tvst

Retina

An In-Depth Single-Gene Worldwide Carrier Frequency and Genetic Prevalence Analysis of *CYP4V2* as the Cause of Bietti Crystalline Dystrophy

Mor Hanany[1,*], Richard Rui Yang[2,*], Chun Man Lam[2], Avigail Beryozkin[1], Yogapriya Sundaresan[1], and Dror Sharon[1]

[1] Department of Ophthalmology, Hadassah Medical Center, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem, Israel
[2] Reflection Biotechnologies Limited, Hong Kong, China

Correspondence: Dror Sharon, Department of Ophthalmology, Hadassah-Hebrew University Medical Center, Jerusalem 91120, Israel.
e-mail: dror.sharon1@mail.huji.ac.il
Richard Rui Yang, Reflection Biotechnologies Limited, Unit 601, 6/F, Core Building 1, No. 1 Science Park East Avenue, Pak Shek Kok, New Territories, Hong Kong, China.
e-mail: richard@reflectionbio.com

Received: March 30, 2022
Accepted: January 19, 2023
Published: February 16, 2023

Keywords: carrier frequency; disease-causing mutation; Bietti crystalline dystrophy; *CYP4V2*; human genome; inherited retinal diseases; genetic prevalence; gnomAD database

Citation: Hanany M, Yang RR, Lam CM, Beryozkin A, Sundaresan Y, Sharon D. An in-depth single-gene worldwide carrier frequency and genetic prevalence analysis of *CYP4V2* as the cause of Bietti crystalline dystrophy. Transl Vis Sci Technol. 2023;12(2):27, https://doi.org/10.1167/tvst.12.2.27

Purpose: Bietti crystalline dystrophy (BCD) is a rare monogenic autosomal recessive (AR) chorioretinal degenerative disease caused by biallelic mutations in *CYP4V2*. The aim of the current study was to perform an in-depth calculation of worldwide carrier frequency and genetic prevalence of BCD using gnomAD data and comprehensive literature *CYP4V2* analysis.

Methods: *CYP4V2* gnomAD data and reported mutations were used to calculate carrier frequency of each variant. An evolutionary-based sliding window analysis was used to detect conserved protein regions. Potential exonic splicing enhancers (ESEs) were identified using ESEfinder.

Results: We identified 1171 *CYP4V2* variants, 156 of which were considered pathogenic, including 108 reported in patients with BCD. Carrier frequency and genetic prevalence calculations confirmed that BCD is more common in the East Asian population, with ∼19 million healthy carriers and 52,000 individuals who carry biallelic *CYP4V2* mutations and are expected to be affected. Additionally, we generated BCD prevalence estimates of other populations, including African, European, Finnish, Latino, and South Asian. Worldwide, the estimated overall carrier frequency of *CYP4V2* mutation is 1:210, and therefore, ∼37 million individuals are expected to be healthy carriers of a *CYP4V2* mutation. The estimated genetic prevalence of BCD is about 1:116,000, and we predict that ∼67,000 individuals are affected with BCD worldwide.

Conclusions: Our analysis estimates BCD prevalence and revealed large differences among various populations. Moreover, it highlights advantages and limitations of the gnomAD database.

Translational Relevance: This analysis is likely to have important implications for genetic counseling in each studied population and for developing clinical trials for potential BCD treatments.

Introduction

Inherited retinal diseases (IRDs) are a large group of retinal phenotypes showing large clinical variability and extreme genetic heterogeneity. Most IRDs are heterogeneous and can be caused by mutations in one out of many causative genes. The most common IRD is retinitis pigmentosa (RP), which is one of the most heterogeneous conditions in humans, caused by mutations in over 60 genes that can be inherited in every known pattern. The prevalence of RP was studied in specific geographic regions, yielding variable results, with an average of 1:5000 individuals.[1–4] Disease prevalence can also be estimated by calculating the genetic prevalence (GP) using data extracted from genetic databases (such as the Genome Aggregation Database [gnomAD]), allowing one to predict the prevalence of individuals carrying a genotype that is expected to result in a disease. Since most IRDs are noncongenital, genetic prevalence is expected to be higher than reported disease prevalence. In a previous comprehensive IRD analysis, we used carrier frequency (CF) and genetic prevalence data to predict that the number of individuals affected with autosomal recessive (AR) nonsyndromic RP is over 1 million.[5,6]

One of the most studied monogenic IRDs is Bietti crystalline dystrophy (BCD; also known as Bietti crystalline corneoretinal dystrophy, OMIM 210370), a rare AR chorioretinal degenerative disease first reported by G. B. Bietti.[7] To date, over 100 publications have reported patients with BCD (for an updated list of papers: https://reflectionbio.com/about-bcd/bcd-around-the-world), with variable age of onset and progression.[8] BCD is caused by severe retinal degeneration, including severe atrophy of the retinal pigment epithelium and loss of the outer retina, with near-total degeneration of all functional elements of the retina by the late stage of the disease,[9] and there is currently no approved treatment for BCD.

Most patients with BCD notice the first symptoms between the second and fourth decades of life, and progressive visual loss and constriction of the visual fields lead to legal blindness usually in the fifth or sixth decade.[10] Abnormal retinal function evident from diminished electroretinograms (ERGs) has been documented in the early stage preceding loss of central vision, and extinction of ERG occurs during the intermediate stage, long before legal blindness occurrence.[11] The disease hallmark is the accumulation of small yellow/white retinal crystals and, in some cases, in the cornea as well (Fig. 1). Retinal crystals may not always

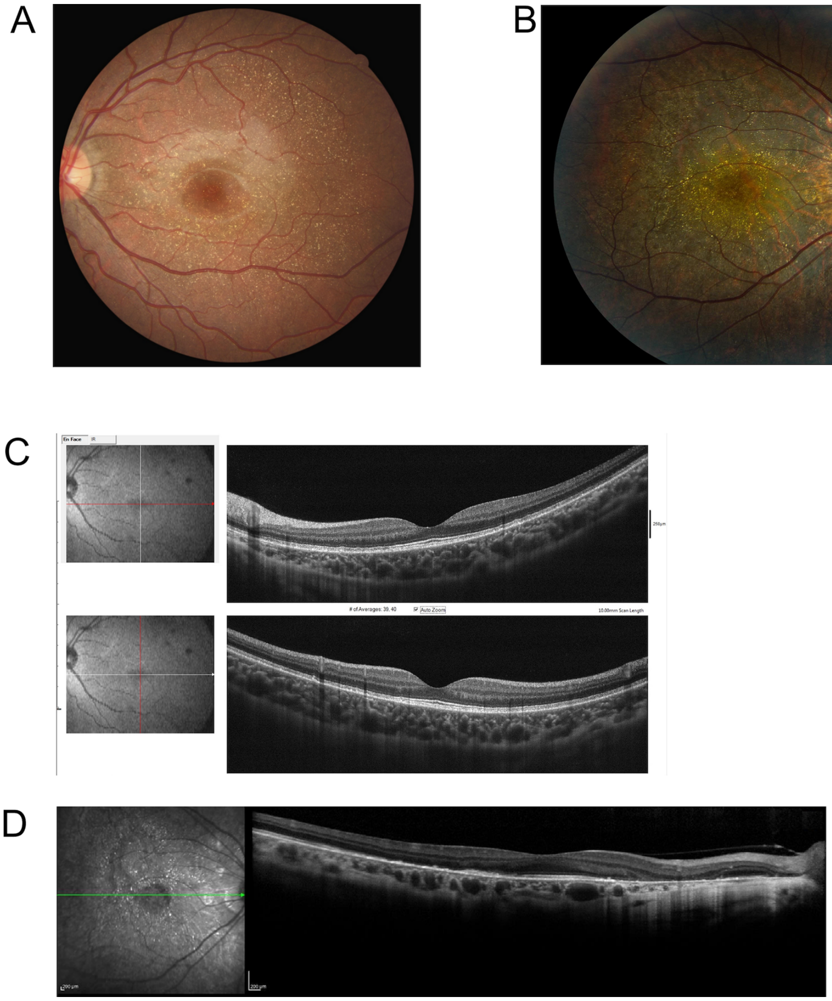

Figure 1. The sparkling retina: representative retinal images of BCD. Retinal crystals are seen in fundus photography (A, B) and spectral domain optical coherence tomography images (C, D) of patients with BCD. In late disease stages, retinal pigment epithelium (RPE) thinning is severe and dilated choroid vasculature is visible (D), indicating that the RPE has degenerated significantly. In early-stage BCD, the outer nuclear layer (ONL) is intact (C), while in advanced-stage BCD, the RPE and ONL have degenerated (D). *Image credit:* BCD fundus and OCT images were provided by Invincible Vision (www.InvincibleVision.org), a BCD patient organization.

be visible. BCD symptoms are similar to other IRDs and therefore can be misdiagnosed, making it even more difficult to estimate disease prevalence. In 2004, mutations in *CYP4V2*,[10] a novel gene encoding a 525–amino acid protein that is a member of the cytochrome P450 protein family, were reported to cause BCD.

The prevalence of BCD is currently unknown, and no well-designed prevalence study has been performed. Two methods can be used to estimate BCD prevalence: assessing disease prevalence based on the number of diagnosed patients in a specific population and genetic prevalence analysis based on genetic databases. The first method is based on relatively small cohorts of patients who belong to the same population, allowing one to assess disease prevalence in this particular population. Prevalence rough estimates have been reported and show a high level of variability, mainly due to population differences and difficulties in obtaining the correct diagnosis. BCD is relatively prevalent in East Asians, particularly the Chinese, Japanese, and Korean populations. An epidemiologic survey of genetic diseases in the Chinese population estimated the gene frequency to be 0.005.[12,13] In Europe, this method generated different estimates varying from 2.5% of all patients with RP and 10% of those with nonsyndromic AR RP,[11] which translates to an incidence rate of 1:67,000 to 1:4.5 million individuals.[14] We have previously reported seven cases of Yemenite-Jewish origin who are biallelic for *CYP4V2* mutations in Israel,[15] yielding an expected prevalence of 1:527,647 individuals. The second method is based on the analysis of genetic information that is available in large genomic databases (mainly gnomAD; https://gnomad.broadinstitute.org),[16] as well as mutation databases (mainly HGMD [http://www.hgmd.cf.ac.uk/ac/index.php] and LOVD [https://databases.lovd.nl/shared/diseases/01680]). Such databases include variants identified in the studied cohorts, while in the gnomAD database, only a minority of these are disease causing. Using this method, based on the gnomAD exome database (V2), we previously estimated genetic prevalence and carrier frequency of all known IRD genes, including *CYP4V2*, which was estimated to have a genetic prevalence of 1:57,600, which can be translated to about 132,000 individuals worldwide who are expected to be affected with BCD.[5] Similarly, CF was calculated for each suspected pathogenic variant that appears in the gnomAD database, and the accumulated carrier frequency for *CYP4V2* is 1:181, and therefore, 43 million carriers worldwide are expected to carry a *CYP4V2* pathogenic variant.

In the current study, we performed an in-depth worldwide *CYP4V2* pathogenic variants carrier frequency and BCD genetic prevalence calculation in different subpopulations, which is based on our previous analysis[5] using updated gnomAD data and comprehensive literature *CYP4V2* analysis to more accurately assess the number of individuals affected with the disease.

Methods

GnomAD Data Analysis

All *CYP4V2* variants in gnomAD (V2 and V3) were downloaded (May 25, 2021). The genomic positions in gnomAD-V2 (based on GRCh37-hg19) and V3 (based on GRCh37-hg38) were translated using LiftOver (https://genome.ucsc.edu/cgi-bin/hgLiftOver). After translation, we added up the total number of identified individuals and the size of each population in gnomAD. The combination of gnomAD-V2 and V3 enabled us to obtain an accurate analysis of variants and to determine their pathogenicity. The allele frequency calculation is based on the weighted mean of both gnomAD versions.

In cases when a variant was found only in one gnomAD version, we considered the frequency in the other version as zero out of the size of the population. In other words, for example, if a variant was not found in a specific population in V2 (with a total allele count of 150,000) but was found once in V3 (with a total allele count of 100,000), the weighted mean would be $\frac{0+1}{250{,}000}$.

Published Mutations and Their Integration into the Database

Not all previously published disease-causing mutations appear in the gnomAD cohort. In those instances, we decided to track down the population that the disease-causing mutations were found in and assess their allele frequency based on the size of that population in gnomAD. We calculated the carrier frequency as if there was one individual in that population found in gnomAD and divided it by 2, meaning that carrier frequency for those disease-causing mutations in that population would be between 0 and 1 out of the size of the population in gnomAD.

For mutations in which the specific population was not published, we included them in an unknown population and used the entire gnomAD cohort size to calculate the allele frequency.

Carrier Frequency Calculation

Carrier frequency was calculated as reported previously.[17] The gnomAD database provides the following values for each subpopulation: “allele count” (representing the number of detected alleles in a given subpopulation), “allele number” (total number of genotyped alleles at the genomic position of the variant considered), and “homozygote count” (total number of homozygous individuals for that specific allele). Based on these values, we calculated the following parameters (Supplementary Table S1): allele frequency, the total number of individuals, the number of heterozygous individuals, and the number of wild-type individuals. In addition, using the Hardy–Weinberg equation and the abovementioned parameters, we calculated carrier frequency as $2pq$, where $p = 1 - q$, and q was calculated as the root square of the number of homozygous patients plus half the number of compound heterozygous patients divided by the population size, as shown in the following equation (see also Supplementary Table S1): $q = \sqrt{\frac{NHo+0.5*NHe}{NI\ in\ subpopulation}}$.

When carrier frequency was calculated for a specific mutation, only individuals heterozygous for a single mutation in *CYP4V2* were included, whereas homozygotes were excluded.

Genetic Prevalence Calculation

Genetic prevalence was calculated as reported previously.[5] In short, we calculated the genetic prevalence of affected individuals for BCD in the worldwide subpopulations using a product-based algorithm for allele matrices. In order to calculate BCD prevalence, we created a matrix of all the possible genetic combinations of different mutations in *CYP4V2* and multiplied the carrier frequency of the two mutations in each pair, including both homozygous and compound heterozygous combinations. We aggregated all the sums of each multiplication based on the following equation: $(\sum_{i,j} x_{i,j} + \sum_{i} x_{i,i})/4$.

Demographic Worldwide Data

Demographic data on the worldwide populations (in terms of the number of individuals per each subpopulation) are based on the Department of Economic and Social Affairs of the United Nations Secretariat 2019 published on August 2019 (https://population.un.org/wpp/Download/Standard/Population/) as follows: total worldwide population size, 7.8 billion; Africans, 1,340,598,113; East Asians, 1,678,089,627; South Asians, 1,940,369,605; Finnish, 5,540,781; European (non-Finnish), 742,095,327; and Latino, 653,962,332 (other populations that are not represented in the gnomAD database: Oceania, 42,677,809; North America, 368,869,644; West Asia, 279,636,774; Southeastern Asia, 668,619,854; Central Asia, 74,338,926).

See Supplementary Methods for additional information.

Results

Genetic Landscape of *CYP4V2*

Aiming to generate a comprehensive list of all likely pathogenic variants in *CYP4V2*, we collected information regarding 1171 variants from the scientific literature (https://reflectionbio.com/about-bcd/bcd-literature-list) and gnomAD (versions V2 and V3) and performed an analysis of exonic sequence enhancers (ESEs). Of the 1171 variants, 156 were considered pathogenic (Supplementary Fig. S1), including 108 that were reported in patients with BCD and 48 that appear in gnomAD (but not reported in the literature) and considered as truncating. We collected the relevant information regarding these pathogenic variants to create the “*CYP4V2* global variants database” (CGVD; Supplementary Fig. S1 and Supplementary Table S2) data set. Twelve additional variants were previously reported pathogenic, but available genetic information (e.g., high population frequency, part of a complex allele) excludes this possibility in nine of these variants (Supplementary Table S3), while the remaining two variants are unlikely to be pathogenic unless they act as hypomorphic alleles (Supplementary Table S4), and one variant (c.802-8_810del17insGC) was listed using two different nomenclatures.

Among the 156 CGVD mutations (Supplementary Table S2), 108 have been reported in patients, of which 60 were not found in gnomAD. Of the 96 mutations found in gnomAD, 41 appear in both V2 and V3, 37 appear only in V2, and 18 appear only in V3.

The CYP4V2 protein contains various known domains, including the transmembrane and cytoplasmatic domains, coiled-coil domain, and the cytochrome P450 cysteine heme–iron ligand domain (Fig. 2A), as well as a large number of phosphorylation sites. We examined the distribution of pathogenic mutations in CGVD along these domains. The most common mutation types were missense (58 variants, 37% of disease-causing mutations), followed by frameshift (41 variants, 26%), stop gained (25 variants, 16%), and splice-site (26 variants, 17%) (Supplementary Fig. S2). While null mutations are spread along the protein (Fig. 2B), missense are centralized and show some level of aggregation at specific regions, mainly the C-terminal portion within the cytoplasmic domain and the cytochrome P450 cysteine heme–iron

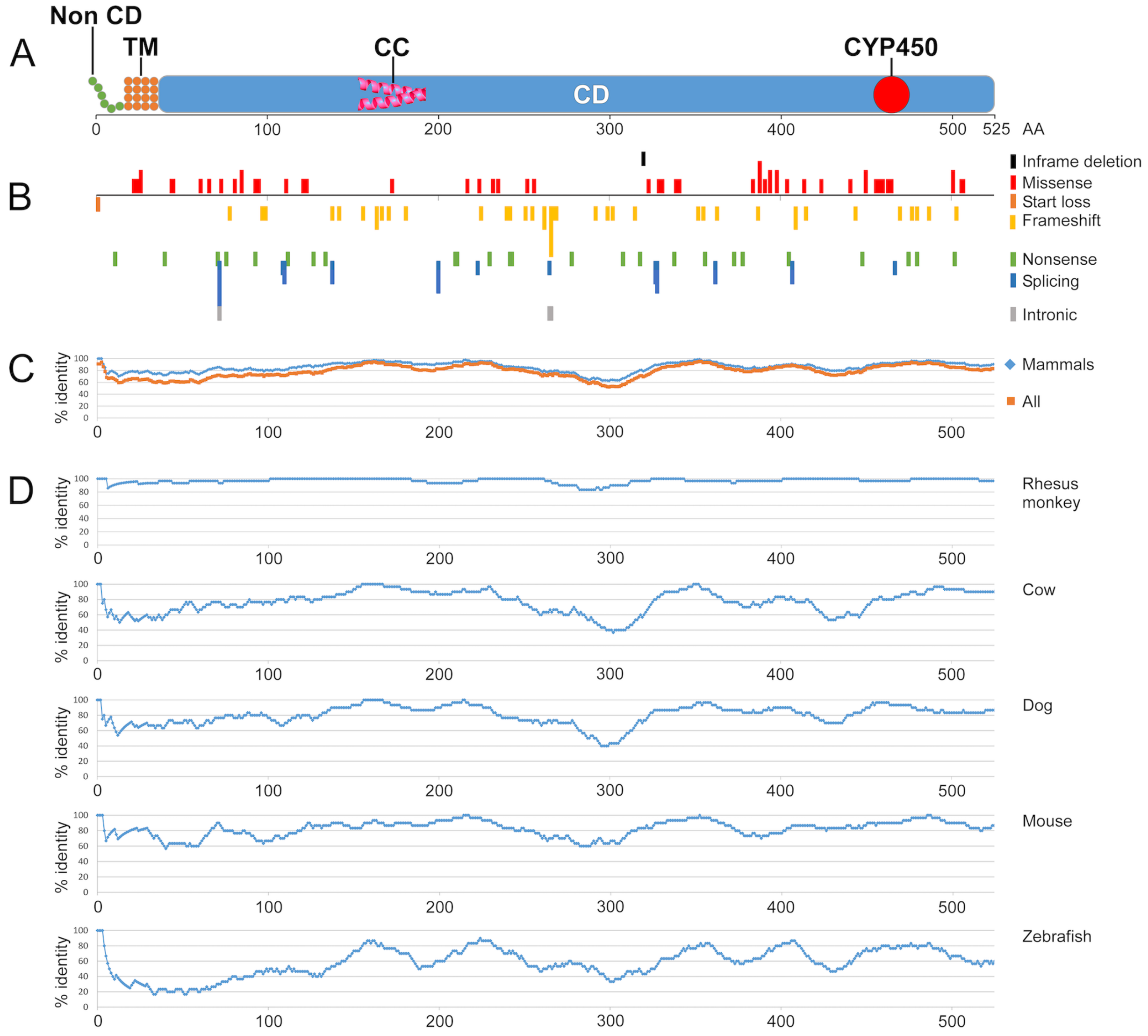

Figure 2. CYP4V2 protein structure, mutation location, and evolutionary analysis. (A) A schematic representation of the CYP4V protein and the protein's domains. CC, coiled-coil domain; CD, cytoplasmic domain; CYP450, cytochrome P450; Non CD, noncytoplasmic domain; TM, transmembrane domain. (B) Mutation types and their location along the protein. In-frame variants (missense and deletion) are above the *horizontal black line* while truncating variants (frameshift, nonsense, splicing, and intronic) are below the line. (C) A summary of amino acid sliding window (length of 50 amino acids) comparing the human protein sequence to selected orthologs (chimpanzee, rhesus monkey, rabbit, horse, dog, cow, rat, mouse, chicken, frog, and zebrafish) is presented. Average of only mammalian sequences (*blue*) versus all the sequences (*orange*) is presented. (D) Representative images of amino acid sliding window comparing the human protein sequence to representative orthologs (rhesus monkey, cow, dog, mouse, and zebrafish) is shown. x-axis: amino acid number; y-axis: percentage of amino acid identity in a 50–amino acid window. Accession numbers are as follows: human (NP_997235.3), chimpanzee (XP_001165629.1), rhesus monkey (NP_001180767.1), rabbit (XP_002709379.1), rat (NP_001129072), mouse (NP_598730.1), horse (XP_023486491.1), dog (XP_038546059.1), cow (NP_001029545), chicken (NP_001001879), frog (NP_001072667.1), and zebrafish (NP_001071070.1). The average percentage of amino acid identity for each studied sequence: chimpanzee (99%), rhesus monkey (96%), rabbit (85%), rat (83%), mouse (82%), horse (80%), dog (81%), cow (77%), chicken (60%), frog (80%), and zebrafish (57%).

ligand domain (Fig. 2B). We subsequently performed a sliding window analysis comparing the human CYP4V2 amino acid sequence to various orthologues (Fig. 2C). A few highly conserved regions were identified: the coiled-coil domain (amino acids 163–183), amino acids 330 to 370 and 390 to 420 in the cytoplasmic domain, and the cytochrome P450 cysteine heme–iron ligand domain (460–469). A relatively large number of missense mutations occur in the last three conserved regions. We therefore predict that these areas are highly conserved and cannot tolerate alterations.

Table. Worldwide Carrier Frequency of *CYP4V2* Mutations and Genetic Prevalence of Biallelic Cases

Population	Population Size	Carrier Frequency	Total Number of Carriers	Genetic Prevalence	Total Number of Biallelic Cases	Top 3 *CYP4V2* Mutations and Carrier Frequency
African	1,340,598,113	1:291	4,603,731	1:332,661	4030	c.987+3A>G: 1.87E-03 c.990del, p.His331Thrfs*8: 1.39E-04 c.1355G>A, p.Arg452His: 2.11E-04
East Asian	1,678,089,627	1:90	18,743,394	1:32,014	52,417	c.802-8_810del17insGC: 3.27E-03[a] c.992A>C, p.His331Pro: 1.70E-03 c.1091-2A>G: 9.29E-04
European	742,095,327	1:277	2,675,220	1:301,508	2461	c.130T>A, p.Trp44Arg: 3.96E-04 c.400G>T, p.Gly134*: 5.98E-04 c.1198C>T, p.Arg400Cys: 2.43E-04
Finnish	5,540,718	1:257	21,598	1:263,566	21	c.1A>G, p.Met1?: 2.84E-04 c.414-1G>A: 2.81E-03 c.1167del, p.Arg390Alafs*25: 5.60E-04
Latino	653,962,332	1:164	3,978,575	1:108,064	6052	c.130T>A, p.Trp44Arg: 2.05E-03 c.254G>A, p.Arg85His: 1.77E-03 c.1338del, p.Glu447Argfs*22: 6.02E-04
South Asian	1,940,369,605	1:566	3,427,060	1:1,282,417	1513	c.197T>G, p.Met66Arg: 6.39E-04 c.1169G>A, p.Arg390His: 1.19E-04 c.1199G>A, p.Arg400His: 2.82E-04
Worldwide	7,794,798,729	1:210	37,123,684	1:116,051	67,167	

The genetic prevalence and carrier frequency data are presented graphically in Figure 3.

[a] Listed as c.802-8_807del in genomAD. For relationships among c.802-8_807del, c.802-8_810del17insGC, and c.810del mutations, see the Discussion section.

Analysis of Potential ESEs and Splice-Altering Intronic Variants

Aiming to identify *CYP4V2* variants that might be located within ESEs and therefore affect splicing, we analyzed the 1171 variants and identified 126 that are located within the open reading frame of coding exons (1–11). ESE analysis of these 126 variants using ESEfinder revealed 38 (7 null, 25 missense, 5 silent, and 1 inframe deletion) that result in an ESE score that was below the threshold (Supplementary Table S5). Three out of the 31 inframe variants were reported in patients as pathogenic but were not considered potential ESE-related mutations. The remaining 28 variants (not reported in patients and not expected to cause protein truncation; Supplementary Table S6) were excluded from the CGVD, and the carrier frequency and genetic prevalence results are shown in Table. Recently, two *CYP4V2* missense variants among the 28 suspected ESE variants (c.1382C>T and c.1199G>T) were reported in patients with BCD as pathogenic.[18,19]

In addition, we analyzed all gnomAD *CYP4V2* variants, identified 176 noncanonical intronic variants with allele frequency (AF) <0.5%, and performed splice artificial intelligence analysis on each of these variants. The analysis revealed 105 variants with a zero score, 55 variants with a benign score, 8 with an uncertain score, 6 with a low splice-altering score, and 2 with a splice-altering score of >0.5 (Supplementary Table S7). One of these variants, c.987+3A>G, unique to the African population, is included in our analysis as a pathogenic variant. The second suspected splice-altering variant, c.605-6T>G, is unique to the Latino population, with a low AF of 0.00001, and if indeed proved pathogenic (either by a splicing assay or by biallelic presence in patients with BCD), it will have an extremely small effect on carrier frequency and genetic prevalence.

Carrier Frequency and Genetic Prevalence Analysis

Carrier frequency and genetic prevalence analysis of the CGVD in various populations showed large variability, similar to previous disease prevalence results. The population in whom BCD was reported to be most common is East Asian, and our analysis shows that ∼19 million (1:90) healthy individuals are expected to be carriers for a *CYP4V2* mutation, and 52,000 individuals (GP of 1:32,014) carry biallelic *CYP4V2* mutations and are expected to be affected (Figs. 3, 4 and Table). In other words, over 1% of the East Asian population are carriers of a *CYP4V2* mutation. We therefore predict that 50% of healthy carriers and 78% of affected individuals worldwide are of East Asian origin. The relatively high contribution of the East Asian population is mostly due to a number of founder mutations that are common in this population (c.802-8_810del17insGC

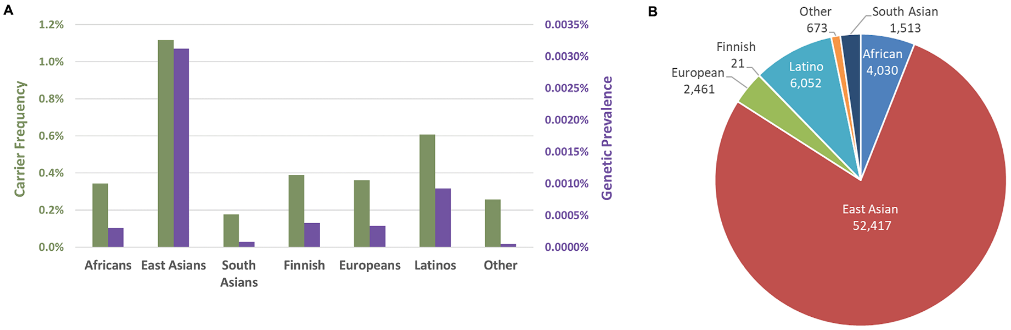

Figure 3. Worldwide carrier frequency and genetic prevalence values based on *CYP4V2* pathogenic variants. (A) A dual *bar graph* showing carrier frequency of *CYP4V2* mutations and BCD genetic prevalence in various worldwide populations. The *green bars* and *left-hand axis* represent the carrier frequency values. The *purple bars* and the *right-hand axis* represent the genetic prevalence values. (B) The number of expected biallelic *CYP4V2* cases in various populations.

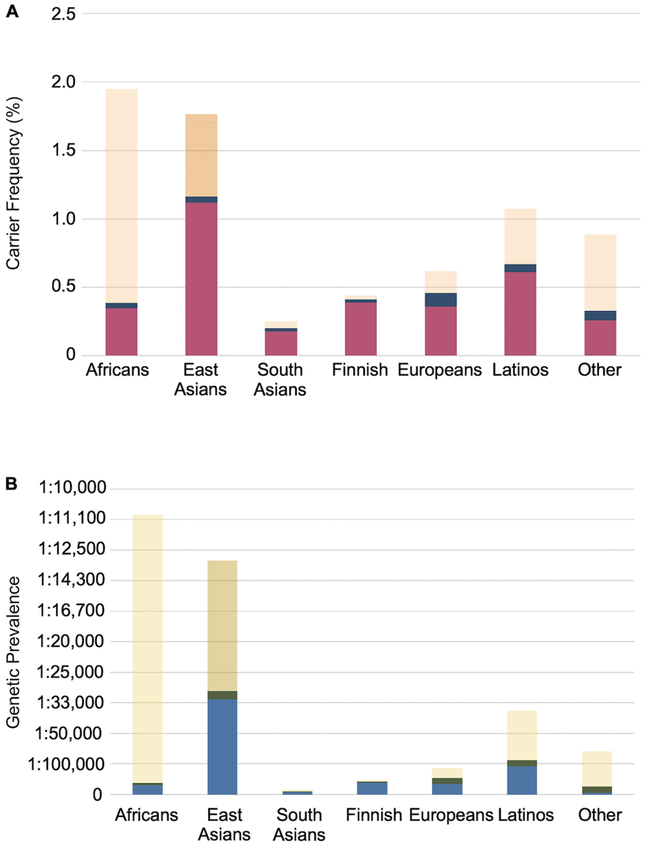

Figure 4. Worldwide carrier frequency and genetic prevalence values based on *CYP4V2* pathogenic variants, ESE-related variants, and hypomorphic variants. (A) A *bar graph* showing carrier frequency of *CYP4V2* mutations in various worldwide populations. The *fuchsia bars* represent data from the CGVD database, *blue bars* represent the added values of suspected 28 ESE variants, and *light* and *dark orange* represent the hypomorphic variants c.1328G>A and c.237G>T, respectively. (B) A *bar graph* showing genetic prevalence of biallelic *CYP4V2* mutations in various worldwide populations. The *blue bars* represent data from the CGVD database, *dark green bars* represent the added values of suspected 28 ESE variants, and *light* and *dark yellow* represent the hypomorphic variants c.1328G>A and c.237G>T, respectively.

[listed as c.802-8_807del in gnomAD V3 or c.802-8_807delTCATACAGGTCATC in gnomAD-V2] and p.His331Pro).

The African population has a genetic prevalence of 1:333,000 and about 4000 expected affected individuals, as well as a carrier frequency of 1:291 with ~4.6 million healthy carriers for BCD (Fig. 3A). The most common variant in the African population is c.987+3A>G, which was previously reported in a homozygous state only in a single patient with BCD from the United Kingdom[20] in *cis* with a missense variant (p.Gly26Asp). While the missense variant is considered nonpathogenic by most analysis tools, c.987+3A>G is predicted to abolish the donor splice-site at the 3′ end of exon 7.[20] This variant was found in 62 individuals in gnomAD in the African population and might represent an example of underrepresentation of the African patient population in previous publications. The population with the smallest carrier frequency is South Asians, with only 3.4 million individuals expected to be healthy carriers and about 1500 individuals expected to be affected. Since the "Other" population in gnomAD corresponds to individuals who are not directly related to any of the specific populations mentioned in gnomAD, very few individuals in gnomAD are considered part of the "Other" population, and variants found in these individuals do not represent the entire 1.4 billion individuals from populations not included in gnomAD. Hence, the carrier frequency and genetic prevalence calculated in this study for the "Other" population resulted in very small values. These values would likely be higher once more populations are included and a larger number of individuals from those populations are sequenced.

Based on the current study, the estimated overall carrier frequency of *CYP4V2* mutations is 1:210, and therefore ~37 million individuals are expected to be healthy carriers of a *CYP4V2* mutation worldwide. The estimated genetic prevalence of BCD is about 1:116,000, and we predict that ∼67,000 individuals are affected with BCD worldwide (Fig. 3B). These values can be higher if the 28 suspected ESE variants are being considered pathogenic (CF, 1:190, 41 million; genetic prevalence, 1:99,600, 78,000).

Hypomorphic BCD Mutations

Two variants that were reported as pathogenic, c.237G>T (p.Glu79Asp) and c.1328G>A (p.Arg443 Gln), show high AF and are unlikely to be pathogenic (see Supplementary note for additional information). If indeed these variants were fully penetrant pathogenic variants, they could have a dramatic effect on carrier frequency and genetic prevalence, as shown in Figures 4A and 4B, respectively. These variants are therefore unlikely to be pathogenic but might have some effect as hypomorphic variants, which are generally defined as variants that cause disease only when paired in *trans* with a variant affecting protein function. The relatively high AF of hypomorphic variants usually does not correlate with the expected number of affected individuals reported with such mutations. In some cases, healthy homozygous individuals for these variants are identified in genetic databases, further increasing their definition as hypomorphic.

Discussion

Estimating disease prevalence in various populations worldwide is an important task, mainly for obtaining decisions regarding development of therapeutic modalities. However, studies in which disease prevalence is actually measured are rare, and such studies would be relevant only for the specific studied population, since genetic variability among populations is pronounced. Therefore, there is currently no reliable estimate of BCD prevalence worldwide. Rough estimations of BCD prevalence range dramatically between extremely rare (1:4.5 million individuals in Spain[14] and 1:∼0.5 million in Israel) and a relatively common disease (1:24,000 individuals among Chinese). One can alternatively calculate genetic prevalence from large data sets of nonaffected individuals, bearing in mind that genetic prevalence should always be larger than reported disease prevalence in noncongenital diseases.

In the current study, we used different sources of information and mainly gnomAD variants and pathogenic variants reported in patients to calculate both carrier frequency and genetic prevalence for BCD. The combined data set we created is more accurate and more reliable compared to using each information source as is. The gnomAD database, which by far is the largest multiethnic genomic database, is a powerful tool for genetic disease epidemiology research, providing many advantages over the traditional approach of estimating disease prevalence in a specific population or region. Compared to BCD prevalence rough estimates from prior studies,[11–14] the gnomAD-based results not only confirm that BCD is more common in East Asians but also elucidate prevalence in other populations, whose prevalence either varies widely in prior studies or has not been studied previously. Further, the gnomAD-based results revealed dramatic differences in genetic prevalence and top *CYP4V2* mutations among different populations. This highlights the importance of including patients from different populations in clinical research and development of potential treatments for BCD.

However, in the course of this study, we noticed that gnomAD also has certain limitations, some of which were discussed in detail by the gnomAD research team.[16] First, gnomAD has currently two data sets, V2 and V3, each with its own pros and cons. The larger one, V2, contains primarily exome data from 141,456 individuals, whereas V3 contains genomic data from 76,156 individuals. These structural differences may lead to materially different results. In the current study, among 96 pathogenic variants in gnomAD, 55 appear in either V2 or V3, most of which ($n = 46$) are rare with only a single allele. However, a few variants had a relatively large number of heterozygotes (up to 12) in V2 while none appear in V3. Among the 41 variants that appear in both data sets, a reasonable correlation coefficient of 0.70 was obtained. However, there are exceptions. For example, the status of the c.802-8_810del17insGC mutation being the most common *CYP4V2* mutation with a founder effect in the East Asian population has been well established by multiple studies in patients.[19,21,22] Its carrier frequency varies dramatically between the two versions ($9.67 * 10^{-3}$ in gnomAD V3 vs. $1.61 * 10^{-3}$ in gnomAD V2, a ∼6-fold difference), a difference that might stem from the variant nature (indel) and its intronic location that might be more accurately called in genome analysis. Indeed, c.802-8_810del17insGC is the most common *CYP4V2* mutation in V3 in the East Asian population, a result that is consistent with its relatively high prevalence among patients with BCD. Second, despite the large sample size of gnomAD, rare mutations are absent. In the current study, 60 of 108 mutations previously reported in patients with BCD are absent from gnomAD, and most (∼63%) of these variants appear in patients of East Asian origin. It is therefore reasonable to predict that the data presented here are more accurate in well-represented populations, such as the European population, but this gap is expected to narrow over time as gnomAD includes more samples with higher-quality data. On the other hand, gnomAD may capture mutations that have not been reported in patients, especially for rare diseases. Based on the current study, gnomAD contains 48 likely pathogenic variants that have not been reported in patients. This is not surprising as new *CYP4V2* mutations are continuously being reported in patients, even in the East Asian population, in whom BCD has been well studied.[19,22] As new mutations are discovered, reanalysis of genomic data can increase genetic diagnosis of patients who harbor mutations that have not been published previously.[23] Third, the small sample size of some populations in gnomAD may result in the absence of mutations from gnomAD, especially for rare ones. For example, V3 has data for the Middle Eastern population based on a sample size of only 158 individuals, leading to a calculated genetic prevalence of zero for BCD, a value that does not reflect disease prevalence in this population, since mutations were reported in the Arab-Muslim,[21] Iranian,[24] Jewish,[25] and Lebanese[26] populations.

Furthermore, our results show that there is significant variability in carrier frequency and genetic prevalence among different populations. Consistent with existing knowledge on reported patients with BCD, our results confirm that carrier frequency and genetic prevalence in East Asians (1:90 and 1:32,014 respectively) are much higher than in other populations. Moreover, our results provide a more accurate estimate of BCD prevalence in other populations. For example, previous studies in Europe have generated dramatically different BCD prevalence, ranging from 1:67,000 to 1:4.5 million individuals. According to our calculation, carrier frequency and genetic prevalence in the European population are estimated as 1:277 and 1:~302,000 respectively, which is within the reported values. Furthermore, this study provides insights on BCD prevalence in populations for whom no prevalence estimates have been performed previously. Importantly, besides significant differences in genetic prevalence among different populations, we found that the top *CYP4V2* mutations with the highest carrier frequency also vary significantly among different populations (Table).

Among all *CYP4V2* mutations, the founder East Asian mutation, c.802-8_810del17insGC, has the highest carrier frequency, and it also appears in Europeans at a much lower carrier frequency and in European patients with BCD.[20] In gnomAD, it is listed as two separate variants: c.802-8_807del and c.810del, and the allele count for c.810del is higher than that of c.802-8_807del by one allele, suggesting that although in most cases, these two mutations appear *in-cis* to form the c.802-8_810del17insGC mutation, c.810del is an independent mutation in very rare cases.

The data and analysis we present here might suffer from some limitations. For example, some mutation types (e.g., structural variants and deep intronic) are not well identified in whole exome sequencing (WES) data while others are not recognized as pathogenic by next generation sequencing (NGS) analysis schemes and therefore will not be presented in our data set. This includes mainly intronic mutations that affect splicing and exonic silent mutations that affect ESEs. Aiming to better address this issue, we performed an ESE analysis of the whole open reading frame (ORF) and identified potential ESE variants that were added to the genetic prevalence and carrier frequency analyses, but their aggregated contribution is relatively low. Additional work is needed to prove that these variants affect ESEs and therefore splicing. In addition, some of the reported variants might have a dual-mutation mechanism or might actually act only as ESEs. Another limitation stems from the lack of knowledge regarding the pathogenicity of mainly missense variants that appear in gnomAD but were not reported to cause disease in patients with IRD and therefore are expected to slightly reduce our estimates.

Another aspect that can affect BCD prevalence is misdiagnosis. BCD symptoms can be similar to other IRDs and therefore BCD might be misdiagnosed and underestimated. A prior study reported that only 1:6 patients with BCD was correctly diagnosed, while the remaining were initially diagnosed with RP.[11] BCD can also be misdiagnosed as choroideremia,[27] late-onset retinal degeneration, other retinal diseases with refractile crystal-like deposits in the retina (Stargardt disease type 3, dominant forms of *RPE65*, *PRPH2*, or other crystalline retinopathies),[28,29] and nongenetic retinal disease. Although often used in diagnosing BCD, retinal crystals are not unique to BCD[29] and may not be visible in early or late disease stages.[30,31] Near-infrared imaging can enhance detection of retinal crystals and improve correct diagnosis.[20,32,33] Therefore, there is in low awareness of BCD among ophthalmologists, and genetic testing for *CYP4V2* mutations is the ultimate tool to confirm BCD diagnosis.[34]

The high carrier frequency and genetic prevalence of BCD in East Asians indicate that BCD is a leading IRD in China, Japan, Korea, and Singapore. Our data indicate that over 52,000 individuals of East Asian origin are biallelic for *CYP4V2* mutations (about 78% of the total number of biallelic individuals worldwide). Moreover, the prevalence of BCD should not be overlooked outside of East Asia since thousands of biallelic individuals are expected in the African (6% of worldwide biallelic individuals), European (3.7%), Latino (9%), and South Asian (2.3%) populations.

Taken together, these findings indicate BCD is a genetic disease, affecting different populations throughout the world. Because the top *CYP4V2* mutations vary among different populations, it would be helpful to include patients from multiple worldwide populations in future clinical trials aiming to better assess the clinical benefits to patients with different *CYP4V2* mutations.

Acknowledgments

The authors thank Invincible Vision for providing the fundus photography and spectral domain optical coherence tomography images of patients with BCD with patients' consent. Invincible Vision (www.InvincibleVision.org) is a California-based, patient-led nonprofit organization focused on serving the BCD community.

Supported by Reflection Biotechnologies Limited (ReflectionBio) with its BY PATIENTS, FOR PATIENTS approach. ReflectionBio is a patient-driven biotechnology company focusing on vision research and developing gene and cell therapies for ocular diseases. BY PATIENTS, FOR PATIENTS is a trademark of Reflection Biotechnologies Limited.

Disclosure: **M. Hanany**, None; **R.R. Yang**, ReflectionBio (O, P, S); **C.M. Lam**, ReflectionBio (E); **A. Beryozkin**, None; **Y. Sundaresan**, None; **D. Sharon**, ReflectionBio (F)

[*] MH and RRY contributed equally to this article.

References

1. Rosenberg T. Epidemiology of hereditary ocular disorders. *Dev Ophthalmol*. 2003;37:16–33.
2. Bundey S, Crews SJ. A study of retinitis pigmentosa in the City of Birmingham. I. Prevalence. *J Med Genet*. 1984;21(6):417–420.
3. Bunker CH, Berson EL, Bromley WC, Hayes RP, Roderick TH. Prevalence of retinitis pigmentosa in Maine. *Am J Ophthalmol*. 1984;97(3):357–365.
4. Sharon D, Banin E. Nonsyndromic retinitis pigmentosa is highly prevalent in the Jerusalem region with a high frequency of founder mutations. *Mol Vis*. 2015;21:783–792.
5. Hanany M, Rivolta C, Sharon D. Worldwide carrier frequency and genetic prevalence of autosomal recessive inherited retinal diseases. *Proc Natl Acad Sci*. 2020;117(5):2710–2716.
6. Schneider N, Sundaresan Y, Gopalakrishnan P, et al. Inherited retinal diseases: linking genes, disease-causing variants, and relevant therapeutic modalities. *Prog Retin Eye Res*. 2022;89:101029.
7. Bietti G. Ueber faxmiliares Vorkommen von "Retinitis punctata albescens" (verbunden mit "dystrophis marginalis cristallinea cornea"), glitzern, des glaskorpers und anderen degenerativen augenveranderungen. *Klin Monbl Augenheilkd*. 1937;99:737–756.
8. Lai TYY, Ng TK, Tam POS, et al. Genotype phenotype analysis of Bietti's crystalline dystrophy in patients with CYP4V2 mutations. *Invest Ophthalmol Vis Sci*. 2007;48(11):5212–5220.
9. Furusato E, Cameron JD, Chan C-C. Evolution of cellular inclusions in Bietti's crystalline dystrophy. *Ophthalmol Eye Dis*. 2010;2010(2): 9–15.
10. Li A, Jiao X, Munier FL, et al. Bietti crystalline corneoretinal dystrophy is caused by mutations in the novel gene CYP4V2. *Am J Hum Genet*. 2004;74(5):817–826.
11. Mataftsi A, Zografos L, Millá E, Secrétan M, Munier FL. Bietti's crystalline corneoretinal dystrophy: a cross-sectional study. *Retina*. 2004;24(3):416–426.
12. Hu DN. Genetic aspects of retinitis pigmentosa in China. *Am J Med Genet*. 1982;12(1):51–56.
13. Hu D-N. Ophthalmic genetics in China. *Ophthalmic Paediatr Genet*. 1983;2(1):39–45.
14. García-García GP, Martínez-Rubio M, Moya-Moya MA, Pérez-Santonja JJ, Escribano J. Current perspectives in Bietti crystalline dystrophy. *Clin Ophthalmol*. 2019;13:1379–1399.
15. Sharon D, Ben-Yosef T, Goldenberg- Cohen N, et al. A nation-wide genetic analysis of inherited retinal diseases in Israel as assessed by the Israeli inherited retinal disease consortium (IIRDC). *Hum Mutat*. 2019;41(1):140–149.
16. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*. 2020;581(7809):434–443.
17. Hanany M, Allon G, Kimchi A, et al. Carrier frequency analysis of mutations causing autosomal-recessive-inherited retinal diseases in the Israeli

population. *Eur J Hum Genet*. 2018;26(8):1159–1166.

18. Zhang S, Wang L, Liu Z, et al. Observation of the characteristics of the natural course of Bietti crystalline dystrophy by fundus fluorescein angiography. *BMC Ophthalmol*. 2021;21(1):239.
19. Murakami Y, Koyanagi Y, Fukushima M, et al. Genotype and long-term clinical course of Bietti crystalline dystrophy in Korean and Japanese patients. *Ophthalmol Retin*. 2021;5(12):1269–1279.
20. Halford S, Liew G, Mackay DS, et al. Detailed phenotypic and genotypic characterization of Bietti crystalline dystrophy. *Ophthalmology*. 2014;121(6):1174–1184.
21. Jiao X, Li A, Jin Z-B, et al. Identification and population history of CYP4V2 mutations in patients with Bietti crystalline corneoretinal dystrophy. *Eur J Hum Genet*. 2017;25(4):461–471.
22. Meng XH, He Y, Zhao TT, Li SY, Liu Y, Yin ZQ. Novel mutations in CYP4V2 in Bietti corneoretinal crystalline dystrophy: next-generation sequencing technology and genotype-phenotype correlations. *Mol Vis*. 2019;25:654–662.
23. Robertson AJ, Tan NB, Spurdle AB, Metke-Jimenez A, Sullivan C, Waddell N. Re-analysis of genomic data: an overview of the mechanisms and complexities of clinical adoption. *Genet Med*. 2022;24(4):798–810.
24. Darki F, Fekri S, Farhangmehr S, Ahmadieh H, Dehghan MH, Elahi E. CYP4V2 mutation screening in an Iranian Bietti crystalline dystrophy pedigree and evidence for clustering of CYP4V2 mutations. *J Curr Ophthalmol*. 2019;31(2):172–179.
25. Beryozkin A, Shevah E, Kimchi A, et al. Whole exome sequencing reveals mutations in known retinal disease genes in 33 out of 68 Israeli families with inherited retinopathies. *Sci Rep*. 2015;5:13187.
26. Haddad NMN, Waked N, Bejjani R, et al. Clinical and molecular findings in three Lebanese families with Bietti crystalline dystrophy: report on a novel mutation. *Mol Vis*. 2012;18:1182–1188.
27. Katagiri S, Hayashi T, Gekka T, Tsuneoka H. A novel homozygous CYP4V2 variant (p.S121Y) associated with a choroideremia-like phenotype. *Ophthalmic Genet*. 2017;38(3):286–287.
28. Tabatabaei A, Soleimani M, Moghimi S, Kiarudi MY. A case of Bietti crystalline dystrophy with preserved visual acuity and extinguished electroretinogram: a case report. *Cases J*. 2009;2:7100.
29. Kovach JL, Isildak H, Sarraf D. Crystalline retinopathy: unifying pathogenic pathways of disease. *Surv Ophthalmol*. 2019;64(1):1–29.
30. Astuti GDN, Sun V, Bauwens M, et al. Novel insights into the molecular pathogenesis of CYP4V2-associated Bietti's retinal dystrophy. *Mol Genet Genomic Med*. 2015;3(1):14–29.
31. Ameri H, Su E, Dowd-Schoeman TJ. Autofluorescence of choroidal vessels in Bietti's crystalline dystrophy. *BMJ Open Ophthalmol*. 2020;5(1):e000592.
32. Yanagi Y, Tamaki Y, Fukushima H. Fine retinal crystalline deposits observed by confocal scanning laser ophthalmoscopic examination using infrared light. *Br J Ophthalmol*. 2003;87(4):509–510.
33. Brar VS, Benson WH. Infrared imaging enhances retinal crystals in Bietti's crystalline dystrophy. *Clin Ophthalmol*. 2015;9:645–648.
34. Ghosh A. Molecular diagnosis of inherited retinal diseases with non-specific clinical phenotypes using whole exome sequencing. *J Bioinforma Proteomics Rev*. 2016;2(2):1–3.