Molecular Vision 2014; 20:1579-1593 <http://www.molvis.org/molvis/v20/1579>
Received 27 May 2014 | Accepted 4 November 2014 | Published 6 November 2014

Congenital cataracts: *de novo* gene conversion event in *CRYBB2*

Sarah J. Garnai,[1] Jeroen R. Huyghe,[2] David M. Reed,[1] Kathleen M. Scott,[1] Jeffrey M. Liebmann,[4,5] Michael Boehnke,[2] Julia E. Richards,[1,3] Robert Ritch,[4] Hemant Pawar[1]

[1]Department of Ophthalmology and Visual Sciences, University of Michigan, Ann Arbor, MI; [2]Department of Biostatistics, University of Michigan, Ann Arbor, MI; [3]Department of Epidemiology, University of Michigan, Ann Arbor, MI; [4]Einhorn Clinical Research Center, New York Eye and Ear Infirmary at Mount Sinai School of Medicine, New York, NY; [5]New York University School of Medicine, New York, NY

Purpose: To identify the cause of congenital cataracts in a consanguineous family of Ashkenazi Jewish ancestry.
Methods: We performed genome-wide linkage analysis and whole-exome sequencing for the initial discovery of variants, and we confirmed the variants using gene-specific primers and Sanger sequencing.
Results: We found significant evidence of linkage to chromosome 22, under an autosomal dominant inheritance model, with a maximum logarithm of the odds (LOD) score of 3.91 (16.918 to 25.641 Mb). Exome sequencing identified three nonsynonymous changes in the *CRYBB2* exon 5 coding sequence that are consistent with the sequence of the corresponding region of the pseudogene *CRYBB2P1.* The identification of these changes was complicated by possible mismapping of some mutated *CRYBB2* sequences to *CRYBB2P1.* Sequencing with gene-specific primers confirmed that the changes—rs2330991, c.433 C>T (p.R145W); rs2330992, c.440A>G (p.Q147R); and rs4049504, c.449C>T (p.T150M)—present in all ten affected family members are located in *CRYBB2* and are not artifacts of cross-reaction with *CRYBB2P1.* We did not find these changes in six unaffected family members, including the unaffected grandfather who contributed the affected haplotype, nor did we find them in the 100 Ashkenazi Jewish controls.
Conclusions: Our data are consistent with a de novo gene conversion event, transferring 270 base pairs at most from *CRYBB2P1* to exon 5 of *CRYBB2.* This study highlights how linkage mapping can be complicated by *de novo* mutation events, as well as how sequence-analysis pipeline mapping of short reads from next-generation sequencing can be complicated by the existence of pseudogenes or other highly homologous sequences.

Congenital cataracts (OMIM 601547) are opacities of the crystalline lens that appear in the first year of life and affect between 1 and 3 in 10,000 births in industrialized countries [1]. Congenital cataracts can occur as a simple ocular trait or as part of a multisystem disorder. The most common mode of inheritance is autosomal dominant, but other modes of inheritance have been reported [2].

More than 200 loci or genes have been associated with cataracts (Cat-Map) [3], so far including about 45 different loci and 38 cataract genes that are involved in nonsyndromic forms of cataracts [4]. Crystallins are involved in about half the families with known mutations [2]. Crystallin stability and order are critical to the transparency of the lens [5]. Mutations in crystallins that are severe enough to cause aggregation can lead to congenital cataracts in a highly penetrant Mendelian manner, while mutations that merely increase susceptibility to environmental influences can contribute to age-related cataracts in a multifactorial manner [2]. Other important categories of cataract genes include connexins, membrane proteins, beaded filament proteins, and growth and transcription factors [5]. The same mutation, either within a family [6-9] or in different families [6-12], can result in different morphologies and severities of the cataracts, while mutations in completely different genes [13] can cause cataracts that appear clinically similar [5].

Here, we report the mapping of a congenital cataract locus in a consanguineous Ashkenazi Jewish family and demonstrate that the gene *CRYBB2* (OMIM 123620) has been altered in ways that have been predicted to have unfavorable effects on its protein product, βB2-crystallin. We discuss a probable transfer of information from the pseudogene *CRYBB2P1* (OMIM 123620) to the active gene *CRYBB2*, implying that a gene conversion event covering a region of less than 271 bp has occurred. The transfer of information from *CRYBB2P1* to *CRYBB2* has implications for the development of *CRYBB2* mutation screening programs and raises questions about the rate at which multiple sequence variants are introduced into the *CRYBB2* gene.

METHODS

Sample collection and clinical examination: We recruited 16 individuals from three generations of family 581 (Figure 1) for this study after obtaining informed consent according to

Correspondence to: Julia E. Richards, University of Michigan, Ophthalmology and Visual Sciences, 1000 Wall St., Ann Arbor, MI 48105; Phone: (734) 936-8966; FAX: (734) 615-0542; email: richj@med.umich.edu

a protocol approved by the Institutional Review Board of the University of Michigan and in accordance with the tenets of the Declaration of Helsinki. Participants underwent ocular examinations at the New York Eye and Ear Infirmary. We extracted genomic DNA from peripheral blood using the Gentra Puregene Blood Kit (QIAgen, Valencia, CA). The Ashkenazi Jewish control DNAs consisted of 90 samples from Tel Aviv University and 10 samples from the Coriell Institute (Camden, NJ). As shown in Figure 1, the family is consanguineous. The family history indicates that V:4 came from a different European country than the rest of the family, suggesting that V:4 is not closely related to his wife. Assuming complete penetrance and V:4 being unaffected, simulation via FastSLINK [14,15] indicated that this family had powers of 88.4% and 88.1% to detect a logarithm of the odds (LOD) score greater than 3 under dominant and recessive inheritance models, respectively (based on 10,000 replications).

Linkage analysis: We performed genome-wide linkage analysis on the 16 family members using single-nucleotide polymorphism (SNP) data from the Human Omni1-Quad v1.0 DNA BeadChip (Illumina, San Diego, CA). To obtain the most accurate SNP positions and to detect problematic SNPs, Illumina probe sequences were mapped to the hg19 genome assembly using Burrows-Wheeler Aligner (BWA) [16]. This led to the exclusion of 121,108 SNPs due to alignment problems including no alignment, multiple alignments, or the presence of known variants near the 3′ ends of probes; 1,013,406 SNPs remained for analysis after this step. We used PLINK [17] for quality control of the genotype data. The sample genotype call rates were >95% for each sample, indicating no evidence of genotyping issues. We confirmed sex based on the homozygosity rates of the X chromosome and we verified familial relationships by plotting estimates of the genome-wide proportion of SNPs sharing one allele based on identity by descent (IBD = 1) versus estimates of the genome-wide proportion of SNPs with no alleles IBD (IBD = 0) for all pairwise comparisons between subjects.

Starting from the set of 1,013,406 SNPs with confidently mapped probe sequences, we obtained a set of SNPs informative for linkage by sequentially applying the following filtering steps: SNPs with any missing genotypes (95,633 SNPs), SNPs with <8 copies of the minor allele in the 16 genotyped subjects (minor allele frequency <25%; 606,442

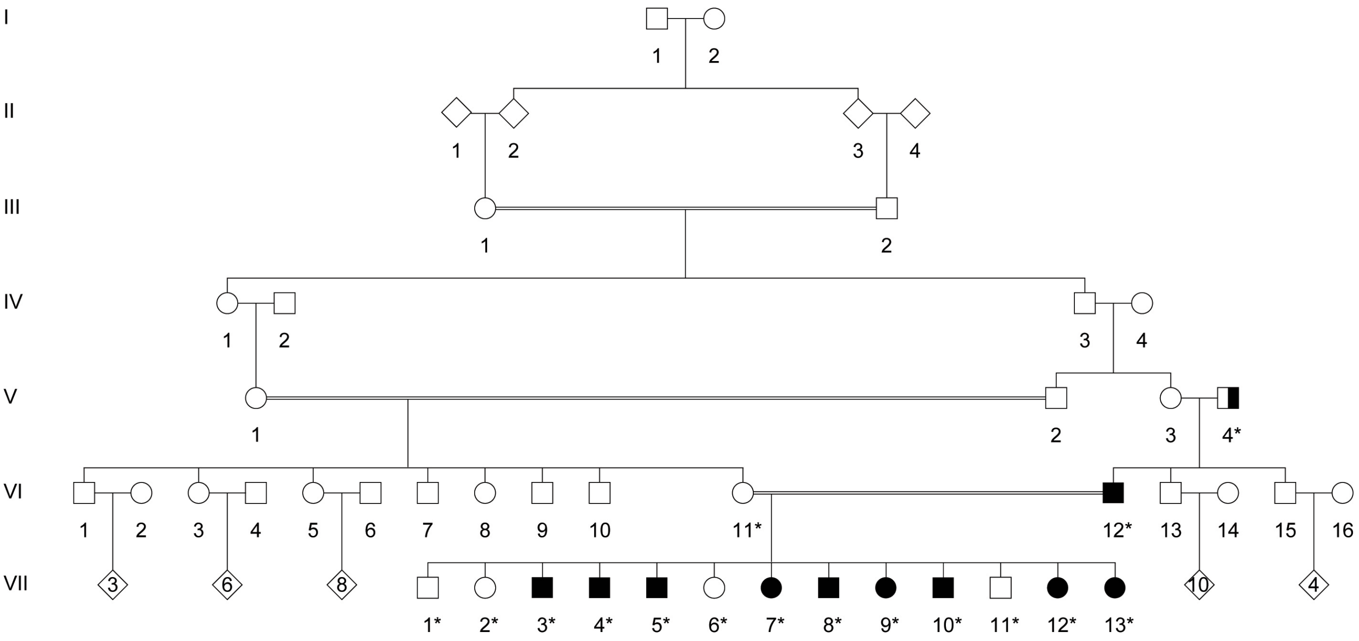


Figure 1. Family 581 pedigree. The filled symbols indicate the individuals affected with congenital cataracts; the half-filled symbols indicate the individuals affected with senile cataracts. The circles indicate females and the squares indicate males. The numbers inside of the diamonds indicate the number of children in that sibship. The individuals genotyped for this study are marked with an asterisk. For simulations and analyses, we only considered the genotyped individuals and one untyped founder (individual V:3) who was unavailable for genotyping. While it is reasonable to presume that the older members of the pedigree have passed away, we do not have clear information regarding which members of the earlier generations are still alive; therefore, the symbols have not been modified to indicate deceased status. The consanguinity indicated by these earlier generations has been included in the figure to help indicate where the concept of autosomal recessive inheritance originated.

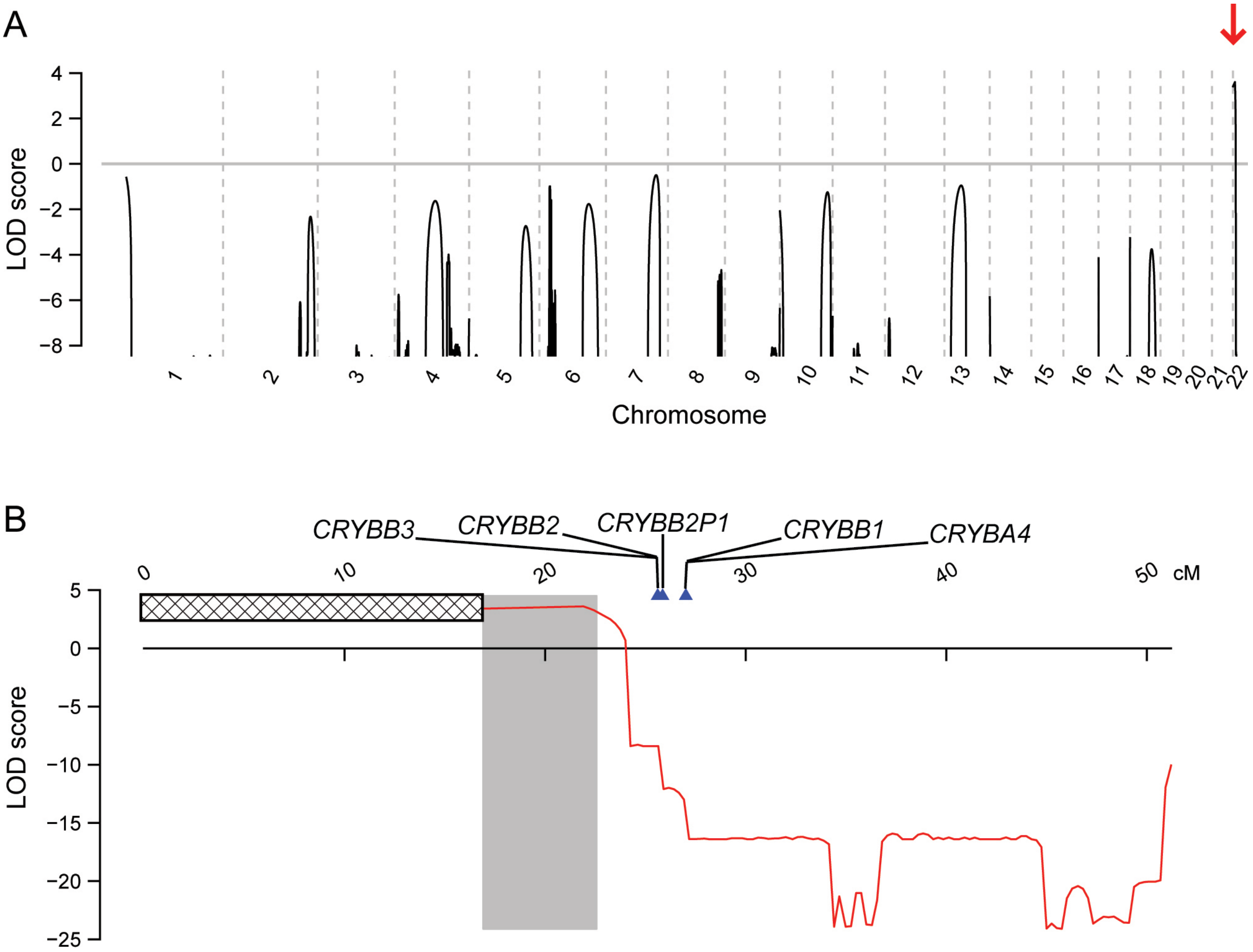


Figure 2. Multipoint logarithm of the odds (LOD) score plots for **A**: the entire genome and **B**: chromosome 22 under the autosomal recessive model with the phenotype of V:4 set to missing. On the genome-wide plot (**A**), the region corresponding to the maximum LOD score is marked with a red arrow. The blue triangles on the chromosome 22 plot (**B**) mark the positions of the crystallin genes. Note that in **B**, *CRYBB2* and *CRYBB3* are outside of the genetic inclusion interval. The hatched bar in **B** marks the p-arm repetitive region that is not represented among the SNPs in the panel that was screened.

SNPs), and SNPs showing Mendelian inconsistencies identified using Pedstats [18] (5,470 SNPs) were excluded from further analyses. These steps resulted in a final set of 305,861 SNPs.

Next, we divided the genome into segments of 250 kb and selected the first SNP in every segment that was located at least 200 kb away from the previous adjacent SNP, yielding a map of 10,223 autosomal SNPs. Because linkage disequilibrium (LD) may lead to upward biases in multipoint linkage analysis, the analysis was also performed using a second set of 2,660 autosomal SNPs with intermarker distances of approximately 1 Mb (minimum distance between adjacent SNPs was 750 kb). We obtained the same conclusions showing that LD did not impact our results, and we only discuss the results for the denser 250 kb map.

Prior to linkage analysis, we identified possible remaining genotyping errors by looking for unlikely double recombinants using Merlin's [19] error option, and we removed 29 (0.02%) and 26 (0.06%) genotypes flagged as unlikely for the 250 kb and 1 Mb maps, respectively.

We performed a multipoint parametric linkage analysis using Merlin [19] under both autosomal dominant and autosomal recessive models of inheritance, assuming a disease allele frequency of 0.0001. To assess the robustness of the results to model parameters, analyses were repeated with varying penetrance parameters. For these analyses, we

Table 1. Prioritization of non-synonymous variants in the linkage interval found by whole-exome sequencing in transcripts that are in Ensembl release 74.

AD or AR	Chr 22 Position	dbSNP132	Alleles		Bases for prioritization					Gene/Locus
					Evolutionary conservation		Allele frequency		Protein prediction	
			Ref	Alt	PhyloP	GERP	ESP	1000G	PolyPhen	
AD/AR	18,354,783	rs5992128	T	G	0.594	**2.19**	0.024825	0.02	benign	MICAL3|XXbac-B461K10.4
AD/AR	19,471,506	rs9606030	A	G	0.226	1.73	0.001852	0.0014	benign	CDC45
AD/AR	22,049,783	rs12484060	C	T	0.694	1.82	**0.337418**	**0.39**	benign	PPIL2
AD/AR	22,318,354	rs75602167	T	C	1.096	**2.79**	0.02151	0.01	benign	TOP3B
AD	24,035,970	rs2070446	C	T	0.167	1.45	**0.705556**	**0.76**	benign	RGL4
AD	24,038,847	rs1007298	T	C	−0.18	−1.35	**0.689513**	**0.75**	benign	RGL4
AD	24,179,922	rs3177243	G	C	**5.424**	**3.92**	**0.15812**	**0.15**	benign	DERL3
AD	24,199,704	rs16986337	C	T	0.28	−0.222	0.032479	**0.08**	possibly damaging	SLC2A11
AD	24,226,890	rs60699980	G	A	**2.124**	1.05	0.02963	0.03	probably damaging	SLC2A11

Chr=chromosome, SNP=single-nucleotide polymorphism, Ref=reference allele, Alt=alternate allele, PhyloP=phylogenetic p value, GERP=genomic evolutionary rate profiling, ESP=NHLBI Exome Sequencing Project, 1000G=1000 Genomes Project, AR=autosomal recessive, AD=autosomal dominant, bold indicates values suggesting reduced priority as a candidate for further evaluation. PolyPhen result is from Ensembl. Chromosome 22 positions are reported using hg19/GRCh37 coordinates.

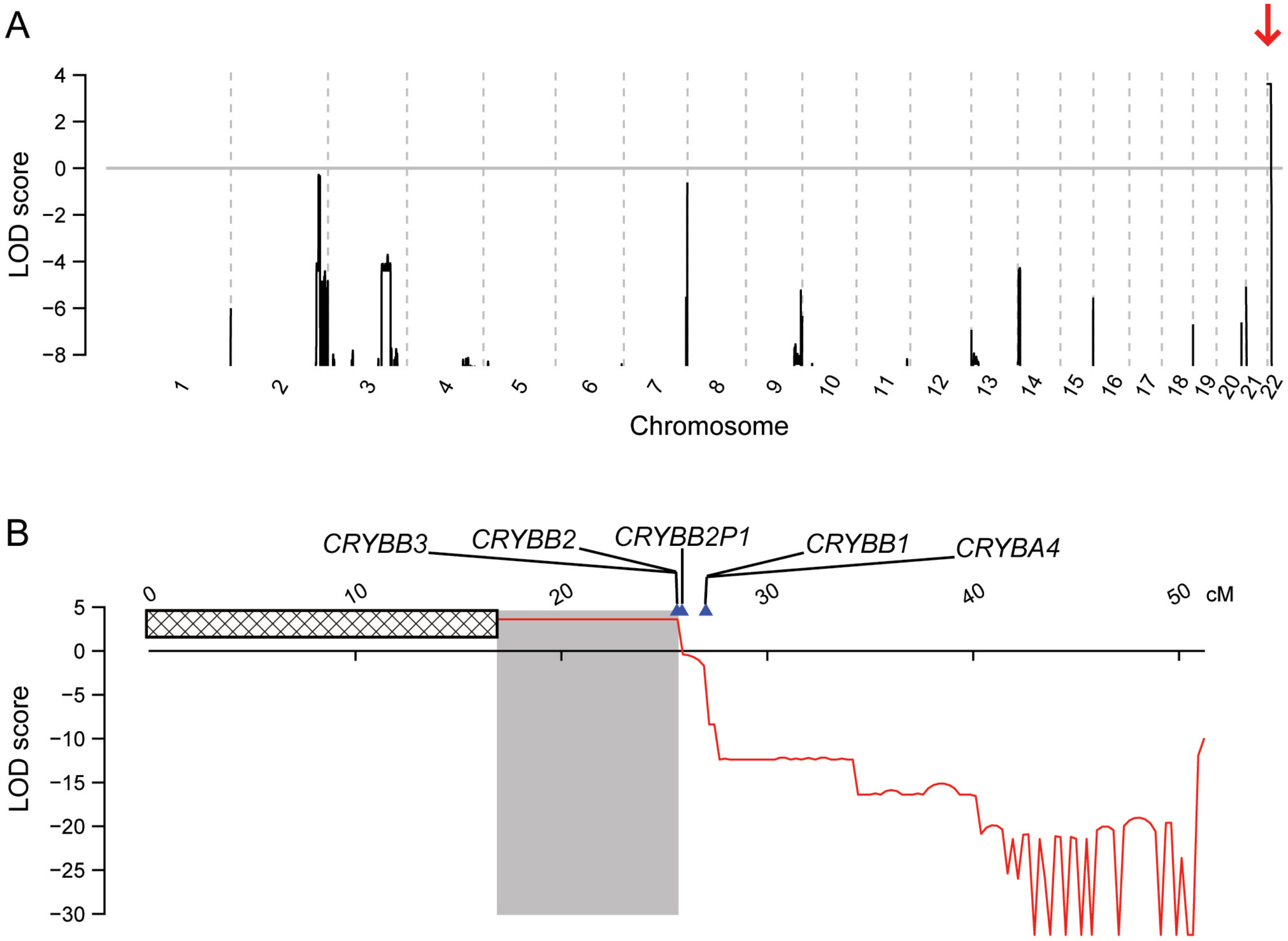
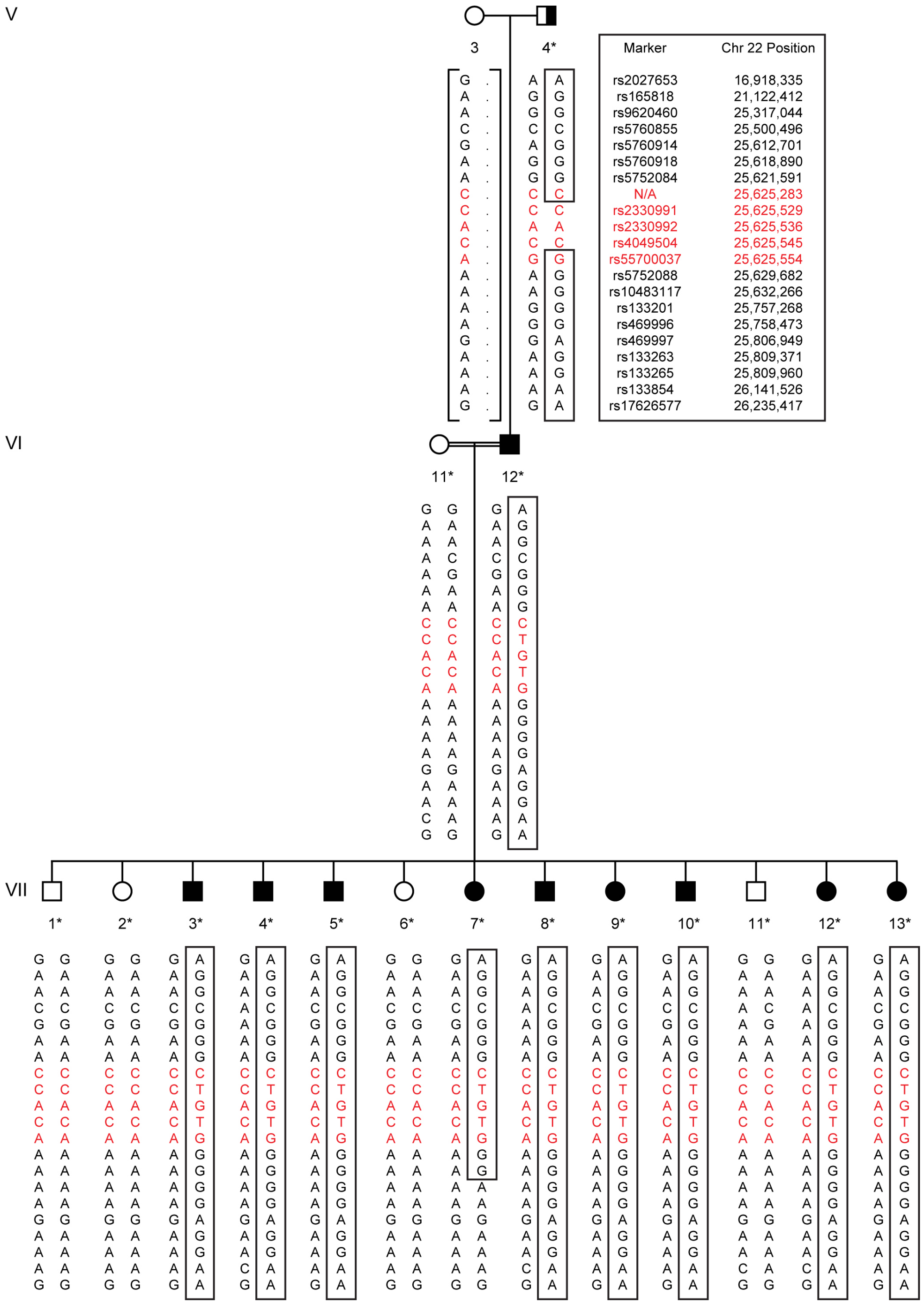


Figure 3. Multipoint LOD score plots for **A**: the entire genome and **B**: chromosome 22 under the autosomal dominant model with the phenotype of V:4 set to missing. On the genome-wide plot (**A**), the region corresponding to the maximum LOD score is marked with a red arrow. The blue triangles on the chromosome 22 plot (**B**) mark the positions of the crystallin genes. Note that in **B**, *CRYBB2* and *CRYBB3* are inside of the genetic inclusion interval. The hatched bar in **B** marks the p-arm repetitive region that is not represented among the SNPs in the panel that was screened.

approximated the genetic map by 1 cM to 1 Mb. We also performed homozygosity mapping to detect possible microdeletions. All genomic positions are reported using hg19/GRCh37 coordinates.

Whole-exome sequencing: Whole-exome sequencing was performed at the University of Michigan DNA Sequencing Core using DNA samples from three affected (VI:12, VII:3, and VII:10) and three unaffected (VI:11, VII:2, and VII:6) family members. Sequencing was done with the TruSeq Exome Enrichment Kit (Illumina) and an Illumina HiSeq 2000 sequencer, using 100×100 paired-end reads. We performed sequence alignment using BWA [16] and variant calling using SAMtools/BCFtools [20]. We excluded from further analysis the variants that were common (minor allele frequency greater than 5%), and we considered other nonsynonymous variants in transcripts in Ensembl Release 74 [21] to be of reduced interest based on phylogenetic p values (PhyloP) [22] and genetic evolutionary rate profiling (GERP) [23]. For the one variant that passed this filtering, we used PolyPhen-2 [24] to determine that the SNP was not predicted to have deleterious effects on the protein. We annotated variants using ANNOVAR [25] and the UCSC Known Gene database [26]. We inspected mapped data using the Integrated Genomics Viewer [27,28], and we examined visually the assembled sequence for regions including *CRYBB2* and *CRYBB3* (OMIM 123630).

PCR amplification and Sanger sequencing: We amplified a 753-bp region containing exon 5 of *CRYBB2* by polymerase


V
3
G A A C G A A C C A C A A A A A G A A A G
. . . .
4*
A G G C A G G C C A C G A A G G G A A A G
A G G C G G G C C A C G G G G G A G G A A
VI
11*
G A A A A A A C C A C A A A A A G A A C G
G A A C G A A C C A C A A A A A G A A A G
12*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A
VII
1*
G A A C G A A C C A C A A A A A G A A A G
G A A C G A A C C A C A A A A A G A A A G
2*
G A A C G A A C C A C A A A A A G A A A G
G A A C G A A C C A C A A A A A G A A A G
3*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A
4*
G A A A A A A C C A C A A A A A G A A C G
A G G C G G G C T G T G G G G G A G G A A
5*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A
6*
G A A C G A A C C A C A A A A A G A A A G
G A A C G A A C C A C A A A A A G A A A G
7*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G A A G A A A G
8*
G A A A A A A C C A C A A A A A G A A C G
A G G C G G G C T G T G G G G G A G G A A
9*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A
10*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A
11*
G A A A A A A C C A C A A A A A G A A C G
G A A C G A A C C A C A A A A A G A A A G
12*
G A A A A A A C C A C A A A A A G A A C G
A G G C G G G C T G T G G G G G A G G A A
13*
G A A C G A A C C A C A A A A A G A A A G
A G G C G G G C T G T G G G G G A G G A A

Marker	Chr 22 Position
rs2027653	16,918,335
rs165818	21,122,412
rs9620460	25,317,044
rs5760855	25,500,496
rs5760914	25,612,701
rs5760918	25,618,890
rs5752084	25,621,591
N/A	25,625,283
rs2330991	25,625,529
rs2330992	25,625,536
rs4049504	25,625,545
rs55700037	25,625,554
rs5752088	25,629,682
rs10483117	25,632,266
rs133201	25,757,268
rs469996	25,758,473
rs469997	25,806,949
rs133263	25,809,371
rs133265	25,809,960
rs133854	26,141,526
rs17626577	26,235,417

Figure 4. Family 581 haplotypes from SNP chip data (black lettering) and Sanger sequencing using gene-specific primers (red lettering). The affected haplotype is boxed. The symbol definitions are the same for Figure 1. The inferred haplotype is shown for V:3 (not genotyped). More than 100 markers were analyzed to generate these haplotypes. The chromosome 22 positions are reported using hg19/GRCh37 coordinates.

chain reaction (PCR) from DNA collected from 16 family members and 100 Ashkenazi Jewish controls. The PCR reaction used Amplitaq Gold (Carlsbad, CA) and primers in the introns flanking exon 5 (left primer 5′-AGT GGT CAT AGA CAC GTA GTG GGT GCA C-3′, previously described by Santhiya et al. [29], and right primer 5′-AGG GTC GAT GTG CCC AGG AAC TTT-3′). The 3′ position in each primer was complementary to the template strand of the sequence flanking *CRYBB2* exon 5, but that 3’ position was not complementary to the corresponding sequence of *CRYBB2P1*. We examined additional SNPs in the fourth intron of *CRYBB2* to map the end of the pseudogene-like alleles in a similarly gene-specific manner using the left primer 5′-TGT GTG TGC ATG TGT GGG TGT TCA C-3′ and the right primer

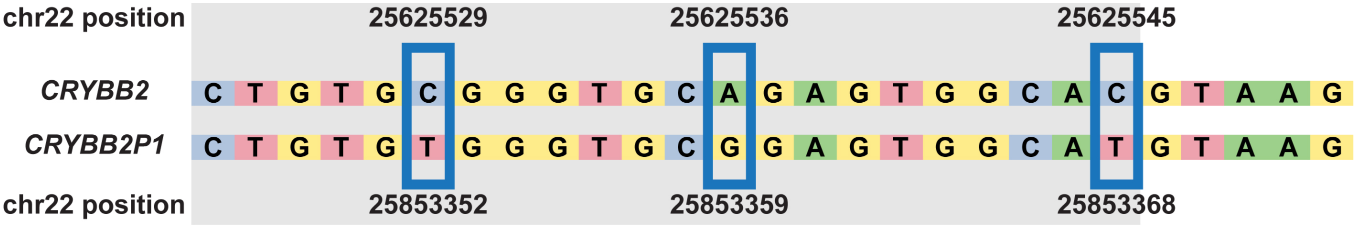

chr22 position
25625529
25625536
25625545
CRYBB2
C T G T G C G G G T G C A G A G T G G C A C G T A A G
CRYBB2P1
C T G T G T G G G T G C G G A G T G G C A T G T A A G
chr22 position
25853352
25853359
25853368

Figure 5. Alignment of the *CRYBB2* and *CRYBB2P1* sequences. The blue boxes highlight three coding sequence changes consistent with the sequence of the pseudogene that were modestly represented in the exome sequence reads. Nucleotides within the exon 5 coding sequence appear inside the gray box. The chromosome 22 positions are reported using hg19/GRCh37 coordinates.

5′-AGG GTC GAT GTG CCC AGG AAC TTT-3′. We amplified *CRYBB2P1*-specific sequences in all 16 family members using the left primer 5′-CTT GTG TGG GTG TGC GTG TGT GTG TA-3′ and the right primer 5′-AGG GTC GAT GTG CCC AGG AAC TTC-3′. We sequenced the PCR products at the University of Michigan DNA Sequencing Core using enzyme-directed incorporation of chain-terminating dideoxynucleotides (Sanger sequencing) and primer(s) from the PCR reaction; in the case of the pseudogene, we substituted the left primer with 5′- AGA GTG ATG TGT GGG ACA TG-3′. We analyzed the sequence visually for variations.

Sequence analysis: We aligned DNA sequences of exon 5 of *CRYBB2* (NM_000496.2) and the corresponding region of *CRYBB2P1* (NR_033733.1) using BLAST [30]. To evaluate cross-species conservation, we performed multiple protein sequence alignments in MacVector (MacVector, Cary, NC) based on ClustalW version 1.83 using the gonnet similarity matrix. The βB2-crystallin sequences used were from *Homo sapiens* (NP_000487.1), *Bos taurus* (NP_777232.1), *Macaca mulatta* (NP_001116366.1), *Canis lupus familiaris* (NP_001041578.1), *Oryctolagus cuniculus* (NP_001082786.1), *Cavia porcellus* (NP_001166542.1), *Rattus norvegicus* (NP_037069.1), *Mus musculus* (NP_031799.1), *Gallus gallus* (NP_990506.2), *Xenopus (Silurana) tropicalis* (NP_001238879.1), *Danio rerio* (NP_001018138.1), *Sus scrofa* (ENSSSCP00000010625.2), *Equus caballus* (XP_001500104), *Felis catus* (XP_003994803), and *Tetraodon nigroviridis* (ENSTNIP00000014410.1).

We produced hydrophobicity plots using ProtScale [31]. We used PROVEAN [32] and PolyPhen-2 [24] to predict the effects of the SNPs rs2330991, rs2330992, and rs4049504 on the βB2-crystallin protein.

RESULTS

Ten members of a large consanguineous Ashkenazi Jewish family were found to be affected with congenital cataracts (Figure 1). As most cataract extractions were performed at other institutions before our recruitment of the family, detailed information about the type of cataract is not available for anyone other than individual VII:13, whose medical records state that the cataract was nuclear. Individuals with congenital cataracts exhibited additional phenotypes, including microphthalmia in all individuals affected with congenital cataracts; most of the congenital cataract cases subsequently developed glaucoma or ocular hypertension (VII:3, VII:4, VII:5, VII:8, VII:9, and VII:13). The individuals VI:11, VII:1, VII:2, VII:6, and VII:11 do not have congenital cataracts; two of these individuals, VII:1 and VII:11, exhibited developmental delay. The grandfather V:4 had primary open angle glaucoma and senile cataracts, but he had neither congenital cataracts nor microphthalmia, raising questions about whether he might represent a case of delayed penetrance or reduced expressivity. None of the other family members were reported as having congenital cataracts, including grandmother V:3, the siblings of VI:11 and VI:12, and their children.

Evaluating the autosomal recessive model of inheritance: Given that the father VI:12 and his children are affected with congenital cataracts despite there being no other family history, and given that the parents VI:11 and VI:12 are inside a consanguinity loop, we evaluated an autosomal recessive model of inheritance and found significant evidence of linkage on chromosome 22 from 16.918 to 22.437 Mb (LOD = 3.61, Figure 2), and no other regions in the genome reached significance. Figure 2B shows the genetic inclusion interval and the locations of the chromosome 22 crystallin genes (*CRYBB3*, *CRYBB2*, *CRYBB1*, and *CRYBA4*, as well as the pseudogene *CRYBB2P1*) that are all located outside of the autosomal recessive genetic inclusion interval. Given that more than 100 genes are located inside that interval, whole-exome sequencing was a more inexpensive approach to screening candidate genes than targeted sequencing.

Analysis of paired-end whole-exome sequencing data identified a small number of nonsynonymous changes in the three affected individuals that were absent in the controls.

Table 2. Read depths for *CRYBB2* and *CRYBB2P1* in whole-exome sequencing data.

	Pseudogene allele	Chr 22 position (Build 37/ hg19)	Affected			Unaffected		
			VI-12	VII-3	VII-10	VI-11	VII-2	VII-6
Mapped back to *CRYBB2*	T	25,625,529	6/21	11/34	0/7	0/40	0/41	1/59
	G	25,625,536	4/19	9/31	0/5	0/36	0/36	0/51
	T	25,625,545	4/17	6/25	1/8	0/30	0/31	0/43
Mapped back to *CRYBB2P1*	T	25,853,352	30/30	79/80	65/65	58/58	44/44	68/70
	G	25,853,359	29/29	73/73	68/68	52/53	37/37	62/62
	T	25,853,368	28/28	70/70	65/65	52/52	33/34	52/53

Fractions listed represent pseudogene allele calls/total calls mapped to that location

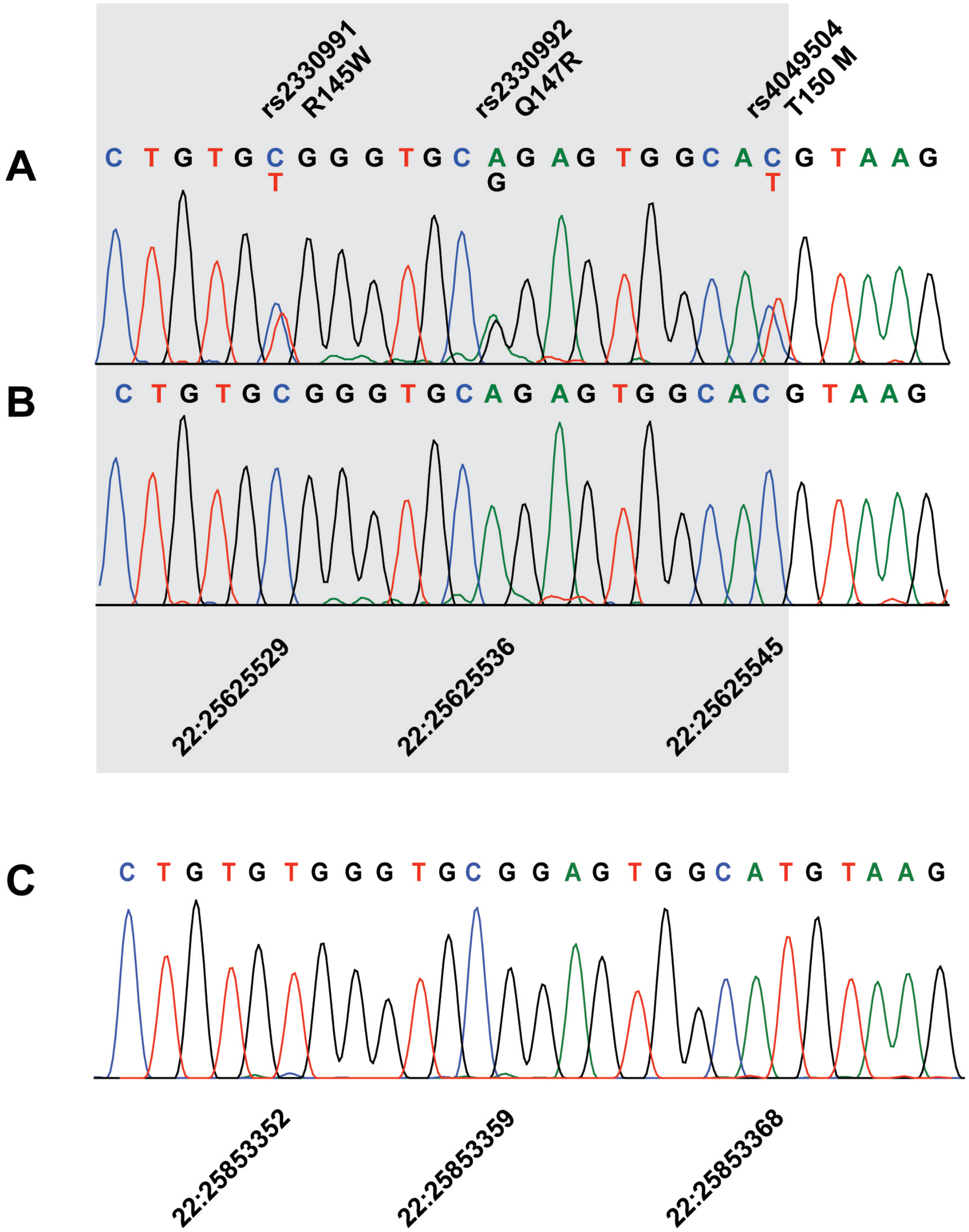


Figure 6. Sanger sequence traces from *CRYBB2* exon 5 and the corresponding region of C*RYBB2P1*. The representative chromatograms for *CRYBB2* are shown for **A:** VI:12 (affected) and **B:** VI:11 (unaffected). The section of the sequence that includes the coding sequence is highlighted in gray. All affected family members are heterozygous for all three nonsynonymous changes. In **C**, the corresponding region of the pseudogene in individual VII:13 is shown for comparison. Chromosome 22 positions are reported using hg19/GRCh37 coordinates.

These changes were assigned reduced priority as candidate genes based on a combination of allele frequencies, cross-species conservation [22,23], and informatic predictions of deleterious changes to the protein [24] (Table 1). Other changes identified were synonymous or were in noncoding or extragenic sequences.

Evaluating the autosomal dominant model of inheritance: In the absence of strong candidates in the autosomal recessive genetic inclusion interval, we also considered a model involving autosomal dominant inheritance with possible nonpenetrance or reduced expressivity on the part of the grandfather V:4, who has senile—but not congenital—cataracts. With the phenotype of the grandfather V:4 set to missing, we found significant evidence of linkage to a region of chromosome 22 (16.918 to 25.641 Mb, LOD = 3.61, Figure 3) that overlaps with the region identified under the autosomal recessive model. When the phenotype of V:4 was set to affected, the maximum LOD score was 3.91; if set to unaffected, all linkage evidence disappeared. Changing the phenotype did not affect the result for the recessive model. Sensitivity analyses using varied penetrance parameters

Table 3. Frequencies of fifth intron *CRYBB2* SNPs in 100 Ashkenazi Jewish population controls.

SNP	Minor allele	Ashkenazi Jewish		dbSNP[a]		
		Minor allele frequency (MAF)	Minor allele count	MAF	Minor allele count	MAF source
rs55700037	G	0.330	66	0.218	475	1000 Genomes
rs138678957	T	0.005	1	0.008	17	1000 Genomes
rs4049507	A	0.010	2	N/A	N/A	N/A
rs116926807	A	0.140	28	0.030	65	1000 Genomes
rs117284842	T	0.140	28	0.030	65	1000 Genomes

[a]Accessed 8/14/2013

did not result in a substantial qualitative change in the LOD scores (data not shown).

Haplotype analysis (Figure 4, black lettering) identified an affected haplotype present on one copy of chromosome 22 in all of the affected individuals. Haplotype analysis supports an autosomal dominant mode of inheritance. Some affected individuals in generation VII received one of the maternal chromosomes from VI:11, while some received the other copy from her. The fact that V:3 is unaffected, that her ancestors and siblings are unaffected, and that her children are affected whether they received her maternal copy of chromosome 22 or her paternal copy all support an autosomal dominant mode of inheritance of a defect located on a chromosome that did not originate in the maternal branch of the family. Tracing transmission of the affected haplotype suggests that it originated with V:4, the unaffected paternal grandfather, whose haplotype matches one that is present in all affected individuals and none of the other unaffected individuals.

Initial discovery of the variants: Figure 3B shows the region within chromosome 22 that is co-segregating with the disease phenotype under the autosomal dominant model. A recombination event detected in individual VII:7 between rs133201 and rs10483117 indicates that the congenital cataract locus is located centromeric to rs133201, and it encompasses a larger region of chromosome 22 than was encompassed by the autosomal recessive genetic inclusion interval. This interval includes the known cataract genes includes the known cataract genes *CRYBB2* [10] and *CRYBB3* [33]. Analysis of the paired-end whole-exome sequencing data identified no additional attractive candidates (Table 1), including no nonsynonymous coding variants in *CRYBB2* or *CRYBB3*.

A visual inspection of the sequence data for the *CRYBB2* and *CRYBB3* regions using the Integrated Genomics Viewer identified three changes in *CRYBB2* represented at a low level that had not been picked up by the sequence-analysis pipeline. These sequence variants convert the normal *CRYBB2* exon 5

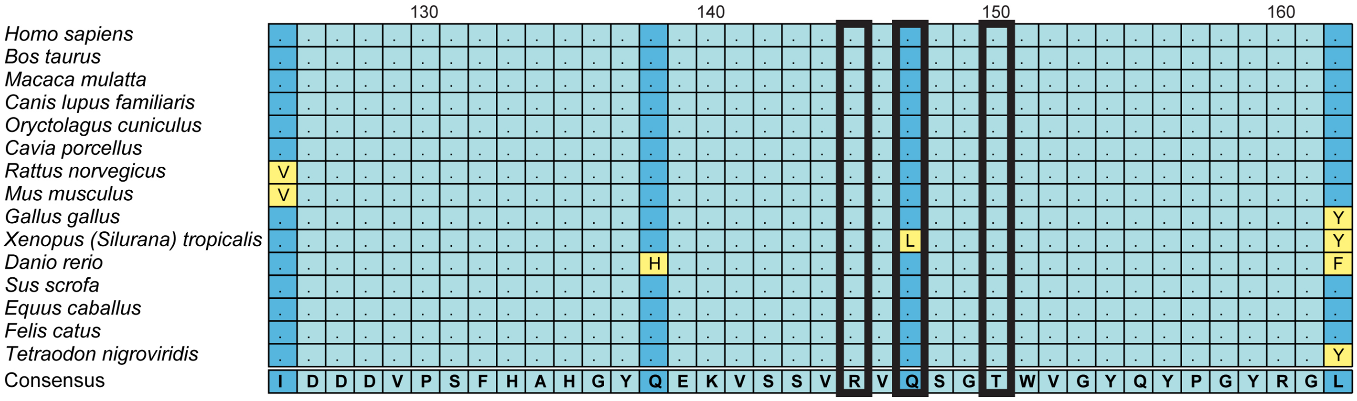


Figure 7. Multiple sequence alignments of βB2-crystallins. The protein sequences were aligned from the same region of the βB2-crystallins from 15 different species using MacVector. The numbering across the top is based on the human sequence. The light-blue boxes show 100% sequence identity in all species, the darker blue boxes show the consensus sequence, and the yellow boxes show mismatches. The amino acids that are altered in this family are boxed in thick black. Two of the amino acids that were altered in our congenital cataract family are completely conserved among the species studied, and one is highly conserved.

Table 4. Prediction of the effect of each of the changes on the bB2-crystallin protein.

dbSNP	rs2330991	rs2330992	rs4049504
Allele	C/T	A/G	C/T
Protein position	145	147	150
Chr position (GRCh37.p5)	22: 25,625,529	22: 25,625,536	22: 25,625,545
Residue change	R [Arg] → W [Trp]	Q [Gln] → R [Arg]	T [Thr] → M [Met]
Allele change	CGG → TGG	CAG → CGG	ACG → ATG
PROVEAN score	−5.605	−1.069	−4.830
PROVEAN prediction	deleterious	neutral	deleterious
PolyPhen-2 score	0.999	0.006	0.997
PolyPhen-2 interpretation	probably damaging	benign	probably damaging

RCSB Protein Data Bank ID 1YTQ

coding sequence to look like the corresponding sequence of the pseudogene *CRYBB2P1* (Figure 5). When comparing exon 5 of *CRYBB2* and the corresponding region of *CRYBB2P1*, we find 98% DNA sequence homology; the variant form of *CRYBB2* exon 5 has 100% sequence homology to the corresponding region of the pseudogene.

Upon analysis of the mapped sequence data, we noticed that the read depths were especially low for this region of *CRYBB2* in the affected individuals compared to the read depths across the same region of *CRYBB2P1* (Table 2). This raised the concern that mismapping of sequences between *CRYBB2P1* and *CRYBB2* might be complicating our efforts to detect causative mutations.

Validation of the variants: We elected to use gene-specific PCR amplification of exon 5 followed by Sanger sequencing to evaluate the hypothesis that the variants were actually present in *CRYBB2* exon 5, but the sequence-analysis pipeline was mismapping the reads containing the nonsynonymous variants to the pseudogene. All ten affected family members were found to have three nonsynonymous variants in the coding sequence of *CRYBB2*: rs2330991, c.433 C>T (p.R145W); rs2330992, c.440A>G (p.Q147R); and rs4049504, c.449C>T (p.T150M; Figure 6A,B); these three variants were unseen in the six unaffected family members. The scoring of all SNPs from the Sanger sequencing data appears in red lettering in Figure 4. Sequencing the corresponding region of the pseudogene with pseudogene-specific primers in all 16 members of the family revealed that all are homozygous for the *CRYBB2P1* alleles shown in Figure 5, so both alleles of *CRYBB2P1* match the converted region in variant *CRYBB2* (Figure 6C). Using gene-specific amplification, we did not observe any of the three exon 5 *CRYBB2* nonsynonymous changes in the 100 Ashkenazi Jewish controls. These changes

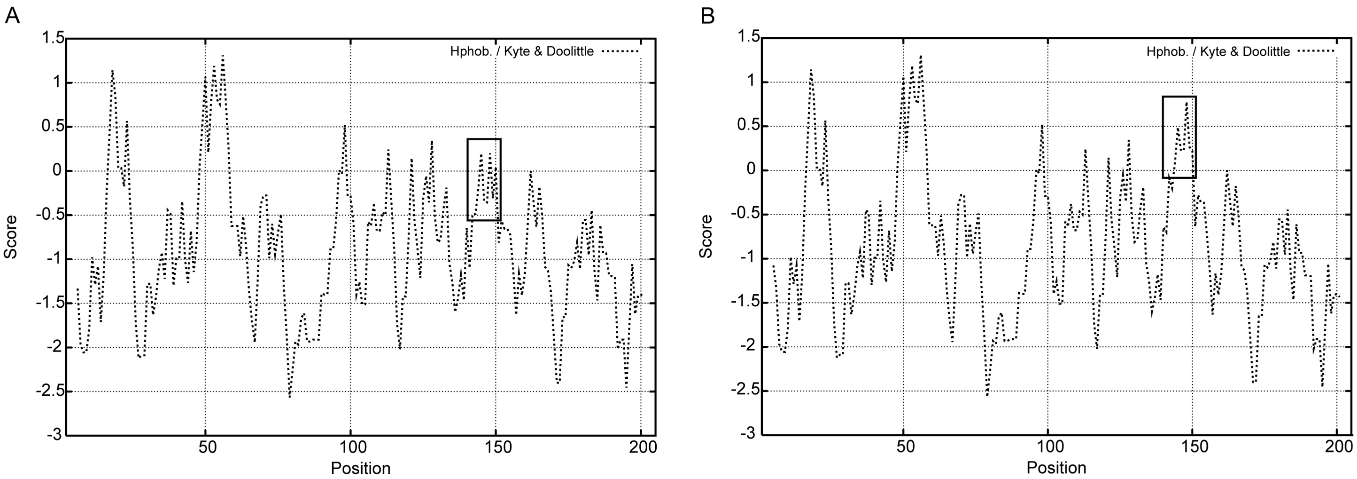


Figure 8. Hydrophobicity plots for **A**: wild-type and **B**: mutant βB2-crystallins with all three coding sequence mutations present. Introducing the three amino acid changes seen in this family increases the hydrophobicity in the boxed region of the protein containing the mutations.

were neither reported in the 1,092 individuals representing populations across four continents who were screened in the 1000 Genomes Project [34] nor in the exomes of the 6,503 individuals sequenced in the National Heart, Lung, and Blood Institute (NHLBI) Exome Sequencing Project (Exome Variant Server-EVS). Sanger sequencing of this region in the population controls also covered the intronic SNPs rs55700037, rs138678957, rs4049507, rs116926807, and rs117284842, which are variable in the general population (Table 3).

Determination of the length of the region altered: To determine the boundaries of the region of *CRYBB2* containing pseudogene alleles, we used gene-specific primers to PCR-amplify and sequence the regions flanking exon 5. The three changes in the coding sequence cover a region of 17 bp. The next nucleotide that differs between *CRYBB2* and *CRYBB2P1* in the 5′ direction from the three changes is 246 bp upstream of rs2330991 at position 25,625,283, where all 16 individuals are homozygous for the *CRYBB2* allele and do not contain the *CRYBB2P1* allele. In the 3′ direction, individual V:4 is homozygous for G at rs55700037 in *CRYBB2*, 9 bp downstream from the last *CRYBB2* coding sequence change at rs4049504.

Since the corresponding position in the pseudogene is reported to be polymorphic (rs4049505), we sequenced the pseudogene across this region to determine whether this marker was informative for the end of the region of the pseudogene alleles. All 16 family members, including the grandfather, father, mother, and children, were homozygous for A at rs4049505 in *CRYBB2P1*. This indicates that the altered sequence does not extend as far as rs4049505, since all affected individuals appear to inherit a *CRYBB2* G allele from V:4, who has a G in *CRYBB2* but not in *CRYBB2P1*. Therefore, the minimum length of the sequence of the gene-converted region in *CRYBB2* containing the pseudogene-like alleles is 17 bp; the maximum is 270 bp.

Altered protein sequence: The three nonsynonymous changes in exon 5 alter residues that are highly conserved across species and reside in a highly conserved region of βB2-crystallin (Figure 7). At two of the three positions (R145 and T150), the amino acid is completely conserved among 15 species; in our comparison, only one species (the Western clawed frog, *Xenopus [Silurana] tropicalis*) differs from the consensus at position 147.

Hydrophobicity plots show altered hydrophobicity in the mutant βB2-crystallin protein in the region surrounding the three nonsynonymous variants when compared to the normal protein (Figure 8). Predictions of the effects of each of the changes on the βB2-crystallin protein with PROVEAN [32] and PolyPhen-2 [24] indicated that changes R145W and T150M are deleterious/probably damaging, while Q147R is predicted to be neutral/benign (Table 4).

DISCUSSION

We performed linkage analysis using an autosomal recessive model of inheritance suggested by the transmission of the phenotypes in the family; this identified linkage to chromosome 22 (16.918 to 22.437 Mb, LOD = 3.61). Whole-exome sequence analysis yielded no likely pathogenic variants inside of the genetic inclusion interval, and the likely candidates—the crystallin genes—were excluded from the genetic inclusion interval.

Exploration of an autosomal dominant model showed significant evidence of linkage to chromosome 22 (maximum LOD = 3.91). Linkage and haplotype analyses place the gene between 16.918 and 25.641 Mb on chromosome 22, a larger region that contains the two known cataract genes *CRYBB2* [10] and *CRYBB3* [33] that were excluded by the autosomal recessive model. Haplotype analysis confirmed that all the affected individuals in the large sibship carry a single copy of chromosome 22 that can be traced to the father with congenital cataracts (VI:12) and the grandfather with senile cataracts (V:4; Figure 4), while they do not all share the same copy of the maternal chromosome from the unaffected mother VI:11. Thus, the gene responsible for congenital cataracts in this family is likely inherited in an autosomal dominant manner. Whole-exome sequence analysis yielded no likely pathogenic variants inside the genetic inclusion interval for the autosomal dominant model, including in *CRYBB2* and *CRYBB3*. However, a visual inspection of the mapped sequence data revealed three variants present at a low level in exon 5 of *CRYBB2* not identified by the sequence-analysis pipeline. Using gene-specific primers, we confirmed the three changes in one copy of exon 5 of *CRYBB2* in all affected individuals and showed that those changes are absent from *CRYBB2* in all unaffected individuals, as well as in the population of 100 ethnically matched controls.

Our results are consistent with the hypothesis that a *de novo* gene conversion event, either in the germline of V:4 or at an early embryonic stage in VI:12, transferred a region of less than 271 bp from *CRYBB2P1* to exon 5 of *CRYBB2;* this transfer happened between individuals V:4 and VI:12. The maximum LOD score (3.91) was obtained when grandfather V:4 was scored as affected, even though he lacks the causative congenital cataract genotype or phenotype. Relationship testing, which used the whole genome SNP panel but did not include the gene-converted alleles, indicates that individual V:4 is related as purported, and V:4 provided the affected haplotype. However, the gene-specific sequence data show that V:4 lacks the three sequence changes found in his affected son and grandchildren, resolving the apparent discrepancy between his affected haplotype and his unaffected phenotype. This type of *de novo* event is an expected occasional complication in linkage studies, as we have seen previously [35].

These *CRYBB2* variants were initially detected by a visual inspection of mapped exome sequencing information and were apparently missed by the sequence-analysis pipeline because *CRYBB2* variants were mapped to *CRYBB2P1*. This mismapping appears to be occurring despite the use of paired-end sequencing, and it likely occurred because the short sequence reads in this region of the *CRYBB2* variants shared a greater homology with the *CRYBB2P1* sequence. The presence of more than 11,000 pseudogenes sharing homology with more than 3,000 parent genes present in the GENCODE version 7 data set [36] suggests that mismapping of next-generation sequencing might be a widespread problem, even when paired-end sequencing is used. This type of mismapping is reflected in the Short Genetic Variations (dbSNP) database, where as many as 8.32% of biallelic coding SNPs are thought to be artifacts generated by sequence variations between two homologous genes rather than by actual inter-individual differences [37]. Other researchers have reported non-uniform coverage of whole-exome sequencing causing them to initially miss a potentially causative mutation that was later confirmed by Sanger sequencing [38]. Our study underscores the potential for these types of problems with next-generation sequencing when a highly homologous gene or pseudogene is present.

The gene conversion event and the sequence mismapping both arise from the existence of a highly homologous pseudogene *CRYBB2P1* near *CRYBB2*. Gene conversion is known to occur where there is a high sequence homology, usually greater than 95%, and it is more common when the interacting sequences are closely linked [39]. In exon 5, *CRYBB2* shares 98% DNA sequence homology with the corresponding region of *CRYBB2P1* and is located just 228 kb downstream, making gene conversion a likely mechanism for introducing these changes. While the region transferred is too small (at most, 270 bp) to be expected in a double-recombination event, intervals this small have been reported in gene conversion before [9]. Other independent gene conversion events between *CRYBB2* and *CRYBB2P1* have been proposed in *CRYBB2* exon 6 in Chilean [11], Chinese [12], and Indian [9] families with congenital cataracts. In exon 5, the same three changes that we report here were reported in a Danish family [40], but the results were subsequently reported to have resulted from nonspecific amplification of the pseudogene [41,42]. Repeated observations of gene conversion events between *CRYBB2* and *CRYBB2P1* suggest that the highly homologous sequences are susceptible to gene conversion events; the fact that such regions are also susceptible to cross-reaction in some genotyping systems has implications for the design of mutation screening programs in cataract patients.

The sequence changes identified in *CRYBB2* alter amino acids in a highly conserved region of a known congenital cataract gene product, are predicted to alter hydrophobicity in the region of the crystallin protein encoded by exon 5 and are the likely cause of congenital cataracts in this family. In a gene conversion event that transfers more than one variation, the process of determining what is or is not causative becomes complex. Future functional assays will be required to determine whether one or more of these *CRYBB2* variants can cause disease when occurring alone or whether only a combination of these alleles will cause disease.

ACKNOWLEDGMENTS

This work was supported in part by R01EY011671 (J.E.R.), R01EY09580 (J.E.R.), P30EY007003 (core), and Research to Prevent Blindness. This work was presented in part at the Association for Research in Vision and Ophthalmology in Seattle, WA in 2013 and in Orlando, FL in 2014.

REFERENCES

1. Foster A, Gilbert C, Rahi J. Epidemiology of cataract in childhood: a global perspective. J Cataract Refract Surg 1997; 23:Suppl 1601-4. [PMID: 9278811].
2. Shiels A, Hejtmancik JF. Genetic origins of cataract. Arch Ophthalmol 2007; 125:165-73. [PMID: 17296892].
3. Shiels A, Bennett TM, Hejtmancik JF. Cat-Map: putting cataract on the map. Mol Vis 2010; 16:2007-15. [PMID: 21042563].
4. Shiels A, Hejtmancik JF. Genetics of human cataract. Clin Genet 2013; 84:120-7. [PMID: 23647473].
5. Hejtmancik JF. Congenital cataracts and their molecular genetics. Semin Cell Dev Biol 2008; 19:134-49. [PMID: 18035564].
6. Gill D, Klose R, Munier FL, McFadden M, Priston M, Billingsley G, Ducrey N, Schorderet DF, Heon E. Genetic heterogeneity of the Coppock-like cataract: a mutation in CRYBB2 on chromosome 22q11.2. Invest Ophthalmol Vis Sci 2000; 41:159-65. [PMID: 10634616].
7. Yao K, Tang X, Shentu X, Wang K, Rao H, Xia K. Progressive polymorphic congenital cataract caused by a CRYBB2 mutation in a Chinese family. Mol Vis 2005; 11:758-63. [PMID: 16179907].
8. Li FF, Zhu SQ, Wang SZ, Gao C, Huang SZ, Zhang M, Ma X. Nonsense mutation in the CRYBB2 gene causing autosomal

dominant progressive polymorphic congenital coronary cataracts. Mol Vis 2008; 14:750-5. [PMID: 18449377].

9. Vanita SV. Reis A, Jung M, Singh D, Sperling K, Singh JR, Burger J. A unique form of autosomal dominant cataract explained by gene conversion between beta-crystallin B2 and its pseudogene. J Med Genet 2001; 38:392-6. [PMID: 11424921].

10. Litt M, Carrero-Valenzuela R, LaMorticella DM, Schultz DW, Mitchell TN, Kramer P, Maumenee IH. Autosomal dominant cerulean cataract is associated with a chain termination mutation in the human beta-crystallin gene CRYBB2. Hum Mol Genet 1997; 6:665-8. [PMID: 9158139].

11. Bateman JB, von-Bischhoffshaunsen FR, Richter L, Flodman P, Burch D, Spence MA. Gene conversion mutation in crystallin, beta-B2 (CRYBB2) in a Chilean family with autosomal dominant cataract. Ophthalmology 2007; 114:425-32. [PMID: 17234267].

12. Wang L, Lin H, Gu J, Su H, Huang S, Qi Y. Autosomal-dominant cerulean cataract in a chinese family associated with gene conversion mutation in beta-B2-crystallin. Ophthalmic Res 2009; 41:148-53. [PMID: 19321936].

13. Reddy MA, Francis PJ, Berry V, Bhattacharya SS, Moore AT. Molecular genetic basis of inherited cataract and associated phenotypes. Surv Ophthalmol 2004; 49:300-15. [PMID: 15110667].

14. Ott J. Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 1989; 86:4175-8. [PMID: 2726769].

15. Cottingham RW Jr, Idury RM, Schaffer AA. Faster sequential genetic linkage computations. Am J Hum Genet 1993; 53:252-63. [PMID: 8317490].

16. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics 2010; 26:589-95. [PMID: 20080505].

17. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007; 81:559-75. [PMID: 17701901].

18. Wigginton JE, Abecasis GR. PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. Bioinformatics 2005; 21:3445-7. [PMID: 15947021].

19. Abecasis GR, Cherny SS, Cookson WO, Cardon LR. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 2002; 30:97-101. [PMID: 11731797].

20. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009; 25:2078-9. [PMID: 19505943].

21. Flicek P, Ahmed I, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Garcia-Giron C, Gordon L, Hourlier T, Hunt S, Juettemann T, Kahari AK, Keenan S, Komorowska M, Kulesha E, Longden I, Maurel T, McLaren WM, Muffato M, Nag R, Overduin B, Pignatelli M, Pritchard B, Pritchard E, Riat HS, Ritchie GR, Ruffier M, Schuster M, Sheppard D, Sobral D, Taylor K, Thormann A, Trevanion S, White S, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Harrow J, Herrero J, Hubbard TJ, Johnson N, Kinsella R, Parker A, Spudich G, Yates A, Zadissa A, Searle SM. Ensembl 2013. Nucleic Acids Res 2013; 41:Database issueD48-55. [PMID: 23203987].

22. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res 2010; 20:110-21. [PMID: 19858363].

23. Cooper GM, Stone EA, Asimenos G, Green ED, Batzoglou S, Sidow A. Distribution and intensity of constraint in mammalian genomic sequence. Genome Res 2005; 15:901-13. [PMID: 15965027].

24. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. A method and server for predicting damaging missense mutations. Nat Methods 2010; 7:248-9. [PMID: 20354512].

25. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 2010; 38:e164 [PMID: 20601685].

26. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes. Bioinformatics 2006; 22:1036-46. [PMID: 16500937].

27. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. Integrative genomics viewer. Nat Biotechnol 2011; 29:24-6. [PMID: 21221095].

28. Thorvaldsdottir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 2013; 14:178-92. [PMID: 22517427].

29. Santhiya ST, Manisastry SM, Rawlley D, Malathi R, Anishetty S, Gopinath PM, Vijayalakshmi P, Namperumalsamy P, Adamski J, Graw J. Mutation analysis of congenital cataracts in Indian families: identification of SNPS and a new causative allele in CRYBB2 gene. Invest Ophthalmol Vis Sci 2004; 45:3599-607. [PMID: 15452067].

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10. [PMID: 2231712].

31. Gasteiger E, Hoogland C, Gattiker A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. In: Walker JM, editor. The Proteomics Protocols Handbook. Totowa, NJ: Humana Press; 2005. p. 571–607.

32. Choi Y, Sims GE, Murphy S, Miller JR, Chan AP. Predicting the functional effect of amino acid substitutions and indels. PLoS ONE 2012; 7:e46688 [PMID: 23056405].

33. Riazuddin SA, Yasmeen A, Yao W, Sergeev YV, Zhang Q, Zulfiqar F, Riaz A, Riazuddin S, Hejtmancik JF. Mutations in betaB3-crystallin associated with autosomal recessive cataract in two Pakistani families. Invest Ophthalmol Vis Sci 2005; 46:2100-6. [PMID: 15914629].

34. Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, Handsaker RE, Kang HM, Marth GT, McVean GA. An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491:56-65. [PMID: 23128226].

35. Krafchak CM, Pawar H, Moroi SE, Sugar A, Lichter PR, Mackey DA, Mian S, Nairus T, Elner V, Schteingart MT, Downs CA, Kijek TG, Johnson JM, Trager EH, Rozsa FW, Mandal MN, Epstein MP, Vollrath D, Ayyagari R, Boehnke M, Richards JE. Mutations in TCF8 cause posterior polymorphous corneal dystrophy and ectopic expression of COL4A3 by corneal endothelial cells. Am J Hum Genet 2005; 77:694-708. [PMID: 16252232].

36. Pei B, Sisu C, Frankish A, Howald C, Habegger L, Mu XJ, Harte R, Balasubramanian S, Tanzer A, Diekhans M, Reymond A, Hubbard TJ, Harrow J, Gerstein MB. The GENCODE pseudogene resource. Genome Biol 2012; 13:R51 [PMID: 22951037].

37. Musumeci L, Arthur JW, Cheung FS, Hoque A, Lippman S, Reichardt JK. Single nucleotide differences (SNDs) in the dbSNP database may lead to errors in genotyping and haplotyping studies. Hum Mutat 2010; 31:67-73. [PMID: 19877174].

38. Sirmaci A, Edwards YJ, Akay H, Tekin M. Challenges in whole exome sequencing: an example from hereditary deafness. PLoS ONE 2012; 7:e32000 [PMID: 22363784].

39. Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 2007; 8:762-75. [PMID: 17846636].

40. Hansen L, Mikkelsen A, Nurnberg P, Nurnberg G, Anjum I, Eiberg H, Rosenberg T. Comprehensive mutational screening in a cohort of Danish families with hereditary congenital cataract. Invest Ophthalmol Vis Sci 2009; 50:3291-303. [PMID: 19182255].

41. Kumar KD, Kumar GS, Santhiya ST. Nonspecific PCR amplification of CRYBB2-pseudogene leads to misconception of natural variation as mutation. Invest Ophthalmol Vis Sci 2012; 53:5770 [PMID: 22915216].

42. Hansen L, Rosenberg T. Author response: Nonspecific PCR amplification of CRYBB2-pseudogene leads to misconception of natural variation as mutation. Invest Ophthalmol Vis Sci 2012; 53:6666 [PMID: 23011185].