



A case–control collapsing analysis identifies retinal dystrophy genes associated with ophthalmic disease in patients with no pathogenic *ABCA4* variants

Charles J. Wolock, BA¹, Nicholas Stong, PhD¹, Chu Jian Ma, PhD², Takayuki Nagasaki, PhD², Winston Lee, MA², Stephen H. Tsang, MD, PhD², Sitharthan Kamalakaran, PhD¹, David B. Goldstein, PhD¹ and Rando Allikmets, PhD^{2,3}

Purpose: Variants in the *ABCA4* gene are causal for a variety of retinal dystrophy phenotypes, including Stargardt disease (STGD1). However, 15% of patients who present with symptoms compatible with STGD1/*ABCA4* disease do not have identifiable causal *ABCA4* variants. We hypothesized that a case–control collapsing analysis in *ABCA4*-negative patients with compatible symptoms would provide an objective measure to identify additional disease genes.

Methods: We performed a genome-wide enrichment analysis of “qualifying variants”—ultrarare variants predicted to impact protein function—in protein-coding genes in 79 unrelated cases and 9028 unrelated controls.

Results: Despite modest sample size, two known retinal dystrophy genes, *PRPH2* and *CRX*, achieved study-wide significance ($p < 1.33 \times 10^{-6}$) under a dominant disease model, and eight additional known retinal dystrophy genes achieved nominal significance ($p <$

0.05). Across these ten genes, the excess of qualifying variants explained up to 36.8% of affected individuals. Furthermore, under a recessive model, the cone–rod dystrophy gene *CERKL* approached study-wide significance.

Conclusion: Our results indicate that case–control collapsing analyses can efficiently identify pathogenic variants in genes in non-*ABCA4* retinal dystrophies. The genome-wide collapsing analysis framework is an objective discovery method particularly suitable in settings with overlapping disease phenotypes.

Genetics in Medicine (2019) 21:2336–2344; <https://doi.org/10.1038/s41436-019-0495-0>

Keywords: retinal dystrophy; *ABCA4* disease; collapsing analysis; exome sequencing

INTRODUCTION

The *ABCA4* gene is the most widely studied gene involved in retinal/macular dystrophies, with pathogenic variants in this adenosine triphosphate (ATP)-binding cassette transporter linked to a variety of phenotypes. These include recessive Stargardt disease (STGD1) (ref. ¹), cone–rod dystrophy (CRD),^{2,3} bull’s eye maculopathy (BEM),⁴ pattern-like dystrophy (PD),⁴ and retinitis pigmentosa (RP),² collectively now referred to as “*ABCA4* disease.” Previous studies have identified over 800 disease-associated *ABCA4* variants,⁵ the most frequent of which are detected in ~20% of STGD1 patients of European descent.⁶ Some variants are seen at much higher frequencies in other ethnic groups.^{7–9} In about 15% of patients with clinical diagnosis compatible with STGD1/*ABCA4* disease, sequencing of the entire *ABCA4* gene¹⁰ or even the entire genomic locus¹¹ reveals no disease-associated variants. It is likely that most of these cases represent STGD1-like phenotypes caused by variants in one

or more other genes.¹⁰ The quality of clinical diagnosis is an important factor in these analyses because the phenotypic overlap between STGD1, RP, CRD, PD, BEM, macular dystrophy (MD), and other types of retinal dystrophy can be substantial. However, the unequivocal diagnosis is achieved only by determining the genetic causality of the disease phenotype. In patients with *ABCA4*-like phenotypes and no causal *ABCA4* variants, uncovering the genetic cause of disease is not always straightforward.

One promising approach to identifying disease-causing genetic variants is exome sequencing (ES), which has been used to study many retinal dystrophies, including STGD1, RP, and CRD.¹² Most of these studies have examined families or small numbers of sporadic cases. In both the familial and sporadic frameworks, searching for variants in known disease genes is the primary method of analysis, though novel retinal disease genes, such as *RDH11* (ref. ¹³) have also been implicated.

¹Institute for Genomic Medicine, Columbia University, New York, NY, USA; ²Department of Ophthalmology, Columbia University, New York, NY, USA; ³Department of Pathology and Cell Biology, Columbia University, New York, NY, USA. Correspondence: Rando Allikmets (rla22@cumc.columbia.edu)

Submitted 30 May 2018; accepted: 12 March 2019
Published online: 30 March 2019

The increasing availability of ES data has allowed for the development of population-genetic approaches to detect disease associations in case-control, rather than familial, settings. For example, collapsing analyses test the association of the burden of rare, presumably deleterious variants between cases and controls, generally on a genic level. In collapsing analysis studies, these “qualifying variants” are considered to have identical effects on gene function regardless of where they occur in the gene. In recent years, collapsing studies have successfully pinpointed disease genes in various contexts, including amyotrophic lateral sclerosis (ALS)¹⁴ and idiopathic pulmonary fibrosis.¹⁵ It has also been shown that the collapsing analysis framework can identify epileptic encephalopathy genes implicated using trio de novo mutation analysis.¹⁶ The success of collapsing analysis thus far suggests that it could be a useful method for agnostically exploring associations in rare diseases, when causal variants are likely to be highly penetrant but very rare. Additionally, collapsing analysis offers an objective measure of statistical enrichment of putatively causal variants in genes in which disease associations are observed—a feature that we used in this study to catalog genes that are most important to interrogate in *ABCA4*-negative STGD1-like patients.

We implemented exome-wide gene-based collapsing analysis comparing ES data from 79 patients against a large cohort of unrelated control individuals to identify the genes contributing to disease status. Importantly, all 79 affected individuals were diagnosed with phenotypes compatible with *ABCA4* disease but were all negative for any disease-causing, pathogenic variants in the *ABCA4* locus, which eliminates the most likely genetic explanation for the observed phenotypes. Despite the modest sample size, under a dominant collapsing analysis model we identified two known retinal disease genes achieving study-wide significance ($p < 1.33 \times 10^{-6}$) and an additional eight known retinal disease genes at nominal significance ($p < 0.05$). Under a recessive model, there were no genes at study-wide significance, but the known CRD gene *CERKL* was the top hit with a p value of 3.29×10^{-5} . These results demonstrate the power of collapsing analysis to detect disease associations even with extremely limited sample size. Furthermore, we have shown that, even after discounting the possible effects of known *ABCA4* pathogenic variants, the diverse phenotypes of the patients in this study are potentially explainable by rare variants in other known retinal disease genes.

MATERIALS AND METHODS

Subjects and sequencing

We started with ES data generated from 96 cases and 14,090 controls, with controls selected from other Institute for Genomic Medicine (IGM) studies. Controls were known to not have ophthalmic disease, liver disease, kidney disease, metabolic disease, or ALS. We collected written informed consent at the time of recruitment. Patient collection and sharing of anonymized specimens for research was approved

by site-specific institutional review boards and ethics committees.

For the cases, patients whose phenotypes were compatible with those caused by pathogenic variants in the *ABCA4* gene (STGD1, CRD, PD, BEM, RP-like) were recruited and clinically examined during a 20-year period at Columbia University. All study-related procedures adhered to tenets set out in the Declaration of Helsinki. Patients presenting with *ABCA4* disease were enrolled into Columbia University Medical Center–approved protocol #AAAI9906 after providing written consent at the conclusion of a complete ophthalmic exam by a retina specialist. Clinical determination of *ABCA4* disease phenotype was based on the assessment of disease features on slit-lamp examination, fundus autofluorescence (AF, 488-nm and 787-nm), spectral domain–optical coherence tomography (SD-OCT), and full-field electroretinogram (ffERG) testing. Table 1 summarizes the criteria by which patients were initially diagnosed and classified as exhibiting characteristics consistent with *ABCA4* disease. Disease onset in *ABCA4* disease is defined as the decade within which patients report visual symptoms associated with their condition. Disease inheritance was determined through patient-reported family histories or through clinical exams if relatives were available. The clinical manifestation of *ABCA4* disease is profoundly heterogeneous and can often exhibit phenotypic overlap with other inherited retinal conditions such as PD, BEM, occult macular dystrophy, and CRD, among others. Clinically ubiquitous features most often associated with *ABCA4* disease include increased levels of fundus autofluorescence, presence of yellow flecks across the posterior pole, progressive but confined chorioretinal atrophy, and peripapillary sparing. Variation in these fundus characteristics provided the most significant means by which patients were categorized. The presence of flecks on clinical exam was indicative of STGD1 or PD but their absence was not contraindicative, particularly in younger patients. Most diagnoses varied in the preservation of cone-specific response on ffERG testing; however, cases who exhibited reduced rod-specific relative to cone responses were excluded.

Sequencing of the *ABCA4* gene in all cases prior to ES was performed using the Illumina TruSeq Custom Amplicon protocol (Illumina, San Diego, CA).^{10,11} Sequencing of the entire *ABCA4* genomic locus at an average depth of 100× was performed in some patients using the Illumina TruSeq Custom Amplicon protocol as described previously.¹¹ Selected patients were also screened for copy-number variation by custom array comparative genomic hybridization (aCGH) arrays (Agilent Technology), in an 8×60 K format, as described previously.¹¹ All cases included in this study did not carry any possible disease-causing *ABCA4* variants and were subjected to ES as described below.

For ES, all samples were sequenced after exome capture using SeqCap EZ Exome v2 and SeqCap EZ Exome v3. Sequencing was performed at multiple sites, but we processed all data using the alignment and annotation pipeline at the IGM.

Table 1 Diagnostic summary of patient groups categorized within the clinical spectrum of ABCA4 disease

Diagnosis	Cases	Disease onset (decade)	Reported inheritance	Fundus characteristics		fERG		
				Spatial extent	Pigment deposition	Flecks	Cone	Rod
Stargardt disease	34	1st–3rd	AR	Macula–periphery	Late stage	Present ^a	WNL→↓↓	WNL ^b
Cone–rod dystrophy	24	1st–2nd	AR, AD	Macula–periphery	Late stage	Absent	WNL→↓↓	WNL→↓ ^b
Pattern dystrophy	9	3rd–5th	AD	Macula	None	Present ^a	WNL	WNL
Bull's eye maculopathy	8	NS	AR, AD, XL	Macula	None	Absent	WNL→↓↓	WNL→↓
Maculopathy	2	NS	NS	Macula	None	Absent	NS	NS
Other	2	NS	NS	NS	NS	Absent	NS	NS

AD autosomal dominant, AR autosomal recessive, fERG full-field electroretinogram, NS nonspecific, WNL within normal limits, XL X-linked recessive, ↓↓ severe waveform attenuation, ↓ moderate waveform attenuation. ^aThe absence of flecks may also be a characteristic feature at early stage presentation in this condition. ^bModerate to severe attenuation of rod-specific fERG responses may occur late in the disease although never lower relative to cone-specific responses.

IGM bioinformatics pipeline

We quality-filtered the raw sequence data using CASAVA (Illumina). We processed all data using a consistent alignment and variant calling pipeline, which consisted of primary alignment to the Human Reference Genome (NCBI Build37/hg19) and duplicate marking using the Dynamic Read Analysis for Genomics (Edico Genome, San Diego, CA) followed by variant calling using best practices outlined in the Genome Analysis Toolkit¹⁷ (GATK v3.6) (Supplemental Methods for further details). Variants for analysis were restricted to the consensus coding sequencing public transcripts (CCDS release 20) plus 2-bp intronic extensions. We further required variants to meet certain quality standards (Supplemental Methods). We annotated variants using ClinEff (DnaMiner) with Ensembl-HGRCh37.73 annotations.¹⁸

Finally, we excluded variants if they demonstrated imbalanced sequencing coverage between cases and controls according to a binomial test of independence between case–control status and coverage. We used the proportion of total samples that were cases as *p* in the binomial framework and the total number of samples that had at least 10× coverage at the genomic site in question as *n*. We performed a two-sided binomial test on the number of case samples with 10× coverage at that site, excluding variants at sites with binomial test *p* value < 0.05. Unlike the site-coverage harmonization in previous collapsing analysis studies,^{15,16} which is performed on all sites across CCDS regions (approximately 33 million), the binomial method is only performed at sites with an otherwise qualifying variant.

Quality control, relatedness check, and principal component analysis

We removed exomes with ambiguous sex as determined by X:Y coverage ratios. We also removed samples with contamination 7% or greater as determined by VerifyBamID¹⁹ and samples with less than 90% of CCDS bases covered at 10× or greater. After these steps, we retained 92 of 96 cases and 10,963 of 14,090 controls.

Before running the collapsing analyses, we implemented sample-level pruning procedures to minimize systemic bias in the data that might lead to spurious association or reduced power to detect real association.

First, we removed related individuals. To do this, we generated genotype data on an LD-pruned (*r*² threshold 0.1) list of 12,840 single-nucleotide polymorphism (SNPs) extracted from exome sequencing data and used KING²⁰ to calculate pairwise kinship coefficients for all subjects. We used a kinship coefficient of 0.0884 (corresponding to the minimum kinship coefficient for second-degree relations or closer) as a threshold and removed samples introducing relatedness to the cohort. We preferentially retained cases over controls, with ties broken randomly. After this step, we retained 92 of 92 cases and 9545 of 10,963 controls.

Next, we performed principal component analysis (PCA) using EIGENSTRAT²¹ on the 92 cases and 9545 controls with

the same SNP list as above. For five iterations, we removed outliers with a distance of six standard deviations or more from the mean on the top ten principal components. At this step, we retained 79 of 92 cases and 9028 of 9545 controls. Plots for the first three principal components calculated on the remaining samples are shown in Figure S1.

Collapsing analysis

To identify genes associated with retinal dystrophy under the case-control association analysis framework, we performed a genome-wide search for enrichment of “qualifying variants” (QVs) in either cases or controls across all protein-coding genes (using the same CCDS regions plus 2-bp extensions described above). We admitted QVs to analysis according to a set of criteria (in addition to quality filters described above) based on allele frequency and functional predictions, designed to capture pathogenic variants. For variant functional effect, we defined loss-of-function (LoF) variants to be variants with any one of the following annotations: `stop_gained`, `frame_shift`, `splice_site_acceptor`, `splice_site_donor`, `start_lost`, `exon_deleted`. The dominant and recessive models included both LoF and missense variants. For the dominant model, we determined missense variants to be qualifying if they were assessed as not benign (“probably damaging,” “possibly damaging,” “unknown”) by PolyPhen-2 (HumVar).²² We also required QVs to meet certain allele frequency (AF) thresholds. We calculated the “leave one out” (LOO) AF in our cohort by excluding the variant in question and then computing the AF of that variant in the remaining samples (both cases and controls). We required the LOO AF of the variant to fall at or below the given threshold (0.0001 for the dominant model and 0.01 for the recessive model). For the external data set AF threshold, we required the variant’s AF to fall at or below the cutoff (0.0001 for the dominant model and 0.01 for the recessive model) in each of eight gnomAD populations (global, African/African-American, Latino/Admixed American, East Asian, Finnish, Non-Finnish European, South Asian, and Other (population not assigned)).²³ After applying all of the filtering criteria, the dominant model included 383,452 total variants for analysis (348,653 unique variants). For the recessive model, there were 7275 qualifying variants (5001 unique variants).

For each gene, we assigned a qualification state indicator variable (1/0 states) to each individual based on the presence of at least one qualifying genotype in that gene (state 1) or no qualifying genotype in that gene (state 0). For the dominant model, we defined a qualifying genotype as either heterozygous or homozygous for a QV. For the recessive model, we defined a qualifying genotype as homozygous for a QV. Under this framework, for a given gene, a qualifying sample (case or control) was defined to be a sample (case or control) with an indicator variable status of 1. For the dominant model, we used the two-tailed Fisher’s exact test (FET) (SciPy module,²⁴ Python 2.7.7) to evaluate the statistical significance of association on each gene. For the recessive model, to address the potential confounding effect of population

stratification at higher MAF thresholds, we constructed a logistic regression model using the first ten EIGENSTRAT principal components as covariates. To account for bias due to small numbers of QVs in these recessive logistic regression models, we applied a Firth correction with profile likelihood confidence intervals.^{25,26} For both models, with 18,852 genes being tested under two models, we used the Bonferroni multiple-test correction to set a study-wide significance threshold of $p = 1.33 \times 10^{-6}$. We performed all logistic regression analyses using R version 3.2.5.

Quantile-quantile plots and genomic inflation factor (λ)

We generated quantile-quantile plots using a permutation-based empirical expected probability distribution. For the dominant model’s qualification state matrix (genes \times samples), we randomly permuted the case-control labels associated with each column of qualification states. After each permutation of the 79 case and 9028 control labels, we recomputed the FET p values for each gene and ordered the p values from lowest to highest. We repeated this process 1000 times and calculated the mean of each rank-ordered value across the 1000 permutations, i.e., the average 1st rank p value, the average 2nd rank p value, etc. These averages represent the empirical estimates of the ordered expected p values. This empirical expected p value distribution does not depend on the assumption that the p values are uniformly distributed under the null.

To compute the permutation-based expected p value distribution for the recessive model Firth logistic regression, due to the presence of the EIGENSTRAT-determined ancestry covariates, we implemented permutation using the R package “BiasedUrn” (cran.r-project.org/web/packages/BiasedUrn/) to maintain the confounding role of covariates in each permuted data set while the association between genotype and disease was broken.²⁷ We performed the permutation 1000 times and calculated the empirical expected p value distribution as described above.

We estimated the λ inflation factor in a manner similar to the procedure defined in the “estlambda” method in the genABEL R package (cran.r-project.org/web/packages/GenABEL/). Following the removal of data points corresponding to observed p values of study-wide significance and data points where either the observed or expected p value was 1, we calculated the “inverse” of each p value by subtracting it from 1. We then transformed these “inverse” observed and expected p values using χ^2 quantile function (with one degree of freedom). We performed a least-squares linear regression on the transformed observed and expected p values. The λ factor was taken to be the slope of this regression line.

RESULTS

To identify genes associated with rare eye disease, we used ES data from 79 cases and 9028 controls. The affected individuals varied in clinical phenotype (Table 1), but all received diagnoses compatible with the clinical presentation of ABCA4 disease. Nearly three-quarters of cases were initially classified

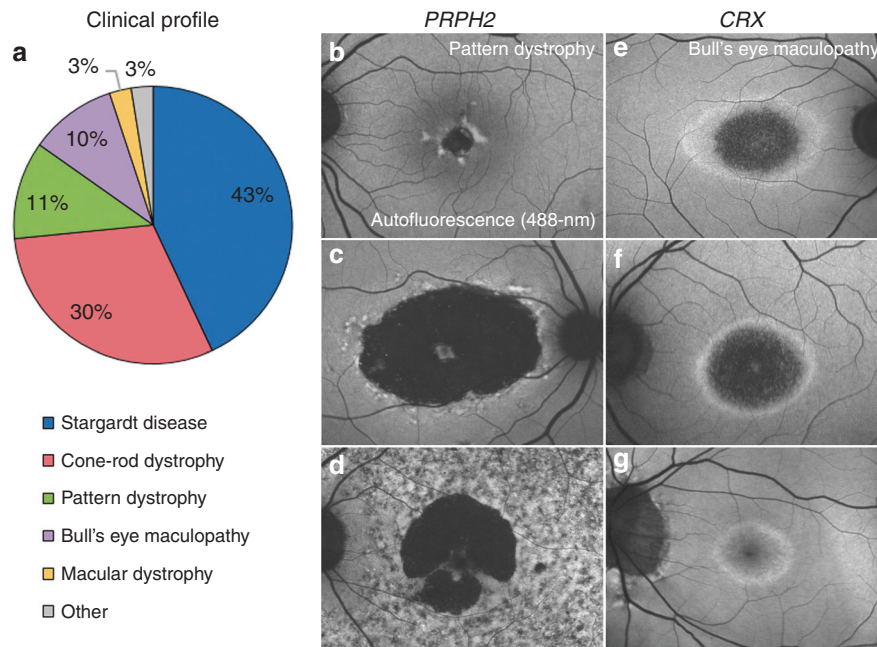


Fig. 1 Clinical and diagnostic profile of *ABCA4*-negative study cases. The study cohort consisted predominantly of Stargardt disease (43%) and cone-rod dystrophy (30%). Approximately one-fifth of cases exhibited features consistent with pattern dystrophy (11%) and bull's eye maculopathy (10%) (a). Cases presenting with nonspecific, early-onset macular disease classified as macular dystrophy (3%) and those with suspected *ABCA4*-related retina-wide degeneration or retinitis pigmentosa-like phenotypes were classified as "other" (3%). Cases in whom disease-causing variants in *PRPH2* were identified by collapsing analysis exhibited a range of phenotypes on autofluorescence (488-nm excitation) consistent with pattern or butterfly macular dystrophy (b–d). All cases with causal variants identified in *CRX* presented with elliptical bull's eye maculopathies and a cone-rod pattern of attenuation on full-field electroretinogram (e–g).

as either STGD1 or CRD (combined 73%), although BEM (10%), PD (11%), MD (3%), and other cases with generally overlapping *ABCA4* features (3%) were included (Fig. 1). All cases were determined negative for any disease-causing *ABCA4* pathogenic variants. We ran a standard gene-level collapsing analysis comparing these two groups. We coded individuals using the presence/absence of qualifying genotypes, where the qualifying criteria were defined according to either a dominant or a recessive model (see "Materials and methods.")

In the dominant model, two genes, *PRPH2* and *CRX*, demonstrated enrichment of QVs at a level that achieved study-wide significance ($p = 1.33 \times 10^{-6}$) using FET p values (Fig. 2a). We identified qualifying genotypes in *PRPH2* in 5.06% of cases versus 0.022% of controls (odds ratio [OR] >240; FET p value 7.77×10^{-8}). Previously, *PRPH2* has been reported to exhibit a dominant effect in PD,²⁸ RP,²⁹ MD,³⁰ CRD,³¹ and central areolar choroidal dystrophy.³² In *CRX*, we identified qualifying genotypes in 5.06% of cases versus 0.067% of controls (OR >80; FET p value 1.06×10^{-6}). Previous studies have implicated *CRX* in dominant forms of CRD,^{33,34} RP,³⁴ and Leber congenital amaurosis.³⁴ Cases in which *PRPH2* and *CRX* were indicated to be causal presented with late-onset STGD1 and BEM (with cone-rod dysfunction on fERG), respectively (Fig. 1). We also identified eight additional known retinal disease genes reaching nominal significance ($p = 0.05$). Of these, *PROM1* (ref.³⁵) and *RPE65* (ref.³⁶) have been reported to harbor variants exhibiting

dominant disease effects. To ascertain the proportion of disease burden in our cohort that could be explained by genetic variation in these genes, we calculated the excess of qualified individuals in each of the ten genes (Table 2). We subtracted the frequency of qualified controls from the frequency of qualified cases to quantify the potential percent of cases explained by each gene. Taken together, the qualified individual excess in *PRPH2* and *CRX* potentially explains 10.03% of the affected individuals, while the excess across all ten nominally significant retinal disease genes potentially explains 36.78% of affected individuals.

We also performed a recessive model collapsing analysis (see "Materials and methods") with a more permissive minor allele frequency (MAF) threshold. To better account for artifacts arising from population stratification effects at the higher MAF threshold, we used logistic regression with ancestry covariates (the first ten EIGENSTRAT principal components) in place of FET (Fig. 2b). This analysis identified *CERKL* as the top gene, which qualified in 2.53% of cases and 0.01% of controls (OR >234; logistic regression p value 3.29×10^{-5}). Both qualified *CERKL* cases carried the same homozygous stop-gain variant, and both were successfully identified by the diagnostic analysis. *CERKL* has been implicated in recessive cone-rod dystrophy,³⁷ which was the clinical diagnosis in both qualified cases. The top four identified genes and specific, determined pathogenic, variants are detailed in Table 3. Causality was determined by a combination of in silico methods including previous

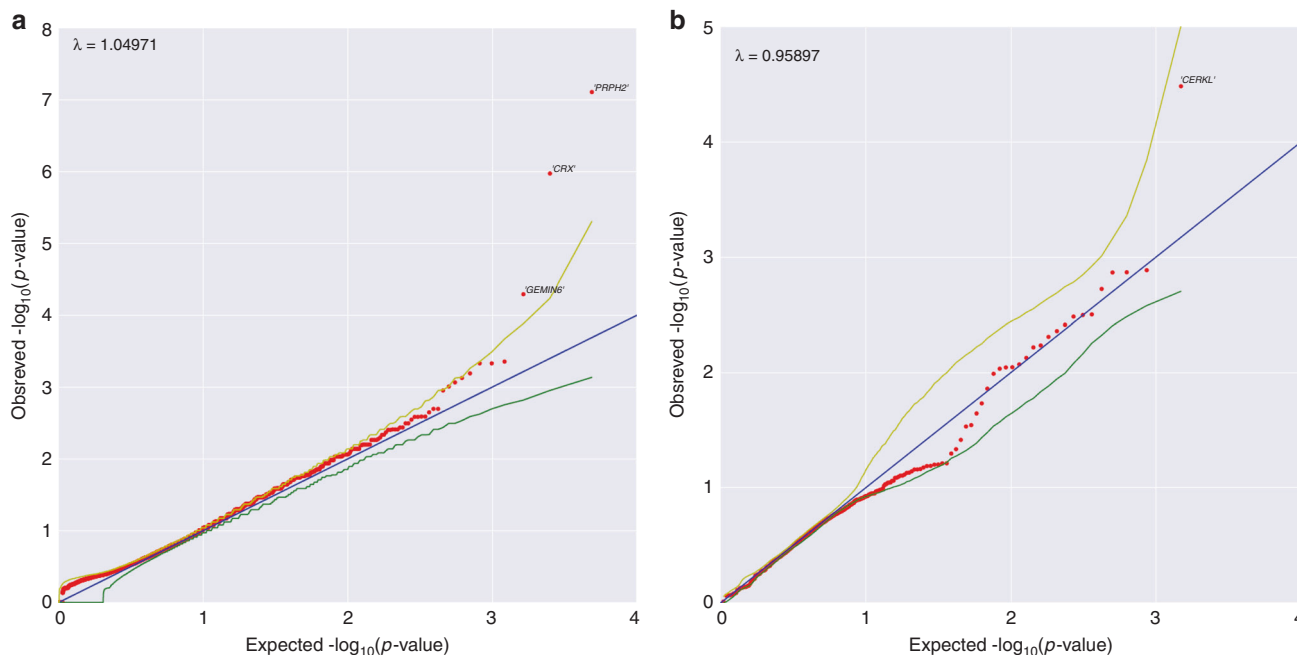


Fig. 2 Dominant model and recessive nonsynonymous model collapsing analysis results. (a) Quantile–quantile (QQ) plot for exome-wide gene-based collapsing analysis under the dominant genetic model. The y-axis represents the $-\log_{10}$ of the observed Fisher’s exact test (FET) p values (sorted). The x-axis represents the $-\log_{10}$ of the permutation-based expected FET p values (sorted). The red dots represent the data points, while the blue line is the diagonal with slope 1. The green and yellow lines represent permutation-based 95% confidence intervals. Data points falling outside the 97.5th percentile bound are labeled with corresponding gene symbols. *PRPH2* and *CRX* reached study-wide significance. (b) QQ plot for the collapsing analysis under the recessive n model. The y-axis represents the $-\log_{10}$ of the observed p values (sorted) evaluated in the Firth logistic regression adjusting for the first ten ancestry principal components. The x-axis represents the $-\log_{10}$ of the expected p values (sorted) evaluated in the same logistic regression model with permutation (“BiasedUrn”). The red dots represent the data points, while the blue line is the diagonal with slope 1. The green and yellow lines represent permutation-based 95% confidence intervals.

Table 2 Collapsing analysis results for nominally significant retinal disease genes

Rank	Gene	FET p value	Qualified case frequency	Qualified control frequency	QV excess ^a
1	<i>PRPH2</i>	7.77E-08	0.0506	2.22E-04	0.0504
2	<i>CRX</i>	1.06E-06	0.0506	6.65E-04	0.0499
11	<i>RPE65</i>	0.0011	0.038	0.0023	0.0357
17	<i>CDHR1</i>	0.0026	0.038	0.0032	0.0348
23	<i>C2orf71</i>	0.0037	0.038	0.0037	0.0343
33	<i>PROM1</i>	0.0053	0.038	0.0042	0.0338
84	<i>EYS</i>	0.0131	0.0633	0.0174	0.0459
115	<i>GPR179</i>	0.0179	0.038	0.0068	0.0312
141	<i>CC2D2A</i>	0.0219	0.038	0.0073	0.0307
299	<i>INVS</i>	0.0469	0.0253	0.0042	0.0211
Total					0.3678

FET Fisher’s exact test, QV qualifying variant.

^aQV excess is calculated as the qualified control frequency subtracted from the qualified case frequency.

knowledge of causality, MAF in matched general population, analysis with predictive software (e.g., CADD), and segregation with the disease in families if family members were available.

DISCUSSION

Comparing 79 *ABCA4*-negative retinal dystrophy cases and 9028 controls using gene-based collapsing analysis, we identified *PRPH2* and *CRX*, two known retinal disease genes, at study-wide significance. Combining these with eight additional known retinal disease achieving nominal significance, we observed a 36% excess of qualifying genotypes in cases. Therefore, in this prescreened cohort, over one-third of non-*ABCA4* retinal dystrophy is attributable to damaging variants in previously characterized retinal disease. The collapsing analysis framework is intended to capture the effect of rare protein-altering or protein-truncating variants that, unlike most standard diagnostic analyses, which mine previously reported qualifying pathogenic variants, does not depend on a priori knowledge of disease association and is therefore not dependent on curated databases such as ClinVar and HGMD. This is especially important because of the phenotypic variability of the affected individuals in this study: we analyzed a group of patients under six distinct clinical diagnoses. Within the group of eight patients who harbored qualifying *CRX* and *PRPH2* variants, the clinical diagnoses included STGD1 (3/8), BEM (3/8), CRD (1/8), and PD (1/8) diagnoses. The genetic overlap between retinal dystrophy phenotypes is widely acknowledged (e.g.,³⁸), and our study underlines the value of grouping various retinal dystrophy phenotypes to maximize the power to detect disease

Table 3 All identified pathogenic variants in *PRPH2*, *CRX*, *PROM1*, and *CERKL* genes in study cases^a

Patient ID	Clinical phenotype	Gene	Transcript	cDNA	Protein	Zygoty	gnomAD	MAF ^c	CADD ^d	Mode	Segregation	Reported	Pathogenic criteria ^f
3239	STGD1	<i>PRPH2</i>	ENST00000230381	c.584G>A	p.Arg195Gln	het	—	—	32.0	AD	+	+	PM2, PM5
3459	PD	<i>PROM1</i>	ENST00000447510	c.1117C>T	p.Arg373Cys	het	—	—	16.9	AD	+	+	PS1, PM2
3672	STGD1	<i>PRPH2</i>	ENST00000230381	c.246C>A	p.Cys82*	het	—	—	36.0	AD	+	+	PM2, PM4
3746	STGD1	<i>PRPH2^b</i>	ENST00000221996	c.262A>G	p.Lys88Glu	het	—	—	26.9	AD	+	+	PM2, PM5
3830	BEM	<i>CRX</i>	ENST00000230381	c.749_769delinsCATGA	p.Cys250SerfsTer8	het	—	—	0.0	AD	+	+	PM2, PM4
3922	STGD1	<i>PROM1</i>	ENST00000221996	c.268C>T	p.Arg90Trp	het	1.3E-04	—	34.0	AD	+	+	PS1
4246	PD	<i>PROM1</i>	ENST00000447510	c.303+1G>A	p.?	het	0	—	24.6	AD	+	+	PM1, PP4
4337	STGD1	<i>PROM1</i>	ENST00000447510	c.303+2T>C	p.?	het	—	—	23.8	AD	+	+	PM1, PP4
4369	PD	<i>PRPH2</i>	ENST00000447510	c.400C>T	p.Arg134Cys	het	3.6E-05	—	35.0	AD	+	+	PM3, PP3
4416	BEM	<i>CRX</i>	ENST00000230381	c.582-1G>A	p.?	het	—	—	27.1	AD	+	+	PS1
4539	CRD	<i>PROM1</i>	ENST00000221996	c.258G>T	p.Trp86Cys	het	—	—	27.5	AD	+	+	PM1, PM2
4548	BEM	<i>CRX</i>	ENST00000447510	c.1157T>A	p.Leu386*	het	—	—	35.0	AR	+	+	PS1, PM3
4619	CRD	<i>CERKL</i>	ENST00000221996	c.1557C>A	p.Trp519*	het	7.0E-05	—	35.0	AR	+	+	PS1, PM3
4645	CRD	<i>CERKL</i>	ENST00000339098	c.847C>T	p.Arg283*	hom	5.5E-04	—	27.7	AR	—	+	PS1, PM2
		<i>CERKL</i>	ENST00000339098	c.847C>T	p.Arg283*	hom	5.5E-04	—	28.6	AR	+	+	PS1, PM2

AD autosomal dominant, AR autosomal recessive, BEM bull's eye maculopathy, CADD Combined Annotation Dependent Depletion (<http://cadd.gs.washington.edu/home>), cDNA complementary DNA, CRD cone-rod dystrophy, het heterozygous, hom homozygous, MAF minor allele frequency, MD macular dystrophy, PD pattern dystrophy, PR pattern dystrophy, RP retinitis pigmentosa, STGD1 Stargardt disease.

^aNone of these variants were observed in study controls.

^bPredicted change resulting from three frameshift variants on the same allele.

^cMAF in the non-Finnish European population as all presented cases are of non-Finnish European ancestry and represent the highest MAF except for 2 variants labeled as 0. The *PROM1* variant c.303+1G>A is detected once in the "Other" population and the *CRX* p.Arg69His is detected once in the African population. Variants absent from gnomAD database are indicated by a dash (—).

^dCADD score ≥ 20 suggests pathogenicity.

^eNew pathogenic variants at the same amino acid position as previously reported.

^fBased on guidelines for interpretation of sequence variants (Richards et al., Genet Med. 2015).

associations. In the clinic and in diagnostic analyses, these diagnosed conditions may be treated as distinct, but their shared genetic etiology becomes clear in a pooled collapsing analysis. Our findings support the hypothesis that genes besides *ABCA4* cause disease in many of the 15–20% of patients with STGD1-like phenotypes but no identifiable pathogenic *ABCA4* variant.¹⁰ Of course, over 60% of the cases are not directly explained by qualifying variant excess in nominally significant known eye disease genes. Many causal single variants/genes (e.g., *PROM1*, Table 3) do not reach statistical significance due to the small cohort size and fraction of each gene/variant. Furthermore, some of these cases likely harbor causal variants in other, still unknown, retinal disease genes.

In the context of retinal dystrophies, many of which are known to be recessive, the recessive analysis provides a useful complement to the dominant analysis. Performing collapsing analyses under a recessive model necessitates the use of a more permissive allele frequency threshold, which often introduces inflation into the analysis due to population stratification within the cohort subpopulations. However, our recessive analysis shows no inflation and a clear, though not study-wide, enrichment of qualified samples for *CERKL*, a gene sometimes known to cause recessive CRD. This result, especially with such a small case cohort, is a favorable sign for the potential utility of recessive model collapsing analysis. Dominant effects are more evident in this study; however, it is likely that many of the unsolved cases harbor very rare biallelic disease-causing variants.

Among the noncarriers of *PRPH2*, *CRX*, and *CERKL* qualifying genotypes, the cause of disease could also be more complex than damaging variants in single genes. However, it is also worth noting that collapsing analysis increases in statistical power with larger sample sizes. The 79 cases in this study represent an extremely limited sample size for this method. Given the magnitude of signal observed in *PRPH2*, *CRX*, and *CERKL*, we anticipate that the genes in this analysis demonstrating strong, non-study-wide significance could reach the study-wide threshold under investigation of a larger cohort. Furthermore, we might expect to uncover additional retinal dystrophy genes that explain even smaller fractions of the patient population. Our investigation indicates that, at least in this setting, collapsing analysis can allow researchers to leverage a large data set of unaffected individuals to draw conclusions from fewer than 80 phenotypically heterogeneous cases. Though larger sample sizes are desirable, they are not absolutely necessary, which has clear implications for discovery strategies in a variety of rare genetic diseases.

In conclusion, our analyses indicate that rare missense and LoF variants in *PRPH2* and *CRX* are significantly associated with a broad retinal dystrophy phenotype (encompassing STGD1, CRD, PD, BEM, MD, and other diagnoses) in patients without a disease-causing *ABCA4* variants. These results demonstrate the power of collapsing analysis in a phenotypically heterogeneous disease setting and suggest that a substantial portion of non-*ABCA4* retinal dystrophy can be

attributed to previously known eye disease genes. Furthermore, we demonstrate the general utility of using collapsing analysis as an objective method to understand the heterogeneity and interplay of genetics and disease phenotypes. Collapsing analyses leverage the power derived from a large control population to identify enrichment of qualifying signal in the genes associated with disease phenotypes and therefore are not dependent on published literature or curated gene and variant lists. Even in this study of retinal disease, where there is an abundance of clinically reported pathogenic variants, collapsing analyses provide the initial, significant explanatory evidence of disease burden in known retinal disease genes, which can be then analyzed in detail in follow-up studies. In disease phenotypes that are poorly understood beyond the primary disease driver gene, collapsing analyses may offer a powerful and economical approach to expand our understanding of the genetic basis of the disease with limited sample sizes.

SUPPLEMENTARY INFORMATION

The online version of this article (<https://doi.org/10.1038/s41436-019-0495-0>) contains supplementary material, which is available to authorized users.

ACKNOWLEDGEMENTS

This study was supported, in part, by grants from the National Eye Institute/NIH EY028203 and EY019007 (Core Support for Vision Research) and unrestricted funds from Research to Prevent Blindness (New York, NY) to the Department of Ophthalmology, Columbia University.

DISCLOSURE

DBG is a founder and holds equity in Pairnomix and Praxis, serves as a consultant to AstraZeneca, and has research supported by Janssen, Gilead, Biogen, and UCB.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

REFERENCES

- 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.* 1997;15:236–246.
- Cremers FPM, Van De Pol DJR, 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.
- Maugeri A, Klevering BJ, Rohrschneider K, et al. Report mutations in the ABCA4 (ABCR) gene are the major cause of autosomal recessive cone-rod dystrophy. *Am J Hum Genet.* 2000;67:960–966.
- Duncker T, Tsang SH, Lee W, et al. Quantitative fundus autofluorescence distinguishes ABCA4-associated and non-ABCA4-associated bull's-eye maculopathy. *Ophthalmology.* 2015;122:345–355.
- Cornelis SS, Bax NM, Zernant J, et al. In silico functional meta-analysis of 5,962 ABCA4 variants in 3,928 retinal dystrophy cases. *Hum Mutat.* 2017;38:400–408.
- Burke TR, Fishman GA, Zernant J, et al. Retinal phenotypes in patients homozygous for the G1961E mutation in the ABCA4 gene. *Invest Ophthalmol Vis Sci.* 2012;53:4458–4467.
- Zernant J, Collison FT, Lee W, et al. Genetic and clinical analysis of ABCA4-associated disease in African American patients. *Hum Mutat.* 2014;35:1187–1194.
- Lee W, Schuerch K, Zernant J, et al. Genotypic spectrum and phenotype correlations of ABCA4-associated disease in patients of south Asian descent. *Eur J Hum Genet.* 2017;25:735–743.
- Guymier RH, Héon E, Lotery AJ, et al. Variation of codons 1961 and 2177 of the Stargardt disease gene is not associated with age-related macular degeneration. *Arch Ophthalmol.* 2001;119:745–51.
- Zernant J, Schubert C, Im KM, et al. Analysis of the ABCA4 gene by next-generation sequencing. *Investig Ophthalmol Vis Sci.* 2011;52:8479–8487.
- Zernant J, Xie YA, Ayuso C, et al. Analysis of the ABCA4 genomic locus in Stargardt disease. *Hum Mol Genet.* 2014;23:6797–6806.
- Gupta S, Chatterjee S, Mukherjee A, Mutsuddi M. Whole exome sequencing: uncovering causal genetic variants for ocular diseases. *Exp Eye Res.* 2017;164:139–150.
- Xie Y, Lee W, Cai C, et al. New syndrome with retinitis pigmentosa is caused by nonsense mutations in retinol dehydrogenase RDH11. *Hum Mol Genet.* 2014;23:5774–5780.
- Cirulli ET, Lasseigne BN, Petrovski S, et al. Exome sequencing in amyotrophic lateral sclerosis identifies risk genes and. *Science.* 2015;347:1436.
- Petrovski S, Todd JL, Durheim MT, et al. An exome sequencing study to assess the role of rare genetic variation in pulmonary fibrosis. *Am J Respir Crit Care Med.* 2017;196:82–93.
- Zhu X, Padmanabhan R, Copeland B, et al. A case-control collapsing analysis identifies epilepsy genes implicated in trio sequencing studies focused on de novo mutations. *PLoS Genet.* 2017;13:e1007104
- McKenna A, Hanna M, Banks E, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* 2010;20:1297–1303.
- Flicek P, Amode MR, Barrell D, et al. Ensembl 2014. *Nucleic Acids Res.* 2014;42(D1):749–755.
- Jun G, Flickinger M, Hetrick KN, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet.* 2012;91:839–848.
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM. Robust relationship inference in genome-wide association studies. *Bioinformatics.* 2010;26:2867–2873.
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006;38:904–909.
- Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods.* 2010;7:248–249.
- Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature.* 2016;536:285–291.
- Jones E, Oliphant T, Peterson P. SciPy: open source scientific tools for Python. 2001. <http://www.scipy.org/>.
- Firth D. Bias reduction of maximum likelihood estimates. *Biometrika.* 1993;80:27–38.
- Wang X. Firth logistic regression for rare variant association tests. *Front Genet.* 2014;5:1–2.
- Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am J Hum Genet.* 2012;91:215–223.
- Nichols BE, Sheffield VC, Vandenburgh K, Drack AV, Kimura AE, Stone EM. Butterfly-shaped pigment dystrophy of the fovea caused by a point mutation in codon 167 of the RDS gene. *Nat Genet.* 1993;3:202–207.
- Farrar GJ, Kenna P, Jordan SA, et al. A three-base-pair deletion in the peripherin-RDS gene in one form of retinitis pigmentosa. *Nature.* 1991;354:478–480.
- Khani SC, Karoukis AJ, Young JE, et al. Late-onset autosomal dominant macular dystrophy with choroidal neovascularization and nonexudative maculopathy associated with mutation in the RDS gene. *Invest Ophthalmol Vis Sci.* 2003;44:3570–3577.
- Nakazawa M, Kikawa E, Chida Y, Wada Y, Shiono T, Tamai M. Autosomal dominant cone-rod dystrophy associated with mutations in codon 244 (Asn244His) and codon 184 (Tyr184Ser) of the Peripherin/RDS gene. *Arch Ophthalmol.* 1996;114:72–78.
- Boon CJF, Klevering BJ, Cremers FPM, et al. Central areolar choroidal dystrophy. *Ophthalmology.* 2009;116:771–782.

33. Freund CL, Gregory-Evans CY, Furukawa T, et al. Cone-rod dystrophy due to mutations in a novel photoreceptor-specific homeobox gene (CRX) essential for maintenance of the photoreceptor. *Cell*. 1997;91:543–553.
34. Sohocki MM, Sullivan LS, Mintz-Hittner HA, et al. A range of clinical phenotypes associated with mutations in CRX, a photoreceptor transcription-factor gene. *Am J Hum Genet*. 1998;63:1307–1315.
35. Kniazeva M, Chiang MF, Morgan B, et al. A new locus for autosomal dominant Stargardt-like disease maps to chromosome 4. *Am J Hum Genet*. 1999;64:1394–1399.
36. Bowne SJ, Humphries MM, Sullivan LS, et al. A dominant mutation in RPE65 identified by whole-exome sequencing causes retinitis pigmentosa with choroidal involvement. *Eur J Hum Genet*. 2011;19:1074–1081.
37. Aleman TS, Soumitra N, Cideciyan AV, et al. CERKL mutations cause an autosomal recessive cone-rod dystrophy with inner retinopathy. *Invest Ophthalmol Vis Sci*. 2009;50:5944–5954.
38. Chiang JPW, Trzuppek K. The current status of molecular diagnosis of inherited retinal dystrophies. *Curr Opin Ophthalmol*. 2015;26:346–351.