Author/s:
Ma, A;Grigg, JR;Flaherty, M;Smith, J;Minoche, AE;Cowley, MJ;Nash, BM;Ho, G;Gayagay, T;Lai, T;Farnsworth, E;Hackett, EL;Slater, K;Wong, K;Holman, KJ;Jenkins, G;Cheng, A;Martin, F;Brown, NJ;Leighton, SE;Amor, DJ;Goel, H;Dinger, ME;Bennetts, B;Jamieson, RV

Title:
Genome sequencing in congenital cataracts improves diagnostic yield

Date:
2021-09-01

Citation:
Ma, A., Grigg, J. R., Flaherty, M., Smith, J., Minoche, A. E., Cowley, M. J., Nash, B. M., Ho, G., Gayagay, T., Lai, T., Farnsworth, E., Hackett, E. L., Slater, K., Wong, K., Holman, K. J., Jenkins, G., Cheng, A., Martin, F., Brown, N. J. ,... Jamieson, R. V. (2021). Genome sequencing in congenital cataracts improves diagnostic yield. Human Mutation, 42 (9), pp.1173-1183. https://doi.org/10.1002/humu.24240.

Alan Ma ORCID iD: 0000-0002-9293-4753

Mark Cowley ORCID iD: 0000-0002-9519-5714

Benjamin Nash ORCID iD: 0000-0003-4235-5144

David Amor ORCID iD: 0000-0001-7191-8511

Genome sequencing in congenital cataracts improves diagnostic yield

Alan Ma, [1, 2, 3]

John R. Grigg, [1, 4, 5, 10]

Maree Flaherty, [4,5]

James Smith 4,5

Andre E. Minoche, [6]

Mark J. Cowley, [6, 7, 8]

Benjamin M. Nash, [1, 3, 9]

Gladys Ho, [3, 9]

Thet Gayagay,[9]

Tiffany Lai, [9]

Elizabeth Farnsworth, [9]

This is the author manuscript accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/humu.24240.

Emma L. Hackett, [9]

Katrina Slater,[9]

Karen Wong, [9]

Katherine J. Holman, [9]

Gemma Jenkins, [9]

Anson Cheng, [1]

Frank Martin, [4,5]

Natasha J. Brown, [11,12]

Sarah E Leighton, [13]

David J. Amor, [11, 12]

Himanshu Goel, [14]

Marcel E. Dinger,[6,15]

Bruce Bennetts,[3,9]

Robyn V. Jamieson, [1, 2, 3]

1. Eye Genetics Research Unit, The Children's Hospital at Westmead, Save Sight Institute, Children's Medical Research Institute, University of Sydney, Sydney, NSW, Australia

2. Department of Clinical Genetics, Western Sydney Genetics Program, The Children's Hospital at Westmead, Sydney, NSW, Australia

3. Specialties of Genomic Medicine & Child and Adolescent Health, Faculty of Medicine and Health, University of Sydney, NSW, Australia

4. Department of Ophthalmology, The Children's Hospital at Westmead, Sydney, NSW, Australia

5. Specialty of Ophthalmology, University of Sydney, Sydney, NSW, Australia

6. Kinghorn Centre for Clinical Genomics, Garvan Institute of Medical Research, Darlinghurst NSW, Australia

7. Children's Cancer Institute, Lowy Cancer Research Centre, University of New South Wales, Randwick, New South Wales, Australia

8. St Vincent's Clinical School, Faculty of Medicine, UNSW Australia, Sydney, New South Wales, Australia

9. Sydney Genome Diagnostics, Western Sydney Genetics Program, The Children's Hospital at Westmead, Sydney, NSW, Australia

10. Save Sight Institute, Sydney Eye Hospital, Sydney NSW Australia

11. Department of Paediatrics, University of Melbourne, Parkville, Victoria, Australia

12. Victorian Clinical Genetics Services, Murdoch Children's Research Institute, Melbourne, Australia

13. Tasmanian Clinical Genetics Service, Hobart, Australia

14. Hunter Genetics, Newcastle, NSW, Australia

15. School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW, Sydney, NSW, Australia

RVJ and JRG acknowledge support from NHMRC grant APP1116360, the NSW Office of Health and Medical Research, Costco and the Ophthalmic Research Institute of Australia.

Corresponding Author: Prof R.V. Jamieson, Eye Genetics Research Unit, The Children's Hospital at Westmead, Save Sight Institute, and Children's Medical Research Institute, University of Sydney, Sydney, NSW, Australia 2145; (Email: rjamieson@cmri.org.au; Tel: +61 2 98453273; Fax: +61 2 98453204)

Abstract

Congenital cataracts are one of the major causes of childhood-onset blindness around the world. Genetic diagnosis provides benefits through avoidance of unnecessary tests, surveillance of extraocular features and genetic family information. In this study, we demonstrate the value of genome sequencing in improving diagnostic yield in congenital cataract patients and families. We applied genome sequencing to investigate 20 probands with congenital cataracts. We examined the added value of genome sequencing across a total cohort of 52 probands, including 14 unable to be diagnosed using previous microarray and exome or panel-based approaches. While exome or genome sequencing would have detected the variants in 35/52 (67%) of the cases, specific advantages of genome sequencing led to additional diagnoses in 10% (5/52) of the

overall cohort, and we achieved an overall diagnostic rate of 77% (40/52). Specific benefits of genome sequencing were due to detection of small CNVs (2), indels in repetitive regions (2) or SNVs in GC rich regions (1), not detectable on previous microarray, exome sequencing or panel-based approaches. In other cases, single nucleotide variants were identified in cataract disease genes, including those newly identified since our previous study. This study highlights the additional yield of genome sequencing in congenital cataracts.

Keywords: genome sequencing, genomics, cataracts, microphthalmia

Introduction

Congenital cataracts are one of the commonest causes of childhood visual impairment. While traditionally, patients born with congenital cataracts underwent a number of generally low yield metabolic and infection screens, recent studies have found a high diagnostic yield from next-generation sequencing approaches(Gillespie et al., 2014; A. S. Ma et al., 2016). In congenital cataract patients, diagnosis through next-generation sequencing strategies provides significant clinical utility by allowing avoidance of unnecessary investigations, initiation of surveillance and management for additional relevant manifestations, reduction of prognostic uncertainty and provision of reproductive confidence for families(Javadiyan, Craig, Souzeau, et al., 2017; Lenassi et al., 2020; A. S. Ma et al., 2016).

Despite these advances, there remain 30-40% of congenital cataract sporadic cases and families without a genetic diagnosis following microarray and panel-based or exome sequencing approaches (Javadiyan, Craig, Souzeau, et al., 2017; Lenassi et al., 2020; A. S. Ma et al., 2016). Since genome sequencing (GS) with use of various tools including those for copy number variant (CNV) and structural variant (SV) analysis, has become more readily available and robust, we wished to test the impact of GS on diagnostic yield and clinical utility. This is the first large international report in a cohort of congenital cataract patients undergoing GS, and we demonstrate additional diagnoses and benefits of using GS.

Methods

Editorial Policies and Ethical Considerations

All experiments were approved by the Human Research Ethics Committee of Sydney Children's Hospital Network, Sydney, Australia. Informed consent was obtained from all patients participating in this study.

Twenty probands with apparently non-syndromic congenital cataracts diagnosed in the first year of life were investigated. They included 6 previously untested patients recruited and consented from a genetic eye clinic in a major paediatric referral hospital, (Supp. Table S1) as well as 14 unsolved probands from our previous study (A. S. Ma et al., 2016). These 14 unsolved patients had previously had a panel of 32 congenital cataract genes analysed on two platforms – the Illumina TruSeq Custom Amplicon

(Version 1.5; Illumina Inc., 2011–2013, CA, USA), and TruSight One Clinical Exome (Illumina Inc., 2013–2014) and CNV analysis on CMA (chromosome microarray) and no genetic cause was found. The majority (18/20) were from a Caucasian background, with a small minority with Middle Eastern heritage (2). The majority (12/20) were sporadic cases on referral, while 8/20 were familial, 7 with an autosomal dominant pedigree, and one which was suggestive of autosomal recessive inheritance (Supp. Table S1). Ophthalmological details and samples for genomic DNA extraction were collected from family members when available.

Genome Sequencing, Variant Filtering and Analysis

Patients underwent GS with the TruSeq Nano HT kit on an Illumina HiSeq X sequencer and 2x150bp paired-end sequencing was performed (Illumina Inc.) at the Kinghorn Centre for Clinical Genomics, Garvan Institute, Sydney Australia. The genomic alignment, variant calling, annotation and filtering were performed as in our previous studies, with specific examination of an expanded panel of 55 congenital cataract genes (A. Ma et al., 2020; A. S. Ma et al., 2016; A. S. Ma et al., 2018; Nash et al., 2018; Prokudin et al., 2014). The gene list of 55 was updated from our initial 32, based on recent reviews of the cataract gene literature and the addition of 12 non-syndromal, and 11 syndrome-associated cataract genes (Reis & Semina, 2018) (Supp. Table S2). Single nucleotide variant (SNV) and short indel analysis utilized QIAGEN's Ingenuity Variant Analysis software (IVA - www.qiagen.com/ingenuity - QIAGEN Redwood City) and Alamut-Visual (Version 2.6, Jan 2015; Interactive Biosoftware,

http://www.interactive-biosoftware.com/alamut-visual/). Variants were first filtered for confidence, with variants with coverage <20X or a quality score of <20 removed. Further filtration occurred based on minor allele frequencies in the dbSNP [Smigielski et al., 2000] (http://www.ncbi.nlm.nih.gov/projects/SNP), 1000genomes [Abecasis et al., 2012] (http://www.1000genomes.org/), Exome Variant Server (NHLBI GO Exome Sequencing Project [ESP], Seattle, WA; http://evs.gs.washington.edu/EVS/) and Exome Aggregation Consortium (ExAC) and GnomAD databases (Cambridge, MA, http://exac.broadinstitute.org/, http://gnomad.broadinstitute.org/). Heterozygous variants with minor allele frequencies >0.01 were filtered out, unless they were an established Human Gene Mutation Database (HGMD) or ClinVar pathogenic or likely pathogenic variant. Variants were then manually filtered for possible pathogenic clinical significance according to the 2015 American College of Medical Genetics and Genomics (ACMG) guidelines (Richards et al., 2015). For assessing missense mutations, Alamut Visual (Version 2.6, Jan 2015; Interactive Biosoftware, http://www.interactive-biosoftware.com/alamut-visual/) was used, providing computational algorithms for SIFT (Version 1, http://sift.jcvi.org/), PolyPhen-2 (Version 2.2.2, 2012, http://genetics.bwh.harvard.edu/pph2/) and Mutation Taster (Version 2, 2012, http://www.mutationtaster.org/). Conservation with PhyloP scores [Pollard et al., 2010] was used in prioritizing variants. Variants were also inspected using the Integrated Genomics Viewer (IGV) (Robinson et al., 2011) (Supp. Table S3, Supp. Figure S1).

The final list of SNVs was annotated as per ACMG guidelines(Richards et al., 2015), and any pathogenic or likely pathogenic variants were validated using Sanger sequencing in the proband and segregation was performed in family members when available (Table 1, Supp. Table S3). All variants reported in this manuscript have been submitted to ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/). The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Structural variant (SV), including CNV analysis, was undertaken with a bioinformatics structural variant analysis pipeline ClinSV (Minoche et al., 2021) utilizing evidence from split-reads, discordant pairs and depth-of-coverage to obtain rare, high confidence SV and CNV calls, in accordance with best practice guidelines. SVs in cataract genes were reviewed on IGV and had further confirmation by either Sanger sequencing or chromosome microarray to confirm CNVs (Illumina Infinium CytoSNP-850K microarray with mean effective resolution of 0.06Mb, and Agilent SurePrint G3 Human Microarray 2x400K targeted to Chromosome X, with mean effective resolution of 0.006Mb).

Results

Specific advantages of GS reveals diagnoses in 5 cases, with an expanded gene list contributing to other diagnosed cases

In the 20 patients examined in this study, four had causative small structural variants shown to be difficult to detect on exome sequencing (Patients 2,6,19 and 20), while one had a variant in a GC rich region with a poor detection rate on previous exome sequencing (Patient 18) (A. S. Ma et al., 2016). In two cases diagnoses were made due to the extended gene list used in this study (Patients 1 and 5), while an SNV in a known disease gene was detected in a previously uninvestigated case (Patient 17). Hence, eight of twenty cases were diagnosed in this study providing four new diagnoses in the 14 previously unsolved cases from our original cohort of 46 patients, and four diagnoses were made in the six newly investigated cases in this study. Of our original 46-patient cohort study, 36 are now solved, and with four of the six additional cases solved, this gives a combined overall detection rate of 77% (40/52) (Table 1, Fig. 1A). Of the diagnosed cases, 67% (35/52) had SNVs or small indels in cataract disease genes either detected or expected to be able to be detected on exome or genome sequencing. An additional 10% (5/52) of diagnosed cases were attributable to the value of the GS platform, especially for detection of indels in repetitive regions, small CNVs, and SNVs in high GC content regions (Fig. 1A). This is a much higher yield than previous studies including our own (Javadiyan, Craig, Souzeau, et al., 2017; Lenassi et al., 2020; A. S. Ma et al., 2016).

GS improves detection of common *PITX3* insertion and *MAF* variants

Two patients harboured a recurrent pathogenic 17bp *PITX3* duplication c.657_673dup17 p.(Leu225Profs*90) (Anand, Agrawal, Slavotinek, & Lachke, 2018; Zazo Seco et al., 2018), not detected on clinical exome (Table 1, Fig. 2A, Supp. Figure S1). The first patient (Patient 2) was part of our original cataract cohort, from an autosomal dominant family comprising a mother and two daughters with congenital cataracts, sclerocornea and microphthalmia (Fig. 5A). They were negative on clinical exome and amplicon-based approaches (Illumina TruSight and TruSeq), but GS identified the insertion in this region (Supp. Figure S1), which was confirmed on Sanger sequencing. This variant segregated appropriately in both affected daughters and their affected mother on Sanger sequencing (Supp. Table S3). The second patient (Patient 19) was from an autosomal dominant family with cataracts, also referred for a clinical exome which was reported as negative, and GS identified the same insertion, which led to a frameshift and premature truncation of PITX3. This result was also confirmed on Sanger sequencing, while additional family members were unavailable for study (Supp. Table S3).

We also report an additional novel *MAF* missense variant (c.916C>G, p.(Arg306Gly)) (Table 1, Fig. 2B&C). This patient (patient 18) was born with bilateral congenital cataracts, and had lensectomies performed at the age of 2 months. His father, paternal aunt and grandmother also had congenital cataracts. This basic region of *MAF* was difficult to sequence in our past study using ES mainly due to the very GC rich nature of the

region, requiring gap filling, and we also identified this as a 'hotspot' for missense variants in the gene (A. S. Ma et al., 2016) (Fig. 2C).

GS with CNV/SV analysis in cataracts detects small novel deletions in *MIP* and *NHS* undetectable by chromosome microarray

A 3.2kb deletion including exon 4 of the known cataract disease gene *MIP*, was found on GS, in an autosomal dominant family with congenital cataracts and microphthalmia (Patient 6, Fig. 5B). This deletion removes the coding regions of exon 4, comprising the final transmembrane (TM6) and cytoplasmic domains of the MIP protein, as well as the 3'UTR (Fig. 3A, Supp. Figure S1). *MIP* contains six transmembrane domains, acts as a hemipore for intercellular water transport, and forms part of the thin junction between lens fiber cells (Gu et al., 2007). More than 20 pathogenic variants, including missense, splice site, frameshift and nonsense variants, have been reported in this gene, including in exon 4, but no exonic deletions to date. The 3.2kb deletion extends from intron 3 and the coding exon 4 and 3'UTR region of *MIP*, and includes exon 1 in the 5'UTR of the adjacent gene *TIMELESS* (NM_003920.5), which is not a known cataract or other disease-causing gene. The deletion was beyond the mean effective resolution of the 850K Illumina cytoSNP microarray (0.06Mb). GS was valuable in resolving this and Sanger sequencing was used to confirm the deletion region.

One patient (patient 20) was also diagnosed by GS with a 3Kb exonic deletion in the X-linked syndromal cataract gene *NHS* (Table 1, Fig. 3B, Supp. Figure S1). This patient was born with congenital bilateral posterior

cortical cataracts, diagnosed at 3 months of age, and operated for removal at age 6 months. He also had superior iris hypoplasia, aphakic glaucoma, microcornea and nystagmus. Other symptoms included supernumerary teeth, and a patent ductus arteriosus requiring surgical closure. This exonic deletion (part of exon 6, and all of exon 7) was beyond the standard resolution of the 850kb SNP microarray. Further studies using a high-density custom X-chromosome microarray confirmed the deletion found on GS, and identified that it was heterozygous in his carrier mother. Similar deletions, and deleterious variants in *NHS* have been reported in the Nance Horan syndrome literature (Huang et al., 2007; Sharma et al., 2006).

Novel variants found in *LONP1, COL4A1* and *GJA8*

A recently reported syndromal gene, *LONP1*, was included in this analysis, and a novel homozygous variant was found in one of our original cataract cohort families (Patient 1 this study, Family 1, ref: Ma et al, 2016) (Table 1, Fig. 4A, 5C). This family was initially referred due to congenital cataracts in two siblings. The sister was intellectually normal, while her brother had more severe cataracts, and was born with imperforate anus and mild intellectual disability. Both had subtle dysmorphic facial features with flattened helices and a prominent long grooved nasal tip. A novel homozygous missense variant (c.1939G>A; p.(Glu647Lys)) was found in both siblings, and both parents were confirmed to be heterozygous carriers. This gene, initially reported in the mitochondrial literature, is now a recognized cause of syndromal congenital cataracts (Inui et al., 2017;

Khan & AlBakri, 2018; Peter et al., 2018). Our report indicates the phenotypic variability and subtlety of syndromic clinical features that may be present in association with variation in this gene.

A novel likely pathogenic heterozygous variant was found in *COL4A1*, in Patient 5 (Patient 9 in Ma et al 2016) (Table 1, Figs. 4B & 5D), who had bilateral congenital cataracts and microphthalmia. The missense variant (c.2336G>A, p.(Gly779Val)) affected a key glycine (G-X-Y) residue of the *COL4A1* protein and such variants are well reported in the literature (Mao, Kiss, Ou, & Gould, 2017; Meuwissen et al., 2015).

A heterozygous novel likely pathogenic variant in the key cataract gene *GJA8* (c.175C>A; p.(Pro59Thr)) was found, in the region encoding the first extracellular domain (EC1), in Patient 17 (Table 1, Fig. 4C & 5E). There are multiple reported pathogenic missense variants in this domain (Fig. 4D), highlighting its important role in the function of the protein (A. S. Ma et al., 2018).

Discussion

This is the first study applying GS to a large cohort of patients with congenital cataracts, and highlights the additional diagnostic yield compared to conventional microarray with panel or exome sequencing. Overall we found a high diagnostic yield of 77% across our entire cohort (40/52) (Fig. 1A). Increased diagnostic yield that could be specifically attributed to GS, was due to the enhanced ability of GS to detect small CNVs and indels in repetitive regions, as well as variants in GC rich

regions of the genome. This accounted for 5/52 (10%) of cases across our full cohort (Fig. 1A) (Patients 2, 6, 18, 19 and 20). An extended gene list of additional identified disease genes compared with our previous study (A. S. Ma et al., 2016), also provided genetic diagnoses in two cases in this study (Patients 1 and 5).

Overall, with use of a comprehensive list of cataract disease genes for investigation, exome or genome sequencing across our cohort of 52 patients would be expected to find genetic diagnoses in 67% (35/52) of cases. In our cohort, additional diagnoses were able to made in 5 cases, due to the specific benefits of the GS platform for small CNVs, repetitive region indels and GC rich regions, leading to an overall diagnostic rate of 77% (40/52). The platform to be used will depend upon local availability and cost, and it is noteworthy that the use of ES or GS in patients with congenital cataracts has many benefits. Firstly, it assists in avoidance of additional unnecessary investigations of low yield, such as metabolic and infection screens. Secondly, it provides accurate diagnostic and family recurrence risk information that helps inform further reproductive planning, access to prenatal genetic counselling, and preimplantation genetic diagnosis for couples. Finally, it removes diagnostic uncertainty and provides management and surveillance options, especially in syndromal cases, that would not be available based on clinical diagnosis alone.

GS in congenital cataracts assisting with genetic diagnosis in small duplications, CNVs and GC rich regions

In *PITX3*, a short pathogenic duplication was found by genome sequencing in patients 2 and 19. Previous panel-based and ES methods both failed to identify this variant in the two probands (Supp. Figure S1). The 17bp duplication involves an 11bp repetitive motif (CAGGGCCTGGG) flanked with a 6bp sequence GCCCTG (Fig. 2A). Visual re-inspection revealed 0/900 supporting reads in the TruSeq amplicon panel-based data, and just 4/45 supporting reads with significant strand bias in the original TruSight capture ES data, which at the time was discarded due to insufficient evidence. We hypothesize that the improved performance of GS over this difficult region may be due to more accurate duplication calling, better performance in GC-rich regions, and no capture-bias towards the reference allele. This duplication is well reported (Anand et al., 2018; Zazo Seco et al., 2018). A corresponding deletion has been reported once, in homozygous form, causing severe ASD/microphthalmia (Aldahmesh, Khan, Mohamed, & Alkuraya, 2011). This finding has important implications about the need for 'gapfill' Sanger sequencing to look specifically for this duplication in ES-negative patients, which would be recommended if GS is not available.

Similarly, in *MAF*, there were many initial difficulties in the NGS platform, and Sanger sequencing of the variants found in the bZIP domain in our original cohort (A. S. Ma et al., 2016). In fact, one of the missense variants reported in that paper in this domain was only diagnosed after Sanger

gap-filling, and due to the GC rich region this was problematic and difficult to cover. This issue was aided in resolution through use of GS and here we report an additional likely pathogenic variant in this hotspot (Fig. 2B). This gene has a high rate of variation in this region in our cohort (4/52 patients, 8% of the total cohort – Fig. 1B), which is higher than the reported literature. This may indicate that *MAF* could be a more frequent cause of congenital cataracts than currently recognised.

Including our results, there are now at least 20 variants reported in *MAF* since its discovery in 2002 (Hansen, Eiberg, & Rosenberg, 2007; Jamieson et al., 2002; Javadiyan, Craig, Sharma, et al., 2017; Narumi et al., 2014; Perveen, Favor, Jamieson, Ray, & Black, 2007; Si et al., 2019; Vanita, Guo, Singh, Ott, & Robinson, 2014), including the three cataract-related pathogenic variants we reported previously (A. S. Ma et al., 2016). Also, the variant we have identified lies amongst others in a hotspot within the highly conserved bZIP domain, and has a role in binding to target DNA and dimerization with other proteins with a bZIP domain (Fig. 2B&C). Recent work (Si et al., 2019) has demonstrated significantly impaired transactivation of four important crystallin genes (*CRYGA, CRYAA, CRYBA1* and *CRYBA4*) due to variants in this region.

Our finding of two small (~3kb) deletions in *MIP* and *NHS,* (Figs. 3A&B*)* demonstrates the improved detection of small deletions using GS. Both deletions were smaller than the resolution of conventional microarrays or CNV-calling algorithms for clinical ES and had important implications for

diagnosis and genetic counselling for recurrence risk(A. S. Ma et al., 2016).

Additional variants in rarely reported and syndromal genes in congenital cataracts

The ability to reinterrogate data with an improved/updated gene list shows the importance of both gene/literature review, as well as the importance of phenotypic information and consideration of syndromal genes. An additional 23 novel and syndromal genes were examined in this study, since our original study which included 32 genes for analysis (Supp. Table S2), resulting in diagnoses in 2/14 negative families from our first study (*COL4A1* and *LONP1*). The diagnosis of *LONP1*- associated recessive syndromal cataracts raises some lessons regarding the inclusion of syndromal genes in cataract panels, and also the blurring of boundaries between syndromal and non-syndromal presentations and has important implications for management. Initially this gene was reported in the literature as the cause of CODAS (cerebral ocular dental auricular skeletal) syndrome (Strauss et al., 2015), a mitochondrial disorder causing characteristic cataracts, vertebral and dental abnormalities, with dysmorphic facial features. Of note, two of the originally reported patients also had imperforate anus. Additional autosomal recessive families were reported with a much broader range of clinical features, ranging from just isolated congenital cataracts and mild dysmorphism alone, without any other CODAS systemic features, to severe mitochondrial disease and most variants are missense and located in the highly conserved AAA

domain, as was our family's (Fig. 4A) (Inui et al., 2017; Khan & AlBakri, 2018; Peter et al., 2018).

While variants in *COL4A1* were first reported in patients with porencephaly, they have now been implicated in a variety of non-cerebral phenotypes including eye (ASD, retinal vessel tortuosity, cataract) as well as renal, cardiac and other vascular malformations (Mao et al., 2017; Meuwissen et al., 2015). Ophthalmic features, including cataracts or anterior segment dysgenesis, are now the second most common association with *COL4A1* variants after porencephaly, and this may highlight the importance of this extracellular-matrix protein in the formation of the eye structures(A. Ma et al., 2020). Although our patient's p.Gly779Val variant is novel, similar glycine residue missense mutations have been reported nearby (Fig. 4B) in congenital cataract patients including Gly782Ala, Gly773Arg, and Gly708Arg, (Shah et al., 2012; Xia et al., 2014) (Deml et al., 2014). Reported phenotypes associated with these variants include: microphthalmia, cataract and developmental delay in one family, cataract, glaucoma and Peters anomaly in a second, and isolated dominant cataract in a third. All of these findings together demonstrate that this region of *COL4A1* may indeed be a 'hotspot' for missense variants affecting the glycine residues leading to congenital cataracts.

A more comprehensive diagnostic picture with GS in congenital cataracts

Our study demonstrates that GS has a high yield, and as the cost and accessibility improves compared to ES, it may indeed become the first line

investigation in congenital cataracts. A similar theme of increased diagnostic yield from using GS after negative ES has also been found in the retinal dystrophy literature (Ellingford et al., 2016) and is likely to be repeated elsewhere, although there are relatively few studies looking at this issue. Certainly some are arguing that GS should be first line testing such as in cardiomyopathies and dystonia (Bagnall et al., 2018; Kumar et al., 2019; Minoche et al., 2019).

The increased diagnostic rate in this study, contributed to by GS, has also given us a clearer picture of the 'missing' genetic diagnoses in congenital cataracts. In our original congenital cataract cohort study (A. S. Ma et al., 2016) the crystallin and gap junction genes were found to be responsible for over 60% of the genetic diagnoses, and these genes have traditionally been the most studied and sequenced in congenital cataract patients. This was also consistent with another study at the time, where the majority of genetic diagnoses were also in the crystallin and gap junction genes(Javadiyan, Craig, Souzeau, et al., 2017). In this study, where negative cases were analysed and further cases examined, additional variants were mostly found in more rarely reported and new syndromal genes. While crystallins and gap junction genes still contribute to approximately half the reported variants (Fig. 1B), there is an increasing recognition of variants in syndromal genes (*NHS, LONP1*), additional structural proteins (*MIP, COL4A1*), and transcription factors (*MAF, PITX3*). This breakdown of genes is also consistent with a more recent report in the literature (Lenassi et al., 2020) and may represent a truer representation of the genes involved in congenital cataracts. Our group

found a similar theme in the anterior segment disorders, where GS uncovered a more comprehensive representation of the genes involved in the pathogenesis of the disorder, and indicated gene families and interactions underlying these conditions (A. Ma et al., 2020). In the remaining 23% unsolved patients, there may be additional hidden variants within the untranslated, intronic, promoter and regulatory regions of the genome and these regions warrant further study and reanalysis of GS data, as the functional information related to these regions improves over time.

Conclusion

We have demonstrated the high diagnostic utility of ES or GS in a cohort of congenital cataract patients. In this study, GS provided an overall improved diagnostic yield of 10% compared to conventional analysis with clinical ES and chromosome microarray. This is due to the benefit of GS in terms of diagnosing small deletions beyond the resolution of standard chromosome microarray, as well as improved detection in genomically complex regions. Overall our diagnostic rate of 77% in congenital cataracts is very high and demonstrates the added benefit of GS in helping to accurately diagnose and inform families with congenital cataracts.

Data Availability Statement

All variants reported in this manuscript have been submitted to ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/).

Acknowledgements

RVJ and JRG acknowledge support from NHMRC grant APP1116360, the NSW Office of Health and Medical Research, Costco and the Ophthalmic Research Institute of Australia.

On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Aldahmesh, M. A., Khan, A. O., Mohamed, J., & Alkuraya, F. S. (2011). Novel recessive BFSP2 and PITX3 mutations: insights into mutational mechanisms from consanguineous populations. *Genet Med, 13*(11), 978-981. doi:10.1097/GIM.0b013e31822623d5

Anand, D., Agrawal, S. A., Slavotinek, A., & Lachke, S. A. (2018). Mutation update of transcription factor genes FOXE3, HSF4, MAF, and PITX3 causing cataracts and other developmental ocular defects. *Hum Mutat, 39*(4), 471-494. doi:10.1002/humu.23395

Bagnall, R. D., Ingles, J., Dinger, M. E., Cowley, M. J., Ross, S. B., Minoche, A. E.,... Semsarian, C. (2018). Whole Genome Sequencing Improves Outcomes of Genetic Testing in Patients With Hypertrophic Cardiomyopathy. *J Am Coll Cardiol, 72*(4), 419-429. doi:10.1016/j.jacc.2018.04.078

Deml, B., Reis, L. M., Maheshwari, M., Griffis, C., Bick, D., & Semina, E. V. (2014). Whole exome analysis identifies dominant COL4A1 mutations in patients with complex ocular phenotypes involving microphthalmia. *Clin Genet, 86*(5), 475-481. doi:10.1111/cge.12379

Ellingford, J. M., Barton, S., Bhaskar, S., Williams, S. G., Sergouniotis, P. I., O'Sullivan, J.,... Black, G. C. (2016). Whole Genome Sequencing Increases Molecular Diagnostic Yield Compared with Current Diagnostic Testing for Inherited Retinal Disease. *Ophthalmology, 123*(5), 1143-1150. doi:10.1016/j.ophtha.2016.01.009

Gillespie, R. L., O'Sullivan, J., Ashworth, J., Bhaskar, S., Williams, S., Biswas, S.,... Lloyd, I. C. (2014). Personalized diagnosis and management of congenital cataract by next-generation sequencing. *Ophthalmology, 121*(11), 2124-2137 e2121-2122. doi:10.1016/j.ophtha.2014.06.006

Gu, F., Zhai, H., Li, D., Zhao, L., Li, C., Huang, S., & Ma, X. (2007). A novel mutation in major intrinsic protein of the lens gene (MIP) underlies autosomal dominant cataract in a Chinese family. *Mol Vis, 13*, 1651-1656. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17893667

Hansen, L., Eiberg, H., & Rosenberg, T. (2007). Novel MAF mutation in a family with congenital cataract-microcornea syndrome. *Mol Vis, 13*, 2019-2022. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17982426

Huang, K. M., Wu, J., Brooks, S. P., Hardcastle, A. J., Lewis, R. A., & Stambolian, D. (2007). Identification of three novel NHS mutations

in families with Nance-Horan syndrome. *Mol Vis, 13*, 470-474. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/17417607

Inui, T., Anzai, M., Takezawa, Y., Endo, W., Kakisaka, Y., Kikuchi, A.,... Haginoya, K. (2017). A novel mutation in the proteolytic domain of LONP1 causes atypical CODAS syndrome. *J Hum Genet, 62*(6), 653-655. doi:10.1038/jhg.2017.11

Jamieson, R. V., Perveen, R., Kerr, B., Carette, M., Yardley, J., Heon, E.,... Black, G. C. (2002). Domain disruption and mutation of the bZIP transcription factor, MAF, associated with cataract, ocular anterior segment dysgenesis and coloboma. *Hum Mol Genet, 11*(1), 33-42. doi:10.1093/hmg/11.1.33

Javadiyan, S., Craig, J. E., Sharma, S., Lower, K. M., Casey, T., Haan, E.,... Burdon, K. P. (2017). Novel missense mutation in the bZIP transcription factor, MAF, associated with congenital cataract, developmental delay, seizures and hearing loss (Ayme-Gripp syndrome). *BMC Med Genet, 18*(1), 52. doi:10.1186/s12881-017-0414-7

Javadiyan, S., Craig, J. E., Souzeau, E., Sharma, S., Lower, K. M., Mackey, D. A.,... Burdon, K. P. (2017). High-Throughput Genetic Screening of 51 Pediatric Cataract Genes Identifies Causative Mutations in Inherited Pediatric Cataract in South Eastern Australia. *G3 (Bethesda), 7*(10), 3257-3268. doi:10.1534/g3.117.300109

Khan, A. O., & AlBakri, A. (2018). Clinical features of LONP1-related infantile cataract. *J AAPOS, 22*(3), 229-231. doi:10.1016/j.jaapos.2017.10.012

Kumar, K. R., Davis, R. L., Tchan, M. C., Wali, G. M., Mahant, N., Ng, K.,... Sue, C. M. (2019). Whole genome sequencing for the genetic diagnosis of heterogenous dystonia phenotypes. *Parkinsonism Relat Disord, 69*, 111-118. doi:10.1016/j.parkreldis.2019.11.004

Lenassi, E., Clayton-Smith, J., Douzgou, S., Ramsden, S. C., Ingram, S., Hall, G.,... Sergouniotis, P. I. (2020). Clinical utility of genetic testing in 201 preschool children with inherited eye disorders. *Genet Med, 22*(4), 745-751. doi:10.1038/s41436-019-0722-8

Ma, A., Yousoof, S., Grigg, J. R., Flaherty, M., Minoche, A. E., Cowley, M. J.,... Jamieson, R. V. (2020). Revealing hidden genetic diagnoses in the ocular anterior segment disorders. *Genet Med*. doi:10.1038/s41436-020-0854-x

Ma, A. S., Grigg, J. R., Ho, G., Prokudin, I., Farnsworth, E., Holman, K.,... Jamieson, R. V. (2016). Sporadic and Familial Congenital Cataracts: Mutational Spectrum and New Diagnoses Using Next-Generation Sequencing. *Hum Mutat, 37*(4), 371-384. doi:10.1002/humu.22948

Ma, A. S., Grigg, J. R., Prokudin, I., Flaherty, M., Bennetts, B., & Jamieson, R. V. (2018). New mutations in GJA8 expand the

phenotype to include total sclerocornea. *Clin Genet, 93*(1), 155-159. doi:10.1111/cge.13045

Mao, M., Kiss, M., Ou, Y., & Gould, D. B. (2017). Genetic dissection of anterior segment dysgenesis caused by a Col4a1 mutation in mouse. *Dis Model Mech, 10*(4), 475-485. doi:10.1242/dmm.027888

Meuwissen, M. E., Halley, D. J., Smit, L. S., Lequin, M. H., Cobben, J. M., de Coo, R.,... Mancini, G. M. (2015). The expanding phenotype of COL4A1 and COL4A2 mutations: clinical data on 13 newly identified families and a review of the literature. *Genet Med, 17*(11), 843-853. doi:10.1038/gim.2014.210

Minoche, A. E., Horvat, C., Johnson, R., Gayevskiy, V., Morton, S. U., Drew, A. P.,... Fatkin, D. (2019). Genome sequencing as a first-line genetic test in familial dilated cardiomyopathy. *Genet Med, 21*(3), 650-662. doi:10.1038/s41436-018-0084-7

Minoche, A. E., Lundie, B., Peters, G. B., Ohnesorg, T., Pinese, M., Thomas, D. M.,... Cowley, M. J. (2021). ClinSV: clinical grade structural and copy number variant detection from whole genome sequencing data. *Genome Med, 13*(1), 32. doi:10.1186/s13073-021-00841-x

Narumi, Y., Nishina, S., Tokimitsu, M., Aoki, Y., Kosaki, R., Wakui, K.,... Kosho, T. (2014). Identification of a novel missense mutation of MAF in a Japanese family with congenital cataract by whole exome

sequencing: A clinical report and review of literature. *Am J Med Genet A, 164*(5), 1272-1276. doi:10.1002/ajmg.a.36433

Nash, B. M., Symes, R., Goel, H., Dinger, M. E., Bennetts, B., Grigg, J. R., & Jamieson, R. V. (2018). NMNAT1 variants cause cone and cone-rod dystrophy. *Eur J Hum Genet, 26*(3), 428-433. doi:10.1038/s41431-017-0029-7

Perveen, R., Favor, J., Jamieson, R. V., Ray, D. W., & Black, G. C. (2007). A heterozygous c-Maf transactivation domain mutation causes congenital cataract and enhances target gene activation. *Hum Mol Genet, 16*(9), 1030-1038. Retrieved from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=17374726

Peter, B., Waddington, C. L., Olahova, M., Sommerville, E. W., Hopton, S., Pyle, A.,... Lightowlers, R. N. (2018). Defective mitochondrial protease LonP1 can cause classical mitochondrial disease. *Hum Mol Genet, 27*(10), 1743-1753. doi:10.1093/hmg/ddy080

Prokudin, I., Simons, C., Grigg, J. R., Storen, R., Kumar, V., Phua, Z. Y.,... Jamieson, R. V. (2014). Exome sequencing in developmental eye disease leads to identification of causal variants in GJA8, CRYGC, PAX6 and CYP1B1. *Eur J Hum Genet, 22*(7), 907-915. doi:10.1038/ejhg.2013.268

Reis, L. M., & Semina, E. V. (2018). Genetic landscape of isolated pediatric cataracts: extreme heterogeneity and variable inheritance patterns within genes. *Hum Genet*. doi:10.1007/s00439-018-1932-x

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J.,... 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*(5), 405-424. doi:10.1038/gim.2015.30

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., & Mesirov, J. P. (2011). Integrative genomics viewer. *Nat Biotechnol, 29*(1), 24-26. doi:10.1038/nbt.1754

Shah, S., Ellard, S., Kneen, R., Lim, M., Osborne, N., Rankin, J.,... Jardine, P. (2012). Childhood presentation of COL4A1 mutations. *Dev Med Child Neurol, 54*(6), 569-574. doi:10.1111/j.1469-8749.2011.04198.x

Sharma, S., Ang, S. L., Shaw, M., Mackey, D. A., Gecz, J., McAvoy, J. W., & Craig, J. E. (2006). Nance-Horan syndrome protein, NHS, associates with epithelial cell junctions. *Hum Mol Genet, 15*(12), 1972-1983. doi:10.1093/hmg/ddl120

Si, N., Song, Z., Meng, X., Li, X., Xiao, W., & Zhang, X. (2019). A novel MAF missense mutation leads to congenital nuclear cataract by

impacting the transactivation of crystallin and noncrystallin genes. *Gene, 692*, 113-118. doi:10.1016/j.gene.2019.01.011

Strauss, K. A., Jinks, R. N., Puffenberger, E. G., Venkatesh, S., Singh, K., Cheng, I.,... Suzuki, C. K. (2015). CODAS syndrome is associated with mutations of LONP1, encoding mitochondrial AAA+ Lon protease. *Am J Hum Genet, 96*(1), 121-135. doi:10.1016/j.ajhg.2014.12.003

Vanita, V., Guo, G., Singh, D., Ott, C. E., & Robinson, P. N. (2014). Differential effect of cataract-associated mutations in MAF on transactivation of MAF target genes. *Mol Cell Biochem, 396*(1-2), 137-145. doi:10.1007/s11010-014-2150-z

Xia, X. Y., Li, N., Cao, X., Wu, Q. Y., Li, T. F., Zhang, C.,... Xue, C. Y. (2014). A novel COL4A1 gene mutation results in autosomal dominant non-syndromic congenital cataract in a Chinese family. *BMC Med Genet, 15*, 97. doi:10.1186/s12881-014-0097-2

Zazo Seco, C., Plaisancie, J., Lupasco, T., Michot, C., Pechmeja, J., Delanne, J.,... Chassaing, N. (2018). Identification of PITX3 mutations in individuals with various ocular developmental defects. *Ophthalmic Genet, 39*(3), 314-320. doi:10.1080/13816810.2018.1430243

Table

Table 1: Patients with likely causative variants

Patient no	Inheritance, Before/ After testing	Gene (NM)	Nucleotide change	Amino acid change	Segregation	Novel^
1*	AR/AR	LONP1 NM_004793.4	hom c.1939G>A;	p.(Glu647Lys)	Yes, heterozygous in parents, homozygous in affected sister	Yes
2*	AD/AD	PITX3 NM_005029.3	c.657_673dup17	p.(Leu225Profs*90)	Yes, present in affected sister and mother	No
5*	Sporadic/ likely new AD	COL4A1 NM_001845.6	c.2336G>A	p.(Gly779Glu)	N/K: parental samples unavailable	Yes
6*	AD/AD	MIP NM_012064.4	Chr12: 56842291-56845532 del		Yes, present in affected father, absent in unaffected mother and sib	Yes
17#	Sporadic/ likely new AD.	GJA8 NM_005267.5	c.175C>A	p.(Pro59Thr)	N/K: parental samples unavailable	yes
18#	Sporadic/ likely new AD.	MAF NM_005360.5	c.916C>G,	p.(Arg306Gly)	N/K: parental samples unavailable	yes
19#	AD/AD	PITX3 NM_005029.3	c.657_673dup17	p.(Leu225Profs*90)	N/K: parental samples unavailable	no
20#	Sporadic/ X-linked	NHS NM_001291	ChrX: 17744374-17748930		Yes, heterozygous in unaffected mother	yes

		867.2	del			

* Patient from previous study who had previously tested negative, # Patient newly tested in this study, ^ More details on the variant interpretation, prioritization, and IGV screenshots are provided in Supp. Table S3 and Supp. Figure S1.

Figure 1: Overall yield and gene breakdown in combined cataract cohort of 52 patients

Figure 1A: demonstrates the overall yield in the congenital cataract cohort studied in this and our previous publication. In 52 patients, pathogenic variants were found in an additional 10% of cases using the genome sequencing platform due to detection of small CNVs (2), indels in repetitive regions (2) and SNVs in GC rich regions (1). Additional SNVs were identified in this study due to use of an extended gene list (2) and identification of an SNV in a known disease gene in a previously uninvestigated case (1). 62% were identified previously on clinical ES.

Figure 1B: Chart demonstrates the overall breakdown, by gene category, in the entire cataract cohort. Over half the responsible variants are still in the frequently reported crystallin and gap junction genes. However, an increasing role of rarely reported genes such as the transcription factors and lens structural proteins has been found. Also, syndromal genes are increasingly reported, even in cases referred initially as non-syndromal cataracts.

Figure 1A: Overall yield in cataract cohort of 52 patients

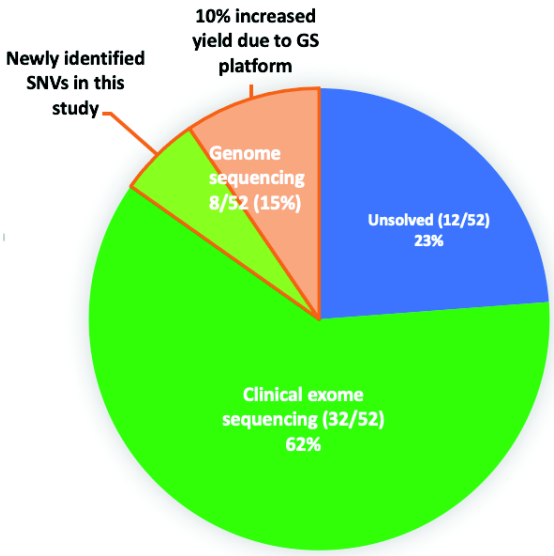


Figure 1B: Overall breakdown, by gene category, of combined cataract cohort

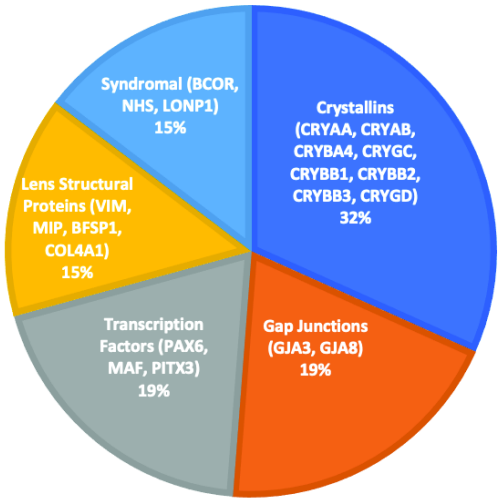


Figure 2: GS detects a common *PITX3* insertion and *MAF* variant in GC rich and repetitive regions of the genome

This figure demonstrates the difficult to sequence regions of PITX3 and MAF. 2A: In PITX3, a repetitive and GC rich region as highlighted harbours a recurrent insertion variant, with a repetitive 11bp motif (in red) flanked by a 6bp sequence (in blue) which has been duplicated. This 17bp duplication was filtered out by the variant call pipeline on amplicon and capture exome based platforms, but retained in the genome sequencing alignment. 2B: Similarly, the b-ZIP region of *MAF* (blue) contains a highly GC rich region that has been traditionally difficult to sequence, requiring gap filling, but also harbouring a majority of missense variants responsible for congenital cataracts in the basic region (shown in 2C with conservation). The p.Arg306Gly variant found in our patient is also highlighted in grey.

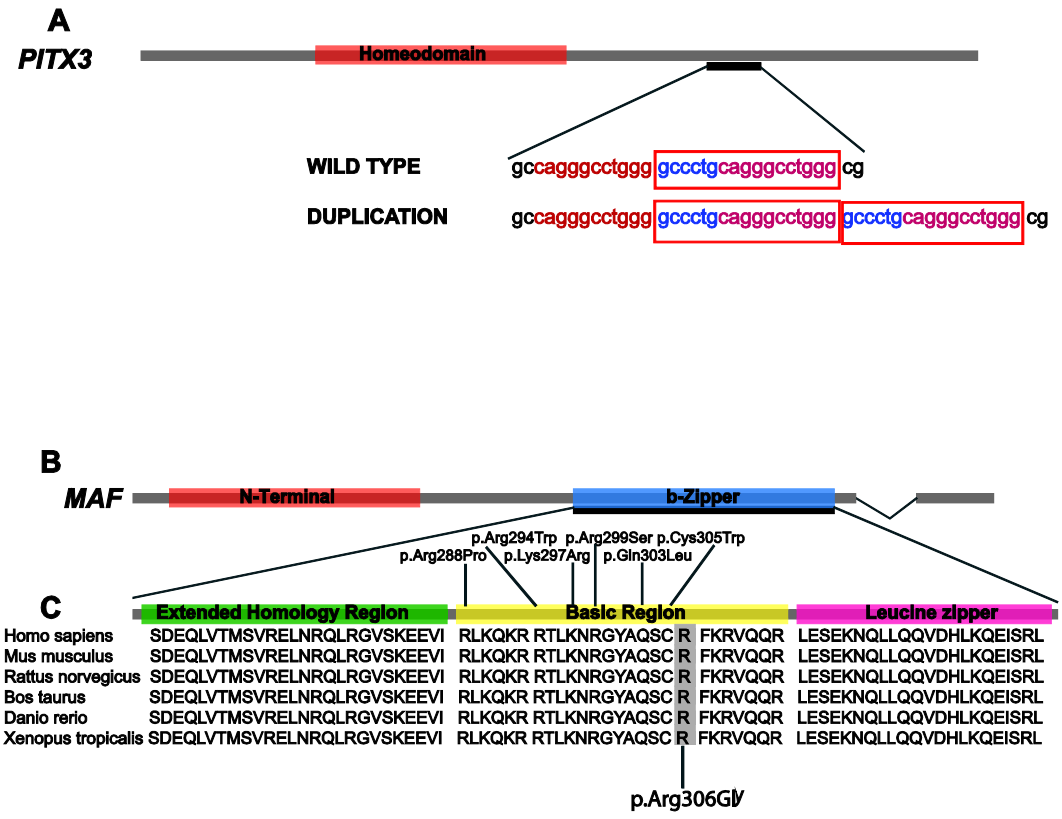


Figure 3: GS detects small novel deletions in *MIP* and *NHS*

undetectable by chromosome microarray

A: Diagram of the protein domain structure (top) and exons/introns (below) of *MIP*. The protein domains are labelled as four cytoplasmic (C) domains, six transmembrane (TM) domains, and three extracellular (EC) domains coded by its four exons. Two pathogenic variants found in the previous cataract paper are illustrated above the gene, along with other variants reported in the literature below the gene (^ nonsense, * frameshift, # splice site ~ missense). Exons are indicated by grey shading, and introns by the thin black line. The 3.2kb deletion found in this paper is illustrated in the thick black line below the gene extending from intron 3 and including the coding exon 4 and 3'UTR region of *MIP*, and 5'UTR region of *TIMELESS*.

B: Pathogenic variants in *NHS*, including previously reported variants marked by symbols below the gene (^ nonsense, * frameshift, # splice site ~ missense), and two pathogenic variants reported in our previous cataract paper demonstrated above the gene. The thick black line represents the 4.6kb deletion of part of exon 6, and exon 7 found in this paper.

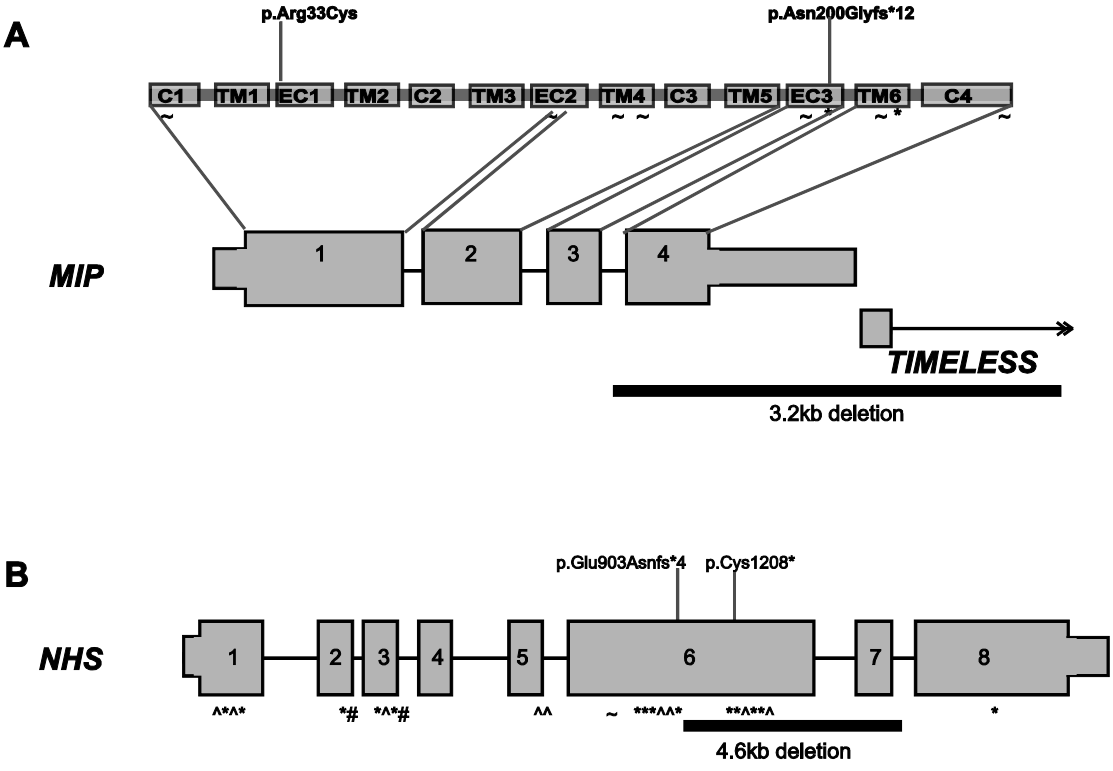


Figure 4: GS in cataracts finds novel variants in syndromal cataract gene *LONP1, COL4A1* and *GJA8*

A: Domains of LONP1 illustrated with previously reported variants below the gene, and the homozygous p.Glu647Lys variant from this paper illustrated above the gene.

B: Domain structure of *COL4A1*, demonstrating its main triple helical domain where most variants reside. The p.Gly779Val variant reported in this paper is illustrated above the gene, and nearby missense variants reported in the literature and cataracts, showing this may be a 'hotspot' of missense glycine variants responsible for cataract formation.

C. Demonstrates the domain structure of *GJA8* which has a signal peptide (SP) domain, four transmembrane (TM), two extracellular (EC), and two intracellular (IC) domains. The p.Pro59Thr variant found in this paper is illustrated above the gene, and the EC1 domain is expanded in 4D with conservation demonstrated Additional missense variants in this domain are illustrated above the alignment diagram with stars, demonstrating that this is a variation hotspot within the gene.

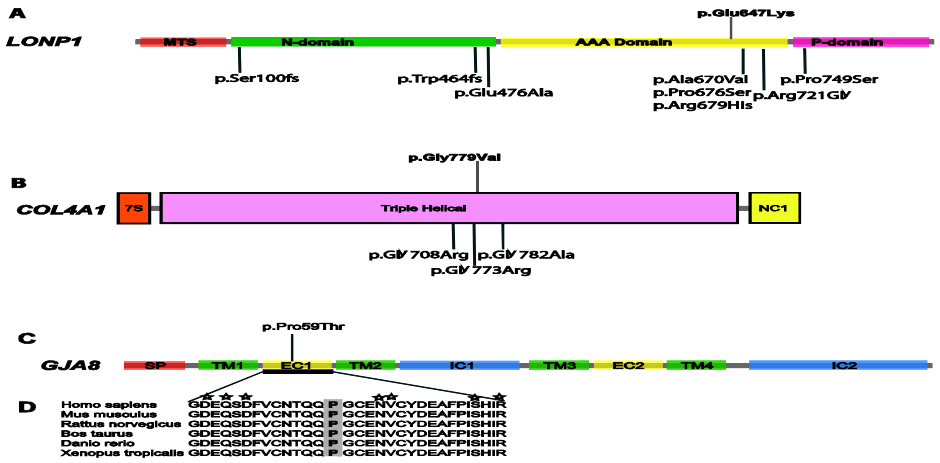
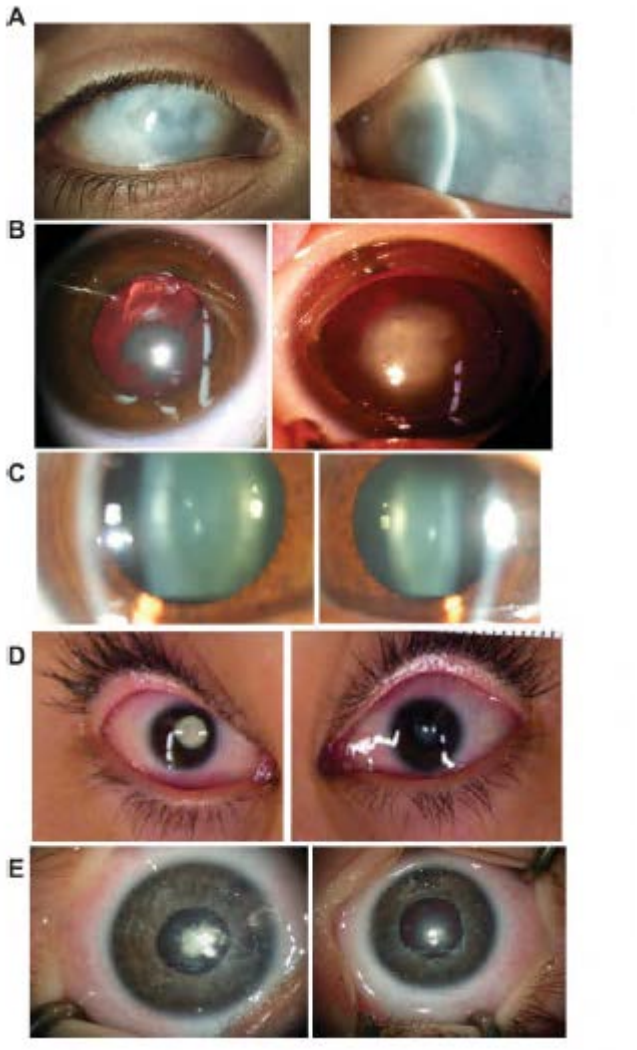


Figure 5: Clinical images of congenital cataract patients solved by GS

Clinical eye photographs of congenital cataract patients solved by GS. The right and left eyes are presented together and images from patients 2 (A) with congenital cataracts and sclerocornea, as well patients 6 (B), 1 (C), 5 (D), and 17 (E) with congenital cataracts.



Graphical Abstract

Genome sequencing in congenital cataracts leads to additional diagnoses compared to an exome approach, with a 10% increased yield in a large Australian cohort. We report an overall diagnostic rate of 77% using both exome and genome sequencing in 52 Australian congenital cataract patients.