

**Developing variant interpretation
pipelines for inherited retinal
diseases and ciliopathies: using
medical genomics to improve
diagnostic yield**



UNIVERSITY OF LEEDS

Sunayna Kathleen Best

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

The University of Leeds
School of Medicine and
Health

November 2022

Intellectual Property and Publication Statements

Jointly authored publications have been used in this thesis. The contribution of the candidate and the other authors to this work has been explicitly indicated below. The candidate confirms that appropriate credit has been given within the thesis where reference has been made to the work of others.

Jointly authored publications are included within the following chapters:

Chapter 2: Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project

Publication included (thesis section 2.4): Best S, Lord J, Roche M, Watson CM, Poulter JA, Bevers RPJ, Stuckey A, Szymanska K, Ellingford JM, Carmichael J, Brittain H, Toomes C, Inglehearn C, Johnson CA, Wheway G; Genomics England Research Consortium. Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project. *J Med Genet.* 2022 Aug;59(8):737-747. doi: 10.1136/jmedgenet-2021-108065. Epub 2021 Oct 29. PMID: 34716235

Research contributions

I designed the research study, undertook the analyses, and wrote and edited the manuscript. Alongside my PhD supervisors Prof. Colin Johnson, Prof. Chris Inglehearn and Dr. Carmel Toomes, I received additional supervision from Dr. Gabrielle Wheway, a Lecturer in Functional Genomics within the Faculty of Medicine at the University of Southampton. In particular, she helped to train me in navigation of the Genomics England (GEL) research environment and addressed specific variant queries that I found during analysis.

Dr. Matthew Roche, a general practitioner (GP) and experienced medical coder, helped me to write a Python script to batch run identified variants through the Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) and then filter down variants of interest (see additional methodology, thesis section 2.2). I received additional training and advice on practical 100K analyses from Dr. Jamie Ellingford, a Research Fellow at the Manchester Centre for Genomic Medicine, and from Dr. Jenny Lord, a Postdoctoral Research Fellow within the Faculty of Medicine at the University of Southampton. Dr. Lord also gave permission to use her custom Python scripts for un-tiered variant

analysis (*find_variants_by_gene_and_consequence.py*); available at https://github.com/JLord86/Extract_variants) and SpliceAI analysis (*find_variants_by_gene_and_SpliceAI_score.py*); available at https://github.com/JLord86/Extract_variants). These are both provided in full in Appendix section 6.2.

Locally, I received additional training and variant interpretation support from Dr. Chris Watson, a registered Clinical Scientist leading NHS research and development activities in the Translational Genomics Unit in Leeds, and Dr. James Poulter, a UKRI Future Leaders Fellow in Molecular Neuroscience at the University of Leeds. Finally, Dr. Kasia Szymanska, a post-doctoral research fellow in the ciliopathy research group, shared her list of locally identified candidate ciliopathy disease genes for the analysis.

I disseminated all potential new diagnoses I identified to recruiting clinicians via the Genomics England (GEL) Airlock system. I have included the clinicians Jenny Carmichael and Helen Brittain who collaborated on this research as authors on the research paper. Finally, I received technical support from two GEL bioinformaticians via their online helpdesk who have been credited as authors (Roel Bevers and Alex Stuckey).

Chapter 3: Uncovering the burden of hidden ciliopathies in the 100 000 Genomes Project: a reverse phenotyping approach

Publication included (thesis section 3.2): Best S, Yu J, Lord J, Roche M, Watson CM, Bevers RPJ, Stuckey A, Madhusudhan S, Jewell R, Sisodiya SM, Lin S, Turner S, Robinson H, Leslie JS, Baple E; Genomics England Research Consortium, Toomes C, Inglehearn C, Wheway G, Johnson CA. Uncovering the burden of hidden ciliopathies in the 100 000 Genomes Project: a reverse phenotyping approach. *J Med Genet.* 2022 Jun 28;jmedgenet-2022-108476. doi: 10.1136/jmedgenet-2022-108476. Epub ahead of print. PMID: 35764379

Research contributions

I designed the research study, undertook the analyses, and wrote and edited the manuscript. I received supervision from Dr. Gabrielle Wheway from the University of Southampton for practical research environment and specific variant queries, as well as from my main PhD supervisors. I collaborated again with the experienced medical coder Dr. Matthew Roche to write the custom Python script called *filter_gene_variant_workflow.py* (available from

https://github.com/sunaynabest/filter_100K_gene_variant_workflow and in section 6.2.2 of the Appendix). Roel Bevers and Alex Stuckey are GEL bioinformaticians who wrote and provided support for use of the *Gene-Variant Workflow* script (available from <https://research-help.genomicsengland.co.uk/display/GERE/GeneVariant+Workflow>). Dr. Jing Yu, a senior bioinformatician with the Nuffield Department of Clinical Neurosciences at the University of Oxford wrote and provided support for use of the SVRare script (available from <https://github.com/Oxford-Eye/SVRare-GEL>) used for structural variant analysis. Dr. Jenny Lord, a Postdoctoral Research Fellow within the Faculty of Medicine at the University of Southampton, wrote and provided support for use of the *find_variants_by_gene_and_SpliceAI_score.py* script (available at https://github.com/JLord86/Extract_variants and provided in full in Appendix section 6.2.4) for splice variant analysis.

Dr. Chris Watson (Lead Clinical Scientist in the Translational Genomics Unit in Leeds) once again helped with variant interpretation queries and undertook the functional laboratory work required to validate the *BBS1* mobile element insertion (thesis section 3.2, manuscript Figure 3E.i) (Best et al., 2022c). The remaining credited authors are clinicians who collaborated with us to interpret identified variants amongst their recruited 100K participants (Savita Madhusudhan, Rosalyn Jewell, Sanjay Sisodiya, Siying Lin, Stephen Turner, Hannah Robinson, Joseph Leslie, and Emma Baple).

Chapter 4: Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*

Publication included (thesis section 4.5): Lange KI, Best S, Tsiropoulou S, Berry I, Johnson CA, Blacque OE. Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*. *Hum Mol Genet.* 2022 May 19;31(10):1574- 1587. doi: 10.1093/hmg/ddab344. PMID: 34964473; PMCID: PMC9122650.

Research contributions

This laboratory-based project was designed in collaboration with colleagues from Professor Oliver Blacque's cilium disease research group at University College Dublin (UCD). This project was jointly led by me within Professor Johnson's laboratory group, and the post- doctoral researcher Dr. Karen Lange in Professor Blacque's group.

Dr. Lange was responsible for the design and deployment of all *C. elegans*

experiments, helped by another UCD post-doctoral researcher, Sofia Tsiropoulou. I was responsible for all human cell line experiments. Briefly, this included generation and characterisation of the *TMEM67* knockout hTERT RPE-1 cell lines, generation and validation of variant *TMEM67* c- myc-tagged plasmid constructs by site-directed mutagenesis and development of our *in vitro* human cell culture-based assay of *TMEM67* function. I also undertook the initial variant selection in collaboration with the clinical scientist Ian Berry (detailed in thesis section 4.1) and protein structure prediction modelling presented in Figure 1B of the manuscript (Lange et al., 2022).

Dr. Lange and I jointly wrote and edited the manuscript introduction and discussion. I was the lead author for the following sections of the results: selection of *TMEM67* variants for analysis; Figure 1; *in vitro* genetic complementation assay of *TMEM67* VUS function in human cell culture and Figure 4. Within the materials and methods section, I was the lead author on the following sections: modelling of protein secondary structure; cell culture; *TMEM67* cloning, plasmid constructs and transfections; CRISPR/Cas9 genome editing in cell culture; PCR and sequence validation of crispant cell-lines; whole cell extract preparation and western immunoblotting. In the supplementary material, I created Figure S2 (RaptorX predicted structures of *TMEM67* and MKS-3) and Figure S4 (characterization of *TMEM67* crispant). Dr. Lange was the lead author for all remaining sections and figures.

Chapter 5: Discussion

Publications included

Thesis section 5.6: Best S, Inglehearn CF, Watson CM, Toomes C, Wheway G, Johnson CA. Unlocking the potential of the UK 100,000 Genomes Project - lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort. *Am J Med Genet C Semin Med Genet.* 2022 Mar;190(1):5-8. doi: 10.1002/ajmg.c.31965. Epub 2022 Mar 15. PMID: 35289502; PMCID: PMC9315030.

Thesis section 5.7: Brown MA, Wigley C, Walker S, Lancaster D, Rendon A, Scott R. Re: Best et al., 'Unlocking the potential of the UK 100,000 Genomes Project - Lessons learned from analysis of the "Congenital malformations caused by ciliopathies" cohort'. *Am J Med Genet A.* 2022 Nov;188(11):3376-3377. doi: 10.1002/ajmg.a.62909. Epub 2022 Jul 21. PMID: 35861231.

Research contributions

I wrote and edited this commentary article, with supervision and support from the other authors. These include my PhD supervisors Prof. Chris Inglehearn, Dr. Carmel Toomes and Prof. Colin Johnson, Dr. Gabrielle Wheway (Lecturer in Functional Genomics within the Faculty of Medicine at the University of Southampton), who provided additional supervision for all my 100,000 Genomes Project research projects, and Dr. Chris Watson ((lead Clinical Scientist in the Translational Genomics Unit in Leeds), who provided particular insight into the realities of 100,000 Genomes Project analysis from within an NHS diagnostic laboratory.

We received a response to our commentary article from senior staff within Genomics England (Brown et al, 2022). Although I did not contribute to the writing of this manuscript, I include it to provide insight from the other side, adding to the discussion about lessons learned from 100,000 Genomes Project data analysis.

Rationale for submitting for PhD by publication

I am submitting for PhD by publication because the included articles meet the threshold for examination through this alternative route. They were peer-reviewed prior to publication, demonstrating that they have been considered of publishable quality by the journals selected. It therefore seems fitting to include this research in its peer-reviewed format.

I have structured the results chapters to introduce the original research manuscripts; starting with the two 100,000 Genomes Project articles (Best et al., 2022b, Best et al., 2022c), followed by the functional *TMEM67* variant interpretation project article (Lange et al., 2022). In each chapter introduction, I have provided the research rationale, any additional methodology and any additional results not covered within the manuscript. These chapter introductions are followed by the published manuscripts and any accompanying supplementary material is provided in thesis appendix section 6.1. My published 100,000 Genomes Project commentary article (Best et al., 2022a), as well as the response received from GEL (Brown et al., 2022) supplement the new material in the Discussion (thesis section 5).

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

Acknowledgements

I am indebted to my PhD supervisors Prof. Colin Johnson, Prof Chris Inglehearn and Dr. Carmel Toomes, for their enduring enthusiasm, patience, support, and encouragement. You have been there to help me through every research challenge and query, no matter how small, for which I am extremely grateful. Thank you for being especially supportive through the anxieties and uncertainties of the Covid pandemic. Your reassurances, flexibility and calm nurturing were hugely appreciated.

Thank you to the 4Ward North Clinical PhD Academy for funding this PhD and providing support during the process.

I am also grateful for the additional supervision from Dr. Gabrielle Wheway, who provided invaluable online support and guidance for the 100,000 Genomes Project analysis. I have benefitted hugely from working with our key bioinformatics collaborators Dr. Jenny Lord, Dr. Jing Yu and Roel Bevers, and thank them for their help and generosity. I look forward to someday meeting you all in person now that we are through the restrictions of the pandemic.

Thank you to my colleagues in Team Meckel and the Vision Research Group. It has been great to be part of these teams, who have provided solidarity and support every step of the way. Thanks in particular go to Dr. Basudha Basu, Dr. Katarzyna Szymanska, Dr. Claire Smith, Rowan Taylor, Ewa Jaworska and Dr. James Poulter for their practical help in the labs.

Huge thanks to my patient and proactive husband, Dr. Matthew Roche, who not only provided emotional support during my PhD, but hands-on help in writing critical scripts for my 100,000 Genomes Project analyses. I could not have done it without you. I am grateful also to my parents and sisters for their patience and kindness during my PhD.

This thesis is dedicated to my babies. Thank you for bringing me comfort, company, and joy during this time. You are the light of my life.

Abstract

Primary ciliopathies are a group of rare inherited disorders caused by defects in the structure or function of primary cilia (the 'cell's antenna'). This thesis describes approaches to improve molecular diagnosis rates for primary ciliopathy patients over the ~40-80% currently achieved, through whole genome sequencing (WGS) analysis and functional variant interpretation.

Firstly, I analysed WGS data from the 100,000 Genomes Project (100K) for participants who were clinically suspected to have primary ciliopathies. I identified a molecular diagnosis rate for $n=45/83$ (54.2%), providing a 21.7% diagnostic uplift compared to results previously reported by Genomics England (GEL).

I then performed a reverse phenotyping study, starting by looking for pathogenic variants in nine multisystemic ciliopathy disease genes across the 100K rare disease dataset. This was linked back to available clinical data, aiming to identify participants with "hidden" ciliopathy diagnoses recruited to alternative categories. I identified 18 new, reportable diagnoses and 44 previously reported by GEL. I also found 11 un-reportable molecular diagnoses, lacking key clinical features to provide a confident fit for phenotype. This shows that the quality of entered phenotypic data is critical to allow accurate genotype-phenotype correlation.

In a third study, I developed strategies for functional interpretation of eight *TMEM67* missense variants of uncertain significance (VUSs) with collaborators in Ireland, using CRISPR/Cas9 gene editing in a human ciliated cell-line (RPE-1) and *C. elegans*. These assays provided interpretation of three VUS as benign and five as pathogenic.

The two 100K studies show that diagnosis rates for ciliopathies can be improved through WGS analysis, especially structural and splice variant analysis. We are a long way from delivering a high-throughput system for VUS interpretation that could provide clinical utility in the diagnostic setting. Overall, we have provided benefit for ciliopathy patients through additional molecular diagnoses, accompanied by transferable skills applicable to wider patient groups.

Table of Contents

<i>Intellectual Property and Publication Statements</i>	2
<i>Acknowledgements</i>	7
<i>Abstract</i>	8
<i>Table of Contents</i>	9
<i>Abbreviation list</i>	13
<i>List of Figures</i>	18
<i>List of Tables</i>	19
1 Introduction	20
1.1 Genetic variant interpretation	20
1.1.1 Rare disease diagnostics.....	20
1.1.2 Genetic variation.....	21
1.1.3 American College of Medical Genetics and Genomics (ACMG) variant classification.....	23
1.1.4 NGS-based genomic testing strategies.....	28
1.1.5 Sources of missed genetic diagnoses.....	29
1.1.6 The 100,000 Genomes project (100K).....	33
1.1.7 Challenges in genomic testing.....	35
1.2 Cilia	37
1.2.1 Cilia types.....	37
1.2.2 Cilia structure and ciliogenesis.....	40
1.2.3 Ciliary proteins.....	46
1.3 Ciliopathies	47
1.3.1 Clinical features.....	47
1.3.2 Genetics of ciliopathies.....	55
1.3.3 Molecular diagnosis rates.....	61
1.4 Primary cilia in cell signalling	61
1.4.1 Sonic Hedgehog pathway (Shh).....	61
1.4.2 Wnt signalling.....	64
1.5 Modelling genetic variants	70
1.5.1 Cilia research models.....	70
1.5.2 CRISPR-Cas genome editing.....	72
1.6 TMEM67	77
1.6.1 Encoded protein.....	77

1.6.2	TMEM67 knockout models	80
1.6.3	Disease associations	80
1.6.4	Variant pathogenicity	81
1.7	Hypothesis	81
1.8	Overall objective	81
1.9	Specific chapter aims	82
1.9.1	Chapter 2	82
1.9.2	Chapter 3	82
1.9.3	Chapter 4	83
2	<i>Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project.....</i>	84
2.1	Research Rationale	84
2.2	Additional methodology	85
2.3	Additional results	87
2.4	Manuscript	90
3	<i>Uncovering the burden of hidden ciliopathies in the 100,000 Genomes Project: a reverse phenotyping approach.....</i>	101
3.1	Research Rationale	101
3.2	Manuscript	104
4	<i>Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in Caenorhabditis elegans.....</i>	118
4.1	Research Rationale	118
4.2	Additional methodology	119
4.2.1	Polymerase Chain Reaction (PCR)	119
4.2.2	Agarose gel electrophoresis	119
4.2.3	Exonuclease I – Shrimp Alkaline Phosphatase (ExoSAP) PCR purification	120
4.2.4	Sanger Sequencing	120
4.2.5	Bacterial transformation of variant plasmids generated by site- directed mutagenesis .	121
4.2.6	Small interfering RNA (siRNA) knockdown experiments.....	121
4.2.7	Whole cell extract (WCE) preparation and Western Blotting	122
4.2.8	High-content imaging	123
4.3	Additional results	125
4.3.1	Generation of <i>TMEM67</i> knockout cell lines	125

4.3.2	Attempt at variant interpretation by high-content imaging	129
4.4	Conclusion	131
4.5	Manuscript	132
5	Discussion	146
5.1	Research output summary	146
5.2	Motivation for the PhD and overall take-home messages	146
5.3	Lessons learned: 100,000 Genomes Project analyses	147
5.3.1	Diagnostic uplift achieved by 100K rare disease cohort research analyses	147
5.3.2	Time commitments and strategy development	151
5.3.3	Reverse phenotyping as a source of missed diagnoses	153
5.3.4	Added value of structural variant analysis in 100K	154
5.3.5	Added value of splice variant analysis in 100K	156
5.4	Lessons learned: functional VUS analyses	158
5.5	Looking to the future	161
5.5.1	Clinical genomics era	161
5.5.2	Increased use of long-range sequencing	162
5.5.3	Newborn Genomes Programme	162
5.6	Manuscript: Unlocking the potential of the UK 100,000 Genomes Project - lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort	165
5.7	Manuscript: Re: Best et al., 'Unlocking the potential of the UK 100,000 Genomes Project - Lessons learned from analysis of the "Congenital malformations caused by ciliopathies" cohort	169
6	Appendices	171
6.1	Published manuscript supplementary materials	171
6.1.1	Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project	171
6.1.2	Uncovering the burden of hidden ciliopathies in the 100,000 Genomes Project: a reverse phenotyping approach	188
6.1.3	Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in <i>Caenorhabditis elegans</i>	204
6.2	Custom python scripts	215
6.2.1	Filter_vep_output_variants.py	215
6.2.2	filter_gene_variant_workflow.py	217
6.2.3	find_variants_by_gene_and_consequence.py	221

6.2.4	find_variants_by_gene_and_SpliceAI_score.py.....	225
6.3	Reagents	227
6.3.1	Suppliers.....	227
6.3.2	Reagents.....	229
6.3.3	Buffers and Solutions	232
6.3.4	Cells lines.....	234
6.3.5	Antibodies and cell stains.....	234
6.4	Plasmid map: TMEM67_myc_HisA wild type	237
7	References.....	238

Abbreviation list

Abbreviation	Expansion
100K	100,000 Genomes Project
ABI	Association of British Insurers
ACGS	Association for Clinical Genomic Science
ACMG	American College of Medical Genetics and Genomics
ADPKD	Autosomal dominant polycystic kidney disease
ALMS	Alström syndrome
AONs	Antisense oligonucleotides
APC	Adenomatous polyposis coli
array-CGH	Array-comparative genomic hybridisation
BB	Basal body
BBS	Bardet-Biedl syndrome
bp	Base pair
BSA	Bovine serum albumin
CADD	Computer Annotation Dependent Depletion
Cas	CRISPR associated protein
CC	Connecting cilium
CK1	Caesin kinase 1
CMC	Congenital malformations caused by ciliopathies
CNS	Central nervous system
CNV	Copy number variant
COACH	Cerebellar vermis hypoplasia, Oligophrenia (developmental delay/mental retardation), Ataxia, Coloboma, and Hepatic fibrosis
CRD	Cysteine rich domain
CRISPR	Clustered regularly interspaced short palindromic repeats
crRNA	Crispr-RNA
dCas9	Dead Cas9

DDD	Deciphering Developmental Disorders study
DDG2P	Developmental Disorders Genotype-to-Phenotype
Dhh	Desert hedgehog
DSB	Double stranded break
Dvl	Disheveled
ERG	Electroretinogram
EU	European Union
EVC	Ellis Van Creveld
ExoSAP	Exonuclease I – shrimp alkaline phosphatase
FACS	Fluorescence-activated cell sorting
FDA	Food and Drug Authority
FFPE	Formalin-fixed paraffin embedded
FITC	Fluorescein isothiocyanate
FORGE	Finding Of Rare Disease genes
GEL	Genomics England
GliA	Gli activator form
GMC	Genomic Medicine Centre
gnomAD	Genome Aggregation Database
GP	General Practitioner
GPCR	G-protein coupled receptor
gRNA	Guide RNA
HDR	Homology Directed Repair
HFEA	Human Fertilisation and Embryology Authority
HPO	Human Phenotype Ontology
HRP	Horseradish peroxidase
hTERT	Human telomerase reverse transcriptase
IDA	Inner dynein arm
IFT	Intraflagellar transport
IGV	Integrative Genomics Viewer

Ihh	Indian hedgehog
iPSC	Induced pluripotent stem cell
IRD	Inherited retinal dystrophy
IRDiRC	International Rare Diseases Research Consortium
IS	Inner segments
JATD	Jeune asphyxiating thoracic dystrophy
Jbn	Joubertin
JBTS	Joubert syndrome
Kb	Kilobase
LCA	Leber congenital amaurosis
LEF	Lymphocyte enhancer factor
LOF	Loss of function
MAF	Minor allele frequency
Mb	Megabase
mIMCD3	Murine inner medullary collecting duct
MKS	Meckel Gruber syndrome
MRI	Magnetic resonance imaging
mRNA	Messenger RNA
MTS	Molar tooth sign
N-DRC	Nexin–dynein regulatory complexes
N/a	Not applicable
NGS	Next generation sequencing
NHEJ	Non-homologous end joining
NHS	National Health Service
NMD	Nonsense mediated decay
NTD	Neural tube defect
ODA	Outer dynein arm
OFD	Oral-facial-digital syndrome
ONT	Oxford Nanopore Technologies

OS	Outer segments
OTE	Off target effect
PacBio	Pacific Biosciences
PAM	Protospacer associated motif
PBS	Phosphate buffered saline
PBST	Phosphate buffered saline with Tween
PCD	Primary ciliary dyskinesia
PCP	Planar cell polarity
PCR	Polymerase chain reaction
PDGF	Platelet-derived growth factor
PGT-M	Pre-implantation genetic testing for monogenic or single-gene disorders
PKD	Polycystic kidney disease
PVDF	Polyvinylidene difluoride
RMCD	Rare multisystem ciliopathy disorders
RP	Retinitis pigmentosa
RPE	Retinal pigment epithelium
SBS	Sequencing by synthesis
Shh	Sonic hedgehog
siRNA	Small interfering RNA
Smo	Smoothened
SMRT	Single-molecule real-time
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SpCas9	Streptococcus pyogenes Cas9
SRTD	Short-rib thoracic dysplasia
ssODN	Single-stranded oligodeoxynucleotide
STR	Short tandem repeat
SuFu	Suppressor of Fused

SV	Structural variant
TALEN	Transcription activator-like effector nucleases
TCF	T-cell factor
TF	Transition fibres
tracrRNA	Transactivating CRISPR RNA
TZ	Transition zone
UCD	University College Dublin
UCSC	University of California Santa Cruz
UPD	Uniparental disomy
UPS	Ubiquitin-proteasome system
USH	Usher syndrome
UTR	Untranslated regions
UV	Ultraviolet
V	Volts
VCF	Variant call format
VEP	Variant effect predictor
VUS	Variant of uncertain significance
WAD	Weyer's acrofacial dysostosis
w/v	Weight for volume
WCE	Whole cell extract
WES	Whole exome sequencing
WGS	Whole genome sequencing
ZFN	Zinc-finger nucleases
EDTA	Ethylenediaminetetraacetic acid

List of Figures

Figure 1. Schematic of transverse cross section of cilia.....	38
Figure 2. Schematic of the structure of the primary cilium.	42
Figure 3. Overlapping disease features of the ciliopathies.....	48
Figure 4. Typical external features for a fetus with MKS at 16 weeks' gestation.	51
Figure 5. Shh signalling at the primary cilium.....	63
Figure 6. The non-canonical Wnt signalling pathway in primary cilia.	65
Figure 7 Canonical Wnt signalling at the primary cilium.....	68
Figure 8. Repair pathways for Cas9-induced DSBs.....	74
Figure 9. Schematic of TMEM67 protein with target variants for modelling	79
Figure 10. IGV captures showing homozygous BBS4 deletion in CMC proband 78 and heterozygous deletion in their father.	89

List of Tables

Table 1. ACMG guidelines on strengths of evidence in favour of variant pathogenicity.	24
Table 2. AMCG guidance on combinations of lines of variant pathogenicity evidence required to meet thresholds for classification as pathogenic or likely pathogenic.	25
Table 3. Ciliopathy disease genes	57
Table 4. Filtering steps applied in the custom Python script Filter_VEP_output.variants.py	86
Table 5. Variants identified amongst TMEM67 crisprant RPE-1 cell lines	126
Table 6. Electropherograms showing mutations generated amongst crisprant RPE-1 cell lines.	127
Table 7. Example Columbus well views comparing cells transfected with TMEM67_myc wild-type plasmid and un-transfected cells in both wild-type RPE-1 and the TMEM67 knockout RPE-1 cell line C16.....	130
Table 8. List of suppliers for reagents used	228
Table 9. List of reagents used.	229
Table 10. List of buffers and solutions used.....	232
Table 11. List of cell lines used.	234
Table 12. List of primary antibodies used.....	234
Table 13. List of secondary antibodies used.	235
Table 14. List of cell stains used.	235

1 Introduction

1.1 Genetic variant interpretation

Advances in next generation sequencing (NGS) technologies have facilitated the widespread introduction of genomic tests in mainstream clinical settings, as well as in the research environment. These include multi-gene panel, whole exome and whole genome testing strategies. These genomic tests offer exciting new opportunities, as well as challenges.

1.1.1 Rare disease diagnostics

In the European Union (EU), rare diseases are defined as those that affect fewer than 1 in 2,000 people in the general population (Eurordis, 2005). Frequently cited estimates of the number of rare diseases are between 5000-8000, with 70-80% being genetic in origin (Ferreira, 2019, Nguengang Wakap et al., 2020). More recently, RARE-X, the Rare Disease Database Platform based in the United States, estimated that this burden was much higher, with nearly 11,000 unique rare diseases, of which approximately 87% are known to be genetic and a further 13% suspected to be genetic (Lamoreaux et al., 2022). Although individually rare, the cumulative population prevalence of rare disease is estimated at 6.2%, equating to 473 million people affected globally (reference world population 7.6 billion) (Ferreira, 2019).

Accurate diagnosis for patients with rare disorders is essential for their optimal medical management. This is especially important because around half of rare diseases have childhood or prenatal onset, amongst which 30% of those affected will die by their fifth birthday (Global Genes, 2021). The necessity of identifying diagnoses for patients with rare diseases is recognised in the vision of the International Rare Diseases Research Consortium (IRDIRC) 2017 – 2027: to “enable all people living with a rare disease to receive an accurate diagnosis, care and available therapy within one year of coming to medical attention” (Austin et al., 2018).

Achieving a molecular genetic diagnosis is defined as identifying the precise molecular cause (genotype) that explains the clinical features (phenotype) (Wright et al., 2018a). Identifying a diagnosis for every individual with a rare disease is a considerable challenge because of the genetic and phenotypic variability associated with these

conditions, and our incomplete knowledge about their genetic origins (Wright et al., 2018a). Medical professionals face diagnostic challenges in recognising previously unencountered or ultra-rare conditions, and for those with non-specific features with several possible genetic and non-genetic causes. At least half of patients with rare diseases remain undiagnosed despite multiple tests, and for those that do receive a diagnosis, it takes 4.8 years on average to identify (Mattick et al., 2018, Global Genes, 2021). Appropriate selection of a genomic test, facilitating analysis of multiple potentially causative genes at once, can curtail the “diagnostic odyssey” experienced by many patients with rare disorders (Sawyer et al., 2016). Data analysis can be iterative, with new genes and DNA regions analysed if the answer (identification of causative, pathogenic genetic variant(s)) is not identified from initial attempts. This reduces the costs of expensive and sometimes invasive serial testing, including molecular, imaging, and other pathological investigations such as biopsies (Wright et al., 2018a).

Determining the underlying genotype for a patient’s phenotype allows provision of accurate information about their condition, including potential current and future associated features for which screening or treatment may be available. It allows counselling about the mode of inheritance, the chances of family members and future children being affected and facilitates cascade and prenatal testing to those at risk. For conditions approved by the Human Fertilisation and Embryology Authority (HFEA), pre-implantation genetic testing for monogenic or single-gene disorders (PGT-M) may be possible. A molecular diagnosis allows a clearer prognosis to be inferred from previous cases with the same condition. It also enables direction to disorder-specific support groups, reducing the sense of isolation and anxiety for families affected by rare disorders. Having a molecular genetic diagnosis is a pre-requisite for the increasing number of targeted therapeutics becoming available for rare diseases. Despite significant efforts in the field, less than 10% of rare diseases have an approved therapy (Tambuyzer et al., 2020, Lamoreaux et al., 2022).

1.1.2 Genetic variation

The human genome is made of 3.055 billion nucleotides packaged into 23 pairs of chromosomes (Nurk et al., 2022). Variation in DNA sequence from the reference sequence is responsible for normal individual variation but can also cause disease when it disrupts critical gene function.

Genetic variation can occur at every level of DNA resolution: from abnormalities in chromosome number (aneuploidy), large chromosomal structural variants (SVs) (e.g. translocations, inversions), gains or losses of chromosome material (copy number variation (CNV)), all the way down to single base pair (bp) alterations or small insertions and deletions (indels). A typical human genome has 4.1 – 5 million variants from the reference sequence (Genomes Project et al., 2015). Over 99% of these are single nucleotide variants (SNVs), including superficially alarming numbers of seemingly damaging variants identified in healthy individuals (149–182 protein truncating variants, 10,000 - 12,000 missense variants and 459,000 - 565,000 variants overlapping known regulatory regions). Furthermore, a typical human genome contains 2,100 to 2,500 SVs, including ~1,000 large deletions and ~160 CNVs (Genomes Project et al., 2015). Therefore, identifying the single genetic variant responsible for an individual's genetic disease amongst the huge number present is the fundamental challenge of clinical genetics, analogous to finding a needle in a haystack.

Different testing strategies must be adopted to detect different types of genetic variation. Large chromosomal abnormalities are usually detected through cytogenetic tests. Karyotyping has a resolution limit of 5-10 megabases (Mb), and array-comparative genomic hybridisation (array-CGH) around 500 kilobases (kb). Array-CGH is frequently used as a first line investigation for suspected genetic disorders, including structural fetal abnormalities, unexplained learning disability and/or developmental delay, dysmorphism and multiple congenital abnormalities (NHS England, 2022). This is largely related to its low cost and straightforward processing (Nurchis et al., 2022). Karyotyping is usually reserved for specific clinical indications where a cytogenetic abnormality is suspected and if array CGH is uninformative, such as recurrent miscarriage. It is more labour intensive, requiring specialist training, than array CGH.

Sequencing technologies must be used to detect SNVs and CNVs smaller than the resolution achievable through cytogenetic tests, accountable for a significant proportion of disease-causing variants. NGS can be broadly divided into short and long-read strategies. Illumina has emerged as the dominant provider of next-generation sequencers in the last decade due to their lower cost, higher speed, and higher yield than other systems (Midha et al., 2019). Illumina sequencers use a short read, sequencing by synthesis (SBS) approach (Goodwin et al., 2016). They provide short reads of <300bp with an error rate of <1% across their sequencing platforms (Stoler and Nekrutenko, 2021). Short-read NGS is massively parallel, sequencing millions of

fragments simultaneously per run. Parallel individual reads are aligned to the reference sequence, distinguishing true variation from sequencing artefacts through repeated appearance across reads. In contrast, traditional Sanger sequencing only sequences a single fragment at a time. Therefore, Sanger sequencing is now more commonly used to confirm variants identified through NGS than as a first-line diagnostic strategy. Sanger sequencing is also used clinically to perform cascade testing when a pathogenic variant is known in a family, or for conditions caused by a single pathogenic variant, such as achondroplasia (Legare, 2022).

Newer long-read sequencing strategies, such as single-molecule real-time (SMRT) sequencing from Pacific Biosciences (PacBio) and the MinION nanopore sequencer from Oxford Nanopore Technologies (ONT), overcome the read-length limitations of short-read sequencing, but are considerably more expensive and have lower accuracy levels, so far limiting widespread adoption of these technologies (Goodwin et al., 2016).

1.1.3 American College of Medical Genetics and Genomics (ACMG) variant classification

Instructed by the Association for Clinical Genomic Science (ACGS), UK diagnostic laboratories use guidelines from the ACMG to classify the pathogenicity of genetic variants according to the type(s) of available evidence about the variant, and how strongly that evidence is graded (Ellard et al., 2018, Richards et al., 2015). The ACMG guidelines provide scores about the strength of evidence in favour of pathogenicity or benign impact (summarised in Table 1), as well as rules for combining criteria to classify sequence variants into one of five categories: class 1 (benign), class 2 (likely benign), class 3 (uncertain significance), class 4 (likely pathogenic) and class 5 (pathogenic). Combinations of evidence required to meet the threshold for pathogenic or likely pathogenic classification are summarised in Table 2. This process of evidence gathering and collective consideration is known in genetics as variant interpretation.

Table 1. ACMG guidelines on strengths of evidence in favour of variant pathogenicity.

Adapted with permission from (Richards et al., 2015).

	Strength of evidence in favour of pathogenicity			
Type of evidence	Supporting	Moderate	Strong	Very strong
Population Data		Absent in population databases <i>PM2</i>	Prevalence in affecteds statistically increased over controls <i>PS4</i>	
Computational and Predictive Data	Multiple lines of computation evidence support a deleterious effect on the gene / gene product <i>PP3</i>	Novel missense change at an amino acid residue where a different pathogenic missense change has been seen before <i>PM5</i> Protein length changing variant <i>PM4</i>	Same amino acid change as an established pathogenic variant <i>PS1</i>	Predicted null variant in a gene where LOF is a known mechanism of disease <i>PVS1</i>
Functional Data	Missense in gene with low rate of benign missense variants and pathogenic missenses common <i>PP2</i>	Mutational hot spot or well-studied functional domain without benign variation <i>PM1</i>	Well-established functional studies show a deleterious effect <i>PS3</i>	
Segregation Data	Co-segregation with disease in multiple affected family members <i>PP1</i>			
<i>De novo</i> Data		<i>De novo</i> (without paternity & maternity confirmed) <i>PM6</i>	<i>De novo</i> (paternity & maternity confirmed) <i>PS2</i>	
Allelic Data		For recessive disorders, detected <i>in trans</i> with a pathogenic variant <i>PM3</i>		
Other Databases	Reputable source – pathogenic <i>PP5</i>			
Other Data	Patient's phenotype or family history highly specific for gene <i>PP4</i>			

Table 2. AMCG guidance on combinations of lines of variant pathogenicity evidence required to meet thresholds for classification as pathogenic or likely pathogenic.

Adapted with permission from (Richards et al., 2015).

	Pathogenic	Likely pathogenic
Combination 1	1 Very Strong (PVS1) AND a) ≥ 1 Strong (PS1–PS4) OR b) ≥ 2 Moderate (PM1–PM6) OR c) 1 Moderate (PM1–PM6) and 1 Supporting (PP1–PP5) OR d) ≥ 2 Supporting (PP1–PP5)	1 Very Strong (PVS1) AND 1 Moderate (PM1–PM6)
Combination 2	≥ 2 Strong (PS1–PS4)	1 Strong (PS1–PS4) AND 1–2 Moderate (PM1–PM6)
Combination 3	1 Strong (PS1–PS4) AND a) ≥ 3 Moderate (PM1–PM6) OR b) 2 Moderate (PM1–PM6) AND ≥ 2 Supporting (PP1–PP5) OR c) 1 Moderate (PM1–PM6) AND ≥ 4 Supporting (PP1–PP5)	1 Strong (PS1–PS4) AND ≥ 2 Supporting (PP1–PP5)
Combination 4	-	≥ 3 Moderate (PM1–PM6)
Combination 5	-	2 Moderate (PM1–PM6) AND ≥ 2 Supporting (PP1–PP5)
Combination 6	-	1 Moderate (PM1–PM6) AND ≥ 4 Supporting (PP1–PP5)

To undertake this classification, extensive literature and database review is required, approved by an accredited clinical scientist in the diagnostic setting. Clinical information must be integrated into the variant interpretation pipeline before definitive classification and subsequent decision making. This requires consideration of whether pathogenic variants in the gene of interest could be compatible with the patient's phenotype, the mode of inheritance and the functional consequence of the variant's mutational mechanism (e.g. haploinsufficiency, dominant negative effects) (Strande et al., 2017).

1.1.3.1 ACMG strengths of evidence of pathogenicity

Very strong (PVS1)

To qualify as 'very strong' evidence of pathogenicity, the variant must have a "null effect (nonsense, frameshift, canonical ± 1 or ± 2 splice sites, initiation codon, single or multi-exon deletion) in a gene where loss of function (LOF) is a known mechanism of disease".

Strong

There are four lines of 'strong' evidence of pathogenicity:

- 1) PS1: "The variant causes the same amino acid change as a previously established pathogenic variant, regardless of the nucleotide change".
- 2) PS2: "*De novo* (both maternity and paternity confirmed) in a patient with the disease and no family history".
- 3) PS3: "Well-established *in vitro* or *in vivo* functional studies supportive of a damaging effect on the gene or gene product".
- 4) PS4: "The prevalence of the variant in affected individuals is significantly increased compared with the prevalence in controls".

Moderate

'Moderate' evidence of pathogenicity falls into six categories:

- 1) PM1: The variant is in a "mutational hot spot and/or critical and well-established functional domain".
- 2) PM2: The variant is "absent from controls or is found at extremely low frequency if recessive". The huge, publicly available Genome Aggregation Database (gnomAD) (available from <https://gnomad.broadinstitute.org>) makes finding this evidence very straightforward (Karczewski et al., 2020). The latest version span 125,748 exomes and 15,708 genomes (GRCh37) (v2) and 76,156

genomes (GRCh38) (v3.1) sequenced as part of various disease-specific and population genetic studies.

- 3) PM3: “For recessive disorders, detected *in trans* with a pathogenic variant”.
- 4) PM4: The encoded “protein length changes as a result of in-frame deletions/insertions in a nonrepeat region or stop-loss variants”.
- 5) PM5: “Novel missense change at an amino acid residue where a different missense change determined to be pathogenic has been seen before”.
- 6) PM6: The variant is “assumed to be *de novo*, but without confirmation of paternity and maternity” (relevant only to dominant disorders).

Supporting

There are five lines of ‘supporting’ evidence:

- 1) PP1: “Co-segregation with disease in multiple affected family members in a gene definitively known to cause the disease”.
- 2) PP2: “Missense variant in a gene that has a low rate of benign missense variation, and where missense variants are a common mechanism of disease”.
- 3) PP3: “Multiple lines of computational evidence support a deleterious effect on the gene or gene product”. For missense variants, *in silico* predictive software tools perform physical and evolutionary comparative considerations, to assess the impact of amino acid substitutions on the structure or function of a protein. To predict the pathogenicity of a missense substitution, these *in silico* programs consider the evolutionary conservation of an amino acid/nucleotide, its location, and the biochemical consequence of the amino acid substitution. Many tools are publicly available, with PolyPhen-2 (available from <http://genetics.bwh.harvard.edu/pph2/>) (Adzhubei et al., 2013) and SIFT (available from <https://sift.bii.a-star.edu.sg>) (Sim et al., 2012) in regular use in UK diagnostic laboratories. *In silico* splicing tools can be used either as stand-alone programs or as interfaces integrating multiple algorithms. Programs including Human Splicing Finder (available from <http://www.umd.be/HSF3/>) (Desmet et al., 2009) and MaxEntScan (available from http://hollywood.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html) (Yeo and Burge, 2004) incorporate data about splicing signals, splicing regulatory elements, the spliceosome and other trans-acting elements to predict the effects of variants on splicing signals or to identify splicing motifs.
- 4) PP4: “Patient’s phenotype or family history is highly specific for a disease with a single genetic aetiology”.
- 5) PP5: “Reputable source recently reports variant as pathogenic, but the evidence is not available to the laboratory to perform an independent

evaluation”.

1.1.4 NGS-based genomic testing strategies

NGS-based genomic tests are those that use NGS to sequence large stretches of DNA, which can include both coding and non-coding regions. The sequenced DNA can then be selectively analysed according to the clinical or research question through a strategy called virtual gene panel analysis. This means that variants that are in genes on pre-approved lists (panels) relevant to the suspected medical condition are extracted for analysis from the whole dataset. This allows greater opportunities for molecular diagnosis than traditional single- gene testing for conditions demonstrating locus heterogeneity, or where the condition is difficult to recognise clinically and therefore may have different genetic causes.

In whole genome sequencing (WGS), the genome (coding and non-coding regions) is sequenced without prior selection. In whole exome sequencing (WES), DNA regions containing the protein-coding exons are first selectively captured before sequencing. These ~20,000 genes make up only 1% to 2% of the genome but contain >85% of all disease-causing variants (van Dijk et al., 2014). In clinical exome sequencing, only the ~5,000 genes known to have a clinical association with human disease are sequenced.

To date, WES and clinical exome sequencing have been more commonly used than WGS in the clinical context, related to lower costs, generated data volume for storage and the fact that non-coding regions are usually not analysed in standard diagnostic pipelines. However, with falling costs and improving processing capabilities, a move towards WGS is in underway. WGS is less prone to technical artifacts than WES due to fewer preparation steps being required before sequencing (Belkadi et al., 2015). Technical limitations of target-probe hybridisation and/or high GC content in WES can lead to inadequate sequencing depth for some regions leading to uneven capture or complete exon skipping. This can lead to poor accuracy scores of potentially pathogenic variants which are either missed completely, or excluded from further analysis, producing false negatives (Mattick et al., 2018). This is a particular problem in potential diagnostic blind spots including SVs, pseudogenes, and repetitive regions (Hannan, 2018, Mallawaarachchi et al., 2016, Noll et al., 2016). WGS, particularly PCR-free WGS, provides the opportunity to identify CNVs and SVs due to even coverage of the whole genome, not available through WES (see thesis section 1.1.5.2). Some previously undetectable complex SVs, such as balanced inversions, are detectable through WGS with or without complementary cytogenetic tests (Schuy et

al., 2022). Furthermore, WGS facilitates the opportunity to analyse intronic regions not covered in WES, likely to be a source of hidden pathogenic variants (see thesis section 1.1.5.1).

Compared with traditional testing approaches, WES and WGS have provided improved diagnostic rates for patients with rare disorders. In the UK, the Deciphering Developmental Disorders Study (DDD) (<https://www.ddduk.org>) undertook WES for children with developmental disorders undiagnosed through targeted single-gene and cytogenetic testing, identifying an underlying genetic diagnosis in 40% (Wright et al., 2018b, Wright et al., 2015). A 2018 meta-analysis of 37 studies, comprising over 20,000 children, revealed a diagnostic yield of 25-30% from WES and WGS in children with suspected genetic disorders, with both providing significantly greater diagnostic utility than array CGH (Clark et al., 2018). This study reported that the diagnostic utility of WGS and WES were not significantly different from one another. In the same year, a meta-analysis of 29 studies undertaking WES and nine studies performing WGS for a much broader range of suspected genetic conditions, including paediatric and adult-onset disorders, reported that WGS nearly doubles the diagnosis rate compared to WES (range from WES = 25-35%, weighted average = 28%; vs WGS range = 40-60%, weighted average = 49%) (Mattick et al., 2018). However, the authors acknowledge that the range of conditions and cohorts between studies were very different, meaning direct comparison is difficult, and the technical improvements in WES over the inclusion period led to better diagnosis rates from WES later in the study window (weighted average diagnostic yield 2013-2014 = 26% vs 2017-2018 = 31%).

1.1.5 Sources of missed genetic diagnoses

Despite advances in sequencing technologies and available diagnostic tests, clearly a significant proportion of individuals with genetic disorders still have undetected molecular diagnoses. There are several reasons underlying this, which can be split into practical problems detecting the causative variant (inappropriate test selection or sequence quality), and problems recognising the right variant as causative. The latter includes variants in genes not previously associated with human disease and variants with functional consequences that are not captured by standard variant filtering approaches. Recent estimates are that there are more than 1000 developmental disease genes yet to be identified, which get harder and harder to find due to reduced penetrance and high levels of pre- and perinatal mortality (Kaplanis et al., 2020).

Genomic tests provide the opportunity to analyse any variant identified, but the sheer numbers generated understandably drive a practical need for filtering prior to analysis. Common filtering strategies in rare disease diagnostics include removal of variants common in the general population (see description of gnomAD, section 1.1.3.1), removal of non-coding variants, removal of seemingly benign variant types (e.g. synonymous, in-frame indels) and removal of variants in genes not on selected virtual gene panels.

1.1.5.1 Missed pathogenic splice and non-coding variants

Splice variants are already recognised as an important contributor to genetic disease. The public variant pathogenicity database ClinVar (Landrum et al., 2016) contains 209,123 pathogenic and/or likely pathogenic variants, of which 24,825 (11.9%) are entered as affecting splice sites (data correct as of 09/11/2022). However, this number is likely to be hugely under-representative of the total pathogenic variant burden due to limitations in the recognition and interpretation of splice variants.

Splicing is a complex, tightly regulated process that transforms freshly transcribed pre-messenger RNA (mRNA) to mature mRNA, ready to be translated into protein. This involves removing non-coding sequence (introns) by cleavage at conserved sequences called splice sites and splicing the remaining coding sequence (exons) back together. Splicing occurs in the nucleus coordinated by the *trans*-acting spliceosome protein-RNA complex (Anna and Monika, 2018). This interacts with *cis*-acting elements, which are DNA sequences that define exons, introns, and regulatory sequences necessary for proper splicing. These include the two-nucleotide canonical splice sites at either side of each exon: an “AG” motif upstream of the acceptor (3’), and “GT” motif downstream of the donor (5’). Other important *cis*-acting elements include the polypyrimidine tract, branch point and auxiliary elements such as splicing silencers and enhancers (Lord and Baralle, 2021).

Splicing can be disturbed by disruption to any *cis* or *trans* acting element. Incorrect splicing can lead to exon skipping or the introduction of novel splice sites, which alter the reading frame of protein-coding genes. Alternatively, missed splicing causing intron retention incorporates non-coding DNA into mature mRNA, which often contains stop codons and therefore leads to premature protein truncation.

Canonical splice site variants are already recognised as causing LOF with a null effect (Richards et al., 2015). The complex mechanisms of splicing are not yet fully understood, impairing our ability to determine whether identified variants outside the canonical splice sites will disrupt splicing and how damaging this may be. Cryptic splice variants disrupt mRNA splicing despite lying outside the canonical splice sites and have been recognised as having an important role in genetic disease for many years (Cooper et al., 2009). These may be deep- intronic, near splice-site or exonic and labelled as alternative variant types (particularly missense and synonymous). However, tools to interpret cryptic splice variants are still in development. Data from *in silico* prediction tools can only provide supporting evidence for ACMG variant assessments (see thesis section 1.1.3.1), preventing classification as anything more definitive than a variant of uncertain significance (VUS) (see thesis section 1.1.7.1) in the absence of functional experiments which are usually limited to the research setting (Richards et al., 2015).

Non-coding DNA, historically labelled as “junk”, is proving to have much more important roles in gene regulation and expression than previously thought. Gene transcription is mediated by a promotor element directly upstream of a gene as well as through binding of transcription factors to more distal enhancer and repressor elements. Gene expression is also regulated at the post-transcriptional level, controlled by the 5' and a 3' untranslated regions (UTRs) in mature mRNA, which regulate RNA stability, trafficking, and the rate at which it is translated into protein. Collectively, disruptive variation in non-coding regions has been shown to cause severe disease by affecting splicing, transcription, translation, chromatin stability and RNA processing and stability (Ellingford et al., 2022). This contribution to disease burden has been shown to be significant for some conditions. For example, non-coding region variants causing LOF represent 23% of likely diagnoses identified in *MEF2C* in the DDD cohort (Wright et al., 2021).

Several founder non-coding pathogenic variants have been known for many years. These are ancestral variants which rose to relatively high frequency in a given population and are now shared by families with the resultant phenotype. For example, the *CEP290* deep-intronic variant c.2991+1655 A>G accounts for up to 15% of the early onset blindness condition Leber Congenital Amaurosis (LCA) (Sallum et al., 2020, Coppieters et al., 2010a), and represents the majority of the *CEP290* pathogenic variant burden for patients with LCA (Testa et al., 2021, Feldhaus et al., 2020). Functional experiments showed that this variant creates a cryptic splice donor site,

resulting in the insertion of an aberrant pseudoexon with a premature stop codon into ~50% of all *CEP290* transcripts (den Hollander et al., 2006). Correct identification of splicing variants is extremely important, not only to provide molecular diagnoses to unsolved patients but because they are an area for development of new targeted therapeutics such as antisense oligonucleotides (AONs). The RNA AON Sepofarsen targeted to the *CEP290* variant c.2991+1655 A>G is in clinical trials, demonstrating significant improvements in visual acuity and retinal sensitivity and a manageable safety profile (Russell et al., 2022, Xue and MacLaren, 2020).

1.1.5.2 Missed structural variants

SVs are usually defined as changes of at least 50 nucleotides. They can be balanced or unbalanced and defined as canonical (two breakpoints) or complex (three or more breakpoints) (Quinlan and Hall, 2012). It is very difficult to ascertain the contribution of SVs to rare disease that are undetectable on cytogenetic tests (if undertaken) because systems to prioritise them for analysis from genomic data are not well established. These include balanced or unbalanced SVs smaller than the resolution of array CGH (500kb) and balanced changes that are too small or complex to be seen on karyotyping. Many software packages are available to call CNVs and SVs from short read WGS data, such as Manta (Chen et al., 2016) and Canvas (Ivakhno et al., 2018). Manta is an SV caller that based on breakpoint analysis, whereas Canvas is a CNV caller mainly based on coverage. The two work well in parallel; Manta is better for picking up SVs like translocations and CNVs <10kb, whereas Canvas is better at picking up larger CNVs. However, high false positive rates and a lack of consistent filtering strategies make accurate identification and interpretation of pathogenic SVs challenging.

Published data from WGS studies has already demonstrated the value added through SV analysis. A study led by a Swedish team evaluated WGS with SV analysis as a first-line investigation for intellectual disability (Lindstrand et al., 2019). They undertook WGS for three cohorts: (i) a retrospective cohort with validated CNVs (n = 68); (ii) individuals referred for monogenic multi-gene panels (n = 156); (iii) prospective cases referred for array CGH (n=100). As well as validating 92 previously known SVs through WGS, they detected 11 new SVs, improving the diagnostic yield. More recently, the same group undertook explicit non-SNV analysis for 285 patients undergoing WGS for multiple clinical indications, finding 35 (12%) with non-SNV variants (Stranneheim et al., 2021). As the non-SNV analyses were implemented gradually, it is impossible to ascertain the exact number analysed for these variants in their cohort. However, the

authors reported that 45 identified non-SNVs are CNVs (70% of cases), with five balanced rearrangements, two complex SVs, ten short tandem repeat (STR) expansions and one maternal uniparental disomy (UPD) also reported. In the UK, analysis of WGS data from 650 unsolved inherited retinal dystrophy patients revealed 33 pathogenic SVs from 31 individuals (4.8% diagnostic uplift) (Carss et al., 2017).

1.1.6 The 100,000 Genomes project (100K)

The 100,000 Genomes project (100K) is a hybrid clinical/research initiative, launched in 2012 by Genomics England (GEL) as part of the UK's Life Sciences Strategy (Turnbull et al., 2018). The project aimed to sequence 100,000 genomes from individuals with rare diseases and cancer alongside their family members in a trio testing approach and link this sequence data to clinical data from longitudinal patient records. To take part, participants had to consent to the clinical arm of the project, i.e. to receive a diagnosis should one be identified, and to the research arm, including access to their past, present, and future genetic and medical records for approved academic and commercial researchers. They also had the option for an opportunistic search for additional findings, such as for inherited cancer predispositions and reproductive carrier risks. Short-read genome sequencing was performed using Illumina 'TruSeq' library preparation kits for read lengths 100 bp and 125 bp (Illumina HiSeq 2500 instruments), or 150 bp reads (HiSeq X). These generated a mean read depth of 32× (range, 27–54) and a depth >15× for at least 95% of the reference human genome (Wheway et al., 2019).

GEL adopted a "tiering" system which prioritised variants for analysis by regional National Health Service (NHS) diagnostic laboratories. This is described in detail under "tiering issues" (Best et al., 2022a) (manuscript section 2) in our published 100K commentary article. In summary, clinical assessment was only expected for prioritized Tier 1 (protein damaging) and Tier 2 (protein altering) SNVs affecting coding sequences and splice donor or acceptor sites, in genes on selected panel(s). All other SNVs, CNVs and SVs were not systematically analysed in the whole cohort.

GEL also provide PanelApp (available from <https://panelapp.genomicsengland.co.uk>), a crowdsourcing tool for sharing and evaluation of curated gene panels by the scientific community (Martin et al., 2019). PanelApp provides a traffic light system for genes: 'green' genes are diagnostic grade, 'amber' genes are borderline and 'red' genes have a low level of evidence.

Recruitment to the main 100K program was delayed due to oncology sample problems identified during the pilot project. Sequenced DNA from traditional Formalin-Fixed Paraffin- Embedded (FFPE) tissue specimens produced noisy profiles with large numbers of artefacts, requiring a move to using only frozen tissue samples. This heralded a nationwide sea-change in pathology departments, including rapid transfer and processing of samples between operating theatres and pathology departments.

Recruitment to 190 different rare disease domains took place between 2016 and 2018 across 85 NHS Trusts, coordinated by 13 Genomic Medicine Centres (GMCs). A preliminary report from the pilot study of 4660 rare disease participants reports a genetic diagnosis in 25% of probands (The 100,000 Genomes Project Pilot Investigators et al., 2021). Interestingly, 14% of these diagnoses were made through a combination of GEL's automated tiering system and research collaborations, proving especially important in identification of pathogenic non-coding and structural variants. Data from the main program is still emerging and has not been comprehensively summarised.

The longer-term aim of the 100K is to fully integrate genomic testing for eligible patients within existing NHS healthcare pathways. In October 2018, the new NHS Genomic Medicine service was established as a follow on from the 100,000 Genomes Project (Department of Health and Social Care, 2019). This provides a curated National Genomic Test Directory, specifying which genomic tests are commissioned by NHS England (NHS England, 2022). The Test Directory sets out the technology by which tests are available, including WES and WGS where appropriate, and the patients who are eligible to access commissioned tests. It is subject to extensive annual review by national clinical and scientific experts, existing genetic laboratory staff, patient and public representatives and organisations.

The UK Government policy paper "Genome UK: the future of healthcare", published in September 2020, committed to sequence at least 500,000 whole genomes in England by 2024, and to offer WGS for "seriously ill children who are likely to have a rare genetic condition, children with cancer, and adults suffering from certain rare conditions or specific cancers" (Gov.uk, 2020). However, in the latest version of the National Genomic Test Directory (v3.1, August 2022), WGS is only recommended for 33/592

(5.6%) of clinical indications and WES for 63/592 (10.6%) (NHS England, 2022). Despite the promises of the 100K, the House of Commons Science and Technology Committee reported that the roll out of the NHS Genomic Medicine Service has been held up by delays in digital infrastructure, insufficient training and a lack of qualified staff, and ethical concerns over use of patient data (Parliament.uk, 2018).

1.1.7 Challenges in genomic testing

Although genomic tests are more successful in detecting clinically significant findings compared to traditional cytogenetic or single-gene tests, they also increase the chance of detecting variants of uncertain significance (VUS) and incidental findings, providing challenges to both clinicians and patients.

The tools and technologies used to interpret DNA sequence variants are not as advanced as the NGS tools used to generate the sequence in the first place. Using readily available tools, a significant proportion of genetic variants remain difficult to interpret, limiting their clinical utility.

1.1.7.1 VUS

Class 3 variants are also known as VUS. These are genomic variants which cannot be definitively classified as pathogenic or benign because of inadequate or conflicting available evidence. With genomic tests now including hundreds to thousands of genes, generation of VUS results is dramatically increasing (Hoffman-Andrews, 2017). Missense and non-canonical splice variants pose particular challenges. Often, the only lines of available evidence are 'moderate' (absent from population controls) and 'supporting' (*in silico* tools support a deleterious effect), which together are not enough to meet the threshold for a 'likely pathogenic' classification. The ACMG advises that 'efforts to resolve the classification of the variant as pathogenic or benign should be undertaken' when VUS are identified (Richards et al., 2015). However, it is unclear how far this should be pursued by the clinical team, and the expenditure of time and resources to ensure this classification can be prohibitive (Feldman, 2016). Currently, functional work to provide additional 'strong' evidence is largely limited to the research setting, done on a case-by-case basis where resources are available and interested researchers are involved.

The inherent uncertainty of a VUS result is challenging both for clinicians and patients. The ACMG advises that a VUS result cannot be used in clinical decision making (Richards et al., 2015). Not only does this apply to the index patient, but to cascade testing for other family members, and to prenatal testing. If reported to patients, VUS can cause significant anxiety and make decision-making challenging (Han et al., 2017, Makhnoon et al., 2019).

Clearly, better tools are required to allow more definitive interpretation of genetic variants. Given the wealth of genomic data now available, there is a pressing clinical need to provide systematic functional interpretation of VUS since this is essential for accurate molecular diagnosis. For functional testing to be incorporated into standard variant interpretation and to be deliverable in the mainstream clinical setting, it would need to be accurate, quick, affordable, and easily interpretable.

1.1.7.2 Incidental findings

Incidental findings are results of potential clinical significance that are unexpectedly discovered and unrelated to the purpose of the test. The discovery of a result inferring an unanticipated medical condition or predisposition can cause significant anxiety and upset. Deciding whether to return such findings is a challenge for both clinicians and patients, that must be carefully accommodated within the testing consent procedure. The ACMG advises to only report incidental findings of conditions that are medically actionable, meaning that there is effective screening and/or treatment available for that condition, amongst which there are high levels of concordance amongst specialists (Kearney et al., 2011, Green et al., 2012). Challenging incidental findings are typically discussed at multi-disciplinary meetings of relevant clinicians and scientists, and a consensus is reached on whether they should be returned. The chances of identifying incidental findings can be reduced by applying phenotype-specific gene panels or targeted interpretation of sequence data, restricting the subsequent debate about whether to return identified results.

As well as revealing unexpected medical conditions, trio testing approaches can also reveal unforeseen issues such as non-paternity or parental consanguinity. Genomic tests can identify misattributed relationships with greater certainty than single-gene tests. The unintentional disclosure of such findings creates an ethical dilemma between our duty to inform and the value of truthfulness, and our reluctance to disrupt

relationships within a family. Both good clinical practice and good research governance stipulate consideration of the relative benefits and harms of disclosing information beyond the realm of the original inquiry (Wright et al., 2019). The 100,000 Genomes Project and the DDD study both had explicit statements that they would never reveal information about misattributed parentage. However, the hybrid clinical/research natures of these studies can cause scenarios where such promises clash with opinions about good clinical practice, for example where realities about familial relationships are directly relevant to clinical care.

It is important that the possibilities of identifying parental consanguinity or misattributed parentage are discussed clearly and explicitly during the consent procedure when undertaking genomic tests. If identified, responsible clinicians must use their judgement on a case-by-case basis, usually involving a multidisciplinary team to reach a consensus decision.

1.2 Cilia

The cilium was the first identified cellular organelle, described in protozoa in 1675 by Antony van Leeuwenhoek as “incredibly thin feet, or little legs” (Dobell, 1932). Cilia are microtubule-based, hair-like organelles. They have highly conserved structure and function and are found ubiquitously across species from nematodes to ancient protozoa (Mitchison and Valente, 2017). Despite being considered vestigial for decades, recent studies have shown that cilia are essential for multiple key biological processes.

1.2.1 Cilia types

Eukaryotic cells contain both primary and motile cilia, which have distinct structures and functions. Primary and motile cilia are distinguished from one another by the number of cilia found on the cell, and on the microtubular structure observed on cross-section (See Figure 1). Primary cilia are found as a single monocilia on the cell surface, in contrast to motile cilia, which are found as multiple cilia.

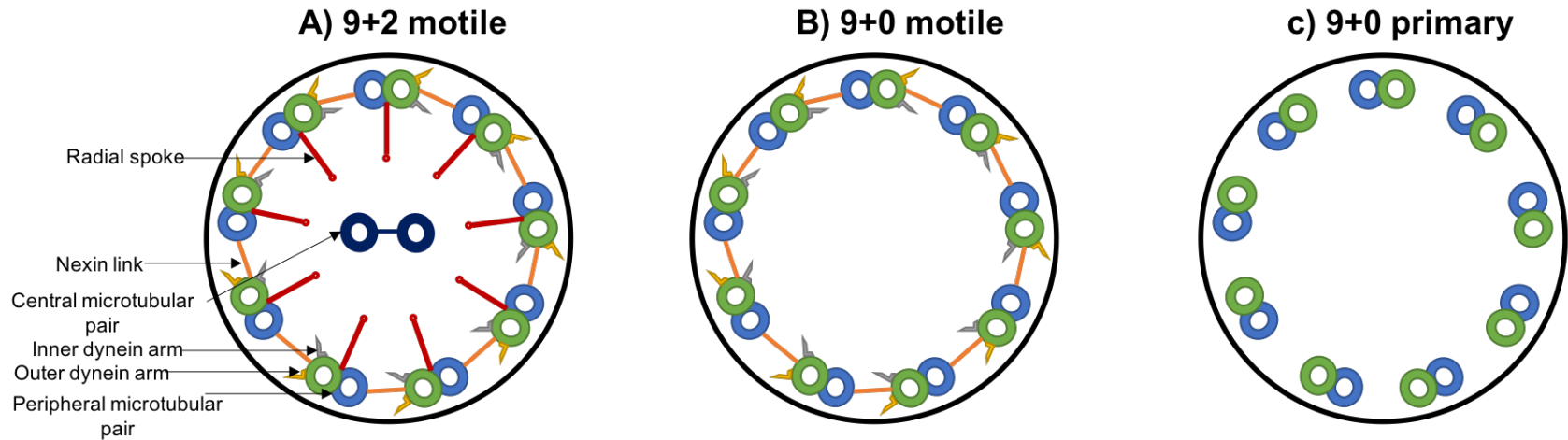


Figure 1. Schematic of transverse cross section of cilia

A) 9+2 motile cilium cross section. As well as 9 peripheral microtubular pairs and a central pair, 9+2 motile cilia contain accessory elements including inner and outer dynein arms, nexin links and radial spokes. B) 9+0 motile cilia, found transiently in the embryonic node, lack the radial spokes and central pair seen in 9+2 motile cilia. C) 9+0 primary cilia lack all accessory elements, consisting simply of 9 peripheral microtubular pairs.

1.2.1.1 Primary cilia

Primary cilia act as cellular 'antennae', transducing diverse signals from the extracellular environment and other cells to their cell body. These signals include proteins, low molecular weight chemicals, mechanical stimuli, and light (Malicki and Johnson, 2017). A single primary cilium projects from the surface of most vertebrate cells contained within a specialised extension of the cell plasma membrane (Malicki and Johnson, 2017). They are dynamically regulated through the cell cycle, present in G0 and G1 cells and usually in S/G2 cells, resorbed before mitotic entry and then reappear after cytokinesis (Plotnikova et al., 2009). Primary cilia assemble and specialise in function when cells differentiate (Sanchez and Dynlacht, 2016).

The immotile (9+0) primary cilium contains nine outer microtubular doublets, and lacks the other accessory elements found in motile cilia (see Figure 1.C). Therefore, it cannot generate its own movement. Primary cilia have diverse roles in homeostasis, embryonic development and sensory perception (see thesis section 1.4) (Malicki and Johnson, 2017, Valente et al., 2014, Reiter and Leroux, 2017). Cells lacking a primary cilium include hepatocytes, mature adipocytes, and skeletal muscle (Sanchez and Dynlacht, 2016). However, regeneration of vertebrate skeletal muscles requires a type of stem cell called satellite cells, for which primary cilia provide an intrinsic cue essential for self-renewal (Jaafar Marican et al., 2016).

1.2.1.2 Motile cilia

The motile ciliary axoneme has a canonical 9+2 microtubular pattern, composed of nine peripheral microtubular doublets surrounding a central pair of microtubules (see figure 1.A). The peripheral microtubular doublets are studded along their length with inner dynein arms (IDA) and outer dynein arms (ODA). These dynein arms contain kinase domains with adenosine triphosphate (ATP)-ase activities that act as molecular motors to allow sliding of adjacent peripheral microtubular pairs during ciliary beating. Other accessory elements include radial spokes which extend from the peripheral doublets to the central pair, and nexin links which connect adjacent microtubular pairs, and help to coordinate dynein arm activity within the nexin–dynein regulatory complexes (N-DRC) (Bower et al., 2013). Together, these accessory elements provide a scaffold for the 9+2 structure, which allows the cilia to bend and govern the waveform (Shoemark and Hogg, 2013). Embryonic nodal motile cilia have slightly different ultrastructure to motile cilia found elsewhere in the body in that they lack a central pair and the radial spokes and are therefore called 9+0 motile cilia (see figure 1.B) (Basu

and Brueckner, 2008).

Motile cilia beating plays an essential role in cell motility and transport of fluids over mucosal surfaces on the surface of epithelial cells lining the respiratory tract and inner ear, the ventricles of the brain and the Fallopian tubes. Embryonic nodal motile cilia are present transiently during early development in the embryonic node, where they provide a rotary motion to direct the establishment of the body's left-right axis, and subsequent laterality of organ positioning (Basu and Brueckner, 2008, Nonaka et al., 1998, Best et al., 2019).

1.2.2 Cilia structure and ciliogenesis

The cilium has a core structure called the axoneme, formed of nine parallel microtubular doublets (Malicki and Johnson, 2017). The axoneme can vary from 1-9 μ m in length, depending on the cell type (Dummer et al., 2016). The diameter of the ciliary membrane is approximately 250–300 nm (Yang et al., 2015). The axoneme extends from a centriolar- anchor called the basal body (BB), located at the base of the cilium (see Figure 2). In the BB, the mother and daughter centrioles align at 90° to one another, and the mother centriole acts as a matrix for microtubule nucleation during formation of the cilium.

Ciliogenesis occurs in quiescent cells in a set of ordered steps. Firstly, the centrosome, consisting of mother and daughter centrioles, migrates to the cell membrane. It docks onto the actin-rich framework via fibrous distal and sub-distal appendages and matures into the BB. The orientation and positioning of the BB determines the alignment of the resulting cilium (Ishikawa and Marshall, 2011). After docking, the BB nucleates outgrowth of axonemal microtubules that protrude beneath the cell membrane, giving rise to the cilium. Extension of the cilium through assembly of outer doublets occurs exclusively at the distal tip.

As synthesis of proteins needed for elongation of the cilium is restricted to the cytoplasm, the required proteins must be selectively imported and transported to the tip through a process called intraflagellar transport (IFT). Protein cargo is transported bidirectionally with anterograde and retrograde IFT mediated by kinesin-2 and cytoplasmic dynein-2 motors respectively (Taschner and Lorentzen, 2016). IFT mediates both the assembly and resorption of the cilium, and the trafficking of key

components of signalling cascades.

To enter the cilium, proteins are imported through the ciliary gate region found just distal to the BB, consisting of transition fibres (TFs) and the transition zone (TZ) (Figure 2) (Satir and Christensen, 2007, Garcia-Gonzalo and Reiter, 2012). TFs anchor the mature mother centriole to the cell plasma membrane. It is in the TZ that triplet microtubules become doublets (Gibbons, 1961). The highly organised TZ contains a selective barrier to protein trafficking into and out of the cilium. This is connected to the doublet microtubules by Y-shaped linkers (Figure 2) (Gilula and Satir, 1972, Satir, 2017). Further detail about the TZ can be found in thesis section 1.2.3.1.

primary cilium: ultrastructure

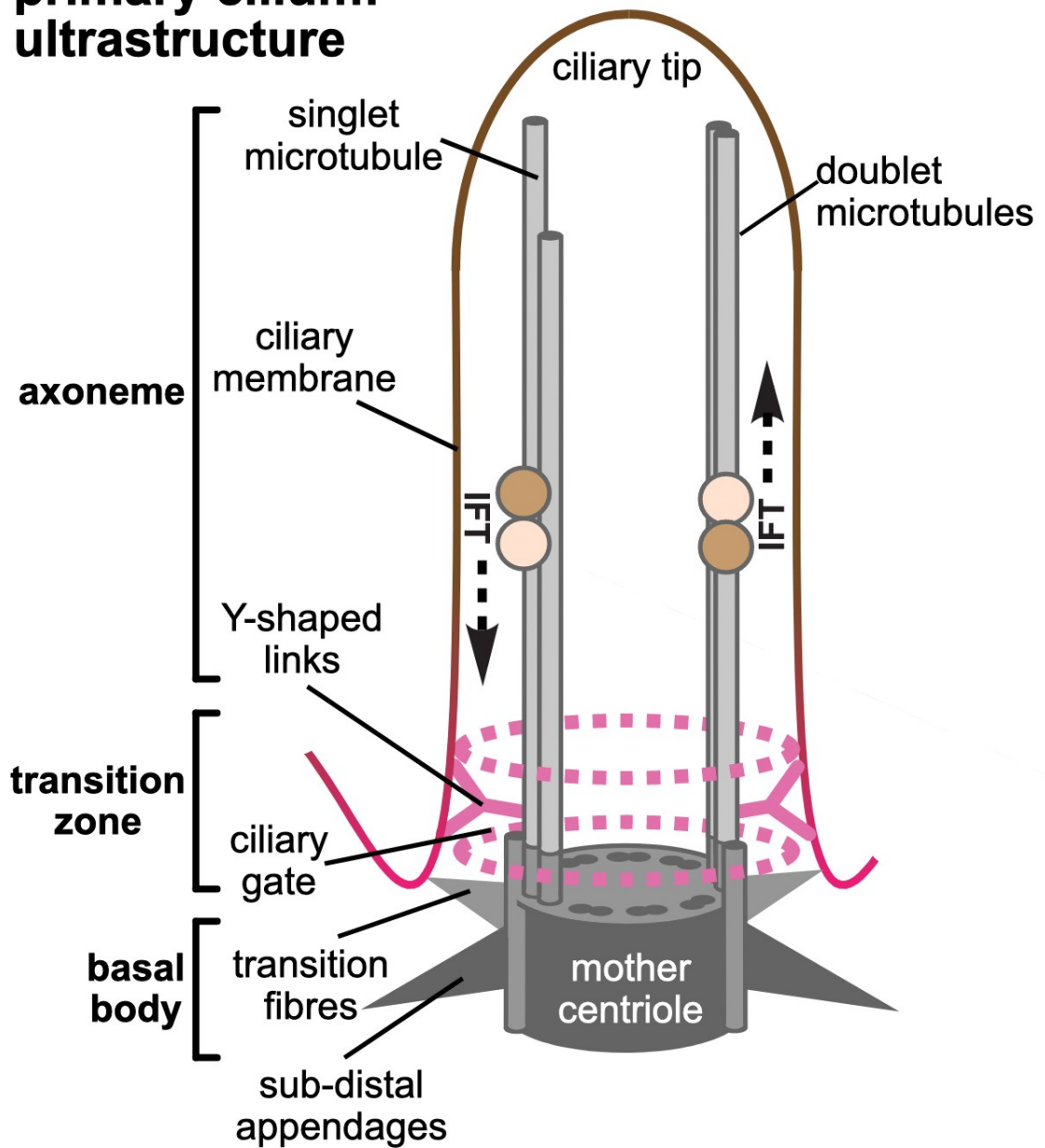


Figure 2. Schematic of the structure of the primary cilium.

The main substructures of the cilium are the axoneme, transition zone (TZ), and basal body (BB). Within the axoneme, selectively imported proteins are transported to the ciliary tip by intraflagellar transport (IFT).

1.2.3 Ciliary proteins

An abundance of proteins is known to be involved in the structure and function of ciliary structures and associated signalling pathways. CiliaCarta, a comprehensive online ciliary compendium, contains 956 putative ciliary genes based on systematically integrated genomic, proteomic, transcriptomic, and evolutionary data (van Dam et al., 2019). The authors estimate the total size of the human ciliome to be approximately 1200 genes.

The location and function of known protein components of the cilium is not yet fully understood, and extensive research is ongoing to determine this. Many proteins are found in multiple locations at different points of ciliogenesis.

1.2.3.1 The Transition Zone

The TZ acts as a diffusion barrier that restricts entrance and exit of membrane and soluble proteins to regulate ciliogenesis and receptor localisation for essential signalling pathways (see thesis section 1.4) (Garcia-Gonzalo et al., 2011). It is thought to control the composition of essential ciliary compartments including the ciliary membrane, axoneme, and associated proteins (Williams et al., 2011). It is clearly an important ciliary region, as this tiny structure is home to a significant number of ciliopathy-related proteins (see thesis section 1.3.1.2) (Szymanska and Johnson, 2012). Disruption of ciliary TZ architecture has been shown to cause Joubert Syndrome (JBTS), but the molecular mechanisms by which this disruption leads to ciliopathy phenotypes remains a subject of ongoing research (Shi et al., 2017).

1.2.3.2 Photoreceptors

A specialised type of primary cilia is found in the retinal photoreceptors in the eye. Photoreceptors, the light sensing cells of the eye, are divided into inner segments (IS) and outer segments (OS), which are joined by a connecting cilium (CC). The photoreceptor OS develops from a primitive primary cilium and consists of stacked membrane discs that contain components of the phototransduction cascade, organised around an axoneme (Bachmann- Gagescu and Neuhauss, 2019, Wheway et al., 2014). The OS is anchored inside the IS of the cell body through the CC, homologous to the TZ of a primary cilium. The OS lacks translational machinery, therefore all proteins required for phototransduction are made in the IS and regulated

and trafficked through the CC, enabling biochemical purification of OS components (Szymanska and Johnson, 2012).

1.3 Ciliopathies

Inherited pathogenic variants leading to abnormalities of motile and primary cilia structure or function result in a group of genetic conditions known as ciliopathies (Reiter and Leroux, 2017). Although individually rare, collectively ciliopathies are thought to affect up to 1 in 2000 people based on three frequent clinical features: renal cysts (1 in 500 adults), retinal degeneration (1 in 3000), and polydactyly (1 in 500) (Quinlan et al., 2008). There is considerable phenotypic and genetic heterogeneity between the 35 individual ciliopathy syndromes (Mitchison and Valente, 2017, Reiter and Leroux, 2017).

Unfortunately, very few treatment options are currently available for the majority of ciliopathies (Molinari and Sayer, 2017). To be able to deliver better diagnostic rates, prognostic information and targeted therapies, further work must be done to understand the genetic aetiology of ciliopathies.

1.3.1 Clinical features

Ciliopathy syndromes exist on a clinical spectrum, related to the strength and nature of the underlying causative variant. They range from relatively, common single-system disorders such as retinal or renal ciliopathies, through to rare, complex, multi-system syndromes. The variety in systems involvement reflects the critical role of cilia in development and health (Wheway et al., 2019). There is extensive overlap in clinical features that are characteristic of ciliopathies between different syndromes, shown in Figure 3. Frequently affected systems include the central nervous system (developmental delay and structural brain abnormalities), the skeletal system (polydactyly and thoracic dystrophy), ophthalmic system (pigmentary retinopathy) and the renal system (cystic kidneys).

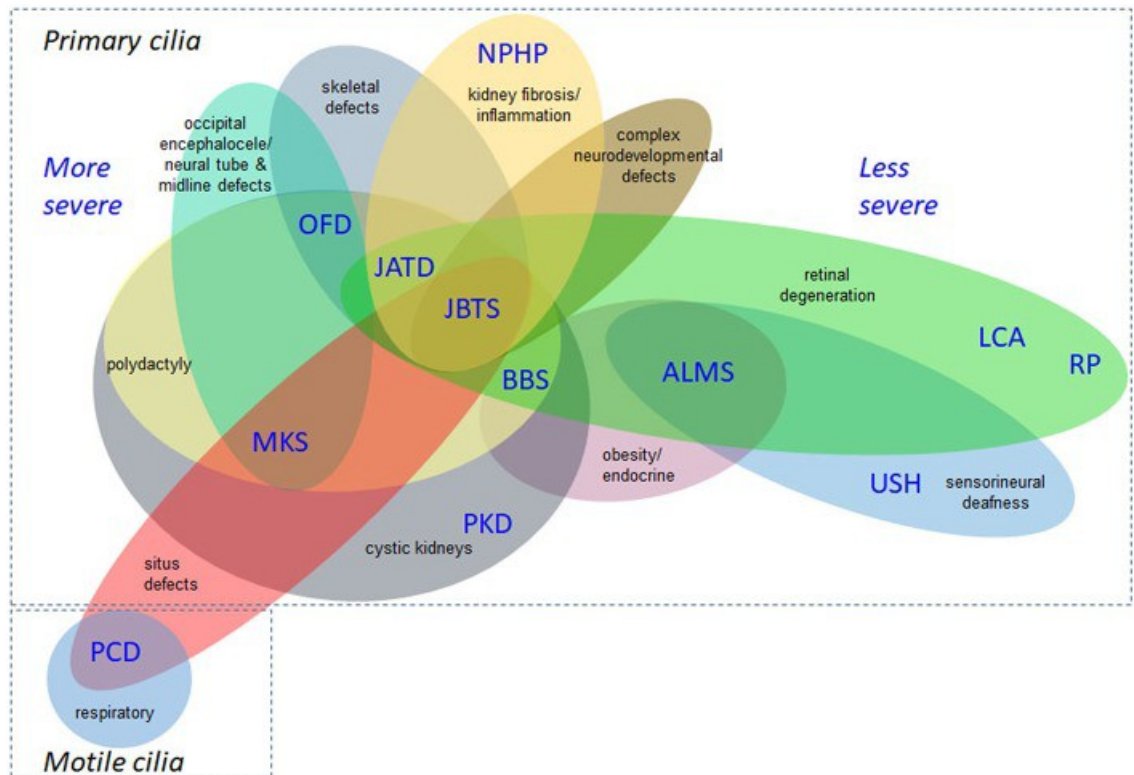


Figure 3. Overlapping disease features of the ciliopathies.

Copied from (Whewey et al., 2019). Copyright © 2019 Whewey, Genomics England Research Consortium and Mitchison. This image is re-used under the Open Access CC-BY 3.0 license.

This illustrates the complex and overlapping features of the different ciliopathy syndromes. Note: the severity indicator reflects the number of clinical features involved in each condition, and how life limiting/threatening they are. For example, MKS and OFD can include encephalocele/neural tube defects as well as several other features associated with significant morbidity from multiple organ systems, so are found at the severe end of the ciliopathy disease spectrum. Although LCA can be considered a severe condition in that it is associated with blindness from birth or early childhood, it is an eye-specific disorder, not involving other organ systems, and compatible with a normal lifespan. By this measure, it therefore is classified as a less severe ciliopathy syndrome.

The extreme genetic heterogeneity of ciliopathies is demonstrated by different variants in individual genes causing dramatically different phenotypes. For example, *CEP290* variants can cause the perinatal lethal multisystem Meckel Gruber Syndrome (MKS), through to non-syndromic LCA, a blinding disorder affecting just the retina (see Table 3) (Coppieters et al., 2010b, Drivas et al., 2013). The variable expressivity of ciliopathy phenotypes complicates diagnostic and prognostic testing, since it extends to intra-familial variation for individuals that carry the same pathogenic variant within families (Valente et al., 2010) and even between monozygotic twins (Hsia et al., 1971).

1.3.1.1 Motile ciliopathies

The motile ciliopathy primary ciliary dyskinesia (PCD) arises from dysfunction of motile cilia (see thesis section 1.2.1.2) (Lucas et al., 2014, Boon et al., 2014). PCD has an estimated prevalence of one in 10,000 (Lucas et al., 2014). The incidence has been observed to be higher in some ethnic groups, particularly among consanguineous populations (O'Callaghan et al., 2010). PCD is characterised clinically by oto-sinus disease, chronic lung disease, reduced fertility, and organ laterality defects in approximately 50% of patients (Best et al., 2019). There are also syndromic forms of PCD, such as X-linked PCD associated with retinitis pigmentosa (RP), due to pathogenic variants in the *RPGR* gene (Moore et al., 2006). To date, pathogenic variants in almost 50 genes are known to cause motile ciliopathies (Horani and Ferkol, 2021), providing molecular diagnoses for up to 75% of cases (Wheway et al., 2019, Marshall et al., 2015, Paff et al., 2018).

1.3.1.2 Primary ciliopathies

1.3.1.2.1 Neurodevelopmental ciliopathies

The severe multi-system primary ciliopathies MKS (Hartill et al., 2017), JBTS, (Bachmann- Gagescu et al., 2015), Bardet Biedl Syndrome (BBS) (Forsythe and Beales, 2013) and oral- facial-digital syndrome (OFDS) (Gurrieri et al., 2007) feature neurodevelopmental phenotypes, alongside combinations of other features such as retinal dystrophy, skeletal abnormalities, and renal dysplasia (see Figure 3 & Table 3) (Waters and Beales, 2011). The frequent association of syndromic ciliopathies with a retinal dystrophy phenotype is reflective of the importance of the specialised photoreceptor cilium in the function of the retina.

Meckel Gruber Syndrome (MKS)

MKS is the most severe ciliopathy syndrome. It is characterised clinically by posterior fossa abnormalities (most frequently occipital encephalocele), bilateral enlarged cystic kidneys, postaxial polydactyly and liver defects including ductal plate malformation associated with hepatic fibrosis and cysts (Hartill et al., 2017) (see Figure 4). MKS is lethal *in utero* or immediately after birth, usually due to pulmonary hypoplasia. It has autosomal recessive inheritance and is more common amongst consanguineous populations including in Saudi Arabia and Kuwait (Teebi et al., 1992, Teebi and Teebi, 2005). The worldwide incidence of MKS has been estimated at 1 in 135,000 live births, although it is more common in certain populations such as Finnish (1:9000) and Gujarati Indians (1:1300) (Auber et al., 2007, Salonen and Norio, 1984, Young et al.,

1985). The OMIM MKS Phenotypic Series PS249000 contains ten morbid genes and two provisional disease genes (data accessed 20/12/2022), detailed in Table 3. All these genes localise to the TZ, apart from TXNDC15, which has unknown localisation (Van De Weghe et al., 2022).



Figure 4. Typical external features for a fetus with MKS at 16 weeks' gestation.

Image courtesy of the Robert J Gorlin Slide Collection.

Visible clinical features include occipital encephalocele, postaxial polydactyly of both hands and feet, massive flank masses due to bilateral renal cystic dysplasia, and typical Potter's facies caused by oligohydramnios (slanting forehead, flattened nose).

Joubert Syndrome (JBTS)

JBTS is a congenital cerebellar ataxia with autosomal recessive or X-linked inheritance, characterised clinically by hypotonia, developmental delay and a distinctive cerebellar and brain stem malformation observed on axial magnetic resonance imaging (MRI) called the molar tooth sign (MTS) (Romani et al., 2013). The MTS consists of hypoplastic cerebellar vermis with hypoplasia of the superior cerebellar peduncle (Nag et al., 2013). The phenotypic variability of this condition has led to sub-classification into classical JBTS and Joubert syndrome-related disorders, which can also feature retinal dystrophy, ocular colobomas, occipital encephalocele, renal disease, polydactyly, oral hamartomas, hepatic fibrosis, polydactyly, oral hamartomas, and endocrine abnormalities (Parisi and Glass, 2017). A recognised JBTS variant phenotype is COACH syndrome (Cerebellar vermis hypoplasia, Oligophrenia (developmental delay/mental retardation), Ataxia, Coloboma, and Hepatic fibrosis). The prevalence is unknown, with reports of 1:80,000 to 1:100,000 likely to be an under-estimate (Parisi and Glass, 2017). The OMIM JBTS phenotypic series PS213300 contains 38 morbid genes and two provisional disease genes (data accessed 20/12/2022), of which 31 are contained in PanelApp's 'green' list on the RMCD super panel version 4.151 used in this study (detailed in Table 3). Around half of JBTS genes localise to the TZ, with the rest localising to various other ciliary sub-compartments (Van De Weghe et al., 2022).

Oral-facial-digital syndrome (OFDS)

OFDS is characterized by abnormalities of the face, oral cavity, and digits, and can also feature abnormalities of the central nervous system (CNS) and kidney (Franco and Thauvin- Robinet, 2016). Several subtypes are delineated, most of which have autosomal recessive inheritance. OFDS Type 1 is most common, with an X-linked dominant, male-lethal pattern of inheritance in familial cases. OFDS Type 1 has an estimated incidence of 1:50,000 live births (Wahrman et al., 1966). The OMIM OFDS phenotypic series PS311200 contains seven morbid genes and three provisional disease genes (data accessed 20/12/2022), of which seven are contained in PanelApp's 'green' list on the RMCD super panel version 4.151 (see Table 3).

Bardet-Biedl syndrome (BBS)

BBS is characterized by rod-cone dystrophy, cognitive impairment, renal abnormalities, truncal obesity, postaxial polydactyly, male hypogonadotropic

hypogonadism, and complex female genitourinary malformations (Forsythe and Beales, 2013). It has autosomal recessive inheritance. Amongst non-consanguineous populations of European descent, BBS has a prevalence of 1 in 100,000 – 160,000. Rates are higher amongst consanguineous populations, such as 1 in 13,500 Bedouin peoples of Kuwait (Frag and Teebi, 1989).

The OMIM BBS phenotypic series PS209900 contains 19 morbid genes and four provisional disease genes (data accessed 20/12/2022), of which 20 are contained in PanelApp's 'green' list on the RMCD super panel version 4.151. *TMEM67* is listed as a modifier gene for BBS. Eight BBS genes encode proteins that assemble into the BBSome protein complex, important for primary ciliary homeostasis. The BBSome works together with the BB and TZ to orchestrate formation and maintenance of the cilium (Waters and Beales, 2011). It functions as a cargo adapter that recognises a diverse set of membrane-bound ciliary proteins, and links them to the IFT machinery (Klink et al., 2020).

Alström syndrome (ALMS)

ALMS is characterised by cone-rod dystrophy, obesity, insulin resistance/type two diabetes mellitus, cardiomyopathy, progressive fatty liver disease, chronic kidney disease and sensorineural hearing impairment (Paisey et al., 2019). The prevalence in the general population of 1-9 per 1,000,000 is probably an underestimate, given the likelihood of missed diagnoses (Orphanet, 2022). It is caused by pathogenic variants in just one ciliopathy gene, *ALMS1*, the encoded protein of which is found in the BB and centrosomes of ciliated cells (Marshall et al., 2011).

1.3.1.2.2 Renal ciliopathies

Renal ciliopathies include the relatively common autosomal dominant polycystic kidney disease (ADPKD) and nephronophthisis. Nephronophthisis, characterised by chronic tubulointerstitial nephritis, can be found in isolation or as part of Senior-Løken syndrome (nephronophthisis, retinal degeneration, hepatic fibrosis, and situs defects) (Tsang et al., 2018). Nephronophthisis is the leading cause of kidney failure in children. The OMIM nephronophthisis phenotypic series PS256100 contains 15 morbid genes and one provisional disease gene (data accessed 20/12/2022), of which 14 are contained in PanelApp's 'green' list on the RMCD super panel version 4.151 (see Table 3).

ADPKD is one of the most common inherited disorders. Epidemiological studies report that it affects up to 1:4000 people in the EU (Willey et al., 2017), but population whole-genome sequencing suggests a higher-than-expected prevalence of ADPKD-associated variants, affecting up to 1 in 1000 (Lanktree et al., 2018). The OMIM polycystic kidney disease phenotypic series PS173900 contains seven entries, amongst which five genes have autosomal dominant inheritance and two have autosomal recessive inheritance (data accessed 20/12/2022). Ciliopathy genes linked to ADPKD include *PKD1* (75–85% of cases) and *PKD2* (15%) (Rossetti et al., 2007). *PKD1* and *PKD2* encode polycystin-1 (PC1) and polycystin-2 (PC2), respectively. PC1 has features of both an ion channel and a G-protein coupled receptor, and PC2 acts as an ion channel in the primary cilium (Barroso-Gil et al., 2021). Along with fibrocystin, the protein product of the autosomal recessive polycystic kidney disease (ARPKD) gene *PKHD1*, *PC1* and *PC2* form a heteromeric complex in the primary cilium (Ta et al., 2020). This was thought to regulate calcium signalling in response to urine flow (DeCaen et al., 2013), but this has been refuted (Delling et al., 2016). Therefore, the pathomechanism linking *PKD1/2* variants and ADPKD remains unclear.

1.3.1.2.3 Ciliopathies with major skeletal involvement

Skeletal ciliopathies comprise at least 16 different subtypes ranging from conditions compatible with life such as Jeune Thoracic Asphyxiating Dystrophy (JATD), through to lethal short-rib thoracic dysplasia (SRTD) types I-V (Mitchison and Valente, 2017). Skeletal phenotypes in ciliopathies are mainly due to IFT defects that affect the Sonic Hedgehog (Shh) and Indian Hedgehog (Ihh) signalling pathways (see thesis section 1.4.1), impairing the growth of bones and cartilage (Bangs et al., 2011).

1.3.1.2.4 Inherited retinal diseases (IRDs)

As well as retinal dystrophy being a feature of several syndromic ciliopathies, around a third of non-syndromic IRDs, including RP and LCA, are associated with a retinal cilium defect, collectively called retinal ciliopathies (Bujakowska et al., 2017). Although most cases of RP are non-syndromic, 20–30% of patients have an associated non-ocular condition (Verbakel et al., 2018). IRDs are a leading cause of blindness and visual loss in the UK working age population, with the annual cost to the UK economy estimated at £523.3 million in 2019 (Galvin et al., 2020).

RP initially causes degeneration of rod photoreceptors, leading to progressive night blindness typically presenting in adolescence, followed by concentric visual field loss. Cone dysfunction usually lags the onset of rod dysfunction; if it manifests clinically it causes loss of central vision later in life. Fundus examination typically reveals peripheral bony spicule pigmentation, attenuation of retinal vessels, and pallor of the optic nerve head. The worldwide prevalence of RP is estimated at 1:4000 (Pagon, 1988), although rates amongst populations with higher levels of consanguinity are greater, such as 1:750 in rural, central India (Nangia et al., 2012).

LCA causes early-onset retinal dystrophy, with severe visual impairment from birth or the first few months of life (Tsang and Sharma, 2018). Affected individuals have wandering nystagmus, poor pupillary light responses, the oculodigital sign (poking, rubbing, and/or pressing of the eyes), and undetectable or severely abnormal full-field electroretinogram (ERG). The prevalence is approximately 1:80,000.

Pathogenic variants in 280 genes are causative for IRDs (Daiger, 2022). Molecular diagnostic rates vary between testing centres and strategies. WES and WGS testing approaches have successfully identified molecular diagnoses for around 60% of IRD patients (Ellingford et al., 2016, Jespersgaard et al., 2019, Zampaglione et al., 2020). Additional CNV analysis has been shown to boost diagnostic yields; pathogenic CNVs were found in 7% of 550 UK IRD patients (Ellingford et al., 2018) and 8.8% of 500 American IRD patients (Zampaglione et al., 2020). However, clearly a significant proportion remain unsolved.

There has been significant progress made in the pre-clinical development and clinical trials of gene-directed targeted therapies, and the first *in vivo* gene therapy drug voretigene neparvovec (trade name “Luxturna”) was approved by the Food and Drug Authority (FDA) for *RPE65*-related retinal dystrophies in 2017 (Leroy et al., 2022).

1.3.2 Genetics of ciliopathies

PanelApp (Martin et al., 2019) (see thesis section 1.1.6) contains a Rare Multisystem Ciliopathy Super Panel, reviewed by 22 genetics and ciliopathy experts. Version 4.151 was used at the time of writing (available from <https://panelapp.genomicsengland.co.uk/panels/728/>). It contains four sub-panels:

renal ciliopathies (v1.64), neurological ciliopathies (v1.31), ophthalmological ciliopathies (v1.30) and skeletal ciliopathies (v1.17). It contains 167 genes known to be associated with human ciliopathy syndromes, of which 99 are diagnostic grade “green” on PanelApp’s traffic light grading system. Most ciliopathies have autosomal recessive inheritance, although there are a few autosomal dominant forms, and *OFD1* has X-linked recessive inheritance. The 99 green PanelApp ciliopathy genes and their OMIM disease associations, including eleven of the major ciliopathy syndromes, are summarised in Table 3 (Amberger et al., 2019).

Table 3. Ciliopathy disease genes

PanelApp diagnostic grade “green” ciliopathy disease genes and known OMIM gene-phenotype relationships(s), including 11 major ciliopathy syndromes. PanelApp Rare multisystem ciliopathy Super panel version 4.151. MKS = Meckel Gruber Syndrome, JBTS = Joubert Syndrome, OFD = Oral-facial- digital syndrome, BBS = Bardet Biedl Syndrome, ALMS = Alström Syndrome, JATD = Jeune Asphyxiating Thoracic Dystrophy, STRD = short rib thoracic dysplasia, RP = retinitis pigmentosa, LCA = Leber Congenital Amaurosis, PKD = polycystic kidney disease, ? = provisional gene-phenotype relationship, Mod = modifier

Gene			OMIM gene-phenotype relationship(s)										Other
			Major ciliopathy syndrome										
			Neurodevelopmental					Skeletal	Retinal	Renal			
Ensembl ID (GrCh38)	RefSeq Transcript	MKS	JBTS	OFD	BBS	ALMS	JATD/SRTD	RP	LCA	Nephronophthisis	PKD		
AHI1	ENSG00000135541	NM_001134831.2	x	✓	x	x	x	x	x	x	x	x	
ALMS1	ENSG00000116127	NM_001378454.1	x	x	x	x	✓	x	x	x	x	x	
ANKS6	ENSG00000165138	NM_173551.5	x	x	x	x	x	x	x	x	✓	x	
ARL13B	ENSG00000169379	NM_001174150.2	x	✓	x	x	x	x	x	x	x	x	
ARL6	ENSG00000113966	NM_001278293.3	x	x	x	✓	x	x	✓	x	x	x	
ARMC9	ENSG00000135931	NM_001352754.2	x	✓	x	x	x	x	x	x	x	x	
B9D2	ENSG00000123810	NM_030578.4	?	✓	x	x	x	x	x	x	x	x	
BBS1	ENSG00000174483	NM_024649.5	x	x	x	✓	x	x	x	x	x	x	
BBS10	ENSG00000179941	NM_024685.4	x	x	x	✓	x	x	x	x	x	x	
BBS12	ENSG00000181004	NM_152618.3	x	x	x	✓	x	x	x	x	x	x	
BBS2	ENSG00000125124	NM_031885.5	x	x	x	✓	x	x	x	x	x	x	
BBS4	ENSG00000140463	NM_033028.5	x	x	x	✓	x	x	x	x	x	x	
BBS5	ENSG00000163093	NM_152384.3	x	x	x	✓	x	x	x	x	x	x	
BBS7	ENSG00000138686	NM_176824.3	x	x	x	✓	x	x	x	x	x	x	
BBS9	ENSG00000122507	NM_198428.3	x	x	x	✓	x	x	x	x	x	x	
C21orf2	ENSG00000160226	NM_004928.3	x	x	x	x	x	x	✓	x	x	x	Axial Spondylometaphyseal Dysplasia
C2CD3	ENSG00000168014	NM_001286577.2	x	x	✓	x	x	x	x	x	x	x	
C5orf42	ENSG00000197603	NM_001384732.1	x	✓	✓	x	x	x	x	x	x	x	
C8orf37	ENSG00000156172	NM_177965.4	x	x	x	✓	x	x	✓	x	x	x	
CC2D2A	ENSG00000048342	NM_001378615.1	✓	✓	x	x	x	x	✓	x	x	x	COACH syndrome
CENPF	ENSG00000117724	NM_016343.4	x	x	x	x	x	x	x	x	x	x	Stromme Syndrome
CEP104	ENSG00000116198	NM_014704.4	x	✓	x	x	x	x	x	x	x	x	
CEP120	ENSG00000168944	NM_001375405.1	x	✓	x	x	x	✓	x	x	x	x	
CEP164	ENSG00000110274	NM_014956.5	x	x	x	x	x	x	x	x	✓	x	
CEP290	ENSG00000198707	NM_025114.4	✓	✓	x	?	x	x	x	✓	✓	x	
CEP41	ENSG00000106477	NM_018718.3	x	✓	x	x	x	x	x	x	x	x	
CEP83	ENSG00000173588	NM_016122.3	x	x	x	x	x	x	x	x	✓	x	
CRB2	ENSG00000148204	NM_173689.7	x	x	x	x	x	x	x	x	x	x	Ventriculomegaly with Cystic Kidney Disease;

													Focal Segmental Glomerulosclerosis
CSPP1	ENSG00000104218	NM_001382391.1	x	✓	x	x	x	x	x	x	x	x	
DDX59	ENSG00000118197	NM_001031725.6	x	x	✓	x	x	x	x	x	x	x	
DHCR7	ENSG00000172893	NM_001360.3	x	x	x	x	x	x	x	x	x	x	Smith-Lemli-Opitz syndrome
DLG5	ENSG00000151208	NM_004747.4	x	x	x	x	x	x	x	x	x	x	No OMIM morbid disease associations; PanelApp reports association with DLG5-associated developmental disorder
DYNC2H1	ENSG00000187240	NM_001377.3	x	x	x	x	x	✓	x	x	x	x	
DYNC2LI1	ENSG00000138036	NM_016008.4	x	x	x	x	x	✓	x	x	x	x	
EVC	ENSG00000072840	NM_153717.3	x	x	x	x	x	x	x	x	x	x	Ellis-van Creveld syndrome
EVC2	ENSG00000173040	NM_147127.5	x	x	x	x	x	x	x	x	x	x	Ellis-van Creveld syndrome
GLI3	ENSG00000106571	NM_000168.6	x	x	x	x	x	x	x	x	x	x	Greig cephalopolysyndactyly syndrome; Pallister-Hall syndrome; Polydactyly, postaxial, types A1 and B; Polydactyly, preaxial, type IV
HNF1B	ENSG00000275410	NM_000458.4	x	x	x	x	x	x	x	x	x	x	Renal cysts and diabetes syndrome; Type 2 diabetes mellitus
HYLS1	ENSG00000198331	NM_001134793.2	x	x	x	x	x	x	x	x	x	x	Hydrolethalus syndrome
ICK	ENSG00000112144	NM_014920.5	x	x	x	x	x	x	x	x	x	x	Endocrine-cerebroosteodysplasia
IFT122	ENSG00000163913	NM_052989.3	x	x	x	x	x	x	x	x	x	x	Cranioectodermal dysplasia
IFT140	ENSG00000187535	NM_014714.4	x	x	x	x	x	✓	✓	x	x	x	
IFT172	ENSG00000138002	NM_015662.3	x	x	x	✓	x	✓	✓	x	x	x	
IFT27	ENSG00000100360	NM_001177701.3	x	x	x	✓	x	x	x	x	x	x	
IFT43	ENSG00000119650	NM_001102564.3	x	x	x	x	x	✓	?	x	x	x	
IFT52	ENSG00000101052	NM_016004.5	x	x	x	x	x	✓	x	x	x	x	
IFT74	ENSG00000096872	NM_025103.4	x	✓	x	✓	x	x	x	x	x	x	Spermatogenic failure
IFT80	ENSG00000068885	NM_020800.3	x	x	x	x	x	✓	x	x	x	x	
IFT81	ENSG00000122970	NM_014055.4	x	x	x	x	x	✓	x	x	x	x	
INPP5E	ENSG00000148384	NM_019892.6	x	✓	x	x	x	x	x	x	x	x	Mental retardation, truncal obesity, retinal dystrophy, and micropenis
INVS	ENSG00000119509	NM_014425.5	x	x	x	x	x	x	x	x	✓	x	
IQCB1	ENSG00000173226	NM_001023570.4	x	x	x	x	x	x	x	x	x	x	Senior-Loken syndrome
IQCE	ENSG00000106012	NM_152558.5	x	x	x	x	x	x	x	x	x	x	Polydactyly, postaxial, type A7
KIAA0586	ENSG00000100578	NM_001329943.3	x	✓	x	x	x	✓	x	x	x	x	

KIAA0753	ENSG00000198920	NM_014804.3	x	?	?	x	x	✓	x	x	x	x	
KIF7	ENSG00000166813	NM_198525.3	x	✓	x	x	x	x	x	x	x	x	Acrocallosal syndrome
LBR	ENSG00000143815	NM_002296.4	x	x	x	x	x	✓	x	x	x	x	
LZTFL1	ENSG00000163818	NM_020347.4	x	x	x	✓	x	x	x	x	x	x	
MAPKBP1	ENSG00000137802	NM_014994.3	x	x	x	x	x	x	x	x	✓	x	
MKKS	ENSG00000125863	NM_170784.3	x	x	x	✓	x	x	x	x	x	x	Mckusick-Kaufman syndrome
MKS1	ENSG00000011143	NM_017777.4	✓	✓	x	✓	x	x	x	x	x	x	
NEK1	ENSG00000137601	NM_001199397.3	x	x	x	x	x	✓	x	x	x	x	
NEK8	ENSG00000160602	NM_178170.3	x	x	x	x	x	x	x	x	?	x	Renal-Hepatic-Pancreatic Dysplasia
NPHP1	ENSG00000144061	NM_001128178.3	x	✓	x	x	x	x	x	x	✓	x	Senior-Loken syndrome
NPHP3	ENSG00000113971	NM_153240.5	✓	x	x	x	x	x	x	x	✓	x	Renal-Hepatic-Pancreatic Dysplasia
NPHP4	ENSG00000131697	NM_015102.5	x	x	x	x	x	x	x	x	✓	x	Senior-Loken syndrome
OFD1	ENSG00000046651	NM_003611.3	x	✓	✓	x	x	x	?	x	x	x	
PIBF1	ENSG00000083535	NM_006346.4	x	✓	x	x	x	x	x	x	x	x	
PIK3C2A	ENSG00000011405	NM_002645.4	x	x	x	x	x	x	x	x	x	x	Oculoskeletodental syndrome
PKD1	ENSG00000008710	NM_001009944.3	x	x	x	x	x	x	x	x	x	✓	
PKD2	ENSG00000118762	NM_000297.4	x	x	x	x	x	x	x	x	x	✓	
PKHD1	ENSG00000170927	NM_138694.4	x	x	x	x	x	x	x	x	x	✓	
PMM2	ENSG00000140650	NM_000303.3	x	x	x	x	x	x	x	x	x	x	Congenital disorder of glycosylation
RPGRIP1L	ENSG00000103494	NM_015272.5	✓	✓	x	x	x	x	x	x	x	x	
SBDS	ENSG00000126524	NM_016038.4	x	x	x	x	x	x	x	x	x	x	Shwachman-Diamond syndrome
SCLT1	ENSG00000151466	NM_144643.4	x	x	x	x	x	x	x	x	x	x	No OMIM morbid disease association; PanelApp reports association with Oro-facio-digital syndrome type IX
SDCCAG8	ENSG00000054282	NM_006642.5	x	x	x	✓	x	x	x	x	x	x	Senior-Loken syndrome
TCTEX1D2	ENSG00000213123	NM_152773.5	x	x	x	x	x	✓	x	x	x	x	
TCTN1	ENSG00000204852	NM_001082538.3	x	✓	x	x	x	x	x	x	x	x	
TCTN2	ENSG00000168778	NM_024809.5	?	✓	x	x	x	x	x	x	x	x	
TCTN3	ENSG00000119977	NM_015631.6	x	✓	✓	x	x	x	x	x	x	x	
TMEM107	ENSG00000179029	NM_183065.4	✓	?	✓	x	x	x	x	x	x	x	
TMEM138	ENSG00000149483	NM_016464.5	x	✓	x	x	x	x	x	x	x	x	
TMEM216	ENSG00000187049	NM_001173990.3	✓	✓	x	x	x	x	x	x	x	x	
TMEM218	ENSG00000150433	NM_001258244.2	x	✓	x	x	x	x	x	x	x	x	
TMEM231	ENSG00000205084	NM_001077418.3	✓	✓	x	x	x	x	x	x	x	x	
TMEM237	ENSG00000155755	NM_001044385.3	x	✓	x	x	x	x	x	x	x	x	
TMEM67	ENSG00000164953	NM_153704.6	✓	✓	x	Mod	x	x	x	x	✓	x	COACH syndrome; RHYNS syndrome
TRAF3IP1	ENSG00000204104	NM_015650.4	x	x	x	x	x	x	x	x	✓	x	Senior-Loken syndrome

TTC21B	ENSG00000123607	NM_024753.5	x	x	x	x	x	✓	x	x	✓	x	
TTC8	ENSG00000165533	NM_144596.4	x	x	x	✓	x	x	✓	x	x	x	
TXNDC15	ENSG00000113621	NM_024715.4	✓	x	x	x	x	x	x	x	x	x	
VPS13B	ENSG00000132549	NM_152564.5	x	x	x	x	x	x	x	x	x	x	Cohen syndrome
WDPCP	ENSG00000143951	NM_015910.7	x	x	x	?	x	x	x	x	x	x	Congenital heart defects, hamartomas of tongue, and polysyndactyly
WDR19	ENSG00000157796	NM_025132.4	x	x	x	x	x	✓	✓	x	✓	x	Cranioectodermal dysplasia; Senior-Loken syndrome
WDR34	ENSG00000119333	NM_052844.4	x	x	x	x	x	✓	x	x	x	x	
WDR35	ENSG00000118965	NM_020779.4	x	x	x	x	x	✓	x	x	x	x	Cranioectodermal dysplasia
WDR60	ENSG00000126870	NM_018051.5	x	x	x	x	x	✓	x	x	x	x	
ZSWIM6	ENSG00000130449	NM_020928.2	x	x	x	x	x	x	x	x	x	x	Acromelic frontonasal dysostosis; Neurodevelopmental disorder with movement abnormalities, abnormal gait, and autistic features

1.3.3 Molecular diagnosis rates

Diagnostic rates for ciliopathies vary between individual syndromes, testing centres and testing strategies. A molecular diagnostic rate of 62% was reported for severe primary neurodevelopmental ciliopathies (see thesis section 1.3.1.2.1) using targeted gene panel sequencing and single nucleotide polymorphism (SNP) array testing (Knopp et al., 2015). 44% of ciliopathy patients enrolled in the Finding Of Rare Disease GENes (FORGE) Canada project, who had already received standard-of-care genetics evaluation and diagnostic testing, received a molecular diagnosis from WES (Sawyer et al., 2016). A research study adopting a genomic approach identified likely causal variants in 85% of 371 families with phenotypes expanding the full ciliopathy spectrum, including in seven novel candidate genes and a novel morbid gene (*TXNDC15*) (Shaheen et al., 2016). The diagnostic rate for motile ciliopathies is up to 68% using targeted gene panels (Paff et al., 2018) and 76% using WES with targeted CNV analysis (Marshall et al., 2015). Clearly, the genetic cause for a significant proportion of ciliopathies remains unknown.

1.4 Primary cilia in cell signalling

Multiple ion channels and receptors are found within the ciliary membrane, which initiate signalling cascades on detection of various mechanical stimuli and chemical messengers. This involves multiple pathways including Hedgehog, Wnt, Platelet-Derived Growth Factor (PDGF), Notch, Hippo, G-Protein Coupled Receptor (GPCR), mTOR, and TGF-beta (Wheway et al., 2018, Anvarian et al., 2019). The role of primary cilia in Wnt and Hedgehog signalling is discussed in more detail below.

1.4.1 Sonic Hedgehog pathway (Shh)

The primary cilium is the key organelle for transduction of the Shh signalling pathway in vertebrates, with functional cilia and IFT essential for normal Shh signalling (Wheway et al., 2018). Hedgehogs are a family of secreted proteins that are essential during vertebrate embryogenesis, homeostasis, and regeneration (Bangs and Anderson, 2017). There are three mammalian Hedgehog proteins: Shh, Indian-Hedgehog (Ihh), and Desert-Hedgehog (Dhh). Ihh has important roles in skeletal development, mainly endochondral ossification. Dhh is restricted to the gonads including granulosa cells of ovaries and Sertoli cells of testis (Carballo et al., 2018).

Shh is required for limb patterning, as well as specification of cell types in the nervous

system. Differentiation of neural progenitors is determined by a gradient of Shh secreted from the notochord to the floor plate of the neural tube (Dessaud et al., 2008). Defective Shh signalling during embryonic development of neuroectodermal lineages is therefore associated with neural tube defects (NTDs). Other clinical features include holoprosencephaly, microcephaly, craniofacial defects, skeletal abnormalities, and polydactyly (Murdoch and Copp, 2010, Briscoe and Therond, 2013). Abnormal development of the cerebellum in severe ciliopathies (vermis hypoplasia, foliation defects) has been ascribed to defective response of granule cell progenitors to Shh produced by the adjacent Purkinje cells. Abnormally active Shh signalling (due to acquired mutations) can lead to multiple cancer types including basal cell carcinomas, medulloblastomas, meningiomas, rhabdomyosarcomas, and odontogenic tumours (Briscoe and Therond, 2013).

Activation of the Shh pathway can happen in two ways: by ligand-dependent interaction or receptor-induced signalling in canonical signalling, or downstream of Smoothed (Smo) in non-canonical signalling (Carballo et al., 2018). Shh signalling at the primary cilium is summarised in Figure 5. The 12 transmembrane domain receptor of Shh ligand, Patched (Ptc1), is located within the ciliary membrane. In the unstimulated state, Ptc1 keeps the canonical Shh pathway off by repressing and excluding the seven transmembrane-domain protein Smo from the cilium (Bangs and Anderson, 2017). This causes the sequestration and suppression of glioblastoma (Gli) transcription factor by Suppressor of Fused (SuFu) at the tip of the primary cilium, blocking the formation and translocation of the activated isoform of Gli to the nucleus and the subsequent transcription of Hh target genes (Haycraft et al., 2005, Zeng et al., 2010).

Binding of Shh to Ptc1 inhibits its activity, relieving the repression of Smo which translocates out through the ciliary membrane into a vesicular compartment prior to regulated degradation (Rohatgi et al., 2007). This then allows Smo to repress SuFu, relieving repression of Gli at the tip of the cilium. This is then free to be post-translationally modified to the Gli activator form (GliA), which is transported out of the cilium to the nucleus, activating expression of downstream Hh target genes. Movement of Hh signalling intermediates in and out of the cilium is facilitated by IFT proteins and IFT motor proteins.

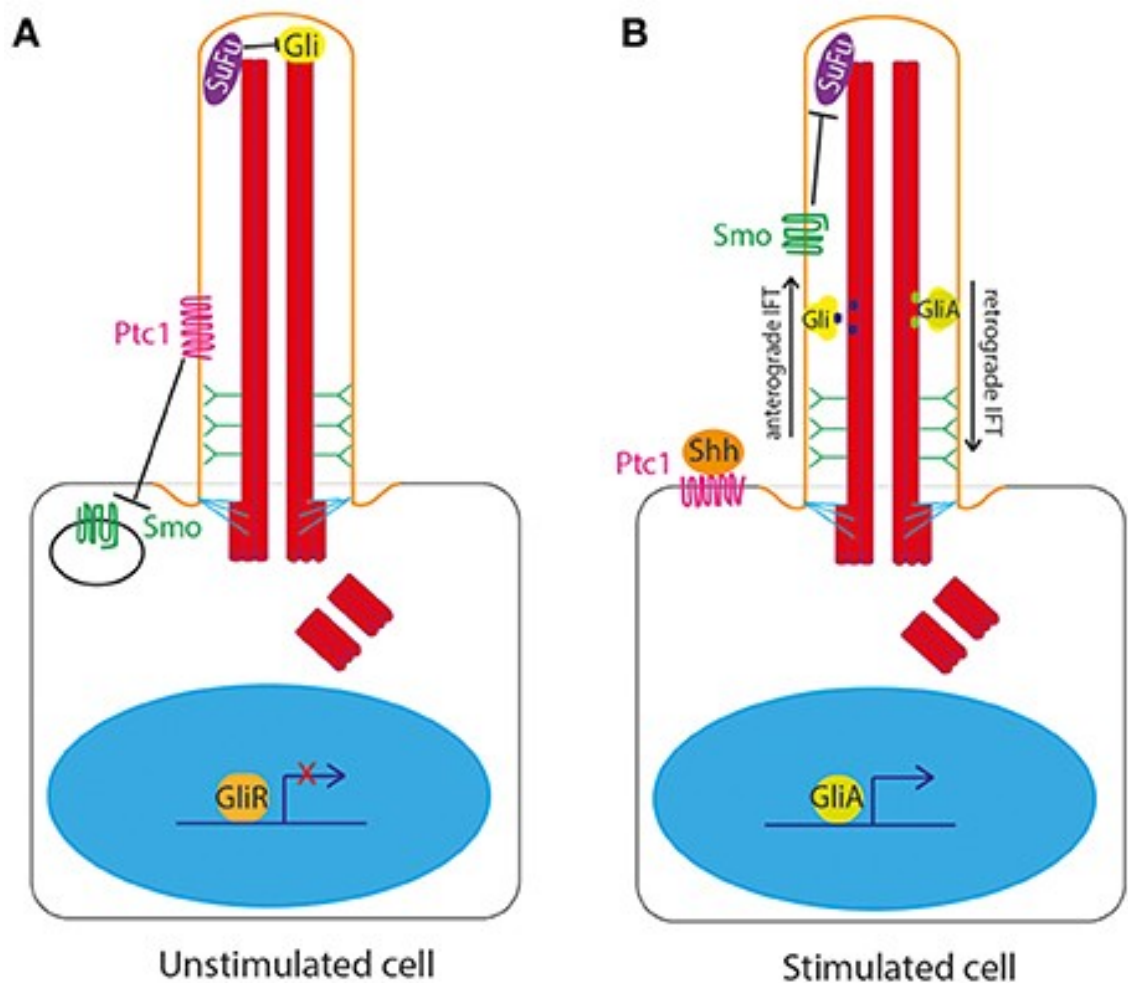


Figure 5. Shh signalling at the primary cilium

Copied with permission from (Wheway et al., 2018) (CC BY 4.0). Copyright © 2018 Wheway, Nazlamova and Hancock.

(A) In the unstimulated state, Ptc1 sits in the cilium membrane, repressing and excluding Smo from the cilium. At the tip of the cilium, SuFu sequesters and suppresses Gli transcription factors. (B) In the stimulated state, the repression of Smo by Ptc1 is relieved upon binding of Shh to Ptc1, allowing Smo to enter and Ptc1 to leave the cilium. This allows Smo to repress SuFu, relieving repression of Gli at the tip of the cilium. Gli is then free to be post-translationally modified to the Gli activator form (GliA), which is transported out of the cilium to the nucleus to activate expression of downstream target genes

Ciliary localisation of Shh signalling molecules, such as Smo, is adversely affected by pathogenic variants in genes encoding several TZ proteins (Yang et al., 2015).

Non-canonical Shh activation occurs through Gli-independent mechanisms. It remains relatively poorly characterised but is thought to be mainly dependent on Smo. Research is ongoing to better understand this pathway, and how Smo selects between canonical and non-canonical routes.

1.4.2 Wnt signalling

Wnt signalling is involved in cell migration, planar cell polarity (PCP), neural patterning, skeletal system development, and organogenesis (Pala et al., 2017). There are two signalling pathways in mammals driven by Wnt proteins: the canonical (β -catenin-dependent) and non-canonical (β -catenin independent) Wnt pathways (Wheway et al., 2018). Both are initiated by the binding of a Wnt ligand to a Frizzled (Fzd) receptor.

1.4.2.1 Non-canonical Wnt signalling

The non-canonical Wnt signalling pathway in primary cilia is summarised in Figure 6. In the stimulated state, non-canonical Wnt ligand binds to the Fzd 3 receptor (Fzd3), which triggers asymmetric localisation of Vangl2 in the cell. Recruitment of Disheveled (Dvl) to the plasma membrane activates RhoA and the JNK pathway, triggering Ca^{2+} release and stimulating remodelling of the actin cytoskeleton. Dvl regulates the migration of the BB, along with TZ proteins TMEM67 and TMEM216, Inversin and BB protein MKS1 (Wheway et al., 2018). Inversin inhibits the canonical Wnt pathway by targeting cytoplasmic Dvl for degradation. Inversin is particularly important in regulating the balance between canonical and non-canonical Wnt signalling.

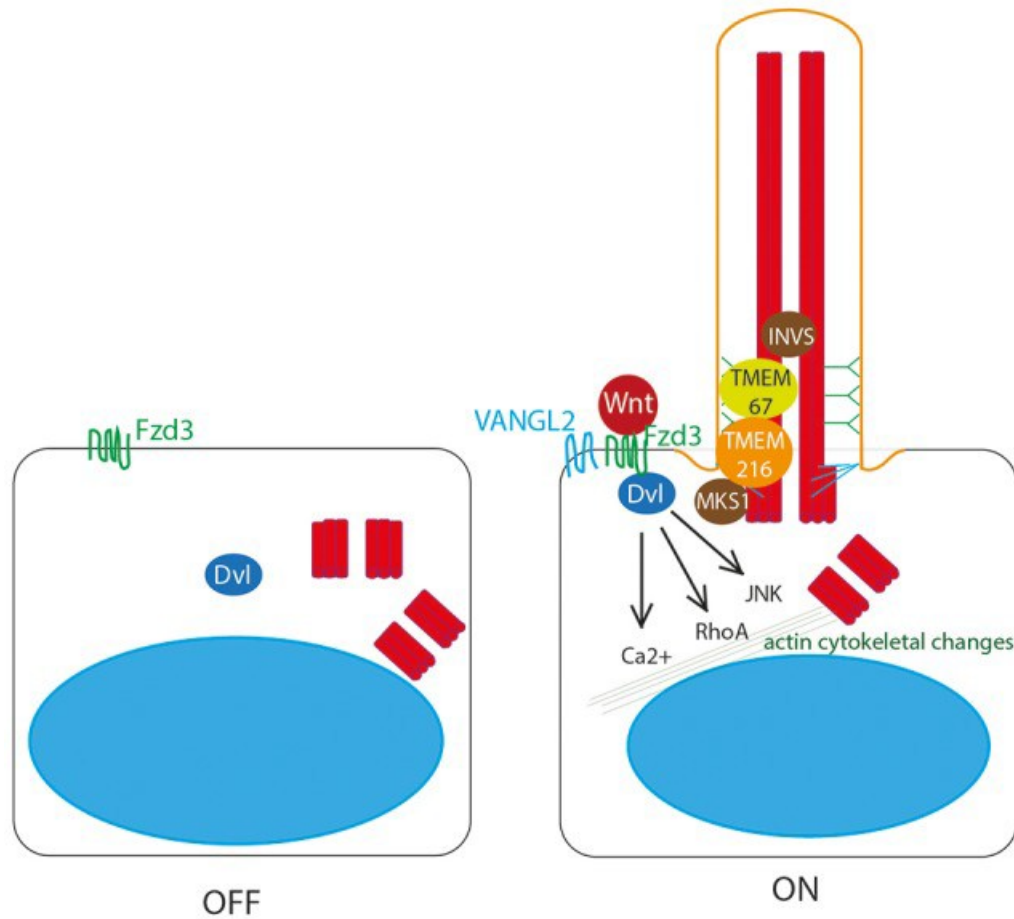


Figure 6. The non-canonical Wnt signalling pathway in primary cilia.

Copied with permission from (Wheway et al., 2018) (CC BY 4.0). Copyright © 2018 Wheway, Nazlamova and Hancock.

Non-canonical Wnt ligands bind to the Fzd3 receptor, which triggers asymmetric localisation of Vangl2 in the cell. Remodelling of the actin cytoskeleton is stimulated by Ca²⁺ release, triggered by Dvl activation of RhoA and the JNK pathway. This is dependent upon correct definition of cell polarity by BB migration to the apical cell surface. This migration is regulated by Dvl, TMEM67, TMEM216, MKS1 and Inversin.

Non-canonical Wnt signalling is involved in controlling tissue functions and maintaining tissue architectures by modulating cell migration and orientation.

Normal ciliogenesis is required for the non-canonical Wnt signalling pathway, which depends on PCP being correctly established (Gomez-Orte et al., 2013). PCP, initially identified through genetic studies of *Drosophila*, is an important organizer of tissues during morphogenesis, whereby distinct polarity is established within the plane of a cell sheet (Butler and Wallingford, 2017, Yang and Mlodzik, 2015). Wnt-PCP results in cytoskeletal actin rearrangements, mediated by Rho proteins, important in regulating cell morphology, migration, and correctly oriented cell division.

Inherited defects in proteins regulating ciliogenesis and BB migration therefore result in complex PCP defects, including abnormalities in dorsal axis organisation. These can manifest clinically as NTDs as well as inner ear defects due to failure of correct orientation of stereocilia in the cochlear hair cells. Consequently, inherited pathogenic variants of ciliary proteins can lead to a combination of congenital deafness and RP in the condition Usher syndrome (Ush) (Sorusch et al., 2014).

1.4.2.2 Canonical Wnt signalling

The canonical Wnt signalling pathway is summarised in Figure 7. In the absence of Wnt ligand in the unstimulated state, a “destruction complex” is formed including Axin, Adenomatous Polyposis Coli (APC), casein kinase 1 (CK1) and GSK3 β . The β -catenin destruction complex operates within the β -TrCP/SCF-dependent ubiquitin-proteasome pathway. In the absence of Wnt, the phosphorylation of β -catenin by CK1 and GSK-3 at the BB acts as a trigger for degradation by the proteasome, preventing it from entering the nucleus.

In the presence of a canonical Wnt signal in the stimulated state, cytosolic levels of β -catenin rise due to the inhibition of the β -catenin destruction complex. Wnt ligand binds to a membrane bound Fzd receptor, which then binds LRP5/6, allowing it to recruit and sequester Axin. The Wnt signal is transduced via Dvl, which is recruited to the membrane, inhibits GSK-3 β and binds Axin upon stimulation. Without Axin, the destruction complex is unable to degrade β -catenin, leaving it free to translocate into

the nucleus. There, aided by Joubertin (Jbn), encoded by *AHI1*, β -catenin functions as a transcriptional coactivator. β -catenin associates with the nuclear transcription factors T-cell factor (TCF) and lymphocyte enhancer factor (LEF) and induces transcription of Wnt target genes that are under the control of TCF/LEF promoters, such as cMYC, AXIN2 and L1CAM (Pala et al., 2017).

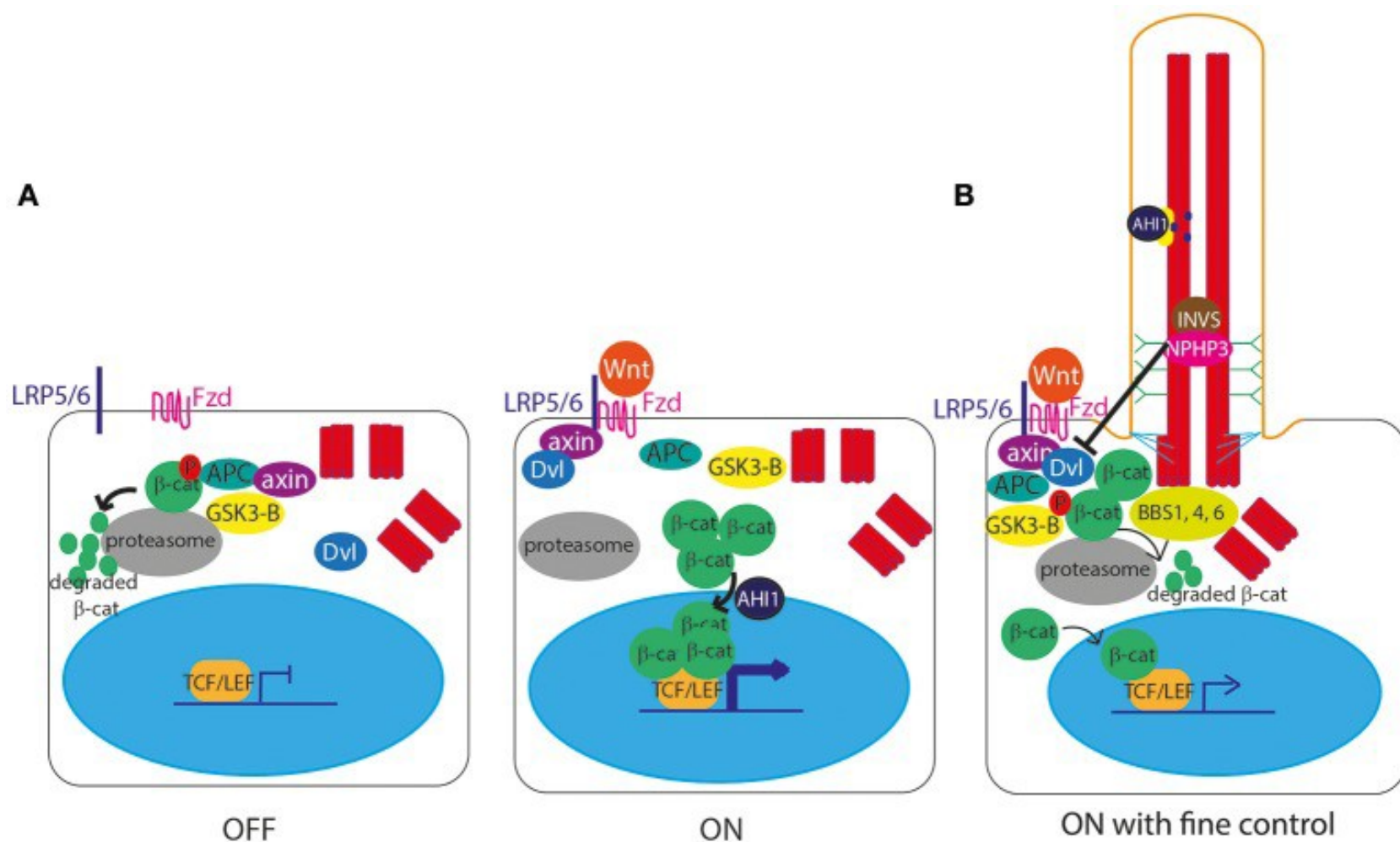


Figure 7 Canonical Wnt signalling at the primary cilium.

Copied with permission from (Wheway et al., 2018) (CC BY 4.0). Copyright © 2018 Wheway, Nazlamova and Hancock.

A: In the unstimulated state, the Axin/APC/GSK3-β “destruction complex” targets β-catenin to the proteasome for degradation. In the stimulated state, Wnt ligands bind to a Fzd receptor which then binds LRP5/6, allowing it to recruit Axin. With Axin sequestered by LRP5/6, the destruction complex can no longer degrade β-catenin, leaving it free to enter the nucleus, aided by Jbn, to activate transcription of Wnt

target genes under TCF/LEF promoters. The Wnt signal is transduced via Disheveled (Dvl), which is recruited to the membrane and binds Axin upon stimulation.

B: The primary cilium controls the level of expression of Wnt target genes, via controlled degradation of Dvl by cilia proteins INVS and NPHP3, and by sequestering Jbn at the cilium so it cannot aid translocation of β -catenin into the nucleus.

The role of the primary cilium in canonical Wnt signalling remains somewhat controversial. Primary cilia are proposed to mediate a negative modulatory effect on the canonical Wnt/ β -catenin pathway (Lancaster et al., 2011). Supporting evidence suggests that the ciliary proteins NPHP3 and Inversin control degradation of Dvl, which transduces the Wnt signal. Jbn, which normally shuttles β -catenin between the cytosol and nucleus, can be sequestered at the cilium so it cannot aid translocation of β -catenin into the nucleus, therefore influencing the level of expression of Wnt target genes. Furthermore, recent data shows that the ciliary protein MKS1 acts as a novel substrate-adaptor that interacts with β -catenin and ubiquitin-proteasome system (UPS) components, thereby regulating levels of β -catenin through normal degradation during Wnt signalling (Szymanska et al., 2022).

1.5 Modelling genetic variants

1.5.1 Cilia research models

Great insights into ciliary biology have been gained through work in model organisms (Vincensini et al., 2011). Many model systems have been used in ciliary research, from simple nematodes through to genome-edited human organoids.

1.5.1.1 *Caenorhabditis elegans* (*C. elegans*)

The *C. elegans* nematode worm has been a vital model organism for biomedical research for over 50 years (Brenner, 1974). Their small size (approximately 1mm), rapid life cycle (from egg to adult in 3.5 days), ease of culture, large brood size (~300), low maintenance costs and amenity to long-term cryopreservation make them an attractive small animal model system (Ganner and Neumann-Haefelin, 2017). Self-fertilization means that after hermaphrodites are mutagenised, mutant alleles (except dominant lethals) can be maintained through self-propagation in subsequent generations without mating (Greenwald, 2016). The adult hermaphrodite has a transparent body that can be visualized by live-imaging, allowing *in vivo* studies of cell morphology, protein sub-localisation and microarchitecture. Many functional experimental methods have proven insightful, including behavioural, fluorescence and transport assays.

The *C. elegans* and human genomes have almost the same number of genes (~20,000) and share a surprisingly high proportion of cellular and molecular processes.

60–80% of human genes have a *C. elegans* homolog (Kaletta and Hengartner, 2006). This high conservation of human disease genes and evolutionary pathways between *C. elegans* and mammals, accompanied by relatively easy access for genetic manipulations, has made them invaluable to biomedical research.

Studies of gene expression and protein localisation are straightforward in *C. elegans* through DNA transformation and microinjection techniques. The Nobel Prize-winning discovery of gene silencing by RNA interference was first described in *C. elegans* (Fire et al., 1998). Clustered regularly interspaced short palindromic repeats (CRISPR) – CRISPR associated protein (Cas) gene editing has also been applied successfully in *C. elegans* in multiple projects, facilitating genetic variant interpretation and offering potential for targeted therapeutics (see thesis section 1.5.2) (Kim and Colaiacovo, 2016).

The only ciliated cell type in *C. elegans* is the sensory neurons, which detect and transduce extracellular and internal signals, and mediate a range of behaviours (Inglis et al., 2007). These include chemo-sensation, mechano-sensation, male copulation, thermo-sensation, and adaptation (Bae and Barr, 2008).

One of the easiest ways to assay the structural integrity of sensory cilia in *C. elegans* is to test their ability to take up a fluorescent dye. This is done by placing living worms in a solution containing dyes such as fluorescein isothiocyanate (FITC), DiI, DiO and DiD, and observing the filling of amphid sensory neurons in the head and phasmid sensory neurons in the tail through their exposed, ciliated endings (Tong and Burglin, 2010). Several other tests of behavioural phenotypes are available to identify worms with defective cilia, including the osmotic avoidance abnormal (Osm) phenotype, chemotaxis (Che) phenotype and mechanosensory (Mec) phenotypes (Inglis et al., 2007).

1.5.1.2 Cell lines

Immortalised cell lines include cells that have been artificially ‘immortalised’ through the forced expression of the human telomerase reverse transcriptase (hTERT) gene, and tumorous cells that do not stop dividing because they lack cell cycle checkpoint controls. Many cell lines are available commercially, derived from animals and

humans. Immortalised cell lines are mostly well characterised, genetically identical populations, facilitating consistent and reproducible results. They are easier to culture than primary cultures, growing quickly and continuously. This makes it possible to extract large amounts of proteins for biochemical assays.

The major disadvantage to using cell-lines is that their abnormal culture conditions means that they are not truly reflective of what their cell type would do, or look like, in a normal living system. They lack normal cell-cell contacts and positional signals that tell them what they should do, be and make, and are exposed to abnormal levels of oxygen and carbon dioxide. They divide indefinitely, and sometimes express unique gene patterns not found in any cell type *in vivo*. Some are extremely genetically and phenotypically different to their living cell counterparts.

Some cell lines have proven particularly suitable for cilia research. The diploid immortalised retinal pigment epithelial (RPE) cell line (hTERT RPE-1) displays horizontal cilia that are well- suited for high-content imaging and has largely typical RPE functions and morphology (Wheway et al., 2015, Kuznetsova et al., 2014). Similarly, the spontaneously arising human RPE cell line, ARPE-1, has been used extensively (Dunn et al., 1998). The spontaneously arising murine inner medullary collecting duct (mIMCD3) cell line is easy to culture, forms polarised monolayers and display long cilia suitable for immunofluorescence microscopy and protein localisation studies (Rauchman et al., 1993).

1.5.2 CRISPR-Cas genome editing

The advent of new genome editing technologies, particularly the CRISPR-Cas system, has provided an exciting opportunity to model ciliopathy gene variants and gain functional insight into their effects.

CRISPR is derived from a prokaryotic adaptive immune system that provides defence against foreign genetic elements such as plasmids and 'phages (Barrangou and Doudna, 2016, Marraffini and Sontheimer, 2008). It is significantly easier and cheaper to implement and more efficient at editing than the older gene editing technologies such as meganucleases, zinc- finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) (Cui et al., 2018). These attributes, along with its

specificity and versatility, have made CRISPR the leading genome editing tool.

The basic CRISPR system consists of two parts: a guide RNA (gRNA) which targets a specific sequence, and a Cas DNA endonuclease which cleaves the targeted sequence. The CRISPR guide RNA (crRNA) base-pairs with a transactivating CRISPR RNA (tracrRNA), which directs the Cas endonuclease to a specific location of the genome, complementary to the first 20 nucleotides of the crRNA. This region must be adjacent to a protospacer associated motif (PAM) for the Cas to recognise the position and bind. 'NGG' is the PAM for the most commonly used *S. pyogenes*-derived Cas9. The identification of several other Cas proteins with different PAM recognition sites has facilitated targeting of virtually every site in the genome. Once bound, the Cas induces a double stranded break (DSB) 3bp upstream of the PAM. The cell predominantly employs error prone non-homologous end joining (NHEJ), that competes with the less efficient homology directed repair (HDR) pathway to repair the break (Figure 8) (Maruyama et al., 2015).

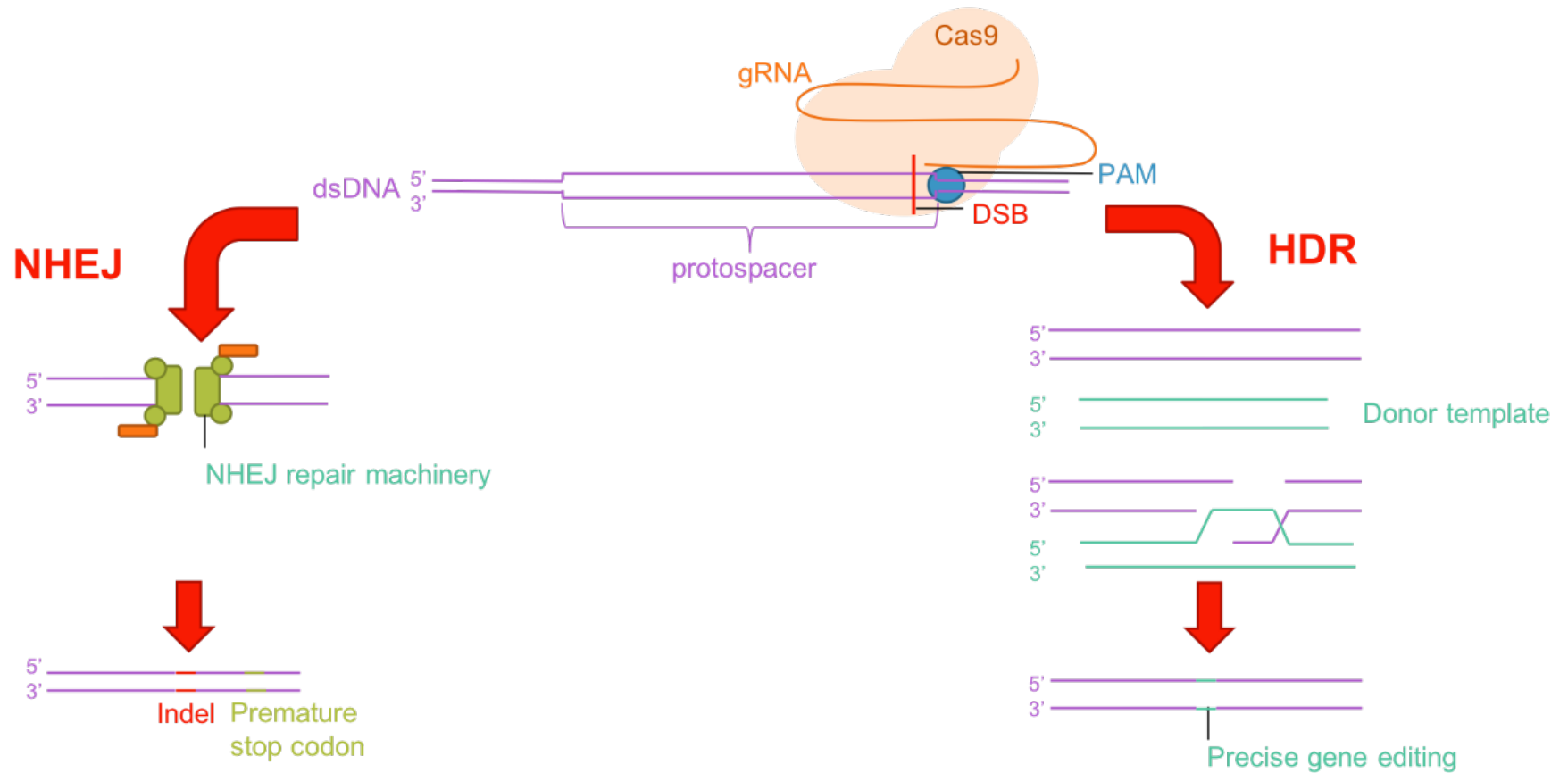


Figure 8. Repair pathways for Cas9-induced DSBs

DSBs are caused by Cas9 when guided to target DNA by gRNA. They can be repaired by two pathways. The NHEJ pathway is error prone. The re-joining of the broken ends of the DSB by the NHEJ machinery can result in indels at the junction site. These can produce frameshifts, leading to premature stop codons and therefore gene knockouts. When a donor repair template is provided, and repair occurs down the HDR pathway, precise insertions or modifications can be engineered. However, repair occurs less efficiently down this pathway

1.5.2.1 Delivery strategies

Both viral and non-viral methods can be used for the delivery of CRISPR/Cas9 components into cell lines and animal models. Non-viral hosts include plasmid DNA, Cas9/gRNA ribonucleoprotein complexes and donor nucleic acid templates, which can be delivered via lipid-mediated transfection, electroporation, induced osmocytosis and hydrodynamic delivery (Zuris et al., 2015, D'Astolfo et al., 2015).

1.5.2.2 Off target effects (OTEs)

A major concern about genome editing is the potential OTEs of editing enzymes, which could lead to unexpected mutations and genomic instabilities (Duan et al., 2014). Genome-wide chromatin immunoprecipitation sequencing (ChIP-seq) experiments have shown that that, depending on the gRNA used, sgRNA:dCas9 complexes have tens to thousands of off-target binding sites (Kuscu et al., 2014). A few mismatches between the 5'-end 20-nt sequence in the gRNA and the target DNA sequence have been shown to be tolerated (Lin et al., 2014). Evidence of cleavage by wild-type Cas9 at some of these off-target sites highlights the importance of taking careful steps to reduce OTEs (Duan et al., 2014, Kuscu et al., 2014, Fu et al., 2013). Strategies are being developed to overcome this, such as generation of dead Cas9 (dCas9) through D10A and H840A mutations at RuvC and HNH endonuclease domains of wild type *Streptococcus pyogenes* Cas9 (SpCas9) (Kanafi and Tavallaei, 2022). This does not cut target DNA, but still can still bind to target DNA based on the gRNA targeting sequence.

1.5.2.3 gRNA design

Choosing an appropriately specific and efficient gRNA for the target DNA sequence is an essential step in avoiding OTEs. Although, in theory, a gRNA-Cas9 complex should bind and cleave any target DNA sequence if the 5'-end 20-nt sequence in the gRNA is complementary to the target DNA sequence, cutting efficiency has been shown to vary significantly between different gRNAs (Cui et al., 2018).

Several computational tools have been developed to design gRNAs with high efficacy and specificity, reviewed in (Cui et al., 2018). Factors considered in gRNA design include the location of the cleave site within the gene, non-canonical PAM sequences, guanine content, and numbers and positions of mismatches between the gRNA and

protospacer sequence non- canonical PAM sequences (Han et al., 2020).

Designing gRNAs for CRISPR genome editing is relatively easy compared with other editing tools, due to computational determination of OTEs based on genomic sequences with high similarity to the target locus. Many online gRNA design tools provide on-target and off-target predictions based on custom algorithms that may be species and/or nuclease specific (Han et al., 2020). Off target binding sites are enriched in open genomic regions, suggesting that chromatin structure is a major determinant of Cas9 binding (Gilbert et al., 2014). By incorporating chromatin context into computational off target prediction tools, better guide design can be possible, but, in general, most workflows recommend that gRNA efficiencies are determined empirically.

1.5.2.4 Disease modelling through CRISPR mediated knockout strategies

The CRISPR toolkit allows us to precisely replace, rearrange, silence, activate and remodel genomic elements efficiently, cheaply, and relatively easily. In 'knockout' experiments, imperfect repair of DSBs by endogenous NHEJ machinery is exploited to generate disruptive random insertions and deletions (indels), which can lead to LOF variants through a shift in the reading frame or a premature stop codon. The resulting edited cell line, organoid or organism must then be characterised as a knockout through functional experiments as well as sequencing. This approach has been highly successful in generating knockout alleles in protein-coding genes and disrupting transcription factor binding sites (Hanna and Doench, 2020). Pairs of programmed DSBs have also been used to generate custom larger deletions or chromosomal rearrangements (Choi and Meyerson, 2014). Knockout CRISPR experiments offer greatest flexibility in guide selection because most of the exonic region of a gene is usually a viable target (Doench et al., 2014).

CRISPR/Cas technology has sparked a great deal of excitement within the scientific community because it has significantly reduced the time required to generate genetically modified animal and cellular models (Jacinto et al., 2020). For example, a single CRISPR editing step by zygote injection can now generate mice carrying mutations in multiple genes, without the need for ES cell derivation or complex genetic crosses. CRISPR has been used for the generation of *C. elegans*, *Drosophila*, zebrafish, mice, pigs, and non-human primate model organisms, as well as human cell lines and organoids (Dow, 2015). CRISPR-mediated gene knockouts can provide

insights into disease mechanisms. Mutant disease models can also provide a platform for identifying therapeutics that demonstrate phenotypic rescue.

1.5.2.5 CRISPR “editing” experimental strategies

In ‘editing’ experiments, specifically designed base changes are generated in target DNA (Hanna and Doench, 2020). These variants may be introduced through provision of an exogenous template DNA co-delivered with the nuclease, which can be inserted via the HDR repair pathway (Liang et al., 2017, Yang et al., 2013). This is often in the form of single- stranded oligodeoxynucleotides (ssODNs) for point mutation corrections.

Unfortunately, the efficiency of repairing DSBs by HDR is relatively low (Ran et al., 2013). As an alternative to relying on this HDR pathway, several modified Cas proteins have been engineered that act directly on endogenous DNA to make prescribed DNA modifications. These include C to T (Komor et al., 2016) and A to G (Gaudelli et al., 2017) base editors, both optimised for mutation in mammalian systems (Koblan et al., 2018). Over 80% of pathogenic ClinVar SNPs that arise from transition mutations are editable by at least one base editor (Hanna and Doench, 2020). However, the narrow window for the edit site for base editors limits the number of available guides per target site.

There is also the newer prime editor which can theoretically induce targeted insertions, deletions and all 12 types of point mutation at virtually every site in the genome (Anzalone et al., 2019). In prime editing, a modified Cas9 “nicks” a single strand of the double helix, instead of cutting both strands. A modified guide, called a pegRNA, contains an RNA template for a new DNA sequence, to be added to the genome at the target location. Attached to the Cas9 is a reverse transcriptase enzyme, which can make a new DNA strand from the RNA template and insert it at the nicked site. In principle, prime editing could correct up to 89% of known genetic variants associated with human diseases (Anzalone et al., 2019).

1.6 TMEM67

1.6.1 Encoded protein

TMEM67, found at chromosome 8q22.1, encodes the 995-amino acid transmembrane

protein 67, also known as meckelin. The TMEM67 protein contains an extracellular N-terminal domain with a highly conserved cysteine-rich domain (CRD), a predicted β -pleated sheet region, seven predicted transmembrane regions and an intracellular C-terminus including a coiled-coil domain (see Figure 9) (Abdelhamed et al., 2015).

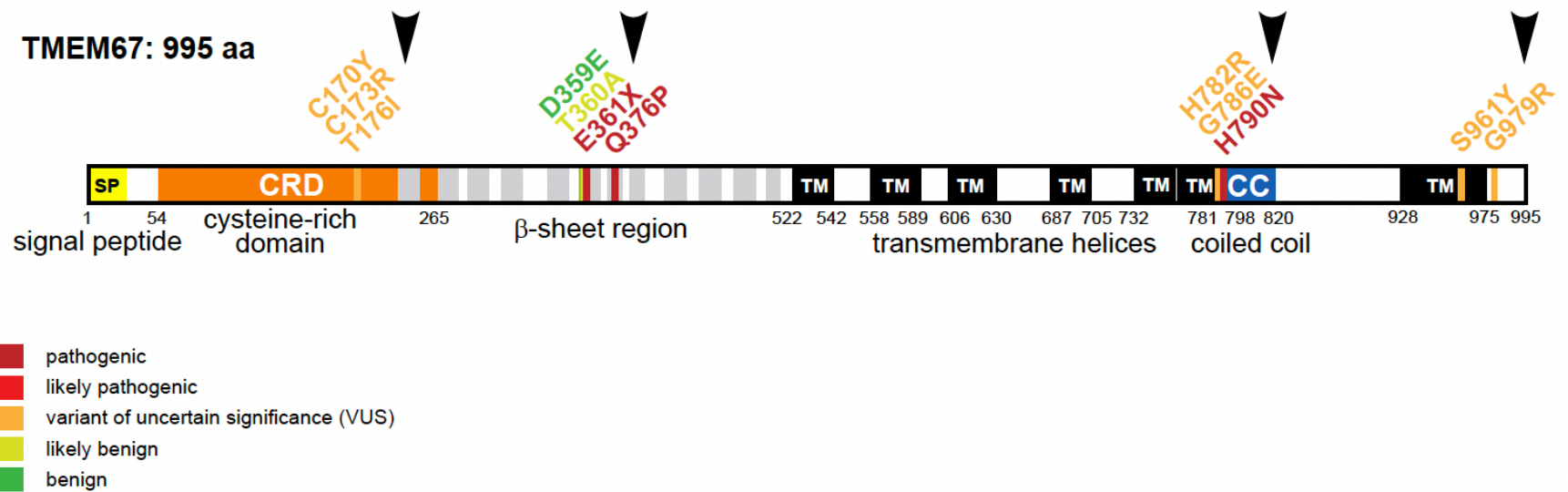


Figure 9. Schematic of TMEM67 protein with target variants for modelling

Pathogenic variants in TMEM67 are associated with MKS, JBTS, nephronophthisis and COACH syndrome. Marked variants are coloured according to their ClinVar status: pathogenic variants are red; VUS are orange and likely benign/benign are green. Variants are clustered in the cysteine-rich domain of TMEM67, the β -sheet region, and in the terminal transmembrane helices

The extracellular CRD has structural similarity to the CRD of Frizzled, which is implicated in canonical and non-canonical Wnt signalling (Smith et al., 2006). The coiled-coil domain is thought to interact with other proteins such as Nesprin-2, an important scaffold protein for maintenance of the actin cytoskeleton, nuclear positioning, and nuclear-envelope architecture (Dawe et al., 2009).

As well as being present in the TZ (see thesis section 1.2.3.1), TMEM67 can also be found within the ciliary membrane (Dawe et al., 2007). In the non-canonical Wnt signalling pathway, TMEM67 is required for centriolar migration to the apical membrane, regulation of actin cytoskeleton remodelling and RhoA activity (Dawe et al., 2009, Dawe et al., 2007). TMEM67 functionally interacts with the ligand Wnt5a, which activates the non-canonical pathway, and inhibits the canonical Wnt pathway (Abdelhamed et al., 2015).

1.6.2 TMEM67 knockout models

TMEM67 knockout rodents have combinations of brain abnormalities including cerebellar hypoplasia and hydrocephalus, limb defects, cardiac abnormalities, polycystic kidney disease and pulmonary hypoplasia (Cook et al., 2009, Gattone et al., 2004). The *TMEM67*^{tm1(Dgen)/H} knock-out mouse model has a variable MKS-like phenotype early in embryonic development (NTDs, occipital meningocele, midbrain-hind brain exencephaly and frontal encephalocele), and develops a JBTS-phenotype at later gestations (cerebellar vermis hypoplasia or aplasia, deep interpeduncular fossa and posterior fossa defects) (Abdelhamed et al., 2015). The severe cerebellar hypoplasia phenotype seen in this model is due to complex Wnt signalling, ciliogenesis and rostral hindbrain patterning defects which impact on downstream Shh signalling events (Abdelhamed et al., 2019). TMEM67 is essential for optimal levels of canonical Wnt/ β -catenin signalling and the formation of primary cilia required for responsiveness to Shh signalling. Tmem67 has been shown to regulate canonical Wnt/ β -catenin signalling in the developing cerebellum via Hoxb5, providing new mechanistic insights into ciliopathy cerebellar hypoplasia phenotypes (Abdelhamed et al., 2019).

1.6.3 Disease associations

Pathogenic variants in *TMEM67* are the most frequent cause of MKS, accounting for 16% of cases (Hartill et al., 2017, Iannicelli et al., 2010). Several founder mutations are known, including two splice variants identified in families of Pakistani origin (c.1546 + 1

G>A and c.870-2A>G) (Smith et al., 2006, Szymanska et al., 2012). Patients with *TMEM67*-mutated MKS have less frequent polydactyly and CNS malformations than those with pathogenic variants in another major disease gene, *MKS1*, demonstrating genotype-phenotype correlation within the condition (Consugar et al., 2007).

TMEM67 pathogenic variants are also a leading cause of JBTS, accounting for 6-20% of total JBTS cases within different populations (Parisi and Glass, 2017). In particular, *TMEM67* variants are associated with the JBTS variant phenotype COACH syndrome, responsible for 57–83% of total COACH cases (Doherty et al., 2010, Iannicelli et al., 2010, Brancati et al., 2009). Pathogenic *TMEM67* variants are also reported to cause nephronophthisis with hepatic fibrosis (Otto et al., 2009), BBS (Leitch et al., 2008) and RHYNS syndrome (Brancati et al., 2018).

1.6.4 Variant pathogenicity

The public variant pathogenicity database ClinVar lists 693 *TMEM67* variants, of which 256 (37%) are classified as VUS and 37 (5%) have conflicting pathogenicity interpretations (accessed 26/10/2022) (Landrum et al., 2016). There are 121 variants with at least one pathogenic ClinVar entry, of which 84 (69%) are short variants (<50bp) and 19 are structural variants (>50bp). Amongst the pathogenic ClinVar short variants, the most common type is missense (n=31; 37%) followed by nonsense (n=22; 26%), frameshift (n=14; 17%), splice site (n=13; 15%), non-coding RNA (n=3; 4%) and untranslated region variants (UTR) (n=1; 1%).

1.7 Hypothesis

Molecular genetic diagnosis rates can be improved for patients with ciliopathies by detecting previously missed pathogenic variants, and by reducing the proportion of variants categorised as VUSs. Previously undetected pathogenic variants can be found through WGS data analysis. Definitive variant pathogenicity interpretation for ciliopathy gene VUSs can be achieved through functional assays in ciliated cell lines that reveal alterations in ciliary phenotype.

1.8 Overall objective

To enhance the existing pathways of genetic variant interpretation and functional validation, aiming to deliver a rapid and translatable system for improved molecular

diagnostics that has potential for use in mainstream diagnostic centres. Ciliopathy patients are selected as an exemplar, but the objective is to develop systems applicable to other disease groups.

1.9 Specific chapter aims

1.9.1 Chapter 2

To undertake detailed genomic analysis for participants recruited to 100K with suspected primary ciliopathies, with the aim of improving molecular genetic diagnosis rates. For unsolved participants, the gene search included:

- a) Analysis of known ciliopathy disease genes
- b) Analysis of candidate ciliopathy genes
- c) Analysis of genes outside of ciliopathy gene panels according to entered phenotypic features

The variant search included:

- a) Analysis of coding variants
- b) Analysis of non-coding variants potentially impacting splicing
- c) Analysis of potentially pathogenic structural variants

1.9.2 Chapter 3

To undertake a search for un-diagnosed ciliopathy patients entered to non-ciliopathy recruitment categories in 100K through a reverse phenotyping strategy. This included:

- a) Selection of key ciliopathy genes, representative of the full multi-systemic ciliopathy disease spectrum.
- b) Search of the whole 100K rare disease dataset for potential molecular diagnoses in these key ciliopathy genes.
- c) Link back to the entered clinical features and any additional available clinical data for participants with potentially pathogenic variants to undertake

genotype-phenotype correlation analyses.

- d) Definition of key ciliopathy clinical features required for potential new diagnoses to justify reporting to recruiting clinicians.

1.9.3 Chapter 4

To develop functional missense variant interpretation strategies in the human ciliated cell line hTERT RPE-1 and in the *C. elegans* worm model for an exemplar ciliopathy disease gene (*TMEM67*), to reduce the proportion classified as VUSs and therefore prohibiting definitive molecular diagnosis. The human RPE-1 work included:

- a) Development and characterisation of a knockout *TMEM67* RPE-1 cell line using CRISPR-Cas9 gene editing.
- b) Engineering of *TMEM67* plasmids containing variants with a range of predicted effects. This includes VUS from fetuses with MKS as well as known benign and pathogenic variants.
- c) Development of a functional system following transfection of variant plasmids into the knockout cell line that allows determination of variant pathogenicity and subsequent interpretation of VUS.

2 Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project

2.1 Research Rationale

This project was designed to be a comprehensive genotype-phenotype analysis of participants recruited to 100K with a prior clinical suspicion of a primary ciliopathy. At the time of starting this project, no comprehensive cohort analysis of any disease group entered to 100K had been published, and overall diagnosis rates for the main project remain unreported. Participants entered to three representative recruitment categories were selected for analysis: Bardet-Biedl syndrome (BBS) (n=45), Joubert syndrome (JBTS) (n=14) and 'Rare Multisystem Ciliopathy Disorders' (RMCD) (n=24), given our research group's expertise in primary ciliopathies. The rationale was to analyse the molecular diagnosis rates achieved by GEL and see whether we could add any further diagnostic uplift through additional genomic analysis and domain specific knowledge.

We were interested in how much value could be added by opening up a diagnostic search beyond GEL's prioritised variants for mainstream diagnostic analysis (Tier 1 and 2 variants in genes on selected PanelApp panels). Given the limitations of the Tiering system (discussed extensively under 'Tiering Issues' (manuscript section 2) in our commentary article (Best et al., 2022a), we suspected that many missed molecular diagnoses would be identifiable amongst Tier 3 and un-tiered variants. We were curious about how much time and effort would be required to improve molecular diagnoses beyond that possible from prioritised Tier 1 and 2 variant analysis, already a huge workload for diagnostic labs in the UK, and whether any strategy in particular could be flagged as a worthwhile and achievable addition to usual service testing. We wanted to know how easy this may or may not be for mainstream clinicians and clinical scientists using available software in the GEL Research Environment, who largely lack training in command-line coding and big data analytics. Personally, I hoped to gain useful transferable skills to take back to my Clinical Genetics training from this experience, to benefit my own patients in the future.

We were interested in missed diagnoses caused by problems in variant detection and

interpretation (e.g. missense variants, non-coding variants, structural variants) and by problems of gene selection driven by panel application, dependent upon the clinical features entered for the participants. We suspected that some patients had been mistakenly recruited as ciliopathies, so would have identifiable molecular diagnoses in non-ciliopathy disease genes that wouldn't have been on the applied panels. We also wanted to use the opportunity to look for potential new diagnoses in candidate ciliopathy genes.

Upon completion of our search for molecular diagnoses, we aimed to present a diagnosis rate from WGS for the cohort, including any diagnostic uplift from undertaking our additional analyses. We also wanted to analyse the distribution of clinical diagnoses identified amongst this previously un-studied cohort. We thought that it would provide an interesting insight into how well the phenotypes associated with primary ciliopathies are recognised in the clinical setting given the known variable expressivity of ciliopathy phenotypes, even amongst relatives. We predicted that this may inform clinicians about differential diagnoses to consider when molecular diagnoses are not easily identified amongst analysed ciliopathy genes. We also wanted to explore the contribution of non-coding and structural variants (SVs) to molecular diagnoses, as the wealth of WGS data available through projects such as 100K offers opportunities to detect these previously difficult-to-detect causative variants.

2.2 Additional methodology

Detailed methodology is provided in the manuscript (thesis section 2.4) and supplementary material (thesis section 6.1.1) (Best et al., 2022b). In addition, variants of interest were run through Ensembl Variant Effect Predictor (VEP) (McLaren et al., 2016) from the command line with additional *in silico* prediction tool plugins from Combined Annotation Dependent Depletion (CADD) (Kircher et al., 2014) and SpliceAI (Jaganathan et al., 2019). This provided an output csv file containing VEP annotated variants. This could be manually analysed in Excel, or else filtered using a further custom Python script written by myself and Dr. Matthew Roche, called *Filter_VEP_output.variants.py*. This is provided in Appendix section 6.2.1. *Filter_VEP_output.variants.py* produced csv files of filtered lists of annotated variants of interest for more concise and focussed analysis as follows in Table 4.

Table 4. Filtering steps applied in the custom Python script *Filter_VEP_output.variants.py*

Note: a CADD_PHRED cutoff score of 15 was selected as this is the recommended threshold for analysis on the CADD website (available from <https://cadd.gs.washington.edu/info>)

Filtering step	Input file	Output file name	Consequence
1	VEP annotated variant list Note: separate input file required for variants called on each chromosome build (GrCh37 and GrCh38)	VEP_filtered_rare.csv	Excludes variants with Minor Allele Frequency (MAF) >0.1% in gnomAD to leave only rare variants for further analysis
2	VEP_filtered_rare.csv	VEP_filtered_high_impact.csv	Creates a sub-file of rare variants annotated by VEP as high impact (stop gain, frameshift, start loss, canonical splice donor, canonical splice acceptor) for focused analysis
3	VEP_filtered_rare.csv	VEP_filtered_ClinVar_pathogenic.csv	Creates a sub-file of rare variants with ClinVar pathogenic or likely_pathogenic entries for focused analysis
4	VEP_filtered_rare.csv	VEP_filtered_missense_all.csv	Creates a sub-file of rare missense variants
5	VEP_filtered_missense_all.csv	VEP_filtered_missense_CADD.csv	Creates a sub-file of rare missense variants with a CADD_PHRED score of >15 for focused analysis
6	VEP_filtered_rare.csv	VEP_filtered_splice_region.csv	Creates a sub-file of rare variants predicted by VEP to be in splice regions for focused analysis

2.3 Additional results

The accompanying manuscript (thesis section 2.4) (Best et al., 2022b) includes research molecular diagnosis for n=43/83 (51.8%) probands in the 100K CMC cohort. Two further diagnoses have been identified post-publication, taking the overall diagnosis rate up to n=45/83 (54.2%).

Participant #59, entered to the BBS category with classical sounding BBS features, had a ClinVar known pathogenic, maternally inherited missense variant in *BBS1*: NM_024649.5:c.1169T>G, NP_078925.3:p.(Met390Arg) identified during the CMC cohort analysis. This was un-tiered by GEL. It is, in fact, a pathogenic founder variant that accounts for approximately 27% of all cases of BBS (Cox et al., 2012). The identification of this heterozygous variant guided us to review the *BBS1* locus on IGV, where we found an unusual region within exon 13 that could not be characterised through visual analysis alone. The IGV trace is presented in our subsequent manuscript “Uncovering the burden of hidden ciliopathies in the 100,000 Genomes Project: a reverse phenotyping approach” (thesis section 3.2, manuscript Figure 3E.i) (Best et al., 2022c). This soft-clipped read signature in exon 13 was consistent with a recently described mobile SVA F family element insertion of size 2.4kb (Delvallee et al., 2021). It was characterised through additional laboratory work by colleagues in the Northeast and Yorkshire Genomic Laboratory Hub. This includes a duplex PCR screening assay (manuscript Figure 3E.iii) and Sanger sequencing of upstream (manuscript Figure 3E.iv) and downstream (manuscript Figure 3E.v) junction fragments, confirming that the 2.4kb mobile element insertion was present in the proband and his father in the same form as previously reported (Best et al., 2022c, Delvallee et al., 2021). Further detail about this diagnosis is available in the “Reportable diagnoses” section of the manuscript discussion (thesis section 3.2) (Best et al., 2022c).

Participant #78, entered to the BBS category with pigmentary retinopathy and obesity only, was diagnosed post-publication with a homozygous, multi-exon *BBS4* deletion of approximately 5.5kb. This diagnosis was made through the SVRare script written by our collaborator Dr. Jing Yu, a senior bioinformatician with the Nuffield Department of Clinical Neurosciences at the University of Oxford (Yu et al., 2022). I used SVRare to search for SVs in the reverse phenotyping study (thesis section 3.2) (Best et al., 2022c). I was not aware of the SVRare approach at the time of the main CMC cohort analysis. SVRare uses a database of 554,060 SVs called by Manta (Chen et al., 2016) and Canvas (Ivakhno et al., 2018) aggregated from 71,408 participants in the rare disease arm of 100K (Yu et al., 2022).

Upon completion of the reverse phenotyping study, I asked Dr. Yu to extract rare SV calls (made in <10 participants) from SVRare that overlapped coding regions of diagnostic grade “green” PanelApp genes from the RMCD super panel version 4.151 (see Table 3 for genes included) amongst the remaining unsolved participants from the CMC cohort. This returned a homozygous deletion of 5524bp on chromosome 15 called by Manta (Chen et al., 2016) in participant #78, that included coding regions of *BBS4*. This SV had been called four times in the 100K rare dataset: two calls in proband #78, one in their father and one in an unrelated proband from the hereditary spastic paraplegia recruitment category. The region was manually inspected on IGV for the proband #78 and their father; no sequence was available for their mother. An IGV capture of *BBS4* including regions from exon 2 to exon 6 is provided in Figure 10. This shows the homozygous *BBS4* deletion including the whole of exons 4 and 5 in the proband and heterozygosity for the deletion in the father, consistent with the SVRare findings. This result was submitted to the GEL Airlock system for return to the recruiting clinician, but no response was received.

Chr 15

Proband 78

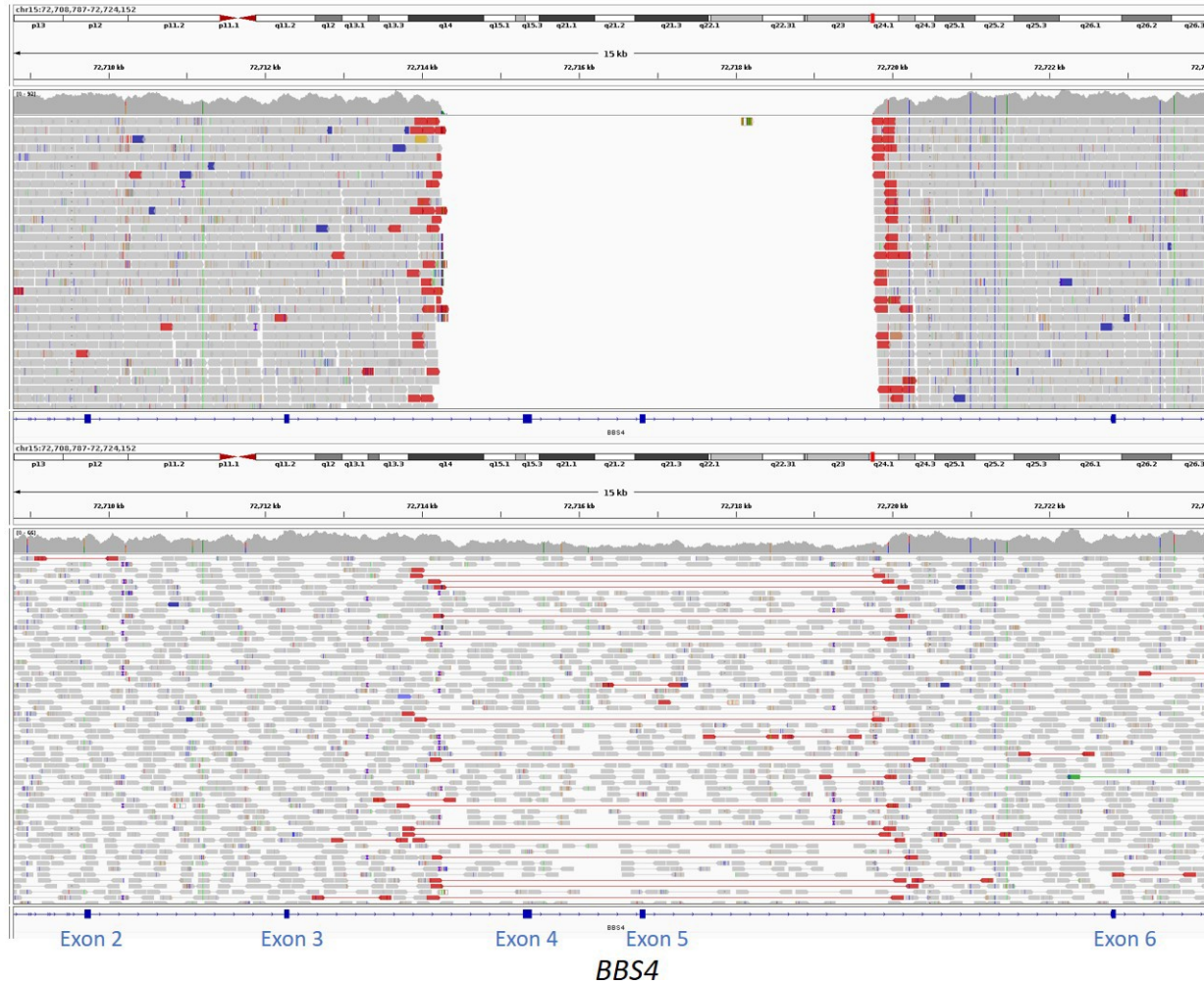


Figure 10. IGV captures showing homozygous BBS4 deletion in CMC proband 78 and heterozygous deletion in their father. The deletion includes the whole of exons 4 and 5



Original research

Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project

Sunayna Best,^{1,2} Jenny Lord,^{3,4} Matthew Roche,⁵ Christopher M Watson ^{6,7}, James A Poulter ¹, Roel P J Bevers,⁸ Alex Stuckey,⁸ Katarzyna Szymanska,¹ Jamie M Ellingford ^{9,10}, Jenny Carmichael,¹¹ Helen Brittain,⁸ Carmel Toomes,¹ Chris Inglehearn,¹ Colin A Johnson,¹ Gabrielle Wheway ^{3,12} Genomics England Research Consortium

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/jmedgenet-2021-108065>).

For numbered affiliations see end of article.

Correspondence to

Dr Gabrielle Wheway, Department of Human Development and Health, University of Southampton Faculty of Medicine, Southampton, SO17 1BJ, UK; g.wheway@soton.ac.uk

CJ and GW are joint senior authors.

Received 1 July 2021
Accepted 27 August 2021
Published Online First 29 October 2021



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY. Published by BMJ.

To cite: Best S, Lord J, Roche M, et al. *J Med Genet* 2022;**59**:737–747.

ABSTRACT

Background Primary ciliopathies represent a group of inherited disorders due to defects in the primary cilium, the ‘cell’s antenna’. The 100,000 Genomes Project was launched in 2012 by Genomics England (GEL), recruiting National Health Service (NHS) patients with eligible rare diseases and cancer. Sequence data were linked to Human Phenotype Ontology (HPO) terms entered by recruiting clinicians.

Methods Eighty-three prescreened probands were recruited to the 100,000 Genomes Project suspected to have congenital malformations caused by ciliopathies in the following disease categories: Bardet-Biedl syndrome (n=45), Joubert syndrome (n=14) and ‘Rare Multisystem Ciliopathy Disorders’ (n=24). We implemented a bespoke variant filtering and analysis strategy to improve molecular diagnostic rates for these participants.

Results We determined a research molecular diagnosis for n=43/83 (51.8%) probands. This is 19.3% higher than previously reported by GEL (n=27/83 (32.5%)). A high proportion of diagnoses are due to variants in non-ciliopathy disease genes (n=19/43, 44.2%) which may reflect difficulties in clinical recognition of ciliopathies. n=11/83 probands (13.3%) had at least one causative variant outside the tiers 1 and 2 variant prioritisation categories (GEL’s automated triaging procedure), which would not be reviewed in standard 100,000 Genomes Project diagnostic strategies. These include four structural variants and three predicted to cause non-canonical splicing defects. Two unrelated participants have biallelic likely pathogenic variants in *LRRC45*, a putative novel ciliopathy disease gene.

Conclusion These data illustrate the power of linking large-scale genome sequence to phenotype information. They demonstrate the value of research collaborations in order to maximise interpretation of genomic data.

INTRODUCTION

Ciliopathies represent a group of inherited genetic disorders that arise as a result of defects in the primary cilium, the ‘cell’s antenna’,¹ or motile cilia, organelles responsible for the movement of fluid over the surface of cells.² They encompass a range of severe developmental and degenerative diseases

that are individually rare but collectively common, affecting an estimated 15.8 million people worldwide including an estimated 133 000 people in the UK. Cilia have also been implicated in conditions such as diabetes, cancer, congenital heart disease and osteoarthritis.^{3–5} As cilia have a near-ubiquitous anatomical distribution, genetic defects affecting the structure or function of cilia cause a range of conditions that can affect multiple organs. Ciliopathies are typically classified into: retinal ciliopathies that exclusively or predominantly affect the eye⁶; renal ciliopathies, which include autosomal dominant polycystic kidney disease affecting around 1:500 people⁷; skeletal ciliopathies that cause a diverse range of skeletal dysplasias and cranio-facial dysmorphism⁸; metabolic or ‘obesity’ ciliopathies⁹; neurodevelopmental ciliopathies¹⁰; and the respiratory motile ciliopathies.¹¹

It is estimated that around 1000 genes contribute to ciliogenesis and cilium function,^{12–15} and ciliopathies are highly genetically heterogeneous.^{16–17} Approximately one-third of the around 270 genes implicated in inherited retinal dystrophies are cilia genes,¹⁸ whereas roughly 20 genes have been associated with renal ciliopathies (PKD OMIM phenotypic series PS173900; nephronophthisis OMIM PS256100). The short-rib polydactyly syndromes, which encompass most of the skeletal ciliopathies, have 22 known genetic causes (OMIM PS208500). There are 24 known genetic causes of the metabolic/obesity ciliopathy Bardet-Biedl syndrome (BBS) (OMIM PS209900). In this same series, Alström syndrome is unusual, because it is a single gene ciliopathy (caused by pathogenic variants in *ALMS1*). There is extensive genetic overlap between neurodevelopmental ciliopathies Joubert syndrome (JBTS) and Meckel-Gruber syndrome (MKS), with 37 known JBTS genes (OMIM PS213300) and 13 MKS genes (OMIM PS249000), many of which also cause JBTS. Several MKS and JBTS disease genes also overlap with the nine genes known to cause complex multiorgan ciliopathy orofacial digital syndrome (OFD) (OMIM PS311200). OFD is considered by some to be a skeletal ciliopathy, involving malformations of the face, mouth and digits, while OFD type 1, which specifically includes

Diagnosics

polycystic kidney disease, may be considered a renal ciliopathy. In total, at least 220 different genes have been shown to cause a single (or multiple) ciliopathy when mutated.

The number of identified ciliopathy disease genes has advanced rapidly since the early to mid-2010s following the ubiquitous implementation of next-generation sequencing (NGS) technologies. Using targeted gene panel, or whole exome sequencing (WES) approaches, genetic diagnosis rates for syndromic primary (non-motile) ciliopathies are typically 40%–70% and for motile (respiratory ciliopathies) are approximately 70% (studies summarised in online supplemental table 1). A recent large whole genome sequencing (WGS) study in 125 families with ciliopathies achieved an 87% diagnosis rate,¹⁶ and a further increase was achieved following the inclusion of structural variant (SV) analysis and RNA sequencing in carefully phenotyped cohorts.¹⁹

The 100,000 Genomes project is a hybrid clinical/research initiative, launched in 2012 and overseen by Genomics England Ltd (GEL), a company set up and wholly owned by the UK Government Department of Health and Social Care.²⁰ The project aimed to sequence 100 000 genomes from 70 000 individuals with rare diseases and cancer. Rare disease patients' genomes were sequenced alongside their family members in a trio testing approach. Cancer patients' germline and somatic genomes were sequenced from matched tumour and normal tissue. Genome sequence data were linked to clinical data from longitudinal patient records and Human Phenotype Ontology (HPO) terms entered by recruiting clinicians. Participants consented to receive a diagnosis for the specific condition they were recruited to the project for and to allow access to their fully anonymised genome sequence data and phenotype information for approved academic and commercial researchers. Recruitment to 190 different rare disease domains took place between 2016 and 2018 across 85 NHS Trusts, coordinated by 13 Genomic Medicine Centres (GMCs). In the data release used in this study (Main Programme Release 11 (17 December 2020)), data were available for 88 918 individuals: 71 682 in the rare diseases arm of the 100,000 Genomes Project and 17 236 in the cancer arm. In the rare diseases arm, 33 329 participants were entered as probands and 38 352 as relatives.

GEL also developed PanelApp (available from <https://panelapp.genomicsengland.co.uk>), a crowdsourcing tool for sharing and evaluation of gene panels by the scientific community.²¹ Virtual gene panels were applied to WGS data to facilitate focused analysis, returning variants in selected genes on curated lists with convincing evidence of an association with the disease(s) of interest. Not only does this shorten the list of variants to analyse, but it also reduces the risk of unwanted incidental findings.

As part of the effort to integrate NGS into standard of care (SOC) testing in the UK's National Health Service (NHS), ciliopathy patients who had previously undergone existing SOC testing (typically gene panel testing) were recruited to the 100,000 Genomes Project to undergo WGS.²² Patients recruited under congenital malformations caused by ciliopathies (CMC) categories (subdivided into BBS, JBTS and rare multisystem ciliopathy disorders (RMCD) or respiratory ciliopathies) accounted for just under 1% of the total rare disease cohort. There were no dedicated recruitment categories for retinal ciliopathies, renal ciliopathies or skeletal ciliopathies, and these were recruited under subcategories of ophthalmological disorders, renal and urinary tract disorders or other categories, and so there are likely to be many further ciliopathy participants in the rare disease cohort. In this study, we aimed to optimise strategies to improve

molecular diagnostic rates for probands recruited to the CMC category within the 100,000 Genomes Project.

MATERIALS AND METHODS

Participant selection and phenotypic classification

Participants recruited under CMC categories were extracted from the GEL Main Programme Release 11 (17 December 2020) using the user interface 'LabKey' within the GEL secure research environment. All data analysis was conducted within the GEL Research Environment. We exported anonymised data for publication through the Airlock system, after review by the GEL Airlock Review Committee. HPO terms recorded for each participant by their recruiting clinicians were assessed within the research environment prior to genetic analysis to determine the most likely clinical diagnosis for each proband based on phenotypic features alone. For selected cases, further clinical information was obtained through the 'Participant Explorer' interface.

Variant filtering and analysis

The GEL data processing pipeline, which includes an automated variant triaging algorithm to classify variants into a series of 'Tiered' categories (as defined by the Genomics England Rare Disease Tiering Process), has been described previously.²² Variants were tiered against 'green' genes listed in PanelApp panels selected according to entered HPO terms. PanelApp provides a traffic light system for genes: 'green' genes are diagnostic grade, 'amber' genes are borderline and 'red' genes have a low level of evidence. In instances where tiered variants did not indicate the cause of disease, untiered single nucleotide variants (SNVs) including heterozygous variants were extracted from participant genomes using a custom Python script ('find_variants_by_gene_and_consequence.py'; available at https://github.com/JLord86/Extract_variants). The script extracts variants in diagnostic grade 'green' genes from provided PanelApp panels and candidate genes with the variant effect predictor (VEP) annotations stop_gained, splice_acceptor, splice_donor, frameshift, missense and splice_region (if the variant was within either the terminal 1–3 bases of the exon or terminal 3–8 bases of the intron).

The script was first run using the RMCD Super Panel V.4.91 (available from <https://panelapp.genomicsengland.co.uk/panels/728/>) (green genes recorded in online supplemental table 2) and ciliopathy candidate genes from several sources. These include all 'red' and 'amber' genes from the PanelApp RMCD panel, genes of interest highlighted by local research teams and all genes on the curated SYSCILIA gold standard (SCGSv1) (online supplemental table 3). If a single potentially pathogenic heterozygous SNV in a recessive gene was identified through this strategy, manual inspection of the whole gene locus was undertaken using the Integrative Genomics Browser (IGV)²³ to determine if a potential SV could be identified as the second biallelic variant. SVs were considered potentially causative if present in >30% of reads.

For those cases that remained unsolved, untiered SNVs were then extracted using further panels compatible with the participant's phenotype. These included: the Retinal Disorders panel V.2.172 for those with retinal dystrophy only (available from <https://panelapp.genomicsengland.co.uk/panels/307/>), the Developmental Disorders Genotype-to-Phenotype database (DDG2P) panel V.2.21 for those with multisystemic developmental disorders (<https://panelapp.genomicsengland.co.uk/panels/484/>), the Laterality Disorders and Isomerism panel V.1.21 for those with a laterality defect (<https://panelapp.genomicsengland.co.uk/panels/549/>) and the Broad Renal Super

panel V.2.346 for those with isolated renal anomalies (<https://panelapp.genomicsengland.co.uk/panels/902/>).

For all remaining unsolved participants, variants potentially affecting splicing (SpliceAI delta scores >0.5) in diagnostic grade 'green' genes) from the PanelApp RMCD panel were extracted with a further custom Python script ('find_variants_by_gene_and_SpliceAI_score.py'; available at https://github.com/JLord86/Extract_variants).²⁴ Finally, the find_variants_by_gene_and_SpliceAI_score.py Python script was run again using the DDG2P panel V.2.21 for all remaining unsolved participants.

Bespoke research variant analysis pipeline

All data analysis was conducted within the secure online Research Environment including interrogation of BAM, VCF, SV and HPO information files. The Ensembl VEP was used to obtain variant information for interpretation of variant pathogenicity.²⁵ Information about associations between genes and disease phenotypes was obtained from the OMIM database (<https://www.omim.org>). The mode of inheritance was defined according to the literature and OMIM for each gene. Variant evidence was reviewed using ACMG/AMP guidelines for clinical variant interpretation,²⁶ and each variant of interest was assigned a pathogenicity score according to current (Association for Clinical Genomic Science (ACGS) guidelines.²⁷

The research analysis workflow comprised steps to filter genomic data (figure 1A), assess putative pathogenic variants (figure 1B), then classify and assign diagnostic confidence (figure 1C).

Variant classification and diagnostic confidence

To benchmark our ability to appropriately classify and interpret identified variants, first-pass analysis was blinded to previous results, and then verified against the GEL reported findings in the GMC exit questionnaires. These were completed by regional NHS GMCs for each analysed participant. Recruiting clinicians were contacted through the GEL secure airlock system for notification of a research molecular diagnoses, if they did not have a consistent completed GMC exit questionnaire. Additional clinical data were requested, where required, using the 'contact the clinician' form. All diagnoses identified through this blinded research strategy were termed 'research molecular diagnoses'. The interpretation of these findings was subdivided into 'confident', 'probable' or 'possible' according to the ACMG classification for each variant, the inheritance pattern of the identified condition and the match to the proband's phenotypic features (summarised in figure 1C).

RESULTS

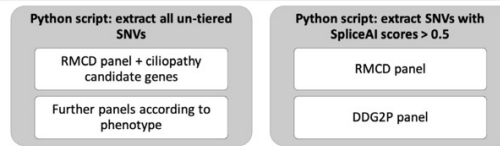
Congenital malformations caused by ciliopathies cohort

A total of 83 probands were identified in the CMC cohort. This was subdivided into 45 in the BBS category, 14 in the JBTS category and 24 in the RMCD category. Fifteen participants were recruited as singleton cases, and for 68 individuals at least one additional family member underwent WGS. Including probands and relatives, genomic data were available for 211 individuals.

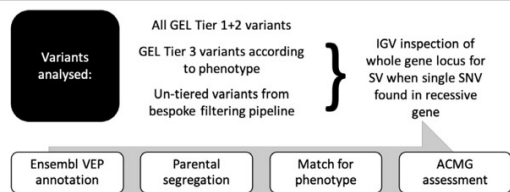
HPO term analysis

Analysis of HPO terms for the 83 probands shows that for 51 cases, phenotypes were consistent with their disease recruitment category. The remaining 32 probands lack recorded phenotypes suggestive of a syndromic ciliopathy (table 1). This suggests that participants were either frequently misdiagnosed as having ciliopathies or HPO terms were not entered accurately.

A: Bespoke research variant filtering pipeline



B: Bespoke variant analysis pipeline



C. Variant classification and diagnostic confidence

Full or partial match to phenotype in OMIM morbid gene			
Mode of inheritance	Confident diagnosis	Probable diagnosis	Possible diagnosis
Recessive	2 pathogenic / likely pathogenic variants	1 pathogenic / likely pathogenic + 1 VUS	2 VUSs
Dominant / X-linked	1 pathogenic / likely pathogenic variant	N/a	1 VUS

Inconsistent match to phenotype / variant in candidate gene			
Mode of inheritance	Confident diagnosis	Probable diagnosis	Possible diagnosis
Recessive	N/a	N/a	2 pathogenic / likely pathogenic variants
Dominant / X-linked	N/a	N/a	1 pathogenic / likely pathogenic variant

Figure 1 Research analysis workflow that (A) describes steps to filter genomic data, (B) analyse putative pathogenic variants and (C) classify variants then assign diagnostic confidence. ACMG, Association for Clinical Genomic Science; DDG2P, Development Disorder Genotype - Phenotype Database; GEL, Genomics England; IGV, Integrative Genomics Browser; RMCD, rare multisystem ciliopathy disorders; SNV, single nucleotide variant; SV, structural variant; VEP, variant effect predictor; VUS, variant of uncertain significance.

Tiered variants

Thirty-eight tier 1 variants were identified in 28 different genes among 29 different probands in the CMC cohort. Two hundred and sixteen tier 2 variants were identified in 142 different genes among 53 different probands. A total of 8777 tier 3 variants were identified in 5220 different genes among all 83 probands. No SVs had been tiered.

GEL reported molecular diagnoses

GMC exit questionnaires were completed for 67/83 (80.7%) patients by Release 11 (released 17 December 2020) (table 1). Twenty-three participants (27.7%) had GMC exit questionnaires reporting causative tier 1 or tier 2 variants, with one case partially solved and 22 fully solved. Four GMC exit questionnaires reported variants of uncertain significance (VUS) (figure 2A).

We identified that one of the cases previously reported as solved was a false positive. The GMC questionnaire reported compound heterozygous *ALMS1* variants in participant #32 including an untiered heterozygous exon 11 deletion. The deletion was not visible using the IGV or detectable in the patients

Diagnostics

Table 1 Anonymised phenotypic and research molecular diagnosis data for the probands in the congenital malformations caused by ciliopathies cohort

Research number	Recruitment category	Most likely clinical diagnosis based on HPO terms	Does recruitment category match most likely clinical diagnosis?	GEL GMC exit report	Research molecular diagnosis	Gene	Is identified diagnosis a ciliopathy?	Diagnostic confidence
1	JBTS	JBTS	Yes	Sol	CHARGE Syn	<i>CHD7</i>	No	Conf
2	BBS	Non-cil MS cond	No	Sol	Alström Syn	<i>ALMS1</i>	Yes	Conf
3	BBS	BBS	Yes	Sol	BBS + RP	<i>ARL6 + IMPG2</i>	Yes	Conf
4	BBS	BBS	Yes	Sol	RP	<i>RPGR</i>	Yes	Conf
5	BBS	Non-cil MS cond	No	Sol	Retinal cil, possibly syndromic	<i>CEP290</i>	Yes	Conf
6	JBTS	JBTS	Yes	Sol	JBTS	<i>KIAA0586</i>	Yes	Conf
7	RMCD	OFD-like cil	Yes	Sol	OFD1, PKD +inherited cataract	<i>OFD1, PKD1, CRYBB1</i>	Yes (<i>OFD1</i>)	<i>OFD1</i> Conf, <i>PKD1 + CRYBB1</i> Poss
8	BBS	Isol RD	No	Sol	RP	<i>PRPF8</i>	No	Conf
9	RMCD	JBTS-like MS cil	Yes	Uns	Seckel Syn	<i>CEP152</i>	No	Poss
10	JBTS	JBTS	Yes	Sol	JBTS	<i>CEP290</i>	Yes	Conf
11	RMCD	Jeune-like cil	Yes	Unr	Feingold Syn	<i>MYCN</i>	No	Conf
12	JBTS	JBTS	Yes	Unr	JBTS	<i>ARMC9</i>	Yes	Conf
13	BBS	BBS	Yes	Unr	Tubulinopathy	<i>TUBA1A</i>	No	Poss
14	RMCD	Jeune-like cil	Yes	Unr	Jeune Syn	<i>WDR19</i>	Yes	Conf
15	BBS	Isol RD	No	Unr	RP	<i>RHO</i>	No	Conf
16	RMCD	Non-cil MS cond	No	VUS	STAG1 syndromic ID syn	<i>STAG1</i>	No	Prob
17	BBS	BBS	Yes	Sol	BBS	<i>BBS1</i>	Yes	Conf
18	BBS	BBS	Yes	Sol	Neurodevelopmental disorder	<i>RERE</i>	No	Conf
19	BBS	BBS	Yes	Sol	Alström Syn	<i>ALMS1</i>	Yes	Conf
20	BBS	Isol eye cond (not RD)	No	Sol	BBS	<i>BBS2</i>	Yes	Conf
21	JBTS	JBTS	Yes	Unr	Poretti-Boltshauser Syn+ArboledaTham Syn	<i>LAMA1, KAT6A</i>	No	<i>LAMA1</i> Prob, <i>KAT6A</i> Poss
22	BBS	BBS	Yes	Sol	BBS	<i>MKKS</i>	Yes	Conf
23	JBTS	JBTS	Yes	Sol	JBTS	<i>CEP290</i>	Yes	Prob
24	BBS	Non-cil MS cond	No	Uns				Uns
25	BBS	BBS	Yes	Sol	Smith Magenis Syn	<i>RAI1</i>	No	Conf
26	BBS	BBS	Yes	Sol	Cone-rod dystrophy	<i>PROM1</i>	No	Conf
27	JBTS	Non-cil MS cond	No	Unr	Luscan-Lumish Syn	<i>SETD2</i>	No	Conf
28	BBS	Non-cil MS cond	No	Sol	Optic Atrophy	<i>OPA1</i>	No	Conf
29	BBS	Non-cil MS cond	No	Sol	Alström Syn	<i>ALMS1</i>	Yes	Conf
30	BBS	BBS	Yes	Sol	Chung-Jansen Syn	<i>PHIP</i>	No	Conf
31	BBS	Isol RD	No	Sol	Cone-rod dystrophy	<i>RAB28</i>	Yes	Conf
32	BBS	BBS	Yes	Sol	None: Unsolved	<i>ALMS1</i>	N/a	False+ve
33	RMCD	Non-cil MS cond	No	Uns				Uns
34	RMCD	Non-cil MS cond	No	Uns	Van Esch-O'Driscoll Syn	<i>POLA1</i>	No	Poss
35	JBTS	JBTS	Yes	Uns				Uns
36	JBTS	JBTS	Yes	Uns				Uns
37	RMCD	Non-cil MS cond	No	Uns				Uns
38	BBS	BBS	Yes	Uns				Uns
39	BBS	BBS	Yes	Uns				Uns
40	BBS	BBS	Yes	Uns				Uns
41	JBTS	JBTS	Yes	Uns	JBTS	<i>CSPP1</i>	Yes	Prob
42	JBTS	JBTS	Yes	Unr	JBTS	<i>PIBF1</i>	Yes	Prob
43	BBS	BBS	Yes	Uns				Uns
44	RMCD	Non-cil MS cond	No	Uns				Uns
45	BBS	Isol polydactyly	No	Uns				Uns
46	RMCD	MKS/JBTS-like MS cil	Yes	Uns				Uns
47	BBS	Non-cil MS cond	No	Unr				Uns
48	RMCD	BBS-like MS cil	Yes	Uns	Candidate cil	<i>LRRC45</i>	Candidate	Poss
49	RMCD	Non-cil MS cond	No	Unr				Uns
50	BBS	BBS	Yes	Unr				Uns
51	RMCD	DM	DM	Unr				Uns

Continued

Table 1 Continued

Research number	Recruitment category	Most likely clinical diagnosis based on HPO terms	Does recruitment category match most likely clinical diagnosis?	GEL GMC exit report	Research molecular diagnosis	Gene	Is identified diagnosis a ciliopathy?	Diagnostic confidence
52	RMCD	JBTS-like MS cil	Yes	Unr				Uns
53	RMCD	Isol GI disorder	No	Unr				Uns
54	RMCD	Non-cil MS cond	No	Uns				Uns
55	JBTS	JBTS	Yes	Uns				Uns
56	BBS	Isol eye cond (not RD)	No	VUS	BBS	<i>BBS9</i>	Yes	Poss
57	JBTS	JBTS	Yes	Uns				Uns
58	RMCD	JBTS-like MS cil	Yes	Uns				Uns
59	BBS	BBS	Yes	Uns				Uns
60	BBS	BBS	Yes	Uns				Uns
61	RMCD	Non-cil MS cond	No	Unr	WT1-related disorder	<i>WT1</i>	No	Conf
62	RMCD	Non-cil MS cond	No	Uns				Uns
63	RMCD	Non-cil MS cond	No	Uns				Uns
64	RMCD	JBTS-like MS cil	Yes	Uns				Uns
65	BBS	BBS	Yes	Uns				Uns
66	RMCD	BBS-like MS cil	Yes	Uns				Uns
67	BBS	Non-cil MS cond	No	VUS	Alström Syn	<i>ALMS1</i>	Yes	Poss
68	JBTS	JBTS	Yes	Uns				Uns
69	BBS	BBS	Yes	Sol	BBS	<i>BBS1</i>	Yes	Conf
70	BBS	Non-cil MS cond	No	Uns				Uns
71	RMCD	Non-cil MS cond	No	Unr	Shukla-Vernon Syn	<i>BCORL1</i>	No	Poss
72	BBS	BBS	Yes	Unr	Sifrim-Hitz-Weiss Syn	<i>CHD4</i>	No	Poss
73	RMCD	Isol GI disorder	No	Uns				Uns
74	BBS	Non-cil MS cond	No	Uns				Uns
75	BBS	DM	DM	Unr	BBS	<i>BBS4</i>	Yes	Poss
76	BBS	BBS	Yes	VUS	BBS	<i>BBS10</i>	Yes	Poss
77	BBS	BBS	Yes	Uns				Uns
78	BBS	BBS	Yes	Uns				Uns
79	BBS	BBS	Yes	Uns				Uns
80	BBS	BBS	Yes	Uns				Uns
81	BBS	BBS	Yes	Uns				Uns
82	BBS	Non-cil MS cond	No	Unr	Attenuated mucopolysaccharidosis 1	<i>IDUA</i>	No	Prob
83	BBS	BBS	Yes	Uns				Uns

Table includes the recruitment category, designated 'most likely' clinical diagnosis based on entered HPO terms alone, GEL GMC exit questionnaire reporting outcome, research molecular diagnosis (determined by genotype), responsible gene, whether the identified diagnosis is a ciliopathy and diagnostic confidence. Note: individual variant information, including data taken into consideration in forming ACMG classifications, can be found in online supplemental table 4.

BBS, Bardet-Biedl syndrome; Cil, ciliopathy; Cond, condition; Conf, confident; DM, data missing; GEL, Genomics England; GI, gastrointestinal; GMC, Genomic Medicine Centres; HPO, Human Phenotype Ontology; Isol, isolated; JBTS, Joubert syndrome; MKS, Meckel Gruber syndrome; MS, multisystemic; PKD, polycystic kidney disease; Poss, possible; Prob, probable; RD, retinal dystrophy; RMCD, rare multisystem ciliopathy disorders; RP, retinitis pigmentosa; Sol, solved; Syn, syndrome; Unr, unreported; Uns, unsolved.

VCF file; following correspondence with the GEL helpdesk. the variant was confirmed to be a false positive.

Identification of research molecular diagnoses

Our bespoke variant-to-diagnosis pipeline shows that 43 of the 83 probands (51.8%) have a research molecular diagnosis that is compatible with their phenotypic features (table 1). Individual variant information, including data taken into consideration in performing ACMG classification, is recorded in online supplemental table 4. Twenty-eight of the 83 participants (33.7%) are classified as having a confident diagnosis, 5/83 (6%) a probable diagnosis and 10/83 (12%) only a possible diagnosis (figure 2B). Overall, 34/83 participants (41%) had a research molecular diagnosis that fully accounted for their entered phenotypic features and 9/83 (10.8%) that partially accounted for their entered features (online supplemental table

4). No phenotypic features were entered for proband #75, but the possible molecular diagnosis of BBS matches their BBS recruitment category. Diagnoses according to recruitment category are shown in figure 2C.

Seventeen of the 43 research molecular diagnoses (39.5%) can be considered novel findings. Fourteen diagnoses are new findings in probands with no completed GMC exit questionnaire (unreported) and three are in probands with negative GMC outcome questionnaires (reported as 'unsolved'). Interestingly, a significant proportion of research molecular diagnoses have been made in non-ciliopathy genes. Only 23 of the 43 potentially solved participants (53.5%) have variants in genes known to be causative of ciliopathy syndromes. The remaining 19/43 potentially solved probands (44.2%) have variants identified in non-ciliopathy genes.

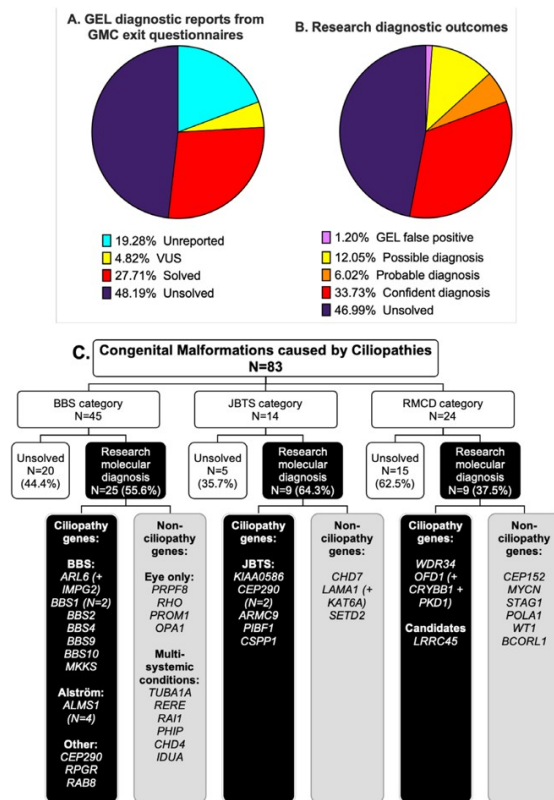


Figure 2 Comparison of diagnostic reporting outcomes between gel GMC exit reports (A) and research diagnostic outcomes (B) for the 83 probands in the CMC cohort. (C) Research molecular diagnoses according to recruitment category. Genes with identified potentially causative variants are grouped according to whether they are known to be associated with ciliopathies or not. A '+' is used where participants had potentially causative variants in more than one gene contributing to their clinical features (additional gene(s) are included in brackets). Diagnostic confidence for each research molecular diagnosis is shown in table 1. Detailed variant information, including whether the gene variant(s) are thought to be a full or partial match to phenotype, is provided in online supplemental table 4. BBS, Bardet-Biedl syndrome; CMC, congenital malformations caused by ciliopathies; GEL, Genomics England; GMC, Genomic Medicine Centre; JBTS, Joubert syndrome; RMCD, rare multisystem ciliopathy disorder.

Research molecular diagnoses made outside GEL tiers 1 and 2

Thirty-two of the 83 probands (38.5%) have research molecular diagnoses made from tier 1 and 2 variants only. The remaining 11/83 probands (13.3%) with research molecular diagnoses have at least one variant outside of tiers 1 and 2 (variant information provided in online supplemental table 4). These diagnoses would have been missed by the standard 100,000 Genomes Project diagnostic pipeline, which routinely inspects only tier 1 and 2 variants. Five tier 3 variants and 12 untiered variants contribute to the diagnoses for these 11 participants. Three of the untiered variants are SVs (IGV captures shown in figure 3); the other nine are SNVs identified through our bespoke filtering pipeline. Interestingly, a variant annotated by GEL as a tier 2 *ALMS1* missense was discovered via IGV inspection to be an indel (92 nucleotide

deletion and 31 nucleotide insertion) leading to a splice acceptor change (participant #29, shown in figure 3A).

SpliceAI analysis of variants filtered using our pipeline identified three untiered ciliopathy gene variants predicted to cause splice donor site losses. One is a homozygous synonymous variant in *ARL6* in proband #3, entered with suspected BBS (NM_001278293.3:c.534A>G, NP_001265222.1:p.Gln178=) (online supplemental table 4). The overall allele frequency (AF) on gnomAD is 0.000007960 with zero homozygotes.²⁸ The 100,000 Genomes Project AF is 0.00049985 for participants called on GrCh37 (one heterozygote) and 0.0000571872 for participants called on GrCh38 (three heterozygotes and three homozygotes). On further analysis, the two further homozygous individuals were identified as affected siblings of proband #3. The heterozygous individuals are the parents of proband #3 plus one unrelated participant. This variant has previously been published in association with BBS and proven to cause aberrant splicing in vitro by minigene assay.²⁹ The other two are at +3 and +5 positions in probands #75 (*BBS4* NM_033028.5:c.642+3A>T) and #41 (*CSPP1* NM_001382391.1:c.2968+5G>A). Clinical material was not available for testing to validate splicing effects at the molecular level. Therefore, both have been classified as VUSs.

Putative novel disease genes

Participant #48, entered to the RMCD category and determined most likely to have BBS based on entered HPO terms, has two separate homozygous, protein-truncating variants in candidate ciliopathy genes. Proband #48 has a sibling who was separately entered to the 100,000 Genomes Project in the intellectual disability category, without additional features suggestive of a syndromic ciliopathy. Further phenotypic analysis using the Participant Explorer tool revealed that participant #48 also has clinical features suggestive of a motile ciliopathy. Specific clinical features cannot be provided to protect participant anonymity. There is a recorded history of parental consanguinity in this family.

The first variant of interest identified in participant #48 is a homozygous frameshift variant in *LRRC45* (GrCh38 chromosome 17: 82028260 C>CTG; NM_144999.4:c.1074_1075insTG, NP_659436.1:p.Leu359CysfsTer19). This was also found to be homozygous in the proband's sibling from the intellectual disability category. Segregation analysis is consistent with autosomal recessive inheritance; both parents are confirmed heterozygotes. According to the Illumina Region of Homozygosity (ROH) caller, this *LRRC45* variant is in a 1 359 569 base pair ROH (GrCh38 chromosome 17: 81841582–83201151) containing 797 homozygous and zero heterozygous variants (ROH score 19.92) in the proband and an 1 364 960 base pair ROH (GrCh38 chromosome 17: 81841582–83206542) containing 728 homozygous and zero heterozygous variants (ROH score 18.2) in the sibling. The second variant of interest is a homozygous stop gain variant in *CFAP45* (*CCDC19*) (GrCh38 chromosome 1: 159887996 G>A; NM_012337.3:c.433C>T, NP_036469.2:p.Arg145Ter) (online supplemental table 4). Segregation analysis showed again that the parents are both heterozygotes but the sibling in the intellectual disability category is homozygous for the reference allele. This *CFAP45* variant is in a 8142476 bp ROH (GrCh38 chromosome 1: 158386429–166528905) containing 3821 homozygous and zero heterozygous variants (ROH score 95.53), not present in the sibling.

Next, we searched for other biallelic, potentially causative variants in either *LRRC45* or *CFAP45* across the entire rare disease 100 000 genomes dataset to gain independent replication

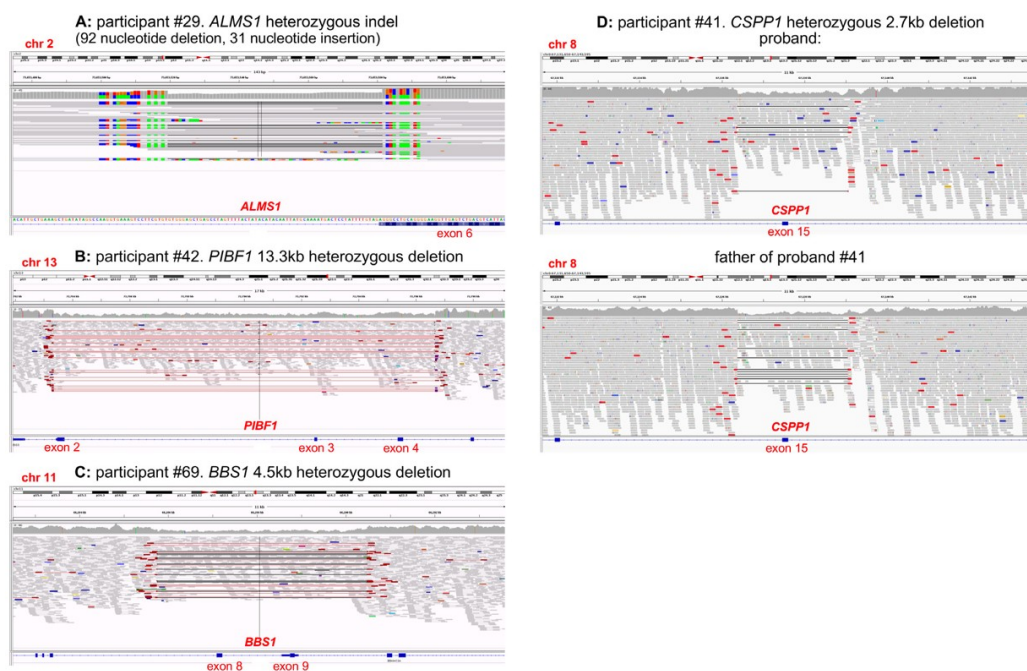


Figure 3 IGV captures of structural variants identified among participants of the congenital malformations caused by ciliopathies cohort. First, an untiered *ALMS1* SV identified in participant #29 was initially called a tier 2 *ALMS1* missense variant. Closer inspection on IGV determined that this was an indel (92 nucleotide deletion and 31 nucleotide insertion) leading to a splice acceptor change at the beginning of exon 6 (A). Our filtering pipeline identified a second untiered *ALMS1* frameshift variant, completing the molecular diagnosis of Alström syndrome. Three larger heterozygous deletions were identified through manual IGV inspection of whole gene loci when searching for second hits in probands with potentially causative SNVs. An untiered 13.3 kb deletion in *PIBF1* (also known as *CEP90*) (B) was identified in a proband with an untiered novel missense variant (proband #42). An untiered 4.5 kb deletion in *BBS1* (C) was found in a proband with an untiered, ClinVar pathogenic missense variant (proband #69). Finally, a 2.7 kb deletion in *CSPP1* (D) was found in a proband with a predicted splice donor loss (SpliceAI DS_DL 0.79) (proband #41). This *CSPP1* deletion was only seen in ~30% of reads in the proband but in ~50% of reads in their father. SNV, single nucleotide variant.

of causality. No additional potentially pathogenic variants were identified for *CFAP45*. However, we identified a second proband with *LRRC45* variants within the cone-rod dystrophy recruitment category and with an ‘unsolved’ GMC exit questionnaire. We identified a heterozygous *LRRC45* start loss variant: NM_144999.4:c.1A>T, NP_659436.1:p.Met1? (absent from gnomAD, GEL 100K MAF 1.271×10^{-5}), and a heterozygous splice acceptor variant: NM_144999.4:c.1126-1G>A (gnomAD allele frequency 8.059×10^{-6} , GEL 100K MAF 2.542×10^{-5}). The proband was entered as a singleton participant, so parental sequence is not available in the 100,000 Genomes Project or on clinician request to establish phase. *LRRC45* therefore remains a putative novel disease gene accounting for the phenotype in these individuals.

DISCUSSION

Diagnosis rate for participants in the CMC cohort of the 100,000 Genomes Project

This study provides a research molecular diagnosis from WGS data for just over half of the participants in the CMC cohort of the 100,000 Genomes Project (43/83, 51.8%), 33 of which are classified as confident or probable (39.8%). Our overall diagnosis rate is 19.3% higher than the 27/83 (32.5%) with GEL reported findings in GMC exit questionnaires (23/83 reported as solved plus 4/83 with VUSs). It is likely that at least nine of the novel research molecular diagnoses would eventually be made

and reported by GEL given that they contain only tier 1 or 2 variants (participants #11, #12, #13, #14, #15, #21, #27, #72 and #75). In identifying and alerting clinical teams, we are providing benefit to participants who have, in some cases, been waiting years for identification of a molecular diagnosis (recruitment to the 100,000 Genomes Project ended in 2018).

There are 11 participants with research molecular diagnoses with at least one variant outside of tiers 1 and 2, which would be missed by the standard diagnostic strategy of inspecting only those variants. Therefore, the added diagnostic value of undertaking analyses outside tiers 1 and 2 is at least 11/83 (13.3%). This highlights the value of research collaborations to investigate unsolved cases and improved diagnosis rates from accessible genomic data.

Unfortunately, major challenges remain in returning research identified diagnoses to recruiting clinicians to ensure they are successfully fed back to participants, which is being addressed with collaborators at GEL. Improved communication between recruiting clinicians and researchers would facilitate better interpretation of variants, but a lack of an automated system for researcher/clinician contact introduces a significant bottleneck, and the long time between recruitment and research identified molecular diagnosis has meant that some recruiting clinicians no longer work in the NHS trust and GMC where they recruited patients to the project, and there is no mechanism of forwarding emails in cases such as this. Recruiting clinician collaboration is

Diagnostics

hugely valuable to provide additional clinical information where required, as well as contacting patients to ask for consent to publication of more detailed clinical data. Furthermore, they can obtain relevant tissue samples to validate variant effects, particularly useful for novel splice variants and SVs.

Conditions identified

Among probands in the CMC cohort with research molecular diagnoses, a surprisingly high proportion have causative variants in non-ciliopathy genes (19/43, 44.2%). This suggests that there are likely to be significant numbers of participants with ciliopathies recruited to other rare disease categories. This misdiagnosis rate may be because primary ciliopathies can be difficult to recognise clinically due to the great diversity of possible disease features. More specific 'hard' phenotypic features can signpost healthcare professionals to the likelihood of a ciliopathy syndrome, but these are not always present. The best example is the molar tooth sign, which is the pathognomonic sign for JBTS-related conditions with no differential diagnoses.³⁰ This is reflected in the highest correlation between recruitment category and identified molecular diagnosis rate being for the JBTS group: 6/14 (42.9%) were recruited as suspected JBTS, and then confirmed to have JBTS at the molecular level. Ten of the 14 patients recruited with suspected JBTS had the HPO term 'Molar Tooth Sign on MRI' entered by the recruiting clinician, including all six that were solved at the molecular level.

Another reason for the high proportion of non-ciliopathy diagnoses could be limitations or difficulties in choosing appropriate recruitment categories for participants of the 100,000 Genomes Project. Categories may have been selected for convenience or lack of awareness of alternative, potentially more appropriate options. The RMCD category may have been treated as a 'catch-all' group for participants with constellations of multisystemic features, not obviously recognisable as a specific syndrome. This is reflected by this group having the lowest diagnosis rate of the three included in the CMC cohort: 9/24 (37.5%) have a research-identified molecular diagnosis, but only two are ciliopathies.

An important outcome to explore further is the relatively high number of participants recruited in the BBS category, found to have variants causative of isolated eye disorders (n=4). It is unclear if recruiting clinicians suspected BBS due to the presence of non-ocular features or whether the participants were inappropriately included in the BBS category. This problem clearly demonstrates the importance of accurate and comprehensive phenotyping to refine the interpretation of sequence variants.

Mutational mechanism of causative variants

Sixty-four individual, potentially causative variants, have been identified in this research study (online supplemental table 4). Of the variants detected, at least four would not have been detectable or accurately described by WES or gene panel, as they are SVs including significant intronic regions (figure 3). Ideally, all SVs of interest should be confirmed by long-range PCR and either third generation nanopore or Sanger sequencing, but DNA samples from these cases could not be obtained from referring clinicians. A recent study of NHS rare diseases patients undergoing WGS, reported 102 large deletions and six complex SVs from 1103 distinct causal variants (9.8% SVs).³¹ Our identified rate of SVs is slightly lower at 4/64 (6.3%). It seems likely that further SVs are responsible for a proportion of the unsolved participants in the CMC cohort, but strategies to detect them are not yet well established.

WGS, particularly PCR-free WGS, offers great advantages in SV analysis over WES, due to even coverage of the whole genome permitting reliable identification of SVs, but we are yet to fully take advantage of these methodologies. The GEL dataset is being used to improve the way we analyse SVs, with a gnomAD-type database of all SVs in GEL with allele frequencies in the cohort having been developed by Jing Yu in Oxford to permit exclusion of SVs from analysis in a patient if that SV appears above a particular minor allele frequency (MAF) in the GEL dataset. PCR-free WGS adds the further benefit of improved coverage of GC rich regions of the genome that are not efficiently amplified in PCR. As many promoter regions are GC rich, this provides an advantage for identifying regulatory region variants.

A further benefit of WGS over WES or gene panel testing is the opportunity to analyse intronic regions. We used the *in silico* tool SpliceAI to find variants predicted to cause novel splicing effects and identified three variants outside the canonical splice sites predicted to cause splice donor site defects. No novel splicing variants were identified in genes from the DDG2P gene panel using our SpliceAI script in unsolved participants of the cohort. However, given the diversity of diagnoses, it is highly likely that further causative splicing variants could be found in non-ciliopathy genes. As well as splice variant identification, intronic WGS data can also be interrogated for regulatory region variants implicated in human disease, using resources such as the UTRannotator tool to annotate high-impact 5' untranslated region variants either creating new upstream opening reading frames (ORFs) or disrupting existing upstream ORFs.³²

Despite the many advantages of WGS over WES, WES remains a popular sequencing strategy as it involves sequencing of only around 2% of the genome, significantly lowering costs of sequencing, permitting sequencing to greater depth on a limited budget, lowering demands on data storage, increasing analysis times and reducing workload for clinical scientists and researchers to process and interpret the significantly smaller number of identified variants. Furthermore, coding region variants are more straightforward to classify, making analysis of WES data more straightforward than analysis of WGS data.

Candidate gene analysis

A list of 302 candidate ciliopathy genes (online supplemental table 3) was used in conjunction with our custom variant filtering pipeline in pursuit of diagnosis for probands unsolved through tiered variant analysis. One proband, participant #48, has two homozygous, protein-truncating variants in the candidate ciliopathy genes *LRRC45*, a protein associated with distal appendages of the basal body that contributes to early steps of axoneme extension during ciliogenesis,³³ and *CFAP45*, a coiled coil domain protein and expressed in nasopharyngeal epithelium and trachea.³⁴

There are various possibilities regarding the potential contribution of these variants to the clinical features of proband #48 and their sibling in the intellectual disability category. The two siblings share neurodevelopmental delay and intellectual disability. Proband #48 also has additional features in keeping with both syndromic primary and motile ciliopathies. *CFAP45* has been recently published as a motile ciliopathy gene,³⁵ so it is possible that the homozygous nonsense *CFAP45* variant present in participant #48 but not their sibling could account for the clinical motile ciliopathy features in participant #48, with the *LRRC45* variants accounting for the neurodevelopmental delay and intellectual disability in both siblings.

Given the phenotypic heterogeneity in ciliopathies even within families with the same variant, another hypothesis is that the two siblings have different presentations of a condition caused by their shared homozygous *LRRC45* frame-shift variant. The putative loss of function (pLoF) gnomAD score for *LRRC45* (pLoF=0.88) suggests that *LRRC45* is not tolerant to loss of function.²⁸ The additional proband from the cone-rod dystrophy category with compound heterozygous high impact *LRRC45* variants adds to the evidence that this may be a ciliopathy gene.

Value of diagnoses

Undertaking broad genomic tests like WES and WGS can curtail the ‘diagnostic odyssey’ experienced by many patients with rare disorders, potentially sparing them multiple invasive tests and misdiagnoses.³⁶ Analysis can be iterative such that the data can be ‘opened up’ beyond the first virtual gene panel without the need for serial testing. Results from this study demonstrate the value of this approach, given the high proportion of participants with non-ciliopathy diagnoses. The NHS Genomic Medicine service, introduced in 2018 as a follow on from the 100,000 Genomes Project, provides a curated National Genomic Test Directory including WES and WGS where appropriate.²⁰ This will embed genomic testing into mainstream care and standardise testing across the country.

Determining the underlying genotype for a patient’s phenotype allows provision of accurate information about their condition, including potential current and future associated features for which screening or treatment may be available. An example of this in action is participant #61, recruited in the RMCD category. An untiered heterozygous missense variant in *WT1* was identified through our filtering that is listed as pathogenic on ClinVar, in keeping with autosomal dominant *WT1*-related disorder. This diagnosis, which was successfully fed back to the recruiting clinician, is considered especially important given the associated risk of Wilms’ tumour and the recommendation for regular screening to facilitate early detection and treatment.³⁷

Lack of a genetic diagnosis can lead to inappropriate management of conditions and delays in accessing specialised services such as the multidisciplinary service for BBS and Alström syndrome in Birmingham Children’s Hospital and Great Ormond Street Hospital in the UK. Without greater awareness and higher diagnosis rates of ciliopathies, it may continue to be difficult to secure funding for additional specialist services for rare ciliopathies.

Perspective on the future of genetic diagnosis

This study prompts reconsideration of approaches to genetic diagnostics, particularly traditional forward genetics in comparison with reverse phenotyping. Classically, clinicians have suggested a possible underlying diagnosis based on the collection of clinical features observed, then the lab have tested for variants in gene(s) associated with that suspected diagnosis. This study demonstrates the utility of a reverse genetics strategy, by going ‘backwards’ from variants that are assessed as pathogenic at the molecular level, to determine if they could match with the patient’s features and the disease’s inheritance pattern. As the cost and availability of large-scale sequencing tests including WES and WGS continues to fall, this reverse phenotyping strategy is becoming increasingly integrated into NHS genetic diagnostics. With this, the current bottleneck is clinical interpretation of variants. To realise the potential of WES and WGS, investment into

dedicated time and resourcing for specialist variant interpretation is essential, as is careful and comprehensive phenotyping and strong communication between clinical scientists, clinical geneticists, mainstream clinicians and researchers. Improved integration of SV and splice variant analysis tools, such as SpliceAI, will be essential to maximise the diagnostic potential of WGS data beyond coding variants in exons of virtual panels of genes. The 19.3% genetic diagnosis uplift achieved in our study demonstrates what can be achieved with additional time and resources invested into WGS analysis. Now that this variant filtering and analysis pipeline has been established, we anticipate that this additional analysis can be achieved within days or weeks rather than months.

Clearly, large-scale genomic studies such as the 100,000 Genomes Project offer huge opportunities to improve diagnostics, understanding of disease mechanisms and identification of novel drug targets. The current challenge is to improve our strategies to analyse sequence data to provide the maximum benefit for patients and the scientific community.

Author affiliations

¹Division of Molecular Medicine, University of Leeds Leeds Institute of Medical Research at St James’s, Leeds, UK

²Department of Clinical Genetics, Leeds Teaching Hospitals NHS Trust, Leeds, UK

³Department of Human Development and Health, University of Southampton Faculty of Medicine, Southampton, UK

⁴University Hospital Southampton NHS Foundation Trust, Southampton, UK

⁵Mid Yorkshire Hospitals NHS Trust, Wakefield, UK

⁶Department of Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Leeds, UK

⁷School of Medicine, University of Leeds, Leeds, UK

⁸Genomics England, Queen Mary University of London, London, UK

⁹Division of Evolution and Genomic Sciences, Faculty of Biology, Medicine and Health, University of Manchester, Manchester, UK

¹⁰Manchester Centre for Genomic Medicine, Manchester, UK

¹¹East Anglian Medical Genetics Service, Addenbrooke’s Hospital, Cambridge, UK

¹²Southampton University Hospitals NHS Trust, Southampton, UK

Twitter Christopher M Watson @ChrisM_Watson, James A Poulter @jamesapoulter and Gabrielle Wheway @gabriellewheway

Acknowledgements This research was made possible through access to the data and findings generated by the 100,000 Genomes Project. The 100,000 Genomes Project is managed by Genomics England Limited (a wholly owned company of the Department of Health and Social Care). The 100,000 Genomes Project is funded by the National Institute for Health Research and National Health Service (NHS) England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100,000 Genomes Project uses data provided by patients and collected by the NHS as part of their care and support.

Collaborators John C Ambrose (Genomics England, London, UK); Prabhu Arumugam (Genomics England, London, UK); Roel Bevers (Genomics England, London, UK); Marta Bleda (Genomics England, London, UK); Freya Boardman-Pretty (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Christopher R Bousted (Genomics England, London, UK); Helen Brittain (Genomics England, London, UK); Mark J Caulfield (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Georgia C Chan (Genomics England, London, UK); Greg Elgar (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Tom Fowler (Genomics England, London, UK); Adam Giess (Genomics England, London, UK); Angela Hamblin (Genomics England, London, UK); Shirley Henderson (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Tim J P Hubbard (Genomics England, London, UK); Rob Jackson (Genomics England, London, UK); Louise J Jones (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Dalia Kasperaviciute (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Melis Kayikci (Genomics England, London, UK); Athanasios Kousathanas (Genomics England, London, UK); Lea Lahnstein (Genomics England, London, UK); Sarah E A Leigh (Genomics England, London, UK); Ivonne

Diagnostics

U S Leong (Genomics England, London, UK); Javier F Lopez (Genomics England, London, UK); Fiona Maleady-Crowe (Genomics England, London, UK); Meriel McEntagart (Genomics England, London, UK); Federico Minnici (Genomics England, London, UK); Loukas Moutsianas (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Michael Mueller (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Nirupa Murugaesu (Genomics England, London, UK); Anna C Need (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Peter O'Donovan (Genomics England, London, UK); Chris A Odhams (Genomics England, London, UK); Christine Patch (Genomics England, London, UK); Mariana Buongiorno Pereira (Genomics England, London, UK); Daniel Perez-Gil (Genomics England, London, UK); John Pullinger (Genomics England, London, UK); Tahrima Rahim (Genomics England, London, UK); Augusto Rendon (Genomics England, London, UK); Tim Rogers (Genomics England, London, UK); Kevin Savage (Genomics England, London, UK); Kushmita Sawant (Genomics England, London, UK); Richard H Scott (Genomics England, London, UK); Afshan Siddiq (Genomics England, London, UK); Alexander Sieghart (Genomics England, London, UK); Samuel C Smith (Genomics England, London, UK); Alona Sosinsky (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Alexander Stuckey (Genomics England, London, UK); Mélanie Tanguy (Genomics England, London, UK); Ana Lisa Taylor Tavares (Genomics England, London, UK); Ellen R A Thomas (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Simon R Thompson (Genomics England, London, UK); Arianna Tucci (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Matthew J Welland (Genomics England, London, UK); Eleanor Williams (Genomics England, London, UK); Katarzyna Witkowska (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK); Suzanne M Wood (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK).

Contributors Conceptualisation: SB, JL, CT, CFI, CAJ, GW; Data curation: SB, JL, MR, RPJB, AS, KS, JAP, JC, HB, Genomics England Research Consortium, G.W; Formal analysis: SB, JL, MR, JAP, CT, CFI, CAJ, GW; Funding acquisition: SB, JL, CT, CFI, CAJ, GW; Investigation: SB, JL, MR, CW, JAP, CT, CFI, CAJ, GW; Methodology: SB, JL, MR, CT, CFI, CAJ, GW; Software: JL, MR, RPJB, AS, JME; Project administration: SB, Genomics England Research Consortium, G.W; Resources: SB, JL, MR, RPJB, AS, KS, JME, JC, HB, Genomics England Research Consortium; Supervision: CT, CFI, CAJ, GW; Validation: JC, HB; Writing – original draft: SB, GW; Writing – review and editing: all authors; Guarantors: CAJ, GW.

Funding SB acknowledges support from the Wellcome Trust 4Ward North Clinical PhD Academy (ref. 203914/Z/16/Z). GW acknowledges support from Wellcome Trust Seed Award (ref. 204378/Z/16/Z). CAJ acknowledges support from MRC project grants MR/M000532/1 and MR/T017503/1. JL is supported by a NIHR Research Professorship awarded to Professor Diana Baralle (DB NIHR RP-2016-07-011). JAP is supported by a UKRI Future Leader Fellowship (MR/T02044X/1).

Competing interests Disclosure: HB, RPJB and AS are employed by Genomics England, UK. GW is employed by Illumina. The other authors declare no conflict of interest.

Patient consent for publication Not applicable.

Ethics approval Written informed consent was obtained from all participants (or from their parent/legal guardian) in the 100,000 Genomes Project (IRAS ID 166046; REC reference 14/EE/1112). Access to the secure online Research Environment within the Genomics England Ltd (GEL) Data Embassy was provided by the GEL Access Review Committee, and research project RR185 'Study of cilia and ciliopathy genes across the 100,000 GP cohort' was registered and approved by GEL. This research study received ethical approval from University of Southampton Faculty of Medicine Ethics Committee (ERGO#54400).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data may be obtained from a third party and are not publicly available. Full data is available in the Secure Genomic England Secure Research Environment.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and

is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Christopher M Watson <http://orcid.org/0000-0003-2371-1844>
James A Pouler <http://orcid.org/0000-0003-2048-5693>
Jamie M Ellingford <http://orcid.org/0000-0003-1137-9768>
Gabrielle Wheway <http://orcid.org/0000-0002-0494-0783>

REFERENCES

- Singla V, Reiter JF. The primary cilium as the cell's antenna: signaling at a sensory organelle. *Science* 2006;313:629–33.
- Oud MM, Lamers JC, Arts HH. Ciliopathies: genetics in pediatric medicine. *J Pediatr Genet* 2017;6:018–29.
- Higgins M, Obaidi I, McMorrow T. Primary cilia and their role in cancer. *Oncol Lett* 2019;17:3041–7.
- Gabriel GC, Young CB, Lo CW. Role of cilia in the pathogenesis of congenital heart disease. *Semin Cell Dev Biol* 2021;110:S1084-9521(19)30166-1:2–10.
- Barsch F, Niedermair T, Mamilos A, Schmitt VH, Grevenstein D, Babel M, Burgoyne T, Shoemark A, Brochhausen C. Physiological and pathophysiological aspects of primary Cilia-A literature review with view on functional and structural relationships in cartilage. *Int J Mol Sci* 2020;21:4959.
- Bujakowska KM, Liu Q, Pierce EA. Photoreceptor cilia and retinal ciliopathies. *Cold Spring Harb Perspect Biol* 2017;9:a028274.
- McConnachie DJ, Stow JL, Mallett AJ. Ciliopathies and the kidney: a review. *Am J Kidney Dis* 2021;77:S0272-6386(20)31013-1:410–9.
- Handa A, Voss U, Hammarsjö A, Grigelioniene G, Nishimura G. Skeletal ciliopathies: a pattern recognition approach. *Jpn J Radiol* 2020;38:193–206.
- Engle SE, Bansal R, Antonellis PJ, Berbari NF. Cilia signaling and obesity. *Semin Cell Dev Biol* 2021;110:S1084-9521(19)30183-1:43–50.
- Hasenpusch-Theil K, Theil T. The multifaceted roles of primary cilia in the development of the cerebral cortex. *Front Cell Dev Biol* 2021;9.
- Wallmeier J, Nielsen KG, Kuehni CE, Lucas JS, Leigh MW, Zariwala MA, Omran H. Motile ciliopathies. *Nat Rev Dis Primers* 2020;6.
- van Dam TJP, Kennedy J, van der Lee R, de Vrieze E, Wunderlich KA, Rix S, Dougherty GW, Lambacher NJ, Li C, Jensen VL, Leroux MR, Hjeij R, Horn N, Texier Y, Wissinger Y, van Reeuwijk J, Wheway G, Knapp B, Scheel JF, Franco B, Mans DA, van Wijk E, Képes F, Slaats GG, Toedt G, Kremer H, Omran H, Szymanska K, Koutroumpas K, Ueffing M, Nguyen T-MT, Letteboer SJF, Oud MM, van Beersum SEC, Schmidts M, Beales PL, Lu Q, Giles RH, Szklarczyk R, Russell RB, Gibson TJ, Johnson CA, Blacque OE, Wolfrum U, Boldt K, Roepman R, Hernandez-Hernandez V, Huynen MA. CiliaCata: an integrated and validated compendium of ciliary genes. *PLoS One* 2019;14:e0216705.
- Boldt K, van Reeuwijk J, Lu Q, Koutroumpas K, Nguyen T-MT, Texier Y, van Beersum SEC, Horn N, Willer JR, Mans DA, Dougherty G, Lamers JC, Coene KLM, Arts HH, Betts MJ, Beyer T, Bolat E, Gloeckner CJ, Haidari K, Hatterschijt L, Iaconis D, Jenkins D, Klose F, Knapp B, Latour B, Letteboer SJF, Marcellis CL, Mitic D, Morleo M, Oud MM, Riemersma M, Rix S, Terhal PA, Toedt G, van Dam TJP, de Vrieze E, Wissinger Y, Wu KM, Apic G, Beales PL, Blacque OE, Gibson TJ, Huynen MA, Katsanis N, Kremer H, Omran H, van Wijk E, Wolfrum U, Képes F, Davis EE, Franco B, Giles RH, Ueffing M, Russell RB, Roepman R, UK10K Rare Diseases Group. An organelle-specific protein landscape identifies novel diseases and molecular mechanisms. *Nat Commun* 2016;7:11491.
- Wheway G, Schmidts M, Mans DA, Szymanska K, Nguyen T-MT, Racher Y, Phelps IG, Toedt G, Kennedy J, Wunderlich KA, Sorusch N, Abdelhamed ZA, Natarajan S, Herdige W, van Reeuwijk J, Horn N, Boldt K, Parry DA, Letteboer SJF, Roosing S, Adams M, Bell SM, Bond J, Higgins J, Morrison EE, Tomlinson DC, Slaats GG, van Dam TJP, Huang L, Kessler K, Giessl A, Logan CV, Boyle EA, Shendure J, Anazi S, Aldahmesh M, Al Hazzaa S, Hegele RA, Ober C, Frosk P, Mhanni AA, Chodirker BN, Chudley AE, Lamont R, Bernier FP, Beaulieu CL, Gordon P, Pon RT, Donahue C, Barkovich AJ, Wolf L, Toomes C, Thiel CT, Boycott KM, McKibbin M, Inglehearn CF, Stewart F, Omran H, Huynen MA, Sergouniotis PI, Alkuraya FS, Parboosingh JS, Innes AM, Willoughby CE, Giles RH, Webster AR, Ueffing M, Blacque O, Gleeson JG, Wolfrum U, Beales PL, Gibson T, Doherty D, Mitchison HM, Roepman R, Johnson CA. An siRNA-based functional genomics screen for the identification of regulators of ciliogenesis and ciliopathy genes. *Nat Cell Biol* 2015;17:1074–87.
- van Dam TJP, Wheway G, Slaats GG, Huynen MA, Giles RH, SYSCILIA Study Group. The SYSCILIA gold standard (SCGSV1) of known ciliary components and its applications within a systems biology Consortium. *Cilia* 2013;2:7.
- Shamseldin HE, Shaheen R, Ewida N, Bubsait DK, Alkuraya H, Almadawi E, Howaidi A, Sabr Y, Abdalla EM, Alfaifi AY, Alghamdi JM, Alsagheir A, Alfares A, Morsy H, Hussein MH, Al-Muhaizea MA, Shagrani M, Al Sabban E, Salih MA, Meriki N, Khan R, Almugbel M, Qari A, Tulba M, Mahnashi M, Alhazmi K, Alsalamah AK, Nowliaty

- SR, Alhashem A, Hashem M, Abdulwahab F, Ibrahim N, Alshidi T, AlObeid E, Alenazi MM, Alzaidan H, Rahbeeni Z, Al-Owain M, Sogaty S, Seidahmed MZ, Alkuraya FS. The morbid genome of ciliopathies: an update. *Genet Med* 2020;22:1051–60.
- 17 Shaheen R, Szymanska K, Basu B, Patel N, Ewida N, Faqeih E, Al Hashem A, Derar N, Alsharif H, Aldahmesh MA, Alazami AM, Hashem M, Ibrahim N, Abdulwahab FM, Sonbul R, Alkuraya H, Alnemer M, Al Tala S, Al-Husain M, Morsy H, Seidahmed MZ, Meriki N, Al-Owain M, AlShahwan S, Tabarki B, Sali H, Faqih T, El-Kalioby M, Ueffing M, Boldt K, Logan CV, Parry DA, Al Tassan N, Monies D, Megarbane A, Abouelhoda M, Halees A, Johnson CA, Alkuraya FS, Ciliopathy Working Group. Characterizing the morbid genome of ciliopathies. *Genome Biol* 2016;17:242.
 - 18 Estrada-Cuzcano A, Roepman R, Cremers FPM, den Hollander AJ, Mans DA. Non-Syndromic retinal ciliopathies: translating gene discovery into therapy. *Hum Mol Genet* 2012;21:R111–24.
 - 19 Hammarsjö A, Pettersson M, Chitayat D, Handa A, Anderlid B-M, Bartocci M, Basel D, Batkovskytė D, Beleza-Meireles A, Conner P, Eisfeldt J, Girisha KM, Chung BH-Y, Horemuzova E, Hyodo H, Kornejeva L, Lagerstedt-Robinson K, Lin AE, Magnusson M, Moosa S, Nayak SS, Nilsson D, Ohashi H, Ohashi-Fukuda N, Stranneheim H, Taylan F, Traberg R, Voss U, Wirta V, Nordgren A, Nishimura G, Lindstrand A, Grigelioniene G. High diagnostic yield in skeletal ciliopathies using massively parallel genome sequencing, structural variant screening and RNA analyses. *J Hum Genet* 2021;66:995–1008.
 - 20 Turnbull C, Scott RH, Thomas E, Jones L, Murugaesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereau A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100,000 Genomes Project. The 100,000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* 2018;361:k1687.
 - 21 Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, Leong IUS, Smith KR, Gerasimenko O, Haraldsdottir E, Thomas E, Scott RH, Baple E, Tucci A, Brittain H, de Burca A, Ibañez K, Kasperaviciute D, Smedley D, Caulfield M, Rendon A, McDonagh EM. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet* 2019;51:1560–5.
 - 22 Wheway G, Mitchison HM, Genomics England Research Consortium. Corrigendum: opportunities and challenges for molecular understanding of Ciliopathies-The 100,000 genomes project. *Front Genet* 2019;10:569.
 - 23 Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178–92.
 - 24 Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li Yi, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Xia A, Batzoglu S, Sanders SJ, Farh KK-H. Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:S0092-8674(18)31629-5:535–48.
 - 25 McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. *Genome Biol* 2016;17:122.
 - 26 Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehml HL, ACMG Laboratory Quality Assurance Committee. 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* 2015;17:405–23.
 - 27 Ellard SB, Berry I, Forrester N, Turnbull C, Owens M, Eccles DM, Abbs S, Scott R, Deans Z, Lester T, Campbell J, Newman W, McMullan D. ACGS best practice guidelines for variant classification in rare disease 2020. Available: <https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>
 - 28 Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alfoldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferreira S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG, Genome Aggregation Database Consortium. Author correction: the mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2021;590:E53.
 - 29 Maria M, Lamers IJC, Schmidts M, Ajmal M, Jaffar S, Ullah E, Mustafa B, Ahmad S, Nazmutdinova K, Hoskins B, van Wijk E, Koster-Kamphuis L, Khan MI, Beales PL, Cremers FPM, Roepman R, Azam M, Arts HH, Qamar R. Genetic and clinical characterization of Pakistani families with Bardet-Biedl syndrome extends the genetic and phenotypic spectrum. *Sci Rep* 2016;6:34764.
 - 30 Poretti A, Boltshauser E, Valente EM. The molar tooth sign is pathognomonic for Joubert syndrome! *Pediatr Neurol* 2014;50:S0887-8994(13)00666-8:e15–16.
 - 31 Turro E, Astle WJ, Megy K, Gräf S, Greene D, Sharmardina O, Allen HL, Sanchez-Juan A, Frontini M, Thys C, Stephens J, Mapeta R, Burren OS, Downes K, Haimel M, Tuna S, Deevi SVV, Aitman TJ, Bennett DL, Calleja P, Cars K, Caulfield MJ, Chinnery PF, Dixon PH, Gale DP, James R, Koziell A, Laffan MA, Levine AP, Maher ER, Markus HS, Morales J, Morrell NW, Mumford AD, Ormondroyd E, Rankin S, Rendon A, Richardson S, Roberts I, Roy NBA, Saleem MA, Smith KGC, Stark H, Tan RYY, Themistocleous AC, Thrasher AJ, Watkins H, Webster AR, Wilkins MR, Williamson C, Whitworth J, Humphray S, Bentley DR, Kingston N, Walker N, Bradley JR, Ashford S, Penkett CJ, Freson K, Stirrups KE, Raymond FL, Ouwehand WH, NHRI BioResource for the 100,000 Genomes Project. Whole-Genome sequencing of patients with rare diseases in a national health system. *Nature* 2020;583:96–102.
 - 32 Zhang X, Wakeling M, Ware J, Whiffin N. Annotating high-impact 5' untranslated region variants with the UTRannotator. *Bioinformatics* 2021;37:1171–3.
 - 33 Kurtulmus B, Yuan C, Schuy J, Neuner A, Hata S, Kalamakis G, Martin-Villalba A, Pereira G. LRR45 contributes to early steps of axoneme extension. *J Cell Sci* 2018;131:jcs223594.
 - 34 Li Z, Yao K, Cao Y. Molecular cloning of a novel tissue-specific gene from human nasopharyngeal epithelium. *Gene* 1999;237:235–40.
 - 35 Dougherty GW, Mizuno K, Nöthe-Menchen T, Ikawa Y, Boldt K, Ta-Shma A, Aprea I, Minegishi K, Pang Y-P, Pennekamp P, Loges NT, Raidt J, Hjejir R, Wallmeier J, Mussaffi H, Perles Z, Elpeleg O, Rabert F, Shiratori H, Letteboer SJ, Horn N, Young S, Strünker T, Stumme F, Werner C, Olbrich H, Takaoka K, Ide T, Twan WK, Biebach L, GroBe-Onnebrink J, Klinckenbusch JA, Praveen K, Bracht DC, Höben IM, Junger K, Gütlzaff J, Cindrić S, Aviram M, Kaiser T, Memari Y, Dzeja PP, Dworniczak B, Ueffing M, Roepman R, Bartscherer K, Katsanis N, Davis EE, Amirav I, Hamada H, Omran H. CFAP45 deficiency causes situs abnormalities and asthenospermia by disrupting an axonemal adenine nucleotide homeostasis module. *Nat Commun* 2020;11:5520.
 - 36 Sawyer SL, Hartley T, Dymont DA, Beaulieu CL, Schwartzentruber J, Smith A, Bedford HM, Bernard G, Bernier FP, Brais B, Bulman DE, Warman Chardon J, Chitayat D, Deladoëy J, Fernandez BA, Frosk P, Geraghty MT, Gerull B, Gibson W, Gow RM, Graham GE, Green JS, Heon E, Horvath G, Innes AM, Jabado N, Kim RH, Koenekoop RK, Khan A, Lehmann OJ, Mendoza-Londono R, Michaud JL, Nikkel SM, Penney LS, Polychronakos C, Richer J, Rouleau GA, Samuels ME, Siu VM, Suchowersky O, Tarnopolsky MA, Yoon G, Zahir FR, Majewski J, Boycott KM. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin Genet* 2016;89:275–84.
 - 37 Lipska-Ziętkiewicz BS. WT1 Disorder. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LH, Stephens K, eds. GeneReviews. Seattle (WA): University of Washington, Seattle, 2020.

3 Uncovering the burden of hidden ciliopathies in the 100,000 Genomes Project: a reverse phenotyping approach

3.1 Research Rationale

This study was designed upon completion of our first 100K project (Molecular diagnoses in the congenital malformations caused by ciliopathies (CMC) cohort of the 100,000 Genomes Project – Thesis Chapter 2; (Best et al., 2022b)). One of our main findings in that study was the high proportion of participants entered to 100K under suspected primary ciliopathy recruitment categories but who proved to have alternative diagnoses caused by variants in non-ciliopathy disease genes (n=19/43, 44.2%). Reflecting this finding, we thought it was reasonable to assume that there are also ‘hidden’ patients with ciliopathies who have been recruited to alternative categories, and we wanted to design a strategy to identify them.

Reverse phenotyping emerged as the most promising approach; presenting an interesting opportunity to analyse 100K variant data without prior prejudice about clinical associations. In reverse phenotyping, the search begins with the identification of potentially pathogenic variants, which are then mapped in a reverse strategy against the clinical features of patients. Patients with potentially causative variants in the selected genes are assessed to see if their clinical features match the associated disease phenotype and inheritance pattern reported in the medical literature (genotype-to-phenotype model). We were especially interested to undertake this project given group discussions about “agnostic” approaches to rare disease diagnostics with access to large genomic datasets.

Given the limitations in the tiering system and the previously observed poor quality phenotyping data for a substantial proportion of cases in 100K, we knew that there were very likely to be pathogenic variants that would have been missed from mainstream diagnostic pipelines simply because the right panels had not been selected for analysis. We hypothesised that some of these would be easy to detect via appropriate filtering strategies once variants in the right genes were extracted (e.g. previously reported as pathogenic on ClinVar and/or high impact variant types).

Beyond these “low-hanging fruit” pathogenic variants, we also wanted to look again for causative variants that would be routinely missed from mainstream diagnostic strategies, as we did in the 100K CMC cohort analysis (thesis chapter 2) (Best et al., 2022b). In particular, we wanted to look for missed SVs and non-coding variants. This was another opportunity to make the most of the available WGS data to boost diagnosis rates for unsolved participants.

Through dialogue with the GEL Bioinformatician Roel Bevers via the Research Environment helpdesk, we were made aware of the soon-to-be released workflow called “Gene-Variant Workflow’ written by himself and Alex Stuckey (now available from <https://research-help.genomicsengland.co.uk/display/GERE/GeneVariant+Workflow>). This could be used to extract all variants in up to ten genes at a time from the 100K dataset, including all intronic and exonic variants within the specified gene region. We therefore set out to use this script to perform a reverse phenotyping study. We decided to focus on multi-systemic ciliopathy genes because we thought that 100K participants with pathogenic variants in those genes would be more likely to be recruited to alternative categories (clinically mis-diagnosed) than those with single-system disorders (e.g. renal or retinal ciliopathies) so our pickup rate would be higher. We decided to set a limit of 10 ciliopathy disease genes for analysis, partly to allow the script to run in one batch, and partly to try to maintain to a workable output volume.

We started by selecting a list of key multisystemic ciliopathy disorders that we suspected may be identifiable in alternative disease categories, then performed a literature search to define a list of genes causative of $\geq 10\%$ of the total syndrome burden. Our syndrome list includes Bardet-Biedl syndrome (BBS) and Alström syndrome (metabolic/obesity ciliopathies); Joubert syndrome (JBTS), Meckel Gruber syndrome (MKS) and orofacioidigital syndrome (OFD) (neurodevelopmental ciliopathies); the skeletal ciliopathy Jeune asphyxiating thoracic dystrophy (JATD) and nephronophthisis (isolated or syndromic renal ciliopathy). The accompanying gene list contains nine genes, pathogenic variants in which are a frequent cause of these conditions: *BBS1*, *BBS10*, *ALMS1*, *OFD1*, *DYNC2H1*, *WDR34*, *NPHP1*, *TMEM67* and *CEP290*. Further detail about selection of these genes is provided in the published supplementary material (thesis section 6.1.2)

In the 100K CMC cohort analysis (thesis chapter 2) (Best et al., 2022b), we undertook SV analysis in pursuit of “second hit” pathogenic variants only in the presence of a “first-hit” SNV with the suspicion of compound heterozygosity. We therefore knew which gene to look at and carried out manual searches of the entire gene locus looking for visible SVs on the Integrative Genomics Viewer (IGV) within the research environment. Not only was this a slow, laborious, and unsystematic strategy, but it could not be used to look for first-hit or homozygous variants in our reverse phenotyping study. I was aware of the available Manta (Chen et al., 2016) and Canvas (Ivakhno et al., 2018) structural variant calls on the 100K dataset, but had no strategy to filter them.

For this reverse phenotyping project, I was put in touch with our collaborator, Dr. Jing Yu, a senior bioinformatician with the Nuffield Department of Clinical Neurosciences at the University of Oxford, by my PhD supervisors. He was in the process of developing the SVRare script, which used a database of 554,060 SVs called by Manta and Canvas aggregated from 71,408 participants in the rare disease arm of 100K (Yu et al., 2022). Dr. Yu and I collaborated to extract rare SVs (≤ 10 SVRare database calls) that overlapped coding regions of our nine selected ciliopathy disease genes, which were then analysed manually. I also worked again with Dr. Jenny Lord, Postdoctoral Research Fellow within the Faculty of Medicine at the University of Southampton, to do SpliceAI analysis on coding and non-coding genomic variants using her publicly available script (`find_variants_by_gene_and_SpliceAI_score.py`; available at https://github.com/JLord86/Extract_variants).

We hoped that this project would not only boost diagnostic rates for previously missed ciliopathy patients but would also provide some useful insight into alternative strategies for genomic analysis. We suspected that it would provoke interesting dialogue about the required links between genotype and clinical data to provide confident diagnoses for patients and the consent procedures that would need to be in place to both look for and report unexpected molecular diagnoses. We thought it could also expand known genotype-phenotype correlations for our ciliopathy disease genes of interest.



Original research

Uncovering the burden of hidden ciliopathies in the 100 000 Genomes Project: a reverse phenotyping approach

Sunayna Best,^{1,2} Jing Yu,³ Jenny Lord,^{4,5} Matthew Roche,⁶ Christopher Mark Watson ,^{1,7} Roel P J Bevers,⁸ Alex Stuckey,⁸ Savita Madhusudhan,⁹ Rosalyn Jewell,² Sanjay M Sisodiya,^{10,11} Siying Lin,^{12,13} Stephen Turner,¹² Hannah Robinson,¹³ Joseph S Leslie,¹⁴ Emma Baple,^{14,15} Genomics England Research Consortium, Carmel Toomes,¹ Chris Inglehearn,¹ Gabrielle Whewey ,^{4,5} Colin A Johnson ¹

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/jmedgenet-2022-108476>).

For numbered affiliations see end of article.

Correspondence to

Prof Colin A Johnson, Division of Molecular Medicine, Leeds Institute of Medical Research, University of Leeds, Leeds LS9 7TF, UK; c.johnson@leeds.ac.uk

Received 26 January 2022
Accepted 7 June 2022

ABSTRACT

Background The 100 000 Genomes Project (100K) recruited National Health Service patients with eligible rare diseases and cancer between 2016 and 2018. PanelApp virtual gene panels were applied to whole genome sequencing data according to Human Phenotyping Ontology (HPO) terms entered by recruiting clinicians to guide focused analysis.

Methods We developed a reverse phenotyping strategy to identify 100K participants with pathogenic variants in nine prioritised disease genes (*BBS1*, *BBS10*, *ALMS1*, *OFD1*, *DYNC2H1*, *WDR34*, *NPHP1*, *TMEM67*, *CEP290*), representative of the full phenotypic spectrum of multisystemic primary ciliopathies. We mapped genotype data 'backwards' onto available clinical data to assess potential matches against phenotypes. Participants with novel molecular diagnoses and key clinical features compatible with the identified disease gene were reported to recruiting clinicians.

Results We identified 62 reportable molecular diagnoses with variants in these nine ciliopathy genes. Forty-four have been reported by 100K, 5 were previously unreported and 13 are new diagnoses. We identified 11 participants with unreportable, novel molecular diagnoses, who lacked key clinical features to justify reporting to recruiting clinicians. Two participants had likely pathogenic structural variants and one a deep intronic predicted splice variant. These variants would not be prioritised for review by standard 100K diagnostic pipelines.

Conclusion Reverse phenotyping improves the rate of successful molecular diagnosis for unsolved 100K participants with primary ciliopathies. Previous analyses likely missed these diagnoses because incomplete HPO term entry led to incorrect gene panel choice, meaning that pathogenic variants were not prioritised. Better phenotyping data are therefore essential for accurate variant interpretation and improved patient benefit.

INTRODUCTION

The 100 000 Genomes Project (100K) is a combined diagnostic and research initiative managed by Genomics England Ltd (GEL). It aimed to sequence

WHAT IS ALREADY KNOWN ON THIS TOPIC

⇒ Whole genome sequencing and targeted gene-panel analysis have improved molecular diagnosis rates for patients with multisystemic ciliopathies.

WHAT THIS STUDY ADDS

⇒ Reverse phenotyping from 100 000 Genomes Project data has identified 62 reportable molecular diagnoses with variants in nine prioritised ciliopathy genes, of which 18 are new diagnoses not reported by Genomics England Ltd.
⇒ Furthermore, we identified 11 unreportable molecular diagnoses in these genes, but these lacked adequate clinical data to justify returning the findings to recruiting clinicians.

HOW THIS STUDY MIGHT AFFECT RESEARCH, PRACTICE AND/OR POLICY

⇒ Reverse phenotyping can improve molecular diagnosis rates from large-scale genomic projects.
⇒ Comprehensive phenotypic data are essential to facilitate accurate variant interpretation.

100 000 genomes from 70 000 participants seen within the UK National Health Service (NHS) with either selected rare diseases or cancers, the latter allowing comparison of matched germline and somatic tumour genomes.^{1 2} To take part in 100K, participants consented to receive a result 'relevant to the explanation, main diagnosis or treatment of the disease for which the patient was selected for testing' (the 'pertinent finding'), if identified.³ Furthermore, they consented to allow access to their fully anonymised genome sequence data and phenotype information for approved academic and commercial researchers. Short-read genome sequencing was performed using Illumina 'TruSeq' library preparation kits for read lengths 100bp and 125bp (Illumina HiSeq 2500 instruments), or 150bp reads (HiSeq X). These generated a mean



© Author(s) (or their employer(s)) 2022. Re-use permitted under CC BY. Published by BMJ.

To cite: Best S, Yu J, Lord J, et al. *J Med Genet* Epub ahead of print: [please include Day Month Year]. doi:10.1136/jmedgenet-2022-108476

Developmental defects

read depth of 32× (range, 27–54) and a depth >15× for at least 95% of the reference human genome.² In the Main Programme Data Release 12 (5 June 2021) used in this study, data were available for 88 844 individuals: 71 597 in the rare diseases arm (33 208 probands and 33 388 relatives) and 17 247 in the cancer arm.

Large-scale genomic studies such as the 100K offer the opportunity to perform reverse phenotyping for genes of interest. In traditional forward genetics, observation of clinical features prompts differential diagnoses and the subsequent evaluation of genes with potentially pathogenic variants (phenotype-to-genotype model). In reverse phenotyping, the search begins with the identification of potentially pathogenic variants, which are then mapped in a reverse strategy against the key clinical features of patients in order to guide phenotyping. Patients with potential causative variants in the selected genes are assessed to see if their clinical features match the associated disease phenotype and inheritance pattern reported in the medical literature (genotype-to-phenotype model).

Reverse phenotyping strategies have been especially successful for diseases characterised by high heterogeneity and complex phenotypes. For example, reverse phenotyping is helping to uncover the genetic architecture of pulmonary arterial hypertension.⁴ Reverse phenotyping allowed diagnosis of 18/64 previously unsolved patients with steroid-resistant nephrotic syndrome through analysis of 298 causative genes after whole exome sequencing (WES). This was followed by multidisciplinary team (MDT) discussion and recommended additional examinations to detect previously overlooked signs or symptoms of the syndromic genetic disorder that was guided by knowledge of the identified pathogenic variants.⁵ Reverse phenotyping also provides an opportunity to extend or refine the phenotype for disease-associated genes, as demonstrated for a family with an *INPPE*-related ciliopathy.⁶

Ciliopathies are a group of rare inherited disorders caused by abnormalities of structure or function of primary cilia (the ‘cell’s antenna’) or motile cilia (organelles responsible for the movement of fluid over the surface of cells).^{8,9} Ciliopathy syndromes present as a clinical spectrum, ranging from relatively common single-system disorders such as retinal or renal ciliopathies, through to rare, complex, multisystem syndromes. There is considerable phenotypic and genetic heterogeneity between the >35 reported ciliopathy syndromes.^{9,10} Common, shared clinical features include renal malformations and/or renal dysfunction, retinal dystrophy, developmental delay, intellectual disability, cerebellar abnormalities, obesity and skeletal abnormalities.¹¹ Collectively, ciliopathies are thought to affect up to 1 in 2000 people based on three common frequent clinical features: renal cysts (1 in 500 adults), retinal degeneration (1 in 3000) and polydactyly (1 in 500).¹² Multisystemic ciliopathies can be grouped into metabolic/obesity ciliopathies, neurodevelopmental ciliopathies and skeletal ciliopathies. The variety in systems involvement reflects the critical role of cilia in development and health.²

We recently published a study determining a research molecular diagnosis for n=43/83 (51.8%) of probands recruited under primary ciliopathy categories by GEL, comprising the ‘Congenital Malformations caused by Ciliopathies’ cohort.¹³ We noted that a high proportion of diagnoses were caused by variants in non-ciliopathy disease genes (n=19/43, 44.2%). We hypothesised that this reflects difficulties in the clinical recognition of ciliopathies, as well as practical challenges in recruiting participants to 100K under appropriate rare disease domains. It is therefore reasonable to assume that there are

also ‘hidden’ patients with ciliopathies recruited to alternative categories.

METHODS

In order to improve the rate of successful molecular diagnosis for unsolved 100K participants with known or suspected ciliopathies, we developed a reverse phenotyping strategy for selected exemplar genes that are most frequently mutated as a cause of primary multisystemic ciliopathies.

Selection of common multisystemic ciliopathy genes to assess

A literature review was undertaken to determine the most common genetic causes of multisystemic primary ciliopathies: Bardet-Biedl syndrome (BBS) and Alström syndrome (metabolic/obesity ciliopathies); Joubert syndrome (JBTS), Meckel-Gruber syndrome (MKS) and orofacioidigital syndrome (OFD) (neurodevelopmental ciliopathies); the skeletal ciliopathy Jeune asphyxiating thoracic dystrophy (JATD) and nephronophthisis (isolated or syndromic renal ciliopathy).² Disease genes causative of ≥10% of the total syndrome burden were selected for inclusion in the reverse phenotyping analysis and are summarised alongside referenced literature (online supplemental table 1). Where disease genes are known to cause multiple ciliopathy syndromes, all associated conditions are included in the table. On this basis, nine disease genes were selected as exemplars that span the extensive phenotypic range of primary multisystemic ciliopathies: *BBS1*, *BBS10*, *ALMS1*, *OFD1*, *DYNC2H1*, *WDR34*, *NPHP1*, *TMEM67* and *CEP290*. All have autosomal recessive inheritance except *OFD1* which is associated with X linked dominant OFD type 1 (OFD-1) and X linked recessive JBTS.¹³ Almost all individuals with OFD-1 are female; the few affected males are reported to be malformed fetuses delivered by an affected female.

Identification of solved participants with causative variants in representative ciliopathy disease genes

All analysis on the GEL datasets were performed within a secure workspace called the ‘Research Environment’. Clinical and participant data were integrated and analysed using ‘LabKey’ data management software. Previously reported diagnoses were identified using data in the NHS Genomics Medical Centres (GMC) ‘Exit Questionnaire’. The Exit Questionnaire is completed by the clinicians at the GMC for each closed case, and summarises the extent to which a participant’s diagnosis can be explained by the combined variants reported to the GMC from GEL and clinical interpretation providers. Data in Exit Questionnaires were filtered for reports containing variants in the nine ciliopathy disease genes, where the ‘case solved family’ was annotated as ‘yes’ (solved) or ‘partially’ (partially solved).

Selection of key clinical terms associated with selected ciliopathy genes

A literature search of review articles prioritised the key clinical terms for each of the nine selected ciliopathy genes. This assessed the potential match against phenotype and justification for reporting new molecular findings. Approved researchers submit a ‘Researcher Identified Diagnosis’ (RID) form using the secure GEL ‘Airlock’ system. This is then sent to the participant’s recruiting clinician for consideration of the fit to phenotype and the interpretation of variant pathogenicity, followed by decisions about whether the finding should be reported back to the participant. Usually, such cases are discussed at multiMDT meetings involving clinical scientists, researchers and clinicians. Variants

classified as likely pathogenic or pathogenic and felt to be a good clinical match for phenotype, must be molecularly confirmed and formally reported by an NHS-accredited diagnostic laboratory before being fed back to the participant by the clinician responsible for their care.³ Decisions about feedback of variants of uncertain clinical significance (VUS) to participants are the

responsibility of individual clinicians following MDT discussion, but are usually not fed back.

The rationale for selection of key features is presented in table 1, supported by key references from the literature. To allow easier categorisation and to protect participant anonymity, they are grouped into 11 body systems. Without specific participant

Table 1 Key clinical features for ciliopathy syndromes associated with the nine selected ciliopathy genes of interest

	Ciliopathy syndrome	BBS	ALMS	JATD	OFD-1	Nephronophthisis	JBTS	MKS	LCA/EOSRD
	Reference(s)	40	30	41	13 39	42	43	44	45
System	Chosen ciliopathy gene(s) associated with syndrome	<i>BBS1, BBS10, TMEM67, CEP290</i>	<i>ALMS1</i>	<i>DYNC2H1, WDR34</i>	<i>OFD1</i>	<i>NPHP1</i> (isolated+syndromic), <i>TMEM67+CEP290</i> (syndromic)	<i>TMEM67, CEP290, NPHP1, OFD1</i>	<i>TMEM67, CEP290</i>	<i>CEP290</i>
Ophthalmic	Retinal dystrophy	M	M	M		m*†‡	m*†‡		M
	Abnormality of eye movement					m*†‡	M		M
	Lens opacities								M
	Keratoconus								M
Gastrointestinal	Abnormality of the liver		m	M	m	m*†‡	m*†‡	M	
	Abnormality of the gut	m		m					
Renal	Abnormal renal morphology/dysfunction	M	M	M	M	M	m*†	M	
Genitourinary	Abnormality of the genitourinary system	M	m					m	
Cardiovascular	Cardiomyopathy		M						
	Laterality defect	m				m*†	m*†	m	
	Congenital heart disease	m		m				m	
	Hypertension		m						
Sensory	SNHL	m	M						
	Glue ear		m						
	Chronic otitis media		m		m				
	Abnormality of the sense of smell	M							
Endocrine/Metabolic	Hypogonadotrophic hypogonadism	M	M						
	Glucose intolerance		M						
	Obesity	M	M						
	Hypertriglyceridemia		M						
	Thyroid abnormality	m	m					m	
	Polycystic ovarian syndrome	m			m				
Neurological	Intellectual disability	M	m		M	m*†	M		
	Neurodevelopmental delay	M	m				M		
	Hypotonia		m				M		
	Ataxia		m				M		
	Abnormality of brain morphology			m	M	m*†	M	M	
	Seizures		m						
	Unusual sleep patterns		m						
Skeletal	Polydactyly	M		m	M		m	M	
	Short stature			M					
	Narrow chest			M					
	Brachydactyly			M	M				
	Micromelia			M	M			m	
	Leg cramps		M						
Facial/Oral	Dental abnormalities	M							
	Abnormal oral morphology	M			M		m	m	
	Dysmorphic facial features				M				
Respiratory	Abnormal pattern of respiration						M		
	Chronic airway infection		m						
	Asthma		m						
	Pulmonary hypoplasia							m	
	Cystic lung							m	

Key features are grouped into 11 body systems. Clinical features marked 'M' are major features (present in >50% and/or listed as major diagnostic or characteristic feature in the literature cited). Features marked with 'm' are minor features (present in <50% and/or listed as a minor diagnostic feature in the literature cited).
*Feature of *NPHP1*-associated JBTS-plus syndrome (Senior-Loken syndrome).
†Feature of *CEP290*-associated JBTS-plus syndrome (Senior-Loken syndrome, Joubert syndrome with retinal disease, Joubert syndrome with renal disease, COACH syndrome).
‡Feature of *TMEM67*-associated JBTS-plus syndrome (COACH syndrome).
ALMS, Alström syndrome; BBS, Bardet-Biedl syndrome; COACH syndrome, Cerebellar vermis hypoplasia, Oligophrenia, Ataxia, Coloboma and Hepatic fibrosis; EOSRD, early-onset severe retinal dystrophy; JATD, Jeune asphyxiating thoracic dystrophy; JBTS, Joubert syndrome; LCA, Leber congenital amaurosis; m, minor clinical feature; M, major clinical feature; MKS, Meckel-Gruber syndrome; OFD-1, orofaciodigital syndrome 1; SNHL, sensorineural hearing loss.

Developmental defects

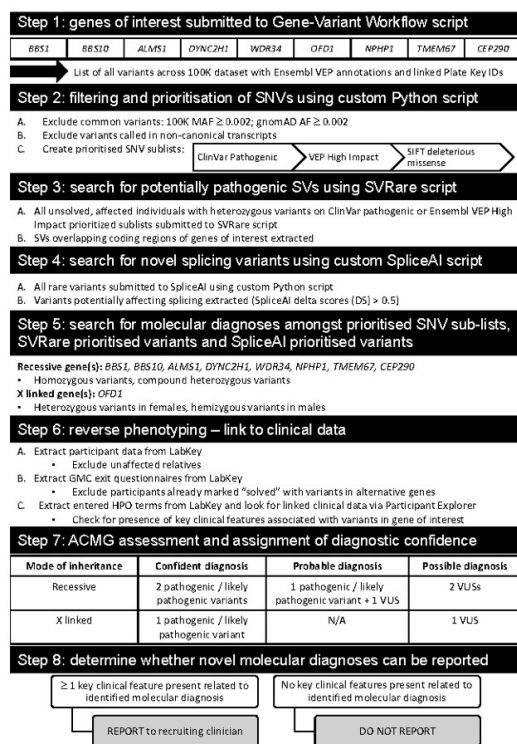


Figure 1 Reverse phenotyping diagnostic research workflow. ACMG, American College of Medical Genetics and Genomics; AF, allele frequency; GMC, Genomics Medical Centres; HPO, Human Phenotyping Ontology; MAF, major allele frequency; N/A, not available; SNV, single nucleotide variant; SV, structural variant; VEP, Variant Effect Predictor; VUS, variant of uncertain significance.

consent for research studies, we are unable to present clinical features that would potentially identify individuals to within five participants in 100K.³ Major features (M) are those present in > 50% of affected individuals and/or listed as major diagnostic or characteristic features in the cited literature. Minor features (m) are those present in < 50% of affected individuals and/or listed as minor diagnostic features. The EMBL-EBI Ontology Lookup Service was used to supplement linked Human Phenotyping Ontology (HPO) terms for each key clinical term, to facilitate capture of a wider selection of appropriate HPO terms that were entered by recruiting clinicians (available from <https://www.ebi.ac.uk/ols/index>). The list of acceptable linked HPO terms is available in online supplemental table 2.

Development of a research diagnostic workflow to identify new diagnoses

The full diagnostic workflow developed, from extraction through to reporting of variants, is represented in figure 1.

Steps 1 and 2: single nucleotide variant filtering and prioritisation

The script ‘Gene-Variant Workflow’ (available from <https://research-help.genomicsengland.co.uk/display/GERE/Gene-Variant+Workflow>) was used to extract all variants in the nine genes in the 100K dataset from Illumina variant call format (VCF) files, aggregate them together and annotate them using

the Ensembl Variant Effect Predictor (VEP).¹⁴ This includes all intronic and exonic variants within the specified gene region. A custom Python script called `filter_gene_variant_workflow.py` (available from https://github.com/sunaynabest/filter_100K_gene_variant_workflow) was used to exclude common variants using the following criteria: 100K major allele frequency (MAF) \geq 0.002; gnomAD allele frequency (AF) \geq 0.002¹⁵ and variants called in non-canonical transcripts. The allele frequency threshold of 0.002 was calculated using the ImperialCardioGenetics frequency filter calculator (available from <https://cardiodb.org/allelefrequencyapp/>),¹⁶ as recommended by the Association for Clinical Genomic Science Best Practice Guidelines.¹⁷ Parameters were set as follows: biallelic inheritance, prevalence 1 in 500, allelic heterogeneity 0.1, genetic heterogeneity 0.2, penetrance 1, confidence 0.95, reference population size 121 412 (based on the Exome Aggregation Consortium cohort).

Finally, prioritised sublists of SNVs were extracted using `filter_gene_variant_workflow.py` as follows: (i) ClinVar pathogenic (variants annotated by ClinVar as ‘pathogenic’ or ‘likely pathogenic’)¹⁸; (ii) high impact (variants annotated by VEP as ‘high impact’ (stop_gained, stop_lost, start_lost, splice_acceptor_variant, splice_donor_variant, frameshift_variant, transcript_ablation, transcript_amplification)¹⁴; (iii) SIFT deleterious missenses (missense variants predicted ‘deleterious’ by the in silico prediction tool SIFT).¹⁹ Additional in silico missense variant predictions were obtained via the Ensembl VEP web interface (available from <https://www.ensembl.org/Tools/VEP>) from Combined Annotation Dependent Depletion²⁰ and PolyPhen-2.²¹

Step 3: SVRare script to prioritise potentially pathogenic structural variants

Heterozygous variants in the nine selected genes in either the ‘ClinVar pathogenic’ or ‘high impact’ SNV sublists were then analysed by the SVRare script.²² This uses a database of 554 060 structural variants (SVs) called by Manta²³ and Canvas²⁴ aggregated from 71 408 participants in the rare disease arm of 100K. Common SVs (\geq 10 database calls) were excluded, and the remaining rare SVs that overlapped coding regions of the selected genes were extracted and analysed manually. BAM files for prioritised SVs were inspected in the Integrative Genomics Browser (IGV).²⁵ SVs were considered potentially causative if present in > 30% of reads. Participants with heterozygous variants identified as ‘deleterious missense’ by SIFT were excluded from further manual analysis by SVRare because of the very high number of such variants and likelihood that they would be classified as VUS. Online supplemental table 4 summarises the numbers of SIFT deleterious missense variant calls in each gene, for example, there are 810 calls in *ALMS1* alone.

Step 4: SpliceAI script to prioritise potentially pathogenic splice defects

All rare variants called by the Gene-Variant Workflow script in the nine representative ciliopathy disease genes (100K MAF \leq 0.002; gnomAD AF \leq 0.002) were run through SpliceAI prediction software with an additional custom Python script (`find_variants_by_gene_and_SpliceAI_score.py`; available at https://github.com/JLord86/Extract_variants). Variants predicted to affect splicing according to the recommended cut-off (SpliceAI delta scores > 0.5) were extracted and analysed manually.²⁶ Variants previously annotated by ClinVar as ‘benign’ were excluded.

Step 5: search for molecular diagnoses among prioritised variants

All prioritised variant lists were manually analysed for each gene: these comprised ClinVar pathogenic, high impact and SIFT deleterious missense SNV, SVRare and SpliceAI prioritised variant lists. For recessive genes (all except *OFD1*), homozygous or compound heterozygous variants were pursued. Heterozygous variants called in female participants and hemizygous variants called in male participants were pursued for X linked *OFD1*.

Step 6: link to clinical data and reverse phenotyping

The Gene-Variant Workflow output files contain 'plate key' identifiers (IDs; unique identifiers used by GEL for DNA sample tracking and logistics) for all participants in whom each variant was called. These unique IDs for participant samples were used to obtain participant data via LabKey, including GMC exit questionnaires reporting outcomes and participant status. Participants were excluded if recruited as unaffected relatives or 'solved' or 'partially solved' with variants in alternative genes. For remaining participants (all unsolved probands or affected relatives), parental data were analysed where available, to determine variant segregation. HPO terms entered at the time of recruitment were also extracted. Further linked clinical data were obtained using the GEL user interface 'Participant Explorer'. This links to the source data in LabKey to identify participants with particular clinical phenotypes, determine longitudinal phenotypic and clinical data for any participant and allow comparison between multiple participants. From these, the number of key clinical features related to the identified ciliopathy gene was recorded for each participant, as well as the bodily system(s) involved.

Step 7: decision on reporting of novel molecular diagnoses

We reasoned that the presence of at least one major key clinical feature that was compatible with the implicated gene would be sufficient to report any newly identified potential molecular diagnoses to recruiting clinicians. If no major key clinical features were present, we were unable to justify reporting because they could not be considered a potential match for patients' clinical features, the so-called 'pertinent findings'.

Step 8: ACMG classification and assignment of diagnostic confidence categories for reportable diagnoses

Variant interpretation was reviewed using the American College of Medical Genetics and Genomics (ACMG)/Association for Molecular Pathology guidelines²⁷ and each variant of interest among participants with reportable diagnoses was assigned an ACMG pathogenicity score.¹⁷ Phenotype specificity is a key factor in variant interpretation, so only those deemed potentially pertinent findings, in the presence of at least one major key feature and therefore reportable, underwent variant interpretation and diagnostic confidence scoring. Diagnostic confidence categories were assigned as 'confident', 'probable' or 'possible' based on the assigned ACMG variant classifications (figure 1). A 'confident' diagnosis required two pathogenic or likely pathogenic variants in genes with recessive inheritance, or one pathogenic or likely pathogenic variant in *OFD1*. A 'probable' diagnosis required one pathogenic/likely pathogenic and one VUS in genes with recessive inheritance; no 'probable' classification was possible for *OFD1* variants. A 'possible' diagnosis was assigned in the presence of two VUS in recessive genes or one VUS in *OFD1*.

We exported anonymised data for publication through the Airlock system, after review by the GEL Airlock Review Committee. We present only information about the body systems with key features for each participant rather than specific HPO terms, in order to protect participant anonymity.

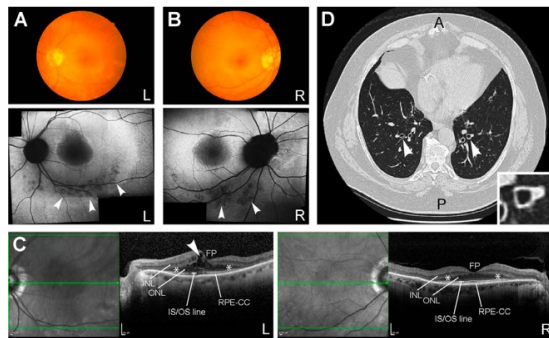


Figure 2 Clinical features of participant #32 consistent with a syndromic ciliopathy. (A) (left eye) and (B) (right eye): upper panels, colour funduscopy of retina; lower panels, fundus autofluorescence images showing perimacular pigment changes (arrowheads) and relatively hypofluorescent central macula. (C) Optical coherence tomography (OCT) for left eye (L; left panel) and right eye (R; right panel), with the plane of OCT shown by green arrows in left-hand regions of each panel, showing loss of ellipsoid zone outside of the central macula with disruption of the outer nuclear layer (*) indicative of rod-cone photoreceptor dystrophy. Arrowhead indicates cystoid macular oedema for the left retina. Scale bars=200 μ m. (D) CT axial section of chest showing 'signet ring' signs (arrowheads; detail shown in inset) typical of bronchiectasis.⁴⁶ A, anterior; FP, foveal pit; INL, inner nuclear layer; IS/OS, inner segment/outer segment; L, left; ONL, outer nuclear layer; P, posterior; R, right; RPE-CC, retinal pigment epithelium-choriocapillaris complex.

RESULTS**100K participants previously solved with causative variants in representative ciliopathy disease genes**

Forty-four participants have previously been reported to have 'solved' or 'partially solved' molecular diagnoses in GMC exit questionnaires with variants in the nine representative ciliopathy disease genes (online supplemental table 3). Seven of these reported cases overlap with participants described in 'Congenital Malformations caused by Ciliopathies' cohort analyses.²⁸ Interestingly, male participant #32 was reported 'solved' with a pathogenic hemizygous *OFD1* frameshift variant in exon 20/23 (NM_003611.3:c.2680_2681del, NP_003602.1:p.(Glu894ArgfsTer6)). Participant #32 was recruited to the 'rod-cone dystrophy' category with an apparently milder non-syndromic form of retinal dystrophy that was only identified in late adulthood (online supplemental table 3). Further clinical information from the recruiting clinicians revealed that the participant had a rod-cone dystrophy that lacked bone spicules typical for retinitis pigmentosa but was similar to Bardet-Biedl syndrome (figure 2A, B,C). Participant #32 also had intellectual disability, truncal obesity, evidence of renal failure, short fingers and chronic respiratory disease with mild bronchiectasis ('signet ring' signs on CT scan of the chest; figure 2DD). These are clinical features consistent with a syndromic ciliopathy, and we are not aware of any previous reports of males with hemizygous *OFD1* variants having this combination of features.

Molecular details for two reported variants are incomplete, described as a heterozygous 'large delins' in *ALMS1* (participant

Developmental defects

#6) and a ‘whole gene deletion’ of *NPHP1* (participant #33). Data are also incomplete for participant #43, reported solved with a single heterozygous variant, classified as a VUS, in the recessive disease gene *CEP290*.

New reportable diagnoses identified through the reverse phenotyping research diagnostic workflow

We prioritised a total number of 3666 variants from the SNV, SV and SpliceAI outputs (online supplemental table 4) through our research diagnostic workflow; 30 variants led to potential reportable diagnoses in 18 previously unsolved participants through reverse phenotyping (table 2). However, on further investigation, n=5/18 participants (#45, #47, #48, #50 and #51) had causative variants that were already included in their GMC Exit Questionnaires, but had reporting outcomes annotated as ‘unknown’ or without listing the ciliopathy disease genes of interest. Although these outcomes may be due to inadvertent coding errors, we did not include the data from these participants for further analysis. Our workflow therefore identified a total of n=13/18 participants with new reportable diagnoses.

Identification of reportable SVs

Two participants have been identified with new potentially causative SVs through the SVRare script (figure 3). Participant #45 had a maternally inherited, 116969bp chr2 inversion and a 63550bp gain (identified using Manta and Canvas, respectively), both including coding regions of *ALMS1*. After a careful inspection of the IGV plot, we also observed a monoallelic, complex SV in the *ALMS1* gene spanning from chr2: g.73424245 to chr2: g.73544334 (GRCh38). We interpreted this as a paired-duplication inversion (figure 3A–B). Ideally, this would be confirmed experimentally; we have contacted the recruiting clinician about performing these studies but no response has been received. Participant #45 also has a paternally inherited, known pathogenic *ALMS1* frameshift variant (NM_015120.4:c.10775del, NP_055935.4:p.Thr3592LysfsTer6). Therefore, segregation analysis is consistent with autosomal recessive inheritance as expected. Participant #45 was recruited to the cone dysfunction category and has one *ALMS1* key feature involving the ophthalmic system that allowed this research finding to be reported to the recruiting clinician.

Participant #70 had a maternally inherited, 56371 bp chromosome 11 deletion (identified by Canvas), including the terminal four exons of *DYNC2H1* (figure 3C). This individual also has a ClinVar ‘likely pathogenic’ paternally inherited *DYNC2H1* synonymous variant (NM_001377.3: c.11049G>A, NP_001368.2: p.Pro3683=). This variant is predicted to cause a splice acceptor loss by SpliceAI (DS_AL 0.51). No clinical detail is provided with the ClinVar entry (from the Rare Disease Group, Karolinska Institutet), but the ‘likely pathogenic’ listing in association with Jeune syndrome provides some confidence in this assessment of pathogenicity. Participant #70, recruited to the proteinuric renal disease category, has two Jeune syndrome key features from the renal and skeletal systems, allowing this research finding to be reported to the recruiting clinician. Furthermore, the participant’s affected sibling, also recruited to 100K with three Jeune syndrome clinical key features from the renal and skeletal systems, was found to have the same two variants, strengthening the confidence in the diagnosis.

Identification of reportable non-canonical splice defects

One new homozygous *CEP290* intronic variant has been identified by using our SpliceAI script, predicted to cause a splice acceptor

gain (SpliceAI DS_AG 0.64) (NM_025114.4:c.6011+874G>T) and gain of a potential splice acceptor site (Alamut screenshot; figure 3D). This variant was identified in participant #49, recruited to the cystic kidney disease category. The proband’s father is heterozygous for the variant, but there is no maternal sample available in 100K. The recruiting clinician has been contacted and relevant tissues (blood, urinary renal epithelial cells) requested to perform functional splicing assays, but no response has been received. Therefore, the variant has been called a VUS, allowing classification of only a ‘possible’ diagnosis to be made.

Novel unreportable diagnoses identified through research workflow

Eleven participants have unreportable, novel diagnoses in the nine ciliopathy disease genes (table 3). These participants have no major key clinical features among their entered HPO terms, or identifiable among the additional clinical data available on Participant Explorer, that can justify reporting to recruiting clinicians as potentially pertinent clinical findings. Four of these 11 have novel missense variants, which can only be classified as VUS. The other seven (#60, #61, #64, #65, #71, #72, #73) have at least one more definitively damaging variant, including high impact frameshifts, stop gains, splice acceptors and ClinVar pathogenic missenses.

DISCUSSION

Reportable diagnoses

We have used a reverse phenotyping strategy to identify 62 reportable molecular diagnoses with variants in 9 prioritised, multisystemic ciliopathy genes (*BBS1*, *BBS10*, *ALMS1*, *OFD1*, *DYNC2H1*, *WDR34*, *NPHP1*, *TMEM67*, *CEP290*). The nine genes chosen were representative exemplars that, from the literature review, span the extensive phenotypic range of ciliopathies. The addition of other ciliopathy genes (such as *CPLANE1* for JBTS) would, of course, further increase diagnostic yield. Forty-four have been previously reported by 100K in GMC Exit Questionnaires, 5 were previously unreported and 13 represent new diagnoses that are compatible with the entered clinical features for unsolved participants (table 2). Based on ACMG classifications of underlying variants, 6 are classified as confident diagnoses, 2 as probable diagnoses and 10 as only possible diagnoses. In summary, 14 molecular diagnoses are in *ALMS1*, 13 in *BBS1*, 2 in *BBS10*, 16 in *CEP290*, 3 in *DYNC2H1*, 7 in *OFD1*, 4 in *NPHP1* and 3 in *TMEM67*. No molecular diagnoses have been made in *WDR34*. These ciliopathy findings fit with what has previously been reported for reverse phenotyping studies; namely, that this approach proves particularly useful in conditions with high genetic heterogeneity and/or complex phenotypes.^{4–6}

We have reported VUS results to recruiting clinicians in this project by using RID forms submitted through the secure GEL Airlock. The ACMG advises that VUS results cannot be used in clinical decision-making.²⁷ This applies to the index patient, and to cascade testing of other family members and to prenatal testing. If reported to patients, VUS can cause significant anxiety and make decision-making challenging.^{22,29} We do not anticipate that VUS results identified through this study will be immediately reported back to patients by recruiting clinicians, but there is a high probability that at least some are the correct molecular diagnosis. Therefore, we believe it is important to report them from the research setting for current and future consideration, especially with the emergence of improved functional

Table 2 Reportable new diagnoses identified via reverse phenotyping research diagnostic workflow

Research ID	Dx confidence	Reported sex	Recruitment category	Gene(s)	Variant zygosity	Consequence	HGVS	HEVSy	gromAD AF	100K MAF	SIFT	Polyphen	CAED	PubMed	GINVar listing	Segregation	ACMG Classification	# of key features	System(s) involved	
45	Conf	Ma	Cone dysfunction syndrome	<i>ALMS1</i>	Het	FS	NM_015120.4:c.1075del	NP_055935.4:p.Trp3592>ys1Ter6	5.23E-05	4.77E-04				11941369, 11941370, 17594715	Path	Pat	Path	1M	M. Oph	
47	Poss	Fe	Bardele-Bianchi syndrome	<i>ABSF1</i>	Hom	Ms	NM_024668.4:c.1790G>A	NP_078961.3:p.Gly574>asp 0	2.54E-05		Delet	Prob_Jam	35		Abs	Bi-par	VUS	4M	M. Ren, oph, slet, endomet	
48	Conf	Ma	Ret dysfunction syndrome	<i>MPHF1</i>	Hom	Ms	NM_001128178.3:c.1027G>A	NP_000263.2:p.Gly434>arg	0.0001155	0.00034312	Delet	Prob_Jam	35	10839884	Path	1 par (other unkl)	Path	1M, 1m	M. Ren m: Oph	
49	Poss	Fe	Cystic kidney disease	<i>CEP290</i>	Hom	Intr	NM_025114.4:c.6011+874G>T	-	0	3.81E-05					Abs	1 par (other unkl)	VUS	1M, 1m	M. Ren m: CVS	
50	Poss	Fe	Syndromic cleft lip and/or cleft palate	<i>ORF1</i>	Het	Ms	NM_003611.3:c.65G>C	NP_003602.1:p.Arg212>pro 0	1.27E-05		Delet	Pos_Jam	21.2		Abs	De novo	VUS	3M	M. Faciona (n=2), slet	
51	Prob	Ma	Joubert syndrome	<i>CEP290</i>	Het	Ms	NM_025114.4:c.1041T>S	NP_078960.3:p.Val55>gly 0	1.00E-04	2.50E-04	Delet	Prob_Jam	33		Abs	Abs	Mat	VUS	4M	M. Oph, neu (n=3)
53	Poss	Ma	Cystic kidney disease	<i>ALMS1</i>	Het	Ms	NM_015120.4:c.8759A>G	NP_055935.4:p.Gln2912>arg 0	5.00E-05		Delet	Pos_Jam	19.03		Abs	Abs	Unkl	VUS	2M, 2m	M. Ren, endo met m: GI, CVS
54	Poss	Fe	Ret cone dystrophy	<i>ALMS1</i>	Het	Ms	NM_015120.4:c.10831A>G	NP_055935.4:p.Arg2471>gly 0	5.00E-05		Delet	Prob_Jam	25.9		Abs	Abs	Unkl	VUS		
55	Conf	Fe	Single autosomal recessive mutation in rate disease	<i>ALMS1</i>	Het	FS	NM_015120.4:c.10377C>G	NP_055935.4:p.Arg3611>gly 3.22E-05	5.00E-05		Delet	Pos_Jam	23.5		Abs	Abs	Unkl	VUS	2M, 1m	M. Oph, ren m: Resp
56	Prob	Ma	Intellectual disability	<i>ALMS1</i>	Het	FS	NM_015120.4:c.11794del	NP_055935.4:p.Glu3392>lys1Ter18	3.99E-06	1.27E-05					26104872, 3281362, 17594715, 24462884	Path	Unkl	Path	4M	M. Oph, endo met (n=2), CVS
57	Poss	Ma	Compromised hearing impairment	<i>ALMS1</i>	Het	Ms	NM_015120.4:c.7510G>T	NP_055935.4:p.Ala2504>Ser	8.89E-05	0.00019062	Delet	Prob_Jam	25		Abs	VUS	Unkl	VUS	1M	M. Senc
58	Poss	Fe	Syndromic congenital heart disease	<i>ABSF1</i>	Het	Ms	NM_024668.5:c.730C>T	NP_078961.3:p.Pro245>Leu 7.16E-05	6.35E-05		Delet	Ben	23.4		Abs	VUS	Unkl	VUS	1M, 1m	M. Neu m: CVS
62	Conf	Fe	Epilepsy plus other features	<i>CEP290</i>	Het	FS	NM_025114.4:c.5484L>545del	NP_078962.3:p.Trp438>arg 7.96E-05	4.45E-05	4.45E-05	Delet	Prob_Jam	25.3		Abs	VUS	Mat	VUS	2M	M. Oph, Neu
63	Conf	Fe	Cystic kidney disease	<i>CEP290</i>	Het	SL	NM_025114.4:c.21>A	NP_078960.3:p.Met17	4.07E-05	2.54E-05					20201475, 17964524, 20201500, 16909894, 17564867, 17345804	Path	Pat	Path	2M	M. Ren, oph
					Het	SG	NM_025114.4:c.4866G>T	NP_078960.3:p.Glu656>Ter	3.80E-05	1.95E-04					23559409, 25525159, 16909894, 20079881	Path	Mat	Path		

Continued

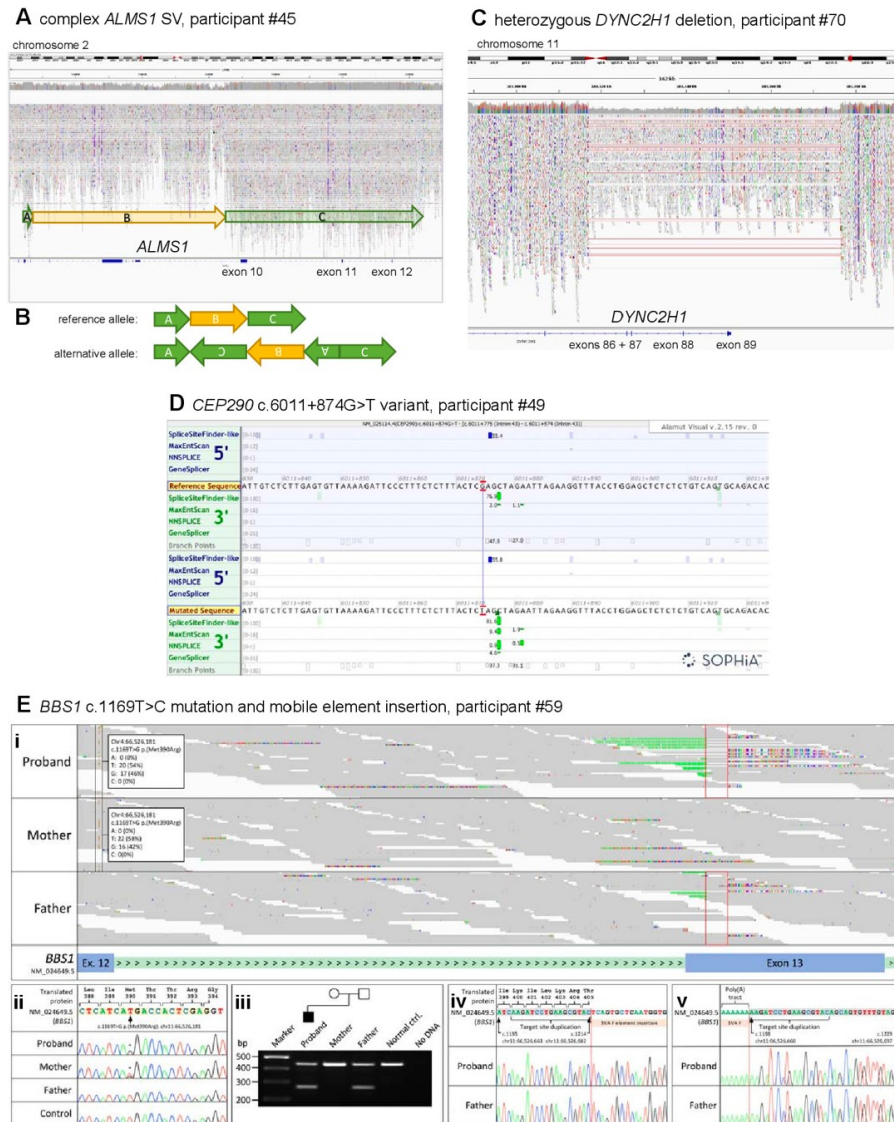


Figure 3 Likely pathogenic structural variants and other variants in selected ciliopathy genes identified through the reverse phenotyping research diagnostic workflow. (A) IGV plot of *ALMS1* (NM_015120.4) in participant #45. We observed a monoallelic complex SV in the *ALMS1* gene spanning from chr2:g.73424245 to chr2:g.73544334 (GRCh38). (B) Diagrammatic representation of complex *ALMS1* SV in participant #45. After inspection of the IGV plots, we surmised that the alternative allele is a paired-duplication inversion, with block A at chr2:g.73424245_73427355, covering exons 4 and 5 (NM_015120.4), block B at chr2:g.73427355_73484777, covering exons 6–9 and block C at chr2:g.73484777_73544334. Note that the boundary between block B and C is an estimate as it is within a region with relatively low alignment quality. (C) IGV plot of heterozygous 56 kb deletion identified in *DYNC2H1* (NM_001377.3) in participant #70. The terminal four exons (86–89) have been deleted. (D) Alamut screenshot for *CEP290* c.6011+874G>T variant in participant #49. Top tracks are donor/acceptor splice site predictions for the reference sequence and the bottom tracks are donor/acceptor predictions for the mutated sequence. Green highlighting identifies increasing scores for a potential splice acceptor site in the non-reference mutated sequence track. (E) Analysis of the *BBS1* locus for Congenital Malformations caused by Ciliopathies (CMC) cohort participant #59 following trio whole genome sequencing. (i) The maternally inherited pathogenic variant, NM_024649.5:c.1169T>G, NP_078925.3:p.(Met390Arg) (highlighted by the black frames) is in trans with a paternally inherited mobile element insertion for which the target site duplication sequence is highlighted (red frames). Soft-clipped junction spanning reads, showing inserted nucleotides and the terminal poly(A) tract, are visible. (ii) Sanger sequencing confirmation of the maternally inherited c.1169T>C mutation. Exon 12 coding sequence is highlighted in peach. (iii) Duplex screening assay³² confirming that the mobile element insertion was present in the proband and his father (270 bp band). Upstream (iv) and downstream (v) junction fragments confirm that the target site duplication sequence is as previously reported.³² Exon 13 coding sequence is highlighted in grey. Genomic coordinates are according to Human Genome build Hg38. Variant nomenclature is according to transcript NM_024649.5. IGV, Integrative Genomics Browser; SV, structural variant.

Table 3 Novel, unreportable diagnoses identified via reverse phenotyping research diagnostic workflow

Research ID	Recruitment category	Gene	Variant zygosity	Consequence	HGVSc	HGVSp	gnomADAF	100K:MAF	SIFT	PolyPhen	CADD	PubMed	ClinVar listing	Segregation	# of key features
52	Intellectual disability	ALMS1	Het	Ms	NM_015170.4:c.738A>T	NP_059395.4:p.Ile2580Phe	4.01E-06	0.00009697	Delet	Ben	22.5	Abx	Abx	Mat	OM, 2m
59	Hereditary spastic paraplegia	HBS1	Het	Ms	NM_015170.4:c.346C>T	NP_059395.4:p.His1161Yr	2.41E-05	0.00019594	Delet	Ben	16.05	Abx	Abx	Pat	OM, 0m
			Het	Ms	NM_024649.5:c.235G>A	NP_078925.3:p.Glu79Asp	0.000756	0.00120728	Delet	Poss_dam	23.8	Abx	Abx	Unk	OM, 0m
60	Primary immunodeficiency	CEP290	Het	Fs	NM_034648.5:c.1714G>T	NP_078925.3:p.Gly572Cys	-	1.27E-05	Delet	Prob_dam	32	Abx	Abx	Unk	OM, 0m
			Het	Fs	NM_025114.4:c.6154_G161del	NP_079390.3:p.Asp2022Leu616Ter17	0	1.27E-05	Delet	Prob_dam	32	Abx	Abx	Unk	OM, 0m
61	Primary lymphoedema	CEP290	Het	Sg	NM_025114.4:c.7048C>T	NP_079390.3:p.Gln2301Irr	1.30E-06	1.91E-05	Delet	Prob_dam	32	Abx	Abx	Unk	OM, 0m
			Het	Ms	NM_025114.4:c.4083C>T	NP_079390.3:p.Arg1355Cys	4.97E-05	9.53E-05	Delet	Prob_dam	32	Abx	Abx	Unk	OM, 0m
64	Limb-girdle muscular dystrophy	CEP290	Hom	Ms	NM_025114.4:c.4805C>T	NP_079390.3:p.Thr1602Met	0.000226	0.0002756	Delet	Poss_dam	27.7	Abx	Abx	Unk	OM, 0m
			Het	Ms	NM_025114.4:c.5908C>A	NP_079390.3:p.Thr1970Asn	0	1.27E-05	Delet	Prob_dam	25.6	Abx	Abx	Unk	OM, 0m
65	Undiagnosed monogenic disorders	CEP290	Het	Sg, Fs	NM_025114.4:c.7283_7286dup	NP_079390.3:p.Tyr429Irr	2.11E-05	2.54E-05	Delet	Prob_dam	25.6	Abx	Abx	Unk	OM, 0m
			Het	Ms	NM_025114.4:c.31As>G	NP_079390.3:p.Met117Val	7.72E-05	6.35E-05	Delet	Ben	23.2	-	-	Unk	OM, 0m
68	Early onset dementia	CEP290	Het	Ms	NM_025114.4:c.2447G>A	NP_079390.3:p.Arg816His	3.13E-05	6.35E-05	Delet	Prob_dam	26.4	-	-	Unk	OM, 0m
			Het	Ms	NM_025114.4:c.2486C>T	NP_079390.3:p.Arg816Cys	5.04E-05	2.54E-05	Delet	Prob_dam	32	25741868	25741868	Unk	OM, 0m
69	Epilepsy plus other features	CEP290	Het	Ms	NM_025114.4:c.4741C>T	NP_079390.3:p.Leu1581Phe	1.32E-05	1.27E-05	Delet	Poss_dam	25.1	25741868	25741868	Unk	OM, 0m
			Het	Ms	NM_001377.3:c.10140C>T	NP_001368.2:p.Pro3381Ileu	3.62E-05	1.00E-04	Delet	Prob_dam	31	Abx	Abx	Unk	OM, 0m
71	Hereditary ataxia	DYNCH1	Het	Ms	NM_001377.3:c.3419G>T	NP_001368.2:p.Gly1140Val	0.000898	0.0004987	Delet	Ben	23.2	Abx	Abx	Unk	OM, 0m
			Het	Spl,A	NM_008611.3:c.935-1G>A	-	-	0	6.35E-05	Delet	Ben	23.2	Abx	Abx	Unk
72	Early onset dementia	OPD1 (S)	Het	Fs	NM_008611.3:c.1911del	NP_008602.1:p.Glu673Asp1917er29	0	6.35E-05	Delet	Ben	23.2	Abx	Abx	Unk	OM, 0m
73	Early onset dystonia	OPD1 (S)	Het	Fs	NM_008611.3:c.1911del	NP_008602.1:p.Glu673Asp1917er29	0	6.35E-05	Delet	Ben	23.2	Abx	Abx	Unk	OM, 0m

Abx, absent; ACMG, American College of Medical Genetics and Genomics; AF, allele frequency; CADD, Combined Annotation Dependent Uplication; F, female; FS, frameshift; Het, heterozygous; HGVS, Human Genome Variation Society coding; HGVS, Human Genome Variation Society protein; Hom, homozygous; Int, intronic; LOD, 100,000 Genomes Project; Lik_path, likely pathogenic; M, male; M, major clinical feature; MAF, maximum allele frequency; Mat, maternal; Ms, missense; Pat, paternal; Path, pathogenic; Spl, splice region; Sg, splice acceptor; SV, structural variant; Syn, synonymous; Unk, unknown; VUS, variant of uncertain significance.

those variants affecting coding sequences, and splice donor or acceptor sites. The standard 100K pipeline requires diagnostic labs to analyse variants that are triaged into tier 1 or 2. Tier three variants (rare coding SNVs in genes not included in the selected panel or panels) and untiered variants are not routinely analysed in the diagnostic setting. The selection of incorrect panels that prevents appropriate tiering of causative variants, and the fact that certain types of variant are not routinely tiered, will therefore both contribute to missed diagnoses. Furthermore, inaccurate or incomplete HPO term entries at the time of recruitment will lead to inappropriate virtual gene panel selections that will not allow the analysis of the correct causative disease gene. These problems of missed diagnoses for both the present reverse phenotyping study and our previous analysis of the ‘CMC’ cohort,¹³ suggests that a change in protocol should be considered. This would permit further gene panel selection in the absence of good phenotyping data, or when the answer is not found from the first panel(s) applied.

SVs and single heterozygous SNVs in recessive disease genes are not routinely tiered, even when the genes are on the panel(s) applied. Filtering of all variants in our selected genes independent of the GEL tiering system, followed by independent annotation and analysis, has allowed us to identify SNVs most likely to be pathogenic, even when they are a single hit in a recessive disease gene. If the second variant in the same gene is difficult to find, for example, if it is an SV or intronic variant, then their identification in our pipeline could improve diagnostic yield. In particular, the introduction of the SVRare script,²² permitting exclusion of SV calls from analysis if they appear in >10 100K participants, has facilitated diagnosis of two previously unsolved participants (#45 and #70) with untiered, likely pathogenic SVs. SVRare provides a fast and systematic approach to SV analysis, which will be invaluable for future genomic studies. All 100K participants have SV.vcf files available in the Research Environment, called using the Manta and Canvas pipelines.^{23 24} To date, strategies to filter the huge number of SVs from these outputs, most of which are common and benign, have been limited. Alongside manual IGV inspection, the SVRare pipeline also allowed more accurate definition of the complex *ALMS1* SV found in participant #45, since it was called as both a rare inversion (Manta) and duplication (Canvas).

A further source of untiered, potentially pathogenic variants is our custom SpliceAI script. Currently, novel intronic variants are not routinely tiered. SpliceAI has provided one possible new diagnosis in participant #49, with the identification of a rare, homozygous intronic variant predicted to cause a *CEP290* splice acceptor gain (NM_025114.4:c.6011+874G>T, SpliceAI_DS_AG 0.64; figure 3D).

These sources of potentially missed causative variants shows the value of research collaborations to make the most of available genomic data. In particular, comprehensive SV and intronic variant analysis facilitates diagnoses not easily achievable through WES and gene-panel testing, but the standard 100K diagnostic pipelines do not yet take full advantage of these analyses.

The challenge of poor phenotyping data that prevents accurate variant interpretation

The quality of phenotyping has proven highly significant in determining the accuracy of variant interpretation in this study. At the time of recruitment to 100K, the HPO term entry for participants was frequently sparse, comprising one or two terms only, often from just one organ system. The Participant Explorer user interface can provide additional clinical data from longitudinal

patient records, which summarise medical history, and timelines for inpatient and outpatient observations, treatments and procedures. However, these data are of variable quality, and clinical features are not collated in a form amenable for genotype-phenotype correlation analyses. Given the frequently sparse clinical data available, we decided to report identified molecular diagnoses among participants with at least one major key clinical feature. This was to maximise the number of potential new diagnoses. With the limited data and systems available, we must pass responsibility on to the recruiting clinicians to refine any phenotypic fit in light of any additional clinical data to which they have access.

Effective communication with recruiting clinicians, providing additional clinical information not entered at the time of recruitment to 100K, has proven invaluable for accurate variant interpretation. However, of the 20 researcher-identified diagnosis forms and clinical collaboration request forms submitted via the GEL Airlock in the last 3 months, we have only received responses from four recruiting clinicians. Participant #62, recruited under the ‘epilepsy plus other features’ category with an ‘unsolved’ status on their GMC exit questionnaire, illustrates the value of effective researcher-clinician collaboration. We identified a ClinVar pathogenic *CEP290* frameshift variant (NM_025114.4:c.5434_5435del, NP_079390.3:p.Glu1812LysfsTer5) and a deep intronic *CEP290* variant known to cause a strong splice-donor site and insertion of a cryptic exon (NM_025114.4:c.2991+1655A>G).³⁴ Participant #62 had one *CEP290*-related key clinical feature from the ophthalmic system category (keratoconus), permitting us to report the finding. The recruiting clinician confirmed the presence of key ophthalmological features not entered during recruitment to 100K, comprising a formal diagnosis of Leber Congenital Amaurosis (bilateral keratoconus and cataracts, no detectable ERG responses to light) that was not previously specified. This strengthened confidence that the molecular diagnosis is correct and that this participant is highly likely to have a *CEP290*-related syndromic ciliopathy. It is unclear if the neurological features reported for participant #62 (diffuse cerebellar atrophy confirmed by MRI, but no evidence of structural brain abnormalities or intellectual disability), in addition to epilepsy, are associated with syndromic ciliopathy or comprise a separate phenotype. Nevertheless, reporting the molecular diagnosis is especially important in this instance, because the *CEP290* c.2991+1655A>G variant is a target for the development of antisense oligonucleotides that may offer a personalised therapy for patients.^{35 36}

Reverse phenotyping facilitates expansion of ciliopathy disease-gene associations

As was previously demonstrated for a family with an *INPP5E*-related ciliopathy,⁶ this study widens the phenotypic spectrum of known ciliopathy disease-gene associations through reverse phenotyping. For example, male participant #32 was reported ‘solved’ with a pathogenic hemizygous *OFD1* frameshift variant in exon 20/23 (NM_003611.3:c.2680_2681del, NP_003602.1:p.Glu894ArgfsTer6). Although participant #32 was recruited to the ‘rod-cone dystrophy’ category with apparently non-syndromic retinal dystrophy, reverse phenotyping revealed that he had clinical features that were consistent with a syndromic ciliopathy. Truncating variants in the C-terminal end of *OFD1* (exons 20–21) have recently been associated with the motile ciliopathy primary ciliary dyskinesia (PCD) without the characteristic skeletal, neurological or renal features of other *OFD1*-related disorders.^{32 37} The OFD1 protein is a component

Developmental defects

of ciliary basal bodies and centrioles, and has been shown to be essential for both primary and motile ciliogenesis.³⁸ Therefore, it is entirely plausible that pathogenic *OFD1* variants could cause features compatible with both motile and primary ciliopathies, therefore accounting for participant #32's full constellation of features (retinal dystrophy, renal failure and intellectual disability in keeping with primary ciliopathies and PCD-like respiratory disease with motile ciliopathies). Further reports of patients with both motile and primary ciliopathy features that carry pathogenic *OFD1* variants would strengthen this potential broadening of associated phenotypes. It is possible that the exon 20 frame-shift variant identified in participant #32 could just explain part of his phenotype, for example, his PCD-like respiratory disease, in keeping with the published literature.^{32,37} Conversely, retinal dystrophy may be an additional feature, as has been reported in association with X linked recessive JBTS caused by pathogenic *OFD1* variants in affected males.³⁹ We therefore suggest that individuals with a suspected *OFD1*-associated ciliopathy undergo a formal ophthalmological assessment to strengthen the diagnosis.

Unreportable diagnoses

As well as the 18 reportable molecular diagnoses, we also identified 11 unreportable molecular diagnoses for the 9 ciliopathy disease genes (table 3). Parental sequence is not available for any of the participants with unreportable diagnoses apart from one (#52). Lack of segregation analyses hamper accurate variant interpretation. Nevertheless, it is highly likely that some of these molecular diagnoses are correct and clinically actionable, with implications for the proband and for their relatives. The inability to report these findings is likely to be driven by inaccurate HPO term entry, which is a great loss to the participants. A review of reporting guidelines, given this important observation, may prove beneficial. For example, a system could be devised that marks potential pathogenic variants of interest that then requests further clinical information, but these remain unreportable until further, actionable data are available.

Conclusion

This study reveals the power of reverse phenotyping approaches to improve diagnosis rates for rare disease participants entered into large-scale genomic studies such as the 100K. Through the application of additional novel screening methodologies such as the SVRare suite, and with domain-specific knowledge, we have confirmed existing ciliopathy diagnoses and identified additional ones in a series of 100K participants who were not originally recruited as having a primary ciliopathy. Our findings suggest that diagnoses may be missed when screening of limited gene panels is directed by incorrect or incomplete HPO term entry, and that inaccurate phenotyping may prevent participants from accessing clinically valuable findings. We have discussed the challenges of 100K analyses more extensively in our recent commentary article and suggest potential improvements for future use of 100K data.³³ Clearly, open dialogue between researchers, clinicians and clinical scientists is essential to fully exploit the available data for patient benefit in the postgenomic era.

Author affiliations

¹Division of Molecular Medicine, Leeds Institute of Medical Research, University of Leeds, Leeds, UK

²Yorkshire Regional Genetics Service, Leeds Teaching Hospitals NHS Trust, Leeds, UK

³Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, University of Oxford, Oxford, UK

⁴University Hospital Southampton NHS Foundation Trust, Southampton, UK

⁵Faculty of Medicine, Human Development and Health, University of Southampton, Southampton, UK

⁶Windsor House Group Practice, Mid Yorkshire Hospitals NHS Trust, Leeds, UK

⁷North East and Yorkshire Genomic Laboratory Hub, Central Lab, Leeds Teaching Hospitals NHS Trust, Leeds, UK

⁸Genomics England, Queen Mary University of London, London, UK

⁹St. Paul's Eye Unit, Royal Liverpool University Hospital, Liverpool, UK

¹⁰University College London (UCL) Queen Square Institute of Neurology, London, UK

¹¹Chalfont Centre for Epilepsy, Chalfont, UK

¹²Department of Ophthalmology, Torbay and South Devon NHS Foundation Trust, Torquay, UK

¹³Exeter Genomics Laboratory, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK

¹⁴RILD Wellcome Wolfson Centre, University of Exeter Medical School, Exeter, UK

¹⁵Peninsula Clinical Genetics Service, Royal Devon and Exeter NHS Foundation Trust, Exeter, UK

Twitter Christopher Mark Watson @ChrisM_Watson and Gabrielle Wheway @gabriellewheway

Collaborators The Genomics England Research Consortium: Ambrose, J C (Genomics England, London, UK); Arumugam, P (Genomics England, London, UK); Bevers, R (Genomics England, London, UK); Bleda, M (Genomics England, London, UK); Boardman-Pretty, F (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Boustred, C R (Genomics England, London, UK); Brittain, H (Genomics England, London, UK); Brown, MA (Genomics England, London, UK); Caulfield, M J (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Chan, G C (Genomics England, London, UK); Fowler, T (Genomics England, London, UK); Giess A (Genomics England, London, UK); Hamblin, A (Genomics England, London, UK); Henderson, S (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Hubbard, T J P (Genomics England, London, UK); Jackson, R (Genomics England, London, UK); Jones, L J (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Kasperaviciute, D (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Kayikci, M (Genomics England, London, UK); Kousathanas, A (Genomics England, London, UK); Lahnstein, L (Genomics England, London, UK); Leigh, S E A (Genomics England, London, UK); Leong, I U S (Genomics England, London, UK); Lopez, F J (Genomics England, London, UK); Maleady-Crowe, F (Genomics England, London, UK); McEntagart, M (Genomics England, London, UK); Minnedi F (Genomics England, London, UK); Moutsianas, L (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Mueller, M (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Murugaesu, N (Genomics England, London, UK); Need, A C (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; O'Donovan P (Genomics England, London, UK); Odhams, CA (Genomics England, London, UK); Patch, C (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Perez-Gil, D (Genomics England, London, UK); Pereira, M B (Genomics England, London, UK); Pullinger, J (Genomics England, London, UK); Rahim, T (Genomics England, London, UK); Rendon, A (Genomics England, London, UK); Rogers, T (Genomics England, London, UK); Savage, K (Genomics England, London, UK); Sawant, K (Genomics England, London, UK); Scott, R H (Genomics England, London, UK); Siddiq, A (Genomics England, London, UK); Sieghart, A (Genomics England, London, UK); Smith, S C (Genomics England, London, UK); Sosinsky, A (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Stuckey, A (Genomics England, London, UK); Tanguy M (Genomics England, London, UK); Taylor Tavares, A L (Genomics England, London, UK); Thomas, E R A (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Thompson, S R (Genomics England, London, UK); Tucci, A (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Welland, M J (Genomics England, London, UK); Williams, E (Genomics England, London, UK); Witkowska, K (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK; Wood, S M (Genomics England, London, UK); William Harvey Research Institute, Queen Mary University of London, London, EC1M 6BQ, UK.

Contributors Conceptualisation: SB, CT, CI, CAJ, GW; data curation: SB, JY, Genomics England Research Consortium; formal analysis: SB, JY, JL, MR, CMW, SM,

RJ, CT, CI, CAJ, GW: funding acquisition: SB, JL, CT, CI, CAJ, GW; investigation: SB, JY, JL, MR, CMW, SMS, RJ, CT, CI, CAJ, GW; methodology: SB, JY, JL, MR, CT, CI, CAJ, GW; software: JY, JL, MR, RPJB, AS, Genomics England Research Consortium; project administration: SB, JY, Genomics England Research Consortium, GW; resources: SB, SMS, SL, ST, EB, Genomics England Research Consortium; supervision: CT, CFI, CAJ, GW; validation: SMS, SL, ST, EB; writing—original draft: SB, GW; writing—review and editing: all authors; guarantor: CAJ

Funding SB acknowledges support from the Wellcome Trust 4Ward North Clinical PhD Academy (ref. 203914/Z/16/Z). JY acknowledges support from Retina UK (grant HMRO3950). GW acknowledges support from Wellcome Trust Seed Award (ref. 204378/Z/16/Z). CAJ acknowledges support from Medical Research Council project grants MR/M000532/1 and MR/T017503/1. JL is supported by a National Institute for Health Research (NIHR) Research Professorship awarded to Professor Diana Baralle (DB NIHR RP-2016-07-011). SMS is supported by the Epilepsy Society and the NIHR University College London Hospitals Biomedical Research Centre. This research was made possible through access to the data and findings generated by the 100 000 Genomes Project. The 100 000 Genomes Project is managed by Genomics England Ltd (a wholly owned company of the Department of Health and Social Care). The 100 000 Genomes Project is funded by the National Institute for Health Research and NHS England. The Wellcome Trust, Cancer Research UK and the Medical Research Council have also funded research infrastructure. The 100 000 Genomes Project uses data provided by patients and collected by the National Health Service as part of their care and support.

Competing interests RPJB and AS are employed by Genomics England Ltd, UK. GW is employed by Illumina. The other authors declare no conflict of interest.

Patient consent for publication Not applicable.

Ethics approval Access to the secure online Research Environment within the Genomics England Ltd (GEL) Data Embassy was provided by the GEL Access Review Committee, and research project RR185 'Study of cilia and ciliopathy genes across the 100 000 GP cohort' was registered and approved by GEL. This research study received ethical approval from University of Southampton Faculty of Medicine Ethics Committee (ERGO 54400). Written informed consent was obtained from all participants (or from their parent/legal guardian) in the 100 000 Genomes Project (IRAS ID 166046; REC reference 14/EE/1112).

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. Data are available on reasonable request. Full data are available in the Genomic England Secure Research Environment. All datasets are available in the re_gcep shared folder of the GEL research environment for approved researchers. Access to our folder containing variant data (re_gcep/shared_allGCE/CIPs/GW_SB) can be requested from the GEL Helpdesk.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution 4.0 Unported (CC BY 4.0) license, which permits others to copy, redistribute, remix, transform and build upon this work for any purpose, provided the original work is properly cited, a link to the licence is given, and indication of whether changes were made. See: <https://creativecommons.org/licenses/by/4.0/>.

ORCID iDs

Christopher Mark Watson <http://orcid.org/0000-0003-2371-1844>

Gabrielle Wheway <http://orcid.org/0000-0002-0494-0783>

Colin A Johnson <http://orcid.org/0000-0002-2979-8234>

REFERENCES

- Tumbull C, Scott RH, Thomas E, Jones L, Murugesu N, Pretty FB, Halai D, Baple E, Craig C, Hamblin A, Henderson S, Patch C, O'Neill A, Devereau A, Smith K, Martin AR, Sosinsky A, McDonagh EM, Sultana R, Mueller M, Smedley D, Toms A, Dinh L, Fowler T, Bale M, Hubbard T, Rendon A, Hill S, Caulfield MJ, 100 000 Genomes Project. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ* 2018;361.
- Wheway G, Mitchison HM, Genomics England Research Consortium. Opportunities and challenges for molecular understanding of Ciliopathies-The 100,000 genomes project. *Front Genet* 2019;10:127–27.
- The National genomic research library v5.1 2020.
- Swietlik EM, Prapa M, Martin JM, Pandya D, Auckland K, Morrell NW, Gräf S. 'There and back Again'-Forward genetics and reverse phenotyping in pulmonary arterial hypertension. *Genes* 2020;11. doi:10.3390/genes11121408. [Epub ahead of print: 26 11 2020].
- Landini S, Mazzinghi B, Becherucci F, Allinovi M, Provenzano A, Palazzo V, Ravaglia F, Artuso R, Bosi E, Stagi S, Sansavini G, Guzzi F, Cirillo L, Vaglio A, Murer L, Peruzzi L, Pasini A, Materassi M, Roberto RM, Anders H-J, Rotondi M, Giglio SR, Romagnani P. Reverse phenotyping after whole-exome sequencing in steroid-resistant nephrotic syndrome. *Clin J Am Soc Nephrol* 2020;15:89–100.
- de Goede C, Yue WW, Yan G, Ariyaratnam S, Chandler KE, Downes L, Khan N, Mohan M, Lowe M, Banka S. Role of reverse phenotyping in interpretation of next generation sequencing data and a review of INPP5E related disorders. *Eur J Paediatr Neurol* 2016;20:286–95.
- Singla V, Reiter JF. The primary cilium as the cell's antenna: signaling at a sensory organelle. *Science* 2006;313:629–33.
- Oud MM, Lamers IJC, Arts HH. Ciliopathies: genetics in pediatric medicine. *J Pediatr Genet* 2017;6:018–29.
- Reiter JF, Leroux MR. Genes and molecular pathways underpinning ciliopathies. *Nat Rev Mol Cell Biol* 2017;18:533–47.
- Mitchison HM, Valente EM. Motile and non-motile cilia in human pathology: from function to phenotypes. *J Pathol* 2017;241:294–309.
- Waters AM, Beales PL. Ciliopathies: an expanding disease spectrum. *Pediatr Nephrol* 2011;26:1039–56.
- Quinlan RJ, Tobin JL, Beales PL. Modeling ciliopathies: primary cilia in development and disease. *Curr Top Dev Biol* 2008;84:249–310.
- Toriello HV, Franco B, Bruel AL, Thauvin-Robinet C. Oral-Facial-Digital Syndrome Type I. In: Adam MP, Ardinger HH, Pagon RA, eds. *GeneReviews*®. Seattle (WA): University of Washington, 1993.
- McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. The Ensembl variant effect predictor. *Genome Biol* 2016;17:122.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferreira S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Neale BM, Daly MJ, MacArthur DG, Genome Aggregation Database Consortium. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 2020;581:434–443.
- Whiffin N, Minikel E, Walsh R, O'Donnell-Luria AH, Karczewski K, Ing AY, Barton PJR, Funke B, Cook SA, MacArthur D, Ware JS. Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet Med* 2017;19:1151–8.
- Ellard SBE, Berry I, Forrester N, Turnbull C, Owens M, Eccles DM, Abbs S, Scott R, Deans Z, Lester T, Campbell J, Newman W, McMullan D. ACGS Best Practice Guidelines for Variant Classification in Rare Disease, 2020. Available: <https://www.acgs.uk.com/media/11631/uk-practice-guidelines-for-variant-classification-v4-01-2020.pdf>
- Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, Jang W, Katz K, Ovetzky M, Riley G, Sethi A, Tully R, Villamarin-Salomon R, Rubinstein W, Maglott DR. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* 2016;44:D862–8.
- 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:W452–7.
- Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. Cadd: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* 2019;47:D886–94.
- Adzhubei I, Jordan DM, Sunyaev SR. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* 2013;Chapter 7.
- Yu J, Szabo A, Pagnamenta AT, Shalaby A, Giacomuzzi E, Taylor J, Shears D, Pontikos N, Wright G, Michaelides M, Halford S, Downes S, Genomics England Research Consortium. SVRare: discovering disease-causing structural variants in the 100K genomes project. *medRxiv* 2021;10.15121265069.
- Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* 2016;32:1220–2.
- Ivakhno S, Roller E, Colombo C, Tedder P, Cox AJ. Canvas SPW: calling de novo copy number variants in pedigrees. *Bioinformatics* 2018;34:516–8.
- Thorvaldsdóttir H, Robinson JT, Mesirov JP, Viewler IG. Integrative genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 2013;14:178–92.
- Jaganathan K, Kyriazopoulou Panagiopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, Kosmicki JA, Arbelaez J, Cui W, Schwartz GB, Chow ED, Kanterakis E, Gao H, Kia A, Batzoglu S, Sanders SJ, Farh KK-H. Predicting splicing from primary sequence with deep learning. *Cell* 2019;176:535–48.
- Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, Hegde M, Lyon E, Spector E, Voelkerding K, Rehml HL, ACMG Laboratory Quality Assurance Committee. 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* 2015;17:405–24.
- 28 Best S, Lord J, Roche M, Watson CM, Poulter JA, Bevers RP, Stuckey A, Szymanska K, Ellingford JM, Carmichael J, Brittain H, Toomes C, Inglehearn C, Johnson CA, Wheway G, Genomics England Research Consortium. Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 genomes project. *J Med Genet* 2021. doi:10.1136/jmedgenet-2021-108065. [Epub ahead of print: 29 Oct 2021].
- 29 Han PKJ, Umstead KL, Bernhardt BA, Green RC, Joffe S, Koenig B, Krantz I, Waterston LB, Biesecker LG, Biesecker BB, PKJ H, Biesecker BB. A taxonomy of medical uncertainties in clinical genome sequencing. *Genet Med* 2017;19:918–25.
- 30 Paisey RB, Steeds R, Barrett T, Williams D, Geberhiwot T, Gunay-Aygun M, Adam MP, Ardinger HH, Pagon RA. Alström Syndrome. In: *GeneReviews*(®). Seattle (WA): University of Washington, Seattle, 1993.
- 31 Myktyyn K, Nishimura DY, Searby CC, Shastri M, Yen H-jan, Beck JS, Braun T, Streb LM, Cornier AS, Cox GF, Fulton AB, Carmi R, Lüleci G, Chandrasekharappa SC, Collins FS, Jacobson SG, Heckenlively JR, Weleber RG, Stone EM, Sheffield VC. Identification of the gene (BBS1) most commonly involved in Bardet-Biedl syndrome, a complex human obesity syndrome. *Nat Genet* 2002;31:435–8.
- 32 Bukowy-Bierylo Z, Rabiś A, Dabrowski M, Pogorzelski A, Wojda A, Dmenska H, Grzela K, Sroczynski J, Witt M, Zietkiewicz E. Truncating mutations in exons 20 and 21 of *OFD1* can cause primary ciliary dyskinesia without associated syndromic symptoms. *J Med Genet* 2019;56:769–77.
- 33 Best S, Inglehearn CF, Watson CM, Toomes C, Wheway G, Johnson CA. Unlocking the potential of the UK 100,000 Genomes Project—lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort. *Am J Med Genet C Semin Med Genet* 2022;190:5–8.
- 34 den Hollander AI, Koenekoop RK, Yzer S, Lopez I, Arends ML, Voeseek KEJ, Zonneveld MN, Strom TM, Meitinger T, Brunner HG, Hoyng CB, van den Born LI, Rohrschneider K, Cremers FPM. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet* 2006;79:556–61.
- 35 Dulla K, Aguila M, Lane A, Jovanovic K, Parfitt DA, Schulkens I, Chan HL, Schmidt I, Beumer W, Vorthoren L, Collin RWJ, Garanto A, Duijkers L, Brugulat-Panes A, Semo Ma'ayan, Yugler AA, Biasutto P, Adamson P, Cheatham ME. Splice-Modulating Oligonucleotide QR-110 Restores CEP290 mRNA and Function in Human c.2991+1655A>G LCA10 Models. *Mol Ther Nucleic Acids* 2018;12:730–40.
- 36 Duijkers L, van den Born LI, Neidhardt J, Bax NM, Pierrache LHM, Klevering BJ, Collin RWJ, Garanto A. Antisense Oligonucleotide-Based Splicing Correction in Individuals with Leber Congenital Amaurosis due to Compound Heterozygosity for the c.2991+1655A>G Mutation in CEP290. *Int J Mol Sci* 2018;19. doi:10.3390/ijms19030753. [Epub ahead of print: 07 Mar 2018].
- 37 Guo Z, Chen W, Wang L, Qian L. Clinical and genetic spectrum of children with primary ciliary dyskinesia in China. *J Pediatr* 2020;225:157–65.
- 38 Ferrante MI, Zullo A, Barra A, Bimonte S, Messaddeq N, Studer M, Dollé P, Franco B. Oral-Facial-Digital type I protein is required for primary cilia formation and left-right axis specification. *Nat Genet* 2006;38:112–7.
- 39 Bruel A-L, Franco B, Duffout Y, Thevenon J, Jégo L, Lopez E, Deleuze J-F, Doummar D, Giles RH, Johnson CA, Huynen MA, Chevrier V, Burglen L, Morleo M, Desguerres I, Pierquin G, Doray B, Gilbert-Dussardier B, Reversade B, Steichen-Gersdorf E, Baumann C, Panigrahi I, Fargeot-Espaliat A, Dieux A, David A, Goldenberg A, Bongers E, Gaillard D, Argente J, Aral B, Gigot N, St-Onge J, Birnbaum D, Phadke SR, Cormier-Daire V, Eguether T, Pazour GJ, Herranz-Pérez V, Goldstein JS, Pasquier L, Loget P, Saunier S, Mégarbané A, Rosnet O, Leroux MR, Wallingford JB, Blacque OE, Nachury MV, Attie-Bitach T, Rivière J-B, Fairve L, Thauvin-Robinet C. Fifteen years of research on oral-facial-digital syndromes: from 1 to 16 causal genes. *J Med Genet* 2017;54:371–80.
- 40 Forsyth R, Gunay-Aygun M. Bardet-Biedl Syndrome Overview. In: Adam MP, Ardinger HH, Pagon RA, eds. *GeneReviews*(®). Seattle (WA): University of Washington, 1993.
- 41 Keppler-Noreuil KM, Adam MP, Welch J, Muilenburg A, Willing MC. Clinical insights gained from eight new cases and review of reported cases with Jeune syndrome (asphyxiating thoracic dystrophy). *Am J Med Genet A* 2011;155:1021–32.
- 42 Stokman M, Lilien M, Knoers N. Nephronophthisis. In: Adam MP, Ardinger HH, Pagon RA, eds. *GeneReviews*(®). Seattle (WA): University of Washington, 1993.
- 43 Parisi M, Glass I. Joubert Syndrome. In: Adam MP, Ardinger HH, Pagon RA, eds. *GeneReviews*(®). Seattle (WA): University of Washington, 1993.
- 44 Hartill V, Szymanska K, Sharif SM, Wheway G, Johnson CA. Meckel-Gruber syndrome: an update on diagnosis, clinical management, and research advances. *Front Pediatr* 2017;5.
- 45 Kumaran N, Pennesi ME, Yang P, Trzupek KM, Schlechter C, Moore AT, Weleber RG, Michaelides M. Leber Congenital Amaurosis / Early-Onset Severe Retinal Dystrophy Overview. In: Adam MP, Ardinger HH, Pagon RA, eds. *GeneReviews*(®). Seattle (WA): University of Washington, 1993.
- 46 Brown DE, Pittman JE, Leigh MW, Fordham L, Davis SD. Early lung disease in young children with primary ciliary dyskinesia. *Pediatr Pulmonol* 2008;43:514–6.

4 Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*

4.1 Research Rationale

This laboratory-based project was designed in collaboration with colleagues from Professor Oliver Blacque's cilium disease research group at University College Dublin (UCD). The UCD group has expertise in ciliary biology through the study of the nematode worm *C. elegans*. The rationale was to develop functional strategies for definitive ciliopathy gene missense variant interpretation, to reduce the proportion of variants that were classified as VUSs. The plan was to develop parallel variant interpretation strategies in two model systems: *C. elegans* and a human ciliated cell line. We used CRISPR to perform variant modelling as it was emerging as a relatively rapid and straightforward genome editing strategy compared to previous options. The aim was to develop a high-throughput variant interpretation pipeline, providing proof of principle that functional tests involving CRISPR can be useful in the diagnostic setting.

TMEM67 was selected as an *exemplar* gene for two main reasons. Firstly, the *C. elegans* worm protein mks-3 is orthologous with human *TMEM67* and has highly conserved structure and function, unlike other ciliary proteins such as CEP290. Secondly, discussion with clinical scientist colleagues in the Leeds NHS Genetics Diagnostics Laboratory revealed that a long list of *TMEM67* missense VUSs had been identified amongst local fetuses with clinical features of the lethal ciliopathy syndrome MKS, preventing their definitive molecular diagnosis. We wanted to undertake functional research for some of these VUSs with the aim of providing clinical benefit to our local patients, as well as the wider ciliopathy community.

I was initially responsible for selection of *TMEM67* variants to model. I obtained the list of local *TMEM67* VUSs through discussion with Ian Berry, a Clinical Scientist within the Leeds NHS Genetics Diagnostics Laboratory. I then accessed medical records for the patients amongst whom these VUSs were identified to collate relevant clinical data. I also used the ClinVar database to select *TMEM67* variants with a full range of predicted effects, including known pathogenic and known benign, to provide a range of expected cellular phenotypes with which to test VUS interpretation. We aimed to prioritise variants that were adjacent to one another, so the same CRISPR guides utilizing the same PAM

sequence could be used to generate multiple different variants. Only variants in residues conserved between worm mks-3 and human TMEM67 were considered.

4.2 Additional methodology

The Materials and Methods section of the accompanying manuscript contains most of the methodology applied in this study (Lange et al., 2022). The detail below contains methodology either not applied in the experiments included in the accompanying manuscript or that was considered un-necessarily detailed. Further information about reagents used is provided in Appendix section 6.3, including suppliers (6.3.1), reagents (6.3.2), buffers and solutions (6.3.3), cell lines (6.3.4) and antibodies and cell stains (6.3.5).

4.2.1 Polymerase Chain Reaction (PCR)

4.2.1.1 Primer design

PCR primers were designed using AutoPrimer 3 software (available from (<https://github.com/gantzgraf/autoprimer3>)). This retrieves gene information from the University of California Santa Cruz (UCSC) genome browser and uses primer3 (<http://primer3.ut.ee>) to automatically design primers to genes or genomic coordinate targets. Primers were selected when specific parameters were met, including an optimum annealing temperature of 58 – 65°C, GC content 40-60%, and excluding common SNPs. Primers are presented in Table S7 of the Supplementary Material (thesis section 6.1.3) (Lange et al., 2022).

4.2.1.2 PCR reaction

PCR amplification of target regions was performed using 5µl of HotShot Diamond 2x PCR Mastermix, 0.5µl of 10µM forward and reverse primers, 25ng of DNA and nuclease free H₂O to make up final volume of 10µl. Reactions were cycled on a Veriti Dx Thermal Cycler with an initial denaturation at 95°C (10 minutes), then 35 cycles of: denaturation at 95°C (30 seconds), annealing with temperature optimised to primers (~59-64°C) (30 seconds) and extension at 72°C (1 minute). After the 35 cycles, there was a final extension at 72°C (5 minutes). Completed reactions were held at 10°C. PCR products were analysed by gel electrophoresis (see thesis section 4.2.2).

4.2.2 Agarose gel electrophoresis

Samples for visualisation were mixed in a 5:1 ratio with 6x loading buffer and run on 1-2% weight for volume (w/v) agarose gels stained with 1x Midori Green Advance, alongside appropriate size standard (Easy Ladder or Quick-Load® Purple 1 kb DNA Ladder). Gels were run at 100-150V, for between 30 minutes and 2 hours according to expected product size, in an electrophoresis tank with 1x TAE buffer. Products were visualised on a GelDoc Ultraviolet (UV) transillumination station (BioRad) and displayed on Image Lab (v4.0) software for analysis (BioRad).

4.2.3 Exonuclease I – Shrimp Alkaline Phosphatase (ExoSAP) PCR purification

PCR products were purified by enzymatic treatment with ExoSAP-IT™ Express to digest excess primer and dephosphorylate nucleotides to allow for downstream sequencing reactions. 2.5µl of PCR product was treated with 1µl of ExoSAP according to manufacturer's instructions.

4.2.4 Sanger Sequencing

Sequencing reactions were made up of 0.5µl BigDye Terminator Kit V3.1, 2µl BigDye Sequencing Buffer (5x), 0.5µl 0.2µM sequencing primer, 1µl purified PCR product or 100ng purified plasmid DNA and dH₂O to make up final volume of 10µl. Primers are presented in Table S7 of the Supplementary Material (thesis section 6.1.3) (Lange et al., 2022). Sequencing reactions underwent initial denaturation at 96°C (1 minute), then 45 cycles of denaturation at 96°C (10 seconds), annealing at 50°C (5 seconds) and extension at 60°C (4 minutes). Completed reactions were held at 10°C.

Sequencing products were transferred to 96-well sequencing plates for precipitation. All spins were performed at 4°C. 5µl of 125mM Ethylenediaminetetraacetic acid (EDTA) and 60µl of 100% ethanol were added to each well before centrifugation at 2750 x g for 30 minutes. Plates were inverted onto tissue and centrifuged at 10 x g for 10 seconds to remove the supernatant. Contents were washed in 70% ethanol and spun for 15 minutes at 2750 x g. They were inverted onto tissue and again centrifuged at 10 x g for 10 seconds to remove residual ethanol. Pellets were dried on a 95°C hot plate until all visible ethanol had gone (around 2 minutes). 10µl of deionised HiDi™ formamide was applied to each well, then sequencing reactions run on an ABI 3130xl Genetic Analyzer. Base calling was done using Sequencing Analysis software v5.2 (Applied Biosystems™) and sequence data analysed using SeqScape software v2.5 (Applied Biosystems™) and

SnapGene software (GSL Biotech LLC).

4.2.5 Bacterial transformation of variant plasmids generated by site-directed mutagenesis

A *TMEM67_myc/HisA* plasmid was used for complementation assays. Detail about how this was designed and generated is provided in the “*TMEM67* cloning, plasmid constructs and transfections” section within the Materials and Methods of the accompanying manuscript (thesis section 4.5) (Lange et al., 2022). The wild-type plasmid was fully sequence verified; the complete plasmid map is presented in section 6.4 of the Appendix. Variant plasmids were generated from the wild-type using a QuikChange II XL Site-Directed Mutagenesis Kit (Agilent) then transformed into either *E. coli* XL10-Gold Ultracompetent Cells or Alpha-Select Chemically Competent Cells according to manufacturer’s instructions.

Following bacterial transformation, four individual cell colonies for each variant plate were picked using a pipette tip and transferred to separate 15ml falcon tubes containing 5ml of 100µg/ml ampicillin Luria-Bertani (LB) media. These were transferred to the 37°C shaking incubator for 16 hours, before being stored at 4°C. 1ml of cell solution was used for DNA extraction using a QIAprep Spin Miniprep Kit (Qiagen) according to the manufacturer’s protocol. DNA concentration was determined using a Nanodrop™ 2000 spectrophotometer. DNA from each miniprep sample was Sanger sequenced using at least two internal *TMEM67* primers that covered the site-directed mutagenesis targeted site. Primers are presented in Table S7 of the Supplementary Material (thesis section 6.1.3) (Lange et al., 2022). Once the targeted variants were verified by sequencing, 1ml of cell solution was grown in 200ml of 100µg/ml ampicillin LB media, before bulk DNA extraction and purification using a Plasmid endonuclease free Maxi Kit (Qiagen). Again, all maxi-prepped variant DNA was sequence verified prior to experimental use.

4.2.6 Small interfering RNA (siRNA) knockdown experiments

siRNA knockdown experiments were conducted as forward transfections within 6-well tissue culture plates. 3×10^5 cells per well were plated and ready for transfection when they reached ~70% confluence. *TMEM67* siRNA stock (100µM) was diluted in siRNA buffer (Dharmacon Inc.) to a final amount of 5nmol per well. Forward transfection reactions were prepared according to manufacturer’s protocols with 3µl of Lipofectamine 2000 and 5µl of diluted siRNA solution. Media was changed after 3-5 hours, and

transfections left for 24-72 hours depending on estimated transfection efficiency.

4.2.7 Whole cell extract (WCE) preparation and Western Blotting

WCEs were prepared from confluent cells in 6-well tissue culture plates or scaled as appropriate. All steps were undertaken on ice, to prevent protein degradation by proteases. Cells were washed twice with cold 1x phosphate buffered saline (PBS) and lysed with 50µl ice-cold NP40 lysis buffer for 5 minutes. Cells were scraped from the plates using chilled plastic cell scrapers, into pre-chilled tubes, then frozen at -80°C for at least 1 hour. Samples were thawed on ice, then agitated for 30 minutes in the orbital rotator at 4°C. Cells were spun down at 12,000 x g for 15 minutes at 4°C and the supernatant transferred to new tubes.

A RC DC™ Protein Assay Kit was used to measure protein concentration according to the manufacturer's protocol. Absorbance was measured on a spectrophotometer at 750nm. A bovine serum albumin (BSA) assay was used to produce a standard curve to infer protein concentration of WCE samples.

WCE samples were diluted in NP40 lysis buffer with 1% protease +/- phosphatase inhibitor (Promega) to produce equal concentrations for loading. Maximum protein concentration for loading was determined from the size of the gel being used. 4x loading dye with 2.5% beta-mercaptoethanol was added in a 1:3 ratio to give 1x loading dye in all samples. Samples were boiled at 90°C for 10 minutes to denature proteins. Samples were loaded into NuPAGE™ 4-12% MES SDS gradient gels alongside protein marker and run in 200ml of 1x MES-SDS running buffer for 90 minutes at 120V.

A polyvinylidene difluoride (PVDF) was activated in 100% methanol for 20 seconds. Proteins were immunoblotted onto the activated membrane in transfer buffer supplemented with 10% methanol for 60-90 minutes at 40V. The gel tank was placed on ice, and outer chambers filled with iced water during transfer.

Membranes were blocked in 5% w/v Marvel dried milk or 5% BSA diluted in 1x PBST for one hour. They were then incubated in primary antibodies diluted in 5ml of blocking agent in 50ml Falcon tubes on a roller at room temperature for 1 hour, or overnight on a roller

at 4°C. Membranes were washed 6 times in 1x PBST for 2 minutes per wash, then incubated for 1 hour with the appropriate horseradish peroxidase (HRP)-conjugated secondary antibody at 1:10,000 dilution on a roller at room temperature. Again, the membrane was washed 6 times in PBST. The West Femto immunoblot detection system was used to reveal bands by enhanced chemiluminescence. Bands were visualised on a GelDoc station (Bio-Rad) and processed in ImageLab software (Bio-Rad). Protein levels were quantified against a reference band and normalised to the loading control quantifications. Primary and secondary antibodies used are summarised in Tables 13 and 14 in Appendix section 6.3.5.

If membranes needed to be probed with another antibody, they were either stripped for 10 minutes in Restore™ PLUS Western Blot Stripping Buffer, then washed and re-blocked before the next stain, or the area of the membrane that had already been probed was cut off.

4.2.8 High-content imaging

4.2.8.1 Transfection

For high-content imaging, reverse transfections were set up in tissue culture treated CellCarrier-96 Ultra Microplates (Perkin-Elmer). First, wells were coated with 0.67µl Matrigel® in 50µl of ice-cold OptiMEM and left to set for an hour. This was washed in room temperature OptiMEM prior to setting up experiments. A 20µl transfection reagent mix was prepared for each well in Eppendorf tubes. 0.2µl of transfection reagent (PEI or Lipofectamine 2000) was added to OptiMEM and allowed to equilibrate for five minutes. 70ng of plasmid was added and incubated for 20-30 minutes. This transfection mix was added to the wells, then 80µl of suspended cells at 2×10^5 concentration applied on top (16,000 cells per well). Media was changed after 3-5 hours, and transfections left for 24-72 hours before fixation depending on estimated transfection efficiency.

4.2.8.2 Fixing and staining

A FluidX XRD-384 dispenser on slow speed (100rpm) was used for all steps to fix and stain plates, with solutions dispensed to the left side of the wells. Wells were first washed in sticky PBS then fixed with 50µl of ice-cold methanol for 5-7 minutes at -20°C. Plates were inverted and blotted to remove methanol then wells washed in 50µl of PBS. Plates were blocked for at least 10 minutes with 50µl of 1% Marvel dried milk/PBS [w/v], that

had previously been cleared of particulates by centrifugation at 3000 x g for 10 minutes. All antibodies and stains were diluted in blocking solution and clarified by centrifugation at 16K x g for one minute. 50µl of primary antibody solution was applied per well (see Table 12 for dilutions) and incubated for one hour at room temperature. Plates were washed 5 times in 1x PBS with 0.05% Triton™-X100, and once in 1x PBS, with inversion and blotting of plates between washes. 50µl of secondary antibody solutions (see Table 13 for dilutions) with 1:1000 DAPI and 1:4000 TOTO-3 were incubated for 1 hour, under foil wrapping to prevent light bleaching. Again, plates were washed five times in 1x PBS with 0.05% Triton™-X100 and left in the final wash of 1x PBS.

4.2.8.3 Imaging

A PerkinElmer Operetta high-content wide-field fluorescence imaging system, linked to Harmony software, was used to image and process plates. Wells were imaged using the 20x objective lens. Up to four fluorescent colours could be detected in different focal planes to provide maximum resolution for each, as well as bright-field imaging. The Operetta infra-red focussing laser was used to detect the bottom of each well automatically, and focal planes of detection for each colour calculated relative to this value. Image acquisition was optimised on negative controls. A consistent pattern of at least six fields of view were imaged per well, positioned in a ring around the central dispense area, with an approximate total of 4,000 cells detected and analysed per well.

4.2.8.4 Image analysis

Image data was imported into Columbus™ Image Data Storage and Analysis System for high-throughput analysis. Recognition protocols were written with the in-built software building blocks. The 'find nuclei' protocol recognition block was used in the DAPI (blue) channel as fluorescent regions $> 30\mu\text{m}^2$. The cell body was defined using the 'find cytoplasm' protocol recognition block by recognising far-red (TOTO-3) fluorescent regions surrounding nuclei. To make sure only whole cells were analysed, border objects were removed. A "find spots" algorithm was used to detect cilia on whole cells, identifying green (Alexa Fluor 488) or red (Alexa Fluor 568) fluorescent spots with radius < 8.3 pixels, contrast > 0.11 , uncorrected spot to region intensity > 0.5 . Key output parameters were number of whole cells, percentage of cells with a single or double cilium and intensity of anti-myc staining. These were calculated as an average across all fields of field per well.

4.2.8.5 Statistical analysis

Wells that had passed preliminary analysis had robust z-scores calculated. These were used instead of a standard z-score as they take experimental variation into account. Robust $Z_{score} = (x-m)/M$ where m = median values of the measured phenotype of the negative controls and M = median absolute deviation of the measured phenotype of the negative controls. On a normal distribution curve, of data point x compared to the negative controls, $-1.96 \geq \text{Robust } Z \text{ of } x \leq +1.96$ is equivalent to $p = 0.05$.

4.3 Additional results

4.3.1 Generation of *TMEM67* knockout cell lines

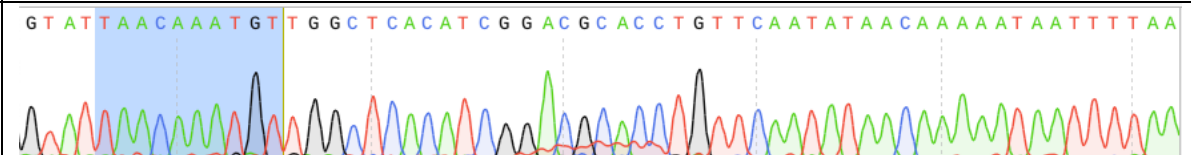
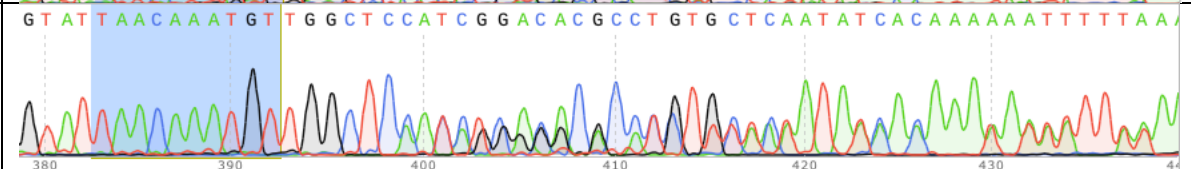
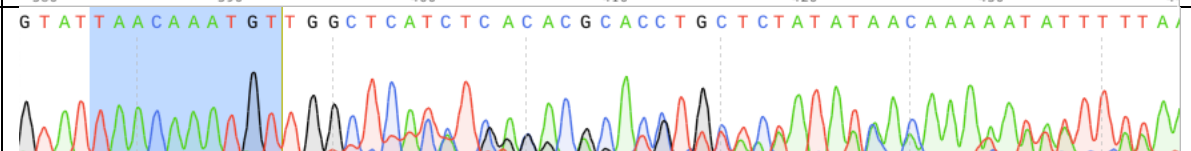
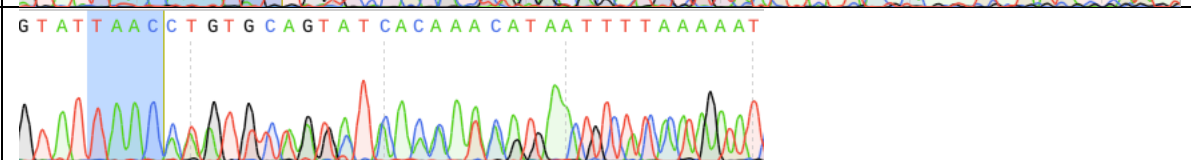
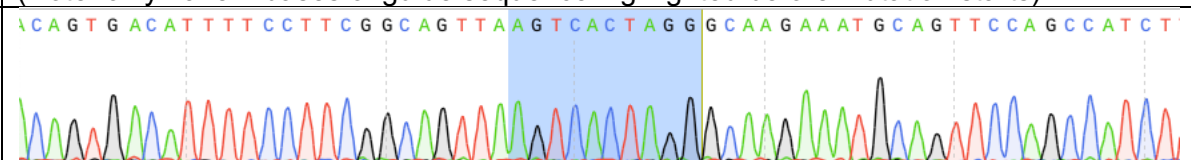
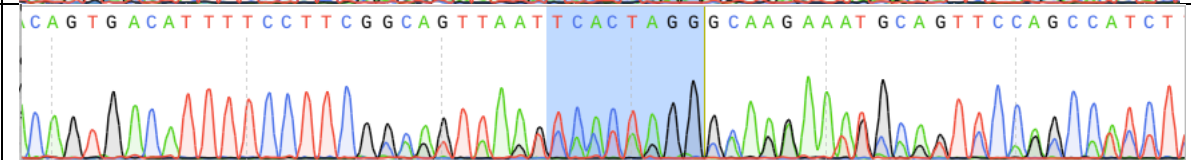
48-hours after transfection with crRNA:tracrRNA duplexes, 1440 GFP-expressing RPE-1 cells were index sorted by fluorescence-activated cell sorting (FACS) into 96-well plates. After three weeks of clonal growth, the healthiest surviving clones were taken forward for onward growth in fresh plates and DNA was extracted for sequencing. These represent ~10% of the original population. PCR encompassing gRNA target sequences of *TMEM67* (exons 3 and 5) was performed on extracted DNA, which was purified with ExoSapIT Express prior to Sanger sequencing. DNA from 70 clones that had successfully amplified on PCR, as visualised on gel electrophoresis, was Sanger sequenced. From these 70 sequenced clones, 5 were identified as CRISPR knockout cell lines through sequence analysis (7% efficiency rate).

The bi-allelic knockout *TMEM67* crisprant cell line clone 16 is presented in manuscript (thesis section 4.5) (Lange et al., 2022). It was selected for further characterisation and subsequent experimental use because it had the most straightforward bi-allelic variants identified, that were predicted to cause nonsense mediated decay (NMD). Sequence analysis revealed a one base-pair deletion on one allele and a one base-pair insertion at the same position on the other allele, corresponding to biallelic frameshift variants: c.519delT, p.(Cys173Trpfs*20) and c.519dupT, p.(Glu174*). In addition, four further crisprant cell lines were characterised by Sanger sequencing, presented in Table 5. Electropherograms are presented in Table 6. Three clones (C62, C78, C85) had bi-allelic variants identified, while C40 had one heterozygous variant. The heterozygous cell line C40 was also characterised by western blotting and high content imaging alongside C16 (see Supplementary Material Figure S4 - Characterization of *TMEM67* crisprant) (thesis section 6.1.3) (Lange et al., 2022). The biallelic lines C62, C78 and C85 have been frozen down for characterisation and application in future projects.

Table 5. Variants identified amongst TMEM67 crisprant RPE-1 cell lines

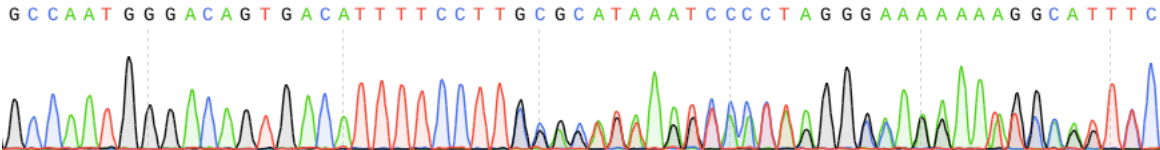
Clone number	Allele 1			Allele 2		
	Nucleotide Change	Protein change	Predicted molecular consequence	Nucleotide change	Protein change	Predicted molecular consequence
C16	c.519delT	p.Cys173Trpfs*20	Nonsense mediated decay (NMD)	c.519dupT	p.Glu174*	NMD
C40	N/a (wild type)	N/a	None	c.369delC	p.Glu124Lysfs*12	NMD
C62	c.364delA	p.Thr122fs*14	NMD	c.369_370delGC	p.Glu124Argfs*17	NMD
C78	c.509-9insT	n/a (intronic)	Unknown	c.514_515delGC	p.Arg172Metfs*2	NMD
C85	c.516_517insT	p.Cys173Leufs*2	NMD	c.507_532delGT GCGTCCGATG TGAGCCAACATTT G	P.Arg169Serfs*2	NMD

Table 6. Electropherograms showing mutations generated amongst crispant RPE-1 cell lines.

Clone number	gRNA region	target	Electropherogram Note: highlighted sequence = first 10 bases of gRNA target Exon 5: TAACAAATGTTGGCTCACAT Exon 3 (reverse complement): TCGGCAGTTAAGTCACTAGG
C21 (mock-transfected wild type control)	Exon 5		
C16	Exon 5		
C78	Exon 5		
C85	Exon 5		 (Note: only have 4 bases of guide sequence highlighted before mutation starts)
C21 (mock-transfected wild type control)	Exon 3		
C40	Exon 3		 Note: only have 8 bases of guide sequence highlighted before mutation starts)

C62

Exon 3



Note: last "T" before mutation starts corresponds to last letter of exon 3 CRISPR guide reverse complement. Highlighted region is encompassed within the mutation.

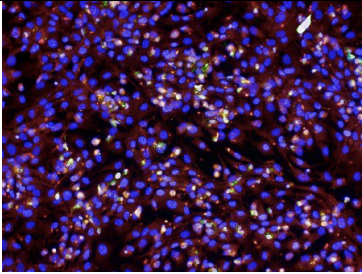
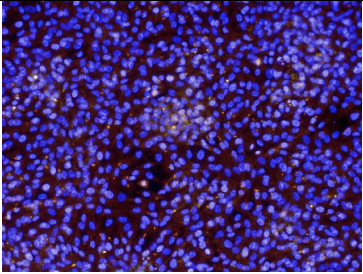
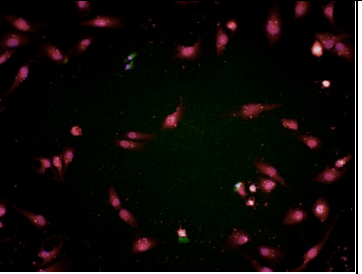
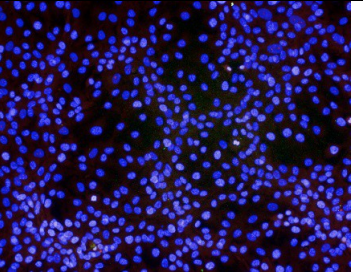
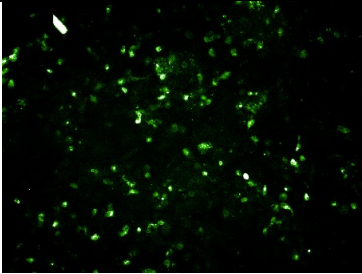

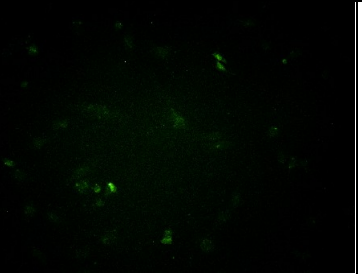
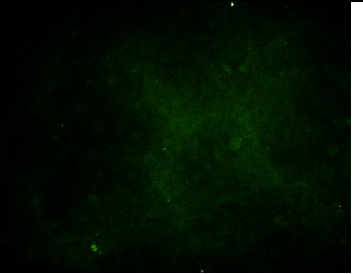
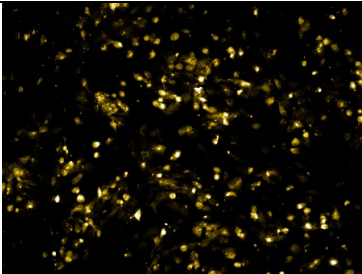
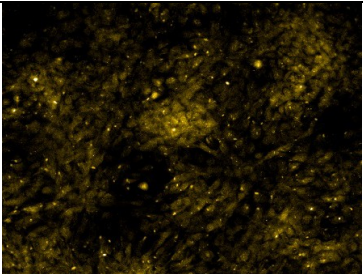
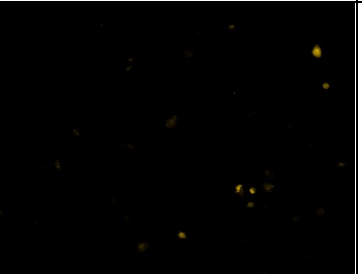
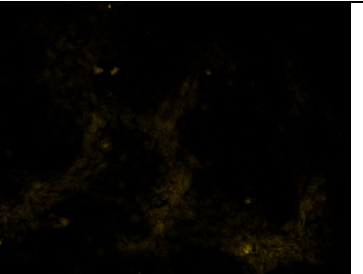
4.3.2 Attempt at variant interpretation by high-content imaging

The original plan was to undertake *TMEM67* variant interpretation through high-content imaging complementation assays. We hypothesised that transfection of the wild-type *TMEM67*-myc plasmid into the null C16 *TMEM67* knockout RPE-1 cell line would restore cilia number, measurable via high-throughput analysis on the Columbus™ Image Data Storage and Analysis System. Conversely, we predicted that transfection of pathogenic variant plasmids would not restore cilia number. By testing cilia number on a range of known variant effects (known benign, wild-type, known pathogenic), we predicted that we could develop a high-throughput system to infer the pathogenicity of *TMEM67* VUSs.

Experimental conditions were optimised on wild-type RPE-1 cells then run alongside the C16 cell line to see how that compared. The optimal concentration of *TMEM67*-myc plasmid to allow clear visualisation for high-content imaging without causing significant toxicity was 140ng, transfected in a ratio of 1:3 with PEI. Example Columbus well images comparing transfection of the *TMEM67*_myc wild type plasmid and untransfected controls between wild-type RPE-1 and the C16 knockout line are shown in Table 7. Although some cell loss can be seen in the transfected wild-type RPE-1 compared with negative controls, the cell loss in the transfected C16 cells is visibly much more significant. Cell loss as measured by the effect on whole cell number by the Columbus recognition protocol was statistically significant for C16 cells transfected with *TMEM67*_myc_WT plasmid (robust z-score -3.85).

Further optimisation experiments were conducted, reducing the concentration of *TMEM67*_myc_WT plasmid transfected to reduce toxicity, but this made detection of transfected cells by the automated software very challenging.

Table 7. Example Columbus well views comparing cells transfected with TMEM67_myc wild-type plasmid and un-transfected cells in both wild-type RPE-1 and the TMEM67 knockout RPE-1 cell line C16

Cell line	Wild-type RPE (passage 26)		TMEM67 knockout C16 RPE-1 (passage 30)	
Channel	140ng TMEM67_myc wild type transfection	Un-transfected control	140ng TMEM67_myc wild type transfection	Un-transfected control
Merge: Red: TOTO-3 (cytoplasm) Blue: DAPI (nucleus) Green: C-myc (transfected cells) Gold: ARL13B (cilia)				
Green: C-myc (transfected cells)				
Gold: ARL13B (cilia)				

Note: some bleed through between the green and gold channels is noted, particularly seen in wild-type RPE-1 cells transfected with TMEM67_myc. However, given the toxic effects of TMEM67_myc_WT plasmid transfection observed, further optimisation to reduce bleed through was not undertaken.

4.4 Conclusion

We concluded that knocking out both alleles of *TMEM67* made the C16 cell line too fragile to tolerate the insult of transfection of plasmids at high enough concentrations to allow the high-content imaging assay to work. Furthermore, restoration of cilia number was never observed in C16 cells transfected with *TMEM67_myc_WT* plasmid. Without this fundamental step, we could not continue to pursue this complementation high-content imaging strategy to interpret *TMEM67* VUS. Therefore, we changed tactics, developing the successful functional cell-signalling assay published in the accompanying manuscript (thesis section 4.5) (Lange et al., 2022).

4.5 Manuscript

OXFORD

Human Molecular Genetics, 2022, Vol. 31, 10, 1574–1587

<https://doi.org/10.1093/hmg/ddab344>

Advance access publication date 20 November 2021

Original Article

Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*

Karen I. Lange^{1,*}, Sunayna Best², Sofia Tsiropoulou¹, Ian Berry³, Colin A. Johnson² and Oliver E. Blacque^{1,*}

¹School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, Ireland

²Division of Molecular Medicine, Leeds Institute of Medical Research, University of Leeds, Leeds, West Yorkshire, UK

³Bristol Genetics Laboratory, Pathology Sciences, Southmead Hospital, Bristol BS10 5NB, UK

*To whom correspondence should be addressed: School of Biomolecular and Biomedical Science, University College Dublin, Belfield, Dublin 4, D04 V1W8, Ireland. Tel: +353 -1-716-6953; Fax: +353-1-716-6701; Email: karen.lange@ucd.ie, oliver.blacque@ucd.ie

Abstract

Better methods are required to interpret the pathogenicity of disease-associated variants of uncertain significance (VUS), which cannot be actioned clinically. In this study, we explore the use of an animal model (*Caenorhabditis elegans*) for *in vivo* interpretation of missense VUS alleles of *TMEM67*, a cilia gene associated with ciliopathies. CRISPR/Cas9 gene editing was used to generate homozygous knock-in *C. elegans* worm strains carrying *TMEM67* patient variants engineered into the orthologous gene (*mks-3*). Quantitative phenotypic assays of sensory cilia structure and function (neuronal dye filling, roaming and chemotaxis assays) measured how the variants impacted *mks-3* gene function. Effects of the variants on *mks-3* function were further investigated by looking at MKS-3::GFP localization and cilia ultrastructure. The quantitative assays in *C. elegans* accurately distinguished between known benign (Asp359Glu, Thr360Ala) and known pathogenic (Glu361Ter, Gln376Pro) variants. Analysis of eight missense VUS generated evidence that three are benign (Cys173Arg, Thr176Ile and Gly979Arg) and five are pathogenic (Cys170Tyr, His782Arg, Gly786Glu, His790Arg and Ser961Tyr). Results from worms were validated by a genetic complementation assay in a human *TMEM67* knock-out hTERT-RPE1 cell line that tests a *TMEM67* signalling function. We conclude that efficient genome editing and quantitative functional assays in *C. elegans* make it a tractable *in vivo* animal model for rapid, cost-effective interpretation of ciliopathy-associated missense VUS alleles.

Introduction

Exome and genome sequencing have revolutionized our ability to identify the genetic causes of disease, interrogate disease mechanisms and pinpoint gene targets for therapy. Missense variants (single codon altered to encode a different amino acid) are the most numerous class of protein-altering variants (1) but only a subset are associated with disease (2). Based on disease features and patterns of inheritance, identified variants are classified as benign, likely benign, uncertain significance, likely pathogenic or pathogenic (as defined by the Association for Clinical Genomic Science; ACGS) (3). For novel or previously uncharacterized variants, the only evidence available to assess their pathogenicity is population allele frequency and analysis by *in silico* tools (e.g. SIFT/PolyPhen/CADD), which are not sufficient to meet the threshold for a 'likely pathogenic' classification according to best practices established by ACGS (3). A VUS (variant of uncertain significance, VUS) classification is made when there is insufficient evidence to conclude on pathogenicity (3,4). Currently, most missense variants are classified as VUS (118 864/206 594 = 57.5%, accessed from ClinVar (5) November 2021). Since VUS

designations cannot be acted upon clinically, a VUS classification can delay or prohibit accurate disease management and/or genetic counselling, and prevent patients from accessing gene-specific therapies and clinical trials (6,7). Given the pressing clinical need to reclassify VUS as benign or pathogenic, it is clear that new effective experimental strategies for VUS interpretation are required (8). With the emergence of advanced genetics tools such as CRISPR-Cas9 gene editing, non-rodent model organisms such as zebrafish, *Drosophila* and *C. elegans* are emerging as robust *in vivo* experimental platforms for functional interpretation of variant pathogenicity (9–12).

Ciliopathies are a heterogeneous group of at least 25 inherited disorders with clinically overlapping phenotypes, caused by pathogenic variants in >200 genes (13,14). Ciliopathies affect many organ systems, causing a broad range of clinical phenotypes of varied severity and penetrance that include cystic kidneys, retinal dystrophy, bone abnormalities, organ laterality defects, respiratory tract defects, infertility, obesity, neurodevelopmental defects and cognitive impairment (15). Due to the extreme heterogeneity of ciliopathy phenotypes, it can be difficult to accurately diagnose ciliopathies.

Received: October 4, 2021. Revised: November 16, 2021. Accepted: November 17, 2021

© The Author(s) 2021. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

For example, in the UK 100 000 genomes project, >20% of patients recruited in the ciliopathy cohort were subsequently diagnosed with non-ciliopathy disorders (14). Several gene therapies targeting ciliopathy genes are currently in development (16–19), highlighting the need to increase accurate genetic diagnoses for these disorders. Ciliopathies are caused by defects in cilia which are 2–20 micron-long microtubule-based organelles that extend from the surfaces of most cell types. Motile cilia propel cells through a fluid or push fluid across a tissue surface. Primary cilia act as cellular ‘antennae’ (20), transducing a wide variety of extrinsic chemical and physical (e.g. light, odorants) signals into the cell (21). Primary cilia are also especially important for coordinating cell–cell communication signalling pathways (e.g. Shh, Wnt, PDGF- α) that are essential for development and homeostasis (22).

The nematode *Caenorhabditis elegans* (*C. elegans*) is a leading model for investigating cilia biology, with many ciliopathy genes and associated pathways conserved in worms (23,24). In *C. elegans*, primary cilia are only found on 60 sensory neurons, extending from the distal tips of dendrite processes. Most nematode cilia are found in the animal’s head and are environmentally exposed via pores in the nematode cuticle (25). Recently, we used CRISPR-Cas9 knock-in technology and quantitative read-outs of gene function to show that pathogenic missense mutations in the Joubert syndrome gene *B9D2* are also pathogenic in the context of the *C. elegans* orthologue (26). Having established this proof-of-principle for modelling ciliopathy variants in worms, we have now examined the use of *C. elegans* for interpreting ciliopathy missense VUS. In this study, we have focussed on *TMEM67* (also called *MKS3*), which is associated with several ciliopathies including Meckel Syndrome (27–30) (OMIM #607361), COACH Syndrome (31) (OMIM #216360) and Joubert Syndrome (7,32–35) (OMIM #610688). Most missense variants reported in *TMEM67* have uncertain clinical significance (74/142 = 52.1%, accessed from ClinVar (5) November 2021), and their abundance makes *TMEM67* an excellent candidate to explore VUS interpretation in *C. elegans*.

TMEM67/MKS3 is a transmembrane protein that functions at the ciliary transition zone (TZ), which corresponds to the proximal-most 0.2–1.0 μm of the ciliary axoneme (27,36). Defined by unique structural features such as Y-linkers that connect the ciliary microtubules with the membrane, the TZ acts as a diffusion barrier, or ‘gate’, to facilitate the cilium as a compartmentalized organelle (37). Indeed, the TZ is a ciliopathy hotspot, with at least two dozen ciliopathy proteins found there (38). Work from multiple model systems, including major input from the nematode system, has identified several ciliopathy-associated genetic and molecular assemblies within the TZ such as the MKS and NPHP modules (39–41). In *C. elegans*, the *TMEM67* orthologue *MKS-3* forms part of the MKS module, along with at least 10 other ciliopathy proteins, whereas the NPHP module consists of just two proteins (NPHP-1, NPHP-4) (38). The relationship

organisms. *Drosophila* has a simplified TZ organization that lacks the NPHP module (42,43). Whereas *C. elegans* MKS and NPHP module genes function redundantly to regulate cilia and TZ formation (44), this is typically not the case in vertebrates and mammals where loss of individual MKS or NPHP module components, such as *TMEM67*, results in mild to severe ciliogenesis defects and lethality in some cases (45–52). Although the MKS and NPHP module genes are not strictly redundant in vertebrates, they do genetically interact (48,53).

Here, we used CRISPR-Cas9 technology to engineer eight *TMEM67* missense VUS at the orthologous position in the *C. elegans* *mks-3* orthologue. Using quantifiable assays of cilium structure and sensory function, as well as protein localization at the TZ, we determined that three of the variants are benign and five are damaging. We then validated the worm findings using a genetic complementation-based approach in *TMEM67* null human cells. Our study indicates that *C. elegans* is a tractable model system that can provide evidence of pathogenicity for ciliopathy-associated variants.

Results

Selection of *TMEM67* variants for analysis

TMEM67 variants were selected using two criteria: (i) conservation of the mutated amino acid in the worm orthologue (*MKS-3*) (Supplementary Material, Fig. S1), and (ii) presence of an adjacent Cas9 PAM site in the *C. elegans* genome to facilitate CRISPR gene editing. Using these criteria, we modelled eight missense *TMEM67* VUS in *C. elegans* *mks-3* (Fig. 1A, Supplementary Material, Table S1). We also included two known benign and two pathogenic variants as controls. Benign1(Asp359Glu) and VUS4 (His782Arg) were identified on the Ensembl variation database (54). Benign2(Thr360Ala), Pathogenic1 (Glu361Ter), Pathogenic2(Gln376Pro), VUS2(Cys173Arg) and VUS5(Gly786Glu) were identified on ClinVar (5). VUS1(Cys170Tyr), VUS3(Thr176Ile), VUS6(His790Arg), VUS7(Gly979Arg) and VUS8(Ser961Tyr) were identified from clinical exome sequencing of Meckel syndrome fetuses. All variants analysed in this study are recessive alleles. For simplicity, we refer to the variants using a shorthand notation (e.g. Benign1, Pathogenic1, VUS1). A *de novo* protein structure prediction program, Raptor X, revealed that the human and worm *TMEM67* proteins show remarkable similarity in their overall predicted domain organization and secondary structure (Fig. 1B). The targeted pathogenic, benign and VUS residues are present in comparable regions of secondary structure (Fig. 1B). For example, the known benign variant residues are in exposed loops and the known pathogenic residues are buried in β -sheets (Fig. 1B, Supplementary Material, Fig. S2).

Quantitative phenotypic analysis of *mks-3* VUS alleles in *C. elegans*

We employed a CRISPR/Cas9 genome editing strategy to

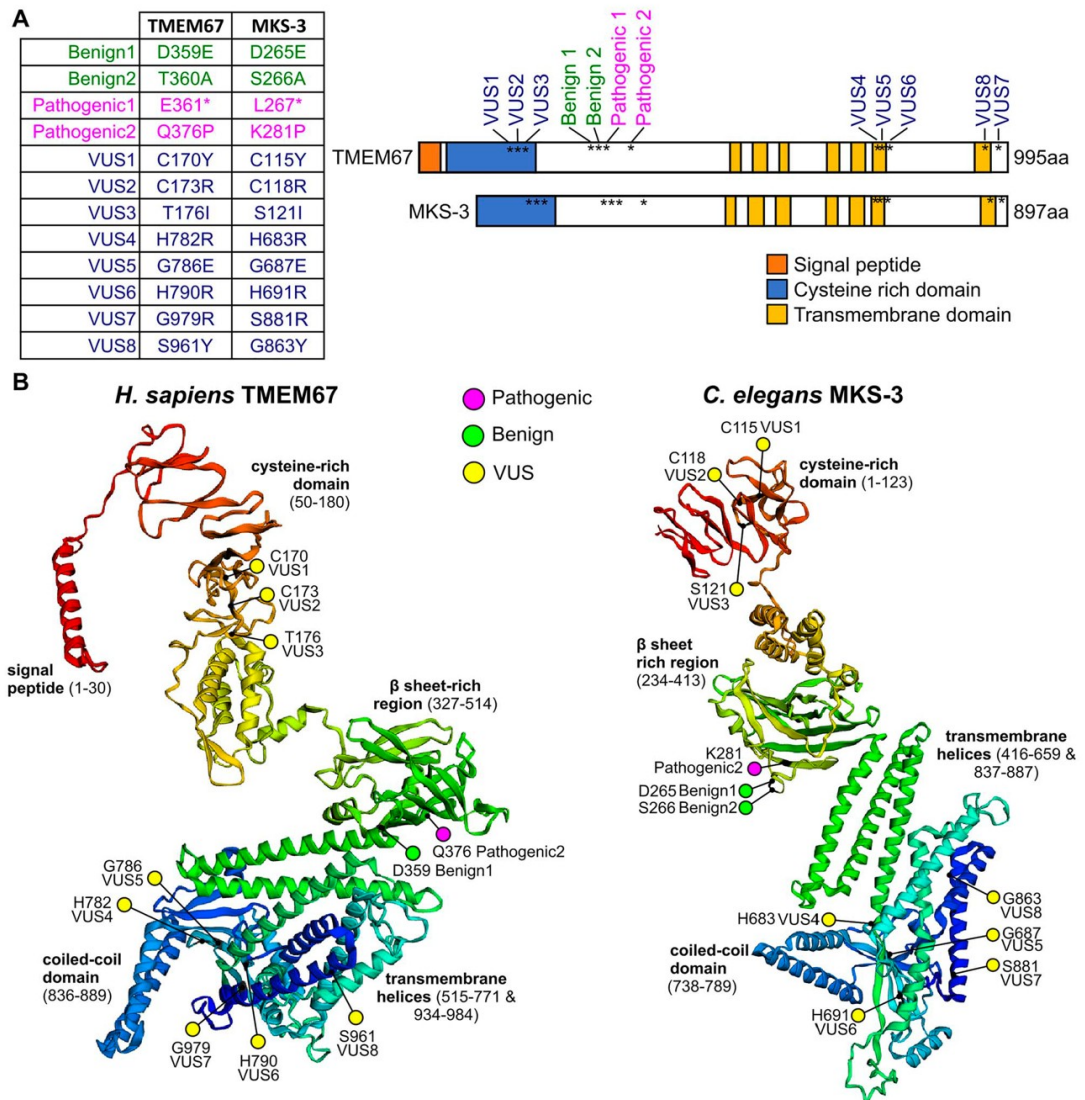


Figure 1. *TMEM67* variants analysed in this study. (A) Twelve variants in *TMEM67*/*mks-3* were generated by CRISPR-Cas9 gene editing and characterized in *C. elegans*. The schematic shows the relative positions of the variants along the length of the proteins. The domains are conserved between humans and worms, but the worm protein lacks the N-terminal signal peptide. (B) RaptorX protein structure and domain organization predictions for the full-length *TMEM67* and *MKS-3* proteins. RaptorX is a deep learning algorithm that predicts secondary and tertiary structures of proteins without close homologues or known structures in the protein data bank. Ribbon diagrams of proteins are rainbow-coloured (red at N-terminus to dark blue at C-terminus) with variants indicated (magenta—known pathogenic; yellow—VUS; green—known benign) within the predicted protein domains. Note that transmembrane helices 1–5 (green to cyan) are followed by a single coiled-coil domain (light blue) and then two C-terminal transmembrane helices 6–7 (dark blue).

many TZ proteins belong to two genetically redundant entities termed the ‘MKS’ and ‘NPHP’ modules that regulate cilia formation and function (44). Relatively minor cilia-dependent phenotypes are observed in *mks-3* and *nphp-4* single mutants (such as a subtle chemotaxis defect), whereas severe cilia defects are observed in *mks-3*; *nphp-4* double mutants (36,55,56). Since *mks-3* functions redundantly with NPHP module genes, we generated the *mks-3* knock-in variants in an

nphp-4(tm925) mutant background to facilitate phenotypic analysis. *nphp-4(tm925)* is a 1109 bp deletion, subsequently referred to as *nphp-4(Δ)*. In phenotypic assays, double mutant *mks-3(variant)*; *nphp-4(Δ)* phenotypes were compared to *mks-3(+)*; *nphp-4(Δ)* (wild-type, positive control) and *mks-3(Δ)*; *nphp-4(Δ)* (949 bp deletion of *mks-3*, negative control). We hypothesized that pathogenic *mks-3* patient alleles would be phenotypically similar to the *mks-3(Δ)* allele.

To assess cilia structure and function we performed three quantitative assays: neuronal dye filling, roaming/foraging and chemotaxis. The dye filling assay indirectly assesses the structural integrity of cilia that are exposed to the environment via their location in head and tail sensory pores (57,58). Specifically, we assessed lipophilic dye (DiI/DiO) uptake in the four ciliated phasmid sensory neurons in the tail. Wild-type and *mks-3(+); nphp-4(Δ)* positive controls display robust dye filling, whereas the *mks-3(Δ); nphp-4(Δ)* negative control is dye filling defective (Fig. 2A). Benign1 and Benign2-containing strains show robust dye uptake whereas strains with Pathogenic1 and Pathogenic2 are defective (Fig. 2A). For strains with the VUS alleles, five cause a severe dye filling defect (VUS1/4/5/6/8), whereas three (VUS2/3/7) do not (Fig. 2A).

C. elegans foraging behaviour is dependent on sensory cilia (57,59). A single young adult worm is placed on a lawn of bacteria for 20 h and the extent of its roaming across the plate is quantified (Fig. 2B). *mks-3(+); nphp-4(Δ)* positive control worms show a slight decrease in roaming compared to wild-type worms, whereas *mks-3(Δ); nphp-4(Δ)* negative controls exhibit a severe roaming defect (Fig. 2B). As expected, Benign1 and Benign2-containing strains exhibit normal roaming behaviour, whereas Pathogenic1 and Pathogenic2-containing worm strains are roaming defective (Fig. 2B). For strains with the VUS alleles, those with VUS4/5/6/8 exhibit a roaming defect, whereas those with VUS1/2/3/7 are roaming normal (Fig. 2B).

C. elegans chemotaxis towards benzaldehyde is also dependent on sensory cilia (57,60). A population of 50–300 worms is placed in the centre of a plate, equidistant from spots of control (ethanol) and benzaldehyde (1:200 in ethanol). *mks-3(+); nphp-4(Δ)* positive control worms show a slight reduction in chemotaxis compared to wild-type, whereas *mks-3(Δ); nphp-4(Δ)* negative controls exhibit a chemotaxis defective phenotype (Fig. 2C). As expected, Benign1 and Benign2-containing strains are chemotaxis normal whereas Pathogenic1 and Pathogenic2-containing strains are defective (Fig. 2C). For the VUS-containing strains, those with VUS1/4/5/6 show a severe chemotaxis defect, those with VUS3/8 exhibit an intermediate phenotype, and those with VUS2/7 are chemotaxis normal (Fig. 2C).

To derive a predictive 'interpretation' score for the VUS alleles, we integrated the results from the three phenotypic assays into a single value (equal weighting; averages normalized to the *mks-3(+); nphp-4(Δ)* positive control; maximum score of 1.0 per assay) (Fig. 2D). A score of <2.5 is considered pathogenic. The Benign1 and Benign2 variants score similarly to the *mks-3(+); nphp-4(Δ)* positive control, whereas Pathogenic1 and Pathogenic2 variants score similarly to the *mks-3(Δ); nphp-4(Δ)* negative control (Fig. 2D). Strains with VUS2, VUS3, or VUS7 received high scores comparable to the benign variants, whereas those with VUS1, VUS4, VUS5, VUS6 or VUS8 received scores comparable to the

pathogenic variants. Therefore, in *C. elegans*, we conclude that VUS2/3/7 are benign variants and VUS1/4/5/6/8 are pathogenic variants.

Effect of VUSs on TZ localization of MKS-3 and cilia ultrastructure

To provide further insight into the damaging or benign nature of the TMEM67 VUS alleles, we first examined the effect of the variants on the subcellular localization of MKS-3. In *C. elegans* sensory neurons, transmembrane MKS-3 localizes to the ciliary TZ, which corresponds to the most proximal ~1 μm of the ciliary axoneme adjacent to the basal body (36). Fusion PCR was used to generate linear *mks-3::gfp* fragments (Supplementary Material, Fig. S3A) containing each of the engineered variants. Constructs were then expressed as extrachromosomal arrays in *C. elegans*, and a fluorescent lipophilic red dye (DiI) employed to co-stain the ciliary membrane. Pathogenic1 was excluded from this analysis because it is a nonsense allele with a premature stop codon. As expected, MKS-3(+):GFP, Benign1::GFP, and Benign2::GFP exhibit very specific TZ localization in the sensory neurons (Fig. 3A, Supplementary Material, Fig. S3B). In contrast, Pathogenic2::GFP showed no detectable fluorescence at the TZ, consistent with the finding that the human Pathogenic2 variant (Q376P) disrupts TMEM67 plasma membrane localization in cell culture (52). Our conclusion that VUS2, VUS3 and VUS7 are benign (Fig. 2D) predicts that the proteins should localize normally. Consistent with this hypothesis, VUS2::GFP, VUS3::GFP and VUS7::GFP display robust TZ localizations in transgenic worms, although Benign2 and VUS7 do show some modest reduction in signal levels (Fig. 3A, Supplementary Material, Fig. S3B). In contrast, VUS1::GFP, VUS5::GFP, VUS6::GFP and VUS8::GFP show no detectable TZ-localization, consistent with these variants being pathogenic. Mislocalized GFP signal elsewhere in the neurons is not observed for these proposed pathogenic VUS suggesting that these variant proteins may be misfolded, unstable and/or degraded. Interestingly, despite a predicted pathogenic classification (Fig. 2D), VUS4::GFP was TZ-localized in most transgenic worms, although signal levels were reduced by ~50% (Fig. 3A, Supplementary Material, Fig. S3B). This observation highlights that TZ-localization alone is not sufficient to interpret pathogenicity.

Using serial section transmission electron microscopy (TEM), we also assessed the effects of the TMEM67 VUS variants on the ultrastructure of amphid neuron cilia and TZs in the nose of the worms. Specifically, we analysed cross-sections of one predicted pathogenic VUS (*mks-3(VUS1)*), one predicted benign VUS (*mks-3(VUS2)*), along with positive (*mks-3(+)*) and negative (*mks-3(Δ)*, *mks-3(Pathogenic2)*) controls. All strains were also homozygous for the *nphp-4(Δ)* allele. Wild-type amphidial pores contain 10 rod-shaped ciliary axonemes emanating from the dendritic tips of eight sensory neurons (two neurons possess a pair of rods); each

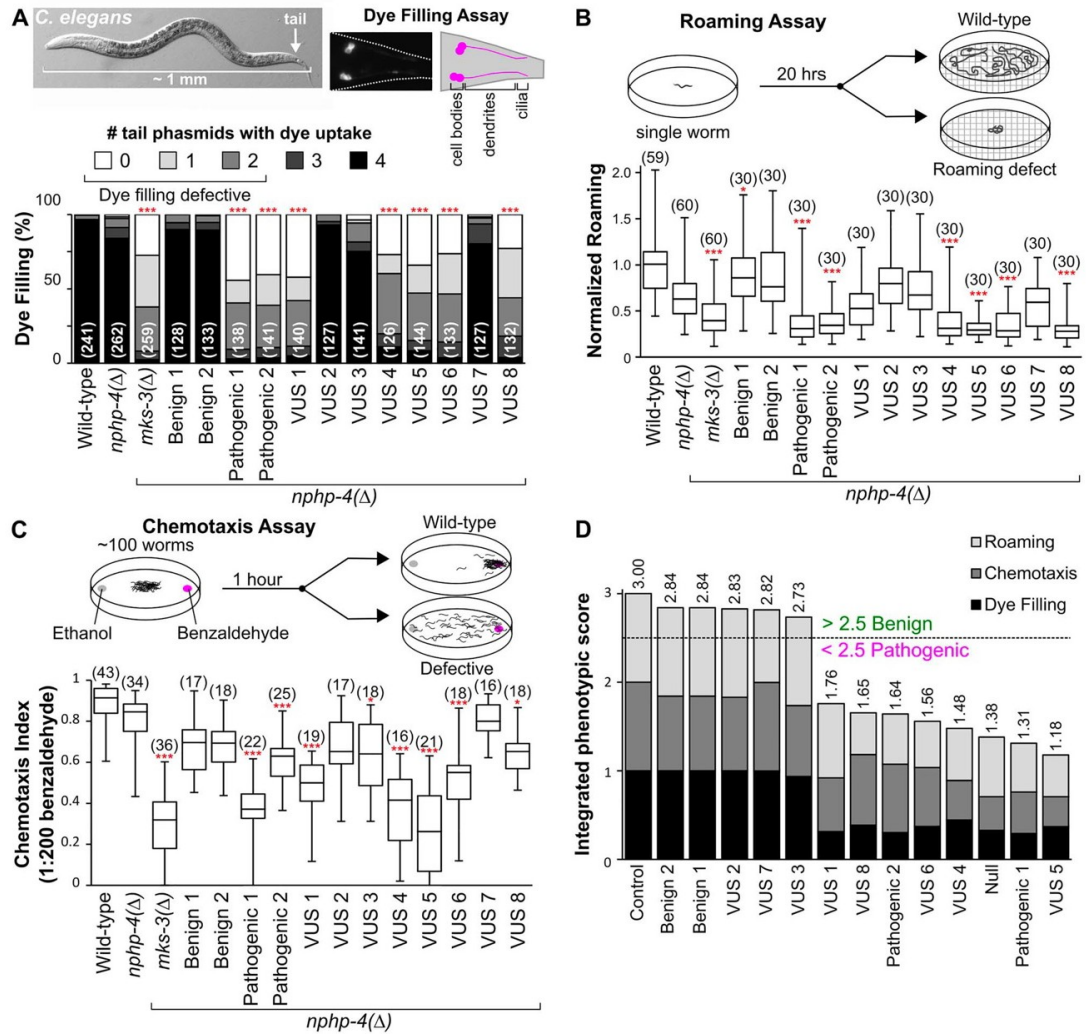


Figure 2. Quantitative phenotyping of cilia-dependent phenotypes in *C. elegans*. All assays were performed blind with at least three independent biological replicates. (A) Lipophilic dye (DiI or DiO) filling assay of the four phasmid (tail) neurons. The number of cell bodies which uptake dye was counted (values range from 0 to 4). The bar graph indicates the proportion of the population with dye uptake in 0 (white) to 4 (black) phasmid neurons. The number of worms is shown in brackets. Statistical significance according to Kruskal–Wallis followed by Schaich–Hammerle *post hoc* test. (B) Assessment of worm roaming behaviour normalized to wild-type. A single young adult hermaphrodite was placed on a food-rich plate for 20 h and the roaming activity was quantified. The number of worms is shown in brackets. Box plots indicate the maximum and minimum values (bars), median, lower quartile and upper quartile. Statistical significance according to Kruskal–Wallis followed by Dunn's *post hoc* test. (C) Quantification of worm chemotaxis towards benzaldehyde after 60 min. Assay is performed on a population of 50–300 worms. The number of assays is shown in brackets. Statistical significance according to ANOVA followed by Tukey's *post hoc* test. Box plots indicate the maximum and minimum values (bars), median, lower quartile and upper quartile. * and *** refer to *P*-values of <0.05 and <0.001, respectively. (D) Integration of the phenotypic results from the three quantitative assays (panels A–C) into one value. Averages from each assay were normalized to the *nphp-4*(Δ) control (with a maximum score of 1.0 for each assay) and summed. Values were ranked from highest (benign) to lowest (pathogenic).

axoneme consists of middle (doublet microtubules) and distal (singlet microtubules) segments, and a proximal TZ compartment that emerges from a swelling at the distal dendrite tip called the periciliary membrane compartment (PCMC) (Fig. 3B). Analysis of cross sections taken from the mid region of the pore shows that *nphp-4*(Δ) worms containing *mks-3*(+) or *mks-3*(VUS2) display an almost full complement of 10 cilia. In contrast, at least two axonemes are missing in the corresponding

cross sections of *nphp-4*(Δ) worms with *mks-3*(Δ), *mks-3*(Pathogenic2) or *mks-3*(VUS1), indicating that some axonemes are either truncated or missing entirely in these strains (Fig. 3C). In cross-sections of the TZ and PCMC regions, ~50% of the TZs of *nphp-4*(Δ) worms with *mks-3*(+) or *mks-3*(VUS2) show a wild-type phenotype, where the TZ membrane and microtubules are in close apposition, connected by electron dense Y-linkers (Fig. 3B). In contrast, *nphp-4*(Δ) worms with *mks-3*(Δ),

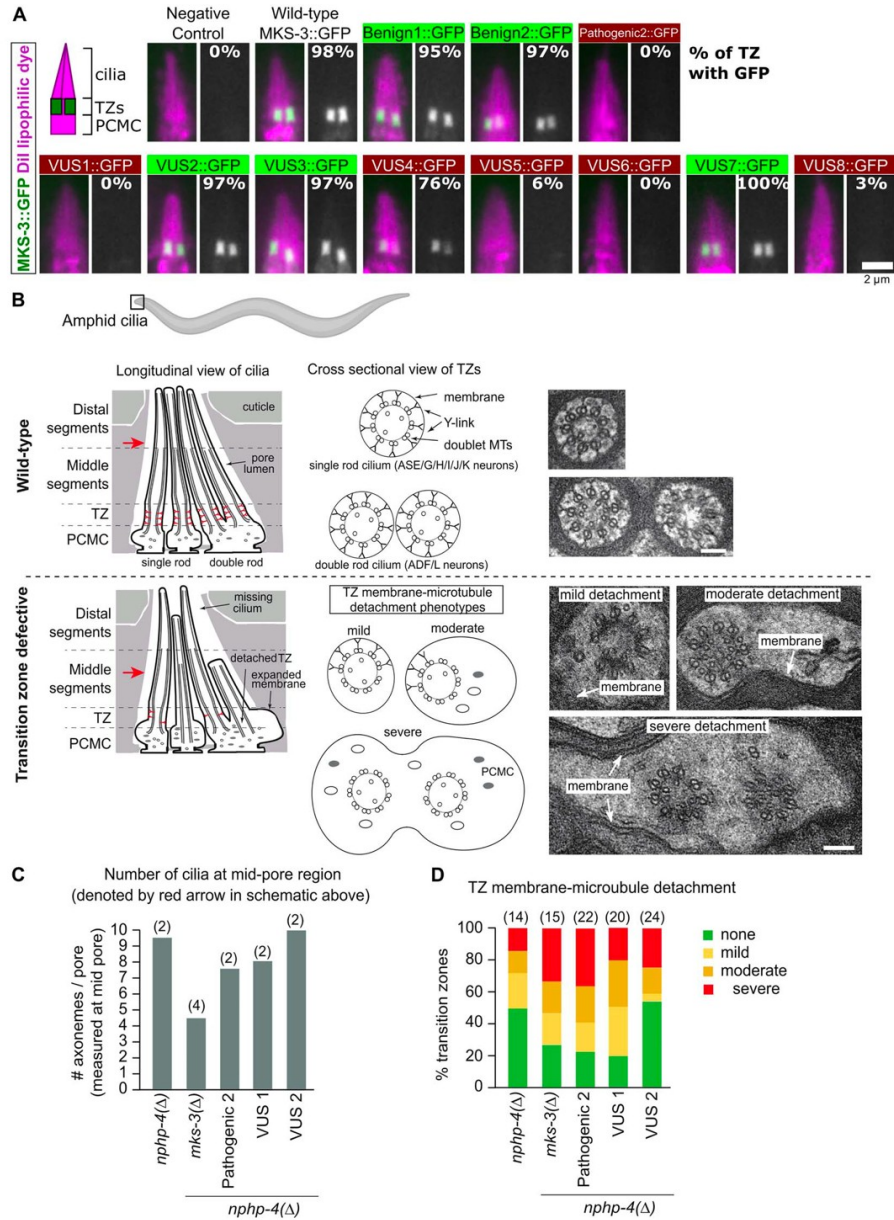


Figure 3. Testing *mks-3* variant predictions via analysis of GFP reporter localization and cilia/transition zone ultrastructure. (A) Transgenic worm strains containing *MKS-3::GFP* extrachromosomal arrays were generated in the *mks-3(Δ)* genetic background. Dil (magenta) lipophilic dye stains the cilia and periciliary membrane compartment (PCMC). Grayscale image shows the green channel alone. GFP levels from 34 to 40 transition zone (TZ) pairs were quantified for each variant. The percentage of cilia with *MKS-3::GFP* localized to the TZ is indicated. Scale bar is 2 μ m. (B–D) Ultrastructure of amphid channel ciliary axonemes and the TZ compartment. TEM images in (B) show examples of wild-type and disrupted TZs, in cross section, for neurons with single and double rod cilia (scale bar is 100 nm). Disrupted TZs show loss of Y-links, resulting in varying degrees of microtubule detachment from the membrane, along with frequent expansion of the membrane. Mild—slight detachment of TZ microtubules from the membrane, which is expanded to a small degree; moderate—many TZ microtubules are detached from the membrane, which is extensively expanded, indicating ectopic docking of the TZ within the PCMC cytoplasm. Schematics show the amphid channel pore and cilia in longitudinal orientation (only 4 of the 10 ciliary axonemes are shown for simplicity), and TZs in radial orientation (cross-section), and indicate the phenotypes shown in the TEM images. The histogram in (C) shows the mean number of cilia observed in electron micrographs from cross sections of the mid-pore region (red arrows in B) taken from the indicated genotypes (number of pores analysed shown in brackets). The chart in (D) shows the quantification of the TZ membrane-microtubule detachment phenotypes for the indicated genotypes (number of TZs analysed shown in brackets).

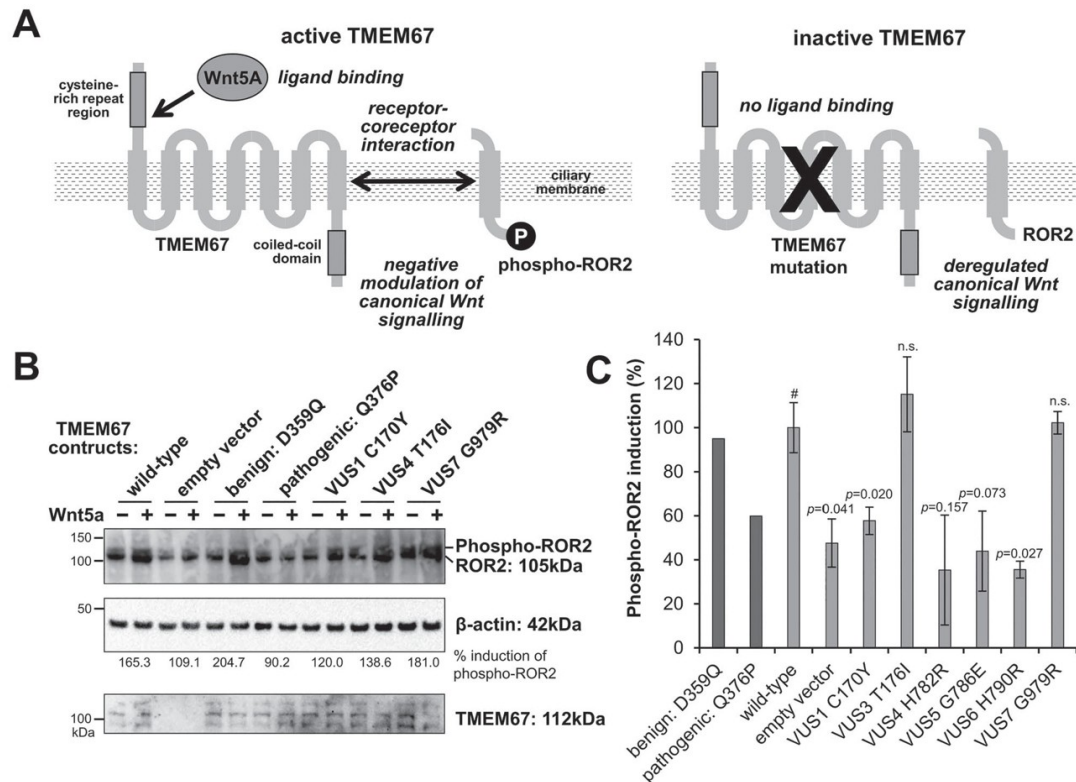


Figure 4. Validation of *C. elegans* predictions of TMEM67 VUS pathogenicity in cell culture. (A) Schematic summarizes the genetic complementation assay in hTERT-RPE1 TMEM67 knock-out cells. Left: in the presence of TMEM67, phosphorylation of the co-receptor ROR2 is stimulated by exogenous treatment with the non-canonical ligand Wnt5a in comparison to control treatment. Right: if TMEM67 is lost or disrupted, ROR2 phosphorylation is not stimulated by this treatment. (B) Western blots of ROR2, with upper phosphorylated isoform indicated (top panel), following transfection and expression (bottom panel) of TMEM67 constructs (wild-type, empty vector negative control, known benign variant control, known pathogenic variant control, and a selection of VUS alleles). Transfected TMEM67^{-/-} knock-out cells were treated with control conditioned medium (-) or Wnt5a-conditioned medium (+). Loading control for normalization is β -actin. (C) Densitometry scans of the phosphorylated ROR2 isoform (for $n = 3$ biological replicates) were quantitated in the bar graph for percentage induction of Wnt5a-stimulated response compared to control response, normalized against responses to the wild-type TMEM67 construct. Statistical significance was determined in pairwise t-tests with wild-type (#) for a minimum of $n = 3$ biological replicates. P-values are listed, n.s., not significant. Error bars indicate standard error of the mean. Statistical significance is not included for Benign(D359Q) and Pathogenic(Q376P) because these values were derived from a single biological replicate.

mks-3(Pathogenic2) or *mks-3(VUS1)* show a more disrupted phenotype, with severe loss of Y-linkers and expansion of the surrounding ciliary/periciliary membrane (Fig. 3D). Together, the ultrastructure data for axoneme number and TZ integrity shows that VUS1 phenocopies the defects observed in the negative controls, whereas VUS2 phenocopies the positive control, thereby confirming the pathogenic and benign nature of VUS1 and VUS2, respectively.

In vitro genetic complementation assay of TMEM67 VUS function in human cell culture

To further validate our findings from *C. elegans*, we utilized an in vitro human cell culture-based assay of TMEM67 function. Previously, we demonstrated that TMEM67 is required for phosphorylation of the ROR2 co-receptor and subsequent activation of non-canonical Wnt signalling (61) (Fig. 4A). Here, we developed a

hTERT-RPE1 crisprant cell-line that has compound heterozygous (biallelic) null mutations in TMEM67 (Supplementary Material, Fig. S4). In the absence of TMEM67, phosphorylation of ROR2 was not stimulated by exogenous treatment with the non-canonical ligand Wnt5a. Transient transfection with full-length wild-type TMEM67 fully rescued ROR2 phosphorylation following Wnt5a treatment (Fig. 4B). These responses allowed us to determine the relative effects of VUS on TMEM67 biological function. In this assay, transfection of Benign1 allowed 204.7% induction of phospho-ROR2 levels by Wnt5a relative to control (Fig. 4B). In contrast, Pathogenic2 did not rescue biological function (90.2% induction) (Fig. 4B). Comparison of all VUS, normalized to wild-type TMEM67 responses across three independent biological replicates, enabled us to interpret VUS1, VUS4, VUS5 and VUS6 as pathogenic, and VUS3 and VUS7 as benign (Fig. 4C). VUS2 and VUS8 were not tested in

in *C. elegans* are corroborated by similar findings in mammalian cells.

Discussion

Ciliopathies are multisystem disorders that affect many organs including kidneys, liver and retina. Although the organ systems affected by cilia dysfunction are not present in *C. elegans*, the basic biology of primary cilia is shared across species. Furthermore, despite undoubted context-specific distinctions, the cilia proteins themselves are functionally conserved (23,24), therefore allowing us to model and characterize patient missense variants in worms.

In this study, we exploited efficient genome editing and quantitative phenotypic analysis of cilium structure and function in *C. elegans* to determine the pathogenicity of TMEM67 variants. This approach accurately classified known pathogenic and known benign variants. We also generated a pathogenicity prediction for all eight missense VUS alleles analysed. Three VUS were phenotypically benign (VUS2(Cys173Arg), VUS3(Thr176Ile), VUS7(Gly979Arg)) and five were phenotypically pathogenic (VUS1(Cys170Tyr), VUS4(His782Arg), VUS5(Gly786Glu), VUS6(His790Arg), VUS8(Ser961Tyr)). We validated these predictions using TEM and localization assay data, showing that pathogenic missense mutations abrogate TZ ultrastructure and prevent MKS-3 from localizing at the TZ. The one exception was VUS4, which although severely pathogenic for cilium structure and function, can localize at the TZ, albeit at a reduced level. Thus, the VUS4 patho-mechanism is likely due to loss of function at the TZ (such as disrupting a protein-protein interaction), rather than disruption of upstream MKS-3 trafficking to the compartment. Finally, we validated our nematode-based predictions in a human cell culture assay of TMEM67 function. When taken together, our data show that *C. elegans* can interpret the pathogenicity of VUS and provide evidence towards their reclassification as benign or pathogenic.

Several *in silico* algorithms have been developed to predict the pathogenicity of missense variants. However, their accuracy is inconsistent (62–64). Indeed, we tested five *in silico* tools (MISTIC (65), SIFT (66), Poly-Phen (67), CADD (68) and REVEL (69)) and found that they return deleterious/damaging predictions for all eight of the VUS examined in this project (Supplementary Material, Fig. S5). The only exceptions are VUS1/3/8, where one or two of the tools returned non-deleterious or intermediate scores (Supplementary Material, Fig. S5). Therefore, the algorithm predictions do not align with our observations in *C. elegans* and cell culture experiments. We conclude that *in vivo* modelling of missense variants in *C. elegans* more accurately predicts results in human cells than currently available prediction algorithms.

An additional advantage of using nematodes to interpret genetic variation is the availability of quantitative assays that are suitable for high-throughput analysis. For

example, live animal fluorescence-activated cell sorting (FACS) can be used for assessing cilium structure (via dye filling) and automated worm tracking can be applied to measure cilium function (roaming, chemotaxis) (70–72). Furthermore, machine learning can streamline the analysis of complex datasets to predict VUS pathogenicity (73). Another advantage of the nematode approach is that engineered patient alleles can be used for high-throughput small molecule suppressor screens to identify potential therapeutics. Indeed, *C. elegans* is emerging as an excellent model for whole animal large-scale drug screens and such strategies have already been used to investigate a variety of metabolic and neuromuscular disorders (74–78). Despite these advantages, one limitation to modelling patient variants in worms is that a substantial number of residues mutated in disease are not conserved in the *C. elegans* orthologue. However, a potential solution to this problem is to create 'humanized' worms, where the entirety or specific domains of the nematode orthologue are replaced with the human sequence (79,80). If the human protein retains functionality in the worm context, humanized strains can be used to model all missense human variants in the corresponding gene.

The utility of the nematode approach to interpreting human gene variants goes well beyond TMEM67 and the ciliopathy gene class. Indeed, a humanized nematode model was very recently employed to interpret the pathogenicity of 29 missense VUS in an epilepsy gene (73). Thus, for genes functioning in conserved molecular pathways, worms offer a powerful system to generate *in vivo* evidence towards reclassifying VUS as pathogenic or benign. With ever increasing throughput in generating the knock-in alleles, *C. elegans* can therefore make a significant contribution to interpreting the huge numbers of VUS deposited in ClinVar, and bring us closer to the ambition that the clinical relevance of all encountered genomic variants will be more readily predictable (81).

In summary, this study highlights that *C. elegans* is a practical model for variant interpretation of ciliary genes. Analysis of ciliopathy-associated VUS in *C. elegans* is accurate, quick, affordable and easily interpretable. Although this study focussed on TMEM67, we anticipate that VUS alleles of any conserved cilia genes can be modelled and characterized in *C. elegans* using the approach described here.

Materials and Methods

Modelling of protein secondary structure

Human TMEM67 (NP_714915.3) and *C. elegans* MKS-3 (NP_495591.2) protein sequences were analysed by the RaptorX protein structure prediction server, using default settings for the deep dilated convolutional residual neural networks method (82). Absolute model quality was assessed by ranking Global Distance Test (GDT) scores defined as $1 \times N(1) + 0.75 \times N(2) + 0.5 \times N(4) + 0.25 \times N(8)$, where $N(x)$ is the number of residues with estimated

modelling error (in Å) smaller than x , divided by protein length and multiplied by 100. GDT scores >50 indicate a good quality model. However, the highest ranking models for TMEM67 (GDT=28.968) and MKS-3 (GDT=19.269) suggest that portions of these models are lower quality. Models in the .pdb format were visualized and annotated in EzMol (<http://www.sbg.bio.ic.ac.uk/ezmol/>).

C. *elegans* maintenance

All *C. elegans* strains in this study were maintained at 20°C or 15°C on nematode growth medium (NGM) seeded with OP50 *E. coli* using standard techniques (83). Young adult hermaphrodites were synchronized by selecting L4 larvae and incubating at 20°C for 16–20 h or by alkaline hypochlorite treatment of gravid hermaphrodites at 20°C ~65–70 h before the assay. All worm strains are listed in Supplementary Material, Table S2.

CRISPR/Cas9 to engineer *mks-3* mutants in *C. elegans*

CRISPR protocols were performed as previously described (26) in a *nphp-4(tm925)* genetic background using an *unc-58* co-CRISPR strategy (84). Cas9 enzyme (IDT, #1081058), tracrRNA (IDT, #1072533), and custom synthesized crRNA were obtained from Integrated DNA Technologies. Suitable PAM sites were selected based on Azimuth 2.0 scores (85) and distance from the desired edit (<10 nucleotides). crRNA are listed in Supplementary Material, Table S3. Injection mixes were prepared on ice as follows: 1 μ l crRNA (0.3 nmol/ μ l), 1 μ l tracrRNA (0.425 nmol/ μ l), 0.25 μ l *unc-58* crRNA (1 nmol/ μ l), 0.25 μ l *unc-58* ssODN (500 ng/ μ l), 0.5 μ l each variant specific ssODN (1 μ g/ μ l), 2 μ l 1 M KCl, 0.4 μ l HEPES (200 mM, pH 7.4), 0.2 μ l Cas9 (10 μ g/ μ l) and RNase-free water up to 10 μ l. The injection mix was mixed gently, centrifuged at ~15000 *g* for 2 min, and incubated at 37°C for 15 min before injection. All Unc F1 were screened in pools of three hermaphrodites and engineered alleles were detected with variant specific PCR primers. All primers are listed in Supplementary Material, Table S4. The CRISPR efficiency (defined as the percent of F1 pools that were positive for the edit) varied from 1% to 35% with an average 15%. One CRISPR mutant was isolated and characterized for each variant. Accuracy of the engineered variants was confirmed with Sanger sequencing. The co-CRISPR marker, *unc-58*, was also sequenced and unintended *unc-58* mutations (86) were outcrossed.

C. *elegans* quantitative phenotyping assays

Assays to assess cilia structure and function were performed with young adult hermaphrodites (57). The phenotypic assays were performed blinded to genotype with at least three independent biological replicates. Quantitative dye filling assays were performed with DiO (Invitrogen, D275) and dye uptake of phasmid (tail) neurons was assessed on a wide-field epifluorescence microscope. For each variant, dye filling in 125–145 worms was quantified. Roaming activity of

worms was quantified by placing a single young adult hermaphrodite on a fully seeded NGM plate for 20 h at 20°C. A 5 × 5 mm grid was used to count the number of squares the worm entered. The roaming activity of 30 worms was quantified for each variant. Values were normalized to wild-type (N2) for each replicate. Chemotaxis plates were prepared 16–24 h before the assay was performed (9 cm petri dishes with 10 ml of chemotaxis agar: 2% agar, 5 mM KPO₄ pH6, 1 mM CaCl₂, 1 mM MgSO₄). Two points were marked at opposite sides of the plate 1.5 cm from the edge and 1 μ l of 1M sodium azide (Sigma, S2002) was applied to the spots. Then 1 μ l of ethanol (Honeywell, 32294) or 1:200 benzaldehyde (Sigma, B1334) diluted in ethanol was added to the spots. Young adult hermaphrodites were washed three times in M9 (22 mM KH₂PO₄, 42 mM Na₂HPO₄, 85.5 mM NaCl, 1 mM MgSO₄) and once with deionized water and 50–300 worms were placed in the centre of the plate and excess water was removed. After 1 h the worms were counted. The chemotaxis index was calculated as follows: $(b-c)/n$ where b is the number of worms within 1.5 cm of the benzaldehyde spot, c is the number of worms within 1.5 cm of the ethanol control, and n is the total number of worms on the plate. For each variant a total 15–25 assays were performed.

Integration of phenotypic data to predict variant pathogenicity

The results from the dye filling, roaming and chemotaxis assays were consolidated to generate a value to predict variant pathogenicity. Averages from each assay were normalized to the *nphp-4(A)* control (with a maximum value of 1.0 for each assay). These averages were then summed to generate the integrated phenotypic score. The *nphp-4(A)* control received a score of 3.0 whereas the *mks-3* null allele received a score of 1.38. Variants that scored <2.5 were predicted to be pathogenic.

Generating transgenic worms expressing extrachromosomal MKS-3::GFP

mks-3::gfp transgenes were generated with PCR-based fusion (87) of *mks-3* gDNA (including 485 bp of 5' UTR sequence) with GFP and the *unc-54* 3' UTR (pPD95_77, a gift from Andrew Fire, Addgene plasmid #1495). All primers are listed in Supplementary Material, Table S4. *mks-3(tm2547)* hermaphrodites were injected with 0.25 ng/ μ l *mks-3::gfp* and 100 ng/ μ l *coel::dsRed* (a gift from Piali Sengupta, Addgene plasmid #8938) to generate extrachromosomal arrays (1–7 lines each). PCR was used to confirm the presence of *mks-3::gfp* transgene in the stable extrachromosomal arrays.

C. *elegans* wide-field imaging and quantification of fluorescence

Young adult hermaphrodites were immobilized on 4% agarose pads in 40 mM tetramisole (Sigma, L9756). Images were acquired with a 100× (1.40 NA) oil objective on an upright Leica DM5000B epifluorescence

microscope and captured with an Andor iXon+ camera. Image analysis was performed with FIJI/ImageJ (NIH). MKS-3::GFP fluorescence was quantified as previously described (26). Briefly, a 40 × 40 pixel box was drawn around a TZ pair and the integrated signal intensity was measured. The box size was increased by one pixel in each direction and the signal intensity of this 42 × 42 pixel box was used to calculate the background fluorescence. Background fluorescence was subtracted and values were normalized to wild-type.

Transmission electron microscopy

Young adult hermaphrodites were processed as previously described (57). Briefly, worms were fixed in 2.5% glutaraldehyde (Merck) in Sørensen's phosphate buffer (0.1 M, pH 7.4) for 48 h at 4°C, post-fixed in 1% osmium tetroxide (EMS) for 1 h, and dehydrated through an increasing ethanol gradient. Samples were treated with propylene oxide (Sigma) and embedded in EPON resin (Agar Scientific) for 24 h at 60°C. Serial, ultra-thin (90 nm) sections of the worm nose tissue were cut using a Leica EM UC6 Ultramicrotome, collected on copper grids (EMS), stained with 2% uranyl acetate (Agar Scientific) for 20 min followed by 3% lead citrate (LabTech) for 5 min, and imaged on a Tecnai 12 (FEI software) with an acceleration voltage of 120 kV.

Cell culture

Human hTERT-immortalized retinal pigmentary epithelial (hTERT-RPE1, American Type Culture Collection; ATCC) wild-type and crispant cell-lines were grown in Dulbecco's minimum essential medium (DMEM)/Ham's F12 medium supplemented with GlutaMAX (Gibco #10565018) and 10% foetal bovine serum (FBS). For selected experiments involving cilia, cells following passage were serum-starved in DMEM/F-12 media containing 0.2% FBS. Cells were cultured in an incubator at 37°C with 5% CO₂.

TMEM67 cloning, plasmid constructs and transfections

Full-length *H. sapiens* TMEM67 isoform 1 (RefSeq JF432845, plasmid ID HsCD00505975, DNASU Plasmid Repository) was cloned into pENTR223. The ORF was Gateway cloned into a C-terminal GFP-tagged Gateway pcDNA-DEST47 vector (ThermoFisher Scientific), sequence verified, and sub-cloned into pcDNA3.1 myc/HisA vector with HiFi cloning (New England Biolabs). The construct was also engineered to contain an endogenous Kozak sequence prior to the start site, the first 30 nucleotides of the main transcript that were missing from the DNASU sequence, and a GS linker between the ORF and myc tag. A QuikChange II XL Site-Directed Mutagenesis Kit (Agilent) was used according to the manufacturer's protocol to generate TMEM67 variants. Primer sequences are listed in Supplementary Material, Table S5. The final constructs were verified by sequencing. Cells at 80% confluency were transfected with plasmids using Lipofectamine

CRISPR/Cas9 genome editing in cell culture

GFP-expressing pSpCas9(BB)-2A-GFP (PX458) was a gift from Feng Zhang, Addgene plasmid #48138. Three crRNAs targeting human TMEM67 (RefSeq NM_153704.5) were designed using Benchling (<https://benchling.com>), selected for the highest ranking on- and off-target effects. crRNAs were ordered as HPLC-purified oligos from Integrated DNA Technologies in addition to Alt-R CRISPR-Cas9 tracrRNA-ATTO550 conjugates. crRNA sequences are listed in Supplementary Material, Table S6. Lyophilised pellets were resuspended in Tris-EDTA (TE) buffer (Qiagen) to give 100 mM stocks. crRNA and tracrRNA were mixed (1:1), and incubated in nuclease-free duplex buffer (30 mM HEPES, pH 7.5, 100 mM potassium acetate) to make 300 nM guide RNA master mixes. Before transfecting into cells the crRNA:tracrRNA duplexes were incubated with Lipofectamine 2000 at an RNA:Lipofectamine ratio of 2:1 in 200 µl/well Opti-MEM for 20 min. One millilitres of media from 6-well plate wells was removed, the transfection reagents applied, and the cells incubated overnight. Media was changed to fresh DMEM/F-12 with 10% FBS after 16 h, and cells incubated for 48 h. Following transfection, FACS was performed to enrich cells expressing GFP and to produce clonal populations. Ninety-six well plates (Corning) were treated with 200 µl of 4% bovine serum albumin (BSA) per well for 1 h. Wells were then filled with 100 µl of filter-sterilized collection buffer (20% FCS, 1% penicillin-streptomycin, 50% conditioned media, 29% fresh DMEM/F-12 media). Transfected cells were prepared for FACS by removing media, washing in PBS, and treating with trypsin for 5 min before resuspending in filter-sterilized sorting buffer (1× Ca²⁺/Mg²⁺-free PBS, 5 mM EDTA, 25 mM HEPES pH 7.0). A 70 µm filter was used to disperse cells into 4% BSA-treated polystyrene FACS tubes. A BD Influx 6-way cell sorter (BD Biosciences) was used to index sort GFP-positive cells, calibrated against un-transfected control cells. When an abundance of GFP-positive cells were present, the top 5% were targeted for index sorting. After sorting, cells were incubated for 3 weeks at 37°C with 5% CO₂, with weekly checks for growing colonies.

PCR and sequence validation of crispant cell-lines

To extract DNA from colonies within the 96-well plates, cells were washed with 1× Ca²⁺/Mg²⁺-free PBS and resuspended in 50 µl of DirectPCR Lysis Reagent (Viagen Biotech) containing 0.4 mg/ml Proteinase K (Sigma, # P4850). Suspensions were incubated at 55°C for 5 h, followed by 85°C for 45 min. One microlitre of DNA extracts were used in PCR reactions. Primers are listed in Supplementary Material, Table S7. Variants were identified by Sanger sequencing (GeneWiz Inc.) followed by analysis using the Synthego ICE v2 (88). The following clones were chosen for further study: clone 40, heterozygous for c.369delC (p.Glu124Lysfs*12); and clone 16 carrying biallelic variants [c.519delT]+[c.519dupT]

predicted to result in nonsense mediated decay by the Ensembl Variant Effect Predictor (89). Clone 21 was a negative control cell-line that was mock-transfected, underwent FACS, but was verified to carry wild-type TMEM67.

Whole cell extract preparation and western immunoblotting

Whole cell extracts containing total soluble proteins were prepared from hTERT-RPE1 cells that were transiently transfected with 1.0 μ g plasmid constructs in 90 mm tissue culture dishes, or scaled down as appropriate. Ten micrograms total soluble protein was analysed by SDS-PAGE (4–12% polyacrylamide gradient) and western blotting performed according to standard protocols. Primary antibodies used: mouse anti- β actin (1:10000, clone AC-15, Abcam Ltd., Cambridge, UK); rabbit polyclonal anti-TMEM67 (1:500, 13975-1-AP; Protein-Tech Inc., Rosemont, IL, USA); goat anti-ROR2 (1:1000, AF2064; R&D Systems Inc., Minneapolis, MN, USA). Appropriate HRP-conjugated secondary antibodies (Dako UK Ltd.) were used (final dilutions of 1:10000–25000) for detection by the enhanced chemiluminescence 'Femto West' western blotting detection system (Thermo Fisher Scientific Inc., Rockford, IL, USA) and visualized using a ChemiDoc MP imaging system (BioRad Inc., Hercules, CA, USA). Ratios of active phosphorylated ROR2: unphosphorylated ROR2 isoforms were calculated by quantitating band intensity using ImageLab 5.2.1 software (BioRad Inc.) for three biological replicates, as described previously (61).

Statistical analyses

All *C. elegans* statistical analyses were performed in Microsoft Excel with the Real Statistics Resource Pack Version 7.2 (www.real-statistics.com). A Shapiro–Wilk test determined if data were normally distributed. Statistical significance of normally distributed datasets was determined with an ANOVA followed by Tukey's post hoc (chemotaxis and GFP quantification) or Kruskal–Wallis followed by Dunn's (roaming) for non-parametric datasets. Statistical significance of dye filling was determined using a Kruskal–Wallis followed by Schaich–Hammerle post hoc test using a chi-squared distribution. For cell culture results, a normal distribution of data were confirmed using the Kolmogorov–Smirnov test (GraphPad Prism). Pairwise comparisons were analysed with Student's two-tailed t-test using InStat (GraphPad Software Inc.). *, **, and *** refer to P-values of <0.05, <0.01 and <0.001, respectively.

Supplementary Material

Supplementary Material is available at HMG online.

Acknowledgements

We thank Joseph McNicholl for assistance with construction of the *mks-3::gfp* transgenes. We thank the

University College Dublin Conway Institute imaging facility (Dimitri Scholz, Tiina O'Neill and Niamh Stephens) for assistance with transmission electron microscopy. Some figures were created with BioRender.com.

Conflict of Interest statement. None declared.

Funding

Science Foundation Ireland (SFI) in partnership with the Biotechnology and Biological Sciences Research Council (BBSRC) under grant number 16/BBSRC/3394 (to O.E.B.) and BB/P007791/1 (to C.A.J.). SB acknowledges support from a Wellcome Trust clinical training fellowship (203914/Z/16/Z).

References

- Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P. et al. (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, **581**, 434–443.
- MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B. et al. (2012) A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, **335**, 823–828.
- Ellard, S., Baple, E.L., Berry, I., Forrester, N., Turnbull, C., Owens, M., Eccles, D.M., Abbs, S., Scott, R., Deans, Z.C. et al. (2020) ACGS best practice guidelines for variant classification in rare disease 2020. *Assoc. Clin. Genomic Sci*, **4**, 1–32.
- Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E. et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.*, **17**, 405–424.
- Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W. et al. (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**, D1062–D1067.
- Splinter, K., Adams, D.R., Bacino, C.A., Bellen, H.J., Bernstein, J.A., Cheattle-Jarvela, A.M., Eng, C.M., Esteves, C., Gahl, W.A., Hamid, R. et al. (2018) Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N. Engl. J. Med.*, **379**, 2131–2139.
- Bachmann-Gagescu, R., Dempsey, J.C., Bulgheroni, S., Chen, M.L., D'Arrigo, S., Glass, I.A., Heller, T., Héon, E., Hildebrandt, H., Joshi, H. et al. (2020) Healthcare recommendations for Joubert syndrome. *Am. J. Med. Genet.*, **182**, 229–249.
- Brnich, S.E., Abou Tayoun, A.N., Couch, F.J., Cutting, G.R., Greenblatt, M.S., Heinen, C.D., Kanavy, D.M., Luo, X., McNulty, S.M., Starita, L.M. et al. (2019) Recommendations for application of the functional evidence P53/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. *Genome Med.*, **12**, 3.
- Wangler, M.F., Yamamoto, S., Chao, H.T., Posey, J.E., Westerfield, M., Postlethwait, J., Members of the Undiagnosed Diseases (UDN), Hieter, P., Boycott, K.M., Campeau, P.M. et al. (2017) Model organisms facilitate rare disease diagnosis and therapeutic research. *Genetics*, **207**, 9–27.
- Baldrige, D., Wangler, M.F., Bowman, A.N., Yamamoto, S., Undiagnosed Diseases Network, Schedl, T., Pak, S.C., Postlethwait, J.H.,

- Shin, J., Solnica-Krezel, L. et al. (2021) Model organisms contribute to diagnosis and discovery in the undiagnosed diseases network: current state and a future vision. *Orphanet J. Rare Dis.*, **16**, 206.
11. Kropp, P.A., Bauer, R., Zafra, I., Graham, C. and Golden, A. (2021) *Caenorhabditis elegans* for rare disease modeling and drug discovery: strategies and strengths. *Dis. Model. Mech.*, **14**, dmm049010.
 12. Silverman, G.A., Luke, C.J., Bhatia, S.R., Long, O.S., Vetica, A.C., Perlmutter, D.H. and Pak, S.C. (2009) Modeling molecular and cellular aspects of human disease using the nematode *Caenorhabditis elegans*. *Pediatr. Res.*, **65**, 10–18.
 13. Wheway, G., Mitchison, H.M. and Genomics England Research Consortium (2019) Opportunities and Challenges for Molecular Understanding of Ciliopathies-The 100,000 Genomes Project. *Front. Genet.*, **10**, 127.
 14. Best, S., Lord, J., Roche, M., Watson, C.M., Poulter, J.A., Bevers, R.P.J., Stuckey, A., Szymanska, K., Ellingford, J.M., Carmichael, J. et al. (2021) Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project. *J. Med. Genet.*, Published Online First: 29 October 2021. doi: 10.1136/jmedgenet-2021-108065.
 15. Waters, A.M. and Beales, P.L. (2011) Ciliopathies: an expanding disease spectrum. *Pediatr. Nephrol.*, **26**, 1039–1056.
 16. Aguirre, G.D., Cideciyan, A.V., Dufour, V.L., Ripolles-García, A., Sudharsan, R., Swider, M., Nikonov, R., Iwabe, S., Boye, S.L., Hauswirth, W.W. et al. (2021) Gene therapy reforms photoreceptor structure and restores vision in NPHP5-associated Leber congenital amaurosis. *Mol. Ther.*, **29**, 2456–2468.
 17. Molinari, E. and Sayer, J.A. (2021) Gene and epigenetic editing in the treatment of primary ciliopathies. *Prog. Mol. Biol. Transl. Sci.*, **182**, 353–401.
 18. Chiu, W., Lin, T.Y., Chang, Y.C., Mulyadi Lai, H.I.A., Lin, S.C., Ma, C., Yarmishyn, A.A., Lin, S.C., Chang, K.J., Chou, Y.B., et al. (2021) An update on gene therapy for inherited retinal dystrophy: experience in Leber congenital Amaurosis clinical trials. *Int. J. Mol. Sci.*, **22**, 4534.
 19. Kenny, J., Forsythe, E., Beales, P. and Bacchelli, C. (2017) Toward personalized medicine in Bardet-Biedl syndrome. *Per. Med.*, **14**, 447–456.
 20. Malicki, J.J. and Johnson, C.A. (2017) The Cilium: cellular antenna and central processing Unit. *Trends Cell Biol.*, **27**, 126–140.
 21. Satir, P., Pedersen, L.B. and Christensen, S.T. (2010) The primary cilium at a glance. The primary cilium at a glance. *J. Cell Sci.*, **123**, 499–503.
 22. Anvarian, Z., Mykytyn, K., Mukhopadhyay, S., Pedersen, L.B. and Christensen, S.T. (2019) Cellular signalling by primary cilia in development, organ function and disease. *Nat. Rev. Nephrol.*, **15**, 199–219.
 23. Kim, W., Underwood, R.S., Greenwald, I. and Shaye, D.D. (2018) OrthoList 2: a new comparative genomic analysis of human and *Caenorhabditis elegans* genes. *Genetics*, **210**, 445–461.
 24. van Dam, T.J.P., Kennedy, J., van der Lee, R., de Vrieze, E., Wunderlich, K.A., Rix, S., Dougherty, G.W., Lambacher, N.J., Li, C., Jensen, V.L. et al. (2019) CiliaCarta: An integrated and validated compendium of ciliary genes. *PLoS One*, **14**, e0216705.
 25. Inglis, P.N., Ou, G., Leroux, M.R. and Scholey, J.M. (2007) The sensory cilia of *Caenorhabditis elegans*. *WormBook*, 1–22.
 26. Lange, K.I., Tsiropoulou, S., Kucharska, K. and Blacque, O.E. (2021) Interpreting the pathogenicity of Joubert Syndrome missense variants in *Caenorhabditis elegans*. *Dis. Model. Mech.*, **14**, dmm046631.
 27. Smith, U.M., Consugar, M., Tee, L.J., McKee, B.M., Maina, E.N., Whelan, S., Morgan, N.V., Goranson, E., Gissen, P., Lilliquist, S. et al. (2006) The transmembrane protein meckelin (MKS3) is mutated in Meckel-Gruber syndrome and the wpk rat. *Nat. Genet.*, **38**, 191–196.
 28. Iannicelli, M., Brancati, F., Mougou-Zerelli, S., Mazzotta, A., Thomas, S., Elkhartoufi, N., Travaglini, L., Gomes, C., Ardisino, G.L., Bertini, E. et al. (2010) Novel TMEM67 mutations and genotype-phenotype correlates in meckelin-related ciliopathies. *Hum. Mutat.*, **31**, E1319–E1331.
 29. Tallila, J., Salonen, R., Kohlschmidt, N., Peltonen, L. and Kestila, M. (2009) Mutation spectrum of Meckel syndrome genes: one group of syndromes or several distinct groups? *Hum. Mutat.*, **30**, E813–E830.
 30. Szymanska, K., Berry, I., Logan, C.V., Cousins, S.R., Lindsay, H., Jafri, H., Raashid, Y., Malik-Sharif, S., Castle, B., Ahmed, M. et al. (2012) Founder mutations and genotype-phenotype correlations in Meckel-Gruber syndrome and associated ciliopathies. *Cilia*, **1**.
 31. Brancati, F., Iannicelli, M., Travaglini, L., Mazzotta, A., Bertini, E., Boltshauser, E., D'Arrigo, S., Emma, F., Fazzi, E., Gallizzi, R. et al. (2009) MKS3/TMEM67 mutations are a major cause of COACH Syndrome, a Joubert Syndrome related disorder with liver involvement. *Hum. Mutat.*, **30**, E432–E442.
 32. Suzuki, T., Miyake, N., Tsurusaki, Y., Okamoto, N., Alkindy, A., Inaba, A., Sato, M., Ito, S., Muramatsu, K., Kimura, S. et al. (2016) Molecular genetic analysis of 30 families with Joubert syndrome. *Clin. Genet.*, **90**, 526–535.
 33. Consugar, M.B., Navarro-Gomez, D., Place, E.M., Bujakowska, K.M., Sousa, M.E., Fonseca-Kelly, Z.D., Taub, D.G., Janessian, M., Wang, D.Y., Au, E.D. et al. (2015) Panel-based genetic diagnostic testing for inherited eye diseases is highly accurate and reproducible, and more sensitive for variant detection, than exome sequencing. *Genet. Med.*, **17**, 253–261.
 34. Fleming, L.R., Doherty, D.A., Parisi, M.A., Glass, I.A., Bryant, J., Fischer, R., Turkbey, B., Choyke, P., Daryanani, K., Vemulapalli, M. et al. (2017) Prospective Evaluation of Kidney Disease in Joubert Syndrome. *Clin. J. Am. Soc. Nephrol.*, **12**, 1962–1973.
 35. Huynh, J.M., Galindo, M. and Laukaitis, C.M. (2018) Missense variants in TMEM67 in a patient with Joubert syndrome. *Clin Case Rep.*, **6**, 2189–2192.
 36. Williams, C.L., Masyukova, S.V. and Yoder, B.K. (2010) Normal ciliogenesis requires synergy between the cystic kidney disease genes MKS-3 and NPHP-4. *J. Am. Soc. Nephrol.*, **21**, 782–793.
 37. Garcia-Gonzalo, F.R. and Reiter, J.F. (2017) Open sesame: how transition fibers and the transition zone control ciliary composition. *Cold Spring Harb. Perspect. Biol.*, **9**, a028134.
 38. Gonçalves, J. and Pelletier, L. (2017) The ciliary transition zone: finding the pieces and assembling the gate. *Mol. Cells*, **40**, 243–253.
 39. Sang, L., Miller, J.J., Corbit, K.C., Giles, R.H., Brauer, M.J., Otto, E.A., Baye, L.M., Wen, X., Scales, S.J., Kwong, M. et al. (2011) Mapping the NPHP-JBTS-MKS protein network reveals ciliopathy disease genes and pathways. *Cell*, **145**, 513–528.
 40. Garcia-Gonzalo, F.R., Corbit, K.C., Sirerol-Piquer, M.S., Ramaswami, G., Otto, E.A., Noriega, T.R., Seol, A.D., Robinson, J.F., Bennett, C.L., Josifova, D.J. et al. (2011) A transition zone complex regulates mammalian ciliogenesis and ciliary membrane composition. *Nat. Genet.*, **43**, 776–784.
 41. Chih, B., Liu, P., Chinn, Y., Chalouni, C., Komuves, L.G., Hass, P.E., Sandoval, W. and Peterson, A.S. (2011) A ciliopathy complex at the transition zone protects the cilia as a privileged membrane domain. *Nat. Cell Biol.*, **14**, 61–72.
 42. Pratt, M.B., Titlow, J.S., Davis, I., Barker, A.R., Dawe, H.R., Raff, J.W. and Roque, H. (2016) Drosophila sensory cilia lacking MKS proteins exhibit striking defects in development but only subtle defects in adults. *J. Cell Sci.*, **129**, 3732–3743.

43. Barker, A.R., Renzaglia, K.S., Fry, K. and Dawe, H. (2014) Bioinformatic analysis of ciliary transition zone proteins reveals insights into the evolution of ciliopathy networks. *BMC Genomics*, **15**, 531.
44. Williams, C.L., Li, C., Kida, K., Inglis, P.N., Mohan, S., Semene, L., Biala, N.J., Stupay, R.M., Chen, N., Blacque, O.E. et al. (2011) MKS and NPHP modules cooperate to establish basal body/transition zone membrane associations and ciliary gate function during ciliogenesis. *J. Cell Biol.*, **192**, 1023–1041.
45. Shim, J.W., Territo, P.R., Simpson, S., Watson, J.C., Jiang, L., Riley, A.A., McCarthy, B., Persohn, S., Fulkerson, D. and Blazer-Yost, B.L. (2019) Hydrocephalus in a rat model of Meckel Gruber syndrome with a TMEM67 mutation. *Sci. Rep.*, **9**, 1069.
46. Leightner, A.C., Hommerding, C.J., Peng, Y., Salisbury, J.L., Gainullin, V.G., Czarnecki, P.G., Sussman, C.R. and Harris, P.C. (2013) The Meckel syndrome protein meckelin (TMEM67) is a key regulator of cilia function but is not required for tissue planar polarity. *Hum. Mol. Genet.*, **22**, 2024–2040.
47. Cook, S.A., Collin, G.B., Bronson, R.T., Naggert, J.K., Liu, D.P., Akeson, E.C. and Davison, M.T. (2009) A mouse model for Meckel syndrome type 3. *J. Am. Soc. Nephrol.*, **20**, 753–764.
48. Louie, C.M., Caridi, G., Lopes, V.S., Brancati, F., Kispert, A., Lancaster, M.A., Schlossman, A.M., Otto, E.A., Leitges, M., Gröne, H.J. et al. (2010) AH11 is required for photoreceptor outer segment development and is a modifier for retinal degeneration in nephronophthisis. *Nat. Genet.*, **42**, 175–180.
49. Won, J., Marin de Esvikova, C., Smith, R.S., Hicks, W.L., Edwards, M.M., Longo-Guess, C., Li, T., Naggert, J.K. and Nishina, P.M. (2011) NPHP4 is necessary for normal photoreceptor ribbon synapse maintenance and outer segment formation, and for sperm development. *Hum. Mol. Genet.*, **20**, 482–496.
50. Bentley-Ford, M.R., LaBonty, M., Thomas, H.R., Haycraft, C.J., Scott, M., LaFayette, C., Croyle, M.J., Parant, J.M. and Yoder, B.K. (2021) Evolutionarily conserved genetic interactions between *nphp-4* and *bbs-5* mutations exacerbate ciliopathy phenotypes. *Genetics*, *ijab209*.
51. Fliegauf, M., Horvath, J., von Schnakenburg, C., Olbrich, H., Müller, D., Thumfart, J., Schermer, B., Pazour, G.J., Neumann, H.P.H., Zentgraf, H. et al. (2006) Nephrocystin specifically localizes to the transition zone of renal and respiratory cilia and photoreceptor connecting cilia. *J. Am. Soc. Nephrol.*, **17**, 2424–2433.
52. Dawe, H.R., Smith, U.M., Cullinane, A.R., Gerrelli, D., Cox, P., Badano, J.L., Blair-Reid, S., Sriram, N., Katsanis, N., Attie-Bitach, T. et al. (2007) The Meckel-Gruber Syndrome proteins MKS1 and meckelin interact and are required for primary cilium formation. *Hum. Mol. Genet.*, **16**, 173–186.
53. Yee, L.E., Garcia-Gonzalo, F.R., Bowie, R.V., Li, C., Kennedy, J.K., Ashrafi, K., Blacque, O.E., Leroux, M.R. and Reiter, J.F. (2015) Conserved genetic interactions between ciliopathy complexes cooperatively support ciliogenesis and ciliary signaling. *PLoS Genet.*, **11**, e1005627.
54. Hunt, S.E., McLaren, W., Gil, L., Thormann, A., Schuilenburg, H., Sheppard, D., Parton, A., Armean, I.M., Trevanion, S.J., Flicek, P. et al. (2018) Ensembl variation resources. *Database*.
55. Winkelbauer, M.E., Schafer, J.C., Haycraft, C.J., Swoboda, P. and Yoder, B.K. (2005) The *C. elegans* homologs of nephrocystin-1 and nephrocystin-4 are cilia transition zone proteins involved in chemosensory perception. *J. Cell Sci.*, **118**, 5575–5587.
56. Jauregui, A.R. and Barr, M.M. (2005) Functional characterization of the *C. elegans* nephrocystins NPHP-1 and NPHP-4 and their role in cilia and male sensory behaviors. *Exp. Cell Res.*, **305**, 333–342.
57. Sanders, A.A.W.M., Kennedy, J. and Blacque, O.E. (2015) Image analysis of *Caenorhabditis elegans* ciliary transition zone structure, ultrastructure, molecular composition, and function. *Methods Cell Biol.*, **127**, 323–347.
58. Perkins, L.A., Hedgecock, E.M., Nichol Thomson, J. and Culotti, J.G. (1986) Mutant sensory cilia in the nematode *Caenorhabditis elegans*. *Dev. Biol.*, **117**, 456–487.
59. Starich, T.A., Herman, R.K., Kari, C.K., Yeh, W.H., Schackwitz, W.S., Schuyler, M.W., Collet, J., Thomas, J.H. and Riddle, D.L. (1995) Mutations affecting the chemosensory neurons of *Caenorhabditis elegans*. *Genetics*, **139**, 171–188.
60. Bargmann, C.I., Hartwig, E. and Horvitz, H.R. (1993) Odorant-selective genes and neurons mediate olfaction in *C. elegans*. *Cell*, **74**, 515–527.
61. Abdelhamed, Z.A., Natarajan, S., Wheway, G., Inglehearn, C.F., Toomes, C., Johnson, C.A. and Jagger, D.J. (2015) The Meckel-Gruber syndrome protein TMEM67 controls basal body positioning and epithelial branching morphogenesis in mice via the non-canonical Wnt pathway. *Dis. Model. Mech.*, **8**, 527–541.
62. Thusberg, J., Olatubosun, A. and Vihinen, M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.
63. Shulman, C., Liang, E., Kamura, M., Udwan, K., Yao, T., Cattran, D., Reich, H., Hladunewich, M., Pei, Y., Savige, J. et al. (2021) Type IV Collagen Variants in CKD: Performance of Computational Predictions for Identifying Pathogenic Variants. *Kidney Med.*, **3**, 257–266.
64. Miosge, L.A., Field, M.A., Sontani, Y., Cho, V., Johnson, S., Palkova, A., Balakishnan, B., Liang, R., Zhang, Y., Lyon, S. et al. (2015) Comparison of predicted and actual consequences of missense mutations. *Proc. Natl. Acad. Sci. U.S.A.*, **112**, E5189–E5198.
65. Chennen, K., Weber, T., Lornage, X., Kress, A., Böhm, J., Thompson, J., Laporte, J. and Poch, O. (2020) MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants. *PLoS One*, **15**, e0236962.
66. Sim, N.L., Kumar, P., Hu, J., Henikoff, S., Schneider, G. and Ng, P.C. (2012) SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res.*, **40**, W452–W457.
67. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S. and Sunyaev, S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
68. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. and Kircher, M. (2019) CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.*, **47**, D886–D894.
69. Ioannidis, N.M., Rothstein, J.H., Pejaver, V., Middha, S., McDonnell, S.K., Baheti, S., Musolf, A., Li, Q., Holzinger, E., Karyadi, D. et al. (2016) REVEL: an ensemble method for predicting the Pathogenicity of Rare Missense Variants. *Am. J. Hum. Genet.*, **99**, 877–885.
70. Swierczek, N.A., Giles, A.C., Rankin, C.H. and Kerr, R.A. (2011) High-throughput behavioral analysis in *C. elegans*. *Nat. Methods*, **8**, 592–598.
71. Churgin, M.A. and Fang-Yen, C. (2015) An imaging system for *C. elegans* behavior. *Methods Mol. Biol.*, **1327**, 199–207.
72. Fernandez, A.G., Bargmann, B.O.R., Mis, E.K., Edgley, M.L., Birnbaum, K.D. and Piano, F. (2012) High-throughput fluorescence-based isolation of live *C. elegans* larvae. *Nat. Protoc.*, **7**, 1502–1510.
73. McCormick, K., Brock, T., Wood, M., Guo, L., McBride, K., Kim, C., Resch, L., Pop, S., Bradford, C., Kendrick, P. et al. (2021) A gene

- replacement humanization platform for rapid functional testing of clinical variants in Epilepsy-associated STXBP1. *bioRxiv*, 2021.08.13, 453827.
74. Kukhtar, D., Rubio-Peña, K., Serrat, X. and Cerón, J. (2020) Mimicking of splicing-related retinitis pigmentosa mutations in *C. elegans* allow drug screens and identification of disease modifiers. *Hum. Mol. Genet.*, **29**, 756–765.
 75. Iyer, S., Mast, J.D., Tsang, H., Rodriguez, T.P., DiPrimio, N., Prangle, M., Sam, F.S., Parton, Z. and Perlstein, E.O. (2019) Drug screens of NGLY1 deficiency in worm and fly models reveal catecholamine, NRF2 and anti-inflammatory-pathway activation as potential clinical approaches. *Dis. Model. Mech.*, **12**, dmm040576.
 76. Patten, S.A., Aggad, D., Martinez, J., Tremblay, E., Petrillo, J., Armstrong, G.A., La Fontaine, A., Maios, C., Liao, M., Ciura, S. et al. (2017) Neuroleptics as therapeutic compounds stabilizing neuromuscular transmission in amyotrophic lateral sclerosis. *JCI Insight*, **2**, e97152.
 77. Gosai, S.J., Kwak, J.H., Luke, C.J., Long, O.S., King, D.E., Kovatch, K.J., Johnston, P.A., Ying Shun, T., Lazo, J.S., Perlmutter, D.H. et al. (2010) Automated high-content live animal drug screening using *C. elegans* expressing the aggregation prone serpin α 1-antitrypsin Z. *PLoS One*, **5**, e15460.
 78. Di Rocco, M., Galosi, S., Lanza, E., Tosato, F., Caprini, D., Folli, V., Friedman, J., Bocchinfuso, G., Martire, A., Di Schiavi, E. et al. (2021) *Caenorhabditis elegans* provides an efficient drug screening platform for GNAO1-related disorders and highlights the potential role of caffeine in controlling dyskinesia. *Hum. Mol. Genet.*, **30**, ddab296.
 79. McDiarmid, T.A., Au, V., Loewen, A.D., Liang, J., Mizumoto, K., Moerman, D.G. and Rankin, C.H. (2018) CRISPR-Cas9 human gene replacement and phenomic characterization in *Caenorhabditis elegans* to understand the functional conservation of human genes and decipher variants of uncertain significance. *Dis. Model. Mech.*, **11**, dmm036517.
 80. Zhu, B., Mak, J.C.H., Morris, A.P., Marson, A.G., Barclay, J.W., Sililas, G.J. and Morgan, A. (2020) Functional analysis of epilepsy-associated variants in STXBP1/Munc18-1 using humanized *Caenorhabditis elegans*. *Epilepsia*, **61**, 810–821.
 81. Green, E.D., Gunter, C., Biesecker, L.G., Di Francesco, V., Easter, C.L., Feingold, E.A., Felsenfeld, A.L., Kaufman, D.J., Ostrander, E.A., Pavan, W.J. et al. (2020) Strategic vision for improving human health at the forefront of genomics. *Nature*, **586**, 683–692.
 82. Xu, J. (2019) Distance-based protein folding powered by deep learning. *Proc. Natl. Acad. Sci. U.S.A.*, **116**, 16856–16865.
 83. Brenner, S. (1974) The genetics of *Caenorhabditis elegans*. *Genetics*, **77**, 71–94.
 84. Paix, A., Folkmann, A., Rasoloson, D. and Seydoux, G. (2015) High efficiency, homology-directed genome editing in *Caenorhabditis elegans* using CRISPR-Cas9 ribonucleoprotein complexes. *Genetics*, **201**, 47–54.
 85. Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R. et al. (2016) Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. *Nat. Biotechnol.*, **34**, 184–191.
 86. Rawsthorne-Manning, H., Calahorra, F., Izquierdo, P.G., Holden-Dye, L., O'Connor, V. and Dillon, J. (2021) Confounds of using the *unc-58* selection marker highlights the importance of genotyping co-CRISPR genes. *bioRxiv*, 2021.05.26, 445785.
 87. Hobert, O. (2002) PCR fusion-based approach to create reporter gene constructs for expression analysis in transgenic *C. elegans*. *Biotechniques*, **32**, 728–730.
 88. Hsiau, T., Conant, D., Rossi, N., Maures, T., Waite, K., Yang, J., Joshi, S., Kelso, R., Holden, K., Enzmann, B.L. et al. (2019) Inference of CRISPR edits from sanger trace data. *bioRxiv*, 251082.
 89. McLaren, W., Gil, L., Hunt, S.E., Singh Riat, H., Ritchie, G.R.S., Thormann, A., Flicek, P. and Cunningham, F. (2016) The Ensembl variant effect predictor. *Genome Biol.*, **17**, 122.

5 Discussion

5.1 Research output summary

During this PhD, I have had two original research projects published (Best et al., 2022c, Best et al., 2022b) based on analysis of 100K data. Both outputs contribute to improved molecular diagnosis rates for ciliopathy patients and provide transferable skills and lessons about WGS analysis, particularly from 100K, applicable to wider patient groups. I have also had a commentary article published reflecting upon lessons learned from the two 100K analyses, provided at the end of this discussion (thesis section 5.6) (Best et al., 2022a). GEL published a response letter to this commentary, also included (thesis section 5.7) (Brown et al., 2022). My third, published original research project was development of a functional *TMEM67* VUS interpretation strategy, done in collaboration with colleagues from UCD (Lange et al., 2022).

5.2 Motivation for the PhD and overall take-home messages

I was motivated to undertake this PhD by experiencing real clinical challenges facing patients and clinicians during my paediatric and clinical genetics training. The major issue I wanted to focus on is the need to improve genetic diagnosis rates for patients with rare diseases, having witnessed the anxieties and frustrations generated by negative or VUS results. I wanted to achieve this through genomic variant analyses to find previously “hidden” molecular diagnoses, and through functional variant interpretation to help move VUS results out of their diagnostic grey area into definitive categories.

The clearest overall message from this PhD is that our ability to interpret genomic sequence data is far behind our ability to generate it. Our inability to definitively interpret many variant types without bespoke research input, particularly novel rare missense variants and non-coding variants, is a major bottleneck preventing us from improving genetic diagnosis rates for patients with genetic diseases. Without a molecular diagnosis, patients cannot access targeted therapeutics or benefit from family, prenatal or preimplantation genetic testing. This research experience demonstrates that a huge amount more time and money need to be directed into providing training and resources for genomic variant interpretation, if we have any chance of offering patients better diagnosis rates and subsequent benefits than is currently being achieved.

Without undertaking SV and non-coding variant analysis, there doesn't seem to be an overall clinical benefit for doing WGS over WES yet. We must hope that with time and better diagnostic pipelines, iterative retrospective WGS analyses can be performed for historically unsolved cases to find previously un-detected pathogenic variants. However, reports from my clinical colleagues suggest that workload pressures on diagnostic labs have forced them to "close" unsolved 100K cases, leading to those patients having repeated chemistry and fresh analyses if further genetic tests are requested. This seems to defeat the purpose of undertaking WGS, which should be the gold-standard in genomic testing, curtailing the "diagnostic odyssey" of serial, more limited testing.

From the bioinformatics part of the project, I have acquired several transferable skills in genomic variant analysis which will be directly applicable to my clinical work. From the laboratory part, I have mostly learned a new appreciation for the huge amounts of work, time and resources that go into functional variant analyses. When reporting diagnostic results to patients, we must err on the side of caution to avoid returning false negative results, hence the high rates of VUS results. I now understand much better how important it is to develop high-throughput, simple and inexpensive strategies that could be applicable in the diagnostic setting to facilitate definitive variant classification.

5.3 Lessons learned: 100,000 Genomes Project analyses

Lessons learned from the two 100K projects undertaken are extensively covered in the manuscript discussions and our published commentary article: "Unlocking the potential of the UK 100,000 Genomes Project - lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort" (thesis section 5.6) (Best et al., 2022a). Some further points for discussion are presented below.

5.3.1 Diagnostic uplift achieved by 100K rare disease cohort research analyses

At the time of writing, our analysis of 83 probands with suspected primary ciliopathies recruited in the BBS, JBTS and RMCD categories (the so-called congenital malformations caused by ciliopathies (CMC) cohort) was the first cohort study reporting diagnosis rates from 100K rare disease participants with a range of molecular and clinical diagnoses. Released 100K publications to date were focused on individual genes or

variants (listed at <https://www.genomicsengland.co.uk/research/publications>), so we did not have an idea of the diagnostic uplift that we could expect to achieve through our analysis.

We determined a research molecular diagnosis for n=45/83 (54.2%) probands. 43 of these are published in the CMC cohort analysis manuscript (thesis section 2.4) (Best et al., 2022b), and two further diagnoses were made post-publication. Participant #78 was diagnosed post-publication with a homozygous *BBS4* deletion through the SVRare script, detailed in the additional results section of the CMC cohort study chapter (thesis section 2.3). Participant #59 was solved through additional laboratory work including a duplex PCR screening assay and sequencing to characterise a 2.4kb insertion in *BBS1*, in *trans* with a known pathogenic *BBS1* founder variant (detailed in the discussion section of the reverse phenotyping manuscript (thesis section 3.2) (Best et al., 2022c). Overall, we provided a 21.7% diagnostic uplift compared to results previously reported by GEL (n=45/83 (54.2%) vs n=27/83 (32.5%)), although we recognise that 10/83 (12%) had to be classified as possible diagnoses as they contained only VUSs. Although these cannot be acted upon clinically, we think they are still important to report in the research setting, as new tools for functional validation are emerging which may allow definitive classification in the future.

Our most significant source of alternative diagnoses was from analysis of non-ciliopathy disease genes (n=19/45 diagnoses), which we hypothesised reflected difficulties in the clinical recognition of ciliopathy syndromes or selection of appropriate recruitment categories to recruit participants into 100K. The next most important source of otherwise missed diagnoses was SV analysis, which contributed to five participants' diagnoses (three published in the manuscript, two made post-publication). However, it was first important to do the un-biased SNV analysis we did, independent of the tiering system, to identify the 'first-hit' SNV alleles that signposted us to IGV to look for the second-hit SVs that we found. As discussed in "Tiering Issues" (manuscript section 2) of our commentary article (Best et al., 2022a), a major limitation of the 100K tiering system is the failure to flag single heterozygous variants in recessive genes for further analysis, where the second variant completing the biallelic inheritance model is harder to find (e.g. intronic or SV). Other important sources of new potentially pathogenic variants, which mostly had to be classified as VUSs, were non-coding variant and candidate gene analyses.

Since our publication, additional 100K cohort studies have been released. The preliminary report from the 100K pilot study of 4660 rare disease participants provided a molecular diagnosis rate for 25% of probands (The 100,000 Genomes Project Pilot Investigators et al., 2021). In this pilot project report, only 60% of molecular diagnoses contained SNVs in genes on the automatically applied panels, with 26% coming from other disease genes. These diagnoses were made through expert review and phenotype-based prioritisation with additional clinical data by the study clinicians or the clinical genetics teams from industry partners (Congenica and Fabric Genomics). A further 14% of diagnoses were made from phenotype-agnostic research analyses looking for variants beyond coding SNVs on applied panels (mitochondrial DNA (1%), non-coding SNVs on applied panels (4%), SVs on applied panels (8%), SNVs in newly discovered disease genes (1%)). However, in the main 100K program (data awaiting release), these additional analyses are not guaranteed.

Amongst probands entered to 100K with diagnostically challenging primary mitochondrial disease phenotypes, only 17/102 (16.7%) had a molecular diagnosis identified through standard analysis pipelines (Macken et al., 2022). The unsolved cases were reviewed by a specialist multi-disciplinary team led by a genomic medicine clinician and bioinformatician. They identified an additional 15/102 diagnoses, almost doubling the diagnostic rate to 31.4%, with an additional 3.9% (4/102) candidate diagnoses (highly suspicious VUSs in known or newly established genes). This was achieved through a comprehensive review of phenotypes and pedigrees leading to analysis of genes on alternative panels, analysis of PanelApp 'amber' and 'red' genes, reassessment of VUSs, search for second hits in recessive genes in the presence of a single strong heterozygous candidate, a search for pathogenic CNVs and custom analysis of mitochondrial DNA. Three diagnoses required additional functional validation. Their most significant sources of additional diagnoses were missed intronic second hits in recessive genes (5/15) and analysis of genes on alternative panels (7/15).

Another parallel cohort study to which we can compare our outcomes is of 100K participants recruited with craniosynostosis (Hyder et al., 2021). This research group evaluated the performance of the automated GEL panel-based pipelines, reporting a diagnostic sensitivity of only 47%. Through their analyses, they identified 18 pathogenic or likely pathogenic variants in addition to the 16 reported by GEL, amongst 114 probands recruited to 100K with craniosynostosis. Their sources of missed diagnoses include variants on applied panels that were mis-called or filtered out, for example due

to being a single hit in a recessive gene (n=6), genes on alternative, unapplied PanelApp panels (n=7), SVs (n=2) and variants in research genes (n=2) (Hyder et al., 2021).

The diagnostic uplift achieved through additional clinical and research efforts in these three projects (The 100,000 Genomes Project Pilot Investigators et al., 2021, Macken et al., 2022, Hyder et al., 2021) are similar to what we reported in the CMC cohort analysis (Best et al., 2022b), with overlapping sources of missed diagnoses. Clearly, the automated tiering system is going to miss huge numbers of diagnoses, many of which are not too hard to find with additional clinical information and time to explore the data. Engagement with researchers who have the time, funding, and motivation to find these additional diagnoses is critical, to make the most of this resource. (Macken et al., 2022) go a step further, suggesting that establishment of specialist genomics MDTs are required to improve diagnosis rates for complex cases, as individual research groups cannot offer a systematic or equitable solution to the challenge of unsolved WGS. This reflects their frequent lack of access to detailed clinical data, discrepancies in research interests and funding, and variable patient involvement in research.

GEL provided a response our to commentary article (thesis section 5.7) (Brown et al., 2022). In summary, they acknowledged the issues we identified for 100K phenotyping, tiering issues, difficulties in using the GEL research environment and reporting problems. They provided reassurance that measures are in place to improve the quality and quantity of clinical data available and means to return results to clinicians. They reported that there is a new, cloud- based research environment designed for ease of use, particularly by researchers who are not skilled programmers. I am so far yet to notice any major changes since I started my 100K research in 2019. They provided information about regular live online training sessions for 100K researchers but acknowledged that these were not available during the pandemic. They also reported a new analytical bioinformatics pipeline with improved variant calling performance, which will prioritise all Tier 1 and 2 variants, *de novo* Tier 3 variants, the best candidate variants called by Exomiser (Robinson et al., 2014), CNVs and short tandem repeat expansions for routine diagnostic laboratory review. I would hope that these measures improve diagnosis rates for 100K participants and make research from this dataset easier in the future.

5.3.2 Time commitments and strategy development

Analysis of the CMC cohort took significantly longer than the reverse phenotyping project. This is largely due to my unfamiliarity with the difficult GEL research environment, Linux command line entry and lack of an established diagnostic strategy. In total, the CMC cohort analysis took around a year, whereas the reverse phenotyping project was completed within six weeks. I was able to apply many of the same strategies that I had learned in the CMC cohort, navigate the environment quickly and comfortably and had established connections with key collaborators that helped me to develop the diagnostic pipeline much faster than I was able to in my first 100K project.

In the CMC cohort analysis, I began by manually inspecting all the Tier 1, 2 and 3 variants for the first few participants before realising that this would not be practical given the sheer volume of Tier 3's. I needed to come up with a systematic and un-biased strategy to extract variants in both ciliopathy and non-ciliopathy genes for analysis for filtering and analysis, independent of the tiering system. I initially planned to use available variant data within the research environment from Exomiser (Robinson et al., 2014), as was applied in the Pilot 100K project (The 100,000 Genomes Project Pilot Investigators et al., 2021). Exomiser is a Java program that comprises a suite of algorithms for prioritising rare, segregating, and predicted pathogenic variants from WES or WGS data. However, it is dependent upon both VCF file and HPO term input data. Therefore, given our awareness of the frequently poor phenotyping data in 100K, I decided not to pursue this strategy.

An introduction to Dr. Jenny Lord, a Postdoctoral Research Fellow within the Faculty of Medicine at the University of Southampton, made my unbiased approach possible through sharing of her *find_variants_by_gene_and_consequence.py* script, which allowed identification of all variants, independent of the tiering system, in multiple PanelApp panels at a time. The output file was in a VCF format, allowing it to be submitted to Ensembl VEP from the command line with selected additional plugins. I could then do variant filtering and analysis from un-biased data.

I undertook serial panel analysis, beginning with the RMCD panel, then applying additional panels according to the entered HPO terms. In hindsight it would have been more scientific and streamlined to have started with the DDG2P panel of 1193 diagnostic grade “green genes” (signed off version 2.2 available from <https://nhsgms->

panelapp.genomicsengland.co.uk/panels/484/v2.2) for all participants, rather than undertake several smaller panel analyses, given the frequently poor phenotyping data available. It would have also saved me time wasted in repeating the variant extraction, filtering, and analysis steps multiple times.

The other strategy that was hugely time-consuming and un-systematic, was the post-hoc manual inspection of IGV that I undertook for SV analysis in the CMC cohort analysis. I manually inspected the entire gene locus for all participants with a single heterozygous variant of interest in a recessive gene in pursuit of a second “hit” SV to complete a biallelic molecular diagnosis. I was aware of the SV variant call data from Manta (Chen et al., 2016) and Canvas (Ivakhno et al., 2018) available within the GEL research environment. However, neither I, nor any of my more experienced colleagues or collaborators at the time, had a confident strategy to filter the SV.vcf files based on any quality or frequency metrics. The number of SV calls made was too large to be manually analysed without prior filtering (~ 5000-11,000 SV calls per CMC cohort participant from sample size n=5). Unfortunately, SV discovery tools still report large numbers of false positives (false discovery rates from short-read WGS data reported as high as 89-91%) (Bertolotti et al., 2020, Mills et al., 2011), and many researchers still depend upon visual inspection to identify real SVs.

An introduction to Dr. Jing Yu, a senior bioinformatician with the Nuffield Department of Clinical Neurosciences at the University of Oxford, transformed my ability to prioritise SVs quickly and accurately for analysis for the reverse phenotyping project, through access to his SVRare script. It also allowed detection of homozygous SVs, which would not have been achievable through manual visual analysis in absence of a “signposting” first hit variant to guide me to the right gene. This contributed to the post-publication diagnosis in participant #78 of the CMC cohort, revealing a homozygous *BBS4* deletion (detailed in thesis section 2.3). It is completely impractical to manually inspect every ciliopathy gene, or even more than one or two candidate genes, on IGV, so this shows the value of systematic and quick strategies for SV analyses to boost diagnosis rates from WGS data.

The final research collaboration that allowed me to conduct the reverse phenotyping study as efficiently as I did was with Roel Bevers, the GEL bioinformatician who wrote and tutored me through the Gene-Variant Workflow script. This extracted all variants in up to ten genes at a time from the 100K rare disease dataset, with accompanying

frequency data and links to the clinical data for participants in which the extracted variants had been called. Once I had this script up and running, data extraction was complete within hours and then analysis could begin.

These experiences of learning from generous experienced researchers and bioinformaticians around the country demonstrate the power of collaboration. Platforms like GitHub for the sharing of established scripts are invaluable. I am determined to promote ongoing collaboration to maximise research outputs and patient benefit from the groundwork already established, rather than expecting new 100K researchers to “reinvent the wheel” in a difficult research environment. Our scripts are already being used by clinical and research colleagues in Leeds, and since our diagnostic pipelines have been published in journal articles and GitHub, we hope they will be used more broadly.

5.3.3 Reverse phenotyping as a source of missed diagnoses

We developed a reverse phenotyping strategy, looking for “hidden” ciliopathy patients recruited to alternative 100K categories, with pathogenic variants in nine disease genes representative of the multi-systemic primary ciliopathy spectrum (*BBS1*, *BBS10*, *ALMS1*, *OFD1*, *DYNC2H1*, *WDR34*, *NPHP1*, *TMEM67*, *CEP290*) (thesis chapter 3) (Best et al., 2022c). It proved to be a successful approach, allowing me to report 18 molecular ciliopathy diagnoses identified amongst unsolved 100K participants (13 new findings and 5 un-reported by GEL), as well as finding 44 previously identified and reported by GEL.

We also identified 11 participants with potential molecular diagnoses that we could not justify reporting to recruiting clinicians because the clinical detail available within the GEL research environment was not compatible with the major clinical features for the associated ciliopathy gene. We will never know whether these diagnoses are real, and consistent with clinical features that were not entered into 100K during recruitment, or are spurious findings. Therefore, the main outcome of this project is that the quality of phenotyping data is critical to allow accurate genotype-phenotype correlation. Use of the Human Phenotype Ontology (HPO) to standardise the vocabulary of phenotypic abnormalities for 100K and other genotype-phenotype correlation studies is extremely helpful, but we have observed a frequent lack of detail, especially regarding multi-systemic problems. Without taking the time to do comprehensive, multi-systemic phenotyping, variant-level data cannot be accurately interpreted. This message must

be disseminated to both clinical geneticists and mainstream clinicians for future studies and genetic tests.

The only other example of reverse phenotyping from 100K data is from (Macken et al., 2022), who report that reverse phenotyping contributed to new diagnoses in 5/102 unsolved primary mitochondrial disease 100K cases. This involved review of pre-existing clinical data, undertaking of additional clinical history taking and examination and further investigations to validate whether identified variants were relevant to the patients. This included review of brain MRI, facial dysmorphology, skeletal survey and muscle biopsy.

Having completed this project relatively quickly, with the pipeline in place, I would be eager to extend the approach to other ciliopathy genes to look for further “hidden” ciliopathy patients in 100K. It would also be interesting to reflect upon other conditions that may be clinically difficult to recognise, with high heterogeneity and complex phenotypes, for which this approach may also be useful. Alternatively, this approach could be used agnostically for larger sets of developmental genes to look for missed diagnoses that are “low-hanging fruit”, by, in the first instance, looking for straightforward molecular diagnoses (for example, high impact variants or those previously listed as pathogenic on ClinVar).

5.3.4 Added value of structural variant analysis in 100K

We were successful in providing previously missed diagnoses through SV analysis from WGS data for participants in both the CMC cohort analysis (n=5; three published plus two additional retrospective diagnosis) and the reverse phenotyping project (n=2). In the CMC cohort, our diagnostic uplift from SV analysis is 5/83 (6%). This is very similar to previous reports of the diagnostic uplift from SV analysis from WGS data (4.8% from unsolved British inherited retinal dystrophy patients (Carss et al., 2017), 8% from pilot 100K participants (The 100,000 Genomes Project Pilot Investigators et al., 2021)).

Use of the SVRare script in the reverse phenotyping project made SV analysis quick, streamlined, and straightforward. Prior to the introduction of SVRare, the inability to efficiently merge SVs from different individuals had prevented the discovery of disease-causing SVs in 100K. The high number of false positives from Manta (Chen et

al., 2016) and Canvas (Ivakhno et al., 2018), and the inability to estimate allele frequency, made systematic strategies for SV filtering and analysis extremely challenging. SVRare aggregated 554,060,126 SVs called by Manta and Canvas amongst all 71,408 participants in the rare-disease arm of 100K. This provided a database from which only rare potentially pathogenic SVs overlapping coding regions in genes of interest could be identified, akin to use of the gnomAD database for rare SNV filtering. The output data from SVRare did still contain some false positives, but at a more manageable volume for visual inspection. For example, extraction of rare SVs from SVRare with <10 calls across the 100K rare disease dataset, that overlapped coding regions of our nine multi-systemic ciliopathy disease genes of interest in the reverse phenotyping study, returned two real SVs (*ALMS1* paired-duplication inversion and *DYNC2H1* deletion) and two false positives (Best et al., 2022c).

We are aware of other platforms for automated SV filtering, for example Samplot (Belyeu et al., 2021), DeepSVFilter (Liu et al., 2021) and AquilaDeepFilter (Hu et al., 2022), which all use deep learning algorithms to determine between true and false positive SV calls. However, the closed research environment makes it difficult and cumbersome to import external scripts, and SVRare's use of the 100K dataset and opportunity to collaborate with its author who was familiar with the environment made it a more appealing option.

The manuscript from the authors of the SVRare script is not yet peer reviewed or published, however a pre-print available on medRxiv reports that SVRare identified 36 novel protein-coding disrupting SVs from a pilot study of 4313 100K families on diagnostic grade "green" genes that explain the probands' phenotype (Yu et al., 2022). Prior to SVRare analysis, only nine disease-causing SVs had been identified amongst these 4313 pilot participants, of which four were found outside the GEL automated diagnostic pipeline. Therefore, the authors estimate that SVRare can increase SV-based diagnosis yield at least 4-fold. All the SVs detected prior to SVRare analysis in the pilot patients were deletions. SVRare was successful in detecting multiple SV types, including three inversions and seven complex SVs. This is important, as one of the major benefits of WGS vs WES or clinical exome analysis is the opportunity to identify these otherwise undetectable SVs. In our own reverse phenotyping study, we found a paired-duplication inversion through the SVRare calls and manual inspection of the IGV plot (manuscript Figure 3A and 3B) (Best et al., 2022c).

From my limited experience, I think that that application of SVRare to the unsolved 100K cohort would be one of the fastest and easiest ways to boost diagnosis rates from WGS data in the mainstream diagnostic setting. It would not add unachievable workloads for the clinical scientists and, clearly, can add value by detecting a significant burden of pathogenic alleles that have not been systematically assessed. Furthermore, SVs involving multiple exons of disease-causing genes can usually be classified more easily as pathogenic, according to ACMG criteria, than other “hidden” variant types that may require more laborious functional validation to prevent VUS classification. This is an understandable motivation to direct resources towards SV analyses using tools such as SVRare, which are attractively high- throughput and likely to have high clinical utility.

5.3.5 Added value of splice variant analysis in 100K

In both of our 100K studies, we were able to identify new potentially pathogenic variants predicted to affect splicing by SpliceAI software (n=3 in CMC cohort; n=1 in reverse phenotyping study). One of these from the CMC cohort (*ARL6*: c.534A>G, p.Gln178=) turned out to have previously been published in association with BBS and proven to cause aberrant splicing on minigene assay, although it was not listed on ClinVar so it was not prioritised for analysis by any other filtering strategy (Maria et al., 2016). The other three all had to be classified as VUSs in the absence of functional analyses.

SpliceAI was selected as our *in-silico* prediction tool of choice because it was shown to perform as the best single strategy to prioritize rare genomic variants that affected splicing when compared to seven other algorithms (Rowlands et al., 2021). This study showed that by combining results from at least four tools and using a weighted average, accuracy can be further slightly improved. However, we did not think that this small improvement was worth the complication of compiling multiple predictions, as it was unlikely to significantly change our overall outcomes.

One strategy that could provide functional evidence of novel splicing effects for our identified splicing VUSs would require acquisition of patient RNA samples from tissues relevant for the disorder which could be analysed. This involves conversion of reverse transcription (RT) of RNA to cDNA before PCR amplification (RT-PCR) with primers designed to capture the impact of the variants on splicing. The products can then be compared to controls through gel electrophoresis or Sanger sequencing. Urinary renal

epithelial cells have proven to be a useful source to demonstrate alternative splicing events when derived from patients with renal ciliopathies (Molinari et al., 2020) and multi-systemic ciliopathies, such as JBTS (Ramsbottom et al., 2018). They are easily and readily accessible, and do not require any invasive procedures. Obtaining more inaccessible tissues is less practical, for example retinal samples for retinal ciliopathies, however blood-based RT-PCR has been used to successfully assay splicing in genes for which blood would not be an obvious disease-relevant tissue (Wai et al., 2020). The residual transcription of tissue-specific transcripts in blood cells reflects a phenomenon originally termed “illegitimate transcription,” which is thought to occur in virtually all cells (Chelly et al., 1989). Because an RT-PCR strategy is dependent upon provision of the right samples from the recruiting clinicians, our ~20% clinician response rate would have significantly limited our ability to achieve these research findings. Indeed, we did request participant blood and urine for all cases with predicted novel splice defects, but no samples were received.

Another option to functionally validate predicted splice variants, not requiring patient samples, would be minigene or midigenes assays. Minigenes and midigenes are circular plasmids, into which a region of interest can be inserted in both wild-type and variant forms. For splicing assessments, this usually includes an exon and flanking intronic sequence. Different versions can be generated by site-directed mutagenesis. When the plasmids are expressed in a cell-line, the splicing of the wild-type and variant forms can be compared to assess whether the variant has an effect on splicing (Lord and Baralle, 2021). This strategy is less reflective of the true splicing circumstances within the patient than the RT-PCR method, since the artificiality of the construct removes much of the larger context in which the variant occurs.

A systematic analysis of 38,688 individuals in the Rare Disease arm of 100K has recently been published, searching for potentially pathogenic new splice variants (Blakes et al., 2022). The authors looked for unsolved 100K participants with *de novo* SNVs at constrained regions near exon–intron boundaries and at putative splicing branchpoints in known disease genes. Variants were annotated with VEP by using the SpliceAI plugin, GEL tiering data, available phenotype data and participant outcome data to allow filtering and genotype-phenotype correlation analysis. From 258 candidate *de novo* splicing variants, they extracted 84 variants that were already considered to be diagnostic by GEL and 35 new likely diagnoses. At the time of publication, they had functionally confirmed a new diagnosis for four out of five cases for which RT-PCR studies from participant blood samples were conducted. This

interesting study shows that non-canonical splice defects are likely to be a significant source of missed diagnoses, but the bespoke, low-throughput functional validation step currently required is likely to hinder application of this strategy in the mainstream diagnostic setting until a faster system emerges.

RNA-sequencing (RNA-seq) offers a high-throughput and unbiased alternative to gene-specific splice variant validation, which can simultaneously detect and functionally characterise splicing variants in a transcriptome-wide manner (Putscher et al., 2021). Indeed, we understand that a whole-transcriptome RNA-seq pilot study is underway for unsolved 100K participants. Around 40% of 100K probands had RNA as well as DNA extracted from blood at recruitment, which has been frozen since the time of ascertainment. GEL are currently performing blood-based RNA-seq for around 5000 unsolved 100K participants, as well as a limited number of positive controls with known splice defects. When this data gets released, it will provide a rich resource for additional diagnoses and validation of existing VUSs predicted to impact splicing.

As for RT-PCR experiments, selection of appropriate tissue types is important because RNA-seq only provides meaningful results when sufficient levels of sequence coverage of a relevant gene transcript are found in the sampled tissue(s). A metric called the minimum required sequencing depth (MRSD) has been developed to determine the depth of sequencing required from RNA-seq to achieve user-specified sequencing coverage of a transcript, individual gene, or group of genes (Rowlands et al., 2022). Application of the MRSD metric across cultured fibroblasts, whole blood, lymphoblastoid cell lines and skeletal muscle showed that it can overcome transcript region-specific sequencing biases with high precision (90.1%–98.2%).

5.4 Lessons learned: functional VUS analyses

We developed parallel strategies with colleagues at UCD for functional interpretation of *TMEM67* missense VUSs by using CRISPR/Cas9 gene editing in a human hTERT-RPE-1 cell line and in *C. elegans*. Two known pathogenic, two known benign and eight VUSs from fetuses with the lethal ciliopathy MKS were selected for modelling. I generated a biallelic knockout *TMEM67* RPE-1 cell line using CRISPR/Cas9 and characterised it as a knockout by sequencing, western blotting, and high-content imaging. This, along with *TMEM67*-myc epitope-tagged plasmids that contained the selected variants of interest (generated by site-directed mutagenesis), were used for a genetic complementation assay that tested *TMEM67* signalling function and allowed

determination between benign and pathogenic alleles. This complimented quantitative phenotypic assays of sensory cilia structure and function in *C. elegans* done by our Irish colleagues. Together, these assays provided interpretation of three VUSs as benign and five as pathogenic.

On the promises of CRISPR editing being quick and easy, we had optimistically hoped to develop a high-throughput strategy for *TMEM67* VUS interpretation that could be transferable to the diagnostic setting. Although we did manage to interpret the eight *TMEM67* variants that we selected, we did not achieve anything close to a high-throughput system. In fact, generation and characterisation of the knockout cell-lines using CRISPR was probably the most straightforward part of this project, but still took several months. The development of the variant interpretation assay was much more challenging. It took two years of optimisation and troubleshooting to conclude that our initial plan to interpret variants through the effect on cilia number by high content imaging was not going to work, given the fragility of the *TMEM67* knockout line with accompanying cytotoxicity caused by transfection. Fortunately, we were able to develop an alternative strategy for VUS interpretation through the functional signalling assay, allowing us to complete the project and compliment the successful *C. elegans* interpretation system.

Other studies have been more successful in using high-content imaging to interpret VUSs in CRISPR knock out cell lines, for example for interpretation of *PRPF31* missense variants (Nazlamova et al., 2021). Pathogenic *PRPF31* variants cause RP11; the second most common cause of the dominant form of the degenerative retinal ciliopathy retinitis pigmentosa. Their methodology was very similar to ours; they generated a stable knockout *PRPF31*^{+/-} RPE-1 cell line using CRISPR and transfected in myc-DDK tagged variant *PRPF31* plasmids, generated using site-directed mutagenesis (Nazlamova et al., 2021). They included three known benign and three known pathogenic controls, and five *PRPF31* VUSs from ClinVar, and used tests of cilia number from high-content imaging to interpret the variants, However, their assays only provided ACMG supporting evidence in favour of benign impact for one VUS and in favour of pathogenic impact for one VUS, which, in the absence of clinical data, could not change the overall interpretation for any VUSs. This again shows the importance of pairing high-quality phenotyping data with variant data to allow definitive interpretation.

We know that heterozygous pathogenic *PRPF31* variants cause the adult onset, eye-only condition autosomal dominant RP11, whereas biallelic pathogenic *TMEM67* variants cause multi-systemic, severe developmental phenotypes. Based on the more extensive distribution and functional importance of *TMEM67* compared to *PRPF31*, we hypothesise that the heterozygous *PRPF31* knockout cell line was more overall stable and tolerant to transfection than our biallelic *TMEM67* knockout, which was too fragile and damaged.

The main advantages of *C. elegans* as a model system for this variant interpretation project are covered in the manuscript discussion (Lange et al., 2022) and in the introduction (thesis section 1.5.1.1). One of the major differences was the UCD team's ability to generate "knock-in" worms through injection of crRNA:tracrRNA:ssODN complexes directly into the gonads of young adult hermaphrodites, leading to stable expression of *mks-3* (the worm orthologue of the human *TMEM67* protein) in edited progeny containing homozygous variants of interest. They had done this successfully before for interpretation of *mksr-2/B9D2* variants, so already had their methodology optimized (Lange et al., 2021). This was much more streamlined and reflective of true biological conditions, than our knock-out and complementation approach. Another reason that the *C. elegans* system worked so well is that the team had several well-characterised quantitative assays of cilia structure and function established, suitable for high-throughput analysis (Sanders et al., 2015).

We did have one try at generating knock-in RPE-1 cell lines for three of our VUS through provision of an ssODN repair template alongside the crRNA:tracrRNA complexes but found no successful edits after sequencing of 100 clonal cell lines following FACs and therefore abandoned further attempts. We now know more about the difficulties for HDR-mediated gene-editing in this diploid cell-line. Other authors reported a failure of gene editing in immortalised RPE-1 cells during CRISPR-Cas9 'dropout' screens that included a panel of cell lines (Haapaniemi et al., 2018). They went on to show that genome editing by CRISPR-Cas9 induces a p53-mediated DNA damage response and cell cycle arrest, causing a selection against cells with a functional p53 pathway (Haapaniemi et al., 2018).

A downside of using *C. elegans* for broader SNV interpretation is the lack of widespread protein conservation between humans and *C. elegans*, limiting its scope.

20-40% of human genes have no *C. elegans* homolog (Kaletta and Hengartner, 2006), including our other major ciliopathy protein of interest, *CEP290* (Harris et al., 2020, Cunningham et al., 2022). *C. elegans* is a simple organism where the only ciliated cell type is the sensory neuron, making it very different to humans, in whom almost every cell type is ciliated (Malicki and Johnson, 2017). *C. elegans* do not possess any of the cilia-associated developmental signalling pathways of vertebrates, such as Wnt or Shh signalling, meaning they cannot be used to investigate these critical functions.

Another human option that we could have considered for VUS interpretation is stem cells. Stem cells are becoming increasingly popular for disease modelling, as they are a step closer to the model organism than immortalised cell lines. Induced pluripotent stem cell (iPSCs) can be engineered to contain specific genetic variants, facilitating analysis of the resulting phenotype from early embryonic developmental stages. Differentiation of stem cells into inaccessible tissue types, for example retina, provides the opportunity to extract otherwise unobtainable RNA samples for RNA-sequencing and transcriptomic analysis, which is being used to investigate retinal development, normal physiology, and disease (Zerti et al., 2020). Models of mammalian retina are particularly useful for studying ciliopathies with a retinal phenotype. Robust protocols are widely available for culture of human retinal organoids (Chichagova et al., 2019). These organoids form laminated, mature neural retina containing all major retinal cell types, including the elaboration of photoreceptor outer segments, and limited scotopic responsiveness to light. However, stem cells and derived organoids are significantly more expensive to derive, culture and differentiate than immortalised cell lines, which therefore continue to be used extensively for disease-modelling experiments.

5.5 Looking to the future

5.5.1 Clinical genomics era

We are in a time of real transformation from traditional clinical genetics towards mainstream genomics. This certainly offers huge opportunities for patient benefit, but also comes with significant ethical and practical challenges. As I have said repeatedly, much better training and resources are essential for both healthcare professionals ordering genomic tests and returning results to patients, and for clinical scientists interpreting variants and writing reports, to make this work. Genomic tests are being increasingly ordered by mainstream clinicians rather than clinical geneticists, who do not all have the same clinic time available, training in the complex issues required to obtain informed consent or understanding of complex results such as VUSs. This needs to be addressed within undergraduate and postgraduate medical education

programmes. As we move into this genomics era, we as clinical geneticists must embed ourselves within multi-disciplinary teams, to act as liaison between the labs, mainstream clinicians and the patients being tested.

5.5.2 Increased use of long-range sequencing

As a community, we are anticipating increased application of long-range sequencing technologies for cases that cannot be solved by short-read WGS analysis. These can routinely generate reads of up to 10kb. Having longer reads simplifies the task of reconstruction of true DNA molecules through alignment of parallel short reads (Amarasinghe et al., 2020). Therefore, this can improve mapping certainty, especially important for detection and characterisation of structural variants, accurate sequencing in repetitive regions and phasing of variants. However, this will require establishment of expensive new sequencing infrastructure and software analysis tools, and inevitably more training.

5.5.3 Newborn Genomes Programme

One of the major next steps to consider is introduction of the Newborn Genomes Programme, launched as a vision by GEL in 2021 with hopes to begin recruitment in 2023, and currently undergoing a consultation process involving specialists and the general public (Genomics England, 2021). They propose to sequence the genomes of 100,000 newborns born within the NHS, and screen them for a set of actionable genetic conditions with childhood onset. This would expand the current newborn screening programme from the heel-prick spot test, which tests for nine actionable conditions. The list of new actionable conditions under consideration has not yet been released. They also propose to add the genomic data and paired clinical data to the National Genomic Research Library, accessible by vetted academic, clinical, and biopharma healthcare researchers.

Although I understand the potential value added through detection of further actionable findings than is currently achieved, I have several concerns about this proposal. Firstly, I am curious about who will be consenting the patients, how long will be allocated for the consent discussion and how long prospective parents will have to consider whether to take part. If it is going to be midwives or health visitors taking consent, they will need a lot of training about the complexities of genomes analysis and data sharing with academic and commercial partners to obtain informed consent. This is especially

crucial when consenting parents on behalf of their newborn children, who do not have capacity to consent for themselves. If they anticipate a separate workforce to act as consenters, they need to allow time and money to recruit and train this team across the country. I am concerned that the immediate post-natal setting is not ideal for parents to consider this complex opportunity, where most are recovering from delivery, distracted by learning how to look after their newborns, and usually sleep-deprived. I would hope that the subject would be introduced well ahead of delivery, with provision of written and online resources in appropriate lay language to consider with more time.

Having worked briefly as a 100K consentor, I worry that we were not accurate enough in what we were consenting patients for. I certainly promised patients more than has been delivered, stating that we would be analysing their whole genomes to look for the explanation of their rare disease. As this project shows, this certainly has not been the case, especially when researchers have not got involved for unsolved cases. In reality, most 100K patients have received an inferior virtual gene-panel assessment than would have been done in the mainstream diagnostic setting from a clinical exome or WES, which has taken many more years to deliver largely negative, and frequently falsely reassuring results.

The timeline to return results in the Newborn Genomes Programme will be much more important than in 100K as they are medically actionable in childhood. It would need to be very clear in the protocol who would be expected to return the results and facilitate the required medical action to make this project worthwhile and ethical. A survey of referring clinicians to the Next Generation Children's (NGC) project, which performed WGS for 521 young, seriously ill children, reports significant challenges about the additional communication about genetic testing and uncertainty about explaining genetic results to parents, accompanied by an increased workload (French et al., 2022). This demonstrates that additional consultation time and training must be accounted for the clinicians returning results from the Newborn Screening Programme during the of planning this project. Furthermore, with huge backlogs on standard testing already, I worry that time-sensitive analysis and return of additional results from the Newborn Genomes Programme will be an unsustainable pressure for clinical scientists in the diagnostic laboratories without significant additional recruitment and training.

The Newborn Genomes Programme proposal sparks an interesting debate about who

owns our genetic information. Although genomic data can offer great benefits, it is also susceptible to abuse, for example discrimination for health insurance caused by predictive tests for medical conditions that could be costly. This needs to be addressed within the consent procedures. The “Code on Genetic Testing and Insurance” was published in October 2018, containing a voluntary agreement between government and the Association of British Insurers (ABI) to “never require or pressure any applicant to undertake a predictive or diagnostic genetic test, and only consider the result of a predictive genetic test for a very small minority of cases” (Gov.uk, 2019). This agreement is open-ended, with no expiry date, but could certainly change within the lifetime of a newborn child. I would want to see clear guidelines on the age at which the recruited children would be informed that they were part of the study, how that would be explained in age-appropriate language, and how they could choose to opt-out if they wanted to. I would also be very eager to hear that, should these children present with suspected genetic disorders in the future, their existing genome would be accessible for analysis rather than repeating the chemistry and analysis, as is currently happening done for 100K participants. I hope that many of the lessons that we have learned from 100K will be taken forward to improve future studies such as the Newborn Genomes Programme.

Overall, I have benefitted hugely from this extended research experience, which I am confident will make me a better clinician genomicist in the future.

5.6 Manuscript: Unlocking the potential of the UK 100,000 Genomes Project - lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort



Received: 8 December 2021 | Revised: 2 February 2022 | Accepted: 9 March 2022

DOI: 10.1002/ajmg.c.31965

COMMENTARY



Unlocking the potential of the UK 100,000 Genomes Project—lessons learned from analysis of the “Congenital Malformations caused by Ciliopathies” cohort

Sunayna Best^{1,2} | Chris F. Inglehearn¹ | Christopher M. Watson³  |
Carmel Toomes¹ | Gabrielle Wheway^{4,5} | Colin A. Johnson¹ 

¹Division of Molecular Medicine, Leeds Institute of Medical Research at St. James's, University of Leeds, St. James's University Hospital, Leeds, UK

²Yorkshire Regional Genetics Service, Leeds, UK

³North East and Yorkshire Genomic Laboratory Hub, Central Lab, St. James's University Hospital, Leeds, UK

⁴University Hospital Southampton NHS Foundation Trust, Southampton, UK

⁵Faculty of Medicine, Human Development and Health, University of Southampton, Southampton, UK

Correspondence

Colin A. Johnson, Division of Molecular Medicine, Leeds Institute of Medical Research at St. James's, University of Leeds, St. James's University Hospital, Leeds LS9 7TF, UK.

Email: c.johnson@leeds.ac.uk

Funding information

Medical Research Council, Grant/Award Numbers: MR/M000532/1, MR/T017503/1; Wellcome Trust

We reviewed sequencing, variant and clinical data from patients recruited to the “Congenital Malformations caused by Ciliopathies” (CMC) cohort of the UK 100,000 Genomes Project (100K) (Best et al., 2021).¹ By using domain-specific knowledge of ciliopathy genetics (Reiter & Leroux, 2017; Wheway et al., 2019), and examining variants in non-ciliopathy disease genes, we were able to identify potentially causative variants beyond those reported by the triaging process implemented by Genomics England (GEL, the company set up to run 100K). As a result, we increased diagnoses from the 27/83 (32.5%) that were reported by GEL, to 43/83 (51.8%). During this work, we experienced several difficulties in accessing and working with the data and observed several limitations with the currently available datasets. Here, we review these issues, suggest ways in which 100K data could be made more accessible and utilized more fully for patient benefit, and propose lessons that can be learned for future large-scale human genomics studies.

The issues are grouped into four broad categories: those relating to the clinical information available for recruited individuals; issues relating to the triaging and prioritization process for variants (so-called “tiering”); difficulties experienced using the secure GEL research environment; and difficulties in reporting pertinent research findings back to recruiting clinicians.

1 | PHENOTYPING ISSUES

In the early stages of recruitment to 100K, recruiters were required to comply with strict entry criteria. These included pre-screening of the key genes or gene panels relevant to the participant's condition, the recruitment of parent-child trios and adherence to a complex, time-consuming process for the uploading of Human Phenotype Ontology (HPO) terms. However, pressure to recruit from busy NHS clinics led to relaxation of requirements for pre-screening and trio recruitment, and frequently resulted in sparse HPO term usage, with patient phenotypes often described using only one or two terms from one organ system. The choice of organ system may have reflected the interests and expertise of the recruiting clinician: for instance, many participants in the CMC cohort were recruited under solely vision-related terms such as rod-cone dystrophy, with limited or absent information about extra-ocular, syndromic features. As a result, the relevance of HPO terms varies across the cohort, ranging from accurate and highly informative to unhelpful or even misleading. Additional data from longitudinal patient records are accessible using the “Participant Explorer” tool, but these are available only in a proportion of cases, are of variable quality and are not collated in a form that can be readily used for phenotype-genotype correlation and variant prioritization.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2022 The Authors. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics* published by Wiley Periodicals LLC.

The accuracy of HPO descriptions has a direct impact on diagnostic success. 100K was configured to communicate results only from one or more virtual panels of genes that are defined in the GEL PanelApp database as relevant to a participant's suspected condition, based on entered HPO terms (although there is also ethical approval for broader variant screening on a research basis). The selection of gene panels is therefore largely dependent on the HPO descriptions. Thus, inappropriate HPO descriptions will inevitably lead to inappropriate gene panel selection and therefore to missed diagnoses because the correct disease gene(s) have not been analyzed. For example, a participant in the "epilepsy plus other features" category, with keratoconus and epilepsy as the entered HPO terms, was found to have bi-allelic pathogenic *CEP290* variants. In a reverse phenotyping approach following contact with the recruiting clinician, it emerged that this participant had key ophthalmological features that were not entered during recruitment to 100K, comprising a formal diagnosis of Leber Congenital Amaurosis.

2 | TIERING ISSUES

The GEL tiering system prioritizes variants for analysis by regional NHS diagnostic laboratories. Clinical assessment is only expected for prioritized Tier 1 and 2 single nucleotide variants (SNVs) or Tier A structural variants, which are provided in a report. Tiered variants are primarily limited to those variants affecting coding sequences and splice donor or acceptor sites. These are rare protein damaging (Tier 1) or protein altering (Tier 2) variants in genes on selected panel(s) in which the allelic state matches the known mode of inheritance for the gene and disorder, and segregates with disease where familial sequence data is available. Copy number variants and structural variants have been classified Tier A (>10 kb in appropriate PanelApp genes) or Tier Null in cases recruited toward the end of the project, but these have not yet been systematically analyzed in the whole cohort.

Rare SNVs in genes not on the selected panel(s) are classed as Tier 3. These include variants known or predicted to be pathogenic but not in a relevant PanelApp gene, or in a relevant gene but considered insufficient to explain the phenotype, such as a heterozygous variant in a gene implicated in recessive disease. All other variants are un-tiered (although white-listing of known pathogenic variants is an area of active development). Tier 3 and un-tiered variants are not inspected routinely by NHS diagnostic labs, and left to external researchers to consider more fully, if at all. In our own work (Best et al., 2021), we identified 11/83 probands (13.3%) with research molecular diagnoses with at least one variant outside of tiers 1 and 2. Five tier 3 variants and 12 untiered variants contribute to the diagnoses for these 11 participants. Furthermore, no attempt has yet been made to prioritize less obvious splicing defects using SpliceAI (Jaganathan et al., 2019) or a similar program, or to analyze variants in intronic regions. One obvious limitation of reporting based on these partial analyses is that many recessive alleles appear monoallelic

because the second allele is a structural, splice or intronic variant missed by the current GEL pipeline. These single recessive pathogenic alleles in relevant PanelApp genes will then be classed as Tier 3 and not prioritized for analysis because they alone cannot explain the participant's phenotype.

Anecdotally, we understand that some participants were recruited because a diagnostic laboratory had previously identified one variant in a relevant recessive gene, and the referring clinician anticipated that genome analysis would reveal the second. Instead, the eventual report was negative, lacking even the known variant, leading to confusion for clinicians and clinical scientists. Given the ever-increasing demand for genetic testing, there seems little likelihood that NHS laboratories will have the operational flexibility to reassess these data in response to improvements in the GEL variant detection pipeline. In practice, therefore, although participants have a whole genome sequence, variant identification is typically no better than a targeted gene panel analysis.

3 | USING THE GEL SECURE RESEARCH ENVIRONMENT

Given the limitations of the variant identification and triaging carried out by GEL itself, any further screening is dependent on individual researchers revisiting the data on a research basis. Our experience of the GEL secure research environment is that it can be a frustrating and uninviting area within which to work. Service interruption is not infrequent and can lead to work disruption and data loss. Scripts must be self-contained for security reasons and must be security checked before importing, meaning users tend to work with and adapt what is already there rather than importing alternative tools and pipelines that are more fit for purpose. Opportunities for training are limited, meaning the aspiring genomics researcher is often dependent on generous collaborators who are already familiar with the research environment and are willing to share their skills and code. Use of the Linux command line is required for several investigative strategies within the GEL research environment, which is unfamiliar and intimidating to many inexperienced clinicians and scientists and requires significant time investment to master. The lack of a forum for script sharing, advice and learning from others seems a significant omission. An MSc program in Genomic Medicine was intended to address this deficit, but many of the funded programs completed before the data was released, missing an opportunity for hands-on training within the GEL secure research environment. We accept that many of these issues arise from the need to protect patient data, which will in turn limit the scope for changes to the GEL research environment. Nevertheless, these difficulties have the effect of making training and collaboration more difficult and are a further disincentive to those wishing to work with 100K data. That many still do is a testament to the huge potential research value of this resource, but any efforts to make it more accessible could significantly enhance exploitation and patient benefit.

4 | REPORTING PROBLEMS

We encountered significant problems disseminating identified diagnoses to recruiting clinicians, which limited the returning of results to patients and publication of findings. Reporting of research findings must be carried out through the 100,000 Genomes Airlock system using the “Researcher Identified Potential Diagnosis” and “Clinician Contact” process. The researcher submits their findings to this system using a request form, which is sent to the recruiting clinician, who remains anonymous unless and until they choose to respond. In our experience, the response rate is less than 20%. This may reflect the time that has elapsed since recruitment (2013–2018), meaning that some clinicians may have moved post.

Such a low response rate is another major obstacle to research on 100K cohorts. Researchers can publish un-identifiable overview findings without involving recruiting clinicians, but must obtain consent from clinicians and participants before publishing detailed individual phenotypic data. Limited engagement by recruiting clinicians at best restricts, but may even prevent, the publication of findings, a major driver of research activity. Furthermore, researchers are unable to assess detailed phenotyping data or to obtain additional clinical samples from patients or relatives that could help segregation testing or functional analyses of variants. These issues limit researchers’ opportunities to interpret the pathogenicity of variants, further reducing opportunities to benefit patients by making a definitive molecular diagnosis, and to publish.

5 | FUTURE USE OF 100K DATA

During the period 2016–2018, many clinicians were encouraged to recruit to this project in preference to local clinical exome screening on the basis that it was a more comprehensive test. Screening to date has fallen well short of that promise, and despite the predicted 1 to 2-year turnaround, reports have still not been issued for some patients. Nevertheless, the 100K dataset remains a powerful resource of immense value to patients, clinicians and researchers, both in the UK and globally. Whole genome sequence data can be revisited indefinitely, reducing the need for expensive and sometimes invasive serial tests frequently required in the “diagnostic odyssey” for patients with genetic diseases.

We suggest that a more agnostic approach to gene panel selection, like that used by the Deciphering Developmental Disorders project, rather than one driven narrowly by HPO term usage, would be beneficial. This approach would permit analysis of additional panels of genes with broadly overlapping phenotype ranges if an answer is not obtained from the relevant PanelApp gene panel, or if phenotyping data are not well documented. Reanalysis should also include approaches to identify variants likely to alter splicing and likely pathogenic structural variants, for example using SpliceAI and the SVRare suite of programs (Yu et al., 2021). This broader approach could identify “second hits” in relevant genes that appear to be monoallelic for tiered variants, and remain refractory to current strategies. Additionally, increasing accessibility for research teams around the world and reporting of new research-based findings could reap further benefits for patients and clinicians. Updating

the security software could make it easier to access and use, especially for research-minded clinicians, without compromising security risks.

To derive maximum benefit from these efforts, lines of communication between researchers and clinicians should be improved. This may require an overhaul and update of the database of recruiting clinician contact details held by GEL, with new contacts established when clinical responsibility changes hands. In our experience, when the recruiting clinician did respond, the information they supplied proved invaluable in confirming molecular diagnoses. Often, many additional clinical features which had not been listed in the entered HPO terms were provided, which facilitated more accurate genotype-phenotype correlation and greater diagnostic confidence. All new findings, whether generated through reanalysis by GEL or by researchers applying domain-specific knowledge, would still need accredited diagnostic confirmation, so additional staff and resources for service testing are also essential.

As well as addressing issues within the existing study, the experience of those involved in 100K can inform future large-scale human genomics studies. The use of HPO terms to describe and define phenotypes, if applied effectively, could facilitate an AI-based, phenotype-informed variant prioritization approach. A simple, comprehensively applied HPO term entry system could significantly enhance the value of any future human genome resource.

In summary, the ciliopathies provide an exemplar group of disorders that illustrate both the challenges and opportunities of working with 100K datasets (Best et al., 2021; Wheway et al., 2019). 100K remains an immensely valuable clinical and scientific resource with huge potential for patient benefit, but that benefit has not yet been fully realized. There is an urgent need for re-evaluation of the data in light of improvements in genome interpretation technologies. Additional understanding could also be gained from research activity, which would benefit from efforts to simplify access, and train and support more researchers in using the data.

ACKNOWLEDGMENTS

Sunayna Best acknowledges support from the Wellcome Trust 4Ward North Clinical PhD Academy (ref. 203914/Z/16/Z). Gabrielle Wheway acknowledges support from Wellcome Trust Seed Award (ref. 204378/Z/16/Z). Colin A. Johnson acknowledges support from MRC project grants MR/M000532/1 and MR/T017503/1.

CONFLICT OF INTEREST

None.

ENDNOTE

¹Published 2021 with data from Main Program Release 11 (dated December 17, 2020). At that time, 16/83 of the “Congenital Malformations caused by Ciliopathies” cohort (19.3%) did not have a complete GMC exit questionnaire.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Christopher M. Watson  <https://orcid.org/0000-0003-2371-1844>

Colin A. Johnson  <https://orcid.org/0000-0002-2979-8234>

REFERENCES

- Best, S., Lord, J., Roche, M., Watson, C. M., Poulter, J. A., Bevers, R. P. J., ... Genomics England Research Consortium. (2021). Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 genomes project. *Journal of Medical Genetics*. <https://doi.org/10.1136/jmedgenet-2021-108065>
- Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J. F., Darbandi, S. F., Knowles, D., Li, Y. I., ... Farh, K. K. (2019). Predicting splicing from primary sequence with deep learning. *Cell*, *176*(3), 535–548. <https://doi.org/10.1016/j.cell.2018.12.015>
- Reiter, J. F., & Leroux, M. R. (2017). Genes and molecular pathways underpinning ciliopathies. *Nature Reviews Molecular Cell Biology*, *18*(9), 533–547. <https://doi.org/10.1038/nrm.2017.60>
- Wheway, G., Mitchison, H. M., & Genomics England Research Consortium. (2019). Opportunities and challenges for molecular understanding of

- ciliopathies - the 100,000 genomes project. *Frontiers in Genetics*, *10*, 127. <https://doi.org/10.3389/fgene.2019.00127>
- Yu, J., Szabo, A., Pagnamenta, A. T., Shalaby, A., Giacomuzzi, E., Taylor, J., ... Genomics England Research Consortium. (2021). SVRare: Discovering disease-causing structural variants in the 100K genomes project. *MedRxiv*. <https://doi.org/10.1101/2021.10.15.21265069>

How to cite this article: Best, S., Inglehearn, C. F., Watson, C. M., Toomes, C., Wheway, G., & Johnson, C. A. (2022). Unlocking the potential of the UK 100,000 Genomes Project—lessons learned from analysis of the “Congenital Malformations caused by Ciliopathies” cohort. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *190*:5–8. <https://doi.org/10.1002/ajmg.c.31965>

5.7 Manuscript: Re: Best et al., 'Unlocking the potential of the UK 100,000 Genomes Project - Lessons learned from analysis of the "Congenital malformations caused by ciliopathies" cohort

Received: 25 May 2022 | Revised: 21 June 2022 | Accepted: 27 June 2022

DOI: 10.1002/ajmg.a.62909

CORRESPONDENCE

AMERICAN JOURNAL OF PART **medical genetics** **A** WILEY

Re: Best et al., 'Unlocking the potential of the UK 100,000 Genomes Project – Lessons learned from analysis of the “Congenital malformations caused by ciliopathies” cohort’

To the editor,

The Genomics England (GEL) 100,000 Genomes Project is a landmark project in translational genomics research, established in 2014 with the primary goal of assessing the potential of whole genome sequencing to benefit particularly patients and families affected with rare diseases or cancer. Delivered as a partnership between GEL and NHS England, the project was highly ambitious for the time, and unprecedented in its scale and vision. Delivering the project required development of major new sequencing and analytical capabilities, as well as new clinical partnerships and logistical solutions to establish the close clinical interfaces required. The project was performed not as a standalone research project, but alongside busy clinical services. This created its own set of challenges, but also strengths in regards to informing development of future services for example. The project has proven to be a huge success, leading to provision of diagnoses for rare disease families, and genomics-informed personalized medicine recommendations for large numbers of cancer patients, directly leading to health benefits for tens of thousands of people. Perhaps more importantly, it laid the foundations for the establishment of whole genome sequencing in routine care nationally as part of the NHS Genomic Medicine Service in England, and informed the development of other similar initiatives worldwide.

Unsurprisingly there have been major learnings as the project progressed about many aspects of its design and delivery. GEL has benefited immeasurably from the generous support and advice we have received both from our patient partners, and the clinical-research community through our Genomics England Clinical Interpretation Partnerships (GECIPs). Their suggestions and those from others mean that our systems continue to improve and have brought diagnoses that had not yet been made. This model we continue to believe is crucial and generalizable, with 14% of diagnoses coming from additional researcher input in our recent study across the range of rare disease participants in our pilot program (100,000 Genomes Project Pilot Investigators, 2021).

Best and colleagues make numerous helpful suggestions arising from their experience working with 100,000 Genomes Project data, noting that the dataset they were using is now 16 months old, a period in which many changes have been implemented which we believe resolve several of the issues raised (Best et al., 2022).

With recruitment occurring in the context of busy clinical environments, comprehensiveness of data collection such as of HPO terms was less complete than had the study been performed as a formal research

program. GEL continues to put effort into increasing the clinical data available on 100,000 Genomes participants. We are also investigating the utility of additional data such as proteomic and transcriptomic datasets, histopathology and imaging, and novel technologies such as long-read sequencing, to improve diagnostic rates and increase the research value of the data held in the National Genomics Research Library.

GEL works closely with NHS England, NHS laboratories, and diverse clinical and research groups, to benefit patients through assisting with diagnosis, requiring good interfaces between researchers and the caring for clinicians for each patient. Tracking those clinicians is not straightforward given the long period since most of the patients were recruited, and the natural mobility of both patients and their clinicians. We have introduced several initiatives to ease and accelerate researcher-clinician interactions, including employing additional support staff and introducing new software to ease and simplify the processes involved. Researchers are now notified when Clinical Collaboration Requests are made and are asked to notify GEL if no response is received. Perhaps encouraged by these changes, we have seen a significant upswing in researcher-clinician contact requests (from an average of 48 per month in 2021 to an average of 61 per month in 2022).

Learning from the 100000 Genomes Project, GEL has designed and implemented a new analytical bioinformatics pipeline with improved variant calling performance. The variant interpretation strategies used by GEL and its NHS partners have evolved over time, aiming to maximize sensitivity for diagnostic variants while not overloading the service capacity with variants of uncertain clinical significance. Along with improvements in the GEL variant prioritization (tiering) pipeline, current interpretation guidelines include variants that were not routinely considered during the 100,000 Genomes Project. In addition to the variants prioritized as Tier 1 and Tier 2, all de novo variants in Tier 3 and the best candidate variants prioritized by Exomiser (Smedley et al., 2015) are reviewed, along with copy number variants and short tandem repeat expansions. Analysis of diagnoses identified for 100,000 Genomes Project participants that were not originally prioritized in the top two tiers by the GEL interpretation pipeline has demonstrated that improvements in scientific knowledge of genotype-phenotype correlations since the time of the initial interpretation accounts for biggest proportion of variants. This emphasizes the benefits of targeted reanalysis and the critical role of research activities in improving overall diagnostic yield, in identifying both new disease associations and putative new diagnoses.

GEL's research environment has also significantly evolved over the years, and undoubtedly the experience Best et al. report was widespread early on. In the last 2 years GEL has moved our infrastructure to the cloud, and, working with Lifebit, has developed a completely new research environment designed for ease of use, particularly by researchers who are not skilled programmers. There is a dedicated team within GEL to provide comprehensive support for users, which conducts regular live online training sessions and provides comprehensive guidance documentation (URLs for these resources below). Unfortunately the live sessions did not occur over the peak of the pandemic when Best et al.'s work in our research environment was ongoing, but we are delighted to say that these have recently restarted.

These improvements have led to a marked increase in the number of our Rare Disease families receiving diagnoses as a result of new research (from 357 in 2021 to 504 in the first 2 months of 2022). GEL is continually developing its protocols and tools to improve performance of this critical task. We are very appreciative of contributions such as those made by Best and colleagues, supporting us in this challenge, and welcome any further feedback.

URLs for GEL research environment assistance: Tutorials: <https://research-help.genomicsengland.co.uk/display/GERE/Research+Environment+Training+Sessions>. Documentation: <https://research-help.genomicsengland.co.uk/display/GERE/Research+Environment+User+Guide>. Assistance: <https://research-help.genomicsengland.co.uk/display/GERE/8.+Getting+help+and+resolving+problems>

CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Matthew A. Brown 
 Christopher Wigley
 Susan Walker 
 Deborah Lancaster 
 Augusto Rendon 
 Richard Scott 

Genomics England Ltd, London, UK

Correspondence

Matthew Brown, Genomics England Ltd, Charterhouse Square,
 London, UK.
 Email: matt.brown@genomicsengland.co.uk

ORCID

Matthew A. Brown  <https://orcid.org/0000-0003-0538-8211>
 Susan Walker  <https://orcid.org/0000-0002-5016-6426>
 Deborah Lancaster  <https://orcid.org/0000-0001-8512-3808>
 Augusto Rendon  <https://orcid.org/0000-0001-8994-0039>
 Richard Scott  <https://orcid.org/0000-0002-9113-2978>

REFERENCES

- 100000 Genomes Project Pilot Investigators, Smedley, D., Smith, K. R., Martin, A., Thomas, E. A., McDonagh, E. M., Cipriani, V., ... Caulfield, M. J. (2021). 100,000 genomes pilot on rare-disease diagnosis in health care. *New England Journal of Medicine*, 385(20), 1868–1880.
- Best, S., Inglehearn, C. F., Watson, C. M., Toomes, C., Whewey, G., & Johnson, C. A. (2022). Unlocking the potential of the UK100,000 genomes project – Lessons learned from analysis of the “congenital malformations caused by ciliopathies” cohort. *American Journal of Medical Genetics*, 190C, 5–8.
- Smedley, D., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Schubach, M., Siragusa, E., Zemojtel, T., Buske, O. J., Washington, N. L., Bone, W. P., Haendel, M. A., & Robinson, P. N. (2015). Next-generation diagnostics and disease-gene discovery with the exomiser. *Nature Protocols*, 10(12), 2004–2015.

6 Appendices

6.1 Published manuscript supplementary materials

6.1.1 Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

J Med Genet

Supplementary tables

Supplementary Table 1. Summary of genetic diagnosis rates of various ciliopathies using different next generation sequencing approaches.

Patient group	Sequencing approach	Genetic diagnosis rate	Year	Reference
10 families with NPHP	homozygosity mapping + WES	70%	2014	(1)
Syndromic ciliopathies	Gene panel, SNP genotyping, targeted sequencing of candidate genes	62%	2015	(2)
375 families with JBTS	Gene panel	32%	2015	(3)
Ciliopathy patients in a large rare disease cohort	WES	44%	2015	(4)
79 suspected NPHP cases	WES	40.5%	2016	(5)
43 patients with NPHP who had had the 5 most common NPHP genes excluded as a cause of their disease	Gene panel	16.3%	2016	(6)
26 patients with JBTS or MKS	WES with split read mapping	46%	2016	(7)
6 BBS patients	WES	67%	2017	(8)
100 families with JBTS	Gene panel sequencing + WES	94%	2017	(9)
56 families in whom PCD was suspected	WES	68%	2020	(10)
125 families with ciliopathies	WGS	87%	2020	(11)
29 families with skeletal ciliopathies	WGS + structural variant screening + RNA analyses	90%	2021	(12)

Supplementary Table 2. Diagnostic grade “green” genes in Rare Multisystem Ciliopathy Disorders PanelApp panel V1.139

Gene name	Mode of inheritance	Associated phenotypes	Ensembl Id
AHI1	Biallelic	JBTS	ENSG00000135541
ALMS1	Biallelic	BBS, Alström	ENSG00000116127
ANKS6	Biallelic	PKD, NPHP	ENSG00000165138
ARL13B	Biallelic	JBTS	ENSG00000169379
ARL6	Biallelic	BBS	ENSG00000113966
ARMC9	Biallelic	JBTS	ENSG00000135931
B9D2	Biallelic	JBTS, MKS	ENSG00000123810
BBS1	Biallelic	BBS	ENSG00000174483
BBS10	Biallelic	BBS	ENSG00000179941

BBS12	Biallelic	BBS	ENSG00000181004
BBS2	Biallelic	BBS	ENSG00000125124
BBS4	Biallelic	BBS	ENSG00000140463
BBS5	Biallelic	BBS	ENSG00000163093
BBS7	Biallelic	BBS	ENSG00000138686
BBS9	Biallelic	BBS	ENSG00000122507
C21orf2	Biallelic	JATD, Spondylometaphyseal dysplasia, IRD	ENSG00000160226
C2CD3	Biallelic	SRPS, JATD, OFD	ENSG00000168014
C5orf42	Biallelic	JBTS, OFD	ENSG00000197603
CC2D2A	Biallelic	JBTS, MKS, COACH	ENSG00000048342
CENPF	Biallelic	Stromme	ENSG00000117724
CEP104	Biallelic	JBTS	ENSG00000116198
CEP120	Biallelic	SRTD, CED, JATD	ENSG00000168944
CEP164	Biallelic	NPHP, SLS	ENSG00000110274
CEP290	Biallelic	JBTS, MKS, COACH, SLS	ENSG00000198707
CEP41	Biallelic	JBTS	ENSG00000106477
CEP83	Biallelic	NPHP	ENSG00000173588
CRB2	Biallelic	PKD with ventriculomegaly	ENSG00000148204
CSPP1	Biallelic	JBTS, MKS	ENSG00000104218
DDX59	Biallelic	OFD	ENSG00000118197
DHCR7	Biallelic	SLO	ENSG00000172893
DYNC2H1	Biallelic	SRTD, CED, JATD	ENSG00000187240
DYNC2L1	Biallelic	SRTD	ENSG00000138036
EVC	Biallelic	EVC, WAD	ENSG00000072840
EVC2	Biallelic	EVC, WAD	ENSG00000173040
GLI3	Monoallelic	JBTS, SLS	ENSG00000106571
HNF1B	Monoallelic	PKD, NPHP	ENSG00000275410
HYLS1	Biallelic	JBTS, Hydrolethalus syndrome	ENSG00000198331
ICK	Biallelic	Endocrine-cerebro-osteodysplasia	ENSG00000112144
IFT122	Biallelic	CED	ENSG00000163913
IFT140	Biallelic	SRTD, JATD, Mainzer-Saldino	ENSG00000187535
IFT172	Biallelic	RP, SRTD, JATD, Mainzer-Saldino, SRTD	ENSG00000138002
IFT27	Biallelic	? BBS	ENSG00000100360
IFT43	Biallelic	SRTD, CED, Sensenbrennar syndrome	ENSG00000119650
IFT52	Biallelic	SRTD	ENSG00000101052
IFT74	Biallelic	? BBS	ENSG00000096872
IFT80	Biallelic	SRTD, JATD	ENSG00000068885
INPP5E	Biallelic	JBTS	ENSG00000148384
INVS	Biallelic	NPHP, SLS	ENSG00000119509
IQCB1	Biallelic	SLS	ENSG00000173226

KIAA0586	Biallelic	JBTS, SRTD	ENSG00000100578
KIAA0753	Biallelic	OFD, SRTD, JBTS	ENSG00000198920
KIF7	Biallelic	JBTS, Acrocallosal syndrome	ENSG00000166813
LZTFL1	Biallelic	BBS	ENSG00000163818
MAPKBP1	Biallelic	NPHP	ENSG00000137802
MKKS	Biallelic	BBS	ENSG00000125863
MKS1	Biallelic	MKS, BBS, JBTS	ENSG00000011143
NEK1	Biallelic	SRTD	ENSG00000137601
NEK8	Biallelic	NPHP	ENSG00000160602
NPHP1	Biallelic	NPHP, JBTS, SLS	ENSG00000144061
NPHP3	Biallelic	MKS, SLS, NPHP	ENSG00000113971
NPHP4	Biallelic	NPHP, SLS	ENSG00000131697
OFD1	X-linked	JBTS, OFD	ENSG00000046651
PIBF1	Biallelic	JBTS	ENSG00000083535
PKD1	Monoallelic and biallelic	PKD	ENSG00000008710
PKD2	Monoallelic	PKD	ENSG00000118762
PKHD1	Biallelic	Polycystic kidney and hepatic disease	ENSG00000170927
PMM2	Biallelic	Congenital disorder of glycosylation	ENSG00000140650
RPGRIPL1	Biallelic	JBTS, MKS	ENSG00000103494
SBDS	Biallelic	Skeletal ciliopathies	ENSG00000126524
SCLT1	Biallelic	OFD, SLS	ENSG00000151466
SDCCAG8	Biallelic	BBS, SLS	ENSG00000054282
TCTEX1D2	Biallelic	SRTD, JATD	ENSG00000213123
TCTN1	Biallelic	JBTS	ENSG00000204852
TCTN2	Biallelic	JBTS, MKS	ENSG00000168778
TCTN3	Biallelic	OFD, JBTS, MKS, Mohr-Majewski syndrome	ENSG00000119977
TMEM107	Biallelic	MKS, OFD, ? JBTS	ENSG00000179029
TMEM138	Biallelic	JBTS	ENSG00000149483
TMEM216	Biallelic	JBTS, MKS	ENSG00000187049
TMEM231	Biallelic	JBTS, MKS	ENSG00000205084
TMEM237	Biallelic	JBTS	ENSG00000155755
TMEM67	Biallelic	JBTS, MKS, COACH, NPHP, Senior-Boichis syndrome, ? BBS	ENSG00000164953
TRAF3IP1	Biallelic	SLS	ENSG00000204104
TTC21B	Biallelic	NPHP, SRTD, JATD	ENSG00000123607
TTC8	Biallelic	BBS	ENSG00000165533
TXNDC15	Biallelic	MKS	ENSG00000113621
VPS13B	Biallelic	Cohen syndrome	ENSG00000132549
WDPCP	Biallelic	MKS, ?BBS	ENSG00000143951
WDR19	Biallelic	CED, SRTD, JATD, NPHP, SLS	ENSG00000157796

WDR34	Biallelic	SRTD, JATD	ENSG00000119333
WDR35	Biallelic	CED, SRTD	ENSG00000118965
WDR60	Biallelic	SRTD, JATD	ENSG00000126870
ZSWIM6	Monoallelic	Acromelic frontonasal dysostosis	ENSG00000130449
ISCA-37405-Loss	Biallelic	JBTS, NPHP	
ISCA-37432-Loss	Monoallelic	Renal cysts and diabetes syndrome, Autism Spectrum Disorder, Mayer-Rokitansky-Kuster-Hauser syndrome	

Abbreviations: BBS: Bardet Biedl syndrome; CED: Cranioectodermal dysplasia; COACH: cerebellar vermis hypo/aplasia, oligophrenia, ataxia, ocular coloboma, and hepatic fibrosis; EVC: Ellis-Van-Creveld syndrome; IRD: Inherited retinal dystrophy; JATD: Jeune asphyxiating thoracic dystrophy; JBTS: Joubert syndrome; LCA: Leber's congenital amaurosis; MKS: Meckel Gruber Syndrome; NPHP: nephronophthisis; PKD: Polycystic kidney disease; RP: Retinitis Pigmentosa; SLO: Smith Lemli Optiz syndrome; SLS: Senior Loken syndrome; SRPS: Short rib polydactyly syndrome; SRTD: Short rib thoracic dystrophy; WAD: Weyers acrofacial dysostosis

Supplementary Table 2. Candidate gene list provided alongside Rare Multisystem Super Ciliopathy panel v4.91 for Extract_hets Python script.

Gene symbol	Source	Ensembl Id
ACVR2B	PanelApp RMCD red gene	ENSG00000114739
ADAMTS10	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000142303
ADAMTS9	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000163638
ADCY3	On SCGS V1 list	ENSG00000138031
ADGRV1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000164199
AIPL1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000129221
AK7	On SCGS V1 list	ENSG00000140057
AK8	On SCGS V1 list	ENSG00000165695
ARF4	On SCGS V1 list	ENSG00000168374
ARL3	On SCGS V1 list	ENSG00000138175
ARMC4	PanelApp RMCD red gene	ENSG00000169126
ASAP1	On SCGS V1 list	ENSG00000153317
ATXN10	PanelApp RMCD red gene + SCGS V1 list	ENSG00000130638
B9D1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000108641
BBIP1	PanelApp RMCD red gene	ENSG00000214413
C21orf59	PanelApp RMCD red gene	ENSG00000159079
C2orf71	PanelApp RMCD red gene	ENSG00000179270
C8orf37	PanelApp RMCD red gene + SCGS V1 list	ENSG00000156172

CBY1	On SCGS V1 list	ENSG00000100211
CCDC103	PanelApp RMCD red gene + SCGS V1 list	ENSG00000167131
CCDC114	PanelApp RMCD red gene + SCGS V1 list	ENSG00000105479
CCDC151	PanelApp RMCD red gene	ENSG00000198003
CCDC28B	PanelApp RMCD red gene + SCGS V1 list	ENSG00000160050
CCDC39	PanelApp RMCD red gene + SCGS V1 list	ENSG00000145075
CCDC40	PanelApp RMCD red gene + SCGS V1 list	ENSG00000141519
CCDC65	PanelApp RMCD red gene	ENSG00000139537
CCNO	PanelApp RMCD red gene	ENSG00000152669
CCP110	On SCGS V1 list	ENSG00000103540
CDH23	PanelApp RMCD red gene + SCGS V1 list	ENSG00000107736
CENPJ	On SCGS V1 list	ENSG00000151849
CEP131	On SCGS V1 list	ENSG00000141577
CEP135	On SCGS V1 list	ENSG00000174799
CEP250	On SCGS V1 list	ENSG00000126001
CEP72	On SCGS V1 list	ENSG00000112877
CEP89	On SCGS V1 list	ENSG00000121289
CEP97	On SCGS V1 list	ENSG00000182504
CFAP100	On SCGS V1 list	ENSG00000163885
CFAP43	PanelApp RMCD red gene	ENSG00000197748
CFAP44	PanelApp RMCD red gene	ENSG00000206530
CFAP45	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000213085
CFAP53	PanelApp RMCD red gene	ENSG00000172361
CFAP57	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000243710
CFC1	PanelApp RMCD red gene	ENSG00000136698
CFTR	PanelApp RMCD red gene	ENSG00000001626
CLDN2	On SCGS V1 list	ENSG00000165376
CLRN1	PanelApp RMCD red gene	ENSG00000163646
CLTB	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000175416
CLUAP1	On SCGS V1 list	ENSG00000103351
CNGA2	On SCGS V1 list	ENSG00000183862
CNGA4	On SCGS V1 list	ENSG00000132259
CNGB1	On SCGS V1 list	ENSG00000070729
CRB1	PanelApp RMCD red gene	ENSG00000134376
CRB3	On SCGS V1 list	ENSG00000130545
CRELD1	PanelApp RMCD red gene	ENSG00000163703
CROCC	On SCGS V1 list	ENSG00000058453
CRX	PanelApp RMCD red gene	ENSG00000105392
CTNNB1	On SCGS V1 list	ENSG00000168036

DCDC2	PanelApp RMCD red gene + SCGS V1 list	ENSG00000146038
DISC1	On SCGS V1 list	ENSG00000162946
DNAAF1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000154099
DNAAF2	PanelApp RMCD red gene + SCGS V1 list	ENSG00000165506
DNAAF3	PanelApp RMCD red gene + SCGS V1 list	ENSG00000167646
DNAAF4	PanelApp RMCD red gene	ENSG00000256061
DNAAF5	PanelApp RMCD red gene	ENSG00000164818
DNAH1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000114841
DNAH10	On SCGS V1 list	ENSG00000197653
DNAH11	PanelApp RMCD red gene + SCGS V1 list	ENSG00000105877
DNAH2	On SCGS V1 list	ENSG00000183914
DNAH5	PanelApp RMCD red gene + SCGS V1 list	ENSG00000039139
DNAH6	On SCGS V1 list	ENSG00000115423
DNAI1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000122735
DNAI2	PanelApp RMCD red gene + SCGS V1 list	ENSG00000171595
DNAJB13	PanelApp RMCD red gene	ENSG00000187726
DNAL1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000119661
DNAL1	On SCGS V1 list	ENSG00000163879
DNHD1	PanelApp RMCD red gene	ENSG00000179532
DPCD	On SCGS V1 list	ENSG00000166171
DPYSL2	On SCGS V1 list	ENSG00000092964
DRC1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000157856
DRD1	On SCGS V1 list	ENSG00000184845
DRD2	On SCGS V1 list	ENSG00000149295
DRD5	On SCGS V1 list	ENSG00000169676
DVL1	On SCGS V1 list	ENSG00000107404
DYNLT1	On SCGS V1 list	ENSG00000146425
EFHC1	On SCGS V1 list	ENSG00000096093
EXOC3	On SCGS V1 list	ENSG00000180104
EXOC3L2	PanelApp RMCD amber gene	ENSG00000283632
EXOC4	On SCGS V1 list	ENSG00000131558
EXOC5	On SCGS V1 list	ENSG00000070367
EXOC6	On SCGS V1 list	ENSG00000138190
EXOC6B	On SCGS V1 list	ENSG00000144036
EXOC8	PanelApp RMCD red gene	ENSG00000116903
EZH2	On SCGS V1 list	ENSG00000106462
FAM149B1	PanelApp RMCD amber gene	ENSG00000138286
FAM161A	On SCGS V1 list	ENSG00000170264
FBF1	On SCGS V1 list	ENSG00000188878
FLNA	On SCGS V1 list	ENSG00000196924
FOPNL	On SCGS V1 list	ENSG00000133393

FOXH1	PanelApp RMCD red gene	ENSG00000160973
FOXJ1	On SCGS V1 list	ENSG00000129654
FUZ	On SCGS V1 list	ENSG00000010361
GAS8	PanelApp RMCD red gene + SCGS V1 list	ENSG00000141013
GDF1	PanelApp RMCD red gene	ENSG00000130283
GLI1	On SCGS V1 list	ENSG00000111087
GLI2	On SCGS V1 list	ENSG00000074047
GLIS2	PanelApp RMCD red gene + SCGS V1 list	ENSG00000126603
GPR161	On SCGS V1 list	ENSG00000143147
GSK3B	On SCGS V1 list	ENSG00000082701
GUCY2D	PanelApp RMCD red gene	ENSG00000132518
HAP1	On SCGS V1 list	ENSG00000173805
HSD11B1	On SCGS V1 list	ENSG00000117594
HSPA8	On SCGS V1 list	ENSG00000109971
HSPB11	On SCGS V1 list	ENSG00000081870
HTR6	On SCGS V1 list	ENSG00000158748
HTT	On SCGS V1 list	ENSG00000197386
HYDIN	PanelApp RMCD red gene + SCGS V1 list	ENSG00000157423
IFT122	On SCGS V1 list	ENSG00000128581
IFT20	On SCGS V1 list	ENSG00000109083
IFT46	On SCGS V1 list	ENSG00000118096
IFT57	On SCGS V1 list	ENSG00000114446
IFT81	PanelApp RMCD red gene + SCGS V1 list	ENSG00000122970
IFT88	On SCGS V1 list	ENSG00000032742
IMPDH1	PanelApp RMCD red gene	ENSG00000106348
INTU	On SCGS V1 list	ENSG00000164066
JADE1	On SCGS V1 list	ENSG00000077684
KCNJ13	PanelApp RMCD red gene	ENSG00000115474
KIAA0556	PanelApp RMCD red gene	ENSG00000047578
KIF14	PanelApp RMCD red gene	ENSG00000118193
KIF17	On SCGS V1 list	ENSG00000117245
KIF19	On SCGS V1 list	ENSG00000196169
KIF24	On SCGS V1 list	ENSG00000186638
KIF27	On SCGS V1 list	ENSG00000165115
KIF3A	On SCGS V1 list	ENSG00000131437
KIF3B	On SCGS V1 list	ENSG00000101350
KIF3C	On SCGS V1 list	ENSG00000084731
KIF5B	On SCGS V1 list	ENSG00000170759
LBR	PanelApp RMCD red gene	ENSG00000143815
LCA5	PanelApp RMCD red gene + SCGS V1 list	ENSG00000135338
LEFTY2	PanelApp RMCD red gene	ENSG00000143768
LRAT	PanelApp RMCD red gene	ENSG00000121207

LRRC6	PanelApp RMCD red gene + SCGS V1 list	ENSG00000129295
LRRC45	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000169683
MAK	On SCGS V1 list	ENSG00000111837
MAL	On SCGS V1 list	ENSG00000172005
MAPRE1	On SCGS V1 list	ENSG00000101367
MCHR1	On SCGS V1 list	ENSG00000128285
MCIDAS	PanelApp RMCD red gene	ENSG00000234602
MDM1	On SCGS V1 list	ENSG00000111554
MICAL2	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000133816
MLF1	On SCGS V1 list	ENSG00000178053
MNS1	On SCGS V1 list	ENSG00000138587
MUC1	PanelApp RMCD red gene	ENSG00000185499
MYO15A	On SCGS V1 list	ENSG00000091536
MYO7A	PanelApp RMCD red gene + SCGS V1 list	ENSG00000137474
NEK2	On SCGS V1 list	ENSG00000117650
NEK4	On SCGS V1 list	ENSG00000114904
NGFR	On SCGS V1 list	ENSG00000064300
NIN	On SCGS V1 list	ENSG00000100503
NINL	On SCGS V1 list	ENSG00000101004
NKX2-5	PanelApp RMCD red gene	ENSG00000183072
NME5	On SCGS V1 list	ENSG00000112981
NME7	On SCGS V1 list	ENSG00000143156
NME8	PanelApp RMCD red gene + SCGS V1 list	ENSG00000086288
NODAL	PanelApp RMCD red gene	ENSG00000156574
NOTO	On SCGS V1 list	ENSG00000214513
NUP214	On SCGS V1 list	ENSG00000126883
NUP35	On SCGS V1 list	ENSG00000163002
NUP37	On SCGS V1 list	ENSG00000075188
NUP62	On SCGS V1 list	ENSG00000213024
NUP93	On SCGS V1 list	ENSG00000102900
OCRL	PanelApp RMCD red gene + SCGS V1 list	ENSG00000122126
ODF2	On SCGS V1 list	ENSG00000136811
ORC1	On SCGS V1 list	ENSG00000085840
PACRG	On SCGS V1 list	ENSG00000112530
PAFAH1B1	On SCGS V1 list	ENSG00000007168
PARD3	On SCGS V1 list	ENSG00000148498
PARD6A	On SCGS V1 list	ENSG00000102981
PCDH15	PanelApp RMCD red gene + SCGS V1 list	ENSG00000150275
PCM1	On SCGS V1 list	ENSG00000078674
PDE6D	PanelApp RMCD red gene + SCGS V1 list	ENSG00000156973

PDZD7	On SCGS V1 list	ENSG00000186862
PKD1L1	On SCGS V1 list	ENSG00000158683
PLK1	On SCGS V1 list	ENSG00000166851
POC1A	PanelApp RMCD red gene + SCGS V1 list	ENSG00000164087
POC1B	PanelApp RMCD amber gene	ENSG00000139323
PPP3CC	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000120910
PRKCSH	PanelApp RMCD red gene	ENSG00000130175
PTCH1	On SCGS V1 list	ENSG00000185920
PTPDC1	On SCGS V1 list	ENSG00000158079
RAB11A	On SCGS V1 list	ENSG00000103769
RAB11FIP3	On SCGS V1 list	ENSG00000090565
RAB17	On SCGS V1 list	ENSG00000124839
RAB23	On SCGS V1 list	ENSG00000112210
RAB3IP	On SCGS V1 list	ENSG00000127328
RAB8A	On SCGS V1 list	ENSG00000167461
RAN	On SCGS V1 list	ENSG00000132341
RANBP1	On SCGS V1 list	ENSG00000099901
RD3	PanelApp RMCD red gene	ENSG00000198570
RDH12	PanelApp RMCD red gene	ENSG00000139988
RFX3	On SCGS V1 list	ENSG00000080298
RILPL1	On SCGS V1 list	ENSG00000188026
RILPL2	On SCGS V1 list	ENSG00000150977
ROPN1L	On SCGS V1 list	ENSG00000145491
RP1	On SCGS V1 list	ENSG00000104237
RP2	On SCGS V1 list	ENSG00000102218
RPE65	PanelApp RMCD red gene	ENSG00000116745
RPGR	PanelApp RMCD red gene + SCGS V1 list	ENSG00000156313
RPGRIP1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000092200
RSPH1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000160188
RSPH3	On SCGS V1 list	ENSG00000130363
RSPH4A	PanelApp RMCD red gene + SCGS V1 list	ENSG00000111834
RSPH9	PanelApp RMCD red gene + SCGS V1 list	ENSG00000172426
RTTN	On SCGS V1 list	ENSG00000176225
SASS6	On SCGS V1 list	ENSG00000156876
SCNN1A	PanelApp RMCD red gene	ENSG00000111319
SCNN1B	PanelApp RMCD red gene	ENSG00000168447
SCNN1G	PanelApp RMCD red gene	ENSG00000166828
SEC63	PanelApp RMCD red gene	ENSG00000025796
SEPT2	On SCGS V1 list	ENSG00000168385
SEPT7	On SCGS V1 list	ENSG00000122545
SHH	On SCGS V1 list	ENSG00000164690

SLC47A2	On SCGS V1 list	ENSG00000180638
SMCR8	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000283741
SMO	On SCGS V1 list	ENSG00000128602
SNAP25	On SCGS V1 list	ENSG00000132639
SNX10	On SCGS V1 list	ENSG00000086300
SNX17	On SCGS V1 list	ENSG00000115234
SPA17	On SCGS V1 list	ENSG00000064199
SPAG1	PanelApp RMCD red gene	ENSG00000104450
SPAG16	On SCGS V1 list	ENSG00000144451
SPAG17	On SCGS V1 list	ENSG00000155761
SPAG6	On SCGS V1 list	ENSG00000077327
SPATA7	PanelApp RMCD red gene + SCGS V1 list	ENSG00000042317
SPEF2	On SCGS V1 list	ENSG00000152582
SPTBN4	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000160460
SSNA1	On SCGS V1 list	ENSG00000176101
SSTR3	On SCGS V1 list	ENSG00000278195
STIL	On SCGS V1 list	ENSG00000123473
STK36	On SCGS V1 list	ENSG00000163482
STK38L	On SCGS V1 list	ENSG00000211455
STOML3	On SCGS V1 list	ENSG00000133115
STX3	On SCGS V1 list	ENSG00000166900
SUFU	PanelApp RMCD amber gene + SCGS V1 list	ENSG00000107882
SYNE2	On SCGS V1 list	ENSG00000054654
TAPT1	PanelApp RMCD red gene	ENSG00000169762
TBC1D30	On SCGS V1 list	ENSG00000111490
TBC1D32	PanelApp RMCD red gene	ENSG00000146350
TBC1D7	On SCGS V1 list	ENSG00000145979
TEKT2	On SCGS V1 list	ENSG00000092850
TEKT4	On SCGS V1 list	ENSG00000163060
TEKT5	On SCGS V1 list	ENSG00000153060
TEX12	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000150783
TNPO1	On SCGS V1 list	ENSG00000083312
TOPORS	PanelApp RMCD red gene + SCGS V1 list	ENSG00000197579
TPPP2	On SCGS V1 list	ENSG00000179636
TRAPPC10	On SCGS V1 list	ENSG00000160218
TRAPPC3	On SCGS V1 list	ENSG00000054116
TRAPPC9	On SCGS V1 list	ENSG00000167632
TRIM32	PanelApp RMCD red gene + SCGS V1 list	ENSG00000119401
TRIP11	On SCGS V1 list	ENSG00000100815

TSC1	PanelApp RMCD red gene	ENSG00000165699
TSC2	PanelApp RMCD red gene	ENSG00000103197
TTBK2	PanelApp RMCD red gene + SCGS V1 list	ENSG00000128881
TTC12	On SCGS V1 list	ENSG00000149292
TTC26	On SCGS V1 list	ENSG00000105948
TTC29	On SCGS V1 list	ENSG00000137473
TTC30A	On SCGS V1 list	ENSG00000197557
TTC30B	On SCGS V1 list	ENSG00000196659
TTK	On SCGS V1 list	ENSG00000112742
TTLL3	On SCGS V1 list	ENSG00000214021
TTLL6	On SCGS V1 list	ENSG00000170703
TTLL9	On SCGS V1 list	ENSG00000131044
TUBA1A	On SCGS V1 list	ENSG00000167552
TUBA1C	On SCGS V1 list	ENSG00000167553
TUBA4A	On SCGS V1 list	ENSG00000127824
TUBB2A	On SCGS V1 list	ENSG00000137267
TUBB2B	On SCGS V1 list	ENSG00000137285
TUBB3	On SCGS V1 list	ENSG00000258947
TUBB8	Ciliopathy candidate identified through local research team experimental screen(s)	ENSG00000261456
TUBE1	On SCGS V1 list	ENSG00000074935
TUBGCP2	On SCGS V1 list	ENSG00000130640
TUBGCP3	On SCGS V1 list	ENSG00000126216
TUBGCP4	On SCGS V1 list	ENSG00000137822
TUBGCP5	On SCGS V1 list	ENSG00000275835
TUBGCP6	On SCGS V1 list	ENSG00000128159
TULP1	PanelApp RMCD red gene + SCGS V1 list	ENSG00000112041
TULP3	On SCGS V1 list	ENSG00000078246
ULK4	On SCGS V1 list	ENSG00000168038
UMOD	PanelApp RMCD red gene	ENSG00000169344
USH1C	PanelApp RMCD red gene + SCGS V1 list	ENSG00000006611
USH1G	PanelApp RMCD red gene + SCGS V1 list	ENSG00000182040
USH2A	PanelApp RMCD red gene + SCGS V1 list	ENSG00000042781
VANGL2	On SCGS V1 list	ENSG00000162738
VDAC3	On SCGS V1 list	ENSG00000078668
VHL	PanelApp RMCD red gene + SCGS V1 list	ENSG00000134086
WDR63	PanelApp RMCD red gene	ENSG00000162643
WDR78	On SCGS V1 list	ENSG00000152763
WHRN	PanelApp RMCD red gene + SCGS V1 list	ENSG00000095397
XPNPEP3	PanelApp RMCD red gene + SCGS V1 list	ENSG00000196236
ZIC3	PanelApp RMCD red gene	ENSG00000156925
ZMYND10	PanelApp RMCD red gene	ENSG00000004838

ZNF423	PanelApp RMCD amber gene + SCGS V1 list	ENSG00000102935
--------	---	-----------------

Abbreviations: RMCD = Rare multisystem Ciliopathy disorders, SCGS V1 = SYSCILIA gold standard list version 1

Supplementary Table 4. Detailed variant information for all variants identified amongst participants of the congenital malformations caused by ciliopathies cohort with research identified diagnoses.

Proband research number	Gene symbol	HGVSc	HGVSp	Segregation	Zygosity (presumed where segregation data unk)	Mutational mechanism	SIFT	PolyPhen	1000G AF	gnomAD AF	ClinVar	PMID	CADD Phred	SpliceAI max DS	Match for phenotype	Variant ACMG class
1	CHD7	NM_017780.4:c.6955C>T	NP_060250.2:p.Arg2319Cys	De Novo	Het	Mis	Delet (0)	Prob_dam (0.99)	-	-	Path	16400610	27.1		Full	5
2	ALMS1	NM_001378454.1:c.10772del	NP_001365383.1:p.Thr3591LysfsTer6	Pat	Comp het	SG	-	-	-	0.00005227	Path	11941369, 11941370, 17594715	23		Full	5
2	ALMS1	NM_001378454.1:c.11104C>T	NP_001365383.1:p.Arg3702Ter	Mat	Comp het	SG	-	-	-	0.00001205	Path	-	34		Full	5
3	ARL6	NM_001278293.3:c.534A>G	NP_001265222.1:p.Gln178=	Bi-parental	Hom	Syn	-	-	-	0.00000796	-	27708425	24.1	DS_DL 0.87	Full	5
3	IMPG2	NM_016247.4:c.3262C>T	NP_057331.2:p.Arg108Ter	Bi-parental	Hom	SG	-	-	0.0002	0.0000358	Path, Lik_Path	-	42		Partial	4
4	RPGR	NM_000328.3:c.1627del	NP_000319.1:p.Asp543IlefsTer11	Unk	XLR	FS, male	-	-	-	-	-	-	-		Partial	4
5	CEP290	NM_025114.4:c.5668G>T	NP_079390.3:p.Gly1890Ter	Unk	Comp het	SG	-	-	0.0002	0.00009486	Path	18414213, 26092869, 16682970, 16682973, 17564967	35		Full	5
5	CEP290	NM_025114.4:c.322C>T	NP_079390.3:p.Arg108Ter	Unk	Comp het	SG	-	-	-	0.00000402	Path	-	35		Full	5
6	KIAA0586	NM_001329943.3:c.392del	NP_001316872.1:p.Arg131LysfsTer4	Pat	Comp het	FS	-	-	0.0024	0.00305	Path, Lik_Path	25741868, 26096313, 28552196, 26386247	33		Full	5
6	KIAA0586	NM_001329943.3:c.1402_1408del	NP_001316872.1:p.Pro468IlefsTer15	Mat	Comp het	FS	-	-	-	-	-	-	-		Full	5
7	CRYBB1	NM_001887.4:c.193G>A	NP_001878.1:p.Glu65Lys	Unk	Het	Mis	Delet (0)	Prob_dam (0.979)	-	-	-	-	29.2		Partial	3
7	OFD1	NM_003611.3:c.306del	NP_003602.1:p.Glu103LysfsTer42	Unk	XLD	FS, female	-	-	-	-	-	-	-		Full	4

7	<i>PKD1</i>	NM_001009 944.3:c.5890 C>T	NP_00100 9944.3:p.A rg1964Cys	Unk	Het	Mis	Delet (0.03)	poss_dam (0.772)	-	0.00001316	-	-	25.5		Partial	3
8	<i>PRPF8</i>	NM_006445. 4:c.5804G>A	NP_00643 6.3:p.Arg1 935His	De Novo	Het	Mis	Delet (0)	Prob_dam (1)	-	-	Path, Lik_Path	-	28.3		Full	4
9	<i>CEP152</i>	NM_001194 998.2:c.2041 C>T	NP_00118 1927.1:p.H is681Tyr	Pat	Comp het	Mis	Delet (0)	Prob_dam (0.999)	0.0002	0.0001603	VUS	-	25.1		Partial	3
9	<i>CEP152</i>	NM_001194 998.2:c.1499 A>T	NP_00118 1927.1:p.G lu500Val	Mat	Comp het	Mis	Delet (0.01)	Ben (0.159)	0.0002	-	-	-	23.2		Partial	3
10	<i>CEP290</i>	NM_025114. 4:c.5932C>T	NP_07939 0.3:p.Arg1 978Ter	Bi-parental	Hom	SG	-	-	-	0.00001212	Path	26092869	42		Full	5
11	<i>MYCN</i>	NM_005378. 6:c.1006del	NP_00536 9.2:p.Ser3 36.LeufsTer 15	De Novo	Het	FS	-	-	-	-	-	-	-		Full	5
12	<i>ARMC9</i>	NM_001271 466.4:c.879 G>A	NP_00125 8395.2:p.T hr293=	Bi-parental	Hom	Syn	-	-	-	0.00000806 5	Path, Lik_Path	29159890, 17576681, 9536098	26.8	DS_DL 0.9	Full	5
13	<i>TUBA1A</i>	NM_006009. 4:c.641G>T	NP_00600 0.2:p.Arg2 14Leu	De Novo	Het	Mis	Delet_ ow conf (0.03)	Ben (0.023)	-	-	Path, Lik_Path	25741868	23.5		Poor	5
14	<i>WDR19</i>	NM_025132. 4:c.1630_16 39del	NP_07940 8.3:p.Val5 44.LeufsTer 72	Mat	Comp het	Spl A	-	-	-	-	-	-	-		Full	5
14	<i>WDR19</i>	NM_025132. 4:c.817A>G	NP_07940 8.3:p.Asn2 73Asp	Pat	Comp het	Mis	Tol (0.28)	Ben (0.003)	-	0.00001629	VUS, Path	29068549	14.46		Full	5
15	<i>RHO</i>	NM_000539. 3:c.133T>C	NP_00053 0.1:p.Phe4 5Leu	Unk	Het	Mis	Tol (0.53)	Prob_dam (0.979)	0.0002	0.00002386	Path	1862076	23.6		Full	5
16	<i>STAG1</i>	NM_005862. 3:c.1033G>A	NP_00585 3.2:p.Glu3 45Lys	De Novo	Het	Mis	Delet (0)	Prob_dam (0.954)	-	-	-	-	28.6		Full	3
17	<i>BBS1</i>	NM_024649. 5:c.1169T>G	NP_07892 5.3:p.Met3 90Arg	Bi-parental	Hom	Mis	Delet (0)	Ben (0.347)	0.001	0.001512	Path, Lik_Path	15 PMIDs	26.2		Full	5
18	<i>RERE</i>	NM_001042 681.2:c.4286 A>T	NP_00103 6146.1:p.H is1429Leu	De Novo	Het	Mis	Delet_ ow conf (0)	Prob_dam (0.936)	-	-	-	-	28		Full	4
19	<i>ALMS1</i>	NM_001378 454.1:c.4681 _4687dup	NP_00136 5383.1:p.I le1563Asnf sTer20	Bi-parental	Hom	FS	-	-	-	-	-	-	-		Full	5
20	<i>BBS2</i>	NM_031885. 5:c.471+1G> C		Unk	Comp het	Spl d	-	-	-	-	-	-	33	DS_DL 0.95	Full	4
20	<i>BBS2</i>	NM_031885. 5:c.646C>T	NP_11409 1.4:p.Arg2 16Ter	Unk	Comp het	SG	-	-	-	0.00001194	Path, Lik_Path	11567139	37		Full	4

21	KAT6A	NM_006766. 5:c.1121C>T	NP_00675 7.2:p.Ser3 74Leu	Unk	Het	Mis	Delet (0)	Prob_dam (0.966)	-	0.00006369	-	-	28.4	Partial	3
21	LAMA1	NM_005559. 4:c.3397C>T	NP_00555 0.2:p.Arg1 133Ter	Unk	Comp het	SG	-	-	-	0.00001997	Path	-	38	Full	4
21	LAMA1	NM_005559. 4:c.281A>G	NP_00555 0.2:p.Gln9 4Arg	Unk	Comp het	Mis	Delet (0)	Prob_dam (1)	-	-	-	-	26.4	Full	3
22	MKKS	NM_170784. 2:c.1017_10 18del	NP_74075 4.1:p.Ile33 9MetfsTer 3	Bi-parental	Hom	FS	-	-	-	-	-	-	-	Full	5
23	CEP290	NM_025114. 4:c.5668G>T	NP_07939 0.3:p.Gly1 890Ter	Pat	Comp het	SG	-	-	0.0002	0.00009486	Path	18414213, 26092869, 16682970, 16682973, 17564967	36	Full	5
23	CEP290	NM_025114. 4:c.104T>G	NP_07939 0.3:p.Val3 5Gly	Mat	Comp het	Mis	Delet (0)	Prob_dam (0.943)	-	-	-	-	33	Full	3
25	RAI1	NM_030665. 4:c.2479C>T	NP_10959 0.3:p.Gln8 27Ter	De Novo	Het	SG	-	-	-	-	-	-	35	Full	5
26	PROM1	NM_006017. 3:c.1354dup	NP_00600 8.1:p.Tyr4 52LeufsTer 13	Unk	Hom	FS	-	-	-	0.0002195	Path, Lik_Path	-	32	Full	5
27	SETD2	NM_014159. 7:c.5218C>T	NP_05487 8.5:p.Arg1 740Trp	1 parent's unk, absent from other	Het	Mis	Delet (0)	Prob_dam (0.998)	-	-	Path, Lik_Path, VUS	-	32	Full	5
28	OPA1	NM_130837. 3:c.2678del	NP_57085 0.2:p.His89 3LeufsTer9	1 parent's unk, absent from other	Het	FS	-	-	-	-	-	-	-	Partial	5
29	ALMS1	NM_015120. 4:c.11881dup	NP_05593 5.4:p.Ser3 961PhefsT er12	Unk	Comp het	FS	-	-	-	-	-	-	-	Full	5
29	ALMS1	NM_015120. 4:c.1241- 81_1252deli nsCCTGCAG GCCCTCCAC ATATGCTAC AAAATA	-	Unk	Comp het	Spl A	-	-	-	-	-	-	-	Full	4
30	PHIP	NM_017934. 7:c.3202delA insTACCTG	NP_06040 4.4:p.Ile10 68AsnfsTer 3	Unk	Het	FS	-	-	-	-	-	-	-	Full	5
31	RAB28	NM_001017 979.3:c.58dup	NP_00101 7979.1:p.A sp20GlyfsT er62	Unk	Hom	FS	-	-	-	0.00000450 4	-	-	33	Full	5
32	ALMS1	GMC questionnaire reports		Unk	Comp het	Exon del	-	-	-	-	-	-	-	False positive	False positive

		exon 11 Deletion															
32	ALMS1	NM_015120. 4:c.9001C>T	NP_05593 5:4:p.Gln3 001Ter	Unk	Comp het	SG	-	-	-	0.00000400 8	-	-	36		Full	5	
34	POLA1	NM_001330 360:2:c.460 G>T	NP_00131 7289:1:p.A sp154Tyr	Mat	XLR	Mis, male	Delet (0)	Prob_dam (0.935)	-	0.0001049	-	-	32		Partial	3	
41	CSPP1	NM_001382 391:1:c.2968 +5G>A		Mat	Comp het	Spl region	-	-	-	-	VUS	-	-	DS_DL 0.79	Full	3	
41	CSPP1	GrCh38: Chr8:671366 72_6713904 8del		Pat	Comp het	2.7kb del incl exon 15									Full	5	
42	PIBF1	NM_006346. 4:c.1205T>C	NP_00633 7:2:p.Met4 02Thr	Unk	Comp het	Mis	Tol (0.09)	Ben (0.157)	-	-	-	-	22.8		Full	3	
42	PIBF1	GrCh38: chr13:g.7278 3352_72796 671del		Unk	Comp het	13.3kb del incl exons 2- 4									Full	5	
48	LRRCA5	NM_144999. 4:c.1074_10 75insTG	NP_65943 6:1:p.Leu3 59CysfsTer 19	Bi-parental	Hom	FS	-	-	-	-	-	-	-		Candidate gene	Candidat e gene	
56	BBS9	NM_198428. 3:c.1028G>A	NP_94082 0:1:p.Gly3 43Glu	1 parent's unk, present in other	Hom	Mis	Delet (0)	Prob_dam (1)	-	-	-	-	25.9		Full	3	
61	WT1	NM_024426. 6:c.1400G>A	NP_07774 4:4:p.Arg4 67Gln	Unk	Het	Mis	Delet_l ow conf (0.02)	Poss_dam (0.767)	-	-	Path	1302008	29.9		Full	5	
67	ALMS1	NM_015120. 4:c.1612C>G	NP_05593 5:4:p.Leu5 38Val	Pat	Comp het	Mis	Tol_low conf (0.17)	Ben (0.009)	-	0.0003289	lik_ben, VUS	-	0.003		Full	3	
67	ALMS1	NM_015120. 4:c.9613A>G	NP_05593 5:4:p.Ile32 05Val	Mat	Comp het	Mis	Delet (0.03)	Prob_dam (0.997)	-	0.00000802 1	VUS	-	23		Full	3	
69	BBS1	NM_024649. 5:c.1169T>G	NP_07892 5:3:p.Met3 90Arg	Mat	Comp het	Mis	Delet (0)	Ben (0.347)	0.001	0.001512	Path, Lik_Path	15 PMIDs	26.2		Full	5	
69	BBS1	NM_024649. 5:c.592- 1333_831- 449del		Pat	Comp het	4.7kb del incl exons 8 and 9									Full	5	
71	BCORL1	NM_001379 451:1:c.3463 C>A	NP_00136 6380:1:p.P ro1155Thr	Mat	XLR	Mis, male	Tol (0.12)	Poss_dam (0.557)	-	0.00003468	-	-	21.2		Partial	3	
72	CHD4	NM_001273. 5:c.3225G>T	NP_00126 4:2:p.Met1 075Ile	1 parent's unk, absent from other	Het	Mis	Delet (0)	Ben (0.358)	-	-	-	-	26		Full	3	
75	BBS4	NM_033028. 5:c.642+3A> T		1 parent's unk, present in other	Hom	Spl region	-	-	-	-	-	-	22.5	DS_DL 0.82	Data missing	3	

References

1. Gee HY, Otto EA, Hurd TW, Ashraf S, Chaki M, Cluckey A, et al. Whole-exome resequencing distinguishes cystic kidney diseases from phenocopies in renal ciliopathies. *Kidney Int.* 2014;85(4):880-7.
2. Knopp C, Rudnik-Schoneborn S, Eggermann T, Bergmann C, Begemann M, Schoner K, et al. Syndromic ciliopathies: From single gene to multi gene analysis by SNP arrays and next generation sequencing. *Molecular and cellular probes.* 2015;29(5):299-307.
3. Bachmann-Gagescu R, Dempsey JC, Phelps IG, O'Roak BJ, Knutzen DM, Rue TC, et al. Joubert syndrome: a model for untangling recessive disorders with extreme genetic heterogeneity. *Journal of medical genetics.* 2015;52(8):514-22.
4. Sawyer SL, Hartley T, Dymont DA, Beaulieu CL, Schwartzentruber J, Smith A, et al. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clinical genetics.* 2016;89(3):275-84.
5. Braun DA, Schueler M, Halbritter J, Gee HY, Porath JD, Lawson JA, et al. Whole exome sequencing identifies causative mutations in the majority of consanguineous or familial cases with childhood-onset increased renal echogenicity. *Kidney Int.* 2016;89(2):468-75.
6. Kang HG, Lee HK, Ahn YH, Joung JG, Nam J, Kim NK, et al. Targeted exome sequencing resolves allelic and the genetic heterogeneity in the genetic diagnosis of nephronophthisis-related Ciliopathy. *Exp Mol Med.* 2016;48(8):e251.
7. Watson CM, Crinnion LA, Berry IR, Harrison SM, Lascelles C, Antanaviciute A, et al. Enhanced diagnostic yield in Meckel-Gruber and Joubert syndrome through exome sequencing supplemented with split-read mapping. *BMC Medical Genetics.* 2016;17(1):1.
8. Castro-Sánchez S, Álvarez-Satta M, Tohamy MA, Beltran S, Derdak S, Valverde D. Whole exome sequencing as a diagnostic tool for patients with Ciliopathy-like phenotypes. *PLoS One.* 2017;12(8):e0183081.
9. Vilboux T, Doherty DA, Glass IA, Parisi MA, Phelps IG, Cullinane AR, et al. Molecular genetic findings and clinical correlations in 100 patients with Joubert syndrome and related disorders prospectively evaluated at a single center. *Genet Med.* 2017;19(8):875-82.
10. Shamseldin HE, Al Mogarri I, Alqwaiee MM, Alharbi AS, Baqais K, AlSaadi M, et al. An exome-first approach to aid in the diagnosis of primary ciliary dyskinesia. *Hum Genet.* 2020;139(10):1273-83.
11. Shamseldin HE, Shaheen R, Ewida N, Bubshait DK, Alkuraya H, Almardawi E, et al. The morbid genome of ciliopathies: an update. *Genet Med.* 2020.
12. Hammarsjö A, Pettersson M, Chitayat D, Handa A, Anderlid BM, Bartocci M, et al. High diagnostic yield in skeletal ciliopathies using massively parallel genome sequencing, structural variant screening and RNA analyses. *J Hum Genet.* 2021.
13. Forsythe E, Beales PL. Bardet-Biedl syndrome. *European journal of human genetics : EJHG.* 2013;21(1):8-13.
14. Bachmann-Gagescu R, Dempsey JC, Bulgheroni S, Chen ML, D'Arrigo S, Glass IA, et al. Healthcare recommendations for Joubert syndrome. *Am J Med Genet A.* 2020;182(1):229-49.
15. Hartill V, Szymanska K, Sharif SM, Wheway G, Johnson CA. Meckel-Gruber Syndrome: An Update on Diagnosis, Clinical Management, and Research Advances. *Front Pediatr.* 2017;5(244).
16. Quinlan RJ, Tobin JL, Beales PL. Modeling ciliopathies: Primary cilia in development and disease. *Curr Top Dev Biol.* 2008;84:249-310.

6.1.2 Uncovering the burden of hidden ciliopathies in the 100,000 Genomes Project: a reverse phenotyping approach

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance placed on this supplemental material which has been supplied by the author(s)

J Med Genet

Supplementary Table 1: Selection of leading multi-systemic ciliopathy disease genes from the medical literature

Ciliopathy syndrome	Leading genetic cause(s)	Mode of inheritance	Further ciliopathies associated with gene	Reference(s)
Bardet-Biedl syndrome (BBS)	<i>BBS1</i> (23.4% of all BBS)	Recessive	N/A	(1-3)
	<i>BBS10</i> (14.5% of all BBS)	Recessive	N/A	
Alström Syndrome (ALMS)	<i>ALMS1</i> (only causative gene)	Recessive	-Non-syndromic retinal dystrophy -Non-syndromic cardiomyopathy	(4-8)
Joubert syndrome (JBTS) and Meckel Gruber syndrome (MKS)	<i>TMEM67</i> (6-26% of all JBTS; 16% of all MKS)	Recessive	-NPHP with hepatic fibrosis -COACH syndrome (cerebellar vermis hypo/aplasia, oligophrenia, ataxia, ocular coloboma, and hepatic fibrosis)	(9-17)
	<i>CEP290</i> (6-22% of all JBTS, 2 nd most common cause of MKS)	Recessive	-Leber Congenital Amaurosis (LCA) / Early-Onset Severe Retinal Dystrophy (EOSRD) (15-20% of LCA / EOSRD cases) -NPHP -BBS -Senior-Løken syndrome -COACH syndrome	(14, 18-24)
Jeune Asphyxiating Thoracic Dystrophy (JATD)	<i>DYNC2H1</i> (~50% of all JATD)	Recessive	N/A	(25-28)
	<i>WDR34</i> (~10% of all JATD)	Recessive		
Nephronophthisis (NPHP)	<i>NPHP1</i> (20-25% of all NPHP)	Recessive	JBTS	(29-31)
Oral-facial-digital syndrome (OFD) Type 1	<i>OFD1</i> (only genetic cause)	X-linked dominant	JBTS (X-linked recessive)	(9, 32)

Supplementary Table 2: HPO terms linked to clinical key terms for ciliopathy syndromes

Key term	HPO ID	HPO descriptor	Linked HPO terms included in analysis
Retinal dystrophy	HP:0000556	Breakdown of light-sensitive cells in back of eye	<ul style="list-style-type: none"> • Cone/cone-rod dystrophy + sub-terms • Rod-cone dystrophy + sub-terms • Pattern dystrophy of the retina + sub-terms
Abnormality of eye movement	HP:0000496	An abnormality in voluntary or involuntary eye movements or their control	<ul style="list-style-type: none"> • Oculomotor apraxia (JBT5) • Nystagmus (LCA) • Roving eye movements (LCA)
Abnormal renal morphology / renal insufficiency	HP:0012210	Any structural anomaly of the kidney	<ul style="list-style-type: none"> • Abnormal localisation of kidney + sub-terms • Abnormal renal cortex morphology + sub-terms • Abnormal renal echogenicity + sub-terms • Abnormal renal medulla morphology + sub-terms • Abnormal renal pelvis morphology + sub-terms • Renal cyst + sub-terms • Renal dysplasia + sub-terms • Renal fibrosis + sub-terms • Renal hypoplasia/aplasia + sub-terms
	HP:0000083	A reduction in the level of performance of the kidneys in areas of function comprising the concentration of urine, removal of wastes, the maintenance of electrolyte balance, homeostasis of blood pressure, and calcium metabolism	<ul style="list-style-type: none"> • Chronic kidney disease + sub-terms
Abnormality of the liver	HP:0001392	An abnormality of the liver	<ul style="list-style-type: none"> • Abnormal liver morphology + sub-terms • Abnormal liver physiology + sub-terms • Abnormality of the biliary system + sub-terms
Abnormality of the genitourinary system	HP:0000119	The presence of any abnormality of the genitourinary system	<ul style="list-style-type: none"> • Abnormality of the genital system + sub-terms • Abnormality of the urinary system + sub-terms
Cardiomyopathy	HP:0001638	A myocardial disorder in which the heart muscle is structurally and functionally abnormal, in the absence of coronary artery disease, hypertension, valvular disease and congenital heart disease sufficient to cause the observed myocardial abnormality.	<ul style="list-style-type: none"> • All sub-terms
Sensorineural hearing impairment	HP:0000407	A type of hearing impairment in one or both ears related to an abnormal functionality of the cochlear nerve.	<ul style="list-style-type: none"> • All sub-terms

Abnormality of the sense of smell	HP:0004408	An anomaly in the ability to perceive and distinguish scents (odors).	<ul style="list-style-type: none"> All sub-terms
Abnormal pattern of respiration	HP:0002793	An anomaly of the rhythm or depth of breathing	<ul style="list-style-type: none"> Apnoea + sub-terms Tachypnoea + sub-terms
Hypogonadotrophic hypogonadism	HP:000044	Hypogonadotropic hypogonadism is characterized by reduced function of the gonads (testes in males or ovaries in females) and results from the absence of the gonadal stimulating pituitary hormones: follicle stimulating hormone (FSH) and luteinizing hormone (LH).	<ul style="list-style-type: none"> All sub-terms
Glucose intolerance	HP:0001952	Glucose intolerance (GI) can be defined as dysglycemia that comprises both prediabetes and diabetes. It includes the conditions of impaired fasting glucose (IFG) and impaired glucose tolerance (IGT) and diabetes mellitus (DM).	<ul style="list-style-type: none"> Type II diabetes mellitus + sub-terms Impaired glucose tolerance + sub-terms
Obesity	HP:0001513	Accumulation of substantial excess body fat.	<ul style="list-style-type: none"> All sub-terms
Hypertriglyceridemia	HP:0002155	An abnormal increase in the level of triglycerides in the blood	<ul style="list-style-type: none"> All sub-terms
Intellectual disability	HP:0001249	Subnormal intellectual functioning which originates during the developmental period. Intellectual disability, previously referred to as mental retardation, has been defined as an IQ score below 70.	<ul style="list-style-type: none"> All sub-terms
Neurodevelopmental delay	HP:0012758	None listed	<ul style="list-style-type: none"> All sub-terms
Hypotonia	HP:0001252	Hypotonia is an abnormally low muscle tone (the amount of tension or resistance to movement in a muscle). Even when relaxed, muscles have a continuous and passive partial contraction which provides some resistance to passive stretching. Hypotonia thus manifests as diminished resistance to passive stretching. Hypotonia is not the same as muscle weakness, although the two conditions can co-exist.	<ul style="list-style-type: none"> All sub-terms
Ataxia	HP:0001251	Cerebellar ataxia refers to ataxia due to dysfunction of the cerebellum. This causes a variety of elementary neurological deficits including asynergy (lack of coordination between muscles, limbs and joints), dysmetria (lack of ability to judge distances that can lead to under- or overshoot in grasping movements), and dysidiadochokinesia (inability to perform	<ul style="list-style-type: none"> All sub-terms

		rapid movements requiring antagonizing muscle groups to be switched on and off repeatedly).	
Abnormality of brain morphology	HP:0012443	A structural abnormality of the brain, which has as its parts the forebrain, midbrain, and hindbrain.	<ul style="list-style-type: none"> • Abnormal brainstem morphology + sub-terms • Abnormal cerebral ventricle morphology + sub-terms • Abnormal midbrain morphology + sub-terms • Abnormality of forebrain morphology + sub-terms • Abnormality of hindbrain morphology + sub-terms
Polydactyly	HP:0010442	A congenital anomaly characterized by the presence of supernumerary fingers or toes.	<ul style="list-style-type: none"> • All sub-terms
Short stature	HP:0004322	A height below that which is expected according to age and gender norms. Although there is no universally accepted definition of short stature, many refer to "short stature" as height more than 2 standard deviations below the mean for age and gender (or below the 3rd percentile for age and gender dependent norms).	<ul style="list-style-type: none"> • All sub-terms
Thoracic hypoplasia	HP:0005257	None listed	<ul style="list-style-type: none"> • All sub-terms
Brachydactyly / micromelia	HP:0001156	Digits that appear disproportionately short compared to the hand/foot.	<ul style="list-style-type: none"> • All sub-terms
Micromelia	HP:0002983	The presence of abnormally small extremities.	<ul style="list-style-type: none"> • All sub-terms
Abnormality of dentition	HP:0000164	Any abnormality of the teeth	<ul style="list-style-type: none"> • All sub-terms
Abnormal oral morphology	HP:0031816	Any structural anomaly of the mouth, which is also known as the oral cavity.	<ul style="list-style-type: none"> • All sub-terms
OFD1-specific facial dysmorphic features	HP:0000316	Hypertelorism: Interpupillary distance more than 2 SD above the mean (alternatively, the appearance of an increased interpupillary distance or widely spaced eyes)	<ul style="list-style-type: none"> • This term only
	HP:0000430	Underdeveloped nasal alae: Thinned, deficient, or excessively arched ala nasi.	<ul style="list-style-type: none"> • This term only
	HP:0000347	Micrognathia: Developmental hypoplasia of the mandible.	<ul style="list-style-type: none"> • This term only

Supplementary Table 3: Participants reported solved or partially solved in GMC exit questionnaires with variants in ciliopathy genes of interest

RESEARCH ID	GMC exit report outcome	Reported Sex	100K Recruitment Category	Gene	Variant Zygosity	Consequence	HGVSc	HGVSp	GMC exit questionnaire ACMG Class
1	Solved	MALE	BBS	ALMS1	Het	FS	NM_015120.4:c.10775del	NP_055935.4:p.Thr3592LysfsTer6	Path
					Het	SG	NM_015120.4:c.11107C>T	NP_055935.4:p.Arg3703Ter	Path
2	Solved	FEMALE	CDS	ALMS1	Het	SG	NM_015120.4:c.10975C>T	NP_055935.4:p.Arg3659Ter	Path
					Het	SG; FS	NM_015120.4:c.4571dup	NP_055935.4:p.Tyr1524Ter	Path
3	Solved	MALE	RCD	ALMS1	Het	FS	NM_015120.4:c.284del	NP_055935.4:p.Pro95ArgfsTer19	Path
					Het	FS	NM_015120.4:c.1793del	NP_055935.4:p.Glu598GlyfsTer3	Path
4	Solved	FEMALE	LCA or EOSRD	ALMS1	Het	SG	NM_015120.4:c.10483C>T	NP_055935.4:p.Gln3495Ter	Path
					Het	FS	NM_015120.4:c.6590del	NP_055935.4:p.Lys2197SerfsTer10	Path
5	Solved	FEMALE	ID; RCD	ALMS1	Het	FS	NM_015120.4:c.6570del	NP_055935.4:p.Ser2191HisfsTer16	Path
					Het	FS	NM_015120.4:c.10831_10832del	NP_055935.4:p.Arg3611AlafsTer6	Path
6	Solved	MALE	BBS	ALMS1	Het	FS	NM_015120.4:c.11881dup	NP_055935.4:p.Ser3961PhefsTer12	Path
					Het	"Large delins"	Data missing	Data missing	Likely path
7	Solved	MALE	URUMD	ALMS1	Hom	FS	NM_015120.4:c.2515dup	NP_055935.4:p.Ser839PhefsTer8	Path
8	Solved	MALE	BBS	ALMS1	Hom	FS	NM_015120.4:c.4684_4690dup	NP_055935.4:p.Ile1564AsnfsTer20	Path
9	Solved	FEMALE	RCD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
10	Solved	FEMALE	RCD	BBS1	Hom	Mis	19)	NP_078925.3:p.Met390Arg	Path
11	Solved	FEMALE	RCD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
12	Solved	MALE	RCD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
13	Solved	FEMALE	RCD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
14	Solved	FEMALE	SEOO +/- OEF + SS	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
15	Solved	FEMALE	ID	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
16	Solved	MALE	BBS	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
17	Solved	MALE	RCD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
18	Solved	MALE	CKD	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
19	Partially	MALE	ID	BBS1	Het	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
					Het	SG	NM_024649.5:c.871C>T	NP_078925.3:p.Gln291Ter	Path
20	Solved	FEMALE	Mito D	BBS1	Hom	Mis	NM_024649.5:c.1169T>G	NP_078925.3:p.Met390Arg	Path
21	Partially	MALE	RDS	BBS10	Het	Mis	NM_024685.4:c.1230T>G	NP_078961.3:p.His410Gln	Likely path
					Het	FS	NM_024685.4:c.271dup	NP_078961.3:p.Cys91LeufsTer5	Path
22	Solved	MALE	CAKUT	CEP290	Het	FS	NM_025114.4:c.2848dup	NP_079390.3:p.Gln950ProfsTer6	Path
					Het	Mis	NM_025114.4:c.2817G>T	NP_079390.3:p.Lys939Asn	Likely path

23	Solved	FEMALE	JBTS	CEP290	Hom	SG	NM_025114.4:c.5932C>T	NP_079390.3:p.Arg1978Ter	Path
24	Solved	MALE	LCA or EOSRD	CEP290	Hom	In-frame deletion	NM_025114.4:c.4661_4663del	NP_079390.3:p.Glu1554del	Likely path
25	Solved	FEMALE	LCA or EOSRD	CEP290	Het	FS	NM_025114.4:c.5434_5435del	NP_079390.3:p.Glu1812LysfsTer5	Path
					Het	SG	NM_025114.4:c.5668G>T	NP_079390.3:p.Gly1890Ter	Path
26	Solved	FEMALE	CAKUT	CEP290	Hom	SG	NM_025114.4:c.4174G>T	NP_079390.3:p.Glu1392Ter	Likely path
27	Partially	MALE	ID	CEP290	Het	SG	NM_025114.4:c.322C>T	NP_079390.3:p.Arg108Ter	Path
					Het	FS	NM_025114.4:c.3422dup	NP_079390.3:p.Leu1141PhefsTer5	Path
28	Solved	MALE	RCD	CEP290	Het	SG	NM_025114.4:c.1984C>T	NP_079390.3:p.Gln662Ter	Path
					Het	SG	NM_025114.4:c.7048C>T	NP_079390.3:p.Gln2350Ter	Path
29	Solved	FEMALE	BBS	CEP290	Het	SG	NM_025114.4:c.5668G>T	NP_079390.3:p.Gly1890Ter	Path
					Het	SG	NM_025114.4:c.322C>T	NP_079390.3:p.Arg108Ter	Path
30	Solved	MALE	RCD	DYNC1H1	Hom	SG	NM_001080463.2:c.9836C>A	NP_001073932.1:p.Ser3279Ter	Path
31	Solved	MALE	USD	DYNC1H1	Het	Spl A	NM_001080463.2:c.10834-1G>A	-	Path
					Het	Spl Reg	NM_001080463.2:c.6140-5A>G	-	Likely path
32	Solved	MALE	RCD	OFD1	Hemi	FS	NM_003611.3:c.2680_2681del	NP_003602.1:p.Glu894ArgfsTer6	Path
33	Solved	FEMALE	RCD	NPHP1	Het	Mis	NM_001128178.3:c.1882C>T	NP_001121650.1:p.Arg628Trp	Likely path
					Het	"Whole gene deletion"	Data missing	Data missing	Not specified
34	Solved	MALE	UKFIYP	NPHP1	Hom	Mis	NM_001128178.3:c.859G>A	NP_001121650.1:p.Gly287Arg	Path
35	Solved	MALE	UKFIYP	NPHP1	Hom	SG	NM_001128178.3:c.1142G>A	NP_001121650.1:p.Trp381Ter	Path
36	Solved	FEMALE	UKFIYP	OFD1	Het	FS	NM_003611.3:c.1651_1654del	NP_003602.1:p.Thr551ProfsTer2	Path
37	Solved	FEMALE	SARMIRD	OFD1	Het	Mis	NM_003611.3:c.1363A>C	NP_003602.1:p.Lys455Gln	VUS
38	Solved	FEMALE	Craniosyn S	OFD1	Het	Spl Reg	NM_003611.3:c.382-4A>G	-	VUS
39	Solved	FEMALE	CKD	OFD1	Het	Spl A	NM_003611.3:c.112-1G>A	-	Path
40	Partially	FEMALE	RMCD	OFD1	Het	FS	NM_003611.3:c.306del	NP_003602.1:p.Glu103LysfsTer42	Likely path
					Het	FS	NM_153704.6:c.103del	NP_714915.3:p.Gln35ArgfsTer52	Path
41	Solved	MALE	CKD	TMEM67	Het	FS	NM_153704.6:c.415_416del	NP_714915.3:p.Asp139HisfsTer2	Path
					Het	Mis	NM_153704.6:c.1319G>A	NP_714915.3:p.Arg440Gln	Path
42	Partially	MALE	ID	TMEM67	Het	Mis	NM_153704.6:c.2498T>C	NP_714915.3:p.Ile833Thr	Likely path
					Het	Mis	NM_153704.6:c.2498T>C	NP_714915.3:p.Ile833Thr	Likely path
43	Solved	MALE	RCD	CEP290	Het	FS	NM_025114.4:c.254dup	NP_079390.3:p.Asn85LysfsTer6	Likely path
44	Solved	MALE	LCA or EOSRD	CEP290	Hom	Mis	NM_025114.4:c.21G>T	NP_079390.3:p.Trp7Cys	Likely path

Abbreviations: 100K = 100,000 Genomes Project, GMC = Genomic Medicine Centre, ACMG = American College of Medical Genetics and Genomics, BBS = Bardet-Biedl syndrome, CDS = cone dysfunction syndrome, RCD = rod-cone dystrophy, LCA or EOSRD = Leber Congenital Amaurosis or Early-Onset Severe Retinal Dystrophy, ID = intellectual disability, URUMD = Ultra-rare undescribed monogenic disorders, SEOO +/- OEF + SS = Significant early-onset obesity with or without other endocrine features and short stature, CKD = cystic kidney disease, Mito D = mitochondrial disorders, RDS = rod-dysfunction syndrome, CAKUT = Congenital Anomaly of the Kidneys and Urinary Tract, JBTS = Joubert

syndrome, USD = Unexplained skeletal dysplasia, UKFIYP = Unexplained kidney failure in young people, SARMIRD = Single autosomal recessive mutation in rare disease, Craniosyn S = craniosynostosis syndromes, RMCD = Rare multisystem ciliopathy disorders, Het = heterozygous, Hom = homozygous, Hemi = hemizygous, FS = frameshift, SG = stop gain, Mis = missense, Spl A = splice acceptor, Spl Reg = splice region, Path = pathogenic, Likely path = likely pathogenic, VUS = variant of uncertain significance

Supplementary Table 4: Prioritised variants extracted through reverse phenotyping diagnostic research workflow

Step 2 workflow inputs and outputs: filtering and prioritisation of SNVs using custom Python script																		
INPUTS																		
INPUT SNV DATA: All SNVs from the 100K dataset for each selected ciliopathy gene generated by Gene-Variant Workflow. Separate lists for participants called on GrCh37 and GrCh38																		
Gene	ALMS1		BBS1		BBS10		DYNC2H1		WDR34		OFD1		NPHP1		TMEM67		CEP290	
Build	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38	GrCh37	GrCh38
# un-filtered Gene-Variant Workflow variants	52420	287121	24050	71969	166	601	80615	284569	7636	234958	2122	27257	30997	104051	28384	95596	19436	96000
<p>PROCESS: filter using custom python script filter_gene_variant_workflow.py</p> <p>A: Exclude common variants: 100K MAF \geq 0.002; gnomAD AF \geq 0.002</p> <p>B: Exclude variants called in non-canonical transcripts</p> <p style="text-align: center;">↓</p>																		
# filtered variants: rare, canonical transcripts only	11862	43098	1217	3802	153	588	16127	59165	1465	4939	279	4365	3399	12254	2810	10226	3740	14200
<p>PROCESS: extract prioritised SNV sub-lists using custom python script filter_gene_variant_workflow.py:</p> <ul style="list-style-type: none"> ClinVar pathogenic/likely pathogenic VEP High Impact (stop_gained, stop_lost, start_lost, splice_acceptor_variant, splice_donor_variant, frameshift_variant, transcript_ablation, transcript_amplification) SIFT deleterious missense 																		
OUTPUTS																		
Gene	ALMS1		BBS1		BBS10		DYNC2H1		WDR34		OFD1		NPHP1		TMEM67		CEP290	
Total ClinVar Pathogenic	13	43	1	14	5	22	16	58	2	9	0	64	3	8	10	36	22	78
Total VEP High Impact	30	130	2	22	5	28	19	141	4	38	0	70	7	35	11	57	36	167
Total SIFT deleterious missense	167	643	33	86	18	86	125	556	32	107	5	75	26	79	33	167	84	344

DISTRIBUTION OF PRIORITISED VARIANTS BETWEEN DIFFERENT PRIORITISED SNV SUB-LISTS																		
Gene	ALMS1		BBS1		BBS10		DYNC2H1		WDR34		OFD1		NPHP1		TMEM67		CEP290	
# ClinVar Pathogenic + VEP High Impact	13	43	0	11	5	17	5	26	1	6	0	58	2	7	4	20	19	73
# ClinVar pathogenic + SIFT deleterious missense	0	0	1	3	0	5	10	30	1	3	0	5	1	1	6	14	2	4
# VEP High Impact (only)	17	87	2	11	0	11	13	115	3	32	0	12	5	28	7	37	17	94
# SIFT deleterious missense (only)	167	643	32	83	18	81	115	526	31	104	5	70	25	78	27	153	82	340
# ClinVar Pathogenic (only)	0	0	0	0	0	0	1	2	0	0	0	1	0	0	0	2	1	1
Total	197	773	35	108	23	114	144	699	36	145	5	146	33	114	44	226	121	512
Step 3 workflow inputs and outputs: search for potentially pathogenic SVs using SVRare script																		
INPUTS																		
INPUT DATA: PlateKey identifiers for all unsolved 100K participants (probands and affected relatives) with heterozygous ClinVar pathogenic or VEP high impact prioritised SNVs in one of the nine ciliopathy genes N = 801 participants																		
PROCESS: Submitted to SVRare script (Yu et al, 2021) Extracts participants with SVs called by Manta and/or Canvas with ≤ 10 calls across the 100K database, overlapping coding regions of the 9 ciliopathy genes ↓																		
OUTPUTS																		
Gene	ALMS1		BBS1		BBS10		DYNC2H1		WDR34		OFD1		NPHP1		TMEM67		CEP290	
# Prioritised SNVs	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
Impression	N/a	LP	N/a	N/a	N/a	N/a	LP	N/a	N/a	N/a	N/a	Excl: 2 nd hit in different gene	N/a	N/a	N/a	N/a	N/a	Excl: alternative diagnosis

Step 4 workflow inputs and outputs: search for novel splicing variants using custom SpliceAI script																		
INPUTS																		
INPUT DATA: all rare variants (100K MAF \leq 0.002; gnomAD AF \leq 0.002) called in canonical transcripts in the nine ciliopathy genes identified in unsolved 100K participants AS PER Step 2: Gene-Variant Workflow rare SNVs called in canonical transcripts filtered through custom python script (filter_gene_variant_workflow.py)																		
PROCESS: Run through custom SpliceAI Python script (find_variants_by_gene_and_SpliceAI_score.py)																		
↓																		
FILTERING: <ul style="list-style-type: none"> • Variants called in unaffected relatives excluded • Variants with SpliceAI delta score (DS) > 0.5 retained • Variants already assessed on other SNV prioritised sub-lists excluded 																		
↓																		
OUTPUTS																		
Gene	ALMS1		BBS1		BBS10		DYNC2H1		WDR34		OFD1		NPHP1		TMEM67		CEP290	
# rare variants with SpliceAI DS >0.5	1	22	3	10	0	1	7	53	1	9	0	10	3	12	2	15	4	34

The number of variants input, filtered and prioritised in steps 2, 3 and 4 of the reverse phenotyping diagnostic research workflow. Note that 100K participants had genomes called on GrCh37 or GrCh38 depending on when they were recruited to the project.

Abbreviations: SNV = single nucleotide variant, 100K = 100,000 Genomes Project, AF = allele frequency, MAF = maximum allele frequency, VEP = Variant Effect Predictor, SV = structural variant, Excl = excluded

Supplementary Data 1: Duplex PCR assay of a *BBS1* exon 13 mobile element insertion

The patient presented with congenital right ptosis, childhood onset high myopia, rod/cone dysfunction, autism, dyspraxia and postaxial polydactyly on the left hand and foot that were removed in childhood. The patient was recruited to the 100,000 Genomes Project (100K) for whole genome sequencing, following identification of a heterozygous pathogenic variant in an autosomal recessive disease gene through mainstream testing. The *BBS1* missense mutation, NM_024649.5:c.1169T>G, NP_078925.3:p.(Met390Arg), was insufficient to confirm the diagnosis in the absence of a second pathogenic variant. 100K tiering failed to identify a second deleterious allele in *BBS1*. Manual inspection of the aligned sequence reads using the Integrative Genome Browser (IGV) v.2.4.10 (<http://software.broadinstitute.org/software/igv/>) (33) and interrogation of soft-clipped reads using BLAT (<http://genome.ucsc.edu/cgi-bin/hgBlat>) (34), revealed a soft-clipped read signature that was consistent with a 2.4 kb insertion of an SVA F family element mobile element (35).

To confirm the *BBS1* heterozygous missense variant, c.1169T>C, a PCR amplicon was first optimised; each reaction comprised 0.5 µL of genomic DNA (~50 ng/µL) 19.3 µL MegaMix PCR reagent (Microzone Ltd., Haywards Heath, UK) and 0.1 µL each of 10 µM forward (dTGTA AACGACGCGCCAGTAAAGGCAGCATTGTGAAGGG) and reverse (dCAGGAAACAGCTATGACCCCTTCACTCCCGACTTCAA) primers. Thermocycling conditions comprised 94°C for 5 minutes then 30 cycles of 94°C for 30 seconds, 55°C for 1 minute and 72°C for 2 minutes before a final extension step at 72°C for 5 minutes. Amplification products were resolved on a 1% Tris-borate-EDTA agarose gel, before being extracted and purified using a QIAquick column (Qiagen GmbH, Hilden, Germany), then Sanger sequenced using an ABI3730 following manufacturer's protocols throughout (Life Technologies Ltd., Paisley, UK). Sequence chromatograms were analysed using 4Peaks v.1.8 (<http://nucleobytes.com/4peaks/index.html>). Universal sequence tags (underlined) were incorporated into primer tails for use with our routine diagnostic workflow.

To verify the apparent *BBS1* exon 13 mobile element insertion, we implemented the duplex PCR assay as described previously (35). Each reaction comprised 0.5 µL of genomic DNA (~50 ng/µL) 19.2 µL of MegaMix PCR reagent and 0.1 µL each of 10 µM primer. These included a common intron 12 forward (dCACAGTACTCCACAAATAACTGCT), an intron 13 reverse

(dATCCCCAGCTTTGCTGT) and insertion-specific reverse (dCAGCCTGGGCACCATTGA) primer. Thermocycling conditions required 35 cycles, but were otherwise as described above. Amplification products specific for the normal (440 bp) and insertion-containing (270 bp) allele were resolved on a 2% TRIS-borate-EDTA agarose gel prior to gel extraction and Sanger sequencing. To determine the precise sequence of the downstream target site duplication a further PCR was optimised for Sanger sequencing, using previously reported forward (F9: dAGTACCCAGGGACAACACT) and reverse (R5: dGTCTTCGGGGCACATTGAG) primers (35). Analysis of parental alignments supported the mobile element insertion being in *trans* with the maternally-inherited c.1169T>C mutation, with Sanger sequencing confirming the presence of the insertion in the proband and his father.

Supplementary references

1. Niederlova V, Modrak M, Tsyklauri O, Huranova M, Stepanek O. Meta-analysis of genotype-phenotype associations in Bardet-Biedl syndrome uncovers differences among causative genes. *Hum Mutat.* 2019;40(11):2068-87.
2. Shamseldin HE, Shaheen R, Ewida N, Bubshait DK, Alkuraya H, Almardawi E, Howaidi A, Sabr Y, Abdalla EM, Alfaifi AY, Alghamdi JM, Alsagheir A, Alfares A, Morsy H, Hussein MH, Al-Muhaizea MA, Shagrani M, Al Sabban E, Salih MA, Meriki N, Khan R, Almugbel M, Qari A, Tulba M, Mahnashi M, Alhazmi K, Alsalamah AK, Nowilaty SR, Alhashem A, Hashem M, Abdulwahab F, Ibrahim N, Alshidi T, AlObeid E, Alenazi MM, Alzaidan H, Rahbeeni Z, Al-Owain M, Sogaty S, Seidahmed MZ, Alkuraya FS. The morbid genome of ciliopathies: an update. *Genet Med.* 2020;22(6):1051-60.
3. Forsyth R, Gunay-Aygun M. Bardet-Biedl Syndrome Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2021, University of Washington, Seattle. *GeneReviews* is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.
4. Paisey RB, Steeds R, Barrett T, Williams D, Geberhiwot T, Gunay-Aygun M. Alström Syndrome. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LJH, Mirzaa G, et al., editors. *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2021, University of Washington, Seattle. *GeneReviews* is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.
5. Aldrees A, Abdelkader E, Al-Habboubi H, Alrwebah H, Rahbeeni Z, Schatz P. Nonsyndromic retinal dystrophy associated with homozygous mutations in the *ALMS1* gene. *Ophthalmic Genet.* 2019;40(1):77-9.
6. Hull S, Kiray G, Chiang JP, Vincent AL. Molecular and phenotypic investigation of a New Zealand cohort of childhood-onset retinal dystrophy. *Am J Med Genet C Semin Med Genet.* 2020;184(3):708-17.
7. Lazar CH, Kimchi A, Namburi P, Mutsuddi M, Zelinger L, Beryozkin A, Ben-Simhon S, Obolensky A, Ben-Neriah Z, Argov Z, Pikarsky E, Fellig Y, Marks-Ohana D, Ratnapriya R, Banin E, Sharon D, Swaroop A. Nonsyndromic Early-Onset Cone-Rod Dystrophy and Limb-Girdle Muscular Dystrophy in a Consanguineous Israeli Family are Caused by Two Independent yet Linked Mutations in *ALMS1* and *DYSF*. *Hum Mutat.* 2015;36(9):836-41.
8. Louw JJ, Corveleyn A, Jia Y, Iqbal S, Boshoff D, Gewillig M, Peeters H, Moerman P, Devriendt K. Homozygous loss-of-function mutation in *ALMS1* causes the lethal disorder mitogenic cardiomyopathy in two siblings. *Eur J Med Genet.* 2014;57(9):532-5.
9. Bachmann-Gagescu R, Dempsey JC, Phelps IG, O'Roak BJ, Knutzen DM, Rue TC, Ishak GE, Isabella CR, Gorden N, Adkins J, Boyle EA, de Lacy N, O'Day D, Alswaid A, Ramadevi AR, Lingappa L, Lourenco C, Martorell L, Garcia-Cazorla A, Ozyurek H, Haliloglu G, Tuysuz B, Topcu M, Chance P, Parisi MA, Glass IA, Shendure J, Doherty D. Joubert syndrome: a model for untangling recessive disorders with extreme genetic heterogeneity. *J Med Genet.* 2015;52(8):514-22.
10. Vilboux T, Doherty DA, Glass IA, Parisi MA, Phelps IG, Cullinane AR, Zein W, Brooks BP, Heller T, Soldatos A, Oden NL, Yildirimli D, Vemulapalli M, Mullikin JC, Nisc Comparative Sequencing P, Malicdan MCV, Gahl WA, Gunay-Aygun M. Molecular genetic findings and

- clinical correlations in 100 patients with Joubert syndrome and related disorders prospectively evaluated at a single center. *Genet Med.* 2017;19(8):875-82.
11. Parisi M, Glass I. Joubert Syndrome. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LH, Mirzaa G, et al., editors. *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2021, University of Washington, Seattle. *GeneReviews* is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.
 12. Suzuki T, Miyake N, Tsurusaki Y, Okamoto N, Alkindy A, Inaba A, Sato M, Ito S, Muramatsu K, Kimura S, Ieda D, Saitoh S, Hiyane M, Suzumura H, Yagyu K, Shiraishi H, Nakajima M, Fueki N, Habata Y, Ueda Y, Komatsu Y, Yan K, Shimoda K, Shitara Y, Mizuno S, Ichinomiya K, Sameshima K, Tsuyusaki Y, Kurosawa K, Sakai Y, Haginoya K, Kobayashi Y, Yoshizawa C, Hisano M, Nakashima M, Saitsu H, Takeda S, Matsumoto N. Molecular genetic analysis of 30 families with Joubert syndrome. *Clin Genet.* 2016;90(6):526-35.
 13. Otto EA, Tory K, Attanasio M, Zhou W, Chaki M, Paruchuri Y, Wise EL, Wolf MT, Utsch B, Becker C, Nurnberg G, Nurnberg P, Nayir A, Saunier S, Antignac C, Hildebrandt F. Hypomorphic mutations in meckelin (MKS3/TMEM67) cause nephronophthisis with liver fibrosis (NPHP11). *J Med Genet.* 2009;46(10):663-70.
 14. Hartill V, Szymanska K, Sharif SM, Whewey G, Johnson CA. Meckel-Gruber Syndrome: An Update on Diagnosis, Clinical Management, and Research Advances. *Front Pediatr.* 2017;5(244).
 15. Iannicelli M, Brancati F, Mougou-Zerelli S, Mazzotta A, Thomas S, Elkhartoufi N, Travaglini L, Gomes C, Ardissino GL, Bertini E, Boltshauser E, Castorina P, D'Arrigo S, Fischetto R, Leroy B, Loget P, Bonniere M, Starck L, Tantau J, Gentilin B, Majore S, Swistun D, Flori E, Lalatta F, Pantaleoni C, Penzien J, Grammatico P, Dallapiccola B, Gleeson JG, Attie-Bitach T, Valente EM. Novel TMEM67 mutations and genotype-phenotype correlates in meckelin-related ciliopathies. *Hum Mutat.* 2010;31(5):E1319-31.
 16. Brancati F, Iannicelli M, Travaglini L, Mazzotta A, Bertini E, Boltshauser E, D'Arrigo S, Emma F, Fazzi E, Gallizzi R, Gentile M, Loncarevic D, Mejaski-Bosnjak V, Pantaleoni C, Rigoli L, Salpietro CD, Signorini S, Stringini GR, Verloes A, Zablocka D, Dallapiccola B, Gleeson JG, Valente EM. MKS3/TMEM67 mutations are a major cause of COACH Syndrome, a Joubert Syndrome related disorder with liver involvement. *Hum Mutat.* 2009;30(2):E432-42.
 17. Doherty D, Parisi MA, Finn LS, Gunay-Aygun M, Al-Mateen M, Bates D, Clericuzio C, Demir H, Dorschner M, van Essen AJ, Gahl WA, Gentile M, Gorden NT, Hikida A, Knutzen D, Ozyurek H, Phelps I, Rosenthal P, Verloes A, Weigand H, Chance PF, Dobyns WB, Glass IA. Mutations in 3 genes (MKS3, CC2D2A and RPGRIP1L) cause COACH syndrome (Joubert syndrome with congenital hepatic fibrosis). *J Med Genet.* 2010;47(1):8-21.
 18. Travaglini L, Brancati F, Attie-Bitach T, Audollent S, Bertini E, Kaplan J, Perrault I, Iannicelli M, Mancuso B, Rigoli L, Rozet JM, Swistun D, Tolentino J, Dallapiccola B, Gleeson JG, Valente EM, Zankl A, Leventer R, Grattan-Smith P, Janecke A, D'Hooghe M, Sznajder Y, Van Coster R, Demerleir L, Dias K, Moco C, Moreira A, Kim CA, Maegawa G, Petkovic D, Abdel-Salam GM, Abdel-Aleem A, Zaki MS, Marti I, Quijano-Roy S, Sigaudy S, de Lonlay P, Romano S, Touraine R, Koenig M, Lagier-Tourenne C, Messer J, Collignon P, Wolf N, Philippi H, Kitsiou Tzeli S, Halldorsson S, Johannsdottir J, Ludvigsson P, Phadke SR, Udani V, Stuart B, Magee A, Lev D, Michelson M, Ben-Zeev B, Fischetto R, Benedicenti F, Stanzial F, Borgatti R, Accorsi P, Battaglia S, Fazzi E, Giordano L, Pinelli L, Boccone L, Bigoni S, Ferlini A, Donati MA, Caridi G, Divizia MT, Faravelli F, Ghiggeri G, Pessagno A, Briguglio M,

- Briuglia S, Salpietro CD, Tortorella G, Adami A, Castorina P, Lalatta F, Marra G, Riva D, Scelsa B, Spaccini L, Uziel G, Del Giudice E, Laverda AM, Ludwig K, Permunian A, Suppiej A, Signorini S, Uggetti C, Battini R, Di Giacomo M, Cilio MR, Di Sabato ML, Leuzzi V, Parisi P, Pollazzon M, Silengo M, De Vescovi R, Greco D, Romano C, Cazzagon M, Simonati A, Al-Tawari AA, Bastaki L, Mégarbané A, Sabolic Avramovska V, de Jong MM, Stromme P, Koul R, Rajab A, Azam M, Barbot C, Martorell Sampol L, Rodriguez B, Pascual-Castroviejo I, Teber S, Anlar B, Comu S, Karaca E, Kayserili H, Yüksel A, Akcakus M, Al Gazali L, Sztriha L, Nicholl D, Woods CG, Bennett C, Hurst J, Sheridan E, Barnicoat A, Hennekam R, Lees M, Blair E, Bernes S, Sanchez H, Clark AE, DeMarco E, Donahue C, Sherr E, Hahn J, Sanger TD, Gallager TE, Dobyns WB, Daugherty C, Krishnamoorthy KS, Sarco D, Walsh CA, McKanna T, Milisa J, Chung WK, De Vivo DC, Raynes H, Schubert R, Seward A, Brooks DG, Goldstein A, Caldwell J, Finsecke E, Maria BL, Holden K, Cruse RP, Swoboda KJ, Viskochil D. Expanding CEP290 mutational spectrum in ciliopathies. *Am J Med Genet A*. 2009;149a(10):2173-80.
19. Kumaran N, Pennesi ME, Yang P, Trzupke KM, Schlechter C, Moore AT, Weleber RG, Michaelides M. Leber Congenital Amaurosis / Early-Onset Severe Retinal Dystrophy Overview. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LH, Mirzaa G, et al., editors. *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2021, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.
 20. Coppieters F, Lefever S, Leroy BP, De Baere E. *CEP290*, a gene with many faces: mutation overview and presentation of *CEP290*. *Human Mutation*. 2010;31(10):1097-108.
 21. Adams M, Simms RJ, Abdelhamed Z, Dawe HR, Szymanska K, Logan CV, Whewey G, Pitt E, Gull K, Knowles MA, Blair E, Cross SH, Sayer JA, Johnson CA. A meckelin-filamin A interaction mediates ciliogenesis. *Human molecular genetics*. 2012;21(6):1272-86.
 22. Chang B, Khanna H, Hawes N, Jimeno D, He S, Lillo C, Parapuram SK, Cheng H, Scott A, Hurd RE, Sayer JA, Otto EA, Attanasio M, O'Toole JF, Jin G, Shou C, Hildebrandt F, Williams DS, Heckenlively JR, Swaroop A. In-frame deletion in a novel centrosomal/ciliary protein CEP290/NPHP6 perturbs its interaction with RPGR and results in early-onset retinal degeneration in the rd16 mouse. *Hum Mol Genet*. 2006;15(11):1847-57.
 23. Leitch CC, Zaghoul NA, Davis EE, Stoetzel C, Diaz-Font A, Rix S, Alfadhel M, Lewis RA, Eyaid W, Banin E, Dollfus H, Beales PL, Badano JL, Katsanis N. Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome. *Nature Genetics*. 2008;40:443.
 24. Brancati F, Camerota L, Colao E, Vega-Warner V, Zhao X, Zhang R, Bottillo I, Castori M, Caglioti A, Sangiolo F, Novelli G, Perrotti N, Otto EA. Biallelic variants in the ciliary gene TMEM67 cause RHYNS syndrome. *European journal of human genetics : EJHG*. 2018;26(9):1266-71.
 25. Schmidts M, Arts HH, Bongers EM, Yap Z, Oud MM, Antony D, Duijkers L, Emes RD, Stalker J, Yntema JB, Plagnol V, Hoischen A, Gilissen C, Forsythe E, Lausch E, Veltman JA, Roeleveld N, Superti-Furga A, Kutkowska-Kazmierczak A, Kamsteeg EJ, Elçioğlu N, van Maarle MC, Graul-Neumann LM, Devriendt K, Smithson SF, Wellesley D, Verbeek NE, Hennekam RC, Kayserili H, Scambler PJ, Beales PL, Knoers NV, Roepman R, Mitchison HM. Exome sequencing identifies DYNC2H1 mutations as a common cause of asphyxiating thoracic dystrophy (Jeune syndrome) without major polydactyly, renal or retinal involvement. *J Med Genet*. 2013;50(5):309-23.
 26. Baujat G, Huber C, El Hokayem J, Caumes R, Do Ngoc Thanh C, David A, Delezoide AL, Dieux-Coeslier A, Estournet B, Francannet C, Kayirangwa H, Lacaille F, Le Bourgeois M,

- Martinovic J, Salomon R, Sigaudy S, Malan V, Munnich A, Le Merrer M, Le Quan Sang KH, Cormier-Daire V. Asphyxiating thoracic dysplasia: clinical and molecular review of 39 families. *J Med Genet.* 2013;50(2):91-8.
27. Huber C, Wu S, Kim AS, Sigaudy S, Sarukhanov A, Serre V, Baujat G, Le Quan Sang KH, Rimoin DL, Cohn DH, Munnich A, Krakow D, Cormier-Daire V. WDR34 mutations that cause short-rib polydactyly syndrome type III/severe asphyxiating thoracic dysplasia reveal a role for the NF- κ B pathway in cilia. *Am J Hum Genet.* 2013;93(5):926-31.
 28. Schmidts M, Vodopiutz J, Christou-Savina S, Cortés CR, McInerney-Leo AM, Emes RD, Arts HH, Tüysüz B, D'Silva J, Leo PJ, Giles TC, Oud MM, Harris JA, Koopmans M, Marshall M, Elçioğlu N, Kuechler A, Bockenhauer D, Moore AT, Wilson LC, Janecke AR, Hurles ME, Emmet W, Gardiner B, Streubel B, Dopita B, Zankl A, Kayserili H, Scambler PJ, Brown MA, Beales PL, Wicking C, Duncan EL, Mitchison HM. Mutations in the gene encoding IFT dynein complex component WDR34 cause Jeune asphyxiating thoracic dystrophy. *Am J Hum Genet.* 2013;93(5):932-44.
 29. Halbritter J, Porath JD, Diaz KA, Braun DA, Kohl S, Chaki M, Allen SJ, Soliman NA, Hildebrandt F, Otto EA. Identification of 99 novel mutations in a worldwide cohort of 1,056 patients with a nephronophthisis-related ciliopathy. *Hum Genet.* 2013;132(8):865-84.
 30. Stokman M, Lilien M, Knoers N. Nephronophthisis. In: Adam MP, Ardinger HH, Pagon RA, Wallace SE, Bean LH, Mirzaa G, et al., editors. *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2021, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.; 1993.
 31. Parisi MA, Bennett CL, Eckert ML, Dobyns WB, Gleeson JG, Shaw DW, McDonald R, Eddy A, Chance PF, Glass IA. The NPHP1 gene deletion associated with juvenile nephronophthisis is present in a subset of individuals with Joubert syndrome. *Am J Hum Genet.* 2004;75(1):82-91.
 32. Bruel AL, Franco B, Duffourd Y, Thevenon J, Jégo L, Lopez E, Deleuze JF, Doummar D, Giles RH, Johnson CA, Huynen MA, Chevrier V, Burglen L, Morleo M, Desguerres I, Pierquin G, Doray B, Gilbert-Dussardier B, Reversade B, Steichen-Gersdorf E, Baumann C, Panigrahi I, Fargeot-Espaliat A, Dieux A, David A, Goldenberg A, Bongers E, Gaillard D, Argente J, Aral B, Gigot N, St-Onge J, Birnbaum D, Phadke SR, Cormier-Daire V, Eguether T, Pazour GJ, Herranz-Pérez V, Goldstein JS, Pasquier L, Loget P, Saunier S, Mégarbané A, Rosnet O, Leroux MR, Wallingford JB, Blacque OE, Nachury MV, Attie-Bitach T, Rivière JB, Faivre L, Thauvin-Robinet C. Fifteen years of research on oral-facial-digital syndromes: from 1 to 16 causal genes. *J Med Genet.* 2017;54(6):371-80.
 33. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform.* 2013;14:178-92.
 34. Kent WJ. BLAT-the BLAST-like Alignment Tool. *Genome Res.* 2002;12:656-64.
 35. Delvallée C, Nicaise S, Antin M, Leuvrey AS, Nourisson E, Leitch CC, et al. A BBS1 SVA F retrotransposon insertion is a frequent cause of Bardet-Biedl syndrome. *Clin Genet.* 2021;99:318-324.

6.1.3 Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*

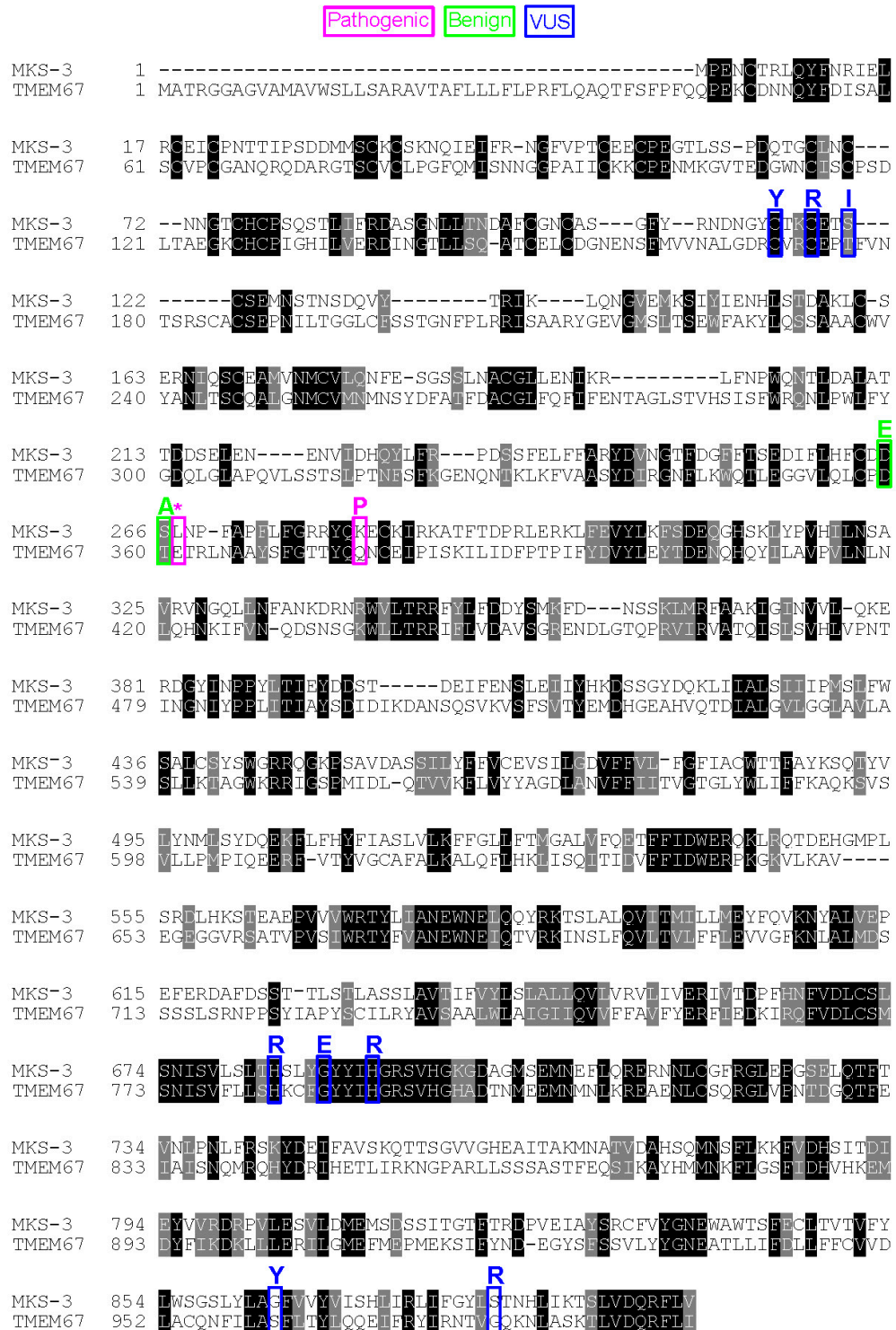


Figure S1. TMEM67 is conserved from humans to worms.

Protein alignment between human TMEM67 (NP_714915) and *C. elegans* MKS-3 (NP_495591.2).

Conservation of identical (black) and similar (grey) amino acids are highlighted. Amino acids that were mutated in this study are labelled green (benign), magenta (pathogenic), and blue (VUS). Amino acid sequences were aligned using Clustal Omega and the figure was generated using BoxShade 3.21.

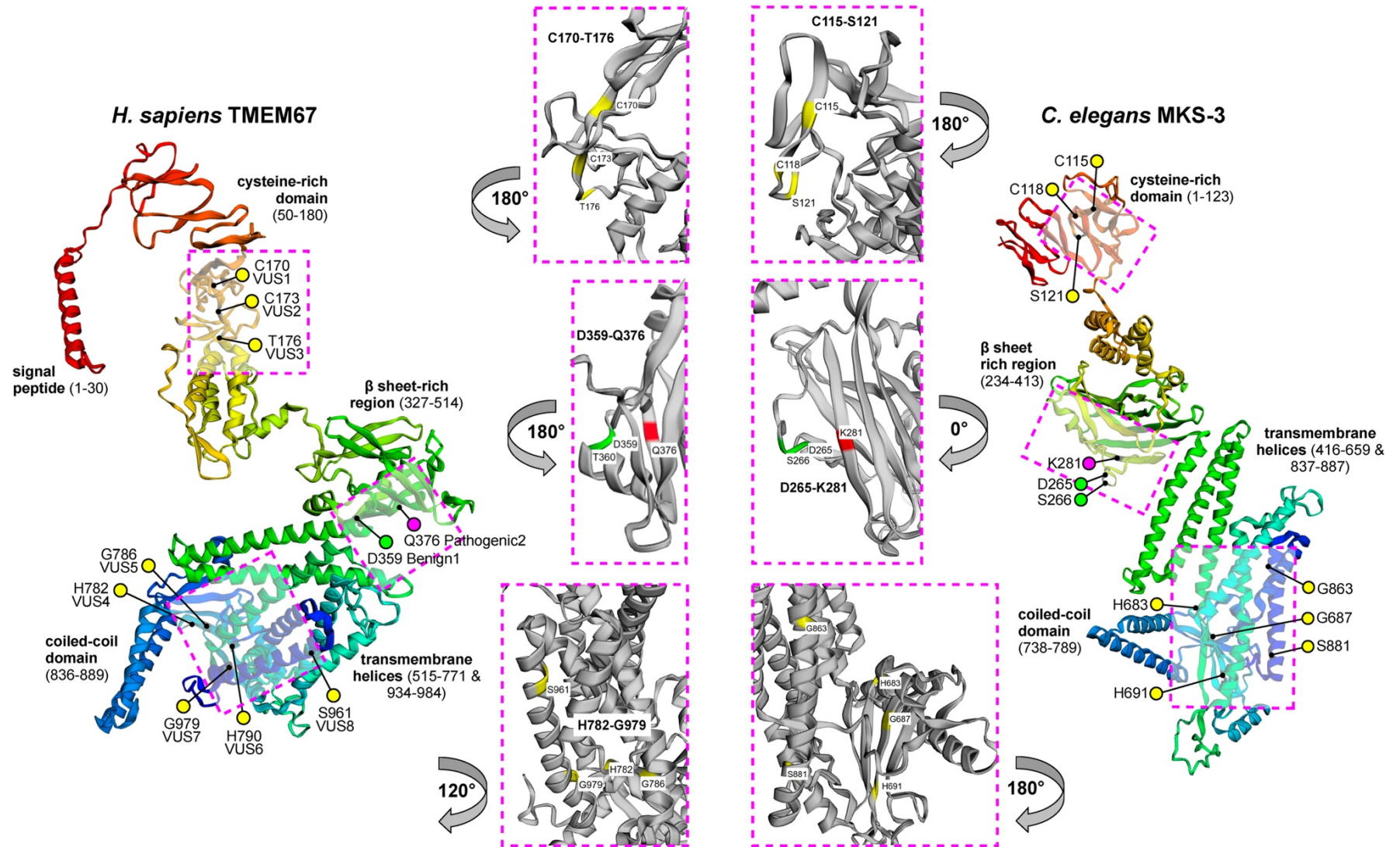


Figure S2. RaptorX predicted structures of TMEM67 and MKS-3

Ribbon diagrams of proteins are rainbow-coloured (red at N-terminus to dark blue at C-terminus) with variants indicated (red, known pathogenic; yellow, VUS; green, known benign). Insets highlight amino acids analyzed in this study.

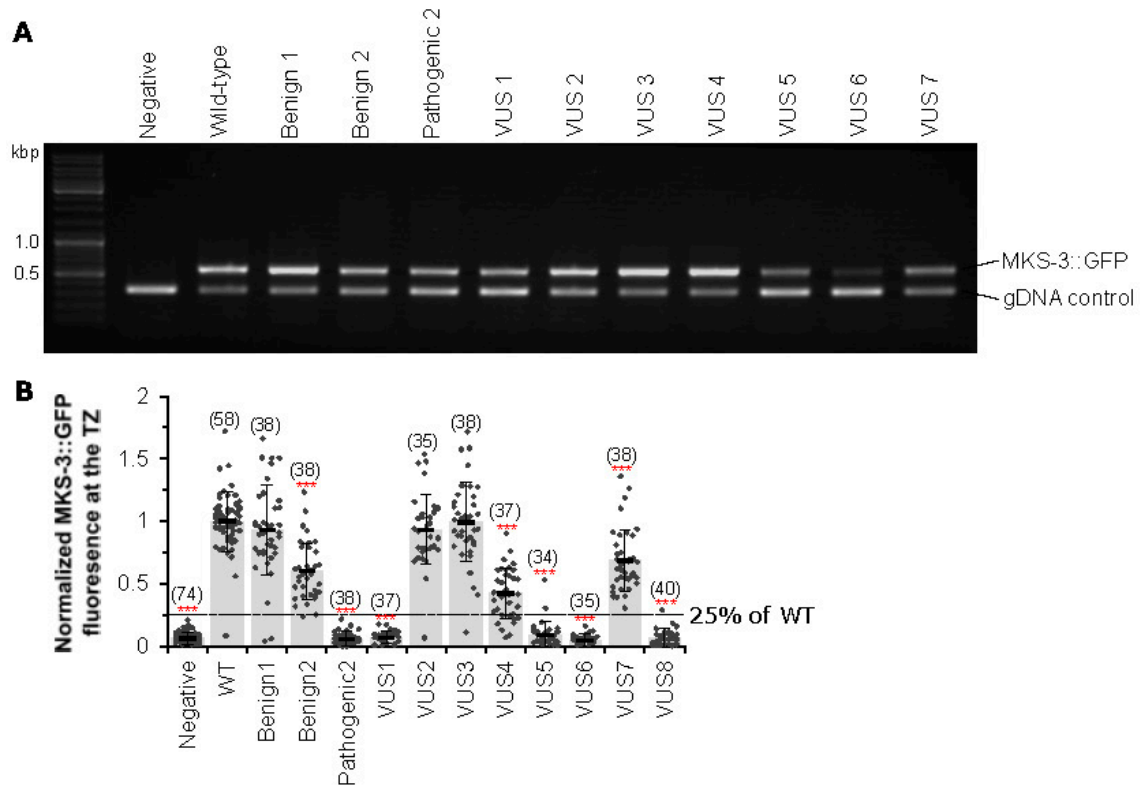


Figure S3. Transgenic MKS-3::GFP

A) PCR products after DNA gel electrophoresis. The upper band is specific to *mks-3::gfp*. The lower band is a gDNA control. Non-injected worms (Negative) do not contain the *mks-3::gfp* product while the transgenic strains do. PCR for *VUS8::gfp* is not shown.

B) Quantification of MKS-3::GFP levels at the transition zone. Background fluorescence was subtracted. Individual dots show each measurement while the bars show the average +/- the standard deviations. The total number of transition zone pairs measured is shown in brackets. If a measurement had more than 25% of the wild-type level (dashed line) we concluded that it was positive for MKS-3::GFP localization to the transition zone.

Statistical significance according to a one way ANOVA followed by Tukey's *post hoc* test and is relative to the wild-type control. *** Indicates a p-value < 0.001.

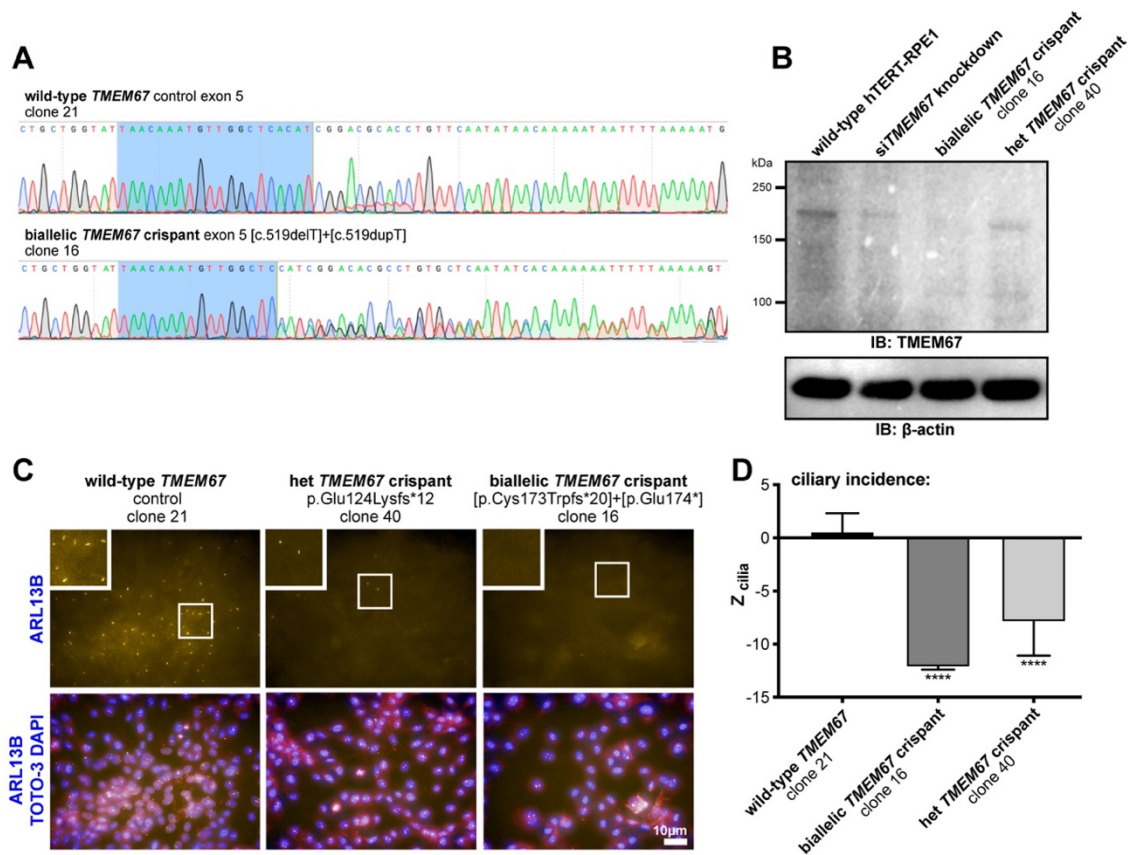


Figure S4. Characterization of *TMEM67* crispr

A) Sanger sequencing electropherograms comparing *TMEM67* exon 5 sequence between wild-type negative control cell-line clone 21 and the bi-allelic crisprant cell-line clone 16. Highlighted sequence in blue indicates the guide RNA sequence used for targeting exon 5. Sequence analysis reveals a one base-pair deletion on one allele and a one base-pair insertion on the other allele corresponding to biallelic frameshift variants: c.519delT, p.(Cys173Trpfs*20) and c.519dupT, p.(Glu174*).

B) Western blotting of protein lysates from untreated wild-type hTERT-RPE-1, wild-type hTERT-RPE-1 following siRNA knockdown of *TMEM67*, the bi-allelic *TMEM67* crisprant clone 16 and the heterozygous *TMEM67* crisprant clone 40, with beta-actin as a loading control. A band visible at ca. 200kDa likely corresponds to post-translationally modified *TMEM67* (expected molecular weight 112kDa). Levels of *TMEM67* expression are most reduced in the bi-allelic knockout crisprant clone 16 (no band visible), with decreased levels observed in the heterozygous crisprant clone 40 and the siRNA *TMEM67* knockdown compared to wild-type.

C) hTERT-RPE-1 cells imaged using an "Operetta" (Perkin-Elmer) high-content imaging system, with representative images from Harmony/Columbus software cilia recognition protocol "find spots". Images show cells stained for the ciliary membrane protein ARL13B (gold), nuclei with DAPI (blue) and cytoplasm with TOTO3 (pink). Significantly fewer cilia were present in the heterozygous *TMEM67* crisprant clone 40 compared to wild-type hTERT-RPE-1, and no cilia are visible in the bi-allelic *TMEM67* crisprant clone 16. Frames indicate the position of magnified insets. Scale bar = 10 μ m.

D) Bar graphs showing mean robust z score for % ciliated cells (z cilia) for wild-type hTERT-RPE-1 (+0.60), the bi-allelic *TMEM67* crisprant clone 16 (-12.15) and the heterozygous *TMEM67* crisprant clone 40 (-7.89). Ciliary incidence is significantly decreased (z cilia < -2.0) in both crisprant clones compared to wild-type.

Figure S5. Prediction of deleteriousness of missense alleles using *in silico* analysis

For each nonsynonymous variant, the text colour denotes if the prediction is tolerated/benign (green), deleterious/damaging (red), or possibly damaging (black). Using these prediction tools we ranked the variants for their overall predicted deleteriousness (1 = benign, 11 = most severe/pathogenic). Variants that are predicted to be more deleterious/damaging are assumed to correlate with disease pathogenesis. The different analyses consistently revealed Benign1 and Benign2 (green background) as benign and the least damaging of all 11 missense mutations. Pathogenic2 (gray background) is identified as deleterious by four of the five prediction tools and ranks as the 3rd most deleterious variant. Overall the *in silico* predictions suggest all eight VUS alleles are likely deleterious/damaging. The only exceptions are VUS3/8 (predicted as probably not damaging by PolyPhen-2 and CADD) and VUS7 (predicted to be tolerated by SIFT).

TMEM67 ^a	Clinical Significance	MISTIC ^b	SIFT ^c	Poly-Phen2 ^d	CADD ^e	REVEL ^f	Overall Rank
D359E	Benign1	0.293	0.27	0.25	17.0	0.338	1
T360A	Benign2	0.437	0.18	0.55	23.2	0.476	2
Q376P	Pathogenic2	0.965	0.14	0.998	26.0	0.935	9
C170Y	VUS1	0.926	0	1	27.1	0.889	6
C173R	VUS2	0.955	0	1	27.2	0.926	10
T176I	VUS3	0.782	0.03	0.681	24.2	0.596	4
H782R	VUS4	0.735	0	0.959	25.3	0.927	5
G786E	VUS5	0.746	0	1	32.0	0.919	8
H790R	VUS6	0.962	0	1	25.7	0.988	11
G979R	VUS7	0.901	0.25	1	28.8	0.958	7
S961Y	VUS8	0.741	0.02	0.495	24.8	0.781	3

a. Amino acid residues correspond to TMEM67 reference sequence NP_714915.

b. MISTIC(MISsense deleTeriousness predICTor) values are from 0-1 with >0.5 being deleterious (Chennen et al. 2020).

c. SIFT(Sorting Intolerant from Tolerant) scores probability of deleteriousness with <0.05 considered significant (Sim et al. 2012). Scores were calculated using an alignment of human TMEM67 with orthologs from mouse, rat, zebrafish, and nematodes.

d. PolyPhen-2(Polymorphism Phenotyping v2) scores range from 0-1 with >0.908 being probably damaging, > 0.446 ≤ 0.908 possibly damaging, and ≤ 0.446 benign (Adzhubei et al. 2010).

e. CADD(Combined Annotation Dependent Depletion) v1.6 phred scores range from 1-99 with higher scores being more deleterious (Rentzsch et al. 2019). We used a cut off of 25.0 because this was the score between the known benign and pathogenic scores.

f. REVEL(Rare Exome Variant Ensemble Learner) scores range from 0-1 with >0.5 being likely pathogenic (Ioannidis et al. 2016).

References

- Adzhubei, Ivan A., Steffen Schmidt, Leonid Peshkin, Vasily E. Ramensky, Anna Gerasimova, Peer Bork, Alexey S. Kondrashov, and Shamil R. Sunyaev. 2010. "A Method and Server for Predicting Damaging Missense Mutations." *Nature Methods* 7 (4): 248–49.
- Chennen, Kirsley, Thomas Weber, Xavière Lornage, Arnaud Kress, Johann Böhm, Julie Thompson, Jocelyn Laporte, and Olivier Poch. 2020. "MISTIC: A Prediction Tool to Reveal Disease-Relevant Deleterious Missense Variants." *PloS One* 15 (7): e0236962.
- Ioannidis, Nilah M., Joseph H. Rothstein, Vikas Pejaver, Sumit Middha, Shannon K. McDonnell, Saurabh Baheti, Anthony Musolf, et al. 2016. "REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants." *American Journal of Human Genetics* 99 (4): 877–85.
- Rentzsch, Philipp, Daniela Witten, Gregory M. Cooper, Jay Shendure, and Martin Kircher. 2019. "CADD: Predicting the Deleteriousness of Variants throughout the Human Genome." *Nucleic Acids Research* 47 (D1): D886–94.
- Sim, Ngak-Leng, Prateek Kumar, Jing Hu, Steven Henikoff, Georg Schneider, and Pauline C. Ng. 2012. "SIFT Web Server: Predicting Effects of Amino Acid Substitutions on Proteins." *Nucleic Acids Research* 40 (Web Server issue): W452–7.

Table S1. Variants analyzed in this study

	Variant	Clinical Significance	GRCh37	cDNA	Protein	Additional information	Conditions	Publications
Benign1	D359E	---	Chr8:94794634	NM_153704.6:c.1077C>G	p.(Asp359Glu)			
Benign2	T360A	Likely Benign	Chr8:94794635	NM_153704.6:c.1078A>G	p.(Thr360Ala)		Joubert syndrome	Consugar, et al. (2015) Genetics in Medicine 17 (4): 253–61.
Pathogenic1	E361*	Pathogenic	Chr8:94794638	NM_153704.6:c.1081G>T	p.(Glu361Ter)	in trans with 5 nt insertion/frameshift	Joubert syndrome	Fleming, et al. (2017) CJASN 12 (12), 1962–1973.
Pathogenic2	Q376P	Pathogenic	Chr8:94794684	NM_153704.6:c.1127A>C	p.(Gln376Pro)	Homozygous	Meckel syndrome	Smith et al. (2006) Nature Genetics 38 (2): 191–96.
VUS1	C170Y	VUS*	Chr8:94777684	NM_153704.5:c.509G>A	p.(Cys170Tyr)	identified in trans with an established pathogenic splice variant (c.224-2A>T)	Meckel syndrome	This study.
VUS2	C173R	VUS	Chr8:94777644	NM_153704.6:c.517T>C	p.(Cys173Arg)	Compound heterozygote with S312P	Joubert syndrome	Huynh, et al. (2018) Clinical Case Reports 6(11), 2189–2192.
VUS3	T176I	VUS*	Chr8:94777654	NM_153704.6:c.527C>T	p.(Thr176Ile)	identified <i>in trans</i> with an established pathogenic variant	Meckel syndrome	This study.
VUS4	H782R	VUS	Chr8:94817012	NM_153704.4:c.2345A>G	p.(His782Arg)		COACH	Brancati, et al. (2009) Human Mutation 30 (2): E432–42.
VUS5	G786E	VUS	Chr8:94817024	NM_153704.6:c.2357G>A	p.(Gly786Glu)	Biallelic with nonsense allele (E848*)	Meckel syndrome	Iannicelliet al. (2010) Human Mutation 31(5), E1319–E1331.
VUS6	H790R	VUS*	Chr8:94817036	NM_153704.5:c.2369A>G	p.(His790Arg)	Identified heterozygous in a parent	Meckel syndrome	This study.
VUS7	G979R	VUS*	Chr8:94828627	NM_153704.5:c.2935G>A	p.(Gly979Arg)	identified in trans with an established pathogenic splice variant(c.507-3C>A)	Meckel syndrome	This study.
VUS8	S961Y	VUS/Likely pathogenic	Chr8:94827650	NM_153704.5:c.2882C>A	p.(Ser961Tyr)	identified <i>in trans</i> with an established pathogenic variant, no functional evidence	Meckel Syndrome	Szymanska, K., et al. (2012). Cilia, 1 (18).

*VUS1(C170Y), VUS3(T176I), VUS6(H790R), and VUS7(G979R) were identified from whole exome sequencing of suspected Meckel syndrome fetuses.

Table S2. Worm strains

	Strain	Genotype	Details
Controls	N2	Wild-type	
		<i>nphp-4(tm925) V</i>	
		<i>mks-3(tm2547) II</i>	949 bp deletion
		<i>mks-3(tm2547) II; nphp-4(tm925) V</i>	
CRISPR mutants	OEB934	<i>mks-3(oq123[K281P]) II; nphp-4(tm925) V</i>	Pathogenic 2, K281P
	OEB935	<i>mks-3(oq124[D265E]) II; nphp-4(tm925) V</i>	Benign 1, D265E
	OEB941	<i>mks-3(oq130[L267*]) II; nphp-4(tm925) V</i>	Pathogenic 1, L267*
	OEB942	<i>mks-3(oq131[S266A]) II; nphp-4(tm925) V</i>	Benign 2, S266A
	OEB943	<i>mks-3(oq132[C115Y]) II; nphp-4(tm925) V</i>	VUS1, C115Y
	OEB944	<i>mks-3(oq133[C118R]) II; nphp-4(tm925) V</i>	VUS2, C118R
	OEB955	<i>mks-3(oq134[H683R]) II; nphp-4(tm925) V</i>	VUS4, H683R
	OEB956	<i>mks-3(oq135[G687E]) II; nphp-4(tm925) V</i>	VUS5, G687E
	OEB957	<i>mks-3(oq136[H691R]) II; nphp-4(tm925) V; unc-58(oq146) X</i>	VUS6, H691R
	OEB1018	<i>mks-3(oq136[H691R]) II; nphp-4(tm925) V</i>	VUS6, H691R, outcrossed 1x
	OEB980	<i>mks-3(oq139[S121I]) II; nphp-4(tm925) V</i>	VUS3, S121I
	OEB981	<i>mks-3(oq140[S881R]) II; nphp-4(tm925) V; unc-58(oq147) X</i>	VUS7, S881R
	OEB1019	<i>mks-3(oq140[S881R]) II; nphp-4(tm925) V</i>	VUS7, S881R, outcrossed 1x
	OEB1001	<i>mks-3(oq145[G863Y]) II; nphp-4(tm925) V</i>	VUS8, G863Y
Extrachromosomal Arrays	OEB990	<i>mks-3(tm2547) II; oqEx122[mks-3p::mks-3::gfp + coel::dsRed]</i>	Wild-type
	OEB991	<i>mks-3(tm2547) II; oqEx123[mks-3p::mks-3(oq124)::gfp + coel::dsRed]</i>	Benign 1, D265E
	OEB992	<i>mks-3(tm2547) II; oqEx124[mks-3p::mks-3(oq131)::gfp + coel::dsRed]</i>	Benign 2, S266A
	OEB993	<i>mks-3(tm2547) II; oqEx125[mks-3p::mks-3(oq123)::gfp + coel::dsRed]</i>	Pathogenic 2, K281P
	OEB994	<i>mks-3(tm2547) II; oqEx126[mks-3p::mks-3(oq132)::gfp + coel::dsRed]</i>	VUS1, C115Y
	OEB995	<i>mks-3(tm2547) II; oqEx127[mks-3p::mks-3(oq133)::gfp + coel::dsRed]</i>	VUS2, C118R
	OEB996	<i>mks-3(tm2547) II; oqEx128[mks-3p::mks-3(oq139)::gfp + coel::dsRed]</i>	VUS3, S121I
	OEB997	<i>mks-3(tm2547) II; oqEx129[mks-3p::mks-3(oq134)::gfp + coel::dsRed]</i>	VUS4, H683R
	OEB998	<i>mks-3(tm2547) II; oqEx130[mks-3p::mks-3(oq135)::gfp + coel::dsRed]</i>	VUS5, G687E
	OEB999	<i>mks-3(tm2547) II; oqEx131[mks-3p::mks-3(oq136)::gfp + coel::dsRed]</i>	VUS6, H691R
	OEB1000	<i>mks-3(tm2547) II; oqEx132[mks-3p::mks-3(oq140)::gfp + coel::dsRed]</i>	VUS7, S881R
	OEB1020	<i>mks-3(tm2547) II; oqEx133[mks-3p::mks-3(oq145)::gfp + coel::dsRed]</i>	VUS8, G863Y

Table S3. Worm crRNA sequences and repair templates

crRNA sequence	Allele	Mutation	ssODN Repair Template Sequence*	Strand
ATCCACGCACATGGTCACTA	---	<i>unc-58</i> co-CRISPR	atTTTgtgtataaaaatagccgagtaggaaacaaatTTTctttcagGT <u>TTT</u> CTGTGCTTACCATGTGCGTGGATCTTGCCTCCACACATCTCAAGGCGTACTT	sense
AAAAGAATGCAAAATAGAA	<i>oq123</i>	Pathogenic 2, K281P	TGAACCCATTGCTCCATTTTATTCGGAAGACGTTATCA <u>CC</u> AGAGTGCAAGAT <u>CCG</u> TAAAGCAACATTACAGATCCACGCTTGGAGAGAAAATATTT	sense
CGAATAAAAATGGAGCAAAAT	<i>oq124</i>	Benign 1, D265E	TTTTTCACTTCGGAAGACATATTTCTTCATTTTGTGAT <u>GAG</u> TCCCTTAATCCATT <u>CG</u> CCCATTTTTATTTCGGAAGACGTTATCAAAAAGAATGCAAAAT	sense
	<i>oq130</i>	Pathogenic 1, L267*	TTTCACTTCGGAAGACATATTTCTTCATTTTGTGATGACTCT <u>TGA</u> AATCCATT <u>CG</u> CCCATTTTTATTTCGGAAGACGTTATCAAAAAGAATGCAAAATAA	sense
	<i>oq131</i>	Benign 2, S266A	TTTTCACTTCGGAAGACATATTTCTTCATTTTGTGATGAC <u>CC</u> CTCAATCCATT <u>CG</u> CCCATTTTTATTTCGGAAGACGTTATCAAAAAGAATGCAAAAT	sense
GGCTTTTACAGAAATGACAA	<i>oq132</i>	VUS1, C115Y	gaatgaacaaacCATTTCGGAACACGACGTTTCACATTCGT <u>ATA</u> GTATCCGTT <u>AT</u> CGTTTCTGTAAAGCCAGAAGCACAGTTCCACAGAATGCATCA	antisense
	<i>oq133</i>	VUS2, C118R	atgtgaatgaacaaacCATTTCGGAACACGACGTTTC <u>ACG</u> CTTCGTGCACTATCCGTT <u>AT</u> CGTTTCTGTAAAGCCAGAAGCACAGTTCCACAGAATGCA	antisense
TCTCTGACCCATTCTCTGTA	<i>oq134</i>	VUS4, H683R	CATCACCCCTTCCATGAACAGAACGACCATGAATATAGTATCCATAGAG <u>GG</u> AG <u>GG</u> GTGAGAGAAAGGACActgaaattgatttaagccgtgaagaa	antisense
	<i>oq135</i>	VUS5, G687E	CCAGCATCACCCCTTCCATGAACAGAACGACCATGAATATAGTAT <u>CT</u> CATAGAG <u>GG</u> AG <u>GG</u> GTGAGAGAAAGGACActgaaattgatttaagccgtgaagaa	antisense
	<i>oq136</i>	VUS6, H691R	TCATTCCAGCATCACCCCTTCCATGAACAGAACGAC <u>CC</u> ATGATAGTATAGTAT <u>ACC</u> ATAGAG <u>GG</u> AG <u>GG</u> GTGAGAGAAAGGACActgaaattgatttaagccgt	antisense
TGTGAATGAACAAACCATTT	<i>oq139</i>	VUS3, S121I	GCTTTTACAGAAATGACAACGGATATTGCACGAAATGTGAAAC <u>ATC</u> TGCTCTGAGATGgTTTgttcattcacatTTtagtgtttctttttggaatact	sense
ACAAGTAACGAAAATCAAA	<i>oq140</i>	VUS7, S881R	TGCTGGTTTTGTTGTGTATGTTATTTCTCACTTATCCG <u>CCT</u> CATCTTCGG <u>AT</u> ACCT <u>CCGT</u> ACGAATCATTGATTAAGACGAGTTTAGTAGATCAACGA	sense
TCTGGATCTTTATATCTTGC	<i>oq145</i>	VUS8, G863Y	ATGTTTAAACAGTTACAGTCTTCTATTTATGGTCTGGATCTTTATACCT <u>CGT</u> <u>TAC</u> TTTGTTGTGTATGTTATTTCTCACTTATCCGTTTgattttCGGT	sense

*Red text indicates mutations and the codon with the amino acid substitution is in bold and underlined. Introns are in lower case.

Table S4. Worm sequencing/PCR primers

ID	Name	Sequence	Purpose
NL132	F35D2.4_R+2720	tggttcaagtcctcggaaatc	Genotype <i>tm2547</i>
NL446	F35D2.4_F+374	tgcccttcacaaagcaactct	Genotype <i>tm2547</i>
NL447	F35D2.4_F+1537	catcatattcaattgttaataacgggtg	Genotype <i>tm2547</i>
KL154	mks-3.F-2	ctATGCCTGAAAATTGTACGAG	Genotype <i>oq132, oq133</i>
KL155	mks-3.R+718	tcaaaatgaaacCTCGCAAC	Genotype <i>oq132, oq133, oq139</i>
KL156	mks-3.C115Y	CGACGTTTCACATTTCTGTatAg	Genotype <i>oq132</i>
KL159	mks-3.C118R	AACACGACGTTTCACGcTTg	Genotype <i>oq133</i>
KL161	mks-3.F+798	TTCACTGAATGCGGTGTGGAC	Genotype <i>oq123, oq124, oq130, oq131</i>
KL162	mks-3.R+1664	AGAGCTGACCAGAACAAATGAC	Genotype <i>oq123, oq131</i>
KL163	mks-3.D265E	gGCgAATGGaTTaAGgGAc	Genotype <i>oq124</i>
KL164	mks-3.S266A	CATTTTTGTGATGACgCcTtc	Genotype <i>oq131</i>
KL165	mks-3.L267Ter	TAAAAATGGgGCgAATGGaTTtc	Genotype <i>oq130</i>
KL166	mks-3.K281P	CTTaCggATcTTGCAcTCTgg	Genotype <i>oq123</i>
KL176	mks-3.R.1123	AAATGGAGCAAAATGGGTTTC	Genotype <i>oq124, oq130, oq131</i>
KL177	mks-3.For-39	CAAATGCTCAGTTTCGTTCCAC	Genotype <i>oq139</i>
KL178	mks-3.F+2936	TCGGGATCGACCAGTCTTG	Genotype <i>oq140</i>
KL179	mks-3.R+3649	caggagatcagtgccaacg	Genotype <i>oq140</i>
KL181	mks-3.F+2125	TGCGAATGAATGGAATGAAC	Genotype <i>oq134, oq135</i>
KL182	mks-3.R+2814	CAGATGTAGTTTGTGGAGAC	Genotype <i>oq134, oq135, oq136</i>
KL184	mks-3.S881R-Rev	TAATCAAATGATTCGTaccgAg	Genotype <i>oq140</i>
KL185	mks-3.H683R-Rev	ATAGTATCCATAgAGgGAgc	Genotype <i>oq134</i>
KL186	mks-3.G687E-For	GACCCAcTcCTcTATGag	Genotype <i>oq135</i>
KL188	mks-3.H691R-For	cTATGGtTACTAcATcCg	Genotype <i>oq136</i>
KL199	mks-3.For+2532	TTCTCTGACCCATTCTCTG	Genotype <i>oq136</i>
KL200	S121I.Rev	caaacCATcTCaGAgCAgat	Genotype <i>oq139</i>
KL183	mks-3.G863Y-Rev	TACACAACAAAgtaAGCgAGg	Genotype <i>oq145</i>
KL230	unc-58. For+3669	GACTCGGAGATATCGTTGTGACTG	<i>unc-58</i> PCR
KL231	unc-58. Rev+4393	CGCGGAGTTCGTTATCCAGGAAG	<i>unc-58</i> PCR
KL232	unc-58. Rev+4367	CGCACATCATTCCATGTAAC	Sequencing primer
NL72	GFP For	AAGCTTGCATGCCTGCAGGTCGACTC	Amplify GFP
NL118	GFP Rev(D-1)	tcacogtcatcaccgaaaacg	Amplify GFP
KL229	mks-3.For-495	tgtctttgactagggcataacccaac	<i>mks-3::gfp</i> stitch (PCR1)
NL441	F35D2.4_F-1196	tggtaatattgctcagtgtttcaattg	<i>mks-3::gfp</i> stitch (PCR1)
NL443	mks-3_GFP_R1	GAGTCGACCTGCAGGCATGCAAGCTTaacaagaaatcgttgatctactaaactcgt	<i>mks-3::gfp</i> stitch (PCR1)
ST54	mks-3_-485_F	taggcataacccaacaatcaac	<i>mks-3::gfp</i> stitch (PCR2)
NL74	GFP Rev (D*)	GGAAACAGTTATGTTTGGTATATTGGG	<i>mks-3::gfp</i> stitch (PCR2)

Table S5. Site Directed Mutagenesis Primers

Target TMEM67 cDNA change	Target TMEM67 protein change	Primer Direction	Sequence (5' to 3')
c.509G>A	Cys170Tyr	Forward	gcttaggagacaggtacgccgatgtgagc
		Complement	gctcacatcggacgtacctgtctcctaaagc
c.515G>A	Arg172Gln	Forward	aaatgttgctcaccatggacgcacctgtctcc
		Complement	ggagacaggtcgcctcaatgtgagccaacatt
c.517T>C	Cys173Arg	Forward	caaatgttgctcaccgtcggacgacacctgtc
		Complement	gacaggtcgcctcggacgtgagccaacattg
c.527C>T	Thr176Ile	Forward	ctgctggtaatacaaatattggctcacatcggacgc
		Complement	gcgtccgatgtgagcacaattttgtaataccagcag
c.1077C>G	Asp359Glu	Forward	cctgtctctgtctctggaacaagctgtaaacac
		Complement	gtgttttacagctttgtccagagacagacacaagg
c.1078A>G	Thr360Ala	Forward	catttagcctgtctctgctctggacaagaagctgta-
		Complement	ttacagctttgtccagacgcagagacaaggctaaatg
c.2935G>A	Gly979Arg	Forward	atgccaaaatctttgtcttctgtattacggataatctaaaatctctgt
		Complement	acaagagattttagataataccgtaatacagtaagacaaaagaaattggcat
c.2882C>A	Ser961Tyr	Forward	tctgtgtagatgtgaggaaatgtgtaaaataaaatttgcaagc
		Complement	gcttgccaaaatatttttagcatacttctacatactcaacaaga
c.1127A>C	Gln376Pro	Forward	gagataggaatctcacaatttggtgtaggtgtccaatg
		Complement	cattggaacaacctaccaacaaattgtgagattcctatctc

Primers were designed using the web-based QuikChange Primer Design Program (<https://www.agilent.com/store/primerDesignProgram.jsp>).

Table S6. TMEM67 Alt-R crRNAs

TMEM67 target exon	crRNA sequence	PAM	Specificity Score	Efficiency Score
2	CAGATGATCTCTAATAATGG	AGG	63.16	71.18
3	CCTAGTGACTTAACTGCCGA	AGG	90.6	63.58
5	TAACAAATGTTGGCTCACAT	CGG	64.11	71.28

Table S7. TMEM67 sequencing/PCR primers

Target	Primer name	Primer sequence (5' to 3')
Internal TMEM67	cDNA1-RV	AACAGTGCTCAGTCCAGCAG
Internal TMEM67	EXON8R	CACACATATTTCCAAGAGCTTGAC
Internal TMEM67	EX4-7R	TTGCAAACCATTCTGAAGTTAAAG
Internal TMEM67	EX4-7F	TTGTGAGCTCTGTGATGGAAA
Internal TMEM67	EXON3F	TGTCCCATTGGCCATATTTT
Internal TMEM67	EX27-3'UTR-R	ACTACACACAATGGGAAAACAGTA
Internal TMEM67	ISH2R	AAAAATATGGCAAACCTAACCTGA
Internal TMEM67	CDNA5-RV	AAAAATATGGCAAACCTAACCTGA
Internal TMEM67	EXONIC 2F	TGTAAAAAGTGCCAGAAAACA
Internal TMEM67	EX27-3'UTR-F	TGTGTTGTGGATTTGGCTTG
Internal TMEM67	ISH3F	TCTTGGCTCCTTCATTGACC
Internal TMEM67	CDNA4-RV	TGGGTACCAAACCTCTCTGG
Internal TMEM67	ISH2F	TCGACAGTTCGTTGATTTATGC
Internal TMEM67	EXON20/21R	CCCACAACCTCCAAAAAGAA
Internal TMEM67	EX19R	GAGCTTATGCAAAAATTGTAGTGC
Internal TMEM67	ISH1R	AGACTGGCTGTTGGCATCTT
Internal TMEM67	CDNA2-RV	TCGAATTACTCTTGGCTGAGTTC
Internal TMEM67	ISH6F	TTTTGGCTGTGCCTGTGTTA
Internal TMEM67	ISH5R	GGAAGCAGCAACAACTTCA
Internal TMEM67	TMEM67_1203_F	GGCTGTGCCTGTGTTAAACC
Internal TMEM67	TMEM67_1481_R	GACTGGCTGTTGGCATCTTTG
Internal TMEM67	TMEM67_1942_F	GTACGAAGTGCCACTGTTCTG
Internal TMEM67	TMEM67_2459_R	GTCTGACCATCTGTGTTGGGT
Internal PX458	T7	TAATACGACTCACTATAGGG
Internal PX458	U6	GACTATCATATGCTTACCGT
Internal PX458	BGHR	TAGAAGGCACAGTCGAGG
Internal PX458	CMV-Forward	CGCAAATGGGCGGTAGGCGTG
Internal PX458	EGFP-C-For	CATGGTCCTGCTGGAGTTCGTG
Internal PX458	EGFP-C-REV	GTTCCAGGGGGAGGTGTG
Internal PX458	EGFP-N	CGTCGCCGTCCAGCTCGACCA
Internal PX458	EXFP-R	GTCTTGTAGTTGCCGTCGTC
Internal PX458	F1ori-F	GTGGACTCTTGTTCCAAACCTGG
Internal PX458	M13 F	TGTA AACGACGGCCAGT
Internal PX458	pBR322ori-F	GGGAAACGCCTGGTATCTTT
Internal PX458	SP6	ATTTAGGTGACACTATAG
Internal PX458	T7	TAATACGACTCACTATAGGG
Internal PX458	SpCas9_1F	GCCAAGGTGGACGACAGCTT
Internal PX458	SpCas9_2F	ACCTACAACCAGCTGTTCCGAGG
Internal PX458	SpCas9_3F	CAAGAACCCTGTCCGACGCC
Internal PX458	SpCas9_4F	GTGAAGCTGAACAGAGAGGACCT

6.2 Custom python scripts

6.2.1 Filter_vep_output_variants.py

```
# Matching VCF file with VEP output to obtain genotype

# WHAT DOES THIS SCRIPT DO
# It takes a CSV file that has been run through VEP and filters based on specified criteria.
#
# STEPS TO USE THIS SCRIPT
# 1) Ensure you are running python in the correct environment.
#   There should be "(idppy3)" at the start of your terminal line
#   If not:
#     1a) ./resources/conda/miniconda3/etc/profile.d/conda.sh
#     1b) conda activate idppy3
# 2)python filter_vep_output_variants.py <enter-the-path-to-your-folder-here> <enter-your-csv-file-name-here>
#
# Note: The csv file must end in ".csv". It must be tab delimited.
# Note: Output files will be placed into the same folder as the source VCF file
#
# ERRORS:
# ERR1: You have not included the file path after running the script
# ERR2: The vcf file you have linked to does not end in ".csv"
# ERR3: The file does not exist at the path you have provided
# ERR4: The VCF file you have provided is not in the expected format. It must be tab delineated.
# ERR5: One of the VCF lines has no VEP output - this should never happen.

# Imports
import pandas as pd # Used for data manipulation
import io # Used to convert the lines of the data files into python-readable data
import os # Used to manipulate file paths
import sys # Used to obtain variables from terminal function call

# Check a file has been provided as an argument
try:
    path_to_folder = sys.argv[1] # Get the file path from the terminal function call
```

```

    csv_file_name = sys.argv[2] # Get the VCF file name from the terminal function call
except:
    raise SystemExit("ERR1: Include the path of the folder and the csv file name when running this script.")

# Check the provided files are on the correct format
if csv_file_name.split('.')[-1] != 'csv':
    raise SystemExit("ERR2: The vcf file to analyse must be in CSV format.")

# Check files specified in the terminal function exist
csv_file_path = os.path.join(path_to_folder, csv_file_name)
if not os.path.isfile(csv_file_path):
    raise SystemExit("ERR3: The csv file provided to analyse does not exist at the following path: \n" + csv_file_path)

# Load CSV file, skipping header rows
try:
    with open(csv_file_path, 'r') as f: # Load the VCF file
        csv_lines = [l for l in f if not l.startswith('##')] # Remove the header lines
    csv = pd.read_csv( # Read the lines into Pandas
        io.StringIO("\n".join(csv_lines)), # Join the lines of the CSV file into data that Pandas can read
        sep='\t' # Tell pandas that our file uses Tabs to separate values
    )
except:
    raise SystemExit("ERR4: The CSV file you have provided is not in the expected format, it must be tab delimited")

print(f'CSV Shape: {csv.shape}')

print(csv.columns.values)

# **** PREPARING COLUMN TYPES FOR FILTERING ****
csv.MAX_AF = csv.MAX_AF.apply(pd.to_numeric, errors='coerce') # Change all the entries in the MAX_AF column to numbers
csv.MAX_AF.fillna(0, inplace=True) # Change empty entries to 0
csv.CADD_PHRED = csv.CADD_PHRED.apply(pd.to_numeric, errors='coerce') # Change all the CADD_PHRED entries to numbers
csv.CADD_PHRED.fillna(0, inplace=True) # Change empty entries to 0

# **** FILTERING ****
filtered_data_1 = csv[csv.MAX_AF <= 0.01] # Remove all entries with gnomAD_AF > 0.01
print(f'MAX_AF filter length: {filtered_data_1.shape}')
filtered_data_1.to_csv(os.path.join(path_to_folder, r'vep_filtered_rare.csv')) # Write the filtered file to the folder

```

```

filtered_data_2 = filtered_data_1[filtered_data_1.IMPACT == "HIGH"]
print(f'Impact filter length: {filtered_data_2.shape}')
filtered_data_2.to_csv(os.path.join(path_to_folder, r'vep_filtered_high_impact.csv')) # Write the filtered file to the folder

filtered_data_3 = filtered_data_1[filtered_data_1.CLIN_SIG.str.contains('pathogenic')]
print(f'Clinvar filter length: {filtered_data_3.shape}')
filtered_data_3.to_csv(os.path.join(path_to_folder, r'vep_filtered_clinvar_pathogenic.csv')) # Write the filtered file to the folder

filtered_data_4 = filtered_data_1[filtered_data_1.Consequence.str.contains('missense')]
print(f'Consequence filter length: {filtered_data_4.shape}')
filtered_data_4.to_csv(os.path.join(path_to_folder, r'vep_filtered_missense_all.csv')) # Write the filtered file to the folder

filtered_data_5 = filtered_data_4[filtered_data_4.CADD_PHRED >= 15]
print(f'Consequence and CADD filter length: {filtered_data_5.shape}')
filtered_data_5.to_csv(os.path.join(path_to_folder, r'vep_filtered_missense_CADD.csv')) # Write the filtered file to the folder

filtered_data_6 = filtered_data_1[filtered_data_1.Consequence.str.contains('splice_region')]
print(f'Splice filter length: {filtered_data_6.shape}')
filtered_data_6.to_csv(os.path.join(path_to_folder, r'vep_filtered_splice_region.csv')) # Write the filtered file to the folder

```

6.2.2 filter_gene_variant_workflow.py

```

# WHAT DOES THIS SCRIPT DO
# It filters a TSV output file generated by the GeneVariantWorkflow script based
# on specified criteria to give several shorter lists of interesting variants
# to analyse.
#
# Filtering steps include removal of common variants from gnomAD and the 100K
# rare disease dataset, then filtering into separate files for rare homozygous
# variants, rare high impact variants, rare ClinVar pathogenic variants and rare
# missense variants.
#
# GeneVariantWorkflow will have extracted all variants across the 100K rare
# disease dataset in a given gene and annotated them with Ensembl VEP
#

```

```

# STEPS TO USE THIS SCRIPT WITHIN THE RESEARCH ENVIRONMENT
# 1) Ensure you are running python in the correct environment.
#   There should be "(idppy3)" at the start of your terminal line
#   If not:
#     1a) ./resources/conda/miniconda3/etc/profile.d/conda.sh
#     1b) conda activate idppy3
# 2) Navigate to the GeneVariantWorkflow folder containing this script
#   (for me this is under /re_gecip/GW_SB/GeneVariantWorkflow)
# 3) To run on the command line enter:
#   python filter_gene_variant_workflow.py <enter-the-path-to-your-folder-here> <enter-your-tsv-file-name-here>
#
# NOTE: path to your folder should be the one containing the GeneVariantWorkflow data output files: for me it's at
# /home/sbest1/re_gecip/shared_allGeCIPs/GW_SB/GeneVariantWorkflow/<gene_name>/v1.7/final_output/data
# Note: The tsv file must end in ".tsv". It must be tab delimited.
# Note: Output files will be placed into a new folder named after the input tsv file within the final_output/data folder that contains your input
# file.
#
# ERRORS:
# ERR1: You have not included the file path after running the script
# ERR2: The input file you have linked to does not end in ".tsv"
# ERR3: The input file does not exist at the path you have provided
# ERR4: The input tsv file you have provided is not in the expected format.
#   It must be tab delineated.

# Imports
import pandas as pd # Used for data manipulation
import io # Used to convert the lines of the data files into python-readable data
import os # Used to manipulate file paths
import sys # Used to obtain variables from terminal function call

# Check a file has been provided as an argument
try:
    path_to_folder = sys.argv[1] # Get the file path from the terminal function call
    tsv_file_name = sys.argv[2] # Get the VCF file name from the terminal function call
except:
    raise SystemExit("ERR1: Include the path of the folder and the tsv file name when running this script.")

# Check the provided files are on the correct format

```

```

if tsv_file_name.split('.')[-1] != 'tsv':
    raise SystemExit("ERR2: The file to analyse must be in TSV format.")

# Check files specified in the terminal function exist
tsv_file_path = os.path.join(path_to_folder, tsv_file_name)
if not os.path.isfile(tsv_file_path):
    raise SystemExit("ERR3: The tsv file provided to analyse does not exist at the following path: \n" + tsv_file_path)

# Load CSV file, skipping header rows
try:
    tsv = pd.read_csv( # Read the lines into Pandas
        tsv_file_path, # Join the lines of the CSV file into data that Pandas can read
        sep='\t' # Tell pandas that our file uses Tabs to separate values
    )
except:
    raise SystemExit("ERR4: The tsv file you have provided is not in the expected format, it must be tab delimited")

# If folder for results doesn't exist then create it
folder_name = tsv_file_name.split('.')[0]
path_to_save_folder = os.path.join(path_to_folder, folder_name)
if not os.path.exists(path_to_save_folder):
    os.mkdir(path_to_save_folder)

print(f'TSV Shape: {tsv.shape}')

print(tsv.columns.values)

# **** PREPARING COLUMN TYPES FOR FILTERING ****
tsv.MAF_variant = tsv.MAF_variant.apply(pd.to_numeric, errors='coerce') # Change all the entries in the MAX_AF column to numbers
tsv.MAF_variant.fillna(0, inplace=True) # Change empty entries to 0
tsv.gnomAD_AF_annotation = tsv.gnomAD_AF_annotation.apply(pd.to_numeric, errors='coerce') # Change all the entries in the
gnomAD_AF column to numbers
tsv.gnomAD_AF_annotation.fillna(0, inplace=True) # Change empty entries to 0
tsv.AC_Hom_variant = tsv.AC_Hom_variant.apply(pd.to_numeric, errors='coerce') # Change all the entries in the AC_Hom_variant column
to numbers
tsv.AC_Hom_variant.fillna(0, inplace=True) # Change empty entries to 0

# **** FILTERING ****

```

```

filtered_data_1 = tsv[tsv.MAF_variant <= 0.002] # Remove all entries with MAF_variant > 0.002. These are common variants (>0.2%) called
in the 100K data set
print(f'Rare 100K dataset filter length: {filtered_data_1.shape}')
filtered_data_1.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_gel_rare.csv')) # Write the filtered file to the new folder
del tsv

# **** FILTERING ****
filtered_data_2 = filtered_data_1[filtered_data_1.gnomAD_AF_annotation <= 0.002] # Remove all entries with gnomAD_AF_annotation >
0.002. These are common variants (>0.2%) called in the gnomAD dataset.
print(f'Rare 100K and gnomAD filter length: {filtered_data_2.shape}')
filtered_data_2.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_gel_gnomAD_rare.csv')) # Write the filtered file to the
folder

# **** FILTERING ****
filtered_data_3 = filtered_data_2[filtered_data_2.CANONICAL_annotation == "YES"] # Retain only variants called in the canonical
transcript.
print(f'Rare canonical transcript filter length: {filtered_data_3.shape}')
filtered_data_3.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_canonical_transcript_rare.csv')) # Write the filtered file to
the folder

# **** FILTERING ****
filtered_data_4 = filtered_data_3[filtered_data_3.AC_Hom_variant > 0] # Remove all entries with AC_Hom_variant = 0. This will leave
variants called at least once as homozygous in the rare disease 100K dataset.
print(f'Homozygous filter length: {filtered_data_4.shape}')
filtered_data_4.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_rare_homozygous.csv')) # Write the filtered file to the
folder

# **** FILTERING ****
filtered_data_5 = filtered_data_3[filtered_data_3.IMPACT_annotation == "HIGH"] # Retain all entries with "HIGH" in the Impact_annotation
column. This will give a file of variants rare in 100K and gnomAD that are high impact.
print(f'High impact filter length: {filtered_data_5.shape}')
filtered_data_5.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_rare_HIGH_impact.csv')) # Write the filtered file to the
folder
del filtered_data_5

filtered_data_8 = filtered_data_4[filtered_data_4.IMPACT_annotation == "HIGH"] # Retain all entries with "HIGH" amongst the homozygous
variants
print(f'Homozygous high impact filter length: {filtered_data_8.shape}')

```



```

filtered_data_8.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_homozygous_rare_HIGH_impact.csv')) # Write the
filtered file to the folder
del filtered_data_8

# **** FILTERING ****
filtered_data_6 = filtered_data_3[filtered_data_3.ClinVar_CLNSIG_annotation.str.contains('[Pp]athogenic$', regex=True)] # Retain all
entries annotated as pathogenic in ClinVar_CLNSIG_annotation column. This catches variants rare in 100K and gnomAD called 'pathogenic'
and 'likely_pathogenic'.
print(f'Pathogenic filter length: {filtered_data_6.shape}')
filtered_data_6.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_rare_ClinVAR_pathogenic.csv')) # Write the filtered file
to the folder
del filtered_data_6

# **** FILTERING ****
filtered_data_7 = filtered_data_3[filtered_data_3.Consequence_annotation.str.contains('missense')] # Retain all entries with missense in
CLIN_SIG_annotation column. Gives output file of rare missenses from 100K and gnomAD. NOTE: don't have CADD scores, may need to
run through VEP with plugins if want this.
print(f'Missense filter length: {filtered_data_7.shape}')
filtered_data_7.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_rare_missense_all.csv')) # Write the filtered file to the
folder

filtered_data_9 = filtered_data_7[filtered_data_7.SIFT_annotation.str.contains('deleterious')] # Retain all entries with SIFT_annotation entry
containing deleterious. NOTE: don't have CADD scores, may need to run through VEP with plugins if want this.
print(f'Missense SIFT deleterious filter length: {filtered_data_9.shape}')
filtered_data_9.to_csv(os.path.join(path_to_save_folder, r'gene_variant_workflow_rare_missense_SIFT_deleterious.csv')) # Write the
filtered file to the folder

print("Complete!")

```

6.2.3 find_variants_by_gene_and_consequence.py

```

## Script to do initial PanelApp filtering to find variants in genes of interest with particular consequences (can be specified by altering the
CQ_dict dictionary)
## Input files are specified on the command line and are as follows:
## --samples: Tab separated list of "ID      vcf location" - should include full path to VCFs (see example_sample_file.txt)
## --panels: Tab separated list of "ID panel name" - should include full paths to panels files (see example_panel_file.txt)

```

```

## --genes: List of other genes of interest - just a list of gene names, one per line (see example_gene_file.txt)
## example running command:
## python find_variants_by_gene_and_consequence.py --samples example_sample_file.txt --panels example_panel_file.txt --genes
example_gene_file.txt

import gzip
import os
import argparse

def get_options():
    parser = argparse.ArgumentParser(description="###")
    parser.add_argument("--samples", required=True, help="# samples file to process")
    parser.add_argument("--panels", required=True, help="# file linking sample to panel(s)")
    parser.add_argument("--genes", required=True, help="# additional gene list to check for all samples")
    args = parser.parse_args()
    return args
args = get_options()

## Set up input and output files:
infile_samples = open(args.samples)
infile_panels = open(args.panels)
infile_genes = open(args.genes)

Outfile = ".join((args.samples, "_variants_out1.txt")) ## output will be named after the samples file specified with _variants_out1.txt
appended
outfile = open(Outfile, 'w')

## Store which panels are relevant for which samples in a dictionary so this file is only parsed once
panel_dict = {}
for line in infile_panels:
    line = line.strip()
    words = line.split("\t")
    if words[0] not in panel_dict: ## If this ID hasn't yet been stored as a key,
        panel_dict[words[0]] = words[1] ## Add the ID to the dictionary with this panel as the value
    else: ## If this ID already has an entry
        panel_dict[words[0]] = panel_dict[words[0]] + ';' + words[1] ## Add this panel to that samples entry separated from previous
panels by ;

```

```
## Any variants with consequences in this dictionary will be pulled out - consequences can be added or removed as needed
CQ_dict = {"stop_gained": 0, "splice_acceptor": 0, "splice_donor": 0, "frameshift": 0, "missense": 0, "splice_region": 0}
```

```
tracking = {} ## setting up a dictionary to store variants that have already been output so we don't get duplicate lines in the output
```

```
for line in infile_samples:
```

```
    line = line.strip()
```

```
    words = line.split('\t')
```

```
    if line.startswith('Participant'): continue ## skip header if present
```

```
    ID = words[0]
```

```
    vcf_file_loc = words[1]
```

```
    gene_dict = {} ## This is to store all the relevant gene names for that individual
```

```
    if ID not in panel_dict: ## If this ID doesn't have any panels stored, print an error then skip it
```

```
        print ID, " - panels unknown"
```

```
    if ID not in panel_dict: continue
```

```
    panels = panel_dict[ID].split(';')
```

```
    for i in panels:
```

```
        panel_file = open(i)
```

```
        for iline in panel_file:
```

```
            iline = iline.strip()
```

```
            iwords = iline.split('\t')
```

```
            if len(iwords) < 14: continue ## skip lines missing info (added to address an error where some files had incomplete
```

```
lines)
```

```
                if iwords[1] != "gene": continue ## skip non-gene entries
```

```
                if "Expert Review Green" not in iwords[3]: continue ## This skips low confidence genes. This can be commented out
```

```
if those are wanted.
```

```
                if iwords[7].startswith("MONOALLELIC") or iwords[7].startswith("BOTH") or iwords[7].startswith("BIALLELIC"): ## this
```

```
can be modified depending on the type of genes you want to include
```

```
                    gene_dict[iwords[2]] = 0
```

```
infile_genes = open(args.genes)
```

```
for lineG in infile_genes: ## go through the additional genes file and add any additional gene names to the gene dictionary
```

```
    lineG = lineG.strip()
```

```
    wordsG = lineG.split('\t')
```

```
    gene_dict[wordsG[0]] = 0
```

```
## Now go through the actual VCFs and start finding variants
```

```
if os.path.exists(vcf_file_loc): ## Check the VCF exists before trying to open it
```

```
    vcf_file = gzip.open(vcf_file_loc) ## If VCFs aren't gzipped, remove "gzip."
```

```
    for Line in vcf_file:
```

```

Line = Line.strip()
Words = Line.split('\t')
if Line.startswith('#'): continue ## Skip vcf headers
if Words[6] != "PASS": continue ## Skip anything that doesn't have a PASS in the filter column
if Words[9].startswith('0/0'): continue ## Skip anything where the proband doesn't actually have a variant here
get_info = Words[7].split(';') ## split up the info field for parsing
count = 0
for i in get_info:
    if i.startswith("CSQT="): ## Pull out the bit of the info field that's got variant annotation in it
        split_info = i.split(',')
        count = int(count) +1
if count == 0: continue ## This skips any lines that don't have VEP variant information
for i in split_info: ## Go through each part of the split annotation information
    for j in CQ_dict: ## Check if any consequences from our dictionary are in it
        if j in i: ## If this variant's CQ is something we're interested in...
            for k in gene_dict: ## See if it's in a gene we're interested in...
                gene = ".join(("|",k,"|")) ## Added this in so it matches on complete gene name (before,
it would have pulled out anything where a dictionary gene was in another gene's name, e.g. CR1 in dictionary would have pulled out variants
in CR1 but also CR1L)
                if gene in i:
                    ## check the depth is at least 5 (this can be adjusted)
                    get_DP = Words[9].split(':')
                    DP = get_DP[3] ## NB this will need to be changed if DP is not always in this
position
                    if int(DP) < 6: continue ## check the depth is at least 5 reads (can be changed)
                    ## Store the ID and variant to prevent duplicates in the output
                    ID_var = '-'.join((words[0], Words[0], Words[1], Words[3], Words[4]))
                    if ID_var in tracking: continue ## Skips any entries that have already been
stored/output
                    tracking[ID_var] = 0
                    outline = ".join((words[0], '\t', Line, '\t', k, '\t', j, '\n')) ## this outputs the ID, the full
VCF line, the gene and the consequence
                    outfile.write(outline)
            else:
                print "File ", vcf_file_loc, "not found" ## prints an error if the VCF doesn't exist which allows it to carry on processing other
samples rather than failing
outfile.close()

```

6.2.4 find_variants_by_gene_and_SpliceAI_score.py

```

## Script to do initial PanelApp filtering to find variants in genes of interest with SpliceAI scores >=0.2
## Input files are specified on the command line and are as follows:
## --samples: Tab separated list of "ID      vcf location" - should include full path to VCFs (see example_sample_file.txt)
## example running command:
## python find_variants_by_gene_and_SpliceAI_score.py --samples example_sample_file.txt
## NB - panel_file = open("/path/to/panel_file.tsv") needs to be changed to your actual panel file
## NB - SAI_snvs = gzip.open("/path/to/spliceai_scores.masked.snv.hg38.vcf.gz") needs to be changed to your SpliceAI file location
## NB - SAI_indels = gzip.open("/path/to/spliceai_scores.masked.indel.hg38.vcf.gz") needs to be changed to your SpliceAI file location

import gzip
import os
import argparse

def get_options():
    parser = argparse.ArgumentParser(description="###")
    parser.add_argument("--samples", required=True, help="# samples file to process")
    args = parser.parse_args()
    return args
args = get_options()

## Set up input and output files:
infile_samples = open(args.samples)
Outfile = ".join((args.samples, "_variants_out_SpliceAI.txt")) ## output will be named after the samples file specified with
_ variants_out_SpliceAI.txt appended
outfile = open(Outfile, 'w')

## Go through the SpliceAI files and store variants in genes of interest with SpliceAI scores >= 0.2 (can be customised)
gene_dict = {}
gene_panel = open("/path/to/panel_file.tsv") ## change this to match the panel you want to use

for line in gene_panel:
    line = line.strip()
    words = line.split('\t')

```

```

    if len(words) <14: continue ## skip lines missing info (added to address an error where some files had incomplete lines)
    if words[1] != "gene": continue ## skip non-gene entries
    if "Expert Review Green" not in words[3]: continue ## Skip unconfirmed/low confidence genes
    if words[7].startswith("MONOALLELIC") or words[7].startswith("BOTH") or words[7].startswith("BIALLELIC"): ## this can be modified
depending on the type of genes you want to include
        gene_dict[words[2]] = 0

```

```

SAI_snvs = gzip.open("/path/to/spliceai_scores.masked.snv.hg38.vcf.gz") ## change this to reflect your SpliceAI file location
SAI_indels = gzip.open("/path/to/spliceai_scores.masked.indel.hg38.vcf.gz") ## change this to reflect your SpliceAI file location

```

```
SAI_dict = {}
```

```
## SpliceAI SNVs file
```

```
for line in SAI_snvs:
```

```
    line = line.strip()
```

```
    words = line.split('\t')
```

```
    if line.startswith('#'): continue ## skip headers
```

```
    info = words[7].split('|') ## split the info for parsing: 2-5 are scores, 6-9 are locations
```

```
    gene = info[1]
```

```
    if gene not in gene_dict: continue ## skip entries that aren't in genes of interest
```

```
    max = 0.00 ## set baseline maximum to 0 to compare scores against
```

```
    for i in info[2:6]: ## for each of the scores
```

```
        if float(i) > float(max): ## see if the score is higher than the current max
```

```
            max = i ## if current score is higher than current max, store the score as max
```

```
    if float(max) >= 0.2: ## if the maximum score is greater than 0.2 we'll output the variant to a temporary SpliceAI subset which can
be deleted after
```

```
        variant = ".join(("chr", words[0], "-", words[1], "-", words[3], "-", words[4]))
```

```
        SAI_dict[variant] = line
```

```
## SpliceAI indels file
```

```
for line in SAI_indels:
```

```
    line = line.strip()
```

```
    words = line.split('\t')
```

```
    if line.startswith('#'): continue ## skip headers
```

```
    info = words[7].split('|') ## split the info for parsing: 2-5 are scores, 6-9 are locations
```

```
    gene = info[1]
```

```
    if gene not in gene_dict: continue ## skip entries that aren't in genes of interest
```

```
    max = 0.00 ## set baseline maximum to 0 to compare scores against
```

```

for i in info[2:6]: ## for each of the scores
    if float(i) > float(max): ## see if the score is higher than the current max
        max = i ## if current score is higher than current max, store the score as max
    if float(max) >= 0.2: ## if the maximum score is greater than 0.2 we'll output the variant to a temporary SpliceAI subset which can
be deleted after
        variant = ".join(("chr", words[0], "-", words[1], "-", words[3], "-", words[4]))
        SAI_dict[variant] = line

## Go through VCFs and see if any probands have variants which were stored from SpliceAI files
for line in infile_samples:
    line = line.strip()
    words = line.split('\t')
    if line.startswith('Participant'): continue ## skip header if present
    ID = words[0]
    vcf_file_loc = words[1]
    if os.path.exists(vcf_file_loc): ## Check the VCF exists before trying to open it
        vcf_file = gzip.open(vcf_file_loc) ## If VCFs aren't gzipped, remove "gzip."
        for Line in vcf_file:
            Line = Line.strip()
            Words = Line.split('\t')
            if Line.startswith('#'): continue ## Skip vcf headers
            if Words[6] != "PASS": continue ## Skip anything that doesn't have a PASS in the filter column
            if Words[9].startswith('0/0'): continue ## Skip anything where the proband doesn't actually have a variant here
            variant = '-'.join((Words[0], Words[1], Words[3], Words[4]))
            if variant not in SAI_dict: continue
            get_DP = Words[9].split(':')
            DP = get_DP[3] ## NB this will need to be changed if DP is not always in this position
            if int(DP) < 6: continue ## check the depth is at least 5 reads (can be changed)
            outline = ".join((ID, '\t', Line, '\t', SAI_dict[variant], '\n'))
            outfile.write(outline)
        else:
            print "File ", vcf_file_loc, "not found"
outfile.close()

```

6.3 Reagents

6.3.1 Suppliers

Table 8. List of suppliers for reagents used

Company Name	Address
Abcam plc.	Discovery Drive Cambridge Biomedical Campus, Cambridge, CB2 0AX, U.K.
Addgene	490 Arsenal Way, Suite 100, Watertown, MA 02472, U.S.A
American Type Culture Collection® (ATCC®)	10801 University Boulevard, Manassas, VA 20110, U.S.A
Applied Biosystems™	120 Birchwood Blvd, Birchwood, Warrington WA3 7QH, U.K.
Bio-Rad	The Junction 3rd And 4th Floor, Station Road, Watford, WD17 1ET, U.K.
Bioline	Edge Business Centre, Humber Rd, London NW2 6EW, U.K.
Clent Life Science	Suite 3, Faraday House, King William St, Amblecote, Stourbridge DY8 4HD, U.K.
Corning	Elwy House, Lakeside Business Village, St Davids Park Ewloe, Flintshire, CH5 3XD, U.K.
Dako, Agilent Technologies	5301 Stevens Creek Blvd., Santa Clara, CA 95051, U.S.A
Dharmacon	Horizon Discovery Ltd. 8100 Cambridge Research Park, Waterbeach, Cambridge, CB25 9TL, U.K.
FluidX Ltd	Northbank Industrial Park, Gilchrist Road, Irlam, Manchester, M44 5AY, U.K.
Gibco™, Life Technologies	3 Fountain Drive, Inchinnan Business Park, Paisley, PA4 9RF, U.K.
Invitrogen™	3 Fountain Drive, Inchinnan Business Park, Paisley, PA4 9RF, U.K.
Melford Laboratories Ltd	Bildeston Rd, Ipswich IP7 7LE
Merck Millipore	Suite 21, Building 6, Croxley Green Business Park, Watford, Hertfordshire, WD18 8YH, U.K.
New England Biolabs	75-77 Knowl Piece, Wilbury Way, Hitchin, Hertfordshire, SG4 0TY, U.K.
Nippon Genetics Europe	Binsfelder Street 77, 52351 Dueren, Germany
Perkin Elmer	Chalfont Road Buckinghamshire, Seer Green, HP9 2FX, U.K.

Polysciences Inc	Badener Str. 13, 69493 Hirschberg an der Bergstrasse, Germany
Premier Foods Plc	Premier House, Griffiths Way, St Albans, Hertfordshire, AL1 2RE, U.K.
Qiagen	Skelton House Lloyd Street North, Manchester, M15 6SH, U.K.
Scientific Laboratory Supplies	Wilford Industrial Estate, Ruddington Lane, Wilford, Nottingham, NG11 7EP, U.K.
Sigma-Aldrich	The Old Brickyard, New Rd, Gillingham, Dorset, SP8 4XT, U.K.
Thermo-Fisher Scientific TM	Bishop Meadow Rd, Loughborough LE11 5RG, U.K.

6.3.2 Reagents

Table 9. List of reagents used.

Category	Reagent	Supplier
General reagents	Nuclease free water	Merck Millipore
	Methanol	Sigma-Aldrich
	Ethanol	Sigma-Aldrich
	Isopropanol	Sigma-Aldrich
PCR	Hot-shot Diamond PCR Mastermix	Clent Life Science
	Primers, 25nmol (full list in Table S7 of the Supplementary Material (thesis section 6.1.3) (Lange et al., 2022))	Sigma-Aldrich
Genotyping	BigDye TM Terminator v3.1 Cycle Sequencing Kit	Applied Biosystems TM
	Hi-Di TM Formamide	Thermo Fisher Scientific TM
	ExoSAP-IT TM PCR Product Clean-up Reagent	Applied Biosystems TM
Gel Electrophoresis	Agarose	Thermo Scientific TM
	EDTA	Sigma-Aldrich

	Midori Green Advance DNA/RNA stain	Nippon Genetics
	Easyladder I	Bioline
	Quick-Load® Purple 1 kb DNA Ladder	New England Biolabs
DNA extraction	DirectPCR Reagent	Viagen Biotech
	Proteinase K	Sigma-Aldrich
Cloning	Luria-Bertani Medium (LB) (25g dissolved in 1l of water and autoclaved prior to use)	Sigma-Aldrich
	Agar	Sigma-Aldrich
	Ampicillin	Melford Laboratories Ltd
	“α-Select Gold” DH5α Chemically Competent E. coli Cells	Bioline
	QIAprep Spin Miniprep Kit	Qiagen
	QIA filter Plasmid Maxi Kit	Qiagen
	Glycerol	Sigma-Aldrich
	Super Optimal Broth with Catabolite Repression (SOC) Outgrowth Medium	New England Biolabs
	NEBuilder® HiFi DNA Assembly Master Mix/NEBuilder HiFi DNA Assembly Cloning Kit	New England Biolabs
Tissue culture	Dubecco's Modified Essential Medium (DMEM)	Gibco™
	DMEM/F-12, GlutaMAX™ Supplement	Gibco™
	Fetal Bovine Serum	Sigma-Aldrich
	Trypsin	Sigma-Aldrich
	Ca ²⁺ /Mg ²⁺ free phosphate-buffered saline (PBS)	Sigma-Aldrich
	Lipofectamine® 2000 Transfection Reagent	Invitrogen™

	Polyethylenimine (PEI) linear (1 mg/mL) Transfection Reagent	PEI powder from Polysciences Inc PEI reagent home-made as per Cold Spring Harbor Protocols, pdb.rec11323–pdb.rec11323 (2008). doi: 10.1101/pdb.rec11323
	Opti-MEM™ Reduced Serum Medium	Thermo-Fisher Scientific
	0.4% Trypan Blue viability stain	Gibco™
	Dimethyl sulphoxide (DMSO)	Sigma-Aldrich
	siRNA duplexes (5nmol)	Dharmacon
	Matrigel® Matrix	Corning
Site-directed mutagenesis	QuikChange II XL Site-Directed Mutagenesis Kit	Agilent
HiFi cloning	NEBuilder® HiFi DNA Assembly Master Mix/NEBuilder HiFi DNA Assembly Cloning Kit	New England Biolabs
Immunofluorescence	Methanol	Sigma-Aldrich
	“Marvel” Non-fat skimmed dried milk	Premier Foods PLC
	Triton™ X-100	Sigma-Aldrich
	ProLong™ Gold Antifade Mountant	Thermo Fisher Scientific™
Western Blotting	100x Halt™ Protease Inhibitor Cocktail 100x / Halt™ Phosphatase Inhibitor Cocktail	Thermo Fisher Scientific™
	NP40 (IGEPAL)	Thermo Fisher Scientific™
	RC DC™ Protein Assay Kit	Bio-Rad
	NuPAGE™ 4-12% MES SDS gels	Thermo Fisher Scientific™
	NuPAGE™ MES running buffer	Thermo Fisher Scientific™
	NuPAGE™ transfer buffer	Thermo Fisher Scientific™

	Invitrolon™ PVDF filter paper sandwich	Thermo Fisher Scientific™
	Precision Plus Protein™ All Blue Prestained Protein Standard	Bio-Rad
	SuperSignal™ West Femto Maximum Sensitivity Substrate	Thermo Fisher Scientific™
	NuPAGE™ LDS Sample Buffer (4X)	Thermo Fisher Scientific™
	Beta-Mercaptoethanol	Thermo Fisher Scientific™
	Restore™ PLUS Western Blot Stripping Buffer	Thermo Fisher Scientific™
CRISPR/Cas9 genome editing	Green-fluorescent protein (GFP) expressing CRISPR-Cas9 PX458 vector	Addgene
	crRNA	Integrated DNA technologies (IDT)
	Alt-R CRISPR-Cas9 tracrRNA – ATTO™ 550	IDT
	Nuclease free duplex buffer	IDT

6.3.3 Buffers and Solutions

Table 10. List of buffers and solutions used.

Buffer/reagent	Component	Amount	Notes
1X phosphate buffered saline (PBS)	Tablet PBS	x 5	Autoclaved and filter sterilised
	dH ₂ O	1l	
1x PBST	PBS	1x	
	Tween-20	0.1% [v/v]	
	EDTA (pH 8.0)	50mM	

	Glacial acetic acid	0.97M	
NP40 cell lysis buffer	NP40	0.2%	Stored at 4°C. 100x Halt™ Protease Inhibitor Cocktail and/or 100x Halt™ Phosphatase Inhibitor Cocktail were diluted to 1X in the buffer immediately prior to use if required.
	Tris-HCl pH 8.0	50mM	
	NaCl	150mM	
	Protease/phosphatase inhibitors	1x	
	Glycerol	5%	
SDS loading buffer (2x)	SDS	4%	
	Glycerol	20%	
	Beta-Mercaptoethanol	20mM	
	Tris-HCl pH 8.0	100mM	
	Bromophenol blue	0.004%	
Fluorescence Activated Cell Sorting (FACS) sorting buffer	Ca ²⁺ /Mg ²⁺ free PBS	1x	Stored at 4°C
	EDTA	5mM	
	HEPES (pH 7)	25mM	
FACS collection buffer	FCS	20%	Stored at 4°C. Conditioned media was collected from cultured cells during passage.
	Pen-Strep	1%	
	Conditioned media	50%	
	DMEM-F12 media	29%	
Gel loading buffer (1x)	TAE	1x	
	Orange G	0.15% [w/v]	
	Glycerol	60% [v/v]	

6.3.4 Cells lines

All cell lines were sourced from American Type Culture Collection® (ATCC®).

Table 11. List of cell lines used.

Cell line	Origin	Medium	Catalogue Number
hTERT RPE-1	Human telomerase reverse transcriptase (hTERT)-immortalised retinal pigment epithelial cell cultures	DMEM-F12	CRL-4000™

6.3.5 Antibodies and cell stains

6.3.5.1 Primary antibodies

Table 12. List of primary antibodies used.

Ms = mouse, Hu = human, Rb = rabbit, f IF = immunofluorescence, WB = Western Blotting, PFA = paraformaldehyde, MtOH = methanol

Antigen	Raised in	Fixation – PFA	Fixation – MtOH	IF dilution (1/x)	WB dilution (1/x)	Producer	Catalogue number	Clone number
ARL13B	Rb	+	+	8000	5000	Proteintech	17711-1-AP	N/A
TMEM67	Rb	+	+	1000	1000	Proteintech	13975-1-AP	N/a
β-actin	Ms	N/a	N/a	N/a	10,000	Abcam	Ab6276	AC-15

C-myc	Ms	+	+	500	1000	Sigma-Aldrich	M4439	
-------	----	---	---	-----	------	---------------	-------	--

6.3.5.2 Secondary antibodies

Table 13. List of secondary antibodies used.

IF = immunofluorescence, WB = Western Blotting

Target	Raised in	Conjugate	Vendor	Catalogue number	Dilution
Mouse IgG	Goat	Alexa Fluor® 488	Invitrogen	A1102	IF 1:2000
Mouse IgG	Goat	Alexa Fluor® 568	Invitrogen	A11031	IF 1:2000
Mouse IgG	Goat	Alexa Fluor® 647	Invitrogen	A28181	IF 1:2000
Mouse IgG	Goat	Horseradish Peroxidase (HRP)	Dako, Agilent Technologies	P0447	WB 1:10000
Mouse IgG	Donkey	Alexa Fluor® 555	Invitrogen	A31570	IF 1:2000
Rabbit IgG	Goat	Alexa Fluor® 488	Invitrogen	A11034	IF 1:2000
Rabbit IgG	Goat	Alexa Fluor® 568	Invitrogen	A11036	IF 1:2000
Rabbit IgG	Goat	Horseradish Peroxidase (HRP)	Dako, Agilent Technologies	P0448	WB 1:10000
Rabbit IgG	Donkey	Alexa Fluor® 488	Invitrogen	A21206	IF 1:2000
Goat IgG	Donkey	Alexa Fluor® 633	Invitrogen	A21082	IF 1:2000
Goat IgG	Donkey	Alexa Fluor® 350	Invitrogen	A21081	IF 1:2000

6.3.5.3 Cell stains

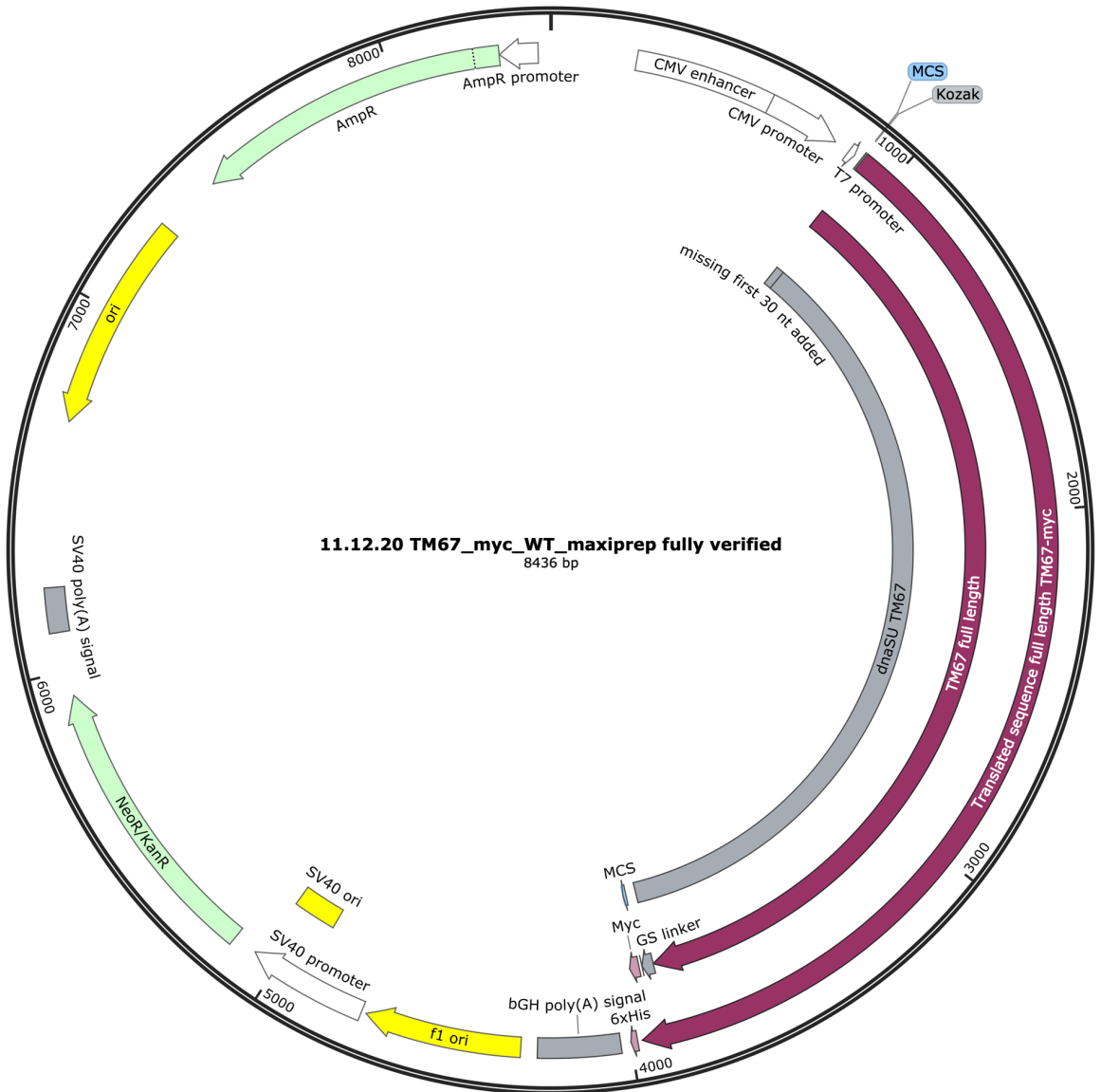
Table 14. List of cell stains used.

Name	Excitation	Sub-Cellular	Vendor	Catalogue #	Dilution
------	------------	--------------	--------	-------------	----------

	/Emission	Localisation			
DAPI	358/461	Nucleus/DNA	Invitrogen™	D1306	IF 1:1000
TOTO®-3 Iodide	642/660	Cytoplasm	Invitrogen™	T3604	IF 1:4000

6.4 Plasmid map: TMEM67_myc_HisA wild type

Created with SnapGene®



7 References

- ABDELHAMED, Z. A., ABDELMOTTALEB, D. I., EL-ASRAG, M. E., NATARAJAN, S., WHEWAY, G., INGLEHEARN, C. F., TOOMES, C. & JOHNSON, C. A. 2019. The ciliary Frizzled-like receptor Tmem67 regulates canonical Wnt/beta-catenin signalling in the developing cerebellum via Hoxb5. *Sci Rep*, 9, 5446.
- ABDELHAMED, Z. A., NATARAJAN, S., WHEWAY, G., INGLEHEARN, C. F., TOOMES, C., JOHNSON, C. A. & JAGGER, D. J. 2015. The Meckel-Gruber syndrome protein TMEM67 controls basal body positioning and epithelial branching morphogenesis in mice via the non-canonical Wnt pathway. *Dis Model Mech*, 8, 527-41.
- ADZHUBEI, I., JORDAN, D. M. & SUNYAEV, S. R. 2013. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet*, Chapter 7, Unit 7.20.
- AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biol*, 21, 30.
- AMBERGER, J. S., BOCCHINI, C. A., SCOTT, A. F. & HAMOSH, A. 2019. OMIM.org: leveraging knowledge across phenotype-gene relationships. *Nucleic Acids Res*, 47, D1038-D1043.
- ANNA, A. & MONIKA, G. 2018. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J Appl Genet*, 59, 253-268.
- ANVARIAN, Z., MYKYTYN, K., MUKHOPADHYAY, S., PEDERSEN, L. B. & CHRISTENSEN, S. T. 2019. Cellular signalling by primary cilia in development, organ function and disease. *Nat Rev Nephrol*, 15, 199-219.
- ANZALONE, A. V., RANDOLPH, P. B., DAVIS, J. R., SOUSA, A. A., KOBLAN, L. W., LEVY, J. M., CHEN, P. J., WILSON, C., NEWBY, G. A., RAGURAM, A. & LIU, D. R. 2019. Search-and-replace genome editing without double-strand breaks or donor DNA. *Nature*, 576, 149-157.
- AUBER, B., BURFEIND, P., HEROLD, S., SCHONER, K., SIMSON, G., RAUSKOLB, R. & REHDER, H. 2007. A disease causing deletion of 29 base pairs in intron 15 in the MKS1 gene is highly associated with the campomelic variant of the Meckel-Gruber syndrome. *Clin Genet*, 72, 454-9.
- AUSTIN, C. P., CUTILLO, C. M., LAU, L. P. L., JONKER, A. H., RATH, A., JULKOWSKA, D., THOMSON, D., TERRY, S. F., DE MONTLEAU, B., ARDIGO, D., HIVERT, V., BOYCOTT, K. M., BAYNAM, G., KAUFMANN, P., TARUSCIO, D., LOCHMULLER, H., SUEMATSU, M., INCERTI, C., DRAGHIA-AKLI, R., NORSTEDT, I., WANG, L., DAWKINS, H. J. S. & INTERNATIONAL RARE DISEASES RESEARCH, C. 2018. Future of Rare Diseases Research 2017-2027: An IRDiRC Perspective. *Clin Transl Sci*, 11, 21-27.
- BACHMANN-GAGESCU, R., DEMPSEY, J. C., PHELPS, I. G., O'ROAK, B. J., KNUTZEN, D. M., RUE, T. C., ISHAK, G. E., ISABELLA, C. R., GORDEN, N., ADKINS, J., BOYLE, E. A., DE LACY, N., O'DAY, D., ALSWAID, A., RAMADEVI, A. R., LINGAPPA, L., LOURENCO, C., MARTORELL, L., GARCIA-CAZORLA, A., OZYUREK, H., HALILOGLU, G., TUYSUZ, B., TOPCU, M., UNIVERSITY OF WASHINGTON CENTER FOR MENDELIAN, G., CHANCE, P., PARISI, M. A., GLASS, I. A., SHENDURE, J. & DOHERTY, D. 2015. Joubert syndrome: a model for untangling recessive disorders with extreme genetic heterogeneity. *J Med Genet*, 52, 514-22.

BACHMANN-GAGESCU, R. & NEUHAUSS, S. C. 2019. The photoreceptor cilium and its diseases. *Curr Opin Genet Dev*, 56, 22-33.

BAE, Y. K. & BARR, M. M. 2008. Sensory roles of neuronal cilia: cilia development, morphogenesis, and function in *C. elegans*. *Front Biosci*, 13, 5959-74.

BANGS, F. & ANDERSON, K. V. 2017. Primary Cilia and Mammalian Hedgehog Signaling. *Cold Spring Harb Perspect Biol*, 9, a028175.

BANGS, F., ANTONIO, N., THONGNUEK, P., WELTEN, M., DAVEY, M. G., BRISCOE, J. & TICKLE, C. 2011. Generation of mice with functional inactivation of *talpid3*, a gene first identified in chicken. *Development*, 138, 3261-72.

BARRANGOU, R. & DOUDNA, J. A. 2016. Applications of CRISPR technologies in research and beyond. *Nat Biotechnol*, 34, 933-941.

BARROSO-GIL, M., OLINGER, E. & SAYER, J. A. 2021. Molecular genetics of renal ciliopathies. *Biochem Soc Trans*, 49, 1205-1220.

BASU, B. & BRUECKNER, M. 2008. Cilia multifunctional organelles at the center of vertebrate left-right asymmetry. *Curr Top Dev Biol*, 85, 151-74.

BELKADI, A., BOLZE, A., ITAN, Y., COBAT, A., VINCENT, Q. B., ANTIPENKO, A., SHANG, L., BOISSON, B., CASANOVA, J. L. & ABEL, L. 2015. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A*, 112, 5473-8.

BELYEU, J. R., CHOWDHURY, M., BROWN, J., PEDERSEN, B. S., CORMIER, M. J., QUINLAN, A. R. & LAYER, R. M. 2021. Samplot: a platform for structural variant visual validation and automated filtering. *Genome Biol*, 22, 161.

BERTOLOTI, A. C., LAYER, R. M., GUNDAPPA, M. K., GALLAGHER, M. D., PEHLIVANOGLU, E., NOME, T., ROBLEDI, D., KENT, M. P., RØSÆG, L. L., HOLEN, M. M., MULUGETA, T. D., ASHTON, T. J., HINDAR, K., SÆGROV, H., FLORØ-LARSEN, B., ERKINARO, J., PRIMMER, C. R., BERNATCHEZ, L., MARTIN, S. A. M., JOHNSTON, I. A., SANDVE, S. R., LIEN, S. & MACQUEEN, D. J. 2020. The structural variation landscape in 492 Atlantic salmon genomes. *Nat Commun*, 11, 5176.

BEST, S., INGLEHEARN, C. F., WATSON, C. M., TOOMES, C., WHEWAY, G. & JOHNSON, C. A. 2022a. Unlocking the potential of the UK 100,000 Genomes Project—lessons learned from analysis of the "Congenital Malformations caused by Ciliopathies" cohort. *Am J Med Genet C Semin Med Genet*, 190, 5-8.

BEST, S., LORD, J., ROCHE, M., WATSON, C. M., POULTER, J. A., BEVERS, R. P. J., STUCKEY, A., SZYMANSKA, K., ELLINGFORD, J. M., CARMICHAEL, J., BRITTAIN, H., TOOMES, C., INGLEHEARN, C., JOHNSON, C. A., WHEWAY, G. & GENOMICS ENGLAND RESEARCH, C. 2022b. Molecular diagnoses in the congenital malformations caused by ciliopathies cohort of the 100,000 Genomes Project. *J Med Genet*, 59, 737-747.

BEST, S., SHOEMARK, A., RUBBO, B., PATEL, M. P., FASSAD, M. R., DIXON, M., ROGERS, A. V., HIRST, R. A., RUTMAN, A., OLLOSSON, S., JACKSON, C. L., GOGGIN, P., THOMAS, S., PENGELLY, R., CULLUP, T., PISSARIDOU, E., HAYWARD, J., ONOUFRIADIS, A., O'CALLAGHAN, C., LOEBINGER, M. R., WILSON, R., CHUNG, E. M., KENIA, P., DOUGHTY, V. L., CARVALHO, J. S., LUCAS, J. S., MITCHISON, H. M. & HOGG, C. 2019. Risk factors for situs defects and congenital heart disease in primary ciliary dyskinesia. *Thorax*, 74, 203-205.

BEST, S., YU, J., LORD, J., ROCHE, M., WATSON, C. M., BEVERS, R. P. J.,

STUCKEY, A., MADHUSUDHAN, S., JEWELL, R., SISODIYA, S. M., LIN, S., TURNER, S., ROBINSON, H., LESLIE, J. S., BAPLE, E., GENOMICS ENGLAND RESEARCH, C., TOOMES, C., INGLEHEARN, C., WHEWAY, G. & JOHNSON, C. A. 2022c. Uncovering the burden of hidden ciliopathies in the 100 000 Genomes Project: a reverse phenotyping approach. *J Med Genet*.

BLAKES, A. J. M., WAI, H. A., DAVIES, I., MOLEDINA, H. E., RUIZ, A., THOMAS, T., BUNYAN, D., THOMAS, N. S., BURREN, C. P., GREENHALGH, L., LEES, M., PICHINI, A., SMITHSON, S. F., TAYLOR TAVARES, A. L., O'DONOVAN, P., DOUGLAS, A. G. L., GENOMICS ENGLAND RESEARCH CONSORTIUM, S., DISEASE WORKING, G., WHIFFIN, N., BARALLE, D. & LORD, J. 2022. A systematic analysis of splicing variants identifies new diagnoses in the 100,000 Genomes Project. *Genome Med*, 14, 79.

BOON, M., WALLMEIER, J., MA, L., LOGES, N. T., JASPERS, M., OLBRICH, H., DOUGHERTY, G. W., RAIDT, J., WERNER, C., AMIRAV, I., HEVRONI, A., ABITBUL, R., AVITAL, A., SOFERMAN, R., WESSELS, M., O'CALLAGHAN, C., CHUNG, E. M., RUTMAN, A., HIRST, R. A., MOYA, E., MITCHISON, H. M., VAN DAELE, S., DE BOECK, K., JORISSEN, M., KINTNER, C., CUPPENS, H. & OMRAN, H. 2014. MCIDAS mutations result in a mucociliary clearance disorder with reduced generation of multiple motile cilia. *Nat Commun*, 5, 4418.

BOWER, R., TRITSCHLER, D., VANDERWAAL, K., PERRONE, C. A., MUELLER, J., FOX, L., SALE, W. S. & PORTER, M. E. 2013. The N-DRC forms a conserved biochemical complex that maintains outer doublet alignment and limits microtubule sliding in motile axonemes. *Mol Biol Cell*, 24, 1134-52.

BRANCATI, F., CAMEROTA, L., COLAO, E., VEGA-WARNER, V., ZHAO, X., ZHANG, R., BOTTILLO, I., CASTORI, M., CAGLIOTI, A., SANGIUOLO, F., NOVELLI, G., PERROTTI, N., OTTO, E. A. & UNDIAGNOSED DISEASE NETWORK, I. 2018. Biallelic variants in the ciliary gene TMEM67 cause RHYSN syndrome. *Eur J Hum Genet*, 26, 1266-1271.

BRANCATI, F., IANNICELLI, M., TRAVAGLINI, L., MAZZOTTA, A., BERTINI, E., BOLTSHAUSER, E., D'ARRIGO, S., EMMA, F., FAZZI, E., GALLIZZI, R., GENTILE, M., LONCAREVIC, D., MEJASKI-BOSNJAK, V., PANTALEONI, C., RIGOLI, L., SALPIETRO, C. D., SIGNORINI, S., STRINGINI, G. R., VERLOES, A., ZABLOKA, D., DALLAPICCOLA, B., GLEESON, J. G., VALENTE, E. M. & INTERNATIONAL, J. S. G. 2009. MKS3/TMEM67 mutations are a major cause of COACH Syndrome, a Joubert Syndrome related disorder with liver involvement. *Hum Mutat*, 30, E432-42.

BRENNER, S. 1974. The genetics of *Caenorhabditis elegans*. *Genetics*, 77, 71-94.

BRISCOE, J. & THEROND, P. P. 2013. The mechanisms of Hedgehog signalling and its roles in development and disease. *Nat Rev Mol Cell Biol*, 14, 416-29.

BROWN, M. A., WIGLEY, C., WALKER, S., LANCASTER, D., RENDON, A. & SCOTT, R. 2022. Re: Best et al., 'Unlocking the potential of the UK 100,000 Genomes Project - Lessons learned from analysis of the "Congenital malformations caused by ciliopathies" cohort'. *Am J Med Genet A*, 188, 3376-3377.

BUJAKOWSKA, K. M., LIU, Q. & PIERCE, E. A. 2017. Photoreceptor Cilia and Retinal Ciliopathies. *Cold Spring Harb Perspect Biol*, 9.

BUTLER, M. T. & WALLINGFORD, J. B. 2017. Planar cell polarity in development and disease. *Nat Rev Mol Cell Biol*, 18, 375-388.

CARBALLO, G. B., HONORATO, J. R., DE LOPES, G. P. F. & SPOHR, T. 2018. A highlight on Sonic hedgehog pathway. *Cell Commun Signal*, 16, 11.

CARSS, K. J., ARNO, G., ERWOOD, M., STEPHENS, J., SANCHIS-JUAN, A., HULL, S., MEGY, K., GROZEVA, D., DEWHURST, E., MALKA, S., PLAGNOL, V., PENKETT, C., STIRRUPS, K., RIZZO, R., WRIGHT, G., JOSIFOVA, D., BITNER-GLINDZICZ, M., SCOTT, R. H., CLEMENT, E., ALLEN, L., ARMSTRONG, R., BRADY, A. F., CARMICHAEL, J., CHITRE, M., HENDERSON, R. H. H., HURST, J., MACLAREN, R. E., MURPHY, E., PATERSON, J., ROSSER, E., THOMPSON, D. A., WAKELING, E., OUWEHAND, W. H., MICHAELIDES, M., MOORE, A. T., CONSORTIUM, N. I.-B. R. D., WEBSTER, A. R. & RAYMOND, F. L. 2017. Comprehensive Rare Variant Analysis via Whole-Genome Sequencing to Determine the Molecular Pathology of Inherited Retinal Disease. *Am J Hum Genet*, 100, 75-90.

CHELLY, J., CONCORDET, J. P., KAPLAN, J. C. & KAHN, A. 1989. Illegitimate transcription: transcription of any gene in any cell type. *Proceedings of the National Academy of Sciences*, 86, 2617-2621.

CHEN, X., SCHULZ-TRIEGLAFF, O., SHAW, R., BARNES, B., SCHLESINGER, F., KALLBERG, M., COX, A. J., KRUGLYAK, S. & SAUNDERS, C. T. 2016. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, 32, 1220-2.

CHICHAGOVA, V., DORGAU, B., FELEMBAN, M., GEORGIU, M., ARMSTRONG, L. & LAKO, M. 2019. Differentiation of Retinal Organoids from Human Pluripotent Stem Cells. *Curr Protoc Stem Cell Biol*, 50, e95.

CHOI, P. S. & MEYERSON, M. 2014. Targeted genomic rearrangements using CRISPR/Cas technology. *Nat Commun*, 5, 3728.

CLARK, M. M., STARK, Z., FARNAES, L., TAN, T. Y., WHITE, S. M., DIMMOCK, D. & KINGSMORE, S. F. 2018. Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. *NPJ Genom Med*, 3, 16.

CONSUGAR, M. B., KUBLY, V. J., LAGER, D. J., HOMMERDING, C. J., WONG, W. C., BAKKER, E., GATTONE, V. H., 2ND, TORRES, V. E., BREUNING, M. H. & HARRIS, P. C. 2007. Molecular diagnostics of Meckel-Gruber syndrome highlights phenotypic differences between MKS1 and MKS3. *Hum Genet*, 121, 591-9.

COOK, S. A., COLLIN, G. B., BRONSON, R. T., NAGGERT, J. K., LIU, D. P., AKESON, E. C. & DAVISSON, M. T. 2009. A mouse model for Meckel syndrome type 3. *J Am Soc Nephrol*, 20, 753-64.

COOPER, T. A., WAN, L. & DREYFUSS, G. 2009. RNA and disease. *Cell*, 136, 777-93.

COPPIETERS, F., CASTEELS, I., MEIRE, F., DE JAEGERE, S., HOOGHE, S., VAN REGEMORTER, N., VAN ESCH, H., MATULEVICIENE, A., NUNES, L., MEERSSCHAUT, V., WALRAEDT, S., STANDAERT, L., COUCKE, P., HOEBEN, H., KROES, H. Y., VANDE WALLE, J., DE RAVEL, T., LEROY, B. P. & DE BAERE, E. 2010a. Genetic screening of LCA in Belgium: predominance of CEP290 and identification of potential modifier alleles in AH11 of CEP290-related phenotypes. *Hum Mutat*, 31, E1709-66.

COPPIETERS, F., LEFEVER, S., LEROY, B. P. & DE BAERE, E. 2010b. CEP290, a gene with many faces: mutation overview and presentation of CEP290base. *Hum Mutat*, 31, 1097-108.

COX, K. F., KERR, N. C., KEDROV, M., NISHIMURA, D., JENNINGS, B. J., STONE, E. M., SHEFFIELD, V. C. & IANNACCONE, A. 2012. Phenotypic expression of Bardet-Biedl

syndrome in patients homozygous for the common M390R mutation in the BBS1 gene. *Vision Res*, 75, 77-87.

CUI, Y., XU, J., CHENG, M., LIAO, X. & PENG, S. 2018. Review of CRISPR/Cas9 sgRNA Design Tools. *Interdiscip Sci*, 10, 455-465.

CUNNINGHAM, F., ALLEN, J. E., ALLEN, J., ALVAREZ-JARRETA, J., AMODE, M. R., ARMEAN, I. M., AUSTINE-ORIMOLOYE, O., AZOV, A. G., BARNES, I., BENNETT, R., BERRY, A., BHAI, J., BIGNELL, A., BILLIS, K., BODDU, S., BROOKS, L., CHARKHCHI, M., CUMMINS, C., DA RIN FIORETTO, L., DAVIDSON, C., DODIYA, K., DONALDSON, S., EL HOUDAIGUI, B., EL NABOULSI, T., FATIMA, R., GIRON, C. G., GENEZ, T., MARTINEZ, J. G., GUIJARRO-CLARKE, C., GYMER, A., HARDY, M., HOLLIS, Z., HOURLIER, T., HUNT, T., JUETTEMANN, T., KAIKALA, V., KAY, M., LAVIDAS, I., LE, T., LEMOS, D., MARUGÁN, J. C., MOHANAN, S., MUSHTAQ, A., NAVEN, M., OGEH, D. N., PARKER, A., PARTON, A., PERRY, M., PILIŽOTA, I., PROSOVETSKAIA, I., SAKTHIVEL, M. P., SALAM, A. I. A., SCHMITT, B. M., SCHUILENBURG, H., SHEPPARD, D., PÉREZ-SILVA, J. G., STARK, W., STEED, E., SUTINEN, K., SUKUMARAN, R., SUMATHIPALA, D., SUNER, M. M., SZPAK, M., THORMANN, A., TRICOMI, F. F., URBINA-GÓMEZ, D., VEIDENBERG, A., WALSH, T. A., WALTS, B., WILLHOFT, N., WINTERBOTTOM, A., WASS, E., CHAKIACHVILI, M., FLINT, B., FRANKISH, A., GIORGETTI, S., HAGGERTY, L., HUNT, S. E., GR, I. I., LOVELAND, J. E., MARTIN, F. J., MOORE, B., MUDGE, J. M., MUFFATO, M., PERRY, E., RUFFIER, M., TATE, J., THYBERT, D., TREVANION, S. J., DYER, S., HARRISON, P. W., HOWE, K. L., YATES, A. D., ZERBINO, D. R. & FLICEK, P. 2022. Ensembl 2022. *Nucleic Acids Res*, 50, D988-d995.

D'ASTOLFO, D. S., PAGLIERO, R. J., PRAS, A., KARTHAUS, W. R., CLEVERS, H., PRASAD, V., LEBBINK, R. J., REHMANN, H. & GEIJSSEN, N. 2015. Efficient intracellular delivery of native proteins. *Cell*, 161, 674-690.

DAIGER, S. 2022. RetNet: Retinal Information Network [Online]. The University of Texas Health Science Center, Houston, Texas. Available: <http://www.sph.uth.tmc.edu/RetNet/> [Accessed 26/10/2022].

DAWE, H. R., ADAMS, M., WHEWAY, G., SZYMANSKA, K., LOGAN, C. V., NOEGEL, A. A., GULL, K. & JOHNSON, C. A. 2009. Nesprin-2 interacts with meckelin and mediates ciliogenesis via remodelling of the actin cytoskeleton. *J Cell Sci*, 122, 2716-26.

DAWE, H. R., SMITH, U. M., CULLINANE, A. R., GERRELLI, D., COX, P., BADANO, J. L., BLAIR-REID, S., SRIRAM, N., KATSANIS, N., ATTIE-BITACH, T., AFFORD, S. C., COPP, A. J., KELLY, D. A., GULL, K. & JOHNSON, C. A. 2007. The Meckel-Gruber Syndrome proteins MKS1 and meckelin interact and are required for primary cilium formation. *Hum Mol Genet*, 16, 173-86.

DECAEN, P. G., DELLING, M., VIEN, T. N. & CLAPHAM, D. E. 2013. Direct recording and molecular identification of the calcium channel of primary cilia. *Nature*, 504, 315-8.

DELLING, M., INDZHYKULIAN, A. A., LIU, X., LI, Y., XIE, T., COREY, D. P. & CLAPHAM, D. E. 2016. Primary cilia are not calcium-responsive mechanosensors. *Nature*, 531, 656- 60.

DELVALLEE, C., NICAISE, S., ANTIN, M., LEUVREY, A. S., NOURISSON, E., LEITCH, C. C., KELLARIS, G., STOETZEL, C., GEOFFROY, V., SCHEIDECKER, S., KEREN, B., DEPIENNE, C., KLAR, J., DAHL, N., DELEUZE, J. F., GENIN, E., REDON, R., DEMURGER, F., DEVRIENDT, K., MATHIEU-DRAMARD, M., POITOU-BERNERT, C., ODENT, S., KATSANIS, N., MANDEL, J. L., DAVIS, E. E., DOLLFUS, H. & MULLER, J. 2021. A BBS1 SVA F retrotransposon insertion is a frequent cause of Bardet-Biedl

syndrome. *Clin Genet*, 99, 318-324.

DEN HOLLANDER, A. I., KOENEKOOP, R. K., YZER, S., LOPEZ, I., ARENDS, M. L., VOESENEK, K. E., ZONNEVELD, M. N., STROM, T. M., MEITINGER, T., BRUNNER,

H. G., HOYNG, C. B., VAN DEN BORN, L. I., ROHRSCHEIDER, K. & CREMERS, F. P. 2006. Mutations in the CEP290 (NPHP6) gene are a frequent cause of Leber congenital amaurosis. *Am J Hum Genet*, 79, 556-61.

DEPARTMENT OF HEALTH AND SOCIAL CARE. 2019. The UK Strategy for Rare Diseases. 2019 update to the Implementation Plan for England. [Online]. Department of Health and Social Care. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/781472/2019-update-to-the-rare-diseases-implementation-plan-for-england.pdf [Accessed 26/10/2022].

DESMET, F. O., HAMROUN, D., LALANDE, M., COLLOD-BEROUD, G., CLAUSTRES, M. & BEROUD, C. 2009. Human Splicing Finder: an online bioinformatics tool to predict splicing signals. *Nucleic Acids Res*, 37, e67.

DESSAUD, E., MCMAHON, A. P. & BRISCOE, J. 2008. Pattern formation in the vertebrate neural tube: a sonic hedgehog morphogen-regulated transcriptional network. *Development*, 135, 2489-503.

DOBELL, C. 1932. Antony van Leeuwenhoek and his "Little Animals." Being Some Account of the Father of Protozoology and Bacteriology and his Multifarious Discoveries in these Disciplines. Collected, Translated, and Edited, from his Printed Works, Unpublished Manuscripts, and Contemporary Records. Published on the 300th Anniversary of his Birth.

DOENCH, J. G., HARTENIAN, E., GRAHAM, D. B., TOTHOVA, Z., HEGDE, M., SMITH, I., SULLENDER, M., EBERT, B. L., XAVIER, R. J. & ROOT, D. E. 2014. Rational design of highly active sgRNAs for CRISPR-Cas9-mediated gene inactivation. *Nat Biotechnol*, 32, 1262-7.

DOHERTY, D., PARISI, M. A., FINN, L. S., GUNAY-AYGUN, M., AL-MATEEN, M., BATES, D., CLERICUZIO, C., DEMIR, H., DORSCHNER, M., VAN ESSEN, A. J., GAHL, W. A., GENTILE, M., GORDEN, N. T., HIKIDA, A., KNUTZEN, D., OZYUREK, H., PHELPS, I., ROSENTHAL, P., VERLOES, A., WEIGAND, H., CHANCE, P. F., DOBYNS, W. B. & GLASS, I. A. 2010. Mutations in 3 genes (MKS3, CC2D2A and RPGRIP1L) cause COACH syndrome (Joubert syndrome with congenital hepatic fibrosis). *J Med Genet*, 47, 8-21.

DOW, L. E. 2015. Modeling Disease In Vivo With CRISPR/Cas9. *Trends Mol Med*, 21, 609- 621.

DRIVAS, T. G., HOLZBAUR, E. L. & BENNETT, J. 2013. Disruption of CEP290 microtubule/membrane-binding domains causes retinal degeneration. *J Clin Invest*, 123, 4525-39.

DUAN, J., LU, G., XIE, Z., LOU, M., LUO, J., GUO, L. & ZHANG, Y. 2014. Genome-wide identification of CRISPR/Cas9 off-targets in human genome. *Cell Res*, 24, 1009-12.

DUMMER, A., POELMA, C., DERUITER, M. C., GOUMANS, M. J. & HIERCK, B. P. 2016. Measuring the primary cilium length: improved method for unbiased high-throughput analysis. *Cilia*, 5, 7.

DUNN, K. C., MARMORSTEIN, A. D., BONILHA, V. L., RODRIGUEZ-BOULAN, E., GIORDANO, F. & HJELMELAND, L. M. 1998. Use of the ARPE-19 cell line as a model of RPE polarity: basolateral secretion of FGF5. *Invest Ophthalmol Vis Sci*, 39, 2744-9.

ELLARD, S., BAPLE, E., OWENS, M., ECCLES, D., TURNBULL, C., ABBS, S., SCOTT, R., DEANS, Z., LESTER, T., CAMPBELL, J., NEWMAN, W. & MCMULLAN, D. 2018. ACGS Best Practice Guidelines for Variant Classification 2018 [Online]. Association for Clinical Genomic Science website: Association for Clinical Genomic Science. Available: <https://www.acgs.uk.com/quality/best-practice-guidelines/> [Accessed 26/10/2022].

ELLINGFORD, J. M., AHN, J. W., BAGNALL, R. D., BARALLE, D., BARTON, S., CAMPBELL, C., DOWNES, K., ELLARD, S., DUFF-FARRIER, C., FITZPATRICK, D. R., GREALLY, J. M., INGLES, J., KRISHNAN, N., LORD, J., MARTIN, H. C., NEWMAN, W. G., O'DONNELL-LURIA, A., RAMSDEN, S. C., REHM, H. L., RICHARDSON, E., SINGER-BERK, M., TAYLOR, J. C., WILLIAMS, M., WOOD, J. C., WRIGHT, C. F., HARRISON, S. M. & WHIFFIN, N. 2022. Recommendations for clinical interpretation of variants found in non-coding regions of the genome. *Genome Med*, 14, 73.

ELLINGFORD, J. M., BARTON, S., BHASKAR, S., O'SULLIVAN, J., WILLIAMS, S. G., LAMB, J. A., PANDA, B., SERGOUNIOTIS, P. I., GILLESPIE, R. L., DAIGER, S. P., HALL, G., GALE, T., LLOYD, I. C., BISHOP, P. N., RAMSDEN, S. C. & BLACK, G. C. M. 2016. Molecular findings from 537 individuals with inherited retinal disease. *J Med Genet*, 53, 761-767.

ELLINGFORD, J. M., HORN, B., CAMPBELL, C., ARNO, G., BARTON, S., TATE, C., BHASKAR, S., SERGOUNIOTIS, P. I., TAYLOR, R. L., CARSS, K. J., RAYMOND, L. F. L., MICHAELIDES, M., RAMSDEN, S. C., WEBSTER, A. R. & BLACK, G. C. M. 2018. Assessment of the incorporation of CNV surveillance into gene panel next-generation sequencing testing for inherited retinal diseases. *J Med Genet*, 55, 114- 121.

EURORDIS. 2005. Rare Diseases: Understanding this Public Health Priority. . Available: https://www.eurordis.org/IMG/pdf/princeps_document-EN.pdf.

FARAG, T. I. & TEEBI, A. S. 1989. High incidence of Bardet Biedl syndrome among the Bedouin. *Clin Genet*, 36, 463-4.

FELDHAUS, B., WEISSCHUH, N., NASSER, F., DEN HOLLANDER, A. I., CREMERS, F. P. M., ZRENNER, E., KOHL, S. & ZOBOR, D. 2020. CEP290 Mutation Spectrum and Delineation of the Associated Phenotype in a Large German Cohort: A Monocentric Study. *Am J Ophthalmol*, 211, 142-150.

FELDMAN, G. L. 2016. 2016 ACMG Annual Meeting presidential address: the practice of medical genetics: myths and realities. *Genet Med*, 18, 957-9.

FERREIRA, C. R. 2019. The burden of rare diseases. *Am J Med Genet A*, 179, 885-892.

FIRE, A., XU, S., MONTGOMERY, M. K., KOSTAS, S. A., DRIVER, S. E. & MELLO, C. C. 1998. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391, 806-11.

FORSYTHE, E. & BEALES, P. L. 2013. Bardet-Biedl syndrome. *Eur J Hum Genet*, 21, 8-13.

FRANCO, B. & THAUVIN-ROBINET, C. 2016. Update on oral-facial-digital syndromes (OFDS). *Cilia*, 5, 12.

FRENCH, C. E., DOLLING, H., MÉGY, K., SANCHIS-JUAN, A., KUMAR, A., DELON, I., WAKELING, M., MALLIN, L., AGRAWAL, S., AUSTIN, T., WALSTON, F., PARK, S. M.,

PARKER, A., PIYASENA, C., BRADBURY, K., ELLARD, S., ROWITCH, D. H. & RAYMOND, F. L. 2022. Refinements and considerations for trio whole-genome sequence analysis when investigating Mendelian diseases presenting in early childhood. *HGG Adv*, 3, 100113.

FU, Y., FODEN, J. A., KHAYTER, C., MAEDER, M. L., REYON, D., JOUNG, J. K. & SANDER, J. D. 2013. High-frequency off-target mutagenesis induced by CRISPR-Cas nucleases in human cells. *Nat Biotechnol*, 31, 822-6.

GALVIN, O., CHI, G., BRADY, L., HIPPERT, C., DEL VALLE RUBIDO, M., DALY, A. & MICHAELIDES, M. 2020. The Impact of Inherited Retinal Diseases in the Republic of Ireland (ROI) and the United Kingdom (UK) from a Cost-of-Illness Perspective. *Clin Ophthalmol*, 14, 707-719.

GANNER, A. & NEUMANN-HAEFELIN, E. 2017. Genetic kidney diseases: *Caenorhabditis elegans* as model system. *Cell Tissue Res*, 369, 105-118.

GARCIA-GONZALO, F. R., CORBIT, K. C., SIREROL-PIQUER, M. S., RAMASWAMI, G., OTTO, E. A., NORIEGA, T. R., SEOL, A. D., ROBINSON, J. F., BENNETT, C. L., JOSIFOVA, D. J., GARCIA-VERDUGO, J. M., KATSANIS, N., HILDEBRANDT, F. & REITER, J. F. 2011. A transition zone complex regulates mammalian ciliogenesis and ciliary membrane composition. *Nat Genet*, 43, 776-84.

GARCIA-GONZALO, F. R. & REITER, J. F. 2012. Scoring a backstage pass: mechanisms of ciliogenesis and ciliary access. *J Cell Biol*, 197, 697-709.

GATTONE, V. H., 2ND, TOURKOW, B. A., TRAMBAUGH, C. M., YU, A. C., WHELAN, S., PHILLIPS, C. L., HARRIS, P. C. & PETERSON, R. G. 2004. Development of multiorgan pathology in the wpk rat model of polycystic kidney disease. *Anat Rec A Discov Mol Cell Evol Biol*, 277, 384-95.

GAUDELLI, N. M., KOMOR, A. C., REES, H. A., PACKER, M. S., BADRAN, A. H., BRYSON, D. I. & LIU, D. R. 2017. Programmable base editing of A*T to G*C in genomic DNA without DNA cleavage. *Nature*, 551, 464-471.

GENOMES PROJECT, C., AUTON, A., BROOKS, L. D., DURBIN, R. M., GARRISON, E. P., KANG, H. M., KORBEL, J. O., MARCHINI, J. L., MCCARTHY, S., MCVEAN, G. A. & ABECASIS, G. R. 2015. A global reference for human genetic variation. *Nature*, 526, 68-74.

GENOMICS ENGLAND. 2021. Newborn Genomes Programme Vision. Available: https://files.genomicsengland.co.uk/documents/Newborns-Vision-Final_SEP_2021-11-02-122418_jjne.pdf [Accessed 26/10/2022].

GIBBONS, I. R. 1961. The relationship between the fine structure and direction of beat in gill cilia of a lamellibranch mollusc. *J Biophys Biochem Cytol*, 11, 179-205.

GILBERT, L. A., HORLBECK, M. A., ADAMSON, B., VILLALTA, J. E., CHEN, Y., WHITEHEAD, E. H., GUIMARAES, C., PANNING, B., PLOEGH, H. L., BASSIK, M. C., QI, L. S., KAMPMANN, M. & WEISSMAN, J. S. 2014. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. *Cell*, 159, 647-61.

GILULA, N. B. & SATIR, P. 1972. The ciliary necklace. A ciliary membrane specialization. *J Cell Biol*, 53, 494-509.

GLOBAL GENES. 2021. RARE disease facts [Online]. Global Genes. Available: <https://globalgenes.org/rare-disease-facts/> [Accessed 26/10/2022].

GOMEZ-ORTE, E., SAENZ-NARCISO, B., MORENO, S. & CABELLO, J. 2013. Multiple

functions of the noncanonical Wnt pathway. *Trends Genet*, 29, 545-53.

GOODWIN, S., MCPHERSON, J. D. & MCCOMBIE, W. R. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet*, 17, 333-51.

GOV.UK. 2019. Code on genetic testing and insurance: the government's annual report 2019. Available: <https://www.gov.uk/government/publications/code-on-genetic-testing-and-insurance-annual-report-2019/code-on-genetic-testing-and-insurance-the-governments-annual-report-2019> [Accessed 20/10/2022].

GOV.UK 2020. Genome UK: the future of healthcare. In: DEPARTMENT FOR BUSINESS, E., & INDUSTRIAL STRATEGY, CARE, D. H. S. & SCIENCES, O. F. L. (eds.). Gov.uk.

GREEN, R. C., BERG, J. S., BERRY, G. T., BIESECKER, L. G., DIMMOCK, D. P., EVANS, J. P., GRODY, W. W., HEGDE, M. R., KALIA, S., KORF, B. R., KRANTZ, I., MCGUIRE, A. L., MILLER, D. T., MURRAY, M. F., NUSSBAUM, R. L., PLON, S. E., REHM, H. L. & JACOB, H. J. 2012. Exploring concordance and discordance for return of incidental findings from clinical sequencing. *Genet Med*, 14, 405-10.

GREENWALD, I. 2016. WormBook: Worm Biology for the 21st Century. *Genetics*, 202, 883-4. GURRIERI, F., FRANCO, B., TORIELLO, H. & NERI, G. 2007. Oral-facial-digital syndromes: review and diagnostic guidelines. *Am J Med Genet A*, 143A, 3314-23.

HAAPANIEMI, E., BOTLA, S., PERSSON, J., SCHMIERER, B. & TAIPALE, J. 2018. CRISPR-Cas9 genome editing induces a p53-mediated DNA damage response. *Nat Med*, 24, 927-930.

HAN, H. A., PANG, J. K. S. & SOH, B. S. 2020. Mitigating off-target effects in CRISPR/Cas9-mediated in vivo gene editing. *J Mol Med (Berl)*, 98, 615-632.

HAN, P. K. J., UMSTEAD, K. L., BERNHARDT, B. A., GREEN, R. C., JOFFE, S., KOENIG, B., KRANTZ, I., WATERSTON, L. B., BIESECKER, L. G. & BIESECKER, B. B. 2017. A taxonomy of medical uncertainties in clinical genome sequencing. *Genet Med*, 19, 918-925.

HANNA, R. E. & DOENCH, J. G. 2020. Design and analysis of CRISPR-Cas experiments. *Nat Biotechnol*, 38, 813-823.

HANNAN, A. J. 2018. Tandem Repeats and Repeatomes: Delving Deeper into the 'Dark Matter' of Genomes. *EBioMedicine*, 31, 3-4.

HARRIS, T. W., ARNABOLDI, V., CAIN, S., CHAN, J., CHEN, W. J., CHO, J., DAVIS, P., GAO, S., GROVE, C. A., KISHORE, R., LEE, R. Y. N., MULLER, H. M., NAKAMURA, C., NUIN, P., PAULINI, M., RACITI, D., RODGERS, F. H., RUSSELL, M., SCHINDELMAN, G., AUKEN, K. V., WANG, Q., WILLIAMS, G., WRIGHT, A. J., YOOK, K., HOWE, K. L., SCHEDL, T., STEIN, L. & STERNBERG, P. W. 2020. WormBase: a modern Model Organism Information Resource. *Nucleic Acids Res*, 48, D762-d767.

HARTILL, V., SZYMANSKA, K., SHARIF, S. M., WHEWAY, G. & JOHNSON, C. A. 2017. Meckel-Gruber Syndrome: An Update on Diagnosis, Clinical Management, and Research Advances. *Front Pediatr*, 5, 244.

HAYCRAFT, C. J., BANIZS, B., AYDIN-SON, Y., ZHANG, Q., MICHAUD, E. J. & YODER, B.K. 2005. Gli2 and Gli3 localize to cilia and require the intraflagellar transport protein polaris for processing and function. *PLoS Genet*, 1, e53.

HOFFMAN-ANDREWS, L. 2017. The known unknown: the challenges of genetic variants of uncertain significance in clinical practice. *J Law Biosci*, 4, 648-657.

HORANI, A. & FERKOL, T. W. 2021. Understanding Primary Ciliary Dyskinesia and Other Ciliopathies. *J Pediatr*, 230, 15-22 e1.

HSIA, Y. E., BRATU, M. & HERBORDT, A. 1971. Genetics of the Meckel syndrome (dysencephalia splanchnocystica). *Pediatrics*, 48, 237-47.

HU, Y., MANGAL, S., ZHANG, L. & ZHOU, X. 2022. Automated filtering of genome-wide large deletions through an ensemble deep learning framework. *Methods*, 206, 77-86.

HYDER, Z., CALPENA, E., PEI, Y., TOOZE, R. S., BRITAIN, H., TWIGG, S. R. F., CILLIERS, D., MORTON, J. E. V., MCCANN, E., WEBER, A., WILSON, L. C., DOUGLAS, A. G.L., MCGOWAN, R., NEED, A., BOND, A., TAVARES, A. L. T., THOMAS, E. R. A., GENOMICS ENGLAND RESEARCH, C., HILL, S. L., DEANS, Z. C., BOARDMAN- PRETTY, F., CAULFIELD, M., SCOTT, R. H. & WILKIE, A. O. M. 2021. Evaluating the performance of a clinical genome sequencing program for diagnosis of rare genetic disease, seen through the lens of craniosynostosis. *Genet Med*, 23, 2360-2368.

IANNICELLI, M., BRANCATI, F., MOUGOU-ZERELLI, S., MAZZOTTA, A., THOMAS, S., ELKHARTOUFI, N., TRAVAGLINI, L., GOMES, C., ARDISSINO, G. L., BERTINI, E., BOLTSHAUSER, E., CASTORINA, P., D'ARRIGO, S., FISCHETTO, R., LEROY, B., LOGET, P., BONNIERE, M., STARCK, L., TANTAU, J., GENTILIN, B., MAJORE, S., SWISTUN, D., FLORI, E., LALATTA, F., PANTALEONI, C., PENZIEN, J., GRAMMATICO, P., INTERNATIONAL, J. S. G., DALLAPICCOLA, B., GLEESON, J. G., ATTIE-BITACH, T. & VALENTE, E. M. 2010. Novel TMEM67 mutations and genotype-phenotype correlates in meckelin-related ciliopathies. *Hum Mutat*, 31, E1319-31.

INGLIS, P. N., OU, G., LEROUX, M. R. & SCHOLEY, J. M. 2007. The sensory cilia of *Caenorhabditis elegans*. *WormBook*, 1-22.

THE 100,000 GENOMES PROJECT PILOT INVESTIGATORS, G. P. P., SMEDLEY, D., SMITH, K. R., MARTIN, A., THOMAS, E. A., MCDONAGH, E. M., CIPRIANI, V., ELLINGFORD, J. M., ARNO, G., TUCCI, A., VANDROVCOVA, J., CHAN, G., WILLIAMS, H. J., RATNAIKE, T., WEI, W., STIRRUPS, K., IBANEZ, K., MOUTSIANAS, L., WIELSCHER, M., NEED, A., BARNES, M. R., VESTITO, L., BUCHANAN, J., WORDSWORTH, S., ASHFORD, S., REHMSTROM, K., LI, E., FULLER, G., TWISS, P., SPASIC-BOSKOVIC, O., HALSALL, S., FLOTO, R. A., POOLE, K., WAGNER, A., MEHTA, S. G., GURNELL, M., BURROWS, N., JAMES, R., PENKETT, C., DEWHURST, E., GRAF, S., MAPETA, R., KASANICKI, M., HAWORTH, A., SAVAGE, H., BABCOCK, M., REESE, M. G., BALE, M., BAPLE, E., BOUSTRED, C., BRITAIN, H., DE BURCA, A., BLEDA, M., DEVEREAU, A., HALAI, D., HARALDSDOTTIR, E., HYDER, Z., KASPERAVICIUTE, D., PATCH, C., POLYCHRONOPOULOS, D., MATCHAN, A., SULTANA, R., RYTEN, M., TAVARES, A. L. T., TREGIDGO, C., TURNBULL, C., WELLAND, M., WOOD, S., SNOW, C., WILLIAMS, E., LEIGH, S., FOULGER, R. E., DAUGHERTY, L. C., NIBLOCK, O., LEONG, I. U. S., WRIGHT, C. F., DAVIES, J., CRICHTON, C., WELCH, J., WOODS, K., ABULHOUL, L., AURORA, P., BOCKENHAUER, D., BROOMFIELD, A., CLEARY, M. A., LAM, T., DATTANI, M., FOOTITT, E., GANESAN, V., GRUNEWALD, S., COMPEYROT-LACASSAGNE, S., MUNTONI, F., PILKINGTON, C., QUINLIVAN, R., THAPAR, N., WALLIS, C., WEDDERBURN, L. R., WORTH, A., BUESER, T., COMPTON, C., et al. 2021. 100,000 Genomes Pilot on Rare-Disease Diagnosis in Health Care - Preliminary Report. *N Engl J Med*, 385, 1868-1880.

ISHIKAWA, H. & MARSHALL, W. F. 2011. Ciliogenesis: building the cell's antenna. *Nat Rev Mol Cell Biol*, 12, 222-34.

IVAKHNO, S., ROLLER, E., COLOMBO, C., TEDDER, P. & COX, A. J. 2018. Canvas SPW: calling de novo copy number variants in pedigrees. *Bioinformatics*, 34, 516-518.

JAAFAR MARICAN, N. H., CRUZ-MIGONI, S. B. & BORYCKI, A. G. 2016. Asymmetric Distribution of Primary Cilia Allocates Satellite Cells for Self-Renewal. *Stem Cell Reports*, 6, 798-805.

JACINTO, F. V., LINK, W. & FERREIRA, B. I. 2020. CRISPR/Cas9-mediated genome editing: From basic research to translational medicine. *J Cell Mol Med*, 24, 3766-3778.

JAGANATHAN, K., KYRIAZOPOULOU PANAGIOTOPOULOU, S., MCRAE, J. F., DARBANDI, S. F., KNOWLES, D., LI, Y. I., KOSMICKI, J. A., ARBELAEZ, J., CUI, W., SCHWARTZ, G. B., CHOW, E. D., KANTERAKIS, E., GAO, H., KIA, A., BATZOGLOU, S., SANDERS, S. J. & FARH, K. K. 2019. Predicting Splicing from Primary Sequence with Deep Learning. *Cell*, 176, 535-548 e24.

JESPERGAARD, C., FANG, M., BERTELSEN, M., DANG, X., JENSEN, H., CHEN, Y., BECH, N., DAI, L., ROSENBERG, T., ZHANG, J., MOLLER, L. B., TUMER, Z., BRONDUM-NIELSEN, K. & GRONSKOV, K. 2019. Molecular genetic analysis using targeted NGS analysis of 677 individuals with retinal dystrophy. *Sci Rep*, 9, 1219.

KALETTA, T. & HENGARTNER, M. O. 2006. Finding function in novel targets: *C. elegans* as a model organism. *Nat Rev Drug Discov*, 5, 387-98.

KANAFI, M. M. & TAVALLAEI, M. 2022. Overview of advances in CRISPR/deadCas9 technology and its applications in human diseases. *Gene*, 830, 146518.

KAPLANIS, J., SAMOCHA, K. E., WIEL, L., ZHANG, Z., ARVAI, K. J., EBERHARDT, R. Y., GALLONE, G., LELIEVELD, S. H., MARTIN, H. C., MCRAE, J. F., SHORT, P. J., TORENE, R. I., DE BOER, E., DANECEK, P., GARDNER, E. J., HUANG, N., LORD, J., MARTINCORENA, I., PFUNDT, R., REIJNDERS, M. R. F., YEUNG, A., YNTEMA, H. G., DECIPHERING DEVELOPMENTAL DISORDERS, S., VISSERS, L., JUUSOLA, J., WRIGHT, C. F., BRUNNER, H. G., FIRTH, H. V., FITZPATRICK, D. R., BARRETT, J. C., HURLES, M. E., GILISSEN, C. & RETTERER, K. 2020. Evidence for 28 genetic disorders discovered by combining healthcare and research data. *Nature*, 586, 757-762.

KARCZEWSKI, K. J., FRANCIOLI, L. C., TIAO, G., CUMMINGS, B. B., ALFOLDI, J., WANG, Q., COLLINS, R. L., LARICCHIA, K. M., GANNA, A., BIRNBAUM, D. P., GAUTHIER, L. D., BRAND, H., SOLOMONSON, M., WATTS, N. A., RHODES, D., SINGER-BERK, M., ENGLAND, E. M., SEABY, E. G., KOSMICKI, J. A., WALTERS, R. K., TASHMAN, K., FARJOUN, Y., BANKS, E., POTERBA, T., WANG, A., SEED, C., WHIFFIN, N., CHONG, J. X., SAMOCHA, K. E., PIERCE-HOFFMAN, E., ZAPPALA, Z., O'DONNELL-LURIA, A. H., MINIKEL, E. V., WEISBURD, B., LEK, M., WARE, J. S., VITTAL, C., ARMEAN, I. M., BERGELSON, L., CIBULSKIS, K., CONNOLLY, K. M., COVARRUBIAS, M., DONNELLY, S., FERRIERA, S., GABRIEL, S., GENTRY, J., GUPTA, N., JEANDET, T., KAPLAN, D., LLANWARNE, C., MUNSHI, R., NOVOD, S., PETRILLO, N., ROAZEN, D., RUANO-RUBIO, V., SALTZMAN, A., SCHLEICHER, M., SOTO, J., TIBBETTS, K., TOLONEN, C., WADE, G., TALKOWSKI, M. E., GENOME AGGREGATION DATABASE, C., NEALE, B. M., DALY, M. J. & MACARTHUR, D. G. 2020. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581, 434-443.

KEARNEY, H. M., THORLAND, E. C., BROWN, K. K., QUINTERO-RIVERA, F., SOUTH, S. T. & WORKING GROUP OF THE AMERICAN COLLEGE OF MEDICAL GENETICS LABORATORY QUALITY ASSURANCE, C. 2011. American College of Medical

- Genetics standards and guidelines for interpretation and reporting of postnatal constitutional copy number variants. *Genet Med*, 13, 680-5.
- KIM, H. M. & COLAIACOVO, M. P. 2016. CRISPR-Cas9-Guided Genome Engineering in *C. elegans*. *Curr Protoc Mol Biol*, 115, 31 7 1-31 7 18.
- KIRCHER, M., WITTEN, D. M., JAIN, P., O'ROAK, B. J., COOPER, G. M. & SHENDURE, J. 2014. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*, 46, 310-5.
- KLINK, B. U., GATSOGIANNIS, C., HOFNAGEL, O., WITTINGHOFER, A. & RAUNSER, S. 2020. Structure of the human BBSome core complex. *Elife*, 9, e53910.
- KNOPP, C., RUDNIK-SCHONEBORN, S., EGGERMANN, T., BERGMANN, C., BEGEMANN, M., SCHONER, K., ZERRES, K. & ORTIZ BRUCHLE, N. 2015. Syndromic ciliopathies: From single gene to multi gene analysis by SNP arrays and next generation sequencing. *Mol Cell Probes*, 29, 299-307.
- KOBLAN, L. W., DOMAN, J. L., WILSON, C., LEVY, J. M., TAY, T., NEWBY, G. A., MAIANTI, J. P., RAGURAM, A. & LIU, D. R. 2018. Improving cytidine and adenine base editors by expression optimization and ancestral reconstruction. *Nat Biotechnol*, 36, 843-846.
- KOMOR, A. C., KIM, Y. B., PACKER, M. S., ZURIS, J. A. & LIU, D. R. 2016. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature*, 533, 420-4.
- KUSCU, C., ARSLAN, S., SINGH, R., THORPE, J. & ADLI, M. 2014. Genome-wide analysis reveals characteristics of off-target sites bound by the Cas9 endonuclease. *Nat Biotechnol*, 32, 677-83.
- KUZNETSOVA, A. V., KURINOV, A. M. & ALEKSANDROVA, M. A. 2014. Cell models to study regulation of cell transformation in pathologies of retinal pigment epithelium. *J Ophthalmol*, 2014, 801787.
- LAMOREAUX, K., LEFEBVRE, S., LEVINE, D. S., ERLER, W. & HUME, T. 2022. The Power of Being Counted Available: <https://rare-x.org/case-studies/the-power-of-being-counted/> [Accessed 05/09/2022].
- LANCASTER, M. A., SCHROTH, J. & GLEESON, J. G. 2011. Subcellular spatial regulation of canonical Wnt signalling at the primary cilium. *Nat Cell Biol*, 13, 700-7.
- LANDRUM, M. J., LEE, J. M., BENSON, M., BROWN, G., CHAO, C., CHITIPIRALLA, S., GU, B., HART, J., HOFFMAN, D., HOOVER, J., JANG, W., KATZ, K., OVETSKY, M., RILEY, G., SETHI, A., TULLY, R., VILLAMARIN-SALOMON, R., RUBINSTEIN, W. & MAGLOTT, D. R. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res*, 44, D862-8.
- LANGE, K. I., BEST, S., TSIROPOULOU, S., BERRY, I., JOHNSON, C. A. & BLACQUE, O. E. 2022. Interpreting ciliopathy-associated missense variants of uncertain significance (VUS) in *Caenorhabditis elegans*. *Hum Mol Genet*, 31, 1574-1587.
- LANGE, K. I., TSIROPOULOU, S., KUCHARSKA, K. & BLACQUE, O. E. 2021. Interpreting the pathogenicity of Joubert syndrome missense variants in *Caenorhabditis elegans*. *Dis Model Mech*, 14.
- LANKTREE, M. B., HAGHIGHI, A., GUIARD, E., ILIUTA, I. A., SONG, X., HARRIS, P. C., PATERSON, A. D. & PEI, Y. 2018. Prevalence Estimates of Polycystic Kidney and Liver Disease by Population Sequencing. *J Am Soc Nephrol*, 29, 2593-2600.
- LEGARE, J. M. 2022. Achondroplasia. In: ADAM, M. P., EVERMAN, D. B., MIRZAA, G.

M., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., GRIPP, K. W. & AMEMIYA, A. (eds.) GeneReviews®. Seattle (WA): University of Washington, Seattle
Copyright © 1993-2022, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.

LEITCH, C. C., ZAGHLOUL, N. A., DAVIS, E. E., STOETZEL, C., DIAZ-FONT, A., RIX, S., ALFADHEL, M., LEWIS, R. A., EYALD, W., BANIN, E., DOLLFUS, H., BEALES, P. L., BADANO, J. L. & KATSANIS, N. 2008. Hypomorphic mutations in syndromic encephalocele genes are associated with Bardet-Biedl syndrome. *Nat Genet*, 40, 443-8.

LEROY, B. P., FISCHER, M. D., FLANNERY, J. G., MACLAREN, R. E., DALKARA, D., SCHOLL, H. P. N., CHUNG, D. C., SPERA, C., VIRIATO, D. & BANHAZI, J. 2022. Gene therapy for inherited retinal disease: long-term durability of effect. *Ophthalmic Res*.

LIANG, X., POTTER, J., KUMAR, S., RAVINDER, N. & CHESNUT, J. D. 2017. Enhanced CRISPR/Cas9-mediated precise genome editing by improved design and delivery of gRNA, Cas9 nuclease, and donor DNA. *J Biotechnol*, 241, 136-146.

LIN, Y., CRADICK, T. J., BROWN, M. T., DESHMUKH, H., RANJAN, P., SARODE, N., WILE, B. M., VERTINO, P. M., STEWART, F. J. & BAO, G. 2014. CRISPR/Cas9 systems have off-target activity with insertions or deletions between target DNA and guide RNA sequences. *Nucleic Acids Res*, 42, 7473-85.

LINDSTRAND, A., EISFELDT, J., PETTERSSON, M., CARVALHO, C. M. B., KVARNUNG, M., GRIGELIONIENE, G., ANDERLID, B. M., BJERIN, O., GUSTAVSSON, P., HAMMARSJO, A., GEORGII-HEMMING, P., IWARSSON, E., JOHANSSON-SOLLER, M., LAGERSTEDT-ROBINSON, K., LIEDEN, A., MAGNUSSON, M., MARTIN, M., MALMGREN, H., NORDENSKJOLD, M., NORLING, A., SAHLIN, E., STRANNEHEIM, H., THAM, E., WINCENT, J., YGBERG, S., WEDELL, A., WIRTA, V., NORDGREN, A., LUNDIN, J. & NILSSON, D. 2019. From cytogenetics to cytogenomics: whole-genome sequencing as a first-line test comprehensively captures the diverse spectrum of disease-causing genetic variation underlying intellectual disability. *Genome Med*, 11, 68.

LIU, Y., HUANG, Y., WANG, G. & WANG, Y. 2021. A deep learning approach for filtering structural variants in short read sequencing data. *Brief Bioinform*, 22.

LORD, J. & BARALLE, D. 2021. Splicing in the Diagnosis of Rare Disease: Advances and Challenges. *Front Genet*, 12, 689892.

LUCAS, J. S., BURGESS, A., MITCHISON, H. M., MOYA, E., WILLIAMSON, M., HOGG, C. & NATIONAL PCD SERVICE, U. K. 2014. Diagnosis and management of primary ciliary dyskinesia. *Arch Dis Child*, 99, 850-6.

MACKEN, W. L., FALABELLA, M., MCKITTRICK, C., PIZZAMIGLIO, C., ELLMERS, R., EGGLETON, K., WOODWARD, C. E., PATEL, Y., LABRUM, R., PHADKE, R., REILLY, M. M., DEVILE, C., SARKOZY, A., FOOTITT, E., DAVISON, J., RAHMAN, S., HOULDEN, H., BUGIARDINI, E., QUINLIVAN, R., HANNA, M. G., VANDROVCOVA, J. & PITCEATHLY, R. D. S. 2022. Specialist multidisciplinary input maximises rare disease diagnoses from whole genome sequencing. *Nat Commun*, 13, 6324.

MAKHNOON, S., SHIRTS, B. H. & BOWEN, D. J. 2019. Patients' perspectives of variants of uncertain significance and strategies for uncertainty management. *J Genet Couns*, 28, 313-325.

MALICKI, J. J. & JOHNSON, C. A. 2017. The Cilium: Cellular Antenna and Central

Processing Unit. *Trends Cell Biol*, 27, 126-140.

MALLAWAARACHCHI, A. C., HORT, Y., COWLEY, M. J., MCCABE, M. J., MINOCHE, A., DINGER, M. E., SHINE, J. & FURLONG, T. J. 2016. Whole-genome sequencing overcomes pseudogene homology to diagnose autosomal dominant polycystic kidney disease. *Eur J Hum Genet*, 24, 1584-1590.

MARIA, M., LAMERS, I. J., SCHMIDTS, M., AJMAL, M., JAFFAR, S., ULLAH, E., MUSTAFA, B., AHMAD, S., NAZMUTDINOVA, K., HOSKINS, B., VAN WIJK, E., KOSTER- KAMPHUIS, L., KHAN, M. I., BEALES, P. L., CREMERS, F. P., ROEPMAN, R., AZAM, M., ARTS, H. H. & QAMAR, R. 2016. Genetic and clinical characterization of Pakistani families with Bardet-Biedl syndrome extends the genetic and phenotypic spectrum. *Sci Rep*, 6, 34764.

MARRAFFINI, L. A. & SONTHEIMER, E. J. 2008. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. *Science*, 322, 1843-5.

MARSHALL, C. R., SCHERER, S. W., ZARIWALA, M. A., LAU, L., PATON, T. A., STOCKLEY, T., JOBLING, R. K., RAY, P. N., KNOWLES, M. R., CONSORTIUM, F. C., HALL, D. A., DELL, S. D. & KIM, R. H. 2015. Whole-Exome Sequencing and Targeted Copy Number Analysis in Primary Ciliary Dyskinesia. *G3 (Bethesda)*, 5, 1775-81.

MARSHALL, J. D., MAFFEI, P., COLLIN, G. B. & NAGGERT, J. K. 2011. Alström syndrome: genetics and clinical overview. *Curr Genomics*, 12, 225-35.

MARTIN, A. R., WILLIAMS, E., FOULGER, R. E., LEIGH, S., DAUGHERTY, L. C., NIBLOCK, O., LEONG, I. U. S., SMITH, K. R., GERASIMENKO, O., HARALDSDOTTIR, E., THOMAS, E., SCOTT, R. H., BAPLE, E., TUCCI, A., BRITAIN, H., DE BURCA, A., IBANEZ, K., KASPERAVICIUTE, D., SMEDLEY, D., CAULFIELD, M., RENDON, A. & MCDONAGH, E. M. 2019. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*, 51, 1560-1565.

MARUYAMA, T., DOUGAN, S. K., TRUTTMANN, M. C., BILATE, A. M., INGRAM, J. R. & PLOEGH, H. L. 2015. Increasing the efficiency of precise genome editing with CRISPR-Cas9 by inhibition of nonhomologous end joining. *Nat Biotechnol*, 33, 538- 42.

MATTICK, J. S., DINGER, M., SCHONROCK, N. & COWLEY, M. 2018. Whole genome sequencing provides better diagnostic yield and future value than whole exome sequencing. *Med J Aust*, 209, 197-199.

MCLAREN, W., GIL, L., HUNT, S. E., RIAT, H. S., RITCHIE, G. R., THORMANN, A., FLICEK, P. & CUNNINGHAM, F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol*, 17, 122.

MIDHA, M. K., WU, M. & CHIU, K. P. 2019. Long-read sequencing in deciphering human genetics to a greater depth. *Hum Genet*, 138, 1201-1215.

MILLS, R. E., WALTER, K., STEWART, C., HANDSAKER, R. E., CHEN, K., ALKAN, C., ABYZOV, A., YOON, S. C., YE, K., CHEETHAM, R. K., CHINWALLA, A., CONRAD, D. F., FU, Y., GRUBERT, F., HAJIRASOULIHA, I., HORMOZDIARI, F., IAKOUCHEVA, L. M., IQBAL, Z., KANG, S., KIDD, J. M., KONKEL, M. K., KORN, J., KHURANA, E., KURAL, D., LAM, H. Y., LENG, J., LI, R., LI, Y., LIN, C. Y., LUO, R., MU, X. J., NEMESH, J., PECKHAM, H. E., RAUSCH, T., SCALLY, A., SHI, X., STROMBERG, M. P., STÜTZ, A. M., URBAN, A. E., WALKER, J. A., WU, J., ZHANG, Y., ZHANG, Z. D., BATZER, M. A., DING, L., MARTH, G. T., MCVEAN, G., SEBAT, J., SNYDER, M., WANG, J., YE, K., EICHLER, E. E., GERSTEIN, M. B., HURLES, M. E., LEE, C., MCCARROLL, S. A. & KORBEL, J. O. 2011. Mapping copy number variation by

population-scale genome sequencing. *Nature*, 470, 59-65.

MITCHISON, H. M. & VALENTE, E. M. 2017. Motile and non-motile cilia in human pathology: from function to phenotypes. *J Pathol*, 241, 294-309.

MOLINARI, E. & SAYER, J. A. 2017. Emerging treatments and personalised medicine for ciliopathies associated with cystic kidney disease. *Expert Opinion on Orphan Drugs*, 5, 785-798.

MOLINARI, E., SRIVASTAVA, S., DEWHURST, R. M. & SAYER, J. A. 2020. Use of patient derived urine renal epithelial cells to confirm pathogenicity of PKHD1 alleles. *BMC Nephrol*, 21, 435.

MOORE, A., ESCUDIER, E., ROGER, G., TAMALET, A., PELOSSE, B., MARLIN, S., CLEMENT, A., GEREMEK, M., DELAISI, B., BRIDOUX, A. M., COSTE, A., WITT, M., DURIEZ, B. & AMSELEM, S. 2006. RPGR is mutated in patients with a complex X linked phenotype combining primary ciliary dyskinesia and retinitis pigmentosa. *J Med Genet*, 43, 326-33.

MURDOCH, J. N. & COPP, A. J. 2010. The relationship between sonic Hedgehog signaling, cilia, and neural tube defects. *Birth Defects Res A Clin Mol Teratol*, 88, 633-52.

NAG, C., GHOSH, M., DAS, K. & GHOSH, T. 2013. Joubert syndrome: the molar tooth sign of the mid-brain. *Ann Med Health Sci Res*, 3, 291-4.

NANGIA, V., JONAS, J. B., KHARE, A. & SINHA, A. 2012. Prevalence of retinitis pigmentosa in India: the Central India Eye and Medical Study. *Acta Ophthalmol*, 90, e649-50.

NAZLAMOVA, L., THOMAS, N. S., CHEUNG, M. K., LEGEBEKE, J., LORD, J., PENGELLY,

R. J., TAPPER, W. J. & WHEWAY, G. 2021. A CRISPR and high-content imaging assay compliant with ACMG/AMP guidelines for clinical variant interpretation in ciliopathies. *Hum Genet*, 140, 593-607.

NGUENGANG WAKAP, S., LAMBERT, D. M., OLRYS, A., RODWELL, C., GUEYDAN, C., LANNEAU, V., MURPHY, D., LE CAM, Y. & RATH, A. 2020. Estimating cumulative point prevalence of rare diseases: analysis of the Orphanet database. *Eur J Hum Genet*, 28, 165-173.

NHS ENGLAND. 2022. National Genomic Test Directory [Online]. NHS England. Available: <https://www.england.nhs.uk/publication/national-genomic-test-directories/> [Accessed 26/10/2022].

NOLL, A. C., MILLER, N. A., SMITH, L. D., YOO, B., FIEDLER, S., COOLEY, L. D., WILLIG, L. K., PETRIKIN, J. E., CAKICI, J., LESKO, J., NEWTON, A., DETHERAGE, K., THIFFAULT, I., SAUNDERS, C. J., FARROW, E. G. & KINGSMORE, S. F. 2016. Clinical detection of deletion structural variants in whole-genome sequences. *NPJ Genom Med*, 1, 16026.

NONAKA, S., TANAKA, Y., OKADA, Y., TAKEDA, S., HARADA, A., KANAI, Y., KIDO, M. & HIROKAWA, N. 1998. Randomization of left-right asymmetry due to loss of nodal cilia generating leftward flow of extraembryonic fluid in mice lacking KIF3B motor protein. *Cell*, 95, 829-37.

NURCHIS, M. C., RICCARDI, M. T., RADIO, F. C., CHILLEMI, G., BERTINI, E. S., TARTAGLIA, M., CICCETTI, A., DALLAPICCOLA, B. & DAMIANI, G. 2022.

Incremental net benefit of whole genome sequencing for newborns and children with suspected genetic disorders: Systematic review and meta-analysis of cost-effectiveness evidence. *Health Policy*, 126, 337-345.

NURK, S., KOREN, S., RHIE, A., RAUTIAINEN, M., BZIKADZE, A. V., MIKHEENKO, A., VOLLGER, M. R., ALTEMOSE, N., URALSKY, L., GERSHMAN, A., AGANEZOV, S., HOYT, S. J., DIEKHANS, M., LOGSDON, G. A., ALONGE, M., ANTONARAKIS, S. E., BORCHERS, M., BOUFFARD, G. G., BROOKS, S. Y., CALDAS, G. V., CHEN, N. C., CHENG, H., CHIN, C. S., CHOW, W., DE LIMA, L. G., DISHUCK, P. C., DURBIN, R., DVORKINA, T., FIDDES, I. T., FORMENTI, G., FULTON, R. S., FUNGTAMMASAN, A., GARRISON, E., GRADY, P. G. S., GRAVES-LINDSAY, T. A., HALL, I. M., HANSEN, N. F., HARTLEY, G. A., HAUKNES, M., HOWE, K., HUNKAPILLER, M. W., JAIN, C., JAIN, M., JARVIS, E. D., KERPEDJIEV, P., KIRSCH, M., KOLMOGOROV, M., KORLACH, J., KREMITZKI, M., LI, H., MADURO, V. V., MARSCHALL, T., MCCARTNEY, A. M., MCDANIEL, J., MILLER, D. E., MULLIKIN, J. C., MYERS, E. W., OLSON, N. D., PATEN, B., PELUSO, P., PEVZNER, P. A., PORUBSKY, D., POTAPOVA, T., ROGAEV, E. I., ROSENFELD, J. A., SALZBERG, S. L., SCHNEIDER, V. A., SEDLAZECK, F. J., SHAFIN, K., SHEW, C. J., SHUMATE, A., SIMS, Y., SMIT, A. F. A., SOTO, D. C., SOVIC, I., STORER, J. M., STREETS, A., SULLIVAN, B. A., THIBAUD-NISSEN, F., TORRANCE, J., WAGNER, J., WALENZ, B. P., WENGER, A., WOOD, J. M. D., XIAO, C., YAN, S. M., YOUNG, A. C., ZARATE, S., SURTI, U., MCCOY, R. C., DENNIS, M. Y., ALEXANDROV, I. A., GERTON, J. L., O'NEILL, R. J., TIMP, W., ZOOK, J. M., SCHATZ, M. C., EICHLER, E. E., MIGA, K. H. & PHILLIPPY, A. M. 2022. The complete sequence of a human genome. *Science*, 376, 44-53.

O'CALLAGHAN, C., CHETCUTI, P. & MOYA, E. 2010. High prevalence of primary ciliary dyskinesia in a British Asian population. *Arch Dis Child*, 95, 51-2.

ORPHANET. 2022. Orphanet: Alström syndrome [Online]. Orphanet. Available: [https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=1328&Disease_Disease_Search_disease/Group=Alström&Disease_Disease_Search_diseaseType=Pat&Disease\(s\)/group%20of%20diseases=Alström_syndrome&title=Alstr%F6m%20syndrome&search=Disease_Search_Simple](https://www.orpha.net/consor/cgi-bin/Disease_Search.php?lng=EN&data_id=1328&Disease_Disease_Search_disease/Group=Alström&Disease_Disease_Search_diseaseType=Pat&Disease(s)/group%20of%20diseases=Alström_syndrome&title=Alstr%F6m%20syndrome&search=Disease_Search_Simple) [Accessed 26/10/2022].

OTTO, E. A., TORY, K., ATTANASIO, M., ZHOU, W., CHAKI, M., PARUCHURI, Y., WISE, E. L., WOLF, M. T., UTSCH, B., BECKER, C., NURNBERG, G., NURNBERG, P., NAYIR, A., SAUNIER, S., ANTIGNAC, C. & HILDEBRANDT, F. 2009. Hypomorphic mutations in meckelin (MKS3/TMEM67) cause nephronophthisis with liver fibrosis (NPHP11). *J Med Genet*, 46, 663-70.

PAFF, T., KOOI, I. E., MOUTAOUAKIL, Y., RIESEBOS, E., SISTERMANS, E. A., DANIELS, H., WEISS, J. M. M., NIESSEN, H., HAARMAN, E. G., PALS, G. & MICHA, D. 2018. Diagnostic yield of a targeted gene panel in primary ciliary dyskinesia patients. *Hum Mutat*, 39, 653-665.

PAGON, R. A. 1988. Retinitis pigmentosa. *Surv Ophthalmol*, 33, 137-77.

PAISEY, R. B., STEEDS, R., BARRETT, T., WILLIAMS, D., GEBERHIWOT, T. & GUNAY-AYGUN, M. 2019. Alström Syndrome. In: ADAM, M. P., ARDINGER, H. H., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., STEPHENS, K. & AMEMIYA, A. (eds.) *GeneReviews*(®). Seattle (WA): University of Washington, Seattle
Copyright © 1993-2020, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.

PALA, R., ALOMARI, N. & NAULI, S. M. 2017. Primary Cilium-Dependent Signaling Mechanisms. *Int J Mol Sci*, 18. PARISI, M. & GLASS, I. 2017. Joubert Syndrome. In: ADAM, M. P., EVERMAN, D. B., MIRZAA, G. M., PAGON, R. A., WALLACE, S. E., BEAN, L. J. H., GRIPP, K. W. & AMEMIYA, A. (eds.) *GeneReviews*(®). Seattle (WA): University of Washington, Seattle. Copyright © 1993-2022, University of Washington, Seattle. GeneReviews is a registered trademark of the University of Washington, Seattle. All rights reserved.

PARLIAMENT.UK. 2018. Genomics and genome editing in the NHS. Available: <https://publications.parliament.uk/pa/cm201719/cmselect/cmsctech/349/34902.htm> [Accessed 26/10/2022].

PLOTNIKOVA, O. V., PUGACHEVA, E. N. & GOLEMIS, E. A. 2009. Primary cilia and the cell cycle. *Methods Cell Biol*, 94, 137-60.

PUTSCHER, E., HECKER, M., FITZNER, B., LORENZ, P. & ZETTL, U. K. 2021. Principles and Practical Considerations for the Analysis of Disease-Associated Alternative Splicing Events Using the Gateway Cloning-Based Minigene Vectors pDESTsplice and pSpliceExpress. *Int J Mol Sci*, 22.

QUINLAN, A. R. & HALL, I. M. 2012. Characterizing complex structural variation in germline and somatic genomes. *Trends Genet*, 28, 43-53.

QUINLAN, R. J., TOBIN, J. L. & BEALES, P. L. 2008. Modeling ciliopathies: Primary cilia in development and disease. *Curr Top Dev Biol*, 84, 249-310.

RAMSBOTTOM, S. A., MOLINARI, E., SRIVASTAVA, S., SILBERMAN, F., HENRY, C., ALKANDERI, S., DEVLIN, L. A., WHITE, K., STEEL, D. H., SAUNIER, S., MILES, C. G. & SAYER, J. A. 2018. Targeted exon skipping of a CEP290 mutation rescues Joubert syndrome phenotypes in vitro and in a murine model. *Proc Natl Acad Sci U S A*, 115, 12489-12494.

RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc*, 8, 2281-2308.

RAUCHMAN, M. I., NIGAM, S. K., DELPIRE, E. & GULLANS, S. R. 1993. An osmotically tolerant inner medullary collecting duct cell line from an SV40 transgenic mouse. *Am J Physiol*, 265, F416-24.

REITER, J. F. & LEROUX, M. R. 2017. Genes and molecular pathways underpinning ciliopathies. *Nat Rev Mol Cell Biol*, 18, 533-547.

RICHARDS, S., AZIZ, N., BALE, S., BICK, D., DAS, S., GASTIER-FOSTER, J., GRODY, W. W., HEGDE, M., LYON, E., SPECTOR, E., VOELKERDING, K., REHM, H. L. & COMMITTEE, A. L. Q. A. 2015. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*, 17, 405-24.

ROBINSON, P. N., KOHLER, S., OELLRICH, A., SANGER MOUSE GENETICS, P., WANG, K., MUNGALL, C. J., LEWIS, S. E., WASHINGTON, N., BAUER, S., SEELOW, D., KRAWITZ, P., GILISSEN, C., HAENDEL, M. & SMEDLEY, D. 2014. Improved exome prioritization of disease genes through cross-species phenotype comparison. *Genome Res*, 24, 340-8.

ROHATGI, R., MILENKOVIC, L. & SCOTT, M. P. 2007. Patched1 regulates hedgehog signaling at the primary cilium. *Science*, 317, 372-6.

ROMANI, M., MICALIZZI, A. & VALENTE, E. M. 2013. Joubert syndrome: congenital cerebellar ataxia with the molar tooth. *Lancet Neurol*, 12, 894-905.

ROSSETTI, S., CONSUGAR, M. B., CHAPMAN, A. B., TORRES, V. E., GUAY-WOODFORD, L. M., GRANTHAM, J. J., BENNETT, W. M., MEYERS, C. M., WALKER, D. L., BAE, K., ZHANG, Q. J., THOMPSON, P. A., MILLER, J. P., HARRIS, P. C. & CONSORTIUM, C. 2007. Comprehensive molecular diagnostics in autosomal dominant polycystic kidney disease. *J Am Soc Nephrol*, 18, 2143-60.

ROWLANDS, C., THOMAS, H. B., LORD, J., WAI, H. A., ARNO, G., BEAMAN, G., SERGOUNIOTIS, P., GOMES-SILVA, B., CAMPBELL, C., GOSSAN, N., HARDCASTLE, C., WEBB, K., O'CALLAGHAN, C., HIRST, R. A., RAMSDEN, S., JONES, E., CLAYTON-SMITH, J., WEBSTER, A. R., AMBROSE, J. C., ARUMUGAM, P., BEVERS, R., BLEDA, M., BOARDMAN-PRETTY, F., BOUSTRED, C. R., BRITAIN, H., CAULFIELD, M. J., CHAN, G. C., FOWLER, T., GIESS, A., HAMBLIN, A., HENDERSON, S., HUBBARD, T. J. P., JACKSON, R., JONES, L. J., KASPERAVICIUTE, D., KAYIKCI, M., KOUSATHANAS, A., LAHNSTEIN, L., LEIGH, S. E. A., LEONG, I. U. S., LOPEZ, F. J., MALEADY-CROWE, F., MCENTAGART, M., MINNECI, F., MOUTSIANAS, L., MUELLER, M., MURUGAESU, N., NEED, A. C., O'DONOVAN, P., ODHAMS, C. A., PATCH, C., PEREZ-GIL, D., PEREIRA, M. B., PULLINGER, J., RAHIM, T., RENDON, A., ROGERS, T., SAVAGE, K., SAWANT, K., SCOTT, R. H., SIDDIQ, A., SIEGHART, A., SMITH, S. C., SOSINSKY, A., STUCKEY, A., TANGUY, M., TAYLOR TAVARES, A. L., THOMAS, E. R. A., THOMPSON, S. R., TUCCI, A., WELLAND, M. J., WILLIAMS, E., WITKOWSA, K., WOOD, S. M., DOUGLAS, A. G. L., O'KEEFE, R. T., NEWMAN, W. G., BARALLE, D., BLACK, G. C. M., ELLINGFORD, J. M. & GENOMICS ENGLAND RESEARCH, C. 2021. Comparison of in silico strategies to prioritize rare genomic variants impacting RNA splicing for the diagnosis of genomic disorders. *Scientific Reports*, 11, 20607.

ROWLANDS, C. F., TAYLOR, A., RICE, G., WHIFFIN, N., HALL, H. N., NEWMAN, W. G., BLACK, G. C. M., O'KEEFE, R. T., HUBBARD, S., DOUGLAS, A. G. L., BARALLE, D., BRIGGS, T. A. & ELLINGFORD, J. M. 2022. MRSD: A quantitative approach for assessing suitability of RNA-seq in the investigation of mis-splicing in Mendelian disease. *Am J Hum Genet*, 109, 210-222.

RUSSELL, S. R., DRACK, A. V., CIDECIYAN, A. V., JACOBSON, S. G., LEROY, B. P., VAN CAUWENBERGH, C., HO, A. C., DUMITRESCU, A. V., HAN, I. C., MARTIN, M., PFEIFER, W. L., SOHN, E. H., WALSHIRE, J., GARAFALO, A. V., KRISHNAN, A. K., POWERS, C. A., SUMAROKA, A., ROMAN, A. J., VANHONSEBROUCK, E., JONES, E., NERINCKX, F., DE ZAEYTIJD, J., COLLIN, R. W. J., HOYNG, C., ADAMSON, P., CHEETHAM, M. E., SCHWARTZ, M. R., DEN HOLLANDER, W., ASMUS, F., PLATENBURG, G., RODMAN, D. & GIRACH, A. 2022. Intravitreal antisense oligonucleotide seprofarsen in Leber congenital amaurosis type 10: a phase 1b/2 trial. *Nat Med*, 28, 1014-1021.

SALLUM, J. M. F., MOTTA, F. L., ARNO, G., PORTO, F. B. O., RESENDE, R. G. & BELFORT, R., JR. 2020. Clinical and molecular findings in a cohort of 152 Brazilian severe early onset inherited retinal dystrophy patients. *Am J Med Genet C Semin Med Genet*, 184, 728-752.

SALONEN, R. & NORIO, R. 1984. The Meckel syndrome in Finland: epidemiologic and genetic aspects. *Am J Med Genet*, 18, 691-8.

SANCHEZ, I. & DYNLACHT, B. D. 2016. Cilium assembly and disassembly. *Nat Cell Biol*, 18, 711-7.

SANDERS, A. A., KENNEDY, J. & BLACQUE, O. E. 2015. Image analysis of *Caenorhabditis elegans* ciliary transition zone structure, ultrastructure, molecular

composition, and function. *Methods Cell Biol*, 127, 323-47.

SATIR, P. 2017. CILIA: before and after. *Cilia*, 6, 1.

SATIR, P. & CHRISTENSEN, S. T. 2007. Overview of structure and function of mammalian cilia. *Annu Rev Physiol*, 69, 377-400.

SAWYER, S. L., HARTLEY, T., DYMENT, D. A., BEAULIEU, C. L., SCHWARTZENTRUBER, J., SMITH, A., BEDFORD, H. M., BERNARD, G., BERNIER, F. P., BRAIS, B., BULMAN, D. E., WARMAN CHARDON, J., CHITAYAT, D., DELADOEY, J., FERNANDEZ, B. A., FROSK, P., GERAGHTY, M. T., GERULL, B., GIBSON, W., GOW, R. M., GRAHAM, G. E., GREEN, J. S., HEON, E., HORVATH, G., INNES, A. M., JABADO, N., KIM, R. H., KOENEKOOP, R. K., KHAN, A., LEHMANN, O. J., MENDOZA-LONDONO, R., MICHAUD, J. L., NIKKEL, S. M., PENNEY, L. S., POLYCHRONAKOS, C., RICHER, J., ROULEAU, G. A., SAMUELS, M. E., SIU, V. M., SUCHOWERSKY, O., TARNOPOLSKY, M. A., YOON, G., ZAHIR, F. R., CONSORTIUM, F. C., CARE4RARE CANADA, C., MAJEWSKI, J. & BOYCOTT, K. M. 2016. Utility of whole-exome sequencing for those near the end of the diagnostic odyssey: time to address gaps in care. *Clin Genet*, 89, 275-84.

SCHUY, J., GROCHOWSKI, C. M., CARVALHO, C. M. B. & LINDSTRAND, A. 2022. Complex genomic rearrangements: an underestimated cause of rare diseases. *Trends Genet*, 38, 1134-1146.

SHAHEEN, R., SZYMANSKA, K., BASU, B., PATEL, N., EWIDA, N., FAQEIH, E., AL HASHEM, A., DERAR, N., ALSHARIF, H., ALDAHMEH, M. A., ALAZAMI, A. M., HASHEM, M., IBRAHIM, N., ABDULWAHAB, F. M., SONBUL, R., ALKURAYA, H., ALNEMER, M., AL TALA, S., AL-HUSAIN, M., MORSY, H., SEIDAHMED, M. Z., MERIKI, N., AL-OWAIN, M., ALSHAHWAN, S., TABARKI, B., SALIH, M. A., CILIOPATHY, W., FAQUIH, T., EL-KALIOBY, M., UEFFING, M., BOLDT, K., LOGAN, C. V., PARRY, D. A., AL TASSAN, N., MONIES, D., MEGARBANE, A., ABOUELHODA, M., HALEES, A., JOHNSON, C. A. & ALKURAYA, F. S. 2016.

Characterizing the morbid genome of ciliopathies. *Genome Biol*, 17, 242.

SHI, X., GARCIA, G., 3RD, VAN DE WEGHE, J. C., MCGORTY, R., PAZOUR, G. J., DOHERTY, D., HUANG, B. & REITER, J. F. 2017. Super-resolution microscopy reveals that disruption of ciliary transition-zone architecture causes Joubert syndrome. *Nat Cell Biol*, 19, 1178-1188.

SHOEMARK, A. & HOGG, C. 2013. Electron tomography of respiratory cilia. *Thorax*, 68, 190-1. SIM, N. L., KUMAR, P., HU, J., HENIKOFF, S., SCHNEIDER, G. & NG, P. C. 2012. SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res*, 40, W452-7.

SMITH, U. M., CONSUGAR, M., TEE, L. J., MCKEE, B. M., MAINA, E. N., WHELAN, S., MORGAN, N. V., GORANSON, E., GISSEN, P., LILLIQUIST, S., ALIGIANIS, I. A., WARD, C. J., PASHA, S., PUNYASHTHITI, R., MALIK SHARIF, S., BATMAN, P. A., BENNETT, C. P., WOODS, C. G., MCKEOWN, C., BUCOURT, M., MILLER, C. A., COX, P., ALGAZALI, L., TREMBATH, R. C., TORRES, V. E., ATTIE-BITACH, T., KELLY, D. A., MAHER, E. R., GATTONE, V. H., 2ND, HARRIS, P. C. & JOHNSON, C. A. 2006. The transmembrane protein meckelin (MKS3) is mutated in Meckel-Gruber syndrome and the wpk rat. *Nat Genet*, 38, 191-6.

SORUSCH, N., WUNDERLICH, K., BAUSS, K., NAGEL-WOLFRUM, K. & WOLFRUM, U. 2014. Usher syndrome protein network functions in the retina and their relation to other retinal ciliopathies. *Adv Exp Med Biol*, 801, 527-33.

STOLER, N. & NEKRUTENKO, A. 2021. Sequencing error profiles of Illumina

sequencing instruments. *NAR Genom Bioinform*, 3, lqab019.

STRANDE, N. T., RIGGS, E. R., BUCHANAN, A. H., CEYHAN-BIRSOY, O., DISTEFANO, M., DWIGHT, S. S., GOLDSTEIN, J., GHOSH, R., SEIFERT, B. A., SNEDDON, T. P., WRIGHT, M. W., MILKO, L. V., CHERRY, J. M., GIOVANNI, M. A., MURRAY, M. F., O'DANIEL, J. M., RAMOS, E. M., SANTANI, A. B., SCOTT, A. F., PLON, S. E., REHM, H. L., MARTIN, C. L. & BERG, J. S. 2017. Evaluating the Clinical Validity of Gene- Disease Associations: An Evidence-Based Framework Developed by the Clinical Genome Resource. *Am J Hum Genet*, 100, 895-906.

STRANNEHEIM, H., LAGERSTEDT-ROBINSON, K., MAGNUSSON, M., KVARNUNG, M., NILSSON, D., LESKO, N., ENGVALL, M., ANDERLID, B. M., ARNELL, H., JOHANSSON, C. B., BARBARO, M., BJORCK, E., BRUHN, H., EISFELDT, J., FREYER, C., GRIGELIONIENE, G., GUSTAVSSON, P., HAMMARSJO, A., HELLSTROM-PIGG, M., IWARSSON, E., JEMT, A., LAAKSONEN, M., ENOKSSON, S. L., MALMGREN, H., NAESS, K., NORDENSKJOLD, M., OSCARSON, M., PETTERSSON, M., RASI, C., ROSENBAUM, A., SAHLIN, E., SARDH, E., STODBERG, T., TESI, B., THAM, E., THONBERG, H., TOHONEN, V., VON DOBELN, U., VASSILIOU, D., VONLANTHEN, S., WIKSTROM, A. C., WINCENT, J., WINQVIST, O., WREDENBERG, A., YGBERG, S., ZETTERSTROM, R. H., MARITS, P., SOLLER, M. J., NORDGREN, A., WIRTA, V., LINDSTRAND, A. & WEDELL, A. 2021. Integration of whole genome sequencing into a healthcare setting: high diagnostic rates across multiple clinical entities in 3219 rare disease patients. *Genome Med*, 13, 40.

SZYMANSKA, K., BERRY, I., LOGAN, C. V., COUSINS, S. R., LINDSAY, H., JAFRI, H., RAASHID, Y., MALIK-SHARIF, S., CASTLE, B., AHMED, M., BENNETT, C., CARLTON, R. & JOHNSON, C. A. 2012. Founder mutations and genotype-phenotype correlations in Meckel-Gruber syndrome and associated ciliopathies. *Cilia*, 1, 18.

SZYMANSKA, K., BOLDT, K., LOGAN, C. V., ADAMS, M., ROBINSON, P. A., UEFFING, M., ZEQRARAJ, E., WHEWAY, G. & JOHNSON, C. A. 2022. Regulation of canonical Wnt signalling by the ciliopathy protein MKS1 and the E2 ubiquitin-conjugating enzyme UBE2E1. *Elife*, 11.

SZYMANSKA, K. & JOHNSON, C. A. 2012. The transition zone: an essential functional compartment of cilia. *Cilia*, 1, 10.

TA, C. M., VIEN, T. N., NG, L. C. T. & DECAEN, P. G. 2020. Structure and function of polycystin channels in primary cilia. *Cell Signal*, 72, 109626.

TAMBUYZER, E., VANDENDRIESSCHE, B., AUSTIN, C. P., BROOKS, P. J., LARSSON, K., MILLER NEEDLEMAN, K. I., VALENTINE, J., DAVIES, K., GROFT, S. C., PRETI, R., OPREA, T. I. & PRUNOTTO, M. 2020. Therapies for rare diseases: therapeutic modalities, progress and challenges ahead. *Nat Rev Drug Discov*, 19, 93-111.

TASCHNER, M. & LORENTZEN, E. 2016. The Intraflagellar Transport Machinery. *Cold Spring Harb Perspect Biol*, 8.

TEEBI, A. S., AL SALEH, Q. A. & ODEH, H. 1992. Meckel syndrome and neural tube defects in Kuwait. *J Med Genet*, 29, 140.

TEEBI, A. S. & TEEBI, S. A. 2005. Genetic diversity among the Arabs. *Community Genet*, 8, 21-6.

TESTA, F., SODI, A., SIGNORINI, S., DI IORIO, V., MURRO, V., BRUNETTI-PIERRI, R., VALENTE, E. M., KARALI, M., MELILLO, P., BANFI, S. & SIMONELLI, F. 2021.

Spectrum of Disease Severity in Nonsyndromic Patients With Mutations in the CEP290 Gene: A Multicentric Longitudinal Study. *Invest Ophthalmol Vis Sci*, 62, 1.

TONG, Y. G. & BURGLIN, T. R. 2010. Conditions for dye-filling of sensory neurons in *Caenorhabditis elegans*. *J Neurosci Methods*, 188, 58-61.

TSANG, S. H., AYCINENA, A. R. P. & SHARMA, T. 2018. Ciliopathy: Senior-Løken Syndrome. *Adv Exp Med Biol*, 1085, 175-178.

TSANG, S. H. & SHARMA, T. 2018. Leber Congenital Amaurosis. In: TSANG, S. H. & SHARMA, T. (eds.) *Atlas of Inherited Retinal Diseases*. Cham: Springer International Publishing.

TURNBULL, C., SCOTT, R. H., THOMAS, E., JONES, L., MURUGAESU, N., PRETTY, F. B., HALAI, D., BAPLE, E., CRAIG, C., HAMBLIN, A., HENDERSON, S., PATCH, C., O'NEILL, A., DEVEREAU, A., SMITH, K., MARTIN, A. R., SOSINSKY, A., MCDONAGH, E. M., SULTANA, R., MUELLER, M., SMEDLEY, D., TOMS, A., DINH, L., FOWLER, T., BALE, M., HUBBARD, T., RENDON, A., HILL, S., CAULFIELD, M. J. & GENOMES, P. 2018. The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. *BMJ*, 361, k1687.

UNIPROT, C. 2019. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res*, 47, D506-D515.

VALENTE, E. M., LOGAN, C. V., MOUGOU-ZERELLI, S., LEE, J. H., SILHAVY, J. L., BRANCATI, F., IANNICELLI, M., TRAVAGLINI, L., ROMANI, S., ILLI, B., ADAMS, M., SZYMANSKA, K., MAZZOTTA, A., LEE, J. E., TOLENTINO, J. C., SWISTUN, D., SALPIETRO, C. D., FEDE, C., GABRIEL, S., RUSS, C., CIBULSKIS, K., SOUGNEZ, C., HILDEBRANDT, F., OTTO, E. A., HELD, S., DIPLAS, B. H., DAVIS, E. E., MIKULA L.M., STROM, C. M., BEN-ZEEV, B., LEV, D., SAGIE, T. L., MICHELSON, M., YARON, Y., KRAUSE, A., BOLTSHAUSER, E., ELKHARTOUFI, N., ROUME, J., SHALEV, S., MUNNICH, A., SAUNIER, S., INGLEHEARN, C., SAAD, A., ALKINDY, A., THOMAS, S., VEKEMANS, M., DALLAPICCOLA, B., KATSANIS, N., JOHNSON, C. A., ATTIE-BITACH, T. & GLEESON, J. G. 2010. Mutations in TMEM216 perturb ciliogenesis and cause Joubert, Meckel and related syndromes. *Nat Genet*, 42, 619-25.

VALENTE, E. M., ROSTI, R. O., GIBBS, E. & GLEESON, J. G. 2014. Primary cilia in neurodevelopmental disorders. *Nat Rev Neurol*, 10, 27-36.

VAN DAM, T. J. P., KENNEDY, J., VAN DER LEE, R., DE VRIEZE, E., WUNDERLICH, K. A., RIX, S., DOUGHERTY, G. W., LAMBACHER, N. J., LI, C., JENSEN, V. L., LEROUX, M. R., HJEIJ, R., HORN, N., TEXIER, Y., WISSINGER, Y., VAN REEUWIJK, J., WHEWAY, G., KNAPP, B., SCHEEL, J. F., FRANCO, B., MANS, D. A., VAN WIJK, E., KEPES, F., SLAATS, G. G., TOEDT, G., KREMER, H., OMRAN, H., SZYMANSKA, K., KOUTROUMPAS, K., UEFFING, M., NGUYEN, T. T., LETTEBOER, S. J. F., OUD, M. M., VAN BEERSUM, S. E. C., SCHMIDTS, M., BEALES, P. L., LU, Q., GILES, R. H., SZKLARCZYK, R., RUSSELL, R. B., GIBSON, T. J., JOHNSON, C. A., BLACQUE, O. E., WOLFRUM, U., BOLDT, K., ROEPMAN, R., HERNANDEZ-HERNANDEZ, V. & HUYNEN, M. A. 2019. CiliaCarta: An integrated and validated compendium of ciliary genes. *PLoS One*, 14, e0216705.

VAN DE WEGHE, J. C., GOMEZ, A. & DOHERTY, D. 2022. The Joubert-Meckel-Nephronophthisis Spectrum of Ciliopathies. *Annu Rev Genomics Hum Genet*, 23, 301-329

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. 2014. Ten years of next-generation sequencing technology. *Trends Genet*, 30, 418-26.

VERBAKEL, S. K., VAN HUET, R. A. C., BOON, C. J. F., DEN HOLLANDER, A. I., COLLIN, R. W. J., KLAVER, C. C. W., HOYNG, C. B., ROEPMAN, R. & KLEVERING, B. J. 2018. Non-syndromic retinitis pigmentosa. *Prog Retin Eye Res*, 66, 157-186.

VINCENSINI, L., BLISNICK, T. & BASTIN, P. 2011. 1001 model organisms to study cilia and flagella. *Biol Cell*, 103, 109-30.

WAHRMAN, J., BERANT, M., JACOBS, J., AVIAD, I. & BEN-HUR, N. 1966. The oral-facial-digital syndrome: a male-lethal condition in a boy with 47/xy chromosomes. *Pediatrics*, 37, 812-21.

WAI, H. A., LORD, J., LYON, M., GUNNING, A., KELLY, H., CIBIN, P., SEABY, E. G., SPIERS-FITZGERALD, K., LYE, J., ELLARD, S., THOMAS, N. S., BUNYAN, D. J., DOUGLAS, A. G. L., BARALLE, D., SPLICING & DISEASE WORKING, G. 2020. Blood RNA analysis can increase clinical diagnostic rate and resolve variants of uncertain significance. *Genet Med*, 22, 1005-1014.

WATERS, A. M. & BEALES, P. L. 2011. Ciliopathies: an expanding disease spectrum. *Pediatr Nephrol*, 26, 1039-56.

WHEWAY, G., MITCHISON, H. M. & GENOMICS ENGLAND RESEARCH, C. 2019. Opportunities and Challenges for Molecular Understanding of Ciliopathies-The 100,000 Genomes Project. *Front Genet*, 10, 127.

WHEWAY, G., NAZLAMOVA, L. & HANCOCK, J. T. 2018. Signaling through the Primary Cilium. *Front Cell Dev Biol*, 6, 8.

WHEWAY, G., PARRY, D. A. & JOHNSON, C. A. 2014. The role of primary cilia in the development and disease of the retina. *Organogenesis*, 10, 69-85.

WHEWAY, G., SCHMIDTS, M., MANS, D. A., SZYMANSKA, K., NGUYEN, T. T., RACHER, H., PHELPS, I. G., TOEDT, G., KENNEDY, J., WUNDERLICH, K. A., SORUSCH, N., ABDELHAMED, Z. A., NATARAJAN, S., HERRIDGE, W., VAN REEUWIJK, J., HORN, N., BOLDT, K., PARRY, D. A., LETTEBOER, S. J. F., ROOSING, S., ADAMS, M., BELL, S. M., BOND, J., HIGGINS, J., MORRISON, E. E., TOMLINSON, D. C., SLAATS, G. G., VAN DAM, T. J. P., HUANG, L., KESSLER, K., GIESSL, A., LOGAN, C. V., BOYLE, E. A., SHENDURE, J., ANAZI, S., ALDAHMEH, M., AL HAZZAA, S., HEGELE, R. A., OBER, C., FROSK, P., MHANNI, A. A., CHODIRKER, B. N., CHUDLEY, A. E., LAMONT, R., BERNIER, F. P., BEAULIEU, C. L., GORDON, P., PON, R. T., DONAHUE, C., BARKOVICH, A. J., WOLF, L., TOOMES, C., THIEL, C. T., BOYCOTT, K. M., MCKIBBIN, M., INGLEHEARN, C. F., CONSORTIUM, U. K., UNIVERSITY OF WASHINGTON CENTER FOR MENDELIAN, G., STEWART, F., OMRAN, H., HUYNEN, M. A., SERGOUNIOTIS, P. I., ALKURAYA, F. S., PARBOOSINGH, J. S., INNES, A. M., WILLOUGHBY, C. E., GILES, R. H., WEBSTER, A. R., UEFFING, M., BLACQUE, O., GLEESON, J. G., WOLFRUM, U., BEALES, P. L., GIBSON, T., DOHERTY, D., MITCHISON, H. M., ROEPMAN, R. & JOHNSON, C. A. 2015. An siRNA-based functional genomics screen for the identification of regulators of ciliogenesis and ciliopathy genes. *Nat Cell Biol*, 17, 1074- 1087.

WILLEY, C. J., BLAIS, J. D., HALL, A. K., KRASA, H. B., MAKIN, A. J. & CZERWIEC, F. S. 2017. Prevalence of autosomal dominant polycystic kidney disease in the European Union. *Nephrol Dial Transplant*, 32, 1356-1363.

WILLIAMS, C. L., LI, C., KIDA, K., INGLIS, P. N., MOHAN, S., SEMENEC, L., BIALAS, N. J., STUPAY, R. M., CHEN, N., BLACQUE, O. E., YODER, B. K. & LEROUX, M. R. 2011. MKS and NPHP modules cooperate to establish basal body/transition zone membrane associations and ciliary gate function during ciliogenesis. *J Cell Biol*, 192,

1023-41.

WRIGHT, C. F., FITZGERALD, T. W., JONES, W. D., CLAYTON, S., MCRAE, J. F., VAN KOGELBERG, M., KING, D. A., AMBRIDGE, K., BARRETT, D. M., BAYZETINOVA, T., BEVAN, A. P., BRAGIN, E., CHATZIMICHALI, E. A., GRIBBLE, S., JONES, P., KRISHNAPPA, N., MASON, L. E., MILLER, R., MORLEY, K. I., PARTHIBAN, V., PRIGMORE, E., RAJAN, D., SIFRIM, A., SWAMINATHAN, G. J., TIVEY, A. R., MIDDLETON, A., PARKER, M., CARTER, N. P., BARRETT, J. C., HURLES, M. E., FITZPATRICK, D. R., FIRTH, H. V. & STUDY, D. D. D. 2015. Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. *Lancet*, 385, 1305-14.

WRIGHT, C. F., FITZPATRICK, D. R. & FIRTH, H. V. 2018a. Paediatric genomics: diagnosing rare disease in children. *Nat Rev Genet*, 19, 253-268.

WRIGHT, C. F., MCRAE, J. F., CLAYTON, S., GALLONE, G., AITKEN, S., FITZGERALD, T. W., JONES, P., PRIGMORE, E., RAJAN, D., LORD, J., SIFRIM, A., KELSELL, R., PARKER, M. J., BARRETT, J. C., HURLES, M. E., FITZPATRICK, D. R., FIRTH, H. V. & STUDY, D. D. D. 2018b. Making new genetic diagnoses with old data: iterative reanalysis and reporting from genome-wide data in 1,133 families with developmental disorders. *Genet Med*, 20, 1216-1223.

WRIGHT, C. F., PARKER, M. & LUCASSEN, A. M. 2019. When genomic medicine reveals misattributed genetic relationships-the debate about disclosure revisited. *Genet Med*, 21, 97-101.

WRIGHT, C. F., QUAIFE, N. M., RAMOS-HERNANDEZ, L., DANECEK, P., FERLA, M. P., SAMOCHA, K. E., KAPLANIS, J., GARDNER, E. J., EBERHARDT, R. Y., CHAO, K. R., KARCZEWSKI, K. J., MORALES, J., GALLONE, G., BALASUBRAMANIAN, M., BANKA, S., GOMPERTZ, L., KERR, B., KIRBY, A., LYNCH, S. A., MORTON, J. E. V., PINZ, H., SANBURY, F. H., STEWART, H., ZUCCARELLI, B. D., GENOMICS ENGLAND RESEARCH, C., COOK, S. A., TAYLOR, J. C., JUUSOLA, J., RETTERER, K., FIRTH, H. V., HURLES, M. E., LARA-PEZZI, E., BARTON, P. J. R. & WHIFFIN, N. 2021. Non-coding region variants upstream of MEF2C cause severe developmental disorder through three distinct loss-of-function mechanisms. *Am J Hum Genet*, 108, 1083-1094.

XUE, K. & MACLAREN, R. E. 2020. Antisense oligonucleotide therapeutics in clinical trials for the treatment of inherited retinal diseases. *Expert Opin Investig Drugs*, 29, 1163-1170.

YANG, L., GUELL, M., BYRNE, S., YANG, J. L., DE LOS ANGELES, A., MALI, P., AACH, J., KIM-KISELAK, C., BRIGGS, A. W., RIOS, X., HUANG, P. Y., DALEY, G. & CHURCH, G. 2013. Optimization of scarless human stem cell genome editing. *Nucleic Acids Res*, 41, 9049-61.

YANG, T. T., SU, J., WANG, W. J., CRAIGE, B., WITMAN, G. B., TSOU, M. F. & LIAO, J. C. 2015. Superresolution Pattern Recognition Reveals the Architectural Map of the Ciliary Transition Zone. *Sci Rep*, 5, 14096.

YANG, Y. & MLODZIK, M. 2015. Wnt-Frizzled/planar cell polarity signaling: cellular orientation by facing the wind (Wnt). *Annu Rev Cell Dev Biol*, 31, 623-46.

YEO, G. & BURGE, C. B. 2004. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol*, 11, 377-94.

YOUNG, I. D., RICKETT, A. B. & CLARKE, M. 1985. High incidence of Meckel's

syndrome in Gujarati Indians. *J Med Genet*, 22, 301-4.

YU, J., SZABO, A., PAGNAMENTA, A. T., SHALABY, A., GIACOPUZZI, E., TAYLOR, J., SHEARS, D., PONTIKOS, N., WRIGHT, G., MICHAELIDES, M., HALFORD, S. & DOWNES, S. 2022. SVRare: discovering disease-causing structural variants in the 100K Genomes Project. *medRxiv*, 2021.10.15.21265069.

ZAMPAGLIONE, E., KINDE, B., PLACE, E. M., NAVARRO-GOMEZ, D., MAHER, M., JAMSHIDI, F., NASSIRI, S., MAZZONE, J. A., FINN, C., SCHLEGEL, D., COMANDER, J., PIERCE, E. A. & BUJAKOWSKA, K. M. 2020. Copy-number variation contributes 9% of pathogenicity in the inherited retinal degenerations. *Genet Med*, 22, 1079-1087.

ZENG, H., JIA, J. & LIU, A. 2010. Coordinated translocation of mammalian Gli proteins and suppressor of fused to the primary cilium. *PLoS One*, 5, e15900.

ZERTI, D., COLLIN, J., QUEEN, R., COCKELL, S. J. & LAKO, M. 2020. Understanding the complexity of retina and pluripotent stem cell derived retinal organoids with single cell RNA sequencing: current progress, remaining challenges and future prospective. *Curr Eye Res*, 45, 385-396.

ZURIS, J. A., THOMPSON, D. B., SHU, Y., GUILINGER, J. P., BESSEN, J. L., HU, J. H., MAEDER, M. L., JOUNG, J. K., CHEN, Z. Y. & LIU, D. R. 2015. Cationic lipid-mediated delivery of proteins enables efficient protein-based genome editing in vitro and in vivo. *Nat Biotechnol*, 33, 73-80.