3: *Comparison of sub-phenotypic clinical characteristics of POAG. Patients who were Caucasian with a ‘POAG’ diagnosis ($n=1033$) were compared with a subset which had $21 \text{ mm Hg} \leq \text{IOP} \leq 40 \text{ mm Hg}$, cup:disc ratio ≥ 0.6 and visual field mean deviation (VFMD) ≤ -3 (and ten family member pairs which did not meet all sub-phenotypic criteria, $n=569$).*

2.5.2 Quality control

2.5.2.1 Quality control on kit design

The concierge service provided metrics to summarise the final quality control of the NRCC test run (Table 2.3). It was found that the uniformity of the data (for which $> 0.2X$ of the 2108X mean was found for the target bases) was 95.79%. Picard metrics indicated that the mean coverage was 2108X for the target regions and 85.13% of total bases enriched were mapped to the target region. The metrics also reported that 99.80% of target bases had a depth of at least 20X.

Table 2.3: *Quality control summary test run provided by the concierge service.*

Metric	Result	Definition
Uniformity	95.79%	% bases covered at $>0.2X$ of mean
Enrichment	89.00%	% of padded reads on target
Dropouts	0.00%	% of targets covered at $<1X$
Mean coverage (Picard)	2108X	Average coverage of target
On Target Bases (Picard)	85.13%	% of enriched bases that are on target
Target Bases covered at 20X (Picard)	99.80%	% of targeted bases that are covered at $>20X$

2.5.2.2 Quality control of NGS data

Coverage and contamination

For the 372 sequenced patient samples, coverage across both the *MYOC* gene and all targets were determined (Table 2.4). Depth was highest for batch 1 samples (720X) and lowest in batch 3 (100% reagents) samples (381X). Batch 4 mean average depth was high at 717X, however, batch 4 had the greatest standard deviation of depth. A batch 4 sample (GB057) had no coverage at 100X. Freemix indicated only minor contamination as no sample exceeded the 0.02 threshold for possible contamination. There was no excess heterozygosity in any batch. Gender was not determined due to a

lack of coverage and variants on the X chromosome.

Shared variation

Unexpected relatedness was identified by comparing each samples' variant calls in the coding region (Figure 2.4). Shared variation between unrelated Caucasian individuals averaged at 61.0% concordance. Upper and lower boundaries were defined as extreme outlying boundaries (1st Quartile - 3IQR) to (3rd Quartile+3IQR) which translates to 35.0% - 87.1%. This broad range was necessary as the target region for 66 genes was small and difference in coverage between samples (and therefore, number of variants compared) can have a greater impact on error and variability. The extreme outlier upper threshold of 3rd Quartile+3IQR was used to detect cases of abnormal shared variation for very closely related individuals or sample duplications.

Extreme high levels of shared variation

Three pairs of samples exhibited elevated levels of shared variation exceeding the extreme outlier threshold (87.1%). A suspected duplication of samples was identified with sample pairs GL464 & GLF001 and FG418 & GL418. This was later confirmed when following up patient clinical records. In such cases, the sample with the least coverage at 20X depth was omitted from downstream analyses. A pair of highly similar samples (89.8-92.2%) were identified (QGF015 & QGF016) which were confirmed as monozygotic twins by patient records. Monozygotic twins shared up to 92.2% of their genetic variants. Due to differences in coverage between the samples, the shared variation between the monozygotic twins was not 100%. As the target region was relatively small, modest differences in coverage between samples would incur larger effects in calculating shared variation.

Table 2.4: **Summary of quality control statistics for all four batches.** Each of the four batches are displayed over five rows (including 72 patients in batch 3 with 100% of reagents used and 12 patients in batch 3 with 50% of reagents used). Coverage at 1X, 5X, 50X and 100X depth shows that batch effects exist with high variation of data quality within each batch. There was no conclusive evidence for contamination when observing heterozygosity % and VerifyBamID freemix.

Batch	No. of samples	Coverage										Contamination			
		Mean depth	Min depth	Max depth	SD depth	Mean % 1X	Mean % 5X	Mean % 50X	Mean % 100X	Min % 100X	Max % 100X	Mean % hets	Min % hets	Max % hets	Max Freemix
1	96	719.81	434	1359	123	99.98	99.88	99.02	98.24	96.47	99.29	61.85	50.62	72.24	0.0082
2	96	554.87	263.81	828	138	99.97	99.87	98.93	97.94	95.21	98.76	61.82	45.71	71.32	0.0033
3 (100%)	72	380.76	146	983	132	99.94	99.75	97.94	95.04	80.94	98.79	61.41	45.59	73.12	0.0036
3 (50%)	12	437.82	307	627	100	99.93	99.69	98.00	96.05	93.16	98.34	64.93	54.29	74.21	0.0010
4	96	717.38	11	1068	191	99.98	99.84	98.29	97.65	0.00	99.34	60.35	48.06	70.87	0.0036

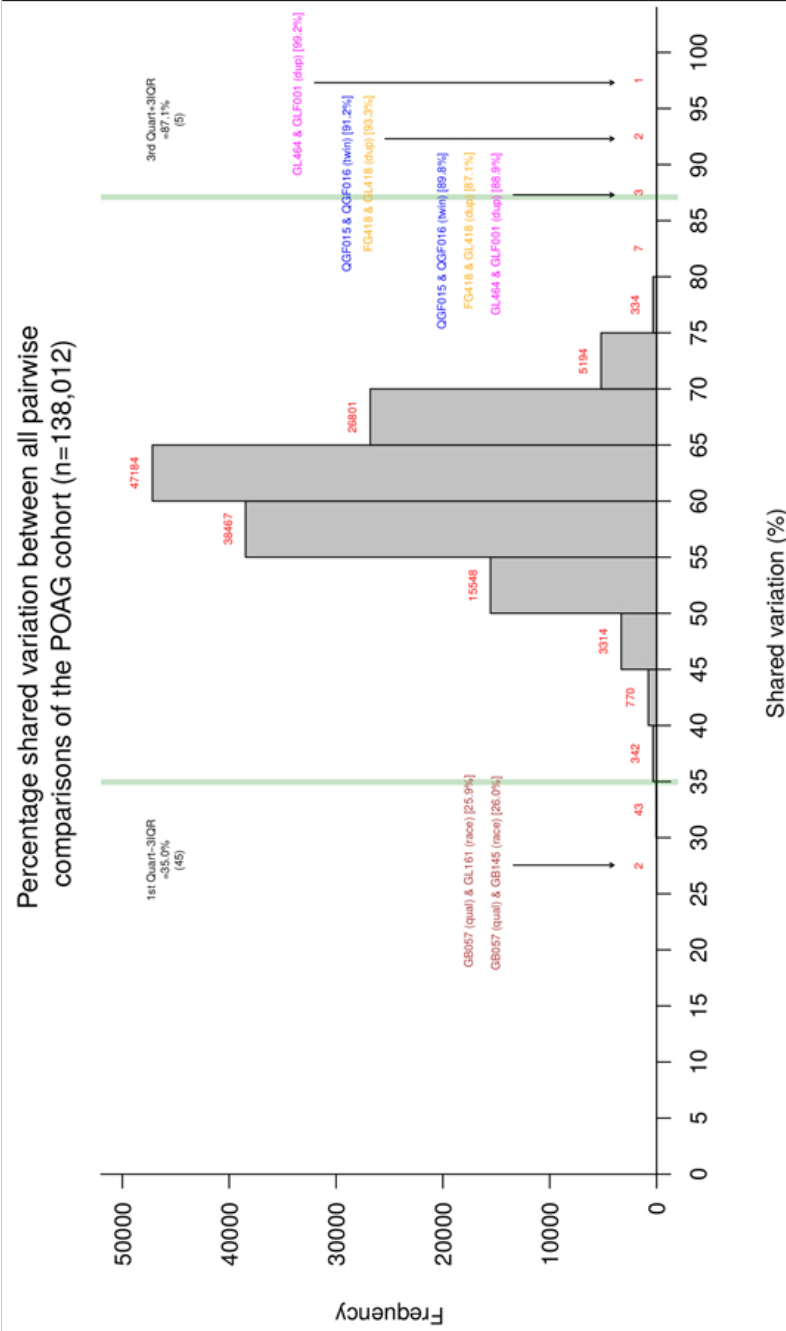


Figure 2.4: **All pairwise comparisons between samples (n=138012).** Reference lines indicate the thresholds for extreme outliers ('abnormal' shared variation variation).

Lower levels of shared variation

Samples GL161 and GB145 were identified as having 30-45% shared variation with the cohort. This range was partially under the threshold for extreme low shared variation outliers. Sample GL145 was confirmed by patient notes to be mixed race (Caribbean/Caucasian). GL161 and GB145 share 63.0% shared variation which was close to the mean average for Caucasian-Caucasian shared variation for unrelated patients (61.0%) which suggested that GL161 was possibly also mixed race.

Extreme low levels of shared variation

Sample GB057 had two instances of extremely low outlying shared variation (25.9-26.0%), however, this was due to substantial differences in coverage.

A further eight samples belonged to patients for which ‘Age at Diagnosis’ was under 40 and, therefore, did not comply with ethics exclusion criteria.

Summary

Overall, 14 samples were omitted for downstream analysis. One sample (GB057) was omitted due to insufficient depth ($\leq 20X$). For instances where sample duplications or monozygotic twins were identified, the sample with lower coverage at 20X depth across the target region was omitted (QGF015, GL464 and GL418). Two samples were omitted for not conforming to the selected Caucasian ethnicity (GL161 and GB145). A further eight samples were omitted due to the patient age at diagnosis being less than 40 years (FG010 FG396, GB026, QGF027, QGF077, GL459, GL572 and GL439). Following the omission of the 14 samples, the minimum depth is improved from 11X to 185X and contamination metrics were unchanged (Table 2.5).

Table 2.5: *Quality statistics before sample omissions and after sample omissions. Following sample omissions, the minimum depth is improved.*

QC	Number of samples	Coverage		Contamination				
		Mean depth	Min depth	Mean % 100x	% 100x Min	% hets mean	% hets max	Freemix Max
Prior to QC	372	600	11	97.30	0.00	62.1	74.2	0.008
Post QC	358	566	185	97.27	88.74	62.0	74.2	0.008

2.5.3 MYOC variant contextualisation

The coordinates of the *MYOC* gene (5' promoter - 3' UTR) are chr1:171,652,737-171,635,416 (hg38). Coverage across *MYOC* was determined to be 100% at 20X depth for all samples. The mean depths within the four batches were 717X, 552X, 389X, and 726X across target regions respectively. The poorest coverage was observed in batch 3 which was speculated to be caused by erroneous tagmentation and clean up steps in the Nextera Rapid Capture enrichment library preparation. A consistent coverage pattern was seen for all four batches (Figure 2.5B) and was found to be correlated with mappability, repetitive context, conservation and GC content ($R^2 = 0.3358$, p-value $< 2.2 \times 10^{-16}$). Conservation scores derived from PhastCons (Figure 2.5C) and PhyloP (Figure 2.5D) were highest across exonic regions, with smaller regions of high conservation in intron 1 and a region upstream of the Eukaryotic Promoter Database (EPD) defined promoter region [185]. Figure 2.5E shows that repetitive regions are not located on the exonic regions of *MYOC* and are mainly found on intron 1. Figure 2.5F highlights regions far upstream of the EPD defined promoter [185] which are predicted to be candidate enhancing regions. These regions were captured, however, they were not selected as regions of interest to interrogate.

CHAPTER 2

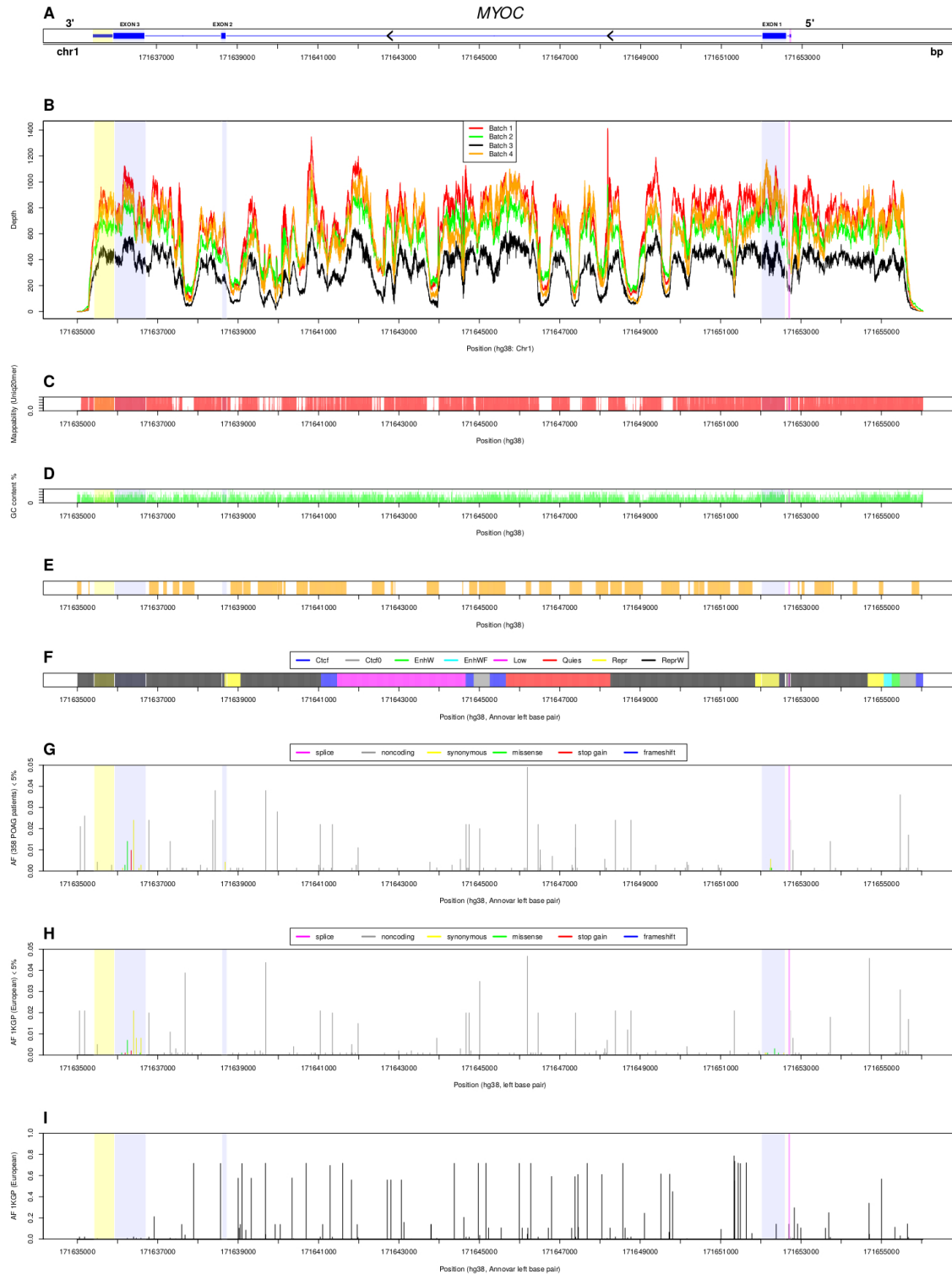


Figure 2.5: *Per base analysis of variants and regions across the MYOC gene.* *MYOC* gene structure (A), Per base read depth (B), 20mer windows of uniqueness (1=unique, 0=occurs more than 4 times in the genome) (C), GC content in 5-base windows (D), repetitive region (RepeatMasker) (E), ENCODE chromatin state segmentations (F), less than 5% AF of variants across the POAG patient cohort (n=358) (G), less than 5% AF of variants identified across 1000gEUR population (H), all AF of variants identified across 1000gEUR population (I).

Table 2.6: *Summary of the number of variants identified across all features of the MYOC gene.*

Gene feature	No. variants		
	No. unique variants	SNPs	Indels
Intergenic us	21	21	0
Promoter	2	2	0
Promoter-5' UTR	0	0	0
5' UTR	0	0	0
Exon 1	4	4	0
Intron 1	118	110	8
Exon 2	1	1	0
Intron 2	15	12	3
Exon 3	7	7	0
3' UTR	2	2	0
Intergenic ds	2	1	1
All	172	160	12

A total of 172 annotated variants were identified comprising 160 SNPs and 12 indels in the POAG cohort of 358 individuals (Appendix Table A.2). These variants were distributed across the myocilin gene with 21 variants upstream intergenic, two variants in the promoter region, four in exon 1, 118 in intron 1, one in exon 2, 15 in intron 2, seven in exon 3, two in the 3' UTR and two variants in the downstream intergenic region (Table 2.6). 156 SNPs were identified in the non-coding regions of the *MYOC* gene in the POAG cohort. The majority (70.5%) of non-coding variants were located in the largest intron, intron 1, which spans 13,285 bp (76.7%) of the 17,321 bp length of *MYOC*. For comparison, there were 574 total SNPs in the 1000 Genomes Project European population (1000gEUR) across the *MYOC* gene. There were 105 rare (AF<5%) variants in POAG and 134 rare (AF<5%) variants in 1000gEUR. Both variants called across the POAG cohort and in 1000gEUR had a similar distribution across the *MYOC* gene (Figure 2.5F & G).

2.5.4 Exonic variants

A total of 12 exonic variants were called (Table 2.7). Four SNPs were detected in exon 1, one SNP in exon 2 and seven SNPs in exon 3. Four variants had CADD

Phred scores greater than 15 suggesting the variants were possibly pathogenic. Variants NM_000261:exon3:c.C1102T (p.Q368*) and NM_000261:exon1:c.C376T (p.R126W) had previously been identified as ‘Glaucoma-causing’ by the ‘myocilin allele-specific glaucoma phenotype database’. Variant p.R126W also had a high MutPred splice score of 0.605, indicating that this variant was likely to affect splicing.

Two variants NM_000261:exon2:c.G648A (p.K216K) and NM_000261:exon3:c.A1255G (p.T419A) not previously identified as glaucoma-causing, had pathogenicity scores indicating they may be of importance. In three individuals p.K216K had a high PhastCons score of 0.992 indicating that it was within a highly conserved element. The variant was also more common in the POAG cohort (AF=0.0042) than in the ExAC Non-Finnish European (NFE) population (AF=0.0005).

The missense variant p.T419A had no known rsID and was not found in ExAC NFE. In the POAG cohort it had an allele frequency of 0.0028, and identified as heterozygous in two individuals. It was located in exon 3 and had a SIFT score of 0, PolyPhen HDIV score of 0, GERP++ score of 4.52 and CADD Phred of 37 which indicate that this variant was likely to be highly pathogenic. However, this variant was found to be present on the same read pair as the p.Q368* variant in both patients (see Appendix Figure A.1).

There was no significant difference in sub-phenotypes between patients with candidate causal *MYOC* variants (p.Q368*, p.R126W, p.K216K or p.T419A) and patients with no candidate causal *MYOC* variants (t-test, IOP p-value=0.766, CDR p-value=0.626, VFMD p-value=0.211). However, hypertension was treated in five of the 11 patients with candidate causal *MYOC* variants. Therefore, a significant difference in IOP was unlikely as patients with treatment for hypertension are likely to have reduced IOP.

Three variants, p.E115K, p.G122G and p.R126W are clustered within the coiled-coil located at aa118-aa186 [199] (Figure 2.6). The aa117-aa166 region contains lysine residues responsible for dimerisation of *MYOC* [200]. The p.K216K variant was located in a linker region whilst p.T285T, p.D302D, p.Y347Y, p.Q368*, p.K398R, p.T419A and p.T438T are all located within the large olfactomendin-like domain.

Table 2.7: **Annotation of all coding sequence variants.** No., variant number; Exon, exon number; Feature, genetic feature within MYOC; Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternate allele; Variant type, type of variant observed; Amino Acid, amino acid single letter abbreviation of reference amino acid and the amino acid substituted to; dbSNP144, rs ID if the variant was known in dbSNP v144; 1000G EUR, allele frequency from 1000 Genomes Project (European ethnic sub-group); ExAC NFE, allele frequency from ExAC Non-Finnish European ethnic sub-group; Sample count, number of patients with the variant in the n=358 POAG cohort; Study AF, allele frequency of the variant within the n=358 POAG cohort; myocDB, known myocilin variants database [89]; CLINSIG, pathogenicity of the variant in ClinVar; SIFT, sorts intolerant from tolerant substitutions; gerp++, Genomic Evolutionary Rate Profiling; PhastCons, conservation scoring and identification of conserved elements; CADD Phred, Combined Annotation Dependent Depletion on a Phred scale; MutPred Splice, machine learning-based predictor of exonic splice variants. **Bold indicates variants which are causal candidates.**

No.	Exon	Chrom	POS	Ref	Alt	Variant type	Amino acid	dbSNP144	1000G EUR	ExAC NFE	Sample count	Study AF	myocDB	CLINSIG	SIFT	gerp++	PhastCons	CADD Phred	MutPred Splice
1	1	chr1	171,652,385	C	T	nonsynonymous	R76K	rs2234926	0.14120	0.13650	92	0.13700	Neutral	Benign	0.049	3.43	0.945	9.00	0.116
2	1	chr1	171,652,269	C	T	nonsynonymous	E115K	rs75751979	-	0.00003	1	0.00140	-	-	0.589	3.53	0.268	9.37	0.159
3	1	chr1	171,652,246	G	A	synonymous	G122G	rs145354114	-	0.00300	4	0.00559	Neutral	Uncertain	-	-	0.000	0.17	0.494
4	1	chr1	171,652,236	G	A	nonsynonymous	R126W	rs200120115	-	0.00007	1	0.00140	Glaucoma	-	0.019	-1.13	0.008	23.50	0.805
5	2	chr1	171,638,679	C	T	synonymous	K216K	rs141584495	-	0.00050	3	0.00419	Neutral	-	-	-	0.992	15.63	0.165
6	3	chr1	171,636,585	C	A	synonymous	T285T	rs146606638	0.00800	0.00480	2	0.00279	Neutral	Benign	-	-	0.591	14.00	0.126
7	3	chr1	171,636,534	G	A	synonymous	D302D	rs148433908	-	0.00030	1	0.00140	Neutral	-	-	-	0.000	0.07	0.137
8	3	chr1	171,636,399	A	G	synonymous	Y347Y	rs61730974	0.02000	0.03050	17	0.02400	Neutral	-	-	-	0.024	0.00	0.140
9	3	chr1	171,636,338	G	A	stopgain	Q368Ter	rs74315329	0.00200	0.00150	7	0.00978	Glaucoma	Pathogenic	-	4.52	0.283	37.00	0.374
10	3	chr1	171,636,247	T	C	nonsynonymous	K398R	rs56314834	0.00700	0.00480	10	0.01400	Neutral	-	0.618	-1.17	0.843	3.87	0.268
11	3	chr1	171,636,185	T	C	nonsynonymous	T419A	-	-	-	2	0.00279	-	-	0.000	5.04	0.945	23.50	0.196
12	3	chr1	171,636,126	G	A	synonymous	T438T	rs375235405	-	0.00004	1	0.00140	-	-	-	-	0.898	11.66	0.125

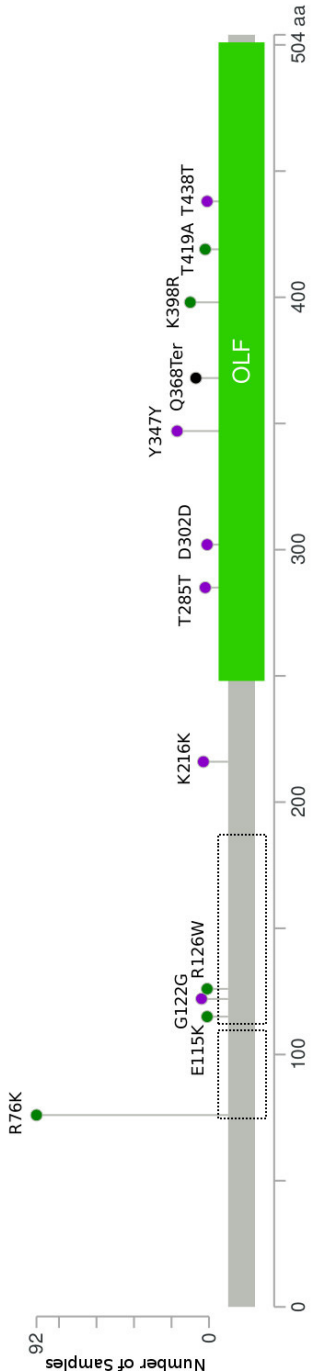


Figure 2.6: **Human Myocilin protein structure with its domains and variants mapped [201, 202].** Additional annotation of coiled-coil domains are outlined in black dotted rectangles (aa74-110 and aa118-186 [199]) and the offactomedin-like domain is highlighted as 'OLF'. Green dots indicate a missense variant, purple dots indicate synonymous variants, whilst black dots indicate stop-gains.

2.5.5 Non-coding variants

There were 160 variants identified in non-coding regions (see Appendix Table A.2). The majority of variants (118) were identified within the largest region, intron 1. Using a FATHMM threshold of 0.5 to prioritise the non-coding variants, one variant upstream of the promoter, three intron 1 and one intron 2 variants remain (Table 2.8). The highest FATHMM score of 0.86 was seen for a common variant with an allele frequency of 10% in the 1000gEUR, and similar (8.2%) in the POAG cohort. A single variant in intron 1, NM_000261.1:c.605-5949C>T, had a CADD Phred score greater than 15, however, there was no significant difference in frequency between this variant in the POAG cohort compared with the 1000gEUR cohort (allelic chi-squared test, p-value=0.716). A second intron 1 variant, NM_000261.1:c.604+5942G>A, had a CADD Phred score of 12.78, a GERP++ score of 3.57 and PhastCons score of 0.504. Although pathogenicity scores are in favour of a potentially pathogenic effect, allele frequencies show no significant difference between the POAG cohort and 1000gEUR (allelic chi-squared test, p-value=0.132). The novel intergenic upstream variant identified, NM_000261.1:c.-2851C>T, had a CADD Phred score of 11.19, a GERP++ score of 3.22 and low conservation PhastCons score of 0.0079 providing ambiguous indications of pathogenicity. This variant was not present in the 1000gEUR but was observed as heterozygous in one individual within the POAG cohort.

Table 2.8: **Five non-coding variants remain following initial filtering of the 160 non-coding variants with FATHMM ≥ 0.5 .** Feature, genetic feature within MYOC; Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternate allele; dbSNP144, rsID if the variant was known; 1000G EUR, allele frequency from 1000 Genomes Project (European ethnic sub-group); Sample count, number of patients with the variant in the n=358 POAG cohort; Study AF, allele frequency of the variant within the n=358 POAG cohort; gerp++, Genomic Evolutionary Rate Profiling; FATHMM, Functional Analysis through Hidden Markov Models; CADD Phred, Combined Annotation Dependent Depletion on a Phred scale; myocDB, known myocilin variants database [89]; Repeat, repetitive region as defined by RepeatMasker; PhastCons, conservation scoring and identification of conserved elements; Regulatory build, Ensembl Regulatory Build containing regions that are likely to be involved in gene regulation.

No.	Feature	Chrom	POS	Ref	Alt	dbSNP144	1000G EUR	Sample count	Study AF	gerp++gt2	FATHMM	CADD Phred	myocDB	Repeat	PhastCons	Regulatory build
1	INTERGENIC_US	chr1	171655462	C	T	-	-	1	0.00140	3.22	0.593	11.19	-	0	0.008	CTCF Binding Site
2	INTRON1	chr1	171,646,066	C	T	rs12035960	0.10440	55	0.08200	3.57	0.860	12.78	-	0	0.504	Promoter Flanking Region
3	INTRON1	chr1	171,644,671	G	A	rs75953590	0.01990	16	0.02200	-	0.551	15.38	-	0	0.244	-
4	INTRON1	chr1	171,643,942	G	C	rs144750384	0.00800	1	0.00150	-	0.623	10.31	-	1	0.061	Open chromatin
5	INTRON2	chr1	171,637,310	A	G	rs79263903	0.01090	10	0.01400	-	0.570	4.669	-	0	0.280	-

Twenty variants were flagged as ‘potentially splice altering’ by human splice finder (HSF) version 3.0, 18 of which have rsIDs (see Appendix Table A.2). Seventeen splice variants were detected in intron 1 (15 SNPs, and 2 insertions) and three in intron 2 (three SNPs). Of these, 19 variants have the potential to introduce a new splice acceptor/splice donor site and one potentially breaks a branch point.

No copy number variants (CNVs) were detected within the *MYOC* gene region. There were no losses in copy number (CN) across the entire region, however, some (N=113) samples were called as a single copy gain with a CN=3 across the entire region.

2.6 Discussion

Targeted next-generation sequencing had been used on the full region of the *MYOC* gene (promoter, UTRs, coding exons, introns and intergenic regions) on 358 POAG patients with the severe POAG sub-phenotypes. All variants detected across the region and were reported and analysed to assess pathogenicity. A known pathogenic stop-gain in exon 3, a known pathogenic missense variant in exon 1, a known synonymous variant not previously considered pathogenic in exon 2, and an unknown missense variant in exon 3 were identified. These variants collectively account for 11/358 (3.07%) of patients within our POAG cohort.

NM_000261:exon3:c.C1102T (p.Q368*) was the most common causal variant in POAG, accounting for 31.2% of disease-causing variants in myocilin [89]. This stop-gain was 6.5 times more common in our POAG cohort than in the 1000gEUR (allelic Fisher’s Exact test, p-value=0.0336) and was seen in seven patients, accounting for 63.6% of the candidate causal variants identified. Shepard *et al* have previously shown that *MYOC* monomers with this variant do not contain a cryptic peroxisomal signalling sequence (PTS1) and that it likely exposes the PTS1 in its *MYOC* dimer partner [184]. This was

believed to cause the mutant *MYOC* dimer to associate with the PTS1R and ultimately cause deleterious trabecular meshwork cell function [184]. This mechanism has been supported by other studies [203, 204]. Previous studies have shown that Caucasian patients with the p.Q368* variant have a mean IOP ranging between 27.3-35.4 mm Hg [79, 205, 206, 207], higher than in our study (mean IOP=26.3 mm Hg). In our study, patients with this variant had a mean CDR of 0.84, a mean VFMD of -13.84, and a mean age of 69.6 years. These findings agree with Graul *et al* who found that p.Q368* patients did not have an earlier onset nor did they have a higher IOP [205]. Similarly, Nag *et al* identified a lower penetrance of p.Q368* in ocular hypertension [181].

The known glaucoma-causing variant, NM_000261:exon1:c.C376T (p.R126W) was found in one of the 358 POAG patients and had been previously reported as a late-onset familial variant [208]. It was a variant which was located on the protein dimer region of a coiled-coil. Gobeil *et al* have shown cell adhesion properties were unaffected [203] by this variant. NM_000261:exon1:c.C376T (p.R126W) had damaging SIFT and CADD Phred scores of 0.019 and 23.5 respectively. There was very little evidence for splicing variation leading to POAG and only one previously known instance in *MYOC* of a predicted cryptic splice site reported within intron 1 [209]. However, a MutPred Splice score of 0.605 implicated that this variant contributes to the creation of a new donor splice site and a subsequent loss of 372 nucleotides from exon 1. This finding suggests that splice variants could be more important in POAG than previously known. Faucher *et al* had previously shown that patients with this variant were found to have a mean IOP of 28.3 mm Hg and an age of onset of 74 [208]. The patient with this variant in our cohort showed similar traits with a maximum IOP of 27 mm Hg and an age at diagnosis of 74 (see Appendix Table A.3).

The synonymous variant NM_000261:exon2:c.G648A (p.K216K) was not previously

considered a pathogenic variant. This variant was found in exon 2 which contains just one known pathogenic variant and was believed to translate to a linker region within *MYOC* [89, 210]. Synonymous variants in *MYOC* have been suggested to have a role affecting *MYOC* mRNA structure and subsequently the translated protein stability [211]. Variant p.L215Q, on the preceding codon of p.K216, was believed to be glaucoma-causing on the basis of an in-silico damaging SIFT score [210]. Similarly, p.K216K had strong in silico pathogenicity scores to suggest possible pathogenic status (PhastCons of 0.992 and CADD Phred of 15.63). Furthermore, this variant was found in the gnomAD Non-Finnish European (NFE) population [152] significantly less frequently than the POAG cohort (allelic Fisher’s Exact test, p-value=0.0109). This heterozygous variant was found in three patients from the University Hospital Southampton site. No evidence of relatedness was identified, however, there was a possibility that there was some distant relatedness for which it was not possible to detect.

The missense variant NM_000261:exon3:c.A1255G (p.T419A) does not have an associated rsID, nor was it found within 1000gEUR or ExAC. However, it was found at an allele frequency of 8.952E-6 in the gnomAD NFE population. This variant had never been observed in a glaucoma context before but was seen as a heterozygote in two patients in this study (AF=0.0028). This was a substantially higher frequency than gnomAD NFE (allelic Fisher’s Exact test, p-value=9.4E-6). This variant had a SIFT score of 0, GERP++ score of 5.04, PhastCons of 0.945 and CADD Phred score of 23.5 which indicated further support for pathogenicity. However, it was found that in both patients p.T419A was co-inherited with the upstream p.Q368* variant (see Appendix Table A.3), therefore the protein will be truncated before translation of the possibly pathogenic substitution. Although these two patients had the earliest onset (50 & 56 years) of those carrying the p.Q368* variant, it was not possible to provide a plausible mechanism by which this variant could have a modifying effect.

When distinguishing between candidate causal variants in the intronic regions, repetitive region context could be useful to distinguish between a more likely functionally relevant variant and a less damaging variant. Transposable elements are believed to have a vital role in accelerating evolution of genes by increasing gene versatility or by truncating the coding region [212]. Human evolution has strongly been affected by transposable elements, most namely Alu repeats [213, 214]. Lin *et al* discusses that coding region variants located in exons derived from Alu repeats could have regulatory roles in RNA metabolism including translation and degradation. Therefore, such variants can have an impact on protein function. Alu exons can also produce human-specific protein isoforms [215]. Therefore, variants within repeats should not be completely ruled out for contribution to causality. However, repeats are known to be problematic in Illumina sequencing due to AT-rich repetitive sequences causing a drop in coverage [216, 217]. Repeats also give reads less specificity and can lead to ambiguous alignment [218]. Overall, although repeats have potentially interesting variation that should be investigated, the methods employed by this study may not be best suited for analysing them.

Whilst there were no clear possibly pathogenic variants in the non-coding region of the gene, NM_000261.1:c.-2851C>T which was located in the upstream intergenic region (44694bp from the neighbouring *VAMP4* gene) was found to be of potential interest. Whilst it was not within a conserved element in PhastCons, it had damaging GERP++ and FATHMM scores. It had an allele frequency in the POAG cohort of 0.0014 (one heterozygous patient). This variant was not found within the 1000gEUR and was at a position not currently covered by gnomAD. The Ensembl regulatory build indicates that this variant could be functionally important as it was located at a potential CTCF binding site. All other variants with a FATHMM score ≥ 0.5 were seen at similar frequencies in both the POAG cohort and 1000gEUR. Genotyping this non-coding

variant across a wider POAG cohort could prove informative. Variants located up to 1000 bp upstream of *MYOC* have been implicated as potentially functionally important for controlling IOP [219, 220].

Five rare variants were present in six individuals that potentially affect splicing. However, the presence of these variants in the European population cannot be confirmed due to lack of coverage in allele frequency databases for these sites.

There was no evidence of sub-gene copy number changes, and no whole gene deletions. There were however some whole gene single copy gains and it was suspected that the predicted gain reflects within-batch depth variation.

Patient selection criteria for this study used strict sub-phenotype parameters in order to select most severe POAG sub-phenotypes with a greater chance of an accurate POAG diagnosis. However, such criteria hinders genotype-phenotype analyses within the selected cohort. Genotyping of a larger POAG cohort not selected on sub-phenotypes was necessary in order to perform robust genotype-phenotype analyses. The *MYOC* gene accounts for $\sim 3\%$ of patients with POAG, therefore a larger cohort would also have greater power to detect rarer causal variants.

Conclusion

Two known pathogenic variants and two high pathogenic scoring variants were identified which may cause POAG in 11 patients. Synonymous and non-coding variants were identified as having pathogenic qualities using in silico pathogenicity predictions, and a known glaucoma-causing variant had been implicated as a potential deep exonic splice variant. This work expands the known allelic diversity of myocilin in POAG which was useful for diagnosis, genetic counselling and cascade genetic testing in families.

Additional sequencing of *MYOC* interacting partners [221] and other POAG-causing genes could reveal rare causal variants and provide further insight into the genetic basis of POAG.

3 Next-generation sequencing analysis of 66 POAG genes across a selected cohort of severe primary open-angle glaucoma patients

3.1 Synopsis

In this chapter, all associated POAG genes are interrogated across a cohort of homogeneously selected POAG patients. The variants are reported which are most likely to be involved in the causality of POAG using single variant (within and between genes) and whole-gene pathogenicity analyses.

Luke O’Gorman performed the design of the gene panel in March 2016, followed by the processing and interpretation the NGS data. Ms Helen Griffiths performed library preparation and sequencing of all samples. Dr Roshan Sood ran samples through the standard CNVkit pipeline. Dr Enrico Mossotto processed the Inflammatory bowel disease (IBD) patient NGS data and ran all samples through GenePy/v.1.2. Prof Sarah Ennis, Dr Jane Gibson and Ms Angela Cree acted as supervisors overseeing the project and provided guidance in the analysis and interpretation of the data. Prof Andrew Lotery provided clinical guidance for the project and provided supervision of the project.

3.2 Background

In 2011 Fingert *et al* described POAG genes as either familial ‘Mendelian-like’ genes (typically identified through linkage studies) or as assumed complex genes (identified through association studies) [77]. Similar terminology has been used in subsequent

reviews and studies [78, 222, 223, 224]. The ‘Mendelian-like’ and assumed complex POAG genes are described below.

3.2.1 Mendelian-like genes

Genetic linkage analysis is used for Mendelian traits or traits typically caused by variants in a single gene. Mendelian-like POAG genes identified through linkage studies include *MYOC* [82, 83] (which was previously described in detail in Chapters 1.3.4 and 2.2), *OPTN* [225], *TBK1* [226, 227], *WDR36* [80], *ASB10* [13, 14]. Initially, the GLC3A locus (harbouring *CYP1B1* candidate gene) was found to be linked to congenital glaucoma [228] and was then further implicated in POAG [229]. The *NTF4* gene in the GLC1O locus was previously listed as a Mendelian-like candidate gene for POAG without family-based studies or linkage analysis being performed [224, 230].

Sarfarazi *et al* mapped the GLC3 locus, which harbours the *CYP1B1* candidate gene (lod score of 11.50), by using a group of 17 congenital glaucoma families [228] and was subsequently implicated in POAG [229]. Cytochrome P450 family 1 subfamily B member 1 (*CYP1B1*) catalyses reactions involved in the metabolism of compounds including 17 β -estradiol, retinals, arachidonic acid and melatonin [176, 231]. It is reported to be expressed ubiquitously in human tissues [232] and localises to the endoplasmic reticulum [176, 231]. Dysfunctional metabolism of 17 β -estradiol can cause *MYOC* up regulation, which contributes to POAG pathogenesis [233]. The metabolism of retinoic acid by *CYP1B1* (critical for ocular development) has implicated its involvement in primary congenital glaucoma (PCG) [234, 235].

The locus (GLC1E) was identified in a large British family with NTG [225] and the genetic cause was further identified as the optineurin (*OPTN*) gene [97]. *OPTN* is an adaptor protein involved in regulating vesicular trafficking from the Golgi to plasma

membrane, endocytic trafficking, and signalling leading to NF-kappa-B [236]. OPTN participates in xenophagy, aggrephagy and mitophagy through the LC3-interacting region (LIR) and the Ubiquitin-binding motif (UBAN) (Figure 3.1). It is also understood to recruit TBK1 at the Golgi apparatus [237] through its N-terminal coiled-coil domain. OPTN contains a cargo-associating domain and a LC3-interacting region which recruits ATG8 family proteins. Within the eye, *OPTN* is expressed in the trabecular meshwork, non-pigmented ciliary epithelium, aqueous humor and retina [97].

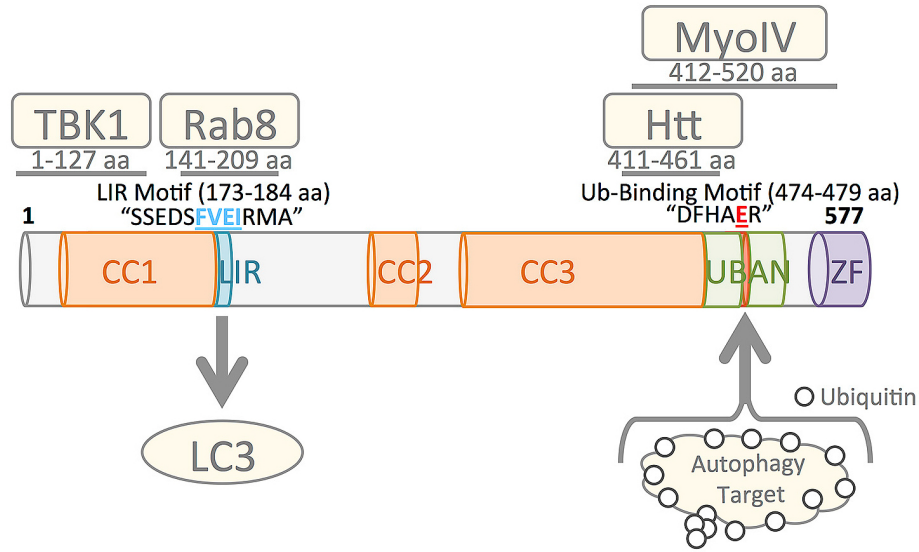


Figure 3.1: **Human OPTN protein structure and interacting binding interfaces.** The three coiled-coil (CC1, CC2 and CC3) domains comprise 70% of the OPTN protein. The LC3-interacting region (LIR) motif and the Ubiquitin-binding motif (UBAN) are shown. There are also several binding interfaces which recruit OPTN-interacting proteins such as TBK1, Rab8, Htt and MyosinIV (adapted from Minegishi et al, 2016) [238].

Fingert *et al* identified the GLC1P (further characterised as TANK-Binding Kinase 1 gene, *TBK1*) in normal tension glaucoma (NTG) on chromosome 12q14 in an African-American pedigree (lod score of 2.7) [227]. *TBK1* encodes a serine/threonine kinase protein that has an essential role in regulating inflammatory responses to foreign agents. *TBK1* is specifically expressed within the retinal ganglion cells, nerve fiber layer and microvasculature of the retina [239, 227]. Duplication of the *TBK1* gene leads to a significant increase in transcription, leading to glaucoma pathogenesis [227]. Direct

interactions are known to occur between *OPTN* and *TBK1*. Morton *et al* identified the binding interface between *OPTN* and *TBK1* in OPTN:1-127aa residues which were then further characterised to the OPTN:26-119aa and TBK1:677-729aa regions by Li *et al* [240]. The interaction, which occurs through the TBK1 C-terminal domain (CTD) [240], is believed to enhance OPTN binding affinity with LC3 [241].

Monemi *et al* performed a linkage analysis on a POAG family and mapped the 5q21.3-5q22.1 region, harbouring the WD repeat domain 36 gene (*WDR36*) gene [80]. The *WDR36* gene encodes a member of the WD repeat protein family. WD repeats are conserved regions of approximately 40 amino acids typically flanked by GH and WD amino acids, which facilitate formation of heterotrimeric or multiprotein complexes. It is reported to be expressed in the heart, placenta, liver, skeletal muscle, kidney and pancreas [80]. In ocular tissues, high expression was previously reported in the iris, sclera, ciliary muscle, ciliary body, retina and optic nerve [80]. Little is known regarding the exact function of *WDR36*, however, it is believed to be involved in the nucleolar processing of small subunit 18S rRNA and in T-cell activation [242]. It is thought that *WDR36* plays an important functional role in the retina homeostasis and mutations to this gene can lead to retinal damage [243]. Passutto *et al* identified that heterozygous *WDR36* variants led to elevated IOP and reduced neurite outgrowth [81]. However, Skarie *et al* concluded that haploinsufficiency alone is was insufficient for POAG causality [244]. This was further supported by a further study that suggested that *WDR36* mutants act in a polygenic molecular pathology to exacerbate and bring about earlier onset of glaucoma [245].

A linkage analysis by Wirtz *et al* studied a POAG family with locus mapping to 7q35-q36 with a (lod score of 4.06). Ankyrin repeat and SOCS box protein 10 gene (*ASB10*) is located within the POAG linkage locus GLC1F on chromosome 7q36 [13]. The

SOCS box suppresses cytokine signalling (SOCS) proteins and their binding partners with the elongin B and C complex, possibly by targeting them for degradation [176]. It is suggested that *ASB10* may play a role in ubiquitin-mediated degradation pathways through an interaction with *HSP70*. It is expressed in heart and skeletal muscle but is most highly expressed in the trabecular meshwork, retinal ganglion cells and ciliary body [13]. In glaucoma, dysfunctional *ASB10* is thought to impair the trabecular meshwork outflow which leads to retinal ganglion cell degeneration [13]. Although there is ambiguity surrounding the mode of inheritance in POAG for the *ASB10* gene [13, 246], associations with POAG continue to be made [247].

Passutto *et al* found that 1.68% of European POAG, NTG and JOAG cases were identified to have mutations in the *NTF4* gene at the GLC10 locus (19q13.33) [81]. Neurotrophin 4 (*NTF4*) is a member of a family of neurotrophic factors that control survival and differentiation of mammalian neurons. It is expressed in embryonic and adult tissues with the highest levels of expression occurring in the prostate and lower levels in thymus, placenta and skeletal muscle [232]. Knock-outs of many neurotrophins including nerve growth factor, brain-derived neurotrophic factor and neurotrophin 3 have proven lethal during early postnatal development, whilst NTF5-deficient mice only show minor cellular deficits and develop normally to adulthood [176, 248]. Functional studies of *NTF4* knockout mice (-/-) have shown mice to be long-lived with no obvious neurological defects [249]. The exact role of *NTF4* in glaucoma is currently not well understood.

3.2.2 POAG genes identified through GWAS

GWAS up until the year 2016 (when this study was designed) found significant associations in POAG patients with variants in *CAV1*, *CAV2* (Icelandic population) [112],

CDKN2B-AS1, *TMCO1* (Austrian population) [113], *SIX6* (US European population) [250], *CDKN2B-AS1*, *SIX6* (Japanese population) [251], *ABCA1*, *AFAP1*, *GMDS* (Austrian population) [252], *ABCA1*, *PMM2* (Chinese population) [253], *TGFBR3*, *FND3B* (multi-ethnic populations) [254], *ARHGEF12* (European population) [255], *TXNRD2*, *ATXN2*, *FOXC1*, and *GAS7* (US European population) [256]. Several loci have been associated at the genome-wide level in both Asian and European populations which include *CDKN2B-AS1* [113, 251], *SIX6* [250, 251], and *ABCA1* [252, 253].

No single molecular pathway encompasses the pathophysiology of POAG [257]. One study attempted to identify the pathways most enriched by POAG associated genes. It was identified that the most over-represented pathways according to ConsensusPathDB (CPDB, a molecular functional interaction database) was ‘extracellular matrix organization’ followed by ‘cytokine-cytokine receptor interactions’, ‘senescence and autophagy in cancer’, and ‘spinal cord injury’ [257]. This demonstrates the vast difference in gene functions and complexity underpinning POAG molecular pathology.

3.2.3 Paucity of associations in POAG GWAS

For various reasons discussed below, GWAS in POAG have reported associations with modest effect sizes which explain a limited proportion of the underlying genetic cause [120].

GWAS sample sizes may have previously been underpowered [258]. Increasing the sample size has been crucial to provide sufficient statistical power and allow replication of identified SNPs [259, 260]. To help address this issue, consortia such as UK Biobank, NEIGHBOR, GLAUGEN, NEIGHBORHOOD, IGGC, EPIC-Norfolk and ANZRAG have acquired larger sample sizes for POAG studies. Performing meta-analyses of GWAS has enabled larger sample sizes and success in identifying genetic variants

associated with POAG sub-phenotypes [261].

Heterogeneity of the POAG phenotypes cause further loss of power in identifying significant associations in genome-wide studies. Therefore, it is essential that standardised, systematic, and accurate phenotyping which are consistent between centres takes place.

Approximately 5% of POAG is accounted for by mutations in known familial Mendelian-like genes, mainly in the *MYOC* *OPTN* genes [77]. The majority of POAG is believed to be caused by genetically complex backgrounds. Previous studies suggested that known familial POAG genes can be involved in digenic/oligogenic inheritance in POAG (see Chapter 3.2.1 for further details). It has also been suggested that gene-gene interactions can help resolve part of the missing heritability [262]. Therefore, whilst phenotypes can vary between POAG patients and between clinics, the underlying genetic mechanisms involved can also be complex and variable between patients.

3.2.4 Database of disease-associated variants

The Human Gene Mutation Database (HGMD) is a database of germline mutations with functional or statistical evidence for suggestive of causality in human inherited disease which are compiled from literature [263, 264]. This database can therefore implicitly provide a list of genes derived from disease causing/associated variants. Whilst OMIM offers an alternative database for this [264, 265], it does not provide a list of associated genes with a phenotype and would require an inefficient manual compilation of gene lists. Similarly, ClinVar does not return a list of genes upon entering a phenotype and only gives annotations for variants pertaining to classification of pathogenicity [266].

3.2.5 NGS to identify causal genes and variants

WES has previously been applied to large cohorts of POAG and JOAG patients where variants were filtered on a candidate gene list of the seven known familial Mendelian-like genes [267, 268]. However, there have been no studies to date involving custom designed targeted NGS panels for POAG. Major drawbacks for WES involve the high costs and the lower coverage across genes [269]. Through targeted sequencing a greater coverage can be generated for specific target genes at a lower cost than WES. With lower costs, a greater sample sizes can be more feasible in order to identify rare genetic variants [269]. Variants can be filtered by prioritising higher impact variant types and less common allelic frequencies, which can be further prioritised by in-silico pathogenicity scores and functional evidence in literature [270].

3.2.6 Aim

This study aims to identify genes implicated in POAG causality by generating a customised POAG gene panel for sequencing. NGS data from this gene panel will be interrogated across a cohort of POAG patients that are selected to have homogeneous phenotypic manifestation. Variants will be interrogated in single genes, across whole genes and, where possible, interacting genes.

3.3 Methods

3.3.1 Patient selection

A total of 372 patients were initially selected from a glaucoma database of 1679 individuals across Wessex, West Midlands, Devon and Cambridgeshire regions (UK). Patient

selection criteria was previously described in Chapter 2.4.1 (Age at diagnosis ≥ 40 years of age, Caucasian ethnicity, $21 \text{ mm Hg} \leq \text{IOP} \leq 40 \text{ mm Hg}$, cup:disc ratio ≥ 0.6 and visual field mean deviation (VFMD) ≤ -3). Patients were selected in accordance with REC reference 05/Q1702/8 (see Chapter 2.4.1 for further details).

3.3.2 Gene selection

Genes were selected as described in Chapter 2.4.2.

3.3.3 Target gene capture and sequencing

The Nextera Rapid Custom Capture (NRCC) used to target sequencing. A ‘dense’ probe spacing of 120bp distances was selected to ensure comprehensive coverage across the target genes. The NRCC data was run on the NextSeq 500 (Mid Output) platform.

3.3.4 Bioinformatic pipeline and filtering parameters

The bioinformatic pipeline was used as per the processing of *MYOC* NGS data in Chapter 2 for alignment, variant calling, annotation and CNV calling.

3.3.5 Filtering of variants

There were three groups of genes (seven Mendelian-like coding genes, 56 complex coding genes, and three complex ncRNA genes) requiring three different filtering approaches (Figure 3.2).

Standard filtering was applied to the seven Mendelian-like coding variants: (1) exclude synonymous *and* (2) allele frequency in ExAC (NFE) $\leq 5\%$ *and* (3) had CADD Phred ≥ 15 [196] or a MaxEntScan score ≥ 3 [263]. The MaxEntScan tool utilises the Maximum Entropy algorithm and determines the difference between reference and alternate allele scores at the splice acceptor/donor site [271].

Following (1) exclusion of synonymous variants *and* (2) variants with an allele frequency in ExAC (NFE) $\leq 5\%$, additional **stringent** filters were applied to the 56 coding (complex) genes in order to identify the more likely pathogenic variants: (3) ‘high impact’ variant types (splice site, frameshift or stop-gain); *or* (3) missense and possible splice altering variants with ExAC NFE $\leq 1\%$, (4) CADD Phred ≥ 30 [196] or a MaxEntScan score ≥ 3 [263].

The **ncRNA** genes were filtered according to Chapter 2.2 for *MYOC* non-coding variants: (1) allele frequencies $\leq 5\%$ in 1000g EUR *and* (2) had FATHMM score ≥ 15 [196] or a MaxEntScan score ≥ 3 [263].

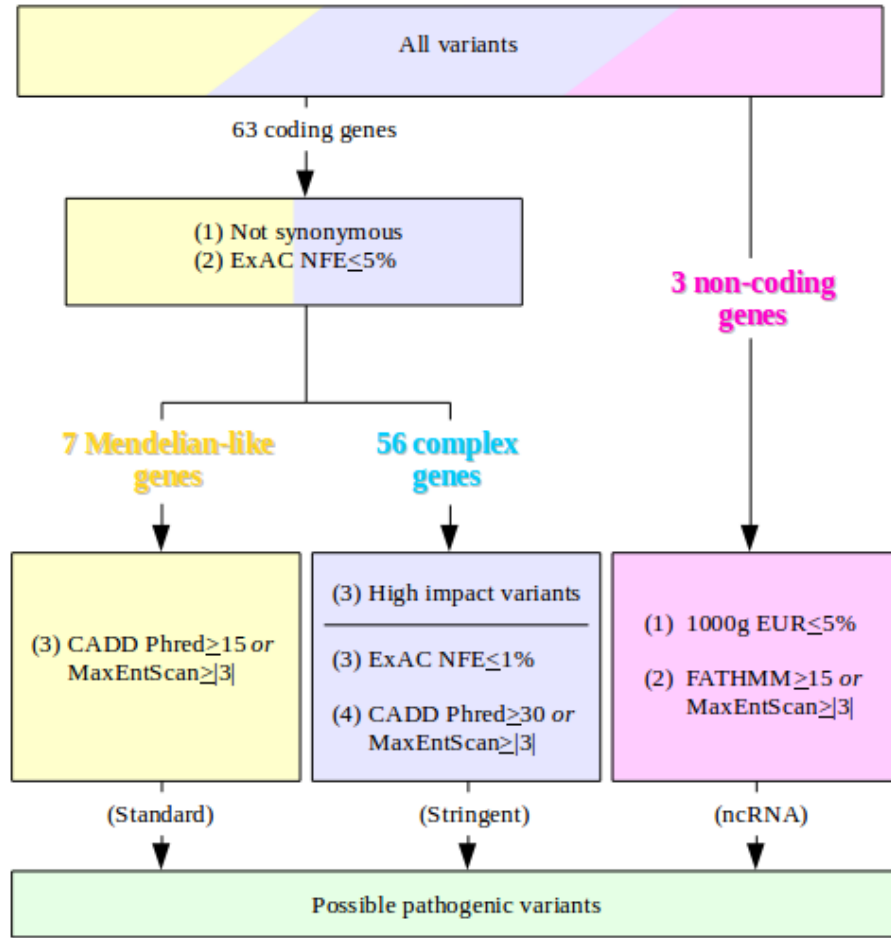


Figure 3.2: **Filtering methods used for three groups of genes.** Filtering inclusion criteria is shown for standard filtering (yellow) used for the seven Mendelian-like coding genes, stringent filtering (blue) used for the 56 coding (complex) genes, and ncRNA filtering (pink) used for the three ncRNA (complex) genes. The three filtering strategies were used to identify possible pathogenic variants (green). Upon shortlisting the respective lists of possible pathogenic variants, manual scrutinising of variant zygosity with known inheritance patterns and supporting literature should take place to infer a likely causal variant in POAG.

3.3.6 Interacting genes

From this filtered ‘possible pathogenic’ subset of variants, the potential for protein-protein disruption were considered and analysed. All genes were submitted to the Digenic Diseases Database (DIDA) which provided curated details of genes involved in digenic disease inheritance [272]. All genes were also submitted to STRINGv11.0 [273] to identify directly interacting genes based on experimental evidence and using

the default confidence threshold (0.4) [273]. For proteins which were identified to be directly interacting, the binding interface regions of the proteins and their coding genes were provided by Interactome INSIDER [274]. Variants were cross-checked with the binding interface regions to highlight variants potentially disruptive to the protein-protein interactions.

3.3.7 Whole-gene pathogenicity

An analysis was performed to identify genes and networks which have multiple variants with different predicted deleteriousness, which may accumulate to collectively cause the POAG phenotype. Filtering methods performed in previous analyses do not have the ability to identify such small effect variants. Therefore, the GenePy/v.1.2 tool was employed to generate a score for each gene for each patient sample based on known deleteriousness metrics, allele frequency and variant zygosity [275]. The scores were corrected for gene length (using the target BED file) and gene damage index (GDI) [276]. Using this score, it was possible to identify genes which have a high burden of deleteriousness that may underpin disease aetiology.

Variants were also viewed using IGV and contextualised with UCSC GC% (5 bp window) and mappability (Umap24) to aid assessment of variants clusters.

A total of 403 non-eye disease controls were used to capture GenePy scores. Control samples originated from a patient cohort of inflammatory bowel disease (IBD) patients with no known ophthalmic disease. Control samples were prepared with Agilent SureSelect capture kits v4,v5 and v6 and were sequenced on the HiSeq sequencing platform. The NGS data were aligned with BWA-mem and variants were called using GATKv3.7 Haplotype Caller. Variants from both the POAG cohort and the non-eye disease cohorts were filtered with VCFtools quality thresholds of $GQ \geq 20$ and missing individuals $\geq 30\%$

in order to prevent whole gene scores incorporating lower quality variant calls. Variants were also restricted to the intersection between the NRCC BED file of 66 genes and the Agilent SureSelect BED file using BEDtools Intersect [277].

CADD scores were selected as the chosen deleteriousness metric for GenePy score calculations as this was considered the compound score with the highest true positive pick-up rate for variants within the coding region [198].

Statistical analyses were performed using R v3.4.2. Batch effects can occur from laboratory conditions which changed from processing one batch of samples to another. This may lead to differences in quality between batches and mislead downstream statistical analyses. Therefore, a PCA analysis of GenePy scores (grouped by batch) was plotted with the ggbiplot R package. Mann-Whitney tests assume independence between all cohort samples. Therefore, where related pairs of individuals occurred, the individual with the lowest VFMD was excluded from these analyses. For each gene the most extreme highest 5% of GenePy scores were compared between the POAG cohort and non-eye disease control cohort of 403 samples.

3.4 Results

3.4.1 Patient clinical traits

The final set of 372 POAG patients prioritised for sequencing had a mean age of 66 years, and mean sub-phenotype measurements of 27.9 mmHg, 0.82 and -14.52 for IOP, CDR and VFMD respectively as previously summarised in Table 2.2.

3.4.2 Gene selection for customised panel

Thirty-one genes associated with ‘POAG’ phenotypes in the Human Gene Mutation Disease (HGMD) database (18/01/2016) [172] were targeted. Thirty-five genes were identified from a literature review on the basis of significant association with POAG causality in a Caucasian population, followed by extensive curation with clinical colleagues (see Table 3.1). In total, 66 genes were targeted for the bespoke gene panel.

3.4.3 Capture kit selection

As shown in Table 3.1, the coverage across WES of Agilent SureSelect v4 and 5 capture kits targeted less than 80% of collapsed RefSeq transcript coding regions for 53 target genes. The Nimblegen SeqCap EZ v3 captures less than 80% of 33 genes, whilst the TruSight One capture kit captured less than 80% of 19 genes. In order to address these deficiencies in target gene capture, the Nextera Rapid Capture Custom (NRCC) kit for targeted sequencing was selected to generate a customised panel. The enrichment based chemistry which NRCC utilised is described in more detail in Chapter 1.4.3.8.

The *MYOC* gene capture was previously detailed in Chapter 2.4.2. Exons including UTRs were captured for RNA genes *FAM27L*, *CDKN2A-AS1* and *CDKN2B-AS1*, and the coding sequence was captured for the 62 remaining genes. The total target region amounted to 204,523 bp for sequencing.

The NRCC capture was described by Illumina documentation to be based on the UCSC RefSeq (RefGene) database. Therefore, a BED file was created based on collapsing the UCSC RefSeq (RefGene) transcripts and coverage was determined across the samples. Coverage for all captured genes based on the NRCC capture kit coordinates was comprehensive with average gene coverage of 98.5% at 20X depth following sequencing

CHAPTER 3

Table 3.1: **A gene list was formed of POAG associated genes that could prospectively be sequenced. Thirty-one genes were identified from HGMD/ Phenomizer and 35 additional genes identified as associating with POAG in Caucasians in literature. Priority was given to the 31 HGMD genes followed by the 35 literature ascertained sub-set ranked by p-value.**

Capturing the whole MYOC transcript rather than just the exonic regions (2045 bp) for targets to be sequenced increased the total target size of the gene list from 282,132 to 297,304 bp.

† Illumina Design Studio derived data based on collapsed UCSC RefSeq transcripts (hg19).

GENE	PURMED ID	SOURCE	PHENOTYPE REPORTED	MUTATION/ rsID	ETHNICITY ASSOCIATED	P-VALUE	TARGET REGION	CHROM	CDS/ EXON COUNT†	TOTAL TARGET LENGTH†	+/-	AGILENT V4 (%)	AGILENT V5 (%)	SEQCAP EZ V3	TSO (%)
ATOH7	N/a	HGMD	POAG				CDS	10	1/3	456	-	57.9	100	86.5	0.800
CD5	N/a	HGMD	POAG				CDS	11	10/12	1485	+	50.5	50.5	50.5	99.5
CDKN2B	N/a	HGMD	POAG				CDS	9	3/5	648	-	28.2	100	34.4	97.4
CNTN4	N/a	HGMD	POAG				CDS	3	24/26	3289	+	62.8	90.4	90.1	93.5
COL1A1	N/a	HGMD	POAG				CDS	17	51/53	4392	-	78.7	100	93.8	98.3
COL3A2	N/a	HGMD	POAG				CDS	1	2/4	2109	-	51.9	100	51.2	72.8
CYP11B1	N/a	HGMD	POAG				CDS	2	2/4	1629	-	35.9	92.2	50.1	93.6
DMXL1	N/a	HGMD	POAG				CDS	5	43/45	9081	+	84.6	100	95.4	97.9
FAM7L	N/a	HGMD	POAG				Exons	17	0/2	609	+	33.7	94.9	85.4	0.000
GALC	N/a	HGMD	POAG				CDS	14	18/20	2172	-	77.8	100	64.8	91.7
GAS7	N/a	HGMD	POAG				CDS	17	17/19	1648	-	21.7	22.5	40.6	0.000
IMMT	N/a	HGMD	POAG				CDS	2	17/19	2472	-	90.5	100	87.9	95.6
LTBP2	N/a	HGMD	POAG				CDS	14	36/38	5463	-	65.4	100	72.6	98.0
MUTYH	N/a	HGMD	POAG				CDS	1	20/22	2235	-	100	100	97.0	97.7
MYOC	N/a	HGMD	POAG				Full region	1	3/5	20287	-	84.3	100	83.7	86.2
NOS3	N/a	HGMD	POAG				CDS	7	29/31	3870	+	86.6	94.5	78.7	99.5
NPH1	N/a	HGMD	POAG				CDS	2	22/24	2327	-	91.0	100	90.6	91.8
NT5C1B	N/a	HGMD	POAG				CDS	2	13/15	2334	-	78.0	100	75.7	99.1
NTF4	N/a	HGMD	POAG				CDS	19	1/3	630	-	63.6	63.7	74.6	66.1
OPTC	N/a	HGMD	POAG				CDS	1	6/8	996	-	73.8	73.8	73.8	94.8
PAK7	N/a	HGMD	POAG				CDS	20	8/10	2157	-	49.3	88.8	88.1	83.5
RPGRI1	N/a	HGMD	POAG				CDS	14	24/26	3858	+	100	100	100	89.3
SH3PXD2B	N/a	HGMD	POAG				CDS	5	13/15	2733	-	46.7	100	48.9	90.9
SIX6	N/a	HGMD	POAG				CDS	14	2/4	738	+	74.7	100	73.8	100.0
SOX2	N/a	HGMD	POAG				CDS	4	5/7	666	-	63.0	100	58.0	99.3
TBR1	N/a	HGMD	POAG				CDS	12	20/22	2187	+	77.5	95.9	91.6	98.0
TLR4	N/a	HGMD	POAG				CDS	9	5/7	4574	+	50.3	89.3	47.2	95.4
TMC01	N/a	HGMD	POAG				CDS	1	10/12	950	-	32.3	53.1	27.6	99.4
TULP3	N/a	HGMD	POAG				CDS	12	13/15	1634	+	58.7	100	63.5	100.0
WDR36	N/a	HGMD	POAG				CDS	5	23/25	2853	+	48.6	100	51.8	88.7
XRC1	N/a	HGMD	POAG				CDS	19	17/19	1899	-	99.3	99.4	100	99.3
TXNRD2	26752265 [256]	Nature Genetics	POAG	rs35934224	Australian, European, Singaporean Chinese	4.05E-11	CDS	22	17/19	1572	-	85.2	100	85.2	92.0
ABO	25173106 [278]	Nature Genetics	POAG (IOP)	rs8176693	Asian, European	6.39E-11	CDS	9	7/9	1061	-	72.9	72.8	71.0	96.9
FOXO1	26752265 [256]	Nature Genetics	POAG	rs2745572	Australian, European, Singaporean Chinese	1.76E-10	CDS	6	1/3	1659	+	54.9	100	100	82.9
ATXN2	26752265 [256]	Nature Genetics	POAG	rs7137828	Australian, European, Singaporean Chinese	4.44E-10	CDS	12	25/27	3939	-	94.3	100	99.2	93.9
AFAP1	25173105 [252]	Nature Genetics	POAG	rs4619890	Australian	7.00E-10	CDS	4	17/19	2442	-	38.0	100	53.8	0.000
GMD5	25173105 [252]	Nature Genetics	POAG	rs1969985	US	7.70E-10	CDS	6	12/14	1128	-	89.8	100	83.9	0.000
ARHGEF12	25637523 [255]	Br J Ophthal	POAG (IOP)	rs58073046	Netherlands, UK	2.81E-09	CDS	11	41/43	4632	+	52.7	99.3	68.8	98.7
ZNF469	20719862 [279]	Hum Mol Genet	POAG (CCT)	rs12447690	Croatian, Scottish	4.40E-09	CDS	16	2/4	11775	+	90.3	90.3	88.9	99.8
CDKN2B-AS1	21532571 [113]	Nature Genetics	POAG	rs4977756	Australian New Zealand	4.70E-09	Exons	9	0/21	4208	+	3.80	3.80	3.80	0.000
CDKN2A-AS1	21532571 [113]	Nature Genetics	POAG	rs4977756	Australian New Zealand	4.70E-09	Exons	9	0/2	615	+	N/a	N/a	N/a	0.000
CDKN2A	21532571 [113]	Nature Genetics	POAG	rs4977756	Australian New Zealand	4.70E-09	CDS	9	7/9	1256	-	22.3	84.4	91.2	90.7
AKAP13	20719862 [279]	Hum Mol Genet	POAG (CCT)	rs6496932	Croatia, Scotland	1.40E-08	CDS	15	39/41	8650	+	64.0	98.4	72.9	95.3
CD07	25861811 [254]	Hum Mol Genet	POAG (CDR)	rs1192415	Asian, African European	1.60E-08	CDS	1	11/13	1722	+	60.9	91.6	58.9	0.000
TGFB3	25861811 [254]	Hum Mol Genet	POAG (CDR)	rs1192415	Asian, African European	1.60E-08	CDS	1	17/19	2888	-	45.8	61.7	54.4	88.4
FND03B	25173106 [278]	Nature Genetics	POAG (IOP)	rs6445055	Asian, European	4.19E-08	CDS	3	25/27	3612	+	53.6	97.0	87.0	0.000
SALL1	20548946 [280]	Hum Mol Genet	POAG (CCT)	rs1362756	Croatia, Scotland	6.48E-08	CDS	16	4/6	7219	-	78.8	96.0	95.0	99.5
COL5A1	20719862 [279]	Hum Mol Genet	POAG (CCT)	rs1536482	Croatia, Scotland	7.10E-08	CDS	9	66/68	5514	+	69.6	100	90.6	97.5
CAV1	24572674 [281]	Ophthal	POAG	rs10256914	Caucasian, Chinese Japanese	3.69E-07	CDS	7	4/6	636	+	45.7	91.4	84.1	71.8
TP53	23049825 [250]	ProS One	POAG	codon 72 PRO/PRO	Australian, US Caucasians	1.30E-06	CDS	17	14/16	1675	-	65.2	97.9	62.0	85.9
RFTN1	20395239 [282]	Hum Mol Genet	Glaucoma (CDR)	rs690037	Australian, UK	1.60E-06	CDS	3	9/11	1734	-	65.2	90.9	76.1	0.000
CAV2	24572674 [281]	Ophthal	POAG	rs1052990	Caucasian, Chinese Japanese	1.09E-05	CDS	7	6/8	872	+	33.8	93.7	45.5	0.0
PAR2	25525164 [253]	IOVS	POAG (IOP)	rs4931170	US Caucasian	1.20E-05	CDS	12	12/14	1619	+	83.1	88.4	79.8	0.0
GGA3	25525164 [253]	IOVS Sci	POAG (IOP)	rs52809447	US Caucasian	6.70E-05	CDS	17	20/22	2392	-	61.3	59.5	80.1	0.0
SBBD1	22605021 [115]	Mol Vis	POAG	rs11884064	UK	6.70E-05	CDS	2	20/22	2985	-	86.5	97.9	85.1	0.0
PKDREJ	25525164 [253]	IOVS	POAG (IOP)	rs7291441	US Caucasian	7.40E-05	CDS	22	1/3	6739	-	90.6	100	90.1	0.0
NTM	22661486 [283]	IOVS	POAG (CCT)	rs7481514	US Caucasian	9.00E-04	CDS	11	11/13	1398	+	47.1	98.1	68.5	0.0
IL1B	23800300 [284]	Acta Ophthal	POAG	-511 T/T	Caucasian	2.00E-03	CDS	2	6/8	807	-	68.0	95.2	65.2	87.2
MMP9	23800300 [284]	Acta Ophthal	POAG	-1562 C/T	Caucasian	6.00E-03	CDS	20	13/15	2121	+	99.8	100	99.8	90.3
ASB10	22156576 [13]	Hum Mol Genet	POAG	rs104886478	Caucasian	8.00E-03	CDS	7	6/8	1672	-	88.4	89.4	89.0	96.8
MMP1	23800300 [284]	Acta Ophthal	POAG	-1607 2G/2G	Caucasian	1.40E-02	CDS	11	11/13	1559	-	88.9	100	78.0	97.5
ABCA1	25173105 [252]	Nature Genetics	POAG	rs2472493	Australian New Zealand	1.46E-02	CDS	9	49/51	6783	+	67.2	97.0	83.3	93.2
CNTNAP1	22661486 [283]	IOVS	POAG (CCT)	rs1428758	Caucasian	1.80E-02	CDS	16	25/27	3933	+	87.0	91.3	88.2	99.0
OPTN	22605021 [115]	Mol Vis	POAG	N/a	UK	2.00E-02	CDS	10	13/15	1731	+	57.7	88.6	57.7	91.7
APOE	24073598 [285]	Ophthal Gen	POAG	-491T	Polish Caucasian	2.00E-02	CDS	19	3/5	951	+	100	100	100	100.0
FBN1	20360993 [286]	Mol Vis	POAG (CCT)	rs17352842	Australian Caucasian	2.00E-02	CDS	15	65/67	8613	-	75.4	98.2	87.3	93.7

Genes are annotated with additional information which includes PubMed ID, PubMed citation of the association; Source, Journal name or HGMD which was the source of the association; Phenotype Reported; Mutation/rsID; Methods, Method by which association was derived; Ethnicity associated; p-value, p-value for the association; Details regarding the transcript and exons for the target gene derived from UCSC; Coverage of the gene (collapsed RefSeq transcripts) for Agilent SureSelect v4 and v5, TruSeq custom amplicon, Nimblegen SeqCap EZ v3 and the TruSight One (TSO) capture kits.

and alignment of NGS data (Table 3.2). For *SOD2*, *NTM*, *APOE*, *TXNRD2*, and *SH3PXD2B* the percentage of the target region captured was less than expected when compared against the UCSC RefGene BED file at 20X depth. Through an inspection of the two BED files (NRCC and UCSC RefGene), it was clear that single or even multiple exons from the UCSC RefGene BED file were not included in the NRCC capture BED file (Appendix Table B.1). Consultation with the Illumina support staff revealed that the RefSeq transcripts of the database dated to the May 2015 release. Therefore, although the custom gene panel was designed in 2016, *SOD2*, *NTM*, *APOE*, *TXNRD2*, and *SH3PXD2B* genes (bold in Figure 3.2) were not captured as planned due to Illumina Design Studio utilising outdated databases in the automation of BED file generation. However, despite this, even these genes had good gene capture with coverage of approximately 90% at 20X depth for all but the *SOD2* gene.

Table 3.2: *Coverage proportion at 1X, 20X, 50X and 100X using the Illumina Design Studio BED file for NRCC design and the UCSC RefSeq (RefGene) BED file. Genes are highlighted in bold which were covered less than expected (<3% at 20X depth).*

**Coordinates were entered manually and not dependent on Illumina Design Studio automated generation of UCSC RefGene coordinates.*

Gene	NRCC BED				UCSC RefSeq BED				Δ a-b
	1X	20X ^a	50X	100X	1X	20X ^b	50X	100X	
<i>ABCA1</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001
<i>ABO</i>	0.993	0.993	0.993	0.993	0.991	0.991	0.991	0.991	0.002
<i>AFAP1</i>	0.993	0.993	0.984	0.949	0.991	0.991	0.985	0.960	0.002
<i>AKAP13</i>	0.996	0.996	0.996	0.996	0.995	0.995	0.995	0.995	0.001
<i>APOE</i>	0.997	0.997	0.997	0.923	0.920	0.920	0.920	0.852	0.077
<i>ARHGEF12</i>	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.991	0.000
<i>ASB10</i>	0.996	0.996	0.995	0.932	0.995	0.995	0.993	0.930	0.001
<i>ATOH7</i>	0.998	0.803	0.783	0.743	0.991	0.797	0.778	0.739	0.006
<i>ATXN2</i>	0.994	0.961	0.923	0.846	0.971	0.971	0.966	0.930	-0.010
<i>CAV1</i>	0.995	0.995	0.995	0.995	0.991	0.991	0.991	0.991	0.004
<i>CAV2</i>	0.995	0.995	0.995	0.995	0.991	0.991	0.991	0.991	0.004
<i>CD5</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001

CHAPTER 3

<i>CDC7</i>	0.994	0.994	0.994	0.994	0.992	0.992	0.992	0.992	0.002
<i>CDKN2A</i>	0.996	0.996	0.996	0.996	0.985	0.985	0.985	0.985	0.011
<i>CDKN2A-AS1</i>	0.998	0.998	0.998	0.998					
<i>CDKN2B</i>	0.997	0.997	0.997	0.997	0.984	0.984	0.984	0.984	0.013
<i>CDKN2B-AS1</i>	0.995	0.982	0.952	0.941	0.995	0.982	0.952	0.941	0.000
<i>CNTN4</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001
<i>CNTNAP4</i>	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.001
<i>COL1A1</i>	0.988	0.967	0.937	0.852	0.988	0.966	0.937	0.851	0.001
<i>COL5A1</i>	0.988	0.988	0.988	0.982	0.975	0.975	0.975	0.970	0.013
<i>COL8A2</i>	0.999	0.938	0.880	0.807	0.998	0.937	0.878	0.806	0.001
<i>CYP1B1</i>	0.999	0.999	0.999	0.999	0.997	0.997	0.997	0.997	0.002
<i>DMXL1</i>	0.995	0.995	0.995	0.995	0.988	0.988	0.988	0.988	0.007
<i>FAM27L</i>	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.997	0.000
<i>FAR2</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001
<i>FBN1</i>	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.992	0.000
<i>FNDC3B</i>	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.993	0.000
<i>FOXC1</i>	0.999	0.964	0.931	0.851	0.998	0.963	0.930	0.850	0.001
<i>GALC</i>	0.992	0.992	0.992	0.991	0.990	0.990	0.990	0.990	0.002
<i>GAS7</i>	0.990	0.990	0.990	0.990	0.988	0.988	0.988	0.988	0.002
<i>GGA3</i>	0.992	0.973	0.910	0.894	0.989	0.970	0.902	0.886	0.003
<i>GMDS</i>	0.989	0.989	0.989	0.989	0.987	0.987	0.987	0.987	0.002
<i>IL1B</i>	0.993	0.993	0.993	0.993	0.989	0.989	0.989	0.989	0.004
<i>IMMT</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001
<i>LTBP2</i>	0.993	0.975	0.969	0.954	0.993	0.974	0.969	0.953	0.001
<i>MMP1</i>	0.993	0.993	0.993	0.993	0.991	0.991	0.991	0.991	0.002
<i>MMP9</i>	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.001
<i>MUTYH</i>	0.991	0.991	0.991	0.991	0.988	0.988	0.988	0.988	0.003
<i>MYOC</i>	1.000	1.000	1.000	0.999					
<i>NOS3</i>	0.993	0.989	0.972	0.948	0.990	0.987	0.970	0.946	0.002
<i>NPHP1</i>	0.991	0.991	0.991	0.991	0.990	0.990	0.990	0.990	0.001
<i>NT5C1B</i>	0.996	0.996	0.996	0.996	0.993	0.993	0.993	0.993	0.003
<i>NTF4</i>	0.998	0.856	0.808	0.781	0.994	0.852	0.804	0.777	0.004
<i>NTM</i>	0.993	0.993	0.993	0.993	0.899	0.899	0.899	0.899	0.094
<i>OPTC</i>	0.994	0.994	0.994	0.994	0.992	0.992	0.992	0.992	0.002
<i>OPTN</i>	0.992	0.992	0.992	0.992	0.991	0.991	0.991	0.991	0.001
<i>PAK7</i>	0.996	0.996	0.996	0.996	0.995	0.995	0.995	0.995	0.001
<i>PKDREJ</i>	1.000	0.991	0.971	0.944	0.999	0.991	0.970	0.943	0.000

<i>RFTN1</i>	0.995	0.995	0.995	0.969	0.995	0.995	0.995	0.956	0.000
<i>RPGRIP1</i>	0.994	0.994	0.994	0.994	0.993	0.993	0.993	0.993	0.001
<i>SALL1</i>	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.001
<i>SH3PXD2B</i>	0.995	0.962	0.962	0.962	0.957	0.925	0.925	0.925	0.037
<i>SIX6</i>	0.997	0.997	0.997	0.997	0.995	0.995	0.995	0.995	0.002
<i>SOD2</i>	0.992	0.992	0.992	0.992	0.763	0.763	0.763	0.763	0.229
<i>SRBD1</i>	0.993	0.993	0.993	0.993	0.992	0.992	0.992	0.992	0.001
<i>TBK1</i>	0.991	0.991	0.991	0.991	0.990	0.990	0.990	0.990	0.001
<i>TGFB3</i>	0.994	0.994	0.994	0.989	0.993	0.993	0.993	0.987	0.001
<i>TLR4</i>	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.001
<i>TMCO1</i>	0.991	0.991	0.991	0.991	0.985	0.985	0.985	0.985	0.006
<i>TP53</i>	0.992	0.992	0.992	0.992	0.983	0.983	0.983	0.983	0.009
<i>TULP3</i>	0.992	0.992	0.992	0.992	0.990	0.990	0.990	0.990	0.002
<i>TXNRD2</i>	0.989	0.989	0.989	0.989	0.935	0.935	0.935	0.935	0.054
<i>WDR36</i>	0.992	0.992	0.992	0.992	0.991	0.991	0.991	0.991	0.001
<i>XRCC1</i>	0.991	0.957	0.952	0.952	0.989	0.955	0.951	0.951	0.002
<i>ZNF469</i>	1.000	1.000	1.000	0.999	0.993	0.993	0.993	0.992	0.007

3.4.4 Quality control

Of the 372 sequenced samples, 14 samples were omitted following QC (as previously described in Chapter 2.5.2.2).

3.4.5 Filtering of variants

Initially, 17,148 variants were called within all aligned NGS data. When restricting to only target genes (n=66), 15,705 variants remained which were further reduced to 1506 when selecting only exonic, splicing, ncRNA exonic or ncRNA splicing variants. As detailed in the methods section (Chapter 3.3.5), known familial Mendelian-like genes, complex coding and ncRNA genes were filtered with standard, stringent and ncRNA filters respectively. With exception to one family member pair which both had the

WDR36:p.H85Q variant, no filtered variants were found to be present across both family member pairs.

3.4.5.1 Standard filtering (Mendelian-like genes, n=7)

A total of 34 variants were identified across the ‘Mendelian-like’ POAG genes as likely causal (Table 3.3). Six variants were identified as novel: *WDR36*:p.P17R, *WDR36*:p.H85Q, *WDR36*:p.T793P, *ASB10*:p.E198K and *TBK1*:p.L277V (Table 3.3). No variants were identified to be likely splice damaging using a MaxEntScan difference threshold of |3|. The likely causal *MYOC* variants identified in eight patients were previously discussed in Chapter 2. Two heterozygous *WDR36* variants (A449T and A509S) in one patient and a heterozygous *OPTN* variant (E92V) in another patient were identified as likely causes for POAG. An additional two heterozygous variants in the *NTF4* gene were identified in respective patients and identified as likely causal. A total of 20 patients across the POAG cohort were found to have possible pathogenic heterozygous missense variants in *TBK1*. The *ASB10* gene was indicated to have a haploinsufficiency score of 61% indicating that *ASB10* is less likely to exhibit haploinsufficiency [287]. No patients with homozygous or multiple within-gene heterozygous variants were identified in either *CYP11B1* or *ASB10* genes.

Table 3.3: *Thirty-four filtered variants across the seven POAG-causing genes in POAG.*

Chr	Position (hg38)	Ref	Alt	Gene	HI (%)	Variant type	Amino Acid	dbSNP144	1000g (all)	1000g (eur)	ExAC (ALL)	ExAC (NFE)	Study Cohort AF	Study Cohort AC	SIFT	GERP++	CADD Phred	CLINSIG	CLNDEN	HGMD
1	171636185	T	C	MYOC	48.4	nonsynonymous	T419A						0.00279	2	0.000	5.04	23.5			
1	171636338	G	A	MYOC	48.4	stopgain	Q368*	rs74315329	0.00060	0.00200	0.00110	0.00150	0.00978	7	.	4.52	37.0	P	JOAG	DM
1	171652236	G	A	MYOC	48.4	nonsynonymous	R126W	rs200120115	0.00020	.	0.00010	0.00007	0.00140	1	0.019	-1.13	23.5			DM
2	38071195	C	T	CYP11B1	2.18	nonsynonymous	E387K	rs55989760	.	.	0.00030	0.00060	0.00140	1	.	5.85	33.0	P	Glaucoma	DM
2	38071251	C	T	CYP11B1	2.18	nonsynonymous	R368H	rs79204362	0.00419	0.00200	0.00620	0.00290	0.00140	1	.	5.65	33.0	P	Glaucoma	DM
2	38074539	G	A	CYP11B1	2.18	nonsynonymous	R284W	rs368249322	.	.	0.00009	0.00020	0.00140	1	.	1.70	24.4			
2	38074704	C	T	CYP11B1	2.18	nonsynonymous	E229K	rs57865060	0.01038	0.00400	0.01420	0.00980	0.00838	6	.	3.77	26.7			DM
2	38075148	A	T	CYP11B1	2.18	nonsynonymous	Y81N	rs9282671	0.00220	0.00700	0.00640	0.00700	0.00140	1	.	4.42	25.6	P	POAG	DM
2	38075234	G	A	CYP11B1	2.18	nonsynonymous	P52L	rs201824781	.	.	0.00050	0.00070	0.00140	1	.	4.42	27.3	P	JOAG	DM
5	111092338	C	G	WDR36	30.25	nonsynonymous	P17R	.	.	.	.	.	0.00140	1	.	4.55	23.0			
5	111092543	C	A	WDR36	30.25	nonsynonymous	H85Q	.	.	.	.	.	0.00279	2	.	5.11	28.0			
5	111094935	C	A	WDR36	30.25	nonsynonymous	L116M	rs768394760	.	.	.	.	0.00140	1	.	4.07	23.6			
5	111097096	G	A	WDR36	30.25	nonsynonymous	D126N	rs115541547	0.00040	0.00200	0.00050	0.00090	0.00279	2	.	4.69	21.8			DM
5	111098750	C	T	WDR36	30.25	nonsynonymous	A163V	rs62376783	0.00060	0.00300	0.00410	0.00650	0.00559	4	.	5.72	22.7			
5	111100646	A	C	WDR36	30.25	nonsynonymous	H212P	rs142088179	0.00359	0.01090	0.01150	0.01700	0.02100	15	.	5.24	26.7			
5	111104338	G	A	WDR36	30.25	nonsynonymous	D354N	rs201449456	.	.	0.00020	0.00030	0.00140	1	.	5.68	33.0			
5	111104342	A	G	WDR36	30.25	nonsynonymous	N355S	rs118204022	.	.	0.00030	0.00040	0.00140	1	.	5.68	25.6	P	Glaucoma	DM
5	111106140	G	A	WDR36	30.25	nonsynonymous	A449T	rs35703638	0.01038	0.00700	0.00850	0.00490	0.00279	2	.	4.81	23.2		OAG	
5	111110219	G	T	WDR36	30.25	nonsynonymous	A509S	rs371742632	.	.	0.00002	0.00003	0.00140	1	.	5.16	27.1			
5	111110280	G	A	WDR36	30.25	nonsynonymous	R529Q	rs116529882	0.00020	0.00100	0.00080	0.00130	0.00279	2	.	5.28	34.0	P	Glaucoma	DM
5	111119021	A	G	WDR36	30.25	nonsynonymous	D658G	rs34595252	0.00140	0.00500	0.00420	0.00650	0.01300	9	.	6.07	29.3		OAG	
5	111123865	A	C	WDR36	30.25	nonsynonymous	T793P	.	.	.	.	.	0.00140	1	.	5.65	27.7			
7	151181133	G	A	ASB10	62.24	nonsynonymous	R289C	rs61735130	0.00100	0.00200	0.00280	0.00370	0.00140	1	0.017	5.29	34.0	OAG		
7	151181228	C	T	ASB10	62.24	nonsynonymous	R257H	rs140602973	.	.	0.00040	0.00080	0.00140	1	0.018	1.42	23.1	OAG		
7	151181334	G	C	ASB10	62.24	nonsynonymous	R222G	rs61735708	0.00339	0.00700	0.00440	0.00540	0.00559	4	0.296	3.29	21.7			
7	151181406	C	T	ASB10	62.24	nonsynonymous	E198K	.	.	.	.	.	0.00140	1	0.074	5.14	24.9			
7	151186916	C	T	ASB10	62.24	nonsynonymous	R72H	rs104886488	0.00080	0.00400	0.00210	0.00350	0.00279	2	0.025	1.52	20.5	P	OAG	
10	13110382	A	T	OPTN	20.54	nonsynonymous	E92V	rs202044898	.	.	0.00008	0.00003	0.00140	1	0.002	5.98	26.2			
12	64481858	C	G	TBK1	5.3	nonsynonymous	L277V	.	.	.	.	.	0.00140	1	0.257	5.16	24.0			
12	64481993	C	T	TBK1	5.3	nonsynonymous	H322Y	rs145905497	0.00080	.	0.00050	0.00040	0.00140	1	0.231	4.27	20.2			
12	64484368	T	C	TBK1	5.3	nonsynonymous	I353T	rs753694282	.	.	0.00002	0.00000	0.00279	2	0.002	5.07	23.1			
12	64488537	T	C	TBK1	5.3	nonsynonymous	V464A	rs35635589	0.00359	0.00990	0.01620	0.02420	0.02200	16	0.333	5.25	19.8			
19	49061466	G	A	NTF4	42.63	nonsynonymous	R178W	rs370717457	.	.	.	.	0.00140	1	.	2.19	26.5			
19	49061678	C	G	NTF4	42.63	nonsynonymous	R107P	rs377741315	.	.	0.00001	0.00002	0.00140	1	.	3.54	23.0			

Annotations: Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Gene.refGene, gene symbol; HI, haploinsufficiency score; Variant type, consequence of the variant; Amino acid, amino acid change; avsnp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations), 1000g ALL, 1000 Genomes Project (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar) annotated 'P' if 'pathogenic'; HGMD, annotated as DM for disease-causing mutation; Variant category, 1=assumed pathogenic, 2=assumed likely pathogenic. MYOC and OPTN are the only known dominantly inherited POAG genes.

3.4.5.2 Stringent filtering of complex POAG genes (coding genes, n=56)

In order to identify the most likely functionally important variants, stringent filter criteria were applied which yielded 64 remaining variants from a total of 1,332 coding sequence variants (Table 3.4). The 64 variants were detected across a total of 101 patients. Twelve ‘high impact’ protein truncating variants were identified which included splice site (n=5), frameshift (n=6), and stop-gain (n=1) variants. Fifty-two missense and other potential splice altering variants with very rare allele frequency and high CADD scores were also identified.

Of the 64 variants identified across 33 (of the 56) coding genes, 15 were identified as novel. A variant in the *TLR4* gene (NC_000009.11: 117704472GATdel) caused the deletion of the first two bases of an ATG start codon, resulting in a ‘whole gene deletion’. Of the splice site variants, *NOS3*:p.Q411H (rs145000830) was assigned the highest observed MaxEntScan difference score of 13.4. The highest CADD Phred score acquired by variants in the gene panel was 35 which was assigned to nine variants. Of these nine variants, a novel protein truncating stop-gain, *ABCA1*:p.R227*, was detected. Variants *OPTC*:p.R325W (rs56219555) and *GGA3*:p.S518L (rs146451191) were detected in five and three individuals respectively. The splice site variant *SRBD1*:exon3:c.81-2A>G was identified at substantially higher frequency in the POAG cohort (AF=0.00279) compared with the ExAC NFE reference database (AF=0.00006). Similarly, the variant *MMP9*:p.D165N (rs8125581) was also detected at higher frequency in the POAG cohort (AF=0.00279) than ExAC NFE (AF=0.0004).

Table 3.4: *Stably-four filtered variants across the 56 coding assumed complex genes in POAG. Bold horizontal line separates the 12 variants identified as 'high impact' and the 52 variants identified as very rare with high CADD Phred or MaxEntScan scores.*

Chr	Position (hg38)	Ref	Alt	Gene	HI (%)	ExonicFunc (refGene)	Variant information	dbSNP144	ExAC (ALL)	ExAC (NFE)	Study Cohort AF	Study Cohort AC	SIFT	GERP++	CADD Phred	MaxEnt Scan diff	CLNDSIG	CLNDBN	Interval	HGMD
chr1	45331514	-	CC	MUTYH	44.32	frameshift insertion	E328fs	rs587780078	0.0006	0.0003	0.00140	1	-	-	-	-	-	-	P	
chr2	45602085	T	C	SRBD1	18.97	splice site	exon3:c.81-2A>G	rs139393939	0.0003	0.0006	0.00279	2	-	5.82	25.0	7.95	-	-	-	
chr4	7775843	-	A	AFAP1	70.06	frameshift insertion	G606fs	-	-	-	0.00140	1	-	-	-	-	-	-	-	
chr9	104858563	G	A	ABCA1	9.43	stopgain	R227*	-	-	-	0.00140	1	-	3.26	35.0	-	-	-	P	
chr9	117704472	GAT	-	TLR4	7.76	frameshift deletion	wholegene	-	-	-	0.00140	1	-	-	-	-	-	-	-	
chr11	102796716	A	-	MMP1	34.40	frameshift deletion	I101fs	rs17879749	0.0160	0.0160	0.01100	8	-	-	-	-	-	-	-	
chr11	102797986	A	G	MMP1	34.40	splice site	exon1:c.105+2T>C	rs139018071	0.00120	0.00170	0.00140	1	-	5.78	25.1	7.75	-	-	-	
chr12	2930346	G	A	TULP3	61.13	splice site	exon5:c.492+1G>A	rs145289428	0.00030	0.00020	0.00140	1	-	4.54	23.0	8.18	-	-	-	
chr12	29309185	G	C	PAR2	65.22	splice site	exon6:c.724-1G>C	-	-	-	0.00140	1	-	5.18	25.6	8.06	-	-	P	
chr12	111599267	C	-	ATXN2	4.16	frameshift deletion	G83fs	-	-	-	0.00140	1	-	-	-	-	-	-	-	
chr14	74611682	GGCCCT GGCCGAA	-	LTBP2	38.39	frameshift deletion	L82fs	rs768947464	0.0003	0.0006	0.00140	1	-	-	-	-	-	-	-	
chr17	50195921	A	G	COL1A1	7.09	splice site	exon16:c.1056+2T>C	rs750203677	0.0010	0.00030	0.00140	1	-	5.50	26.3	7.75	-	-	P	
chr17	45322278	C	T	MUTYH	44.32	nonsynonymous	R246Q	rs149866955	0.00030	0.00040	0.00140	1	0.049	5.64	33.0	-	-	-	DM	
chr1	91520145	C	G	CDC7	9.00	nonsynonymous	S390C	rs376716700	0.0003	0.0006	0.00140	1	0.002	5.69	32.0	-	-	-	-	
chr1	91695741	T	A	TGFB3	19.61	nonsynonymous	I789fs	rs137909765	0.00120	0.00200	0.00140	1	0.003	5.50	33.0	-	-	-	-	
chr1	203503694	C	T	OPTC	87.23	nonsynonymous	R325W	rs56219555	0.00650	0.00830	0.00698	5	0.000	3.32	35.0	-	-	-	-	
chr2	18576369	G	A	NT5C1B	-	nonsynonymous	L462F	rs117487981	0.00002	0.00005	0.00140	1	0.001	5.33	32.0	-	-	-	-	
chr2	45148479	C	A	SRBD1	18.97	nonsynonymous	G740V	-	-	-	0.00140	1	0.000	5.96	32.0	-	-	-	-	
chr2	86151395	C	T	IMMT	22.58	nonsynonymous	E435K	rs61731709	0.00260	0.00400	0.00659	4	0.289	4.79	34.0	-	-	-	-	
chr4	7768942	C	T	AFAP1	70.06	nonsynonymous	E774K	-	-	-	0.00140	1	0.007	4.48	33.0	-	-	-	-	
chr4	7786298	G	T	AFAP1	70.06	nonsynonymous	R476S	rs115544900	0.00030	0.00040	0.00140	1	0.087	4.42	32.0	-	-	-	-	
chr4	7886844	G	A	AFAP1	70.06	nonsynonymous	P68L	rs777059475	0.00002	0.00002	0.00140	1	0.012	4.94	33.0	-	-	-	-	
chr6	1610455	C	T	FOXG1	9.01	nonsynonymous	R4C	-	-	-	0.00140	1	0.000	3.33	34.0	-	-	-	-	
chr7	15096878	C	T	NGS3	0.96	nonsynonymous	R179C	rs75697922	0.00002	0.00005	0.00140	1	0.000	1.40	34.0	-	-	-	-	
chr7	15000599	G	C	NGS3	0.96	nonsynonymous	Q411H	rs145000830	0.00003	0.00006	0.00140	1	0.034	3.18	-	13.44	-	-	-	
chr7	151001907	G	A	NGS3	0.96	nonsynonymous	R390Q	rs141787079	0.00010	0.00005	0.00140	1	0.004	4.80	32.0	-	-	-	-	
chr9	22006148	C	T	CDKN2B	17.90	nonsynonymous	D86N	rs148421170	0.00050	0.00240	0.00140	1	0.001	4.78	32.0	-	-	-	-	
chr9	104788041	G	A	ABCA1	9.43	nonsynonymous	A2028V	rs200788099	0.00010	0.00010	0.00140	1	0.019	5.84	34.0	-	-	-	-	
chr9	104793186	A	C	ABCA1	9.43	nonsynonymous	F1874C	rs754410874	0.00002	0.00003	0.00140	1	0.006	5.97	31.0	-	-	-	-	
chr9	104798503	C	T	ABCA1	9.43	nonsynonymous	R1680Q	rs150125857	0.00001	0.00050	0.00140	1	0.003	5.65	35.0	-	-	-	-	
chr9	104818880	C	T	ABCA1	9.43	nonsynonymous	R1082H	rs761945307	0.00001	0.00000	0.00140	1	0.001	5.97	35.0	-	-	-	-	
chr9	134758249	C	T	COL5A1	35.85	nonsynonymous	R630W	rs577618553	0.00008	0.00010	0.00140	1	0.000	4.49	35.0	-	-	-	-	
chr11	102791409	C	T	MMP1	34.40	nonsynonymous	V374M	rs59112365	0.00090	0.00006	0.00140	1	0.014	5.67	33.0	-	-	-	-	
chr11	12047022	G	A	ARHGEF12	22.14	nonsynonymous	R509Q	rs138160103	0.00190	0.00300	0.00279	2	0.062	5.44	34.0	-	-	-	-	
chr14	21310610	A	G	RPRGRP1	59.78	splicing	exon7:c.930+3A>G	rs150107283	0.00880	0.01170	0.00140	1	-	-	3.68	-	-	-	-	
chr14	60509580	C	T	SLX6	5.88	nonsynonymous	A61V	-	-	-	0.00140	1	0.001	5.52	32.0	-	-	-	-	
chr14	60509783	G	A	SIN6	5.88	nonsynonymous	E129K	rs146737847	0.00400	0.00610	0.00140	10	0.000	5.38	34.0	-	-	-	-	
chr14	74525124	C	A	LTBP2	38.39	nonsynonymous	G844C	-	-	-	0.00369	2	0.007	4.75	-	6.06	-	-	-	
chr14	74529009	C	T	LTBP2	38.39	nonsynonymous	V701M	-	-	-	0.00140	1	0.000	5.35	34.0	-	-	-	-	
chr14	74611835	C	A	LTBP2	38.39	nonsynonymous	R37M	rs934996	0.00200	0.00002	0.00140	1	0.000	3.50	32.0	-	-	-	-	
chr14	87984432	C	T	GALC	47.35	nonsynonymous	A166T	-	-	-	0.00140	1	0.016	3.95	31.0	-	-	-	-	
chr15	48412619	G	A	FBN1	2.53	nonsynonymous	R2726W	rs61746008	0.00070	0.00080	0.00140	1	0.001	1.03	32.0	-	-	-	-	
chr15	48492456	T	C	FBN1	2.53	splicing	exon24:c.2854+5A>G	rs762989672	0.00002	0.00003	0.00140	1	-	-	-5.54	-	-	-	-	
chr15	48520779	C	T	FBN1	2.53	nonsynonymous	G343R	rs746726731	0.00020	0.00040	0.00140	1	0.001	5.65	35.0	-	-	-	-	
chr15	48506337	C	T	FBN1	2.53	nonsynonymous	A162T	rs193922210	0.00000	0.00020	0.00140	1	0.003	5.87	33.0	-	-	-	-	
chr15	85543864	G	A	AKAP13	65.04	nonsynonymous	G191R	rs74502151	0.01190	0.01880	0.01500	11	0.033	5.49	33.0	-	-	-	-	
chr15	85717400	A	G	AKAP13	65.04	nonsynonymous	E1949G	rs765716984	0.00001	0.00002	0.00140	1	0.001	6.02	32.0	1.38	-	-	-	
chr15	85729260	C	T	AKAP13	65.04	nonsynonymous	R2229W	rs199748104	0.00007	0.00009	0.00140	1	0.000	3.80	35.0	-	-	-	-	
chr15	85741244	C	G	AKAP13	65.04	nonsynonymous	R2603G	rs749633401	0.00002	0.00004	0.00140	1	0.035	4.24	32.0	-	-	-	-	
chr16	76476051	C	T	CNTNAP2	43.33	splicing	exon10:c.1543+6C>T	rs9935035	0.03230	0.01050	0.00698	5	-	-	-3.64	-	-	-	-	
chr17	9918089	C	T	GAS7	15.42	nonsynonymous	R410Q	rs749184881	0.00002	0.00002	0.00140	1	-	5.08	32.0	-	-	-	-	
chr17	7529386	G	A	GGA3	46.52	nonsynonymous	S518L	rs146451191	0.00160	0.00240	0.00419	3	-	5.32	35.0	-	-	-	-	
chr17	75234371	C	T	GGA3	46.52	nonsynonymous	A62T	rs360684306	0.00006	0.00009	0.00140	1	0.001	5.27	34.0	-	-	-	-	
chr19	43551628	A	T	XRCCL1	43.33	nonsynonymous	V381E	rs75932835	0.00002	0.00005	0.00140	1	0.000	5.16	32.0	-	-	-	-	
chr19	43554684	C	A	XRCCL1	43.33	nonsynonymous	D126Y	-	-	-	0.00140	1	0.000	4.92	33.0	-	-	-	-	
chr19	43554708	G	A	XRCCL1	43.33	nonsynonymous	R118W	rs143881845	0.00070	0.00100	0.00279	2	0.001	3.88	35.0	-	-	-	-	
chr19	43575439	C	A	XRCCL1	43.33	nonsynonymous	R7L	rs2307186	0.00180	0.00300	0.00279	2	0.045	4.66	34.0	-	-	-	-	
chr20	9539598	C	A	PAK7	26.31	nonsynonymous	G675V	rs199976502	0.00003	0.00006	0.00140	1	0.002	4.53	32.0	-	-	-	-	
chr20	9557649	C	T	PAK7	26.31	nonsynonymous	D568N	-	-	-	0.00140	1	0.033	5.69	34.0	-	-	-	-	
chr20	40010604	G	A	MMP9	2.36	nonsynonymous	D165N	rs8125581	0.00030	0.00040	0.00279	2	0.002	4.72	33.0	-	-	-	-	
chr20	40010960	C	T	MMP9	2.36	nonsynonymous	L187F	rs55789927	0.00260	0.00210	0.00559	4	0.001	4.62	32.0	-	-	-	-	
chr20	40013244	T	A	MMP9	2.36	splicing	exon9:c.1331-11T>A	rs3918257	0.01100	0.00008	0.00140	1	-	-	3.42	-	-	-	-	
chr20	40014147	G	A	MMP9	2.36	nonsynonymous	G592S	-	-	-	0.00140	1	0.045	5.11	32.0	-	-	-	-	
chr22	19931063	C	T	TXNRP2	72.21	nonsynonymous	G47R	rs759298418	0.00001	0.00002	0.00140	1	0.000	3.68	33.0	-	-	-	-	

3.4.5.3 Filtering of non-coding genes (ncRNA, n=3)

A total of 68 variants were called across the three ncRNA genes (*FAM27L*, *CDKN2A-AS1*, *CDKN2B-AS1*). The variants were filtered with ncRNA filter criteria to yield five variants (Table 3.5). One variant had a FATHMM score of 0.59 and four variants had MaxEntScan scores of 8.18, 6.04, 7.95 and 7.75 respectively. Interestingly, all variants were identified within the same ncRNA gene, *CDKN2B-AS1*.

Table 3.5: *Five non-coding RNA variants following the application filter criteria. All variant identified were heterozygous.*

Chrom	Position	Ref	Alt	Gene refGene	Variant type	Variant information	avsnp144	1000g ALL	1000g EUR	Study cohort AF	Study cohort AC	FATHMM	MaxEnt Scan diff
chr9	21995045	T	G	<i>CDKN2B-AS1</i>	ncRNA_exonic	.	rs55797833	0.00759	0.02290	0.03900	28	0.59133	-
chr9	22032988	T	A	<i>CDKN2B-AS1</i>	ncRNA_splicing	exon3:c.846+2T>A	rs140053489	0.00699	.	0.00140	1	0.16933	8.18
chr9	22049228	G	C	<i>CDKN2B-AS1</i>	ncRNA_exonic	.	rs766233091	.	.	0.00279	2	0.16146	6.04
chr9	22066233	A	G	<i>CDKN2B-AS1</i>	ncRNA_splicing	exon12:c.2330-2A>G	rs181385014	0.00060	0.00300	0.00419	3	0.05154	7.95
chr9	22066355	T	C	<i>CDKN2B-AS1</i>	ncRNA_splicing	exon12:c.2448+2T>C	rs117869160	0.00938	0.00300	0.00279	2	0.04931	7.75

3.4.6 Interacting proteins

The purpose of this approach was to assess the evidence for compounding gene-gene interactions across target genes that either directly or indirectly interact.

3.4.6.1 Indirect protein-protein interactions

The DIDA database revealed *MYOC* and *CYP1B1* were identified as indirectly interacting genes in congenital glaucoma and JOAG respectively. When comparing the filtered variants with *MYOC* variants in Chapter 2, patient GL483 was identified to have a heterozygous synonymous *MYOC*:p.K216K variant and a heterozygous missense *CYP1B1*:p.E229K variant.

Blanco-Marchite *et al* describe a possible gene-gene interaction between rare *WDR36* and *TP53* [288]. However, no sample had standard filtered variants called in both genes in the POAG cohort.

3.4.6.2 Direct protein-protein interactions

Filtered variants were interrogated across known and predicted binding interfaces. Three unique direct interaction networks (*OPTN-TBK1-TLR4*, *CAV1-CAV2*, and *CDKN2A-CDC7*) were identified by STRINGv11.0 based on experimental evidence (Figure 3.3).

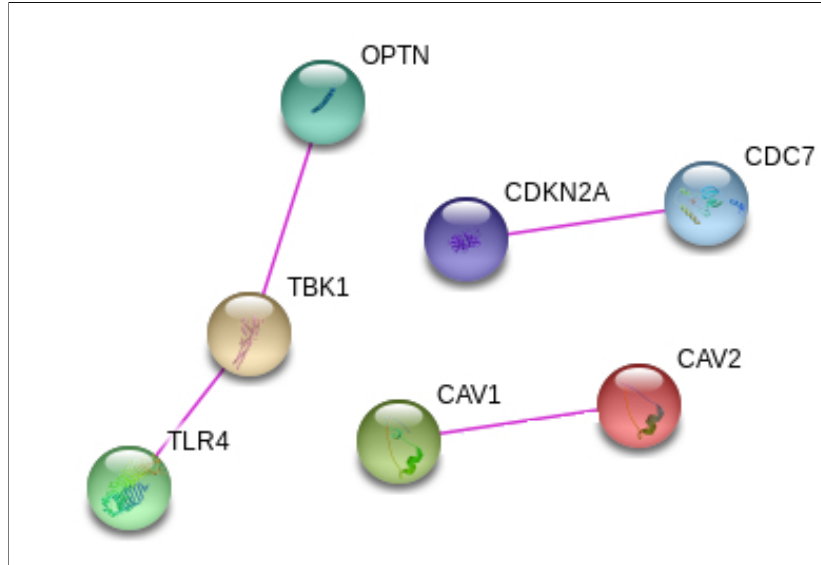


Figure 3.3: **Protein-protein interaction network of the 66 genes selected for targeted sequencing.** Nodes represent a protein (gene product) which are linked by known interactions (three identified) in pink (experimental evidence).

Proteins which were identified as likely interacting from STRING v11.0 were also entered into Interactome INSIDER to predict the likely binding interface. This generated an output BED file of coordinates which highlighted the genomic positions involved in the predicted binding interface of the protein [274]. No binding interface was identified between *CAV1-CAV2* and *TBK1-TLR4* respectively. Binding interfaces were identified by the Interactome INSIDER software between *OPTN-TBK1* (Appendix Tables B.2 & B.3) and *CDC7-CDKN2A* (Appendix Tables B.4 & B.5). No variants were found to reside within the predicted binding interface sites for *CDKN2A-CDC7*. Similarly, no variants were found to reside within the predicted binding interfaces for *OPTN-TBK1*.

However, one *OPTN* variant (p.E92V) was 3 codons from a predicted interface site (hg38 genomic coordinate chr10:13110393) and was within the TBK1 binding region determined by Li *et al* [240]. The single available crystal structure of the OPTN-TBK1 interaction (PDB:5EOF) provided the relevant binding domains of OPTN (N-terminal) and the TBK1 (CTD). Figure 3.4 shows the p.E92 residue not directly interacting with the CTD domain of TBK1.

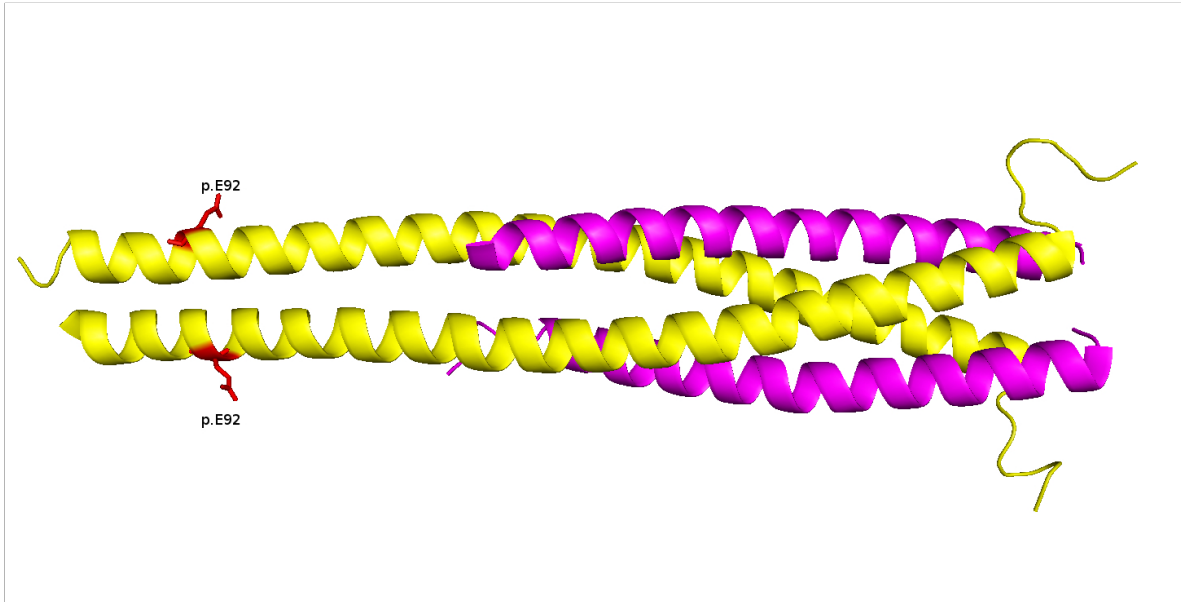


Figure 3.4: *Crystal structure visualisation of the variant in context with the interacting TBK1 protein.* Yellow represents *OPTN* N-terminal residues (aa26-103), magenta represents the *TBK1* CTD (aa677-729), and red sticks represent the p.E92 residue. Visualisation was made with PyMOL/1.7.4.

3.4.7 Whole gene pathogenicity scores

GenePy scores were calculated for each gene for each patient sample. The GenePy score integrates the CADD score, allele frequency and zygosity for each variant across a gene, and combines these into a single gene pathogenicity score that is then corrected for gene length and gene damage index (GDI).

3.4.7.1 Batch effect

A PCA was performed on all 358 POAG samples' GenePy scores (Figure 3.5). The first two components explain 3.4% and 3.0% of the original variance respectively. Clustering of samples by batch was not observed indicating that there was no substantial batch effect affecting whole gene pathogenicity scores.



Figure 3.5: *Principle component analysis of GenePy scores for each sample grouped by batch number. PCA of 358 POAG samples shown as dots which are coloured by batch number. The PCA data points are grouped (with an ellipse) for each batch of samples using a default confidence level of 0.95.*

3.4.7.2 Overview of whole gene pathogenicity scores across the POAG cohort

Across the 358 POAG patients, *TMCO1* did not harbour any genetic variants so raw GenePy scores were generated for 65 genes which were normalised by gene length (Figure 3.6). GDI scores were unavailable for *ABO* gene and the RNA genes (*FAM27L*, *CDKN2A-AS1* and *CDKN2B-AS1*). Therefore, GenePy scores were only able to be normalised by GDI for 61 genes (Appendix Figure B.1).

In Chapter 2, eight patients were identified to harbour likely causal variants. These eight variants were also identified as eight of the nine highest GenePy scoring samples in the *MYOC* gene. This proved that GenePy was sensitive to known causal variants.

While correction for gene length and GDI facilitates direct comparison between genes - this should still be done with caution as some genes/gene families may still be prone to naturally acquiring high scores. We observed three genes (*ILB1*, *CDKN2B-AS1*, and *ABO*) with markedly high GenePy scores which warranted further investigation.

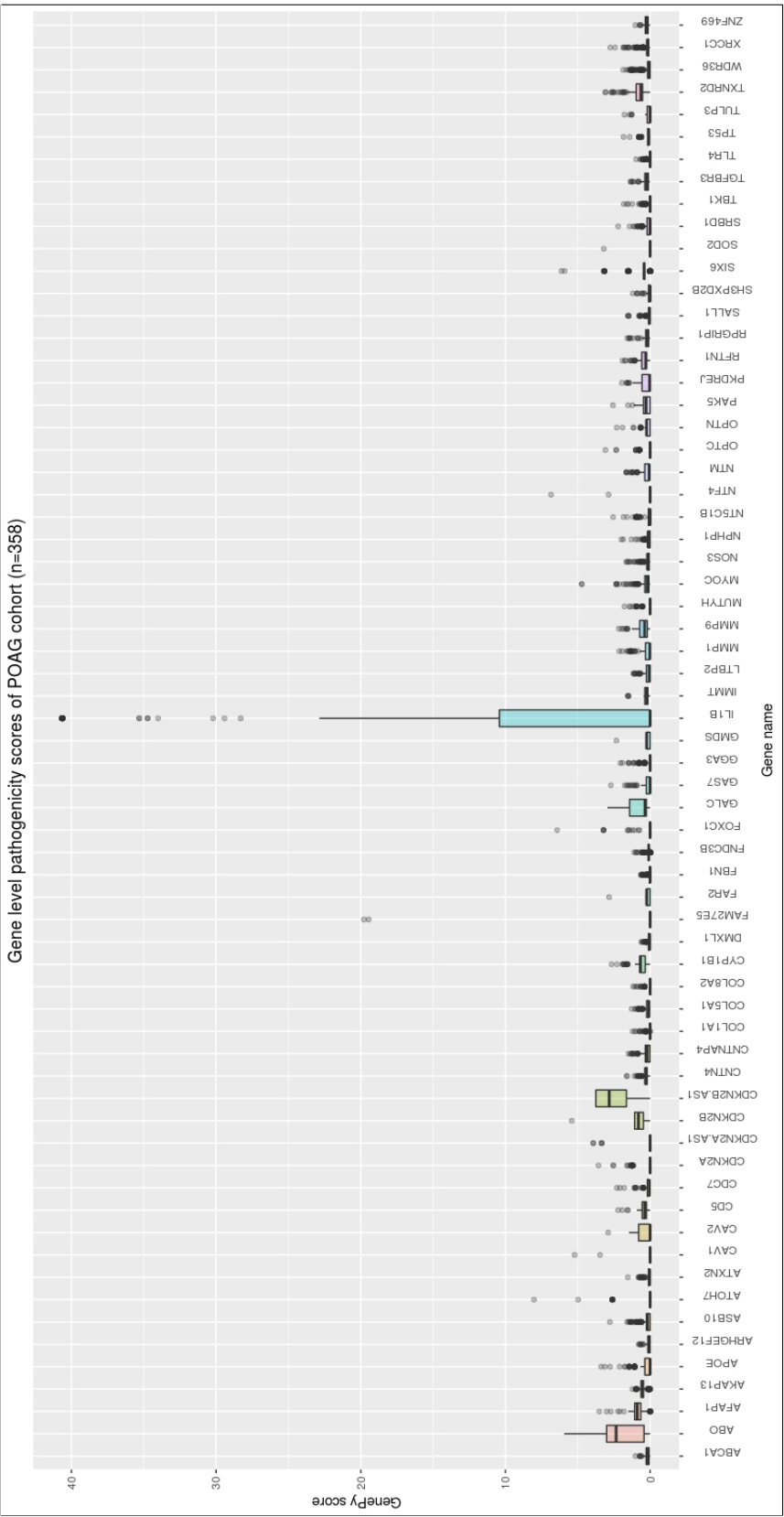


Figure 3.6: 65 POAG gene GenePy scores across the POAG cohort (n=358).

The remarkably high GenePy score in *IL1B* for multiple individuals and a single individual high score in *FAM27E5* warranted further investigation.

IL1B and *FAM27E5* had maximum GenePy scores of 40.63 (for 49 samples) and 19.78 (sample QG195) respectively. Further inspection of the highest *IL1B* GenePy scoring samples revealed that they all harboured five variants in common (c.G630A:p.K210K, c.G627T:p.K209N, c.G624C:p.K208N, c.T616A:p.Y206N, and c.C608A:p.P203H). In contrast, samples that did not have these five variants had a mean and median GenePy score of 3.46 and 0.00 respectively. The five variants collectively produced whole-gene pathogenicity scores ranging 15.89 (all heterozygous) to 31.80 (all homozygous). Therefore, it is clear that this cluster of five variants were key to causing elevated GenePy scores for *IL1B* in the POAG cohort.

Table 3.6: ***GenePy scores for five IL1B variants.*** *ANNOVAR annotation for allele frequencies and CADD scores were provided using hg19 databases (coloured in yellow). GenePy scores were calculated as per the formula described by Mossotto et al [275]. The five variants provide a score total of 15.90 for heterozygous variants and 31.79 for homozygous variants.*

Chrom	Position (hg38)	Position (hg19)	Ref	Alt	Variant type	Variant information	GnomAD ALL	CADD13 Phred	GenePy (het)	GenePy (hom)
2	112830541	113588118	C	T	synonymous	c.G630A:p.K210K	0.00050	15.5	1.85	3.70
2	112830544	113588121	C	A	nonsynonymous	c.G627T:p.K209N	0.00000	24.9	3.69	7.39
2	112830547	113588124	C	G	nonsynonymous	c.G624C:p.K208N	0.00000	26.9	3.95	7.91
2	112830555	113588132	A	T	nonsynonymous	c.T616A:p.Y206N	0.00000	23.0	3.27	6.53
2	112830563	113588140	G	T	nonsynonymous	c.C608A:p.P203H	0.00000	22.2	3.13	6.27

Patient ID GB194 (one of the patients above with the highest *IL1B* GenePy score of 40.63) contained the five variants identified. The five variant cluster had a GC% of 34.76% and were identified to be within a 100% unique region in the human reference genome. Therefore, GC content and mappability were not considered as potential issues that would cause false positives. These variants were also shown to be within a region rich with T alternative alleles and importantly were observed in only a very low proportion of reads $\leq 11\%$ (Figure 3.7) of which were all exclusively found on one

strand of DNA, indicative of strand bias. These likely false positive variant calls of *IL1B* were a major cause of elevated GenePy scores in the POAG cohort, which would provide falsely significant p-values in downstream analyses. Therefore, *IL1B* was excluded from further analyses.

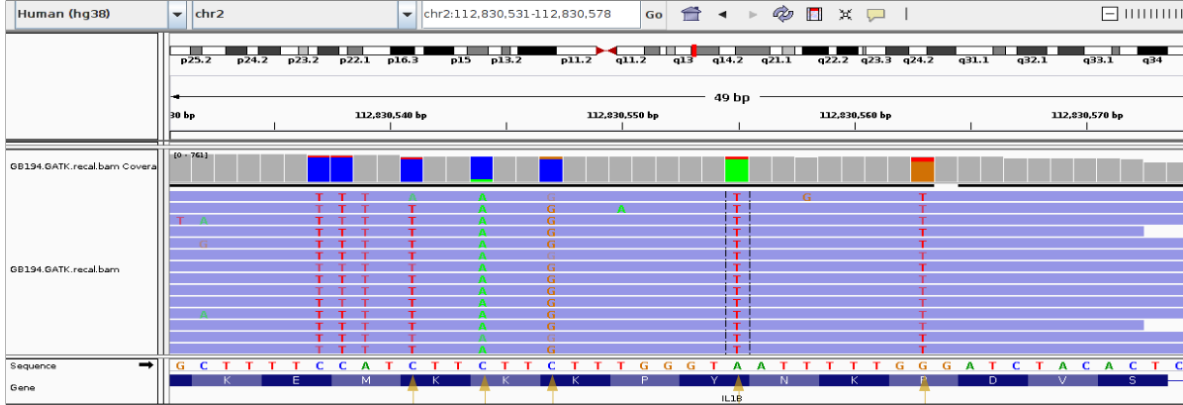


Figure 3.7: IGV image of sample GB194 for the five variant cluster of *c.G630A:p.K210K*, *c.G627T:p.K209N*, *c.G624C:p.K208N*, *c.T616A:p.Y206N*, and *c.C608A:p.P203H* (*chr2:112830541*, *chr2:112830544*, *chr2:112830547*, *chr2:112830555*, and *chr2:112830563* respectively) in the *IL1B* gene. The five variants are indicated with dark yellow arrow heads. Coverage track indicates the proportion of reads (coloured) with the reference and alternative allele.

Patient IDs QG195 and GL515 had the highest GenePy scores (19.78 and 19.46 respectively) for *FAM27E5* in the POAG cohort. QG195 had the variant rs771362103 which was found exclusively in this sample across the cohort and also had the lowest allele frequency of all the *FAM27E5* variants (gnomAD ALL=0.00002). The remaining samples which did not have this variant had mean and median GenePy scores of 0.05 and 0.00 respectively. The context surrounding rs771362103 indicated that it was within a 100% unique 24mer sequence and had a GC content of 60.74%. The alternative T allele existed at 49% and was approximately evenly distributed on the positive (74 reads) and negative (87 reads) strands (Figure 3.8). Therefore, evidence supported that the *FAM27E5* variant responsible for generating a higher GenePy score, was not a false positive. Similarly, patient GL515 had only one variant (rs748295238) which was only observed in this patient across the POAG cohort. This variant had: a low allele

frequency in all populations (gnomAD ALL=0.00003); 45% of patient's reads called as the alternate allele; and was approximately evenly distributed on the positive (86 reads) and negative (103 reads) strand. Therefore, both of the highest GenePy scoring samples for *FAM27E5* were driven by variants that appeared to be reasonably good quality and true calls.

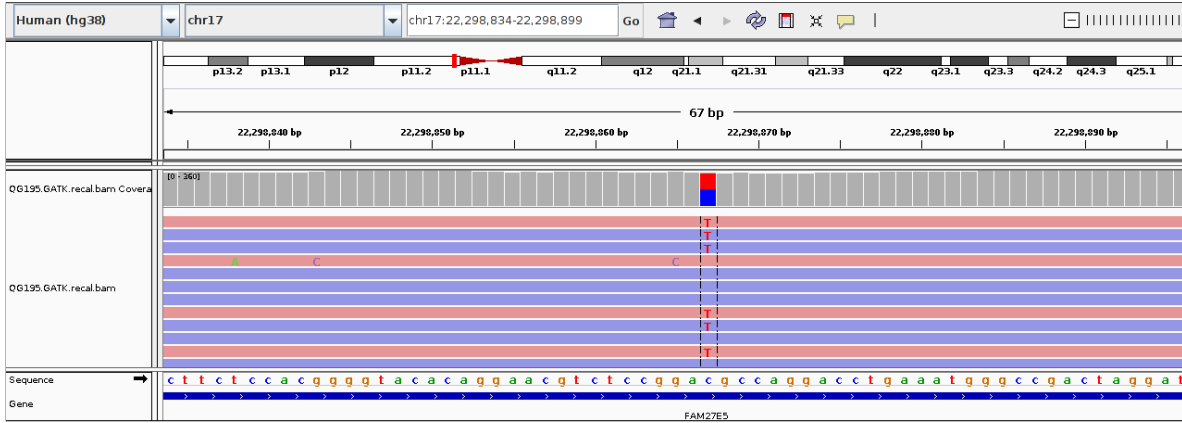


Figure 3.8: *IGV image of sample QG195 for variant rs771362103 in the FAM27E5 gene.* Alignment tracks show visualisation of reads which were coloured based on strand (red, positive; blue, negative). A coverage track for each base pair position is shown to visualise allelic depth.

3.4.7.3 Comparing POAG whole gene pathogenicity scores against a control cohort

A Mann-Whitney test was identified as the most appropriate test for identifying a significant difference in elevated GenePy scores in the POAG cohort compared with the non-eye disease (control) cohort.

In order to avoid bias, ten relatives in the cohort (Appendix Table A.1) with the lowest VFMD were omitted for the Mann-Whitney tests, these omissions were identified as QGF001, QGF033, QGF023, QGF018, QGF011, QGF003, QGF010, and QGF007 (QGF015 and QGF027 were previously omitted by the quality control step, therefore, eight samples were omitted from the statistical analysis).

The top 5% from each cohort were compared which equated to 20 individuals from the

non-eye disease cohort (n=403) and 18 individuals from the POAG cohort (n=350). The non-eye disease cohort had no variants in both the *TMCO1* and *FAM27L* genes. *IL1B* was omitted for reasons stated previously. Therefore, the total number of comparable genes between the POAG and non-eye disease cohorts was 63 genes. Mann-Whitney tests were also performed on normalised GenePy scores summed across the interacting genes between the POAG and non-eye disease cohort. All p-values can be found in Appendix Table B.6 with the significant values displayed in Table 3.7. Seven genes and one interacting pair of genes were nominally significant. Following application of FDR corrections, only *ABO* and *SIX6* maintained significance.

Table 3.7: *Mann-Whitney p-values for comparisons of gene level pathogenicity scores between the POAG (n=350) and non-eye disease (n=403) cohort. Uncorrected p-values were corrected after Bonferroni correction and the false discovery rate (FDR) Benjamini & Hochberg correction. The significant uncorrected p-value genes are shown. Only ABO and SIX6 withstand FDR corrections. IL1B was omitted due to the variants which were responsible for generating elevated GenePy scores being likely false positives. NS was assigned where p-values were 1 and not significant.*

	Genes	Uncorrected p-value	Bonferroni	Benjamini & Hochberg
1	<i>ABO</i>	6.60E-8	4.62E-6	2.31E-6
2	<i>SIX6</i>	3.73E-4	2.61E-2	8.70E-3
3	<i>TLR4</i>	3.14E-3	2.20E-1	5.50E-2
4	<i>CDKN2A-AS1</i>	6.85E-3	4.80E-1	9.59E-2
5	<i>MYOC</i>	9.18E-3	6.42E-1	1.01E-1
6	<i>LTBP2</i>	1.01E-2	7.04E-1	1.01E-1
7	<i>MYOC & CYP1B1</i>	2.49E-2	NS	2.14E-1
8	<i>ATOH7</i>	4.47E-2	NS	2.14E-1

The *ABO* variant n.207G>A (at hg38 position chr9:133259827) was found in three samples in the POAG cohort, including the two samples (FG415 and GL189) with the highest GenePy score (5.94). This heterozygous variant alone generated a GenePy score of 5.082. The mean and median GenePy scores for samples with this variant was 5.46 and 5.94 respectively. In contrast, the mean and median GenePy scores for samples

without this variant was 1.99 and 2.35 respectively. The variants identified in *ABO* were individually of low pathogenicity scores, with the highest CADD Phred score at 13.3 (Table 3.8).

Table 3.8: *GenePy scores for all ABO variants cross the POAG cohort. Variant position in shown in both hg38 (grey) and in hg19 (yellow). ANNOVAR annotation for allele frequencies and CADD scores were provided using hg19 databases. GenePy scores were calculated as per the formula described by Mossotto et al [275]. The variants provide a total score of 15.90 for heterozygous variants and 31.79 for homozygous variants. The ABO variants provide a total possible score of 15.76 for heterozygous variants and 30.92 for homozygous variants.*

Chrom	Position (hg38)	Position (hg19)	Ref	Alt	Variant type	GnomAD ALL	Study cohort AF	CADD13 Phred	GenePy (het)	GenePy (hom)
9	133255670	136131057	G	-	.	.	0.07300	.	1.28	2.55
9	133255801	136131188	C	T	.	0.11780	0.08000	13.6	0.30	0.57
9	133255902	136131289	C	T	.	0.26720	0.24200	8.5	0.20	0.32
9	133255928	136131315	C	G	.	0.11810	0.08100	0.1	0.23	0.43
9	133255929	136131316	C	T	.	0.01440	0.02200	6.0	0.49	0.98
9	133255935	136131322	G	T	.	0.11780	0.08100	0.0	0.20	0.37
9	133255960	136131347	G	A	.	0.26910	0.24200	1.0	0.18	0.28
9	133255963	136131350	G	T	.	0.03030	0.04500	0.8	0.38	0.75
9	133256028	136131415	C	T	.	0.11900	0.08000	6.7	0.26	0.50
9	133256042	136131429	C	T	.	0.00360	0.00279	7.8	0.67	1.34
9	133256050	136131437	C	T	.	0.24420	0.24000	2.4	0.19	0.31
9	133256074	136131461	G	A	.	0.11950	0.08000	2.5	0.25	0.47
9	133256082	136131469	G	A	.	0.00110	0.00140	0.3	0.71	1.42
9	133256085	136131472	A	T	.	0.25220	0.24200	4.2	0.19	0.31
9	133256086	136131473	C	A	.	0.00007	0.00140	8.0	1.15	2.29
9	133256091	136131478	T	GTCCAC	.	.	0.00140	.	1.28	2.55
9	133256136	136131523	G	A	.	0.01370	0.02700	7.7	0.51	1.02
9	133256189	136131576	C	T	.	0.03370	0.02100	13.3	0.45	0.90
9	133256204	136131591	C	T	.	0.00130	0.00140	6.6	0.78	1.55
9	133256205	136131592	G	C	.	0.13260	0.10200	0.2	0.22	0.42
9	133256264	136131651	G	A	.	0.08340	0.07000	5.9	0.30	0.57
9	133257486	136132873	T	C	.	0.40370	0.34600	2.8	0.16	0.20
9	133257521	136132908	-	C	.	.	0.34800	11.0	1.46	2.91
9	133259827	136135231	C	T	.	0.00006	0.00419	2.3	1.07	2.14
9	133259835	136135239	G	A	.	.	0.00140	0.8	1.23	2.46
9	133261370	136136773	C	T	.	0.01430	0.01800	8.1	0.51	1.02
9	133262144	136137547	C	A	.	0.01440	0.02200	4.9	0.49	0.97
9	133275184	136150600	G	A	.	0.00310	0.00140	4.2	0.65	1.31

The *ABO*:n.207G>A variant (hg38 position chr9:133259827) was found in close proximity with two other hg38 based variants which had no recorded allele frequencies in ExAC. The variant was found to have 32% alternative allele (Figure 3.9) and was approximately evenly distributed on the positive (55 reads) and negative (41 reads)

strands. The variant was found to reside within a 100% unique 24mer region with a 5bp window GC content of 56.54%. Therefore, the *ABO* variants which cause the elevated GenePy scores in the POAG cohort are likely to be true variants.

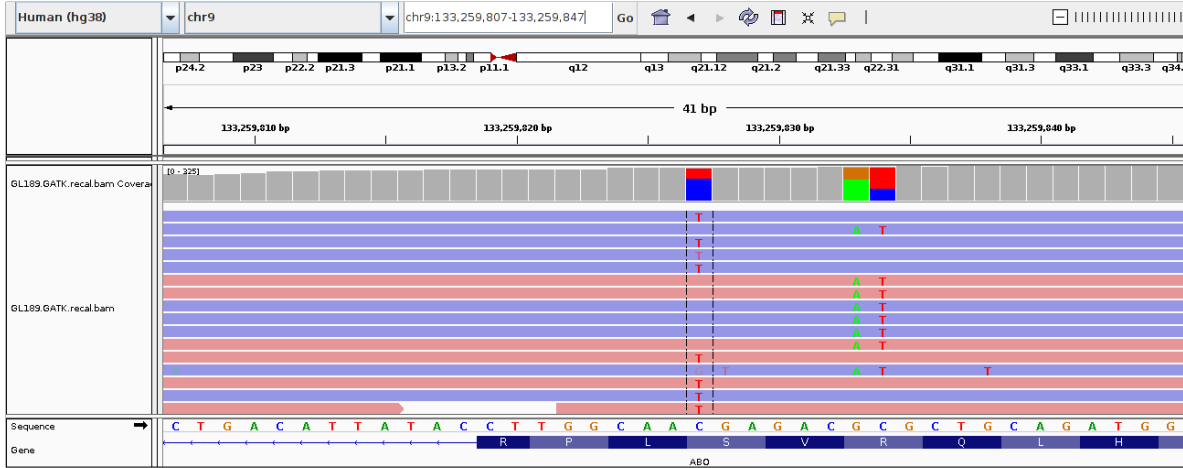


Figure 3.9: *IGV image for POAG sample GL189 of ABO variant n.207G>A (chr9:133259827). The neighbouring variants at chr9:133259833 and chr9:133259834 are not called in the hg19 pipeline, nonetheless, this hg38 aligned variant cluster warranted further investigation. Reads are coloured based on strand (red, positive; blue, negative). A coverage track for each base pair position is shown to visualise allelic depth.*

Sample QG032 had the highest GenePy score in the *SIX6* gene (6.14) and contained only two variants within the gene. One variant (p.H141N) was found in 60.6% of the POAG cohort with a modest CADD Phred score of 12.61, whilst the second variant (p.A61V) passed stringent filter criteria and was exclusive to sample QG032. Samples without the p.A61V variant had mean and median GenePy scores of 0.52 and 0.42 respectively. The variant c.C182T:p.A61V alleles were 50% distributed between the alternate T and reference C alleles (Figure 3.10) and were populated on the positive (224 reads) and negative (155 reads) strand with no clear evidence of strand bias. The variant was found in a 24mer region of 100% uniqueness and a GC content of 56.3%. Therefore, evidence suggests that this variant, which contributed substantially to elevated GenePy scores in *SIX6*, is not impacted by any obvious artefacts and is likely a true variant.

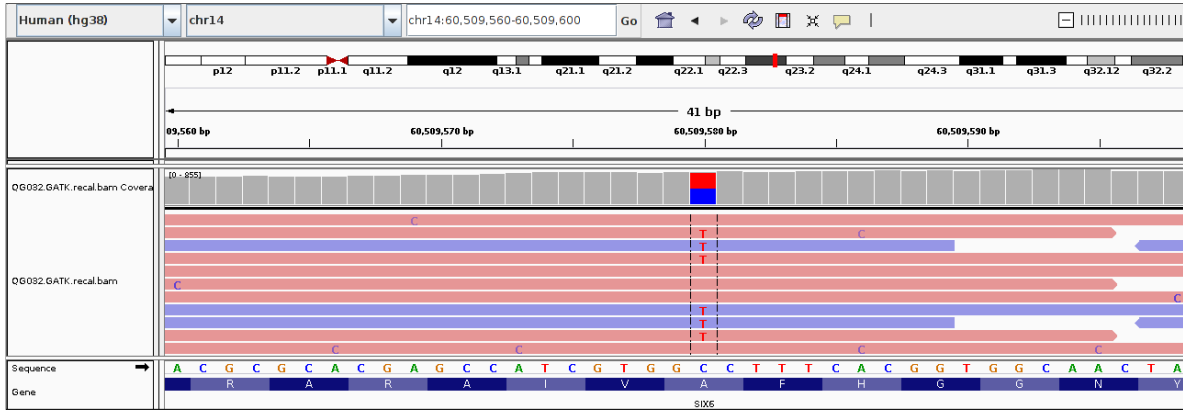


Figure 3.10: *IGV image of sample QG032 for SIX6 variant, c.C182T:p.A61V. Reads are coloured based on strand (red, positive; blue, negative). A coverage track for each base pair position is shown to visualise allelic depth.*

3.4.8 Copy Number Variants (CNVs)

Copy number (CN) was predicted for each sample for each of the 66 POAG genes. All 561 instances where the CN was predicted to be lower or higher than the wild-type CN=2 was listed in Appendix B.7. There were no indels of exonic size (~ 200 bp) in any of the genes (as 1916bp was shortest CNV length detected). There were 16 cases of a predicted single copy deletion (CN=1) detected and one case involving a homozygous deletion (CN=0) of a gene. A total of 541 instances of a single copy number gain (CN=3) and three instances of a double copy number gain (CN=4) were detected (Appendix B.7). The CN=3 (single copy gain) was found for 12 batch 1 samples, 318 batch 2 samples, 159 batch 3 samples and 52 batch 4 samples which suggested there was a possible bias between batch and an overcalling of CNVs gains in batches 2 and 3. Table 3.9 lists the CNV calls with the 541 CN=3 cases omitted. The high variability of within and between batch coverage rendered the calibration for whole-gene gains and losses unreliable for further analysis.

Table 3.9: *CNV gains and losses detected by CNVKit. Whole gene duplications and deletions were detected for one and 15 genes respectively. Single gain CN=3 were omitted due to a likely high false positive rate.*

Batch	Sample	Chrom	Start	End	Length	Gene(s)	CN	Depth
1	FG102	2	110123793	110204968	81175	<i>NPHP1</i>	1	350.6
3	GL192	1	165728025	165768515	40490	<i>TMCO1</i>	1	290.8
3	GL211	1	165728025	165768515	40490	<i>TMCO1</i>	1	283.2
3	GL228	1	165728025	165768515	40490	<i>TMCO1</i>	1	239.4
3	GL260	1	165728025	165768269	40244	<i>TMCO1</i>	1	225.9
3	GL261	1	165728025	165768515	40490	<i>TMCO1</i>	1	274.2
4	GL426	12	884763	2940723	2055960	<i>WNK1,TULP3</i>	1	433.4
3	GL429	10	13109122	13136863	27741	<i>OPTN</i>	1	192.1
3	GL429	12	111452814	111599514	146700	<i>ATXN2,ATXN2,ATXN2-AS</i>	1	143.2
3	GL429	17	7669611	22299893	14630282	<i>TP53,GAS7,FAM27E5</i>	1	190.1
4	GL430	17	75237451	75261587	24136	<i>GGA3</i>	1	291.3
4	GL430	12	64455870	64501378	45508	<i>TBK1</i>	1	311.3
4	GL430	12	884763	2939441	2054678	<i>WNK1,TULP3</i>	1	340.3
2	QG032	15	48596282	48644769	48487	<i>FBN1</i>	1	474.4
2	QG055	1	165728025	165768515	40490	<i>TMCO1</i>	1	306.8
2	QG168	1	36097571	36099487	1916	<i>COL8A2</i>	4	362.3
2	QG171	1	36097571	36099487	1916	<i>COL8A2</i>	4	252.8
2	QG174	1	36097571	36099487	1916	<i>COL8A2</i>	4	313.2
2	SG001	14	87934734	87950748	16014	<i>GALC</i>	1	254.2
4	SG012	2	110123793	110204968	81175	<i>NPHP1</i>	0	118.0

3.5 Discussion

3.5.1 The seven Mendelian-like genes

A customised POAG gene panel was developed for NGS across a homogenous cohort of patients. Chapter 2 detailed how the *MYOC* gene was the underlying genetic cause for 11 patients (3.07%) of the POAG cohort [289]. Analyses across coding sequence of the remaining six known Mendelian-like POAG causing genes revealed the genetic cause in a further four patients (1.17%). This comprised two patients with dominant disease in *NTF4*, one in *OPTN* and a further patient with *WDR36* recessive disease also.

CYP1B1

Assessing *CYP1B1* through an autosomal recessive mode of inheritance in glaucoma, no patient with homozygous or potential compound heterozygous variants were identified. Patient GB483 had both a previously identified pathogenic *MYOC*:p.K216K variant [289] and a heterozygous *CYP1B1*:p.E229K variant. *CYP1B1*:p.E229K had previously been identified in primary congenital glaucoma (PGC) and was concluded to reduce the abundance of the *CYP1B1* enzyme [290]. Theoretically, this could also reduce the metabolism of estradiol and lead to up-regulation of the mutated form of *MYOC* and an earlier onset of glaucoma [233, 291]. Patient GB483 had a later POAG onset with an age at diagnosis of 68 years old. Whilst their IOP appeared to be in the lower range of the POAG cohort (23 mm Hg), their CDR and VFMD were more severe at 0.9 and -12.58 respectively.

OPTN

One heterozygous variant, p.E92V, was found in the autosomal dominant *OPTN* gene for a single patient. The variant resides within a coiled-coil domain of the translated protein (1-127aa) which acts as the binding interface between the *OPTN* and *TBK1*

proteins [240]. Variants in the OPTN-TBK1 binding interface have been implicated in different possible mechanistic outcomes depending on where the variant is located. The *OPTN*:p.E50K variant has been previously studied in depth and was found to decrease retinal thickness and peripheral retinal ganglion cells [238], caused by enhanced binding between OPTN and TBK1 proteins [292]. With the currently available crystal structures of OPTN-TBK1 complexes revealing only partially complete structures, it was unclear how p.E92V mechanistically contributes to the POAG phenotype. However, based on previous studies on variants within the coiled-coil domain which do not directly disrupt the interaction with TBK1, p.E92V is likely to interfere with the oligomeric state of OPTN or its interaction with other unknown binding partners [240].

TBK1

Different *TBK1* mutations can manifest distinct phenotypes [293]. In NTG POAG, it has previously been suggested that *TBK1* undertakes a dominant mode of inheritance with a gain of function effect [293, 227, 294]. In contrast, encephalopathy-8, frontotemporal dementia and amyotrophic lateral sclerosis have been previously identified to be caused by loss of function heterozygous *TBK1* mutations that results in haploinsufficiency [293, 295, 296]. Fingert *et al* previously highlighted the autosomal dominant role of CNVs in NTG POAG [227], however, no copy number gains were identified in our cohort. One case of a whole gene *TBK1* deletion was identified in a single patient. This deletion was similarly reported in a previous study [297]. This might be expected given that CNV gains have been identified in NTG POAG [227] and our study focused on a homogeneously selected high tension POAG cohort. No possible pathogenic variants were located on the interacting domain of the TBK1 protein (667-728aa CTD) with OPTN. A total of four possible pathogenic SNVs were identified within *TBK1*, one of was identified at a frequency in the POAG cohort of 2.2%, although it was similarly

identified at 2.4% in the ExAC NFE population. There is no evidence in the literature suggesting that heterozygous SNVs cause a dominant form of POAG. Without supporting functional evidence, the effects of these variants cannot be speculated upon beyond ‘possible pathogenic’.

WDR36

A single patient was identified with a heterozygous possibly pathogenic p.A449T variant (known to be associated with glaucoma in Caucasian populations [80, 81]) and a heterozygous possibly pathogenic p.A509S variant in the *WDR36* gene. These variants may act as compound heterozygotes in POAG causality. The two variants were located 4079 bp apart and it could not be determined if they were on *trans* alleles by using our NGS data.

One patient (FG181) was found to harbour both a *MYOC*:p.Q368* variant and a *WDR36*:p.D658G variant. The *MYOC*:p.Q368* variant is the most common causal variant in POAG. *WDR36*:p.D658G is a variant previously reported by Footz *et al* to predispose progression of glaucoma when identified with mutations in other POAG genes [245]. Another patient (GL417) had both an *ASB10*:p.R289C variant and the *WDR36*:p.D658G variant. Both patients appeared to have CDRs in the higher range within the POAG cohort (0.90 and 0.85 respectively), however, genotype-phenotype interpretation was limited by the homogeneous severe sub-phenotype patient selection.

NTF4

Heterozygous *NTF4* variants have been suggested to underpin the genetics of NTG POAG specifically [81]. The heterozygous *NTF4* variants p.R178W and p.R107P were each identified in single POAG patients respectively. IOP was in the lower range of the POAG cohort for patients with *NTF4*:p.R178W (25 mmHg) and *NTF4*:p.R107P (21

mmHg). Had patients been selected with more modest IOPs, more possible pathogenic *NTF4* variants may have been identified.

ASB10

When assuming recessive inheritance for *ASB10*, no genotypes within this gene were identified as likely causal. It is therefore likely that the *ASB10* has a complex molecular pathology in POAG [246], or that it is less accountable for POAG in the British population.

Interestingly, Pasutto *et al* previously identified that a heterozygous synonymous variant in *ASB10* (p.T255T; rs104886478), which segregated with glaucoma in a family, altered mRNA splicing to lead to deletion of exon 3 [13]. Although not identified in our POAG cohort, it demonstrates the possible role of synonymous variants in POAG and the risk of excluding them in filtering strategies.

These findings from an analysis of rare variants in the known familial Mendelian-like genes concurs with previous suggestions that POAG is attributed to single-gene or Mendelian-like forms of POAG for approximately 5% of cases [77]. Although a number of variants could not be assigned as likely causal, we cannot exclude a role for these variants perhaps in another gene whereby interacting genes (not identified here) may be affected through epistasis.

3.5.2 Assumed complex genes

Application of standard filter criteria was insufficiently stringent to form a practicable shortlist of possible pathogenic variants. Therefore, more stringent filter criteria were applied across these genes and the most likely pathogenic variants were reported. Due to a lack of known inheritance pattern of the complex coding and non-coding POAG

genes, interpretation of likely causal genotypes was not possible or appropriate.

Assumed complex coding genes

Variants were identified in 33 (of 56) genes. This gene set is enriched for those involved in the degradation of the extracellular matrix pathway (*OPTC*, *MMP1*, *COL5A1*, *COL1A1*, *MMP9*, and *FBN1*) [298]. This reinforces a previous study by Danford *et al* which also found extracellular matrix organisation pathways to be enriched across known POAG genes [257].

After stringently filtering for the most likely pathogenic variants, *ABCA1* (n=5), *AFAP1* (n=4), *AKAP13* (n=4), *FBN1* (n=4) and *XRCC1* (n=4) notably harboured most. It is therefore possible that these genes harboured genetic variants that are involved in Mendelian-like forms of POAG. However, more functional work is needed to prove that mutated forms of these genes lead to a POAG phenotype.

The most interesting variants are discussed below in relation to their deleteriousness and/or allele frequencies.

Notable protein truncating variants

One patient harboured the *TLR4* GAT deletion, NC_000009.11: 117704472GATdel, causing the deletion of the first two bases of the ATG start codon and subsequently, a start codon loss. This resulted in the gene being untranslated and represents the first variant of its kind to be implicated in POAG causality.

A novel protein truncating stop-gain, *ABCA1*:p.R227*, was detected in a gene which was determined by the Residual Variation Intolerance database to be in the top 2.62% genes in the genome most intolerant for variants [299]. Therefore, *ABCA1*:p.R227* could disrupt the management of damaged lipids and damage lipid accumulation in the retinal ganglion cell membrane, possibly leading to apoptosis [300, 301].

Variants with the highest pathogenicity scores

Of the splice site variants, *NOS3*:p.Q411H was determined to be the most splice site disrupting variant across the POAG cohort. The Residual Variation Intolerance database revealed that the *NOS3* gene is amongst the top 22.5% most intolerant genes in the genome for variation [299]. The findings here reinforce previous studies which have demonstrated statistically and functionally that dysfunctional *NOS3* directly leads to development of a glaucoma phenotype [302, 303].

Another example includes the high CADD Phred (35) scoring *OPTC*:p.R325W and *GGA3*:p.S518L variants which were identified in five and three patients respectively. This reinforces previous work which found that rare variants in the *OPTC* gene (which localises to the cornea, iris, ciliary body, vitreous, choroid and retina [304]) are likely to be involved in POAG causality [305].

Variants at a notably higher frequency in POAG

It is not reasonable to conduct association analyses of very rare variants in modestly sized cohorts. However, it is notable that some variants appear to occur substantially more frequently in the POAG cohort compared to what is reported in the ExAC NFE database.

Considering that the most commonly causal POAG variant *MYOC*:p.Q368* variant had an allele frequency of 0.00978 in the POAG cohort compared with 0.00150 in ExAC NFE, it possible that rare and high pathogenic scoring variants with similar differences in allele frequency have roles in POAG molecular pathology. The *SRBD1*:exon3:c.81-2A>G splice site variant was identified in the POAG cohort at an allele frequency of 0.00279 compared with 0.00006 in the ExAC NFE database. However, this gene also had a pLI score of 0.00 suggesting that it was loss of function tolerant [152]. The *MMP9*:p.D165N variant was also detected far higher in the POAG cohort (AF=0.00279)

compared with ExAC NFE (AF=0.00040).

By virtue of the stringent filtering criteria, all variants in Table 3.4 are deemed highly deleterious by at least one metric and would all warrant further follow-up. The variants discussed above may help prioritise functional analyses.

CDKN2B-AS1

The three non-coding RNA genes (ncRNA) known to be associated with POAG were filtered using stringent alternative filter criteria. Five variants remaining following filtering of the ncRNA genes belonged exclusively to the *CDKN2B-AS1* gene which were observed across a total of 36 patients in this cohort. The *CDKN2B-AS1* gene is known to regulate the transcription of cyclin-dependent kinase inhibitor 2A and 2B (*CDKN2A* and *CDKN2B*) [306] which are both upregulated in glaucoma [113]. Further functional research must be completed to provide a stronger case for mutations in this non-coding gene predisposing to POAG.

Whole-gene pathogenicity scores

Filtering criteria excluded variants that were more common or reportedly less deleterious. GenePy provided a method to assess the combined burden of all variants within a gene. It reduces the testing burden and attempts to discriminate genes more often pathogenically mutated in affected cohorts or affected individuals within the cohort.

ABO

The *ABO* gene is responsible for the ABO blood group system. The four major groups in the ABO system (A, B, AB, and O) result from 3 major alleles (A, B, and O) of the *ABO* gene [307]. Although variation in *ABO* was more recently associated with hypertension [278], higher incidences of POAG were previously reported [308, 309]. Patients

in the POAG cohort harboured significantly higher *ABO* whole-gene pathogenicity scores compared to the non-eye disease cohort. Genes identified to be associated with POAG based on whole-gene pathogenicity scores had either (1) a small number of rare but high CADD scoring variants or (2) had a higher frequency of variants which individually had small effect [275]. Variants in the *ABO* gene had modest pathogenicity scores (maximum CADD Phred score of 13.26) which would not pass standard filter criteria on an individual variant basis. Therefore, it is possible that the molecular pathology underpinning *ABO* in POAG is the result of a compound effect from multiple modest effect size variants.

SIX6

SIX6 harboured two stringently filtered high impact or rare high pathogenicity scoring variants across a total of 11 patients. This gene was also identified as having significantly elevated whole gene pathogenicity scores in the POAG cohort over the control cohort. The 11 samples with stringent filtered possibly pathogenic variants were also 11 of the highest 12 GenePy scoring samples in the POAG cohort for this gene.

SIX6 is understood to be involved in ocular development and has previously been associated with the morphology of the optic nerve [310]. The rs33912345 SNP of *SIX6* (p.H141N), which is a common variant in the ExAC NFE population (AF=0.5843), was previously found to be GWAS associated with POAG [310]. Associated cases were also found to have statistically thinner retinal nerve fibre layer [310, 311].

Following stringent filtering, ten patients with p.A61V and one patient with p.E129K were identified in *SIX6*. Interestingly, these 11 patients also harboured the rs33912345 variant (study cohort AF=0.606). Therefore, previous associations for rs33912345 SNP in POAG may have indicated the presence of the more damaging p.A61V and p.E129K variants in *SIX6*. It is possible that *SIX6* follows a very rare autosomal dominant mode

of inheritance in POAG.

FAM27E5

Very little is known about the function of the Family With Sequence Similarity E5 (*FAM27E5*) protein. Although the *FAM27E5* gene did not have assumed statistically higher whole-gene pathogenicity in POAG cases compared to non-eye disease controls (as the control cohort did not harbour any variants across the gene), two patients (QG195 and GL515) possessed nominally higher GenePy scores when compared to the rest of the POAG cohort's genes. Single highly pathogenic scoring variants were responsible for the elevated whole-gene pathogenicity scores, suggesting that this gene may follow a rare autosomal dominant mode of inheritance in POAG.

3.5.3 Limitations

The majority of the stringently filtered variants were identified as heterozygous in one or very few patients which is expected given all previous data on the genetics of POAG. Therefore, the limitations experienced here reinforce the desirability of larger sample sizes to detect likely causal variants in POAG.

While we were motivated to select patients with an unambiguous and strong POAG phenotype, the homogenous cohort impaired our ability to scrutinise genotype-phenotype correlations. Following this work, it would be useful to genotype the entire database of 1679 glaucoma patients for as many of the variants identified here as possible and assess phenotype correlation.

This work focused on Caucasians and excluded other POAG associated genes such as *LMX1B*, *HMGA2*, *MAP3K1*, and *LOXL1* which have recently been associated with POAG in Asians, although they fail to replicate in Europeans and Africans [312]. Gene

panels regularly become obsolete as genes become better characterised and more genes are known to be associated with the phenotype. Since the design of the custom POAG gene panel in 2016, nine genes have been subsequently identified to be associated with POAG in a multi-ethnic GWAS [313]. Similarly, a large meta-analysis of Europeans with POAG identified associations for 26 loci which were not part of the custom POAG gene panel that was used in this study [120].

Although custom targeted sequencing generally captures target genes more comprehensively compared to WES and WGS, a multitude of limitations remain. As this work has shown, the targeted sequencing designs may also be based on out-dated databases which may miss functionally important transcripts.

The nature of the targeted sequencing proved problematic for CNV calling. Over-calling and high false positive rates for exome data CNV detection (which are based on read depth) are a known issue that was previously reported [314, 315]. Due to the punctuate nature of the sequencing data and the evidence suggestive of frequent false positives, the CNV calling herein was suboptimal.

Capture chemistry based artefacts that were specific to the POAG cohort compared to the non-eye disease cohort caused false positive results (e.g. *IL1B*). The POAG and non-eye disease cohorts ideally could have undergone consistent NGS sequencing using the same chemistry and sequencing platforms. However, this was not realistic given the expenses this would have incurred.

3.6 Conclusion

This study reveals further insights into POAG gene mechanisms and expands on the known molecular pathologies, providing new insights into POAG genes. While this

work suggests that targeted gene panels for diagnostic testing in POAG would be of limited impact to determine genetic aetiology, it has implicated a number of possible pathogenic variants in Mendelian-like and complex POAG genes. This work has generated a number of interesting hypotheses that require functional assessment and genotyping in larger, more phenotypically heterozygous glaucoma cohorts.

4 Nystagmus whole exome analysis

4.1 Synopsis

This chapter utilises whole-exome sequencing (WES) data of nine patients to identify likely causal variants for their respective ophthalmic conditions. Using an in-house bioinformatic pipeline to process these data, called variants are filtered and interpreted for likely causality.

Luke O’Gorman processed the raw FastQ files through the bioinformatic pipeline, performed the quality control and interpreted likely causal variants. Mr Jay Self was the referring clinician for the patients and was responsible for phenotyping, recruitment and clinical guidance. Library preparation and sequencing was performed by WISH laboratory technicians.

4.2 Background

Nystagmus is an ophthalmic condition characterised by the oscillation of one or both eyes about one or more axes which ultimately leads to visual impairment of an individual [20]. The overall prevalence of nystagmus in the UK is approximately 0.0024 [19].

Infantile nystagmus syndrome (INS) can be caused by idiopathic INS (IINS) or through a range of ophthalmic diseases including albinism or retinal dystrophies (Figure 1.2). IINS and albinism are the two most common forms of INS [19].

INS is a known clinical feature identified in patients diagnosed with ocular albinism (OA) or oculocutaneous albinism (OCA) [25]. It has also been described as a feature of

many neurological and systemic conditions including in episodic ataxia type 2 [27] and Noonan syndrome [29]. For more information on classes of INS causes from albinism, syndromes or neurological diseases, please see chapter 1.2.1.

Whole genome sequencing (WGS) and whole exome sequencing (WES) offer methods for obtaining next-generation sequencing (NGS) data for analyses to potentially identify likely causal variant(s) [158]. The exome accounts for approximately 1% of the genome [316] and offers the cheaper method for sequencing [155, 317], storage and analytical costs [318, 319]. For more information regarding the different DNA sequencing methods available, please refer to chapter 1.4.3. With these costs and limitations in mind, and since 85% of causal mutations are located within the exome [159, 155], WES is widely used as a method for causal gene/ variant identification in Mendelian diseases [155, 156, 157]. Exome studies have also previously been successful in identifying causal variants in INS patients [320, 321].

4.3 Aim

Within University Hospital Southampton, nine patients which had nystagmus and first degree relatives with nystagmus were selected for whole exome sequencing (WES) to resolve their respective genetic aetiologies. The NGS data obtained from the patient samples were processed with an in-house bioinformatic pipeline. The variants detected for each individual are filtered and interpreted to determine likely causality.

4.4 Methods

4.4.1 Patient selection

Early-onset nystagmus patients were selected for WES which had a strong family history for nystagmus (Figure 4.1). As shown in Table 4.1, this included five patients diagnosed by the referring clinician for IINS diagnoses (samples NG149-2, NG-PE, RNAR1, NG151-2 and NG156). A further four patients were selected with more complex presentations including diagnoses of nystagmus with foveal hypoplasia (sample NRS), nystagmus and retinal dystrophy (sample NG045-1) and nystagmus with ocular albinism and Noonans syndrome (NG222). Patient NG241 had nystagmus with mild albinism (NG241). The mother of NG241 had potentially mild albinism, therefore, NG241 was included for WES.

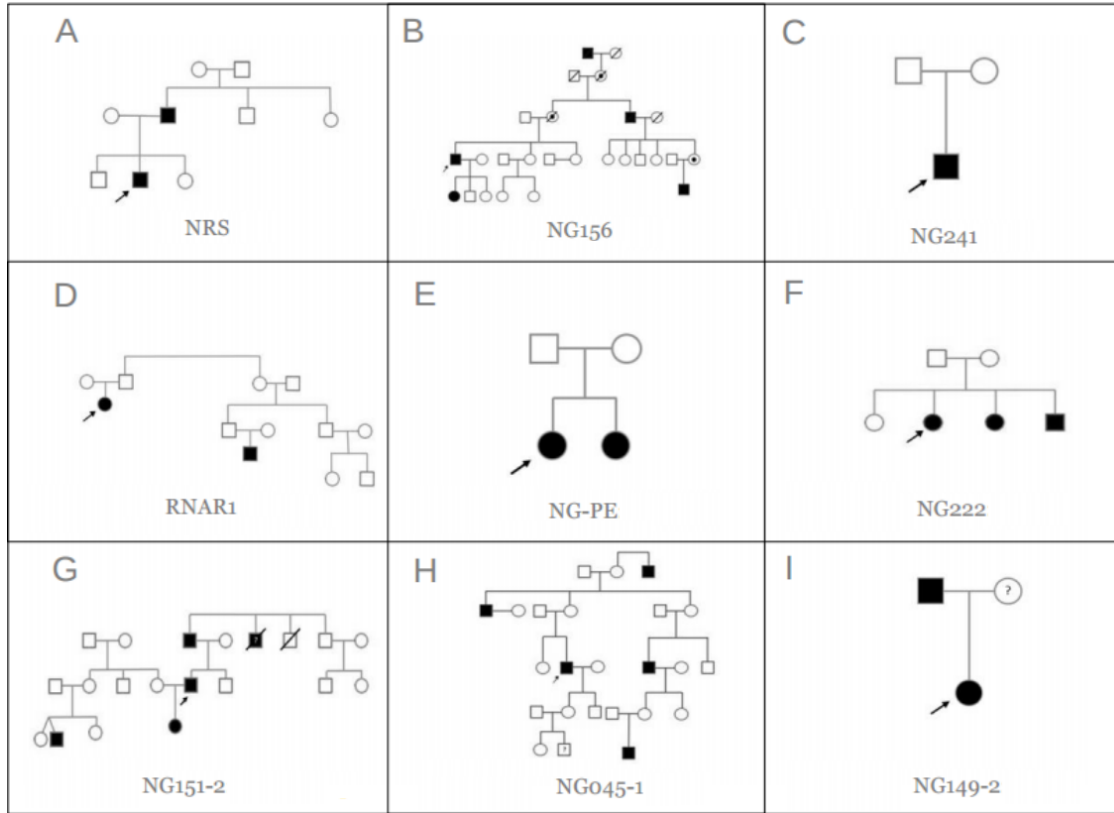


Figure 4.1: *Pedigrees for the 9 unrelated nystagmus patients with their sample ID below. Proband is indicated by a black arrow.*

Table 4.1: *List of nine unrelated nystagmus patients and their phenotypes.*

Sample ID	Gender	Phenotype
NG045-1	Male	XL Nystagmus, possible retinal dystrophy
NG149-2	Male	Nystagmus
NG151-2	Male	Nystagmus
NG156	Male	XL Nystagmus
NG222	Female	Nystagmus, oculocutaneous albinism and Noonans syndrome
NG241	Male	Nystagmus, possible albinism
NG-PE	Male	Nystagmus
NRS	Male	Nystagmus and foveal hypoplasia
RNAR1	Female	Nystagmus

4.4.2 Sequencing and bioinformatic pipeline

Following patient selection and phlebotomy, the main steps in this project are illustrated in Figure 4.2. DNA was extracted using Oragene-DNA kit

(OG-575). Library preparation of the samples were performed using Agilent SureSelect V5 with an enrichment based capture (see Figure 1.9B). The sequencing was performed with Illumina (HiSeq 2000) paired-end sequencing. NGS data was aligned against the human reference genome (hg19) using Novoalign (novoalign/2.08.02). Variant calling was performed using SAMtools v0.1.19 [143] and annotated using ANNOVAR [146] with RefSeq [177], dbSNP v135, Exome Sequencing Project 6500 (ESP6500) [151] and conservation-based pathogenicity scores of SIFT [153] and GERP++ [154] databases. Splice variants were annotated with a maximum entropy algorithm (MaxEnt [322]) to identify variants which disrupt splice sites. Variants were excluded if they had a Phred quality score of <20 or a read depth of <4 . All annotated variants were cross-checked with an in-house database of 816 exome sequenced samples to give estimates of local Wessex population allele frequencies.

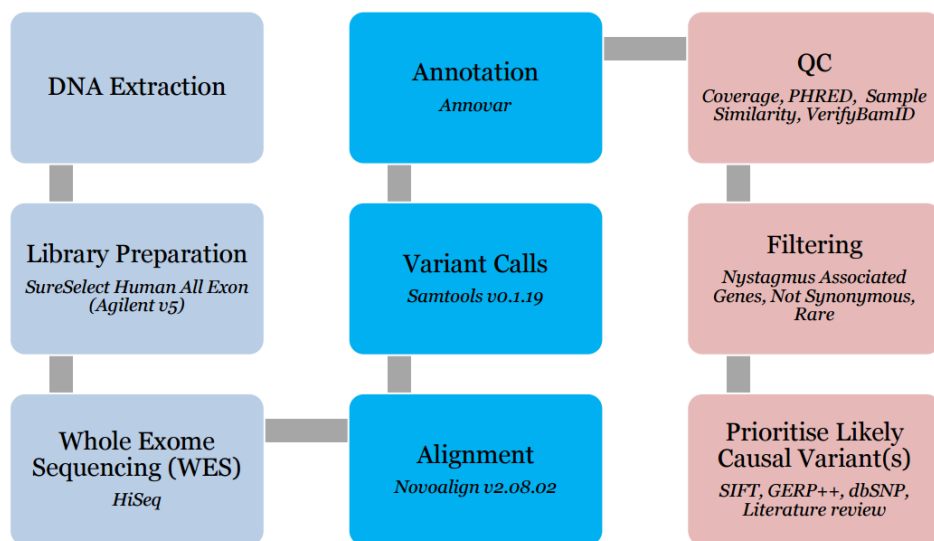


Figure 4.2: **Overview of the stages involved in identifying the causal variants.** Wet-laboratory stages, left; bioinformatic pipeline, middle; bioinformatic QC, filtering and identification, right.

4.4.3 Quality control

Quality control (QC) was performed using the aligned NGS data to inspect coverage and contamination. Coverage across the exome was calculated by using BEDtools [277] at 1X, 5X, 10X and 20X depth across the AgilentSureSelect v5 target region. The ideal depth of coverage has varied as next-generation sequencing chemistry improve reducing the number of low quality base calls and GC bias. Library preparation is also a major factor affecting coverage as GC biases can be introduced during PCR reducing coverage of these regions. Whilst the average mapped depth in genome sequencing is 35X [168], up to 50X depth has been identified as desirable [323]. However, the ideal distribution of horizontal coverage across the desired target is not defined. 20X has been previously used as an acceptable threshold for accurate variant calling [324] whilst 80% at 20X depth has been previously used as a target threshold for horizontal coverage [325].

Samples were also analysed to detect DNA sample contamination and ensure sex concordance by assessing autosomal and X chromosome heterozygosity. An excess of heterozygosity at the X chromosome would indicate that the sample pertains to a female patient. A male patient would be expected to elicit less X chromosome heterozygosity due to the hemizygous status of the X chromosome. VerifyBamID [191] was used to determine the degree of non-reference bases and excessive heterozygosity observed across reference sites. A 0.02 freemix threshold was indicative of contamination [191]. Variant sharing or similarity of the WES samples was checked for appropriate sample relationships. This QC can also highlight differences in ethnicity.

4.4.4 Variant prioritisation

Candidate gene lists were generated through interrogation of the Human Gene Mutation Database (HGMD, 2013 [172]). A first tier of genes were selected using a ‘nystagmus’ phenotype search term (see Table 4.2; $n = 45$). A second tier gene list was generated by selecting nystagmus with ‘albinism’ and ‘ataxia’ phenotype search terms (see Table 4.3; $n = 168$). The referring clinician reported that patient NG045-1 was likely to have a retinal disorder and a previous WES for this sample analysis using nystagmus HGMD candidate genes had been unsuccessful. Therefore, it was necessary to expand the candidate gene list for NG045-1 for retinal disorder genes using HGMD and UKGTN (2015) [171] (see Table 4.4; $n = 242$).

Annotated variants were filtered systematically for each patient. Variants were filtered for presence within the nystagmus associated genes (HGMD), excluding those which were synonymous, occurring less than 5% in the ESP6500 (all populations) reference database whilst also being rare within our in-house database (<10 samples). This systematic filtering process provides a shortlist of variants, each of which were assessed using various in silico functional prediction tools for pathogenicity (e.g. GERP++ or SIFT) and a literature search to identify evidence of causality. All likely causal candidates were verified and confirmed by using Sanger sequencing (see Chapter 1.4.3.1) and segregation analysis performed when possible.

Table 4.2: *List of 45 nystagmus genes. Gene list generated using a HGMD ‘nystagmus’ phenotype search term.*

Tier 1 Genes				
<i>AIPL1</i>	<i>CRYAA</i>	<i>HPS1</i>	<i>PAX6</i>	<i>SETX</i>
<i>ATP1A2</i>	<i>DGUOK</i>	<i>LRP5</i>	<i>PLP1</i>	<i>SLC16A2</i>
<i>BLOC1S3</i>	<i>FRMD7</i>	<i>MTRR</i>	<i>POLG</i>	<i>SLC1A3</i>
<i>CACNA1A</i>	<i>GATA3</i>	<i>NDUFAF2</i>	<i>PRKCG</i>	<i>SLC45A2</i>
<i>CACNA1F</i>	<i>GJC2</i>	<i>NPHP1</i>	<i>ROBO3</i>	<i>SLC4A11</i>
<i>CASK</i>	<i>GNAT2</i>	<i>NUBPL</i>	<i>RPE65</i>	<i>TULP1</i>
<i>CNGA3</i>	<i>GPR143</i>	<i>NYX</i>	<i>RPGRIP1</i>	<i>TYR</i>
<i>CNGB3</i>	<i>GUCY2D</i>	<i>OCA2</i>	<i>SACS</i>	<i>TYRP1</i>
<i>CRX</i>	<i>HEXA</i>	<i>PAX2</i>	<i>SCP2</i>	<i>WT1</i>

Table 4.3: *List of 168 nystagmus with albinism or nystagmus with ataxia genes. Gene list generated using HGMD ‘albinism’ and ‘ataxia’ with nystagmus search terms.*

Tier 2 Genes							
<i>AAAS</i>	<i>CACNA1A</i>	<i>ERCC8</i>	<i>KCNA1</i>	<i>NDUFV1</i>	<i>PEX7</i>	<i>RNF170</i>	<i>SRD5A3</i>
<i>ABCB7</i>	<i>CACNB4</i>	<i>FGF14</i>	<i>KCNC3</i>	<i>NEU1</i>	<i>PHYH</i>	<i>RNF216</i>	<i>SURF1</i>
<i>ABCD1</i>	<i>CC2D2A</i>	<i>FLVCR1</i>	<i>KCND3</i>	<i>NKX2-1</i>	<i>PIK3R5</i>	<i>SACS</i>	<i>SYNE1</i>
<i>ABHD12</i>	<i>CEP290</i>	<i>FTL</i>	<i>KCNH2</i>	<i>NPC1</i>	<i>PLA2G6</i>	<i>SCN1A</i>	<i>SYT14</i>
<i>ADCK3</i>	<i>CLCN1</i>	<i>FXN</i>	<i>KCNJ10</i>	<i>NPC2</i>	<i>PLEKHG4</i>	<i>SCN2A</i>	<i>TDP1</i>
<i>AFG3L2</i>	<i>CNTN1</i>	<i>GAD1</i>	<i>KCNQ2</i>	<i>NPHP3</i>	<i>PLP1</i>	<i>SCN4A</i>	<i>TGM6</i>
<i>AHI1</i>	<i>CNTN4</i>	<i>GAN</i>	<i>KCTD7</i>	<i>NUBPL</i>	<i>PLTP</i>	<i>SCN8A</i>	<i>TMEM67</i>
<i>ANO10</i>	<i>COX20</i>	<i>GBA2</i>	<i>KIAA0226</i>	<i>OCA2</i>	<i>PMM2</i>	<i>SDHA</i>	<i>TPP1</i>
<i>APOA1</i>	<i>CP</i>	<i>GDAP1</i>	<i>KRIT1</i>	<i>OPA1</i>	<i>POLG</i>	<i>SETX</i>	<i>TRAPPC11</i>
<i>APTX</i>	<i>CYP27A1</i>	<i>GFAP</i>	<i>L2HGDH</i>	<i>OPA1TV8</i>	<i>POLR3A</i>	<i>SIL1</i>	<i>TTBK2</i>
<i>ARL13B</i>	<i>DARS2</i>	<i>GJA1</i>	<i>MANBA</i>	<i>OPA3</i>	<i>PPT1</i>	<i>SLC16A2</i>	<i>TTPA</i>
<i>ATCAY</i>	<i>DLAT</i>	<i>GJB1</i>	<i>MAPT</i>	<i>OTUD4</i>	<i>PRF1</i>	<i>SLC17A5</i>	<i>TTR</i>
<i>ATM</i>	<i>DLG</i>	<i>GJC2</i>	<i>MECP2</i>	<i>PANK2</i>	<i>PRICKLE1</i>	<i>SLC1A3</i>	<i>TYR</i>
<i>ATP1A2</i>	<i>DNAJC19</i>	<i>GLB1</i>	<i>MLC1</i>	<i>PARK2</i>	<i>PRKCG</i>	<i>SLC25A15</i>	<i>TYRP1</i>
<i>ATP2B3</i>	<i>DNMT1</i>	<i>GOSR2</i>	<i>MPZ</i>	<i>PAX6</i>	<i>PRND</i>	<i>SLC2A1</i>	<i>UBE3A</i>
<i>ATP8A2</i>	<i>DPM1</i>	<i>GPR143</i>	<i>MRE11A</i>	<i>PDHA1</i>	<i>PRNP</i>	<i>SLC45A2</i>	<i>UQCRCQ</i>
<i>ATR</i>	<i>EEF2</i>	<i>GRM1</i>	<i>MTPAP</i>	<i>PDHB</i>	<i>PRPS1</i>	<i>SLC4A4</i>	<i>UROCI</i>
<i>ATXN8</i>	<i>EIF2B2</i>	<i>IFRD1</i>	<i>MYO7A</i>	<i>PDYN</i>	<i>PRX</i>	<i>SLC6A19</i>	<i>VAMP1</i>
<i>BRCA1</i>	<i>EIF2B3</i>	<i>ISCU</i>	<i>NBN</i>	<i>PEX10</i>	<i>PTEN</i>	<i>SLC9A6</i>	<i>VHL</i>
<i>C10orf2</i>	<i>EIF2B4</i>	<i>ITM2B</i>	<i>NDUFAF2</i>	<i>PEX12</i>	<i>RAD50</i>	<i>SPAST</i>	<i>VLDLR</i>
<i>CA8</i>	<i>EIF2B5</i>	<i>ITPR1</i>	<i>NDUFS1</i>	<i>PEX2</i>	<i>RNF168</i>	<i>SPTBN2</i>	<i>WFS1</i>

Table 4.4: *List of 242 retinal disorder genes. Gene list generated using a HGMD and UKGTN (2015) ‘retinal disorder’ search term.*

<i>Retinal Disorder Genes</i>										
ABCA4	BBS4	CNGB1	DFNB18A	IGFBP7	MAB21L2	PNPLA6	RP	RP36	RP59	TEAD1
ABCC6	BBS5	COD3	DFNB2	ITM2B	MAK	PNPT1	RP1	RP37	RP60	TOPORS
ACHM3	BBS6	COD4	DFNB23	JBTS5	MCOR2	POC1B	RP10	RP38	RP61	TREX1
ACHM4	BBS7	COL2A1	DFNB31	KCNJ13	MCOPS5	PPCRA	RP11	RP39	RP7	TSPAN12
ACO2	BBS8	COL4A1	DHRD	KCNJ13	MERTK	PROM1	RP12	RP4	RP9	TTLL5
ADAM9	BBS9	CORD10	DNM2	KCNV2	MIR204	PRPF31	RP13	RP40	RPE65	TUB
ADAMTS18	BEST1	CORD11	DRAM2	LCA1	MKKS	PRPF6	RP14	RP41	RPGR	TULP1
AHI1	C1QTNF5	CORD12	DYRK1A	LCA10	MKS1	PRPH2	RP17	RP42	RPGRIP1L	USH1
AIPL1	CACD2	CORD13	EFEMP1	LCA11	MKS4	PRPS1	RP18	RP43	RS1	USH1C
APC	CACNA1F	CORD15	EFTUD2	LCA12	MVK	RAX2	RP19	RP44	SACS	USH1D
ARB	CACNA2D4	CORD2	EVR1	LCA13	NDP	RBP3	RP1L1	RP45	SCRA	USH1F
ARMD2	CDHR1	CORD5	EVR2	LCA14	NFRCD	RBP4	RP2	RP46	SDCCAG8	USH1G
ARMD6	CDK11A	CORD7	EVR4	LCA15	NR2E3	RCD3B	RP20	RP47	SIX6	USH2A
ATOH7	CEP290	CORD9	EYS	LCA2	NRL	RCD4	RP25	RP48	SLC24A1	USH2C
BBS1	CERKL	CPHD6	F5	LCA3	NRL	RD3	RP26	RP49	SLC24A2	USH2D
BBS10	CERS2	CRB1	FAM161A	LCA4	OCMD	RDH12	RP27	RP50	SLSN6	USH3A
BBS11	CFH	CRX	GDF5	LCA5	OTX2	RDH5	RP28	RP51	SNRNP200	VEGFA
BBS12	CHM	CSNBAD1	Gene	LCA6	PANK2	RGS9	RP3	RP54	SP4	VMD
BBS13	CHRNA7	CSNBAD2	GRID2	LCA7	PDE6A	RGS9	RP30	RP55	SPATA7	VPS13B
BBS14	CLN3	CTC1	GUCA1A	LCA8	PDE6B	RGS9BP	RP31	RP56	STGD1	VRCP
BBS2	CLRN1	CYP4V2	GUCA1B	LORD	PERRS	RHO	RP33	RP57	STGD3	VSX2
BBS3	CNGA1	DFNA11	IFT140	LRAT	PITPNM3	RLBP1	RP35	RP58	STGD4	ZNF674

4.5 Results

4.5.1 Quality control

Coverage across the exome for 1X, 5X, 10X and 20X depth was shown in Table 4.5. All samples achieved a target region coverage greater than 77% at 20X depth. Although the ideal depth of coverage is arbitrary and dependent on the resources available, it is generally considered for WES that 20X depth is desirable [324, 325].

Table 4.5: *Coverage across the exome for 5X, 10X and 20X depth.*

SampleID	Total reads	Mean depth	20X (%)	10X (%)	5X (%)	1X (%)
NG045-1	77950389	102.33	96.40	98.54	99.13	99.52
NG149-2	40095445	52.67	87.51	96.56	98.62	99.40
NG151-2	28377930	40.36	77.47	92.91	97.65	99.40
NG156	49725619	63.41	92.82	97.94	98.97	99.44
NG222	42470346	58.12	87.71	96.39	98.64	99.49
NG241	41632343	54.95	87.47	96.46	98.64	99.42
NG-PE	41600348	55.54	88.49	96.75	98.64	99.39
NRS	45891587	63.57	91.40	97.49	98.91	99.52
RNAR1	36889993	49.66	85.48	95.92	98.42	99.36

X chromosome heterozygosity was assessed for all 9 samples. The reported gender and determined gender concurred for all samples (Table 4.6). VerifyBamID freemix (degree of non-reference bases and excessive heterozygosity observed across reference sites) was below the 0.02 freemix threshold indicating no evidence of contamination. In some cases, X chromosome variants should be interpreted with caution. Pseudoautosomal regions are homologous regions between the X and Y chromosomes. Genetic variants in these regions can, therefore, be interpreted similarly to autosomal chromosome genetic variants. Aneuploidy diseases such as Klinefelter’s syndrome are chromosomal conditions in which affected individuals have an additional X chromosome, which would allow the affected individual to carry an X-linked recessive allele without presenting with the phenotype.

The degree of shared variants between each annotated sample was consistently $\sim 43\%$ (Table 4.7). This indicates that the patients appear to be ethnically similar and unrelated. As the patients are Caucasian singletons, the result confirms correct uniform ethnicity and unrelatedness.

Table 4.6: *Assessment of X chromosome heterozygosity and the VerifyBamID freemix for 9 samples. Recorded gender reported by clinician and determined gender with <50% heterozygosity of the X chromosome imputing a male gender and >49 imputing a female gender.*

	NG045-1	NG149-2	NG151-2	NG156	NG222	NG241	NG-PE	NRS	RNAR1
No. variants	23787	24767	25150	25150	24639	22409	24688	24837	24782
% Het	60.4111	61.6344	61.1809	61.1809	61.5609	60.1678	62.1881	60.9897	61.4841
% X Het	16.9054	56.7227	16.1202	16.1202	55.1502	18.1818	60.7516	15.5763	59.6950
Recorded Gender	Male	Female	Male	Male	Female	Male	Female	Male	Female
Determined Gender	Male	Female	Male	Male	Female	Male	Female	Male	Female
Freemix	0.00009	0.00065	0.00406	0.00044	0.00014	0.00056	0.00048	0.00226	0.00260

Table 4.7: *Pairwise calculation of to show the percentage of total variants common to both samples.*

	RNAR1	NRS	NG151-2	NG-PE	NG045-1	NG149-2	NG156	NG222	NG241
RNAR1	100	43.33	43.02	43.99	41.97	43.1	42.4	43.26	39.14
NRS	43.23	100	44.95	43.86	42.34	43.14	44.18	43.51	40.38
NG151-2	42.7	44.72	100	43.91	41.45	42.9	43.65	42.98	39.78
NG-PE	44.16	44.12	44.4	100	42.7	43.6	44.03	44.36	39.59
NG045-1	43.73	44.21	43.49	44.32	100	43.94	44.15	44.07	40.07
NG149-2	43.13	43.26	43.24	43.46	42.2	100	43.79	44.17	39.11
NG156	41.78	43.63	43.32	43.22	41.75	43.13	100	43.15	38.98
NG222	43.51	43.86	43.54	44.45	42.54	44.4	44.05	100	39.71
NG241	43.28	44.75	44.31	43.61	42.54	43.22	43.75	43.66	100

4.5.2 Causal variant analysis

A mean of 25661 variants were identified for the 9 samples. Upon restricting variants for those which are in tier 1 and tier 2 candidate genes, the mean number of variants was 353 for the 9 samples. Following applying further filters to exclude synonymous variants, inclusion of variants with less than 5% minor allele frequency in the ESP6500 (all populations) database and occurring in less than 10 control samples within the in-house database, a mean of 10 variants remained.

Strong candidate variants were identified in 5 of the 9 patients analysed (Table 4.8). All causal candidates were unique to their respective patient and were absent or had a very low frequency (0.0002) in the ESP6500 (all populations) reference database.

Table 4.8: *Potential causal variants identified following the applied filtering for each patient.* Each variant was annotated with sample ID, chromosome, zygosity, gene name, variant type, variant information, ESP6500 (all populations allele frequency), SIFT (lower value suggests high conservation) and GERP++ (genomic evolutionary rate profiling, higher value suggests high conservation). Green shading indicates the patients with a resolved aetiology; Orange shading indicates patients that are under follow-up; Red shading indicates patients for whom the analysis did not resolve genetic aetiology for nystagmus; Yellow shading indicates the justification for the variant being a prime candidate for causality.

Sample	Chrom	Het/ Hom	Gene	Variant Type	Variant Information	ESP6500 (EA)	Sift	GERP++
NRS	11	Het	<i>PAX6</i>	stopgain	exon11:c.C116G:p.Y372*	0.0000	-	-
NG156	X	Hom	<i>GPR143</i>	nonsynonymous	exon1:c.T17C:p.L6P	0.0000	0.00	4.35
NG222	11	Het	<i>TYR</i>	nonsynonymous	exon1:c.A56G:p.H19R	0.0000	0.00	6.07
NG-PE	19	Het	<i>CACNA1A</i>	nonsynonymous	exon27:c.T4337A:p.V1446E	0.0000	0.00	4.88
NG241	11	Het	<i>TYR</i>	nonsynonymous	exon1:c.C242T:p.P81L	0.0002	0.00	6.07
RNARI	17	Het	<i>PEX12</i>	splicing	exon2:c.126+1 G	0.0002	-	5.78
NG151-2	13	Het	<i>ATP8A2</i>	nonsynonymous	exon22:c.A1876G:p.T626A	0.0000	0.00	5.93
NG045-1	13	Het	<i>ATP2B3</i>	nonsynonymous	exon8:c.C1211A:p.A404E	0.0000	0.58	5.07
NG149-2	13	Het	<i>CEP290</i>	frameshift deletion	exon40:c.5416_5420del: p.1806_1807del	0.0000	-	-

4.5.3 Patients with verified likely causal variants

Four patients had their genetic aetiology determined for disease following WES analyses. Patient NRS was found to have a premature stop codon in exon 11 of the *PAX6* gene. Mirzayans *et al* [326] found that *PAX6* mutations in exon 11 are strongly associated with keratitis. The *PAX6* variant observed in this patient was a damaging stop-gain which was confirmed with Sanger sequencing. This variant was also found to segregate with the affected father (see Figure 4.1A). Unaffected family members could be tested in future.

Patient NG156 has a pedigree which was suggestive of X-linked inheritance (Figure 4.1B) and was found to have a novel SNP in the *GPR143* gene, which upon translation, substitutes a leucine for a structurally damaging proline. *GPR143* is found in several studies to be involved in X-linked nystagmus. Specifically, mutations in exon 1 have been identified in Chinese populations to be most likely causal for their X-linked

nystagmus [327, 328]. This exon is also known to be causal for ocular albinism in Caucasian patients if mutated [329, 330]. With this supporting evidence and verification by Sanger sequencing, the novel variant was accepted as likely causal. Segregation studies were performed on the proband's affected daughter (with the same phenotype) and results showed that the variant segregated.

Patient NG222 was reported by the referring clinician as having oculocutaneous albinism and Noonans syndrome. A nonsynonymous *TYR* variant was detected which had damaging pathogenicity score in SIFT (0) and GERP (6.07). However, this deleterious variant may not reduce *TYR* activity sufficiently to yield the OCA1B phenotype since the parents of the proband are unaffected [331]. However, when two common missense *TYR* SNPs, S192Y and R402Q are identified on *trans* alleles to the pathogenic variant, this can cause the OCA1B phenotype [332, 333]. Patient NG222 was found to have two common missense *TYR* SNPs, rs1042602 (S192Y) and rs1126809 (R402Q) with ExAC (all populations) minor allele frequencies of 0.25 and 0.18 in all populations respectively [152]. These variants have been confirmed with Sanger sequencing and are the most likely cause of the proband's OCA1B. Sanger has confirmed the verified the variant and segregation will be performed in parents in the future to further verify pathogenicity. This type of 'tri-allelic genotype' in OCA1B is discussed in more detail in Chapter 5.

NG-PE was found to have a SNP mutation within the *CACNA1A* gene. This variant had damaging SIFT (0) and GERP (4.88) scores. Mutations in exon 27 of this gene are known to cause ataxia (SCA6/EA2) [334] for which nystagmus is part of. This variant was found to segregate in the sister of the proband who had the same phenotype. Subsequently, the patient was given clinical diagnosis. Although ataxia was not

presented in the patient, it has previously been shown that this may occur later in life [335].

4.5.4 Patients undergoing follow-up

Patient NG241 was recorded as having subtle albinism phenotype with the nystagmus condition. The *TYR* variant displayed in Table 4.8 has damaging pathogenicity scores in SIFT (0) and GERP (6.07). Giebel *et al* [336] identified albinism patients with a CCT to CTT change in codon 81 resulting in a substitution of proline to leucine. The identified variant's associated rsID, rs28940876, was flagged by dbSNP/ ClinVar to be pathogenic for oculocutaneous albinism (type 1B). A 'tri-allelic genotype' was not found for this patient. Sanger sequencing has confirmed the presence of the known pathogenic variant and it was accepted as partially causal. As this variant is insufficient to cause the phenotype alone, this case requires follow-up analysis for identification of further pathogenic variants within *TYR* or other albinism genes. The mother of the proband was reported as potentially having very subtle albinism. DNA was available for the parents and a segregation analysis could be performed in the parents in the future.

4.5.5 Patients without likely causal variants

There were also four patients with an unresolved genetic aetiology. Patient RNAR1 has a pedigree with a recessive form of inheritance (Figure 4.1D). The most promising variant found was a high MaxEnt scoring (8.5) *PEX12* splice variant which blocks the removal of intron 1 leading premature termination of protein synthesis. The *PEX12* variant was not found in other patients in the Southampton in-house database and was of unknown pathogenic or clinical significance according to dbSNP. Chang & Gould

indicate that a second variant may be required for causality [337], however, no other variant was found for the *PEX12* gene. The referring clinician indicated that only limited phenotyping was possible.

Patient NG151-2 is an ongoing case with no candidates identified through initial filtering parameters, the Southampton database parameters were relaxed and a SNP in the *ATP8A2* gene was identified as a candidate with a low SIFT (0) and a high GERP (5.93) score. However, this candidate was present in 19 individuals in Southampton without nystagmus and was found in 0.08% of individuals in the ExAC (all populations) database.

Patient NG045-1 (Figure 4.1H) was observed from previous tests performed on this patient to have a likely retinal or cone dysfunctionality. Using a tier 1 and 2 candidate gene list, one variant in the *ATP2B3* gene remained following filtering. This variant had a significant GERP score (5.07) but insignificant SIFT score (0.58). Two alternate allele calls were found to originate from the same read pair and, therefore, only 1 of 5 paired reads were of the alternate allele. There was insufficient depth for the alternate allele to have confidence on this variant. Therefore, the *ATP2B3* variant was not considered for further analysis. Expanding the candidate gene list to UKGTN retinal disorder genes (Table 4.4) was not able to identify a candidate causal variant. No further information was able to be collected about the family as they have now left the area.

The patient, NG149-2, had a candidate causal frameshift deletion located in a tier 2 gene, *CEP290*. The variant did not have high quality data with a read depth of 5. This variant was not found to segregate with the father (Figure 4.1I) and was disregarded for further analysis.

4.6 Conclusion

The bioinformatic pipeline and filtering method was used successfully in identifying causal variants in 4 of 9 patients with a further promising variant undergoing follow-up analysis. In future analyses, additional candidate causal variants may be identified in patients by: refining the pipeline to [17]; using updated candidate gene panels and; by refining the clinical data on patients and their relatives. More detailed phenotyping is essential to increase the chance of successful determination of genetic basis for disease. Our results demonstrate success in identifying the genetic aetiology in 44% patients with an additional follow-up study likely to increase this rate. X chromosome genes may present genetic analysts with potentially unexpected results due to unique characteristics and processes which occur uniquely to this chromosome. WES analyses are a powerful methodology in determining causes of disease in patients to directly inform future screening, diagnosis and prognosis.

5 Identification of a functionally significant tri-allelic genotype in the tyrosinase gene causing hypomorphic oculocutaneous albinism (OCA1B)

5.1 Synopsis

This chapter investigates tri-allelic *TYR* genotypes as an underlying genetic cause underpinning a small cohort of hypomorphic albinism patients.

Luke O’Gorman was responsible for the processing and analyses of the patient NGS data. Jay Self was responsible for the recruitment of patients and overseeing phenotyping. Chelsea Norman was responsible for sample library preparation, sequencing and segregation work. This work was supervised by Sarah Ennis, Jane Gibson, Jay Self, Andrew Lotery and Angela Cree.

5.2 Background

Oculocutaneous albinism (OCA) and X-linked ocular albinism (OA) are inherited disorders of melanin biosynthesis which result in varied levels of hypopigmentation of skin, hair, and ocular tissues [25]. Closer examination may reveal foveal hypoplasia (abnormal retinal development determined using optical coherence tomography (OCT)), asymmetry of visual evoked potential (VEP) responses, nystagmus and iris transillumination [25]. Partial phenotypes can occur in which some features are present but others are lacking, however, phenotyping methods have varied significantly and the partial phenotype has currently not yet been described in detail [338, 339, 340]. For more descriptive details of OCA and OA, refer to Chapter 1.2.3. Current management

of albinism focuses on correction of any refractive errors, management of head postures and on the importance of effective sun protection. Albinism is also managed with genetic counselling which highlights the importance of resolving genetic aetiology for patients.

Six genes are known to cause forms of OCA and OA: *TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5*, and *C10orf11* which account for OCA subtypes 1-4 and 6-7 respectively whilst *GPR143* accounts for OA1 (Table 5.1) [341, 39, 340]. OCA5 has been previously found in one Pakistani family and has been mapped to a locus on chromosome 4q24. No gene has been identified for OCA5 [342]. For further descriptions of OCA and OA genes and phenotypes, see Table 1.1. Patients with dominant mutations in the *PAX6* gene can have overlapping phenotypes with partial OCA including foveal hypoplasia, iris transillumination and nystagmus [343].

Table 5.1: *OCA and OA genes targeted to identify genetic causality in eighteen partial albinism patients.*

** Locations are based on RefSeq and hg19 Human reference genome. No gene is known for the OCA5 albinism subtype.*

HGNC symbol	Location *	HGNC name	Albinism subtype	Inheritance
<i>TYR</i>	chr11:88,911,040-89,028,927	Tyrosinase	OCA1A	Autosomal recessive
			OCA1B	
<i>OCA2</i>	chr15:28000023-28344458	OCA2 melanosomal transmembrane protein	OCA2	Autosomal recessive
<i>TYRP1</i>	chr9:12,693,386-12,710,266	Tyrosinase related protein 1	OCA3	Autosomal recessive
<i>SLC45A2</i>	chr5:33,944,721-33,984,780	Solute carrier family 45 member 2	OCA4	Autosomal recessive
-	chr4:101,100,001-107,700,000	-	OCA5	Autosomal recessive
<i>SLC24A5</i>	chr15:48,413,169-48,434,589	Solute carrier family 24 member 5	OCA6	Autosomal recessive
<i>C10orf11</i>	chr10:77,542,519-78,317,126	Chromosome 10 open reading frame 11	OCA7	Autosomal recessive
<i>GPR143</i>	chrX:9,693,453-9,734,005	G protein-coupled receptor 143	OA1	X-linked recessive

As discussed in Chapter 1.2.3, OCA1A is the most severe form of OCA and can be recognised in early infancy by white hair from birth [26]. As many as 30% of OCA1A patients have an unknown genetic aetiology [342, 344] and this percentage may be higher for cases of partial albinism [345].

Two *TYR* variants, rs1126809 (R402Q) and rs1042602 (S192Y) are common in European populations within 1000 Genomes Project reporting minor allele frequencies of 25% and 37% respectively [149, 150]. Functional studies have shown the R402Q causes a 75% reduction in catalytic activity compared to the wild-type [346, 347, 348]; and S192Y results in a 40% reduction of tyrosinase enzymatic activity [349].

R402Q has been proposed as a causal variant when inherited in *trans* with a deleterious *TYR* mutation [332, 333]. However, no OCA phenotype has been observed in the parents of affected individuals even though they carried a combination of a deleterious mutation and R402Q [331]. Two common variants on one allele may produce a further reduction in TYR activity that, when co-inherited with a deleterious *TYR* mutation, causes further loss of activity and an albino phenotype (Figure 5.1) [346, 350]. The term ‘tri-allelic’ has been previously used to describe a similar hypothesis for Bardet-Biedl syndrome by which an affected individual carries three mutant alleles and an unaffected individual can carry two mutant alleles [351]. However, this is yet to be demonstrated in albinism.

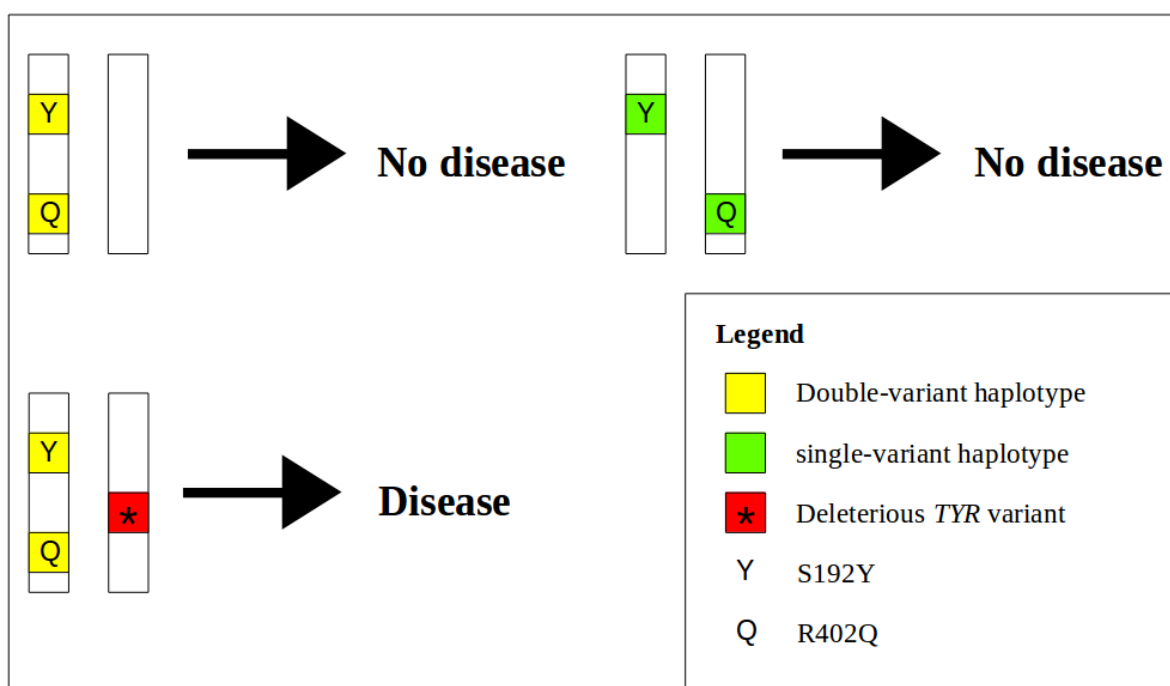


Figure 5.1: *Depiction of ‘tri-allelic’ causal genotype.* The disease causing tri-allelic genotype is shown as two common variants, S192Y and R402Q, as a double-variant haplotype in trans with a deleterious *TYR* variant.

Next-generation sequencing (NGS) offers the opportunity to sequence patient DNA and potentially identify causal variant(s) [158]. The exome accounts for approximately 1% of the genome [316] and ~85% of disease causing mutations are found in the exome [159, 155]. Pre-designed panels such as the Illumina TruSight One (Illumina 5200 Illumina Way San Diego, California USA) targets the exons of 4811 genes associated with known clinical phenotypes. This targeting approach offers a cost-effective alternative to standard whole exome sequencing (WES) [352]. The Illumina TruSight One capture kit uses an enrichment method in library preparation which is described in more detail in chapter 1.4.3.8.

5.3 Aim

In this study, we performed next-generation sequencing on all known albinism genes in patients with possible hypomorphic albinism phenotypes, identified through detailed ocular phenotyping in a tertiary eye clinic. The accountability of a tri-allelic genotype involving common variants is investigated for the first time using detailed phenotyping, NGS analysis and segregation studies to identify the causative genotype.

5.4 Methods

5.4.1 Patients

Patients were recruited following the tenets of the declaration of Helsinki, informed consent was obtained and the research was approved by the Southampton and South West Hampshire Research Ethics Committee. Eighteen patients were identified as having hypomorphic albinism from a regional paediatric nystagmus clinic. The patients had at least two phenotypic features of albinism (low skin and hair pigmentation/ nystagmus/ foveal hypoplasia/ VEP crossing/ iris transillumination) as determined by the referring clinician. Saliva was collected and DNA extracted using Oragene-DNA kit (OG-575)(DNA Genotek).

5.4.2 Sequencing and bioinformatic pipeline

Illumina TruSight One (Illumina 5200 Illumina Way San Diego, California USA) prepared libraries were sequenced with Illumina NextSeq 500 paired-end sequencing. The bioinformatic pipeline which was used for processing the NGS data initially involved alignment against the Human reference genome (hg19) using Novoalign (v2.08.02).

Variant calling was performed using SAMtools v0.1.19 [143]. Variants were excluded if they had a Phred quality score of <20 or a read depth of <4 . Annotation of variants was performed using ANNOVAR [18] against RefSeq transcripts. ANNOVAR annotation also included annotation with dbSNP v135, 1000 Genomes Project [149, 150] minor allele frequencies and conservation-based pathogenicity scores of SIFT [153], PolyPhen2 HumVar [353] and GERP++ [354]. SIFT predicts pathogenicity of missense mutations based on homology [153], PolyPhen2 HumVar predicts pathogenicity based on conservation and protein structure or function [353] and GERP++ measures evolutionary constraint [354]. Additional annotation was applied using a bespoke script to incorporate the Human Gene Mutation Database, HGMD [172].

5.4.3 Quality control of NGS data

Quality control was performed on the aligned data to assess coverage. Coverage was calculated using BEDtools V2.17.0 [277] and the Illumina TruSight One BED file for OCA and OA genes (listed in Table 5.1) and *PAX6*. Samples were also analysed to confirm sex concordance by assessing X chromosome heterozygosity. Higher heterozygosity percentages on the X chromosome are typically indicative of female gender. A male patient would be expected to elicit less X chromosome heterozygosity due to the hemizygous status of the X chromosome. Variant sharing between the NGS samples was checked for consistency with sample known relationships and ethnicities. Verify-BamID [191] software was used to determine the degree of non-reference bases and excessive heterozygosity observed across reference sites and so provide an estimate of contamination. The resultant ‘freemix’ value was examined and a threshold value of >0.02 was applied to indicate contamination as in previous exome studies [192]. Sample provenance was confirmed by application of a validated SNP tracking panel developed

specifically for exome data [188].

5.4.4 Identification of candidate causal variants

Annotated coding region variants were systematically filtered to prioritise variants in *TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5*, *C10orf11*, *GPR143* and *PAX6* genes. Synonymous variants were excluded. Variants were filtered for inclusion using 1000 Genomes Project minor allele frequency <0.05 . Variants were further filtered by excluding variants which had benign pathogenicity scores (SIFT >0.05 , PolyPhen2 Hum-Var <0.43 and GERP++ <2). Sanger sequencing was used to confirm and segregate each *TYR* variant in patients and available family members.

5.5 Results

5.5.1 Quality control

Coverage across OCA and OA genes at 10X, 20X, 50X and 100X depth for the eighteen patients was assessed (Table 5.2). For all samples, more than 77% of the target region had a read depth $\geq 20X$ depth. All OCA and OA genes had a median coverage of at least 85.69% at 20X depth and 78.06% at 50X depth. X chromosome heterozygosity was assessed for the eighteen samples and in all cases recorded and determined gender matched (Table 5.3). NG178 and NG270 had VerifyBamID freemix values marginally over the 0.02 threshold. No substantial levels of heterozygosity were identified in these samples, which suggested that there was no significant sample contamination. The mean percentage of shared variants between samples for samples was 42.3%. Sample NG213 had a mean percentage of shared variants between other patients in the cohort of 36.42%. Follow-up of NG213 patient records identified a mixed race ethnic background

of the patient.

Table 5.2: Coverage for eight genes of interest sequenced in eighteen probands with hypomorphic albinism. Median coverage is presented for each gene at 10X, 20X, 50X and 100X depth. Per proband coverage at 20X depth is presented for each gene of interest.

GENE	Median Coverage Across 18 Probands				Coverage at 20X																	
	10X	20X	50X	100X	NG167	NG178	NG213	NG250	NG251	NG257	NG263	NG265	NG270	NG272	NG296	NG309	NG322	NG327	NG333	NG340	NG344	NG356
<i>TYR</i>	0.9971	0.9835	0.9333	0.7994	0.9788	0.9965	0.9864	0.9947	0.9971	0.9912	0.9794	0.9971	0.9971	0.9788	0.9971	0.9941	0.9953	0.9652	0.9805	0.9569	0.9593	0.9817
<i>OCA2</i>	0.9923	0.9923	0.9513	0.6317	0.9923	0.9923	0.9917	0.9923	0.9923	0.9923	0.9923	0.9923	0.9923	0.9923	0.9923	0.9923	0.9923	0.9907	0.9820	0.9867	0.9923	0.9923
<i>TYRP1</i>	0.9960	0.9750	0.8705	0.6053	0.9419	0.9824	0.9398	0.9886	0.9744	0.9852	0.9676	0.9654	0.9773	0.9648	0.9943	0.9807	0.9761	0.9455	0.9665	0.9546	0.9864	0.9546
<i>SLC45A2</i>	0.9949	0.9652	0.8536	0.5105	0.9339	0.9664	0.9550	0.9960	0.9940	0.9556	0.9595	0.9846	0.9635	0.9726	0.9954	0.9681	0.9664	0.9356	0.9595	0.9402	0.9385	0.9493
<i>SLC24A5</i>	0.9947	0.9947	0.9787	0.8017	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947	0.9947
<i>C10orf11</i>	0.8569	0.8569	0.8569	0.4000	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569	0.8569
<i>GPRI43</i>	0.9024	0.8632	0.7806	0.3775	0.8754	0.8006	0.8746	0.8846	0.8597	0.8725	0.8647	0.8946	0.8561	0.7949	0.8796	0.8739	0.8739	0.8419	0.8953	0.8604	0.8006	0.8462
<i>PAX6</i>	0.9929	0.9890	0.9261	0.7147	0.9760	0.9890	0.9877	0.9877	0.9883	0.9916	0.9870	0.9890	0.9903	0.9916	0.9922	0.9883	0.9903	0.9818	0.9916	0.9780	0.9741	0.9831

Table 5.3: Assessment of heterozygosity and the VerifyBamID freemix for eighteen samples. Recorded gender reported by the referring clinician and determined gender with <50% heterozygosity of the X chromosome imputing a male gender and $\geq 50\%$ imputing a female gender.

Patient ID	X Het (%)	Total Het (%)	Determined Gender	Recorded Gender	Freemix
NG167	61.11	62.46	Female	Female	0.00758
NG178	20.56	60.39	Male	Male	0.02285
NG213	69.33	68.90	Female	Female	0.01209
NG250	58.22	61.85	Female	Female	0.00919
NG251	13.33	62.64	Male	Male	0.00982
NG257	5.77	60.71	Male	Male	0.00912
NG263	68.18	61.44	Female	Female	0.00940
NG265	61.78	62.87	Female	Female	0.00974
NG270	11.61	61.28	Male	Male	0.02779
NG272	13.27	62.67	Male	Male	0.01512
NG296	16.04	61.22	Male	Male	0.00904
NG309	13.18	61.31	Male	Male	0.00944
NG322	17.09	62.44	Male	Male	0.00914
NG327	33.87	60.55	Male	Male	0.01439
NG333	58.82	62.88	Female	Female	0.00902
NG340	65.70	63.25	Female	Female	0.01480
NG344	14.66	61.52	Male	Male	0.01253
NG346	10.62	60.11	Male	Male	0.00772
NG356	3.74	62.57	Male	Male	0.01082

5.5.2 Candidate causal variant identification

Filtering of variants was performed (Figure 5.2) on 2699 unique variants and revealed eighteen potentially causal mutations across five genes in thirteen patients. Five remaining patients had no variants passing filtering (Table 5.4). No patient was found to have more than three variants using this methodology.

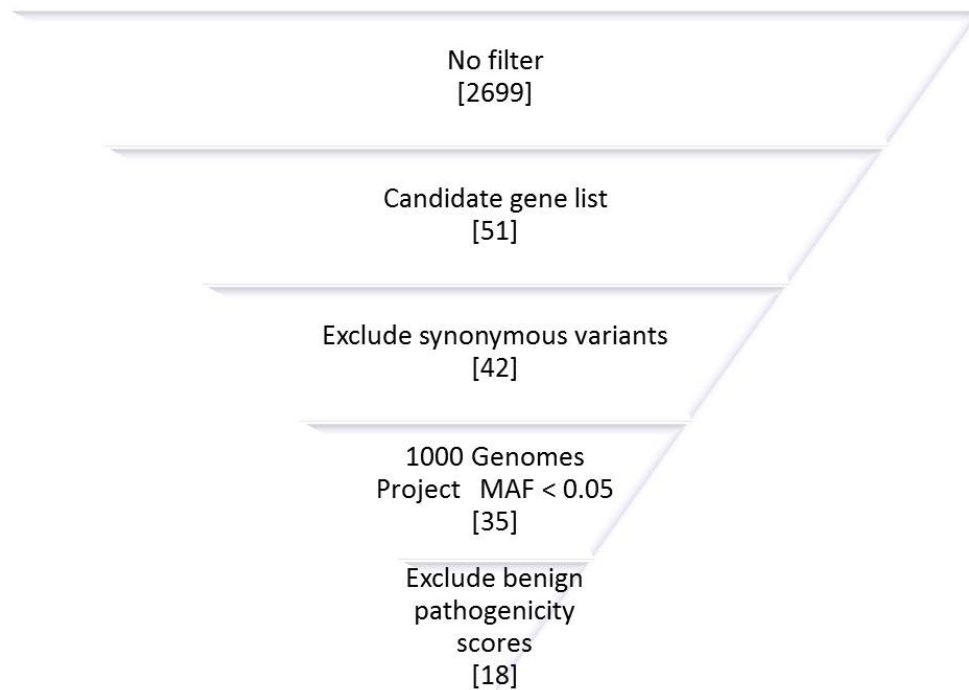


Figure 5.2: ***Filter steps involved in variant prioritisation.*** Filtered were applied to include genes of interest, exclude variants synonymous and exclude variants common within the general population. Variants are further filtered to omit benign scoring variants ($SIFT > 0.05$, $PolPhen HumVar < 0.453$ and $GERP++ < 2$). The number of remaining variants following each successive filter is shown in square brackets.

Table 5.4: *Predicted causal variants, in eighteen probands with phenotypes matching hypomorphic albinism. The prediction scores for nonsynonymous variants are included, for some mutations a prediction score was not available at the time of analysis. Colours highlighting mutations are as follows: TYR gene in cyan, OCA2 gene in dark blue, GPR143 gene in green, PAX6 gene in pink and TYRP1 gene in orange.*

ID	Proband	Variant 1	Variant 2	Variant 3
NG167	1	-	-	-
NG178	2	-	-	-
NG213	3	TYR c.G529T p.V177F (SIFT= . PolyPhen=D GERP=5.16)	OCA2 c.G822C p.W274C (SIFT=0 PolyPhen=D GERP=4.66)	OCA2 c.C1948G p.L650V (SIFT=0.03 PolyPhen=D GERP=5.75)
NG250	4	TYR c.1467dupT p.T489fs	-	-
NG251	5	TYR c.505_507del p.169del	-	-
NG257	6	TYR c.732_733del p.C244*	-	-
NG263	7	TYR c.C1204T p.R402*	-	-
NG265	8	OCA2 c.A1393G p.N465D (SIFT=0.01 PolyPhen=D GERP=5.33)	TYRP1 c.C1037G p.P346R (SIFT=0 PolyPhen=D GERP=5.73)	-
NG270	9	TYR c.C1217T p.P406L (SIFT= . PolyPhen=D GERP=4.68)	PAX6 c.C1264A p.Q422K (SIFT=0 PolyPhen=D GERP=6.16)	-
NG272	10	GPR143 c.485delG p.W162fs	-	-
NG296	11	-	-	-
NG309	12	TYR c.C1217T p.P406L (SIFT= . PolyPhen=D GERP=4.68)	-	-
NG322	13	OCA2 c.C1606T p.R536C (SIFT=0.01 PolyPhen=D GERP=5.8)	-	-
NG327	14	-	-	-
NG333	15	-	-	-
NG340	16	OCA2 c.G1255A p.V419 (SIFT=0.02 PolyPhen=D GERP=5.2)	OCA2 c.A1025G p.Y342C (SIFT=0 PolyPhen=D GERP=5.55)	-
NG344	17	OCA2 c.G1255A p.V419 (SIFT=0.02 PolyPhen=D GERP=5.2)	-	-
NG356	18	TYR c.C1264T p.R422W (SIFT= . PolyPhen=D GERP=2.69)	-	-

Seven patients had candidate causal variants not in the *TYR* gene. Patient NG213 and NG340 each have two compound heterozygous variants in the *OCA2* gene. The two *OCA2* variants identified for NG213 are novel. Patient NG270 was found to have a possible pathogenic variant in the *PAX6* gene and patient NG272 has a deletion resulting in a frameshift mutation in the X-linked gene, *GPR143*. Patient NG265 has a single mutation in *OCA2* and a second mutation in *TYRP1*. Two patients, NG322 and NG344, each have a single heterozygous mutation in the *OCA2* gene with no second mutation identified. These variants would require follow-up investigations and validation before concluding causality.

Six patients (NG250, NG251, NG257, NG263, NG309 and NG356) each had a single heterozygous mutation in the *TYR* gene with no further variants passing the filtering threshold. All six patients also had R402Q and S192Y common *TYR* variants as a tri-allelic genotype. Patients NG270 and NG213 had *TYR* mutations but are likely coincidental carriers. The *TYR* mutation P406L occurs in two patients, as does the

OCA2 mutation V419I.

5.5.3 Segregation analysis

Further investigation of the patients with tri-allelic genotypes and available family members was performed in a segregation analysis. Figure 5.3 depicts the tri-allelic genotype determined for the six patients: NG250, NG251, NG257, NG263, NG309 and NG356. Genotypes were determined for available family members of the patients by Sanger sequencing for the two common *TYR* variants, R402Q (Q) and S192Y (Y), and a deleterious variant (D). In total, twenty patients and family members were phenotyped and genotyped which suggested a total of nine cases of partial albinism (six patients and three affected family members).

Sanger sequencing confirmed the genotypes of the variants in all six affected patients. Observing affected family members of the patients revealed segregation of the tri-allelic genotype in all but one of the families. The mother of patient NG356 (family 18) has both nystagmus and foveal hypoplasia, yet does not have the same deleterious *TYR* mutation as her son. Both parents for patient NG257 have the S192Y and R402Q *TYR* variants so phase of the three variants cannot be deduced for the patient. DNA is unavailable for patient NG356, therefore, a genotype for the three variants was not able to be determined. Without the father's genotype known, phase was not able to be deduced for this patient. It can be deduced that R402Q is on the *trans* allele to the deleterious *TYR* variant in patients NG250, NG251, NG263 and NG309.

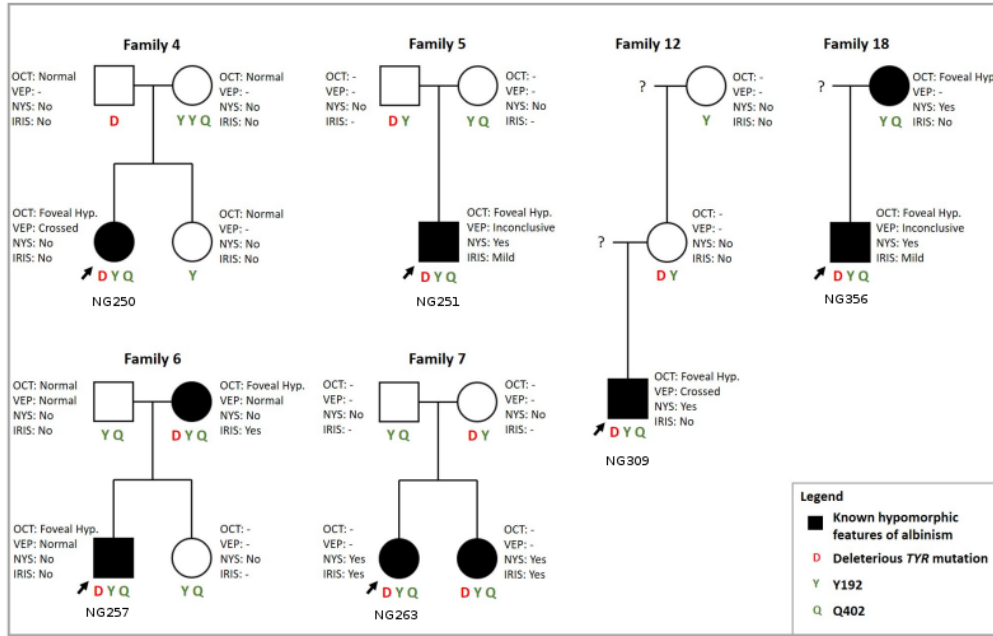


Figure 5.3: *Segregation of the tri-allelic genotype is depicted with family pedigrees.* Six families are shown for the six probands with the tri-allelic genotype in *TYR*: NG250 (family 4), NG251 (family 5), NG257 (family 6), NG263 (family 7), NG309 (family 12) and NG356 (family 18). Family members are described for phenotyping results for foveal hypoplasia (OCT), asymmetry of visual evoked potential (VEP) responses, nystagmus (NYS) and iris transillumination (IRIS). The variants are coloured red for the deleterious *TYR* variant (D) and dark green for the *TYR* common polymorphisms S192Y (Y) and R402Q (Q).

5.6 Discussion

Here detailed phenotyping has been utilised to select a cohort of partial albinism patients and perform next-generation sequencing (NGS) with the Illumina TruSight One pre-designed panel in a cohort of hypomorphic albinism patients. This has given the opportunity to detect genetic variants, assign genotype-phenotype relationships and resolve genetic aetiology for the patients. This study has focused on the causality of the tri-allelic *TYR* genotype to resolve the genetic aetiology of hypomorphic albinism patients.

We identify one novel variant in the *PAX6* gene, one novel frameshift variant in the *GPR143* gene, two novel variants in the *OCA2* gene (both in patient NG213), five previously reported variants in *OCA2* and one previously reported variant in the *TYRP1* gene. These variants will be followed-up with Sanger sequencing to confirm presence in the patient and segregation studies in available family members.

Six patients were identified to have a single heterozygous possible pathogenic variant in the *TYR* gene and no other variants following the application of filter parameters described in the methods (Chapter 5.4). The mutation V177F has been previously reported in an albinism cohort [355]. *TYR* p.T489fs results in a frameshift and has been reported as a causal mutation multiple times [26, 342, 333, 355]. *TYR* p.168del does not cause a frameshift but has been previously reported as a causal mutation [356]. R402* has been reported previously and creates a premature stop codon, considered highly deleterious [333, 355, 357]. R422W has been reported as disease causing [26]. P406L has also been reported many times before in association with albinism [26, 333, 355] and has been shown to reduce enzyme activity to 35% [358]. There is currently no functional evidence of the deleterious effect of the mutations *TYR* p.169del and *TYR* C244* though deletions have been previously been reported as causal and stop-gains are considered highly damaging [356].

Segregation analyses of available family members revealed that unaffected family members also harboured the same deleterious *TYR* variant as the proband. The six patients were found to have two missense *TYR* variants. The two missense *TYR* variants, R402Q and S192Y, have a 1000 Genomes Project minor allele frequency of 25% and 37% in the European populations respectively [149, 150]. As individual SNPs they are considered benign. The common variant R402Q is located in exon 4, near to the CuB catalytic site, and produces a thermolabile enzyme [346, 347] but it has been argued that the reduc-

tion of tyrosinase activity is not enough to produce a phenotype. Segregation of R402Q with a known deleterious variant in *cis* does not confer albinism [331]. The variant S192Y is located in the CuA catalytic site and has been shown to lower enzymatic activity independently to R402Q [346]. There is potential for a double-variant haplotype, p.[S192Y;R402Q], existing on the *trans* allele to the known deleterious *TYR* mutation in affected individuals [332, 333]. The predicted frequency of p.[S192Y;R402Q] in *cis* is 1.1%, using British participants of the 1000 Genomes project (GBR) and the webserver <http://analysistools.nci.nih.gov/LDlink/> [359].

Next-generation sequencing with a broad capture such as the TruSight One clinical exome has allowed genes which overlap phenotypes with OCA and OA to be analysed such as *PAX6*. However, diagnosis of albinism currently focusses on compound mutations in single genes without considering the potential for synergistic relationships between functionally related genes such as that previously suggested for OCA2 and OCA3 genes (*OCA2* and *TYRP1*) [360] and for which there is potentially one example in our cohort. There are also suggestions that missing heritability in hypomorphic albinism could be due to mutations in the *TYR* promoter or an interacting distal gene enhancer which is not targeted with the Illumina TruSight One capture kit [361].

Previously segregation studies would have not been able to verify causality of the deleterious *TYR* variant with unaffected family members having the variant. Missing heritability for partial albinism has been shown to be accounted for by the presence of tri-allelic variants in the *TYR* gene. Further functional analyses of the tri-allelic variants could confirm causality by a double-variant haplotype of common *TYR* variants in *trans* with a deleterious *TYR* variant. This will allow the tri-allelic genotype to be considered for future and retrospective diagnoses of OCA1.

6 The utility of a sequencing panel for clinical use in patients with nystagmus and albinism

6.1 Synopsis

This chapter focuses on an analysis of samples from a cohort of nystagmus and/or albinism patients have undergone basic phenotyping. Variants called across the cohort are restricted to a clinically available 31 gene panel for nystagmus and albinism. Likely causal variants are prioritised using methods which are currently employed by clinical diagnostic laboratories. The diagnostic yield of is reported and the clinical utility in both complete and incomplete phenotyped patients is evaluated.

Luke O’Gorman was responsible for the transfer and demultiplexing of sequence data to the university servers, processing and analyses of data which were supervised by Jay Self, Sarah Ennis, Jane Gibson, Andrew Lotery and Angela Cree. Jay Self was responsible for patient recruitment and identifying patients for sequencing. DNA extraction was performed by staff at the University Hospital Southampton Eye Clinic. Chelsea Norman was responsible for the library preparation of the DNA samples, Sanger sequencing primer design and for the sequencing of all samples.

6.2 Background

6.2.1 Infantile nystagmus syndrome and albinism

Infantile nystagmus syndrome (INS) is a classification of nystagmus in which the condition has an infantile onset, accelerating slow phases of ocular motion and is commonly presented with horizontal-torsional waveforms [23]. INS can be idiopathic (IINS) or

as part of a plethora of ocular or systemic disorders including albinism, retinal disease and neurological disorders. Ocular Albinism (OA) is a form of albinism in which the clinical features are constrained to the eye, whilst Oculocutaneous Albinism (OCA) encompasses a broader phenotypic range, affecting the eyes, hair, skin and in some cases other organ systems [25]. See Chapter 1.2 for more detail regarding nystagmus and albinism.

6.2.2 Genetic basis

Genes involved in the melanin biosynthesis pathway are known to cause forms of both OA and OCA. Examples include *GPR143*, which is causal for OA1 [340], whilst *TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5* and *C10orf11* are associated with OCA subtypes 1-4 and 6-7 respectively [25]. The OCA1 gene, *TYR*, is known to be associated with missing heritability [331]. Our group and others have previously reported a compound heterozygous tri-allelic genotype in *TYR* which involves both rare (AF<5%) and common (AF=28-36%) functionally damaging variants which are likely to be on trans alleles [332, 333, 362]. The two common *TYR* variants p.S192Y and p.R402Q, have previously been shown to cause a 40% reduction in tyrosinase activity and protein misfolding, respectively [363, 364]. However, it has not been conclusively proven that the rare damaging mutation is in trans with the two common variants. Although OCA is considered an autosomal recessive disorder and OA is an X-linked recessive disorder, an individual's overall pigmentary phenotype is known to include a complex interplay between many genes [25, 365, 331]. Additionally, many of the ocular features seen in albinism can be seen in other disorders caused by mutations in genes that are not associated with melanin biosynthesis such as *PAX6* mutations, which can cause a variety of ocular phenotypes including nystagmus and foveal hypoplasia [343].

Similarly, Chediak-Higashi syndrome and Hermansky-Pudlak syndrome, caused by the *LYST* and *HPS* genes respectively, involve many OA and OCA phenotypic traits. Despite the significant systemic health implications of these forms of syndromic albinism, most patients never undergo genetic testing.

6.2.3 Phenotype of INS and albinism

A deficiency of pigmentation in the hair and skin are more conspicuous during phenotyping, however, albinism encompasses a range of ocular traits including INS [25], foveal hypoplasia, abnormal crossing pattern at the optic chiasm and iris transillumination defects [366] which require further examination. Horizontal nystagmus is more common in albinism patients whilst vertical nystagmus is more commonly observed in other neurological syndromes such as ataxia [367, 368].

For the clinician, when presented with a child with nystagmus, typically at the age of 4-6 months, the range of potential diagnoses is very broad [369]. The range of investigations for such patients is similarly broad and typically focussed on excluding the most urgent conditions (such as neoplasms, metabolic disease or blindness). In many cases, once an immediately life-threatening disorder and blindness have been excluded, investigation of molecular aetiology is reduced or ceases. Our group and others have previously shown that a combination of detailed phenotyping and bespoke genetic testing can identify patients that require immediate management (e.g. MRI scans under anaesthesia) from those that do not [335, 370].

Our group and others have previously shown that a combination of detailed phenotyping and genetic testing can identify those cases requiring immediate management

from those where more invasive tests can be avoided (such as MRI scans under anaesthesia) and yield clinically relevant molecular diagnoses leading directly to treatments or other alterations to medical management [271, 172]. As phenotyping becomes more precise and nuanced in children with nystagmus, it is possible that candidate gene lists can become more specific [270]. For example, an abnormal electroretinogram (ERG) can be the only indication that an underlying retinal dystrophy is the cause of the nystagmus. Hence, a retinal gene panel might be the most appropriate genetic testing option. It also allows potential candidate causal variant(s) to be interpreted with greater confidence. Previous studies of next-generation sequencing in INS patients have utilised large gene panels of up to 300 genes whilst identifying candidate causal variants in a recurrent, small subset of genes [371, 372, 373], or in genes for conditions which would have been identified by more detailed phenotyping. This suggests that pre-selecting and interpreting variants in fewer genes in highly phenotyped patients, may provide the most efficient workflow and highest diagnostic yield in routine clinical practice.

6.2.4 Diagnostic gene panels

The UK National Health Service (NHS) presently utilises the UK Genetic Testing Network (UKGTN) as an advisory organisation for clinical genetic testing services. Some clinical laboratories offer UKGTN approved genetic testing for albinism and nystagmus, but the number of genes included can vary widely. A UKGTN gene panel for ‘albinism and nystagmus’ consists of 31 selected genes, including all 14 genes currently approved in Genomics England’s PanelApp for ‘nystagmus’ and ‘ocular and oculocutaneous albinism’ (date accessed 29/05/2018).

The TruSight One capture kit (Illumina 5200 Illumina Way San Diego, California USA) targets the exome of 4811 genes which are relevant to disease. Following the sequencing of sample DNA, researchers are then able to select a candidate gene list amongst the targeted genes in order to prioritise candidate causal variants.

6.2.5 Variant interpretation

The American College of Medical Genetics (ACMG) guidelines stipulate universal guidelines for the interpretation of sequence variants [172] that form a stringent set of criteria, which minimises the likelihood of a variant being reported as causative without sufficient evidence. However, ACMG guidelines are not recommended for the reporting of complex/polygenic genetic causes and are not mandatory for clinical laboratories to follow when reporting causal variants. Identifying causal variants in albinism therefore remains a problem for diagnostic laboratories as the vital role of common pathogenic variants may be missed by stringent ACMG guidelines.

6.3 Aim

The primary aim of this chapter is to determine the underlying genetic cause for nystagmus and albinism using a stringent gene panel in patients which have been phenotyped using equipment available to most paediatric ophthalmology services. With this, I then aim to provide valuable information to clinicians for shaping diagnostic pathways for INS.

6.4 Methods

6.4.1 Patients

Patients were recruited following the tenets of the declaration of Helsinki, informed consent was obtained and the research was approved by the Southampton & South West Hampshire Research Ethics Committee. All patients were referred to a single service (Southampton Nystagmus Clinic) were approached for recruitment. A total of 634 individuals (and their family members) with various ophthalmic diseases (age range 0-18 yrs) were recruited and their demographic and phenotypic data were recorded in an electronic database. 131 patients (and their family members) with nystagmus or albinism were selected for sequencing. Patients with a diagnosis of a condition known to cause nystagmus (such as Down syndrome) but without an INS phenotype (such as Gaze Evoked Nystagmus, GEN due to cerebellar disease) and those who were born before 35/40 weeks gestation were excluded. Saliva samples (ORAGENE) were collected and DNA extracted using Oragene-DNA kit (OG-575) according to the manufacturer's instructions (DNA Genotek).

6.4.2 Phenotyping

All patients had detailed phenotyping as outlined in Chapter 5. Skin and hair were phenotyped relative to family pigmentation. Phenotyping was also performed by orthoptic examination to determine the degree at which the eye turns from various positions. Anterior and posterior segment examinations with a slit-lamp biomicroscope were performed to direct light through the pupil which was reflected off the retina. From this, iris transillumination can be determined if substantial iris thinning had occurred which is typical in albinism. Electroretinogram (ERG) to measure the electrical activity gen-

erated by neural and non-neuronal cells in the retina in response to a light stimulus with a ‘normal’ response expected for IINS and albinism patients. In healthy individuals, optic nerve fibers from each eye cross at the optic chiasm to the opposite side which is important for stereopsis (binocular vision). Albinism is associated with excessive fibre crossing at the optic chiasm a loss of stereopsis. This phenotypic trait was measured by Visual Evoked Potential (VEP) to measure brain activity under stimulus of light. Optical coherence tomography (OCT) was performed with the Leica OCT system or Spectralis OCT (Heidelberg Engineering) and was used to visualise the absence of the foveal pit (foveal hypoplasia). When studied, eye movements were recorded with the EYELink10000 + (SR research) eye tracker to determine the oscillation waveforms of nystagmus.

6.4.3 Grouping by phenotype criteria

Probands were allocated into four phenotype subgroups; clinically IINS with complete phenotyping (group 1), clinically IINS with incomplete phenotyping (group 2), clinical phenotyping consistent with albinism with complete phenotyping (group 3), and clinical features suggestive of albinism with incomplete phenotyping (group 4) (Table 6.1).

Table 6.1: *Selection criteria of the four phenotype groups.* *Must include at least one of the following phenotypes in bold: OCT=‘FH’, VEP-misrouting suggested=‘Yes’ or Iris transillumination=‘Yes’.

Cohort sub-group number	Cohort sub-group	Predominant waveform direction	ERG	OCT	VEP-misrouting suggested	Iris trans illumination
1	Clinically idiopathic nystagmus (with full phenotyping)	Horizontal	Normal	Normal	No	No
2	Clinically idiopathic nystagmus (with incomplete phenotyping)	Horizontal or Equivocal	Normal or Untested or Equivocal	Normal or Untested or Equivocal	Normal or Untested or Equivocal	Normal or Untested or Equivocal
3	Clinically consistent with albinism (with full phenotyping)	Horizontal or Multiplanar	Normal	Foveal hypoplasia	Yes	Yes
4*	Clinical features suggestive of albinism (with incomplete phenotype or missing phenotype data)	Horizontal or Multiplanar	Normal or Untested or Equivocal	FH or Untested or Equivocal or Normal	Yes or Untested or Equivocal or Normal	Yes or Untested or Equivocal or Normal

6.4.4 Library preparation and NGS

A total of 131 samples and three erroneously duplicated samples were prepared for sequencing (Figure 6.1) using the TruSight One capture kit (Illumina 5200 Illumina Way San Diego, California USA) which was completed in six batches. The TruSight One covers 4811 genes associated with disease-causing mutations. Next generation sequencing was performed using the Illumina the NextSeq 500 (mid-output) which was recommended as the platform for TruSight One capture. The reads were paired end and performed in 150 cycles for each read (2 x 150).

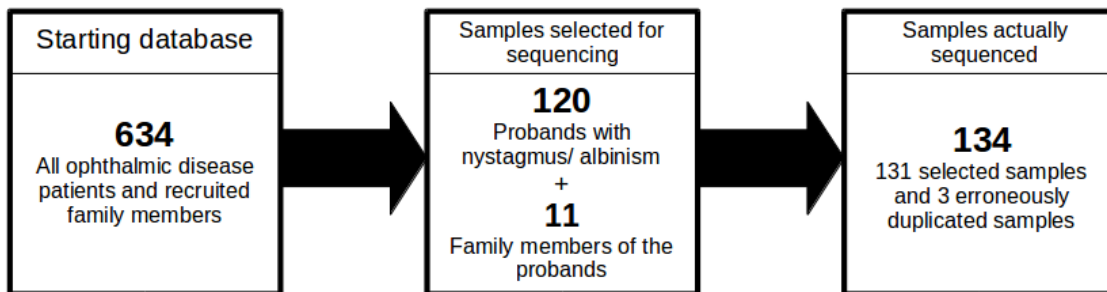


Figure 6.1: *Patient selection for sequencing.* Patients with nystagmus/ albinism and their family members were selected for sequencing. However, there were three known erroneously duplicated samples run resulting in a total of 134 samples sequenced.

6.4.5 UKGTN gene panel

The UK Genetic Testing Network (UKGTN) is an organisation which offers a molecular genetics service for patients and families who require genetic advice or diagnosis. This network allows patients and their families to seek the appropriate tests via local genetics centres. One of the roles it has is the evaluation of new tests for scientific validity and clinical utility that member laboratories wish to provide to NHS patients on a national level. The UKGTN also provide gene panels for a disease of interest and will suggest the laboratories which can carry out a genetic test for the given gene panel. The UKGTN ‘albinism and nystagmus’ gene panel consists of 31 genes (Table 6.2) and was developed by University Hospital Southampton Eye Clinic, University of Southampton and Wessex Regional Genetics Laboratory (WRGL).

Table 6.2: *HGNC approved gene names for genes listed in the UKGTN gene panel for ‘Nystagnus and albinism’ with the associated OMIM inheritance pattern and phenotype.*

Symbol (HGNC)	Loci (HGNC)	Assumed inheritance pattern (OMIM)	Phenotype (OMIM)
<i>AP3B1</i>	5q14.1	AR	Hermansky-Pudlak syndrome 2
<i>BLOC1S3</i>	19q13.32	AR	Hermansky-Pudlak syndrome 8
<i>BLOC1S6</i>	15q21.1	AR	Hermansky-Pudlak syndrome 9
<i>C10orf11</i>	10q22.2-q22.3	AR	Albinism, oculocutaneous, type VII
<i>CACNA1A</i>	19p13.13	AD	Episodic ataxia, type 2
<i>CACNA1F</i>	Xp11.23	XL	Aland Island eye disease
<i>CASK</i>	Xp11.4	XLD	FG syndrome 4
<i>DTNBP1</i>	6p22.3	AR	Hermansky-Pudlak syndrome 7
<i>FRMD7</i>	Xq26.2	XL	Nystagmus 1, congenital, X-linked
<i>GPR143</i>	Xp22.2	XL	Ocular albinism, type I
<i>HPS1</i>	10q24.2	AR	Hermansky-Pudlak syndrome 1
<i>HPS3</i>	3q24	AR	Hermansky-Pudlak syndrome 3
<i>HPS4</i>	22q12.1	AR	Hermansky-Pudlak syndrome 4
<i>HPS5</i>	11p15.1	AR	Hermansky-Pudlak syndrome 5
<i>HPS6</i>	10q24.32	AR	Hermansky-Pudlak syndrome 6
<i>LYST</i>	1q42.3	AR	Chediak-Higashi syndrome
<i>MANBA</i>	4q24	AR	Mannosidosis, beta
<i>MITF</i>	3p13	AR	Tietz albinism-deafness syndrome
<i>MLPH</i>	2q37.3	AR	Griselli syndrome, type 3
<i>MYO5A</i>	15q21.2	AR	Griselli syndrome, type 1
<i>OCA2</i>	15q12-q13.1	AR	Albinism, oculocutaneous, type II
<i>PAX6</i>	11p13	AD	Foveal hypoplasia 1
<i>RAB27A</i>	15q21.3	AR	Griselli syndrome, type 2
<i>SACS</i>	13q12.12	AR	Spastic ataxia, Charlevoix-Saguenay type
<i>SETX</i>	9q34.13	AR	Spinocerebellar ataxia, autosomal recessive 1
<i>SLC24A5</i>	15q21.1	AR	Albinism, oculocutaneous, type VI
<i>SLC45A2</i>	5p13.2	-	Albinism, oculocutaneous, type IV
<i>TULP1</i>	6p21.31	AR	Leber congenital amaurosis 15
<i>TYR</i>	11q14.3	AR	Albinism, oculocutaneous, type I
<i>TYROBP</i>	19q13.12	AR	Nasu-Hakola disease
<i>TYRP1</i>	9p23	AR	Albinism, oculocutaneous, type III
<i>TYRP1</i>	9p23	AR	Albinism, oculocutaneous, type III

The UKGTN gene panel for ‘albinism and nystagmus’ (n=31) was used to prioritise genes for identification of candidate likely causal mutations. The gene panel includes all 14 approved genes in Genomics England’s PanelApp for ‘nystagmus’ and ‘ocular and oculo-cutaneous albinism’ (date accessed 29/05/2018) and also includes all 17 of the known causative genes for OCA [374]. From the data generated with the TruSight One capture kit, the in silico UKGTN ‘albinism and nystagmus’ gene panel of genes was selected for interrogation of variants.

6.4.6 Bioinformatic pipeline

FastQ data were aligned to the hg38 human reference genome with BWA-MEM [141]. GATK v3.7 [144] was used to call SNPs and short indels in a multisample VCF file. Annotation was performed using ANNOVAR v2015Dec[146] to collate variant consequence, dbSNP v144, variant allele frequency (1000 Genomes Project, Exome Sequencing Project and Exome Aggregation Consortium) and pathogenicity scores with CADD [196] and MaxEntScan [271] for splice site variants. Further annotation was included haploinsufficiency scores with DECIPHER v9.23, InterVar (2018) [375] and Human Gene Mutation Database [172]. Coverage was determined using SAMtools v1.3.1 [143] and BEDtools v2.17.0 [277]. Variants were excluded if they had a read depth < 4.

6.4.7 Quality control

Quality control was performed to assess FastQ file quality, coverage of the aligned data, contamination of samples and shared variation between samples to identify unexpected relatedness. FastQ quality was assessed with the FastQC v0.11.3 software. Coverage was determined using SAMtools v1.3.1 and BEDtools v2.17.0. Variant sharing between the targeted NGS samples was checked for consistency with sample relationships and ethnicities. In cases of sample duplicates the sample with superior horizontal coverage across target regions at 20X was retained. VerifyBamID v1.0 software was used to determine the degree of non-reference bases and excessive heterozygosity observed across reference sites and so provide an estimate of the possibility of contamination. The resultant ‘freemix’ value was examined and a threshold value of 0.02 was applied. A coverage threshold were typically set in clinical laboratories by first identifying a sensitivity cut-off for calling variants [376]. However, in cases where a ‘gold standard’ version of a sample was not available for direct comparison, 20X depth was used across gene

panel of 31 albinism and nystagmus genes. Consistent with typical clinical diagnostic reporting, a horizontal coverage threshold of 90% was selected to determine whether or not the gene panel of 31 albinism and nystagmus genes were covered sufficiently.

6.4.8 Variant prioritisation

To create a non-biased analysis, the phenotypes of the patients were not considered when reporting likely causal variants. Likewise, the reported likely causal variants were not given to the clinician prior to phenotyping. This would prevent the clinician's phenotyping or the genetic analysis from being influenced by other factors. Variants were prioritised into two categories of 'assumed pathogenic' and 'assumed likely pathogenic'. 'Assumed pathogenic' was defined as a variant which had a 'pathogenic' annotation in ClinVar, 'pathogenic' annotation by InterVar or 'disease-causing mutation' (DM) in HGMD. 'Assumed likely pathogenic' was defined as a variant which was: (1) not synonymous; (2) had an allele frequency of less than or equal to 5% in 1000 Genomes Project (all populations), Exome Sequencing Project 6500 (all populations) and Exome Aggregation Consortium (all populations) and; (3) had either a CADD Phred ≥ 15 [196] or a MaxEntScan $\geq |3|$ [263]. Variants which form part of a single likely causal genotype were identified as 'likely causal' variants whilst multiple possible causal genotypes were identified as 'reportable likely causal'. Sanger sequencing was performed to verify 'likely causal' variants which were miscalled in $>5\%$ of individuals in the cohort [259, 377, 378].

As discussed previously, the *TYR* tri-allelic causal genotype in OCA involves two common variants and one rare pathogenic variant on trans alleles. As the two common variants were part of a unique molecular pathology, p.R402Q and p.S192Y were analysed separately together with 'assumed pathogenic' or 'assumed likely pathogenic' variants

in any remaining patients which did not have causal diagnostic genotypes previously identified.

6.4.9 Sanger verification

Sanger sequencing was performed to verify ‘assumed pathogenic’ or ‘assumed likely pathogenic’ reported likely causal genotypes which were unsuccessfully called (mis-called) as reference or alternate allele in 5% or more of individuals the cohort (which translates to 4 individuals if the cohort is $n=81$). Primer pairs were designed using the ‘A Plasmid Editor (ApE)’ software and spanned 19bp (forward) and 20bp (reverse) in length (Appendix Table E.1). Primers were checked for ambiguous mapping by performing a genomic BLAST search. Primer binding regions of each sample were examined for variants which could prevent effective annealing of primers to the sample region of interest. The sequence chromatogram and quality data of the sequencing were stored in AP1 files which were viewed in the Sanger ‘Quality Check’ application (version 2.1.2, Thermo Fisher Cloud). Quality thresholds for Sanger sequencing with a base call Phred score of 20 recommended as the threshold for reporting variants [376]. Trace Score quality indicates the average of base call quality values for bases in the clear range (threshold of Phred 20). QV20+ indicates the total number of bases in the entire trace that have basecall quality value ≥ 20 (no threshold set). Signal to Noise shows the average raw signal to noise ratio for all dyes across the trace (software default threshold of 60). Signal strength provides and average relative fluorescent units for all four dyes across the sequence (no threshold set).

6.5 Results

6.5.1 Quality control

6.5.1.1 Sequence quality

There are several aspects to a quality control which should be considered, from the inspection of the sequence data, coverage, contamination and variant concordance between samples. FastQC software was run on the 134 patient samples to inspect the FastQ files and summarise the quality of the sequence data. ‘Per base sequence content’ and ‘Kmer content’ routinely failed FastQC tests and sequence duplication levels were higher across the first 10 bp of reads. These FastQC fails were likely due to a fragmentation bias to transposase restriction enzyme binding sites during library preparation which cause an overrepresentation of the sequence at the start of reads [379]. FastQC results show that batch 5 had a systematic fail of ‘Per base sequence quality’. Degradation of base quality at the tail of reads is typical for Illumina data and FastQC documentation advises that read trimming may be applied for such cases. However, BWA-mem performs soft clipping to take into account the low quality bases during alignment. Due to this feature, BWA-mem developers suggested that low quality tails of reads should not affect SNP calling when using this alignment tool. For exome NGS data which fail ‘sequence base quality’ in FastQC and are to be aligned with BWA-mem, GATK developers state that the removal of low quality bases before alignment should not be performed as these are accounted for when calling variants with GATK Haplotype Caller [144].

6.5.1.2 Coverage, contamination and variant concordance

Following alignment, coverage statistics were calculated for all 134 samples. Coverage at 20X depth for all 134 samples across the 31 gene panel target region was listed in

Table 6.3. The average depth for the cohort was 113X and the average coverage at 20X was 81.8%. VerifyBamID freemix indicates only minor contamination, no sample exceeded the 0.02 freemix threshold for likely contamination.

Table 6.3: *Overall coverage, contamination and file sizes of 134 samples in batches 1-6. The minimum coverage at 20x was less than 90% across the target regions for batches 2, 3, 5 and 6.*

Batch	No. samples	Mean depth	Min depth	Max depth	SD depth	Mean coverage at 20x	Min coverage at 20x	Max coverage at 20x	Mean hets (%)	Min hets (%)	Max hets (%)	Max freemix	Mean FastQ size (GB)	Mean BAM size (GB)	Mean number of coding variants
1	22	166.3	114.2	222.3	30.8	0.980	0.972	0.987	66.1	64.3	72.2	0.002	2.1	6.7	12716
2	20	129.8	1.1	187.4	52.0	0.875	0.000	0.984	63.8	41.1	67.8	0.002	1.3	3.9	11017
3	30	117.1	61.4	214.0	46.3	0.949	0.820	0.988	65.3	63.2	68.6	0.001	1.6	5.7	12024
4	2	110.3	99.0	121.6	16.0	0.957	0.952	0.963	65.2	64.0	66.5	0.001	1.0	3.3	12145
5	14	103.2	44.0	129.5	20.1	0.917	0.605	0.968	62.3	57.4	64.4	0.000	1.4	4.4	10372
6	46	52.0	3.5	129.7	39.7	0.596	0.000	0.972	61.2	48.3	68.1	0.003	1.1	4.6	11260

All samples annotated coding region variants were compared pair-wise to produce variant concordance percentages which could highlight unexpected relationships or ethnicities. Three samples (NG178, NG232 and NG367) had a concordance of genetic variation with their respective duplicate samples of ranging 90.3-95.9% (Appendix Table E.3).

Six family member pairs were identified within the 134 sample cohort (Table 6.4). Five first degree relations were identified which had a minimum concordance of variants of 71% between one another. These were confirmed as true first degree relations by patient records. An additional related pair were identified which was of a lower degree of concordance (minimum of 65% concordant variants) in comparison to the suspected first degree relatives. This related pair of individuals were confirmed as half-brothers by the referring clinician upon inspection of patient notes. There were no known family member relations which were not identified by the analysis.

Table 6.4: *Family member pairs identified from the variant similarity matrix. Analysis was initially conducted with the relationship information blinded. Percentage shared variation is a range of two values as there will be two values for a pair-wise comparison between two samples). Family member pairs are coloured for identification in the original variant similarity matrix (Appendix Table E.3).*

Proband (batch no.)	Family member (batch no.)	Shared variation (%)	Subsequently confirmed relationship
NG178 (1)	NG181 (6)	72-76	Proband-Mother
NG356 (1)	NG357 (3)	71-73	Proband-Mother
NG198 (3)	NG200 (6)	74-75	Proband-Mother
NG412 (6)	NG411 (3)	65-69	Proband-Half Brother
NG558 (6)	NG277 (3)	76-77	Proband-Brother
NG219 (3)	NG218 (2)	72-72	Proband-Father

6.5.1.3 Filtering of the patient cohort

Of the three pairs of sample duplications, the sample with inferior horizontal coverage across the 31 gene panel region regions at 20X was omitted. This resulted in the omission of batch 2 sample ID NG178, batch 3 sample ID NG232 and batch 3 sample ID NG367.

By setting a quality threshold of 90% coverage at 20X depth across the UKGTN ‘nystagmus & albinism’ gene panel, a further 33 samples were removed resulting in n=98 remaining samples. The coverage of the 36 omitted samples are detailed in Table 6.5. It can be seen that for 12 of the omitted batch 6 samples, the % reads mapped to the TruSight One target region is 2.6% or less. This poor mapped to target region was confirmed in IGV as reads appear to be aligning non-specifically to the hg38 reference genome (Appendix Figure E.1). This was later confirmed by Illumina to be due to an ‘incorrectly manufactured reagent causing interference with the capture-based probe application’ (with ‘no compromise to the index sequence’) causing a ‘high % of off-target reads in the alignment data’. The affected index sequence was identified by Illumina to be E505 and the allocated samples for this index sequence matched the 12

samples with $\geq 2.6\%$ reads mapped to target region (Appendix Table E.4). Following these omissions, the minimum coverage at 20X depth across the 31 gene panel was increased from 0.00% to 90.4%.

Table 6.5: *Coverage at 20x depth across the 31 gene panel for the three sample duplicates and 33 samples of coverage less than 90%. Two individuals (from respective family member pairs) of the six family member pairs were omitted due to low coverage (bold).*

Row no.	Batch	Sample ID	Coverage 20x	Omission reason
1	2	NG176	0.0000	Coverage
2	2	NG178	0.9733	Duplicate
3	2	NG302	0.0037	Coverage
4	3	NG219	0.8204	Coverage
5	3	NG232	0.9545	Duplicate
6	3	NG367	0.9472	Duplicate
7	3	NG277	0.8888	Coverage
8	3	NG283	0.8943	Coverage
9	5	NG491	0.6053	Coverage
10	6	NG149-1	0.0138	Coverage
11	6	NG150-2	0.0183	Coverage
12	6	NG210	0.0004	Coverage
13	6	NG236	0.0000	Coverage
14	6	NG267	0.0010	Coverage
15	6	NG325	0.0214	Coverage
16	6	NG412	0.8233	Coverage
17	6	NG504	0.8194	Coverage
18	6	NG556	0.7878	Coverage
19	6	NG561	0.7738	Coverage
20	6	NG563	0.6674	Coverage
21	6	NG566	0.3984	Coverage
22	6	NG571	0.5319	Coverage
23	6	NG574	0.6896	Coverage
24	6	NG575	0.4368	Coverage
25	6	NG580	0.6328	Coverage
26	6	NG582	0.7018	Coverage
27	6	NG584	0.4343	Coverage
28	6	NG587	0.8264	Coverage
29	6	NG592	0.8769	Coverage
30	6	NG594	0.0060	Coverage
31	6	NG597	0.0002	Coverage
32	6	NG598	0.0001	Coverage
33	6	NG602	0.0011	Coverage
34	6	NG607	0.0045	Coverage
35	6	NG610	0.0012	Coverage
36	6	NG613	0.0133	Coverage

A final step of sample omissions was performed at the point of assigning phenotype categories. The cohort was further selected to only include probands to avoid inflation of the diagnostic rate, however, the sequenced relatives can be retrospectively checked

for cascade screening. Patients were excluded which did not have a nystagmus horizontal waveform phenotype and were further excluded if the ERG result was ‘abnormal’ (indicative of a retinal condition). This incurred the loss of 17 samples (Table 6.6) and resulted in a remaining total of n=81 nystagmus and albinism in the final cohort (Table 6.7).

Table 6.6: *Seventeen samples omitted that were not probands, did not have horizontal waveform nystagmus or had ‘abnormal’ ERG results. Four individuals (from respective family member pairs) of the six family member pairs were omitted due to being non-probands (bold). The two other non-probands (NG277 and NG412) were previously omitted for having insufficient coverage.*

No.	Sample ID	Proband	Nystagmus	Nystagmus waveform	ERG
1	NG181	No	No	-	Normal
2	NG200	No	Yes	Horizontal	Normal
3	NG218	No	Yes	Horizontal	Normal
4	NG234	No	Yes	Horizontal	Normal
5	NG244	No	Yes	Horizontal	Normal
6	NG250	Yes	No	-	Normal
7	NG251	Yes	No	-	Normal
8	NG257	Yes	No	-	Normal
9	NG322	Yes	No	-	Normal
10	NG344	Yes	No	-	Normal
11	NG357	No	Yes	Horizontal	Normal
12	NG387	No	Yes	Horizontal	Normal
13	NG403	Yes	No	-	Normal
14	NG451	Yes	No	-	Normal
15	NG557	Yes	No	-	Normal
16	NG320	Yes	Yes	Upbeat	Normal
17	NG285	Yes	Yes	Horizontal	Abnormal

Table 6.7: *Overall coverage, contamination and file sizes of 81 retained samples in batches 1-6. The minimum coverage at 20x is 90.6%.*

Batch	No. samples	Mean depth	Min depth	Max depth	SD depth	Mean coverage at 20x	Min coverage at 20x	Max coverage at 20x	Mean hets (%)	Min hets (%)	Max hets (%)	Max freemix	Mean FastQ size (GB)	Mean BAM size (GB)	Mean number of coding variants
1	16	166.3	114.2	222.3	30.8	0.980	0.972	0.987	66.1	64.3	72.2	0.002	2	6.6	12714
2	13	143.3	80.7	187.4	31.4	0.970	0.939	0.984	65.9	64.2	67.8	0.002	1.5	4.4	12063
3	21	123.5	61.4	214.0	46.3	0.956	0.911	0.986	65.5	63.4	68.6	0.001	1.6	5.4	11944
4	2	110.3	99.0	121.6	16.0	0.957	0.952	0.963	65.2	64.0	66.5	0.001	1.0	3.3	12145
5	13	107.7	89.5	129.5	11.1	0.940	0.915	0.968	62.5	57.4	64.4	0.000	1.4	4.6	10520
6	16	91.8	51.0	129.7	20.7	0.943	0.906	0.972	65.3	63.4	68.1	0.003	1.0	3.5	11849

In summary, a series of sample omissions were performed as depicted in Figure 6.2. The final cohort of $n=81$ samples had a minimum coverage at 20X depth across the 31 gene panel of 90.6%, and no evidence of contamination was identified (maximum freemix value of 0.003).

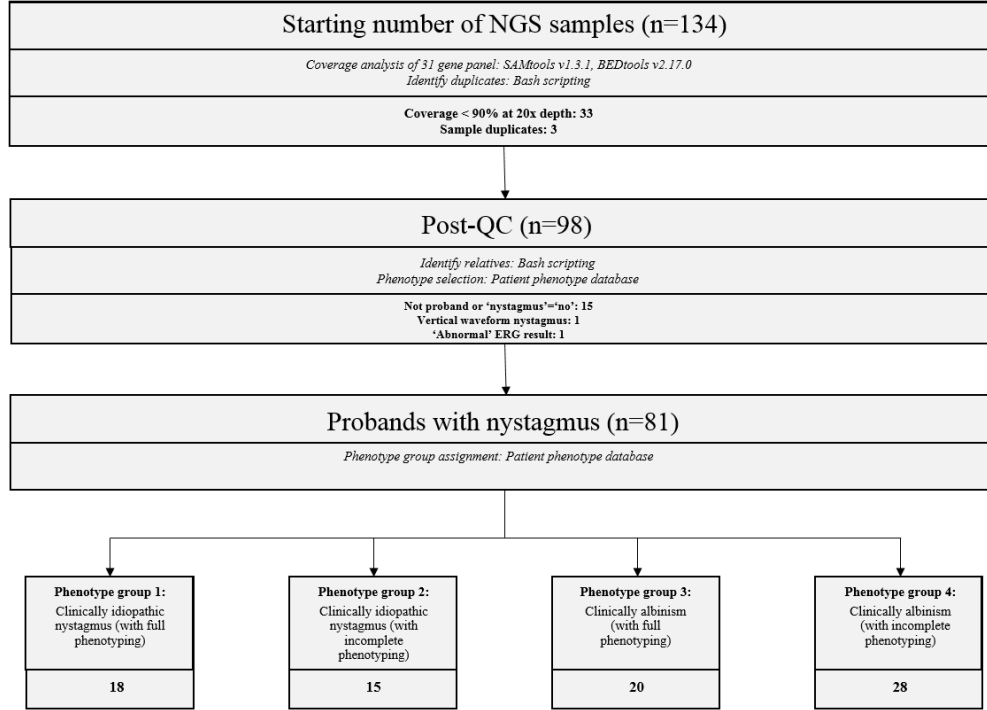


Figure 6.2: *Two steps of sample omissions from the analysis.* Initial number of NGS samples was $n=134$. Poorly covered samples across the 31 gene panel ($<90\%$ at 20X depth) were omitted ($n=33$). Sample duplications of inferior coverage across the 31 gene panel at 20X depth were removed ($n=3$). Samples which were not probands, did not have nystagmus or had ‘abnormal’ ERG results were omitted ($n=17$). The remaining final cohort was $n=81$ samples which consisted 18 phenotype group 1, 15 phenotype group 2, 20 phenotype group 3 and 28 phenotype group 4 patients.

6.5.2 Causal variant analysis

6.5.2.1 Assumed pathogenic variants

A total of 46 variants across the 81 participants met the criteria for assumed likely pathogenic genetic variants. For 17 patients (21.0% of the cohort), a total of 24 vari-

ants were considered to be likely causal variants (Table 6.8). Likely causal diagnoses were identified in 7 genes (*HPS5*, *PAX6*, *TYR*, *OCA2*, *CACNA1A*, *CACNA1F* and *FRMD7*) from the 31 gene panel. Twenty two heterozygous variants that were initially labelled as assumed pathogenic or assumed likely pathogenic were found in genes known to cause recessive disorders in patients without a second identified putative variant.

In phenotype group 1 (IINS with complete phenotyping), NG335 had a stop-gain variant (p.R335*) and NG433 had a hemizygous missense variant (p.G24R) in *FRMD7*. NG477 harboured a hemizygous missense variant (p.R519Q) in the X chromosome *CACNA1F* gene. Patient NG528 was found to have a possible compound heterozygous genotype in *TYR* from two heterozygous missense variants (p.P406L and p.P431T), which was an unexpected likely causal gene when considering the IINS phenotype for this patient.

Phenotype group 2 (IINS with incomplete phenotyping) patient NG315 had a one heterozygous missense variant (p.V443I) and one heterozygous splicing variant (NM_000275:exon6:c. 574-19A>G) in the *OCA2* gene as part of a possible compound heterozygous genotype. Although this likely causal gene is possible due to incomplete phenotype data of this patient, the likely causal gene was not anticipated based on the IINS phenotype group for this patient. NG381 was a male who was hemizygous for a splicing variant (NM_001256790:exon9:c. 1114+1G>C) in the *CACNA1F* gene.

NG280 had two heterozygous missense variants (p.V443I and p.P198L) and one heterozygous splicing variant (NM_000275:exon6:c. 574-19A>G) in the *OCA2* gene. Similarly, NG540 had the identical two heterozygous missense variants in the *OCA2* gene as seen for NG280 (p.V443I and p.P198L). These variants could be acting as a compound heterozygous in a likely causal genotype and were expected causal genes for phenotype

group 3 (albinism with complete phenotyping).

In phenotype group 4 (albinism with incomplete phenotyping), NG195 had a heterozygous missense variant in *PAX6* (p.A113P). NG340 had two heterozygous missense variants in *OCA2* (p.406L and p.G446S) whilst NG391 had a heterozygous splicing variant in *PAX6* (NM_001310159:exon2:c. 10+1G>A). NG394 had a homozygous *HPS5* splicing variant (NM_181508:exon4:c. 135+1G>A) and NG395 had two heterozygous missense *TYR* variants acting as a potential compound heterozygous genotype. NG416 also harboured two heterozygous missense *OCA2* variants. NG498 was male and hemizygous for a missense *CACNA1F* variant (p.R519Q) which was not expected for this phenotype group. NG543 was heterozygous for a missense *CACNA1A* variant (p.A453T) and NG551 was heterozygous for a missense *PAX6* variant (p.R142C).

Table 6.8: **Seventeen of 81 patients with assumed pathogenic variants which were determined to be likely causal genotypes.** Samples, blue background indicates male whilst pink indicates female, orange cells indicate heterozygous variants and red cells indicate homozygous variants. Samples were ordered by phenotypic group, and cells with a 'C' denote assignment of a likely causal variant and 'R' denote a reportable likely causal variant. The TYR variant, NM_000372:exon4:c.G1205A:p.R402Q (bold) would fail the MAF filter detailed above (17.7% AF in ExAC all populations) for putative variants but is highlighted here as it is listed as 'pathogenic' in ClinVar and is of relevance to subsequent work in this publication. Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Variant type, consequence of the variant; Gene.refGene, gene symbol; Omim Inheritance, inheritance as listed on OMIM for the gene in OCA/nystagmus; HI, DECIPHER v9.23 (0-10% likely to exhibit haploinsufficiency, 90-100% unlikely to exhibit haploinsufficiency); Amino acid, amino acid change; amsp144_rsID, dbSNP144 rsID; Failure rate, % of samples in cohort with a miscall at the site; ExAC all, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar) annotated 'p' if 'pathogenic'; InterVar, annotated as 'p' if identified as 'pathogenic' by InterVar; HGMD 2016 CLASS, annotated as DM for disease-causing mutation.

Chrom	Position	Ref	Alt	Variant type	Gene refGene	Omin Inheritance	HI	Amino Acid	rsasp144	ExAC ALL	Failure rate (%)	CADD phred	MaxEnt Scan diff	CLINSIG	Interval	HGMD	Phenotype group			
																	1 (n=18)	2 (n=15)	3 (n=20)	4 (n=28)
2	237493530	G	A	nonsynonymous	MLPH	AR	0.856	R35Q			0.0	33.0	-			DM	NG335	NG433	NG477	NG528
6	35503739	A	C	splicing	TULP1	AR	0.571				0.0	23.8	7.65							
11	18310740	C	T	splicing	HPS5	AR	0.482				0.0	27.2	8.18			P	-			
11	31800832	G	A	nonsynonymous	PAX6	AD	0.008	R142C	rs121907918		0.0	34	-	P		DM				
11	31801623	C	G	nonsynonymous	PAX6	AD	0.008	A113P			0.0	27.8	-			DM				
11	31806401	C	T	splicing	PAX6	AD	0.008				0.0	27.2	8.18			P	DM			
11	89284793	G	A	nonsynonymous	TYR	AR	0.003	R402Q	rs1126809	0.17700	0.0	34.0	-	P						
11	89284805	C	T	nonsynonymous	TYR	AR	0.003	P406L	rs104894313	0.00350	0.0	32.0	-	P		DM				
11	89284879	C	A	nonsynonymous	TYR	AR	0.003	P431T	rs368604842		0.0	28.3	-			DM				
11	89284924	G	A	nonsynonymous	TYR	AR	0.003	G446S	rs104894317	0.00002	0.0	31.0	-	P		DM				
13	23339410	T	C	nonsynonymous	SACS	AR	0.353	N1489S	rs147099630	0.00910	0.0	0.0	-			DM				
15	27983383	T	C	nonsynonymous	OCA2	AR	0.623	N489D	rs121918170	0.00030	0.0	28.2	-	P		DM				
15	27985101	C	T	nonsynonymous	OCA2	AR	0.623	V443I	rs121918166	0.00280	0.0	34.0	-	P		DM				
15	27990579	G	A	synonymous	OCA2	AR	0.623	G371G		0.02060	0.0		-			DM				
15	28014795	T	C	nonsynonymous	OCA2	AR	0.623	Y342C		0.00020	0.0	24.3	-			DM				
15	28022554	G	A	nonsynonymous	OCA2	AR	0.623	P198L		0.00010	0.0	29.0	-			DM				
15	28022592	T	C	splicing	OCA2	AR	0.623			0.00620	0.0		0.72			DM				
15	48121951	T	G	stopgain	SLC24A5	AR	0.347	Y72X	rs142056637	0.00004	0.0	35.0	-			DM				
19	13317310	C	T	nonsynonymous	CACNA1A	AD	0.355	A453T		0.00480	0.0	28.2	-			DM				
X	49222730	T	G	nonsynonymous	CACNA1F	XL	0.400	N746T	rs41270886	0.00170	0.0	26.1	1.35			DM				
X	49226037	C	T	nonsynonymous	CACNA1F	XL	0.400	R519Q		0.03040	0.0	33.0	-			DM				
X	49226936	C	G	splicing	CACNA1F	XL	0.400				0.0	23.2	8.27			P	-			
X	132080053	G	A	stopgain	FRMD7	XL	0.285	R335X	rs137852208	0.00001	0.0	39.0	-	P	P	DM				
X	132100704	C	T	nonsynonymous	FRMD7	XL	0.285	G24I	rs137852210		0.0	29.6	-	P	P	DM				

6.5.2.2 Assumed pathogenic and assumed likely pathogenic variants

For the remaining 64 patients without a likely causal genotype identified, assumed pathogenic variants together with assumed likely pathogenic variants (n=89 unique variants) were interpreted for likely causality. Thirteen of the 64 patients (16.0% of the total cohort) initially had likely causal diagnoses from assumed likely pathogenic variants with a cumulative total of 31 unique variants (Table 6.9). Likely causal diagnoses were initially identified across 7 genes (*TYRP1*, *SETX*, *PAX6*, *SACS*, *OCA2*, *CACNA1A* and *FRMD7*) from the 31 gene panel. The remaining 58 heterozygous variants were found in genes known to cause recessive disorders in patients without a second identified putative variant.

When interpreting phenotype group 1 variants, two reportable likely causal variants were identified in NG299. These variants were identified as a heterozygous p.Q286K variant located in the autosomal dominant *PAX6* gene and a hemizygous p.C264G variant X chromosomal *FRMD7* gene. It is possible the both of these variants were having an effect on contributing to phenotype causality. Patient NG445 harbours two heterozygous variants (p.T1098I and p.G252E) located in the recessive *HPS5* gene. NG534 had a heterozygous p.P1137A variant in the *CACNA1A* gene.

In phenotype group 2, p.Q286K was identified as heterozygous and likely causal in the *PAX6* gene for patient NG296. NG318 had two reportable causal genotypes: two heterozygous *SACS* variants as a potential compound heterozygous genotype (p.K647N and p.Q1247K) and a heterozygous *CACNA1A* variant (p.E1014K). Similarly, NG327 also had multiple reportable likely causal genotypes: a heterozygous *PAX6* variant (p.Q286H), two heterozygous *SACS* variants as a potential compound heterozygous genotype (p.N4573H and p.P3678A) and a heterozygous variant in the *CACNA1A*

gene (p.E731A). NG383 had a likely causal hemizygous variant in the *FRMD7* gene (p.G682S).

For phenotype group 3, NG512 had a possible compound heterozygous likely causal genotype in the *OCA2* gene (p.N781D and p.Y342C). Similarly, NG251 was also found to have a possible compound heterozygous likely causal genotype (p.P27R and p.P346R) in the *TYRP1* gene.

NG386 of phenotype group 4 was found to harbour a likely causal homozygous variant in the *CACNA1A* gene. Patients NG333, NG399 and NG429 were found to have likely causal variants located in the same codon of *PAX6* encoding the amino acid p.Q286, as reported for NG299 (phenotype group 1), NG296 and NG327 (phenotype group 2). As previously stated, the most common causal gene with ‘assumed likely pathogenic’ variants was initially identified as the *PAX6* gene which was not specific to any particular phenotypic group. All of the likely causal variants from *PAX6* were accounted for by two variants which coded for the same codon, p.Q286. The failure rate (number patients with a variant miscall by GATK Haplotype Caller) of the two variants in *PAX6* exceeded the 5% threshold across the cohort. Therefore, these ‘likely causal’ variants were prioritised for verification with Sanger sequencing to determine whether or not they were false positive variant calls.

Table 6.9: *Sixty-four patients investigated for likely causal genotypes as assumed likely pathogenic or as a combination of assumed pathogenic and assumed likely pathogenic. Thirteen patients with assumed likely pathogenic variants which were determined to be likely causal based on the known inheritance pattern. Samples, blue background indicates male whilst pink indicates female, orange cells indicate heterozygous variants and red indicates homozygous variants. Samples were ordered by phenotypic group, and cells with a 'C' denotes a likely causal variant, 'R' denotes a reportable variant, '-' denotes variant miscall. Chrom, chromosome; Pos, position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Variant type, consequence of the variant; Gene.refGene, gene symbol; OMIM, inheritance as listed on OMIM for the gene in OCA/ nyctagmus; Haploinsufficiency, DECIPHER v9.23 (0-10% likely to exhibit haploinsufficiency, 90-100% unlikely to exhibit haploinsufficiency); Amino acid, amino acid change; avsnp144, dbSNP144 rsID; Failure Rate, % of samples in cohort with a miscall at the site; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar) annotated 'p' if 'pathogenic'; InterVar, annotated as 'p' if identified as 'pathogenic' by InterVar; HGMD 2016 CLASS, annotated as DM for disease-causing mutation; Variant category, 1=assumed pathogenic, 2=assumed likely pathogenic.*

Chrom	Position	Ref	Alt	Variant type	Gene refGene	Omn Inheritance	HI	Amino Acid	avsnp144	ExAC ALL	Failure rate (%)	CADD phred	MaxEnt Scan diff	CLINSIG	InterVar	HGMD	Variant category	Phenotype group			
																		1 (n=14)	2 (n=13)	3 (n=18)	4 (n=19)
1	235766255	G	A	nonsynonymous	LYST	AR	0.531	T1982I	rs146591126	0.00630	0.0	17.1						NG299	NG534	NG383	NG333
3	149162256	G	A	nonsynonymous	HP53	AR	0.444	G739R	rs78336249	0.00960	0.0	22.9								NG512	NG399
5	78015546	C	T	nonsynonymous	AP3B1	AR	0.105	V999M	rs146503597	0.00380	0.0	16.4	-0.79								
6	15523217	G	A	nonsynonymous	DTNBP1	AR	0.433	P272S	rs17470454	0.04360	0.0	19.9	0.14			DP					
6	35505739	A	C	splicing	TULP1	AR	0.571				0.0	23.8	7.64		P						
9	12694076	C	G	nonsynonymous	TYRP1	AR	0.218	P27R	rs373327120	0.00002	0.0	24.5								C	
9	12702394	C	G	nonsynonymous	TYRP1	AR	0.218	P346R	rs377679582	0.00007	0.0	31.0								C	
9	12704541	C	T	nonsynonymous	TYRP1	AR	0.218	T366M	rs199823942	0.00003	0.0	24.3									
9	132342716	A	C	nonsynonymous	SETX	AR	0.694	L158V	rs145438764	0.00370	0.0	24.3				DM?					
10	102067257	T	G	nonsynonymous	HP56	AR	0.638	W595G			53.1							-	-	-	-
11	18281986	G	A	nonsynonymous	HP55	AR	0.482	T1098I	rs61884288	0.02360	0.0	18.6		P		DM?		C			
11	18306204	C	T	nonsynonymous	HP55	AR	0.482	G252E	rs755846129	0.00003	0.0	32.0						C			
11	31789937	C	A	nonsynonymous	PAX6	AD	0.008	Q286H			27.2	26						-			
11	31789939	G	T	nonsynonymous	PAX6	AD	0.008	Q286K	Rs751795008	0.00005	24.7	23.0						R	C		C
11	80284793	G	A	nonsynonymous	TYR	AR	0.003	R402Q	rs1126809	0.17700	0.0	34.0		P		DFP		R			
13	23330159	T	G	nonsynonymous	SACS	AR	0.353	N4573H	rs34382952	0.00320	0.0	25.5									
13	23332844	G	C	nonsynonymous	SACS	AR	0.353	P3678A	rs17078601	0.03970	0.0	25.9							R		
13	23340137	G	T	nonsynonymous	SACS	AR	0.353	Q1247K			0.0	21.8							R		
13	23354532	C	T	nonsynonymous	SACS	AR	0.353	A694T	rs17325713	0.02330	0.0	15.0									
13	23354671	C	A	nonsynonymous	SACS	AR	0.353	K647N	rs201021919		0.0	23.6							R		
15	27845050	T	C	nonsynonymous	OCA2	AR	0.623	N781D			0.0	26.7	-0.06							C	
15	27985101	C	T	nonsynonymous	OCA2	AR	0.623	V443I	rs121918166	0.00280	0.0	34.0		P		DM					
15	28014795	T	C	nonsynonymous	OCA2	AR	0.623	Y342C		0.00020	0.0	24.3								C	
15	28022592	T	C	splicing	OCA2	AR	0.623			0.00620	0.0		0.71			DM					
15	52343197	T	A	nonsynonymous	MYO5A	AR	0.158	R1320S	rs61731219	0.03370	0.0	21.7	0.93								
19	13286647	G	C	nonsynonymous	CACNA1A	AD	0.355	P1137A	rs199793367	0.00040	0.0	23.1						C			
19	13298593	C	T	nonsynonymous	CACNA1A	AD	0.355	E1014K	rs16024	0.00260	0.0	16.8							R		
19	13298659	C	T	nonsynonymous	CACNA1A	AD	0.355	E992K			0.0	25.7									C
19	13300637	T	G	nonsynonymous	CACNA1A	AD	0.355	E731A	rs16019	0.01010	0.0	24.8							R		
X	132077973	C	T	nonsynonymous	FRMD7	XL	0.285	G682S			0.0	28.8								C	
X	132082478	A	C	nonsynonymous	FRMD7	XL	0.285	C264G			0.0	25.5						R			

6.5.2.3 PAX6 variant verification

Patients NG299 (phenotype group 1), NG296, NG327 (phenotype group 2), NG333, NG399 and NG429 (phenotype group 4) had ‘assumed likely pathogenic’ likely causal variants located in the p.Q286 codon of *PAX6*. As the failure rate of these two ‘likely causal’ variants in *PAX6* exceeded the 5% threshold across the total cohort (equivalent to a variant miss call in 4 or more patients), prioritised verification of the variants was performed with Sanger sequencing for the six samples.

The two variants within the p.Q286 codon of *PAX6* were located on a 3’ terminal coding exon. Figure 6.3 shows that there was a decrease average depth across the cohort for these variants, NM_001310160:exon10:c.C856A: p.Q286K (chr11:31789937) and NM_001310160:exon10:c.G858T: p.Q286H (chr11:31789939) with 21X and with 24X depth respectively. Furthermore, there was high GC content with repetitive elements spanning the region of the variants. The 20mer sequences (which the variants centre on) were considered unique within the human genome.

Sanger sequencing trace results and quality metrics of the two *PAX6* variants are shown in Figure 6.4. The Sanger trace results for NG327 showed a low signal to noise ratio for all dyes across the trace (software default threshold of 60). This indicates that the peaks in the Sanger chromatogram were not sufficiently higher for one type of dye over another in the trace. Sample NG429 had a trace score of Phred 3 which did not satisfy the quality threshold of Phred 20. These quality results suggested that the Sanger sequencing quality was not sufficient for verification in these particular samples. However, for all remaining Sanger sequencing results, the reference codon ‘CAG’ was confirmed which indicated that the variant calls were false positives. A poly(A) stretch adjacent to the codon of interest likely caused the erroneous variant call and it was therefore

likely that all instances of these variants were false positives. For this reason, the revised number of patients with likely causal diagnoses from ‘assumed likely pathogenic’ variants decreased from 13 to 9 patients (11.1% of the total cohort).

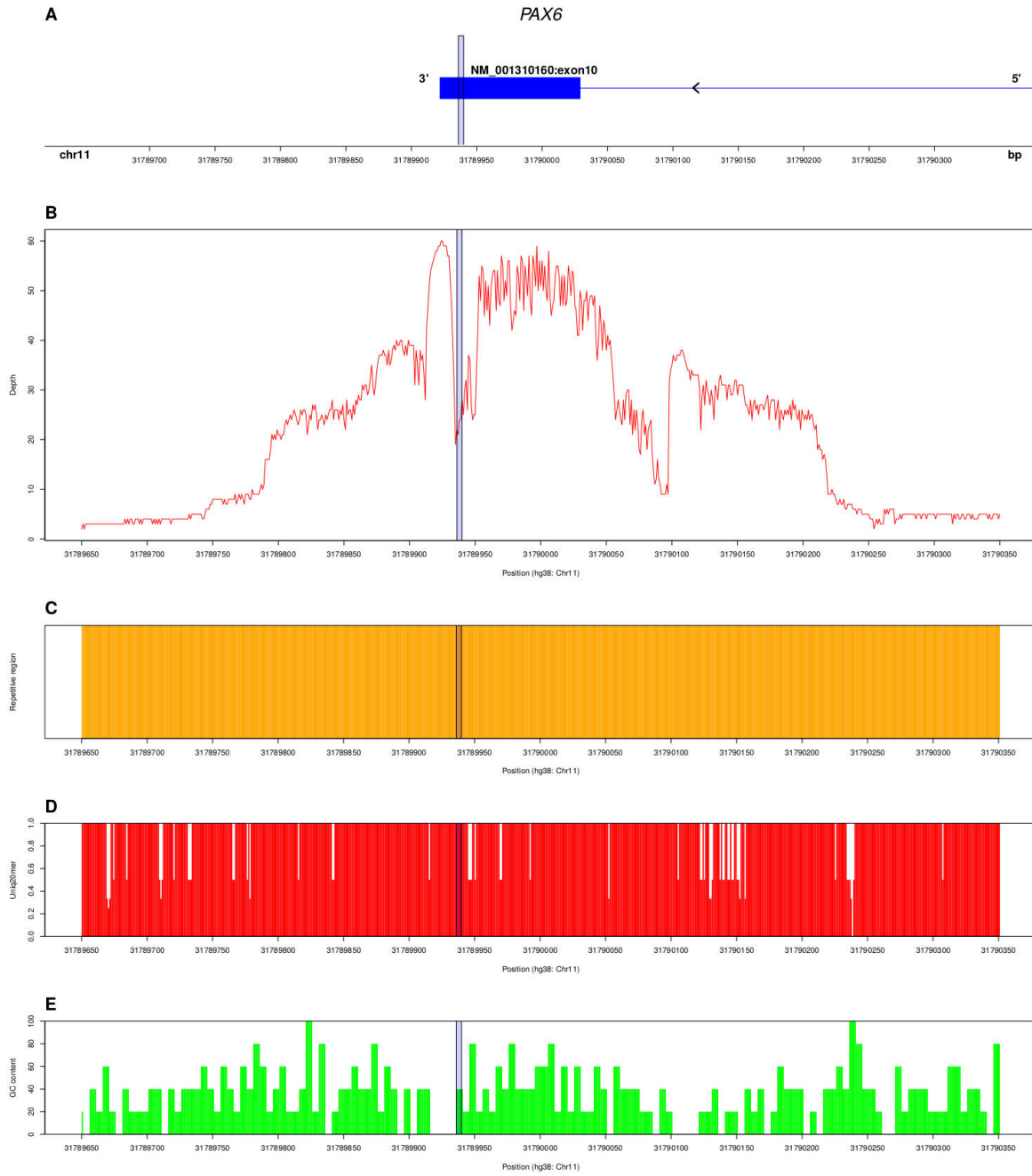


Figure 6.3: *PAX6* gene structure, sequencing depth and genomic context. A: gene structure as targeted by TruSight One, B: base depth, C: repetitive region, D: 20mer uniqueness, E: GC content. The codon containing the two variants NM_001310160:exon10:c.G858T: p.Q286H and NM_001310160:exon10:c.C856A: p.Q286K are highlighted in a blue box.

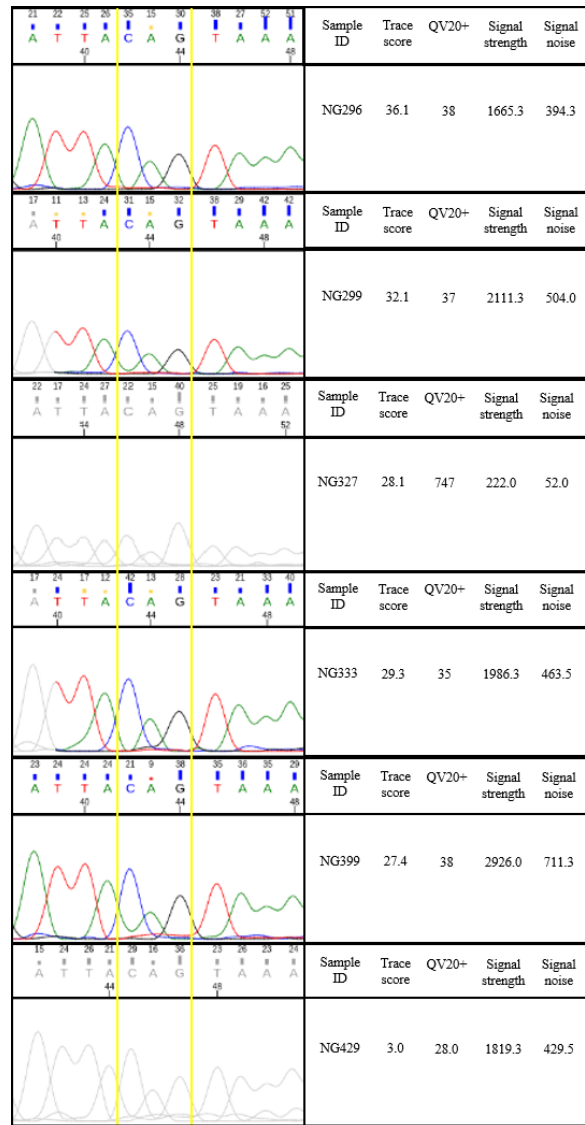


Figure 6.4: *Sanger sequencing chromatograms for the six samples which are queried for the presence of c.G858T and c.C856A (produced by Thermo Fisher Scientific Sanger ‘Quality Check’ application). The codon containing c.G858T and c.C856A was highlighted in the yellow box. Base call Phred scores are shown above each base sequenced. Overall quality metrics calculated from Sanger AP1 output files for the Sanger trace result were displayed on the right of each trace. NG327 and NG429 had insufficient quality to give accurate reportable verification of the variants due the observed trace Phred score and signal noise ratio respectively.*

6.5.2.4 TYR tri-allelic genotypic cause of albinism

The *TYR* variant, NM_000372.4:exon4:c. G1205A:p.R402Q was an exception in previous analyses as it is known to be insufficient to infer compound heterozygosity in causality. Therefore, this variant was interpreted differently as part of a tri-allelic genotype.

Of the 55 remaining undiagnosed patients, nine were identified to have a tri-allelic genotype within *TYR* (Table 6.10). The nine patients originated from phenotype groups 3 (n=6) and 4 (n=3). Each patient had a minimum of two common variants, NM_000372.4:exon1:c. C575A:p.S192Y (25.2% in ExAC) and NM_000372.4:exon4:c. G1205A:p.R402Q (17.7% in ExAC), and one assumed pathogenic/assumed likely pathogenic variant.

Table 6.10: **Fifty-five patients which did not have likely causal genotypes with assumed pathogenic or assumed likely pathogenic variants, nine samples were identified to have a likely causal tri-allelic genotype for albinism within TYR.** All TYR variants identified as assumed pathogenic or assumed likely pathogenic with the addition of S192Y and R402Q are listed. Position, location of 5' base of variant in hg38; REF, reference allele; ALT, alternative allele; Variant type, consequence of the variant; AACHange; aausp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar); InterVar, pathogenicity category according to InterVar interpretation; HGMD 2016 class, HGMD annotation for pathogenicity; Samples, orange indicates heterozygous variants, red indicates homozygous variants. Samples were ordered by phenotypic group and 'c' was used to indicate a likely causal variant or 'R' was used to indicate a variant was reportable, grey highlights the common variants involved in the tri-allelic genotype outlined by Norman et al (S192Y and R402Q).

Position	Alt	Variant type	Amino Acid	avsnpl44	ExAC ALL	Failure rate (%)	CADD phred	MaxEnt Scan diff	CLINSIG	Interval	HGMD	Variant category	Phenotype group									
													3 (n=16)						4 (n=18)			
89178528	A	nonsynonymous	S192Y	rs1042602	0.2518	0.0	25.0				DFP		NG263	NG356*	NG454	NG483	NG530	NG559	NG309	NG441	NG536	
89178602	T	nonsynonymous	R217W	rs63159160	0.0002	0.0	25.1		P		DM	1										C
89178769	T	nonsynonymous	W272C			0.0	29.6				DM	1						C				
89191278	A	nonsynonymous	R299H	rs61754375	0.00007	0.0	33.0		P		DM	1			C							
89227822	A	splicing		rs61754382	0.00001	0.0	25.2	8.75		P	DM	1					C					
89227850	T	nonsynonymous	A355V	rs151206295	0.0002	0.0	26.8			P	DM	1										
89227885	T	nonsynonymous	H367Y	rs776054795	0.00001	0.0	29.5				DM	1					C					C
89284792	T	stopgain	R402*	rs62645917	0.00005	0.0	51.0		P	P	DM	1	C									
89284793	A	nonsynonymous	R402Q	rs1126809	0.17700	0.0	34.0		P	P	DFP	1		C	C	C	C	C	C	C	C	C
89284805	T	nonsynonymous	P406L	rs104894313	0.0035	0.0	32.0		P		DM	1								C		
89284852	T	nonsynonymous	R422W	rs749979474	0.00001	0.0	34.0				DM	1			C							

6.5.2.5 Albinism patients with partially resolved genetic aetiology

Genetic testing in an albinism patient may only yield one heterozygous pathogenic/-likely pathogenic variant in a recessive OCA gene. With the absence of the required second heterozygous variant in the gene to form a causal genotype, inconclusive clinical reporting could be given. In such cases it is possible that through technical limitations the ‘second hit’ is being missed.

There were 46 patients without an assumed pathogenic, assumed likely pathogenic or *TYR* tri-allelic genotype identified. The 46 patients across groups 1 (11), 2 (10), 3 (10) and 4 (15) without genetic diagnoses were investigated for single heterozygous assumed pathogenic and assumed likely pathogenic variants in an OCA/ OA genes (*TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5*, *C10orf11*, *GPR143* [362]). This analysis did not include the *TYR* p.R402Q variant as this was previously highlighted as a unique case.

Sixteen patients had a single assumed pathogenic or assumed likely pathogenic variant which most likely contribute to the albino phenotype (Table 6.11). This corresponds to 2/10 (20.0%), 7/10 (70.0%) and 7/15 (46.6%) patients for of the remaining unresolved cases for phenotype group 2, 3 and 4 respectively. No patients in group 1 had a single assumed pathogenic or assumed likely pathogenic variant in any albinism gene.

Table 6.11: Forty-six patients were analysed for single heterozygous assumed pathogenic or assumed likely pathogenic variants in at least one albinism gene (TYR, OCA2, TYRP1, SLC45A2, SLC24A5, C10orf11 and GPR143). Sixteen patients are shown to have an assumed pathogenic or assumed likely pathogenic variant in at least one albinism gene. TYR variant, NM_000372:c.61205A>p.R402Q (bold) would fail the MAF filter detailed above (17.9% AF in ExAC all populations) for putative variants but is highlighted here as it is listed as 'pathogenic' in ClinVar. Samples, orange indicates heterozygous variants, red indicates homozygous variants. Samples were ordered by phenotypic group. 'C' and 'R' were not used here as these variants alone would not be causal or reportable. Chrom, chromosome; Pos, location of 5' base of variant in hg38; REF, reference allele; ALT, alternative allele; Func.refGene, gene feature; Gene.refGene, gene name; Omim, Inheritance (all populations) as listed on OMIM for the gene in OCA / nyctagmus; Amino acid, amino acid change; avsnp144, dbSNP144; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar); InterVar, pathogenicity category according to InterVar interpretation; HGMD 2016 class, HGMD annotation for pathogenicity.

[illegible]

6.5.2.6 Overview of diagnostic results

Table 6.12 summarises the number of samples harbouring likely causal variants. Overall, a diagnostic rate of 43.2% for the cohort was achieved. The genetic aetiology was determined for 38.9% of the patients in phenotype group 1 (IINS with complete phenotyping), however, the majority were resolved by ‘assumed pathogenic’ variants. Phenotype group 2 (IINS with incomplete phenotyping) had 33.3% of patients’ genetic basis of disease resolved. In contrast, this group had more cases resolved with ‘assumed likely pathogenic’ genetic causes. Groups 3 and 4 had 50.0% and 57.1% of their patients’ underlying genetic cause resolved respectively, however, this high diagnostic rate was mainly contributed by the tri-allelic genotypes in the *TYR* gene.

Table 6.12: *Summary table summarising the number of samples identified with likely causal genotypes. A diagnostic rate is calculated and the likely causal genes are listed for each phenotype group.*

Phenotype Group No.	Cohort sub-group	Samples	Samples with assumed pathogenic diagnostic variants	Samples with assumed likely pathogenic diagnostic variants	Samples with <i>TYR</i> tri-allelic genotype	% Samples with likely causal variants	Likely causal genes reported
1	Idiopathic nystagmus (with complete phenotyping)	18	4	3	0	38.9	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>FRMD7</i> , <i>HPS5</i> , <i>TYR</i>
2	Idiopathic nystagmus (with incomplete phenotyping)	15	2	3	0	33.3	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>FRMD7</i> , <i>OCA2</i> , <i>SACS</i>
3	Albinism (with complete phenotyping)	20	2	2	6	50.0	<i>OCA2</i> , <i>TYR</i> , <i>TYRP1</i>
4	Albinism (with incomplete phenotyping)	28	9	1	3	46.4	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>HPS5</i> , <i>OCA2</i> , <i>PAX6</i> , <i>TYR</i>
All		81	17	9	9	43.2	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>FRMD7</i> , <i>HPS5</i> , <i>OCA2</i> , <i>PAX6</i> , <i>SACS</i> , <i>TYR</i> , <i>TYRP1</i>

6.6 Discussion

This chapter details how a molecular genetic basis of disease identified in 43.2% of an unselected cohort of children with infantile nystagmus syndrome is achievable by utilising detailed phenotyping, a stringent gene panel based on the TruSight One capture kit and causal variant interpretation in line with clinical laboratory diagnostic protocols.

The diagnostic yield was reported for patients falling within the four most common clinical scenarios seen in clinical practice; IINS with complete phenotyping, IINS with incomplete phenotyping, albinism with complete phenotyping and albinism with incomplete phenotyping. A diagnostic rate across 81 patients of 43.2% is substantially higher than other exome diagnostic analyses with the TruSight One capture (approximately 26.3%) [371]. Due to the similar diagnostic yields generated between complete and incomplete phenotyping groups, this study supports the necessity to use a minimum of freely available phenotyping measures (which may not be complete or detailed) along with a stringent selection of clinically callable genes.

In most cases, likely causal genotypes involved genes which were known to cause the phenotype category of the patient. Six patients had assumed pathogenic variants in genes that would not have been previously directly implicated in causing the phenotype presentation of the patient. For example, NG528 from phenotype group 1 (IINS with complete phenotyping) was found to have a likely causal compound heterozygous genotype in the *TYR* gene. Similarly, NG315 from phenotype group 1 was found to have a likely disease-causing compound heterozygous genotype in the *OCA2* gene. These cases reflect the variable, often hypomorphic and overlapping phenotypes seen in children with nystagmus [25, 345].

NG381 of phenotype group 2 was found to be homozygous for an assumed likely pathogenic splicing variant in the *CACNA1F* gene, which is known to cause Aland Island eye disease, cone-rod dystrophy and X-linked incomplete congenital stationary night blindness (CSNB); all of which involve nystagmus and retinal dystrophy. It might be expected that such disorders would be identified by ERGs prior to recruitment to this study, however, this patient's ERG result was initially reported as normal. A sub-

sequent ERG performed at an older age for this patient identified the typical features of CSNB. This case and others identified here support that future gene panels for conditions as broad and variable as nystagmus might include genes for conditions known, in some cases, to be missed by age-appropriate phenotyping methods. They may also support an argument that genomic testing in the future may form an earlier part of the diagnostic workflow, and subsequent, more detailed phenotyping be directed towards proving or disproving diagnoses suggested by putative likely causal variants.

Of the 55 remaining patients, nine had the pathogenic tri-allelic genotype in the *TYR* gene. All tri-allelic *TYR* genotypes were found exclusively within the albinism phenotype group patients (phenotype groups 3 and 4). Incorporating and acknowledging the known tri-allelic in which *TYR* can cause OCA has proven to support suggestions that compound heterozygosity cause much of the OCA phenotypes [346].

Forty-six patients remained without diagnosis across phenotype groups 1-4. No assumed pathogenic or assumed likely pathogenic variants were identified in albinism genes (*TYR* variant p.R402Q not included) in any patient from group 1 but notably, at least one such variant was identified in 20.0%, 70.0% and 46.6% of the remaining patients in groups 2, 3 and 4 respectively. Furthermore, Chiang *et al* have reported that mutation of *TYRP1* (OCA3) can modify the OCA2 phenotype, resulting in red hair of an individual [360]. Similarly, the phenotype of OCA2 can be modified to show albinism with red hair in affected individuals with *OCA2* and *MC1R* mutations [26]. Taken together, this suggests that for a significant proportion of patients with albinism phenotypes, additional causal genetic variant(s) are present but may be currently unidentified because they are either (1) non-exonic, (2) are in unknown genes, (3) are common variants and are filtered out (such as was the case for the *TYR* p.S192Y

variant), (4) are copy number variants, or (5) that there exists poorly understood epistasis between melanogenesis pathway genes. In essence, this reinforces the need to consider more complex genotypes in albinism in order to identify further genetic causes and provide greater diagnostic yields from clinical genetic tests.

The 56.8% of the cohort for which causal genotypes were not identified should be followed up in future analyses. ACMG guidelines (for which the ‘assumed likely pathogenic’ category was loosely based on) were proven to largely be inappropriate for variant prioritisation in albinism. This was likely due to the critical role of common variants in causality of albinism which were incompatible with ACMG guidelines. The InterVar software which attempts to apply ACMG guidelines criteria would typically annotate such variants as ‘uncertain’ or ‘benign’ due to the high allelic frequency of the variants or contradictory pathogenicity annotation from databases such as ClinVar and HGMD. The ClinVar database currently has the highest proportion of pathogenicity misclassifications for low penetrance variants [380, 381]. However, whilst HGMD could be considered a more prudent form of functional annotation due to its manual curation of literature, it has been known to identify low penetrance variants as disease-causing mutations (‘DM’) [345, 382]. Although no such instances were identified in this cohort, databases by which analysts are reliant upon are imperfect and are could be problematic when determining pathogenic variants as per ACMG guidelines, particularly for common (low penetrance) variants. Filter parameters for allele frequency (5%) or pathogenicity scores of CADD Phred ≥ 15 could be relaxed to potentially increase the number of diagnosed cases. However, this would result in an increase in false positives, which would not be appropriate for a clinical report of diagnostic variants [383]. This highlights the need for more functional evidence of variants in nystagmus and albinism genes in order to contribute to more accurate classification of variants.

To perform NGS analyses to determine diagnostic causal variants, the analysis was reliant on the integrity of the capture kits to target the genomic regions of interest. The TruSight One capture kit claims to target the exome of 4811 genes which were relevant to disease. The candidate gene list used to prioritise variants was derived from the ‘albinism and nystagmus’ gene panel from the UKGTN. However, the genes involved in this panel included 22 albinism or pigment affecting genes, resulting in a bias in diagnostic rate towards albinism patients. The gene panel could therefore be expanded to cover other known nystagmus-causing genes such as *SLC38A8* which could account for a proportion of the remaining 56.8% unresolved patients. It was also apparent that the coverage for *C10orf11* was substantially lower (70.5% at 20X depth) than other genes in the 31 gene panel (Appendix Table E.5). Therefore, it is possible that there were causal variants within this gene which were not detected and reported to the clinician. If newly implicated genes in nystagmus and albinism were identified and desired to be included in a candidate gene list for analysis, the new release of the TruSight One Expanded (v3) capture kit would likely be a more suitable option as it captures a more up-to-date list of 6699 genes which have been clinically associated with diseases. The TruSight One capture kit enables economical and efficient capture of clinically relevant genes, from which subpanels can be utilised for a more economically beneficial and improved time efficient determination of a genetic cause. Thus, the TruSight One capture kit and utilisation of specific gene panels could be employed as an ideal ‘first pass’ genetic testing method for children with nystagmus.

It is also possible that a diagnostic report may initially identify false positives, as seen for the two *PAX6* variants (c.G1308T and c.C1306A). However, all identified variants which have been identified as ‘likely causal’ would need to be verified through Sanger

sequencing before reporting as ‘causal’ [376]. Whilst samples with coverage greater than or equal to 90% at 20X depth were selected for analysis, the *PAX6* gene (and *HPS6* gene) had variants with failure rates exceeding 5% across the cohort. Additional filters may be applied to all variants identified to remove those with quality in the ‘grey area’, however, this would result in accepting the loss of potentially diagnosable variants as collateral.

6.7 Conclusion

In conclusion, for clinicians using a standard set of phenotyping criteria and the UKGTN approved 31 gene panel based on the TruSight One platform, a clinically callable genetic diagnosis for 43.2% of children with INS where no other clinical feature suggests a specific underlying diagnosis was achievable. This could significantly reduce the time and number of investigations that many children with nystagmus undergo. This can also permit informed family counselling with regards to recurrence risk in addition to rapid access to tailored management.

7 Determining causal genetic variants for ocular disease in a consanguineous Pakistani cohort

7.1 Synopsis

This chapter involves a collaboration with research colleagues at the University of Exeter who had access to a Pakistani cohort with multiple ophthalmic conditions including nystagmus, albinism, cataract, Waardenburg syndrome, microcornea, Joubert syndrome, and Usher syndrome. Variants detected in the NGS data are interrogated with appropriate UKGTN gene panels and the diagnostic yield is determined.

Luke O’Gorman was responsible for the transfer and demultiplexing of sequence data to the university servers, processing, analyses and interpretation of data which were supervised by Sarah Ennis, Jane Gibson, Angela Cree, Andrew Lotery, and Jay Self. Our University of Exeter collaborators acquired the DNA samples and clinical details (supplied by colleagues at the International Islamic University, Pakistan). Chelsea Norman was responsible for the library preparation and sequencing of the DNA samples.

7.2 Ophthalmic diseases

This section describes the genetic epidemiology and clinical details for a range of ophthalmic diseases that are covered in this chapter, including nystagmus, albinism, cataract, Waardenburg syndrome, microcornea, Joubert syndrome, and Usher syndrome. The clinical details of nystagmus and albinism were previously detailed in Chapter 1.2.

7.2.1 Cataracts

Cataracts are the most common cause of blindness world-wide [1] and has an incidence of 1-6:10,000 [384]. The condition can occur in isolation or as part of a syndrome affecting multiple tissues [385]. A cataract is a clouding of the normally clear lens inside the eye, which leads to vision impairment and blindness [386]. In the short-term, brighter lighting or glasses can help counter the visual impairment caused by cataracts. Cataract surgery with an intraocular lens (IOL) implant is the most prevalent eye operation in the United States [387] and is thought to be one of the most effective surgical procedures in medicine [386]. Age is the most common risk factor associated with cataract formation [386].

Congenital cataracts can be inherited through autosomal recessive, autosomal dominant and X-linked modes of inheritance [388]. Non-syndromic congenital cataracts are more commonly autosomal dominantly inherited [384]. The causal genes for cataracts encode proteins which can be categorised into two main groups: 1.) proteins which are important structural-function aspects of the lens such as crystallins (e.g. *CRYAA*, *CRYBB2* and *CRYGD*), connexins (*GJA3* and *GJA8*) or cytoskeleton proteins (e.g. *MIP* and *BFSP2*); 2.) developmental/regulatory processes involving several transcription factors (e.g. *HSF4* and *PITX3*) and functionally diverse genes (e.g. receptor kinase *EPHA2* gene, mRNA binding *TDRD7* gene and the microtubule vesicle transport *FYCO1* gene) [388].

7.2.2 Waardenburg syndrome

Waardenburg syndrome can present as one of four different types. Type I Waardenburg syndrome is characterised by a white forelock and premature grey hair, changes to the pigment of the iris (e.g. heterochromia iridis and ‘brilliant blue eyes’), congenital

hearing loss and dystopia canthorum (giving an appearance of a widened nasal bridge). Type II is distinguished from type I by the absence of dystopia canthorum. Type III has dystopia canthorum similar to type I, however, it is distinguished by the presence of upper limb abnormalities. Type IV, has the additional feature of Hirschsprung disease (causing chronic constipation) [389, 390]. The four Waardenburg syndrome types collectively have an incidence of 1:20,000-1:40,000 [391] with East African populations having the highest incidence [392, 393].

Waardenburg syndrome types I and II take an autosomal dominant mode of inheritance, whereas type III and IV are known to be inherited in both autosomal dominant and recessive modes of inheritance. Types I and II are the most common forms of Waardenburg syndrome, while types III and IV are considered more rare [389, 390, 394]. Waardenburg syndrome is known to be caused by variants in *PAX3* (types I & III), *MITF* (type II), *SOX10* (type II) [390], *SNAI2* (type II) [395], *EDN3* (type IV) and *EDNRB* (type IV) [390].

7.2.3 Microcornea

Microcornea is a rare congenital cornea dysgenesis and is defined as a cornea less than 11 mm in diameter [396], whereas normally the cornea matures in size to a diameter of 12 mm by 2 years of age. It is hypothesised that microcornea occurs due to overgrowth of the tips of the optic cup leaving less space for the cornea during development [396].

Microcornea can be inherited in an autosomal dominant, autosomal recessive or (in rare cases) sporadic pattern [397]. The condition often in conjunction with other ocular phenotypes including cataract, coloboma and microphthalmos [398, 399]. Genes known to be associated with microcornea include *GJA8* (cataract-microcornea) [400] and *BEST1* (microcornea, rod-cone dystrophy, cataract and posterior staphyloma) [401].

7.2.4 Joubert syndrome

Joubert syndrome is a rare disorder which causes agenesis of the cerebellar vermis presenting episodic hyperpnoea (rapid breathing), abnormal eye movements, ataxia and intellectual disability [402]. Joubert syndrome is inherited in an autosomal recessive manner and has an estimated prevalence in the general population of 1:100,000 and a carrier frequency of 1:160 of individuals [403]. An unaffected sibling of an affected individual with Joubert syndrome has a $2/3$ chance to be a carrier of the disease in a known causal gene. This results in a $1/960$ chance of the unaffected sibling giving birth to a child with Joubert syndrome [403].

Causal genes from Joubert syndrome encode proteins of the primary cilium and are termed ‘ciliopathies’. The primary cilia are critical in the development of multiple cell types including the cerebellum, retinal photoreceptors, neurons, kidney tubules and bile ducts.

7.2.5 Usher syndrome

Usher syndrome is characterised by hearing loss, vision loss and occasional balance problems. Usher syndrome is an autosomal recessive disease [404, 405]. Causal genes for Usher syndrome generally encode proteins known to provide multiple cellular functions such as intracellular transport, organisation of multiprotein complexes, cell adhesion and cell signalling. The Usher syndrome proteins are known to interact and it is suggested that they function in multiprotein complexes in vivo [404, 405]. It can be classified into three types [406]: type I patients (caused by mutations in *MYO7A*, *USH1C*, *CDH23*, *PCDH15*, *USH1G*, *CIB2* genes [404]) are defined as having congenital severe-to-profound deafness, vestibular areflexia and onset of retinitis pigmentosa (RP) within the first decade of life; type II patients (caused by the *USH2A* gene [404]) show

congenital moderate-to-severe hearing loss, normal vestibular function and onset of RP within the second decade of life; type III patients (caused by *CLRN1* and *PDZD7* genes [404]) experience hearing loss, vestibular dysfunction and a variable onset of RP. Early symptoms of RP are night blindness and loss of peripheral vision caused by degeneration of rod photoreceptors. Usher syndrome type II is the most common type and accounts for 50-65% of all Usher syndrome cases. Type I accounts for 10-35% whilst type III is rare and accounts for up to 2-5% of cases, with the exception of Finnish and Ashkenazi Jewish populations where type III accounts for 40% of all cases [404, 406].

It has an overall prevalence between 1:16,000 and 1:50,000, depending on the study ethnicity which have included Scandinavian, Columbian, British and North American populations [407]. Notably, the highest known prevalence of Usher syndrome within a population was identified in the city of Birmingham (UK) which included an undisclosed number of consanguineous Pakistani families in the study [408].

7.3 Population genetics

Genetic drift happens in populations of all sizes, although the effects are stronger in small populations [409]. A founder effect occurs when colonisation by a small number of the original population reduces genetic variation and causes a loss of heterozygosity over time [410]. This can result in divergence of variants between small sub-populations increasing over time particularly when the sub-populations become isolated [410]. As populations diverge they may also have different population-specific variants within the same genes. These differences in population specific variation can lead to different molecular pathologies [411, 412]. Population specific variation may also result in a greater likelihood of developing certain disease. Examples of such disorders include sickle cell anaemia [413], lactose intolerance [414] and Tay-Sachs disease [415].

7.3.1 Consanguinity

Consanguineous relationships are the union between related individuals. In first cousin unions, a couple is predicted to have $1/8$ of their genes being identical by descent (IBD). As a consequence, their progeny are expected to be homozygous for $1/16$ of all loci on average [416]. This leads to an elevation in homozygosity and a greater risk of recessive disease amongst the population [417]. A global overview of consanguinity by Bittles & Black shows highest rates of consanguinity were identified in North Africa, the Middle East and South Asia (Figure 7.1) [417].

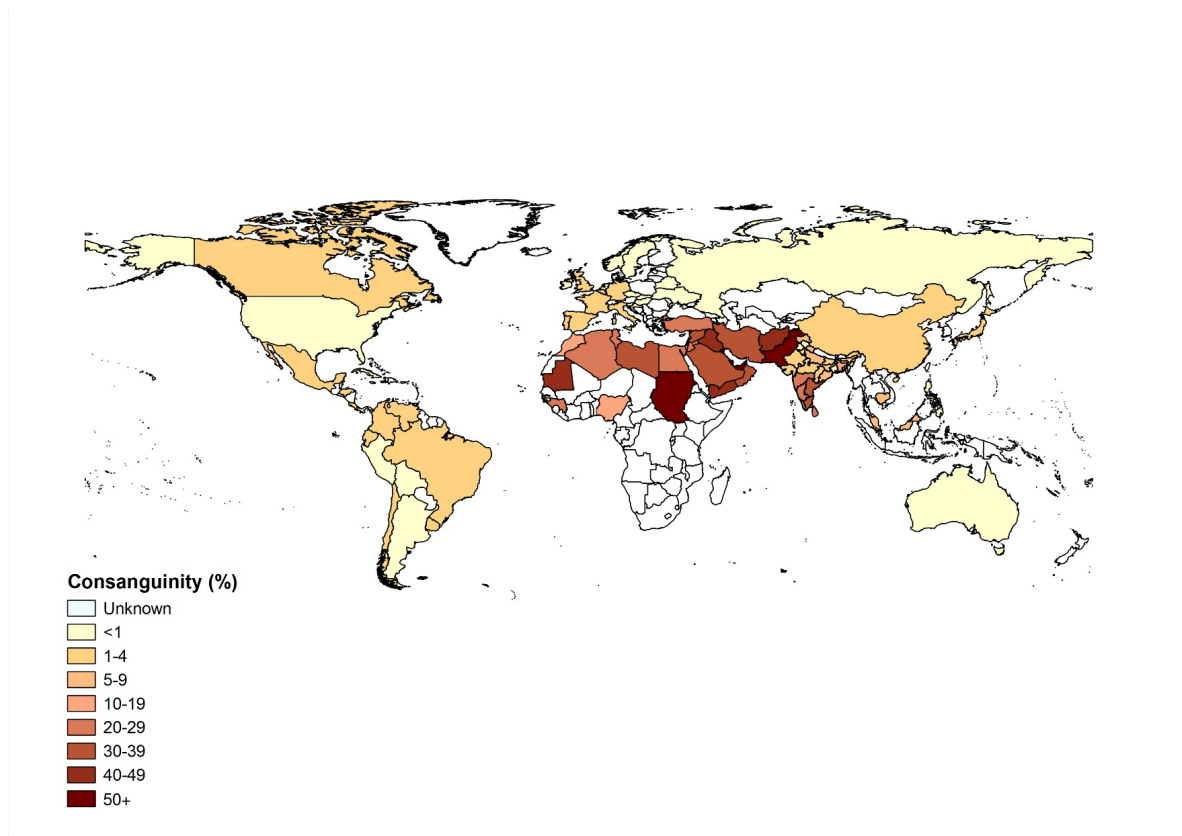


Figure 7.1: *Global map of marriages between couples listed as related second degree relatives or closer. Extracted from the Global Consanguinity website (www.consang.net) [417].*

It is reported that over 50% of Pakistanis are married to a first cousin and more than

60% are married to a known relative [418, 419]. Together with the common practice of endogamy and ‘Biraderi’, Pakistanis are considered amongst the most consanguineous populations world-wide [419, 420]. This cultural practice is also found within the Pakistani community in the UK, which is one of the largest ethnic groups (1.9% of the population [421]) and accounting for 4% of UK total births [420]. It has also been reported that approximately 30% of all recessive disorders in the UK are identified within the British Pakistani population [419, 420].

7.3.2 Ophthalmic disease in consanguineous pedigrees

Recessive ophthalmic diseases are more likely given high levels of consanguinity. Dineen *et al* identified that the most common cause of visual impairment and blindness (defined as $<3/60$ in the better eye on presentation) in Pakistan as cataracts (Figure 7.2) [422]. Cataracts caused 51.5% whilst glaucoma caused 7.1% of blindness. The two major causes of visual impairment ($<6/18$ to $\geq 6/60$) were refractive error (43%) followed by cataract (42%).

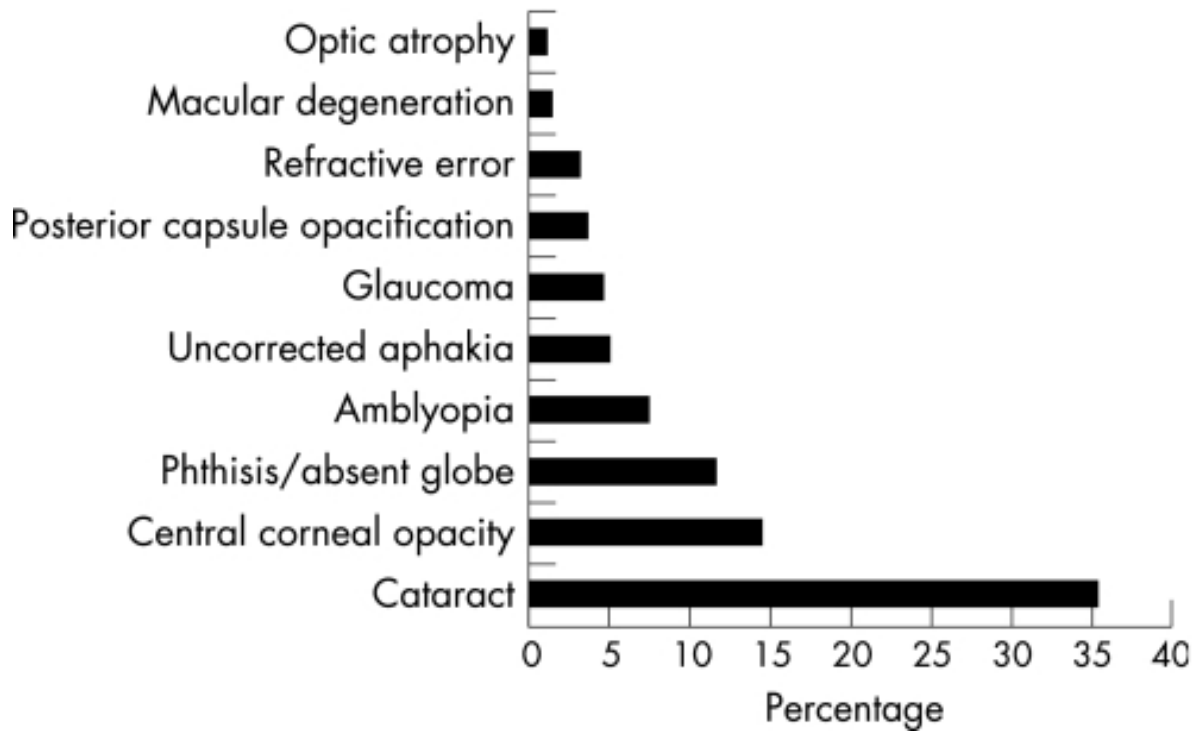


Figure 7.2: *Causes of visual impairment and blindness in Pakistan.* (Extracted from Dineen et al [422]).

Little is known of the prevalence of nystagmus in causing blindness or visual impairment in the Pakistani population, however, some research has been performed on the South Asian populations of Bradford and Leicester (UK) [19, 423]. Research in Leicester investigated the distribution of visual impairment in children and found a higher proportion of patients with nystagmus in the Caucasian population compared with the South Asian population [19]. However, another study in Bradford identified that 20.5% of South Asians registered low vision due to nystagmus compared with only 7.9% of the Caucasian population [423]. The Leicester study had an Asian population consisting of mainly British Indians (25.7% British Indian and 1.5% British Pakistani) whilst the Bradford Asian population had an Asian population consisting of mainly British Pakistanis (2.7% British Indian and 14.5% British Pakistani). The Bradford South Asian population was reported to have 44% of patients with a family history of an

ocular disease and the increase in nystagmus prevalence was likely due to higher levels of consanguinity within this ethnic sub-group [19, 423].

7.4 Aim

To identify likely causal variants amongst a cohort of 57 singleton consanguineous Pakistani patients with a range of ophthalmic diseases.

7.5 Methods

7.5.1 Patient selection

Fifty-seven singleton Pakistani patients born following first cousins unions (but independent of each other) with various ophthalmic diseases were selected for targeted exome sequencing (Table 7.1). Included were 33 OCA, 14 nystagmus, five congenital cataract, one Waardenburg syndrome, one microcornea, one Joubert syndrome, one Usher syndrome and one ‘blind’ patient. The families of the patients belonged to nine different ethnic subgroups (Pathan, Afridi, Khatak, Yousafzai, Punjabi, Rajpoot, Saraki, Virk and Niazi) and were recruited in Pakistan. With the exception of one patient (CBLN112), medical histories from all families were taken and a phenotypic diagnosis in all affected individuals was established by local clinicians. Facial photographs and videos were used to document clinical features and confirm disease status.

Table 7.1: *List of the 57 samples and their given phenotype. Samples were sequenced in two batches ($n=22$ and $n=35$) which is highlighted by a bold horizontal line.*

Row no.	Sample ID.	Reported clinical phenotype
1	30-2	OCA
2	32-1	OCA
3	38-1	OCA
4	39-1	OCA
5	40-1	OCA
6	41-1	OCA
7	CBLN112	Blind
8	PA1011	Congenital cataract
9	PA1031	OCA
10	PA1047	Congenital cataract
11	PA1055	OCA
12	PA1059	OCA
13	PA1074	Congenital cataract
14	PA1079	Congenital cataract
15	PA1084	Congenital cataract
16	PKNYS_01	OCA
17	PKNYS_02	Nystagmus
18	PKNYS_04	OCA
19	PKNYS_05	OCA
20	PKNYS_07	OCA
21	PKNYS_08	OCA
22	PKNYS_09	OCA
23	HC04	Waardenburg syndrome
24	MIC04	Microcornea
25	NG312	OCA
26	NG364	OCA
27	NG377	OCA
28	NG426	OCA
29	NG431	OCA
30	NG444	OCA
31	NG457	OCA
32	NG460	OCA
33	NG463	OCA
34	NG466	OCA
35	NG470	OCA
36	NG615	OCA
37	NG617	OCA
38	NG630	OCA
39	NG633	OCA
40	NG636	OCA
41	NG638	OCA
42	NY204	Nystagmus
43	NYS04	Nystagmus
44	NYS38-15	Nystagmus
45	NYS41-6	Joubert syndrome
46	NYS42-33	Nystagmus
47	PA1374	OCA
48	PA1400	OCA
49	PA1403	OCA
50	PA1410	OCA
51	PA1434	OCA
52	PA1444	OCA
53	PA1459	OCA
54	PA1483	OCA
55	US03	Usher syndrome
56	V-2	OCA
57	VI-14	OCA

7.5.2 Sample processing

Saliva samples were processed at the University of Southampton WISH laboratory and DNA was extracted using the Oragene-DNA kit (OG-575, DNA Genotek). The library preparation was performed across two batches (n=22 and n=35) with the TruSight one capture kit (n=4811 genes). Paired end NGS (2x150 cycles) was accomplished with the NextSeq 500 (mid-output).

7.5.3 Bioinformatic pipeline and quality control

FastQ reads were aligned to the hg38 human reference genome with the BWA-MEM [141] alignment tool. GATK v3.7 [144] was used to call SNPs and short indels. Annotation was performed using ANNOVAR v2015Dec [146] to annotate variants with information regarding variant consequence, dbSNP v144, variant allele frequency (1000 Genomes Project, Exome Sequencing Project and Exome Aggregation Consortium) and pathogenicity scores with SIFT [153], GERP++ [154] and CADD [196]. Further annotation was included from the Human Gene Mutation Database (HGMD) [172] and MaxEntScan [271] for splice site variants. Coverage was determined using SAMtools v1.3.1 [143] and BEDtools v2.17.0 [277]. Variants were excluded if they had a read depth<4.

In order to identify the runs of homozygosity (ROH) in the probands, PLINK v1.9 (LROH) was applied to BED, BIM and FAM derivatives of the VCF files using default parameters [424]. BEDtools v2.17.0 intersect [277] was then applied to identify long regions of homozygosity found in each sample's allocated gene panel.

FastQ quality were assessed with the FastQC v0.11.3 software [379]. Variant concordance between samples was checked for consistency as no relationship between the 57

individuals was expected. All patients were reported as the same Pakistani ethnicity. Samples with $\leq 80\%$ coverage at 20X were flagged as poorly covered but the samples were still analysed as no other DNA was available for resequencing, however, the ability to detect variants may have been compromised. Observation of elevated levels of heterozygosity and/or mismatching between recorded gender and determined gender (based on X chromosome heterozygosity) was used to check sample provenance. VerifyBamID v1.0 software was also used to highlight potential contamination.

7.5.4 Allocation of candidate gene lists

Gene panels were selected for samples based on the assigned phenotype to focus the prioritisation of variants on the most clinically relevant genes for the recorded phenotype. The 47 patients phenotyped as either albinism or nystagmus were allocated the ‘albinism and nystagmus 31 gene panel’ (see Chapter 6.4.5, Table 6.2). Five patients with congenital cataract were assigned the UKGTN ‘cataract disorders, congenital, 112 gene panel’ (Table 7.2). As no gene panel exists for ‘congenital blindness’ and the most common cause of blindness is cataract, the patient with blindness (sample ID: CBLN112) was also assigned the UKGTN ‘cataract disorders, congenital, 112 gene panel’. For patients with Waardenburg syndrome (HC04), Joubert syndrome (NYS41-6) and microcornea (MIC04), a common gene panel covering all three phenotypes was identified in the UKGTN database under the ‘eye malformations, congenital, 204 Gene Exome Panel’ (Table 7.3). However, only Waardenburg type I and III were partially covered by this gene panel. Also included in this gene panel were the genes *JBTS6*, *JBTS7*, *JBTS9*, *JBTS22* (Joubert syndrome), *ADAMTS18* and *CRIM1* (microcornea). Therefore, the Joubert syndrome (NYS41-6) and microcornea (MIC04) samples were assigned this gene panel. The Waardenburg syndrome (HC04) and Usher syndrome (US03) patients

were assigned the UKGTN ‘hearing loss, syndromic and non syndromic, 95 gene panel’ (Table 7.4) as this panel included all currently known Waardenburg syndrome causal genes and the known Usher syndrome type I-III genes.

Table 7.2: *Gene list of UKGTN cataract disorder genes (n=113). LARGE1 is listed in UKGTN, however, the HGNC gene name is LARGE as listed in the analysis. *UKGTN lists MIPEP, TNPO1 and MAFIP. However, these three genes refer to the alias gene name ‘MIP’ as listed in the analysis.*

UKGTN, Cataracts Disorder Genes (n=113)									
ADAMTS10	CHMP4B	CRYGB	EYA1	GALT	LARGE1*	MYH9	PEX1	RAB3GAP2	SRD5A3
AGK	COL18A1	CRYGC	FAM126A	GJA1	LTBP2	NECTIN3	PEX12	RAB3GAP1	SREBF2
AGPS	COL2A1	CRYGD	FTL	GJA3	LTBP3	NF2	PEX13	RECQL4	SC5D
ALDH18A1	COL4A1	CRYGFP	FBN1	GJA8	LRP5	NHS	PEX16	SEC23A	TNPO1*
ATOH7	CRYAA	CRYGS	FLNB	GJE1	LMX1B	OCRL	PEX2	SIX3	TDRD7
B3GLCT	CRYAB	CDKN2A	FOXC1	GJC3	MAF	OPA3	PEX26	SIX5	
BCOR	CRYBA1	CYP27A1	FOXE3	GCNT2	MAFIP*	OTX2	PEX3	SIX6	
BEST1	CRYBA4	DHCR7	FZD4	GNPAT	MIP	PAX6	PEX6	SIL1	
BFSP1	CRYBB1	EPHA2	FKTN	HMX1	MAN2B1	PITX2	PEX7	SLC16A12	
BFSP2	CRYBB2	ERCC6	FKRP	HSF4	MMP1	PITX3	POMT1	SLC2A1	
BUB1B	CRYBB3	ERCC8	FYCO1	HCCS	MIR184	PTCH1	POMT2	SLC33A1	
CAPN15	CRYGA	ESCO2	GALK1	LCT	MIPEP*	PXDN	RAB18	SOX2	

Table 7.3: *Gene list of UKGTN, eye malformations 204 gene exome panel (n=204).*

UKGTN, Eye Malformations Genes (n=204)										
ABHD12	CHST6	CRYGS	FAM111A	GCNT2	LMX1B	OTX2	PIGL	SIL1	TBX22	YAP1
ACTB	CHMP4B	CTDP1	FAM126A	GRIP1	LDLR	PAX2	PDE6D	SIX3	TBC1D20	ZIC2
ACTG1	CHRD1	CBS	FADD	GDF3	KAT6B	PAX3	PIKFYVE	SIX6	TBC1D32	ZEB1
AGK	CHD7	CRIM1	FTL	GDF6	KMT2D	PAX6	PQBP1	SCLT1	TENM3	ZEB2
ADAMTS10	C12orf57	CYP1B1	FBN1	GFER	MAB21L2	PITX2	PORCN	SLC16A12	TMX3	
ADAMTS18	CLDN19	CYP27A1	FOXC1	HMX1	MAF	PITX3	PRDM5	SLC2A1	TFAP2A	
ADAMTSL4	CC2D2A	CYP51A1	FOXE3	HSF4	MIP	PTCH1	P3H2	SLC33A1	TGFB1	
ALDH1A3	COL4A1	DHCR7	FOXL2	HMGB3	MAN2B1	PXDN	PRSS56	SLC38A8	TMEM67	
ALDH18A1	COL8A2	DHX38	FNBP4	HCCS	MFRP	PEX10	RAB18	SLC4A11	TMEM98	
AGPS	COL18A1	DCN	FREM1	IGBP1	MIR184	PEX11B	RAB3GAP2	SLC4A4	TCOF1	
ATOH7	CRYAA	DPYD	FREM2	ITPA	MYOC	PEX12	RAB3GAP1	SHH	TDRD7	
ABCB6	CRYAB	EPG5	FRAS1	ITPR1	MYH9	PEX13	RAX	SALL1	TACSTD2	
AGBL1	CRYBA1	EPHA2	FZD5	JAM3	NAA10	PEX14	RARB	SALL2	UBIAD1	
BCOR	CRYBA4	ERCC1	FYCO1	KRT12	GNPTG	PEX16	RBP4	SALL4	VAX1	
BFSP1	CRYBB1	ERCC2	GALK1	KRT3	NF2	PEX19	POLR1C	SMOC1	VIM	
BFSP2	CRYBB2	ERCC3	GALT	KERA	NHS	PEX2	POLR1D	SOX2	VSX1	
BEST1	CRYBB3	ERCC5	GJA1	LAMB2	NOTCH2	PEX26	RPGRIP1L	SRD5A3	VSX2	
B3GLCT	CRYGB	ERCC6	GJA3	LTBP2	OCRL	PEX3	SEC23A	SC5D	WDR36	
BMP4	CRYGC	ERCC8	GJA8	LCAT	OPA3	PEX5	SEMA3E	STRA6	WRN	
BMP7	CRYGD	EYA1	GSN	LIM2	OPTN	PEX6	SH3PXD2B	SMCHD1	WFS1	

Table 7.4: *Gene list of UKGTN hearing loss, syndromic and non syndromic (n=95).*

UKGTN, Hearing Loss, Syndromic and Non Syndromic Genes (n=95)									
<i>ACTG1</i>	<i>CLRN1</i>	<i>EDN3</i>	<i>GSDME</i>	<i>LRTOMT</i>	<i>MYO1A</i>	<i>PJVK</i>	<i>RDX</i>	<i>SLC4A11</i>	<i>USH1C</i>
<i>ADGRV1</i>	<i>CLDN14</i>	<i>EDNRB</i>	<i>GIPC3</i>	<i>LARS2</i>	<i>MYO3A</i>	<i>PRPS1</i>	<i>RPGR</i>	<i>SOX10</i>	<i>USH1G</i>
<i>ALOXE3</i>	<i>COCH</i>	<i>ESPN</i>	<i>GRXCR1</i>	<i>LHFPL5</i>	<i>MYO6</i>	<i>PNPT1</i>	<i>SERPINB6</i>	<i>STRC</i>	<i>USH2A</i>
<i>ATP2B2</i>	<i>CCDC50</i>	<i>ESRRB</i>	<i>GRHL2</i>	<i>KARS</i>	<i>MYO7A</i>	<i>KCNJ10</i>	<i>SIX1</i>	<i>TPRN</i>	<i>WHRN</i>
<i>BDP1</i>	<i>COL4A6</i>	<i>EYA1</i>	<i>HGF</i>	<i>MARVELD2</i>	<i>MYO15A</i>	<i>KCNQ4</i>	<i>SIX5</i>	<i>TECTA</i>	<i>WFS1</i>
<i>CDH23</i>	<i>COL9A2</i>	<i>EYA4</i>	<i>HARS</i>	<i>MITF</i>	<i>OTOA</i>	<i>POU3F4</i>	<i>SMPX</i>	<i>TJP2</i>	
<i>CIB2</i>	<i>CRYM</i>	<i>GPSM2</i>	<i>HARS2</i>	<i>MSRB3</i>	<i>OTOF</i>	<i>POU4F3</i>	<i>SNAI2</i>	<i>TMC1</i>	
<i>CABP2</i>	<i>DIABLO</i>	<i>GJB2</i>	<i>HSD17B4</i>	<i>MIR96</i>	<i>OTOG</i>	<i>PTPRQ</i>	<i>SLC17A8</i>	<i>TMIE</i>	
<i>CEACAM16</i>	<i>DIAPH1</i>	<i>GJB3</i>	<i>ILDR1</i>	<i>MYH14</i>	<i>PAX3</i>	<i>PCDH15</i>	<i>SLC26A4</i>	<i>TMPRSS3</i>	
<i>CLPP</i>	<i>DIAPH3</i>	<i>GJB6</i>	<i>KIT</i>	<i>MYH9</i>	<i>PDZD7</i>	<i>P2RX2</i>	<i>SLC26A5</i>	<i>TRIOBP</i>	

7.5.5 Variant prioritisation

Variants were filtered to exclude variants which were either synonymous or that exceeded 5% allele frequency in the Exome Aggregation Consortium across all populations (ExAC ALL). Variants were then identified as likely causal if they were consistent with the known inheritance pattern of the gene for the given phenotype and were: (1) known pathogenic (P) for the disease in ClinVar or were a known disease-causing mutation (DM) in HGMD; *or* (2) were ‘high impact’ consequences (stop-gains or frameshift variants [147]); *or* (3) had CADD Phred ≥ 15 [196] or a MaxEntScan $\geq |3|$ [263].

The *TYR* tri-allelic causal genotype in OCA1 involves two ‘common’ variants (p.R402Q and p.S192Y) and one rare pathogenic variant on *trans* alleles [362]. As the two common variants were part of a unique molecular pathology, these were analysed separately for the OCA and nystagmus patients.

7.6 Results

7.6.1 Quality control

7.6.1.1 Sequence

FastQC software was applied to check quality of the FastQ sequence. Quality metrics of ‘Per base sequence content’, ‘Sequence duplication levels’, ‘Per sequence GC content’ and ‘Kmer content’ failed which was not of concern as it was consistent with previous observations with the TruSight One prepared DNA sequenced on the NextSeq. Sequence duplication levels were higher across the first 10 bp of reads and k-mers were most enriched across first 6 bp of reads. These fails were likely due to the library preparation step involving transposase fragmentation which causes a bias in the positions at which reads start.

7.6.1.2 Coverage

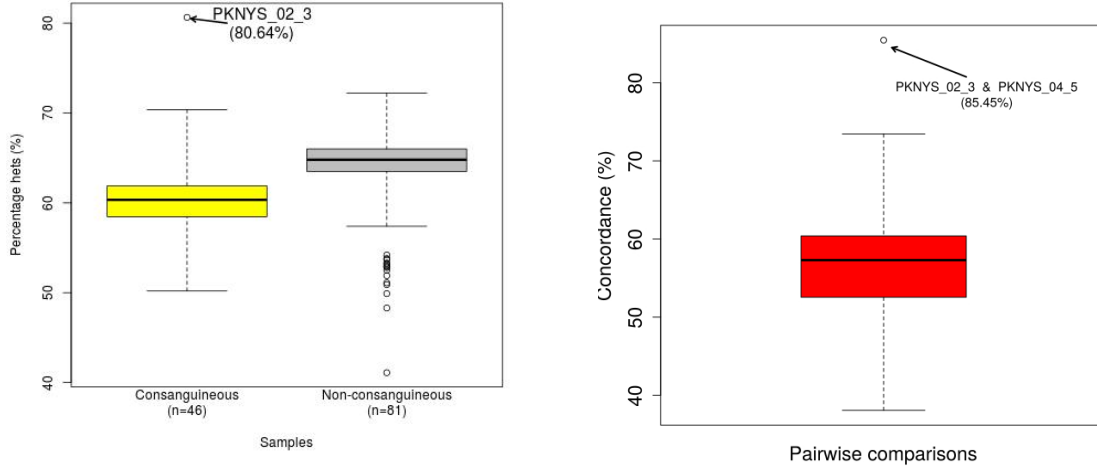
Following completion of alignment and variant calling steps, the coverage, heterozygosity and file size statistics were recorded for the 57 samples (Table 7.5). Eleven samples had $\leq 80\%$ at 20X depth (32-1, 41-1, PA1074, MIC04, NG312, NG460, NG615, NG617, NG630, NG638 and NYS42-33). These poorly covered samples also have small FastQ file sizes and lower total number of reads in comparison to the remaining samples suggesting that the poor coverage was likely due to insufficient quality or quantity of DNA.

Table 7.5: *Coverage statistics, contamination and file size quality control for the Pakistani cohort (n=57). Sample 32-1 has both the lowest coverage at 20X (2.54%) and the lowest FastQ file sizes (0.06GB). PKNYS_02_3 has the highest number of variants (13,262). PKNYS_04_5 X chromosome heterozygosity predicted genders do not match the recorded gender for these patients. Orange coloured samples indicate samples which are flagged as having insufficient coverage or elevated heterozygosity. Yellow coloured samples highlight the samples which have mismatching predicted and recorded genders. Elevated heterozygosity levels are highlighted in bold.*

Row no.	Sample ID	Coverage of TruSight One target region				Contamination					File sizes			
		Total reads	Mapped to target reads +/- 150bp	Mean depth	Target bases 20x	No. variants	Het variants	% Het variants	X variants	X het variants	% Xhets	Predicted gender	Recorded gender	Freeemx VerifyBamID
1	30-2	21421570	18615754	86.90	107.81	91.22	6678	62.65	119	11	9.24	Male	Male	0.00000
2	32-1	875307	804750	91.94	4.8	2.54	641	40.06	18	4	22.22	Male	Male	0.00000
3	38-1	19289452	17955093	93.08	103.79	90.59	5651	56.58	132	11	8.33	Male	Male	0.00000
4	39-1	16849291	16388527	97.27	91.35	89.39	6241	60.97	124	16	12.90	Male	Male	0.00000
5	40-1	27678729	25495864	92.11	149.56	91.76	10799	69.71	121	15	12.40	Male	Male	0.00000
6	41-1	6337882	5946307	93.82	34.23	73.74	8515	49.58	58.23	173	114	Female	Female	0.00000
7	CBLN12	8589622	8248146	96.02	48.55	82.49	8619	46.80	54.30	115	9	Male	Male	0.00000
8	PAI011	19257351	18260758	94.82	105.20	90.27	10028	62.06	61.89	174	100	Female	Female	0.00000
9	PAI031	28503904	25386643	89.06	146.31	92.29	10932	66.80	61.12	140	11	Male	Male	0.00000
10	PAI047	22484672	20822117	92.61	120.04	91.32	10149	56.35	55.52	126	14	Male	Male	0.00000
11	PAI055	25420297	22033830	86.68	126.06	91.93	10464	60.30	57.63	138	19	Male	Male	0.00000
12	PAI059	29929943	25717491	85.93	148.74	92.15	11168	71.90	64.38	215	145	Female	Female	0.00000
13	PAI074	6901345	6462583	93.64	76.07	76.07	8396	47.06	56.05	133	14	Male	Male	0.00000
14	PAI079	21806799	20095679	92.15	115.39	91.21	10419	64.74	62.14	175	30	Male	Male	0.00000
15	PAI084	16642112	15154340	91.06	87.89	89.81	10206	61.91	60.66	127	7	Male	Male	0.00000
16	PKNYS_01_4	30634503	26066522	85.09	151.32	92.15	11235	70.31	62.58	186	105	Female	Female	0.00000
17	PKNYS_02_3	23842974	20703372	86.83	121.54	91.69	13262	106.94	80.64	196	127	Female	Male	0.00086
18	PKNYS_04_5	28548468	24451490	85.65	141.22	91.98	10825	65.70	60.69	165	62	Male	Female	0.00000
19	PKNYS_05_2	24149689	20863423	84.32	115.75	91.9	10805	68.18	63.10	143	22	Male	Male	0.00000
20	PKNYS_07_4	42449925	35767227	84.26	203.92	93.03	11436	72.46	63.36	151	17	Male	Male	0.00000
21	PKNYS_08_3	15738498	13994962	88.92	80.61	89.81	10150	59.81	59.81	122	24	Male	Male	0.00000
22	PKNYS_09_14	18259163	15184173	83.16	87.32	90.50	10360	60.71	61.50	151	20	Male	Male	0.00000
23	HC04	11520413	9685228	84.07	58.58	87.41	8817	5300	60.11	108	7	Male	Male	0.00000
24	MI04	6213819	5376469	86.52	32.88	71.82	7294	4051	55.54	93	9	Male	Male	0.00000
25	NG312	5789314	5233049	90.39	32.87	71.02	6915	4059	58.70	146	84	Female	Female	0.00000
26	NG364	13851910	11998835	86.62	72.91	90.22	9031	5480	60.68	102	8	Male	Male	0.00000
27	NG377	12353552	10770149	87.17	66.02	89.11	8799	5202	59.12	178	93	Female	Female	0.00000
28	NG426	11388761	10163620	89.16	62.05	87.71	8775	5297	60.36	106	19	Male	Male	0.00000
29	NG431	14768982	12917273	87.46	80.05	91.01	9356	5657	60.46	117	12	Male	Male	0.00000
30	NG444	11584804	10418341	89.93	65.2	87.86	8602	5331	61.97	155	88	Female	Female	0.00000
31	NG457	8392783	7390865	88.34	47.05	82.4	8028	4693	58.46	113	9	Male	Male	0.00000
32	NG460	7195542	6347215	88.21	39.02	77.53	7386	4312	58.38	105	11	Male	Male	0.00000
33	NG463	22126556	18800529	84.97	114.28	93.73	10241	6500	63.47	197	137	Female	Female	0.00000
34	NG466	13122184	11414414	86.99	69.83	89.08	9174	5467	59.59	170	102	Female	Female	0.00000
35	NG470	9861110	8539040	88.16	52.74	84.97	8428	5129	60.86	147	86	Female	Female	0.00000
36	NG485	7205243	6532960	90.95	39.85	78.25	7785	4606	59.17	92	17	Male	Male	0.00000
37	NG487	4942679	4622946	93.53	29.01	65.25	6408	3657	57.07	75	21	Male	Male	0.00000
38	NG630	4469944	4208624	94.15	25.97	60.62	5987	3233	54.00	79	6	Male	Male	0.00000
39	NG633	7326887	6997177	92.96	42.55	80.15	7446	4313	57.92	82	4	Male	Male	0.00000
40	NG636	7762694	7207994	92.85	44.35	81.15	7659	4670	60.97	89	9	Male	Male	0.00002
41	NG638	7055069	6411899	90.84	38.75	77.78	6945	3979	57.16	131	70	Female	Female	0.00000
42	NY204	14316675	12772541	89.21	79.78	90.64	9239	5479	59.30	126	14	Male	Male	0.00057
43	NY504	12215221	10733105	87.87	65.64	88.86	9175	5596	60.99	176	104	Female	Female	0.00000
44	NY838-15	12367742	10876872	87.95	66.6	88.87	9472	6665	70.37	133	66	Male	Male	0.00276
45	NY841-6	10164584	88.97	55.45	86.06	8036	4741	59.00	101	5	4.95	Male	Male	0.00000
46	NY842-33	5850519	4836537	82.67	30.5	62.74	6603	3691	55.90	86	5	Male	Male	0.00000
47	PA1374	8743036	7630790	87.28	47.77	81.87	8333	5007	60.09	98	10	Male	Male	0.00000
48	PA1400	12058133	10922146	90.58	67.04	88.09	8658	5168	59.69	172	120	Female	Female	0.00000
49	PA1403	11069632	9025237	89.39	56.09	84.24	8199	4849	59.14	102	13	Male	Male	0.00000
50	PA1410	13028977	11170870	85.74	87.31	87.13	8713	5150	59.11	113	11	Male	Male	0.00000
51	PA1434	10792746	9897600	91.71	61.91	86.12	8386	5058	60.31	138	74	Female	Female	0.00000
52	PA1444	11421747	10472130	91.69	64.41	87.41	8175	4587	56.11	111	11	Male	Male	0.00000
53	PA1459	11464619	10563008	92.14	65.67	87.4	8436	4759	56.41	126	15	Male	Male	0.00000
54	PA1483	8180296	7056777	86.27	43.18	80.87	7756	4171	53.78	169	102	Female	Female	0.00000
55	US03	11507860	10227969	88.88	62.71	87.44	8634	4689	54.31	124	12	Male	Male	0.00000
56	V-2	11455221	10106621	88.23	61.97	88.06	8601	4678	54.39	126	10	Male	Male	0.00000
57	V1-14	10435971	9107494	87.27	55.38	85.96	8321	4176	50.19	104	10	Male	Male	0.00000

7.6.1.3 Contamination

Excessive heterozygosity may be indicative of cross-contamination between samples during sample processing. The median percentage of total variants which were heterozygous calls was 59.59%. Sample PKNYS_02_3 was found to have an excess of heterozygous variant calls (80.64%) (Figure 7.3a). PKNYS_02_3 has 64.80% of X chromosome variants identified as heterozygous inconsistent with a clinically recorded gender of male. It also showed a high level (85.45%) of variant concordance with sample PKNYS4_5 (Figure 7.3b). Evidence points towards PKNYS_02_3 being contaminated with PKNYS_04_5. Sample NYS38-15 had high (70.37%) total heterozygosity and had the highest VerifyBamID freemix value (0.00276) which was suggestive of some degree of contamination.



(a) *Heterozygosity for 46 samples (with sufficient coverage) across all chromosomes and compared with non-consanguineous $n=81$ samples from Chapter 6. Summary statistics of consanguineous samples: mean=60.24%, median=60.33%, SD=4.58, range of 50.19% to 80.63%. PKNYS_02_3 is shown as an upper extreme outlier. Summary statistics of non-consanguineous samples: mean=63.47%, median=64.80%, SD=4.73, range of 41.12% to 72.20%. Comparison between consanguineous and non-consanguineous cohorts confirm expectations that the average % heterozygosity for consanguineous samples will be lower than non-consanguineous samples.*

(b) *Pairwise concordance of variants between all 46 samples with sufficient coverage ($n=2116$ comparisons). Summary statistics: mean=56.32%, median=57.30%, SD=5.60, range of 38.07% to 85.45%. Only one pairwise sample concordance % is shown as an extreme outlier (PKNYS_02_3 & PKNYS_04_5)*

Figure 7.3: *Box plots of heterozygosity and variant concordance between all samples with exception to those with coverage at $20\times < 80\%$ (11 samples).*

The predicted gender of any patient is based on X chromosome heterozygosity with lower levels ($<50\%$) indicative of males and higher levels indicative of females. PKNYS_04_5 had an intermediate 37.58% of X chromosome variants identified as heterozygous and had a clinically recorded gender of female. The resolution of the ROH analysis was not able to detect runs of homozygosity (ROH) in the X chromosome of this sample (Appendix Table F.1) to explain that the lower levels of X chromosome heterozygosity were due to a high level of consanguinity. However, since this sample does not appear

to be contaminated and is well covered, consanguinity appears to be the most plausible reason for their elevated heterozygosity.

7.6.1.4 Quality control summary

Overall, 13 samples were identified to be compromised in quality. 11 samples (32-1, 41-1, PA1074, NG312, NG460, NG615, NG617, NG630, NG638, MIC04 and NYS42-33) were poorly covered across clinical exome target regions and more susceptible to false negatives (type II errors). Sample PKNYS_02_3 was identified to be most likely contaminated with PKNYS_04_5, and NYS38-15 was identified to have elevated heterozygosity levels. Therefore, variants in PKNYS_02_3 and NYS38-15 should be interpreted with caution. These 13 samples were not omitted from the analysis but their compromised data were flagged and considered when scrutinising the potential diagnostic yield.

7.6.2 Causal variant analysis

7.6.2.1 Runs of homozygosity

Runs of homozygosity (ROH > 1,000,000 bp) were identified for 24 of the 57 patients (Appendix Table F.1). Variants which were rare (ExAC_all < 5%) and not synonymous were cross-checked against the ROH which identified three variants within the designated gene panel of three patients (Table 7.6). Shortlisted variants which were found in the intersection of detected ROH supported variant causality of the disease due to the effects of consanguinity. No intersection (between variants and ROH) was identified for the two patients allocated the eye malformations 204 gene exome panel and two patients allocated the hearing loss, syndromic and non syndromic 95 gene exome panel.

Table 7.6: *OCA and cataract sub-cohort filtered variant positions which were found within runs of homozygosity (ROH). Positions are listed in order of patient ID and karyotypically.*

Patient ID	Chrom	Position	UKGTN Gene Panel	Variant identified
PA1059	15	27755447	Albinism and nystagmus	<i>OCA2</i> :S820P
PA1400	1	235766296	Albinism and nystagmus	<i>LYST</i> :c.5923-19G>C
CBLN112	8	144513198	Cataract	<i>RECQL4</i> :c2463+17_2463+19delCC

7.6.2.2 Albinism and nystagmus

For the 47 albinism/nystagmus patients assessed, 294 variants were found in the UKGTN 31 gene panel of ‘albinism and nystagmus’. Excluding synonymous variants reduced the variant list to 202 and further filtering to exclude common variants (ExAC_all $\geq$ 5%) resulted in 124 remaining variants (Table 7.11).

Twenty-six patients of 47 (53.2%) patients with albinism/nystagmus had a total of 35 likely causal variants identified across the UKGTN ‘nystagmus and albinism’ 31 gene panel. The most common genes with likely causal genotypes identified were *OCA2* (12), *TYR* (11), *CACNA1A* (3), *HPS3* (3) and *FRMD7* (2), *MITF* (1), *DTNBP1* (1), *SACS* (1) and *CACNA1F* (1). Ten candidate causal variants were novel and implicated as likely novel causes of nystagmus and OCA. These novel variants were identified across five genes: *TYR* (NM_000372:exon1:c.G629A:p.W210*, NM_000372:exon2:c.1000delA:p.K334fs and NM_000372:exon2:c.C1004G:p.A335G) *OCA2* (NM_000275:exon24:c.T2458C:p.S820P, NM_000275:exon4:c.408_409del:p.R136fs and NM_000275:exon3:c.G286A:p.E96K), *HPS1* (NM_001311345:exon13:c.425+1G>A and NM_182639:exon6:c.G437A:p.W146*), *CACNA1A* (NM_001127221:exon18:c.2282+1G>C) and *FRMD7* (NM_194277:exon6:c.T443A:p.L148*).

Patient PA1059 was previously shown to have ROH which intersected with the variant *OCA2*:S820P. Similarly, patient PA1400 was also found to have ROH which intersected with the splicing *LYST* gene variant c.5923-19G>C.

Samples: blue background indicates male whilst pink indicates female, orange cells indicate a possible causal genotype, red indicates a likely causal genotype. Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Gene.refGene, gene symbol; OMIM Inheritance, inheritance as listed on OMIM for the relevant phenotype where applicable; Variant type, consequence of the variant; Amino acid, amino acid change; avsnp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations), 1000g ALL, 1000 Genomes Project (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; HGMD DM, name of disease with the disease-causing mutation, ClinSig, clinical significance ('p' if 'pathogenic', 'B' if 'benign');

7.6.2.3 TYR tri-allelic genotypes

Filter parameters were relaxed for the *TYR* gene in order to identify the known causal tri-allelic genotypes involving the p.R402Q, p.S192Y and a deleterious variant. [362]. All *TYR* variants are shown in Table 7.8. A further two patients (5.71%) in the cohort of nystagmus and albinism patients (NG444 and NG470) had likely causal tri-allelic *TYR* genotypes.

Table 7.8: *Two albinism/nystagmus patients (NG444 and NG470) not previously identified to have likely causal genotype were identified to have a likely causal tri-allelic genotype for albinism within TYR.*

Position	Ref	Alt	Gene.refGene	OMIM Inheritance	Variant type	Amino Acid	avsnp144	1000g ALL	ExAC ALL	SIFT score	GERP++ RS	CADD phred	CLINSIG	CLNDEN	NG444	NG470
89178528	C	A	TYR	AR	nonsynonymous	S192Y	rs1042602	0.12340	0.25180	0.031	6.07	25.0	P	OCA	HOM	HOM
89227904	C	A	TYR	AR	nonsynonymous	T373K	rs61754388	0.00030	0.00030	0.001	4.45	25.0	P	OCA	HET	
89284793	C	A	TYR	AR	nonsynonymous	R402Q	rs1126809	0.08127	0.17700	0.029	4.68	34.0	P	OCA1B	HET	HET
89284805	C	T	TYR	AR	nonsynonymous	P406L	rs104894313	0.0020	0.00350	0.004	4.68	32.0	P	OCA1B		HET

7.6.2.4 Congenital cataract

For the 6 patients with congenital cataract assessed, 354 variants found across the 113 genes from the UKGTN panel of ‘cataract disorder’. Excluding synonymous variants gave 193 variants and filtering to include only rare variants (ExAC_all<5%) left 58 remaining variants (Table 7.9).

Four patients were identified to have likely causal genotypes in *EPHA2* (heterozygous NM_004431:exon15:c.G2627A:p.R876H), *CRYGC* (heterozygous NM_020989:exon3:c.C502T:p.R168W), *SIL1* (homozygous NM_001037633:exon5:c.C274T:p.R92W) and *ATOH7* (homozygous NM_145178:exon1:c.94delG:p.A32fs) respectively. No variant in these patients was identified in HGMD as disease causing, however, one variant was annotated as pathogenic in ClinVar for cataracts (*CRYGC*:p.R168W).

Table 7.9: *Five congenital cataract and one blind patient were investigated for likely causal genotypes. Four patients were identified to have likely causal variants based on the known inheritance pattern of the candidate genes. Samples: blue background indicates male whilst pink indicates female, orange cells indicate a possible causal genotype, red indicates a likely causal genotype.*

Chrom	Position	Ref	Alt	Gene	RefGene	Omim	Inheritance	Variant type	Amino Acid	rsid	ExAC ALL	1000g ALL	SIFT	GERP++	CADD phred	MaxEnt Scan diff	CLINSIG	CLNDBN	CBLN112*	PA1011	PA1047	PA1074	PA1079	PA1084
1	16123319	C	T	EPHA2	EPHA2	AD	AD	noisy/synonymous	D943N	rs5903225	0.00779	0.017	0.000	4.88	34.0					HET	HET			
1	16130268	T	C	EPHA2	EPHA2	AD	AD	noisy/synonymous	R876H	rs12045417	0.05252	0.004	0.753	5.63	35.0					HET	HET			
1	22017111	T	C	RABGAP2	RABGAP2	AR	AR	noisy/synonymous	T863A	rs201613456	0.00499	0.007	0.308	0.67	0.0					HET	HET			
1	22018922	G	C	RABGAP2	RABGAP2	AR	AR	noisy/synonymous	L670V	rs73098379	0.05471	0.031	0.308	1.72	12.6		B			HET	HET			
1	22019185	T	C	RABGAP2	RABGAP2	AR	AR	splicing		rs72927376	0.0229	0.000	0.077	1.45	9.1					HET	HET			
2	1666298	C	T	PXDN	PXDN	AR	AR	noisy/synonymous	V48I	rs146093600	0.0002	0.000	0.000	0.000	0.000					HET	HET			
2	20812925	G	A	CRYGC	CRYGC	AD	AD	noisy/synonymous	R168W	rs29031904	0.0026	0.007	0.008	-0.342	28.6		P_Cataract			HET	HET			
2	20812950	C	T	CRYGC	CRYGC	AD	AD	noisy/synonymous	R48H	rs80751949	0.01218	0.017	0.259	1.08	22.4									
2	218814154	C	T	CYP27A1	CYP27A1	AR	AR	noisy/synonymous	P384L	rs42728072	0.08859	0.0188	0.002	5.76	33.0		P	CTX		HOM	HET			
3	58105676	C	A	FLNB	FLNB	AD	AD	splicing		rs27074072	0.01957	0.0438	0.000	0.000	0.080									HET
3	58105676	C	A	FLNB	FLNB	AD	AD	noisy/synonymous	T140I	rs754064871	0.00001	0.000	0.006	5.92	34.0									
3	58126739	C	A	FLNB	FLNB	AD	AD	noisy/synonymous	S129I	rs748875467	0.00000	0.000	0.002	2.76	33.0									
4	8871229	C	A	HMX1	HMX1	AR	AR	noisy/synonymous	R92W	rs149242794	0.00559	0.0055	0.001	2.43	27.6					HET	HET			
5	139051017	G	A	SLLI	SLLI	AR	AR	noisy/synonymous	R2W	rs876473597	0.0024	0.007	0.000	0.000	-0.177									
6	42968503	G	A	PEX6	PEX6	AR	AR	splicing	P256L	rs758465154	0.00002	0.000	0.002	4.31	22.4					HET	HET			
6	121447614	C	T	GJA1	GJA1	AR	AR	splicing		rs758465154	0.00002	0.000	0.002	4.31	22.4									
6	136872182	CTT	C	PEX7	PEX7	AR	AR	splicing		rs745199683	0.00839	0.0125	0.000	0.000	-0.411									
6	136872182	CTT	CT	PEX7	PEX7	AR	AR	splicing		rs133699212	0.0018	0.0009	0.000	0.000	0.008									
7	92507133	A	G	PEX1	PEX1	AR	AR	splicing		rs76259565	0.03215	0.0146	0.000	0.000	0.000									
8	71215369	A	G	EYA1	EYA1	AD	AD	splicing		rs747467559	0.00006	0.0006	0.000	0.000	0.000									
8	144512135	C	T	RECQL4	RECQL4	AR	AR	splicing		rs200026534	0.0008	0.007	0.353	4.06	8.5					HOM	HOM			
8	144513198	TGCG	TG	RECQL4	RECQL4	AR	AR	splicing	T1063M	rs16920147	0.03075	0.0171	0.327	3.83	11.4					HOM	HOM			
9	9742891	G	A	PTCH1	PTCH1	AD	AD	noisy/synonymous		rs12115566	0.02697	0.0002	0.340	-3.14	0.1					HOM	HOM			
9	9742891	G	A	TDRD7	TDRD7	AD	AD	splicing		rs745634412	0.00002	0.0002	0.340	-3.14	0.1									
9	9742891	G	A	TDRD7	TDRD7	AD	AD	splicing		rs16920147	0.03075	0.0171	0.327	3.83	11.4									
9	131522077	G	T	POMT1	POMT1	AR	AR	noisy/synonymous	A641V	rs12115566	0.02697	0.0002	0.340	-3.14	0.1		B			HET	HET			
10	27531584	G	A	RAB18	RAB18	AR	AR	noisy/synonymous	S81N	rs12115566	0.02697	0.0002	0.340	-3.14	0.1					HET	HET			
10	49470659	C	G	ERCC6	ERCC6	AR	AR	noisy/synonymous	V1101L	rs745634412	0.00002	0.0002	0.340	-3.14	0.1					HET	HET			
10	68231583	GC	G	ATOH7	ATOH7	AR	AR	frameshift deletion	A326	rs17801742	0.02556	0.0174	0.074	0.000	0.000					HOM	HOM			
11	45915755	C	T	PEX16	PEX16	AR	AR	noisy/synonymous	V103M	rs41317915	0.00819	0.0131	0.000	3.7	24.6					HOM	HOM			
11	61968127	A	A	BEST1	BEST1	AD	AD	splicing		rs1153094	0.01318	0.0234	0.07	5.29	22.9					HOM	HOM			
11	61968127	A	A	BEST1	BEST1	AD	AD	splicing		rs13492158	0.01318	0.0234	0.07	5.29	22.9					HOM	HOM			
11	6840671	G	A	LRP5	LRP5	AD	AD	noisy/synonymous	V667M	rs4988321	0.00479	0.0116	0.039	4.11	23.2					HET	HET			
11	102796715	CA	C	MAP1	MAP1	AR	AR	frameshift deletion	I916s	rs7879749	0.00479	0.0116	0.039	4.11	23.2					HET	HET			
12	47983086	T	C	COL2A1	COL2A1	AD	AD	splicing		rs17801742	0.02556	0.0174	0.074	0.000	0.000					HOM	HOM			
12	47987262	C	T	COL2A1	COL2A1	AD	AD	splicing		rs1317915	0.00819	0.0131	0.000	3.7	24.6					HET	HET			
13	20142287	C	A	GJA3	GJA3	AD	AD	noisy/synonymous	E334D	rs549918653	0.001	0.0004	0.966	4.18	18.6					HOM	HOM			
13	31214332	CTT	CT	B3GLCT	B3GLCT	AR	AR	splicing		rs200252122	0.00699	0.003	0.274	4.66	22.9					HOM	0/2			
13	110187193	G	A	COL4A1	COL4A1	AD	AD	splicing	A588V	rs3832900	0.06889	0.0578	0.274	4.66	22.9					HOM	0/2			
13	110211627	CAAAAT	C	COL4A1	COL4A1	AD	AD	splicing		rs8382900	0.06889	0.0578	0.274	4.66	22.9					HOM	0/2			
14	74509749	C	T	LTBP2	LTBP2	AR	AR	splicing	G1088S	rs61505039	0.0603	0.0335	0.044	3.08	29.3					HET	HET			
14	74521898	G	A	LTBP2	LTBP2	AR	AR	splicing		rs78258030	0.04892	0.0276	0.284	-0.462	13.1					HET	HET			
14	74611925	G	A	LTBP2	LTBP2	AR	AR	splicing	A7V	rs781543497	0.0004	0.0004	0.284	-0.462	13.1					HET	HET			
15	48465869	A	G	FN1	FN1	AD	AD	splicing		rs549918653	0.001	0.0004	0.966	4.18	18.6					HET	HET			
19	8589347	C	T	ADAMTS10	ADAMTS10	AR	AR	noisy/synonymous	V655I	rs732584555	0.00001	0.0004	0.966	4.18	18.6					HET	HET			
19	8605046	C	T	ADAMTS10	ADAMTS10	AR	AR	noisy/synonymous	S134T	rs732584555	0.00001	0.0004	0.966	4.18	18.6					HET	HET			
19	12652453	C	T	MAN2B1	MAN2B1	AR	AR	noisy/synonymous	R612Q	rs1347	0.0002	0.0005	0.32	3.48	21.3					HET	HET			
19	46755970	A	T	FARP	FARP	AR	AR	noisy/synonymous	S174C	rs200900647	0.0002	0.0005	0.32	3.48	21.3					HET	HET			
19	48965920	A	G	FTL	FTL	AD	AD	splicing		rs544433190	0.0002	0.0002	0.02	3.52	21.2					HET	HET			
21	45492526	C	A	COL18A1	COL18A1	AR	AR	splicing		rs200143450	0.0004	0.0004	0.000	0.000	0.000					HOM	HOM			
21	45493563	G	C	COL18A1	COL18A1	AR	AR	splicing		rs22280693	0.0631	0.0367	0.63	0.37	7.5					HOM	HOM			
21	45504511	G	-	COL18A1	COL18A1	AR	AR	nonframeshift deletion	Q780H	rs732584555	0.00001	0.00001	0.000	0.000	0.000					HOM	HOM			
21	45504338	GTCC	GTCC	COL18A1	COL18A1	AR	AR	splicing	942_944del	rs766237690	0.0176	0.0176	0.282	0.405	0					HOM	HOM			
21	45504338	A	C	COL18A1	COL18A1	AR	AR	splicing		rs201318532	0.01837	0.029	0.282	0.405	0					HOM	HOM			
22	25202716	G	A	CRYBB3	CRYBB3	AD	AD	noisy/synonymous	H1107P	rs732584555	0.00001	0.00001	0.000	0.000	0.000					HOM	HOM			
22	25227572	G	A	CRYBB2	CRYBB2	AD	AD	noisy/synonymous	E40K	rs732584555	0.00001	0.00001	0.000	0.000	0.000					HOM	HOM			
22	28616165	G	C	CRYBB2	CRYBB2	AD	AD	noisy/synonymous	A65S	rs16986560	0.02077	0.0008	0.000	0.000	0.000					HOM	HOM			
22	41877929	C	T	SREBF2	SREBF2	AR	AR	splicing	A52E	rs533765538	0.0008	0.0006	0.217	2.8	0.3					HET	HET			
X	17719324	T	C	NHS	NHS	XL	XL	splicing		rs732584555	0.01113	0.0181	0.000	0.000	-1.144					HET	HET			

7.6.2.5 Waardenburg syndrome and Usher syndrome

For the two patients assessed (HC04 and US03), 265 variants found candidate gene list genes. Exclusion of synonymous variants provided 172 variants and further filtering to include only rare variants (ExAC ALL<5%) provided 24 variants (Table 7.10).

One likely causal homozygous variant was identified for the single Usher syndrome patient in the *USH2A* gene (NM_206933:exon39:c.C7334T:p.S2445F). This variant is identified by HGMD as a disease causing variant for ophthalmic disease and has strongly supportive *in silico* scores of SIFT (0.011), GERP++ (6.03) and CADD Phred (29.9). The Waardenburg syndrome case (HC04) was identified to have a likely causal heterozygous variant in the Waardenburg syndrome type IV gene, *EDNRB* (NM_001201397: exon2:c.C512T: p.P171L).

Table 7.10: Two patients with Waardenburg syndrome and Usher syndrome have 24 variants across the UKGTN 'hearing loss, syndromic and Non syndromic, 95 Gene Panel' with synonymous and common (ExAC>5%) variants excluded. Samples: blue background indicates male whilst pink indicates female, orange cells indicate a possible causal genotype, red indicates a likely causal genotype.

Chrom	Position	Ref	Alt	Gene	OMIM	Inheritance	Variant type	Amino Acid	avsnp144	ExAC ALL	1000g ALL	SIFT	GERP++	CADD Phred	HGMD DM	CLINSIG	CLNDBN	HC04	US03
1	160042480	C	T	KCNJ10	AR	AR	nonsynonymous	R18Q	rs115466046	0.01270	0.00439	0.095	5.17	22.6					HET
1	215741542	A	-	USH2A	AR	AR	splicing		rs34565443	0.04110								HET	HET
1	215758581	T	T	USH2A	AR	AR	splicing		rs772697903	0.00230								HET	HET
1	215758582	T	-	USH2A	AR	AR	splicing		rs772697903										
1	215900872	G	A	USH2A	AR	AR	nonsynonymous	S2445F	rs41315579	0.00090	0.00080	0.011	6.03	29.9	Eye disease			HOM	HOM
3	46709584	AAGAAGAAG	-	TMIE	AR	AR	nonframeshift del	123_125del	rs538183178	0.00100	0.00140							HOM	HOM
3	46709587	AAGAAG	-	TMIE	AR	AR	nonframeshift del	124_125del										HOM	HOM
3	46709590	AAG	-	TMIE	AR	AR	nonframeshift del	125_125del										HOM	HOM
6	75844964	G	A	MYO6	AR	AR	nonsynonymous	R295H	rs771731646	0.00002		0.039	5.27	27.5				HOM	HOM
7	81705746	C	T	HGF	-	-	nonsynonymous	V584I	rs538415452	0.00150	0.00220	0.528	-3.76	8.201				HET	HET
8	101558749	A	C	GRHL2	-	-	nonsynonymous	K205N				0.092	12.84	2.6				HET	
8	101558749	G	C	GRHL2	-	-	nonsynonymous	D206H				0.044	24.1	5.3				HET	
9	69205232	G	A	TJP2	AR	AR	nonsynonymous	R24H	rs4493966	0.03820	0.06230	0.346	-2.56	11.2				HET	
9	72754805	A	G	TMC1			nonsynonymous	Y221C	rs533159785	0.00001	0.00020	0.045	6.01	27.3				HET	
9	72792243	T	C	TMC1			nonsynonymous	M486T	rs17058153	0.02490	0.01558	0.248	6.02	13.9				HET	
10	26174123	C	A	MYO3A	AR	AR	nonsynonymous	P128T	rs35575696	0.00990	0.00399	0.121	1.79	7.9				HET	
10	53866871	A	-	PCDH15	AR	AR	splicing		rs530072775	0.00330	0.00359							HET	
10	53866871	-	A	PCDH15	AR	AR	splicing			0.00330								HET	
10	71724114	G	A	CDH23	AR	AR	splicing		rs752929224	0.00005								HET	HET
11	110279763	A	-	RDX	AR	AR	splicing											HOM	HOM
11	110279762	AA	-	RDX	AR	AR	splicing											HOM	HOM
11	121125534	C	T	TECTA			nonsynonymous	P479L	rs35107075	0.00160	0.00319	0.406	5.02	18.6				HET	
13	77918332	G	A	EDNRB	AD	AD	nonsynonymous	P171L	rs756427784	0.00001		0.256	-0.652	15.0				HET	HET
17	18148133	C	T	MYO15A	AR	AR	nonsynonymous	T2205I	rs121908970	0.00420	0.00100	0.002	3.11	23.9		P	Deafness	HET	

7.6.2.6 Eye malformations

For MIC04 and NYS41-6 341 variants found in across the eye malformation gene panel of 204 genes. Excluding synonymous variants gave 191 variants and further filtering to include only rare variants (ExAC_all<5%) resulted in 32 remaining variants. Likely causal variants were given only a ‘possible causal’ status as the gene panel is broad and non-specific to the phenotype.

The microcornea patient (MIC04) had a possible causal heterozygous missense variant in the *CRYGC* gene. The Joubert syndrome patient (NYS41-6) had two possible causal genotypes, a heterozygous missense variant in the *PIKFYVE* and one homozygous splicing variant in the *FRAS1* gene. No relevant pathogenic ClinVar variants or any HGMD DM variants were detected.

Table 7.11: Two patients with microcornea and Joubert syndrome were investigated for likely causal genotypes. No patients were identified to have likely causal variants as the genes did not appear to be specific to the phenotype of the patients. Samples: blue background indicates male and orange cells indicate a possible causal genotype.

Chrom	Position	Ref	Alt	Gene refGene	Omim	Inheritance	Variant type	Amino Acid	avsnp144	ExAC ALL	1000g ALL	SIFT	GERP++	CADD phred	MaxEnt Scan diff	CLINSIG	CLNDEN	MIC04	NYS41-6
1	97883329	A	G	DPYD	AR	AR	nonsynonymous	C29R	rs1801265	0.01538	0.0392		5.84		0.407	P	DPD deficiency	HET	HET
1	220202386	A	G	RAB3GAP2	AR	AR	splicing		rs76473498	0.01538	0.0392		5.84		0.407			HET	HET
2	38071060	G	C	CYP11B1	AR	AR	nonsynonymous	L432V	rs1056836	0.05771	0.0493		5.95		-0.765	B		HET	HET
2	135162538	C	T	RAB3GAP1	AR	AR	splicing		rs56786863	0.0014	0.0009	0.13	3.18	13.94		B		HET	HET
2	135168635	C	G	RAB3GAP1	AR	AR	nonsynonymous	P941A	rs77535003	0.0014	0.0009	0.13	3.18	13.94				HET	HET
2	208129550	C	T	CRYGC	AD	AD	nonsynonymous	R48H	rs61751949	0.01218	0.0174	0.259	1.08	22.40				HET	HET
2	208320333	A	G	PIKFYVE	AD	AD	nonsynonymous	T722A	rs566701537	0.0006	0.0006	0.453	4.87	22.90				HET	HET
2	219214942	G	A	ABCB6	AD	AD	splicing		rs552307539	0.0004	0.0002							HET	HET
3	4663210	G	A	ITPR1	AD	AD	splicing		rs117503975	0.0004	0.0001	0.055	1.6	23.20	-1.414			HET	HET
3	49123113	C	T	LAMB2	AR	AR	splicing		rs114913744	0.00899	0.0115						HOM	HOM	HET
4	6292013	G	A	WFS1	AD	AD	splicing		rs71524367	0.0012	0.0047							HET	HET
4	71547736	A	G	SLC4A4	AR	AR	splicing		rs749179286	0.0014	0.0009						HOM	HOM	HOM
4	78499716	C	A	FRAS1	AR	AR	splicing		rs7695038	0.0014	0.001	0.055	1.6	23.20	2.266			HOM	HOM
4	120818512	T	A	PRDM5	AR	AR	nonsynonymous	K164I	rs146268537	0.00379	0.0104	0.055	1.6	23.20				HET	HET
8	31064454	A	T	WRN	AR	AR	splicing		rs4987239	0.03614	0.0496							HET	HET
8	71215369	A	G	EYA1	AD	AD	splicing		rs76259565	0.03215	0.0146							HET	HET
9	14775919	T	A	FREM1	AD	AD	nonsynonymous	N1576I	rs2101770	0.00519	0.0071	0.167	-5.73	2.54				HET	HET
10	49500592	G	A	ERCC6	AR	AR	nonsynonymous	A544V	rs76663084	0.00479	0.0006	0.06	5.87	24.70	-0.181			HET	HET
10	74972495	A	G	KAT6B	AD	AD	splicing			0.00479	0.0069	0.06	5.87	24.70				HET	HET
10	95626740	G	T	ALDH18A1	AR	AR	nonsynonymous	S372Y	rs3765571	0.02037	0.0172	0.038	5.93	24.80				HET	HET
11	45915755	C	T	PEX16	AR	AR	nonsynonymous	V103M	rs11553094	0.01318	0.0234	0.07	5.29	22.90				HET	HET
12	7202673	T	C	PEX5	AR	AR	nonsynonymous	M272T	rs76708142	0.01777	0.0088	1	4.47	11.28				HET	HET
12	52792705	C	T	KRT3	AD	AD	splicing								0.668			HET	HET
13	31215034	T	-	B3GLCT	AR	AR	splicing											HET	HET
14	74509379	G	A	LTBP2	AR	AR	splicing		rs117613718	0.00479	0.0127				-0.661			HET	HET
15	74182333	T	-	STR6	AR	AR	splicing		rs377631858	0.003	0.0012							HOM	HOM
16	77341684	C	G	ADAMTS18	AR	AR	splicing		rs17620357	0.02536	0.037							HET	HET
19	8005046	C	G	ADAMTS10	AR	AR	nonsynonymous	S134T	rs7255721	0.06769	0.0467	1	5.24					HET	HET
19	11111624	G	A	LDLR	AD	AD	nonsynonymous	A391T	rs11669576	0.06769	0.0467	1	-0.83	1.68				HET	HET
19	45357323	C	T	ERCC2	AR	AR	nonsynonymous	V476I	rs531021258	0.0018	0.0012	0.099	5.27	23.80				HET	HET
21	45491218	C	T	COL18A1	AR	AR	splicing		rs369084150	0.0018	0.0003	0.099	5.27	23.80	0.015			HET	HET
22	36304168	A	G	MYH9	AD	AD	splicing								0.04				HET

7.6.2.7 Potential diagnostic yield

Overall, we demonstrate a potential genetic diagnostic yield of 56.1% across a cohort of 57 consanguineous Pakistani patients with various congenital ophthalmic diseases (Table 7.12). Congenital cataract (80.0%), Waardenburg syndrome & Usher syndrome (100%) and albinism/nystagmus (55.3%) had particularly high potential diagnostic yields. The analysis was unsuccessful in determining likely causal variants for eye malformation patients and the patient diagnosed with blindness.

As identified during quality control, 11 of the 57 samples had poor coverage across the TruSight One capture kit and one nystagmus sample (PKNYS_02_3) had substantial contamination with another sample from the same batch. However, the identified likely causal genotype (a heterozygous *CACNA1A* E731A missense variant) was not found in the contaminating sample (PKNYS_04_5). Similarly, sample NYS38-15 was also identified as being likely contaminated, however, the likely causal variant in *CACNA1F* was also not identified in any other sample in the cohort. Therefore, this was suggestive that the likely causal variants in these two patients flagged as contaminated were likely to be true variants. Of the 11 samples with poor coverage, four nystagmus and albinism patients (PKNYS_04_5, NG460, NG615 and NYS42-33) and the microcornea patient (MIC04) did not have a likely causal variant identified. When recalculating the potential diagnostic yield generated from this analysis with the contaminated samples omitted in addition to the five poorly covered samples without likely causal variants identified, an adjusted diagnostic yield of 60.1% is generated for the cohort.

Table 7.12: **Potential diagnostic yield overview.** Samples with likely causal genotypes and % diagnostic yield is summarised for each phenotype. The potential diagnostic yield is adjusted after omitting the substantially contaminated sample (nystagmus/albinism patient PKNYS_02_3) and any samples with insufficient coverage at 20X (<80%) for which a likely causal variant was not identified in. The adjusted diagnostic yield did not consider nystagmus/albinism patients PKNYS_04_5, NG460, NG615 and NYS42-33, and the eye malformation (microcornea) patient MIC04.

Phenotype	All samples			Samples with compromised quality and no positive result omitted		
	No. Samples	Samples with a likely causal variant(s)	Potential diagnostic yield (%)	No. samples with omissions	Adjusted no. samples with a likely causal variant(s)	Adjusted potential diagnostic yield (%)
Albinism/ nystagmus	47	26	55.3	42	25	59.5
Cataract	5	4	80.0	5	4	80.0
Eye malformations	2	0	0.0	1	0	0.0
Waardenburg syndrome & Usher syndrome	2	2	100.0	2	2	100.0
Blindness	1	0	0.0	1	0	0.0
All	57	33	56.1	51	31	60.1

7.7 Discussion

Overall, 60.1% of patients (with uncompromised samples) of the Pakistani cohort had their genetic basis of disease identified. This translated to 59.5% of albinism/ nystagmus patients, 80% of cataract patients and both the Waardenburg syndrome and Usher syndrome patients. The analysis was however unsuccessful for the Joubert syndrome, microcornea and blindness patients.

Fifty of the 124 variants in the Pakistani albinism and nystagmus cohort (which were filtered to remove synonymous and common variants) were previously identified in the UKGTN cohort of 81 patients of (assumed) Caucasian ethnicity (Chapter 6). Five of these variants met previous criteria for being likely causal (*TYR*:p.P406L, *TYR*:p.G446S, *OCA2*:p.V443I, *SACS*:p.P3678A and *CACNA1A*:p.E731A). Of the 33 identified likely causal variants in the Pakistani cohort, 23 variants were unique to the Pakistani cohort when comparing with the Caucasian cohort of 81 albinism/nystagmus patients in Chapter 6: *OCA2* (p.S820P, p.A787V, p.S736L, p.R588W, p.D486Y, c.1045-15T>G, p.R136fs, p.E96K), *TYR* (p.S44R, p.W210*, p.R278*, p.K334fs, p.A335G, c.1037-7T>A, p.G419R), *HPS1* (c.425+1G>A, p.P324fs, p.W146*), *MITF* (p.G484E), *SACS* (p.I2436L), *CACNA1A* (c.2282+1G>C), *CACNA1F* (p.V752M) and *FRMD7*

(p.L148*). Interestingly, the *TYR* p.R278* and p.G419R alleles were previously identified as amongst the most common causes of OCA in Pakistanis [425]. The GnomAD database shows that p.R278* is substantially more common in the South Asian population (0.00127) compared with any other population including the Non-Finnish European (0.00002) [152]. Similarly, p.G419R is seen at a frequency of 0.00039 in South Asians compared with 0.00003 in Non-Finnish Europeans [152]. It is possible that this is due to a common founder effect in Pakistanis, similar to what has been found in other autosomal recessive diseases such as EAST syndrome [426] and Werner syndrome [427]. The most common causes of nystagmus and albinism in this study were the *OCA2* variant p.D486Y (likely causal in three patients) followed by *TYR*:p.G419R (two patients) and *CACNA1A*:p.E731A (two patients). It was previously known that the p.D486Y and c.1045-15 T>G alleles of *OCA2* were amongst the most common causes of OCA in other Pakistani families [425] along with the *TYR* variant p.C35R which was not identified in this analysis.

The ROH results proved to compliment the identification of likely causal homozygous missense in a patient with OCA (PA1059). Although intersecting the ROH results with the shortlisted variants for patient PA1400 did not prove to be decisive when inferring causality (as the identified splicing variant had a low MaxEnt Scan score of -1.49), it continues to be a highly useful method to highlight regions harbouring likely causal genes in albinism patients [374] but may be more useful with WGS data. The ROH results for patient CBLN112 intersected with a homozygous splicing variant in the *RECQL4* gene. This variants did not have a MaxEntScan score and all patients in the cataract sub-cohort were found to have this variant. Therefore, it is unlikely that this variant is pathogenic and only a possible causal genotype status was given. Although ROH analytical software results could help streamline the analysis to be more efficient, there was limited capability to detect ROH with the punctuate nature of the

clinical exome data [428] and a large number of potentially causal variants could potentially be missed. It should therefore not be used as a hard filter for variants alone and should be used to reinforce the rationale in identifying causality of shortlisted variants. Alternative software such as VCFtools LROH are still currently in development stages [429], whilst other software are dependent on the hg19 human reference genome annotation [143]. Comparative studies of ROH detection software have shown very similar performances [424] and lacking annotation for hg38 human reference genome restricts its utility for data aligned to this reference genome.

Whilst it has not been determined which form of OCA is most prevalent in Pakistan, *OCA2* variants are understood to be more prevalent than *TYR* in Pakistanis with albinism [425] which concurs with our study suggesting that the *OCA2* phenotype is more common. There are other examples of higher prevalences of OCA subtypes in different populations including *OCA1* being more a more common type of OCA in Caucasians than Africans, and *OCA3* has a relatively high incidence in Africans (1:8500) whilst only one case to date is known to be Caucasian [430].

The tri-allelic genotype within *TYR* was less common in the Pakistani cohort (two cases) compared to Chapter 6's study of 81 Caucasians (nine cases) identifying the *TYR* tri-allelic genotype. This is likely due to a far smaller frequency of the 'common' variants p.R402Q and p.S192 in the Pakistani (South Asian) population (GnomAD_SAS AF of 0.06251 and 0.1110 respectively) compared with the Non-Finnish European population (GnomAD_NFE AF of 0.2728 and 0.3640 respectively) [152]. The higher prevalence of p.R402Q and p.S192Y variants in Caucasians appears to be due to convergent evolution of a lighter pigmentation phenotype in Europeans [431].

There were eight likely causal variants in the Pakistani cohort which were also found in the n=81 cohort of Caucasian patients with nystagmus and albinism. These included

three *TYR* variants (p.R77Q, p.G446S and p.P406L), two *OCA2* variants (p.L674V and p.V443I), a *DTNBP1* variant (p.P272S), a *CACNA1A* variant (p.P272S) and a *FRMD7* variant (p.S281L). Interestingly, four of these variants which were deemed likely causal in the Pakistani cohort were not identified as likely causal in the Caucasian cohort. The reasons were identified as either: (1) only heterozygous variants were identified without a second putative variant in the same gene (*DTNBP1*:p.P272S, *TYR*: p.R77Q and *OCA2*:p.L674V); or (2) that the allele frequency filter criteria for this analysis was less stringent (*FRMD7*:p.S281L had AF of 0.057 in 1000g).

Of the six patients with ‘nystagmus’ diagnoses, likely causal variants were identified for two patients in genes which are known to cause albinism, *TYR* and *DTNBP1*. The remaining four patients had likely causal variants identified in *CACNA1F* and *CACNA1A*. No variants were identified in the one known gene to cause idiopathic nystagmus (*FRMD7*) suggesting that all patients diagnosed with nystagmus had a disease which was not detected through phenotyping alone.

The likely causal *CRYGC* variant p.R168W (rs28931604) was a known to be causal for congenital cataract in consanguineous Mexican and Indian families [432, 433]. One likely causal variant appeared to be novel as there were no individuals identified with the *ATOH7* p.A32fs variant in the GnomAD database [152].

Joubert syndrome and many of the diseases known to cause microcornea were all included in the UKGTN panel of eye malformations. Although there were variants which met the likely causal criteria, the genes in which they were identified were not known to be directly related to the respective diseases of the two patients. Therefore, the methodology here could be improved by using alternative gene panel databases [172, 434, 435], re-phenotyping or consideration of a more complex molecular pathology for these patients. If the possible putative causal variants were verified as causal, this would po-

tentially resolve the genetic aetiologies for the eye malformation patients.

Initially, the UKGTN database allocated the phenotype of the single Waardenburg syndrome patient to the eye malformation gene panel. However, this broad gene panel did not cover Waardenburg type II and IV genes. Another search in the UKGTN database for a suitable gene panel for Waardenburg syndrome returned the ‘hearing loss, syndromic and non syndromic, 95 gene panel’. This gene panel covered all known causal genes for the four Waardenburg syndrome types. A likely causal variant was found in the Waardenburg syndrome patient in the Waardenburg type IV gene, *EDNRB* (p.P171L). Therefore, had this patient been assigned the UKGTN eye malformation gene panel, a potentially critical gene and likely causal variant would have been overlooked. This provides an example of how genetic analyses en masse (which attempt to group as many patients as possible under a common candidate gene list) or by utilising obsolete gene panels may overlook important variants in genes which are critical in the molecular pathology of the patient’s disease. The gene panel also proved effective in identifying a likely causal genotype in the Usher syndrome patient in the *USH2A* gene which is known to cause Usher syndrome type II.

For the cohort of consanguineous Pakistani patients (not compromised by poor quality) 40.7% of variants called were homozygous. This was comparably more than the non-consanguineous Caucasian cohort studied in Chapter 6 (35.22% homozygous variants) which was expected due to a larger portion of the genome expected to be autozygous. Sixteen nystagmus/albinism Pakistani patients had homozygous likely causal variants, which accounted for 61.5% of the 26 total patients with likely causal variants identified. This was also comparably higher than the non-consanguineous cohort studied in Chapter 6 which saw only one of 30 nystagmus/albinism patients with a homozygous likely causal variant (0.03%). This was a striking difference in homozygous likely

causal genotypes between the two cohorts which used similar phenotyping methods and identical library preparation, NGS and informatic processing methods. Therefore, the autozygous regions are likely to be key in influencing the elevation of homozygous likely causal variants, especially where family history of the recessive disease is positive in the consanguineous patient.

It was previously stated by Gronskov *et al* that the majority of cases of OCA were attributed to compound heterozygous genotypes [25]. In this cohort of unrelated consanguineous Pakistani patients we have identified the contrary with the majority of OCA cases caused by homozygous genetic variants. Due to the high prevalence of consanguinity as a cultural norm in the Pakistani population [418, 436], the high number of likely causal genotypes in a homozygous (or autozygous) state was also expected [437]. Consanguinity will continue to be a very important consideration of genetic analysis when identifying both causal genes and variants in rare diseases such as albinism [438].

7.8 Conclusion

In summary, 60.1% of patients had the likely underlying genetic cause of disease determined. The UKGTN gene panels, which were primarily designed for UK patients, proved to be highly effective for Pakistani patients with ocular disease. The anticipation of homozygous variants in consanguineous patients which have recessive disorders will continue to be a useful factor to consider when performing future genetic analyses and identifying causal variants.

8 Conclusions

This thesis describes six projects which have commonly involved the processing, analysis and interpretation of targeted NGS data for patients with nystagmus, albinism and POAG.

We show that the nystagmus and albinism gene panels utilised for Caucasians are applicable in other populations world-wide. For Caucasian patients with nystagmus and/or albinism, high diagnostic yields were largely attributed to presumed genetic compound heterozygotes. Without previous functional evidence suggestive that a pathogenic compound heterozygous genotype with common variation in *TYR* was likely causal for OCA, common variants in OCA patients would not have been taken into account when identifying like causal genotypes. Therefore, more functional work is critical to help uncover other similar complex genotypes in nystagmus and albinism. The work performed here will be informative to both genomic analysts and clinicians when forming future diagnostic workflows. These results demonstrate that for carefully selected patients and through the use of selected gene panels with stringent filter criteria, whilst also considering known complex causal variants, a high molecular diagnostic yield can be realised.

This thesis also investigated causal variants in known POAG genes across a homogeneous selection of severe POAG patients. POAG is clearly highly complex at both the phenotypic and genotypic level and only a small proportion of cases were able to be convincingly resolved through identification of rare variants of high predicted pathogenicity in known Mendelian-like POAG genes. This argues that individuals should be carefully reviewed on an individual basis before genetic screening is performed. Those presenting with a strong dominant family history of POAG could undergo genetic screening in the coding region of Mendelian-like genes such as exon 3 of *MYOC*.

Although potentially causal variants were identified in complex POAG genes (such as *ABO*, *SIX6*, *FAM27E5*, *TLR4*, *ABCA1*, *NOS3*, *OPTC*, *SRBD1*, *MMP9*, and *RP-GRIP1*), more functional studies are required to test the causal hypotheses generated by this work. Future expansion of this work could perform targetted genotyping across a larger, more phenotypically heterozygous glaucoma cohort. From this, risk models incorporating possible pathogenic variant status can be developed and allow individuals at increased risk of developing POAG to be identified, as was previously shown for the complex POAG gene, *TMC01* [278, 439, 440].

The analyses in this thesis were highly reliant on the patient selection and the phenotypes which were assigned to each patient. An unclear definition of what is ‘normal’, ‘abnormal’ or ‘equivocal’ may vary between clinicians, hospital sites and countries which will inevitably have an negative impact on the genetic analyses performed. The sharing of clinical systems between hospitals (in both primary and tertiary care) would allow patient phenotypic progress to be monitored and to minimise suboptimal management of patient records. The introduction of Human Phenotype Ontology (HPO) terminology can help maintain consistency between clinical diagnostics to standardise phenotypic descriptions. From this, ambiguous and inconsistent phenotype terminology which can be problematic in downstream analyses can be reduced.

Sequencing of patient DNA and the tools to process and annotate these data are imperfect and are continuously being refined and improved. Short-read lengths issues are able to be largely resolved by implementing long-read sequencing and 10x linked-read technology. Although having a greater financial cost, these types of data allow for more accurate alignments, superior quality of non-coding region data, phasing of variants and a better detection of structural variants. Library preparation kits, alignment and variant calling software continue to be fine-tuned and will undoubtedly help uncover

variants which may have been missed in this work.

To date, there has been modest annotation and understanding of the vast number of variants detected across the non-coding region of the human genome. This presents as a clear limitation for analysts when prioritising likely pathogenic variants. For this reason, the promoter, intronic, enhancer and intergenic regions were largely not captured in this work. Similarly, synonymous variants have mostly been de-prioritised due to a substantially higher quantity of them compared with variants of more likely damaging consequence. There is also currently an annotation bias towards Caucasian populations. This bias in annotation databases may have impacted the ability to identify likely causal variants in the Pakistani cohort.

Whilst the sequencing, processing and annotation of NGS data and variants can be improved, the interpretation of each patient's variants are typically inconsistent between analysts and is a step which remains a bottleneck in NGS analyses. Every patient presents with a unique background and possibly pathogenic variants can have different effects on different patients. Genetic modifiers interacting with disease genes epistatically, environmental exposures, diet, BMI, ethnicity and gender all make each patient a unique case when inferring causality of genetic variants.

Furthermore, variants of uncertain significance (VUS) are typically ignored as benign variants, however, they may have clinical relevance. An example of this is the common *TYR* variants (p.R402Q and p.S192Y), which although by ACMG guidelines are classified as VUS, prove to be key in resolving many OCA cases detailed throughout this thesis. Misclassification of variants as VUS compromises diagnostic inference [441, 442]. There remains much debate on how VUS or secondary variant findings should be reported [443]. Therefore, it is clear that the experienced clinicians and geneticists remain critical in overseeing the determination of causal variants in patients.

NGS data are currently becoming more available at a whole-genome level and organisations such as Genomics England (which has now completed the 100,000 Genomes Project) has allowed UK researchers in Genomics England Clinical Interpretation Partnership (GeCIP) access to whole-genome data to reinvestigate the non-coding region of genomes. As research into these regions becomes more prominent, there is a positive future outlook for a better understanding of variants detected in the non-coding region in disease.

Whilst DNA-seq analyses prove to be highly beneficial in both understanding and diagnosing disease in patients, there are alternative methods which are currently showing promise. Integration of multiple -omics are becoming increasingly affordable and accessible for the whole genome, epigenome, transcriptome, microbiome, and proteome of a patient [444, 445]. This will allow a better understanding of disease, higher precision of molecular diagnosis and a more effective prognosis for patients. Machine learning also presents a powerful tool for enhancing diagnosis, drug discovery and clinical management. Clinicians can now utilise machine learning models based on glaucoma sub-phenotypes [446], eye-tracking results [447], and even OCT and fundus images of patients [448, 449, 450] to help inform diagnoses.

Overall, the work performed in this thesis has provided comprehensive evidence of the value of NGS in ophthalmic diseases. It has provided more insight into the known molecular pathology and allelic diversity in nystagmus, albinism and POAG, which will be useful for diagnosis, genetic counselling and cascade testing in families. Larger cohorts, more functional studies, and refinement of sequencing and informatic technologies in the coming years will enable the testing of new hypotheses and provide a better understanding of disease.

9 Bibliography

- [1] WHO. Global data on visual impairments 2010, 2012.
- [2] C Bunce, W Xing, and R Wormald. Causes of blind and partial sight certifications in England and Wales: April 2007-March 2008. *Eye (London, England)*, 24(11):1692–1699, nov 2010.
- [3] Fruhling V Rijdsdijk and Pak C Sham. Analytic approaches to twin data using structural equation models. *Briefings in bioinformatics*, 3(2):119–133, jun 2002.
- [4] Paul G Sanfilippo, Alex W Hewitt, Chris J Hammond, and David A Mackey. The heritability of ocular traits. *Survey of ophthalmology*, 55(6):561–583, 2010.
- [5] J M Teikari. Genetic factors in open-angle (simple and capsular) glaucoma. A population-based twin study. *Acta ophthalmologica*, 65(6):715–720, dec 1987.
- [6] M S Gottfredsdottir, T Sverrisson, D C Musch, and E Stefansson. Chronic open-angle glaucoma and associated ophthalmic findings in monozygotic twins and their spouses in Iceland. *Journal of glaucoma*, 8(2):134–139, apr 1999.
- [7] Elizabeth A Worthey, Alan N Mayer, Grant D Syverson, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Genetics in medicine : official journal of the American College of Medical Genetics*, 13(3):255–262, mar 2011.
- [8] Timothy P O’Connor and Ronald G Crystal. Genetic medicines: treatment strategies for hereditary disorders. *Nature reviews. Genetics*, 7(4):261–276, apr 2006.
- [9] Thomas Wirth, Nigel Parker, and Seppo Yla-Herttuala. History of gene therapy. *Gene*, 525(2):162–169, aug 2013.

- [10] Slave Petrovski, Vandana Shashi, Steven Petrou, et al. Exome sequencing results in successful riboflavin treatment of a rapidly progressive neurological condition. *Cold Spring Harbor molecular case studies*, 1(1):a000257, oct 2015.
- [11] Mervyn G Thomas, Moira Crosier, Susan Lindsay, et al. The clinical and molecular genetic features of idiopathic infantile periodic alternating nystagmus. *Brain*, 134(Pt 3):892–902, mar 2011.
- [12] Patrick Tarpey, Shery Thomas, Nagini Sarvananthan, et al. Mutations in FRMD7, a newly identified member of the FERM family, cause X-linked idiopathic congenital nystagmus. *Nature genetics*, 38(11):1242–1244, nov 2006.
- [13] Francesca Pasutto, Kate E Keller, Nicole Weisschuh, et al. Variants in ASB10 are associated with open-angle glaucoma. *Human molecular genetics*, 21(6):1336–1349, mar 2012.
- [14] M K Wirtz, J R Samples, K Rust, et al. GLC1F, a new primary open-angle glaucoma locus, maps to 7q35-q36. *Archives of ophthalmology (Chicago, Ill. : 1960)*, 117(2):237–241, feb 1999.
- [15] Michael A Quail, Miriam Smith, Paul Coupland, et al. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC genomics*, 13:341, 2012.
- [16] Segolene Caboche, Christophe Audebert, Yves Lemoine, and David Hot. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. *BMC genomics*, 15:264, 2014.
- [17] Sohyun Hwang, Eiru Kim, Insuk Lee, and Edward M Marcotte. Systematic comparison of variant calling pipelines using gold standard personal exome variants. *Scientific reports*, 5:17875, dec 2015.

- [18] Davis J McCarthy, Peter Humburg, Alexander Kanapin, et al. Choice of transcripts and software has a large effect on variant annotation. *Genome medicine*, 6(3):26, 2014.
- [19] Nagini Sarvananthan, Mylvaganam Surendran, Eryl O Roberts, et al. The prevalence of nystagmus: the Leicestershire nystagmus survey. *Investigative ophthalmology & visual science*, 50(11):4–9, nov 2016.
- [20] Richard V Abadi. Mechanisms underlying nystagmus. *Journal of the Royal Society of Medicine*, 95(5):231–234, may 2002.
- [21] Nahin Hussain. Diagnosis, assessment and management of nystagmus in childhood. *Paediatrics and Child Health*, 26(1):31–36, 2016.
- [22] A Serra and R J Leigh. Diagnostic value of nystagmus: spontaneous and induced ocular oscillations. *Journal of neurology, neurosurgery, and psychiatry*, 73(6):615–618, dec 2002.
- [23] CEMAS Workshop. Classification of Eye Movement Abnormalities and Strabismus (CEMAS) Workshop report. Technical report, National Eye Institute Sponsored Workshop, 2001.
- [24] Xiang He, Feng Gu, Ze Wang, et al. A novel frameshift mutation in FRMD7 causing X-linked idiopathic congenital nystagmus. *Genetic testing*, 12(4):607–613, dec 2008.
- [25] Karen Gronskov, Jakob Ek, and Karen Brondum-Nielsen. Oculocutaneous albinism. *Orphanet journal of rare diseases*, 2:43, nov 2007.
- [26] Richard A King, Rebecca K Willaert, Ramona M Schmidt, et al. MC1R mutations modify the classic phenotype of oculocutaneous albinism type 2 (OCA2). *American journal of human genetics*, 73(3):638–645, sep 2003.

- [27] J Jen, G W Kim, and R W Baloh. Clinical spectrum of episodic ataxia type 2. *Neurology*, 62(1):17–22, jan 2004.
- [28] P L Kramer, Q Yue, S T Gancher, et al. A locus for the nystagmus-associated form of episodic ataxia maps to an 11-cM region on chromosome 19p., jul 1995.
- [29] N B Lee, L Kelly, and M Sharland. Ocular manifestations of Noonan syndrome. *Eye (London, England)*, 6 (Pt 3):328–334, 1992.
- [30] Eleni Papageorgiou, Rebecca J McLean, and Irene Gottlob. Nystagmus in childhood. *Pediatrics and neonatology*, 55(5):341–351, oct 2014.
- [31] W S Oetting, C M Armstrong, A M Holleschau, A T DeWan, and G C Summers. Evidence for genetic heterogeneity in families with congenital motor nystagmus (CN). *Ophthalmic genetics*, 21(4):227–233, dec 2000.
- [32] J B Kerrison, M R Vagefi, M M Barmada, and I H Maumenee. Congenital motor nystagmus linked to Xq26-q27. *American journal of human genetics*, 64(2):600–607, feb 1999.
- [33] C Klein, P Vieregge, W Heide, et al. Exclusion of chromosome regions 6p12 and 15q11, but not chromosome region 7p11, in a German family with autosomal dominant congenital nystagmus. *Genomics*, 54(1):176–177, nov 1998.
- [34] N K Ragge, C Hartley, A M Dearlove, et al. Familial vestibulocerebellar disorder maps to chromosome 13q31-q33: a new nystagmus locus. *Journal of medical genetics*, 40(1):37–41, jan 2003.
- [35] M Michaelides, D M Hunt, and A T Moore. The cone dysfunction syndromes, feb 2004.

- [36] Edwin M Stone. Leber congenital amaurosis - a model for efficient genetic testing of heterogeneous disorders: LXIV Edward Jackson Memorial Lecture., dec 2007.
- [37] Daniel C Chung and Elias I Traboulsi. Leber congenital amaurosis: clinical correlations with genotypes, gene therapy trials update, and future directions. *Journal of AAPOS : the official publication of the American Association for Pediatric Ophthalmology and Strabismus*, 13(6):587–592, dec 2009.
- [38] T Rosenberg and M Schwartz. X-linked ocular albinism: prevalence and mutations—a national study. *European journal of human genetics : EJHG*, 6(6):570–577, 1998.
- [39] C Gail Summers. Albinism: classification, clinical characteristics, and recent findings. *Optometry and vision science : official publication of the American Academy of Optometry*, 86(6):659–662, jun 2009.
- [40] T Kausar, M A Bhatti, M Ali, R S Shaikh, and Z M Ahmed. OCA5, a novel locus for non-syndromic oculocutaneous albinism, maps to chromosome 4q24., jul 2013.
- [41] Ai-Hua Wei, Dong-Jie Zang, Zhe Zhang, et al. Exome sequencing identifies SLC24A5 as a candidate gene for nonsyndromic oculocutaneous albinism. *The Journal of investigative dermatology*, 133(7):1834–1840, jul 2013.
- [42] Karen Gronskov, Christopher M Dooley, Elsebet Ostergaard, et al. Mutations in c10orf11, a melanocyte-differentiation gene, cause autosomal-recessive albinism. *American journal of human genetics*, 92(3):415–421, mar 2013.
- [43] S Kohl, T Marx, I Giddings, et al. Total colourblindness is caused by mutations in the gene encoding the alpha-subunit of the cone photoreceptor cGMP-gated cation channel. *Nature genetics*, 19(3):257–259, jul 1998.

- [44] J A Brody, I Hussels, E Brink, and J Torres. Hereditary blindness among Pingelapese people of Eastern Caroline Islands. *Lancet (London, England)*, 1(7659):1253–1257, jun 1970.
- [45] Susanne Kohl and Christian Hamel. Clinical utility gene card for: Achromatopsia - update 2013. *European journal of human genetics : EJHG*, 21(11), nov 2013.
- [46] H Abouzeid, Y Li, I H Maumenee, S Dharmaraj, and O Sundin. A G1103R mutation in CRB1 is co-inherited with high hyperopia and Leber congenital amaurosis. *Ophthalmic genetics*, 27(1):15–20, mar 2006.
- [47] T P Dryja, S M Adams, J L Grimsby, et al. Null RPGRIP1 alleles in patients with Leber congenital amaurosis. *American journal of human genetics*, 68(5):1295–1298, may 2001.
- [48] R J Tusa. Nystagmus: diagnostic and therapeutic strategies. *Seminars in ophthalmology*, 14(2):65–73, jun 1999.
- [49] Matthew J Thurtell, Anand C Joshi, Alice C Leone, et al. Crossover trial of gabapentin and memantine as treatment for acquired nystagmus. *Annals of neurology*, 67(5):676–680, may 2010.
- [50] Cagla Eroglu, Nicola J Allen, Michael W Susman, et al. Gabapentin receptor alpha2delta-1 is a neuronal thrombospondin receptor responsible for excitatory CNS synaptogenesis. *Cell*, 139(2):380–392, oct 2009.
- [51] Michael A Rogawski and Gary L Wenk. The neuropharmacological basis for the use of memantine in the treatment of Alzheimer’s disease. *CNS drug reviews*, 9(3):275–308, 2003.
- [52] T Shery, F A Proudlock, N Sarvananthan, R J McLean, and I Gottlob. The effects of gabapentin and memantine in acquired and congenital nystagmus: a

- retrospective study. *The British journal of ophthalmology*, 90(7):839–843, jul 2006.
- [53] Jean-Marc Orgogozo, Anne-Sophie Rigaud, Albrecht Stoffler, Hans-Jorgen Mobius, and Francoise Forette. Efficacy and safety of memantine in patients with mild to moderate vascular dementia: a randomized, placebo-controlled trial (MMM 300). *Stroke; a journal of cerebral circulation*, 33(7):1834–1839, jul 2002.
- [54] Louis F Dell’Osso. Development of new treatments for congenital nystagmus. *Annals of the New York Academy of Sciences*, 956:361–379, apr 2002.
- [55] Access Economics and Pty Limited. Future sight loss UK (1): The economic impact of partial sight and blindness in the UK adult population Full report Report prepared for RNIB by Access Economics Pty Limited. Technical Report July, Access Economics Pty Limited, 2009.
- [56] N Gupta and R N Weinreb. New definitions of glaucoma. *Current opinion in ophthalmology*, 8(2):38–41, apr 1997.
- [57] Paul J Foster, Ralf Buhrmann, Harry A Quigley, and Gordon J Johnson. The definition and classification of glaucoma in prevalence surveys. *The British journal of ophthalmology*, 86(2):238–242, feb 2002.
- [58] MAYO. Mayo Foundation for Medical Education and Research, 2016.
- [59] M W Tuck and R P Crick. The age distribution of primary open angle glaucoma. *Ophthalmic epidemiology*, 5(4):173–183, dec 1998.
- [60] Clinical Management Guidelines. Glaucoma (primary open angle) (POAG). Technical Report Version 14, College of Optometrists, 2016.
-

- [61] Royal College of Ophthalmologists. A national research strategy for ophthalmology. *RCOphth publication*, 2002.
- [62] Y Y Kim and H R Jung. Clarifying the nomenclature for primary angle-closure glaucoma. *Survey of ophthalmology*, 42(2):125–136, 1997.
- [63] J T Schwartz, F H Reuling, M Feinleib, R J Garrison, and D J Collie. Twin heritability study of the effect of corticosteroids on intraocular pressure. *Journal of medical genetics*, 9(2):137–143, jun 1972.
- [64] M J Martin, A Sommer, E B Gold, and E L Diamond. Race and primary open-angle glaucoma. *American journal of ophthalmology*, 99(4):383–387, apr 1985.
- [65] R Wilson, T M Richardson, E Hertzmark, and W M Grant. Race as a risk factor for progressive glaucomatous damage. *Annals of ophthalmology*, 17(10):653–659, oct 1985.
- [66] J M Tielsch, J Katz, A Sommer, H A Quigley, and J C Javitt. Family history and risk of primary open angle glaucoma. The Baltimore Eye Survey. *Archives of ophthalmology (Chicago, Ill. : 1960)*, 112(1):69–73, jan 1994.
- [67] P Mitchell, F Hourihan, J Sandbach, and J J Wang. The relationship between glaucoma and myopia: the Blue Mountains Eye Study. *Ophthalmology*, 106(10):2010–2015, oct 1999.
- [68] Mae O Gordon, Julia A Beiser, James D Brandt, et al. The Ocular Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. *Archives of ophthalmology (Chicago, Ill. : 1960)*, 120(6):714–730, jun 2002.
- [69] Douglas R Anderson. Normal-tension glaucoma (Low-tension glaucoma). *Indian journal of ophthalmology*, 59 Suppl:S97—101, jan 2011.

- [70] W L Alward, J H Fingert, M A Coote, et al. Clinical features associated with mutations in the chromosome 1 open-angle glaucoma gene (GLC1A). *The New England journal of medicine*, 338(15):1022–1027, apr 1998.
- [71] Hussein Hollands, Davin Johnson, Simon Hollands, et al. Do findings on routine examination identify patients at risk for primary open-angle glaucoma? The rational clinical examination systematic review. *JAMA*, 309(19):2035–2042, may 2013.
- [72] C Vass, C Hirn, T Sycha, et al. Medical interventions for primary open angle glaucoma and ocular hypertension. *The Cochrane database of systematic reviews*, (4):CD003167, 2007.
- [73] David C Broadway. Visual field testing for glaucoma - a practical guide. *Community eye health*, 25(79-80):66–70, 2012.
- [74] Wallace L M Alward, Young H Kwon, Kazuhide Kawase, et al. Evaluation of optineurin sequence variations in 1,048 patients with open-angle glaucoma. *American journal of ophthalmology*, 136(5):904–910, nov 2003.
- [75] T Aung, N D Ebenezer, G Brice, et al. Prevalence of optineurin sequence variants in adult primary open angle glaucoma: implications for diagnostic testing. *Journal of medical genetics*, 40(8):e101, aug 2003.
- [76] Kristin K McDonald, Karen Abramson, Marco A Beltran, et al. Myocilin and optineurin coding variants in Hispanics of Mexican descent with POAG. *Journal of human genetics*, 55(10):697–700, oct 2010.
- [77] J H Fingert. Primary open-angle glaucoma genes. *Eye (London, England)*, 25(5):587–595, may 2011.

- [78] Khaled Abu-Amero, Altaf A Kondkar, and Kakarla V Chalam. An Updated Review on the Genetics of Primary Open Angle Glaucoma. *International journal of molecular sciences*, 16(12):28886–28911, dec 2015.
- [79] S Ennis, J Gibson, H Griffiths, et al. Prevalence of myocilin gene mutations in a novel UK cohort of POAG patients. *Eye (London, England)*, 24(2):328–333, feb 2010.
- [80] Sharareh Monemi, George Spaeth, Alexander DaSilva, et al. Identification of a novel adult-onset primary open-angle glaucoma (POAG) gene on 5q22.1. *Human molecular genetics*, 14(6):725–733, mar 2005.
- [81] Francesca Pasutto, Tomoya Matsumoto, Christian Y Mardin, et al. Heterozygous NTF4 mutations impairing neurotrophin-4 signaling in patients with primary open-angle glaucoma. *American journal of human genetics*, 85(4):447–456, oct 2009.
- [82] V C Sheffield, E M Stone, W L Alward, et al. Genetic linkage of familial open angle glaucoma to chromosome 1q21-q31. *Nature genetics*, 4(1):47–50, may 1993.
- [83] E M Stone, J H Fingert, W L Alward, et al. Identification of a gene that causes primary open angle glaucoma. *Science (New York, N.Y.)*, 275(5300):668–670, jan 1997.
- [84] Ernst R Tamm. Myocilin and glaucoma: facts and ideas. *Progress in retinal and eye research*, 21(4):395–428, jul 2002.
- [85] J H Fingert, E Heon, J M Liebmann, et al. Analysis of myocilin mutations in 1703 glaucoma patients from five different populations. *Human molecular genetics*, 8(5):899–905, may 1999.

- [86] John H Fingert, Edwin M Stone, Val C Sheffield, and Wallace L M Alward. Myocilin glaucoma. *Survey of ophthalmology*, 47(6):547–561, 2002.
- [87] J R Polansky, D J Fauss, P Chen, et al. Cellular pharmacology and molecular biology of the trabecular meshwork inducible glucocorticoid response gene product. *Ophthalmologica. Journal internationale d’ophtalmologie. International journal of ophthalmology. Zeitschrift fur Augenheilkunde*, 211(3):126–139, 1997.
- [88] R Kubota, S Noda, Y Wang, et al. A novel myosin-like protein (myocilin) expressed in the connecting cilium of the photoreceptor: molecular cloning, tissue expression, and chromosomal mapping. *Genomics*, 41(3):360–369, may 1997.
- [89] Alex W Hewitt, David A Mackey, and Jamie E Craig. Myocilin allele-specific glaucoma phenotype database. *Human mutation*, 29(2):207–211, feb 2008.
- [90] Douglas B Gould, Laura Miceli-Libby, Olga V Savinova, et al. Genetically increasing Myoc expression supports a necessary pathologic role of abnormal proteins in glaucoma. *Molecular and cellular biology*, 24(20):9019–9025, oct 2004.
- [91] Gabriel Beidoe and Shaker A Mousa. Current primary open-angle glaucoma treatments and future directions. *Clinical ophthalmology (Auckland, N.Z.)*, 6:1699–1707, 2012.
- [92] N J Risch. Searching for genetic determinants in the new millennium. *Nature*, 405(6788):847–856, jun 2000.
- [93] Joseph San Filippo, Patrick Sung, and Hannah Klein. Mechanism of eukaryotic homologous recombination. *Annual review of biochemistry*, 77:229–257, 2008.
- [94] N E MORTON. Sequential tests for the detection of linkage. *American journal of human genetics*, 7(3):277–318, sep 1955.

- [95] E Lander and L Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature genetics*, 11(3):241–247, nov 1995.
- [96] J Altmuller, L J Palmer, G Fischer, H Scherb, and M Wjst. Genomewide scans of complex human diseases: true linkage is hard to find. *American journal of human genetics*, 69(5):936–950, nov 2001.
- [97] Tayebbeh Rezaie, Anne Child, Roger Hitchings, et al. Adult-onset primary open-angle glaucoma caused by mutations in optineurin. *Science (New York, N.Y.)*, 295(5557):1077–1079, feb 2002.
- [98] Stephen F Kingsmore, Ingrid E Lindquist, Joann Mudge, Damian D Gessler, and William D Beavis. Genome-wide association studies: progress and potential for drug discovery and development. *Nature reviews. Drug discovery*, 7(3):221–230, mar 2008.
- [99] E S Lander, L M Linton, B Birren, et al. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, feb 2001.
- [100] David M Altshuler, Richard A Gibbs, Leena Peltonen, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, sep 2010.
- [101] S T Sherry, M H Ward, M Kholodov, et al. dbSNP: the NCBI database of genetic variation. *Nucleic acids research*, 29(1):308–311, jan 2001.
- [102] J K Pritchard, M Stephens, N A Rosenberg, and P Donnelly. Association mapping in structured populations. *American journal of human genetics*, 67(1):170–181, jul 2000.

- [103] William S Bush and Jason H Moore. Chapter 11: Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.
- [104] Jonathan Marchini and Bryan Howie. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics*, 11(7):499–511, jul 2010.
- [105] J M Bland and D G Altman. Multiple significance tests: the Bonferroni method. *BMJ (Clinical research ed.)*, 310(6973):170, jan 1995.
- [106] Verner Anttila, Hreinn Stefansson, Mikko Kallela, et al. Genome-wide association study of migraine implicates a common susceptibility variant on 8q22.1. *Nature genetics*, 42(10):869–873, oct 2010.
- [107] Nicole Soranzo, Tim D Spector, Massimo Mangino, et al. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nature genetics*, 41(11):1182–1190, nov 2009.
- [108] Robert J Klein, Caroline Zeiss, Emily Y Chew, et al. Complement factor H polymorphism in age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):385–389, apr 2005.
- [109] Jonathan L Haines, Michael A Hauser, Silke Schmidt, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):419–421, apr 2005.
- [110] Albert O Edwards, Robert 3rd Ritter, Kenneth J Abel, et al. Complement factor H polymorphism and age-related macular degeneration. *Science (New York, N.Y.)*, 308(5720):421–424, apr 2005.
- [111] Masakazu Nakano, Yoko Ikeda, Takazumi Taniguchi, et al. Three susceptible loci associated with primary open-angle glaucoma identified by genome-wide associ-

- ation study in a Japanese population. *Proceedings of the National Academy of Sciences of the United States of America*, 106(31):12838–12842, aug 2009.
- [112] Gudmar Thorleifsson, G Bragi Walters, Alex W Hewitt, et al. Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. *Nature genetics*, 42(10):906–909, oct 2010.
- [113] Kathryn P Burdon, Stuart Macgregor, Alex W Hewitt, et al. Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nature genetics*, 43(6):574–578, jun 2011.
- [114] Akira Meguro, Hidetoshi Inoko, Masao Ota, Nobuhisa Mizuki, and Seiamak Bahram. Genome-wide association study of normal tension glaucoma: common variants in SRBD1 and ELOVL5 contribute to disease susceptibility. *Ophthalmology*, 117(7):1331–8.e5, jul 2010.
- [115] Jane Gibson, Helen Griffiths, Gabriella De Salvo, et al. Genome-wide association study of primary open angle glaucoma risk and quantitative traits. *Molecular vision*, 18(February):1083–1092, 2012.
- [116] Dan Cao, Xiaodong Jiao, Xing Liu, et al. CDKN2B polymorphism is associated with primary open-angle glaucoma (POAG) in the Afro-Caribbean population of Barbados, West Indies. *PloS one*, 7(6):e39278, 2012.
- [117] Evan E Eichler, Jonathan Flint, Greg Gibson, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature reviews. Genetics*, 11(6):446–450, jun 2010.
- [118] Teri A Manolio, Francis S Collins, Nancy J Cox, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, oct 2009.

- [119] Janey L Wiggs, Jae Hee Kang, Brian L Yaspan, et al. Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma in Caucasians from the USA. *Human molecular genetics*, 20(23):4707–4713, dec 2011.
- [120] Anthony P Khawaja, Jessica N Cooke Bailey, Nicholas J Wareham, et al. Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nature genetics*, 50(6):778–782, jun 2018.
- [121] Robert E Furrow, Freddy B Christiansen, and Marcus W Feldman. Environment-sensitive epigenetics and the heritability of complex diseases. *Genetics*, 189(4):1377–1387, dec 2011.
- [122] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, 74(12):5463–5467, dec 1977.
- [123] Mehdi Pirooznia, Melissa Kramer, Jennifer Parla, et al. Validation and assessment of variant calling pipelines for next-generation sequencing. *Human genomics*, 8:14, jul 2014.
- [124] R D Fleischmann, M D Adams, O White, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science (New York, N.Y.)*, 269(5223):496–512, jul 1995.
- [125] Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, oct 2004.
- [126] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *Trends in genetics : TIG*, 24(3):133–141, mar 2008.

- [127] Sara Goodwin, John D McPherson, and W Richard McCombie. Coming of age: ten years of next-generation sequencing technologies., may 2016.
- [128] Jill M Johnsen, Deborah A Nickerson, and Alex P Reiner. Massively parallel sequencing: the new frontier of hematologic genomics. *Blood*, 122(19):3268–3275, nov 2013.
- [129] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456(7218):53–59, nov 2008.
- [130] Jingyue Ju, Dae Hyun Kim, Lanrong Bi, et al. Four-color DNA sequencing by synthesis using cleavable fluorescent nucleotide reversible terminators. *Proceedings of the National Academy of Sciences of the United States of America*, 103(52):19635–19640, dec 2006.
- [131] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in bioinformatics*, 11(5):473–483, sep 2010.
- [132] Martin Kircher, Susanna Sawyer, and Matthias Meyer. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. *Nucleic acids research*, 40(1):e3, jan 2012.
- [133] Matthias Meyer and Martin Kircher. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. *Cold Spring Harbor protocols*, 2010(6):pdb.prot5448, jun 2010.
- [134] Nicholas J Loman, Raju V Misra, Timothy J Dallman, et al. Performance comparison of benchtop high-throughput sequencing platforms. *Nature biotechnology*, 30(5):434–439, may 2012.

- [135] Saumya Pant, Russell Weiner, and Matthew J Marton. Navigating the rapids: the development of regulated next-generation sequencing-based clinical trial assays and companion diagnostics. *Frontiers in oncology*, 4:78, 2014.
- [136] Michael C Schatz, Arthur L Delcher, and Steven L Salzberg. Assembly of large genomes using second-generation sequencing. *Genome research*, 20(9):1165–1173, sep 2010.
- [137] Adam C English, William J Salerno, Oliver A Hampton, et al. Assessing structural variation in a personal genome-towards a human reference diploid genome. *BMC genomics*, 16:286, apr 2015.
- [138] Daniel Branton, David W Deamer, Andre Marziali, et al. The potential and challenges of nanopore sequencing. *Nature biotechnology*, 26(10):1146–1153, oct 2008.
- [139] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, 2009.
- [140] Michael C Schatz. CloudBurst: highly sensitive read mapping with MapReduce. *Bioinformatics (Oxford, England)*, 25(11):1363–1369, jun 2009.
- [141] Heng Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM, 2013.
- [142] P Ferragina and G Manzini. *Opportunistic data structures with applications*. In: Proceedings of the 41st Symposium on Foundations of Computer Science (FOCS 2000), Redondo Beach, CA, USA., 2000.
- [143] Heng Li, Bob Handsaker, Alec Wysoker, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)*, 25(16):2078–2079, aug

2009.

- [144] Aaron McKenna, Matthew Hanna, Eric Banks, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20(9):1297–1303, sep 2010.
- [145] Ferran Casals, Youssef Idaghdour, Julie Hussin, and Philip Awadalla. Next-generation sequencing approaches for genetic mapping of complex diseases. *Journal of neuroimmunology*, 248(1-2):10–22, jul 2012.
- [146] Xiaolei Wang, Jennifer Harmon, Norman Zabrieskie, et al. Using the Utah Population Database to assess familial risk of primary open angle glaucoma. *Vision research*, 50(23):2391–2395, nov 2010.
- [147] William McLaren, Laurent Gil, Sarah E Hunt, et al. The Ensembl Variant Effect Predictor. *Genome biology*, 17(1):122, jun 2016.
- [148] Pablo Cingolani, Adrian Platts, Le Lily Wang, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly*, 6(2):80–92, 2012.
- [149] Goncalo R Abecasis, David Altshuler, Adam Auton, et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, oct 2010.
- [150] Goncalo R Abecasis, Adam Auton, Lisa D Brooks, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, nov 2012.
- [151] NHLBI GO Exome Sequencing Project (ESP). Exome Variant Server.

- [152] Monkol Lek, Konrad J Karczewski, Eric V Minikel, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, 536(7616):285–291, aug 2016.
- [153] Pauline C Ng and Steven Henikoff. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, jul 2003.
- [154] Eugene V Davydov, David L Goode, Marina Sirota, et al. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS computational biology*, 6(12):e1001025, dec 2010.
- [155] Jacek Majewski, Jeremy Schwartzentruber, Emilie Lalonde, Alexandre Montpetit, and Nada Jabado. What can exome sequencing do for you? *Journal of medical genetics*, 48(9):580–589, sep 2011.
- [156] Murim Choi, Ute I Scholl, Weizhen Ji, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(45):19096–19101, nov 2009.
- [157] Chee-Seng Ku, Nasheen Naidoo, and Yudi Pawitan. Revisiting Mendelian disorders through exome sequencing. *Human genetics*, 129(4):351–370, apr 2011.
- [158] Sara Huston Katsanis and Nicholas Katsanis. Molecular genetic testing and the future of clinical genomics. *Nature reviews. Genetics*, 14(6):415–426, jun 2013.
- [159] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, et al. Exome sequencing as a tool for Mendelian disease gene discovery. *Nature reviews. Genetics*, 12(11):745–755, nov 2011.
- [160] Hane Lee, Joshua L Deignan, Naghmeh Dorrani, et al. Clinical exome sequencing for genetic identification of rare Mendelian disorders. *JAMA*, 312(18):1880–1887, nov 2014.

- [161] Yaping Yang, Donna M Muzny, Jeffrey G Reid, et al. Clinical whole-exome sequencing for the diagnosis of mendelian disorders. *The New England journal of medicine*, 369(16):1502–1511, oct 2013.
- [162] Yaping Yang, Donna M Muzny, Fan Xia, et al. Molecular findings among patients referred for clinical whole-exome sequencing. *JAMA*, 312(18):1870–1879, nov 2014.
- [163] Pauline C Ng and Ewen F Kirkness. Whole genome sequencing. *Methods in molecular biology (Clifton, N.J.)*, 628:215–226, 2010.
- [164] Molly Gasperini, Gregory M Findlay, Aaron McKenna, et al. CRISPR/Cas9-Mediated Scanning for Regulatory Elements Required for HPRT1 Expression via Thousands of Large, Programmed Genomic Deletions. *American journal of human genetics*, 101(2):192–205, aug 2017.
- [165] Siddhartha Jaiswal, Pradeep Natarajan, Alexander J Silver, et al. Clonal Hematopoiesis and Risk of Atherosclerotic Cardiovascular Disease. *The New England journal of medicine*, 377(2):111–121, jul 2017.
- [166] G C Black, R Perveen, R Bonshek, et al. Coats’ disease of the retina (unilateral retinal telangiectasis) caused by somatic mutation in the NDP gene: a role for norrin in retinal angiogenesis. *Human molecular genetics*, 8(11):2031–2035, oct 1999.
- [167] Wendy W J de Leng, Christa G Gadellaa-van Hooijdonk, Francoise A S Barendregt-Smouter, et al. Targeted Next Generation Sequencing as a Reliable Diagnostic Assay for the Detection of Somatic Mutations in Tumours Using Minimal DNA Amounts from Formalin Fixed Paraffin Embedded Material. *PloS one*, 11(2):e0149405, 2016.

- [168] David Sims, Ian Sudbery, Nicholas E Illott, Andreas Heger, and Chris P Ponting. Sequencing depth and coverage: key considerations in genomic analyses. *Nature reviews. Genetics*, 15(2):121–132, feb 2014.
- [169] Illumina. Targeted Gene Sequencing.
- [170] Julien Philippe, Mehdi Derhourhi, Emmanuelle Durand, et al. What Is the Best NGS Enrichment Method for the Molecular Diagnosis of Monogenic Diabetes and Obesity? *PloS one*, 10(11):e0143373, 2015.
- [171] UKGTN. UK Genetic Testing Network: Framework for Delivering the UK Genetic Testing Network.
- [172] Peter D Stenson, Matthew Mort, Edward V Ball, et al. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Human genetics*, 133(1):1–9, jan 2014.
- [173] J R Evans, A E Fletcher, and R P L Wormald. Causes of visual impairment in people aged 75 years and older in Britain: an add-on study to the MRC Trial of Assessment and Management of Older People in the Community. *The British journal of ophthalmology*, 88(3):365–370, mar 2004.
- [174] R Rand Allingham, Yutao Liu, and Douglas J Rhee. The genetics of primary open-angle glaucoma: a review. *Experimental eye research*, 88(4):837–844, apr 2009.
- [175] J L Wiggs, R R Allingham, D Vollrath, et al. Prevalence of mutations in TIGR/Myocilin in patients with adult and juvenile primary open-angle glaucoma., nov 1998.

- [176] Nuala A O’Leary, Mathew W Wright, J Rodney Brister, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research*, 44(D1):D733—45, jan 2016.
- [177] Kim D Pruitt, Tatiana Tatusova, and Donna R Maglott. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research*, 33(Database issue):D501—4, jan 2005.
- [178] D H Johnson. Myocilin and glaucoma: A TIGR by the tail? *Archives of ophthalmology (Chicago, Ill. : 1960)*, 118(7):974–978, jul 2000.
- [179] D A Snyder, A M Rivers, H Yokoe, B P Menco, and R R Anholt. Olfactomedin: purification, characterization, and localization of a novel olfactory glycoprotein. *Biochemistry*, 30(38):9143–9153, sep 1991.
- [180] Zachary T Resch and Michael P Fautsch. Glaucoma-associated myocilin: a better understanding but much more to learn. *Experimental eye research*, 88(4):704–712, apr 2009.
- [181] Abhishek Nag, Han Lu, Matthew Arno, et al. Evaluation of the Myocilin Mutation Gln368Stop Demonstrates Reduced Penetrance for Glaucoma in European Populations. *Ophthalmology*, 124(4):547–553, apr 2017.
- [182] Xikun Han, Emmanuelle Souzeau, Jue-Sheng Ong, et al. Myocilin Gene Gln368Ter Variant Penetrance and Association With Glaucoma in Population-Based and Registry-Based Studies. *JAMA ophthalmology*, 137(1):28–35, jan 2019.
- [183] B S Kim, O V Savinova, M V Reedy, et al. Targeted Disruption of the Myocilin Gene (Myoc) Suggests that Human Glaucoma-Causing Mutations Are Gain of Function. *Molecular and cellular biology*, 21(22):7707–7713, nov 2001.

- [184] Allan R Shepard, Nasreen Jacobson, J Cameron Millar, et al. Glaucoma-causing myocilin mutants require the Peroxisomal targeting signal-1 receptor (PTS1R) to elevate intraocular pressure. *Human molecular genetics*, 16(6):609–617, mar 2007.
- [185] Rene Dreos, Giovanna Ambrosini, Rouayda Cavin Perier, and Philipp Bucher. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic acids research*, 43(Database issue):D92—6, jan 2015.
- [186] Rene Dreos, Giovanna Ambrosini, Romain Groux, Rouaida Cavin Perier, and Philipp Bucher. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research*, 45(D1):D51—D55, jan 2017.
- [187] T A Ayoubi and W J Van De Ven. Regulation of gene expression by alternative promoters. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 10(4):453–460, mar 1996.
- [188] Reuben J Pengelly, Jane Gibson, Gaia Andreoletti, et al. A SNP profiling panel for sample tracking in whole-exome sequencing studies. *Genome medicine*, 5(9):89, 2013.
- [189] Adam Siepel, Gill Bejerano, Jakob S Pedersen, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research*, 15(8):1034–1050, aug 2005.
- [190] Eric Talevich, A Hunter Shain, Thomas Botton, and Boris C Bastian. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS computational biology*, 12(4):e1004873, apr 2016.
- [191] Goo Jun, Matthew Flickinger, Kurt N Hetrick, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype

- data. *American journal of human genetics*, 91(5):839–848, nov 2012.
- [192] Vagheesh M Narasimhan, Karen A Hunt, Dan Mason, et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science (New York, N.Y.)*, 352(6284):474–477, apr 2016.
- [193] Donna Karolchik, Angela S Hinrichs, Terrence S Furey, et al. The UCSC Table Browser data retrieval tool. *Nucleic acids research*, 32(Database issue):D493—6, jan 2004.
- [194] Kate R Rosenbloom, Cricket A Sloan, Venkat S Malladi, et al. ENCODE data in the UCSC Genome Browser: year 5 update. *Nucleic acids research*, 41(Database issue):D56–63, jan 2013.
- [195] Chengliang Dong, Peng Wei, Xueqiu Jian, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics*, 24(8):2125–2137, apr 2015.
- [196] Martin Kircher, Daniela M Witten, Preti Jain, et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics*, 46(3):310–315, mar 2014.
- [197] Matthew Mort, Timothy Sterne-Weiler, Biao Li, et al. MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome biology*, 15(1):R19, jan 2014.
- [198] Hashem A Shihab, Mark F Rogers, Julian Gough, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)*, 31(10):1536–1543, may 2015.
- [199] Beatrice Y J T Yue. Myocilin and Optineurin: Differential Characteristics and Functional Consequences. *Taiwan journal of ophthalmology*, 1(1):6–11, nov 2011.
-

- [200] M P Fautsch and D H Johnson. Characterization of myocilin-myocilin interactions. *Investigative ophthalmology & visual science*, 42(10):2324–2331, sep 2001.
- [201] Jianjiong Gao, Bulent Arman Aksoy, Ugur Dogrusoz, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling*, 6(269):p11, apr 2013.
- [202] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery*, 2(5):401–404, may 2012.
- [203] Stephane Gobeil, Marc-Andre Rodrigue, Steve Moisan, et al. Intracellular sequestration of hetero-oligomers formed by wild-type and glaucoma-causing myocilin mutants. *Investigative ophthalmology & visual science*, 45(10):3560–3567, oct 2004.
- [204] Gary Hin-Fai Yam, Katarina Gaplovska-Kysela, Christian Zuber, and Jurgen Roth. Aggregated myocilin induces russell bodies and causes apoptosis: implications for the pathogenesis of myocilin-caused primary open-angle glaucoma. *The American journal of pathology*, 170(1):100–109, jan 2007.
- [205] Thomas A Graul, Young H Kwon, M Bridget Zimmerman, et al. A case-control comparison of the clinical characteristics of glaucoma and ocular hypertensive patients with and without the myocilin Gln368Stop mutation. *American journal of ophthalmology*, 134(6):884–890, dec 2002.
- [206] A Mataftsi, F Achache, E Heon, et al. MYOC mutation frequency in primary open-angle glaucoma patients from Western Switzerland. *Ophthalmic genetics*, 22(4):225–231, dec 2001.

- [207] Colin E Willoughby, Louie Loh Yen Chan, Sarah Herd, et al. Defining the pathogenicity of optineurin in juvenile open-angle glaucoma. *Investigative ophthalmology & visual science*, 45(9):3122–3130, sep 2004.
- [208] Mathieu Faucher, Jean-Louis Anctil, Marc-Andre Rodrigue, et al. Founder TIGR/myocilin mutations for glaucoma in the Quebec population. *Human molecular genetics*, 11(18):2077–2090, sep 2002.
- [209] P J Eswari Pandaranayaka, N Prasanthi, N Kannabiran, et al. Polymorphisms in an intronic region of the myocilin gene associated with primary open-angle glaucoma—a possible role for alternate splicing. *Molecular vision*, 16:2891–2902, dec 2010.
- [210] Wenjing Liu, Yutao Liu, Pratap Challa, et al. Low prevalence of myocilin mutations in an African American population with primary open-angle glaucoma. *Molecular vision*, 18:2241–2246, 2012.
- [211] Deblina Banerjee, Ashima Bhattacharjee, Archisman Ponda, Abhijit Sen, and Kunal Ray. Comprehensive analysis of myocilin variants in east Indian POAG patients. *Molecular vision*, 18:1548–1557, 2012.
- [212] A Nekrutenko and W H Li. Transposable elements are found in a large number of human protein-coding genes. *Trends in genetics : TIG*, 17(11):619–621, nov 2001.
- [213] David J Witherspoon, W Scott Watkins, Yuhua Zhang, et al. Alu repeats increase local recombination rates. *BMC genomics*, 10:530, nov 2009.
- [214] Roy J Britten. Transposable element insertions have strongly affected human evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 107(46):19945–19948, nov 2010.

- [215] Lan Lin, Peng Jiang, Juw Won Park, et al. The contribution of Alu exons to the human proteome. *Genome biology*, 17:15, jan 2016.
- [216] Olivier Harismendy, Pauline C Ng, Robert L Strausberg, et al. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome biology*, 10(3):R32, 2009.
- [217] Monika Zavodna, Andrew Bagshaw, Rudiger Brauning, and Neil J Gemmell. The accuracy, feasibility and challenges of sequencing short tandem repeats using next-generation sequencing platforms. *PloS one*, 9(12):e113862, 2014.
- [218] Todd J Treangen and Steven L Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature reviews. Genetics*, 13(1):36–46, nov 2011.
- [219] E Colomb, T D Nguyen, A Bechetoille, et al. Association of a single nucleotide polymorphism in the TIGR/MYOCILIN gene promoter with the severity of primary open-angle glaucoma. *Clinical genetics*, 60(3):220–225, sep 2001.
- [220] Hui Guo, Minghao Li, Zhe Wang, Qiji Liu, and Xinyi Wu. Association of MYOC and APOE promoter polymorphisms and primary open-angle glaucoma: a meta-analysis. *International journal of clinical and experimental medicine*, 8(2):2052–2064, 2015.
- [221] Myung Kuk Joe, Raquel L Lieberman, Naoki Nakaya, and Stanislav I Tomarev. Myocilin Regulates Metalloprotease 2 Activity Through Interaction With TIMP3. *Investigative ophthalmology & visual science*, 58(12):5308–5318, oct 2017.
- [222] Sarah F Janssen, Theo G M F Gorgels, Wishal D Ramdas, et al. The vast complexity of primary open angle glaucoma: disease genes, risks, molecular mech-

- anisms and pathobiology. *Progress in retinal and eye research*, 37:31–67, nov 2013.
- [223] Matthew A. Miller, John H. Fingert, and Daniel I. Bettis. Genetics and genetic testing for glaucoma. *Current Opinion in Ophthalmology*, 28(2):133–138, 2017.
- [224] Tiger Zhou, Emmanuelle Souzeau, Owen M Siggs, et al. Contribution of Mutations in Known Mendelian Glaucoma Genes to Advanced Early-Onset Primary Open-Angle Glaucoma. *Investigative ophthalmology & visual science*, 58(3):1537–1544, mar 2017.
- [225] M Sarfarazi, A Child, D Stoilova, et al. Localization of the fourth locus (GLC1E) for adult-onset primary open-angle glaucoma to the 10p15-p14 region. *American journal of human genetics*, 62(3):641–652, mar 1998.
- [226] S R Bennett, W L Alward, and R Folberg. An autosomal dominant form of low-tension glaucoma. *American journal of ophthalmology*, 108(3):238–244, sep 1989.
- [227] John H Fingert, Alan L Robin, Jennifer L Stone, et al. Copy number variations on chromosome 12q14 in patients with normal tension glaucoma. *Human molecular genetics*, 20(12):2482–2494, jun 2011.
- [228] M Sarfarazi, A N Akarsu, A Hossain, et al. Assignment of a locus (GLC3A) for primary congenital glaucoma (Buphthalmos) to 2p21 and evidence for genetic heterogeneity. *Genomics*, 30(2):171–177, nov 1995.
- [229] R Melki, E Colomb, N Lefort, A P Brezin, and H-J Garchon. CYP1B1 mutations in French patients with early-onset primary open-angle glaucoma. *Journal of medical genetics*, 41(9):647–651, sep 2004.

- [230] Yutao Liu and R Rand Allingham. Major review: Molecular genetics of primary open-angle glaucoma. *Experimental eye research*, 160:62–84, jul 2017.
- [231] T R Sutter, Y M Tang, C L Hayes, et al. Complete cDNA sequence of a human dioxin-inducible mRNA identifies a new gene subfamily of cytochrome P450 that maps to chromosome 2. *The Journal of biological chemistry*, 269(18):13092–13099, may 1994.
- [232] Rolf Apweiler, Amos Bairoch, Cathy H Wu, et al. UniProt: the Universal Protein knowledgebase. *Nucleic acids research*, 32(Database issue):D115–9, jan 2004.
- [233] Suddhasil Mookherjee, Moulinath Acharya, Deblina Banerjee, Ashima Bhat-tacharjee, and Kunal Ray. Molecular basis for involvement of CYP1B1 in MYOC upregulation and its potential implication in glaucoma pathogenesis. *PloS one*, 7(9):e45077, 2012.
- [234] Gregg Duester. Keeping an eye on retinoic acid signaling during eye development. *Chemico-biological interactions*, 178(1-3):178–181, mar 2009.
- [235] Ales Cvekl and Wei-Lin Wang. Retinoic acid signaling in mammalian eye development. *Experimental eye research*, 89(3):280–291, sep 2009.
- [236] Vipul Vaibhava, Ananthamurthy Nagabhushana, Madhavi Latha Somaraju Chalasani, et al. Optineurin mediates a negative regulation of Rab8 by the GTPase-activating protein TBC1D17. *Journal of cell science*, 125(Pt 21):5026–5039, nov 2012.
- [237] Marie Pourcelot, Naima Zemirli, Leandro Silva Da Costa, et al. The Golgi apparatus acts as a platform for TBK1 activation after viral RNA sensing. *BMC biology*, 14:69, aug 2016.

- [238] Yuriko Minegishi, Mao Nakayama, Daisuke Iejima, Kazuhide Kawase, and Takeshi Iwata. Significance of optineurin mutations in glaucoma and other diseases. *Progress in retinal and eye research*, 55:149–181, nov 2016.
- [239] Y Tojima, A Fujimoto, M Delhase, et al. NAK is an IkappaB kinase-activating kinase. *Nature*, 404(6779):778–782, apr 2000.
- [240] Faxiang Li, Xingqiao Xie, Yingli Wang, et al. Structural insights into the interaction and disease mechanism of neurodegenerative disease-associated optineurin and TBK1 proteins. *Nature communications*, 7:12708, sep 2016.
- [241] Philipp Wild, Hesso Farhan, David G McEwan, et al. Phosphorylation of the autophagy receptor optineurin restricts Salmonella growth. *Science (New York, N. Y.)*, 333(6039):228–233, jul 2011.
- [242] Martin Gallenberger, Dominik M Meinel, Markus Kroeber, et al. Lack of WDR36 leads to preimplantation embryonic lethality in mice and delays the formation of small subunit ribosomal RNA in human cells in vitro. *Human molecular genetics*, 20(3):422–435, feb 2011.
- [243] Zai-Long Chi, Fumie Yasumoto, Yuri Sergeev, et al. Mutant WDR36 directly affects axon growth of retinal ganglion cells leading to progressive retinal degeneration in mice. *Human molecular genetics*, 19(19):3806–3815, oct 2010.
- [244] Jonathan M Skarie and Brian A Link. The primary open-angle glaucoma gene WDR36 functions in ribosomal RNA processing and interacts with the p53 stress-response pathway. *Human molecular genetics*, 17(16):2474–2485, aug 2008.
- [245] Tim K Footz, Jill L Johnson, Stephane Dubois, et al. Glaucoma-associated WDR36 variants encode functional defects in a yeast model system. *Human molecular genetics*, 18(7):1276–1287, apr 2009.

- [246] John H Fingert, Ben R Roos, Frances Solivan-Timpe, et al. Analysis of ASB10 variants in open angle glaucoma. *Human molecular genetics*, 21(20):4543–4548, oct 2012.
- [247] Shazia Micheal, Humaira Ayub, Farrah Islam, et al. Variants in the ASB10 Gene Are Associated with Primary Open Angle Glaucoma. *PloS one*, 10(12):e0145005, 2015.
- [248] C F Ibanez. Neurotrophin-4: the odd one out in the neurotrophin family. *Neurochemical research*, 21(7):787–793, jul 1996.
- [249] J C Conover, J T Erickson, D M Katz, et al. Neuronal deficits, not involving motor neurons, in mice lacking BDNF and/or NT4. *Nature*, 375(6528):235–238, may 1995.
- [250] Janey L Wiggs, Alex W Hewitt, Bao Jian Fan, et al. The p53 codon 72 PRO/PRO genotype may be associated with initial central visual field defects in caucasians with primary open angle glaucoma. *PloS one*, 7(9):e45613, 2012.
- [251] Wael Osman, Siew-Kee Low, Atsushi Takahashi, Michiaki Kubo, and Yusuke Nakamura. A genome-wide association study in the Japanese population confirms 9p21 and 14q23 as susceptibility loci for primary open angle glaucoma. *Human molecular genetics*, 21(12):2836–2842, jun 2012.
- [252] Puya Gharahkhani, Kathryn P Burdon, Rhys Fogarty, et al. Common variants near ABCA1, AFAP1 and GMDS confer risk of primary open-angle glaucoma. *Nature genetics*, 46(10):1120–1125, oct 2014.
- [253] Fei Chen, Alison P Klein, Barbara E K Klein, et al. Exome array analysis identifies CAV1/CAV2 as a susceptibility locus for intraocular pressure. *Investigative ophthalmology & visual science*, 56(1):544–551, dec 2014.

- [254] Zheng Li, R Rand Allingham, Masakazu Nakano, et al. A common variant near TGFBR3 is associated with primary open angle glaucoma. *Human molecular genetics*, 24(13):3880–3892, jul 2015.
- [255] Henriët Springelkamp, Adriana I Iglesias, Gabriel Cuellar-Partida, et al. ARHGEF12 influences the risk of glaucoma by increasing intraocular pressure. *Human molecular genetics*, 24(9):2689–2699, may 2015.
- [256] Jessica N Cooke Bailey, Stephanie J Loomis, Jae H Kang, et al. Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nature genetics*, 48(2):189–194, feb 2016.
- [257] Ian D Danford, Lana D Verkuil, Daniel J Choi, et al. Characterizing the "POAGome": A bioinformatics-driven approach to primary open-angle glaucoma. *Progress in retinal and eye research*, 58:89–114, may 2017.
- [258] Janey L Wiggs, Michael A Hauser, Wael Abdrabou, et al. The NEIGHBOR consortium primary open-angle glaucoma genome-wide association study: rationale, study design, and clinical variables. *Journal of glaucoma*, 22(7):517–525, sep 2013.
- [259] WellcomeTrustConsortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, jun 2007.
- [260] Teri A Manolio, Laura Lyman Rodriguez, Lisa Brooks, et al. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature genetics*, 39(9):1045–1051, sep 2007.
- [261] Henriët Springelkamp, Adriana I Iglesias, Aniket Mishra, et al. New insights into the genetics of primary open-angle glaucoma based on meta-analyses of intraocu-

- lar pressure and optic disc characteristics. *Human molecular genetics*, 26(2):438–453, jan 2017.
- [262] Shefali Setia Verma, Jessica N. Cooke Bailey, Anastasia Lucas, et al. Epistatic Gene-Based Interaction Analyses for Glaucoma in eMERGE and NEIGHBOR Consortium. *PLoS Genetics*, 12(9):1–21, 2016.
- [263] Francois-Olivier Desmet, Dalil Hamroun, Marine Lalande, et al. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research*, 37(9):e67, may 2009.
- [264] Casper Shyr, Maja Tarailo-Graovac, Michael Gottlieb, et al. FLAGS, frequently mutated genes in public exomes. *BMC medical genomics*, 7:64, dec 2014.
- [265] A Hamosh, A F Scott, J Amberger, D Valle, and V A McKusick. Online Mendelian Inheritance in Man (OMIM). *Human mutation*, 15(1):57–61, 2000.
- [266] Melissa J Landrum, Jennifer M Lee, George R Riley, et al. ClinVar : public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, 42(November 2013):980–985, 2014.
- [267] Xiaobo Huang, Miaoling Li, Xiangming Guo, et al. Mutation analysis of seven known glaucoma-associated genes in Chinese patients with glaucoma. *Investigative ophthalmology & visual science*, 55(6):3594–3602, may 2014.
- [268] Chukai Huang, Lijing Xie, Zhenggen Wu, et al. Detection of mutations in MYOC, OPTN, NTF4, WDR36 and CYP1B1 in Chinese juvenile onset open-angle glaucoma using exome sequencing. *Scientific reports*, 8(1):4498, mar 2018.
- [269] Eric Samorodnitsky, Jharna Datta, Benjamin M Jewell, et al. Comparison of custom capture for targeted next-generation DNA sequencing. *The Journal of molecular diagnostics : JMD*, 17(1):64–75, jan 2015.
-

- [270] Eleanor G Seaby, Reuben J Pengelly, and Sarah Ennis. Exome sequencing explained: a practical guide to its clinical application. *Briefings in functional genomics*, 15(5):374–384, sep 2016.
- [271] Gene Yeo and Christopher B Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of computational biology : a journal of computational molecular cell biology*, 11(2-3):377–394, 2004.
- [272] Andrea M Gazzo, Dorien Daneels, Elisa Cilia, et al. DIDA: A curated and annotated digenic diseases database. *Nucleic acids research*, 44(D1):D900–7, jan 2016.
- [273] Damian Szklarczyk, John H Morris, Helen Cook, et al. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic acids research*, 45(D1):D362–D368, jan 2017.
- [274] Michael J Meyer, Juan Felipe Beltran, Siqu Liang, et al. Interactome INSIDER: a structural interactome browser for genomic studies. *Nature methods*, 15(2):107–114, feb 2018.
- [275] E Mossotto, J J Ashton, L O’Gorman, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. *BMC bioinformatics*, 20(1):254, may 2019.
- [276] Yuval Itan, Lei Shang, Bertrand Boisson, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, 112(44):13615–13620, nov 2015.
- [277] Aaron R Quinlan and Ira M Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842,

mar 2010.

- [278] Pirro G Hysi, Ching-Yu Cheng, Henriët Springelkamp, et al. Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. *Nature genetics*, 46(10):1126–1130, oct 2014.
- [279] Veronique Vitart, Goran Bencic, Caroline Hayward, et al. New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. *Human molecular genetics*, 19(21):4304–4311, nov 2010.
- [280] Wishal D Ramdas, Leonieke M E van Koolwijk, M Kamran Ikram, et al. A genome-wide association study of optic disc parameters. *PLoS genetics*, 6(6):e1000978, jun 2010.
- [281] Stephanie J Loomis, Jae H Kang, Robert N Weinreb, et al. Association of CAV1/CAV2 genomic variants with primary open-angle glaucoma overall and by gender and pattern of visual field loss. *Ophthalmology*, 121(2):508–516, feb 2014.
- [282] Stuart Macgregor, Alex W Hewitt, Pirro G Hysi, et al. Genome-wide association identifies ATOH7 as a major gene determining human optic disc size. *Human molecular genetics*, 19(13):2716–2724, jul 2010.
- [283] Megan Ulmer, Jun Li, Brian L Yaspan, et al. Genome-wide analysis of central corneal thickness in primary open-angle glaucoma cases in the NEIGHBOR and GLAUGEN consortia. *Investigative ophthalmology & visual science*, 53(8):4468–4474, jul 2012.
- [284] Lukasz Markiewicz, Ireneusz Majsterek, Karolina Przybylowska, et al. Gene polymorphisms of the MMP1, MMP9, MMP12, IL-1beta and TIMP1 and the risk of primary open-angle glaucoma. *Acta ophthalmologica*, 91(7):e516—23, nov 2013.

- [285] Alicja Nowak, Karolina Przybylowska-Sygut, Mira Gacek, et al. Neurodegenerative Genes Polymorphisms of the -491A/T APOE, the -877T/C APP and the Risk of Primary Open-angle Glaucoma in the Polish Population. *Ophthalmic genetics*, 36(2):105–112, jun 2015.
- [286] David P Dimasi, Kathryn P Burdon, Alex W Hewitt, et al. Candidate gene study to investigate the genetic determinants of normal variation in central corneal thickness. *Molecular vision*, 16:562–569, mar 2010.
- [287] Ni Huang, Insuk Lee, Edward M Marcotte, and Matthew E Hurles. Characterising and predicting haploinsufficiency in the human genome. *PLoS genetics*, 6(10):e1001154, oct 2010.
- [288] Cristina Blanco-Marchite, Francisco Sanchez-Sanchez, Maria-Pilar Lopez-Garrido, et al. WDR36 and P53 gene variants and susceptibility to primary open-angle glaucoma: analysis of gene-gene interactions. *Investigative ophthalmology & visual science*, 52(11):8467–8478, oct 2011.
- [289] Luke O’Gorman, Angela J Cree, Daniel Ward, et al. Comprehensive sequencing of the myocilin gene in a selected cohort of severe primary open-angle glaucoma patients. *Scientific reports*, 9(1):3100, feb 2019.
- [290] Gabriela Chavarria-Soley, Heinrich Sticht, Eleni Aklillu, et al. Mutations in CYP1B1 cause primary congenital glaucoma by reduction of either activity or abundance of the enzyme. *Human mutation*, 29(9):1147–1153, sep 2008.
- [291] Antara Banerjee, Subhadip Chakraborty, Abhijit Chakraborty, Saikat Chakrabarti, and Kunal Ray. Functional and Structural Analyses of CYP1B1 Variants Linked to Congenital and Adult-Onset Glaucoma to Investigate the Molecular Basis of These Diseases. *PloS one*, 11(5):e0156252, 2016.

- [292] Simon Morton, Luke Hesson, Mark Pegg, and Philip Cohen. Enhanced binding of TBK1 by an optineurin mutant that causes a familial form of primary open angle glaucoma. *FEBS letters*, 582(6):997–1002, mar 2008.
- [293] Liyana Ahmad, Shen-Ying Zhang, Jean-Laurent Casanova, and Vanessa Sancho-Shimizu. Human TBK1: A Gatekeeper of Neuroinflammation. *Trends in molecular medicine*, 22(6):511–527, jun 2016.
- [294] John H Fingert, Kathy Miller, Adam Hedberg-Buenz, et al. Transgenic TBK1 mice have features of normal tension glaucoma. *Human molecular genetics*, 26(1):124–132, jan 2017.
- [295] Melina Herman, Michael Ciancanelli, Yi-Hung Ou, et al. Heterozygous TBK1 mutations impair TLR3 immunity and underlie herpes simplex encephalitis of childhood. *The Journal of experimental medicine*, 209(9):1567–1582, aug 2012.
- [296] Cyril Pottier, Kevin F Bieniek, NiCole Finch, et al. Whole-genome sequencing reveals important role for TBK1 and OPTN mutations in frontotemporal lobar degeneration without motor neuron disease. *Acta neuropathologica*, 130(1):77–92, jul 2015.
- [297] Yutao Liu, Melanie E Garrett, Brian L Yaspan, et al. DNA Copy Number Variants of Known Glaucoma Genes in Relation to Primary Open-Angle Glaucoma. *Invest Ophthalmol Vis Sci*, 2014.
- [298] Antonio Fabregat, Konstantinos Sidiropoulos, Guilherme Viteri, et al. Reactome pathway analysis: a high-performance in-memory approach. *BMC bioinformatics*, 18(1):142, mar 2017.
- [299] Slave Petrovski, Ayal B Gussow, Quanli Wang, et al. The Intolerance of Regulatory Sequence to Genetic Variation Predicts Gene Dosage Sensitivity. *PLoS*

- genetics*, 11(9):e1005492, sep 2015.
- [300] Howard R Petty. Frontiers of Complex Disease Mechanisms: Membrane Surface Tension May Link Genotype to Phenotype in Glaucoma. *Frontiers in cell and developmental biology*, 6:32, 2018.
- [301] Liam R Brunham, Janine K Kruit, Jahangir Iqbal, et al. Intestinal ABCA1 directly contributes to HDL biogenesis in vivo. *The Journal of clinical investigation*, 116(4):1052–1062, apr 2006.
- [302] Yuan Lei, Xuejin Zhang, Maomao Song, Jihong Wu, and Xinghuai Sun. Aqueous Humor Outflow Physiology in NOS3 Knockout Mice. *Investigative ophthalmology & visual science*, 56(8):4891–4898, jul 2015.
- [303] Jae Hee Kang, Janey L Wiggs, Bernard A Rosner, et al. Endothelial nitric oxide synthase gene variants and primary open-angle glaucoma: interactions with hypertension, alcohol intake, and cigarette smoking. *Archives of ophthalmology (Chicago, Ill. : 1960)*, 129(6):773–780, jun 2011.
- [304] James S Friedman, Mathieu Faucher, Paul Hiscott, et al. Protein localization in the human eye and genetic screen of opticin. *Human molecular genetics*, 11(11):1333–1342, may 2002.
- [305] Moulinath Acharya, Suddhasil Mookherjee, Ashima Bhattacharjee, et al. Evaluation of the OPTC gene in primary open angle glaucoma: functional significance of a silent change. *BMC molecular biology*, 8:21, mar 2007.
- [306] Eric Pasmant, Audrey Sabbagh, Michel Vidaud, and Ivan Bieche. ANRIL, a long, noncoding RNA, is an unexpected major hotspot in GWAS. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 25(2):444–448, feb 2011.

- [307] Ranadhir Mitra, Nitasha Mishra, and Girija Prasad Rath. Blood groups systems. *Indian journal of anaesthesia*, 58(5):524–528, sep 2014.
- [308] A M Brooks and W E Gillies. Blood groups as genetic markers in glaucoma. *The British journal of ophthalmology*, 72(4):270–273, apr 1988.
- [309] Muhammad Imran Khan, Shazia Micheal, Farah Akhtar, et al. Association of ABO blood groups with glaucoma in the Pakistani population. *Canadian journal of ophthalmology. Journal canadien d’ophtalmologie*, 44(5):582–586, oct 2009.
- [310] Megan Ulmer Carnes, Yangfan P Liu, R Rand Allingham, et al. Discovery and functional annotation of SIX6 variants in primary open- angle glaucoma. *PLoS genetics*, 10(5):e1004372, 2014.
- [311] Anthony P Khawaja, Michelle P Y Chan, Jennifer L Y Yip, et al. A Common Glaucoma-risk Variant of SIX6 Alters Retinal Nerve Fiber Layer and Optic Disc Measures in a European Population: The EPIC-Norfolk Eye Study. *Journal of glaucoma*, 27(9):743–749, sep 2018.
- [312] Yukihiro Shiga, Masato Akiyama, Koji M Nishiguchi, et al. Genome-wide association study identifies seven novel susceptibility loci for primary open-angle glaucoma. *Human molecular genetics*, 27(8):1486–1496, apr 2018.
- [313] Helene Choquet, Seyyedhassan Paylakhi, Stephen C Kneeland, et al. A multi-ethnic genome-wide association study of primary open-angle glaucoma identifies novel risk loci. *Nature communications*, 9(1):2278, jun 2018.
- [314] Pubudu Saneth Samarakoon, Hanne Sormo Sorte, Bjorn Evert Kristiansen, et al. Identification of copy number variants from exome sequence data. *BMC genomics*, 15:661, aug 2014.

- [315] Renjie Tan, Yadong Wang, Sarah E Kleinstein, et al. An evaluation of copy number variation detection tools from whole-exome sequencing data. *Human mutation*, 35(7):899–907, jul 2014.
- [316] Emilie Lalonde, Steffen Albrecht, Kevin C H Ha, et al. Unexpected allelic heterogeneity and spectrum of mutations in Fowler syndrome revealed by next-generation exome sequencing. *Human mutation*, 31(8):918–923, aug 2010.
- [317] Amanda Warr, Christelle Robert, David Hume, et al. Exome Sequencing: Current and Future Perspectives. *G3 (Bethesda, Md.)*, 5(8):1543–1550, jul 2015.
- [318] Elaine R Mardis. The \$1,000 genome, the \$100,000 analysis? *Genome medicine*, 2(11):84, nov 2010.
- [319] Andrea Sboner, Xinmeng Jasmine Mu, Dov Greenbaum, Raymond K Auerbach, and Mark B Gerstein. The real cost of sequencing: higher than you think! *Genome biology*, 12(8):125, aug 2011.
- [320] Tomohiro Kohmoto, Nana Okamoto, Shigeko Satomura, et al. A FRMD7 variant in a Japanese family causes congenital nystagmus. *Human genome variation*, 2:15002, 2015.
- [321] Ruifang Han, Xiaojuan Wang, Dongjie Wang, et al. GPR143 Gene Mutations in Five Chinese Families with X-linked Congenital Nystagmus. *Scientific reports*, 5:12031, 2015.
- [322] Dan L Warren and Stephanie N Seifert. Ecological niche modeling in Maxent: the importance of model complexity and the performance of model selection criteria. *Ecological applications : a publication of the Ecological Society of America*, 21(2):335–342, mar 2011.

- [323] Subramanian S Ajay, Stephen C J Parker, Hatice Ozel Abaan, Karin V Fuentes Fajardo, and Elliott H Margulies. Accurate and comprehensive sequencing of personal genomes. *Genome research*, 21(9):1498–1505, sep 2011.
- [324] Stefan H Lelieveld, Malte Spielmann, Stefan Mundlos, Joris A Veltman, and Christian Gilissen. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Human mutation*, 36(8):815–822, aug 2015.
- [325] Dave Tang, Denise Anderson, Richard W Francis, et al. Reference genotype and exome data from an Australian Aboriginal population for health-based research. *Scientific data*, 3:160023, apr 2016.
- [326] F Mirzayans, W G Pearce, I M MacDonald, and M A Walter. Mutation of the PAX6 gene in patients with autosomal dominant keratitis. *American journal of human genetics*, 57(3):539–548, sep 1995.
- [327] Yuanyuan Peng, Yan Meng, Zheng Wang, et al. A novel GPR143 duplication mutation in a Chinese family with X-linked congenital nystagmus. *Molecular vision*, 15:810–814, 2009.
- [328] Pingtong Zhou, Zhiqiang Wang, Jing Zhang, Landian Hu, and Xiangyin Kong. Identification of a novel GPR143 deletion in a Chinese family with X-linked congenital nystagmus. *Molecular vision*, 14:1015–1019, 2008.
- [329] M V Schiaffino, M T Bassi, L Galli, et al. Analysis of the OA1 gene reveals mutations in only one-third of patients with X-linked ocular albinism. *Human molecular genetics*, 4(12):2319–2325, dec 1995.
- [330] R E Schnur, M Gao, P A Wick, et al. OA1 mutations and deletions in X-linked ocular albinism. *American journal of human genetics*, 62(4):800–809, apr 1998.

- [331] William S Oetting, Jacy Pietsch, Marcia J Brott, et al. The R402Q tyrosinase variant does not cause autosomal recessive ocular albinism. *American journal of medical genetics. Part A*, 149A(3):466–469, mar 2009.
- [332] K Fukai, S A Holmes, N J Lucchese, et al. Autosomal recessive ocular albinism associated with a functionally significant tyrosinase gene polymorphism. *Nature genetics*, 9(1):92–95, jan 1995.
- [333] Saunie M Hutton and Richard A Spritz. A comprehensive genetic study of autosomal recessive ocular albinism in Caucasian patients. *Investigative ophthalmology & visual science*, 49(3):868–872, mar 2008.
- [334] R W Labrum, S Rajakulendran, T D Graves, et al. Large scale calcium channel gene rearrangements in episodic ataxia and hemiplegic migraine: implications for diagnostic testing. *Journal of medical genetics*, 46(11):786–791, nov 2009.
- [335] J Self, C Mercer, E M J Boon, et al. Infantile nystagmus and late onset ataxia associated with a CACNA1A mutation in the intracellular loop between s4 and s5 of domain 3. *Eye (London, England)*, 23(12):2251–2255, dec 2009.
- [336] L B Giebel, M A Musarella, and R A Spritz. A nonsense mutation in the tyrosinase gene of Afghan patients with tyrosinase negative (type IA) oculocutaneous albinism. *Journal of medical genetics*, 28(7):464–467, jul 1991.
- [337] C C Chang and S J Gould. Phenotype-genotype relationships in complementation group 3 of the peroxisome-biogenesis disorders. *American journal of human genetics*, 63(5):1294–1306, nov 1998.
- [338] Brandon K McCafferty, Melissa A Wilk, John T McAllister, et al. Clinical Insights Into Foveal Morphology in Albinism. *Journal of pediatric ophthalmology and strabismus*, 52(3):167–172, 2015.

- [339] Andrew B Wolf, Steven E Rubin, and Sylvia R Kodsi. Comparison of clinical findings in pediatric patients with albinism and different amplitudes of nystagmus. *Journal of AAPOS : the official publication of the American Association for Pediatric Ophthalmology and Strabismus*, 9(4):363–368, aug 2005.
- [340] Lluís Montoliu, Karen Gronskov, Ai-Hua Wei, et al. Increasing the complexity: new genes and new types of albinism. *Pigment cell & melanoma research*, 27(1):11–18, jan 2014.
- [341] H FORSIUS and A W ERIKSSON. [A NEW EYE SYNDROME WITH X-CHROMOSOMAL TRANSMISSION. A FAMILY CLAN WITH FUNDUS ALBINISM, FOVEA HYPOPLASIA, NYSTAGMUS, MYOPIA, ASTIGMATISM AND DYSCHROMATOPSIA]. *Klinische Monatsblätter für Augenheilkunde*, 144:447–457, apr 1964.
- [342] Saunie M Hutton and Richard A Spritz. Comprehensive analysis of oculocutaneous albinism among non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. *The Journal of investigative dermatology*, 128(10):2442–2450, oct 2008.
- [343] Melanie Hingorani, Kathleen A Williamson, Anthony T Moore, and Veronica van Heyningen. Detailed ophthalmologic evaluation of 43 individuals with PAX6 mutations. *Investigative ophthalmology & visual science*, 50(6):2581–2590, jun 2009.
- [344] Annagiusti Gargiulo, Francesco Testa, Settimio Rossi, et al. Molecular and clinical characterization of albinism in a large cohort of Italian patients. *Investigative ophthalmology & visual science*, 52(3):1281–1289, mar 2011.
- [345] Dimitre R Simeonov, Xinjing Wang, Chen Wang, et al. DNA variations in oculo-

- cutaneous albinism: an updated mutation list and current outstanding issues in molecular diagnostics. *Human mutation*, 34(6):827–835, jun 2013.
- [346] Kasturee Jagirdar, Darren J Smit, Stephen A Ainger, et al. Molecular analysis of common polymorphisms within the human Tyrosinase locus and genetic association with pigmentation traits. *Pigment cell & melanoma research*, 27(4):552–564, jul 2014.
- [347] R K Tripathi, L B Giebel, K M Strunk, and R A Spritz. A polymorphism of the human tyrosinase gene is associated with temperature-sensitive enzymatic activity. *Gene expression*, 1(2):103–110, may 1991.
- [348] K Toyofuku, I Wada, J C Valencia, et al. Oculocutaneous albinism types 1 and 3 are ER retention diseases: mutation of tyrosinase or Tyrp1 can affect the processing of both mutant and wild-type proteins. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology*, 15(12):2149–2161, oct 2001.
- [349] Moumita Chaki, Mainak Sengupta, Maitreyee Mondal, et al. Molecular and functional studies of tyrosinase variants among Indian oculocutaneous albinism type 1 patients., jan 2011.
- [350] M Mondal, M Sengupta, and K Ray. Functional assessment of tyrosinase variants identified in individuals with albinism is essential for unequivocal determination of genotype-to-phenotype correlation. *The British journal of dermatology*, 175(6):1232–1242, dec 2016.
- [351] Erica R Eichers, Richard Alan Lewis, Nicholas Katsanis, and James R Lupski. Triallelic inheritance: a bridge between Mendelian and multifactorial traits. *Annals of medicine*, 36(4):262–272, 2004.

- [352] Lidia Feliubadalo, Raul Tonda, Mireia Gausachs, et al. Benchmarking of Whole Exome Sequencing and Ad Hoc Designed Panels for Genetic Testing of Hereditary Cancer. *Scientific reports*, 7:37984, jan 2017.
- [353] Ivan A Adzhubei, Steffen Schmidt, Leonid Peshkin, et al. A method and server for predicting damaging missense mutations., apr 2010.
- [354] Gregory M Cooper, David L Goode, Sarah B Ng, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations., apr 2010.
- [355] Sven Opitz, Barbara Kasmann-Kellner, Markus Kaufmann, Eberhard Schwinger, and Christine Zuhlke. Detection of 53 novel DNA variations within the tyrosinase gene and accumulation of mutations in 17 patients with albinism. *Human mutation*, 23(6):630–631, jun 2004.
- [356] Caroline Rooryck, Fanny Morice-Picard, Nursel H Elcioglu, et al. Molecular diagnosis of oculocutaneous albinism: new mutations in the OCA1-4 genes and practical aspects., oct 2008.
- [357] William S Oetting, James P Fryer, Sabitha Shriram, and Richard A King. Oculocutaneous albinism type 1: the last 100 years. *Pigment cell research*, 16(3):307–311, jun 2003.
- [358] Monika B Dolinska, Nicole J Kus, S Katie Farney, et al. Oculocutaneous albinism type 1: link between mutations, tyrosinase conformational stability, and enzymatic activity. *Pigment cell & melanoma research*, 30(1):41–52, jan 2017.
- [359] Mitchell J Machiela and Stephen J Chanock. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics (Oxford, England)*, 31(21):3555–3557, nov 2015.

- [360] Pei-Wen Chiang, Anne B Fulton, Elaine Spector, and Fuki M Hisama. Synergistic interaction of the OCA2 and OCA3 genes in a family. *American journal of medical genetics. Part A*, 146A(18):2427–2430, sep 2008.
- [361] James P Fryer, William S Oetting, and Richard A King. Identification and characterization of a DNase hypersensitive region of the human tyrosinase gene. *Pigment cell research*, 16(6):679–684, dec 2003.
- [362] Chelsea S Norman, Luke O’Gorman, Jane Gibson, et al. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (TYR) causing hypomorphic oculocutaneous albinism (OCA1B). *Scientific reports*, 7(1):4415, jun 2017.
- [363] J F Berson, D W Frank, P A Calvo, B M Bieler, and M S Marks. A common temperature-sensitive allelic form of human tyrosinase is retained in the endoplasmic reticulum at the nonpermissive temperature. *The Journal of biological chemistry*, 275(16):12281–12289, apr 2000.
- [364] Pei-Wen Chiang, Elaine Spector, and Anne Chun-Hui Tsai. Evidence suggesting the inheritance mode of the human P gene in skin complexion is not strictly recessive. *American journal of medical genetics. Part A*, 146A(11):1493–1496, jun 2008.
- [365] Richard A King, Jacy Pietsch, James P Fryer, et al. Tyrosinase gene mutations in oculocutaneous albinism 1 (OCA1): definition of the phenotype. *Human genetics*, 113(6):502–513, nov 2003.
- [366] S E Dorey, M M Neveu, L C Burton, J J Sloper, and G E Holder. The clinical features of albinism and their correlation with visual evoked potentials. *The British journal of ophthalmology*, 87(6):767–772, jun 2003.

- [367] Elisabeth A H von dem Hagen, Michael B Hoffmann, and Antony B Morland. Identifying human albinism: a comparison of VEP and fMRI. *Investigative ophthalmology & visual science*, 49(1):238–249, jan 2008.
- [368] D Osborne, M Theodorou, H Lee, et al. Supranuclear eye movements and nystagmus in children: A review of the literature and guide to clinical examination, interpretation of findings and age-appropriate norms. *Eye (London, England)*, oct 2018.
- [369] Shery Thomas, Frank A Proudlock, Nagini Sarvananthan, et al. Phenotypical characteristics of idiopathic infantile nystagmus with and without mutations in FRMD7. *Brain : a journal of neurology*, 131(Pt 5):1259–1267, may 2008.
- [370] Ichiro Yabe, Mayumi Kitagawa, Yashio Suzuki, et al. Downbeat positioning nystagmus is a common clinical feature despite variable phenotypes in an FHM1 family. *Journal of neurology*, 255(10):1541–1544, oct 2008.
- [371] S Pajusalu, T Kahre, H Roomere, et al. Large gene panel sequencing in clinical diagnostics-results from 501 consecutive cases. *Clinical genetics*, 93(1):78–83, jan 2018.
- [372] John Hoon Rim, Seung-Tae Lee, Heon Yung Gee, et al. Accuracy of Next-Generation Sequencing for Molecular Diagnosis in Patients With Infantile Nystagmus Syndrome. *JAMA ophthalmology*, 135(12):1376–1385, dec 2017.
- [373] Mervyn G Thomas, Gail D E Maconachie, Viral Sheth, Rebecca J McLean, and Irene Gottlob. Development and clinical utility of a novel diagnostic nystagmus gene panel using targeted next-generation sequencing. *European journal of human genetics : EJHG*, 25(6):725–734, jun 2017.

- [374] Muhammad Waqar Arshad, Gaurav V. Harlalka, Siying Lin, et al. Mutations in TYR and OCA2 associated with oculocutaneous albinism in Pakistani families. *Meta Gene*, 17:48–55, 2018.
- [375] Quan Li and Kai Wang. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *American journal of human genetics*, 100(2):267–280, feb 2017.
- [376] Sian Ellard, Ruth Charlton, Shu Yau, et al. Practice guidelines for Sanger Sequencing Analysis and Interpretation. *Acsq*, (March):1–6, 2016.
- [377] Mark S Silverberg, Judy H Cho, John D Rioux, et al. Ulcerative colitis-risk loci on chromosomes 1p36 and 12q15 found by genome-wide association study. *Nature genetics*, 41(2):216–220, feb 2009.
- [378] Youna Hu, Alena Shmygelska, David Tran, et al. GWAS of 89,283 individuals identifies genetic variants associated with self-reporting of being a morning person. *Nature communications*, 7:10448, feb 2016.
- [379] Simon Andrews. FastQC: a quality control tool for high throughput sequence data., 2010.
- [380] Naisha Shah, Ying-Chen Claire Hou, Hung-Chun Yu, et al. Identification of Misclassified ClinVar Variants via Disease Population Prevalence. *American journal of human genetics*, 102(4):609–619, apr 2018.
- [381] Shan Yang, Stephen E Lincoln, Yuya Kobayashi, et al. Sources of discordance among germ-line variant classifications in ClinVar. *Genetics in medicine : official journal of the American College of Medical Genetics*, 19(10):1118–1126, oct 2017.
- [382] Paris J Vail, Brian Morris, Aric van Kan, et al. Comparison of locus-specific databases for BRCA1 and BRCA2 variants reveals disparity in variant classifica-

- tion within and among databases. *Journal of community genetics*, 6(4):351–359, oct 2015.
- [383] Sue Richards, Nazneen Aziz, Sherri Bale, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genetics in medicine : official journal of the American College of Medical Genetics*, 17(5):405–424, may 2015.
- [384] P J Francis, V Berry, S S Bhattacharya, and A T Moore. The genetics of childhood cataract. *Journal of medical genetics*, 37(7):481–488, jul 2000.
- [385] J Fielding Hejtmancik and Nizar Smaoui. Molecular genetics of cataract. *Developments in ophthalmology*, 37:67–82, 2003.
- [386] Jay Thompson and Naheed Lakhani. Cataracts. *Primary care*, 42(3):409–423, sep 2015.
- [387] Sloan W Rush, Ashley E Gerald, Jason C Smith, J Avery Rush, and Ryan B Rush. Prospective analysis of outcomes and economic factors of same-day bilateral cataract surgery in the United States. *Journal of cataract and refractive surgery*, 41(4):732–739, apr 2015.
- [388] Alan Shiels and J Fielding Hejtmancik. Molecular Genetics of Cataract. *Progress in molecular biology and translational science*, 134:203–218, 2015.
- [389] Veronique Pingault, Dorothee Ente, Florence Dastot-Le Moal, et al. Review and update of mutations causing Waardenburg syndrome. *Human mutation*, 31(4):391–406, apr 2010.
- [390] A P Read and V E Newton. Waardenburg syndrome. *Journal of medical genetics*, 34(8):656–665, aug 1997.

- [391] P J WAARDENBURG. A new syndrome combining developmental anomalies of the eyelids, eyebrows and nose root with pigmentary defects of the iris and head hair and with congenital deafness. *American journal of human genetics*, 3(3):195–253, sep 1951.
- [392] A L Dourmishev, L A Dourmishev, R A Schwartz, and C K Janniger. Waardenburg syndrome. *International journal of dermatology*, 38(9):656–663, sep 1999.
- [393] Chetan S Nayak and Glenn Isaacson. Worldwide distribution of Waardenburg syndrome. *The Annals of otology, rhinology, and laryngology*, 112(9 Pt 1):817–820, sep 2003.
- [394] Sravya Ravi and P V Vsatyanarayana. Waardenburg Syndrome Associated With Nephrotic syndrome and Hypothyroidism – A Rare Case Scenario. *IOSR Journal of Dental and Medical Sciences*, 18(3):60–62, 2019.
- [395] Manuel Sanchez-Martin, Arancha Rodriguez-Garcia, Jesus Perez-Losada, et al. SLUG (SNAI2) deletions in patients with Waardenburg disease. *Human molecular genetics*, 11(25):3231–3236, dec 2002.
- [396] Ribhi Hazin and Arif O Khan. Isolated microcornea: case report and relation to other "small eye" phenotypes. *Middle East African journal of ophthalmology*, 15(2):87–89, apr 2008.
- [397] Panfeng Wang, Wenmin Sun, Shiqiang Li, et al. PAX6 mutations identified in 4 of 35 families with microcornea. *Investigative ophthalmology & visual science*, 53(10):6338–6342, sep 2012.
- [398] Kai Jie Wang, Sha Wang, Ni-Qian Cao, Yong-Bin Yan, and Si Quan Zhu. A novel mutation in CRYBB1 associated with congenital cataract-microcornea syndrome: the p.Ser129Arg mutation destabilizes the betaB1/betaA3-crystallin heteromer

- but not the betaB1-crystallin homomer. *Human mutation*, 32(3):E2050–60, mar 2011.
- [399] Kamron Khan, Ahmed Al-Maskari, Martin McKibbin, et al. Genetic heterogeneity for recessively inherited congenital cataract microcornea with corneal opacity. *Investigative ophthalmology & visual science*, 52(7):4294–4299, jun 2011.
- [400] Ramachandran Ramya Devi and Perumalsamy Vijayalakshmi. Novel mutations in GJA8 associated with autosomal dominant congenital cataract and microcornea. *Molecular vision*, 12:190–195, mar 2006.
- [401] M A Reddy, P J Francis, V Berry, et al. A clinical and molecular genetic study of a rare dominantly inherited syndrome (MRCS) comprising of microcornea, rod-cone dystrophy, cataract, and posterior staphyloma. *The British journal of ophthalmology*, 87(2):197–202, feb 2003.
- [402] M Joubert, J J Eisenring, and F Andermann. Familial dysgenesis of the vermis: a syndrome of hyperventilation, abnormal eye movements and retardation. *Neurology*, 18(3):302–303, mar 1968.
- [403] Melissa A Parisi, Dan Doherty, Phillip F Chance, and Ian A Glass. Joubert syndrome (and related disorders) (OMIM 213300). *European journal of human genetics : EJHG*, 15(5):511–521, may 2007.
- [404] Pranav Mathur and Jun Yang. Usher syndrome: Hearing loss, retinal degeneration and associated abnormalities. *Biochimica et biophysica acta*, 1852(3):406–420, mar 2015.
- [405] Jan Reiners, Erwin van Wijk, Tina Marker, et al. Scaffold protein harmonin (USH1C) provides molecular links between Usher syndrome type 1 and type 2. *Human molecular genetics*, 14(24):3933–3943, dec 2005.
-

- [406] R J Smith, C I Berlin, J F Hejtmancik, et al. Clinical diagnosis of the Usher syndromes. Usher Syndrome Consortium. *American journal of medical genetics*, 50(1):32–38, mar 1994.
- [407] Christine Petit. Usher Syndrome : From Genetics to Pathogenesis. *Annu. Rev. Genomics Hum. Genet.*, 2:271–297, 2001.
- [408] C I Hope, S Bunday, D Proops, and A R Fielder. Usher syndrome in the city of Birmingham—prevalence and clinical classification. *The British journal of ophthalmology*, 81(1):46–53, jan 1997.
- [409] R Frankham. Do island populations have less genetic variation than mainland populations? *Heredity*, 78 (Pt 3):311–327, mar 1997.
- [410] Yi-Fan Lu, David B Goldstein, Misha Angrist, and Gianpiero Cavalleri. Personalized medicine and human genetic diversity. *Cold Spring Harbor perspectives in medicine*, 4(9):a008581, jul 2014.
- [411] E Gonzalez, M Bamshad, N Sato, et al. Race-specific HIV-1 disease-modifying effects associated with CCR5 haplotypes. *Proceedings of the National Academy of Sciences of the United States of America*, 96(21):12004–12009, oct 1999.
- [412] Vijay Kumar, O Prakash, S Manpreet, G Sumedh, and B Medhi. Genetic basis of HIV-1 resistance and susceptibility: an approach to understand correlation between human genes and HIV-1 infection. *Indian journal of experimental biology*, 44(9):683–692, sep 2006.
- [413] Michael Aidoo, Dianne J Terlouw, Margarette S Kolczak, et al. Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet (London, England)*, 359(9314):1311–1312, apr 2002.
- [414] Shinjini Bhatnagar and Rakesh Aggarwal. Lactose intolerance., jun 2007.

- [415] Montgomery Slatkin. A population-genetic test of founder effects and implications for Ashkenazi Jewish diseases. *American journal of human genetics*, 75(2):282–293, aug 2004.
- [416] A H Bittles. A community genetics perspective on consanguineous marriage. *Community genetics*, 11(6):324–330, 2008.
- [417] A H Bittles and M L Black. Global Patterns & Tables of Consanguinity, 2015.
- [418] A Bittles. Consanguinity and its relevance to clinical genetics. *Clinical genetics*, 60(2):89–98, aug 2001.
- [419] Bernadette Modell and Aamra Darr. Science and society: genetic counselling and customary consanguineous marriage. *Nature reviews. Genetics*, 3(3):225–229, mar 2002.
- [420] Mubasshir Ajaz, Nasreen Ali, and Gurch Randhawa. UK Pakistani views on the adverse health risks associated with consanguineous marriages. *Journal of community genetics*, 6(4):331–342, oct 2015.
- [421] Office for National Statistics. 2011 Census: Ethnic group, local authorities in the United Kingdom. Technical report, 2013.
- [422] B Dineen, R R A Bourne, Z Jadoon, et al. Causes of blindness and visual impairment in Pakistan. The Pakistan national blindness and visual impairment survey. *The British journal of ophthalmology*, 91(8):1005–1010, aug 2007.
- [423] S Pardhan and I Mahomed. The clinical characteristics of Asian and Caucasian patients on Bradford’s Low Vision Register. *Eye (London, England)*, 16(5):572–576, sep 2002.

- [424] Tommaso Pippucci, Alberto Magi, Alessandro Gialluisi, and Giovanni Romeo. Detection of runs of homozygosity from whole exome sequencing data: state of the art and perspectives for clinical, population and epidemiological studies. *Human heredity*, 77(1-4):63–72, 2014.
- [425] Thomas J Jaworek, Tasleem Kausar, Shannon M Bell, et al. Molecular genetic studies and delineation of the oculocutaneous albinism phenotype in the Pakistani population. *Orphanet journal of rare diseases*, 7:44, jun 2012.
- [426] Ola Abdelhadi, Daniela Iancu, Horia Stanescu, Robert Kleta, and Detlef Bockenhauer. EAST syndrome: Clinical, pathophysiological, and genetic aspects of mutations in KCNJ10. *Rare diseases (Austin, Tex.)*, 4(1):e1195043, 2016.
- [427] Bidisha Saha, Davor Lessel, Sheela Nampoothiri, et al. Ethnic-specific *<i>WRN</i>* mutations in South Asian Werner syndrome patients: potential founder effect in patients with Indian or Pakistani ancestry. *Molecular Genetics & Genomic Medicine*, 1(1):7–14, 2013.
- [428] Francisco C Ceballos, Scott Hazelhurst, and Michele Ramsay. Assessing runs of Homozygosity: a comparison of SNP Array and whole genome sequence low coverage data. *BMC genomics*, 19(1):106, jan 2018.
- [429] Petr Danecek, Adam Auton, Goncalo Abecasis, et al. The variant call format and VCFtools. *Bioinformatics (Oxford, England)*, 27(15):2156–2158, aug 2011.
- [430] Caroline Rooryck, Christel Roudaut, Eulalie Robine, Jorg Musebeck, and Benoit Arveiler. Oculocutaneous albinism with TYRP1 gene mutations in a Caucasian patient. *Pigment cell research*, 19(3):239–242, jun 2006.
- [431] Heather L Norton, Rick A Kittles, Esteban Parra, et al. Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Molecular*

biology and evolution, 24(3):710–722, mar 2007.

- [432] Luz M Gonzalez-Huerta, Olga M Messina-Baas, and Sergio A Cuevas-Covarrubias. A family with autosomal dominant primary congenital cataract associated with a CRYGC mutation: evidence of clinical heterogeneity. *Molecular vision*, 13:1333–1338, jul 2007.
- [433] S T Santhiya, M Shyam Manohar, D Rawley, et al. Novel mutations in the gamma-crystallin genes cause autosomal dominant congenital cataracts., may 2002.
- [434] Sebastian Kohler, Marcel H Schulz, Peter Krawitz, et al. Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American journal of human genetics*, 85(4):457–464, oct 2009.
- [435] Sebastian Kohler, Nicole A Vasilevsky, Mark Engelstad, et al. The Human Phenotype Ontology in 2017. *Nucleic acids research*, 45(D1):D865—D876, jan 2017.
- [436] A Darr and B Modell. The frequency of consanguineous marriage among British Pakistanis. *Journal of medical genetics*, 25(3):186–190, mar 1988.
- [437] Trevor J Pemberton, Devin Absher, Marcus W Feldman, et al. Genomic patterns of homozygosity in worldwide human populations. *American journal of human genetics*, 91(2):275–292, aug 2012.
- [438] Sajjad Ali Shah, Masroor Ellahi Babar, Tanveer Hussain, et al. Oculocutaneous Albinism in Pakistan : A Review. *Journal of Cancer Science & Therapy*, 10(9):253–257, 2018.
- [439] Anthony P Khawaja and Ananth C Viswanathan. Are we ready for genetic testing for primary open-angle glaucoma? *Eye (London, England)*, 32(5):877–883, may 2018.

- [440] Todd E Scheetz, Ben Faga, Lizette Ortega, et al. Glaucoma Risk Alleles in the Ocular Hypertension Treatment Study. *Ophthalmology*, 123(12):2527–2536, dec 2016.
- [441] Wayne W Grody. The transformation of medical genetics by clinical genomics: hubris meets humility. *Genetics in medicine : official journal of the American College of Medical Genetics*, mar 2019.
- [442] Leonard Berlin. The incidentaloma: a medicolegal dilemma. *Radiologic clinics of North America*, 49(2):245–255, mar 2011.
- [443] Celeste Eno, Pinar Bayrak-Toydemir, Lora Bean, et al. Misattributed parentage as an unanticipated finding during exome/genome sequencing: current clinical laboratory practices and an opportunity for standardization. *Genetics in medicine : official journal of the American College of Medical Genetics*, 21(4):861–866, apr 2019.
- [444] Sijia Huang, Kumardeep Chaudhary, and Lana X Garmire. More Is Better: Recent Progress in Multi-Omics Data Integration Methods. *Frontiers in genetics*, 8:84, 2017.
- [445] Susette Lauwen, Eiko K de Jong, Dirk J Lefeber, and Al den Hollander. Omics Biomarkers in Ophthalmology. *Investigative ophthalmology & visual science*, 58(6):BIO88–BIO98, may 2017.
- [446] Seong Jae Kim, Kyong Jin Cho, and Sejong Oh. Development of machine learning models for diagnosis of glaucoma. *PloS one*, 12(5):e0177726, 2017.
- [447] Raimondas Zemblys, Diederick C Niehorster, Oleg Komogortsev, and Kenneth Holmqvist. Using machine learning to detect events in eye-tracking data. *Behavior research methods*, 50(1):160–181, feb 2018.

- [448] Guangzhou An, Kazuko Omodaka, Kazuki Hashimoto, et al. Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images. *Journal of healthcare engineering*, 2019:4061313, 2019.
- [449] Allan Cerentini, Daniel Welfer, Marcos Cordeiro d’Ornellas, Carlos Jesus Pereira Haygert, and Gustavo Nogara Dotto. Automatic Identification of Glaucoma Using Deep Learning Methods. *Studies in health technology and informatics*, 245:318–321, 2017.
- [450] Stefan Maetschke, Bhavna Antony, Hiroshi Ishikawa, et al. A feature agnostic approach for glaucoma detection in OCT volumes. *arXiv:1807.04855*, pages 1–13, jul 2018.

A Appendix Supplementary Data

Table A.1: *Demographic and clinical information of the ten family member pairs which were included in the study regardless of meeting patient selection criteria. Patients are ordered by most severe VFMD.*

Sample ID	Other family members recruited	Patient gender	Age at diagnosis	Diagnosis	Family history of glaucoma	Ethnic origin (group)	Max CCT	Max IOP	Max CDR	Min VFMD
QGF024	QGF023	Male	50	POAG	Yes	White		42	0.9	-31.54
QGF002	QGF001	Female	83	POAG	Yes	White	522	25	0.9	-20.88
QGF008	QGF007	Female	75	POAG	Yes	White	588	30	0.8	-18.62
QGF034	QGF033	Female	79	POAG	Yes	White		22	0.7	-16.38
QGF001	QGF002	Female	49	POAG	Yes	White	515	32	0.9	-15.49
QGF033	QGF034	Female	51	POAG	Yes	White	537	24	0.6	-15.41
QGF023	QGF024	Female	61	POAG	Yes	White	548	24	0.8	-13.19
QGF016	QGF015	Male	70	POAG	Yes	White			0.8	-12.47
QGF017	QGF018	Female	77	POAG	Yes	White		26	0.8	-11.57
QGF028	QGF027	Female	59	POAG	Yes	White		32	0.8	-11.5
QGF027	QGF028	Male	36	POAG	Yes	White	536	26	0.9	-11.49
QGF012	QGF011	Male	58	POAG	Yes	White	567	38	0.8	-10.06
QGF018	QGF017	Male		POAG	Yes	White		30	0.8	-8.74
QGF015	QGF016	Male	70	POAG	Yes	White	607	26	0.9	-7.96
QGF011	QGF012	Male	59	POAG	Yes	White	533	32	0.8	-5.22
QGF009	QGF010	Female	63	POAG	Yes	White	508	42	0.9	-4.89
QGF004	QGF003	Male	62	POAG	Yes	White		36	0.8	-3.13
QGF003	QGF004	Male		POAG	Yes	White		37	0.8	-2.99
QGF010	QGF009	Female	62	POAG	Yes	White	518	25	0.9	-2.2
QGF007	QGF008	Female	56	POAG	Yes	White	580	28	0.8	-1.07

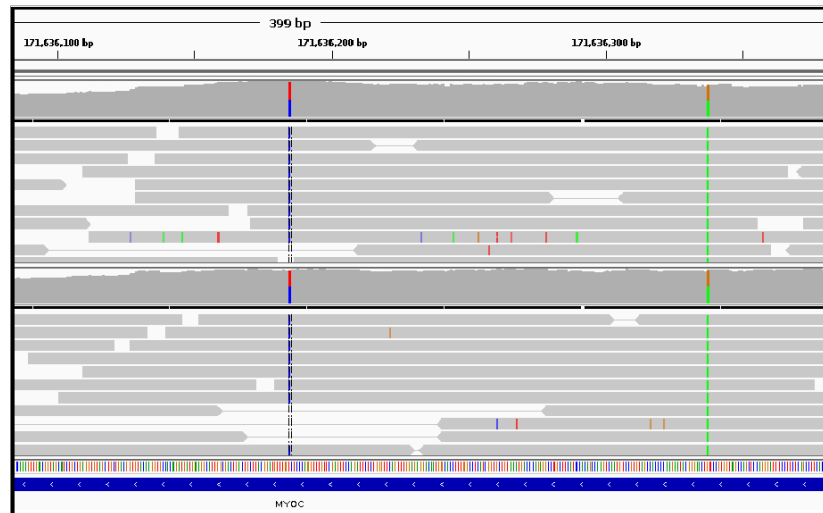


Figure A.1: *The Integrative Genomics Viewer (IGV) image of two samples with two variants, Q368* and T419A. Both variants are found on the same read pair.*

Table A.2: **All variants identified across the entire MYOC gene.** Feature, genetic feature; Start, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; avsnp144, dbSNP144 rsID; ExAC NFE, Alternate allele frequency from ExAC database (NFE populations), 1000g eur, 1000 Genomes Project (European population); FATHMM, Functional Analysis through Hidden Markov Models; CADD Phred, Combined Annotation Dependent Deletion score (Phred scale); myocDB, known myocilin variants database [89]; Rep, repeat region as identified by UCSC RepeatMasker; Acc/Dom impact on splicing, variants identified by HSF to impact splice acceptor or donors; Reg, Regulatory region identified by Ensembl regulatory build; AC, allele count in study cohort; AF, allele frequency in study cohort; Sample count, number of samples with the variant.

FEATURE	Start	Ref	Alt	avsnp144	1000g eur	ExAC NFE	FATHMM nc	CADD13 PHRED	myocDB (Dec2016)	Rep	Acc/Dom impact on splicing	Reg	AC	AF	Sample count
INTERGENIC_UP	171655903	A	G	rs543287223	.	.	0.178	1.24	-	1	CTCF_BS	CTCF_BS	1	0.0014	1
INTERGENIC_UP	171655674	C	T	rs75265199	0.01690	.	0.243	10.72	-	0	CTCF_BS	CTCF_BS	12	0.0170	12
INTERGENIC_UP	171655647	G	A	rs76059084	0.14310	.	0.235	5.32	-	0	CTCF_BS	CTCF_BS	97	0.1350	90
INTERGENIC_UP	171655625	G	C	rs774898634	.	.	0.218	1.48	-	0	CTCF_BS	CTCF_BS	1	0.0014	1
INTERGENIC_UP	171655469	G	T	rs77584896	0.03080	.	0.178	3.93	-	0	CTCF_BS	CTCF_BS	26	0.0360	26
INTERGENIC_UP	171655462	C	T	.	.	.	0.593	11.19	-	0	CTCF_BS	CTCF_BS	1	0.0014	1
INTERGENIC_UP	171655338	C	T	rs79495172	0.11230	.	0.306	9.13	-	0	CTCF_BS	CTCF_BS	66	0.0920	64
INTERGENIC_UP	171655005	C	A	rs2355922	0.56760	.	0.142	3.98	-	1	PFR	PFR	369	0.5150	276
INTERGENIC_UP	171654746	C	T	.	.	.	0.311	7.84	-	0	PFR	PFR	1	0.0014	1
INTERGENIC_UP	171654701	A	G	rs77145826	0.04570	.	0.123	1.83	-	0	PFR	PFR	48	0.0670	47
INTERGENIC_UP	171654693	C	A	rs2075537	0.33900	.	0.064	5.69	-	0	PFR	PFR	197	0.2750	173
INTERGENIC_UP	171654409	G	C	rs147606528	.	.	0.121	1.38	-	0	PFR	PFR	1	0.0014	1
INTERGENIC_UP	171653859	T	C	.	.	.	0.212	4.67	-	0	-	-	1	0.0014	1
INTERGENIC_UP	171653731	G	C	rs235921	0.01790	.	0.134	0.51	-	1	-	-	10	0.0140	10
INTERGENIC_UP	171653690	T	C	rs10798562	0.24950	.	0.124	0.97	-	1	-	-	136	0.1900	121
INTERGENIC_UP	171653609	G	C	rs12035719	0.10340	.	0.124	0.13	-	1	-	-	58	0.0810	54
INTERGENIC_UP	171653334	C	T	rs780725056	.	.	0.137	3.77	-	0	-	-	1	0.0014	1
INTERGENIC_UP	171652996	G	A	rs1490185	0.10540	.	0.188	3.31	-	0	-	-	58	0.0810	54
INTERGENIC_UP	171652917	C	T	rs34324629	0.14310	.	0.172	4.04	N	0	-	-	78	0.1090	74
INTERGENIC_UP	171652835	A	G	rs235920	0.29620	.	0.256	10.09	N	0	-	-	227	0.3170	189
INTERGENIC_UP	171652801	C	A	rs76745622	0.00800	.	0.486	21.30	N	0	-	-	7	0.0098	6
PROMOTER	171652737	A	G	rs34928744	0.02090	.	0.151	1.99	N	0	-	-	17	0.0240	17
PROMOTER	171652694	C	T	rs2075648	0.14120	.	0.086	8.42	N	0	-	-	98	0.1370	92
EXON1	171652385	C	T	rs2234926	0.14120	0.13650	0.961	9.00	N	0	-	-	98	0.1370	92

EXON1	171652269	C	T	rs757551979	.	0.00003	0.291	9.37	-	0	-	1	0.0014	1
EXON1	171652246	G	A	rs145354114	.	0.00300	0.183	0.17	N	0	-	4	0.0056	4
EXON1	171652236	G	A	rs200120115	.	0.00007	0.237	23.50	POAG	0	-	1	0.0014	1
INTRON1	171651780	T	A	rs72717876	0.05370	.	0.134	3.76	N	1	-	37	0.0520	34
INTRON1	171651642	A	G	rs235919	0.72070	.	0.049	0.79	-	1	-	498	0.6960	325
INTRON1	171651494	A	T	rs235918	0.71370	.	0.146	5.87	-	1	-	492	0.6870	323
INTRON1	171651439	A	G	rs235917	0.71770	.	0.061	11.60	-	0	-	494	0.6900	324
INTRON1	171651335	A	C	rs235916	0.78630	.	0.182	7.35	-	0	-	546	0.7630	338
INTRON1	171651027	A	G	rs539211964	.	.	0.131	1.84	-	1	-	1	0.0014	1
INTRON1	171651012	C	T	rs55815575	0.09240	.	0.212	5.36	-	1	-	78	0.1090	75
INTRON1	171650970	C	T	.	.	.	0.108	3.23	-	1	-	1	0.0014	1
INTRON1	171650942	A	G	rs547338880	.	.	0.162	7.18	-	1	-	2	0.0028	2
INTRON1	171650552	C	T	.	.	.	0.057	1.77	-	1	-	1	0.0014	1
INTRON1	171650198	AC	-	.	.	.	.	.	-	1	-	2	0.0028	2
INTRON1	171650163	-	A	.	.	.	.	.	-	0	Fail	3	0.0042	3
INTRON1	171650122	C	G	.	.	.	0.192	3.72	-	0	-	1	0.0014	1
INTRON1	171649862	T	A	rs76868060	.	.	0.106	3.20	-	1	-	1	0.0014	1
INTRON1	171649809	G	A	rs171006	0.44830	.	0.164	1.75	-	1	-	346	0.4830	260
INTRON1	171649736	A	C	rs235914	0.61330	.	0.108	0.87	-	1	-	443	0.6190	306
INTRON1	171649720	G	A	rs4916399	0.06460	.	0.057	0.95	-	1	-	48	0.0670	48
INTRON1	171649516	T	G	rs235913	0.61530	.	0.032	2.19	-	0	-	443	0.6190	306
INTRON1	171649386	T	G	rs539010888	0.00200	.	0.232	4.96	-	0	-	2	0.0028	2
INTRON1	171649106	T	C	rs11586716	0.24450	.	0.137	4.65	-	0	D	193	0.2700	169
INTRON1	171648770	G	A	rs150985414	0.02090	.	0.053	0.70	-	1	-	17	0.0240	17
INTRON1	171648687	G	A	rs12078420	0.01190	.	0.061	2.51	-	1	-	1	0.0014	1
INTRON1	171648631	-	-	TAATAAATT	.	.	.	.	-	1	Fail	59	0.0820	55
INTRON1	171648625	G	A	rs7547721	0.10440	.	0.056	0.41	-	1	-	59	0.0820	55
INTRON1	171648597	C	T	rs372206251	.	.	0.127	1.63	-	0	-	1	0.0014	1
INTRON1	171648568	T	G	rs6425364	0.71570	.	0.031	2.68	-	0	-	497	0.6940	327
INTRON1	171648385	G	C	rs116662210	0.02090	.	0.166	1.20	-	1	AD	17	0.0240	17
INTRON1	171648253	C	T	rs6425363	0.10440	.	0.056	0.22	-	0	-	59	0.0820	55
INTRON1	171648153	C	T	rs12081180	0.00100	.	0.072	1.85	-	1	-	1	0.0014	1
INTRON1	171648152	G	A	rs185718997	.	.	0.053	0.30	-	1	-	1	0.0014	1
INTRON1	171648121	G	A	rs191879785	0.00300	.	0.079	1.30	-	1	-	4	0.0056	4
INTRON1	171648049	C	T	rs10913389	0.61030	.	0.098	2.83	-	1	-	438	0.6120	303
INTRON1	171648019	A	-	.	.	.	.	.	-	1	-	347	0.4850	347

INTRON1	171647682	C	A	rs182907	0.71570	.	0.048	0.89	-	0	-	-	497	0.6940	327
INTRON1	171647464	C	T	rs112422708	0.11130	.	0.069	1.88	-	1	-	-	101	0.1410	95
INTRON1	171647453	G	A	rs235880	0.61030	.	0.071	2.20	-	1	-	-	438	0.6120	303
INTRON1	171647434	G	C	.	.	.	0.124	0.99	-	1	-	-	1	0.0014	1
INTRON1	171647395	G	A	rs71637421	0.00700	.	0.076	1.48	-	1	-	-	8	0.0110	8
INTRON1	171647393	G	A	rs148346371	0.01990	.	0.044	0.10	-	1	-	-	16	0.0220	16
INTRON1	171647381	G	A	rs171003	0.59050	.	0.044	0.14	-	1	-	-	415	0.5800	296
INTRON1	171647323	A	T	.	.	.	0.077	2.02	-	1	D	-	1	0.0014	1
INTRON1	171647322	A	G	.	.	.	0.077	2.13	-	1	-	-	1	0.0014	1
INTRON1	171647307	C	T	rs571141186	.	.	0.062	1.19	-	1	-	-	2	0.0028	2
INTRON1	171647297	G	A	rs144059958	0.10340	.	0.041	0.15	-	1	-	-	68	0.0950	68
INTRON1	171646814	T	A	rs562988873	.	.	0.052	1.02	-	0	-	-	5	0.0070	5
INTRON1	171646796	T	C	rs235879	0.59150	.	0.066	0.29	-	0	A	-	427	0.5960	300
INTRON1	171646717	G	A	rs182546961	0.10440	.	0.031	0.05	-	1	-	-	59	0.0820	55
INTRON1	171646519	C	A	.	.	.	0.056	1.60	-	1	-	-	2	0.0028	2
INTRON1	171646517	G	T	.	.	.	0.044	0.19	-	1	-	-	1	0.0014	1
INTRON1	171646515	G	T	.	.	.	0.060	0.21	-	1	-	-	7	0.0100	6
INTRON1	171646507	T	G	rs567473212	.	.	0.069	0.81	-	1	-	-	1	0.0014	1
INTRON1	171646462	G	A	rs76062269	0.01990	.	0.075	2.52	-	0	D	-	16	0.0220	16
INTRON1	171646279	G	T	rs171002	0.71570	.	0.055	0.18	-	1	-	-	497	0.6940	327
INTRON1	171646206	T	C	rs10913388	0.10440	.	0.168	2.65	-	1	-	-	59	0.0820	55
INTRON1	171646194	G	A	rs78578111	0.04670	.	0.124	4.66	-	1	-	-	35	0.0490	33
INTRON1	171646160	G	T	.	.	.	0.182	1.10	-	1	-	-	1	0.0014	1
INTRON1	171646115	T	C	.	.	.	0.175	2.71	-	0	-	-	1	0.0014	1
INTRON1	171646066	C	T	rs12035960	0.10440	.	0.860	12.78	-	0	-	-	59	0.0820	55
INTRON1	171645992	G	T	rs235878	0.71570	.	0.068	3.45	-	0	A	-	497	0.6940	327
INTRON1	171645800	A	G	rs532051985	.	.	0.150	4.08	-	0	-	-	1	0.0014	1
INTRON1	171645543	G	T	rs236875	0.10640	.	0.062	1.00	-	1	A	-	59	0.0820	55
INTRON1	171645405	C	T	.	.	.	0.092	6.85	-	1	-	-	1	0.0014	1
INTRON1	171645230	T	C	rs7523603	0.10540	.	0.121	0.88	-	1	-	-	59	0.0820	55
INTRON1	171645165	T	C	rs235877	0.71670	.	0.035	0.81	-	1	-	-	497	0.6940	327
INTRON1	171645104	G	A	.	.	.	0.348	8.29	-	1	-	-	1	0.0014	1
INTRON1	171645010	A	G	rs111569622	0.03480	.	0.178	3.95	-	1	-	-	14	0.0200	14
INTRON1	171644974	C	T	rs235876	0.71670	.	0.106	5.08	-	0	AD	-	496	0.6930	327
INTRON1	171644748	T	A	rs74511662	0.01990	.	0.163	4.09	-	0	D	-	16	0.0220	16
INTRON1	171644728	C	T	rs759101717	.	.	0.173	1.54	-	0	-	-	1	0.0014	1
INTRON1	171644695	A	G	.	.	.	0.276	14.58	-	0	-	-	1	0.0014	1
INTRON1	171644671	G	A	rs75953590	0.01990	.	0.551	15.38	-	0	-	-	16	0.0220	16

INTRON1	171644616	C	T	rs235875	0.20580	.	0.119	0.32	-	0	-	-	156	0.2310	140
INTRON1	171644528	C	T	rs181913525	0.00300	.	0.137	5.93	-	0	-	-	4	0.0056	4
INTRON1	171644376	A	G	rs235874	0.71570	.	0.036	1.64	-	0	-	-	498	0.6960	327
INTRON1	171644320	C	T	rs558629603	.	.	0.105	4.23	-	0	-	-	2	0.0028	2
INTRON1	171644264	T	-	rs775068020	.	.	.	.	-	0	-	-	92	0.1280	92
INTRON1	171643942	G	C	rs144750384	0.00800	.	0.623	10.31	-	1	D	OC	1	0.0015	1
INTRON1	171643803	-	C	rs149220264	0.13720	.	.	.	-	1	AD	-	76	0.1060	72
INTRON1	171643769	G	A	rs566356126	0.00100	.	0.237	6.98	-	1	-	-	3	0.0042	3
INTRON1	171643127	G	A	rs16864720	0.15810	.	0.115	8.13	-	0	-	-	93	0.1300	86
INTRON1	171643063	C	T	rs235873	0.55860	.	0.078	0.42	-	0	-	-	410	0.5730	295
INTRON1	171642962	G	A	.	.	.	0.064	0.86	-	0	-	-	1	0.0014	1
INTRON1	171642802	G	A	rs235871	0.55860	.	0.075	1.77	-	1	-	-	410	0.5730	295
INTRON1	171642710	T	A	rs235870	0.55860	.	0.085	6.22	-	0	-	-	410	0.5730	295
INTRON1	171642664	-	T	rs144713010	0.15710	.	.	.	-	0	-	-	93	0.1300	86
INTRON1	171642506	G	C	rs147976258	.	.	0.107	1.02	-	1	-	-	1	0.0014	1
INTRON1	171641985	C	A	rs17587451	0.01490	.	0.127	3.92	-	0	-	-	8	0.0110	8
INTRON1	171641963	A	G	rs76964562	0.13720	.	0.108	3.42	-	0	-	-	72	0.1010	67
INTRON1	171641937	G	A	rs367808238	.	.	0.176	3.50	-	0	-	-	2	0.0028	2
INTRON1	171641830	C	T	.	.	.	0.130	5.77	-	0	-	-	1	0.0014	1
INTRON1	171641819	C	A	rs235869	0.55860	.	0.138	7.38	-	0	-	-	410	0.5730	295
INTRON1	171641600	C	T	rs235868	0.71670	.	0.071	6.91	-	1	A	-	498	0.6960	327
INTRON1	171641344	G	A	rs114289326	0.01990	.	0.251	5.16	-	1	D	-	16	0.0220	16
INTRON1	171641320	C	A	.	.	.	0.195	1.09	-	1	-	-	1	0.0014	1
INTRON1	171641287	C	G	rs171001	0.69580	.	0.177	0.08	-	1	-	-	481	0.6720	324
INTRON1	171641122	G	C	rs552670111	.	.	0.203	0.14	-	1	-	-	1	0.0014	1
INTRON1	171641104	C	T	rs76621934	0.13820	.	0.144	0.33	-	1	-	-	72	0.1010	67
INTRON1	171641045	G	A	rs567010227	.	.	0.094	0.77	-	1	-	-	1	0.0014	1
INTRON1	171641044	C	T	rs10913372	0.02090	.	0.060	1.96	-	1	A	-	16	0.0220	16
INTRON1	171641004	C	T	rs757916564	.	.	0.123	4.51	-	1	A	-	2	0.0028	2
INTRON1	171640694	T	A	rs235867	0.71670	.	0.044	4.23	-	1	-	-	498	0.6960	327
INTRON1	171640453	G	A	.	.	.	0.161	3.20	-	1	A	-	1	0.0014	1
INTRON1	171640341	T	C	rs183532	0.57750	.	0.095	1.53	-	0	-	-	426	0.5950	300
INTRON1	171640047	C	T	rs80185233	0.13820	.	0.050	3.52	-	1	-	-	73	0.1040	67
INTRON1	171639974	G	A	.	.	.	0.037	0.10	-	0	-	-	15	0.0280	13
INTRON1	171639920	G	T	rs112661098	0.13820	.	0.082	1.28	-	1	-	-	72	0.1010	67
INTRON1	171639689	C	G	rs483218	0.04370	.	0.039	0.08	-	1	-	-	27	0.0380	26
INTRON1	171639682	T	C	rs483849	0.71670	.	0.059	1.50	-	1	-	-	497	0.6960	326
INTRON1	171639326	T	C	rs604864	0.57550	.	0.116	3.81	-	0	-	-	427	0.5960	301

INTRON1	171639192	G	A	rs61805425	0.08450	0.084	3.57	-	1	-	-	58	0.0810	57
INTRON1	171639114	-	A	.	.	.	.	-	0	A	-	1	0.0014	1
INTRON1	171639096	G	C	rs603930	0.71370	0.024	0.09	-	1	-	-	498	0.6960	327
INTRON1	171639054	C	T	rs41263718	0.13820	0.081	1.37	-	1	-	-	71	0.0990	67
INTRON1	171639025	G	C	rs41263716	0.10340	0.064	0.12	-	1	-	-	59	0.0820	55
INTRON1	171639017	A	G	rs603501	0.00100	0.056	1.05	-	1	-	-	1	0.0014	1
INTRON1	171639002	A	C	rs603490	0.57550	0.112	2.80	-	1	-	-	427	0.5960	301
INTRON1	171638967	G	A	.	.	0.057	0.89	-	1	-	-	1	0.0014	1
INTRON1	171638811	C	T	rs751233286	.	0.128	0.51	-	1	-	-	1	0.0014	1
EXON2	171638679	C	T	rs141584495	.	0.00050	15.63	N	0	-	-	3	0.0042	3
INTRON2	171638562	C	T	rs2032555	0.71470	0.70970	1.43	N	0	-	-	498	0.6960	327
INTRON2	171638430	AG	-	rs144871239	0.00990	.	.	-	0	-	-	27	0.0380	27
INTRON2	171638370	A	-	rs201642544	0.01990	.	.	-	0	-	-	17	0.0240	17
INTRON2	171638226	C	-	.	.	.	.	-	0	-	OC	1	0.0014	1
INTRON2	171638066	A	T	rs746256075	.	0.218	1.47	-	0	-	OC	2	0.0028	2
INTRON2	171637893	A	G	rs235882	0.71570	0.047	0.07	-	1	-	-	504	0.7040	331
INTRON2	171637715	C	T	rs545761328	.	0.071	2.39	-	1	-	-	1	0.0014	1
INTRON2	171637680	T	C	rs10913370	0.03880	0.042	0.20	-	1	-	-	38	0.0530	38
INTRON2	171637634	A	G	.	.	0.054	0.81	-	1	-	-	1	0.0014	1
INTRON2	171637606	A	T	rs187172709	.	0.089	0.13	-	0	A	-	1	0.0014	1
INTRON2	171637599	T	C	rs7545646	0.13820	0.086	1.26	-	0	D	-	70	0.0980	66
INTRON2	171637310	A	G	rs79263003	0.01090	0.570	4.67	-	0	-	-	10	0.0140	10
INTRON2	171637236	A	G	.	.	0.135	4.77	-	0	-	-	1	0.0014	1
INTRON2	171636914	T	G	rs12076134	0.21170	0.171	2.37	-	1	-	-	168	0.2350	155
INTRON2	171636782	G	A	rs79255460	0.01990	0.176	1.60	N	0	BP	-	17	0.0240	17
EXON3	171636585	C	A	rs146606638	0.00800	0.331	14.00	N	0	-	-	2	0.0028	2
EXON3	171636534	G	A	rs148433908	.	0.00030	0.07	N	0	-	-	1	0.0014	1
EXON3	171636399	A	G	rs61730974	0.02090	0.902	0.00	N	0	-	-	17	0.0240	17
EXON3	171636338	G	A	rs74315329	0.00200	0.00150	37.00	POAG	0	-	-	7	0.0098	7
EXON3	171636247	T	C	rs56314834	0.00700	0.00480	3.87	N	0	-	-	10	0.0140	10
EXON3	171636185	T	C	.	.	0.977	23.50	-	0	-	-	2	0.0028	2
EXON3	171636126	G	A	rs375235405	.	0.00004	11.66	-	0	-	-	1	0.0014	1
UTR3	171635852	C	T	rs74403899	.	0.205	7.76	-	0	-	-	2	0.0028	2
UTR3	171635499	G	A	rs142425726	0.00500	0.106	3.05	-	0	-	-	3	0.0042	3
INTERGENIC_DS	171635181	C	T	rs114295456	0.02090	0.168	3.71	-	0	-	-	18	0.0260	16

	AAAA		.	-	1	-	4	0.0210	2
INTERGENIC_DS	171635075	-	rs781312247	.	.	.	.	.	.
	AAAA								
	AAA								

Table A.3: **11 cases carrying a candidate disease-causing MYOC mutation.** Demographic data and clinical data for each patient is listed including patient gender, ethnicity, age at diagnosis, family history of POAG, diagnosis, intraocular pressure (IOP), cup:disc ratio (CDR), central corneal thickness (CCT), visual field mean deviation (VFMD), cataract status and age-related macular degeneration (AMD) status.

Patient	Exon	Variant	Study Site	Patient gender	Ethnicity	Age at diagnosis	Family history of POAG	Diagnosis	IOP	CDR	CCT	VFMD	Cataract	AMD
1	1	R126W	Frimley Park	Female	Caucasian	74	Yes	POAG	27	0.6	572	-4.8	Yes	No
2	2	K216K	University Hospital Southampton	Male	Caucasian	62	Yes	POAG	32	0.85	Not measured	-7.39	No	No
3	2	K216K	University Hospital Southampton	Male	Caucasian	65	Yes	POAG	34	0.7	569	-4.26	No	No
4	2	K216K	University Hospital Southampton	Female	Caucasian	68	Yes	POAG	23	0.9	588	-12.58	Yes	No
5	3	Q368*	Frimley Park	Female	Caucasian	87	No	POAG	24	0.9	529	-16.63	Yes	Yes
6	3	Q368*	Frimley Park	Female	Caucasian	85	No	POAG	23	0.8	534	-14.74	No	No
7	3	Q368*	University Hospital Southampton	Male	Caucasian	56	Yes	POAG	27	0.8	Not measured	-8.46	No	Yes
8	3	Q368*	Portsmouth	Male	Caucasian	74	No	POAG	24	0.8	487	-30.83	Yes	Yes
9	3	Q368*	Torbay	Female	Caucasian	79	Yes	POAG	30	0.95	Not measured	-13.58	Unknown	Unknown
10	3	Q368* & T419A	Birmingham	Female	Caucasian	50	Yes	POAG	30	0.8	Not measured	-3.34	Yes	No
11	3	Q368* & T419A	Portsmouth	Male	Caucasian	56	Yes	POAG	26	0.8	543	-9.3	Yes	No

B Appendix Supplementary Data

Table B.1: *UCSC RefSeq (RefGene) hg38 coordinates and coverage of at least 1X depth for SH3PXD2B, SOD2, NTM, APOE, and TXNRD2. Red highlights the RefSeq CDS which were missed or partially missed by NRCC capture. *NRCC capture for this CDS had coordinates of chr6:159692660-159692863, which is 197bp less than the UCSC RefSeq CDS length (chr6:159692463-159692863).*

Gene	Chrom	Start	End	Exon size (bp)	Exon covered (bp)	Coverage Proportion
SH3PXD2B	5	172325275	172325380	105	0	0.000
		172338368	172339916	1548	1544	0.997
		172346135	172346261	126	125	0.992
		172347282	172347332	50	49	0.980
		172350362	172350589	227	226	0.996
		172353887	172354005	118	117	0.992
		172358772	172358877	105	104	0.990
		172362734	172362869	135	134	0.993
		172373789	172373815	26	25	0.962
		172382035	172382127	92	91	0.989
		172394562	172394639	77	76	0.987
		172406276	172406352	76	75	0.987
		172422415	172422496	81	80	0.988
SOD2	6	172454277	172454352	75	74	0.987
		159682492	159682638	146	142	0.973
		159684853	159685033	180	179	0.994
		159688125	159688242	117	116	0.991
		159692463	159692863	400	202	0.505*
NTM	11	159693144	159693167	23	22	0.957
		131370806	131370888	82	81	0.988
		131660993	131661048	55	0	0.000
		131911068	131911093	25	0	0.000
		131911481	131911648	167	166	0.994
		132146281	132146514	233	232	0.996
		132212021	132212147	126	125	0.992
		132307688	132307823	135	134	0.993
		132310111	132310232	121	120	0.992
		132314551	132314720	169	166	0.982
		132317654	132317690	36	0	0.000
		132330152	132330185	33	32	0.970
		132335045	132335146	101	98	0.970
APOE	19	44908532	44909250	718	715	0.996
		44905868	44905923	55	0	0.000
		44906601	44906667	66	42	0.636
		44907759	44907952	193	192	0.995
TXNRD2	22	19877104	19877234	130	126	0.969
		19878089	19878187	98	97	0.990
		19878365	19878437	72	71	0.986
		19880178	19880271	93	92	0.989
		19880621	19880717	96	95	0.990
		19883324	19883461	137	136	0.993
		19895132	19895200	68	0	0.000
		19895406	19895581	175	174	0.994
		19898038	19898130	92	91	0.989
		19899048	19899068	20	19	0.950
		19911376	19911447	71	70	0.986
		19915213	19915276	63	62	0.984
		19915764	19915843	79	78	0.987
		19918142	19918217	75	74	0.987
		19918859	19919004	145	144	0.993
		19919542	19919599	57	56	0.982
		19931029	19931098	69	68	0.986
		19932356	19932369	13	0	0.000
		19933433	19933440	7	0	0.000
		19941700	19941803	103	102	0.990

Table B.2: *OPTN predicted interface coordinates (Interactome INSIDER) BED file (hg38).*

Chrom	Start	End
chr10	13110393	13110395
chr10	13125530	13125532
chr10	13125548	13125550
chr10	13132091	13132093
chr10	13132109	13132111
chr10	13132115	13132117
chr10	13132133	13132135
chr10	13133524	13133526
chr10	13133542	13133544
chr10	13136801	13136803
chr10	13136810	13136812
chr10	13136819	13136821
chr10	13136828	13136830

Table B.3: *TBK1 predicted interface coordinates (Interactome INSIDER) BED file (hg38).*

Chrom	Start	End
chr12	64455901	64455903
chr12	64455904	64455906
chr12	64455907	64455909
chr12	64455910	64455912
chr12	64455913	64455915
chr12	64455916	64455918
chr12	64455919	64455921
chr12	64455922	64455924
chr12	64455925	64455927
chr12	64455928	64455930
chr12	64455934	64455936
chr12	64455943	64455945
chr12	64460222	64460224
chr12	64460225	64460227
chr12	64460228	64460230
chr12	64460234	64460236
chr12	64460237	64460239
chr12	64460240	64460242

APPENDIX

chr12	64460243	64460245
chr12	64460246	64460248
chr12	64460249	64460251
chr12	64460252	64460254
chr12	64460258	64460260
chr12	64460261	64460263
chr12	64460270	64460272
chr12	64460282	64460284
chr12	64464337	64464339
chr12	64464340	64464342
chr12	64464367	64464369
chr12	64464373	64464375
chr12	64464376	64464378
chr12	64464388	64464390
chr12	64464391	64464393
chr12	64464400	64464402
chr12	64464403	64464405
chr12	64464406	64464408
chr12	64464409	64464411
chr12	64464415	64464417
chr12	64464418	64464420
chr12	64466927	64466929
chr12	64466942	64466944
chr12	64466951	64466953
chr12	64467017	64467019
chr12	64467020	64467022
chr12	64467026	64467028
chr12	64467029	64467031
chr12	64467032	64467034
chr12	64467035	64467037
chr12	64467041	64467043
chr12	64467044	64467046
chr12	64467047	64467049
chr12	64467050	64467052
chr12	64467053	64467055
chr12	64467056	64467058
chr12	64467059	64467061
chr12	64467062	64467064
chr12	64467065	64467067

APPENDIX

chr12	64467068	64467070
chr12	64467071	64467073
chr12	64467074	64467076
chr12	64467080	64467082
chr12	64474236	64474238
chr12	64474239	64474241
chr12	64474242	64474244
chr12	64474245	64474247
chr12	64474248	64474250
chr12	64474251	64474253
chr12	64474254	64474256
chr12	64474257	64474259
chr12	64474260	64474262
chr12	64474263	64474265
chr12	64474281	64474283
chr12	64474290	64474292
chr12	64474293	64474295
chr12	64474302	64474304
chr12	64474305	64474307
chr12	64474314	64474316
chr12	64474326	64474328
chr12	64474344	64474346
chr12	64474353	64474355
chr12	64474359	64474361
chr12	64474362	64474364
chr12	64474365	64474367
chr12	64474368	64474370
chr12	64474371	64474373
chr12	64474374	64474376
chr12	64474380	64474382
chr12	64474383	64474385
chr12	64474389	64474390
chr12	64480012	64480012
chr12	64480013	64480015
chr12	64480025	64480027
chr12	64480103	64480105
chr12	64481855	64481857
chr12	64481888	64481890
chr12	64481894	64481896

APPENDIX

chr12	64481915	64481917
chr12	64481924	64481926
chr12	64481945	64481947
chr12	64481951	64481953
chr12	64481990	64481992
chr12	64481996	64481998
chr12	64484322	64484324
chr12	64484331	64484333
chr12	64484334	64484336
chr12	64484373	64484375
chr12	64484379	64484381
chr12	64484430	64484432
chr12	64484433	64484435
chr12	64484436	64484438
chr12	64484439	64484441

Table B.4: *CDC7 predicted interface coordinates (Interactome INSIDER) BED file (hg38).*

Chrom	Start	End
chr9	21970903	21970905
chr9	21970924	21970926
chr9	21970927	21970929
chr9	21970990	21970992
chr9	21971023	21971025
chr9	21971026	21971028
chr9	21971029	21971031
chr9	21971056	21971058
chr9	21971083	21971085
chr9	21971086	21971088
chr9	21971089	21971091
chr9	21971092	21971094
chr9	21971095	21971097
chr9	21971098	21971100
chr9	21971107	21971109
chr9	21971128	21971130
chr9	21971131	21971133
chr9	21971134	21971136
chr9	21971143	21971145
chr9	21971158	21971160

APPENDIX

chr9	21971161	21971163
chr9	21971185	21971187
chr9	21971188	21971190
chr9	21971194	21971196
chr9	21971200	21971202
chr9	21971206	21971208
chr9	21974690	21974692
chr9	21974693	21974695
chr9	21974696	21974698
chr9	21974699	21974701
chr9	21974723	21974725
chr9	21974750	21974752
chr9	21974756	21974758
chr9	21974762	21974764
chr9	21974765	21974767
chr9	21974774	21974776

Table B.5: *CDKN2A* predicted interface coordinates (*Interactome INSIDER*) *BED* file (*hg38*).

Chrom	Start	End
chr1	91507892	91507894
chr1	91507895	91507897
chr1	91507904	91507906
chr1	91507925	91507927
chr1	91507928	91507930
chr1	91507931	91507933
chr1	91507934	91507936
chr1	91507937	91507937
chr1	91508262	91508263
chr1	91508264	91508266
chr1	91508267	91508269
chr1	91508276	91508278
chr1	91508339	91508341
chr1	91508342	91508344
chr1	91508345	91508347
chr1	91508351	91508353
chr1	91508354	91508356
chr1	91508357	91508359
chr1	91508366	91508368

APPENDIX

chr1	91511598	91511600
chr1	91511604	91511606
chr1	91511607	91511609
chr1	91511619	91511621
chr1	91511622	91511624
chr1	91511625	91511627
chr1	91511628	91511630
chr1	91511634	91511636
chr1	91511637	91511639
chr1	91511640	91511642
chr1	91511643	91511645
chr1	91511664	91511666
chr1	91511667	91511669
chr1	91511673	91511675
chr1	91511676	91511678
chr1	91511688	91511690
chr1	91511859	91511861
chr1	91511862	91511864
chr1	91511886	91511888
chr1	91511892	91511894
chr1	91511913	91511915
chr1	91511916	91511918
chr1	91511919	91511921
chr1	91511922	91511923
chr1	91513058	91513058
chr1	91513071	91513073
chr1	91513095	91513097
chr1	91513104	91513106
chr1	91515816	91515818
chr1	91515819	91515821
chr1	91515822	91515824
chr1	91515855	91515857
chr1	91515861	91515863
chr1	91515864	91515866
chr1	91520180	91520182
chr1	91520183	91520185
chr1	91520186	91520188
chr1	91520195	91520197
chr1	91520198	91520200

APPENDIX

chr1	91520201	91520203
chr1	91520204	91520206
chr1	91520207	91520209
chr1	91520210	91520212
chr1	91520234	91520236
chr1	91520252	91520254
chr1	91520261	91520263
chr1	91520279	91520279
chr1	91524041	91524042
chr1	91524052	91524054
chr1	91524055	91524057
chr1	91524058	91524060
chr1	91524064	91524066
chr1	91524079	91524081
chr1	91524100	91524102
chr1	91524316	91524318
chr1	91524352	91524354
chr1	91524376	91524378
chr1	91524400	91524402
chr1	91524403	91524405

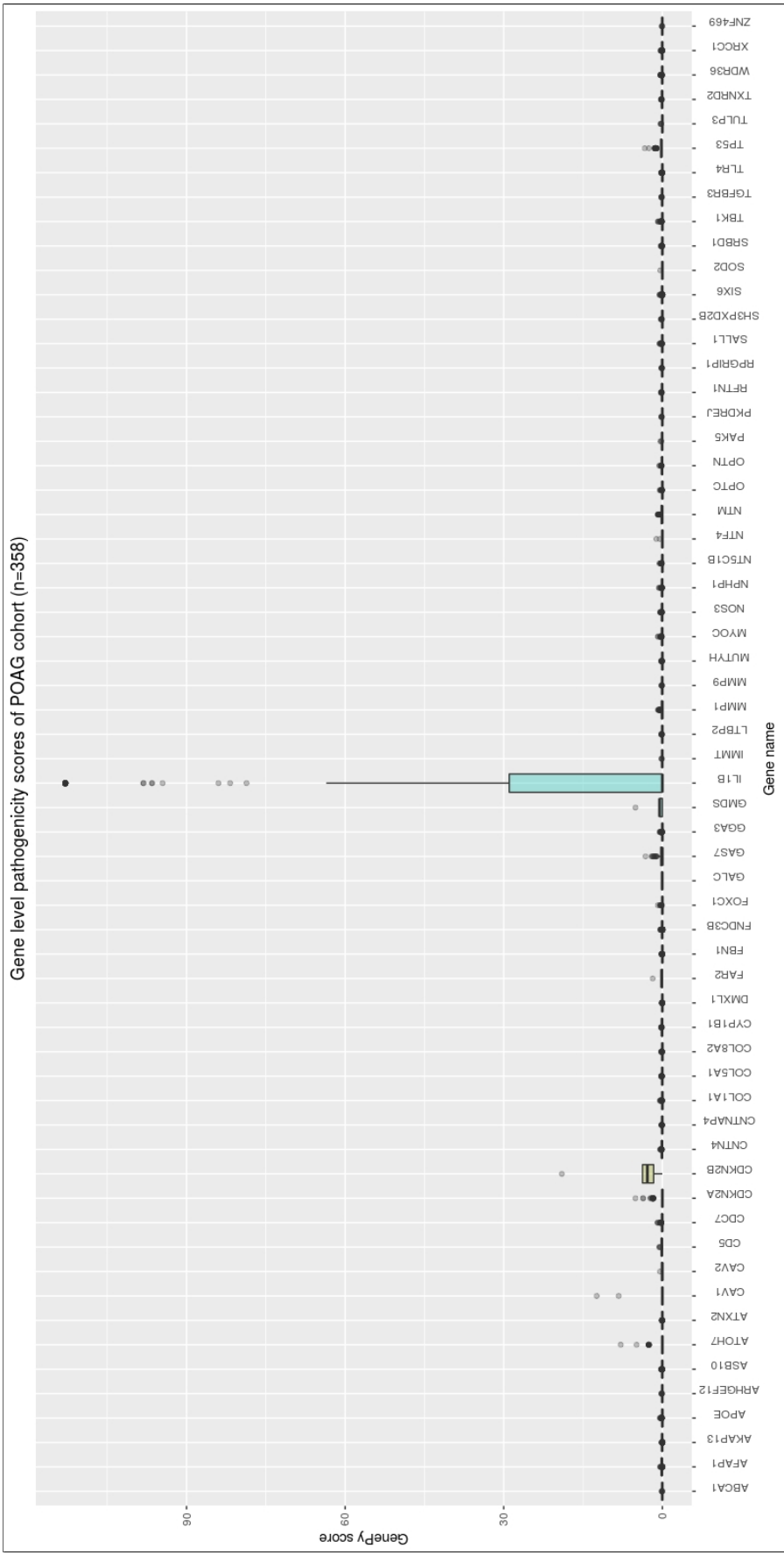


Figure B.1: 61 POAG genes GenePy scores from the POAG cohort (n=358) corrected by GDI and gene length. IL1B has highly elevated GenePy scores in comparison to the other POAG genes.

Table B.6: *Mann-Whitney test p-values between the POAG (n=350) and IBD (n=403) cohort with Bonferroni and False Discovery Rate (FDR) corrected p-values calculated for 64 single genes and six permutations of interacting genes. IL1B, ABO and SIX6 are significant following Bonferroni and False Discovery Rate (FDR) corrections. NS was given to values of 1.*

	Genes	Uncorrected p-value	Bonferroni	Benjamini & Hochberg
1	<i>IL1B</i>	1.09E-9	7.65E-8	7.65E-8
2	<i>ABO</i>	6.60E-8	4.62E-6	2.31E-6
3	<i>SIX6</i>	3.73E-4	2.61E-2	8.70E-3
4	<i>TLR4</i>	3.14E-3	2.20E-1	5.50E-2
5	<i>CDKN2A-AS1</i>	6.85E-3	4.80E-1	9.59E-2
6	<i>MYOC</i>	9.18E-3	6.42E-1	1.01E-1
7	<i>LTBP2</i>	1.01E-2	7.04E-1	1.01E-1
8	<i>MYOC-CYP1B1</i>	2.59E-2	NS	2.14E-1
9	<i>ATOH7</i>	4.47E-2	NS	2.14E-1
10	<i>WDR36</i>	5.09E-2	NS	0.214
11	<i>CNTNAP4</i>	5.22E-2	NS	0.262
12	<i>PKDREJ</i>	6.22E-2	NS	0.335
13	<i>SALL1</i>	9.64E-2	NS	0.535
14	<i>CDC7</i>	1.19E-1	NS	0.535
15	<i>ATXN2</i>	1.24E-1	NS	0.535
16	<i>FOXC1</i>	1.30E-1	NS	0.921
17	<i>FNDC3B</i>	2.13E-1	NS	NS
18	<i>AKAP13</i>	2.37E-1	NS	NS
19	<i>IMMT</i>	2.85E-1	NS	NS
20	<i>GGA3</i>	4.16E-1	NS	NS
21	<i>GMDS</i>	4.70E-1	NS	NS
22	<i>NPHP1</i>	5.30E-1	NS	NS
23	<i>RFTN1</i>	5.87E-1	NS	NS
24	<i>TBK1</i>	6.00E-1	NS	NS
25	<i>CAV1</i>	6.08E-1	NS	NS
26	<i>CDKN2A-CDC7</i>	6.14E-1	NS	NS
27	<i>NT5C1B</i>	6.21E-1	NS	NS
28	<i>AFAP1</i>	6.27E-1	NS	NS
29	<i>FBN1</i>	6.65E-1	NS	NS
30	<i>TBK1-TLR4</i>	6.77E-1	NS	NS
31	<i>FAR2</i>	6.80E-1	NS	NS
32	<i>NTF4</i>	6.80E-1	NS	NS

APPENDIX

33	<i>OPTN-TBK1</i>	7.02E-1	NS	NS
34	<i>GALC</i>	7.07E-1	NS	NS
35	<i>NOS3</i>	7.11E-1	NS	NS
36	<i>TXNRD2</i>	7.21E-1	NS	NS
37	<i>SOD2</i>	7.24E-1	NS	NS
38	<i>CAV2</i>	7.28E-1	NS	NS
39	<i>CYP1B1</i>	7.55E-1	NS	NS
40	<i>TP53</i>	7.57E-1	NS	NS
41	<i>OPTN-TBK1-TLR4</i>	7.68E-1	NS	NS
42	<i>TGFBR3</i>	7.68E-1	NS	NS
43	<i>OPTN</i>	7.82E-1	NS	NS
44	<i>RPGRIP1</i>	7.89E-1	NS	NS
45	<i>COL5A1</i>	8.02E-1	NS	NS
46	<i>MMP9</i>	8.26E-1	NS	NS
47	<i>CAV1-CAV2</i>	8.31E-1	NS	NS
48	<i>MUTYH</i>	8.34E-1	NS	NS
49	<i>GAS7</i>	8.34E-1	NS	NS
50	<i>CDKN2B-AS1</i>	8.42E-1	NS	NS
51	<i>NTM</i>	8.42E-1	NS	NS
52	<i>MMP1</i>	8.72E-1	NS	NS
53	<i>CDKN2A</i>	9.10E-1	NS	NS
54	<i>SRBD1</i>	9.43E-1	NS	NS
55	<i>XRCC1</i>	9.59E-1	NS	NS
56	<i>ABCA1</i>	9.80E-1	NS	NS
57	<i>CNTN4</i>	9.82E-1	NS	NS
58	<i>ARHGEF12</i>	9.84E-1	NS	NS
59	<i>CDKN2B</i>	9.87E-1	NS	NS
60	<i>PAK5</i>	9.87E-1	NS	NS
61	<i>DMXL1</i>	9.91E-1	NS	NS
62	<i>CD5</i>	9.98E-1	NS	NS
63	<i>SH3PXD2B</i>	9.98E-1	NS	NS
64	<i>COL8A2</i>	9.98E-1	NS	NS
65	<i>COL1A1</i>	9.99E-1	NS	2.62E-1
66	<i>APOE</i>	9.99E-1	NS	5.35E-1
67	<i>OPTC</i>	NS	NS	NS
68	<i>ASB10</i>	NS	NS	NS
69	<i>ZNF469</i>	NS	NS	NS
70	<i>TULP3</i>	NS	NS	NS

APPENDIX

Table B.7: *CNVKit output for all samples and genes. Samples and regions which were detected to have CN gains or losses were displayed.*

**3a refers to a subset of 12 samples in batch 3 which were prepared with reagents as per NRCC protocol.*

3b refers to a subset of 72 samples which were prepared with diluted reagents.

Batch	Sample	Chrom	Start	End	Length	Gene(s)	CN	Depth
1	AG003	chr1	171635396	171655416	20020	<i>MYOC</i>	3	773.2
4	FG093	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	984.1
4	FG093	chr15	85578929	85581842	2913	<i>AKAP13</i>	3	1058.5
4	FG093	chr1	171635396	171655416	20020	<i>MYOC</i>	3	826.3
4	FG093	chr16	76558489	88439329	11880840	<i>CNTNAP4,ZNF469</i>	3	1275.8
1	FG099	chr5	83538810	111111278	27572468	<i>VCAN,VCAN-AS1,WDR36</i>	3	1105.5
1	FG100	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	828.1
1	FG102	chr2	110123793	110204968	81175	<i>NPHP1</i>	1	350.6
4	FG178	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	901.1
4	FG178	chr1	171635662	171655416	19754	<i>MYOC</i>	3	647.4
1	FG182	chr10	68231221	98459557	30228336	<i>ATOH7,HPSE2</i>	3	469.7
1	FG185	chr10	68231221	98459557	30228336	<i>ATOH7,HPSE2</i>	3	516.9
1	FG250	chr1	45329308	45340254	10946	<i>MUTYH,MUTYH,TOE1</i>	3	1026.6
4	FG278	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	667.4
4	FG309	chr1	171635396	171655416	20020	<i>MYOC</i>	3	905.3
4	GB036	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	684.5
4	GB066	chr1	171635396	171655149	19753	<i>MYOC</i>	3	801.7
4	GB069	chr1	171635396	171655416	20020	<i>MYOC</i>	3	907.5
1	GB086	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	591.4
1	GB086	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	1058.0
1	GB106	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	490.9
1	GB143	chr5	83538810	111111278	27572468	<i>VCAN,VCAN-AS1,WDR36</i>	3	874.6
1	GB174	chr9	97428497	104794510	7366013	<i>TDRD7,ABCA1</i>	3	1183.9
1	GL162	chr7	151008929	151187722	178793	<i>NOS3,NOS3,ATG9B,ASB10</i>	3	705.1
3a	GL183	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	519.4
3a	GL184	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	528.6
3a	GL184	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	631.5
3a	GL184	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OTPC</i>	3	430.8
3a	GL186	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	337.4
3a	GL186	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	646.9
3a	GL188	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	534.9
3a	GL188	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	359.0
3a	GL188	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	631.4

APPENDIX

3a	GL188	chr1	171635396	171655683	20287	<i>MYOC</i>	3	437.3
3a	GL189	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	547.4
3a	GL189	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	619.1
3a	GL189	chr9	117712388	133256126	15543738	<i>TLR4,ABO</i>	3	655.9
3a	GL189	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPFS2,OPTC</i>	3	441.5
4	GL191	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	801.9
4	GL191	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1055.4
4	GL191	chr1	171635396	171655416	20020	<i>MYOC</i>	3	716.4
3a	GL192	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	631.2
3a	GL192	chr1	165728025	165768515	40490	<i>TMCO1</i>	1	290.8
3a	GL194	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	314.5
3a	GL194	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	597.3
3a	GL202	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	539.0
3a	GL202	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	346.1
3a	GL202	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	628.7
3a	GL202	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	426.7
3a	GL208	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	396.8
3a	GL208	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	268.1
3a	GL208	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	475.8
3a	GL208	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	463.2
3a	GL208	chr1	165768681	203496236	37727555	<i>TMCO1,MYOC,AXDND1, NPFS2,OPTC</i>	3	325.0
3a	GL211	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	512.9
3a	GL211	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	324.7
3a	GL211	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	597.6
3a	GL211	chr1	165728025	165768515	40490	<i>TMCO1</i>	1	283.2
3a	GL219	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	568.4
3a	GL219	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	358.2
3a	GL219	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	649.3
3a	GL219	chr1	165768681	179551371	13782690	<i>TMCO1,MYOC,AXDND1,NPFS2</i>	3	455.2
3a	GL220	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	594.6
3a	GL220	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	685.0
3a	GL228	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	508.5
3a	GL228	chr1	165728025	165768515	40490	<i>TMCO1</i>	1	239.4
3a	GL234	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	950.3
3a	GL234	chr1	171635396	171655416	20020	<i>MYOC</i>	3	1035.1
3a	GL245	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	771.7
3a	GL245	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	633.8

APPENDIX

3a	GL247	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	531.4
3a	GL247	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	330.6
3a	GL247	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	600.9
3a	GL251	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	500.5
3a	GL251	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	318.0
3a	GL251	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	610.2
3a	GL251	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	412.2
3a	GL256	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	400.9
3a	GL260	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	478.8
3a	GL260	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	310.2
3a	GL260	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	555.6
3a	GL260	chr1	165728025	165768269	40244	<i>TMCO1</i>	1	225.9
3a	GL260	chr1	165768481	203503717	37735236	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	389.7
3a	GL261	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	508.2
3a	GL261	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	612.5
3a	GL261	chr1	165728025	165768515	40490	<i>TMCO1</i>	1	274.2
3a	GL261	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	416.8
3a	GL264	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	368.2
3a	GL264	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	413.6
3a	GL264	chr10	68231221	98459557	30228336	<i>ATOX1,HPSE2</i>	3	175.2
3a	GL273	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	633.1
3a	GL273	chr9	117712388	133256126	15543738	<i>TLR4,ABO</i>	3	650.0
3a	GL282	chr16	88427999	88439329	11330	<i>ZNF469</i>	3	276.3
3a	GL282	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	520.2
3a	GL285	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	513.0
3a	GL285	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	613.1
3a	GL291	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	382.8
3a	GL291	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	238.1
3a	GL291	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	450.9
3a	GL291	chr1	171635396	171655416	20020	<i>MYOC</i>	3	307.7
3a	GL293	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	371.6
3a	GL293	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	247.0
3a	GL293	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	424.7
3a	GL293	chr1	171635396	171655416	20020	<i>MYOC</i>	3	292.0
3a	GL294	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	485.2
3a	GL294	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	305.4
3a	GL294	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	556.1

APPENDIX

3a	GL294	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	389.4
3a	GL294	chr9	117712388	133257542	15545154	<i>TLR4,ABO</i>	3	531.8
3a	GL300	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	1003.5
3a	GL300	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	840.8
3a	GL300	chr1	171635396	171655416	20020	<i>MYOC</i>	3	828.3
3a	GL300	chr1	36097571	45333324	9235753	<i>COL8A2,MUTYH</i>	3	761.8
3a	GL301	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	467.7
3a	GL301	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	292.9
3a	GL301	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	547.3
3a	GL301	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	376.8
4	GL317	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1020.3
3a	GL318	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	408.3
3a	GL318	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	275.4
3a	GL318	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	480.4
3a	GL318	chr1	171635662	171655416	19754	<i>MYOC</i>	3	335.5
3a	GL320	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	449.2
3a	GL320	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	270.9
3a	GL320	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	532.1
3a	GL320	chr1	165768681	179551371	13782690	<i>TMCO1,MYOC,AXDND1, NPHS2</i>	3	361.0
3a	GL329	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	451.9
3a	GL329	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	271.9
3a	GL329	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	510.8
3a	GL329	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	357.4
3a	GL329	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	507.9
3a	GL331	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	478.2
3a	GL331	chr16	88427999	88439329	11330	<i>ZNF469</i>	3	280.1
3a	GL331	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	562.6
3a	GL331	chr9	117712388	133256126	15543738	<i>TLR4,ABO</i>	3	578.6
3a	GL334	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	512.4
3a	GL334	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	328.0
3a	GL334	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	584.6
3a	GL334	chr1	171635396	171655683	20287	<i>MYOC</i>	3	387.0
3a	GL336	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	516.8
3a	GL336	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	317.2
3a	GL336	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	586.3
3a	GL337	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	531.2
3a	GL337	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	319.0
3a	GL337	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	622.3

APPENDIX

3a	GL344	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	490.1
3a	GL344	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	347.2
3a	GL344	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	575.3
3a	GL344	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	493.8
3a	GL344	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	403.8
3a	GL346	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	508.7
3a	GL346	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	351.8
3a	GL346	chr1	165768481	171655416	5886935	<i>TMCO1,MYOC</i>	3	403.6
3a	GL347	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	344.2
3a	GL347	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	623.9
3a	GL351	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	465.3
3a	GL351	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	351.4
3a	GL351	chr1	171635396	171655416	20020	<i>MYOC</i>	3	378.8
3a	GL351	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	465.1
3a	GL356	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	470.7
3a	GL356	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	324.4
3a	GL356	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	523.0
3a	GL356	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	385.5
3a	GL358	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	483.3
3a	GL358	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	335.1
3a	GL358	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	390.5
3a	GL360	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	379.5
3a	GL360	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	627.6
3a	GL360	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	446.8
3a	GL361	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	246.3
3a	GL361	chr1	171649276	171655416	6140	<i>MYOC</i>	3	416.9
3a	GL361	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	472.5
3a	GL361	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	384.1
3a	GL361	chr1	171635396	171649276	13880	<i>MYOC</i>	3	396.1
3a	GL361	chr22	19877107	20787012	909905	<i>TXNRD2, TXNRD2, COMT, PI4KA, SERPIND1</i>	3	505.4
3a	GL362	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	437.9
3a	GL362	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	264.3
3a	GL371	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	506.2
3a	GL371	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	334.7
3a	GL373	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	506.9
3a	GL373	chr10	68231221	98459557	30228336	<i>ATOH7, HPSE2</i>	3	225.8
3a	GL385	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	350.4
3a	GL385	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	645.8

APPENDIX

3a	GL385	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	451.9
3a	GL385	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	653.7
4	GL386	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	758.9
4	GL424	chr9	117704472	117714645	10173	<i>TLR4</i>	3	887.2
4	GL424	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	715.3
4	GL424	chr14	50302998	60511249	10208251	<i>L2HGDH,SIX6</i>	3	748.2
4	GL424	chr2	18563798	45389599	26825801	<i>NT5C1B-RDH14,NT5C1B, CYP1B1,SRBD1</i>	3	750.7
4	GL426	chr9	117704472	117714645	10173	<i>TLR4</i>	3	892.7
4	GL426	chr12	884763	2940723	2055960	<i>WNK1,TULP3</i>	1	433.4
4	GL426	chr20	6119440	9644328	3524888	<i>FERMT1,PAK5</i>	3	713.8
4	GL426	chr2	18563798	45418541	26854743	<i>NT5C1B-RDH14,NT5C1B, CYP1B1,SRBD1</i>	3	711.1
3a	GL429	chr10	13109122	13136863	27741	<i>OPTN</i>	1	192.1
3a	GL429	chr12	111452814	111599514	146700	<i>ATXN2,ATXN2,ATXN2-AS</i>	1	143.2
3a	GL429	chr20	6119440	9644328	3524888	<i>FERMT1,PAK5</i>	3	324.1
3a	GL429	chr15	34236746	48644769	14408023	<i>SLC12A6,FBN1</i>	3	335.3
3a	GL429	chr17	7669611	22299893	14630282	<i>TP53,GAS7,FAM27E5</i>	1	190.1
3a	GL429	chr2	18563798	45413293	26849495	<i>NT5C1B-RDH14,NT5C1B, CYP1B1,SRBD1</i>	3	335.5
4	GL430	chr9	117704472	117714645	10173	<i>TLR4</i>	3	794.3
4	GL430	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	592.1
4	GL430	chr17	75237451	75261587	24136	<i>GGA3</i>	1	291.3
4	GL430	chr12	64455870	64501378	45508	<i>TBK1</i>	1	311.3
4	GL430	chr9	21967138	22121094	153956	<i>CDKN2A-AS1,CDKN2A-AS1, CDKN2A,CDKN2A,CDKN2B-AS1, CDKN2B-AS1,CDKN2B</i>	3	557.7
4	GL430	chr12	884763	2939441	2054678	<i>WNK1,TULP3</i>	1	340.3
4	GL430	chr20	6119440	9644328	3524888	<i>FERMT1,PAK5</i>	3	625.8
4	GL430	chr16	70269676	76553501	6283825	<i>AARS,CNTNAP4</i>	3	562.3
4	GL430	chr15	34236746	48644769	14408023	<i>SLC12A6,FBN1</i>	3	630.1
4	GL430	chr2	18563798	45418541	26854743	<i>NT5C1B-RDH14,NT5C1B, CYP1B1,SRBD1</i>	3	619.5
3a	GL466	chr1	171635396	171655416	20020	<i>MYOC</i>	3	586.1
4	GL467	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	324.0
4	GL467	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	287.8
3a	GL510	chr1	171635396	171655683	20287	<i>MYOC</i>	3	490.6
4	GL511	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1224.0

APPENDIX

4	GL517	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1065.8
4	GL517	chr1	171635396	171655416	20020	<i>MYOC</i>	3	769.1
3b	GL531	chr20	46008926	46016365	7439	<i>MMP9,MMP9,LOC100128028</i>	3	485.8
3b	GL531	chr14	87963383	87993499	30116	<i>GALC</i>	3	496.5
4	GL552	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	982.4
4	GL552	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1091.6
4	GL552	chr1	171635396	171655416	20020	<i>MYOC</i>	3	839.5
4	GL557	chr1	171635396	171655149	19753	<i>MYOC</i>	3	590.3
4	GL564	chr1	171635396	171655416	20020	<i>MYOC</i>	3	835.9
4	GL592	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	941.7
4	GL592	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	1000.9
2	GL607	chr1	171635396	203503717	31868321	<i>MYOC,AXDND1,NPHS2,OPTC</i>	3	778.6
2	GL610	chr1	171635396	203498839	31863443	<i>MYOC,AXDND1,NPHS2,OPTC</i>	3	868.3
2	GL611	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	892.5
2	GL611	chr1	171635396	171655416	20020	<i>MYOC</i>	3	825.6
2	GL611	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	952.4
4	GL612	chr1	171635662	171655683	20021	<i>MYOC</i>	3	618.2
2	GL633	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	421.6
2	GL633	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	640.9
2	GL633	chr1	171635396	171655683	20287	<i>MYOC</i>	3	583.4
2	GL635	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	571.9
2	GLF001	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	588.1
2	GLF003	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	534.3
2	QG007	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	936.8
2	QG007	chr1	171635396	171655683	20287	<i>MYOC</i>	3	835.9
4	QG022	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	334.3
2	QG025	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	777.3
2	QG025	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	962.2
2	QG025	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	677.8
2	QG027	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	751.4
2	QG027	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	664.8
2	QG029	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	727.5
2	QG029	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	935.9
2	QG032	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	776.1
2	QG032	chr15	48596282	48644769	48487	<i>FBN1</i>	1	474.4
2	QG036	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	799.2
4	QG040	chr1	171635396	171655416	20020	<i>MYOC</i>	3	592.1

APPENDIX

2	QG041	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	364.3
2	QG041	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	410.3
2	QG041	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	471.4
2	QG041	chr1	171635662	171655683	20021	<i>MYOC</i>	3	339.2
2	QG046	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	505.5
2	QG046	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	520.5
2	QG046	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	651.8
2	QG046	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	469.5
2	QG046	chr9	117712388	133257542	15545154	<i>TLR4,ABO</i>	3	561.1
2	QG055	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	495.4
2	QG055	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	465.1
2	QG055	chr1	165728025	165768515	40490	<i>TMCO1</i>	1	306.8
2	QG055	chr1	165768681	171655416	5886735	<i>TMCO1,MYOC</i>	3	459.6
2	QG055	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	568.4
2	QG055	chr10	68231221	98459557	30228336	<i>ATOX1,HPSE2</i>	3	323.6
2	QG068	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	424.0
2	QG068	chr1	171635396	171655416	20020	<i>MYOC</i>	3	648.0
2	QG088	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	915.1
4	QG092	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	1000.5
2	QG095	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	685.1
2	QG095	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	687.8
2	QG095	chr1	171635396	171655683	20287	<i>MYOC</i>	3	624.5
2	QG095	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	712.8
2	QG119	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	679.5
2	QG122	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	983.1
2	QG122	chr1	171635396	171655416	20020	<i>MYOC</i>	3	861.0
2	QG123	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	981.8
2	QG124	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	450.0
2	QG126	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	355.0
2	QG126	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	577.8
2	QG126	chr1	171635396	171655683	20287	<i>MYOC</i>	3	498.2
2	QG128	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	444.2
2	QG128	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	483.1
2	QG128	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	574.3
2	QG128	chr1	171635396	171655683	20287	<i>MYOC</i>	3	422.3
2	QG128	chr9	117708562	133257542	15548980	<i>TLR4,ABO</i>	3	527.7
2	QG135	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	328.3
2	QG135	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	478.1
2	QG135	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	502.0

APPENDIX

2	QG135	chr1	171635396	171655683	20287	MYOC	3	445.1
2	QG136	chr16	88427470	88439329	11859	ZNF469	3	642.5
2	QG136	chr1	165768681	171655683	5887002	TMCO1,MYOC	3	628.7
2	QG137	chr16	88427470	88439329	11859	ZNF469	3	613.1
2	QG137	chr1	165768681	171655683	5887002	TMCO1,MYOC	3	586.4
2	QG140	chr22	46256563	46263322	6759	PKDREJ	3	630.2
2	QG140	chr16	88427470	88439329	11859	ZNF469	3	511.6
2	QG141	chr22	46256563	46263322	6759	PKDREJ	3	773.3
2	QG141	chr16	88427470	88439329	11859	ZNF469	3	634.9
2	QG141	chr1	171635396	171655149	19753	MYOC	3	675.4
2	QG141	chr9	117708562	133257542	15548980	TLR4,ABO	3	840.1
2	QG156	chr1	36097571	36099487	1916	COL8A2	3	373.6
2	QG156	chr22	46256563	46263322	6759	PKDREJ	3	535.5
2	QG156	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	642.9
2	QG156	chr1	171635396	171655683	20287	MYOC	3	497.5
2	QG156	chr15	85485720	85582107	96387	AKAP13	3	556.3
2	QG156	chr16	76558489	88439329	11880840	CNTNAP4,ZNF469	3	538.1
2	QG159	chr22	46256563	46263322	6759	PKDREJ	3	705.4
2	QG159	chr16	88427470	88439329	11859	ZNF469	3	629.0
2	QG159	chr15	85485720	85582107	96387	AKAP13	3	721.5
2	QG159	chr1	165768681	171655416	5886735	TMCO1,MYOC	3	628.8
2	QG159	chr1	36097571	45333603	9236032	COL8A2,MUTYH	3	600.7
2	QG162	chr1	36097571	36099487	1916	COL8A2	3	336.2
2	QG162	chr22	46256563	46263322	6759	PKDREJ	3	434.8
2	QG162	chr16	88427470	88439329	11859	ZNF469	3	488.2
2	QG162	chr1	171635396	171655683	20287	MYOC	3	407.2
2	QG162	chr9	117712388	133257542	15545154	TLR4,ABO	3	473.6
4	QG163	chr1	36097571	45333603	9236032	COL8A2,MUTYH	3	591.0
2	QG164	chr1	36097571	36099487	1916	COL8A2	3	318.8
2	QG164	chr22	46256563	46263322	6759	PKDREJ	3	503.0
2	QG164	chr16	88427470	88439329	11859	ZNF469	3	472.3
2	QG164	chr15	85485720	85581842	96122	AKAP13	3	510.0
2	QG164	chr9	117712670	133257542	15544872	TLR4,ABO	3	568.7
2	QG164	chr1	165768681	203503717	37735036	TMCO1,MYOC,AXDND1, NPHS2,OPTC	3	446.0
2	QG166	chr1	36097571	36099487	1916	COL8A2	3	320.8
2	QG166	chr22	46256563	46263322	6759	PKDREJ	3	409.4
2	QG166	chr16	88427470	88439329	11859	ZNF469	3	478.2
2	QG166	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	515.3

APPENDIX

2	QG166	chr1	171635396	171655683	20287	MYOC	3	394.8
2	QG168	chr1	36097571	36099487	1916	COL8A2	4	362.3
2	QG168	chr22	46256563	46263322	6759	PKDREJ	3	438.7
2	QG168	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	543.5
2	QG168	chr15	85485720	85582107	96387	AKAP13	3	462.8
2	QG168	chr1	165768681	171655683	5887002	TMCO1,MYOC	3	426.5
2	QG168	chr16	76558489	88439329	11880840	CNTNAP4,ZNF469	3	517.2
2	QG168	chr9	117712388	133256356	15543968	TLR4,ABO	3	507.2
2	QG169	chr22	46256563	46263322	6759	PKDREJ	3	320.3
2	QG169	chr16	88427470	88439329	11859	ZNF469	3	326.1
2	QG169	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	404.4
2	QG169	chr15	85485720	85582107	96387	AKAP13	3	325.7
2	QG169	chr1	165768681	171655416	5886735	TMCO1,MYOC	3	298.7
2	QG169	chr9	117712388	133257542	15545154	TLR4,ABO	3	357.0
2	QG171	chr1	36097571	36099487	1916	COL8A2	4	252.8
2	QG171	chr22	46256563	46263322	6759	PKDREJ	3	318.9
2	QG171	chr16	88427734	88439329	11595	ZNF469	3	377.5
2	QG171	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	377.6
2	QG171	chr15	85485720	85581842	96122	AKAP13	3	323.6
2	QG171	chr1	165768681	171655683	5887002	TMCO1,MYOC	3	294.2
2	QG171	chr9	117712670	133256356	15543686	TLR4,ABO	3	381.2
2	QG173	chr22	46256563	46263322	6759	PKDREJ	3	504.2
2	QG173	chr16	88427734	88439329	11595	ZNF469	3	477.5
2	QG173	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	596.8
2	QG173	chr1	171635662	171655683	20021	MYOC	3	447.6
2	QG173	chr9	117708562	133256356	15547794	TLR4,ABO	3	566.7
2	QG174	chr1	36097571	36099487	1916	COL8A2	4	313.2
2	QG174	chr22	46256563	46263322	6759	PKDREJ	3	409.5
2	QG174	chr16	88427470	88439329	11859	ZNF469	3	432.8
2	QG174	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	484.3
2	QG174	chr15	85485720	85582107	96387	AKAP13	3	420.4
2	QG174	chr1	165768681	171655683	5887002	TMCO1,MYOC	3	382.8
2	QG174	chr9	117712388	133257542	15545154	TLR4,ABO	3	454.1
2	QG194	chr22	46256563	46263322	6759	PKDREJ	3	766.3
2	QG194	chr1	171635396	171655683	20287	MYOC	3	671.3
2	QGF001	chr22	46256563	46263322	6759	PKDREJ	3	667.4
2	QGF001	chr16	88427470	88439329	11859	ZNF469	3	690.4
2	QGF001	chr16	51137114	51151241	14127	MIR548AI,SALL1	3	779.2
2	QGF001	chr1	171635396	171655416	20020	MYOC	3	631.9

APPENDIX

2	QGF001	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	669.0
2	QGF002	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	832.8
2	QGF002	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	682.1
2	QGF002	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	971.2
2	QGF002	chr15	85485720	85581842	96122	<i>AKAP13</i>	3	843.1
2	QGF002	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	906.3
2	QGF002	chr1	165768681	203496236	37727555	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	718.5
2	QGF003	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	838.5
2	QGF003	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	661.2
2	QGF003	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1008.5
2	QGF003	chr1	171635396	171655416	20020	<i>MYOC</i>	3	725.1
2	QGF003	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	845.6
2	QGF003	chr9	117708562	133256356	15547794	<i>TLR4,ABO</i>	3	898.4
2	QGF004	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	841.2
2	QGF004	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	682.1
2	QGF004	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	998.2
2	QGF004	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	822.3
2	QGF004	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	726.5
2	QGF004	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	894.3
2	QGF007	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	565.6
2	QGF007	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	476.4
2	QGF007	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	647.2
2	QGF007	chr1	165768681	179551371	13782690	<i>TMCO1,MYOC,AXDND1, NPHS2</i>	3	498.0
2	QGF009	chr16	88427734	88433150	5416	<i>ZNF469</i>	3	468.0
2	QGF009	chr16	88433150	88439329	6179	<i>ZNF469</i>	3	442.4
2	QGF009	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	530.6
2	QGF009	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	614.3
2	QGF009	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	460.1
2	QGF009	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	581.5
2	QGF010	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	592.4
2	QGF010	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	567.5
2	QGF010	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	708.1
2	QGF010	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	663.5
2	QGF010	chr1	165768681	203496236	37727555	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	525.0
2	QGF012	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	738.2
2	QGF012	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	736.2

APPENDIX

2	QGF016	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	778.2
2	QGF016	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	622.8
2	QGF016	chr1	171635396	203496236	31860840	<i>MYOC,AXDND1,NPHS2,OPTC</i>	3	690.5
2	QGF018	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	974.1
2	QGF018	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	773.8
2	QGF018	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1137.0
2	QGF018	chr1	171635396	171655683	20287	<i>MYOC</i>	3	835.8
2	QGF018	chr9	117708562	133256356	15547794	<i>TLR4,ABO</i>	3	1058.2
2	QGF023	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	931.0
2	QGF023	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	728.6
2	QGF023	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1038.3
2	QGF023	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	815.4
2	QGF023	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	967.3
2	QGF024	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	651.6
2	QGF024	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	571.6
2	QGF024	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	724.7
2	QGF024	chr1	171635396	171655683	20287	<i>MYOC</i>	3	574.8
2	QGF024	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	629.6
2	QGF024	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	686.9
2	QGF025	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	945.7
2	QGF025	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	838.5
2	QGF025	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1066.0
2	QGF025	chr1	171635396	171655416	20020	<i>MYOC</i>	3	827.0
2	QGF025	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	891.7
2	QGF025	chr9	117712388	133256126	15543738	<i>TLR4,ABO</i>	3	1012.7
2	QGF028	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	763.1
2	QGF028	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	668.5
2	QGF028	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	893.4
2	QGF028	chr1	171635396	171655683	20287	<i>MYOC</i>	3	667.6
2	QGF028	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	743.1
2	QGF028	chr9	117708562	133256356	15547794	<i>TLR4,ABO</i>	3	820.9
2	QGF029	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	783.2
2	QGF029	chr16	88427999	88439329	11330	<i>ZNF469</i>	3	641.1
2	QGF029	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	888.9
2	QGF029	chr1	165768681	203496236	37727555	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	688.6
2	QGF033	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	815.5
2	QGF033	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	795.2
2	QGF033	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	970.7

APPENDIX

2	QGF033	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	741.3
2	QGF034	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	898.8
2	QGF034	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	834.2
2	QGF034	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1061.6
2	QGF034	chr1	165768681	203496236	37727555	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	787.8
2	SG001	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	611.6
2	SG001	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	523.9
2	SG001	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	681.2
2	SG001	chr14	87934734	87950748	16014	<i>GALC</i>	1	254.2
2	SG001	chr1	165768681	171655683	5887002	<i>TMCO1,MYOC</i>	3	552.5
4	SG003	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	776.8
2	SG004	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	410.8
2	SG004	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	782.4
2	SG004	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	653.3
2	SG004	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	858.8
2	SG004	chr1	171635396	171655683	20287	<i>MYOC</i>	3	700.5
2	SG004	chr9	117712388	133256356	15543968	<i>TLR4,ABO</i>	3	821.3
4	SG012	chr1	36097571	36099487	1916	<i>COL8A2</i>	3	329.0
4	SG012	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	347.0
4	SG012	chr1	171635396	171655416	20020	<i>MYOC</i>	3	297.4
4	SG012	chr2	110123793	110204968	81175	<i>NPHP1</i>	0	118.0
4	SG012	chr16	76558489	88439329	11880840	<i>CNTNAP4,ZNF469</i>	3	396.9
2	TG007	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	620.4
2	TG007	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	494.9
2	TG007	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	711.2
2	TG007	chr9	117708562	133257542	15548980	<i>TLR4,ABO</i>	3	670.2
2	TG007	chr1	165768681	203503717	37735036	<i>TMCO1,MYOC,AXDND1, NPHS2,OPTC</i>	3	517.8
2	TG008	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	607.7
2	TG008	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	652.9
2	TG009	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	721.8
2	TG009	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	539.4
2	TG009	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	830.3
2	TG009	chr1	165768681	179551371	13782690	<i>TMCO1,MYOC,AXDND1,NPHS2</i>	3	621.2
2	TG009	chr9	117708562	133256356	15547794	<i>TLR4,ABO</i>	3	778.9
2	TG011	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	960.2
2	TG011	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	817.4
2	TG011	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	1170.8

APPENDIX

2	TG011	chr1	171635396	171655416	20020	<i>MYOC</i>	3	832.0
2	TG011	chr9	117708562	133256126	15547564	<i>TLR4, ABO</i>	3	1050.4
2	TG022	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	728.6
2	TG022	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	551.2
2	TG022	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	775.7
2	TG022	chr9	117712388	133256126	15543738	<i>TLR4, ABO</i>	3	756.6
2	TG022	chr1	165768681	203496236	37727555	<i>TMCO1, MYOC, AXDND1, NPHS2, OPTC</i>	3	600.8
2	TG028	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	851.3
2	TG028	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	691.2
2	TG028	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	986.1
2	TG028	chr1	171635396	171655683	20287	<i>MYOC</i>	3	731.6
2	TG028	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	828.3
2	TG028	chr9	117708562	133256356	15547794	<i>TLR4, ABO</i>	3	929.5
2	TG030	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	870.6
2	TG030	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	734.2
2	TG030	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	1054.2
2	TG030	chr1	171635396	171655416	20020	<i>MYOC</i>	3	786.2
2	TG030	chr9	117708562	133256356	15547794	<i>TLR4, ABO</i>	3	946.4
2	TG031	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	1010.9
2	TG031	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	957.7
2	TG031	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	1247.3
2	TG031	chr1	171635396	171655416	20020	<i>MYOC</i>	3	906.4
2	TG031	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	1048.1
2	TG031	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	1090.3
2	TG036	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	1011.9
2	TG036	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	867.2
2	TG036	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	1177.5
2	TG036	chr1	171635662	171655683	20021	<i>MYOC</i>	3	871.3
2	TG042	chr16	88427999	88439329	11330	<i>ZNF469</i>	3	491.0
2	TG042	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	610.0
2	TG045	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	679.0
2	TG045	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	528.5
2	TG045	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	788.3
2	TG045	chr1	165768681	171655683	5887002	<i>TMCO1, MYOC</i>	3	608.6
2	TG045	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	751.0
2	TG049	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	512.9
2	TG049	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	457.4
2	TG049	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	581.0

APPENDIX

2	TG049	chr1	171635396	171655683	20287	<i>MYOC</i>	3	467.3
2	TG049	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	497.8
2	TG049	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	549.8
2	TG074	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	644.4
2	TG074	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	486.5
2	TG074	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	726.3
2	TG074	chr1	165768681	171655683	5887002	<i>TMCO1, MYOC</i>	3	575.6
2	TG074	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	696.1
2	TG078	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	544.6
2	TG078	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	425.9
2	TG078	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	629.3
2	TG078	chr1	165768681	171655683	5887002	<i>TMCO1, MYOC</i>	3	491.6
2	TG078	chr9	117712388	133256126	15543738	<i>TLR4, ABO</i>	3	644.1
2	TG079	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	549.2
2	TG079	chr16	88427734	88439329	11595	<i>ZNF469</i>	3	430.2
2	TG079	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	645.1
2	TG079	chr1	165768681	179551371	13782690	<i>TMCO1, MYOC, AXDND1, NPHS2</i>	3	490.5
2	TG079	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	616.3
2	TG087	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	712.3
2	TG087	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	568.0
2	TG087	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	837.7
2	TG087	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	705.4
2	TG087	chr1	165768681	171655683	5887002	<i>TMCO1, MYOC</i>	3	630.4
2	TG087	chr9	117712388	133256356	15543968	<i>TLR4, ABO</i>	3	782.9
2	TG093	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	776.1
2	TG093	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	613.4
2	TG093	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	917.9
2	TG093	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	728.7
2	TG093	chr1	165768681	171655683	5887002	<i>TMCO1, MYOC</i>	3	675.5
2	TG093	chr9	117712388	133257542	15545154	<i>TLR4, ABO</i>	3	810.5
2	TG095	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	766.6
2	TG095	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	621.9
2	TG095	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	903.3
2	TG095	chr1	171635396	171655683	20287	<i>MYOC</i>	3	674.7
2	TG095	chr9	117708562	133256126	15547564	<i>TLR4, ABO</i>	3	869.4
2	TG100	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	664.2
2	TG100	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	609.3
2	TG100	chr16	51137114	51151241	14127	<i>MIR548AI, SALL1</i>	3	790.3

APPENDIX

2	TG100	chr1	171635396	171655416	20020	<i>MYOC</i>	3	617.2
2	TG100	chr15	85485720	85582107	96387	<i>AKAP13</i>	3	693.5
2	WG015	chr22	46256563	46263322	6759	<i>PKDREJ</i>	3	811.7
2	WG015	chr16	88427470	88439329	11859	<i>ZNF469</i>	3	803.0
2	WG015	chr16	51137114	51151241	14127	<i>MIR548AI,SALL1</i>	3	931.9
2	WG015	chr1	171635396	203496236	31860840	<i>MYOC,AXDND1,NPHS2,OPTC</i>	3	699.8

C Appendix Supplementary Data

Table C.1: *Whole exome sequencing samples' VerifyBamID freemix results.*

SampleID	Freemix
NG045-1	0.00009
NG222	0.00014
NG156	0.00044
NG-PE	0.00048
NG241	0.00056
NG149-2	0.00065
Self-Colob001	0.00189
NRA	0.00226
RNAR1	0.0026
NG151-2	0.00406
NG062-4	0.01786

D Appendix Supplementary Data

Table D.1: *TruSight One batch 1 VerifyBamID freemix results.*

SampleID	Freemix
NG288	0.00761
NG346	0.00840
NG167	0.00873
NG234	0.00893
NG333	0.00931
NG257	0.00971
NG322	0.00971
NG309	0.00999
NG296	0.01008
NG250	0.01017
NG265	0.01047
NG251	0.01082
NG263	0.01092
NG356	0.01164
NG213	0.01284
NG344	0.01406
NG327	0.01486
NG340	0.01511
NG272	0.01661
NG230	0.01975
NG178	0.02391
NG270	0.02708

Table D.2: *TruSight One batch 2 VerifyBamID freemix results.*

SampleID	Freemix
NG244	0.00438
NG285	0.00659
NG299	0.00671
NG195	0.00690
NG172	0.00691
NG190	0.00694
NG289	0.00709
NG218	0.00711
NG232	0.00742
NG178	0.00822
CL036	0.00941
OSTEO	0.00972
CASR002	0.01065
CL033	0.01100
CL035	0.01111
NG191	0.01120
CL039	0.01122
CL027	0.01332
CL021	0.01372
CL028	0.01431
NG367	0.01442
NG320	0.01444
CL040	0.01575
NG338	0.01580
CL038	0.01582
NG318	0.01828
CL030	0.01828
NG264	0.01888
CL022	0.01911
CL037	0.02071
NG335	0.02638
CL025	0.02726
NG306	0.02746
CL024	0.04605
NG302	0.15707

E Appendix Supplementary Data

Table E.1: *Primers designed for Sanger sequencing of NM_001310160:exon10: c.G858T: p.Q286H and NM_001310160: exon10: c.C856A: p.Q286K.*

Primers	5' ————— 3'
Forward	TGTCTAGTTTCTGAAGGTGC
Reverse	CCTTGTTTCAAGTCCATTC

Table E.2: *Percentage coverage at 20x depth across the 31 gene panel using the TruSight One BED file. The total reads for each sample and their percentage mapped to the TruSight One target region are listed. Samples are coloured to represent quality thresholds of $0 < x < 10\%$ (red), $10 < x < 80\%$ (orange) and $90 < x < 100\%$ (green).*

No.	Batch	Sample ID	Coverage 20x	Total reads	Reads mapped to target	%	Reads mapped to target +/- 150bp	%
1	1	NG167	0.9768	28601634	16056134	56.1	20958928	73.3
2	1	NG178	0.9792	32251168	19995766	62.0	26558640	82.4
3	1	NG213	0.9786	31513585	17614483	55.9	23507500	74.6
4	1	NG230	0.9728	26408609	14239229	53.9	18753218	71.0
5	1	NG234	0.9720	28531943	15994475	56.1	20938537	73.4
6	1	NG250	0.9838	35724914	22598237	63.3	29791790	83.4
7	1	NG251	0.9860	41865669	26541762	63.4	35157278	84.0
8	1	NG257	0.9827	34323203	21578920	62.9	28115737	81.9
9	1	NG263	0.9806	37729422	20083153	53.2	26566180	70.4
10	1	NG265	0.9852	34181596	21500121	62.9	28030107	82.0
11	1	NG270	0.9853	39647089	24626388	62.1	32490201	82.0
12	1	NG272	0.9811	30913945	19834134	64.2	25748938	83.3
13	1	NG288	0.9849	33773590	21290754	63.0	28004568	82.9
14	1	NG296	0.9867	45868679	26920993	58.7	35514935	77.4
15	1	NG309	0.9839	36260659	23484728	64.8	30629621	84.5
16	1	NG322	0.9840	39148234	25325656	64.7	32895507	84.0
17	1	NG327	0.9748	29096118	16757129	57.6	21803708	74.9
18	1	NG333	0.9807	35270852	20096112	57.0	26201983	74.3
19	1	NG340	0.9759	25806540	15460338	59.9	20077144	77.8
20	1	NG344	0.9780	33882219	19082609	56.3	24881237	73.4

APPENDIX

21	1	NG346	0.9717	27235895	15307135	56.2	20253363	74.4
22	1	NG356	0.9816	33072678	18514738	56.0	24197832	73.2
23	2	NG172	0.9832	30436760	22075293	72.5	28672979	94.2
24	2	NG176	0.0000	174065	131813	75.7	169858	97.6
25	2	NG178	0.9733	22296700	15997379	71.8	21111825	94.7
26	2	NG190	0.9797	26442220	20284871	76.7	24952095	94.4
27	2	NG191	0.9677	27049768	21473013	79.4	25413539	94.0
28	2	NG195	0.9633	20274878	14086655	69.5	18692647	92.2
29	2	NG218	0.9797	26694820	20060087	75.2	24969254	93.5
30	2	NG232	0.9841	30942230	22163718	71.6	29100867	94.1
31	2	NG244	0.9764	24423267	17644039	72.2	23053348	94.4
32	2	NG264	0.9655	17876500	12925977	72.3	16265642	91.0
33	2	NG285	0.9653	18988955	14140259	74.5	17833654	93.9
34	2	NG289	0.9603	18939814	13792729	72.8	18043148	95.3
35	2	NG299	0.9806	31028188	22527359	72.6	29266414	94.3
36	2	NG302	0.0037	852950	628184	73.7	723604	84.8
37	2	NG306	0.9746	25025463	17595487	70.3	22528478	90.0
38	2	NG318	0.9792	25882777	18672959	72.1	23624813	91.3
39	2	NG320	0.9393	13236088	9586553	72.4	12161571	91.9
40	2	NG335	0.9814	26499077	18445427	69.6	23837025	90.0
41	2	NG338	0.9655	19870887	13871415	69.8	17736815	89.3
42	2	NG367	0.9757	23672748	16696148	70.5	21386263	90.3
43	3	NG198	0.9607	21166022	13463365	63.6	16663030	78.7
44	3	NG219	0.8204	23075800	7512413	32.6	10114838	43.8
45	3	NG232	0.9545	29877343	10890255	36.5	13743925	46.0
46	3	NG277	0.8888	20274455	7680091	37.9	9737531	48.0
47	3	NG283	0.8943	22943196	8005814	34.9	9991188	43.6
48	3	NG348	0.9845	34348882	21546360	62.7	27014373	78.7
49	3	NG357	0.9885	41804731	25599347	61.2	32394045	77.5
50	3	NG361	0.9387	30559368	10086805	33.0	12847053	42.0
51	3	NG367	0.9472	29186118	10221931	35.0	12790496	43.8

APPENDIX

52	3	NG381	0.9107	25317939	8700245	34.4	10913152	43.1
53	3	NG383	0.9235	28789895	9930753	34.5	12708187	44.1
54	3	NG386	0.9494	30492590	11299212	37.1	14273225	46.8
55	3	NG387	0.9118	24557488	8871597	36.1	11112489	45.3
56	3	NG391	0.9477	29774810	11947355	40.1	14590373	49.0
57	3	NG399	0.9474	32171006	11515818	35.8	14682921	45.6
58	3	NG403	0.9863	39597785	24918179	62.9	31376348	79.2
59	3	NG406	0.9447	18795296	12169349	64.8	15083568	80.3
60	3	NG411	0.9723	25022225	15626385	62.5	19812341	79.2
61	3	NG416	0.9864	39682387	24571733	61.9	30851517	77.8
62	3	NG420	0.9517	18848402	12502236	66.3	15401372	81.7
63	3	NG423	0.9631	21492151	14135723	65.8	17260352	80.3
64	3	NG429	0.9686	24465942	15115407	61.8	18944175	77.4
65	3	NG430	0.9478	17944069	11449060	63.8	14133697	78.8
66	3	NG433	0.9518	20607875	12963778	62.9	16062239	77.9
67	3	NG438	0.9701	23787624	15042822	63.2	18887205	79.4
68	3	NG441	0.9614	19999738	12936600	64.7	16109545	80.6
69	3	NG445	0.9539	21264105	13793316	64.9	17110466	80.5
70	3	NG449	0.9603	21236130	13832127	65.1	17204308	81.0
71	3	NG451	0.9850	36089900	22416029	62.1	28115725	77.9
72	3	NG454	0.9851	39873459	24539427	61.5	30831405	77.3
73	4	NG280	0.9516	17652005	11781569	66.7	15366213	87.1
74	4	NG315	0.9626	21305346	14569395	68.4	19114254	89.7
75	5	NG474	0.9526	22013975	15098818	68.6	19511367	88.6
76	5	NG491	0.6053	7799158	5738408	73.6	7472851	95.8
77	5	NG492	0.9447	22845462	14393684	63.0	18871830	82.6
78	5	NG498	0.9425	21351241	13434706	62.9	17617090	82.5
79	5	NG508	0.9468	22613467	14610897	64.6	19252144	85.1
80	5	NG514	0.9305	19595952	12471477	63.6	15909033	81.2
81	5	NG518	0.9218	19008918	11829665	62.2	15558958	81.9
82	5	NG521	0.9598	23492102	14698299	62.6	19275360	82.1

APPENDIX

83	5	NG524	0.9556	22955217	14198087	61.9	18809378	81.9
84	5	NG534	0.9682	26331034	16448506	62.5	21573068	81.9
85	5	NG536	0.9397	21886137	13426393	61.4	17851433	81.6
86	5	NG540	0.9145	19543888	11419337	58.4	15121726	77.4
87	5	NG548	0.9230	20774914	13940366	67.1	18544261	89.3
88	5	NG551	0.9261	19589632	12480982	63.7	16360355	83.5
89	6	NG149-1	0.0138	54438782	1287548	2.4	1978604	3.6
90	6	NG150-2	0.0183	58019382	1325623	2.3	2026623	3.5
91	6	NG181	0.9512	17965919	11369292	63.3	14444151	80.4
92	6	NG200	0.9669	22134950	13885143	62.7	18220813	82.3
93	6	NG210	0.0004	41576159	782569	1.9	1199648	2.9
94	6	NG236	0.0000	23403882	488076	2.1	767065	3.3
95	6	NG267	0.0010	40281564	887152	2.2	1386672	3.4
96	6	NG284	0.9590	21901566	12628927	57.7	16478406	75.2
97	6	NG325	0.0214	1139247	819486	71.9	1007923	88.5
98	6	NG394	0.9558	22431408	13869113	61.8	17350961	77.4
99	6	NG395	0.9650	23276959	14572970	62.6	18615468	80.0
100	6	NG412	0.8233	10572103	6200873	58.7	7926372	75.0
101	6	NG456	0.9562	21573329	13052016	60.5	16916379	78.4
102	6	NG472	0.9687	24011454	14284243	59.5	18359114	76.5
103	6	NG477	0.9457	20040614	12944144	64.6	16065448	80.2
104	6	NG483	0.9724	24945048	16000219	64.1	20292692	81.4
105	6	NG504	0.8194	8776049	5426432	61.8	6926601	78.9
106	6	NG506	0.9065	14469586	8519387	58.9	11134732	77.0
107	6	NG512	0.9196	15971640	9683652	60.6	12842828	80.4
108	6	NG528	0.9107	15198177	8782347	57.8	11434857	75.2
109	6	NG530	0.9304	15439648	9237360	59.8	12056716	78.1
110	6	NG543	0.9393	16835159	9933625	59.0	12819990	76.2
111	6	NG545	0.9517	18971441	11098902	58.5	14588749	76.9
112	6	NG554	0.9246	15855675	9267141	58.5	12093370	76.3
113	6	NG556	0.7878	9154041	5409185	59.1	7032601	76.8

APPENDIX

114	6	NG557	0.9040	13927861	8056313	57.8	10456772	75.1
115	6	NG558	0.9499	18867533	11982097	63.5	15158175	80.3
116	6	NG559	0.9379	16578078	10474217	63.2	13329482	80.4
117	6	NG561	0.7738	8993540	5041083	56.1	6563960	73.0
118	6	NG563	0.6674	8409772	4559240	54.2	5958817	70.9
119	6	NG566	0.3984	4585575	2696902	58.8	3509002	76.5
120	6	NG571	0.5319	6548403	3478539	53.1	4581432	70.0
121	6	NG574	0.6896	8810013	4775679	54.2	6350761	72.1
122	6	NG575	0.4368	5315402	2755341	51.8	3732028	70.2
123	6	NG580	0.6328	6757056	3591004	53.1	4857200	71.9
124	6	NG582	0.7018	8266990	4414378	53.4	5779811	69.9
125	6	NG584	0.4343	4632860	2446439	52.8	3182228	68.7
126	6	NG587	0.8264	10598063	5356846	50.6	7094657	66.9
127	6	NG592	0.8769	11485385	6755215	58.8	8639762	75.2
128	6	NG594	0.0060	48822685	1137511	2.3	1771517	3.6
129	6	NG597	0.0002	34335986	818260	2.4	1271402	3.7
130	6	NG598	0.0001	28258339	703615	2.5	1087344	3.9
131	6	NG602	0.0011	39397207	874326	2.2	1369249	3.5
132	6	NG607	0.0045	43845838	995704	2.3	1556214	3.6
133	6	NG610	0.0012	34367493	797596	2.3	1239935	3.6
134	6	NG613	0.0133	52640345	1225831	2.3	1916169	3.6

Table E.3: *Pair-wise percentage concordance of variants between samples. Family member pairs are shown coloured in accordance with table 6.4. Red font highlights sample IDs which are duplicated in multiple batches.*

	B1-NG178	NG356	B2-NG178	NG218	B2-NG232	B2-NG367	NG198	NG219	B3-NG232	NG277	NG357	B3-NG367	NG411	NG181	NG200	NG412	NG558
B1-NG178		60.1	92.2	59.7	64.2	58.3	57.6	57.5	61.8	55.8	58.8	56.7	58.4	71.8	59.1	56.3	56.7
NG356	60.2		58.6	59.1	61.8	60.8	58.5	57.0	59.5	56.8	71.5	59.2	59.1	58.8	58.7	55.6	56.5
B2-NG178	95.9	60.8		61.6	65.8	60.3	59.7	58.9	63.8	57.7	60.7	58.7	60.5	73.9	60.9	58.1	58.1
NG218	61.6	61.0	61.2		64.4	60.4	60.5	71.6	63.2	59.4	59.4	59.2	59.8	60.3	60.7	57.1	59.7
B2-NG232	63.5	61.0	62.6	61.6		60.4	60.1	57.9	91.8	58.1	59.4	58.6	60.1	60.9	61.0	56.5	58.5
B2-NG367	60.3	62.7	60.0	60.4	63.1		60.4	57.9	61.4	57.7	60.5	92.4	61.1	58.9	59.7	56.8	58.0
NG198	61.1	61.9	60.9	62.1	64.5	62.0		59.3	63.5	60.0	60.1	61.0	62.2	60.6	74.5	58.3	59.7
NG219	59.2	58.7	58.4	71.4	60.3	57.7	57.6		59.3	57.9	56.7	56.8	59.4	56.7	57.2	55.9	57.1
B3-NG232	60.6	58.2	60.1	59.9	90.9	58.3	58.6	56.4		56.8	57.2	57.5	58.4	59.2	59.3	55.5	57.1
NG277	59.1	60.1	58.8	60.9	62.2	59.2	59.9	59.6	61.5		58.9	58.7	60.7	59.0	60.0	57.6	75.8
NG357	59.8	72.6	59.4	58.4	61.1	59.6	57.6	55.9	59.3	56.5		58.2	59.1	57.9	57.9	55.7	56.2
B3-NG367	57.2	59.6	56.9	57.8	59.8	90.3	58.0	55.6	59.2	55.9	57.7		58.3	56.6	57.3	54.9	55.9
NG411	61.7	62.2	61.4	61.1	64.1	62.4	61.8	60.9	62.9	60.4	61.3	61.0		60.8	60.4	65.0	60.5
NG181	76.0	62.1	75.2	61.7	65.1	60.3	60.4	58.2	63.9	58.9	60.2	59.3	61.0		61.1	59.2	59.6
NG200	61.7	61.2	61.1	61.4	64.4	60.3	73.3	57.9	63.2	59.1	59.5	59.3	59.8	60.3		58.7	59.1
NG412	63.5	62.6	63.0	62.3	64.4	62.0	62.0	61.2	63.9	61.3	61.8	61.4	69.5	63.2	63.4	62.1	62.1
NG558	61.1	60.8	60.2	62.2	63.7	60.4	60.6	59.7	62.8	77.0	59.6	59.7	61.8	60.7	61.0	59.3	

APPENDIX

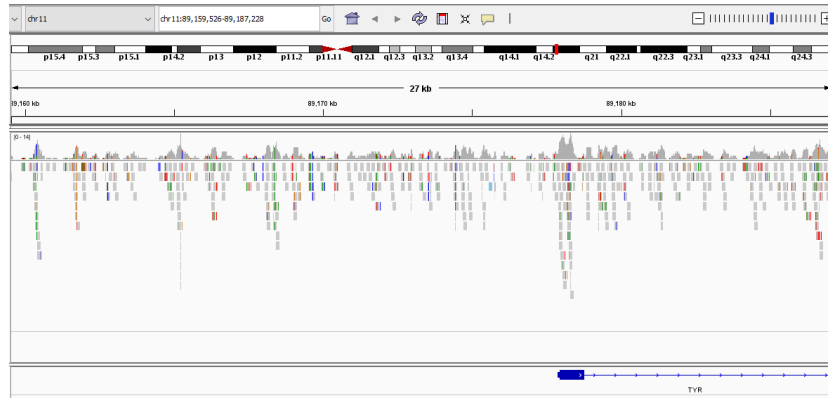


Figure E.1: *IGV image of reads aligning to the TYR gene. Reads align at non-specifically to non-coding regions as they do coding.*

Table E.4: *Extract of the sample sheet ‘SampleSheet.csv’ for batch 6 samples. Compromised index E505 is allocated to the 12 samples (bold) with $\geq 2.6\%$ reads mapped to the target region.*

Sample_ID	Sample_Well	I7_Index_ID	index	I5_Index_ID	index2
NG181	A01	N701	TAAGGCGA	E502	ATAGAGAG
-	A02	N702	CGTACTAG	E502	ATAGAGAG
NG200	A03	N703	AGGCAGAA	E502	ATAGAGAG
NG284	A04	N704	TCCTGAGC	E502	ATAGAGAG
NG325	A05	N705	GGACTCCT	E502	ATAGAGAG
NG394	A06	N706	TAGGCATG	E502	ATAGAGAG
NG395	A07	N707	CTCTCTAC	E502	ATAGAGAG
NG456	A08	N708	CAGAGAGG	E502	ATAGAGAG
-	A09	N709	GCTACGCT	E502	ATAGAGAG
NG472	A10	N710	CGAGGCTG	E502	ATAGAGAG
NG477	A11	N711	AAGAGGCA	E502	ATAGAGAG
NG483	A12	N712	GTAGAGGA	E502	ATAGAGAG
NG504	B01	N701	TAAGGCGA	E503	AGAGGATA
NG506	B02	N702	CGTACTAG	E503	AGAGGATA
NG512	B03	N703	AGGCAGAA	E503	AGAGGATA
NG528	B04	N704	TCCTGAGC	E503	AGAGGATA
NG530	B05	N705	GGACTCCT	E503	AGAGGATA

APPENDIX

NG543	B06	N706	TAGGCATG	E503	AGAGGATA
NG545	B07	N707	CTCTCTAC	E503	AGAGGATA
NG554	B08	N708	CAGAGAGG	E503	AGAGGATA
NG556	B09	N709	GCTACGCT	E503	AGAGGATA
NG557	B10	N710	CGAGGCTG	E503	AGAGGATA
NG558	B11	N711	AAGAGGCA	E503	AGAGGATA
NG559	B12	N712	GTAGAGGA	E503	AGAGGATA
NG561	C01	N701	TAAGGCGA	E504	TCTACTCT
NG563	C02	N702	CGTACTAG	E504	TCTACTCT
NG566	C03	N703	AGGCAGAA	E504	TCTACTCT
NG571	C04	N704	TCCTGAGC	E504	TCTACTCT
NG574	C05	N705	GGACTCCT	E504	TCTACTCT
NG575	C06	N706	TAGGCATG	E504	TCTACTCT
NG580	C07	N707	CTCTCTAC	E504	TCTACTCT
NG582	C08	N708	CAGAGAGG	E504	TCTACTCT
NG584	C09	N709	GCTACGCT	E504	TCTACTCT
NG587	C10	N710	CGAGGCTG	E504	TCTACTCT
NG412	C11	N711	AAGAGGCA	E504	TCTACTCT
NG592	C12	N712	GTAGAGGA	E504	TCTACTCT
NG594	D01	N701	TAAGGCGA	E505	CTCCTTAC
NG597	D02	N702	CGTACTAG	E505	CTCCTTAC
NG598	D03	N703	AGGCAGAA	E505	CTCCTTAC
NG602	D04	N704	TCCTGAGC	E505	CTCCTTAC
NG607	D05	N705	GGACTCCT	E505	CTCCTTAC
NG610	D06	N706	TAGGCATG	E505	CTCCTTAC
NG613	D07	N707	CTCTCTAC	E505	CTCCTTAC
NG210	D08	N708	CAGAGAGG	E505	CTCCTTAC
NG236	D09	N709	GCTACGCT	E505	CTCCTTAC
NG267	D10	N710	CGAGGCTG	E505	CTCCTTAC
NG149-1	D11	N711	AAGAGGCA	E505	CTCCTTAC
NG150-2	D12	N712	GTAGAGGA	E505	CTCCTTAC

Table E.5: *Cohort (n=81) mean percentage coverage at 20x depth across the 31 gene panel using collapsed RefSeq (curated) transcripts from UCSC database.*

Gene	Coverage 20x
<i>BLOC1S3</i>	0.9425
<i>CACNA1A</i>	0.9343
<i>GPR143</i>	0.8436
<i>HPS4</i>	0.9772
<i>DTNBP1</i>	0.9931
<i>TYROBP</i>	0.9883
<i>SLC24A5</i>	0.9947
<i>SETX</i>	0.9927
<i>MYO5A</i>	0.9872
<i>CACNA1F</i>	0.9848
<i>CASK</i>	0.9799
<i>MLPH</i>	0.9917
<i>LYST</i>	0.9688
<i>TULP1</i>	0.9810
<i>TYRP1</i>	0.9535
<i>C10orf11</i>	0.7047
<i>HPS1</i>	0.9290
<i>HPS6</i>	1.0000
<i>SLC45A2</i>	0.9391
<i>AP3B1</i>	0.9924
<i>MANBA</i>	0.9905
<i>HPS5</i>	0.9746
<i>TYR</i>	0.9597
<i>BLOC1S6</i>	0.8360
<i>MITF</i>	0.9804
<i>RAB27A</i>	0.9940

APPENDIX

<i>HPS3</i>	0.9944
<i>OCA2</i>	0.9913
<i>FRMD7</i>	0.9613
<i>PAX6</i>	0.9455
<i>SACS</i>	0.9994

F Appendix Supplementary Data

Table F.1: *Runs of homozygosity (ROH) identified for each sample. The ROH are ordered karyotypically and information of chromosome, start position, end position, length of ROH and the sample ID harbouring the ROH are provided.*

Chr	Start	End	Length	SampleID
1	11092466	15586699	4494234	PA1055
1	20212607	27369520	7156914	V-2
1	33024821	39466755	6441935	PKNYS-09-14
1	36938357	58629926	21691570	PKNYS-01-4
1	43197705	47021717	3824013	30-2
1	58631496	63715784	5084289	PKNYS-07-4
1	62652525	77774432	15121908	PKNYS-01-4
1	92111736	102789388	10677653	PKNYS-07-4
1	108518048	114005751	5487704	PKNYS-07-4
1	114104141	118884844	4780704	PKNYS-07-4
1	155776019	156906193	1130175	PKNYS-07-4
1	155783885	160764443	4980559	PA1047
1	161624310	167848401	6224092	PA1031
1	169960823	186407154	16446300	PA1031
1	196957826	201645114	4687289	PA1031
1	203180248	209539793	6359546	PA1031
1	204477185	208867041	4389857	PA1055
1	227574302	230103510	2529209	PA1031
1	228276547	233985741	5709195	PKNYS-09-14
1	232525311	237911452	5386142	39-1
1	235361938	238492705	3130768	PA1483
1	235755388	238315291	2559904	PA1400
1	237500457	242427196	4926740	PKNYS-01-4
1	244614530	248902068	4287539	PKNYS-01-4
2	3613162	7623864	4010703	PKNYS-07-4
2	15354028	18409442	3055415	PA1059
2	28357283	30735602	2378320	PKNYS-07-4
2	135876392	139687463	3811072	PKNYS-07-4
2	157538567	170465229	12926663	PKNYS-09-14
2	164179351	178071887	13892537	PA1047
2	165420043	170359030	4938988	PA1055
2	197643126	209145341	11502216	PA1031
2	205841165	210648394	4807230	PA1055
2	209147238	215336184	6188947	PA1031

APPENDIX

2	214167327	217995665	3828339	PA1047
2	214729147	216480300	1751154	PA1079
2	216939347	219570312	2630966	PKNYS-09-14
2	227007466	237764060	10756595	PA1079
2	227057388	235602815	8545428	38-1
3	361662	4681620	4319959	PKNYS-08-3
3	10101265	12452364	2351100	PA1059
3	10141653	13771376	3629724	PKNYS-05-2
3	48919714	55831938	6912225	PA1079
3	149155004	155141792	5986789	PA1055
3	156591392	169527953	12936562	PKNYS-07-4
3	176241139	181858589	5617451	PKNYS-07-4
3	184055033	193324987	9269955	PKNYS-09-14
4	1816867	6858212	5041346	PA1444
4	31296581	37191167	5894587	PKNYS-07-4
4	153658641	168364070	14705430	PA1059
4	170179069	174495815	4316747	PKNYS-01-4
4	177439042	182819918	5380877	PA1059
4	178128648	182829621	4700974	PKNYS-07-4
4	185535702	189280323	3744622	PA1059
4	187649398	189901805	2252408	PKNYS-07-4
5	7908888	13677116	5768229	PA1055
5	8032579	13716897	5684319	PA1059
5	35998398	40407048	4408651	PA1031
5	96765516	101469037	4703522	PA1031
5	111534042	116124520	4590479	PA1059
5	126712630	135067868	8355239	PA1031
5	163245011	168273085	5028075	PKNYS-07-4
6	5545583	10460460	4914878	PKNYS-01-4
6	25777159	29830682	4053524	38-1
6	29957444	31270378	1312935	38-1
6	31112655	32518803	1406149	VI-14
6	31377294	32484271	1106978	38-1
6	32757896	35796009	3038114	VI-14
6	36292669	42356953	6064285	PKNYS-01-4
6	36639586	44031332	7391747	PKNYS-04-5
6	36729127	43226781	6497655	PKNYS-09-14
6	43583322	52493222	8909901	PKNYS-09-14
6	118470856	129143875	10673020	PKNYS-07-4

APPENDIX

6	157365939	162153950	4788012	PKNYS-08-3
7	346808	2679481	2332674	PKNYS-01-4
7	6855705	17542334	10686630	PA1031
7	132134108	151692235	19558128	PKNYS-04-5
7	142863497	150474749	7611253	PKNYS-05-2
8	389433	6170865	5781433	PKNYS-05-2
8	7541128	12178871	4637744	PKNYS-08-3
8	12599755	17755130	5155376	PKNYS-08-3
8	23202743	33574787	10372045	PKNYS-01-4
8	33896230	43306151	9409922	PKNYS-01-4
8	68270602	73131887	4861286	PKNYS-07-4
8	73351503	80931634	7580132	PKNYS-07-4
8	86667884	94959130	8291250	PKNYS-07-4
8	104381588	108256576	3874989	PKNYS-07-4
8	120789538	132656934	11867397	PKNYS-01-4
8	134890900	142604319	7713420	PKNYS-01-4
8	135941813	139509891	3568079	PKNYS-07-4
8	136805942	142469760	5663819	PKNYS-05-2
8	141708531	144517495	2808965	NY204
8	142681306	144870446	2189141	CBLN112
9	12934	2544610	2531677	PKNYS-04-5
9	12934	3227125	3214192	PKNYS-07-4
9	30586490	34717208	4130719	PA1055
9	75400138	78956075	3555938	PKNYS-01-4
9	102725522	109806401	7080880	PKNYS-05-2
9	128608477	136108906	7500430	PA1031
9	136269760	138176567	1906808	PA1031
10	65633	4846013	4780381	PKNYS-05-2
10	388048	3697610	3309563	PKNYS-01-4
10	12606150	20657724	8051575	PKNYS-05-2
10	26308088	29585583	3277496	PA1079
10	42094029	48805357	6711329	PA1059
10	47775354	62317371	14542018	PKNYS-09-14
10	48555373	52770934	4215562	30-2
10	50880116	59673534	8793419	PA1059
10	58523248	65966756	7443509	PA1031
10	59947596	69844452	9896857	PA1059
10	67521943	72213223	4691281	CBLN112
10	78764615	89141369	10376755	PKNYS-04-5

APPENDIX

10	108096744	119448383	11351640	PKNYS-09-14
10	122643428	133423556	10780129	PKNYS-09-14
11	1025987	5994153	4968167	US03
11	1031395	3224927	2193533	PKNYS-08-3
11	3676362	6853937	3177576	PKNYS-08-3
11	5860244	9405647	3545404	PA1055
11	18597572	47635111	29037540	PA1055
11	22578058	27338633	4760576	PKNYS-07-4
11	25458482	43006889	17548408	PKNYS-01-4
11	41653136	47580425	5927290	38-1
11	43262919	47605263	4342345	PKNYS-01-4
11	61883614	67881891	5998278	PKNYS-01-4
11	68148630	73605559	5456930	PKNYS-01-4
11	79176968	108114514	28937547	30-2
11	99671293	105070890	5399598	PA1047
11	100666676	108222481	7555806	US03
11	112144038	116790097	4646060	PKNYS-04-5
11	112192862	115699348	3506487	PKNYS-07-4
11	112225058	117730161	5505104	30-2
11	112263196	116473687	4210492	PKNYS-01-4
11	116830378	119182117	2351740	PKNYS-01-4
11	119096814	123280192	4183379	PA1055
11	119638915	125134247	5495333	PKNYS-01-4
11	119638915	126293229	6654315	38-1
11	121574375	126023608	4449234	PKNYS-05-2
11	125908179	129457703	3549525	PKNYS-01-4
12	13076	3964048	3950973	30-2
12	13514	2664689	2651176	PA1047
12	13740	6956413	6942674	VI-14
12	2449134	9200312	6751179	CBLN112
12	37444642	40323341	2878700	PA1059
12	49863312	53328344	3465033	NYS04
12	49877714	52302003	2424290	PA1047
12	52306316	57581787	5275472	PA1047
12	102632056	106638173	4006118	PKNYS-07-4
12	108485690	109808578	1322889	PKNYS-07-4
12	117778808	123407318	5628511	30-2
12	119357711	123257229	3899519	PA1059
12	124028956	131399860	7370905	PA1059

APPENDIX

13	18335023	19383508	1048490	PKNYS-07-4
13	19666603	23062040	3395438	PKNYS-07-4
13	25699400	31317609	5618210	PKNYS-05-2
13	31802581	46517778	14715198	PKNYS-05-2
13	60448049	65265004	4816956	PA1059
13	93378440	99641753	6263314	30-2
13	105983282	111609234	5625953	39-1
13	110306784	112530113	2223330	PKNYS-04-5
14	19508915	22992647	3483733	PKNYS-07-4
14	80998224	87935098	6936875	PKNYS-07-4
14	95743850	105455153	9711304	PKNYS-08-3
14	96638570	105643800	9005231	PKNYS-05-2
14	97027191	104510917	7483727	30-2
15	20418570	24347397	3928828	PA1059
15	21601660	28146074	6544415	PA1031
15	23062386	26138204	3075819	PA1055
15	23893932	28565103	4671172	PKNYS-08-3
15	24395238	28299213	3903976	VI-14
15	24899796	28299921	3400126	PA1059
15	31795583	35081334	3285752	NY204
15	32731469	47866714	15135246	PA1055
15	56173687	64490193	8316507	PKNYS-04-5
15	68351101	82442886	14091786	PKNYS-04-5
15	72135101	77800023	5664923	PA1059
15	84655131	92764696	8109566	PKNYS-04-5
16	29914	14820577	14790664	38-1
16	12959330	16257716	3298387	PKNYS-07-4
16	15129796	21072037	5942242	38-1
16	67154540	81936583	14782044	PKNYS-04-5
16	75194071	81018071	5824001	PKNYS-01-4
16	79305634	85636776	6331143	PA1055
16	81083175	88429246	7346072	38-1
16	82091171	85634339	3543169	PKNYS-04-5
17	184485	3690982	3506498	38-1
17	746607	3869178	3122572	PA1055
17	1584139	7415077	5830939	PKNYS-04-5
17	3950225	10251960	6301736	PA1047
17	4049808	6537844	2488037	PA1055
17	6657277	10508315	3851039	PA1055

APPENDIX

17	10650620	14077033	3426414	PA1055
17	50167819	63872824	13705006	PKNYS-01-4
17	68514683	72234148	3719466	PA1055
17	76620721	81652958	5032238	PA1011
17	78947947	80643639	1695693	PA1055
17	79126888	83189278	4062391	PA1047
18	1056175	6386215	5330041	PA1059
18	2313439	9117869	6804431	PKNYS-08-3
18	31377109	47510766	16133658	PA1059
18	64434158	69357618	4923461	PKNYS-07-4
18	67842475	74263669	6421195	PKNYS-01-4
18	69896112	74431179	4535068	PA1059
19	572878	3902724	3329847	39-1
19	9884209	13212008	3327800	PA1055
19	10285007	13689503	3404497	PKNYS-05-2
19	10293551	13576605	3283055	39-1
19	13927777	16587731	2659955	PA1055
19	17191054	22501942	5310889	PA1055
19	35587974	40613367	5025394	38-1
19	38326846	40840084	2513239	PA1031
19	40876703	46113651	5236949	US03
19	43233890	46113651	2879762	PA1031
19	43356103	50725489	7369387	V-2
20	72450	5304055	5231606	PKNYS-09-14
20	6935638	19813215	12877578	PKNYS-04-5
20	32474894	45909788	13434895	PKNYS-05-2
21	40189189	46602462	6413274	PA1055
22	16401674	19918279	3516606	PA1031
22	37182920	45112525	7929606	38-1
X	150099394	154929926	4830533	PKNYS-07-4

SCIENTIFIC REPORTS

OPEN

Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (*TYR*) causing hypomorphic oculocutaneous albinism (OCA1B)

Chelsea S. Norman¹, Luke O’Gorman², Jane Gibson³, Reuben J. Pengelly⁴, Diana Baralle⁵, J. Arjuna Ratnayaka¹, Helen Griffiths¹, Matthew Rose-Zerilli⁵, Megan Ranger⁶, David Bunyan^{2,7}, Helena Lee^{1,6}, Rhiannon Page¹, Tutte Newall¹, Fatima Shawkat⁶, Christopher Mattocks^{2,8}, Daniel Ward⁷, Sarah Ennis⁴ & Jay E. Self^{1,6}

Oculocutaneous albinism (OCA) and ocular albinism (OA) are inherited disorders of melanin biosynthesis, resulting in loss of pigment and severe visual deficits. OCA encompasses a range of subtypes with overlapping, often hypomorphic phenotypes. OCA1 is the most common cause of albinism in European populations and is inherited through autosomal recessive mutations in the Tyrosinase (*TYR*) gene. However, there is a high level of reported missing heritability, where only a single heterozygous mutation is found in *TYR*. This is also the case for other OCA subtypes including OCA2 caused by mutations in the *OCA2* gene. Here we have interrogated the genetic cause of albinism in a well phenotyped, hypomorphic albinism population by sequencing a broad gene panel and performing segregation studies on phenotyped family members. Of eighteen probands we can confidently diagnose three with OA and OCA2, and one with a *PAX6* mutation. Of six probands with only a single heterozygous mutation in *TYR*, all were found to have the two common variants S192Y and R402Q. Our results suggest that a combination of R402Q and S192Y with a deleterious mutation in a ‘tri-allelic genotype’ can account for missing heritability in some hypomorphic OCA1 albinism phenotypes.

Oculocutaneous albinism (OCA) and X-linked ocular albinism (OA) are inherited disorders of melanin biosynthesis which result in varied levels of hypopigmentation of skin, hair, and ocular tissues¹. Characteristic ophthalmic features include reduced visual acuity, nystagmus, strabismus, and photophobia. Closer examination may reveal foveal hypoplasia (abnormal retinal development), asymmetry of visual evoked potential (VEP) responses, and iris transillumination¹. Foveal hypoplasia for instance, can be determined using Spectral-Domain Optical Coherence Tomography (SD-OCT) and then graded on a scale of 1–4 (Thomas *et al.*²), and the asymmetry of visual-evoked potentials documents the excessive decussation at the optic chiasm seen in albinism³. Partial phenotypes are described widely in the literature in which some features are present but others are lacking (e.g. nystagmus or foveal hypoplasia), however, phenotyping methods have varied significantly and the partial phenotype has never before been described in detail^{4–6}. Current management of albinism focusses on correction of any

¹Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. ²Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK. ³Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, UK. ⁴Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, Southampton, UK. ⁵Cancer Sciences Unit, Faculty of Medicine, University of Southampton, Southampton, UK. ⁶Eye Unit, University Hospital Southampton, Southampton, UK. ⁷Molecular Genetics Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury, UK. ⁸Wessex Investigational Science Hub, University Hospital Southampton, Southampton, UK. Correspondence and requests for materials should be addressed to J.E.S. (email: jes3@soton.ac.uk)

HGNC symbol	HGNC name	Albinism subtype	Mode of inheritance
TYR	Tyrosinase	OCA1A	Autosomal recessive
		OCA1B	
OCA2 (P gene)	OCA2 melanosomal transmembrane protein	OCA2	Autosomal recessive
TYRP1	Tyrosinase related protein 1	OCA3	Autosomal recessive
SLC45A2	Solute carrier family 45 member 2	OCA4	Autosomal recessive
—	Chromosomal location 4q24	OCA5	Autosomal recessive
SLC24A5	Solute carrier family 24 member 5	OCA6	Autosomal recessive
C10orf11	Chromosome 10 open reading frame 11	OCA7	Autosomal recessive
GPR143	G protein-coupled receptor 143	OA1	X-linked recessive

Table 1. Table to describe HGNC approved gene names associated with the subtypes of OCA and OA. OCA5 has been attributed to a chromosomal location but does not yet have an associated gene⁴⁶.

refractive errors, management of head postures/strabismus and on the importance of effective sun protection. Another important factor in the management albinism is genetic counselling; therefore it is important to confirm a genetic diagnosis for patients.

Six genes involved in melanin biosynthesis pathway are known to cause forms of OCA and OA: *TYR* (tyrosinase), *OCA2*, *TYRP1* (tyrosinase-like protein 1), *SLC45A2* (solute carrier family 45 member 2), *SLC24A5* (solute carrier family 24 member 5), and *C10orf11* (chromosome 10 open reading frame 11) accounting for OCA subtypes 1–4 and 6–7 respectively, and *GPR143* accounting for OA1⁶, see Table 1. All of the OCA subtypes are understood to be inherited as autosomal recessive disorders but the subtypes are heterogeneous in pigmentary phenotype^{1,7,8}. OCA1 has a mixed phenotype and is further split into OCA1A and OCA1B. OCA1A describes complete loss of tyrosinase activity (previously described as ‘tyrosine negative’ albinism) and is characterised by an apparent total lack of pigment. Some tyrosinase function is retained in OCA1B, allowing pigment to accumulate and generate a phenotype of minimal to near normal skin pigmentation, as is also the case for the other described OCA and OA phenotypes^{1,8}. Phenotypes of partial OCA also overlap with those seen in patients with dominant mutations in the *PAX6* gene, which is involved in ocular development, where a variety of phenotypes have been described including foveal hypoplasia, iris trans-illumination and nystagmus⁹.

As the most severe form of OCA, OCA1A is often recognised in early infancy. King *et al.* proposed that white hair from birth can be used to predict OCA1⁸, with 85% of patients identified in this way testing positive for pathogenic *TYR* mutations. However, 15% of OCA cases identified in this way had no accountable genetic mutation, and 29% of those confirmed as OCA1 had only one identifiable *TYR* mutation⁸. It is widely recognised that the OCA genes do not account for all non-syndromic cases, as many as 30% of OCA1A occurrences have an unknown genetic origin^{10,11} and this percentage may be higher for cases of partial albinism¹². It is also important to note that a variety of techniques have been employed to screen for tyrosinase gene mutations in these studies and no method has 100% sensitivity.

An individual's pigmentary phenotype depends on polymorphisms in many genes, including polymorphisms in the OCA genes^{13–15}. Ethnic background may play a large role in an individual's susceptibility to the albinism phenotype, with hypomorphic mutations having a more damaging effect on a less active pigmentary pathway^{16,17}. It has been suggested that inheritance of OCA2 is not purely recessive, with the example of haploinsufficiency noticeably affecting skin complexion in a Hispanic family, arguably due to the already fair skin tone¹³. It has also been suggested that a synergistic interaction between genes throughout the pigment pathway may exist in albinism phenotypes, evidenced by one family exhibiting an OCA2 phenotype that is modified by a mutation in the gene for OCA3¹⁴ and a correlation between *OCA2* and *MC1R* variants in a small albinism cohort¹⁸. The quantitative effect of pigmentation also has relevance to OCA1B, particularly the notion of autosomal recessive ocular albinism (ARO), an arbitrary characterisation that has been used previously to describe cases with clinically mild OCA1B^{19,20}.

ARO sparked a debate over the possible pathogenicity of two *TYR* polymorphisms, rs1126809 (p.R402Q) and rs1042602 (p.S192Y), common in Caucasian populations with allele frequencies ~28–36%²¹. Functional studies have shown the R402Q polymorphism produces a thermolabile enzyme, retained by the cells endoplasmic reticulum, with a 75% reduction in catalytic activity compared to the wild-type^{15,22,23}; and S192Y results in a 40% reduction of tyrosinase enzymatic activity²⁴. Multiple OCA1 studies have shown the R402Q allele is strongly associated with albinism patients with only one mutation^{12,17,20}.

R402Q has been proposed as a causal variant, though only when inherited on the trans allele to a null activity *TYR* mutation^{19,20}. However this was disputed with evidence of no OCA phenotype in the parents of affected probands even when they carried a combination of null mutation and R402Q²⁵. This has led to the question of whether it is possible for an additional variant to be necessary for manifestation of the ocular phenotype. The combination of two common variants may produce a reduction in *TYR* activity that, when co-inherited with a deleterious *TYR* mutation, provides sufficient loss of activity to cause an albino phenotype^{15,16}. A similar tri-allelic hypothesis has been demonstrated in Bardet-Biedl syndrome²⁶, but is yet to be demonstrated in albinism.

In this study, we have sequenced all the known albinism genes in patients with possible hypomorphic albinism phenotypes, identified through detailed ocular phenotyping in a tertiary eye clinic. Probands with some, but not all of the typical cutaneous and ocular features of OCA1A were defined as having a likely hypomorphic albinism phenotype. For the first time, we investigate common variants in tri-allelic pattern of inheritance using detailed phenotyping and segregation studies in relatives to identify the causative genotype.

Methods

Patients were recruited following the tenets of the declaration of Helsinki, informed consent was obtained and the research was approved by the Southampton & South West Hampshire Research Ethics Committee.

We investigated the genetic cause of eighteen probands categorized as having hypomorphic albinism. Probands were identified from a regional paediatric nystagmus clinic. All patients seen in this clinic underwent detailed phenotyping of skin and hair tone in context of family pigmentation, orthoptic examination, anterior and posterior segment examinations on a slit-lamp biomicroscope, electrodiagnostics including an electroretinogram (ERG) and visual evoked potential (VEP), and optical coherence tomography (OCT) of the macula using either a Leica OCT system or a Spectralis OCT (Heidelberg Engineering). Eye movement recordings were made on an EYELink10000 + (SR research) eye tracker and refraction was measured. Saliva was collected and DNA extracted using Oragene-DNA kit (OG-575)(DNA Genotek).

Probands with at least two phenotypic features of albinism (skin and hair pigmentation deemed to be low within the family context/nystagmus/foveal hypoplasia/VEP crossing/iris transillumination) as determined by a consultant ophthalmologist (JES), were chosen from a larger database containing approximately 300 probands with albino and/or nystagmus phenotypes. Probands were additionally excluded if they had complete characteristics of OCA1A or where DNA quality was poor.

The DNA samples were enriched using the TruSight One capture platform (Illumina 5200 Illumina Way San Diego, California USA). TruSight One has been dubbed a “clinical exome”, covering 4813 genes associated with disease-causing mutations. The panel targets and captures most of the coding regions of OCA genes 1–4 & 6, the OA1 gene, all syndromic albinism genes and PAX6, coverage of genes is shown in Supplementary Table 1. Prepared libraries underwent paired-end sequencing on an Illumina NextSeq 500 machine.

Next generation sequencing (NGS) data was aligned against the human reference genome (hg19) using Novoalign (v2.08.02). The mean read depth across all samples was 167 (Supplementary Table 1) with 97.2% of all target regions achieving a depth of 20X or greater. Variant calling was performed using SAMtools v0.1.19²⁷ and variant annotation using ANNOVAR²⁸ against RefSeq transcripts. Additional annotation was applied using the Human Gene Mutation Database, HGMD²⁹. The mean depth and percentage of target captured at a read depth of 20X for each sample is listed in Supplementary Table 2.

Variants within the genes of interest were filtered using 1000 Genomes Project Minor Allele Frequency (MAF) (<0.05) and the *in silico* pathogenicity prediction tools SIFT (<0.05), PolyPhen2 HumVar (possibly damaging and probably damaging) and GERP++ (>2). SIFT predicts pathogenicity of missense mutations based on homology³⁰, PolyPhen2 HumVar predicts pathogenicity based on conservation and protein structure/function³¹ and GERP++ measures evolutionary constraint³². The six probands with only a single heterozygous *TYR* mutation were further investigated. Sanger sequencing was used to confirm and segregate each *TYR* variant in probands and family members, primers used are listed in Supplementary Table 3. Primers designed by Chaki *et al.* were used to for amplification of *TYR* exon 4 to avoid amplification of the highly homologous *TYRL* gene³³.

Multiple ligation-dependent probe amplification (MLPA) was carried out for the *TYR* and *OCA2* genes as according to the manufacturer's instructions with the current SALSA MLPA P325 *OCA2* probe mix at the time of testing (MRC-Holland, the Netherlands). Partial albinism probands and control individuals were compared. Subsequent data were analysed using the MLPA analysis function of the GeneMarker (version 1.85) software (SoftGenetics, USA)³⁴.

Results

Diagnosis of hypomorphic albinism. The hypomorphic albinism phenotype varied in both ocular phenotype and pigment level between probands and between family members. For example proband and mother in family 3 both have a phenotype consistent with partial albinism, however the proband exhibits a severe loss of cutaneous pigment but no iris transillumination, whereas the cutaneous pigment in the proband's mother is within that of the family context but ocular investigations revealed trans-illumination defects. The level of foveal hypoplasia also varied between patients and within families. Example OCT images taken from the cohort are in Fig. 1, demonstrating the broad range of foveal developmental anomalies identified.

NGS data for OCA genes 1–4 & 6, the OA1 gene, and PAX6 were initially filtered using predictive scores from SIFT and PolyPhen. GERP++ was also noted, and variants with a MAF >5% were considered benign and were filtered using the 1000 Genomes Project dataset. This revealed eighteen potentially causal mutations across five genes in thirteen probands, leaving five probands with no variants passing the filtering threshold. No proband was found to have more than three variants using this methodology, results in Table 2.

Proband 9 was found to have a likely pathogenic mutation in the PAX6 gene and proband 10 has a deletion resulting in a frameshift mutation in the X-linked gene, *GPR143*. Probands 3 and 16 each have two compound heterozygous mutations in the *OCA2* gene, these putative variants would explain autosomal recessive inheritance of *OCA2*. Proband 8 has a single mutation in *OCA2* and a second mutation in *TYRP1* which would require further investigation before concluding causality. Two probands, 13 and 17, each have a single heterozygous mutation in the *OCA2* gene with no second mutation identified. Furthermore, six probands each had a single heterozygous mutation in the *TYR* gene with no further variants passing the filtering threshold. Probands 3 and 9 also have *TYR* mutations, but due to potentially causal variants in other genes, these single recessive mutations may afford probands 3 and 9 carrier status only. The *TYR* mutation P406L occurs in two probands, as does the *OCA2* mutation V419I.

MLPA of *TYR* and *OCA2* was carried out in probands 1–6 and 8–12 to search for large deletions that would be missed in the NGS data. MLPA results revealed no abnormal copy numbers, ruling out whole gene/exon deletions.

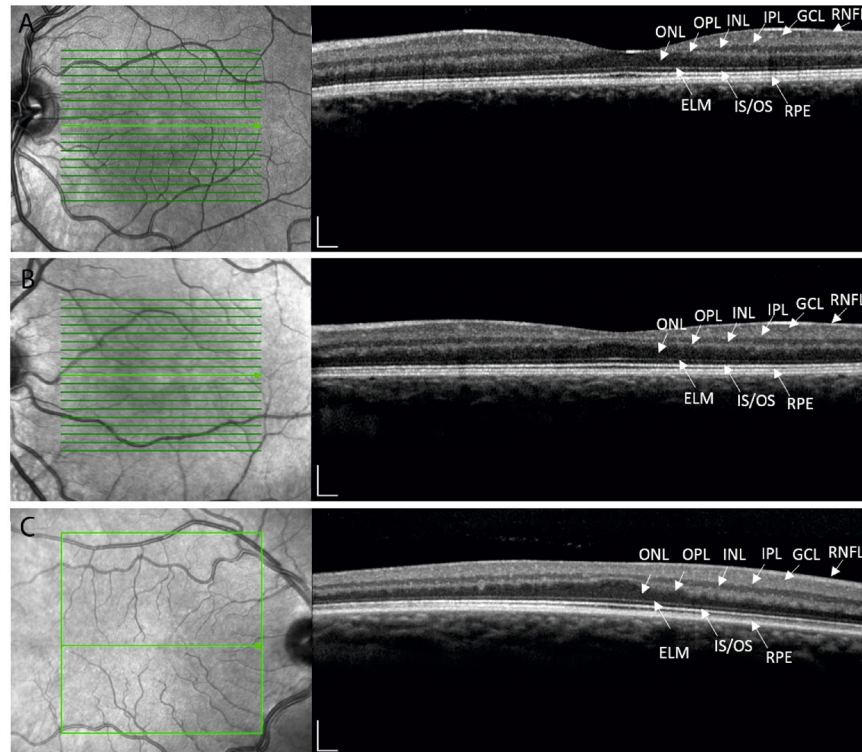


Figure 1. OCT images using the Heidelberg Spectralis Diagnostic imaging platform. (a) Normal fovea (Mother of proband 13) (b) Foveal hypoplasia grade 1 (brother of proband 13) (c) Foveal hypoplasia grade 3 (Mother of proband 18). Foveal grading according to the Thomas *et al.* grading system². Outer nuclear layers (ONL), outer plexiform layers (OPL), inner nuclear layers (INL), inner plexiform layers (IPL), ganglion cell layers (GCL) and retinal nerve fibre layers (RNFL) are labelled.

Segregation of the OCA1 tri-allelic genotype. We further investigated the single *TYR* variants in both probands and family members (families 4–7, 12 and 18 in Table 3) using Sanger sequencing to confirm and determine segregation of variants. In total, twenty probands and family members were phenotyped and genotyped, results in Table 3. The phenotyping results of these six families suggests a total nine cases of partial albinism (six probands and three affected family members). Sanger sequencing confirmed the predicted causal variants in probands and revealed variants segregated with affected family members in every case, with three unaffected family members as carriers.

To explore the apparent missing heritability in these cases we investigated the potential pathogenicity of common variants R402Q and S192Y. The NGS data was examined in probands with *TYR* mutations. All six probands were found to have both common variants. These variants were confirmed in probands with Sanger sequencing and variant segregation was determined across available members of the six pedigrees, shown in Fig. 2. The combined presence of both common polymorphisms and a putative *TYR* mutation in a tri-allelic genotype segregates with affected family members.

It can be deduced that the R402Q variant is on the *trans* allele to the deleterious *TYR* mutation in probands 4, 5, 7 and 12. In family 4 we can also be certain the S192Y variant is on the *trans* allele. The mother of proband 18 has both nystagmus and foveal hypoplasia, yet does not have the same deleterious *TYR* mutation as her son.

Discussion

We have combined high resolution phenotyping, a broad NGS technique, segregation analysis and MLPA studies in a cohort of presumed partial albinism patients. This allows us the opportunity to perform a detailed genotype-phenotype correlation in this group of patients for the first time. In this study we identified one novel variant in the *PAX6* gene, a novel frameshift variant in the *GPR143* gene, two novel variants in the *OCA2* gene (both in probands 3), five previously reported variants in *OCA2*, three novel and four previously reported variants in the *TYR* gene, and one previously reported variant in *TYRP1* in eighteen probands. When combined, these variants provide a convincing genetic diagnosis for only 22% of our original hypomorphic albinism cohort if those with missing variants, in a presumed recessive condition (*OCA1*), are excluded.

The novel variant in *GPR143*, c.485delG, causes a frameshift mutation likely resulting in ocular albinism in proband 10. Of the six different mutations found in *OCA2*; N465D⁸, V419I, Y342C³⁵ and L650V³⁶ have been reported previously in association with albinism. The variants R536C and W274C are both predicted to be deleterious by SIFT, PolyPhen2 and GERP++, described in Table 2.

The probands revealed seven different mutations in the *TYR* gene: V177F, c.1467dup, c.505_507del, C244Ter, R422W, R402Ter and P406L. The mutation V177F has been previously reported in an albinism cohort³⁷. *TYR*

Proband	Variant 1	Variant 2	Variant 3
1	—	—	—
2	—	—	—
3	TYR c.529 G > T p.V177F (SIFT = . PolyPhen = D GERP = 5.16)	OCA2 c.822 G > C p.W274C (SIFT = 0 PolyPhen = D GERP = 4.66)	OCA2 c.1948C > G p.Loo650V (SIFT = 0.03 PolyPhen = D GERP = 5.75)
4	TYR c.1467dup p.T489fs	—	—
5	TYR c.505_507del p.D169del	—	—
6	TYR c.732_733del p.C244Ter	—	—
7	TYR c.1204 C > T p.R402Ter	—	—
8	OCA2 c.1393 A > G p.N465D (SIFT = 0.01 PolyPhen = D GERP = 5.33)	TYRP1 c.1037 C > G p.P346R (SIFT = 0 PolyPhen = D GERP = 5.73)	—
9	TYR c.1217 C > T p.P406L (SIFT = . PolyPhen = D GERP = 4.68)	PAX6 c.1264 C > A p.Q422K (SIFT = 0 PolyPhen = D GERP = 6.16)	—
10	GPR143 c.485del p.W162fs	—	—
11	—	—	—
12	TYR c.1217 C > T p.P406L (SIFT = . PolyPhen = D GERP = 4.68)	—	—
13	OCA2 c.1606C > T p.R536C (SIFT = 0.01 PolyPhen = D GERP = 5.8)	—	—
14	—	—	—
15	—	—	—
16	OCA2 c.1255 G > A p.V419I (SIFT = 0.02 PolyPhen = D GERP = 5.2)	OCA2 c.1025 A > G p.Y342C (SIFT = 0 PolyPhen = D GERP = 5.55)	—
17	OCA2 c.1255 G > A p.V419I (SIFT = 0.02 PolyPhen = D GERP = 5.2)	—	—
18	TYR c.1264 C > T p.R422W (SIFT = . PolyPhen = D GERP = 2.69)	—	—

Table 2. Predicted causal variants, in eighteen probands with phenotypes matching hypomorphic albinism. Pathogenicity determined by filtering all variants in the genes; *TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5*, *C10orf11* and *PAX6*, with the parameters MAF < 0.05, SIFT < 0.05, PolyPhen2 = possibly damaging or probably damaging. The prediction scores for non-synonymous variants are included, for some mutations a prediction score was not available at the time of analysis. Gene accessions number: *TYR* NM_000372, *OCA2* NM_001300984, *PAX6* NM_001258465, *TYRP1* NM_000550, *GPR143* NM_000273.

c.1467dup results in a frameshift and has been reported as a causal mutation multiple times^{8, 10, 20, 37}. R402Ter has been reported previously and creates a premature stop codon, considered highly deleterious^{20, 37, 38}. The mutation P406L has also been reported many times before in association with albinism^{8, 37}, and it has been shown to reduce enzyme activity to 35%³⁹. R422W has been reported as disease causing⁸, however functional studies of this mutation have conflicting results. Mondal *et al.* assayed the tyrosine hydroxylase and DOPA oxidase activity of the R422W mutant and found that the enzyme retained no activity¹⁶, whereas, Dolinska *et al.* assessed only DOPA oxidase activity and found that the R422W mutant retained 95% of wild-type enzyme activity. Dolinska *et al.* also state that R422W is temperature sensitive and the immature glycoprotein is degraded more quickly at 37 °C³⁹, potentially accounting for the difference between assays. Reported literature ascribes many variants as disease causing throughout the coding regions of both *OCA2* and *TYR*, however recent functional studies have questioned the deleterious effect of some of these variants, particularly in the *TYR* gene^{16, 39}. There is currently no functional evidence of the deleterious effect of the mutations TYR c.505_507del and TYR C244Ter though the deletions have been previously been reported as causal mutations, and the introduction of a premature stop codon is considered highly deleterious^{40, 41}. It is likely that further functional analyses are necessary to produce a curated list of mutations for accurate genetic diagnosis⁴².

Six probands within our cohort were found to have single *TYR* variant previously identified in albinism patients, but no variant in another known gene. As there is no functional evidence for the variants in family 5 and family 6 there remains the possibility of another causal gene mutation. It has been suggested that this high level of missing heritability could be due to mutations in the *TYR* promoter or an interacting distal gene enhancer⁴³. Notably, all six had also inherited R402Q and S192Y common *TYR* variants producing a tri-allelic genotype.

The common variant R402Q is located in exon 4, near to the CuB catalytic site, and produces a thermolabile enzyme^{16, 22}, but it has been argued that the reduction of tyrosinase activity is not enough to produce a phenotype. The controversy over the R402Q variant stems from a paper by Oetting *et al.* which argues that segregation of R402Q with a known pathogenic variant on the homologous allele does not confer albinism²⁵.

The variant S192Y is located in the CuA catalytic site of tyrosinase and has been shown to lower enzymatic activity independently to R402Q¹⁵. Previous studies have had stringent criteria for an OCA1 phenotype (white hair and skin and translucent irides from birth)²⁵, whereas, here we have considered hypomorphic presentations that do not appear as severe but result in ocular deficits nonetheless. Here we suggest that a combination of a pathogenic mutation inherited with both variants in a tri-allelic genotype may cause a large enough reduction in tyrosinase activity for a partial OCA1 phenotype.

ID	Relation to proband	Abnormal pigment	Nystagmus	OCT	Trans-illumination	VEP	Genotype		
							TYR Variant 1	R402Q	S192Y
Family 4	Proband	Yes - OCA1A	No	FH	No	Crossed	c.1467dup p.T489fs^{8, 10, 20, 37}	Het	Het
	Father	No	No	Normal	No	—	c.1467dup p.T489fs ^{8, 10, 20, 37}	WT	WT
	Mother	No	No	Normal	No	—	WT	Het	Hom
	Sister	No	No	Normal	No	—	WT	WT	Het
Family 5	Proband	Yes	No	FH	Yes	Abnormal	c.505_507del p.D169del⁴⁰	Het	Het
	Mother	No	No	—	—	—	WT	Het	Het
	Father	No	No	—	—	—	c.505_507del p.D169del ⁴⁰	WT	Het
Family 6	Proband	Yes - OCA1A	No	FH	No	Normal	c.732_733del p.C244Ter⁴¹	Het	Het
	Mother	No	No	FH	Yes	—	c.732_733del p.C244Ter⁴¹	Het	Het
	Father	No	No	Normal	No	—	WT	Het	Het
	Sister	No	No	—	—	—	WT	Het	Het
Family 7	Proband	Yes	Yes	—	Yes	—	c.1204 C > T p.R402Ter^{20, 37, 38}	Het	Het
	Sister	Yes	Yes	—	Yes	—	c.1204 C > T p.R402Ter^{20, 37, 38}	Het	Het
	Mother	No	No	—	—	—	c.1204 C > T p.R402Ter ^{20, 37, 38}	WT	Het
	Father	No	No	—	—	—	WT	Het	Het
Family 12	Proband	No	Yes	FH	No	Crossed	c.1217 C > T p.P406L^{8, 20, 37}	Het	Het
	Mother	No	No	—	—	—	c.1217 C > T p.P406L ^{8, 20, 37}	WT	Het
	Grandmother	No	No	—	—	—	WT	Het	WT
Family 18	Proband	Yes	Yes	FH	Mild	Inconclusive	c.1264 C > T p.R422W^{8, 16, 39}	Het	Het
	Mother	No	Yes	FH	No	—	WT	Het	Het

Table 3. Phenotype-genotype table of families with Sanger-confirmed TYR variants. Family number corresponds with proband number. Phenotype information (from left to right): cutaneous and hair pigmentation in context of family background, presence of nystagmus, foveal hypoplasia (FH), iris trans-illumination, and VEP asymmetry indicating (over)crossing of the optic nerve. Those with partial albinism are in bold.

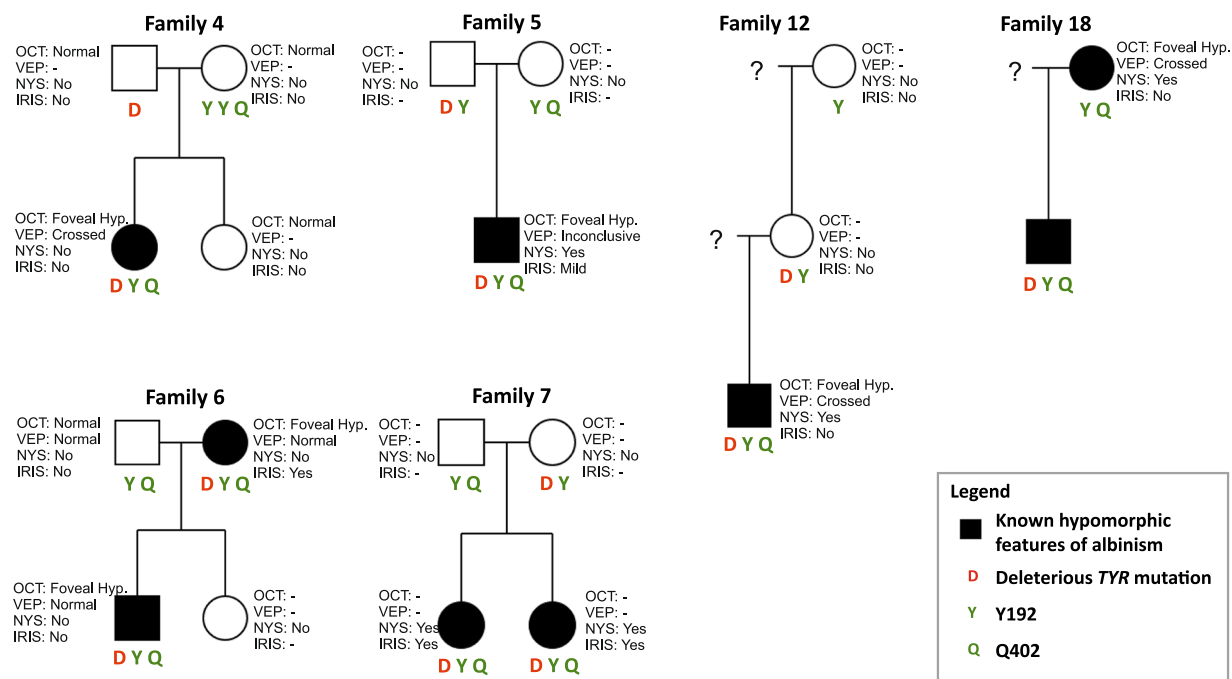


Figure 2. Pedigree diagrams for six families with a single TYR pathogenic mutation and common polymorphism phenotyping. TYR variants are listed beneath each family. Sanger sequencing was performed on family members as opposed to the full exonic region sequenced in probands. Family number corresponds with proband number.

AROA is not an appropriate diagnosis for probands in this cohort as cutaneous and hair pigment is noticeably decreased in most probands and many family members and there is a lot of variation in ocular phenotype. Background level of pigmentation may determine the severity of the mutations as lower pigment levels will be affected more severely by the same dosage loss of tyrosinase. Therefore, our results support the theory of a causal tri-allelic genotype may go some way to account for many cases of OCA1 with apparent missing heritability. Functional studies would assist in confirming pathogenicity, thus allowing the tri-allelic genotype to be considered for both future and retrospective genetic diagnosis of OCA1.

There is potential for a double-variant haplotype, p.[S192Y;R402Q], existing on the *trans* allele to the known *TYR* mutation in affected individuals. A combination of the common variants R402Q and S192Y in *cis* may have a compound effect, producing a great enough loss of function equal to a deleterious *TYR* mutation. Each of the common variants R402Q and S192Y have a MAF of greater than 20%, and as individual SNPs they are considered benign (shown in our cohort in unaffected family members). In contrast, the predicted frequency of p.[S192Y;R402Q] in *cis* is 1.1%, using 'British in England and Scotland' participants of the 1000 Genomes project (GBR) and the webserver <http://analysistools.nci.nih.gov/LDlink/>⁴⁴. Currently, a single variant is considered benign if the MAF is >5%⁴⁵. Our findings suggest standards and guidelines could be revised to consider the combined impact of variants, particularly for more complex disorders such as albinism. Furthermore, the diagnosis of albinism currently focusses on compound mutations in single genes without considering the potential for synergistic relationships between functionally related genes such as that previously suggested for OCA2 and OCA3 genes (*OCA2* and *TYRP1*)¹⁴ and for which there is potentially one example in our cohort.

If our proposed tri-allelic genotype hypothesis is correct, this would increase the diagnostic yield of genetic testing from 22% as described earlier, to 56% in our cohort. Given that hypomorphic albinism is a difficult cohort to diagnose clinically, evidenced by the PAX6 mutation found in the atypical case (proband 9), further exome-seq is suitable for the genetic diagnosis. A sequencing technique with broad capture allows for the pickup of genetic variants which may have resulted in an overlapping ocular phenotype.

There is no current treatment for the underlying molecular anomaly in albinism and present treatments are supportive. Therapeutics are under development but an effective treatment for any of the underlying molecular defects has not yet reached clinical practice. Our work and that of others appears to suggest that small variations in melanin biosynthesis between related family members dictate the extent of the phenotype in OCA pedigrees. Furthermore, the net loss of *TYR* function (caused by cumulative effects of multiple variants, each of which reduce *TYR* function by differing amounts), appear to result in a continuum of clinical features.

Our work supports the assertion that small modulations in components of the melanin biosynthesis pathways, through therapeutic means, may be sufficient to rescue some of the visual disability seen in patients with albinism phenotypes.

References

- Grønskov, K., Ek, J. & Brøndum-Nielsen, K. Oculocutaneous albinism. *Orphanet J. Rare Dis.* **2**, b25 (2007).
- Thomas, M. G. *et al.* Structural grading of foveal hypoplasia using spectral-domain optical coherence tomography: a predictor of visual acuity? *Ophthalmology* **118**, 1653–1660 (2011).
- Dorey, S., Neveu, M., Burton, L., Sloper, J. & Holder, G. The clinical features of albinism and their correlation with visual evoked potentials. *Br. J. Ophthalmol.* **87**, 767–772 (2003).
- McCafferty, B. K. *et al.* Clinical Insights Into Foveal Morphology in Albinism. *J. Pediatr. Ophthalmol. Strabismus* **52**, 167–172, doi:10.3928/01913913-20150427-06 (2015).
- Wolf, A. B., Rubin, S. E. & Kodsi, S. R. Comparison of Clinical Findings in Pediatric Patients With Albinism and Different Amplitudes of Nystagmus. *Journal of American Association for Pediatric Ophthalmology and Strabismus* **9**, 363–368, doi:http://dx.doi.org/10.1016/j.jaapos.2005.03.003 (2005).
- Montoliu, L. *et al.* Increasing the complexity: new genes and new types of albinism. *Pigment Cell & Melanoma Research* **27**, 11–18, doi:10.1111/pcmr.12167 (2014).
- Oetting, W. S. & King, R. A. Molecular basis of albinism: mutations and polymorphisms of pigmentation genes associated with albinism. *Hum. Mutat.* **13**, 99–115, doi:10.1002/(sici)1098-1004(1999)13:2<99::aid-humu2>3.0.co;2-c (1999).
- King, R. *et al.* Tyrosinase gene mutations in oculocutaneous albinism 1 (OCA1): definition of the phenotype. *Hum. Genet.* **113**, 502–513, doi:10.1007/s00439-003-0998-1 (2003).
- Hingorani, M., Williamson, K. A., Moore, A. T. & van Heyningen, V. Detailed ophthalmologic evaluation of 43 individuals with PAX6 mutations. *Invest. Ophthalmol. Vis. Sci.* **50**, 2581–2590, doi:10.1167/iops.08-2827 (2009).
- Hutton, S. M. & Spritz, R. A. Comprehensive analysis of oculocutaneous albinism among non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. *J. Invest. Dermatol.* **128**, 2442–2450 (2008a).
- Gargiulo, A. *et al.* Molecular and clinical characterization of albinism in a large cohort of Italian patients. *Invest. Ophthalmol. Vis. Sci.* **52**, 1281–1289, doi:10.1167/iops.10-6091 (2011).
- Simeonov, D. R. *et al.* DNA Variations in Oculocutaneous Albinism: An Updated Mutation List and Current Outstanding Issues in Molecular Diagnostics. *Hum. Mutat.* **34**, 827–835, doi:10.1002/humu.22315 (2013).
- Chiang, P.-W., Spector, E. & Tsai, A. C.-H. Evidence suggesting the inheritance mode of the human P gene in skin complexion is not strictly recessive. *American Journal of Medical Genetics Part A* **146A**, 1493–1496, doi:10.1002/ajmg.a.32321 (2008a).
- Chiang, P.-W., Fulton, A. B., Spector, E. & Hisama, F. M. Synergistic interaction of the OCA2 and OCA3 genes in a family. *American Journal of Medical Genetics Part A* **146A**, 2427–2430, doi:10.1002/ajmg.a.32453 (2008b).
- Jagirdar, K. *et al.* Molecular analysis of common polymorphisms within the human Tyrosinase locus and genetic association with pigmentation traits. *Pigment cell & melanoma research* **27**, 552–564, doi:10.1111/pcmr.12253 (2014).
- Mondal, M., Sengupta, M. & Ray, K. Functional assessment of tyrosinase variants identified in individuals with albinism is essential for unequivocal determination of genotype to phenotype correlation. *Br. J. Dermatol.* doi:10.1111/bjd.14977 (2016).
- Chiang, P. W., Spector, E. & Tsai, A. C. Oculocutaneous albinism spectrum. *Am. J. Med. Genet. A* **149a**, 1590–1591, doi:10.1002/ajmg.a.32939 (2009).
- Preisling, M. N., Forster, H., Gonser, M. & Lorenz, B. Screening of *TYR*, *OCA2*, *GPR143*, and *MC1R* in patients with congenital nystagmus, macular hypoplasia, and fundus hypopigmentation indicating albinism. *Mol. Vis.* **17**, 939 (2011).
- Fukai, K. *et al.* Autosomal recessive ocular albinism associated with a functionally significant tyrosinase gene polymorphism. *Nat. Genet.* **9**, 92–95 (1995).

20. Hutton, S. M. & Spritz, R. A. A Comprehensive Genetic Study of Autosomal Recessive Ocular Albinism in Caucasian Patients. *Invest. Ophthalmol. Vis. Sci.* **49**, 868–872, doi:10.1167/iov.07-0791 (2008b).
21. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, doi:10.1038/nature15393 (2015).
22. Tripathi, R. K., Giebel, L. B., Strunk, K. M. & Spritz, R. A. A polymorphism of the human tyrosinase gene is associated with temperature-sensitive enzymatic activity. *Gene Expr* **1**, 103–110 (1991).
23. Toyofuku, K. *et al.* Oculocutaneous albinism types 1 and 3 are ER retention diseases: mutation of tyrosinase or Tyrp1 can affect the processing of both mutant and wild-type proteins. *FASEB J.* **15**, 2149–2161, doi:10.1096/fj.01-0216com (2001).
24. Chaki, M. *et al.* Molecular and functional studies of tyrosinase variants among Indian oculocutaneous albinism type 1 patients. *J. Invest. Dermatol.* **131**, 260–262, doi:10.1038/jid.2010.274 (2011).
25. Oetting, W. S. *et al.* The R402Q tyrosinase variant does not cause autosomal recessive ocular albinism. *American Journal of Medical Genetics Part A* **149A**, 466–469, doi:10.1002/ajmg.a.32654 (2009).
26. Eichers, E. R., Lewis, R. A., Katsanis, N. & Lupski, J. R. Triallelic inheritance: a bridge between Mendelian and multifactorial traits. *Ann. Med.* **36**, 262–272 (2004).
27. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, doi:10.1093/bioinformatics/btp352 (2009).
28. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164, doi:10.1093/nar/gkq603 (2010).
29. Stenson, P. D. *et al.* The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum. Genet.* **133**, 1–9, doi:10.1007/s00439-013-1358-4 (2014).
30. Ng, P. C. & Henikoff, S. Predicting Deleterious Amino Acid Substitutions. *Genome Res.* **11**, 863–874, doi:10.1101/gr.176601 (2001).
31. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nature methods* **7**, 248–249, doi:10.1038/nmeth0410-248 (2010).
32. Cooper, G. M. *et al.* Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nature methods* **7**, 250–251, doi:10.1038/nmeth0410-250 (2010).
33. Chaki, M., Mukhopadhyay, A. & Ray, K. Determination of variants in the 3'-region of the tyrosinase gene requires locus specific amplification. *Hum. Mutat.* **26**, 53–58, doi:10.1002/humu.20171 (2005).
34. Schouten, J. P. *et al.* Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. *Nucleic Acids Res* **30**, e57 (2002).
35. Grønskov, K. *et al.* Birth Prevalence and Mutation Spectrum in Danish Patients with Autosomal Recessive Albinism. *Invest. Ophthalmol. Vis. Sci.* **50**, 1058–1064, doi:10.1167/iov.08-2639 (2009).
36. Mondal, M., Sengupta, M., Samanta, S., Sil, A. & Ray, K. Molecular basis of albinism in India: evaluation of seven potential candidate genes and some new findings. *Gene* **511**, 470–474, doi:10.1016/j.gene.2012.09.012 (2012).
37. Opitz, S., Käsmann-Kellner, B., Kaufmann, M., Schwinger, E. & Zühlke, C. Detection of 53 novel DNA variations within the tyrosinase gene and accumulation of mutations in 17 patients with albinism. *Hum. Mutat.* **23**, 630–631 (2004).
38. Oetting, W. S., Fryer, J. P., Shriram, S. & King, R. A. Oculocutaneous albinism type 1: the last 100 years. *Pigment Cell Res* **16**, 307–311 (2003).
39. Dolinska, M. B. *et al.* Oculocutaneous Albinism Type 1: Link between Mutations, Tyrosinase Conformational Stability, and Enzymatic Activity. *Pigment cell & melanoma research*, doi:10.1111/pcmr.12546 (2016).
40. Rooryck, C. *et al.* Molecular diagnosis of oculocutaneous albinism: new mutations in the OCA1-4 genes and practical aspects. *Pigment cell & melanoma research* **21**, 583–587, doi:10.1111/j.1755-148X.2008.00496.x (2008).
41. Oetting, W. S. *et al.* Three different frameshift mutations of the tyrosinase gene in type IA oculocutaneous albinism. *Am. J. Hum. Genet.* **49**, 199–206 (1991).
42. Oetting, W. S., Garrett, S. S., Brott, M. & King, R. A. P gene mutations associated with oculocutaneous albinism type II (OCA2). *Hum. Mutat.* **25**, 323, doi:10.1002/humu.9318 (2005).
43. Fryer, J. P., Oetting, W. S. & King, R. A. Identification and characterization of a DNase hypersensitive region of the human tyrosinase gene. *Pigment Cell Res* **16**, 679–684 (2003).
44. Machiela, M. J. & Chanock, S. J. LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557, doi:10.1093/bioinformatics/btv402 (2015).
45. Richards, S. *et al.* 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**, 405–423, doi:10.1038/gim.2015.30 (2015).
46. Kausar, T., Bhatti, M., Ali, M., Shaikh, R. & Ahmed, Z. OCA5, a novel locus for non-syndromic oculocutaneous albinism, maps to chromosome 4q24. *Clin. Genet.* **84**, 91–93 (2013).

Acknowledgements

We thank the families for their participation in this study.

Author Contributions

C.S.N. and L.O.G. have contributed equally, to the wet lab and bioinformatics work in addition to manuscript preparation. They will share first authorship. J.G. and R.J.P. performed some of the bioinformatics analysis in addition to manuscript preparation and study design. D.B. and J.A.R. contributed to manuscript preparation and study design. H.G., M.R.Z., D.B., R.P., T.N., C.M. and D.W. all assisted in wet lab experiments and manuscript preparation. M.R., H.L. and F.S. all performed clinical aspects of the study and contributed to manuscript preparation. S.E. and J.E.S. contributed to study design, project oversight, bioinformatics work and manuscript preparation.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-04401-5

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2017

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/323954868>

Mutations in TYR and OCA2 associated with oculocutaneous albinism in Pakistani families

Article in *Meta Gene* · March 2018

DOI: 10.1016/j.mgene.2018.03.007

CITATION

1

READS

335

19 authors, including:



Muhammad waqar Arshad

International Islamic University, Islamabad

10 PUBLICATIONS 8 CITATIONS

[SEE PROFILE](#)



Gaurav V Harlalka

University of Exeter Medical School, Exeter, Devon, UK

38 PUBLICATIONS 337 CITATIONS

[SEE PROFILE](#)



Siying Lin

University of Exeter

22 PUBLICATIONS 12 CITATIONS

[SEE PROFILE](#)



Ilaria D'Atri

University of Exeter

12 PUBLICATIONS 22 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



Human molecular genetics [View project](#)



Genetics of Neurological disorder in Oman (Rare Disorders) [View project](#)

Accepted Manuscript

Mutations in TYR and OCA2 associated with oculocutaneous albinism in Pakistani families.



Muhammad Waqar Arshad, Gaurav V. Harlalka, Siying Lin, Ilaria D'Atri, Sarmad Mehmood, Muhammad Shakil, Muhammad Jawad Hassan, Barry A. Chioza, Jay E. Self, Sarah Ennis, Luke O'Gorman, Chelsea Norman, Tahir Aman, Shaheer Sabz Ali, Haiba Kaul, Emma L. Baple, Andrew H. Crosby, Muhammad Ikram Ullah, Muhammad Imran Shabbir

PII: S2214-5400(18)30034-3
DOI: doi:[10.1016/j.mgene.2018.03.007](https://doi.org/10.1016/j.mgene.2018.03.007)
Reference: MGENE 419
To appear in: *Meta Gene*
Received date: 23 December 2017
Revised date: 1 March 2018
Accepted date: 21 March 2018

Please cite this article as: Muhammad Waqar Arshad, Gaurav V. Harlalka, Siying Lin, Ilaria D'Atri, Sarmad Mehmood, Muhammad Shakil, Muhammad Jawad Hassan, Barry A. Chioza, Jay E. Self, Sarah Ennis, Luke O'Gorman, Chelsea Norman, Tahir Aman, Shaheer Sabz Ali, Haiba Kaul, Emma L. Baple, Andrew H. Crosby, Muhammad Ikram Ullah, Muhammad Imran Shabbir, Mutations in TYR and OCA2 associated with oculocutaneous albinism in Pakistani families.. The address for the corresponding author was captured as affiliation for all authors. Please check if appropriate. Mgene(2017), doi:[10.1016/j.mgene.2018.03.007](https://doi.org/10.1016/j.mgene.2018.03.007)

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Mutations in *TYR* and *OCA2* associated with oculocutaneous albinism in Pakistani families.

Muhammad Waqar Arshad¹, Gaurav V Harlalka², Siying Lin², Ilaria D'Atri², Sarmad Mehmood³, Muhammad Shakil⁴, Muhammad Jawad Hassan³, Barry A Chioza², Jay E Self⁵, Sarah Ennis⁶, Luke O'Gorman⁷, Chelsea Norman⁵, Tahir Aman¹, Shaheer Sabz Ali¹, Haiba Kaul⁸, Emma L Baple², Andrew H Crosby², Muhammad Ikram Ullah⁴, Muhammad Imran Shabbir¹

1. Department of Bioinformatics and Biotechnology, International Islamic University Islamabad, Pakistan
2. RILD Wellcome Wolfson Centre, Royal Devon & Exeter NHS Foundation Trust, Barrack Road, Exeter, EX2 5DW, UK
3. Atta ur Rahman School of Applied Biosciences, National University of Sciences & Technology, Sector H-12, Islamabad, 44000, Pakistan
4. Department of Biochemistry, University of Health Sciences, Lahore-54600, Pakistan.
5. Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton
6. Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton
7. Human Development and Health, Faculty of Medicine, University of Southampton
8. Genetics Division, Department of Livestock Production, Faculty of Animal Production and Technology, University of Veterinary and Animal Sciences, Ravi Campus, Pattoki, Pakistan

Abstract

Background: Oculocutaneous albinism (OCA) is a genetically heterogeneous disorder of abnormal melanin synthesis, resulting in decreased or absent pigmentation of eyes, skin and hair. OCA has been classified based on genetic findings into seven subtypes (OCA 1-7). OCA1 is the most common subtype, accounting for 50% of cases worldwide (Hutton and Spritz, 2008; Rooryck et al., 2008), and is caused by mutations in the tyrosinase (*TYR*) gene. This study describes genetic investigations in 11 families from Pakistan with individuals with OCA.

Methods: Whole genome SNP genotyping for autozygosity mapping was undertaken using the Illumina Human CytoSNP-12 array, and exome sequencing performed using the Illumina TruSight One sequencing panel. For individuals putatively linked to the *TYR* gene, dideoxy sequencing of *TYR* was performed using primers targeting all five coding exons and intron-exon splice sites to identify mutations in individuals diagnosed with OCA. Dideoxy sequencing was also performed to confirm the presence and cosegregation of *TYR* and *OCA2* variants identified via exome sequencing.

Results: We identified new and previously reported variations in *TYR* and *OCA2* genes in 11 OCA families from Pakistan. One novel missense variant in *TYR* (c.240G>C; p.Trp80Cys), and three novel variants in *OCA2* (missense variants c.2458T>C; p.Ser820Pro and c.1762C>T; p.Arg588Trp, as well as a frameshift variant c.408_409delTT; p.Arg137Ilefs*83), were observed in five OCA families. In addition, four previously identified variants in *TYR* (c.649C>T; p.Arg217Trp, c.1255G>A; p.Gly419Arg, c.832C>T; p.Arg278Term, and c.132T>A p.Ser44Arg) and three previously identified variants in *OCA2* (c.1045-15T>G, c.2020C>G; p.Leu674Val and c.1327G>A; p.Val443Ile) were identified in nine OCA families. All affected individuals displayed the cardinal features of OCA with white hair, pale skin, nystagmus and decreased vision.

Conclusions: Our findings broaden the molecular spectrum associated with *TYR* and *OCA2* mutations in Pakistani families, aiding the development and refinement of genetic diagnostic and counselling services in Pakistan.

Key words: TYR; oculocutaneous albinism; OCA; OCA2; Pakistan; mutations

ACCEPTED MANUSCRIPT

Background

Oculocutaneous albinism (OCA) is a genetically heterogeneous disorder of abnormal melanin synthesis, resulting in absent or decreased pigmentation of eyes, skin and hair. OCA can be further classified into syndromic and non-syndromic forms with various inheritance patterns displayed. Over 450 different pathogenic sequence variants have been documented to cause the nonsyndromic form of OCA [HGMD Professional 2017.1, https://portal.biobase-international.com/hgmd/pro/search_gene.php], with the majority of these mutations located in the *TYR* gene responsible for the commonest OCA1 subtype.

TYR (MIM# 606933) encodes the enzyme tyrosinase, which has a central role in regulating the biosynthesis of melanin in melanocytes. Tyrosinase catalyses multiple steps in melanin synthesis, including the critical first and second reactions: the hydroxylation of tyrosine to L-DOPA and the oxidation of L-DOPA to DOPA quinone. Mutations in *TYR* can cause complete (tyrosine-negative) or partial (tyrosine-positive) OCA depending on residual enzyme activity (Simeonov et al., 2013). The severe clinical phenotype of OCA1A (MIM# 203100), characterised by an almost complete absence of hair, iris and skin pigmentation, correlates with pathogenic or null alleles in *TYR* (Schnur et al., 1996). Hypomorphic *TYR* mutations instead cause OCA1B (MIM# 606952), where residual (albeit reduced) tyrosinase activity results in a milder phenotype with reduced levels of pigmentation in affected individuals (Norman et al., 2017).

The *OCA2* gene (MIM# 203200) encompasses 23 coding exons along encoding a polypeptide product of ~110 kDa, with 12 transmembrane helices. The molecule belongs to the Na⁺/H⁺ antiporter family and functions in maintaining the pH of melanosomes, resulting in the regulation of tyrosinase activity (Gershoni-Baruch et al., 1994; Matsunaga et al., 1998; Puri et al., 2000; Preising et al., 2007). The *OCA2* protein also participates in the sorting and

transportation of tyrosinase and tyrosinase-related protein 1 (TYRP1) to the plasma membrane (Rosemlat et al., 1998; Orlow and Brilliant, 1999; Chen et al., 2002).

TYRP1 mutations (MIM# 115501) have also been shown to cause OCA (OCA3; MIM 203290). This gene has seven coding exons (GenBank NM_000550) encoding a tyrosinase-related protein of ~61 kDa in size, displaying 58% similarity and 41% sequence identity to tyrosinase. The *TYRP1* gene has a dual role in the oxidation of 5,6-dihydroxyindole-2-carboxylic acid in melanin biosynthesis as well as in tyrosinase hydroxylase activity (Cruz-Inigo et al., 2011). Mutations in the *SLC45A2* gene (MIM# 606202) cause OCA4 (MIM# 606574). This gene has seven coding exons transcribed alternatively into four alternatively spliced variants, with 108 pathogenic variants reported in HGMD® Professional 2017.1 database. The longest spliced isoform (GenBank NM_016180) codes for a solute carrier family 45. The function of the *SLC45A2* protein, which comprises of 530 amino acids with a molecular weight of ~58 kDa, is currently unclear but it is assumed to be a melanosomal protein with a role in subcellular transportation (Fukamachi et al., 2001; Fernandez et al., 2008).

In community settings, founder mutations are often present at increased allele frequency. Identifying the spectrum and frequency of disease-causing variants is important to enable the development of population-specific genetic testing strategies targeting variants common to the local population. This will permit rapid and cost efficient screening and diagnostic assays that will allow accurate disease diagnosis, improved carrier detection and appropriate counselling for affected families. Here we report our genetic findings regarding the causes of OCA in families from several communities in Pakistan, which advance our understanding of the relative contribution of pathogenic *TYR* and *OCA2* variants to OCA in Pakistan

Methods

Patients and family members

This report describes genetic and clinical investigations in 11 Pakistani families, undertaken with informed consent according to appropriate regional ethical approvals. The 11 families, belonging to different ethnic groups (Pathan, Afridi, Khatak, Yousafzai, Punjabi, Rajpoot, Saraki, Virk and Niazi) were recruited from several different provinces of Pakistan. Medical histories from all 11 families were taken and a diagnosis of OCA in all the affected individuals was established (Fig 1 and 2). Following results from genetic analyses and the identification of pathogenic variants in the *TYR* and *OCA2* genes, affected members of all families were revisited, and a detailed medical history was ascertained including documentation of symptoms. Facial photographs and videos were used to document clinical features and confirm disease status. Visual acuity testing using Snellen charts, colour vision testing using Ishihara charts and funduscopy examination by direct ophthalmoscopy was performed in the affected individuals examined.

Molecular genetic analysis

Peripheral blood samples were taken from each participating individual for genomic DNA extraction using standard procedures. In order to identify the homozygous regions shared by the affected individuals from each family, a whole genome scan was undertaken using an Illumina Human CytoSNP-12 array comprising of ~330,000 genetic markers, as described previously (Zollo et al., 2017). Families in which affected individuals displayed homozygosity at the *TYR* locus (families 1-5) were used for targeted dideoxy sequencing (ABI3730 sequencer) of all the five coding exons, and associated intron-exon junctions in *TYR* gene.

Next generation sequencing was performed on an affected individual in families 6-11 using the Illumina TruSight One clinical exome sequencing panel as described previously (Norman et al., 2017). This panel provides targeted sequencing for over 4800 genes associated with clinical phenotypes, and includes 17 of the known causative genes for OCA (Table 1).

(Ammann et al., 2016). Sequence reads were aligned to the human genome reference sequence [hg19] to observe base pair changes using CLC sequence viewer (<https://www.qiagenbioinformatics.com/products/clc-sequence-viewer/>) and Chromas Lite (<http://technelysium.com.au/wp/chromas/>) software. Variants identified via next generation sequencing within genes of interest were filtered using 1000 Genomes Project Minor Allele Frequency (MAF)(<0.05) and the *in silico* pathogenicity prediction tools SIFT (<0.05), PolyPhen2 HumVar (possibly damaging and probably damaging) and GERP++ (>2) as previously described (Norman et al., 2017). Dideoxy sequencing was performed to confirm the presence and cosegregation of *TYR* and *OCA2* variants identified in these families via clinical exome sequencing.

Results

Clinical findings

Study subjects from 11 families with congenital nonsyndromic OCA were enrolled from different provinces in Pakistan (families 1, 2, 8 and 10 from the Khyber Pakhtunkhwa province, families 3-5, 7, 9 and 11 from the Punjab province and family 6 from the Balochistan province). All affected individuals were born to consanguineous unaffected parents, and all pedigrees were consistent with an autosomal recessive mode of inheritance. All affected individuals displayed the cardinal clinical features of OCA with white hair, pale skin and nystagmus, and reported symptoms of photophobia and decreased visual acuity. Clinical examination in all affected individuals examined demonstrated features of foveal hypoplasia and a hypopigmented albinotic fundus on direct ophthalmoscopy. Clinical findings are summarised in Tables 5 and 6.

Genetic findings

Sequencing of the coding regions of *TYR* gene revealed a novel missense mutation c.240G>C; p.Trp80Cys in the first coding exon of *TYR* in two families (families 1 and 2) which co-segregated appropriately. Exome sequencing identified three novel *OCA2* variants (missense variants c.2458T>C; p.Ser820Pro and c.1762C>T; p.Arg588Trp, as well as a frameshift variant c.408_409delTT; p.Arg137Ilefs*83) in five families. *In silico* analyses were undertaken using various mutation prediction tools such as SIFT, PolyPhen-2, Mutation Taster and PROVEAN, which suggested that these variants were deleterious (Table 2).

Three further missense mutations (c.649C>T; p.Arg217Trp in family 3, c.1255G>A; p.Gly419Arg in family 4 and c.132T>A; p.Ser44Arg in family 6), and a nonsense mutation (c.832C>T; p.Arg278Term in family 5), were identified in the coding regions of *TYR* (Table 3).

Exome sequencing identified three previously described *OCA2* variants (c.1045-15T>G, c.2020C>G; p.Leu674Val and c.1327G>A; p.Val443Ile) in a further four families (families 7, 8, 10 and 11) (Table 4) (Lee et al., 1994; Jaworek et al., 2012; Mondal et al., 2012; Zhang et al.,

2013; Shahzad et al., 2017). This is the first time that two of the three previously described OCA2 variants identified here (c.2020C>G; p.Leu674Val and c.1327G>A; p.Val443Ile) have been reported in Pakistan.

None of the *TYR* variants are listed in homozygous form in online gnomAD genomic database (<http://gnomad.broadinstitute.org>). The OCA2 variants c.1045-15T>G, c.2458T>C; p.Ser820Pro and c.408_409delTT; p.Arg137Ilefs*83 were also absent in homozygous form in the online gnomAD genomic database. The OCA2 variants c.2020C>G; p.Leu674Val, c.1327G>A; p.Val443Ile and c.1762C>T; p.Arg588Trp are present in homozygous form in gnomAD in one, four and five individuals respectively. Minor allele frequencies for the above variants are listed in Tables 3 and 4. Localisation of missense variants within TYR and OCA2 are shown in figure 3.

Discussion:

Tyrosinase is an oxidase enzyme, and is the rate-limiting step in regulating the synthesis of melanin. This enzyme catalyses two reactions in melanin synthesis; firstly the hydroxylation of a monophenol, and secondly the conversion of an o-diphenol to the corresponding o-quinone, which is subsequently converted to melanin. *TYR* and *OCA2* variants are common causes of OCA in Pakistan, each accounting for approximately a third of all known OCA variants in this community (Shahzad et al., 2017). In this study, we report known and novel mutations in *TYR* and *OCA2* in families from Pakistan. The two families in which a novel *TYR* (c.240G>C;p.Trp80Cys) mutation was observed were of Pakhtoon origin, located in different geographical localities within the same province (Khyber Pakhtunkhwa) in North West Pakistan. Affected individuals in both families presented with typical features including white hair, reddish white to white skin, nystagmus and de-pigmented irides with decreased visual acuity. Two other variants affecting the p.Trp80 amino acid residue have previously been reported; p.Trp80Term and p.Trp80Arg (rs61753188), for which only a single heterozygous gene carrier is listed in gnomAD browser database. Neither variant, nor the p.Trp80Cys variant defined here, is listed as present in the South Asian population in online genome databases indicating that these are likely to be very rare variants in this population. Our study further identified three novel *OCA2* variants (missense variants c.2458T>C; p.Ser820Pro and c.1762C>T; p.Arg588Trp, as well as a frameshift variant c.408_409delTT; p.Arg137Ilefs*83) in five families, as well as three previously described *OCA2* variants in a further four families, and all families displayed the cardinal features of OCA (Tables 5 and 6).

Several of the *TYR* and *OCA2* variants described in our study appear to be commonly associated with OCA in Pakistan, and likely represent regional founder mutations. For example, the c.832C>T;p.Arg278Term, c.1255G>A;p.Gly419Arg and c.649C>T;p.Arg217Trp variants in *TYR* identified in families 3-5, as well as the c.1045-15T>G variant in *OCA2* identified in families 7 and 8, account for 12.9%, 9.5%, 4.3% and 11.3% of all known OCA variants in Pakistan respectively (Shahzad et al., 2017).

This is the also the first time that two of the *OCA2* variants identified in this study have been reported in Pakistani families. The c.2020C>G; p.Leu674Val variant has only previously been reported in an Indian family and the c.1327G>A; p.Val443Ile variant has only been reported in Caucasian and Chinese individuals (Lee et al., 1994; Mondal et al., 2012; Zhang et al., 2013).

Three of the heterozygous *OCA2* variants identified in our study (c.2020C>G; p.Leu674Val, c.1327G>A; p.Val443Ile and c.1762C>T; p.Arg588Trp) are present in gnomAD in homozygous form in one, four and five individuals respectively in the South Asian population only. The c.2020C>G; p.Leu674Val *OCA2* variant has previously been described in association with OCA in 2 Indian individuals in both homozygous and compound heterozygous form (Mondal et al., 2012). Both individuals exhibited an incomplete albinism phenotype, with the individual heterozygous for the c.2020C>G; p.Leu674Val as well as a c.775_776insG variant showing clinical features of nystagmus, hazel irides, light golden-brown hair and pinkish skin. The individual homozygous for the c.2020C>G; p.Leu674Val variant showed clinical features of, brown irides with iris transillumination, silky-brown hair and very fair pinkish skin with no apparent nystagmus (Mondal et al., 2012). We detected the same c.2020C>G; p.Leu674Val variant in compound heterozygous form in family 10, together with a novel frameshift variant with affected individuals displaying an incomplete OCA phenotype with nystagmus, blue irides, light brown hair and pink skin. The c.2020C>G; p.Leu674Val variant may therefore represent a milder *OCA2* mutation contributing to incomplete OCA phenotype in when occurring in conjunction in compound heterozygous form with a more deleterious OCA variant, and with homozygotes exhibiting a mild phenotype, which may account for the single homozygous individual in gnomAD.

Hypomorphic *TYR* variants are well defined in the aetiology of the OCA1B subtype, where reduced tyrosinase activity results in a milder phenotype with reduced level of pigmentation in

affected individuals. The c.1327G>A; p.Val443Ile and c.1762C>T; p.Arg588Trp variants identified here may represent hypomorphic variants that cause only a partial loss of *OCA2* gene function, accounting for the four and five homozygous individuals respectively present in the gnomAD database. The c.1327G>A; p.Val443Ile variant has been previously described in individuals with OCA in two different populations (Northern European and Chinese), suggesting that this is indeed likely a pathogenic variant, and in the individual heterozygous for this variant there were no other candidate variants in any of the known OCA associated genes apart from the c.1762C>T; p.Arg588Trp variant. Ultimately functional characterisation of the variant would be helpful to determine the nature of the mutation and the extent of its biological impairment (Kamaraj and Purohit, 2014).

In family 8, the previously described disease-associated *OCA2* c.1045-15T>G splice variant was detected in six out of seven affected individuals in the family, and in heterozygous form in the seventh affected individual in which there were no other candidate variants in any of the known OCA associated genes detected. Interestingly, this heterozygous *OCA2* c.1045-15T>G individual displayed a somewhat different phenotype compared to affected individuals who were homozygous for the variant; with a darker hair colour (golden brown instead of blonde). There is a high level of missing heritability in OCA, with 25% of patients investigated only having detectable mutations in a single OCA allele (Simeonov et al., 2013). It is known that the phenotype of *OCA2* can be modified by *MC1R* or *TYRP1* mutations, demonstrating a synergistic interaction between genes throughout this pigment pathway (King et al., 2003; Chiang et al., 2008). Consequently the missing heritability in this individual may reflect an undetected mutation in a pigment pathway gene that interacts with *OCA2*, or in the *OCA2* gene promoter or other regulatory region. Interestingly, a causal tri-allelic genotype where a combination of two common hypomorphic *TYR* variants in *trans* to a known *TYR* deleterious mutation has recently been hypothesised to account for cases of OCA1 with apparent missing heritability; this theory may similarly explain the missing heritability in *OCA2* our family (Norman et al., 2017).

In summary, these studies expand current knowledge of the molecular spectrum and specific genetic causes of OCA, which while specific population frequency is lacking, appears to be relatively common in Pakistani communities. In combination with existing datasets, these studies enable accurate genetic testing and provide valuable information to aid the diagnosis and counselling of affected individuals and family members throughout Pakistan.

Ethics approval and consent to participate

This study was approved by the ethical approval committee institutional review board of the International Islamic University, Islamabad, Pakistan.

Consent for publication

Written consent was obtained from all patients or their relatives for publication.

Availability of data and materials

Data supporting the conclusions of this article are included within the article.

Competing interest

The authors declare that they have no competing interests.

Author contributions

MWA, MS, MJH, TA, SSA, HK, MIU and MIS provided samples and clinical details. GH, MWA, SM, BAC, ID, SL, JES, CN, LO and SE performed genetic studies, and analysed data alongside SL, ID and BAC. MIS, JES, AHC, ELB and MIU designed and conceived studies. SL aided compilation and analysis of clinical information, and edited the manuscript with AHC, ELB and JES.

Acknowledgements

Authors would like to thank the family members for their involvement in this study. The authors are also grateful to the Department of Bioinformatics and Biotechnology, IIU, Islamabad, Pakistan, the Higher Education Commission (HEC), Pakistan and the Warman Foundation for funding this research study.

Web resources (URLs)

DNA Nexus, <https://dnanexus.com>

UCSC Genome Browser, <http://genome.ucsc.edu/>

Ensembl, http://www.ensembl.org/Homo_sapiens/Info/Index

Primer3, <http://primer3.ut.ee/>

PolyPhen-2, <http://genetics.bwh.harvard.edu/pph2/>

SIFT, http://sift.jcvi.org/www/SIFT_enst_submit.html

PROVEAN, http://provean.jcvi.org/seq_submit.php

UniProt, <http://www.uniprot.org/>

Exome Variant Server, <http://evs.gs.washington.edu/EVS/>

The Human Gene Mutation Database, <http://www.hgmd.cf.ac.uk/ac/index.php>

Exome Aggregation Consortium, <http://exac.broadinstitute.org/>

dbSNP, <http://www.ncbi.nlm.nih.gov/SNP/>

PROTTER, <http://wlab.ethz.ch/protter/start/>

Form of oculocutaneous albinism	Gene	Associated patterns of inheritance	Chromosome location	Mutations
OCA1	<i>TYR</i>	Autosomal recessive	11q14-q21	379
OCA2	<i>OCA2</i>	Autosomal recessive	15q11.2-q12	205
OCA3	<i>TYRP1</i>	Autosomal recessive	9p23	29
OCA4	<i>SLC45A2</i>	Autosomal dominant/recessive	5p13.3	108
OCA5	Unidentified	Autosomal recessive	4q24	0
OCA6	<i>SLC24A5</i>	Autosomal recessive	15q21.1	13
OCA7	<i>C10ORF11</i>	Autosomal recessive	10q22.2-q22.3	8
CHS	<i>LYST</i>	Autosomal recessive	1q42.1-q42.2	79
HPS1	<i>HPS1</i>	Autosomal recessive	10q23.1-q23.3	42
HPS2	<i>AP3B1</i>	Autosomal recessive	5q14.1	24
HPS3	<i>HPS3</i>	Autosomal recessive	3q24	13
HPS4	<i>HPS4</i>	Autosomal recessive	22cen-q12.3	15
HPS5	<i>HPS5</i>	Autosomal recessive	11p14	14
HPS6	<i>HPS6</i>	Autosomal recessive	10q24.32	25
HPS7	<i>DTNBP1</i>	Autosomal recessive	6p22.3	3
HPS8	<i>BLOC1S3</i>	Autosomal recessive	19q13.32	2
HPS9	<i>BLOC1S6</i>	Autosomal recessive	15q21.1	1
HPS10	<i>AP3D1</i>	Autosomal recessive	19p13.3	1
Source: HGMD® Professional 2017.1, 14 June 2017.				

Table 1: Various forms of oculocutaneous albinism, list of respective OCA genes, associated inheritance patterns, chromosomal locations and number of mutations identified in each gene (compiled 14 June 2017).

Family	Family 1	Family 2	Family 9	Family 10	Family 11
Gene	<i>TYR</i>	<i>TYR</i>	<i>OCA2</i>	<i>OCA2</i>	<i>OCA2</i>
Nucleotide variant	c.240G>C	c.240G>C	c.2458T>C	c.408_409delTT	c.1762C>T
Protein variant	p.Trp80Cys	p.Trp80Cys	p.Ser820Pro	p.Arg137Ilefs*83	p.Arg588Trp
Status	Homozygous	Homozygous	Homozygous	Heterozygous	Heterozygous
Type of Mutation	Missense	Missense	Missense	Frameshift	Missense
PolyPhen-2	Probably Damaging	Probably Damaging	Probably damaging	- [§]	Benign
Mutation Taster	Disease Causing	Disease Causing	Disease causing	- [§]	Polymorphism
SIFT	Damaging	Damaging	Damaging	- [§]	Damaging
PROVEAN	Deleterious	Deleterious	Deleterious	- [§]	Deleterious
Homozygous form in gnomAD	Not present	Not present	Not present	Not present	5 homozygotes
gnomAD MAF	Not present	Not present	Not present	Not present	0.001021

Table 2: Novel *TYR* and *OCA2* variants identified as a part of this study.

[§]In silico predictions not available for frameshift variant

Family	Family 1	Family 2	Family 3	Family 4	Family 5	Family 6
Nucleotide variant	c.240G>C	c.240G>C	c.649C>T	c.1255G>A	c.832C>T	c.132T>A
Protein variant	p.Trp80Cys	p.Trp80Cys	p.Arg217Trp	p.Gly419Arg	p.Arg278Term	p.Ser44Arg
Status	Homozygous	Homozygous	Homozygous	Homozygous	Homozygous	Homozygous
Type of Mutation	Missense	Missense	Missense	Missense	Nonsense	Missense
Previously reported	Novel (this study)	Novel (this study)	Yes (Tripathi et al., 1992)	Yes (Chaki et al., 2011)	Yes (Wang et al., 2016)	Yes (Shah et al., 2015)
PolyPhen-2	Probably Damaging	Probably Damaging	Probably Damaging	Probably Damaging	Probably Damaging	Probably Damaging
Mutation Taster	Disease Causing	Disease Causing	Disease Causing	Disease Causing	Disease Causing	Disease Causing
SIFT	Damaging	Damaging	Damaging	Damaging	Damaging	Damaging
PROVEAN	Deleterious	Deleterious	Deleterious	Deleterious	Deleterious	Deleterious
Homozygous form in gnomAD	Not present	Not present	Not present	Not present	Not present	Not present
gnomAD MAF	Not present	Not present	0.0001917	0.00006155	0.0001770	0.00001625

Table 3: *TYR* variants identified as a part of this study.

Family	Family 7	Family 8	Family 9	Family 10	Family 11
Nucleotide variant	c.1045-15T>G	c.1045-15T>G	c.2458T>C	c.2020C>G c.408_409delTT	c.1327G>A c.1762C>T
Protein variant	Splice site mutation	Splice site mutation	p.Ser820Pro	p.Leu674Val p.Arg137Ilefs*83	p.Val443Ile p.Arg588Trp
Status	Homozygous	Homozygous Heterozygous in 1 individual	Homozygous	Heterozygous	Heterozygous
Type of Mutation	Splice site mutation	Splice site mutation	Missense	Missense Frameshift	Missense Missense
Previously reported	Yes (Jaworek et al., 2012; Shahzad et al., 2017)	Yes (Jaworek et al., 2012; Shahzad et al., 2017)	Novel (this study)	Yes (Mondal et al., 2012) Novel (this study)	Yes (Lee et al., 1994; Zhang et al., 2013) Novel (this study)
PolyPhen-2	- [§]	- [§]	Probably damaging	Possible damaging - [§]	Probably damaging Benign
Mutation Taster	- [§]	- [§]	Disease causing	Disease causing - [§]	Disease causing automatic Polymorphism
SIFT	- [§]	- [§]	Damaging	Damaging - [§]	Damaging Damaging
PROVEAN	- [§]	- [§]	Deleterious	Neutral - [§]	Neutral Deleterious
Homozygous form in gnomAD	Not present	Not present	Not present	1 homozygote Not present	4 homozygotes 5 homozygotes
gnomAD MAF	0.00002445	0.00002445	Not present	0.0003333 Not present	0.003024 0.001021

Table 4: *OCA2* variants identified as a part of this study.[§]In silico predictions not available for frameshift variant

[§]In silico predictions for splice variant not available for PolyPhen-2, Mutation Taster, SIFT and PROVEAN. Instead in silico predictions using MaxEnt and NNSPLICE show variation scores of -15.3% and -0.4% respectively.

Parameters	Family 1: II:4	Family 2: II:1	Family 3: V:2	Family 4: IV:2	Family 5: IV:5	Family 6: II:2
Age (years)	12	10	12	5	26	9
Gender	Female	Male	Male	Female	Male	Male
Region, Province	Khyber Pakhtunkhwa	Khyber Pakhtunkhwa	Punjab	Punjab	Punjab	Balochistan
Caste	Pathan	Pathan	Rajpoot	Virk	Punjabi	Khatak
Hair color	White	White	White	Light blonde	White	White
Skin color	Reddish-white	White	White	White	White	White
Skin rashes	Absent	Absent	Absent	Absent	Absent	Absent
Visual acuity	Decreased	Decreased	Decreased	Decreased	Decreased	Decreased
Iris color	De-pigmented	De-pigmented	De-pigmented	De-pigmented	De-pigmented	De-pigmented
Photophobia	Not Present	Not Present	Present	Present	Present	Present
Nystagmus	Yes	Yes	Yes	Yes	Yes	Yes
Foveal hypoplasia	Unable to determine	Unable to determine	Present	Present	Present	Unable to determine
Fundus	Unable to determine	Unable to determine	Albinotic	Albinotic	Albinotic	Unable to determine

Table 5: Summary of clinical features observed in families 1-6 with pathogenic *TYR* variants

Parameters	Family 7: IV:4	Family 8: VI:5	Family 9: V:3	Family 10: III:1	Family 11: III:1
Age (years)	4	5		2	8
Gender	Male	Male	Female	Female	Male
Region, Province	Punjab	Khyber Pakhtunkhwa	Punjab	Khyber Pakhtunkhwa	Punjab
Caste	Niaz	Afridi	Niaz	Yousafzai	Saraiki
Hair color	Golden/blonde	White/brown	Golden	White/brown	White/brown
Skin color	White	White	Reddish-white	White	White
Skin rashes	Present	Absent	Absent	Absent	Absent
Visual acuity	Decreased	Decreased	Decreased	Decreased	Decreased
Iris color	De-pigmented	De-pigmented	De-pigmented	De-pigmented	De-pigmented
Photophobia	Present	Present	Present	Present	Present
Nystagmus	Yes	Yes	Yes	Yes	Yes
Foveal hypoplasia	Present	Unable to determine	Present	Unable to determine	Unable to determine
Fundus	Albinotic	Unable to determine	Albinotic	Unable to determine	Unable to determine

Table 6: Summary of clinical features observed in families 7-11 with pathogenic OCA2 variants

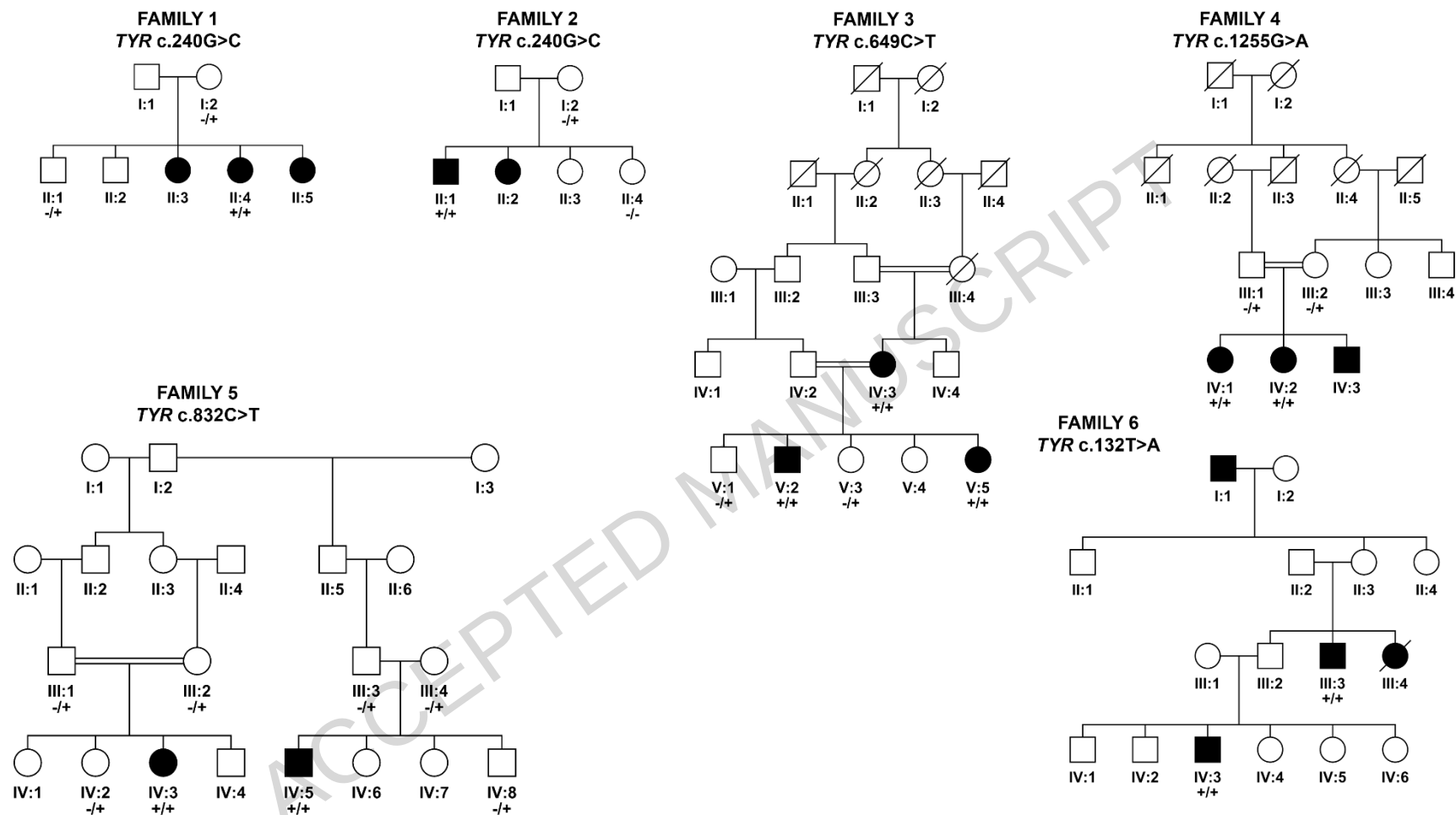


Figure 1: Pedigrees of 6 families with oculocutaneous albinism co-segregating for *TYR* mutations. Presence or absence of the variant is indicated by a + or - sign respectively.

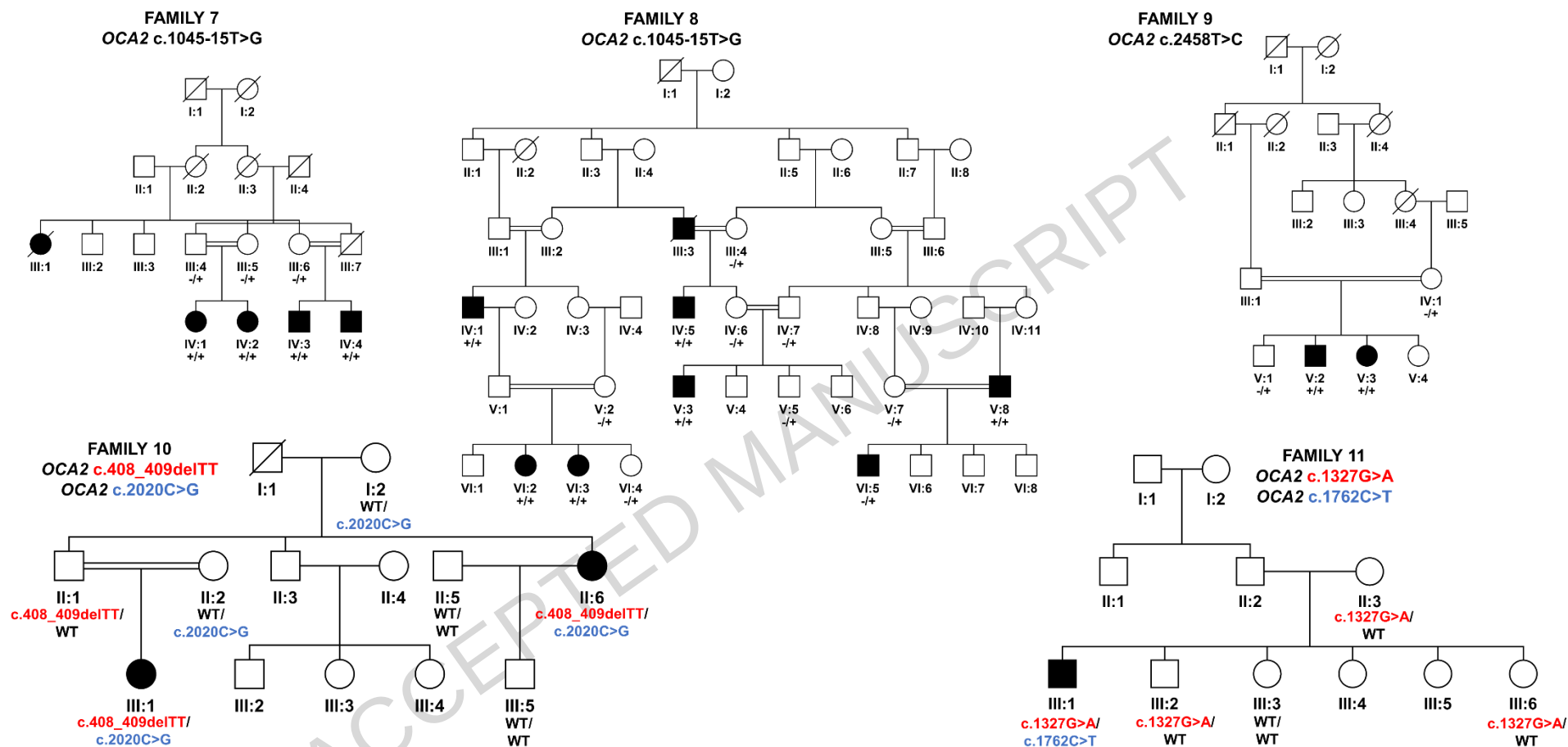


Figure 2: Pedigrees of 5 families with oculocutaneous albinism co-segregating for OCA2 mutations. Presence or absence of the variant is indicated by a + or - sign respectively. For compound heterozygous mutations, the different OCA2 variants within the same family are displayed in different colours.

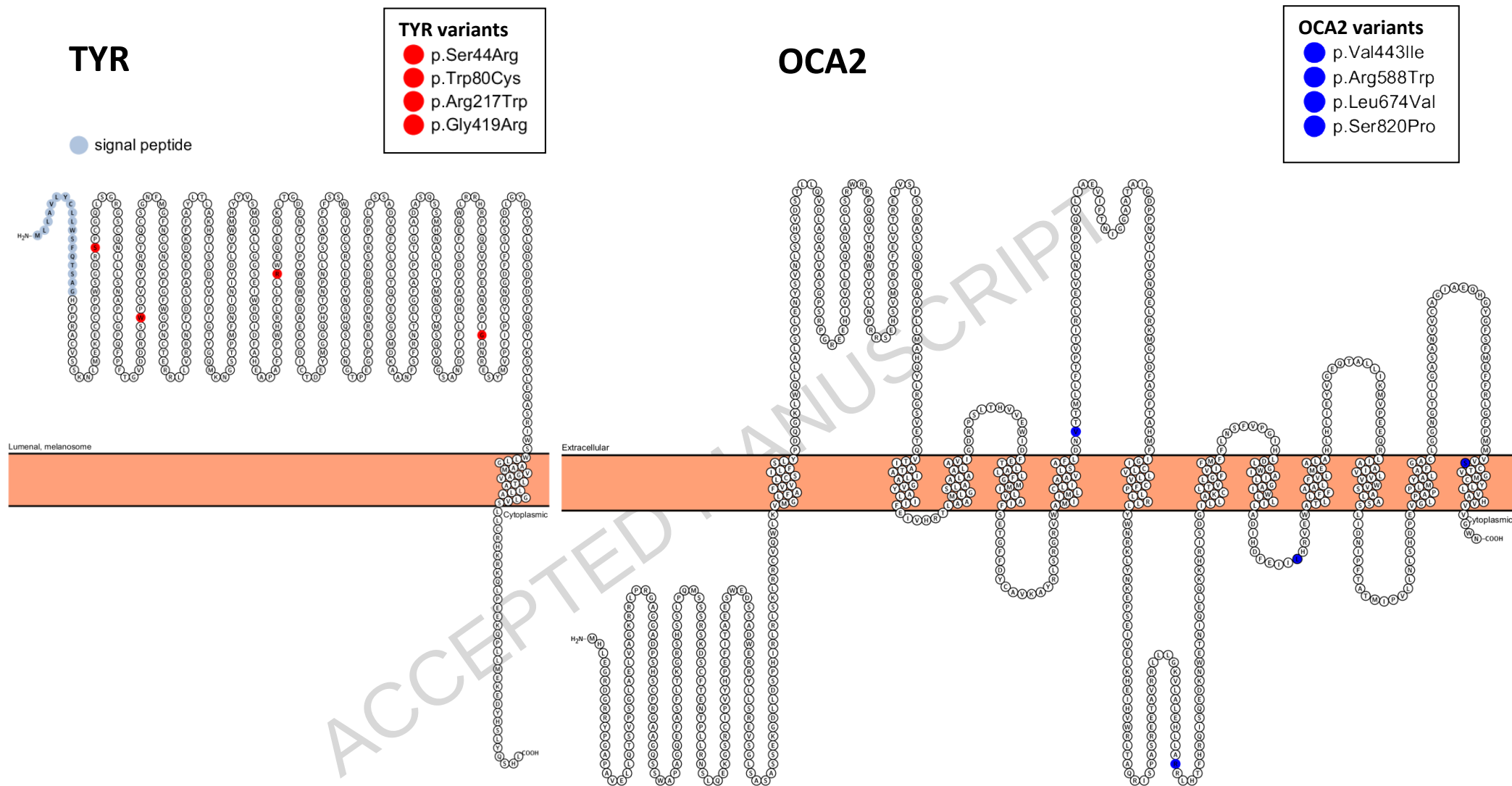


Figure 3: Localisation of the TYR and OCA2 missense variants identified in the study

REFERENCES

- Ammann, S., Schulz, A., Krageloh-Mann, I., Dieckmann, N.M., Niethammer, K., Fuchs, S., Eckl, K.M., Plank, R., Werner, R., Altmüller, J., Thiele, H., Nurnberg, P., Bank, J., Strauss, A., von Bernuth, H., Zur Stadt, U., Grieve, S., Griffiths, G.M., Lehmberg, K., Hennies, H.C. and Ehl, S., 2016. Mutations in AP3D1 associated with immunodeficiency and seizures define a new type of Hermansky-Pudlak syndrome. *Blood* 127, 997-1006.
- Chaki, M., Sengupta, M., Mondal, M., Bhattacharya, A., Mallick, S., Bhadra, R., Indian Genome Variation, C. and Ray, K., 2011. Molecular and functional studies of tyrosinase variants among Indian oculocutaneous albinism type 1 patients. *J Invest Dermatol* 131, 260-2.
- Chen, K., Manga, P. and Orlow, S.J., 2002. Pink-eyed dilution protein controls the processing of tyrosinase. *Mol Biol Cell* 13, 1953-64.
- Chiang, P.W., Fulton, A.B., Spector, E. and Hisama, F.M., 2008. Synergistic interaction of the OCA2 and OCA3 genes in a family. *Am J Med Genet A* 146A, 2427-30.
- Cruz-Inigo, A.E., Ladizinski, B. and Sethi, A., 2011. Albinism in Africa: stigma, slaughter and awareness campaigns. *Dermatol Clin* 29, 79-87.
- Fernandez, L.P., Milne, R.L., Pita, G., Aviles, J.A., Lazaro, P., Benitez, J. and Ribas, G., 2008. SLC45A2: a novel malignant melanoma-associated gene. *Hum Mutat* 29, 1161-7.
- Fukamachi, S., Shimada, A. and Shima, A., 2001. Mutations in the gene encoding B, a novel transporter protein, reduce melanin content in medaka. *Nat Genet* 28, 381-5.
- Gershoni-Baruch, R., Rosenmann, A., Droetto, S., Holmes, S., Tripathi, R.K. and Spritz, R.A., 1994. Mutations of the tyrosinase gene in patients with oculocutaneous albinism from various ethnic groups in Israel. *Am J Hum Genet* 54, 586-94.
- Hutton, S.M. and Spritz, R.A., 2008. Comprehensive analysis of oculocutaneous albinism among non-Hispanic caucasians shows that OCA1 is the most prevalent OCA type. *J Invest Dermatol* 128, 2442-50.
- Jaworek, T.J., Kausar, T., Bell, S.M., Tariq, N., Maqsood, M.I., Sohail, A., Ali, M., Iqbal, F., Rasool, S., Riazuddin, S., Shaikh, R.S. and Ahmed, Z.M., 2012. Molecular genetic studies and delineation of the oculocutaneous albinism phenotype in the Pakistani population. *Orphanet J Rare Dis* 7, 44.
- Kamaraj, B. and Purohit, R., 2014. Mutational analysis of oculocutaneous albinism: a compact review. *Biomed Res Int* 2014, 905472.
- King, R.A., Willaert, R.K., Schmidt, R.M., Pietsch, J., Savage, S., Brott, M.J., Fryer, J.P., Summers, C.G. and Oetting, W.S., 2003. MC1R mutations modify the classic phenotype of oculocutaneous albinism type 2 (OCA2). *Am J Hum Genet* 73, 638-45.

- Lee, S.T., Nicholls, R.D., Bunday, S., Laxova, R., Musarella, M. and Spritz, R.A., 1994. Mutations of the P gene in oculocutaneous albinism, ocular albinism, and Prader-Willi syndrome plus albinism. *N Engl J Med* 330, 529-34.
- Matsunaga, J., Dakeishi-Hara, M., Miyamura, Y., Nakamura, E., Tanita, M., Satomura, K. and Tomita, Y., 1998. Sequence-based diagnosis of tyrosinase-related oculocutaneous albinism: successful sequence analysis of the tyrosinase gene from blood spots dried on filter paper. *Dermatology* 196, 189-93.
- Mondal, M., Sengupta, M., Samanta, S., Sil, A. and Ray, K., 2012. Molecular basis of albinism in India: evaluation of seven potential candidate genes and some new findings. *Gene* 511, 470-4.
- Norman, C.S., O'Gorman, L., Gibson, J., Pengelly, R.J., Baralle, D., Ratnayaka, J.A., Griffiths, H., Rose-Zerilli, M., Ranger, M., Bunyan, D., Lee, H., Page, R., Newall, T., Shawkat, F., Mattocks, C., Ward, D., Ennis, S. and Self, J.E., 2017. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (TYR) causing hypomorphic oculocutaneous albinism (OCA1B). *Sci Rep* 7, 4415.
- Orlow, S.J. and Brilliant, M.H., 1999. The pink-eyed dilution locus controls the biogenesis of melanosomes and levels of melanosomal proteins in the eye. *Exp Eye Res* 68, 147-54.
- Preising, M.N., Forster, H., Tan, H., Lorenz, B., de Jong, P.T. and Plomp, A.S., 2007. Mutation analysis in a family with oculocutaneous albinism manifesting in the same generation of three branches. *Mol Vis* 13, 1851-5.
- Puri, N., Gardner, J.M. and Brilliant, M.H., 2000. Aberrant pH of melanosomes in pink-eyed dilution (p) mutant melanocytes. *J Invest Dermatol* 115, 607-13.
- Rooryck, C., Morice-Picard, F., Elcioglu, N.H., Lacombe, D., Taieb, A. and Arveiler, B., 2008. Molecular diagnosis of oculocutaneous albinism: new mutations in the OCA1-4 genes and practical aspects. *Pigment Cell Melanoma Res* 21, 583-7.
- Rosemlat, S., Sviderskaya, E.V., Easty, D.J., Wilson, A., Kwon, B.S., Bennett, D.C. and Orlow, S.J., 1998. Melanosomal defects in melanocytes from mice lacking expression of the pink-eyed dilution gene: correction by culture in the presence of excess tyrosine. *Exp Cell Res* 239, 344-52.
- Schnur, R.E., Sellinger, B.T., Holmes, S.A., Wick, P.A., Tatsumura, Y.O. and Spritz, R.A., 1996. Type I oculocutaneous albinism associated with a full-length deletion of the tyrosinase gene. *J Invest Dermatol* 106, 1137-40.
- Shah, S.A., Raheem, N., Daud, S., Mubeen, J., Shaikh, A.A., Baloch, A.H., Nadeem, A., Tayyab, M., Babar, M.E. and Ahmad, J., 2015. Mutational spectrum of the TYR and SLC45A2 genes in

- Pakistani families with oculocutaneous albinism, and potential founder effect of missense substitution (p.Arg77Gln) of tyrosinase. *Clin Exp Dermatol* 40, 774-80.
- Shahzad, M., Yousaf, S., Waryah, Y.M., Gul, H., Kausar, T., Tariq, N., Mahmood, U., Ali, M., Khan, M.A., Waryah, A.M., Shaikh, R.S., Riazuddin, S., Ahmed, Z.M. and University of Washington Center for Mendelian Genomics, C., 2017. Molecular outcomes, clinical consequences, and genetic diagnosis of Oculocutaneous Albinism in Pakistani population. *Sci Rep* 7, 44185.
- Simeonov, D.R., Wang, X., Wang, C., Sergeev, Y., Dolinska, M., Bower, M., Fischer, R., Winer, D., Dubrovsky, G., Balog, J.Z., Huizing, M., Hart, R., Zein, W.M., Gahl, W.A., Brooks, B.P. and Adams, D.R., 2013. DNA variations in oculocutaneous albinism: an updated mutation list and current outstanding issues in molecular diagnostics. *Hum Mutat* 34, 827-35.
- Tripathi, R.K., Strunk, K.M., Giebel, L.B., Weleber, R.G. and Spritz, R.A., 1992. Tyrosinase gene mutations in type I (tyrosinase-deficient) oculocutaneous albinism define two clusters of missense substitutions. *Am J Med Genet* 43, 865-71.
- Wang, X., Zhu, Y., Shen, N., Peng, J., Wang, C., Liu, H. and Lu, Y., 2016. Mutation analysis of a Chinese family with oculocutaneous albinism. *Oncotarget* 7, 84981-84988.
- Zhang, L., Xu, B., Zhong, Y., Chen, X., Zheng, H., Jiang, W. and Li, H., 2013. [A de novo mutation of P gene causes oculocutaneous albinism type 2 with prenatal diagnosis]. *Zhonghua Yi Xue Yi Chuan Xue Za Zhi* 30, 318-21.
- Zollo, M., Ahmed, M., Ferrucci, V., Salpietro, V., Asadzadeh, F., Carotenuto, M., Maroofian, R., Al-Amri, A., Singh, R., Scognamiglio, I., Mojarrad, M., Musella, L., Duilio, A., Di Somma, A., Karaca, E., Rajab, A., Al-Khayat, A., Mohan Mohapatra, T., Eslahi, A., Ashrafzadeh, F., Rawlins, L.E., Prasad, R., Gupta, R., Kumari, P., Srivastava, M., Cozzolino, F., Kumar Rai, S., Monti, M., Harlalka, G.V., Simpson, M.A., Rich, P., Al-Salmi, F., Patton, M.A., Chioza, B.A., Efthymiou, S., Granata, F., Di Rosa, G., Wiethoff, S., Borgione, E., Scuderi, C., Mankad, K., Hanna, M.G., Pucci, P., Houlden, H., Lupski, J.R., Crosby, A.H. and Baple, E.L., 2017. PRUNE is crucial for normal brain development and mutated in microcephaly with neurodevelopmental impairment. *Brain* 140, 940-952.

Abbreviations list

OCA, oculocutaneous albinism

ACCEPTED MANUSCRIPT

SCIENTIFIC REPORTS

OPEN

Comprehensive sequencing of the myocilin gene in a selected cohort of severe primary open-angle glaucoma patients

Luke O’Gorman¹, Angela J. Cree², Daniel Ward³, Helen L. Griffiths², Roshan Sood⁴, Alastair K. Denniston⁵, Jay E. Self^{2,6}, Sarah Ennis⁷, Andrew J. Lotery^{2,6} & Jane Gibson⁴

Primary open-angle glaucoma (POAG) is the most common form of glaucoma, prevalent in approximately 1–2% of Caucasians in the UK over the age of 40. It is characterised by an open anterior chamber angle, raised intraocular pressure (IOP) and optic nerve damage leading to loss of sight. The myocilin gene (*MYOC*) is the most common glaucoma-causing gene, accounting for ~2% of British POAG cases. 358 patients were selected for next generation sequencing (NGS) with the following selection criteria: Caucasian ethnicity, intraocular pressure (IOP) 21–40 mm Hg, cup:disc ratio ≥ 0.6 and visual field mean deviation ≤ -3 . The entire *MYOC* gene (17,321 bp) was captured including the promoter, introns, UTRs and coding exons. We identify 12 exonic variants (one stop-gain, five missense and six synonymous variants), two promoter variants, 133 intronic variants, two 3’ UTR variants and 23 intergenic variants. Four known or predicted pathogenic exonic variants (p.R126W, p.K216K, p.Q368* and p.T419A) were identified across 11 patients, which accounts for 3.07% of this POAG cohort. This is the first time that the entire region of *MYOC* has been sequenced and variants reported for a cohort of POAG patients.

Primary open-angle glaucoma

Glaucoma accounts for 7.9% of blindness in the UK¹. It is characterised by a progressive loss of retinal ganglion cells, atrophy of the optic nerve and degradation of the visual field². Glaucoma presents in multiple forms with primary open-angle (POAG) being the most common form^{3,4}.

POAG is characterised by an open anterior chamber angle and raised intraocular pressure (IOP) leading to damage of the optic nerve and visual field loss³. POAG affects at least 1% of Caucasians in the UK over the age of 40 years⁵. Normal tension glaucoma (NTG) is a form of POAG in which optic nerve damage and visual field degradation are characteristic traits, however, IOP is not elevated⁴.

Approximately 5% of POAG is accounted for by monogenic, Mendelian-like variants. The myocilin gene (*MYOC*) accounts for the majority, approximately 2.2% of cases⁶. The optineurin gene (*OPTN*) may contribute to POAG in some populations, but has been implicated in normal tension glaucoma where it accounts for 1.5% of cases⁷. The majority of POAG cases are assumed to be accounted for by combined effects of multiple genetic and non-genetic risk factors. IOP is considered the most important risk factor in POAG. Other important risk factors in POAG include age, race, refractive error, central corneal thickness and family history of POAG^{5,8,9}. However, these risk factors alone do not cause glaucoma⁷.

¹Human Development and Health, Faculty of Medicine, University of Southampton, Southampton, UK. ²Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Southampton, UK. ³Molecular Genetics Wessex Regional Genetics Laboratory, Salisbury NHS Foundation Trust, Salisbury, UK. ⁴Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, UK. ⁵Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. ⁶Eye Unit, University Hospital Southampton, Southampton, UK. ⁷Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, Southampton, UK. Andrew J. Lotery and Jane Gibson jointly supervised this work and will share last authorship. Correspondence and requests for materials should be addressed to S.E. (email: se@soton.ac.uk)

Myocilin

MYOC is understood to have a role in cytoskeletal development and regulation of intraocular pressure (IOP)¹⁰. It is also known as the trabecular meshwork glucocorticoid-inducible response protein (TIGR)^{10,11}. Variants in *MYOC* have also previously been identified as the cause of hereditary juvenile-onset open-angle glaucoma (JOAG), which is an early onset (<40 years) sub-set of POAG^{12–15}.

MYOC is expressed in multiple tissues within the eye including the trabecular meshwork and ciliary body suggesting that it causes an increase in IOP through obstruction of the aqueous outflow^{14,16}. It is also expressed at similar levels in a range of organs and tissue including the heart, skeletal muscle and bone marrow amongst others¹⁷.

The *MYOC* gene is encoded on the negative strand and comprises three exons. *MYOC* has one known RefSeq transcript (NM_000261.1) and spans 17,321 bp^{18,19}. *MYOC* encodes a 504 amino acid polypeptide which consists of an N-terminal helix-turn-helix domain and two coil-coils²⁰ and can homodimerise through leucine zipper interactions²¹. The C-terminal olfactomedin-like domain²² is part of a family of mucus proteins which are mainly found in nasal mucus²³.

Known variants in *MYOC* are curated and made available online via the ‘myocilin allele-specific glaucoma phenotype database’²⁴. Within this database, exons 1, 2 and 3 have 32, 1 and 62 known glaucoma-causing variants respectively. There are no glaucoma-causing variants annotated in the promoter, intronic or intergenic regions currently (27/11/2017)²⁴. The database reports disease-causing variants comprising missense (83.7%), nonsense (5.8%), <21 bp deletion (4.8%), <21 bp insertion (4.8%) and <21 bp indels (1%). The Exome Aggregation Consortium (ExAC)²⁵ scores the probability of loss-of-function intolerance (pLI) as 0.00 indicating that *MYOC* is tolerant of loss-of-function, and *MYOC* homozygous knockout mice experiments have excluded haploinsufficiency as a disease mechanism underlying POAG²⁶. Shepard *et al.* suggested a gain-of-function is the likely cause of POAG and concluded p.Y437H variants in human *MYOC* induce exposure of an N-terminal cryptic peroxisomal targeting signal sequence¹⁶. The majority of pathogenic *MYOC* variants are found in exon 3^{6,27} where the most prevalent pathogenic variants are found to have a penetrance of up to 90%¹⁷. The concentration of known variants in exon 3 may have been exacerbated in recent years due to preferential analysis of this exon. Since the *MYOC* gene involves an autosomal dominant mode of inheritance in POAG¹⁴, pathogenic heterozygous variants would be a sufficient genotype for causality.

Aim

In this study, the entire region of *MYOC* is assessed in 358 individuals with POAG selected from a UK cohort. Through the use of next-generation sequencing (NGS), application of bioinformatic tools and strategic filtering of variants, variants across the intergenic, promoter, UTR, exonic coding sequences and intronic regions are reported for the first time.

Methods

Patients with Primary Open Angle Glaucoma (POAG) were recruited from eye clinics at University Hospital Southampton, Addenbrooks Hospital Cambridge, Frimley Park Hospital Surrey, Queen Elizabeth Hospital Birmingham, Queen Alexandra Hospital Portsmouth, Romsey Hospital, St Mary’s Hospital Isle of Wight, Torbay Hospital Devon and New Cross Hospital Wolverhampton. Patient data was collected including gender, ethnicity, family history of POAG, specific diagnosis of the patient, age at diagnosis, intraocular pressure (IOP), cup:disc ratio (CDR), central corneal thickness and visual field mean deviation (VFMD). Blood samples were collected and DNA was extracted using the salting out method²⁸ and stored at -20°C . Initially, 372 patients were selected for Next Generation Sequencing (NGS) using the following selection criteria: Caucasian ethnicity, $21\text{ mm Hg} \leq \text{IOP} \leq 40\text{ mm Hg}$, cup:disc ratio ≥ 0.6 and visual field mean deviation ≤ -3 . Ten patients who passed all inclusion criteria also had one affected first degree relative recruited to the study, even if they did not meet all inclusion criteria.

The entire *MYOC* gene was targeted for inclusion using a custom sequencing panel to include the intronic, exonic, UTR and promoter regions (Table 1) and an additional 1000 bp upstream of the Eukaryotic Promoter Database defined promoter coordinates (hg38 coordinates: chr1:171652678–171652737)^{29,30}.

Library preparation and sequencing were performed in local laboratories, where DNA was simultaneously fragmented and tagged with sequencing adapters using the Illumina’s Nextera Rapid Capture Custom Enrichment kit (Illumina 5200 Illumina Way San Diego, California USA). Target regions of DNA were bound and amplified with custom capture probes and enriched prior to running on a Illumina NextSeq500 sequencing machine. Sequencing was performed in three batches of 96 samples and one batch of 84 samples.

Next generation sequencing (NGS) data were aligned against the human reference genome (hg38) using BWA-mem³¹. Variant calling was performed using GATK v3.7³². Annotation was performed with ANNOVAR³³ against a database of RefSeq transcripts¹⁹, Exome Aggregation Consortium (ExAC)²⁵, 1000 Genomes Project^{34,35} and conservation-based pathogenicity scores of sort intolerant from tolerant (SIFT)³⁶, PhyloP, PhastCons³⁷ and Genomic Evolutionary Rate Profiling (GERP++)³⁸. Variants were also annotated with MutPred Splice³⁹ and Human Splicing Finder v3.0 (HSF3.0)^{40,41} to evaluate disruption of splicing. Further annotation was performed using non-coding Functional Analysis through Hidden Markov Models (FATHMM)^{42,43} and Combined Annotation Dependent Depletion (CADD)⁴⁴, and incorporation of the ‘Myocilin allele-specific glaucoma phenotype database’²⁴.

For the 372 sequenced patient samples, coverage across both the *MYOC* gene and all targets were determined. The proportion of variants shared between samples was checked for consistency with known sample relationships and ethnicities. VerifyBamID v1.1.13⁴⁵ software was used to estimate possible contamination and a ‘freemix’

	Chromosome	Start	End	Length (bp)
Promoter	1	171652678	171652737	59
Prom-5' UTR	1	171652633	171652678	45
5' UTR	1	171652611	171652633	22
Exon 1	1	171652007	171652611	604
Intron 1	1	171638722	171652007	13285
Exon 2	1	171638596	171638722	126
Intron 2	1	171636709	171638596	1887
Exon 3	1	171635924	171636709	785
3' UTR	1	171635416	171635924	508

Table 1. *MYOC* promoter, intronic, exonic and intergenic region locations within hg38 human reference genome.

value threshold of >0.03 was applied⁴⁶. Coverage statistics were generated using SAMtools v1.3.1⁴⁷ and BEDtools v2.17.0⁴⁸. A minimum threshold of 20X depth was used to distinguish samples with sufficient coverage.

For the entire region of the *MYOC* gene, depth per base was calculated with SAMtools v1.3.1⁴⁷ and conservation scores were downloaded from the University of California Santa Cruz (UCSC) database for PhyloP³⁷ and PhastCons³⁷ databases of 20 mammals. Regions were considered in 'high conservation' if PhastCons ≥ 0.4 ⁴⁹ or PhyloP ≥ 1.5 ⁵⁰. Repetitive regions were annotated using UCSC RepeatMasker data, and rare variants (allele frequency ≤ 0.05) in the POAG cohort and 1000 Genomes Project were plotted.

Variants were considered as previously known glaucoma-causing variants if they were identified as 'Glaucoma-causing' in the 'Myocilin allele-specific glaucoma phenotype database' or ClinVar. Exonic variants were prioritised as candidate causal variants through CADD Phred scores ≥ 15 ⁵¹. Exonic splice variants were prioritised if they exceeded a 0.6 MutPred Splice score threshold³⁹.

Non-coding variants were prioritised using FATHMM which out-performs CADD in the non-coding region whilst CADD has a superior classifier over FATHMM in the coding region. Variants were prioritised if the FATHMM score exceeded the default threshold of ≥ 0.5 ⁴³. Intronic variants were also prioritised if they were flagged as potentially splice affecting in HSF3.0^{40,41}.

CNVkit, which is designed for use with custom target panels and short-read Illumina sequencing, was used to infer copy number⁵². Data were analysed in batches to account for variation in average depth between batches, and a pooled reference of all samples within the batch was used.

Consent was obtained in accordance with the Declaration of Helsinki and was approved by South West Hampshire Local Research Ethics Committee (05/Q1702/8). Informed consent was obtained from all subjects, and all methods were carried out in accordance with the relevant guidelines and regulations of Research Ethics Committees (REC).

Results

Samples analysed. 358 of 372 patients passed inclusion criteria after sample quality control. One sample was omitted due to insufficient depth ($\leq 20X$), five were omitted due to post hoc detection of sample duplication or mixed race individuals. A further eight samples were omitted due to the patient age at diagnosis being less than 40 years. Selected demographic and clinical characteristics for all 358 individuals in the final analysis are summarised in Table 2.

Genomic features and variation across the *MYOC* gene. The coordinates of the *MYOC* gene (5' promoter - 3' UTR) are chr1:171,652,737–171,635,416 (hg38). Coverage was uninterrupted across the entire region with 100% coverage at 20X depth for all samples. The four batches had mean depths of 717X, 552X, 389X, and 726X across target regions respectively, averaged across all samples. The poorest coverage was observed in batch 3. A consistent coverage pattern is seen for all four batches (Fig. 1B) and was found to be correlated with mappability, repetitive context, conservation and GC content ($R^2 = 0.3358$, p -value $< 2.2 \times 10^{-16}$). Conservation scores derived from PhastCons (Fig. 1C) and PhyloP (Fig. 1D) were highest across exonic regions, with smaller regions of high conservation in intron 1 and a region upstream of the Eukaryotic Promoter Database defined promoter region.

We identified a total of 172 annotated variants comprising 160 SNPs and 12 indels in the POAG cohort of 358 individuals (Table 3). These variants were distributed across *MYOC* with 21 variants upstream intergenic, two variants in the promoter region, four in exon 1, 118 in intron 1, one in exon 2, 15 in intron 2, seven in exon 3, two in the 3' UTR and two variants in the downstream intergenic region. 156 SNPs were identified in the non-coding regions of the *MYOC* gene in the POAG cohort. The majority (70.5%) of non-coding variants were located in the largest intron, intron 1, which spans 13,285 bp (76.7%) of the 17,321 bp length of *MYOC*. For comparison, there were 574 total SNPs in the 1000 Genomes Project European population (1000gEUR) in *MYOC*. There were 105 rare (AF $< 5\%$) variants in POAG and 134 rare (AF $< 5\%$) variants in 1000gEUR, and these had a similar distribution across the *MYOC* gene (Fig. 1F and G).

Three SNPs were identified with high conservation in PhastCons (20 mammals), using a threshold of 0.4, and excluding variants in repetitive regions. The variant rs76745622 was located in the upstream intergenic region whilst rs11586716 and rs12035960 were located in intron 1. No variants were identified in the non-coding regions with high conservation using PhyloP (≥ 1.5).

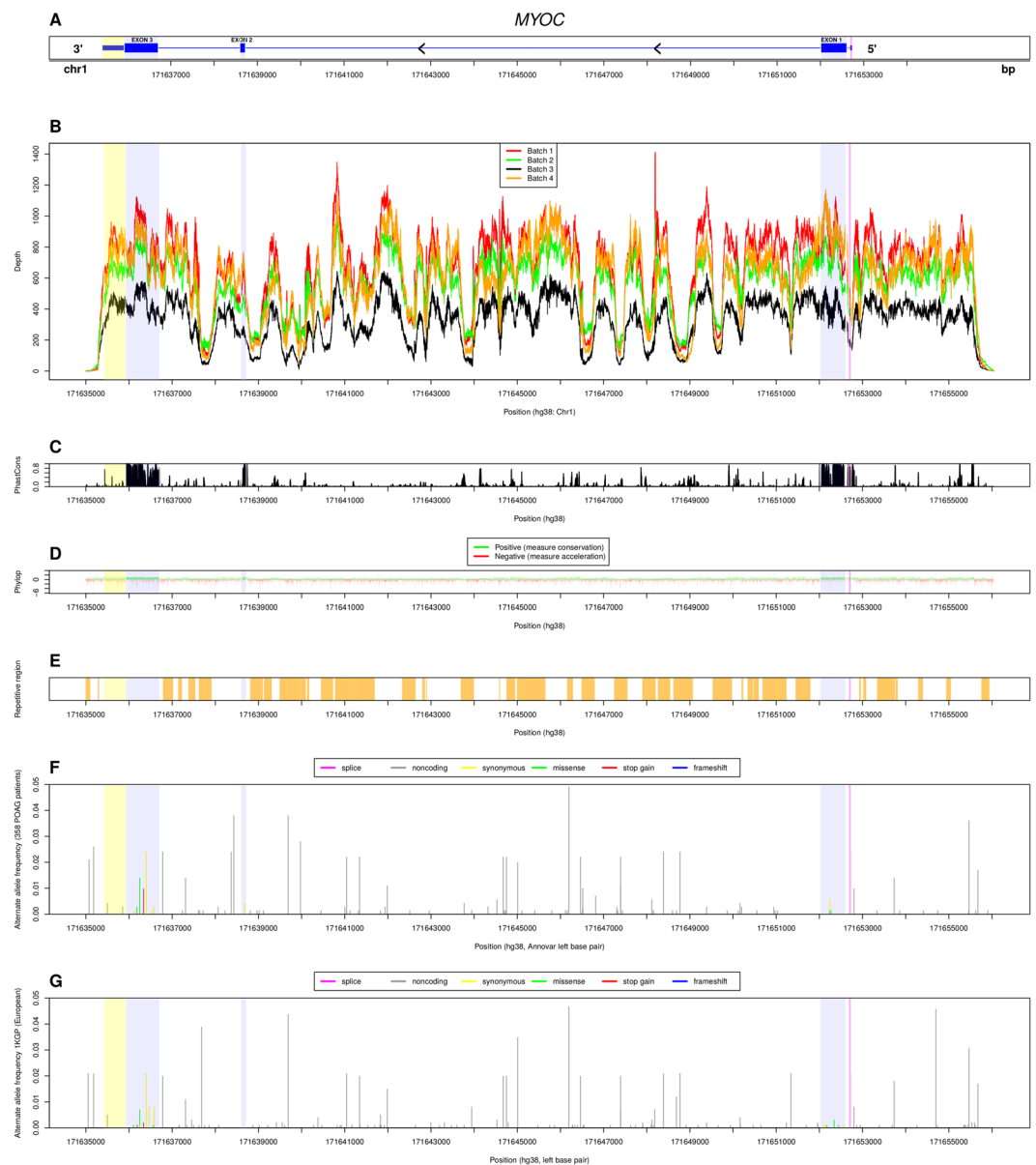


Figure 1. Per base analysis of variants and regions across the *MYOC* gene. Per base analysis of read depth, conservation scores, repetitive region context and allele frequency across the POAG patient data set ($n = 358$) for *MYOC*. (A) Gene structure of *MYOC*. (B) Per base depth for samples of batch 1, red; batch 2, green; batch 3, black; batch 4 orange;. (C) PhastCons 20 way mammals conservation score (20 mammals). (D) PhyloP scores (20 mammals), showing measure of conservation (green) and acceleration (red). (E) Regions identified as repetitive regions by RepeatMasker (orange). (F) Allele frequency for each base's detected alternate allele across 358 POAG samples (left base position). (G) Allele frequency of SNPs in 1000 Genomes Project (European) for corresponding alternate alleles identified in F.

	Min	Max	Mean	Median	SD
Age (years)	42	91	66	66	11
IOP (mmHg)	21.00	42.00	27.90	27.00	4.67
CDR	0.60	1.00	0.82	0.80	0.09
VFMD	-31.54	-1.07	-14.52	-14.51	7.68

Table 2. Demographic and clinical characteristic summaries of age, intraocular pressure (IOP), cup:disc ratio (CDR) and visual field mean deviation (VFMD) for POAG cohort ($n = 358$).

Gene feature	No. variants		
	No. unique variants	SNPs	Indels
Intergenic us	21	21	0
Promoter	2	2	0
Promoter-5' UTR	0	0	0
5' UTR	0	0	0
Exon 1	4	4	0
Intron 1	118	110	8
Exon 2	1	1	0
Intron 2	15	12	3
Exon 3	7	7	0
3' UTR	2	2	0
Intergenic ds	2	1	1
All	172	160	12

Table 3. Summary of the number of variants identified across all features of the *MYOC* gene.

Exonic variants. A total of 12 exonic variants were called (Table 4). Four SNPs were detected in exon 1, one SNP in exon 2 and seven SNPs in exon 3. Four variants had CADD Phred scores greater than 15 suggesting the variants were likely pathogenic. Variants NM_000261:exon3:c.C1102T (p.Q368*) and NM_000261:exon1:c.C376T (p.R126W) had previously been identified as 'Glaucoma-causing' by the 'myocilin allele-specific glaucoma phenotype database'. Variant p.R126W also had a high MutPred splice score of 0.605, indicating that this variant was likely to affect splicing.

Two variants NM_000261:exon2:c.G648A (p.K216K) and NM_000261:exon3:c.A1255G (p.T419A) not previously identified as glaucoma-causing, had pathogenicity scores indicating they may be of importance. In three individuals p.K216K had a high PhastCons score of 0.992 indicating that it is within a highly conserved element. The variant was also more common in the POAG cohort (AF = 0.0042) than in the ExAC Non-Finnish European (NFE) population (AF = 0.0005).

The missense variant p.T419A had no known rsID and was not found in ExAC NFE. In the POAG cohort it has an allele frequency of 0.0028, and identified as heterozygous in two individuals. It is located in exon 3 and has a SIFT score of 0, PolyPhen HDIV score of 0, GERP++ score of 4.52 and CADD Phred of 37 which indicate that this variant is likely to be highly pathogenic. However, this variant was found to be present on the same read pair as the p.Q368* variant in both patients (see Supplementary Fig. S1).

There was no significant difference in sub-phenotypes between patients with candidate causal *MYOC* variants (p.Q368*, p.R126W, p.K216K or p.T419A) and patients with no candidate causal *MYOC* variants (t-test, IOP p-value = 0.766, CDR p-value = 0.626, VFMD p-value = 0.211). However, hypertension was treated in five of the 11 patients with candidate causal *MYOC* variants.

Three variants, p.E115K, p.G122G and p.R126W are clustered within the coiled-coil located at aa118-aa186²⁰ (Fig. 2). The aa117-aa166 region contains lysine residues responsible for dimerisation of *MYOC*⁵³. The p.K216K variant is located in a linker region whilst p.T285T, p.D302D, p.Y347Y, p.Q368*, p.K398R, p.T419A and p.T438T are all located within the large olfactomendin-like domain.

Non-coding variants. There were 160 variants identified in non-coding regions (see Table 3 and Supplementary Fig. S2). The majority of variants (118) were identified within the largest region, intron 1. Using a FATHMM threshold of 0.5 to prioritise the non-coding variants, one variant upstream of the promoter, three intron 1 and one intron 2 variants remain (Table 5). The highest FATHMM score of 0.86 was seen for a common variant with an allele frequency of 10% in the 1000gEUR, and similar (8.2%) in the POAG cohort. A single variant in intron 1, NM_000261.1:c.605-5949C>T, had a CADD Phred score greater than 15, however, there was no significant difference in frequency between this variant in the POAG cohort compared with the 1000gEUR cohort (allelic chi-squared test, p-value = 0.716). A second intron 1 variant, NM_000261.1:c.604+5942G>A, had a CADD Phred score of 12.78, a GERP++ score of 3.57 and PhastCons score of 0.504. Although pathogenicity scores are in favour of a potentially pathogenic effect, allele frequencies show no significant difference between the POAG cohort and 1000gEUR (allelic chi-squared test, p-value = 0.132). The novel intergenic upstream variant identified, NM_000261.1:c.-2851C>T, had a CADD Phred score of 11.19, a GERP++ score of 3.22 and low conservation PhastCons score of 0.0079 providing ambiguous indications of pathogenicity. This variant is not present in the 1000gEUR but is observed as heterozygous in one individual within the POAG cohort.

20 variants were flagged as 'potentially splice altering' by human splice finder (HSF) version 3.0, 18 of which had rsIDs (see Supplementary Table S2). 17 splice variants were in intron 1 (15 SNPs, and 2 insertions) and three in intron 2 (three SNPs). Of these, three variants had potential to introduce both a new splice acceptor and/or a splice donor site (AD), nine introduced a new splice acceptor site only (A), seven a new donor site only (D) and one breaks a branch point (BBP).

No copy number variants (CNVs) were detected within the *MYOC* gene region. There were no losses in copy number (CN) across the entire region, however, some (N = 113) samples were called as a single copy gain with a CN = 3 across the entire region.

No.	Exon	Chrom	POS	Ref	Alt	Variant type	Amino acid	dbSNP144	1000 G EUR	ExAC NFE	Sample count	Study AF	myocDB	CLINSIG	SIFT	gerp++ gt2	Phast Cons	CADD Phred	MutPred Splice
1	1	chr1	171,652,385	C	T	nonsynonymous	R76K	rs2234926	0.14120	0.13650	92	0.13700	Neutral	Benign	0.049	3.43	0.945	9.00	0.116
2	1	chr1	171,652,269	C	T	nonsynonymous	E115K	rs757551979	—	0.00003	1	0.00140	—	—	0.589	3.53	0.268	9.37	0.159
3	1	chr1	171,652,246	G	A	synonymous	G122G	rs145354114	—	0.00300	4	0.00559	Neutral	Uncertain	—	—	0.000	0.17	0.494
4	1	chr1	171,652,236	G	A	nonsynonymous	R126W	rs200120115	—	0.00007	1	0.00140	Glaucoma	—	0.019	-1.13	0.008	23.50	0.605
5	2	chr1	171,638,679	C	T	synonymous	K216K	rs141584495	—	0.00050	3	0.00419	Neutral	—	—	—	0.992	15.63	0.165
6	3	chr1	171,636,585	C	A	synonymous	T285T	rs146606638	0.00800	0.00480	2	0.00279	Neutral	Benign	—	—	0.591	14.00	0.126
7	3	chr1	171,636,534	G	A	synonymous	D302D	rs148433908	—	0.00030	1	0.00140	Neutral	—	—	—	0.000	0.07	0.137
8	3	chr1	171,636,399	A	G	synonymous	Y347Y	rs61730974	0.02090	0.03050	17	0.02400	Neutral	—	—	—	0.024	0.00	0.140
9	3	chr1	171,636,338	G	A	stopgain	Q368*	rs74315329	0.00200	0.00150	7	0.00978	Glaucoma	Pathogenic	—	4.52	0.283	37.00	0.374
10	3	chr1	171,636,247	T	C	nonsynonymous	K398R	rs56314834	0.00700	0.00480	10	0.01400	Neutral	—	0.618	-1.17	0.843	3.87	0.268
11	3	chr1	171,636,185	T	C	nonsynonymous	T419A	—	—	—	2	0.00279	—	—	0.000	5.04	0.945	23.50	0.196
12	3	chr1	171,636,126	G	A	synonymous	T438T	rs375235405	—	0.00004	1	0.00140	—	—	—	—	0.898	11.66	0.125

Table 4. Annotation of all exonic variants in *MYOC*. Exon, exon number; Feature, genetic feature within *MYOC*; Chrom, chromosome; POS, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternate allele; Variant type, type of variant observed; Amino Acid, amino acid single letter abbreviation of reference amino acid and the amino acid substituted to; dbSNP144, rs ID if the variant is known; 1000 G EUR, allele frequency from 1000 Genomes Project (European ethnic sub-group); ExAC NFE, allele frequency from ExAC Non-Finnish European ethnic sub-group; Sample count, number of patients with the variant in the n = 358 POAG cohort; Study AF, allele frequency of the variant within the n = 358 POAG cohort; myocDB, known *MYOC* variants database²⁴; CLINSIG, pathogenicity of the variant in ClinVar; SIFT, sorts intolerant from tolerant substitutions; gerp++, Genomic Evolutionary Rate Profiling; PhastCons, conservation scoring and identification of conserved elements; CADD Phred, Combined Annotation Dependent Depletion on a Phred scale; MutPred Splice, machine learning-based predictor of exonic splice variants. Bold indicates variants which are causal candidates.

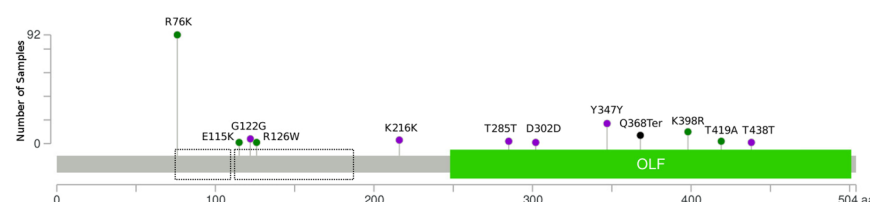


Figure 2. Human *MYOC* protein structure with its domains and variants mapped^{67,68}. Additional annotation of coiled-coil domains are outlined in black (aa74–110 and aa118–186²⁰). Green dots indicate a missense variant, purple dots indicate synonymous variants, whilst black dots indicate stop-gains.

Discussion

We have performed targeted next-generation sequencing on the full region of the *MYOC* gene (promoter, UTRs, coding exons, introns and intergenic regions) on 358 POAG patients with severe POAG sub-phenotypes. We report all variants detected across the region and have performed an in silico analysis to assess pathogenicity. We identified a known pathogenic stop-gain in exon 3, a known pathogenic missense variant in exon 1, a known synonymous variant not previously considered pathogenic in exon 2, and an unknown missense variant in exon 3. Between them these variants account for 11/358 (3.07%) of patients within our POAG cohort.

NM_000261:exon3:c.C1102T (p.Q368*) is the most common causal variant in POAG, accounting for 31.2% of disease-causing variants in *MYOC*²⁴. This stop-gain is 6.5 times more common in our POAG cohort than in the 1000gEUR (allelic Fisher's Exact test, p-value = 0.0336) and is seen in seven patients, accounting for 63.6% of the candidate causal variants identified. Of the seven patients with p.Q368* variants, three did not have a positive family history of POAG in our database. Craig *et al.* have previously shown a 100% family history for this variant, however this was only after retrospective follow up and new diagnoses were made. They found that the index patient was sometimes unaware of their family history of disease and thus family history reported may be underestimated, which may also be that case here. Furthermore, the penetrance of this variant is high but not complete. 82% of carriers of p.Q368* have POAG or OHT at the age of 65 years, and an unaffected carrier has been identified at 74 years of age⁵⁴. Shepard *et al.* have previously shown that protein *MYOC* monomers with this variant do not contain a cryptic peroxisomal signalling sequence (PTS1) and that it likely exposes the PTS1 sequence in its dimer partner¹⁶. This is believed to cause the mutant dimer to associate with the PTS1R and ultimately cause deleterious trabecular meshwork cell function¹⁶. This mechanism has been supported by other studies^{55,56}. Previous studies have shown that Caucasian patients with the p.Q368* variant have a mean IOP ranging between 27.3–35.4 mm Hg^{6,57–59}, higher than in our study (mean IOP = 26.3 mm Hg). In our study, patients with

No.	Feature	Chrom	POS	Ref	Alt	dbSNP144	1000 G EUR	Sample count	Study AF	gerp++ gt2	FATHMM	CADD Phred	myocDB	Repeat	Phast Cons	Regulatory build
1	INTERGENIC_US	chr1	171655462	C	T	—	—	1	0.00140	3.22	0.593	11.19	—	0	0.008	CTCF Binding Site
2	INTRON1	chr1	171,646,066	C	T	rs12035960	0.10440	55	0.08200	3.57	0.860	12.78	—	0	0.504	Promoter Flanking Region
3	INTRON1	chr1	171,644,671	G	A	rs75953590	0.01990	16	0.02200	—	0.551	15.38	—	0	0.244	—
4	INTRON1	chr1	171,643,942	G	C	rs144750384	0.00800	1	0.00150	—	0.623	10.31	—	1	0.061	Open chromatin
5	INTRON2	chr1	171,637,310	A	G	rs79263003	0.01090	10	0.01400	—	0.570	4.669	—	0	0.280	—

Table 5. Five non-coding variants in the *MYOC* region remain following initial filtering of the 160 non-coding variants with FATHMM ≥ 0.5 . Feature, genetic feature within *MYOC*; Chrom, chromosome; POS, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternate allele; dbSNP144, rsID if the variant is known; 1000 G EUR, allele frequency from 1000 Genomes Project (European ethnic sub-group); Sample count, number of patients with the variant in the $n = 358$ POAG cohort; Study AF, allele frequency of the variant within the $n = 358$ POAG cohort; gerp++, Genomic Evolutionary Rate Profiling; FATHMM, Functional Analysis through Hidden Markov Models; CADD Phred, Combined Annotation Dependent Depletion on a Phred scale; myocDB, known *MYOC* variants database²⁴; Repeat, repetitive region as defined by RepeatMasker; PhastCons, conservation scoring and identification of conserved elements; Regulatory build, Ensembl Regulatory Build containing regions that are likely to be involved in gene regulation.

this variant had a mean CDR of 0.84, a mean VFMD of -13.84 , and a mean age of 69.6 years. These findings agree with Graul *et al.* who found that p.Q368* patients did not have an earlier onset nor did they have a higher IOP⁵⁷.

The known glaucoma-causing variant, NM_000261:exon1:c.C376T (p.R126W) was found in one of the 358 POAG patients and had been previously reported as a late-onset familial variant⁶⁰. It is a variant which is located on the protein dimer region of a coiled-coil. Gobeil *et al.* have shown cell adhesion properties were unaffected⁵⁵ by this variant. NM_000261:exon1:c.C376T (p.R126W) had damaging SIFT and CADD Phred scores of 0.019 and 23.5 respectively. There was very little evidence for splicing variation leading to POAG and only one previously known instance in *MYOC* of a predicted cryptic splice site reported within intron 1⁶¹. However, a MutPred Splice score of 0.605 implicated that this variant contributes to the creation of a new donor splice site and a subsequent loss of 372 nucleotides from exon 1. This finding suggests that splice variants could be more important in POAG than previously known. Faucher *et al.* has previously shown that patients with this variant were found to have a mean IOP of 28.3 mm Hg and an age of onset of 74⁶⁰. The patient with this variant in our cohort showed similar traits with a maximum IOP of 27 mm Hg and an age at diagnosis of 74 (see Supplementary Table S3).

The synonymous variant NM_000261:exon2:c.G648A (p.K216K) was not previously considered a pathogenic variant. This variant is found in exon 2 which contains just one known pathogenic variant and is believed to translate to a linker region within the *MYOC* protein^{24,62}. Synonymous variants in *MYOC* have been suggested to have a role affecting *MYOC* mRNA structure and subsequently the translated protein stability⁶³. Variant p.L215Q, on the preceding codon of p.K216, is believed to be glaucoma-causing on the basis of an in-silico damaging SIFT score⁶². Similarly, p.K216K has strong in silico pathogenicity scores to suggest possible pathogenic status (PhastCons of 0.992 and CADD Phred of 15.63). Furthermore, this variant is found in the gnomAD Non-Finnish European (NFE) population²⁵ significantly less frequently than the POAG cohort (allelic Fisher's Exact test, p -value = 0.0109). This heterozygous variant was found in three patients from the University Hospital Southampton site. No evidence of relatedness was identified, however, there is a possibility that there is some distant relatedness which we do not have the capability to detect.

The missense variant NM_000261:exon3:c.A1255G (p.T419A) does not have an associated rsID, nor is it found within 1000gEUR or ExAC. However, it is found at an allele frequency of 8.952×10^{-6} in the gnomAD NFE population. This variant has never been observed in a glaucoma context before but is seen as a heterozygote in two patients in this study (AF = 0.0028). This is a substantially higher frequency than gnomAD NFE (allelic Fisher's Exact test, p -value = 9.4×10^{-6}). This variant had a SIFT score of 0, GERP++ score of 5.04, PhastCons of 0.945 and CADD Phred score of 23.5 which indicated further support for pathogenicity. However, we have found that in both patients p.T419A is co-inherited with the upstream p.Q368* variant (see Supplementary Fig. S1), therefore the protein will be truncated before translation of the potentially pathogenic substitution. Although these two patients had the earliest onset (50 & 56 years) of those carrying the p.Q368* variant, it is not possible to provide a plausible mechanism by which this variant could have a modifying effect.

Whilst there were no clear likely pathogenic variants in the non-coding region of the gene, NM_000261.1:c.-2851C>T which is located in the upstream intergenic region (44694 bp from the neighbouring *VAMP4* gene) was found to be of potential interest. Whilst it is not within a conserved element in PhastCons, it had damaging GERP++ and FATHMM scores. It has an allele frequency in the POAG cohort of 0.0014 (one heterozygous patient). This variant is not found within the 1000gEUR and is at a position not currently covered by gnomAD. The Ensembl regulatory build indicates that this variant could be functionally important as it is located at a potential CTCF binding site. All other variants with a FATHMM score ≥ 0.5 were seen at similar frequencies in both the POAG cohort and 1000gEUR. Genotyping this non-coding variant across a wider POAG cohort could prove informative. Variants located up to 1000 bp upstream of *MYOC* have been implicated as potentially functionally important for controlling IOP^{64,65}.

Five rare variants were present in six individuals that potentially affect splicing. However, the presence of these variants in the European population cannot be confirmed due to lack of coverage in allele frequency databases for these sites.

We found no evidence of sub-gene copy number changes, and no whole gene deletions. We did predict some whole gene single copy gains and we suspect the predicted gain reflects within-batch depth variation. Patient selection criteria for this study used strict sub-phenotype parameters in order to select most severe POAG sub-phenotypes with a greater chance of an accurate POAG diagnosis. However, such criteria hinders genotype-phenotype analyses within the selected cohort. Genotyping of a larger POAG cohort not selected on sub-phenotypes is necessary in order to perform robust genotype-phenotype analyses. The *MYOC* gene accounts for ~3% of patients with POAG, therefore a larger cohort would also have greater power to detect rarer causal variants.

Conclusion

For the first time all regions of *MYOC* have been sequenced and analysed in a POAG cohort. We have identified two known pathogenic variants and two high pathogenic scoring variants, which may cause POAG in 11 patients. Synonymous and non-coding variants have been identified as having pathogenic qualities using in silico pathogenicity predictions, and a known glaucoma-causing variant has been implicated as a potential deep exonic splice variant. This work expands the known allelic diversity of *MYOC* in POAG which is useful for diagnosis, genetic counselling and cascade genetic testing in families. Additional sequencing of *MYOC* interacting partners⁶⁶ and other POAG-causing genes could reveal rare causal variants and provide further insight into the genetic basis of POAG.

Data Availability

Data generated or analysed during this study are included in this published article and its supplementary files.

References

- Evans, J. R., Fletcher, A. E. & Wormald, R. P. L. Causes of visual impairment in people aged 75 years and older in Britain: an add-on study to the MRC Trial of Assessment and Management of Older People in the Community. *The British journal of ophthalmology* **88**, 365–370 (2004).
- Allingham, R. R., Liu, Y. & Rhee, D. J. The genetics of primary open-angle glaucoma: a review. *Experimental eye research* **88**, 837–844, <https://doi.org/10.1016/j.exer.2008.11.003> (2009).
- Gupta, N. & Weinreb, R. N. New definitions of glaucoma. *Current opinion in ophthalmology* **8**, 38–41 (1997).
- Foster, P. J., Buhrmann, R., Quigley, H. A. & Johnson, G. J. The definition and classification of glaucoma in prevalence surveys. *The British journal of ophthalmology* **86**, 238–242 (2002).
- Tuck, M. W. & Crick, R. P. The age distribution of primary open angle glaucoma. *Ophthalmic epidemiology* **5**, 173–183 (1998).
- Ennis, S. *et al.* Prevalence of myocilin gene mutations in a novel UK cohort of POAG patients. *Eye (London, England)* **24**, 328–333, <https://doi.org/10.1038/eye.2009.73> (2010).
- Fingert, J. H. Primary open-angle glaucoma genes. *Eye (London, England)* **25**, 587–595, <https://doi.org/10.1038/eye.2011.97> (2011).
- Tielsch, J. M., Katz, J., Sommer, A., Quigley, H. A. & Javitt, J. C. Family history and risk of primary open angle glaucoma. The Baltimore Eye Survey. *Archives of ophthalmology (Chicago, Ill.: 1960)* **112**, 69–73 (1994).
- Mitchell, P., Hourihan, F., Sandbach, J. & Wang, J. J. The relationship between glaucoma and myopia: the Blue Mountains Eye Study. *Ophthalmology* **106**, 2010–2015 (1999).
- Kubota, R. *et al.* A novel myosin-like protein (myocilin) expressed in the connecting cilium of the photoreceptor: molecular cloning, tissue expression, and chromosomal mapping. *Genomics* **41**, 360–369, <https://doi.org/10.1006/geno.1997.4682> (1997).
- Polansky, J. R. *et al.* Cellular pharmacology and molecular biology of the trabecular meshwork inducible glucocorticoid response gene product. *Ophthalmologica. Journal internationale d'ophtalmologie. International journal of ophthalmology. Zeitschrift fur Augenheilkunde* **211**, 126–139 (1997).
- Alward, W. L. *et al.* Clinical features associated with mutations in the chromosome 1 open-angle glaucoma gene (GLC1A). *The New England journal of medicine* **338**, 1022–1027, <https://doi.org/10.1056/NEJM199804093381503> (1998).
- Alward, W. L. M. *et al.* Evaluation of optineurin sequence variations in 1,048 patients with open-angle glaucoma. *American journal of ophthalmology* **136**, 904–910 (2003).
- Stone, E. M. *et al.* Identification of a gene that causes primary open angle glaucoma. *Science (New York, N.Y.)* **275**, 668–670 (1997).
- Wiggs, J. L. *et al.* Prevalence of mutations in TIGR/Myocilin in patients with adult and juvenile primary open-angle glaucoma, <https://doi.org/10.1086/302098> (1998).
- Shepard, A. R. *et al.* Glaucoma-causing myocilin mutants require the Peroxisomal targeting signal-1 receptor (PTS1R) to elevate intraocular pressure. *Human molecular genetics* **16**, 609–617, <https://doi.org/10.1093/hmg/ddm001> (2007).
- Fingert, J. H., Stone, E. M., Sheffield, V. C. & Alward, W. L. M. Myocilin glaucoma. *Survey of ophthalmology* **47**, 547–561 (2002).
- O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic acids research* **44**, D733–45, <https://doi.org/10.1093/nar/gkv1189> (2016).
- Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic acids research* **33**, D501–4, <https://doi.org/10.1093/nar/gki025> (2005).
- Yue, B. Y. J. T. Myocilin and Optineurin: Differential Characteristics and Functional Consequences. *Taiwan journal of ophthalmology* **1**, 6–11, <https://doi.org/10.1016/j.tjo.2011.08.002> (2011).
- Johnson, D. H. Myocilin and glaucoma: A TIGR by the tail? *Archives of ophthalmology (Chicago, Ill.: 1960)* **118**, 974–978 (2000).
- Tamm, E. R. Myocilin and glaucoma: facts and ideas. *Progress in retinal and eye research* **21**, 395–428 (2002).
- Snyder, D. A., Rivers, A. M., Yokoe, H., Menco, B. P. & Anholt, R. R. Olfactomedin: purification, characterization, and localization of a novel olfactory glycoprotein. *Biochemistry* **30**, 9143–9153 (1991).
- Hewitt, A. W., Mackey, D. A. & Craig, J. E. Myocilin allele-specific glaucoma phenotype database. *Human mutation* **29**, 207–211, <https://doi.org/10.1002/humu.20634> (2008).
- Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291, <https://doi.org/10.1038/nature19057> (2016).
- Kim, B. S. *et al.* Targeted Disruption of the Myocilin Gene (Myoc) Suggests that Human Glaucoma-Causing Mutations Are Gain of Function. *Molecular and cellular biology* **21**, 7707–7713, <https://doi.org/10.1128/MCB.21.22.7707-7713.2001> (2001).
- Fingert, J. H. *et al.* Analysis of myocilin mutations in 1703 glaucoma patients from five different populations. *Human molecular genetics* **8**, 899–905 (1999).

28. Miller, S. A., Dykes, D. D. & Polesky, H. F. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic acids research* **16**, 1215 (1988).
29. Dreos, R., Ambrosini, G., Perier, R. C. & Bucher, P. The Eukaryotic Promoter Database: expansion of EPDnew and new promoter analysis tools. *Nucleic acids research* **43**, D92–6, <https://doi.org/10.1093/nar/gku1111> (2015).
30. Dreos, R., Ambrosini, G., Groux, R., Cavin Perier, R. & Bucher, P. The eukaryotic promoter database in its 30th year: focus on non-vertebrate organisms. *Nucleic acids research* **45**, D51–D55, <https://doi.org/10.1093/nar/gkw1069> (2017).
31. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM (2013).
32. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research* **20**, 1297–1303, <https://doi.org/10.1101/gr.107524.110> (2010).
33. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, <https://doi.org/10.1093/nar/gkq603> (2010).
34. Abecasis, G. R. *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073, <https://doi.org/10.1038/nature09534> (2010).
35. Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65, <https://doi.org/10.1038/nature11632> (2012).
36. Ng, P. C. & Henikoff, S. SIFT: Predicting amino acid changes that affect protein function. *Nucleic acids research* **31**, 3812–3814 (2003).
37. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome research* **15**, 1034–1050, <https://doi.org/10.1101/gr.3715005> (2005).
38. Davydov, E. V. *et al.* Identifying a high fraction of the human genome to be under selective constraint using GERP. *PLoS computational biology* **6**, e1001025, <https://doi.org/10.1371/journal.pcbi.1001025> (2010).
39. Mort, M. *et al.* MutPred Splice: machine learning-based prediction of exonic variants that disrupt splicing. *Genome biology* **15**, R19, <https://doi.org/10.1186/gb-2014-15-1-r19> (2014).
40. Flicek, P. *et al.* Ensembl 2013. *Nucleic acids research* **41**, D48–55, <https://doi.org/10.1093/nar/gks1236> (2013).
41. Desmet, F.-O. *et al.* Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic acids research* **37**, e67, <https://doi.org/10.1093/nar/gkp215> (2009).
42. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation* **34**, 57–65, <https://doi.org/10.1002/humu.22225> (2013).
43. Shihab, H. A. *et al.* An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics (Oxford, England)* **31**, 1536–1543, <https://doi.org/10.1093/bioinformatics/btv009> (2015).
44. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310–315, <https://doi.org/10.1038/ng.2892> (2014).
45. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *American journal of human genetics* **91**, 839–848, <https://doi.org/10.1016/j.ajhg.2012.09.004> (2012).
46. Narasimhan, V. M. *et al.* Health and population effects of rare gene knockouts in adult humans with related parents. *Science (New York, N.Y.)* **352**, 474–477, <https://doi.org/10.1126/science.aac8624> (2016).
47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics (Oxford, England)* **25**, 2078–2079 (2009).
48. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* **26**, 841–842, <https://doi.org/10.1093/bioinformatics/btq033> (2010).
49. Andersen, M. C. *et al.* In silico detection of sequence variations modifying transcriptional regulation. *PLoS computational biology* **4**, e5, <https://doi.org/10.1371/journal.pcbi.0040005> (2008).
50. Nalpathamkalam, T., Derkach, A., Paterson, A. D. & Merico, D. Genetic Analysis Workshop 18 single-nucleotide variant prioritization based on protein impact, sequence conservation, and gene annotation. *BMC proceedings* **8**, S11, <https://doi.org/10.1186/1753-6561-8-S1-S11> (2014).
51. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125–2137, <https://doi.org/10.1093/hmg/ddu733> (2015).
52. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS computational biology* **12**, e1004873, <https://doi.org/10.1371/journal.pcbi.1004873> (2016).
53. Fautsch, M. P. & Johnson, D. H. Characterization of myocilin-myocilin interactions. *Investigative ophthalmology & visual science* **42**, 2324–2331 (2001).
54. Craig, J. E. *et al.* Evidence for genetic heterogeneity within eight glaucoma families, with the GLC1A Gln368STOP mutation being an important phenotypic modifier. *Ophthalmology* **108**, 1607–1620 (2001).
55. Gobeil, S. *et al.* Intracellular sequestration of hetero-oligomers formed by wild-type and glaucoma-causing myocilin mutants. *Investigative ophthalmology & visual science* **45**, 3560–3567, <https://doi.org/10.1167/iiov.04-0300> (2004).
56. Yam, G. H.-F., Gaplovska-Kysela, K., Zuber, C. & Roth, J. Aggregated myocilin induces russell bodies and causes apoptosis: implications for the pathogenesis of myocilin-caused primary open-angle glaucoma. *The American journal of pathology* **170**, 100–109, <https://doi.org/10.2353/ajpath.2007.060806> (2007).
57. Graul, T. A. *et al.* A case-control comparison of the clinical characteristics of glaucoma and ocular hypertensive patients with and without the myocilin Gln368Stop mutation. *American journal of ophthalmology* **134**, 884–890 (2002).
58. Mataftsi, A. *et al.* MYOC mutation frequency in primary open-angle glaucoma patients from Western Switzerland. *Ophthalmic genetics* **22**, 225–231 (2001).
59. Willoughby, C. E. *et al.* Defining the pathogenicity of optineurin in juvenile open-angle glaucoma. *Investigative ophthalmology & visual science* **45**, 3122–3130, <https://doi.org/10.1167/iiov.04-0107> (2004).
60. Faucher, M. *et al.* Founder TIGR/myocilin mutations for glaucoma in the Quebec population. *Human molecular genetics* **11**, 2077–2090 (2002).
61. Pandaranayaka, P. J. E. *et al.* Polymorphisms in an intronic region of the myocilin gene associated with primary open-angle glaucoma—a possible role for alternate splicing. *Molecular vision* **16**, 2891–2902 (2010).
62. Liu, W. *et al.* Low prevalence of myocilin mutations in an African American population with primary open-angle glaucoma. *Molecular vision* **18**, 2241–2246 (2012).
63. Banerjee, D., Bhattacharjee, A., Ponda, A., Sen, A. & Ray, K. Comprehensive analysis of myocilin variants in east Indian POAG patients. *Molecular vision* **18**, 1548–1557 (2012).
64. Colomb, E. *et al.* Association of a single nucleotide polymorphism in the TIGR/MYOCILIN gene promoter with the severity of primary open-angle glaucoma. *Clinical genetics* **60**, 220–225 (2001).
65. Guo, H., Li, M., Wang, Z., Liu, Q. & Wu, X. Association of MYOC and APOE promoter polymorphisms and primary open-angle glaucoma: a meta-analysis. *International journal of clinical and experimental medicine* **8**, 2052–2064 (2015).
66. Joe, M. K., Lieberman, R. L., Nakaya, N. & Tomarev, S. I. Myocilin Regulates Metalloprotease 2 Activity Through Interaction With TIMP3. *Investigative ophthalmology & visual science* **58**, 5308–5318, <https://doi.org/10.1167/iiov.16-20336> (2017).
67. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1, <https://doi.org/10.1126/scisignal.2004088> (2013).
68. Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401–404, <https://doi.org/10.1158/2159-8290.CD-12-0095> (2012).

Acknowledgements

We thank the families for their participation in this research and the International Glaucoma Association (IGA) and Gift of Sight for funding this study. We thank Matthew Mort for his assistance with MutPred Splice. We would also like to thank Nishani Amersinghe and Alex MacLeod (University Hospital Southampton), Ruth Manners (Romsey Hospital), James Kirwan (Queen Alexandra Hospital Portsmouth), Geeta Menon (Frimley Park Hospital), Keith Martin (Addenbrooks Hospital Cambridge), Yit Yang (New Cross Hospital Wolverhampton), Michael Cole Andrew Frost (Torbay Hospital Devon) and Javeed Khan (St Mary's Hospital Isle of Wight) for acting as principle investigators and enrolling glaucoma patients into this study.

Author Contributions

L.O. designed gene target selection, performed data analysis, and manuscript preparation. A.C. performed data analysis and contributed to the manuscript. D.W. assisted in wet lab experiments. H.L.G. performed DNA extraction, targeted sequencing methodology and contributed to the manuscript. R.S. contributed to data analysis. A.D. contributed to collection and phenotyping of patient samples. J.E.S. contributed to manuscript preparation. S.E. contributed to study design, data analysis and interpretation, and manuscript preparation. A.L. organised collection and phenotyping of patient samples, PI of grant funding this work and contributed to the manuscript. J.G. contributed to study design, data analysis and interpretation, and manuscript preparation. All authors approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-38760-y>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019

SOFTWARE

Open Access



GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data

E. Mossotto^{1,2*} , J. J. Ashton^{1,3}, L. O'Gorman¹, R. J. Pengelly^{1,2}, R. M. Beattie³, B. D. MacArthur² and S. Ennis¹

Abstract

Background: Next-generation sequencing is revolutionising diagnosis and treatment of rare diseases, however its application to understanding common disease aetiology is limited. Rare disease applications binarily attribute genetic change(s) at a single locus to a specific phenotype. In common diseases, where multiple genetic variants within and across genes contribute to disease, binary modelling cannot capture the burden of pathogenicity harboured by an individual across a given gene/pathway.

We present GenePy, a novel gene-level scoring system for integration and analysis of next-generation sequencing data on a per-individual basis that transforms NGS data interpretation from variant-level to gene-level. This simple and flexible scoring system is intuitive and amenable to integration for machine learning, network and topological approaches, facilitating the investigation of complex phenotypes.

Results: Whole-exome sequencing data from 508 individuals were used to generate GenePy scores. For each variant a score is calculated incorporating: i) population allele frequency estimates; ii) individual zygosity, determined through standard variant calling pipelines and; iii) any user defined deleteriousness metric to inform on functional impact. GenePy then combines scores generated for all variants observed into a single gene score for each individual. We generated a matrix of ~ 14,000 GenePy scores for all individuals for each of sixteen popular deleteriousness metrics. All per-gene scores are corrected for gene length. The majority of genes generate GenePy scores < 0.01 although individuals harbouring multiple rare highly deleterious mutations can accumulate extremely high GenePy scores.

In the absence of a comparator metric, we examine GenePy performance in discriminating genes known to be associated with three common, complex diseases. A Mann-Whitney U test conducted on GenePy scores for this positive control gene in cases versus controls demonstrates markedly more significant results ($p = 1.37 \times 10^{-4}$) compared to the most commonly applied association tool that combines common and rare variation ($p = 0.003$).

Conclusions: *Per-gene per-individual* GenePy scores are intuitive when assessing genetic variation in individual patients or comparing scores between groups. GenePy outperforms the currently accepted best practice tools for combining common and rare variation. GenePy scores are suitable for downstream data integration with transcriptomic and proteomic data that also report at the gene level.

Keywords: Genome analysis, Mathematical modelling, Next-generation sequencing, Gene score, Pathogenicity score

* Correspondence: Enrico.Mossotto@soton.ac.uk

¹Department of Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK

²Institute for Life Sciences, University of Southampton, Southampton, UK

Full list of author information is available at the end of the article



© The Author(s). 2019 **Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

In the last decade, next-generation sequencing (NGS) has emerged as an effective tool for detecting single nucleotide variants (SNVs) causing rare conditions [1]. Recent retrospective studies have demonstrated an increase of 25–31% in diagnostic yield of rare diseases due to the application of exome or whole genome sequencing in a clinical framework [2, 3]. Through comparison against human genome reference sequence, high quality NGS data on individual patients can be used to identify variation in variant call files (VCF). These files typically contain in excess of 30,000 variants when based on whole exome data that captures sequence on the protein coding region of the genome only and runs to many millions when based on whole genome data. The successful identification of disease causing variation is critically dependent upon annotation and subsequent filtering of these data. Filtering strategies typically focus on very rare variants in panels of genes empirically implicated as related to the clinical manifestation or phenotype of interest. Further exclusion of synonymous variants that have no impact on protein amino acid sequence and variants that occur at a frequency substantially greater than that of the disease of interest are also deprioritised. These steps can reduce the search space for causal variation by orders of magnitude to smaller sets of hundreds or even tens of genetic changes that are then prioritised by *in silico* methods [4].

Many *in silico* tools have been developed in order to estimate the potential impact of genetic variants on gene/protein function. Predicting pathogenicity or deleterious impact can be achieved through a variety of algorithms that focus on one or more specific biological aspect(s). Three broad classes of deleteriousness prediction metrics are: (i) conservation metrics, (ii) function alteration metrics and (iii) composite scores. Conservation metrics such as GERP++ [5], phastCons [6] and phyloP [7] assign a high deleteriousness to variants where the homologous position in other species has remained constrained over evolutionary history. Scores focused on predicting the potential disruption of protein functionality, for example through alteration of resultant protein amino acid sequence, include SIFT [8], FATHMM [9], fathmm-MKL [10], PolyPhen2 [11], MutationTaster [12], PROVEAN [13] and VEST3 [14].

To date, no single *in silico* metric has proven unilateral superiority in estimating consequent severity, despite an expanding list [15] of metrics based on subtly different foundations and assumptions. While individual metrics have the ability to perform well in isolation, discordant evidence when assessing the same data with multiple metrics has led to increased uncertainty in choice of prediction tool [16]. This in turn has led to the development of a range of composite prediction tools

applying statistical and machine learning methodologies that combine metrics assessing both conservation *and* functionality in order to obtain higher accuracy [17]. The most utilised composite scores include CADD [18], MetaSVM and MetaLR [19], M-CAP [20], Eigen [21], hyperSMURF [22] and DANN [23] with no one method emerging as optimal [24]. For this reason, when assessing variant deleteriousness it is still necessary to observe consensus prediction based on multiple scoring metrics rather than focusing on any single score [25]. This remains the case when studying rare Mendelian disease where single gene mutations imparting severe consequence are expected to represent the most extreme set of deleterious variants.

In contrast to rare diseases, common genetic diseases such as ischemic heart disease, asthma, inflammatory bowel disease (IBD) or Alzheimer's disease are caused by the combined action of multiple genetic variants each differentially impacting risk and disease severity while working in combination with environmental exposures [26]. Collectively, common diseases impose an enormous economic burden and arguably have the greatest unmet need for diagnosis and stratified treatment [27]. The set of genes and variants imparting increased susceptibility vary from one patient to the next even when clinical presentation and molecular pathology appear indistinct.

Prior to transformative NGS approaches, genome-wide association studies (GWAS) made substantial advances in explaining the molecular bases of complex diseases. These studies tagged up to a million common single nucleotide markers across the genome and identified statistically significant distributions of biallelic markers in large cohorts of independent patients compared to ethnically matched controls. Genetic regions implicated by GWAS were assumed to harbour genes or regulatory elements underpinning the disease of interest. However, because these genetic breakthroughs were achieved using necessarily huge cohorts of patients compared to controls, while their findings hold true for massive patient groups, they are largely uninformative on an individual patient basis. Importantly, the relevance and value of GWAS findings to individual patients has therefore not translated through to clinical practice in terms of either diagnosis or treatment.

Application of NGS to improve our understanding of common oligogenic diseases have been largely limited to burden tests that extend the association testing framework to integrate information about common and rare variation across discrete genomic regions such as genes. While this approach harnesses the power of NGS through inclusion of rare variants that can only be detected by sequencing approaches, they are most often implemented through collapsing multiple variants into a single value for univariate analysis. The limited success

of these approaches are partly attributed to their intrinsic lack of biological information and inclusion of both causal and benign genetic variation [28, 29]. In order to overcome this limitation, Neale et al. developed the C-alpha test, correcting for both protective and deleterious variants but at the cost of losing statistical power. Currently, SKAT (and SKAT-O optimised for small sample size) [30] represents the most sensitive approach to test for association between a genomic region and a phenotype. SKAT jointly assesses both rare and common variants maximising the statistical power and representing a new class of analysis lying between burden and association tests and has been successfully applied to a large variety of complex diseases [31–35].

While NGS is proving a transformative technology for the diagnosis and treatment of rare diseases, its relatively modest application in common diseases is limited by a lack of analytical approaches that incorporate *individual* profiles of genetic variation ascertained through NGS annotated with biologically meaningful information on their frequency and consequence.

Instead of variant focussed approaches typical for rare disease or large cohort approaches that distinguish GWAS, contemporary analyses of complex polygenic disorders require the development of tools that combine both mutational burden and biological impact of a personalised set of mutations into single scores for discrete sub-genomic units such as genes. A matrix of such a set of scores for any one individual could then be analysed using various methodology including machine learning.

In this study, we describe the development and implementation of GenePy, a novel gene-level scoring system for integration and analysis of next-generation sequencing data on a per-individual basis. The goal of the GenePy scoring system is not to create a statistical tool for burden or association tests, but to generate a novel scoring system that transforms NGS data interpretation from variant level to gene level. The aim is to enable a gene based scoring system for individuals that can be used to compare single gene pathogenicity between individuals or to prioritise genes with high pathogenic loading for scrutiny for any single individual. In addition, GenePy aims to increase the intrinsic biological information content by incorporating data on allele frequency and observed zygosity in addition to any user-defined variant deleteriousness metric. The GenePy scoring system aims to transform typical sequencing data output into a format suitable for integration into downstream network analyses or machine learning approaches for stratification. In the absence of other comparator scoring systems, we validate GenePy performance on three complex diseases: paediatric inflammatory bowel disease (IBD), Parkinson's disease (PD) and primary open angle glaucoma (POAG).

Implementation

Sample data

Whole exome sequencing (WES) data were derived from two sources. This first group comprised 309 patients diagnosed in childhood with IBD. This cohort (further described in [36]) includes unrelated, Caucasian patients ascertained and recruited through Southampton Children's Hospital who were diagnosed under the age of 18 years according to the modified Porto criteria [37]. Additional WES data from a cohort of 199 anonymised individuals diagnosed with an infectious disease but unselected for any form of autoimmune disease were also used to give a total cohort size of 508 individuals with WES data.

Genomic DNA was extracted from peripheral venous blood and fragmented DNA subjected to adaptor ligation and exome library enrichment using the Agilent SureSelect All Exon capture kit versions 4, 5 and 6. Enriched libraries were sequenced on Illumina HiSeq systems.

WES data processing

Raw sequencing fastq sequencing data from all 508 samples were processed using the same custom pipeline. VerifyBamID [38] was utilised to check the presence of DNA contamination across our cohort of 508 individuals. Alignment was performed against the human reference genome (GRCh38/hg38 Dec. 2013 assembly) using BWA [39] (version 0.7.12). Aligned BAM files were sorted and duplicate reads were marked using Picard Tools (version 1.97). Following GATK v3.7 [40] best practice recommendations [41], base qualities were recalibrated in order to correct for systematic errors produced during sequencing. Finally, variants were called using GATK HaplotypeCaller was applied to produce a gVCF file for each sample. Samples were processed on the University of Southampton IRIDIS cluster requiring an average of 4 h run time per sample on a 16-processor node.

While the standard VCF format reports only alternative calls, the gVCF format identifies non-variant blocks of sequencing data and returns reference calls for loci therein. This enables affirmative calling of homozygous reference loci when combining call sets from multiple samples. Multi-sample variant calling was achieved through calling each individual sample separately and then merging all gVCFs using GATK GenotypeGVCFs. Processing efficiency was optimised for the set of 508 individual samples through batching into six subsets using GATK's CombineGVCFs (approx. 6 h/batch on a 16 processor node) and the resultant six gVCF files were merged for genotyping with GenotypeGVCFs (approx 1 h on a 16 proc. node). Annotation of this composite file applied Annovar v2016Feb01 using default databases refSeq gene transcripts (refGene), deleteriousness scores

databases (dbnsfp33a) and dbSNP147). Variant allele frequencies were sourced through Annovar (ExAc03 [42]) or ensembl human variation API [43] where ExAc data were missing.

Quality control framework

In order to reduce heterogeneity, it is necessary to control for bias encountered due to alternative capture kit versions and variant quality. For the entire cohort of 508 samples, exon enrichment was performed using Agilent SureSelect capture kits but at different time-points. For this reason, there is inter-capture kit variability across the 508 cohort with kit versions 4, 5 and 6 being applied. To correct for disparity in the regions targeted by respective versions, all downstream analyses were restricted to the set of overlapping targeted genomic locations (as defined by respective kit BED files) using BEDtools v2.17 [44].

Following GATK best practice guidelines, Haplotype-Caller default settings were utilised, implying that only variants with a minimum Phred base quality score of 20 were called.

GenePy score

Individuals typically have multiple variants across the coding region of genes making the interpretation of their combined effect challenging. We hypothesised that for each individual sample h within our cohort $H = \{h_1, h_2, \dots, h_n\}$, the loss of integrity of any given gene g in the RefGene database $G = \{g_1, g_2, \dots, g_m\}$ can be quantified as the sum of the effect of all (k) variants within its coding region observed in that sample, where each biallelic mutated locus (i) in a gene is weighted according to its predicted allele deleteriousness (D_i), zygosity and allelic frequency (f_i). The GenePy score S_{gh} for a given gene (g) in individual (h) is

$$S_{gh} = - \sum_{i=1}^k D_i \log_{10}(f_{i1} \cdot f_{i2})$$

At any one variant locus (i), we represent both parental alleles using f_{i1} and f_{i2} to embed the population frequency of allele₁ and allele₂ and, in doing so, model observed biological information on both frequency and zygosity. Any homozygous genotype therefore is simply the observed allele frequency squared whereas the product of each of the observed alleles is calculated for heterozygous genotypes. The latter can therefore accommodate variant sites with multiple alleles in addition to the typically encountered biallelic single nucleotide polymorphisms (SNPs). Hemizygotic variation from male X-chromosomes are treated as homozygotic. Where a variant may be novel to an individual or absent from reference databases, we impose a lower frequency limit of 0.00001.

This lower limit is arbitrarily set to conservatively reflect the lowest frequency that can be observed in the largest current repository of human variation (ExAc03). The log function is applied to upweight the biological importance of rare variation.

The GenePy algorithm represents a genetic mixed model, combining the known multiplicative effect of two alleles at a single diploid locus [45] (the frequencies of both observed alleles are multiplied) but with an additive effect at the gene level (variant scores are summed within a gene). The contribution of all variation within a gene is modelled in this additive fashion in order to enable the cumulative pathogenicity incurred from the effects of multiple small/modest effects imposed by individual mutations thus reflecting the non Mendelian inheritance pattern in common diseases. An additive model is assumed to be most universally applicable model particularly in the non-Mendelian situation relevant to many common diseases [46].

Deleteriousness metrics were developed to assess damage induced by nonsynonymous variation, therefore structural variants such as frameshifts or stop mutations that truncate proteins are not routinely assigned deleteriousness values. Due to their highly detrimental impact to function we assign all protein truncating mutations the maximal deleteriousness value of 1. Synonymous and splicing variants are not routinely annotated by ANNOVAR and were not included in the current assessment.

Importantly, the choice of variant deleteriousness score is user-defined, and therefore the GenePy score is able to take into account different definitions of pathogenicity depending on context. Herein we examine the relative attributes of using any one of sixteen of the most commonly applied scores (Table 1). Sixteen of the most common deleteriousness (D) metrics were selected for implementation within the GenePy algorithm. Five of these metrics (shown in bold) are unbounded. In order to implement unbounded metrics in GenePy it was necessary to impose lower and upper limits by applying the respective minimum and maximum values observed in the dbnsfp33a database of 83,422,341 known SNV mutations. These limits were used to transform observed values in our cohort scaled to 0–1.

As a function of their size alone, larger genes have greater opportunity to accrue higher deleterious GenePy scores through having a greater number of variants thus inflating GenePy scores. We therefore generated GenePy scores corrected for the length of targeted gene regions (GenePy_{egl}) by dividing the GenePy score by the targeted length in base pairs and then multiplying by the median observed targeted gene length in our data (1461 base pairs). A final set of 16 deleteriousness metrics, each with a range of 0–1 where highest values were most deleterious, were individually implemented in the model.

Table 1 Pathogenicity scores for SNVs and their reported ranges in the dbsnp database

Metric	Type	Implementation	Actual range	Imposed range for transformation
CADD	Composite	Score	$-\infty$ to $+\infty$	−7.53 to 35.79
DANN	Composite	Score	0 to 1	–
FATHMM ^a	Functionality	1-Score	$-\infty$ to $+\infty$	−16.13 to 10.64
fathmm-MKL	Composite	Score	0 to 1	–
GERP++_RS	Conservation	Score	$-\infty$ to $+\infty$	−12.3 to 6.17
M-CAP	Composite	Score	0 to 1	–
MetaLR	Composite	Score	0 to 1	–
MetaSVM	Composite	Score	$-\infty$ to $+\infty$	−2 to 3
MutationTaster ^a	Functionality	1-Score if N/P; Score if A/D	0 to 1	–
phastCons	Conservation	Score	0 to 1	–
phyloP	Conservation	Score	$-\infty$ to $+\infty$	−13.28 to 1.2
Polyphen2_HDIV	Functionality	Score	0 to 1	–
Polyphen2_HVAR	Functionality	Score	0 to 1	–
PROVEAN ^a	Functionality	1-Score	−14 to 14	–
SIFT ^a	Functionality	1-Score	0 to 1	–
VEST3	Functionality	Score	0 to 1	–

^aIn order to maintain uniform directionality, the complement (1 – score) of a value was taken so that across scores, a value of 0 consistently indicated benign variation and a value of 1 inferred maximal pathogenicity

GenePy score validation on the IBD dataset

In the absence of any comparable gene based scoring system for individuals, GenePy performance was benchmarked by assessing the power to determine significantly different score distributions in disease cases compared to controls for a known causal gene through a Mann-Whitney U test. Using the same variant data, the statistical difference in GenePy scores was compared against that of SKAT-O - the most commonly applied gene level association test. The cohort comprised 309 individuals diagnosed with inflammatory bowel disease (IBD) and 199 controls unselected for autoimmune conditions. The analysis focussed on the *NOD2* gene - the most strongly and repeatedly associated common disease gene conferring strong association specifically with the Crohn's disease (CD) subtype of IBD [47–49]. *NOD2* was selected as a positive control gene, whereby evidence for increased burden of deleterious mutation encoded in CD patient DNA compared to either ulcerative colitis (UC) or control DNA is expected.

The matrix of *NOD2* GenePy scores calculated for all 508 samples was split into controls and cases with the latter further divided into UC and CD subtypes. Statistical significance of GenePy score distribution difference between groups was calculated using the Mann-Whitney U test for unpaired data. Using the same variant input data, the SKAT-O gene based test for association was performed twice using default settings: firstly by considering all variants called within *NOD2* and secondly including only rare variants (MAF < 0.05) as per developer recommendations [30].

Association tests succumb to false positive results due to spurious association brought about by population stratification or systematic differences in case versus control data. We excluded non-Caucasian individuals identified through comparison against the 1000 Genomes Project [50] using Peddy software [51] for ethnic imputation. We enforced parity in sequencing depth (known to impact power to call genetic variation [52]) for case-control data by limiting all score validation data to variants called in gene regions with a minimum read depth of 50X.

GenePy score validation on the Parkinson's disease dataset

A second validation of the GenePy score was performed using WES from the Parkinson's Progression Marker Initiative (PPMI) [53]. Six hundred and ten Caucasian patients diagnosed with Parkinson's disease (PD) were selected from this cohort. No control data were generated within this cohort.

Parkinson's disease is a common complex condition involving the central nervous system. Disease aetiology is complex and only partially understood, but the increased risk of occurrence driven by family history of disease indicates a strong genetic component [54]. To date, several genes have been associated with Parkinson's disease, however only few have been validated as disease causing. In our approach, we focussed on the panel of six genes routinely tested in clinical settings: *LRRK2*, *PRKN* (*PARK2*), *PARK7*, *PINK1*, *SNCA* and *VPS35*. The gene panel and

technical notes are further described the UK Genetic Testing Network database (<https://ukgtn.nhs.uk>).

Whole exome sequencing data for this cohort was generated using Illumina 2500 sequencing machines and Nextera Rapid Capture Expanded Exome Kit. Raw sequencing data were processed as per those for the IBD cohort. GenePy scores, implementing the CADD deleteriousness metric (given CADD's high performance and more complete gene annotation), were generated for 610 PD samples for the six genes included in the panel. GenePy distributions in PD cases were compared using a Mann-Whitney U test against non-PD samples. In the absence of within-cohort control data, IBD and control samples described above were used as non-PD controls for these tests. In order to assure compatibility, GenePy scores were calculated only for common regions targeted by both Nextera and Agilent exon enrichment capture kits used by the respective studies (intersection of bed files). Statistical significance was compared with results obtained through a SKAT-O test as previously described.

We further tested the ability of GenePy to detect extreme gene differences between PD patients and non-PD individuals. A one-tailed Mann-Whitney U test was conducted between the highest 5% of the GenePy distribution scores from the PD patients and the highest 5% of the non-PD cohort for each gene investigated.

GenePy score validation on the primary open angle Glaucoma cohort

The third validation of GenePy was performed on a cohort of Caucasian patients ($n = 358$) affected by primary open angle glaucoma (POAG) [55], a glaucoma subtype characterised by an open and normal anterior chamber angle, increased intraocular pressure and no other concurrent adverse phenotypes [56]. POAG is a common complex condition with a strong genetic component with first-degree relatives of affected individuals harbouring an eight-fold increased risk [57]. Previous studies have established *MYOC* as causative gene in approximately 3% of the POAG diagnoses [58].

Sequencing data for the POAG cohort were generated using Nextera Rapid Capture Custom Enrichment kit, the Nextera 500 sequencing platform and the same best practice bioinformatic pipeline as applied in the IBD cohort [59].

Mann-Whitney U was applied to test whether GenePy was capable of detecting a statistically significant difference between the POAG cohort and non-POAG samples (using IBD and control samples as a proxy for matched controls as above) within the *MYOC* gene. Regions common to the Nextera Rapid Capture Custom Enrichment kit and Agilent SureSelect Capture chemistries were selected using bed file data to ensure compatibility of GenePy scores.

The difference between extreme GenePy scores in the POAG patients compared to non-POAG individuals was assessed. Given the known frequency of *MYOC* pathogenic mutations of 3%, statistically significant differences within the extreme top 3% distribution of both groups was compared as above.

Results

QC results

All WES data ($n = 508$, $n_{\text{ibd}} = 309$, $n_{\text{ctrl}} = 199$) underwent quality control assessment for contamination using VerifyBamID and were confirmed free of contamination (free-mix statistic < 0.01). Out of 508 individuals, we identified three pairs of first degree relatives, one set of monozygotic twins and one mother-father-child trio. In order to correct for relatedness, which would bias association tests, for each pair, the sample with poorest coverage data was excluded. For the trio, the child data were excluded and unrelated parents retained.

GenePy score behaviour – impact of allele frequency and zygosity

Figure 1 shows the results of simulated GenePy score (y-axis) calculated across a range of deleterious metric scores (0.1, 0.5, 0.75, 0.9, 0.95, 0.99) with varying minor allele frequency (x-axis) and further depicts the consequence of heterozygote versus homozygote states. The plot reveals the logarithmic nature of GenePy scores for a single locus only (whereas for any individual, their per gene GenePy score is weighted sum of all variant scores observed in that individual across that gene). For any single variant, the theoretical maximum observable GenePy value of ten occurs only with highest deleteriousness value (D), the lowest minor allele frequency ($\text{MAF} = 0.00001$) and in the homozygous state whereas the upper limit for a heterozygote with the same deleteriousness and frequency settings is five. The logarithmic scale implemented in GenePy algorithm confers rapidly increasing scores as the MAF approaches novelty.

GenePy score behaviour – impact of deleteriousness metric

While there are 27,238 genes annotated in RefSeq, we aimed to generate GenePy scores only for the overlapping subset of 21,577 target genes captured by all versions of the SureSelect capture kits applied. The GenePy scoring algorithm was executed for each of sixteen commonly applied metrics (Table 1). There is fluctuation in the number of genes for which variants were annotated with deleteriousness metric data using ANNOVAR ranging from 12,921 for M-CAP (one of the most recently released scores) to 14,745 genes annotated scores for Polyphen2_HDIV (one of the earliest developed deleteriousness scores) (Table 2). Among the 508 individuals

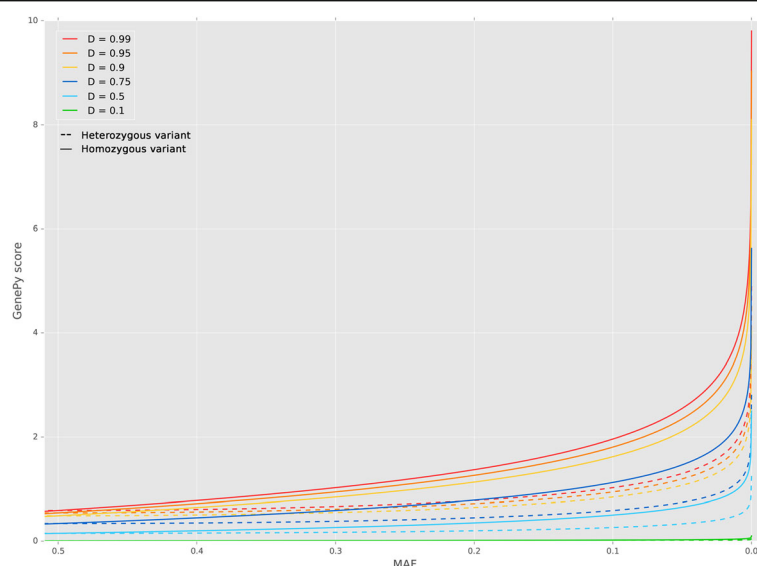


Fig. 1 Single variant GenePy score distribution under fixed deleteriousness values. Impact of varying zygosity and minor allele frequency (MAF)

that underwent GenePy scoring of exome data, the majority of genes are invariant within any one individual (e.g. median 9917 for CADD metric). This is expected for intrinsically sparse genomic data. However, across the cohort, no single gene returns a GenePy score of zero in all individuals indicating all genes have at least one rare variant observed amongst the 508 individuals. The vast majority of genes are scored with GenePy values of less than 0.01 and correction for gene length marginally increases the number of genes achieving lowest scores. More than 97% of genes achieve a score of less than 0.01 when the M-CAP metric is used whereas FATHMM scores approximately 65% of genes in the 0–0.01 range. The inflated percentage of invariant genes observed when implementing M-CAP is explained by its tendency to depress weight for benign variants compared to other tested metrics [20].

Across the ~ 14,000 genes achieving GenePy scores, the observed score mean (uncorrected for length) in our cohort of 508 samples ranges from 0.02 to 0.40 depending on the applied deleteriousness metric. Correction of all scores for gene length has only a modest effect on the range of the mean scores observed (0.02–0.31), however, gene length correction increases the spread of the data reflected by an approximate two-fold increase in the coefficient of variation (CV) for GenePy scores observed across all sixteen deleteriousness metrics. This is despite the fact that for all deleteriousness metrics, correction for gene length subtly increases the proportion of genes with lowest scores confirming that genes of exceptional size incurred inflated scores due to length. GenePy scores generated with M-CAP are least impacted by gene length correction but maintain the largest CV.

In order to further investigate the behaviour of GenePy scores across genes, we calculated the median number of genes exhibiting scores falling within non-overlapping bins across the entire cohort. Figure 2 shows the profiles for the 0.01 to 6 range of GenePy scores and a bin size of 0.01. Genes with scores < 0.01 are overrepresented (Table 2) and not shown. Across most of the sixteen metrics, a distinct pattern characterised by two spikes around uncorrected GenePy scores of 0.6 and 5 represent genes strongly influenced by a single highly deleterious common homozygous variants ($D = 1$, $MAF = 0.5$) or a single highly deleterious very rare heterozygous variant ($D = 1$, $MAF = 0.00001$) respectively. This profile was apparent for most deleteriousness metrics (except CADD, FATHMM, MetaSVM and VEST3, see Additional file 1: Figure S1). These two distinctive spikes are not observable once GenePy scores are corrected for the targeted gene length (Fig. 1, lower panel and Additional file 1: Figure S2). We did not observe further spikes or other anomalies in the long right tail of the distribution of scores greater than 6.

For a subset of 6 patients we plot the gene-level scores for 17 genes across two different molecular pathways important to immune function (Fig. 3). This graphically demonstrates how individual patients diagnosed with the same non-Mendelian condition have unique gene-level deleteriousness score profiles. Individual patients can be genetically compromised within the same or distinct molecular pathways.

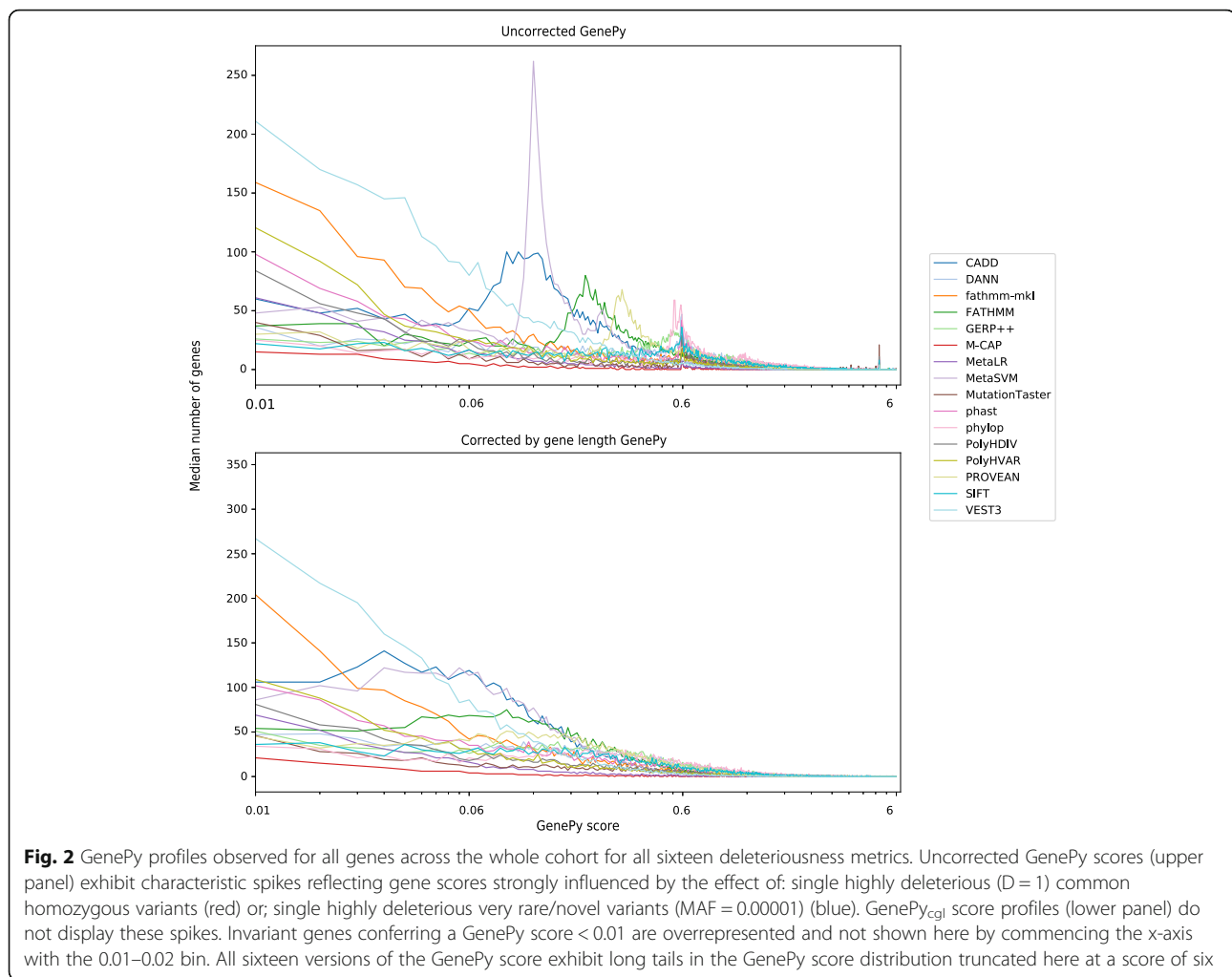
GenePy score validation - IBD cohort

Bias conferred by *NOD2* gene coverage, related samples and non-Caucasian ethnicity (Additional file 1: Figure S3)

Table 2 Statistical attributes of whole gene GenePy scores computed for sixteen deleteriousness metrics. Number of genes for which GenePy scores were calculated, median number of non-variant genes (GenePy = 0), mean GenePy scores, mean and standard deviation across our cohort (n = 508), coefficient of variation (CV, defined as σ/μ) and the median number of genes with a GenePy score < 0.01 as percentage of the total number of genes. The same information is reported for GenePy_{cgl}

Metric	Gene scores calculated	^a Median no. of genes with GenePy = 0 within individuals (%)	Max GenePy	Mean GenePy	CV uncorrected	Median no. of genes with GenePy < 0.01 within individuals (%)	Max GenePy _{cgl}	Mean GenePy _{cgl}	CV _{cgl} corrected	Median no. of genes with GenePy _{cgl} < 0.01(%)
CADD	14,184	9917 (69.92%)	32.15	0.10	3.81	10,231 (72.13%)	74.19	0.08	8.09	10,304 (72.64%)
DANN	14,184	9917 (69.92%)	110.48	0.33	3.37	10,153 (71.58%)	304.15	0.25	6.96	10,196 (71.88%)
FATHMM	13,143	9981 (75.94%)	72.73	0.16	4.15	10,923 (83.11%)	269.62	0.11	6.42	11,092 (84.40%)
fathmm-MKL	14,178	9039 (63.75%)	50.10	0.16	3.29	9282 (65.48%)	131.34	0.12	7.55	9332 (65.84%)
GERP++_RS	14,197	9910 (69.80%)	100.44	0.32	3.35	10,116 (71.25%)	283.69	0.24	6.47	10,143 (71.44%)
M-CAP	12,921	12,577 (97.34%)	24.52	0.02	12.65	12,596 (97.48%)	59.88	0.02	19.05	12,630 (97.74%)
MetaLR	14,063	12,752 (90.68%)	38.14	0.04	8.77	13,146 (93.48%)	87.80	0.04	16.14	13,253 (94.24%)
MetaSVM	14,076	9845 (69.94%)	36.76	0.10	3.95	10,141 (72.04%)	99.44	0.08	8.94	10,207 (72.51%)
MutationTaster	14,039	12,161 (86.62%)	90.86	0.13	5.24	12,521 (89.19%)	332.05	0.09	9.02	12,579 (89.60%)
phastCons	14,197	10,217 (71.97%)	100.64	0.21	3.79	11,018 (77.60%)	324.41	0.14	5.76	11,116 (78.29%)
phyloP	14,202	9910 (69.78%)	118.81	0.40	3.31	10,107 (71.17%)	332.05	0.31	7.15	10,131 (71.34%)
Polyphen2_HDIV	14,745	11,824 (80.19%)	65.48	0.14	4.89	12,558 (85.16%)	257.00	0.12	12.08	12,658 (85.84%)
Polyphen2_HVAR	14,741	11,470 (77.81%)	59.67	0.11	5.47	12,621 (85.62%)	239.71	0.09	14.03	12,778 (86.69%)
PROVEAN	13,888	9733 (70.08%)	74.16	0.23	3.37	9958 (71.70%)	219.39	0.17	7.93	10,003 (72.02%)
SIFT	14,561	11,088 (76.15%)	99.69	0.25	3.69	11,224 (77.08%)	265.64	0.20	7.04	11,257 (77.31%)
VEST3	14,170	9919 (70.00%)	53.36	0.09	5.69	10,528 (74.29%)	136.56	0.08	12.56	10,821 (76.36%)

^aAcross the cohort of 508 individuals assessed, individual samples have a very high median number of invariant genes resulting on GenePy scores of zero



was removed from all IBD cases ($n = 6_{<50x}$, $n = 1_{relative}$ and $n = 20_{non-Caucasian}$) and non-IBD control samples ($n = 16_{<50x}$, $n = 4_{relatives}$ and $n = 13_{non-Caucasian}$) respectively. There remained 282 IBD cases for analysis of which 172 were diagnosed with Crohn's disease, 100 with ulcerative colitis and a further 10 patients had a diagnosis of IBD undetermined (IBDU). There was a corresponding number of 166 controls.

The *NOD2* GenePy scores for the 282 IBD and 166 control individuals were calculated using all sixteen deleteriousness metrics. (Additional file 1: Figure S4). Given *NOD2* gene variant association is specific to the CD subtype of IBD, we calculated GenePy scores for both subtypes and grouped separately (Additional file 1: Table S1).

The Mann-Whitney U test comparison of the distribution of *NOD2* GenePy scores between all IBD, CD and UC subtypes against controls identified statistically significant differences (Table 3). Only modestly significant differences for just three of the implemented deleteriousness metrics (M-CAP, fathmm-mkl and MutTaster)

were observed comparing all IBD against controls in this relatively small sample. When the cases were stratified by disease subtype, UC samples had significantly lower GenePy scores compared to controls but only for two of the implemented deleteriousness metrics (MetaLR, phastCons). As expected, the most significant difference in *NOD2* score distribution was observed when comparing CD patients only against controls. Without exception, a highly significant difference was observed using every deleteriousness metric with M-CAP the most significant ($p = 1.37 \times 10^{-4}$) all of which would withstand correction for the three independent tests performed. Regardless of which deleteriousness metric is used, the mean GenePy score is consistently higher in CD patient when compared with controls.

Interestingly, similar results were observed for the SKAT-O gene test of association when using all variant frequency data but lost significance when restricted to rare variation ($MAF < 0.05$). Importantly, the magnitude of the difference between CD patients and control

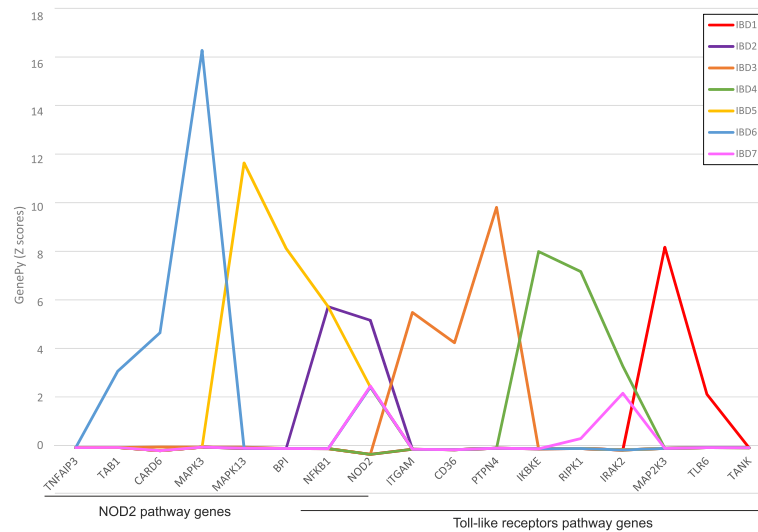


Fig. 3 GenePy score profiles for seven independent patients diagnosed with IBD across selected genes from the NOD2 and TLR pathways. GenePy scores shown were implemented using the M-CAP deleteriousness (D) metric. To facilitate plotting, raw GenePy scores were transformed to Z-scores for each gene. Different colours depict individual patient profiles. Despite being diagnosed with the same disease, all individuals exhibit distinctive profiles across key genes implicated in key immune pathways. Some individuals have evidence of gene pathogenicity within the same pathway (e.g. IBD5 and IBD6) this is conferred through accumulated mutation in different genes – IBD6 has elevated gene-level scores for TAB1, CARD6 and MAPK3 while IBD5 may have impaired function in this pathway due to combined mutation in MAPK13, BP1 and NFKB1. Similarly, IBD1, IBD3 and IBD4 exhibit pathogenic profiles in TLR pathway genes only. These individual level data can be combined with disease phenotype, severity and treatment outcome data in machine learning models to better stratify patient cohorts and realise the promise of personalised medicine

groups was statistically weaker ($p = 0.0346$) and less robust to correction for multiple testing.

Although not the purpose of this comparison, we confirmed GenePy whole gene comparison provided statistical evidence two orders of magnitude greater than any single variant association result (Additional file 1: Table S1).

GenePy score validation - Parkinson's disease cohort

Of the six genes investigated for different GenePy distributions between the PD cohort ($n = 610$) and the non-PD ($n = 465$) cohort, statistically significant results were observed for the *PINK1* gene only ($p = 0.013$) (Table 4). The SKAT-O test did not detect significant associations for any of the six genes.

Restricting the analysis to just the extreme right tail of the GenePy distribution for each of the six PD genes, statistically significant differences were observed between PD and non-PD individuals for *LRRK2* ($p = 0.002$), *PINK1* ($p = 0.010$), *PRKN* ($p = 0.021$) and *VPS35* ($p = 0.036$). Patients with severe *PINK1* and *PRKN* mutations present early onset forms of Parkinson's disease and have been reported in this PD cohort [60]. The most significant result for each gene from traditional single variant association tests reported significant results for two genes only - *LRRK2* (rs10878245, $p = 0.034$) and *PINK1* (rs148871409,

$p = 0.042$) although this required the analysis of multiple SNVs (see Table 4) within each gene.

GenePy score validation - primary open angle glaucoma (POAG) cohort

Comparison of GenePy scores between the POAG cohort ($n = 358$) and the non-POAG cohort ($n = 465$) did not reveal a statistically significant difference for the *MYOC* gene ($p = 0.18$). Similarly, significance was not detected using SKAT-O methodology ($p = 0.66$).

However, performing a Mann-Whitney U test of GenePy scores between the extreme end of the right tail of the GenePy distribution (this time limited to 3% to reflect the known biology) of the POAG cohort and the top 3% of the non-POAG cohort, we observed a statistically significant difference ($p = 0.048$).

In a single variant association test framework, 18 SNVs within the *MYOC* gene were tested for association and only one (rs61730974) reached statistical significance without correcting for multiple testing ($p = 0.0318$).

Discussion

Next generation sequencing is a disruptive technology set to transform biological assessment. Globally, it is rapidly integrating into the medical sector with numerous countries already funding whole genome sequencing of patient

Table 3 NOD2 GenePy score statistics (maxima and means) and Mann-Whitney U tests across groups for all sixteen deleteriousness metrics. *p*-values smaller than 1×10^{-2} or smaller than 5×10^{-2} are highlighted by two (**) or one (*) asterisks respectively. SKAT-O gene association results comparing patient groups against controls provided below thick line

Metric	Controls (<i>n</i> = 166)		IBD (<i>n</i> = 282)			UC (<i>n</i> = 100)			CD (<i>n</i> = 172)		
	max	mean	max	mean	Mann-Whitney U comparison against controls	max	mean	Mann-Whitney U comparison against controls	max	mean	Mann-Whitney U comparison against controls
CADD	2.71	0.28	3.52	0.40	1.04×10^{-1}	2.66	0.20	1.38×10^{-1}	3.52	0.54	4.62×10^{-4} **
DANN	5.92	0.84	7.62	1.06	1.36×10^{-1}	5.62	0.57	1.22×10^{-1}	7.62	1.38	8.16×10^{-4} **
FATHMM	3.33	0.49	4.34	0.66	1.04×10^{-1}	3.14	0.38	1.47×10^{-1}	4.34	0.84	4.84×10^{-4} **
fathmm-MKL	4.53	0.37	6.24	0.55	4.54×10^{-2} *	3.78	0.25	3.15×10^{-1}	6.24	0.76	1.79×10^{-4} **
GERP++_RS	5.30	0.64	7.00	0.87	1.26×10^{-1}	4.95	0.42	1.27×10^{-1}	7.00	1.17	6.95×10^{-4} **
M-CAP	1.87	0.12	3.39	0.22	1.58×10^{-2} *	1.73	0.08	4.62×10^{-1}	3.39	0.32	1.37×10^{-4} **
MetaLR	2.42	0.16	3.39	0.29	2.71×10^{-1}	1.81	0.10	2.34×10^{-2} *	3.39	0.42	1.63×10^{-3} **
MetaSVM	2.67	0.30	3.61	0.43	9.88×10^{-2}	2.50	0.22	1.50×10^{-1}	3.61	0.57	4.39×10^{-4} **
MutationTaster	4.38	0.26	5.10	0.39	4.48×10^{-2} *	2.65	0.13	4.37×10^{-1}	5.10	0.57	7.47×10^{-4} **
phastCons	4.66	0.35	5.24	0.56	2.86×10^{-1}	3.54	0.24	2.70×10^{-2} *	5.24	0.77	2.16×10^{-3} **
phyloP	6.32	1.02	7.93	1.27	1.23×10^{-1}	5.92	0.75	1.38×10^{-1}	7.93	1.62	7.09×10^{-4} **
Polyphen2_HDIV	5.32	0.68	7.03	0.82	2.02×10^{-1}	2.30	0.33	6.22×10^{-2}	7.03	1.13	1.20×10^{-3} **
Polyphen2_HVAR	4.86	0.46	5.31	0.64	1.65×10^{-1}	2.07	0.21	7.22×10^{-2}	5.31	0.92	7.90×10^{-4} **
PROVEAN	4.33	0.66	5.23	0.86	1.04×10^{-1}	4.08	0.49	1.45×10^{-1}	5.23	1.10	4.84×10^{-4} **
SIFT	5.91	0.95	7.61	1.14	1.47×10^{-1}	5.43	0.64	1.16×10^{-1}	7.61	1.47	9.64×10^{-4} **
VEST3	3.28	0.30	4.21	0.44	1.36×10^{-1}	2.24	0.17	1.13×10^{-1}	4.21	0.62	7.48×10^{-4} **
SKAT-O (all variants)	–	–			5.41×10^{-1}			9.76×10^{-2}			3.46×10^{-2} *
SKAT-O (MAF < 0.05)	–	–			4.63×10^{-1}			1.37×10^{-1}			5.02×10^{-2}

samples for diagnosis and treatment of rare disease and cancer. Multiple metrics have emerged that aim to annotate individual mutations with a view to sensitively implicating causal versus non-causal variation. However, for common complex diseases where the action of an unknown number of multiple variants converge to increase susceptibility, the molecular assessment of mutation profiles is necessarily less binary. Furthermore, in order to bring interpretation from bench to bedside, it is important that methodology provides discriminatory evidence for individual patients and not just evidence of modest genetic effects between large cohorts.

We describe the implementation of GenePy representing a novel alternative to examine genomic data that provides a quantitative measure of the combined loading of mutation across each gene for each individual. The scoring system has the freedom to harness the intrinsic properties of any user-defined variant-level deleteriousness metric. By summing across genes, GenePy further integrates biological information on frequency and zygosity and when being used to examine between genes or subsets thereof, should be corrected for gene length.

Different measures of deleteriousness impact the coefficient of variation in the GenePy scoring system but as yet none are proven superior. The logarithmic distribution

Table 4 Comparison of PD versus non-PD individuals. Significant results are shown in bold type. For each gene the most significant result only of all SNV association tests is shown and for each these the rs id is provided. Additionally, the number of SNV association test conducted within each gene is indicated in brackets. No correction is made for testing of six genes nor for testing multiple SNVs within any given gene

Test PD vs non-affected samples	<i>LRRK2</i>	<i>PARK7</i>	<i>PINK1</i>	<i>PRKN</i>	<i>SNCA</i>	<i>VPS35</i>
GenePy	0.178	0.445	0.013	0.983	0.828	0.206
SKAT-O	1	0.557	0.157	0.427	0.712	0.741
Top 5% comparison	0.002	0.107	0.010	0.021	0.347	0.036
Most significant SNV (# tested)	0.034 rs10878245 (88)	0.081 rs71653621 (6)	0.042 rs148871409 (21)	0.051 rs1801582 (27)	0.433 rs548523899 (7)	0.433 rs168745 (17)

confers weight to rare pathogenic variants and these are additive across a gene and theoretically limited only by the number of variant sites within that gene. GenePy returns a score of zero for the majority of genes for any one individual - this reflects the sparse nature of genomic data and is exacerbated when considering whole exome sequencing data where historical negative selection has limited variation in regions that code for proteins.

We provide proof of principle that testing GenePy scores with a non-parametric statistical test improves sensitivity to detect clinically meaningful gene perturbations. Such performance compares favourably against the most commonly applied gene based association test optimised for small data sets (SKAT-O). Superiority to detect the subtle effects of genes in complex disease is likely attributable to the additional modelling of innate biological features of mutations.

Power to determine significant GenePy score differences between IBD patient and control groups was consistent across sixteen different metrics of variant deleteriousness whereby all concordantly reported a similar level of significance despite differing underlying principles. It is noteworthy that the M-CAP deleteriousness metric that enriches for very deleterious, rare variants proved most significant in our specific test case (although this metric annotated fewer genes than other deleteriousness metrics). This result may suggest a more important role for rare variants in the *NOD2* gene that went largely undetected through GWAS studies. Recent publications have similarly evidenced an important role for rare variants in select patients with IBD [61–64]. While GenePy scores generated using M-CAP metric returned the most significant difference in CD patients compared to controls, it is likely that no metric will prove optimal in all situations. The GenePy scoring system can simply accommodate new and improved variant deleteriousness metrics that are constantly evolving with more widespread use and interpretation of NGS data.

We demonstrated the ability of GenePy to model biological variability from next generation sequencing data on two additional common complex disorders, showing its simple implementation and flexible application to different scenarios. In a Parkinson's Disease (PD) cohort of very modest sample size compared to contemporary GWAS studies, GenePy successfully identified association with the *PINK1* gene but failed to reach significance for five other known genes when looking across the entire distribution of scores. SKAT-O did not return significant associations with any of the six genes. Interestingly, restricting the analysis to the extreme distribution scores in the case/control comparison framework, GenePy did detect association for four of the six PD genes. This compares well against the SNV association tests within these known genes where only two genes

(*LRRK2* and *PINK1*) harboured SNVs that achieved nominal significance without correction for the additional tests incurred by such an approach.

When testing GenePy performances against SKAT-O within the glaucoma cohort, neither SKAT-O or comparison of the entire GenePy distribution between cases and controls could discriminate significant differences between the POAG and non-POAG groups. However, by restricting the analysis to the extreme tail of the distribution, GenePy was able to determine a statistical difference presumably driven by only a minority of patients in whom disease is mediated by the *MYOC* gene.

In addition to identifying genes harbouring statistically significant different mutational loadings between case and control groups, selecting samples from the extreme distribution of GenePy scores concurrently identifies the specific individuals whose disease is (partially) explained by these genes and so facilitates clinical translation.

As with all large-scale data, GenePy scoring is dependent upon data integrity and elimination of systematic bias or technical artefacts. High quality individual DNA samples must be sequenced to sufficient depth to return confident variant calls. For larger scale analyses using multiple samples, parity of capture kits, sequencing platforms and informatic pipelines must be ensured. While these pre-processing quality control steps and generation of the multi-calling VCF file represent the highest computational burden, GenePy score calculation on cleaned vcf files is amenable to batching and computationally trivial.

Many of the currently available deleteriousness scores implemented herein fail to annotate synonymous, splicing or protein truncating variation. While we arbitrarily imposed maximum deleteriousness scores to protein truncating mutations, we standardised the set of variants examined across metrics by excluding synonymous and splicing variants from this analysis. Deleteriousness metrics based on conservation alone are calculable for all genomic variation and could be implemented for the assessment sliding windows of non-coding regions derived from whole genome sequencing. Due to association testing in Caucasian samples only, we restricted allele frequency annotation to that ethnic group. Arguably, there is merit in implementation of global allele frequency estimates or those from more ancestrally diverse populations.

Further refinements of the GenePy scoring system might be realised by integration of gene essentiality [65] (and conversely gene redundancy) or gene damage indices (GDI) [66]. Long read NGS data enabling the discrimination of gametic phase would substantially advantage integration of inheritance models and haploinsufficiency.

Conclusions

The key advantage of GenePy is its provision of a continuous quantitative measure of biological integrity of a gene within individuals, resulting in a score that is easily integrated into downstream analyses. GenePy scores are not dependent on cohort size and can be calculated and assessed on per-patient basis. GenePy scores are suited to pathway analyses where scores can be overlaid and summed across defined molecular cascades. This enables users to assess the combinatorial effect of variants in multiple genes involved in complex diseases. For the particular assessment of complex disease, machine learning tools that integrate multi-omic and extensive biomarker ‘big data’ to determine cryptic patterns are increasingly applied. Currently, all machine learning applications are obliged to incorporate genetic data derived from NGS analyses on a variant-by-variant basis and most do so in either a binary (present/absent) manner or through counting for allelic load (0, 1 or 2) [67]. Both approaches ignore much of the additional biological information already available. Furthermore, these methods often impose arbitrary and subjective filters or thresholds for the inclusion of variants (e.g. frequency) that may be incorrect for Mendelian disease and will certainly reduce power for complex disease. GenePy reduces the dimensionality of genomic data from multiple SNVs within a single gene to the resolution of a single gene. This reduces the number of tests to be performed and impacts statistical power in small cohort studies. GenePy facilitates integration with other ‘omics data that also reports at the level and resolution of a gene e.g. transcriptomic, metabolomic proteomic data and so facilitates integration across these contemporary ‘omic approaches in a machine learning and network analysis frameworks. Furthermore, the assessment of individual gene pathogenicity loadings for individual subjects is simple and intuitive in a clinical setting and allows clustering of independent patients each with cumulatively deleterious burden of mutations in a given gene – even when no specific variants are shared between patients – a situation common for sparse genomic data.

Machine learning approaches aim to define patient subgroups on a molecular genetic basis for the advancement of personalised treatment. Such approaches will directly benefit from the refined scores provided by GenePy for the stratification of different patient subgroups. The ability to input biologically rich information and the gene and individual level represents an important step change from the more traditional methods of assessing genetic data at the variant and cohort level.

Availability and requirements Project name: GenePy.

Project home page: <https://github.com/UoS-HGIG/GenePy>

Operating system(s): Unix.

Programming language: Bash, Python 2.7.

Other requirements: GATK 3.x, Annovar.

License: GNU GPL.

Any restrictions to use by non-academics: no licence needed.

Additional file

Additional file 1: Table S1. All single nucleotide variants in the *NOD2* gene used in GenePy validation. **Figure S1.** Median whole gene GenePy_{uncorrected} score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics. **Figure S2.** Median whole gene GenePy_{cgl} score profiles observed across the cohort of 508 patients with WES data depicted separately for each of the sixteen deleteriousness metrics. **Figure S3.** Ethnicity imputation. **Figure S4.** GenePy scores profiles for the *NOD2* gene in the CD and control groups for each of the sixteen implemented deleteriousness metrics. (DOCX 1054 kb)

Abbreviations

CD: Crohn's disease; CGL: Corrected by gene length; CV: Coefficient of variation; GATK: Genome Analysis Toolkit; GWAS: Genome-wide association studies; IBD: Inflammatory bowel disease; IBDU: IBD undetermined; NGS: Next-generation sequencing; PD: Parkinson's disease; POAG: Primary open angle glaucoma; SKAT-O: Sequence kernel association optimal unified test; SNV: Single nucleotide variant; UC: Ulcerative colitis; VCF: Variant calling format; WES: Whole exome sequencing

Acknowledgments

The authors would like to thank Rachel Haggarty for assistance with management of the genetics of PIBD study database. We also would like to acknowledge Nikki Graham for assistance with sample extraction and management. We thank the EUCLIDS consortium, for providing access to anonymised exome data used for comparison and development of our model.

Data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI – a public-private partnership – is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Allergan, Avid, Biogen, BioLegend, Bristol-Myers Squibb, Celgene, Denali, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail, Roche, Sanofi, Servier, Takeda, Teva, Ucb, Verily and Voyager.

The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Funding

This project is supported by the National Institute for Health Research through the NIHR Southampton Biomedical Research Centre; the Hilary Marsden Institute for Life Science Scholarship and the Crohn's in Childhood Research Association. This publication has included data from a project that has received funding from the European Union's seventh Framework program under EC-GA no. 27985 (EUCLIDS). The funding body did not played any roles in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

GenePy algorithm and implementation is available at <https://github.com/UoS-HGIG/GenePy>

Authors' contributions

SE and BDM conceived and designed the study. SE and RMB led the recruitment to the study. EM implemented the algorithm, managed data, performed bioinformatics analyses and wrote the manuscript. LO processed raw glaucoma data. BDM contributed to the mathematical modelling. SE

provided expertise on genomics and data integration. SE contributed substantially to the final version of the manuscript. JJA and RJP advised on model development and manuscript preparation. All authors read and approved the final manuscript.

Ethics approval and consent to participate

The study has ethical approval from Southampton & South West Hampshire Research Ethics Committee (09/H0504/125). Written informed consent was provided by an attending parent or legal guardian for paediatric participants.

Consent for publication

Written consent for publication was provided by the attending parent or legal guardian for paediatric participants.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

¹Department of Human Genetics and Genomic Medicine, University of Southampton, Southampton, UK. ²Institute for Life Sciences, University of Southampton, Southampton, UK. ³Department of Paediatric Gastroenterology, Southampton Children's Hospital, Southampton, UK.

Received: 14 September 2018 Accepted: 6 May 2019

Published online: 16 May 2019

References

- Trujillo D, Bertoli-Avella AM, Kumar Kandaswamy K, Weiss ME, Köster J, Marais A, et al. Clinical exome sequencing: results from 2819 samples reflecting 1000 families. *Eur J Hum Genet*. 2017;25:176–82. <https://doi.org/10.1038/ejhg.2016.146>.
- Shen T, Lee A, Shen C, Lin CJ. The long tail and rare disease research: the impact of next-generation sequencing for rare Mendelian disorders. *Genet Res (Camb)*. 2015;97:e15. <https://doi.org/10.1017/S0016672315000166>.
- Jamuar SS, Tan E-C. Clinical application of next-generation sequencing for Mendelian diseases. *Hum Genomics*. 2015;9:10. <https://doi.org/10.1186/s40246-015-0031-5>.
- Gillissen C, Hoischen A, Brunner HG, Veltman JA. Disease gene identification strategies for exome sequencing. *Eur J Hum Genet*. 2012;20:490–7. <https://doi.org/10.1038/ejhg.2011.258>.
- Cooper GM, Stone EA, Asimenos G, NISC Comparative Sequencing Program ED, Green ED, Batzoglou S, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15:901–13. <https://doi.org/10.1101/gr.3577405>.
- Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034–50. <https://doi.org/10.1101/gr.3715005>.
- Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20:110–21. <https://doi.org/10.1101/gr.097857.109>.
- Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*. 2012;40 Web Server issue:W452–7. <https://doi.org/10.1093/nar/gks539>.
- Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GLA, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat*. 2013;34: 57–65. <https://doi.org/10.1002/humu.22225>.
- Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day INM, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics*. 2015;31:1536–43. <https://doi.org/10.1093/bioinformatics/btv009>.
- Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–9. <https://doi.org/10.1038/nmeth0410-248>.
- Schwarz JM, Cooper DN, Schuelke M, Seelow D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods*. 2014;11:361–2. <https://doi.org/10.1038/nmeth.2890>.
- Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. *PLoS One*. 2012;7:e46688. <https://doi.org/10.1371/journal.pone.0046688>.
- Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics*. 2013;14(Suppl 3):S3. <https://doi.org/10.1186/1471-2164-14-S3-S3>.
- Butkiewicz M, Bush WS. In Silico Functional Annotation of Genomic Variation. *Curr Protoc Hum Genet*. 2016;88:Unit 6.15. doi:<https://doi.org/10.1002/0471142905.hg0615s88>.
- Chun S, Fay JC. Identification of deleterious mutations within three human genomes. *Genome Res*. 2009;19:1553–61. <https://doi.org/10.1101/gr.092619.109>.
- Tang H, Thomas PD. Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation. *Genetics*. 2016;203:635–47. <https://doi.org/10.1534/genetics.116.190033>.
- Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46:310–5. <https://doi.org/10.1038/ng.2892>.
- Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24:2125–37. <https://doi.org/10.1093/hmg/ddu733>.
- Jagadeesh KA, Wenger AM, Berger MJ, Guturu H, Stenson PD, Cooper DN, et al. M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. *Nat Genet*. 2016;48:1581–6. <https://doi.org/10.1038/ng.3703>.
- Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet*. 2016;48:214–20. <https://doi.org/10.1038/ng.3477>.
- Schubach M, Re M, Robinson PN, Valentini G. Imbalance-Aware Machine Learning for Predicting Rare and Common Disease-Associated Non-Coding Variants. *Sci Rep*. 2017;7:2959. <https://doi.org/10.1038/s41598-017-03011-5>.
- Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics*. 2015;31:761–3. <https://doi.org/10.1093/bioinformatics/btu703>.
- Mahmood K, Jung C-H, Philip G, Georgeson P, Chung J, Pope BJ, et al. Variant effect prediction tools assessed using independent, functional assay-based datasets: implications for discovery and diagnostics. *Hum Genomics*. 2017;11:10. <https://doi.org/10.1186/s40246-017-0104-8>.
- Li J, Shi L, Zhang K, Zhang Y, Hu S, Zhao T, et al. VarCards: an integrated genetic and clinical database for coding variants in the human genome. *Nucleic Acids Res*. 2018;46:D1039–48. <https://doi.org/10.1093/nar/gkx1039>.
- Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet*. 2010;11:446–50. <https://doi.org/10.1038/nrg2809>.
- Schork NJ. Personalized medicine: time for one-person trials. *Nature*. 2015; 520:609–11. <https://doi.org/10.1038/520609a>.
- Li B, Leal SM. Methods for Detecting Associations with Rare Variants for Common Diseases: Application to Analysis of Sequence Data. *Am J Hum Genet*. 2008;83:311–21. <https://doi.org/10.1016/j.ajhg.2008.06.024>.
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, et al. Testing for an unusual distribution of rare variants. *PLoS Genet*. 2011;7: e1001322. <https://doi.org/10.1371/journal.pgen.1001322>.
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet*. 2012;91:224–37. <https://doi.org/10.1016/j.ajhg.2012.06.007>.
- Takahashi S, Andreoletti G, Chen R, Muneira Y, Batra A, Afzal NA, et al. De novo and rare mutations in the HSPA1L heat shock gene associated with inflammatory bowel disease. *Genome Med*. 2017;9:8. <https://doi.org/10.1186/s13073-016-0394-9>.
- Tan L, Li Z, Zhou C, Cao Y, Zhang L, Li X, et al. FBN1 mutations largely contribute to sporadic non-syndromic aortic dissection. *Hum Mol Genet*. 2017;26:4814–22. <https://doi.org/10.1093/hmg/ddx360>.
- Ruiz-Pinto S, Pita G, Patiño-García A, Alonso J, Pérez-Martínez A, Cartón AJ, et al. Exome array analysis identifies GPR35 as a novel susceptibility gene for anthracycline-induced cardiotoxicity in childhood cancer. *Pharmacogenet Genomics*. 2017;27:445–53. <https://doi.org/10.1097/FPC.0000000000000309>.

34. Robak LA, Jansen IE, van Rooij J, Uitterlinden AG, Kraaij R, Jankovic J, et al. Excessive burden of lysosomal storage disorder gene variants in Parkinson's disease. *Brain*. 2017;140:3191–203. <https://doi.org/10.1093/brain/awx285>.
35. Wang H, Cade BE, Chen H, Gleason KJ, Saxena R, Feng T, et al. Variants in angiopoietin-2 (ANGPT2) contribute to variation in nocturnal oxyhaemoglobin saturation level. *Hum Mol Genet*. 2016;25:ddw324. <https://doi.org/10.1093/hmg/ddw324>.
36. Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. *Sci Rep*. 2017;7:2427. <https://doi.org/10.1038/s41598-017-02606-2>.
37. Levine A, Koletzko S, Turner D, Escher JC, Cucchiara S, de Ridder L, et al. ESPGHAN revised Porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. *J Pediatr Gastroenterol Nutr*. 2014;58:795–806. <https://doi.org/10.1097/MPG.0000000000000239>.
38. Jun G, Flickinger M, Hetrick KN, Romm JM, Doherty KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and Array-based genotype data. *Am J Hum Genet*. 2012;91:839–48. <https://doi.org/10.1016/j.ajhg.2012.09.004>.
39. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *Genomics*. 2013; <http://arxiv.org/abs/1303.3997>. Accessed 3 Apr 2017.
40. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernysky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20:1297–303. <https://doi.org/10.1101/gr.107524.110>.
41. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–8. <https://doi.org/10.1038/ng.806>.
42. Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016;536:285–91. <https://doi.org/10.1038/nature19057>.
43. Flicek P, Amodé M, Barrell D. Ensembl 2012. In: *Nucleic acids*; 2012.
44. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>.
45. Horita N, Kaneko T. Genetic model selection for a case-control study and a meta-analysis. *Meta gene*. 2015;5:1–8. <https://doi.org/10.1016/j.mgene.2015.04.003>.
46. Marian AJ. Molecular genetic studies of complex phenotypes. *Transl Res*. 2012;159:64–79. <https://doi.org/10.1016/j.TRS.2011.08.001>.
47. Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, et al. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nat Med*. 2015;21:1018–27. <https://doi.org/10.1038/nm.3933>.
48. de Lange KM, Moutsianas L, Lee JC, Lamb CA, Luo Y, Kennedy NA, et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat Genet*. 2017;49:256–61. <https://doi.org/10.1038/ng.3760>.
49. Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cézard J-P, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*. 2001;411:599–603. <https://doi.org/10.1038/35079107>.
50. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491:56–65. <https://doi.org/10.1038/nature11632>.
51. Pedersen BS, Quinlan AR. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with Puddy. *Am J Hum Genet*. 2017;100:406–13. <https://doi.org/10.1016/j.ajhg.2017.01.017>.
52. Ajay SS, Parker SCJ, Abaan HO, Fajardo KV, Margulies EH. Accurate and comprehensive sequencing of personal genomes. *Genome Res*. 2011;21:1498–505. <https://doi.org/10.1101/gr.123638.111>.
53. Marek K, Chowdhury S, Siderowf A, Lasch S, Coffey CS, Caspell-Garcia C, et al. The Parkinson's progression markers initiative (PPMI) – establishing a PD biomarker cohort. *Ann Clin Transl Neurol*. 2018;5:1460–77. <https://doi.org/10.1002/acn3.644>.
54. Farrer MJ. Genetics of Parkinson disease: paradigm shifts and future prospects. *Nat Rev Genet*. 2006;7:306–18. <https://doi.org/10.1038/ng1831>.
55. Norman CS, O'Gorman L, Gibson J, Pengelly RJ, Baralle D, Ratnayaka JA, et al. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (TYR) causing hypomorphic oculocutaneous albinism (OCA1B). *Sci Rep*. 2017;7:4415. <https://doi.org/10.1038/s41598-017-04401-5>.
56. Weinreb RN, Khaw PT. Primary open-angle glaucoma. *Lancet*. 2004;363:1711–20. [https://doi.org/10.1016/S0140-6736\(04\)16257-0](https://doi.org/10.1016/S0140-6736(04)16257-0).
57. Liu Y, Allingham RR. Major review: molecular genetics of primary open-angle glaucoma. *Exp Eye Res*. 2017;160:62–84. <https://doi.org/10.1016/j.exer.2017.05.002>.
58. Fingert JH, Stone EM, Sheffield VC, Alward WL. Myocilin Glaucoma. *Surv Ophthalmol*. 2002;47:547–61. [https://doi.org/10.1016/S0039-6257\(02\)00353-3](https://doi.org/10.1016/S0039-6257(02)00353-3).
59. O'Gorman L, Cree AJ, Ward D, Griffiths HL, Sood R, Denniston AK, et al. Comprehensive sequencing of the myocilin gene in a selected cohort of severe primary open-angle glaucoma patients. *Sci Rep*. 2019;9:3100. <https://doi.org/10.1038/s41598-019-38760-y>.
60. McWilliams TG, Barini E, Pohjola-Pirhonen R, Brooks SP, Singh F, Burel S, et al. Phosphorylation of Parkin at serine 65 is essential for its activation *in vivo*. *Open Biol*. 2018;8:180108. <https://doi.org/10.1098/rsob.180108>.
61. Cho JH, Abraham C. Inflammatory bowel disease genetics: Nod2. *Annu Rev Med*. 2007;58:401–16. <https://doi.org/10.1146/annurev.med.58.061705.145024>.
62. Rivas MA, Beaudoin M, Gardet A, Stevens C, Sharma Y, Zhang CK, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat Genet*. 2011;43:1066–73. <https://doi.org/10.1038/ng.952>.
63. Frade-Proudhon-Clerc S, Smol T, Frenois F, Sand O, Vaillant E, Dhennin V, et al. A Novel Rare Missense Variation of the NOD2 Gene: Evidences of Implication in Crohn's Disease. *Int J Mol Sci*. 2019;20:835. <https://doi.org/10.3390/ijms20040835>.
64. Girardelli M, Loganés C, Pin A, Stacul E, Decleva E, Vozzi D, et al. Novel NOD2 Mutation in Early-Onset Inflammatory Bowel Phenotype. *Inflamm Bowel Dis*. 2018;24:1204–12. <https://doi.org/10.1093/ibd/izy061>.
65. Pengelly RJ, Vergara-Lope A, Alyousfi A, Jabalameli MR, Collins A. Understanding the disease genome: gene essentiality and the interplay of selection, recombination and mutation. *Brief Bioinform*. 2019;20(1):267–3. <https://doi.org/10.1093/bib/bbx110>.
66. Itan Y, Shang L, Boisson B, Patin E, Bolze A, Moncada-Vélez M, et al. The human gene damage index as a gene-level approach to prioritizing exome variants. *Proc Natl Acad Sci*. 2015;112:13615–20. <https://doi.org/10.1073/pnas.1518646112>.
67. Daneshjoui R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. *Hum Mutat*. 2017;38:1182–92. <https://doi.org/10.1002/humu.23280>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



A small gene sequencing panel realises a high diagnostic rate in patients with congenital nystagmus following basic phenotyping

¹Luke O’Gorman, ²Chelsea S. Norman, ³Luke Michaels, ²Tutte Newall, ⁴Andrew H. Crosby,
^{1,5}Christopher Mattocks, ²Angela J. Cree, ^{2,3}Andrew J. Lotery, ⁶Emma L. Baple, ²J. Arjuna Ratnayaka,
¹Diana Baralle, ³Helena Lee, ²Daniel Osborne, ³Fatima Shawkat, ⁷Jane Gibson, ⁸Sarah Ennis⁺, ^{2,3}Jay E
Self⁺⁺

¹Human Development and Health, Faculty of Medicine, University of Southampton, MP808,
Tremona Road, Southampton SO16 6YD, UK

²Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, MP806,
Tremona Road, Southampton, SO16 6YD, UK

³Eye Unit, University Hospital Southampton NHS Foundation Trust, Tremona Road, Southampton,
SO16 6YD, UK

⁴Institute of Biomedical and Clinical Science, University of Exeter Medical School, RILD Wellcome
Wolfson Centre, Exeter, EX2 5DW, UK

⁵Wessex Investigational Science Hub, University Hospital Southampton, Tremona Road,
Southampton SO16 6YD, UK

⁶Medical Research (Level 4), University of Exeter Medical School, RILD Wellcome Wolfson Centre,
Royal Devon and Exeter NHS Foundation Trust, Barrack Road, Exeter, EX2 5DW, UK.

⁷Biological Sciences, Faculty of Natural and Environmental Sciences, University of Southampton,
Southampton, SO17 1BJ, UK

⁸Human Genetics & Genomic Medicine, Faculty of Medicine, University of Southampton, MP 808,
Tremona Road, Southampton SO16 6YD, UK

⁺Joint last authorship. These authors contributed equally to this work.

* Corresponding Author:

Mr Jay Self BM FRCOphth PhD,

Clinical and Experimental Sciences, Faculty of Medicine, University of Southampton, Mail point 806,
Southampton, SO16 6YD, United Kingdom,

Tel: 02381 205049

(J.E.Self@soton.ac.uk)

Abstract

Nystagmus is a disorder of uncontrolled eye movement and can occur as an isolated trait (idiopathic INS, IINS) or as part of multisystem disorders such as albinism, significant visual disorders or neurological disease. Eighty-one unrelated patients with nystagmus underwent routine ocular phenotyping using commonly available phenotyping methods and were grouped into four sub-cohorts according to the level of phenotyping information gained and their findings. DNA was extracted and sequenced using a broad utility next generation sequencing (NGS) gene panel. A clinical subpanel of genes for nystagmus/albinism was utilised and likely causal variants were prioritised according to methods currently employed by clinical diagnostic laboratories.

We determine the likely underlying genetic cause for 43.2% of participants with similar yields regardless of prior phenotyping. This study demonstrates that a diagnostic workflow combining basic ocular phenotyping and a clinically available targeted NGS panel, can provide a high diagnostic yield for patients with infantile nystagmus, enabling access to disease specific management at a young age and reducing the need for multiple costly, often invasive tests. By describing diagnostic yield for groups of patients with incomplete phenotyping data, it also permits the subsequent design of ‘real-world’ diagnostic workflows and illustrates the changing role of genetic testing in modern diagnostic workflows for heterogeneous ophthalmic disorders.

Introduction

Infantile nystagmus syndrome (INS) is a condition which can be present as an isolated trait (idiopathic INS, IINS) or as part of a plethora of ocular or systemic disorders including albinism, retinal disease and neurological disorders. IINS is most commonly seen either in singletons, or in X-linked pedigrees. Tarpey *et al*¹ reported that 57% of putative X-linked pedigrees and 94% of proven X-linked pedigrees harbour causal mutations in the FERM domain containing 7 gene, *FRMD7*¹.

Although *FRMD7* mutations are the only known genetic cause of IINS, many disorders are known to masquerade as IINS in children. These conditions are often missed due to the difficulty in identifying associated phenotypes in children (such as hypomorphic albinism²) or a delay in the onset of additional clinical features (such as spino-cerebellar ataxia type 6³).

Ocular albinism (OA) is a form of albinism in which the clinical features are constrained to the eye, whilst oculocutaneous albinism (OCA) encompasses a broader phenotypic range affecting the eyes, hair and skin⁴. The ocular manifestations of OCA and OA include infantile nystagmus syndrome (INS)⁵, foveal hypoplasia, abnormal crossing pattern at the optic chiasm and iris transillumination defects⁶ all of which can be subtle or incomplete⁷. Additionally, many of the ocular features seen in albinism can be seen in other disorders caused by mutations in genes that are not associated with melanin biosynthesis such as *PAX6* mutations, which can cause a variety of ocular phenotypes including nystagmus and foveal hypoplasia⁸. Similarly, Chediak-Higashi syndrome and Hermansky-Pudlak syndrome, caused by the *LYST* and *HPS* genes respectively, involve many OA and OCA phenotypic traits. Despite the significant systemic health implications of these forms of syndromic albinism, most patients never undergo genetic testing.

Genes involved in the melanin biosynthesis pathway are known to cause forms of both OA and OCA. Examples include *GPR143*, which is causal for OA1⁹, whilst *TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5* and *C10orf11* are associated with OCA subtypes 1–4 and 6–7 respectively⁴. The OCA1 gene, *TYR*, is known to be associated with missing heritability¹⁰¹¹. Our group and others have previously reported a compound heterozygous tri-allelic genotype in *TYR* which involves both rare

(AF < 5%) and common (AF 28–36%) functionally damaging variants which are likely to be on *trans* alleles^{2,12,13}. The two common *TYR* variants p.S192Y and p.R402Q, have previously been shown to cause a 40% reduction in tyrosinase activity and protein misfolding, respectively^{14,15}.

Pigmentary abnormalities in hair and skin may be apparent during phenotyping, although this is not always the case. Consequently, a range of potential diagnoses could be made in children with nystagmus, particularly between the ages of 4-6 months¹⁰.

As phenotyping becomes more precise and nuanced in children with nystagmus, it is possible that candidate gene lists can become more specific¹⁶. For example, an abnormal electroretinogram (ERG) can be the only indication that an underlying retinal dystrophy is the cause of the nystagmus. Hence, a retinal gene panel might be the most appropriate genetic testing option. It also allows potential candidate causal variant(s) to be interpreted with greater confidence. Previous studies of next-generation sequencing in INS patients have utilised large gene panels of up to 300 genes whilst identifying candidate causal variants in a recurrent, small subset of genes^{17–19}, or in genes for conditions which would have been identified by basic phenotyping. This suggests that pre-selecting and interpreting variants in fewer genes for phenotyped patients, may provide the most efficient workflow and highest diagnostic yield in routine clinical practice.

Aim

We describe the diagnostic yield of a clinically available 31 gene panel for nystagmus and albinism. We evaluate its clinical utility in both completely phenotyped and incompletely phenotyped patients in order to reflect the real-world limitations of phenotyping in the clinic and describe its use for patients presenting to specialist and non-specialist centres.

Methods

Ethics and consent

Consent was obtained in accordance with the Declaration of Helsinki and was approved by South West Hampshire Local Research Ethics Committee (LREC 028/04/t). All methods were carried out in

accordance with relevant guidelines and regulations. Informed consent was obtained from all subjects and, if subjects were under 18, from a parent and/or legal guardian.

Patients

Eighty-one individuals (age range 0-18 yrs) were identified from a regional paediatric nystagmus clinic as having INS with or without clinical features suggesting albinism. All patients referred to a single, regional service were offered recruitment.

Phenotyping

All patients underwent basic phenotyping as outlined in Norman *et al* ². Briefly, this included history taking, orthoptic examination, age-appropriate visual acuity testing, anterior and posterior segment examinations and ERG. Visual evoked potential (VEP, using flash VEP for younger children and pattern onset for older patients) and optical coherence tomography (OCT) with the Leica OCT system or Spectralis OCT (Heidelberg Engineering) were performed in most patients. Eye movements were recorded in some subjects with the EYELink1000 + (SR research, Ottawa, Ontario, Canada) eye tracker. Patients with a diagnosis of a condition known to cause nystagmus (such as Down syndrome, congenital cataract or Aniridia) or where a specific diagnosis was strongly suspected (such as a cone disorder in a photophobic patient confirmed by ERG) or without an INS phenotype (such as Gaze Evoked Nystagmus, GEN due to cerebellar disease) and those who were born before 35/40 weeks gestation were excluded. For included patients, saliva samples (ORAGENE) were collected and DNA extracted using Oragene-DNA kit (OG-575) (DNA Genotek).

Probands were allocated into four phenotype subgroups; clinically IINS with complete phenotyping (group 1), clinically IINS with incomplete phenotyping (group 2), clinical phenotyping consistent with albinism with complete phenotyping (group 3), and clinical features suggestive of albinism with incomplete phenotyping (group 4) (Table 1).

Table 1: Selection criteria of the four, clinically relevant phenotype sub-groups. Equivocal results were those deemed insufficient to permit calling typically due to limited patient compliance or borderline responses for example crossed asymmetry identified on only a few runs of monocular VEP testing (as is commonly the case in patients with hypomorphic albinism).

**Must include at least one of the features highlighted in bold.*

Cohort sub-group number	Cohort sub-group	Predominant waveform direction	ERG	OCT	VEP- misrouting suggested	Iris trans illumination	Number of patients
1	Idiopathic nystagmus (with complete phenotyping)	Horizontal	Normal	Normal	No	No	18 (22.2%)
2	Idiopathic nystagmus (with incomplete phenotyping)	Horizontal or Equivocal	Normal or Equivocal	Normal or Untested or Equivocal	Normal or Untested or Equivocal	Normal or Untested or Equivocal	15 (18.5%)
3	Clinically consistent with Albinism (with complete phenotyping)	Horizontal or Multiplanar	Normal	Foveal hypoplasia	Yes	Yes	20 (24.7%)
4	Clinical features suggestive of Albinism (with	Horizontal or Multiplanar	Normal or Equivocal	Foveal hypoplasia or Untested or	Yes or Untested or Equivocal	Yes or Untested or Equivocal	28 (34.6%)

	incomplete phenotyping)*			Equivocal or Normal	or Normal	or Normal	

Next-generation Sequencing

Eighty-one DNA samples were prepared across six batches using the Illumina TruSight One capture kit (Illumina 5200 Illumina Way San Diego, California USA) which targets 4811 clinically relevant genes. Next-generation sequencing was performed on the NextSeq 500 platform.

Gene Panel

The UKGTN gene panel for ‘albinism and nystagmus’ (31 genes, accessed 29/05/2018) was used to prioritise genes for identification of candidate likely causal variants (see S1 Table).

Bioinformatic Pipeline

FastQ data were aligned to the hg38 human reference genome with BWA-MEM²⁰. GATK v3.7²¹ was used to call SNPs and short indels in a multisample VCF file. Annotation was performed using ANNOVAR v2015Dec²² to collate variant consequence, variant allele frequency (1000 Genomes Project, Exome Sequencing Project and Exome Aggregation Consortium) and pathogenicity scores with CADD²³ and MaxEntScan²⁴ for splice site variants. Further annotation was included from InterVar (2018)²⁵ and Human Gene Mutation Database²⁶. Coverage was determined using SAMtools v1.3.1²⁷ and BEDtools v2.17.0²⁸.

Variant Prioritisation

Variants were prioritised into two categories of ‘assumed pathogenic’ and ‘assumed likely pathogenic’. ‘Assumed pathogenic’ was defined as a variant which had a ‘pathogenic’ annotation in ClinVar, ‘pathogenic’ annotation by InterVar or ‘disease-causing mutation’ (DM) in HGMD. ‘Assumed likely pathogenic’ was defined as a variant which was: (1) not synonymous; (2) had an

allele frequency $\leq 5\%$ in 1000 Genomes Project (all populations), Exome Sequencing Project 6500 (all populations) and Exome Aggregation Consortium (all populations) and; (3) had *either* a CADD Phred ≥ 15 ²³ or a MaxEntScan $\geq |3|$ ²⁹. Variants which form part of a single likely causal genotype were identified as ‘likely causal’ variants whilst multiple possible causal genotypes were identified as ‘reportable likely causal’. Sanger sequencing was performed to verify ‘likely causal’ variants which were miscalled in $>10\%$ of individuals in the cohort. A Sanger sequencing primer pair was designed using ‘A Plasmid Editor (ApE)’ software spanning 19bp (forward) and 20bp (reverse) in length.

Results

Quality Control

The mean read depth across all samples was 127X (see S2 Table) with 95.8% coverage at a depth of 20X or greater across the 31 gene panel.

Assumed Pathogenic Variants

A total of 46 variants across the 81 participants met the criteria for assumed likely pathogenic genetic variants. For 17 patients (21.0% of the cohort), a total of 24 variants were considered to be likely causal variants (Table 2). Likely causal diagnoses were identified in 7 genes (*HPS5*, *PAX6*, *TYR*, *OCA2*, *CACNA1A*, *CACNA1F* and *FRMD7*) from the 31 gene panel. Twenty-two heterozygous variants that were initially labelled as assumed pathogenic or assumed likely pathogenic were found in genes known to cause recessive disorders in patients without a second identified putative variant.

Assumed Pathogenic and Likely Pathogenic Variants

For the remaining 64 patients without a likely causal genotype identified, assumed pathogenic variants together with assumed likely pathogenic variants (n=89 unique variants) were interpreted for likely causality (Table 3).

Individuals with likely causal variants within the *PAX6* gene were attributed to two variants within the same codon, p.Q286. However, although these variants initially presented as variants of interest they had high failure rates (miscalls by GATK haplotype caller in 27.2% and 24.7% respectively of the

total cohort). Subsequent verification with Sanger sequencing excluded these variants as false positives. For this reason, the two *PAX6* variants of the p.Q286 codon and an assumed likely pathogenic *HPS6* variant causing a p.W595G substitution (53.1% failure rate) were omitted from Table 3.

Nine of the 64 patients (11.1% of the total cohort) had likely causal diagnoses from assumed likely pathogenic variants with a cumulative total of 29 unique variants. Likely causal diagnoses were identified across 6 genes (*TYRP1*, *HPS5*, *SACS*, *OCA2*, *CACNA1A* and *FRMD7*) from the 31 gene panel. The remaining 58 heterozygous variants were found in genes known to cause recessive disorders in patients without a second identified putative variant.

Table 2: Seventeen of 81 patients with assumed pathogenic variants were determined to harbour likely causal genotypes.

Samples, blue background indicates male whilst pink indicates female, orange cells indicate heterozygous variants and red cells indicate homozygous and hemizygous variants. Samples are ordered by phenotypic group, and cells with a 'C' denote assignment of a likely causal variant. The TYR variant, NM_000372:exon4:c.G1205A:p.R402Q (bold) would fail the MAF filter detailed above (17.7% AF in ExAC all populations) for putative variants but is highlighted here as it is listed as 'pathogenic' in ClinVar and is of relevance to subsequent work in this publication. Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Variant type, consequence of the variant (s=synonymous, ns=nonsynonymous, sp=splicing, sg=stopgain); Gene.refGene, gene symbol; Omim Inheritance, inheritance as listed on OMIM for the gene in OCA/ nystagmus; Amino acid, amino acid change; avsnp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar) annotated 'p' if 'pathogenic'; InterVar, annotated as 'p' if identified as 'pathogenic' by InterVar; HGMD 2016 CLASS, annotated as DM for disease-causing mutation; Variant category, 1=assumed pathogenic, 2=assumed likely pathogenic.

chrom	position	ref	alt	Variant type	Gene refGene	Omim Inheritance	Amino Acid	avsnp144	ExAC_ALL	CADD_phred	MaxEnt Scan_diff	CLINSIG	Intervar	HGMD	Phenotype group																	
															1				2		3		4									
															(n=18)				(n=15)		(n=20)		(n=28)									
															NG335	NG433	NG477	NG528	NG315	NG381	NG280	NG540	NG195	NG340	NG391	NG394	NG395	NG416	NG498	NG543	NG551	
2	237493530	G	A	ns	MLPH	AR	R35Q			33.0				DM																		
6	35505739	A	C	sp	TULP1	AR				23.8	7.65		P																			
11	18310740	C	T	sp	HPS5	AR				27.2	8.18		P																			
11	31800832	G	A	ns	PAX6	AD	R142C	rs121907918		34.0			P		DM																	
11	31801623	C	G	ns	PAX6	AD	A113P			27.8				DM																		
11	31806401	C	T	sp	PAX6	AD				27.2	8.18		P		DM																	
11	89284793	G	A	ns	TYR	AR	R402Q	rs1126809	0.17700	34.0			P																			
11	89284805	C	T	ns	TYR	AR	P406L	rs104894313	0.00350	32.0			P		DM																	
11	89284879	C	A	ns	TYR	AR	P431T	rs368604842		28.3				DM																		
11	89284924	G	A	ns	TYR	AR	G446S	rs104894317	0.00002	31.0			P		DM																	
13	23339410	T	C	ns	SACS	AR	N1489S	rs147099630	0.00910	0.0				DM																		
15	27983383	T	C	ns	OCA2	AR	N489D	rs121918170	0.00030	28.2			P		DM																	
15	27985101	C	T	ns	OCA2	AR	V443I	rs121918166	0.00280	34.0			P		DM																	
15	27990579	G	A	s	OCA2	AR	G371G		0.02060					DM																		
15	28014795	T	C	ns	OCA2	AR	Y342C		0.00020	24.3				DM																		
15	28022554	G	A	ns	OCA2	AR	P198L		0.00010	29.0				DM																		
15	28022592	T	C	sp	OCA2	AR			0.00620		0.71			DM																		

[illegible]

Table 3: Nine patients with assumed likely pathogenic variants determined to be likely causal.

For 64 patients investigated for likely causal genotypes by investigating variants which were assumed likely pathogenic or as a combination of assumed pathogenic and assumed likely pathogenic, 13 patients were determined to have likely causal genotypes.

Samples, blue background indicates male whilst pink indicates female, orange cells indicate heterozygous variants and red indicates homozygous variants. Samples are ordered by phenotypic group, and cells with a 'C' denotes a likely causal variant, 'R' denotes a reportable variant, '-' denotes variant miscall. Chrom, chromosome; Position, location of 5' base of variant in hg38; Ref, reference allele; Alt, alternative allele; Variant type, consequence of the variant (s=synonymous, ns=nonsynonymous, sp=splicing, sg=stopgain); Gene.refGene, gene symbol; Omim Inheritance, inheritance as listed on OMIM for the gene in OCA/nystagmus; Amino acid, amino acid change; avsnp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar) annotated 'p' if 'pathogenic'; InterVar, annotated as 'p' if identified as 'pathogenic' by InterVar; HGMD 2016 CLASS, annotated as DM for disease-causing mutation; Variant category, 1=assumed pathogenic, 2=assumed likely pathogenic.

chrom	position	ref	alt	Variant type	Gene refGene	Omic Inheritance	Amino acid	avsnp144	ExAC ALL	CADD phred	MaxEnt Scan diff	CLINSIG	InterVar	HGMD	Variant category	Phenotype group								
																1 (n=14)			2 (n=13)			3 (n=18)		4 (n=19)
																NG299	NG445	NG534	NG318	NG327	NG383	NG512	NG521	NG386
1	235766255	G	A	ns	LYST	AR	T1982I	rs146591126	0.00630	17.1				DP	2									
3	149162256	G	A	ns	HPS3	AR	G739R	rs78336249	0.00960	22.9					2									
5	78015546	C	T	ns	AP3B1	AR	V999M	rs146503597	0.00380	16.4	-0.79				2									
6	15523217	G	A	ns	DTNBP1	AR	P272S	rs17470454	0.04360	19.9	0.14			DP	2									
6	35505739	A	C	sp	TULP1	AR				23.8	7.64		P		1									
9	12694076	C	G	ns	TYRP1	AR	P27R	rs373327120	0.00002	24.5					2								C	
9	12702394	C	G	ns	TYRP1	AR	P346R	rs377679582	0.00007	31.0					2								C	
11	18281986	G	A	ns	HPS5	AR	T1098I	rs61884288	0.02360	18.6		P		DM?	1		C							
11	18306204	C	T	ns	HPS5	AR	G252E	rs755846129	0.00003	32.0					2		C							

***TYR* tri-allelic Genotypic Cause of Albinism**

The *TYR* variant, NM_000372.4:exon4:c.G1205A:p.R402Q satisfies our assumed pathogenic criteria as it is assigned as assumed pathogenic in ClinVar despite being very common (17.7% AF in ExAC all populations). This variant is thought to form a part of a tri-allelic phenotype^{2,14,15} with another common variant; NM_000372.4:exon1:c.C575A:p.S192Y and any other rare pathogenic *TYR* variant. Therefore, we consider it here as a unique case. Of the 55 remaining undiagnosed patients, nine were identified to have the tri-allelic genotype within *TYR* (Table 4). These nine patients originated from phenotype groups 3 (n=6) and 4 (n=3). Each patient had a minimum of the two common variants, NM_000372.4:exon1:c.C575A:p.S192Y (25.2% in all populations of ExAC) and NM_000372.4:exon4:c.G1205A:p.R402Q (17.7% in all populations of ExAC), and one rare assumed pathogenic variant which is deemed clinically sufficient to call as the molecular basis of albinism.

Table 4: Nine samples were identified to have a likely causal tri-allelic genotype for albinism within TYR. For 55 patients which did not have likely causal genotypes with assumed pathogenic or assumed likely pathogenic variants, nine were identified to have tri-allelic causal genotypes in TYR. All TYR variants identified as assumed pathogenic or assumed likely pathogenic with the addition of S192Y and R402Q are listed. Position, location of 5' base of variant in hg38; REF, reference allele; ALT, alternative allele; Variant type, consequence of the variant (s=synonymous, ns=nonsynonymous, sp=splicing, sg=stopgain); AAchange; avsnp144, dbSNP144 rsID; ExAC ALL, Alternate allele frequency from ExAC database (all populations); CADD Phred, Combined Annotation Dependent Depletion score (Phred scale); MaxEntScan diff, the difference in score between MaxEntScan reference allele and alternative allele; ClinSig, clinical significance (clinvar); InterVar, pathogenicity category according to InterVar interpretation; HGMD 2016 class, HGMD annotation for pathogenicity; Samples, orange indicates heterozygous variants, red indicates homozygous variants. Samples are ordered by phenotypic group. , and 'c' was used to indicate a likely causal variant or 'R' was used to indicate a variant was reportable, grey highlights the common variants involved in the tri-allelic genotype outlined by Norman et al (S192Y and R402Q).

Position	alt	Variant type	AAchange	avsnp144	ExAC ALL	CADD phred	MaxEnt Scan diff	CLINSIG	InterVar	HGMD	Variant category	Phenotype group								
												3						4		
												(n=16)						(n=15)		
												NG263	NG356*	NG454	NG483	NG530	NG559	NG309	NG441	NG536
89178528	A	ns	S192Y	rs1042602	0.25180	25.0						C	C	C	C	C	C	C	C	C
89178602	T	ns	R217W	rs63159160	0.00020	25.1		P		DM	1									C
89178769	T	ns	W272C			29.6				DM	1						C			-
89191278	A	ns	R299H	rs61754375	0.00007	33.0		P		DM	1			C						
89227822	A	sp		rs61754382	0.00001	25.2	8.75		P	DM	1					C				
89227850	T	ns	A355V	rs151206295	0.00020	26.8		P		DM	1								C	
89227885	T	ns	H367Y	rs776054795	0.00001	29.5				DM	1				C					
89284792	T	sg	R402X	rs62645917	0.00005	51.0		P	P	DM	1	C								
89284793	A	ns	R402Q	rs1126809	0.17700	34.0		P			1	C	C	C	C	C	C	C	C	C
89284805	T	ns	P406L	rs104894313	0.00350	32.0		P		DM	1							C		
89284852	T	ns	R422W	rs749979474	0.00001	34.0				DM	1		C							

Albinism Patients with Partially Resolved Genetic Aetiology

In clinical practice it is common for gene testing to yield one well described pathogenic variant in an OCA gene but the absence of a second variant meaning that a molecular diagnosis cannot be made.

There were 46 patients without an assumed pathogenic, assumed likely pathogenic or *TYR* tri-allelic genotype identified. Forty-two patients across groups 1 (11), 2 (10), 3 (10) and 4 (15) without genetic diagnoses were subsequently investigated for single heterozygous assumed pathogenic or assumed likely pathogenic variants in an OCA/ OA genes (*TYR*, *OCA2*, *TYRP1*, *SLC45A2*, *SLC24A5*, *C10orf11*, *GPR143*²). Strikingly, sixteen patients had a single assumed pathogenic or assumed likely pathogenic variant which are likely to contribute to an albino genotype (see S3 Table). This corresponds to 2/10 (20.0%), 7/10 (70.0%) and 7/15 patients (46.6%) for of the remaining unresolved cases for phenotype group 2, 3 and 4 respectively. No patients in group 1 had a single assumed pathogenic or assumed likely pathogenic variant in any albinism gene. This is insufficient for a clinical diagnosis but warrants further investigation as it strongly suggests other missing variants in albinism genes for these cases.

Overview of Diagnostic Results

Table 5 summarises the number of samples harbouring likely causal variants. Overall, a clinically callable diagnostic yield of 48% was identified for the cohort as a whole. Phenotype groups 1 and 2 had similar diagnostic rates of 38% and 40% respectively. Groups 3 and 4 had a 50.0% and 57.1% diagnostic rate respectively. This shows that the additional phenotyping, beyond that of the baseline examinations prior to patient selection, did not significantly increase diagnostic yield.

Table 5: Summary table outlining the number of samples harbouring likely causal variants. A diagnostic rate is calculated and the likely causal genes are listed for each phenotype group.

Group No.	Cohort sub-group	Samples	Samples with assumed pathogenic diagnostic variants	Samples with assumed likely pathogenic diagnostic variants	Samples with <i>TYR</i> tri-allelic genotype	% Samples with likely causal variants	Likely causal genes reported
1	Idiopathic nystagmus (with complete phenotyping)	18	4	3	0	38.9	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>FRMD7</i> , <i>HPS5</i> , <i>TYR</i>
2	Idiopathic nystagmus (with incomplete phenotyping)	15	2	3	0	50.0	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>FRMD7</i> , <i>OCA2</i> , <i>SACS</i>
3	Albinism/ PAX6 disease diagnosis (with complete phenotyping)	20	2	2	6	50.0	<i>OCA2</i> , <i>TYR</i> , <i>TYRP1</i>
4	Albinism/ PAX6 disease diagnosis (with incomplete phenotyping)	28	9	1	3	46.4	<i>CACNA1A</i> , <i>CACNA1F</i> , <i>HPS5</i> , <i>OCA2</i> , <i>PAX6</i> , <i>TYR</i>
		81	17	9	9	43.2	

Discussion

In this study, we have utilised phenotyping methods which are currently employed in most large ophthalmology clinics worldwide. We recruited unselected, sequential patients in order to report on diagnostic yield using a UKGTN approved clinical panel based on the TruSight One ‘clinical exome’ panel which is being utilised in many centres as a cross-specialty, high throughput, sequencing platform. We report diagnostic yield for patients falling within the four most common clinical scenarios encountered in clinical practice; (1) complete phenotyped IINS, (2) likely IINS with incomplete phenotyping, (3) well phenotyped albinism and (4) likely albinism with incomplete phenotyping. We demonstrate a diagnostic rate across 81 patients of 43.2%, which is substantially higher than the majority of exome diagnostic analyses with the TruSight One capture¹⁸ and reflects, in part, the necessity for basic initial phenotyping in order to exclude common, clear clinical presentations and the utility of a subpanel of genes taken from a larger ‘clinical exome’ panel.

Six patients had assumed pathogenic variants in genes that would not have been previously directly implicated in causing the phenotype presentation of the patient according to the available clinical information. For example, NG315 from phenotype group 1 was found to have a likely disease-causing compound heterozygous genotype in the *OCA2* gene. These cases reflect the variable, often hypomorphic and overlapping phenotypes seen in children with nystagmus^{4,30}, and support the argument that basic phenotyping alone prior to panel testing, may be the most efficient clinical diagnostic workflow rather than reducing the gene panel further by assuming that phenotyping has excluded or confirmed an albinism related phenotype.

The variant prioritisation category of assumed likely pathogenic identified some likely causal variants in genes for which the phenotype seen in our patients is unexpected. For example, NG381 (idiopathic nystagmus with incomplete phenotyping) was found to be homozygous for an assumed likely pathogenic splicing variant in the *CACNA1F* gene, which is known to cause Aland Island eye disease, cone-rod dystrophy and X-linked incomplete stationary night blindness (CSNB); all of which cause nystagmus and retinal dystrophy. It might be expected that such disorders would be identified by

ERGs prior to recruitment to this study, however, this patient's ERG result was initially reported as normal. Interestingly, a subsequent ERG performed at an older age for this patient identified the typical features of CSNB. This case and others may support an argument that genomic testing in the future may form an earlier part of the diagnostic workflow. More detailed phenotyping, which has been outlined here and by others ³¹, might then be directed towards proving or disproving diagnoses suggested by putative likely causal variants, but it is clear that clinical evaluation and baseline phenotyping will still be required regardless of the yield from genetic testing.

The high number of cases identified here with likely albinism related phenotypes, but for whom only a single albinism gene pathogenic variant was identified, mirrors that seen in clinical practice. It strongly suggests that other, as yet unknown, variants in albinism genes are contributing or that multiple variants across the melanin biosynthesis pathway may combine to contribute to cause albinism phenotypes (epistasis). Some albinism genes such as the *C10orf11* gene are less well covered (70.5%) which may miss contributing variants for the albinism phenotype. The TruSight One chemistry could be backfilled to provide a more comprehensive and reliable capture of all target regions. Clearly, identification of these 'missing variants' and complex causal genotypes is likely to increase the diagnostic yield still further.

Nystagmus and albinism gene panels can vary from less than 20 to more than 300 genes ^{17,19,32}. Here we have used 31 genes taken as a subpanel from the commonly used TruSight One 'clinical exome' gene panel. The value of employing a large off the shelf target kit such as the TruSight One means that by revisiting the data already generated, the gene panel could be expanded retrospectively. The 'TruSight One Expanded v3.0' has also expanded on the original TruSight One capture to cover other known nystagmus-causing genes such as *SLC38A8*.

It is possible that a diagnostic report may initially include false positives, however, all identified variants which have been identified as 'likely causal' would be verified with Sanger sequencing before clinical reporting.

Conclusions

In conclusion, the work presented here shows that for clinicians using a standard set of phenotyping criteria, the UKGTN approved 31 gene panel based on the TruSight One platform, can provide a clinically callable genetic diagnosis for 43.2% of children with INS regardless of more detailed phenotyping. This could significantly reduce the time and number of investigations that many children with nystagmus undergo and permit informed family counselling with regards to recurrence risk. This could also lead to more rapid access to tailored management, a current priority for health systems worldwide, for conditions such as spinocerebellar ataxia 6/ episodic ataxia type 2 or Hermansky-Pudlak syndrome. For the future planning of diagnostic workflows and genetic testing strategies, it is important to note that the work here has shown similar diagnostic yield across all four patient groups. This suggests that clinical phenotyping (beyond the exclusion of retinal dystrophy, known underlying ophthalmic disease, prematurity or likely neurological cause) is not a key prerequisite for genetic testing, and that detailed phenotyping could be tailored subsequent to genetic testing in order to confirm or refute putative genetic diagnoses. It is also clear that diagnostic workflows for children with nystagmus should include differing genetic strategies according to the basic phenotyping results or availability of phenotyping modalities. For example, a first-line retinal gene panel for children with a retinal dystrophy identified by ERGs and a first line nystagmus/albinism gene panel for children with a clear albinism phenotype prior to ERG and VEP testing.

References

1. Tarpey P, Thomas S, Sarvananthan N, et al. Mutations in FRMD7, a newly identified member of the FERM family, cause X-linked idiopathic congenital nystagmus. *Nat Genet.* 2006;38(11):1242-1244. doi:10.1038/ng1893
2. Norman CS, O’Gorman L, Gibson J, et al. Identification of a functionally significant tri-allelic genotype in the Tyrosinase gene (TYR) causing hypomorphic oculocutaneous albinism

- (OCA1B). *Sci Rep*. 2017;7(1):4415. doi:10.1038/s41598-017-04401-5
3. Self J, Mercer C, Boon EMJ, et al. Infantile nystagmus and late onset ataxia associated with a CACNA1A mutation in the intracellular loop between s4 and s5 of domain 3. *Eye (Lond)*. 2009;23(12):2251-2255. doi:10.1038/eye.2008.389
 4. Gronskov K, Ek J, Brondum-Nielsen K. Oculocutaneous albinism. *Orphanet J Rare Dis*. 2007;2:43. doi:10.1186/1750-1172-2-43
 5. CEMAS Workshop. *Classification of Eye Movement Abnormalities and Strabismus (CEMAS) Workshop Report.*; 2001.
 6. von dem Hagen EAH, Hoffmann MB, Morland AB. Identifying human albinism: a comparison of VEP and fMRI. *Invest Ophthalmol Vis Sci*. 2008;49(1):238-249. doi:10.1167/iovs.07-0458
 7. Osborne D, Theodorou M, Lee H, et al. Supranuclear eye movements and nystagmus in children: A review of the literature and guide to clinical examination, interpretation of findings and age-appropriate norms. *Eye (Lond)*. October 2018. doi:10.1038/s41433-018-0216-y
 8. Hingorani M, Williamson KA, Moore AT, van Heyningen V. Detailed ophthalmologic evaluation of 43 individuals with PAX6 mutations. *Invest Ophthalmol Vis Sci*. 2009;50(6):2581-2590. doi:10.1167/iovs.08-2827
 9. Montoliu L, Gronskov K, Wei A-H, et al. Increasing the complexity: new genes and new types of albinism. *Pigment Cell Melanoma Res*. 2014;27(1):11-18. doi:10.1111/pcmr.12167
 10. Oetting WS, Pietsch J, Brott MJ, et al. The R402Q tyrosinase variant does not cause autosomal recessive ocular albinism. *Am J Med Genet A*. 2009;149A(3):466-469. doi:10.1002/ajmg.a.32654
 11. Gronskov K, Jespersgaard C, Bruun GH, et al. A pathogenic haplotype, common in

- Europeans, causes autosomal recessive albinism and uncovers missing heritability in OCA1. *Sci Rep*. 2019;9(1):645. doi:10.1038/s41598-018-37272-5
12. Hutton SM, Spritz RA. A comprehensive genetic study of autosomal recessive ocular albinism in Caucasian patients. *Invest Ophthalmol Vis Sci*. 2008;49(3):868-872. doi:10.1167/iovs.07-0791
 13. Fukai K, Holmes SA, Lucchese NJ, et al. Autosomal recessive ocular albinism associated with a functionally significant tyrosinase gene polymorphism. *Nat Genet*. 1995;9(1):92-95. doi:10.1038/ng0195-92
 14. Chiang P-W, Spector E, Tsai AC-H. Evidence suggesting the inheritance mode of the human P gene in skin complexion is not strictly recessive. *Am J Med Genet A*. 2008;146A(11):1493-1496. doi:10.1002/ajmg.a.32321
 15. Berson JF, Frank DW, Calvo PA, Bieler BM, Marks MS. A common temperature-sensitive allelic form of human tyrosinase is retained in the endoplasmic reticulum at the nonpermissive temperature. *J Biol Chem*. 2000;275(16):12281-12289.
 16. Seaby EG, Pengelly RJ, Ennis S. Exome sequencing explained: a practical guide to its clinical application. *Brief Funct Genomics*. 2016;15(5):374-384. doi:10.1093/bfgp/elv054
 17. Rim JH, Lee S-T, Gee HY, et al. Accuracy of Next-Generation Sequencing for Molecular Diagnosis in Patients With Infantile Nystagmus Syndrome. *JAMA Ophthalmol*. 2017;135(12):1376-1385. doi:10.1001/jamaophthalmol.2017.4859
 18. Pajusalu S, Kahre T, Roomere H, et al. Large gene panel sequencing in clinical diagnostics- results from 501 consecutive cases. *Clin Genet*. 2018;93(1):78-83. doi:10.1111/cge.13031
 19. Thomas MG, Maconachie GDE, Sheth V, McLean RJ, Gottlob I. Development and clinical utility of a novel diagnostic nystagmus gene panel using targeted next-generation sequencing. *Eur J Hum Genet*. 2017;25(6):725-734. doi:10.1038/ejhg.2017.44

20. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013. <http://arxiv.org/abs/1303.3997>.
21. McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res*. 2010;20(9):1297-1303. doi:10.1101/gr.107524.110
22. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164. doi:10.1093/nar/gkq603
23. Kircher M, Witten DM, Jain P, O’Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892
24. Yeo G, Burge CB. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*. 2004;11(2-3):377-394. doi:10.1089/1066527041410418
25. Li Q, Wang K. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet*. 2017;100(2):267-280. doi:10.1016/j.ajhg.2017.01.004
26. Stenson PD, Mort M, Ball E V, Shaw K, Phillips A, Cooper DN. The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet*. 2014;133(1):1-9. doi:10.1007/s00439-013-1358-4
27. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009;25(16):2078-2079.
28. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*. 2010;26(6):841-842. doi:10.1093/bioinformatics/btq033
29. Desmet F-O, Hamroun D, Lalande M, Collod-Beroud G, Claustres M, Beroud C. Human

- Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res.* 2009;37(9):e67. doi:10.1093/nar/gkp215
30. Simeonov DR, Wang X, Wang C, et al. DNA variations in oculocutaneous albinism: an updated mutation list and current outstanding issues in molecular diagnostics. *Hum Mutat.* 2013;34(6):827-835. doi:10.1002/humu.22315
 31. Kruijt CC, de Wit GC, Bergen AA, Florijn RJ, Schalijs-Delfos NE, van Genderen MM. The Phenotypic Spectrum of Albinism. *Ophthalmology.* 2018;125(12):1953-1960. doi:10.1016/j.ophttha.2018.08.003
 32. Lasseaux E, Plaisant C, Michaud V, et al. Molecular characterization of a series of 990 index patients with albinism. *Pigment Cell Melanoma Res.* 2018;31(4):466-474. doi:10.1111/pcmr.12688

Author Contributions

L.O. and C.S.N. have contributed equally, to the wet lab and bioinformatics work in addition to manuscript preparation. They will share first authorship. L.M. led in the development of the gene panel and contributed to manuscript preparation. D.O. contributed to clinical phenotyping, patient database curation and manuscript preparation. T.N. contributed to sequencing experiments, patient database curation and manuscript preparation. A.H.C., A.J.C., A.J.L., E.L.B., J.A.R. and D.B. contributed to directing and steering the project and manuscript preparation. C.M. contributed gene panel development and manuscript preparation. H.L. contributed to clinical aspects of the study and manuscript preparation. F.S. performed all electro-diagnostic phenotyping in the study and contributed to manuscript preparation. J.G. contributed to study design, analysis and manuscript preparation. S.E. and J.S. contributed to study design, project oversight, analysis and manuscript preparation. They will share last authorship.

Competing interests

The authors declare that they have no competing interests.

Consent

Consent was obtained in accordance with the Declaration of Helsinki and was approved by South West Hampshire Local Research Ethics Committee (LREC 028/04/t).

Data availability

Data generated or analysed during this study are included in this published article and its supplementary files

Acknowledgements

We thank the families for their participation in this research and Gift of Sight for funding this study.