Analysis of the *ABCA4* Gene by Next-Generation Sequencing

Jana Zernant,[1] *Carl Schubert,*[1] *Kate M. Im,*[2] *Tomas Burke,*[1] *Carolyn M. Brown,*[1] *Gerald A. Fishman,*[3] *Stephen H. Tsang,*[1,4] *Peter Gouras,*[1] *Michael Dean,*[2] *and Rando Allikmets*[1,4]

Purpose. To find all possible disease-associated variants in coding sequences of the *ABCA4* gene in a large cohort of patients diagnosed with *ABCA4*-associated diseases.

Methods. One hundred sixty-eight patients who had been clinically diagnosed with Stargardt disease, cone-rod dystrophy, and other *ABCA4*-associated phenotypes were prescreened for mutations in *ABCA4* with the ABCA4 microarray, resulting in finding 1 of 2 expected mutations in 111 patients and 0 of 2 mutations in 57 patients. The next-generation sequencing (NGS) strategy was applied to these patients to sequence the entire coding region and the splice sites of the *ABCA4* gene. Identified new variants were confirmed or rejected by Sanger sequencing and analyzed for possible pathogenicity by in silico programs and, where possible, by segregation analyses.

Results. Sequencing was successful in 159 of 168 patients and identified the second disease-associated allele in 49 of 103 (~48%) of patients with one previously identified mutation. Among those with no mutations, both disease-associated alleles were detected in 4 of 56 patients, and one mutation was detected in 10 of 56 patients. The authors detected a total of 57 previously unknown, possibly pathogenic, variants: 29 missense, 4 nonsense, 9 small deletions and 15 splice-site-altering variants. Of these, 55 variants were deemed pathogenic by a combination of predictive methods and segregation analyses.

Conclusions. Many mutations in the coding sequences of the *ABCA4* gene are still unknown, and many possibly reside in noncoding regions of the *ABCA4* locus. Although the ABCA4 array remains a good first-pass screening option, the NGS platform is a time- and cost-efficient tool for screening large cohorts. (*Invest Ophthalmol Vis Sci.* 2011;52:8479–8487) DOI: 10.1167/iovs.11-8182

Mutations in the *ABCA4* gene are responsible for a wide variety of retinal dystrophy phenotypes, such as autosomal recessive Stargardt disease (STGD1),[1] cone-rod dystrophy (CRD),[2,3] and retinitis pigmentosa (RP).[2,4,5] STGD1 (Mendelian Inheritance in Man 248200) is a predominantly juvenile-onset macular dystrophy associated with rapid central visual impairment, progressive bilateral atrophy of the foveal retinal pigment epithelium (Fig. 1), and frequent appearance of yellowish flecks, defined as lipofuscin deposits, around the macula or in the central and near-peripheral areas of the retina. In a large fraction of STGD1 patients, a "dark " or "silent " choroid is seen on fluorescein angiography that reflects the accumulation of lipofuscin throughout the retina.

More than 600 disease-associated *ABCA4* variants have been identified,[6] and the most frequent disease-associated *ABCA4* alleles have each been described in only approximately 10% of STGD1 patients. Several studies have identified frequent "ethnic group-specific" *ABCA4* alleles, such as the c.2588G>C variant resulting in a dual effect, p.G863A/delG863, as a founder mutation in Northern European patients with STGD1[7] and a complex allele (two variants on the same chromosome), p.L541P/A1038V, in both STGD1 and CRD patients of German origin (Fig. 2B).[3,8] Complex *ABCA4* alleles are not uncommon in STGD1.[9] In fact, they are detected in approximately 10% of all STGD patients.[10]

Allelic heterogeneity has substantially complicated genetic analyses of *ABCA4*-associated retinal disease. Efforts related to mutation detection and genotyping become especially crucial in genotype/phenotype correlation studies, in which screening of thousands of samples is needed to achieve enough statistical power because multiple rare variants and their combinations must be studied.[11] We generated, at the time, a high-throughput and cost-effective screening tool, the *ABCA4* genotyping microarray,[12] using solid-phase arrayed primer extension (APEX) technology. The *ABCA4* microarray, which has been regularly updated, contains all known disease-associated genetic variants (>600) in the *ABCA4* gene. The chip has been used for efficient, systematic screening of patients with *ABCA4*-associated diseases[13,14]; it detects approximately 65% to 75% of all disease-associated alleles. On average, the array screening finds two mutations in approximately 40% of patients diagnosed with "classical" STGD1. Of the rest, one mutation is detected in 40% of patients, whereas no disease-associated allele is found in the *ABCA4* coding region in 20% of screened patients.[6]

Direct Sanger sequencing of the entire *ABCA4* coding region detects between 66% and 80% of the alleles[10,15] but

From the Departments of [1]Ophthalmology and [4]Pathology and Cell Biology, Columbia University, New York, New York; the [2]Laboratory of Experimental Immunology, Cancer and Inflammation Program, Center for Cancer Research, National Cancer Institute at Frederick, Frederick, Maryland; and [3]The Pangere Center for Hereditary Retinal Diseases, Chicago Lighthouse for the Blind and Visually Impaired, Chicago, Illinois.

Supported in part by National Eye Institute/National Institutes of Health Grants EY021163, EY013435, EY019861, and EY019007 (Core Support for Vision Research); Foundation Fighting Blindness; unrestricted funds from Research to Prevent Blindness (Department of Ophthalmology, Columbia University); Intramural Research Program of the National Institutes of Health, National Cancer Institute, Center for Cancer Research; and SAIC-Frederick under Contract NO1-NO1-CO-12400.

Submitted for publication July 7, 2011; revised August 9, 2011; accepted September 6, 2011.

Disclosure: **J. Zernant**, None; **C. Schubert**, None; **K.M. Im**, None; **T. Burke**, None; **C.M. Brown**, None; **G.A. Fishman**, None; **S.H. Tsang**, None; **P. Gouras**, None; **M. Dean**, None; **R. Allikmets**, None

Corresponding author: Rando Allikmets, Department of Ophthalmology, Columbia University, 630 West 168th Street, New York, NY 10032; rla22@columbia.edu.

FIGURE 1. Fundus autofluorescence image of the left eye of patient 3032 harboring *ABCA4* variants c.2300T>A (p.V767D) and c.735T>G (p.Y245*). This combination of *ABCA4* mutations resulted in early disease onset at 5 years of age. At 16 years of age, the patient was found to have extensive hypoautofluorescence, indicative of atrophy of the retinal pigment epithelium throughout the macula with patchy extension of hypoautofluorescence into the extramacular retina. Note the relative "sparing" of uniform hyperautofluorescence in the peripapillary region.

remains prohibitive for large patient cohorts because of time and cost constraints. Studies assessing the fraction of copy number variant (CNV) mutations, which elude PCR-based methods such as direct sequencing, (e.g., large deletions of exons and chromosomal segments), have found these in only approximately 1% of all STGD1 patients.[16]

Overall, variation in the *ABCA4* locus has emerged as the most prevalent cause of Mendelian retinal disease because approximately 1 of 20 people across all populations carry a potential disease-associated variant in this gene.[7,12,15] Recent advances in developing therapeutic applications for STGD1 in preclinical studies[17,18] suggest that more comprehensive and affordable genetic screening technologies have to be implemented for molecular diagnosis and for selection of patients who would benefit from specific (especially gene-based) therapeutic modalities. Here we describe screening by NGS of a large cohort of *ABCA4*-associated patients who had been analyzed by the ABCA4 array and still lacked one or both mutations.

PATIENTS, MATERIALS, AND METHODS

Patients

Patients (n = 168) affected with STGD1 (n = 153), CRD (n = 13), and RP (n = 2) were, after providing written consent, recruited and clinically examined over a 10-year period at the Department of Ophthalmology at Columbia University and at the University of Chicago at Illinois. Patients with CRD and RP were included in this study because they all harbored one *ABCA4* disease-associated variant after array screening. Age of onset was defined as the age at which symptoms were first reported. Visual acuity was measured using the Early Treatment Diabetic Retinopathy Study Chart 1. Clinical examination, fundus photography, fundus autofluorescence (Fig. 1) and spectral domain–optical coherence tomography (SD-OCT) (Spectralis HRA+OCT; Heidelberg Engineering, Heidelberg, Germany) were performed using standard acquisition protocols after pupil dilation with tropicamide 1%. All research was carried out with the approval of the Institutional Review Board of Columbia University and in accordance with the Declaration of Helsinki.

Array Screening

Screening with the ABCA4 array had been performed on all patients, followed by direct sequencing to confirm identified changes, as previ-

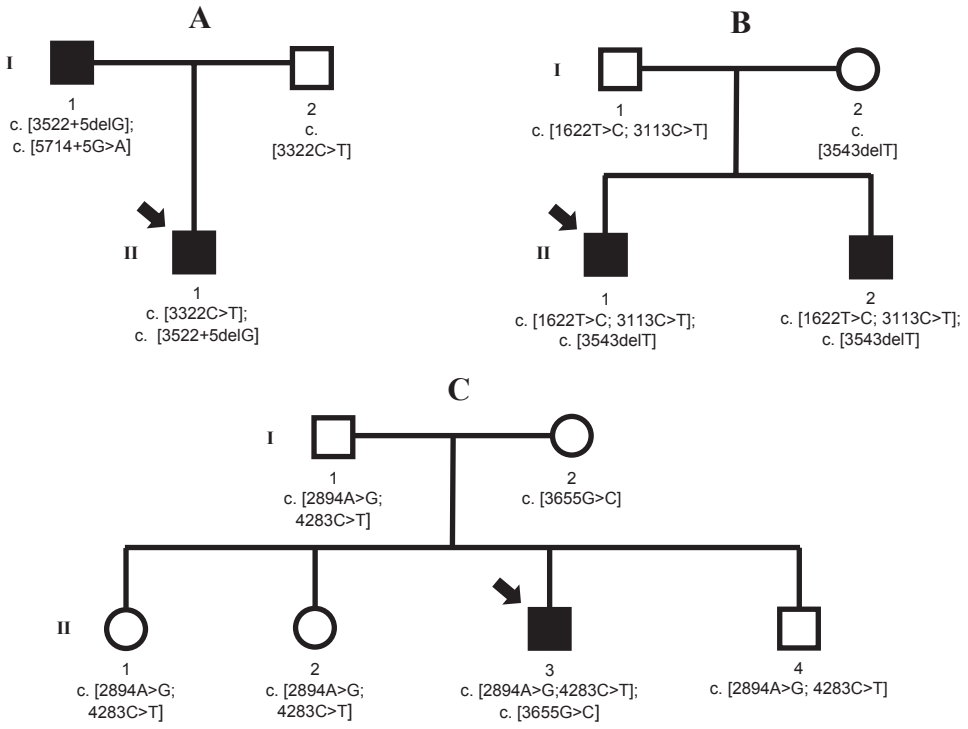

FIGURE 2. Pedigrees segregating Stargardt disease. (**A**) An example of a pedigree with pseudodominant inheritance. Father and son are both affected with arSTGD. Mother is a carrier of a frequent c.3322C>T (p.R1108C) mutation. The new c.3522+5delG variant affecting splicing (Table 1) was detected by NGS. (**B**) An example of a pedigree segregating a frequent complex *ABCA4* allele [c.1622T>C; 3113C>T] (p.L541P;A1038V), in which either mutation separately can cause the disease. The c.3543delT frameshift variant was detected by NGS. (**C**) An example of a pedigree segregating a complex allele in which one variant (c.2894A>G, p.N965S) causes disease and the other, c.4283C>T, p.T1428M, is a benign polymorphism, although it was originally described as a rare mutation in patients of European descent. The new c.3655G>C, p.A1219P variant was detected by NGS.

ously described.[12] Because the array screening had been performed over many years, different versions of the ABCA4 chip had been used, from the least representative (~300 mutations) to the current version (~600 variants).

NGS

All 50 ABCA4 exons were amplified using an amplicon tagging protocol (Access Array; Fluidigm, South San Francisco, CA; http://www.fluidigm.com/products/access-array.html). The integrated fluidic circuit of this system facilitates parallel amplification of 48 unique samples, in effect preparing 48 sequencing libraries. The primers (Supplementary Table S4, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental) for the amplification step were designed in accordance with the guidelines from Fluidigm, using Primer3 software (http://frodo.wi.mit.edu/primer3/input.htm). Universal forward and reverse tag sequences were added to 5′ ends of the designed primer sequences to enable the use of sample-specific barcode primers consisting of 454 sequence tags (all provided by Fluidigm). Every reaction combined both an amplicon tagging and a bar-coding sample tagging (identification) step that enables all 48 amplicons to be multiplexed at the sequencing step, maximizing the usefulness of the sequencer (454 GS-FLX; Fluidigm). Every reaction contained 50 to 100 ng genomic DNA, 5 pmol forward- and reverse-tagged amplification primers, 1 pmol forward and reverse barcode primers, 1× loading reagent (Access Array; Fluidigm), reaction buffer with $MgCl_2$ (FastStart High Fidelity; Roche, Indianapolis, IN), enzyme blend (FastStart High Fidelity; Roche), and 1× PCR-grade nucleotide mix.

Each 48 × 48 array generated 48 amplicons in 48 samples (n = 2304); 168 samples were amplified on 4 (3.5) arrays. The amplicons were tagged with the adaptors (454 Titanium; Roche) during the PCR amplification, pooled, and purified, resulting in one amplicon library per array. The resultant libraries (n = 4) were subjected to emulsion PCR and sequenced using the same chemistry (454 Titanium; Roche). Each library was sequenced in one region of a four-region picotiter plate. All the reactions were carried out in accordance with the manufacturer's protocols. Each region generated between 130,000 and 200,000 high-quality bidirectional reads, resulting in an average 56× coverage per amplicon. The best amplicons were covered with approximately 140 reads; most amplicons yielded, on average, 60 to 80 sequence reads. The total of 112 amplicons with coverage <10 reads (1.46%, 112/7680; 160 samples × 48 amplicons) were resequenced by the Sanger method.

Sequences of barcoded samples were analyzed with the GS Reference Mapper software (http://www.454.com/products-solutions/analysis-tools/gs-reference-mapper.asp), which mapped reads to the reference genome (HG19) and compiled a consensus sequence. All differences compared to the reference sequence are easily viewed with automatic output to separate files for insertions, deletions, and SNPs. High-confidence differences compared to the reference genome are compiled in a separate output file. The output files were converted to .sam files using a Python script kindly provided by Kevin Jacobs and subsequently to sorted .bam files using samtools [PMID: 19505943]. BAM files were then exported into the Integrated Genome Viewer (Broad Institute) for visualization of all identified variants. All identified variants were confirmed by Sanger sequencing.

Sequence Analyses

New missense variants were analyzed with algorithms such as Sorting Intolerant from Tolerant (SIFT; http://sift.jcvi.org/), and Polymorphism Phenotyping (PolyPhen; http://genetics.bwh.harvard.edu/pph/) to predict the impact of variants on the ABCA4 function and, consequently, on disease susceptibility. Variants detected in adjacent to exons intronic sequences were analyzed with splice site prediction programs GeneSplicer (http://www.cbcb.umd.edu/software/GeneSplicer), and Splice Site Finder (www.genet.sickkids.on.ca/~ali/splicesitefinder). All the prediction programs were accessed with bioinformatics software (Alamut 2.0; http://www.interactive-biosoftware.com). Where available, segregation of the new variants with the disease was analyzed in families (Fig. 2).

Results

Discovery of New Disease-Associated Variants by NGS

The entire ORF (all 50 *ABCA4* exons and flanking intronic sequences) was sequenced in 168 *ABCA4*-associated patients (Supplementary Table S1, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental). Of these, 111 had one pathogenic mutation previously identified by the ABCA4 array; in 57 patients array screening had not revealed any mutations, although all of them had been clinically confirmed as affected with *ABCA4*-associated disease. PCR amplification, barcoding, tagging, and pooling were performed on the Fluidigm 48 × 48 Access Arrays. Screening of the *ABCA4* gene is especially amenable for the Access Array system because 50 exons can be amplified as 48 amplicons. Sequencing did not work in eight patients, most likely because of a Fluidigm array error given that all these samples were in one array column. One patient with one identified *ABCA4* mutation represented a false negative for the ABCA4 array (i.e., this patient should not have been included in further sequencing analysis because both mutations in this patient should have been identified by the array). Three additional false-negative samples for the ABCA4 array were also detected; however, because no mutations were found, these samples were still valid for sequencing as one-mutation cases. Therefore, nine patients—seven with one mutation, one with no mutations, and one with two mutations—were excluded from the final analyses (Supplementary Table S1, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental).New variants were analyzed by a combination of predictive in silico methods, statistical analyses, segregation with the disease in the families, and screening of large patient and control cohorts for population frequencies. All variants were first compared against our most up-to-date *ABCA4* mutation database, which includes all published and communicated *ABCA4* variants. As predicted, the previously unknown identified variants were rare and not present in public databases such as dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP) and HapMap (http://hapmap.ncbi.nlm.nih.gov/). Filtering against these databases excluded known frequent variants with frequencies similar to those in the general population.

Of the 103 patients with one mutation, the NGS strategy found the second disease-associated allele in 49 patients (47.6%). Of the 56 patients with no identified mutations, NGS detected both disease-associated alleles in four patients (7.1%) and one mutation in 10 patients (17.9%). We detected a total of 57 previously unknown variants in 64 patients: 29 missense, 4 nonsense, 9 small deletion, and 15 splice-affecting variants from the nucleotide changes 5 base pairs upstream or downstream of the exons (Table 1). All but two (55/57) new variants were deemed pathogenic by a combination of predictive in silico methods and segregation analyses in available families (Table 1). A single-nucleotide deletion, c.885delC, was detected in three patients; the splice site mutation c.3522+5delG was found twice in the same family (parent and child each had an STGD1 diagnosis), and the nonsense mutation p.R2149* was detected in two patients. Two splice-affecting variants, c.6479+1G>A and c.6479+1G>C, targeted the same position. All the other newly identified mutations were detected only once.

The mutations p.D108V, p.Y245*, p.R2149*, and c.4667+1G>A, which were identified by NGS, were already known and included in the current ABCA4 array but were not at the time these samples were screened. Systematic screening of the

TABLE 1. Novel Variants Detected by NGS in the *ABCA4* Gene and Results of Analysis Using Bioinformatics Software

Nucleotide Change	Protein	Splicing Score Original		Splicing Score for New Variant		Average Difference	Polyphen	SIFT
		SpliceSite Finder-like	Gene Splicer	SpliceSite Finder-like	Gene Splicer			
c.91T>C	p.W31R	0	0	0	0	0	Probably damaging (0.999)	W
c.184C>T	p.P62S	0	0	0	0	0	Probably damaging (0.999)	P
c.770T>G	p.L257R	0	0	0	0	0	Possibly damaging (0.308)	m i F L
c.1253T>C	p.F418S	0	0	0	0	0	Probably damaging (0.999)	F
c.1531C>T	p.R511C	0	0	0	0	0	Probably damaging (1.000)	R
c.1745A>G	p.N582S	0	0	0.74	0.82	77.8	Probably damaging (0.894)	d K N
c.1868A>G	p.Q623R	0	0.24	0	0	12.1	Probably damaging (0.937)	Q
c.1964T>G	p.F655C	0	0	0	0	0	Probably damaging (0.999)	F
c.1977G>A	p.M659I	0	0	0.75	0.85	79.8	Probably damaging (0.999)	M
c.2243G>A	p.C748Y	0	0	0	0	0	Probably damaging (0.928)	g S A C
c.2401G>A	p.A801T	0	0	0	0	0	Probably damaging (0.98)	A
c.2893A>T	p.N965Y	0	0	0	0	0	Probably damaging (0.999)	N
c.3148G>A	p.G1050S	0	0	0	0	0	Possibly damaging (0.786)	G
c.3205A>G	p.K1069E	0	0	0	0	0	Probably damaging (0.993)	K
c.3279C>A	p.D1093E	0	0	0	0	0	Probably damaging (0.99)	D
c.3350C>T	p.T1117I	0	0	0	0	0	Probably damaging (0.995)	T
c.3655G>C	p.A1219P	0.77	0	0.74	0	1.5	Probably damaging (0.991)	A
c.3812A>G	p.E1271G	0.8	0.35	0.71	0	21.8	Probably damaging (0.995)	E
c.4177G>A	p.V1393I	0	0	0	0	0	Benign (0.000)	VI
c.4217A>G	p.H1406R	0	0	0	0	0	Probably damaging (0.986)	r p q a t k e g n S D H
c.4248C>A	p.F1416L	0.79	0.1	0.79	0.1	0.27	Probably damaging (0.891)	F
c.4326C>A	p.N1442K	0	0	0	0	0	Possibly damaging (0.374)	a g d s T N
c.4467G>T	p.R1489S	0.85	0.43	0.78	0.24	12.8	Benign (0.047)	p h l s n a e T Q K R
c.4670A>G	p.Y1557C	0.85	0.13	0.80	0	8.8	Probably damaging (0.999)	f W Y
c.5138A>G	p.Q1713R	0	0	0	0	0	Probably damaging (0.997)	Q
c.5177C>A	p.T1726N	0	0	0	0	0	Probably damaging (0.880)	s A T
c.5646G>A	p.M1882I	0	0	0.75	0	37.4	Probably damaging (0.999)	M
c.6306C>A	p.D2102E	0	0	0	0	0	Probably damaging (0.99)	D
c.6718A>G	p.T2240A	0	0	0	0	0	Probably damaging (0.991)	T
c.160+2T>C		0.81	0.86	0.79	0	44.4		
c.1240−2A>G		0.82	0.81	0	0	81.5		
c.2382+1G>A		0.79	0.64	0	0	71.7		
c.2919−2A>G		0.9	0.92	0	0	90.9		
c.3522+5delG		0.87	0.57	0	0.18	63		
c.3523−1G>A		0.9	0.89	0	0	89		Splice site shift of 1 bp
c.3814−2A>G		0.91	0.9	0	0	90.6		
c.4352+1G>A		0.74	0.82	0	0	78		
c.4635−1G>T		0.86	0.89	0	0	87.5		New splice site 7 bp downstream
c.5312+1G>A		0.81	0.91	0	0	86.1		
c.5836−2A>C		0.89	0.87	0	0	88		
c.6387−1G>T		0.77	0.87	0	0	82		
c.6479+1G>A		0.82	0.87	0	0	85		
c.6479+1G>C		0.82	0.31	0	0	56.6		
c.1100−6T>A		0	0	0.9	0.93	91.6		Creates new splice site
c.351_352delAG	p.S119fs						Frameshift	
c.564delA	p.E189Cfs						Frameshift	
c.885delC	p.L296Cfs						Frameshift	
c.1374delA	p.T459Qfs						Frameshift	
c.3543delT	p.K1182Rfs						Frameshift	
c.3846delA	p.G1283Dfs						Frameshift	
c.4734delG	p.L1580*						Stop codon	
c.5932delA	p.T1979Qfs						Frameshift	
c.6317_6323del GCCGCAT	p.R2107_ M2108delfs						Frameshift	
c.121G>A	p.W41*						Stop codon	
c.318T>G	p.Y106*						Stop codon	
c.1906C>T	p.Q636*						Stop codon	
c.4639A>T	p.K1547*						Stop codon	

For SpliceSiteFinder and GeneSplicer, 1 is the highest score for splice site activity and 0 is the lowest. For PolyPhen, 1 is the most damaging and 0 is the least.

For SIFT, tolerated amino acid residues are listed. Uppercase letters indicate tolerated amino acids; lowercase letters indicate less tolerated amino acids.

* Nucleotide positions and protein translation correspond to CCDS747.1 and NP_000341.2, respectively.

STGD1 patients with the ABCA4 array started in 2000 (then with ~300 mutations); the number of mutations on the array has since doubled. The fact that only four mutations from all the array updates reoccurred shows that almost all new findings are rare. Three *ABCA4* mutations in three different samples identified by sequencing—p.P1486L, p.G1961E, and p.R2106C—represented ABCA4 array false negatives. These three samples were screened many years ago when the quality of the ABCA4 array might have not been at its current level. One variant, c.768G>T/p.V256V, identified by the array was not confirmed by NGS; therefore, it represented the only NGS false negative since it was confirmed by Sanger sequencing. From 53 samples in which the combined APEX/NGS analysis had detected two mutations, only two samples carried the same two mutations—p.G1961E and p.R2149*. This proves once more the extraordinary heterogeneity of the *ABCA4* alleles and the necessity of a cost-efficient full gene sequencing platform.

Fifteen benign *ABCA4* missense variants were also detected (Supplementary Table S3, http://www.iovs.org/lookup/suppl/

TABLE 2. Splice Site Analysis of Rare Synonymous and Rare Intronic Variants (>5 bp outside the actual splice site) Using Bioinformatics Software

Nucleotide Change	Protein	Splicing Score Original		Splicing Score for New Variant		Average Difference	Comments
		SpliceSite Finder-like	Gene Splicer	SpliceSite Finder-like	Gene Splicer		
c.141G>A	p.P47P	0.73	0.87	0.73	0.87	0.1	
c.513C>T	p.I171I	0	0	0	0	0	
c.618C>T	p.S206S	0.72	0.85	0.72	0.86	0.3	
c.1029T>C	p.N343N	0	0	0	0	0	
c.1500G>A	p.R500R	0	0.81	0	0.78	1.55	
c.1878G>A	p.A626A	0	0	0	0.76	38	
c.2964C>T	p.L988L	0.69	0	0.69	0	0	
c.4611G>A	p.T1537T	0	0.65	0	0	32.5	
c.6066A>G	p.A2022A	0.72	0	0	0	35.8	
c.6216T>C	p.S2072S	0	0.72	0	0.81	4.1	
c.6333C>T	p.N2111N	0	0.69	0	0.69	0.1	
c.6342G>A	p.V2114V	0	0.80	0.72	0.81	37	
c.6732G>A	p.V2244V	0.75	0	0.73	0.77	37.5	
c.5′UTR−92C>T		0	0	0	0	0	
c.302+20C>T		0.85	0.89	0.87	0.88	0.6	
c.302+68C>T		0.72	0.79	0.72	0.79	0.5	
c.443−45C>T		0	0	0	0	0	
c.570+20C>T		0.81	0	0.73	0	4	
c.859−41T>C		0	0.80	0.75	0.88	41.5	
c.859−9T>C		0.737	0.78	0.71	0.77	1.95	
c.1100−6T>A		0	0	0.9	0.93	91.6	New splice site
c.1239+34T>C		0.76	0.85	0.72	0.83	2.3	
c.1356+7_+8insA		0	0	0	0.65	32.5	
c.1554+83T>C		0	0	0	0	0	
c.1760+22G>T		0.72	0.8	0.72	0.8	0	
c.2383−13C>T		0.78	0.87	0.81	0.87	1.6	
c.2588−12C>G		0	0	0	0.6	30.2	
c.2588−33C>T		0	0	0	0	0	
c.2653+57C>T		0	0	0	0.8	40	
c.3191−20C>T		0	0	0	0	0	
c.3329−23T>C		0.75	0.82	0.74	0.82	0	
c.3329−34A>G		0.79	0	0	0	39.5	
c.3522+31C>A		0	0	0.71	0	35.5	
c.4253+12C>T		0	0	0	0	0	
c.4253+43G>A		0	0	0	0	0	
c.4254−11A>G		0	0	0	0	0	
c.4254−38G>A		0.8	0.84	0.8	0.84	0	
c.4352+32A>G		0	0.74	0	0.77	1.55	
c.4352+54A>G		0.763	0.85	0.76	0.85	0	
c.4352+69A>G		0	0	0	0.75	37.6	
c.4668−34G>A		0	0	0	0.67	33.5	
c.4668−38C>T		0	0.74	0	0.74	0	
c.4849−27G>A		0	0.84	0	0.83	0.1	
c.5018+8A>G		0	0	0	0	0	
c.5461−45G>C		0	0.82	0	0.71	5.6	
c.5585−35C>T		0	0	0	0.72	36	
c.5836−3C>A		0.89	0.87	0.8	0.78	9.7	
c.5836−60C>T		0	0	0	0	0	
c.6730−19G>A		0	0	0	0	0	
c.6817−85C>T		0	0.72	0	0.74	0.9	

For SpliceSiteFinder and GeneSplicer, 1 is the highest score and 0 is the lowest score for splice site activity.
* Nucleotide positions and translation correspond to CCDS747.1 and NP_000341.2, respectively.

doi:10.1167/iovs.11-8182/-/DCSupplemental), including two novel rare variants p.V1393I and p.R1489S (both detected once) that were predicted to be tolerated.

Analysis of Intronic Variants and Splice Sites

In addition to the 14 variants classified as splice affecting, because these occurred within 5 bp of exons (primarily at first or second base in IVS) (Table 1), 70 more variants were detected in intronic sequences adjacent to exons. Of these, 52 were detected in a few, or single, cases (frequency <1% in our cohort) with no frequency data available even for those in dbSNP (Supplementary Table S2, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental). Thirty-seven rare intronic variants, detected in samples with one or no mutation after sequencing, were subjected to in silico analysis (Table 2). As a result, one more intronic variant, c.1100–6T>A, was unequivocally classified as splice affecting. This variant creates a strong new splice acceptor 4 bp upstream of the wt splice site, which would result in frameshift and eventual premature stop codon (Supplementary Fig. S1, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental). Unfortunately, no family members were available for segregation analysis. The remaining variants resulted in no significant difference in splice site predictions (Table 2) and are, therefore, unlikely to be disease associated through this mechanism. Given that one or no disease-associated alleles were identified in many STGD1 patients, it is possible that some of these variants could have affected the ABCA4 protein. The best known example from this category in *ABCA4* is the c.5461–10T>C variant in intron 38, which is the third most frequent variant (found in 7.1% of STGD1 patients) after the p.G1961E and p.L541P/A1038V mutations in our study. The c.5461–10T>C variant always segregates with the disease phenotype in families, is rare in the general population (<0.001), and does not affect splicing.[8] Because there is no mutation in the *ABCA4* coding sequence on the same chromosome with the c.5461–10T>C variant, the latter has to be a disease-associated mutation, although its functional consequences remain unknown.

A similar intronic variant that does not affect splicing, c.4253+43G>A, in intron 28, was present in seven patients, all whom had only one other mutation identified in the *ABCA4* ORF. Screening of additional samples identified this variant in 1 of 45 patients with one mutation (a total of 8/118 patients), 1 of 204 patients with two mutations and 3 of 120 patients with no mutations. The variant was also found in 6 of 364 AMD patients and in 3 of 366 controls (those older than 60 years of age with no retinal pathology). Based on these data indicating that the variant is statistically significantly ($P = 0.0003$) found more often in STGD1 patients (almost exclusively in those with one definite mutation) and is twice as frequent in patients with age-related macular degeneration, it remains a good candidate for a disease-associated allele. Unfortunately, most of the patients harboring the c.4253+43G>A variant had simplex cases or there were not enough family members (affected and unaffected siblings) to perform definitive segregation analyses. Therefore, we could not yet unequivocally call the variant disease associated.

Analysis of Synonymous Variants

We also analyzed rare synonymous variants in the *ABCA4* ORF for possible disease association because these have been predicted to be frequently involved in disease.[19] Two possible ways synonymous variants can affect the protein and, therefore, can be associated with a disease are their effect on splicing and codon use. From 21 synonymous coding variants detected in this cohort (Table 3), 13 rare variants (frequency <1%) were further analyzed with splicing prediction programs. Of these, 12 of 13 are predicted to have no effect on splicing (Table 2). The p.V2244V variant is a result of a GTG>GTA change in the first codon of exon 49 that does not affect the splice acceptor but creates a new splice donor sequence in the very same intron-exon boundary (Supplementary Fig. S2, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental). A similar example is the disease-associated synonymous variant c.768G>T/p.V256V, in which the last codon of exon 6 is changed from GTG to GTT, which, according to prediction analysis, reduces the existing splice donor to an extent comparable to that for the new splice site score of the variant c.6732G>A/p.V2244V (Supplementary Fig. S3, http://www.iovs.org/lookup/suppl/doi:10.1167/iovs.11-8182/-/DCSupplemental). In addition, the p.V2244V

Table 3. Synonymous Variants Found in the *ABCA4* Gene

rs Number (if known)	Nucleotide Change	Protein	Heterozygotes (in 159 samples)	Homozygotes (in 159 samples)	MAF in dbSNP (if available)	MAF This Study
rs1801574	c.5682G>C	p.L1894L	88	15	0.20	0.37
rs2275029	c.5844A>G	p.P1948P	69	2	0.15	0.23
rs4147857	c.5814A>G	p.L1938L	69	2	0.15	0.23
rs1801555	c.6285T>C	p.D2095D	51	6	0.15	0.20
rs1762114	c.6069T>C	p.I2023I	27	3	0.10	0.10
rs1801359	c.6249C>T	p.I2083I	21	1	0.10	0.07
rs1801666	c.4203C>T	p.P1401P	18	2	0.07	0.07
rs4147831	c.1269C>T	p.H423H	17	—	0.10	0.05
rs77293072	c.6732G>A	p.V2244V	3	—	—	0.009
rs61754034	c.2964C>T	p.L988L	3	—	—	0.009
	c.1029T>C	p.N343N	2	—	—	0.006
rs4847281	c.141G>A	p.P47P	1	—	—	0.003
	c.513C>T	p.I171I	1	—	—	0.003
	c.618C>T	p.S206S	1	—	—	0.003
	c.1500G>A	p.R500R	1	—	—	0.003
rs61754023	c.1878G>A	p.A626A	1	—	—	0.003
	c.4611G>A	p.T1537T	1	—	—	0.003
	c.6066A>G	p.A2022A	1	—	—	0.003
	c.6216T>C	p.S2072S	1	—	—	0.003
	c.6333C>T	p.N2111N	1	—	—	0.003
rs61748520	c.6342G>A	p.V2114V	1	—	—	0.003

Nucleotide positions and protein translation correspond to CCDS747.1 and NP_000341.2, respectively.

variant was detected in three samples with only one disease-associated *ABCA4* allele.

One of the proposed mechanisms for functional effect of synonymous variants is the introduction of less frequent codons resulting in delayed translation and folding of the protein.[20] Examples of these mechanisms associated with a specific phenotype are described for other ABC transporters such as MDR1.[20,21] The primary influence on codon use in mammals is the local base composition of the gene.[22] Different codons can affect translation efficiency; it has been shown that G- and C-ending codons are more abundant in constitutive than in alternatively spliced exons in both *Drosophila* and humans.[23]

From the 13 rare, silent ABCA4 variants detected in this study, six resulted in changes from a more frequent codon to a less frequent codon, and two of them, c.6342G>A/p.V2114V and c.6732G>A/p.V2244V (discussed earlier), represent a change from the most frequent codon to the least frequent codon (Table 4). Therefore, the c.6732G>A/p.V2244V change may affect the protein in a dual fashion by affecting splicing and by slowing the protein translation. Both variants were also found in samples with only one *ABCA4* mutation. Because of the absence of family members for the four samples with the p.V2244V and p.V2114V variants, segregation analysis was not possible. Therefore, these two variants are classified as "suggestive" for disease association.

Discussion

Next-generation sequencing of a large cohort of STGD1 patients, who had been screened previously with the ABCA4 array, discovered many new mutations in the coding region of *ABCA4*. NGS revealed 57 new disease-associated variants in 59 patients, and three more possibly pathogenic intronic and synonymous variants in another 11 patients. Therefore, though the (updated) array still remains a cost- and time-efficient first-pass screening tool, sequencing on an NGS platform would be a much more comprehensive approach. Determining the pathogenicity of rare missense (and also splice-affecting and synonymous) variants remains a significant challenge, especially for the *ABCA4* gene, for which close to 700 possibly disease-associated variants have now been identified. Because more than half of these have been detected only once, the unequivocal classification for disease association, which is usually accomplished by segregation and functional analyses, presents an almost impossible task. Given that *ABCA4*-associated diseases are recessive, any given patient often represents the only affected member in a family with no siblings. Although determining the phase is often possible if parental samples are available, unequivocal segregation with the disease is often complicated or impossible. Functional analysis of the *ABCA4* variants is also complicated because *ABCA4* is expressed only in photoreceptors, which means no affected tissue can be obtained from patients, and because no direct functional test is available. Many frequent *ABCA4* variants have been analyzed indirectly, such as by testing their effects on protein yield, folding, and ATP-binding and ATPase activity assays in mostly in vitro systems[10,24] and *Xenopus laevis* models.[25] Performing these experiments for hundreds of rare variants is unrealistic at this time.

In the present study the new variants were analyzed in a multistep process. First, the frequent variants were removed from the analysis by filtering against public databases. Deleterious mutations (nonsense and ins/del) were considered pathogenic based on their truncating or frameshift effect, or both, on the ABCA4 protein. Missense mutations were analyzed with well-known in silico predictive programs, PolyPhen and SIFT, which have shown to be approximately 80% reliable in correctly predicting functional variants.[26] If family members were available, variants were analyzed for segregation with the disease in pedigrees (Fig. 2). Some variants were also screened in a large control cohort (364 persons) from our AMD studies, which includes ethnically matched controls older than 60 years of age without any retinal pathology as documented by thorough eye examination. However, given that all but seven new variants were found only once in 159 STGD1 patients, the screening of 364 controls had limited value. Rare synonymous changes in the *ABCA4* ORF were analyzed for their effect on splicing with in silico programs and for codon use. All rare intronic variants were analyzed with splicing prediction programs.

Screening of the *ABCA4* gene with any method is still far from 100% efficient. Even after complete sequencing of *ABCA4* coding region in patients with definitive clinical diagnosis of STGD1, approximately 25% to 30% of patients remain with one identified pathogenic mutation, and no mutations were found in approximately 15% to 20%. The three most likely reasons for not finding all mutations were that a small subset (~1%) of patients harbored CNVs unde-

Table 4. Codon Use Analysis for Rare Synonymous Variants

Protein	Patient Count	Pathogenic ABCA4 Alleles in Those Patients	Codon Change	Codon Use Change in ABCA4	rs Number (if known)	Conclusion
p.V2244V	3	All with 1 allele	GTG>GTA	70>14	rs77293072	From most frequent to least frequent
p.L988L	3	One with 2 alleles, Two with no alleles	CTC>CTT	65>33	rs61754034	From second frequent to third frequent
p.N343N	2	2 alleles, 1 allele	AAT>AAC	36>54	—	To most frequent
p.P47P	1	1 allele	CCG>CCA	13>44	rs4847281	From least frequent To most frequent
p.I171I	1	1 allele	ATC>ATT	79>43	—	From most frequent to less frequent
p.S206S	1	1 allele	AGC>AGT	49>16	—	From most frequent to less frequent
p.R500R	1	0 alleles	AGG>AGA	29>27	—	Neutral
p.A626A	1	1 allele	GCG>GCA	10>32	rs61754023	To more frequent
p.T1537T	1	0 alleles	ACG>ACA	17>32	—	To more frequent
p.A2022A	1	1 allele	GCA>GCG	32>10	—	To least frequent
p.S2072S	1	2 alleles	AGT>AGC	16>49	—	To most frequent
p.N2111N	1	2 alleles	AAC>AAT	54>36	—	To least frequent
p.V2114V	1	1 allele	GTG>GTA	70>14	rs61748520	From most frequent to least frequent

tected by PCR-based methods, a significant fraction of pathogenic mutations were outside the *ABCA4* coding sequences, and some patients had diagnoses of *ABCA4*-associated diseases that were phenocopies (diseases caused by mutations in other known or yet to be discovered genes).

CNVs are predicted to be rare in the *ABCA4* locus because several studies have found only a few cases (~1% of all patients) with large (entire exon or chromosomal segment) deletions that avoid PCR-based detection methods. However, for complete mutational scanning, CNV analysis with an array comparative genomic hybridization approach, or with multiplex ligation-dependent probe amplification, could be included.

The present study also determined that many pathogenic mutations are likely located outside the *ABCA4* ORF because the second mutation, required for the genetic diagnosis of STGD1, was not found in approximately half of all patients with one mutation. Although one could argue that some of these patients could be carrying the *ABCA4* variant by chance because of the high population frequency of *ABCA4* variants (estimated 1:20), it is highly unlikely that a patient with a clinical diagnosis of STGD1 and carrying one mutation does not have the second pathogenic variant. Detection of disease-associated variants outside the *ABCA4* coding sequences will be accomplished by sequencing of the entire 130-kb *ABCA4* genomic locus in patients with one identified mutation, a study that is in progress.

Finally, in patients in whom no disease-associated variants are found, whole exome or genome approaches can be used to determine new gene mutations that cause STGD-mimicking phenotypes. This approach can be preceded by sequencing known genes, such as *RDS/PRPH2* (gene for multifocal pattern dystrophy, 3 exons),[27,28] *ELOVL4* (dominant STGD-like disease gene, 6 exons),[29,30] *VMD2* (Best disease gene [recessive forms resemble STGD], 11 exons), *RS1* (retinoschisis gene, 6 exons),[31] and *CNGB3* (achromatopsia gene, 17 exons).[32] In our studies, however, we have not found disease-associated mutations in *RDS* or *ELOVL4* genes in 30 to 40 STGD patients with no mutations in *ABCA4*; therefore, the yield by this approach is expected to be limited.

The fraction of genocopies (i.e., clinical misdiagnoses) at a given clinic depends primarily on the depth of clinical analyses. At our centers patients have primarily undergone detailed clinical work-up with all the techniques shown in Materials and Methods; therefore, the fraction of genocopies is expected to be small. However, in an average clinic, diagnosis is based mainly on ophthalmoscopic examination (fundus photography) with only a few additional techniques (e.g., BCVA, OCT, microperimetry) yielding less stringent criteria for final diagnosis. Moreover, even exceptionally extensive clinical data are often not enough for pinpointing the possible genetic cause. As a reminder, depending on the severity of the *ABCA4* mutation and the stage of the disease diagnosed, *ABCA4*-associated pathology presents in a wide range of phenotypes from mild fundus flavimaculatus to CRD and even RP-like phenotypes. The latter two phenotypes are caused by tens of distinct genes.

Given the substantially overlapping phenotypes and several treatment options currently in late stages of preclinical development or in clinical trials, the correct and comprehensive molecular diagnosis of *ABCA4*-associated diseases is crucial. The NGS platform is a time- and cost-efficient tool to analyze large and variable genes simultaneously in large cohorts and could be used for diagnostic applications.

References

1. Allikmets R, Singh N, Sun H, et al. A photoreceptor cell-specific ATP-binding transporter gene (ABCR) is mutated in recessive Stargardt macular dystrophy. *Nat Genet.* 1997a;15:236–246.
2. Cremers FP, van de Pol DJ, van Driel M, et al. Autosomal recessive retinitis pigmentosa and cone-rod dystrophy caused by splice site mutations in the Stargardt's disease gene ABCR. *Hum Mol Genet.* 1998;7:355–362.
3. Maugeri A, Klevering BJ, Rohrschneider K, et al. Mutations in the ABCA4 (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *Am J Hum Genet.* 2000;67:960–966.
4. Martinez-Mir A, Paloma E, Allikmets R, et al. Retinitis pigmentosa caused by a homozygous mutation in the Stargardt disease gene ABCR. *Nat Genet.* 1998;18:11–12.
5. Shroyer NF, Lewis RA, Yatsenko AN, Lupski JR. Null missense ABCR (ABCA4) mutations in a family with Stargardt disease and retinitis pigmentosa. *Invest Ophthalmol Vis Sci.* 2001;42:2757–2761.
6. Allikmets R. Stargardt disease: from gene discovery to therapy. In: Tombran-Tink J, Barnstable CJ, eds. *Retinal Degenerations: Biology, Diagnostics and Therapeutics*. Totowa, NJ: Humana Press; 2007:105–118.
7. Maugeri A, van Driel MA, van de Pol DJ, et al. The 2588G→C mutation in the ABCR gene is a mild frequent founder mutation in the Western European population and allows the classification of ABCR mutations in patients with Stargardt disease. *Am J Hum Genet.* 1999;64:1024–1035.
8. Rivera A, White K, Stohr H, et al. A comprehensive survey of sequence variation in the ABCA4 (ABCR) gene in Stargardt disease and age-related macular degeneration. *Am J Hum Genet.* 2000;67:800–813.
9. Lewis RA, Shroyer NF, Singh N, et al. Genotype/phenotype analysis of a photoreceptor-specific ATP-binding cassette transporter gene, ABCR, in Stargardt disease. *Am J Hum Genet.* 1999;64:422–434.
10. Shroyer NF, Lewis RA, Yatsenko AN, Wensel TG, Lupski JR. Cosegregation and functional analysis of mutant ABCR (ABCA4) alleles in families that manifest both Stargardt disease and age-related macular degeneration. *Hum Mol Genet.* 2001;10:2671–2678.
11. Allikmets R. Simple and complex ABCR: genetic predisposition to retinal disease. *Am J Hum Genet.* 2000;67:793–799.
12. Jaakson K, Zernant J, Kulm M, et al. Genotyping microarray (gene chip) for the ABCR (ABCA4) gene. *Hum Mutat.* 2003;22:395–403.
13. Ernest PJ, Boon CJ, Klevering BJ, Hoefsloot LH, Hoyng CB. Outcome of ABCA4 microarray screening in routine clinical practice. *Mol Vis.* 2009;15:2841–2847.
14. Klevering BJ, Yzer S, Rohrschneider K, et al. Microarray-based mutation analysis of the ABCA4 (ABCR) gene in autosomal recessive cone-rod dystrophy and retinitis pigmentosa. *Eur J Hum Genet.* 2004;12:1024–1032.
15. Yatsenko AN, Shroyer NF, Lewis RA, Lupski JR. Late-onset Stargardt disease is associated with missense mutations that map outside known functional regions of ABCR (ABCA4). *Hum Genet.* 2001;108:346–355.
16. Yatsenko AN, Shroyer NF, Lewis RA, Lupski JR. An ABCA4 genomic deletion in patients with Stargardt disease. *Hum Mutat.* 2003;21:636–644.
17. Kong J, Kim SR, Binley K, et al. Correction of the disease phenotype in the mouse model of Stargardt disease by lentiviral gene therapy. *Gene Ther.* 2008;15:1311–1320.
18. Maiti P, Kong J, Kim SR, Sparrow JR, Allikmets R, Rando RR. Small molecule RPE65 antagonists limit the visual cycle and prevent lipofuscin formation. *Biochemistry.* 2006;45:852–860.
19. Chen R, Davydov EV, Sirota M, Butte AJ. Non-synonymous and synonymous coding SNPs show similar likelihood and effect size of human disease association. *PLoS One.* 2010;5:e13574.
20. Hunt R, Sauna ZE, Ambudkar SV, Gottesman MM, Kimchi-Sarfaty C. Silent (synonymous) SNPs: should we care about them? *Methods Mol Biol.* 2009;578:23–39.
21. Kimchi-Sarfaty C, Oh JM, Kim IW, et al. A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science.* 2007;315:525–528.
22. Kliman RM, Bernal CA. Unusual usage of AGG and TTG codons in humans and their viruses. *Gene.* 2005;352:92–99.

23. Iida K, Akashi H. A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes. *Gene.* 2000;261:93–105.
24. Sun H, Smallwood PM, Nathans J. Biochemical defects in ABCR protein variants associated with human retinopathies. *Nat Genet.* 2000;26:242–246.
25. Wiszniewski W, Zaremba CM, Yatsenko AN, et al. ABCA4 mutations causing mislocalization are found frequently in patients with severe retinal dystrophies. *Hum Mol Genet.* 2005;14:2769–2778.
26. Liu YH, Li CG, Zhou SF. Prediction of deleterious functional effects of non-synonymous single nucleotide polymorphisms in human nuclear receptor genes using a bioinformatics approach. *Drug Metab Lett.* 2009;3:242–286.
27. Grover S, Fishman GA, Stone EM. Atypical presentation of pattern dystrophy in two families with peripherin/RDS mutations. *Ophthalmology.* 2002;109:1110–1117.
28. Boon CJ, van Schooneveld MJ, den Hollander AI, et al. Mutations in the peripherin/RDS gene are an important cause of multifocal pattern dystrophy simulating STGD1/fundus flavimaculatus. *Br J Ophthalmol.* 2007;91:1504–1511.
29. Zhang K, Kniazeva M, Han M, et al. A 5-bp deletion in ELOVL4 is associated with two related forms of autosomal dominant macular dystrophy. *Nat Genet.* 2001;27:89–93.
30. Bernstein PS, Tammur J, Singh N, et al. Diverse macular dystrophy phenotype caused by a novel complex mutation in the ELOVL4 gene. *Invest Ophthalmol Vis Sci.* 2001;42:3331–3336.
31. Tsang SH, Vaclavik V, Bird AC, Robson AG, Holder GE. Novel phenotypic and genotypic findings in X-linked retinoschisis. *Arch Ophthalmol.* 2007;125:259–267.
32. Michaelides M, Aligianis IA, Ainsworth JR, et al. Progressive cone dystrophy associated with mutation in CNGB3. *Invest Ophthalmol Vis Sci.* 2004;45:1975–1982.