bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Partial duplication of the *CRYBB1-CRYBA4* locus is associated with autosomal dominant congenital cataract

Running title: *CRYBB1-CRYBA4* duplication in congenital cataract

Owen M. Siggs[1,*], Shari Javadiyan[1,*], Shiwani Sharma[1], Emmanuelle Souzeau[1], Karen M. Lower[2], Deepa A. Taranath[1], Jo Black[3], John Pater[1,3], John G. Willoughby[1], Kathryn P. Burdon[4,**], Jamie E. Craig[1,**]

[1]Department of Ophthalmology, Flinders University, Flinders Medical Centre, Bedford Park, Australia.

[2]Department of Haematology and Genetic Pathology, Flinders University, Bedford Park, Australia.

[3]Department of Ophthalmology, Women's and Children's Hospital, North Adelaide, Australia.

[4]Menzies Institute for Medical Research, University of Tasmania, Hobart, Australia.

[*]joint first author

[**]joint senior author

Corresponding authors:

Owen Siggs

Department of Ophthalmology

Flinders University

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1 Flinders Drive

Bedford Park SA 5042

Phone: +61 8 8204 4899

Fax: +61 8 8204 6722

Email: sigg0005@flinders.edu.au

Kathryn Burdon

Menzies Institute for Medical Research

University of Tasmania

17 Liverpool St (Private Bag 23)

Hobart TAS 7000

Phone: +61 3 6226 4288

Email: Kathryn.burdon@utas.edu.au

Grant information: Supported by the National Health and Medical Research Council

Conflict of interest: the authors declare no conflict of interest.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Congenital cataract is a rare but severe paediatric visual impediment, often caused by mutations in one of several crystallin genes that produce the bulk of structural proteins in lens. Here we describe a pedigree with autosomal dominant isolated congenital cataract and linkage to the crystallin gene cluster on chromosome 22. No rare single nucleotide variants or short indels were identified by whole exome sequencing, yet copy number variant analysis revealed a duplication spanning both *CRYBB1* and *CRYBA4*. While the *CRYBA4* duplication was complete, the *CRYBB1* duplication was not, with one or both products predicted to create a gain of function allele. This association suggests a new genetic mechanism for the development of isolated congenital cataract.

Keywords: congenital cataract, crystallin, duplication, *CRYBA4*, *CRYBB1*

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Introduction

Cataract is an opacification of the crystalline lens and one of the leading causes of blindness worldwide [1]. Those that occur within the first year of life are categorized as congenital or infantile cataract, with an incidence in the order of 52.8 per 100,000 children [2]. Around 23% of non-syndromic congenital cataracts are familial [3], with around 50% of these associated with mutation of a crystallin gene [4].

Crystallin proteins account for more than 90% of soluble lens protein, and can be divided into α-, β- and γ-crystallin families. With age, the closely related β- and γ-crystallins gradually form insoluble aggregates, which the chaperone-like α-crystallins serve to counteract. Ten human crystallin genes are known to be mutated in congenital cataract, with mutations typically thought to reduce crystallin solubility directly (e.g. by creating an insoluble β- or γ-crystallin) or indirectly (e.g. due to loss of α-crystallin chaperone function) [5]. These genes include both α-crystallins (*CRYAA* and *CRYAB*), two acidic β-crystallins (*CRYBA1* and *CRYBA4*), three basic β-crystallins (*CRYBB1*, *CRYBB2*, and *CRYBB3*), and three γ-crystallins (*CRYGC*, *CRYGD*, and *CRYGS*) [4].

The chromosomal arrangement of human crystallin genes reflects their evolutionary history, with major clusters on chromosomes 2 and 22 [6]. Of a total of eight γ-crystallin genes, six are located on chromosome 2. Similarly, all three basic β-crystallin genes (*CRYBB1*, *CRYBB2*, and *CRYBB3*) are located on chromosome 22, with the acidic β-crystallin *CRYBA4* directly adjacent to

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CRYBB1 (but transcribed in the opposite direction). This *CRYBB1-CRYBA4* arrangement is present in organisms as distant as zebrafish, and likely significant for their coordinate regulation. Either gene can lead to congenital cataract when mutated: *CRYBA4* missense mutations are known to cause dominant cataract {Billingsley:2006fw}, while *CRYBB1* mutations may be dominant {Mackay:2002dg} or recessive {Cohen:2007ge}.

Here we describe an autosomal dominant congenital cataract pedigree associated with a unique duplication of the paired *CRYBB1-CRYBA4* locus. Both genes were found to be duplicated, with a complete duplication of *CRYBA4* and partial duplication of *CRYBB1*.

Materials & Methods

Patients

Clinical information from 19 members of pedigree CSA106 (Caucasian) was collected by referring ophthalmologists (see Table 1), with blood samples taken after informed written consent. Of these, 11 had developed bilateral cataracts. The study was approved by the Southern Adelaide Clinical Human Research Ethics Committee.

Candidate gene panel screening

Individual CSA106.03 (Figure 1A) was screened for mutations in 51 known congenital cataract genes using an Ion AmpliSeq custom amplicon panel (Life

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Technologies). Genomic DNA concentration was measured with a dsDNA high sensitivity Assay Kit on a Qubit fluorometer (Life Technologies). Library preparation (Ion AmpliSeq library kit v2.0) and template preparation (Ion PGM Template OT2 200 Kit) were performed according to the manufacturer's instructions. The clonally amplified library was enriched on the Ion OneTouch enrichment system and quantified with a Bioanalyzer 2100 using the High Sensitivity DNA Kit (Agilent Technologies). Sequencing was performed on an Ion Torrent Personal Genome Machine (PGM) using The Ion PGM Sequencing 200 Kit v2 and an Ion 318 chip (Life Technologies). Torrent Suite (v3.6) was used to align reads to the hg19 reference genome. The number of mapped reads, percentage of on-target reads, and mean read depth were calculated using the Coverage Analysis plugin (v4.0-r77897), and variants were called using the Variant Caller plugin (v4.0-r76860) with germline algorithm. Ion Reporter v4.0 was used for annotation. Variants were prioritized for further analysis if they were predicted to alter protein coding sense, were rare (MAF<0.001 in Exome Variant Server (EVS) (http://evs.gs.washington.edu/EVS/)), and were absent from unaffected controls.

Exome sequencing

A total of 11 members of the CSA106 family (6 affected, 5 unaffected) were subjected to exome sequencing, along with an additional 325 unrelated Australian cases and controls (a mixture of examined normal controls [20], cataract cases [22], keratoconus cases [51], advanced glaucoma cases [195],

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

and primary congenital glaucoma cases [37]). Genomic DNA was extracted from blood samples using a QIAamp Blood Maxi Kit (Qiagen) and subjected to exome capture (Agilent SureSelect v4). Paired-end libraries were then sequenced on an Illumina HiSeq 2000 by an external contractor (Axeq Technologies). Reads were mapped to the human reference genome (hg19) using the Burrows-Wheeler Aligner (v0.7.10), and duplicates were marked and removed using Picard (v1.126). An average of 49,429,126 reads per sample were mapped to capture regions, with a mean read depth of 83.6 and >10X coverage for 97.3% of the capture regions. Variants were called using SAMtools (v1.0) and annotated against RefSeq transcripts using ANNOVAR (2014Nov12) with additional annotations from the NHLBI Exome Sequencing Project (ESP6500SI-V2), 1000 Genomes Project [November 2014 release], ExAC (v0.3), and dbSNP138 databases. Variants were filtered by QUAL score (>20), and considered to be potentially pathogenic if predicted to alter protein coding sense (nonsynonymous, stopgain, stoploss, frameshift, essential splice), and were sufficiently rare (MAF<0.0001). *CRYBB1* variants were annotated according to RefSeq accession numbers NM_001887.3 and NP_001878.1, with *CRYBA4* variants based on NM_001886.2 and NP_001877.1.

Linkage analysis

VCF files were converted to MERLIN input format using the vcf2linkdatagen and linkdatagen scripts [7]. Parametric linkage analysis was then performed using MERLIN (v1.1.2) under a fully penetrant dominant model with a disease frequency of 0.0001.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Exome CNV analysis

Coverage depth across the critical region was extracted from exome BAM files using SAMtools. For copy number variant analysis using CoNIFER (v0.2.2), the same interval was analysed in 343 population-matched control exomes (including 11 from family CSA106) using the following parameters: SVD 5, ZRPKM 1.5.

qPCR CNV analysis

TaqMan Copy Number Assays for duplicated (Hs04088405_cn - chr22:27006444 [hg19], within intron 3) and non-duplicated (Hs00054226_cn - chr22:26995522 [hg19], within exon 6) regions of *CRYBB1* were designed and ordered from Life Technologies. All available CSA106 family members were tested for partial duplication in *CRYBB1* gene using genomic DNA according to manufacturer's protocol. Briefly, each region was amplified in four replicates on a StepOne Plus real-time PCR instrument alongside an endogenous reference gene (TaqMan Copy Number Reference Assay, human, RNase P). The CopyCaller 2.0 software (Life Technologies) was used to predict the copy number of the target genomic DNA. The bars on the graph indicate the minimum and maximum copy number (CN) calculated for the sample replicate group. A total of 118 unaffected and unrelated Caucasian controls were screened using both assays for duplicated and non-duplicated regions (in duplicates). We also screened the duplicated region in an additional 46

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

congenital cataract probands with an unidentified genetic cause (recruited under the same protocol as the CSA106 family).

Lens protein extraction

Cataractous lens material was collected from the proband during phacoemulsification (CSA106.6, aged 13 years), and stored in balanced salt solution with 2mM EDTA pH 8.0 at −80°C. Normal human lens was obtained from an 18 year-old deceased donor (Eye Bank of South Australia, Flinders Medical Centre), collection of which was approved by Southern Adelaide Clinical Human Research Ethics Committee. Lenses were homogenized in 2 mL of extraction buffer containing 50 mM imidazole (pH 7), 50 mM NaCl, 2 mM 6-aminohexanoic acid, 1 mM EDTA, and protease inhibitor cocktail (Roche Diagnostics), and ultracentrifuged at 150,000×*g* for 30 min at 4°C. The soluble fraction was acetone precipitated according to the Thermos Scientific protocol (Thermo Fisher). The insoluble fraction was dissolved in buffer containing 6 M urea, 2% dichloro-diphenyl-trichloroethane (DTT), 2% 3-[C3-cholamidoproyl] dimethyl-ammonio-1-propansulfonat (CHAPS) and 0.1% sodium dodecyl sulfate (SDS). The EZQ Protein Quantitation method (Life Technologies) was used to determine protein concentration.

Denaturing gel electrophoresis and Western blotting

Twenty micrograms of total soluble protein from each lens was size fractionated by SDS-PAGE using a 12% polyacrylamide gel. The precision plus protein standards (BioRad) were used for size comparison. Fractionated proteins were

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

transferred onto Hybond-C Extra nitrocellulose (GE Healthcare), and after blocking in 5% (wt/vol) milk in TBS-Tween (Tris Buffered Saline and 1% Tween-20) was incubated with a mouse monoclonal anti-CRYBB1 primary antibody (1:400, Sigma-Aldrich, WH0001414M3). After washing the membrane was incubated with horseradish peroxidase (HRP) -conjugated goat anti-mouse IgG (1:1000, Jackson ImmunoResearch), and following another wash was treated with Clarity Western ECL Blotting Substrate (BioRad) or Amersham ECL Prime (GE Healthcare) and imaged using an ImageQuant LAS 4000 Imager (GE Healthcare). The same membrane was stripped at 50°C in stripping buffer (100mM β-mercaptoethanol, 2% SDS and 62.5 mM Tris-HCl [pH 7]) then reprobed with polyclonal rabbit anti-CRYBA4 (1:200, Abcam, ab130680) or polyclonal sheep anti-CRYAA (1:1000; Flinders University Antibody Production Facility) primary antibodies, followed by HRP-conjugated goat anti-rabbit IgG or anti-sheep IgG secondary antibodies (1:1000, Jackson ImmunoResearch).

Results

Autosomal dominant congenital cataract

We identified a 6-generation autosomal dominant congenital cataract pedigree (Figure 1A), with affected members diagnosed with nonsyndromic bilateral cataracts between birth and 10 years of age (Table 1). Slit lamp imaging of the lens demonstrated mild to dense fetal nuclear opacification with anterior and posterior sutural involvement (Figure 1B). A custom amplicon sequencing panel, designed to sequence 51 known congenital cataract genes

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

(Supplementary Table 1), was first used to screen an affected family member (CSA106.03). A total of 807,055 reads were mapped to the reference genome, covering 93.21% of the target bases to at least 20X depth (average read depth of 623.5 across 1216 amplicons). A total of 172 variants were identified, none of which was predicted to alter protein sequence, had a mean allele frequency (MAF) of less than 0.001, and was absent in controls.

Linkage to chromosome 22

Given the absence of a candidate pathogenic variant in a known congenital cataract gene, we sequenced the exomes of 11 family members (6 affected, 5 unaffected). Parametric linkage analysis under a rare dominant inheritance model revealed a peak LOD score of 3.3 on chromosome 22 (Figure 2A,B). Haplotype phasing indicated that rs2236005 and rs2347790 were the boundaries of the 3 Mbp critical region (hg19 chr22:26422980-29414001) (Figure 2C), within which lay 17 protein-coding genes (plus 9 ncRNA and 7 pseudogenes) including the known congenital cataract genes *CRYBB1* and *CRYBA4* (Figure 2D). All 17 genes were covered by whole-exome capture sequencing, with *CRYBB1* and *CRYBA4* also having been covered by capillary sequencing previously. We detected two single-nucleotide variants that were shared between affected family members, absent from unaffected family members, and that altered coding sense (Figure 2E). Both variants had a MAF at least an order of magnitude greater than the estimated population incidence of congenital cataract (~0.000528), and therefore were considered extremely unlikely candidates. There were also no synonymous variants in *CRYBB1* or

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CRYBA4 shared between affected members: recently identified as a possible cause of crystallin misfolding [8].

Partial duplication of the CRYBB1-CRYBA4 locus

We next investigated coverage depth across the linkage interval using two techniques. Both the CoNIFER algorithm (Figure 3A) and SAMtools (Figure 3B) revealed an increase in coverage depth (and presumably also copy number) at the *CRYBB1-CRYBA4* locus of all affected individuals. Based on coverage of exons across the locus, this CNV spanned between a minimum of 28.8 kbp (chr22:26997843_27026636) and a maximum of 1.15 Mbp (chr22:26995638_28146902), encompassing only two protein-coding genes (*CRYBB1* and *CRYBA4*). The CNV was present in all six affected family members, absent from all five unaffected family members, and was absent from a further 325 unrelated Australian exomes sequenced contemporaneously. Mean coverage depth analysis revealed that while all five captured exons of *CRYBA4* appeared to be duplicated (HGNC nomenclature NM_001886.2:c.(?_-12)_(*591_?)(2), representing the final five of six total exons), only the first five exons of *CRYBB1* (of a total of six) had been duplicated (HGNC nomenclature NM_001887.3:c.(?_-19)_(575+?)(2)) (Figure 3B). Duplication of *CRYBB1* intron 3 was confirmed by qPCR in all affected family members, with a similar assay confirming that *CRYBB1* exon 6 was not duplicated (Figure 3C). *CRYBB1* and *CRYBA4* variation was also manually inspected via Integrative Genomics Viewer, given that a mutation in one *CRYBB1-CRYBA4* locus of a total of three (two alleles plus the duplicated allele) may not be called as heterozygous by

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

automated algorithms. We also screened an additional 46 unsolved congenital cataract cases for *CRYBB1* duplications, yet did not identify any further duplications (Figure 3D).

Given that exon 6 is absent, partial duplication of *CRYBB1* may lead to read-through into intron 5-6, which would immediately create an ochre (TAA) stop codon (Figure 3E). When transcribed this product would be predicted to avoid nonsense-mediated decay and create a protein lacking the final 60 C-terminal amino acids, including the fourth and final 'Greek key' crystallin domain (Figure 3F). While complete duplication of *CRYBB1* (or *CRYBA4*) could be expected to be benign, a partially duplicated *CRYBB1* may potentially act as a gain of function allele.

To examine the effects of the *CRYBB1-CRYBA4* duplication on protein expression, we prepared protein from the cataractous left lens of CSA106.06 (removed during phacoemulsifcation) and from a non-cataractous control lens. An anti-CRYAA Western blot indicated equivalent loading between the cataract and control samples, with anti-CRYBA4 and anti-CRYBB1 blots showing bands of the appropriate size (22kDa and 28kDa monomers, respectively) and similar density in both samples (Figure 4A). An additional band was detected with the anti-CRYBA4 antibody corresponding to a CRYBA4 dimer (44kDa), although no other bands were apparent. We did not detect any additional anti-CRYBB1-reactive bands in the cataract sample in soluble fractions (Figure 4A), or insoluble fractions (data not shown).

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

Here we describe an autosomal dominant congenital cataract pedigree with unambiguous linkage to chromosome 22. Despite the absence of a candidate single nucleotide variant in the region, we identified a partial duplication of the *CRYBB1-CRYBA4* locus. Although we cannot definitively ascribe this effect to the duplication, there are precedents for both *CRYBB1* and *CRYBA4* causing dominant congenital cataract.

Mutations in *CRYBA4*, for example, have been described in autosomal dominant congenital cataract [9,10]. All three reported variants are missense mutations (G64W, L69P, F94S) which presumably promote cataract formation by creating a less soluble protein. Yet the *CRYBA4* duplication described here covered the complete gene, did not contain any missense variants, and did not lead to any obvious change in protein expression. One caveat is that a 50% increase in protein expression is difficult to quantify by Western blot: a more quantitative method may involve measurement of *CRYBA4* RNA in lymphoblastoid cell lines, although these cells do not share the same transcriptional context as the developing eye.

On the other hand, *CRYBB1* mutations appear to have two distinct pathways to cataractogenesis. The recessive *CRYBB1* variants p.G57Gfs*107 or M1K are likely not expressed at all, and presumably cause cataracts by removing an

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

important structural component altogether [11,12]. Conversely the dominant alleles such as G220* [13], Q223* [14] and *253R [15] are predicted to cause cataract by disrupting the coding sequence of the final exon (exon 6), and creating a protein with reduced solubility. Although we anticipated that the truncated *CRYBB1* duplication product might also fit it into this category, we did not detect an appropriately-sized band by Western blotting.

CRYBA4 shares the same exon/intron structure as *CRYBB1* [6], so it is likely that the segmental duplication described here arose via paralogous recombination. What we have not established is the sequence surrounding the breakpoint, which would be expected to affect the genes on other side. Insertions, deletions or inversions are common at such sites, so it is likely that local transcription is altered. It might also be possible that the terminal exon of *CRYBA4* was inverted, creating a replacement exon for the partially duplicated *CRYBB1* allele. This hybrid protein would not be distinguished by Western blot, but could nevertheless act in a dominant fashion.

The absence of an additional band on Western blot could easily indicate that the duplicated CRYBB1 product was not synthesised at all, or was rapidly degraded, and therefore played no role in cataractogenesis. Given this possibility, we have not excluded the possibility that a complete duplication of *CRYBA4* was responsible for the disease, which could conceivably alter the stoichiometry of crystallin subunits at a critical stage of lens development.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

In any case, copy number variation is largely overlooked in many whole-exome and whole-genome studies, perhaps due to the limited predicted contribution of CNVs to common disease [16]. In cases where CNVs do play a role it is almost always deletions that are responsible, either in *trans* with a second deleterious allele, or by causing haploinsufficiency on their own. Increases in copy number are far less common in a disease setting, and when they do occur, they commonly involve complete genes. Ocular disease is no exception, with duplication or triplication of *TBK1* in normal tension glaucoma being one example [17,18], and a complex *NHS* triplication in the congenital cataract-associated Nance–Horan syndrome being another [19]. *TBK1* CNVs associated with glaucoma cover the entire locus, so a mechanistic explanation has not been immediately obvious. In the case of the Nance–Horan syndrome triplication, disruption of *NHS* transcription is thought to explain the phenotype, which is consistent with the loss-of-function mechanism of other *NHS* mutations [19]. In a third example, both deletions and duplications of the same gene (*FOXC1*) have been associated with anterior segment dysgenesis [20].

Other diseases can be caused by partial gene duplication [21], including ~7% of cases of the X-linked Duchenne and Becker muscular dystrophies (DMD/BMD) [22]. In the case of DMD these mutations invariably cause a frameshift, whereas in BMD the reading frame is maintained [21]. In both cases the predicted result is a loss or reduction in protein function.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Duplication has been integral to the diversification of the crystallin gene family. In the family presented here, we show that crystallin duplication can also be associated with congenital cataract. If validated experimentally, this would represent a previously undescribed genetic mechanism for the development of isolated congenital cataract, with implications for other inherited diseases that appear refractory to whole exome or whole genome sequencing.

Acknowledgements

We thank Angela Chappell for cataract photography. Supported by the National Health and Medical Research Council (NHMRC) grant 1023911. KPB is funded by an NHMRC Senior Research Fellowship, and JEC by an NHMRC Practitioner Fellowship.

References

1 Bourne RRA, Stevens GA, White RA, Smith JL, Flaxman SR, Price H *et al.* Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health* 2013; **1**: e339–49.

2 Haargaard B, Wohlfahrt J, Fledelius HC, Rosenberg T, Melbye M. Incidence and cumulative risk of childhood cataract in a cohort of 2.6 million Danish children. *Invest Ophthalmol Vis Sci* 2004; **45**: 1316–1320.

3 Haargaard B, Wohlfahrt J, Fledelius HC, Rosenberg T, Melbye M. A nationwide Danish study of 1027 cases of congenital/infantile cataracts: etiological and clinical classifications. *Ophthalmology* 2004; **111**: 2292–2298.

4 Shiels A, Bennett TM, Hejtmancik JF. Cat-Map: putting cataract on the map. *Mol Vis* 2010; **16**: 2007–2015.

5 Shiels A, Hejtmancik JF. Genetics of human cataract. *Clin Genet* 2013; **84**: 120–127.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

6 Wistow G. The human crystallin gene families. *Hum Genomics* 2012; **6**: 26.

7 Smith KR, Bromhead CJ, Hildebrand MS, Shearer AE, Lockhart PJ, Najmabadi H *et al.* Reducing the exome search space for mendelian diseases using genetic linkage analysis of exome genotypes. *Genome Biol* 2011; **12**: R85.

8 Buhr F, Jha S, Thommen M, Mittelstaet J, Kutz F, Schwalbe H *et al.* Synonymous Codons Direct Cotranslational Folding toward Different Protein Conformations. *Mol Cell* 2016; **61**: 341–351.

9 Billingsley G, Santhiya ST, Paterson AD, Ogata K, Wodak S, Hosseini SM *et al.* CRYBA4, a novel human cataract gene, is also involved in microphthalmia. *Am J Hum Genet* 2006; **79**: 702–709.

10 Zhou G, Zhou N, Hu S, Zhao L, Zhang C, Qi Y. A missense mutation in CRYBA4 associated with congenital cataract and microcornea. *Mol Vis* 2010; **16**: 1019–1024.

11 Cohen D, Bar-Yosef U, Levy J, Gradstein L, Belfair N, Ofir R *et al.* Homozygous CRYBB1 deletion mutation underlies autosomal recessive congenital cataract. *Invest Ophthalmol Vis Sci* 2007; **48**: 2208–2213.

12 Meyer E, Rahman F, Owens J, Pasha S, Morgan NV, Trembath RC *et al.* Initiation codon mutation in betaB1-crystallin (CRYBB1) associated with autosomal recessive nuclear pulverulent cataract. *Mol Vis* 2009; **15**: 1014–1019.

13 Mackay DS, Boskovska OB, Knopf HLS, Lampi KJ, Shiels A. A nonsense mutation in CRYBB1 associated with autosomal dominant cataract linked to human chromosome 22q. *Am J Hum Genet* 2002; **71**: 1216–1221.

14 Yang J, Zhu Y, Gu F, He X, Cao Z, Li X *et al.* A novel nonsense mutation in CRYBB1 associated with autosomal dominant congenital cataract. *Mol Vis* 2008; **14**: 727–731.

15 Willoughby CE, Shafiq A, Ferrini W, Chan LLY, Billingsley G, Priston M *et al.* CRYBB1 mutation associated with congenital cataract and microcornea. *Mol Vis* 2005; **11**: 587–593.

16 Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* 2010; **464**: 704–712.

17 Awadalla MS, Fingert JH, Roos BE, Chen S, Holmes R, Graham SL *et al.* Copy Number Variations of TBK1 in Australian Patients With Primary Open-Angle Glaucoma. *Am J Ophthalmol* 2015; **159**: 124–130.e1.

18 Fingert JH, Robin AL, Stone JL, Roos BR, Davis LK, Scheetz TE *et al.* Copy number variations on chromosome 12q14 in patients with normal

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

tension glaucoma. *Hum Mol Genet* 2011; **20**: 2482–2494.

19 Coccia M, Brooks SP, Webb TR, Christodoulou K, Wozniak IO, Murday V *et al.* X-linked cataract and Nance-Horan syndrome are allelic disorders. *Hum Mol Genet* 2009; **18**: 2643–2655.

20 Nishimura DY, Searby CC, Alward WL, Walton D, Craig JE, Mackey DA *et al.* A spectrum of FOXC1 mutations suggests gene dosage as a mechanism for developmental defects of the anterior chamber of the eye. *Am J Hum Genet* 2001; **68**: 364–372.

21 Hu X, Worton RG. Partial gene duplication as a cause of human disease. *Hum Mutat* 1992; **1**: 3–12.

22 White SJ, Aartsma-Rus A, Flanigan KM, Weiss RB, Kneppers ALJ, Lalic T *et al.* Duplications in the DMD gene. *Hum Mutat* 2006; **27**: 938–945.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Titles and legends to figures

Figure 1. An autosomal dominant congenital cataract pedigree

(A) CSA106 pedigree, indicating affected (black) and unaffected (white) members. Proband (CSA106.02) is indicated by an arrow.

(B) Direct illumination (top panels) or retroillumination (bottom panels) of the same cataract from four family members, demonstrating mild to dense fetal nuclear opacification with sutural involvement. LE, left eye; RE, right eye.

Figure 2. Linkage of the cataract phenotype to proximal chromosome 22q

(A) Genome-wide LOD scores under a fully penetrant dominant inheritance model.

(B) LOD scores across chromosome 22.

(C) Haplotype analysis and critical recombinants across the interval.

(D) List of genes within the defined linkage interval. Previously known congenital cataract genes are highlighted.

(E) Filtering strategy to identify shared heterozygous coding variants within linkage interval.

Figure 3. Partial duplication of the CRYBB1-CRYBA4 locus

(A) CoNIFER output indicating a copy number gain within the linkage interval.

(B) Average read depth of affected (red) and unaffected (black) family members across the same interval shown in (A). Each peak represents a region covered by exome capture and sequencing, typically an exon. All five coding exons of

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

both *CRYBB1* and *CRYBA4* were covered, with an additional region captured and sequenced for *CRYBB1*. The final peak at the *CRYBB1* locus (i.e. furthest to left) represents the sixth and final exon.

(C) qPCR validation of a duplicated (intron 3) and non-duplicated (exon 6) region of *CRYBB1*.

(D) *CRYBB1* duplication screening in unsolved congenital cataract cases.

(E) Exon structure of *CRYBB1* and predicted translational consequences of read-through into intron 5-6.

(F) Domain structure of protein products of full-length *CRYBB1*, and *CRYBB1* lacking exon 6 ($CRYBB1^{e1-5}$)

Figure 4. Crystallin protein expression in cataractous lens extracts

(A) Control [C] and patient [P] lens extracts were subjected to Western blotting with the indicated antibodies.

bioRxiv preprint doi: https://doi.org/10.1101/042515; this version posted March 5, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Table 1: Clinical details of CSA106 family members. All affected members had bilateral cataracts. y, years; LE, left eye; RE, right eye; NA, not available.

ID	Age diagnosed (y)	Exome sequenced	Disease Status	Age of surgery (y) (RE)	Age of surgery (y) (LE)
CSA106.01	5	Yes	Affected	76	76
CSA106.02	0	Yes	Affected	44	44
CSA106.03	0	Yes	Affected	3	22
CSA106.04	3	Yes	Affected	18	19
CSA106.05	-	Yes	Unaffected	-	-
CSA106.06	4	No	Affected	7	13
CSA106.07	0	No	Affected	NA	NA
CSA106.08	-	Yes	Unaffected	-	-
CSA106.09	-	Yes	Unaffected	-	-
CSA106.10	-	No	Unaffected	-	-
CSA106.11	-	No	Unaffected	-	-
CSA106.12	NA	No	Affected	NA	NA
CSA106.13	NA	No	Affected	NA	NA
CSA106.14	10	Yes	Affected	51	10
CSA106.15	-	Yes	Unaffected	-	-
CSA106.16	NA	No	Affected	24	NA
CSA106.17	-	Yes	Unaffected	-	-
CSA106.18	-	No	Unaffected	-	-
CSA106.19	NA	Yes	Affected	NA	NA

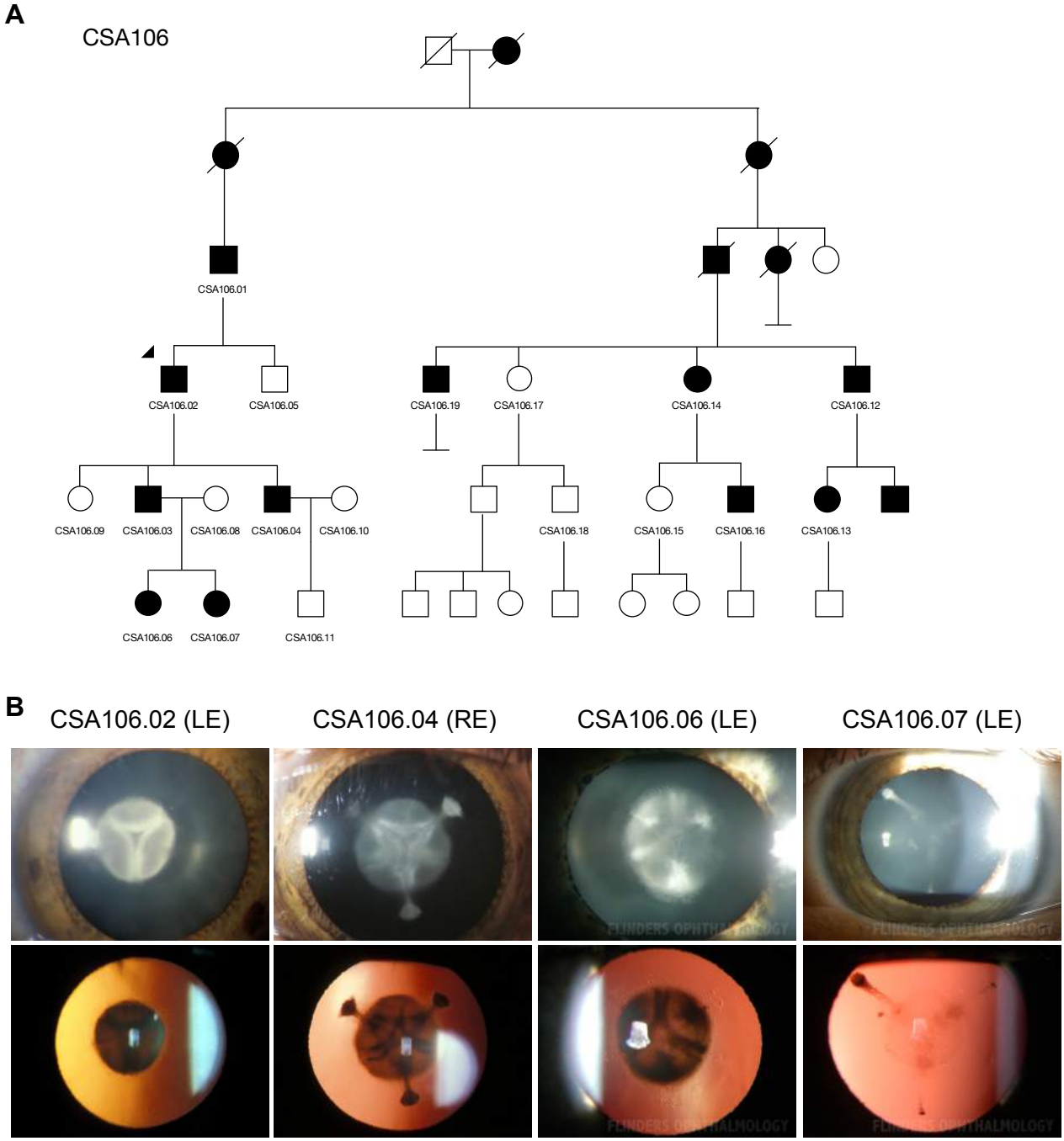
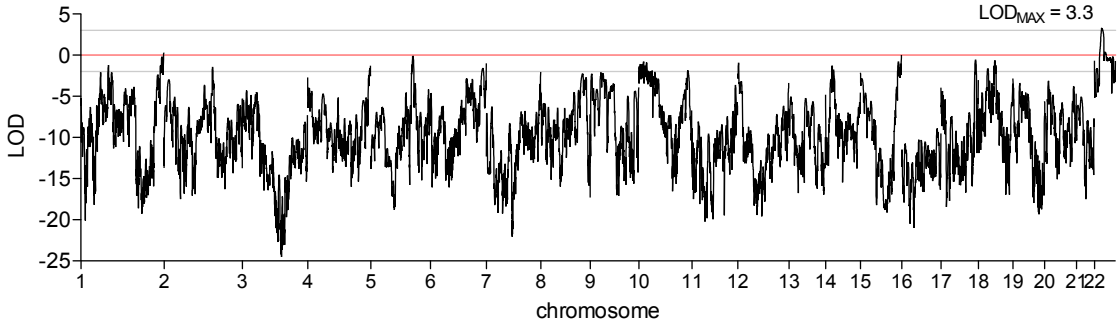
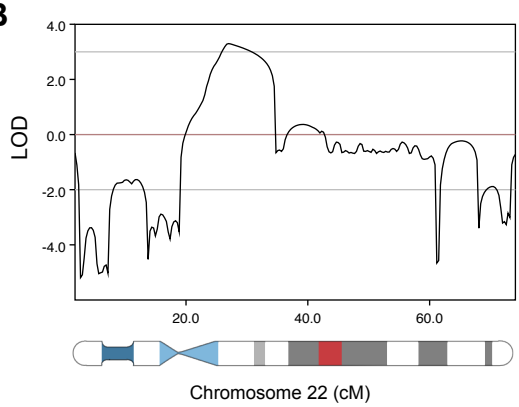

A
CSA106
CSA106.01
CSA106.02
CSA106.05
CSA106.19
CSA106.17
CSA106.14
CSA106.12
CSA106.09
CSA106.03
CSA106.08
CSA106.04
CSA106.10
CSA106.18
CSA106.15
CSA106.16
CSA106.13
CSA106.06
CSA106.07
CSA106.11
B
CSA106.02 (LE)
CSA106.04 (RE)
CSA106.06 (LE)
CSA106.07 (LE)

Figure 1

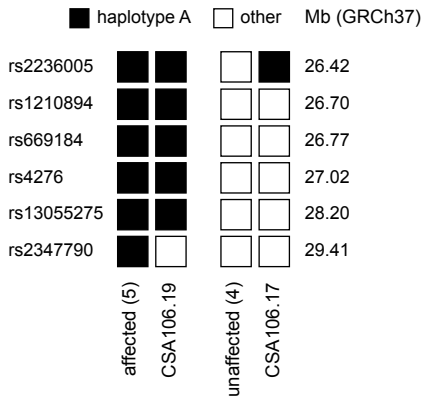
A



B



C



D

symbol	gene	chr	start	end
MYO18B	myosin XVIIIB	22	26138117	26453345
SEZ6L	seizure related 6 homolog (mouse)-like	22	26565440	26779563
ASPHD2	aspartate beta-hydroxylase domain containing 2	22	26825280	26840978
HPS4	Hermansky-Pudlak syndrome 4	22	26846848	26879829
SRRD	SRR1 domain containing	22	26879850	26887904
TFIP11	tuftelin interacting protein 11	22	26887893	26908462
TPST2	tyrosylprotein sulfotransferase 2	22	26921714	26986089
CRYBB1	crystallin, beta B1	22	26995362	27013991
CRYBA4	crystallin, beta A4	22	27017928	27026636
MN1	meningioma (disrupted in balanced translocation) 1	22	28144265	28197486
PITPNB	phosphatidylinositol transfer protein, beta	22	28247657	28315255
TTC28	tetratricopeptide repeat domain 28	22	28374002	29075853
CHEK2	checkpoint kinase 2	22	29083731	29137822
HSCB	HscB mitochondrial iron-sulfur cluster co-chaperone	22	29138008	29153506
CCDC117	coiled-coil domain containing 117	22	29168662	29185283
XBP1	X-box binding protein 1	22	29190548	29196560
ZNRF3	zinc and ring finger 3	22	29279755	29453476

E

total variants	45,842	
shared heterozygous	563	
absent from unaffected	5	
nonsyn/stop/frameshift/splice	3	
chr22:26422980-29414001	2	*SEZ6L* p.W185L (ExAC MAF:0.04381) *HPS4* p.I84V (ExAC MAF:0.004654)

Figure 2

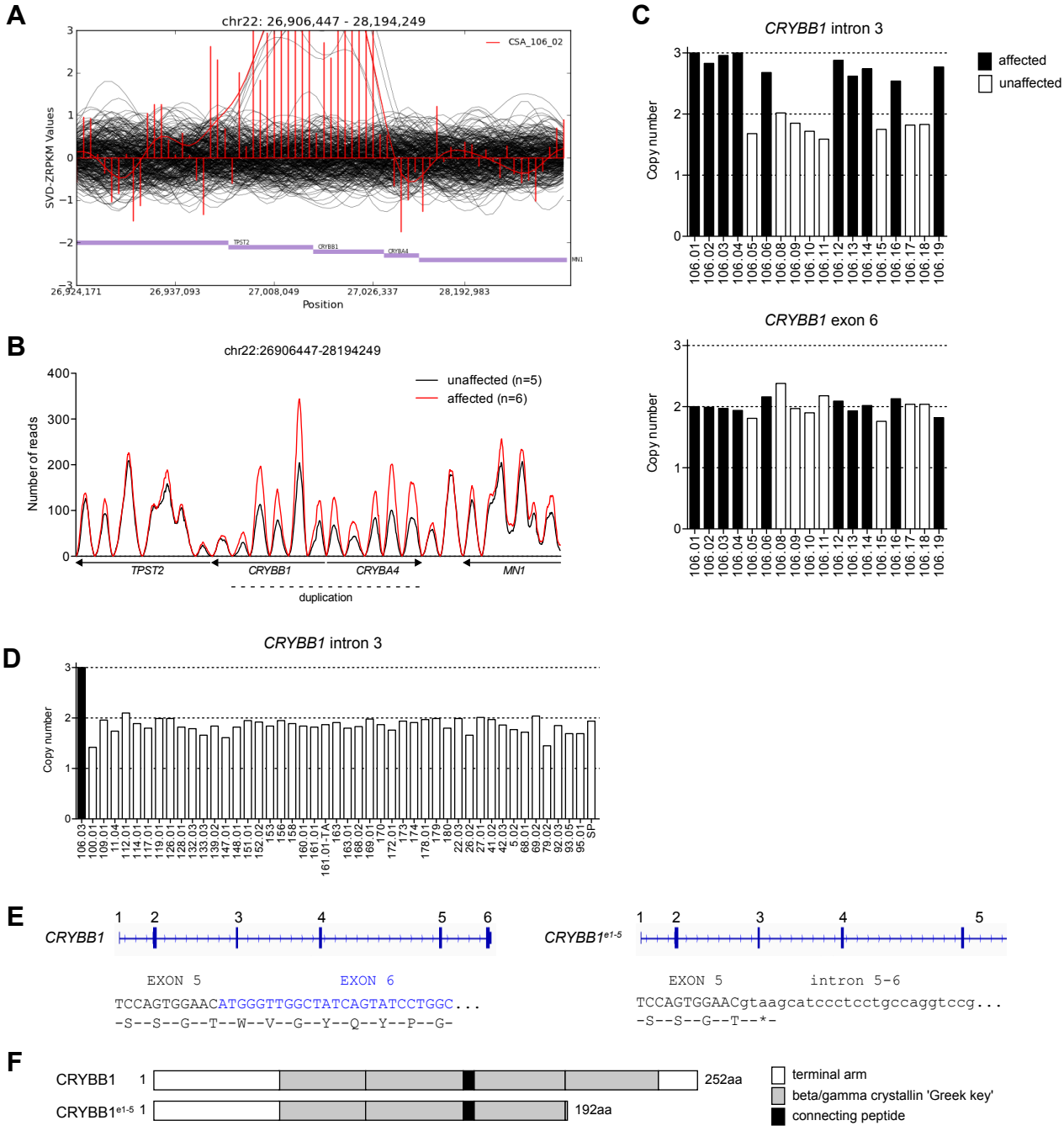

A
chr22: 26,906,447 - 28,194,249
CSA_106_02
SVD-ZRPKM Values
TPST2
CRYBB1
CRYBA4
MN1
26,924,171
26,937,093
27,008,049
27,026,337
28,192,983
Position
B
chr22:26906447-28194249
unaffected (n=5)
affected (n=6)
Number of reads
TPST2
CRYBB1
CRYBA4
MN1
duplication
C
CRYBB1 intron 3
affected
unaffected
Copy number
CRYBB1 exon 6
Copy number
D
CRYBB1 intron 3
Copy number
E
CRYBB1
EXON 5
EXON 6
TCCAGTGGAACATGGGTTGGCTATCAGTATCCTGGC...
-S--S--G--T--W--V--G--Y--Q--Y--P--G-
CRYBB1e1-5
EXON 5
intron 5-6
TCCAGTGGAACgtaagcatccctcctgccaggtccg...
-S--S--G--T--*-
F
CRYBB1
252aa
CRYBB1e1-5
192aa
terminal arm
beta/gamma crystallin 'Greek key'
connecting peptide

Figure 3

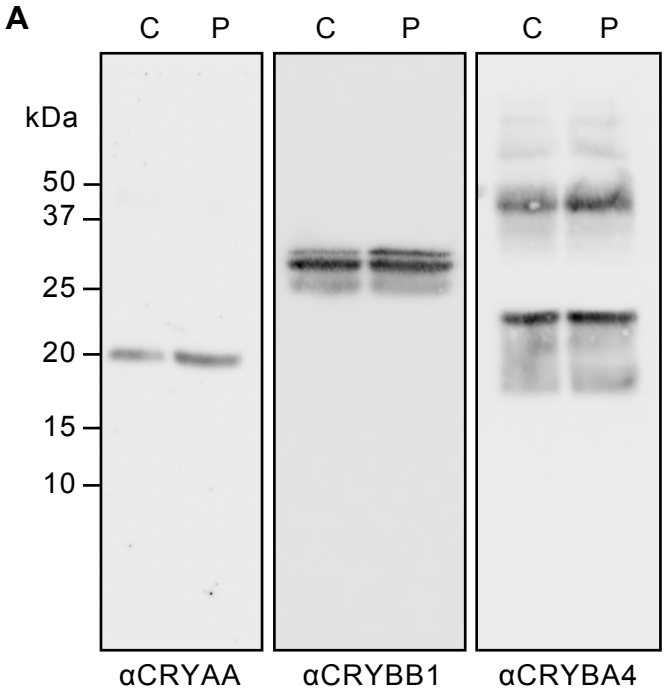


Figure 4