Rare, Potentially Pathogenic Variants in *ZNF469* Are Not Enriched in Keratoconus in a Large Australian Cohort of European Descent

Sionne E. M. Lucas,[1] Tiger Zhou,[2] Nicholas B. Blackburn,[1,3] Richard A. Mills,[2] Jonathan Ellis,[4] Paul Leo,[5] Emmanuelle Souzeau,[2] Bronwyn Ridge,[2] Jac C. Charlesworth,[1] Matthew A. Brown,[4] Richard Lindsay,[6] Jamie E. Craig,[2] and Kathryn P. Burdon[1]

[1]Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania, Australia
[2]Department of Ophthalmology, Flinders University, Adelaide, South Australia, Australia
[3]South Texas Diabetes and Obesity Institute, School of Medicine, University of Texas Rio Grande Valley, Brownsville, Texas, United States
[4]Queensland University of Technology and Translational Research Institute, Princess Alexandra Hospital, Brisbane, Queensland, Australia
[5]Institute of Health and Biomedical Innovation, Queensland University of Technology, Brisbane, Queensland, Australia
[6]Richard Lindsay and Associates, East Melbourne, Victoria, Australia

PURPOSE. The Zinc Finger Protein 469 (*ZNF469*) gene has been proposed as a candidate gene for keratoconus due to the association of an upstream polymorphism (rs9938149) with the disease in two independent studies, and the role of the gene in the autosomal recessive disease Brittle Cornea Syndrome. Coding variants in *ZNF469* have been assessed for association with keratoconus in several small studies, with conflicting results. We assessed rare, potentially pathogenic variants in *ZNF469* for enrichment in keratoconus patients in a cohort larger than all previous studies combined.

METHODS. *ZNF469* was sequenced in 385 Australian keratoconus patients of European descent, 346 population controls, and 230 ethnically matched screened controls by either whole exome sequencing or targeted gene sequencing. The frequency of rare and very rare potentially pathogenic variants was compared between cases and controls using χ^2 or Fisher's exact tests and further explored using a gene based test (Sequence Kernel Association Test [SKAT]), weighting on the rarity of variants.

RESULTS. A total of 49 rare, including 33 very rare, potentially pathogenic variants were identified across all groups. No enrichment of rare or very rare potentially pathogenic variants in *ZNF469* was observed in our cases compared to the control groups following analysis using χ^2 or Fisher's exact tests. This finding was further supported by the SKAT results, which found no significant difference in the frequency of variants predicted to be damaging between cases and either control group ($P = 0.06$).

CONCLUSIONS. Rare variants in *ZNF469* do not contribute to keratoconus susceptibility and do not account for the association at rs9938149.

Keywords: DNA sequencing, genomics, cornea, keratoconus, candidate genes

Keratoconus (OMIM 148300) is a complex disease characterized by the progressive thinning and protrusion of the cornea. This causes high myopia and irregular astigmatism and can severely affect vision and quality of life.[1] Standard treatments such as hard contact lenses are difficult to manage, are often painful for the patient, and corneal transplantation is ultimately required in a large proportion of patients.[2,3] Keratoconus has an incidence of 1 in 50,000[2,4] and usually develops between puberty and early adulthood. The prevalence is reported to be between 17 and 229 per 100,000, and varies greatly between studies, ethnicities, and geographical locations.[2,5–10] In Caucasians, the prevalence is estimated at 54.5 per 100,000.[2] There is strong evidence for the role of genetic risk factors in keratoconus susceptibility; however, largely the functional variants and genes involved in disease development and progression have not been elucidated.[10–12] Identifying genetic factors underlying keratoconus susceptibility would improve our understanding of the disease pathogenesis and aid the development of novel, nonsurgical treatments, which would ultimately improve the quality of life for patients.

As keratoconus progresses, the cornea thins and bulges forward forming a cone shape. Therefore, it has been hypothesized that genes involved in central corneal thickness (CCT) may play a role in the etiology of keratoconus. Under this hypothesis, a genome-wide association study (GWAS) by Lu et al.[13] for CCT assessed genome-wide significant single nucleotide polymorphisms (SNPs) in a large cohort of keratoconus patients.[13] A SNP upstream of the Zinc Finger Protein 469 (*ZNF469*) gene, rs9938149, showed a suggestive association with keratoconus; however, the genotype associated with a

thinner cornea was associated with decreased keratoconus risk.[13] This finding was replicated in an independent cohort showing the same direction of association,[14] indicating that the association is likely to be real, if nonintuitive. As *ZNF469* is the closest gene to this SNP, it has been hypothesized that genetic variation within the gene may account for the association at rs9938149 as well as contribute to CCT and keratoconus susceptibility.

The potential role of *ZNF469* in keratoconus pathogenesis is further supported by the role of this gene in Brittle Cornea Syndrome type 1 (BCS1, OMIM 229200). BCS1 is a rare, autosomal recessive connective tissue disorder, caused by biallelic loss-of-function variants in *ZNF469*. A key feature of this syndrome is extremely thin corneas that are prone to spontaneous rupture. This suggests that *ZNF469* is important for the structural integrity of the cornea.

As an appealing candidate, coding variants in *ZNF469* in keratoconus have since been investigated. Two of these studies reported an association of potentially pathogenic variants in *ZNF469* with keratoconus, while two showed no association with disease.[15-18] Lechner et al.[17] assessed 112 cases and 96 local controls in a combined United Kingdom and Swiss cohort while Vincent et al.[18] examined a largely Polynesian cohort involving 43 cases and 46 controls from New Zealand. Both of these studies concluded that rare potentially pathogenic variants in *ZNF469* were enriched in keratoconus. Conversely, work by Karolak and colleagues[16] did not identify any difference in the frequency of nonsynonymous variants between 42 cases, 49 individuals with high myopia, and 268 controls without eye disease in a Polish cohort. Furthermore, a study by Davidson et al.[15] showed that uncommon and rare potentially pathogenic variants in 11 families with keratoconus (six of Middle Eastern heritage and five of mixed ethnicities) did not segregate with disease. This study also examined four heterozygous carriers of nonsense or frame-shift variants in *ZNF469*, and found no evidence of keratoconus.[15]

The potential involvement of *ZNF469* in keratoconus remains unclear and further assessment of this candidate gene is required. It is likely that the contrasting findings from previous studies are due to their small sample sizes. To address this, our study examined the frequency of rare potentially pathogenic variants in *ZNF469* in 385 keratoconus cases compared to ethnically matched controls. By assessing more cases than all previous studies combined, the present study allows for a more comprehensive analysis of the potential contribution of rare variants in *ZNF469* to keratoconus risk.

METHODS

All investigations adhered to the principles of the Declaration of Helsinki and were approved by the Southern Adelaide Clinical Human Research Committee and the Human Research Ethics committee Tasmania. All participants gave written informed consent.

Case Cohort

Australian keratoconus patients of European descent were referred by treating optometrists and ophthalmologists and recruited through the Flinders Eye Clinic (Adelaide, Australia) and optometry clinics in Adelaide and Melbourne. Additionally, members of Keratoconus Australia were recruited from across Australia via mail. Clinical examinations were performed by an experienced ophthalmologist. Individuals were diagnosed with keratoconus if they had videokeratographic features of keratoconus or any of the following clinical signs: conical corneal protrusion, central or paracentral stromal thinning or

other distinctive features such as Fleischer's ring, Vogt's striae, epithelial or subepithelial scarring, or oil droplet sign and/or scissoring of the retinoscopic reflex. In addition, individuals with a history of corneal transplantation for keratoconus were also classified as cases. Peripheral blood was donated by patients and DNA was extracted using the QiaAmp DNA Maxi Kit (Qiagen, Hilden, Germany).

Control Cohorts

Two Australian control cohorts of European descent were used in this study. The first, the "population controls", were an ethnically matched cohort from the Anglo-Australasian Osteoporosis Genetics Consortium ($n = 346$). These individuals are all female with moderately high or low bone mineral density measurements ($1.5 < |BMD| < 4.0$), but were not examined for eye disease. The second control group, the "screened controls", were a cohort of Australians of European descent involved in other genetic studies who were examined for eye disease and found to be unaffected by experienced ophthalmologists at the Flinders Eye Clinic ($n = 230$). While these individuals had no clinical evidence of keratoconus, the majority of these individuals had advanced glaucoma ($n = 195$). The remaining individuals either had no evidence of eye disease ($n = 22$) or were unaffected individuals from families with congenital cataract ($n = 9$) and nanophthalmos ($n = 4$).

Whole Exome Sequencing

Whole exome sequencing (WES) data were available for 99 keratoconus cases, the screened control cohort ($n = 230$) and the population control cohort ($n = 346$). WES for the keratoconus patients and the screened control cohort was conducted by Macrogen, Inc., using the SureSelect Human All Exon V4 enrichment kit (Agilent Technologies, Inc., Santa Clara, CA, USA) with paired end sequencing on a HiSeq 2000 (Illumina, San Diego, CA, USA). Raw reads were aligned to the hg19 reference genome with BWA-MEM using the Churchill Pipeline[19] and variants were joint-called with SAMtools and BCFtools (versions 1.3.1).[20] The population control cohort was sequenced at the University of Queensland Centre for Clinical Genomics using the TruSeq Exome Enrichment (Illumina) on an Illumina HiSeq 2000. Raw reads were aligned to the reference genome (hg19) using novoalign (version .02.08), and GATK[21] (version 3.2-2) was used for variant calling and quality score calibration according to GATK's "Best Practices Guidelines".[22,23] To assess coverage across *ZNF469* in the WES data (cases and the two control groups), the mean read depth for each variant position was determined using VCFtools[24] and graphed using custom R[25] scripts using the R packages ggplot2[26] and cowplot.[27] For each cohort, regions with a mean depth of <10 were excluded from analyses. These regions were different for the two control cohorts. Variants were only included in analyses if a sequencing depth of ≥ 10 and a quality score ≥ 20 was obtained.

Targeted Gene Screen in Additional Cases

ZNF469 was screened in a total 341 cases as part of a candidate gene screen using the HaloPlex Target Enrichment System (Agilent Technologies, Inc.) with a custom designed probe panel on pooled DNA samples (Supplementary Table S1). A brief summary of the design is presented in Supplementary Figure S1. DNA pools containing equimolar DNA samples from eight keratoconus patients were prepared as published previously.[28] Each DNA pool was indexed with a unique indexing primer cassette, allowing for multiplexed sequencing. Sequencing was conducted in batches of 11 DNA pools on the

TABLE 1. Demographics of the Keratoconus Cases and Control Groups at the Time of Examination

Cohort	No. of Individuals	Mean Age in Years (Range)	% Female	Disease Status
Cases	385	45.2 (14–85)	44.2	Affected
Population controls*	346	69.5 (46–86)	100	Unscreened
Screened controls	230	68.6 (8–92)	56.5	Unaffected

* The demographics were collected at the time of bone densitometry for this group.

MiSeq platform (Illumina) using a MiSeq V2 Reagent kit (300 cycles) with paired-end reads. SureCall (Agilent Technologies, Inc.) was used for sequencing analysis using standard trimmer parameters, the BWA-MEM algorithm to align reads and the SNPPET SNP Caller (part of SureCall) to call variants. Variants were called if a minimum read depth of 10 and quality score of 20 was reached. As variants were called from pooled DNA samples, it was expected that if a single alternate allele was present it would be observed on 6.25% of the reads mapping to that position. To account for this, the minimum allele frequency for heterozygous variants was set to 0.035 and the threshold for indel calling was ≥ 0.04. Suspected artefacts and real variants were selected for validation by direct sequencing in all individuals included in the DNA pool, to identify thresholds for the inclusion of variants. In addition, 55 cases included in this candidate gene screen were included in the WES; providing additional cross-validation between the two sequencing strategies. However, variants from these samples were only counted in the analyses once. An in-depth description of the validation of variant calls from the pooled gene screen and the development of thresholds for variant inclusion is available in the Supplementary Materials.

Variant Validation by Direct Sequencing

To validate variants called in the case data set, primers were designed using Primer3Plus.[29] DNA samples were amplified with either GoTaq (Promega, Madison, WI, USA), HotStar Taq Kit (Qiagen), or MyTaq HS Mix (Bioline, London, UK) and purified using Agencourt AMPure XP (Beckman Coulter, Brea, CA, USA) according to the manufacturer's instructions. Purified amplicons were sequenced using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City, CA, USA) on either an ABI 310 Genetic Analyzer or an ABI 3500 (Applied Biosystems). DNA sequences were aligned to the human reference genome (hg19), and the chromatograms were manually inspected at the position of each variant using Sequencher 4.10.1 (Gene Codes Corporation, Ann Arbor, MI, USA).

Variant Filtering Strategies

While it has recently been suggested that *ZNF469* is a single-exon gene,[30] the capture methods in this study were all designed under the assumption that *ZNF469* has two exons with a short 84bp intron. Therefore, only exonic variants were included in analyses. To ensure high quality sequence data and robust comparisons, only regions with sufficient coverage across the relevant data sets were included in each comparison. ANNOVAR[31] was used to annotate variants with the minor allele frequency (MAF) of the non-Finnish European population in the Exome Aggregation Consortium database[32] (ExAC NFE), as well as the predicted pathogenicity of the variants using Sorting Tolerant from Intolerant (SIFT)[33] and the HumDIV algorithm from Polymorphism Phenotyping v2 (Poly-Phen2).[34] To make our results comparable to previous studies, two filtering strategies were used to define rare and very rare potentially pathogenic variants. Filtering strategy 1 included nonsynonymous variants with a MAF < 0.01 (rare) with a

damaging or probably/possibly damaging prediction by SIFT and/or PolyPhen2. As SIFT and PolyPhen2 do not assess insertions/deletions (indels), all indels with a MAF <1% were included in filtering strategy 1. Filtering strategy 2 was identical to filtering strategy 1, except variants were required to have a MAF < 0.001 (very rare).

Statistical Analyses

For both filtering strategies, the number of alternate alleles was compared between cases and the two control data sets separately using χ^2 or Fisher's exact tests, where appropriate. The odds ratio (OR) and 95% confidence interval (95% CI) were calculated. In addition, the Sequence Kernel Association Test (SKAT)[35] was used to assess if potentially pathogenic variants were enriched in the case WES data compared to the two control cohorts. Using the two control cohorts as a separate comparison, SKATBinary was run using the quantile adjusted moment matching (QA) method and the default weight parameter, Beta (1,25). This weighting applies strong weights to rare variants, nonzero weighting to uncommon variants (MAF 0.01–0.05), and almost zero weights to common variants.[35]

Genetic Power Calculations

Power calculations were conducted using the case-control for discrete traits module of the genetic power calculator (http://zzz.bwh.harvard.edu/gpc/).[36] These calculations assumed an additive module where the prevalence of keratoconus was 0.00067 (1 in 1500) and D' prime was equal to 1. The high-risk allele frequency was set for each filtering strategy separately, using the frequency of variants identified in the screened controls.

Variant Visualization

Variants included in filtering strategy 1 were plotted as a bar plot for each group using custom R[25] scripts and the R package ggplot2.[26] Each variant was plotted along the *x*-axis according to its genomic position, and the frequency of the variant in the study group was indicated by the height of the bar. Indels were plotted according to the position of the first affected base. To indicate the conservation of each variant position, the bars were colored using a color gradient corresponding to the 100-way vertebrate PhastCons[37,38] score as available from the University of California Santa Cruz (UCSC) genome browser.[39] The plots for each group were aligned to a schematic of *ZNF469* using the R package cowplot,[27] to allow for comparison of the variants across the gene. The position of the zinc finger motifs was plotted onto the gene schematic based on the positions obtained from UniProt[40] (entry Q96JG9).

RESULTS

A total of 385 cases, 346 population controls, and 230 screened controls were included in analyses. Demographic details are provided in Table 1.
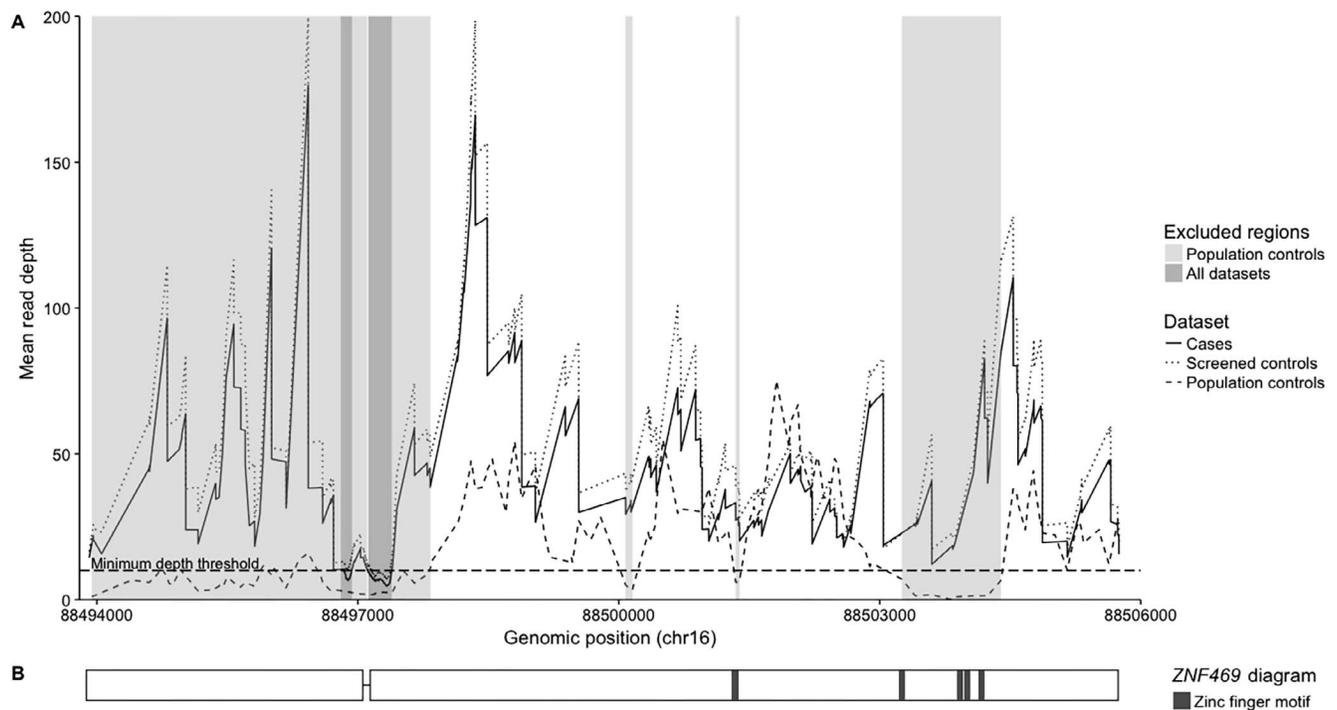
FIGURE 1. Sequencing coverage in the WES data sets. (A) Coverage across ZNF469 in each data set, based on the mean read depth at variant positions. The *horizontal dashed line* indicates the minimum depth threshold accepted (10 reads). Regions where the mean read depth in the population controls was below this threshold are shaded *light gray* and were excluded from analyses in which these controls were included. Regions with insufficient coverage in all cohorts are shaded *darker gray*. (B) A schematic of ZNF469 where *dark gray boxes* indicate the position of the zinc finger motifs and the small intron is indicated by the *horizontal line*.

As we applied a stringent coverage filter to all sequence data sets, regions containing the majority of the first exon, the beginning of exon two, and the regions encoding the zinc finger domains were excluded from analyses that compared the cases to the population control cohort (Fig. 1). Therefore, comparisons using the population controls included a total of 6700 bases across *ZNF469* with high quality sequence data, which corresponds to 56.5% of the coding regions. In contrast, a total of 11,448 bases (96.5% of the coding regions) obtained sufficient coverage across the gene in both the cases and screened controls. For comparisons using the screened controls, only two small regions (85 bp and 279 bp) either side of the intron were excluded from analysis due to poor coverage (Fig. 1). By removing these regions from analysis, the single region that did not meet the minimum depth threshold in the pooled gene screen data was also excluded. A summary of the coverage metrics for the gene screen data set is presented in Supplementary Table S2.

Across all cohorts, 49 variants (46 single nucleotide variants and three small deletions) fulfilled the criteria for rare potentially pathogenic variants in filtering strategy 1 (Table 2). Of these, 19 variants were observed in cases only, while 17 were unique to the control cohorts and the remaining 13 were in both cases and controls. Eleven variants had been observed in previous studies in either cases or controls, with an additional two variants that are not identical, but occur at the same amino acid as a previously reported variant. A total of 33 rare potentially pathogenic variants were observed 73 times in cases and 23 variants were called 48 times in the screened controls, which was not significantly different ($P = 0.66$; Table 3). When considering the comparison between the cases and the population controls, 17 variants were identified 31 times in cases and 10 variants were observed 22 times in the controls,

which was not significantly different between groups ($P = 0.47$; Table 3).

Filtering strategy 2 identified 33 very rare potentially pathogenic variants across all cohorts (Table 2). Sixteen of these variants were unique to keratoconus patients and two, p.(C1693F) and p.(P2548L), were identified in two cases each. Fifteen variants were observed in the control cohorts only, including p.(P3372L), which was identified in two individuals. Two variants, p.(P626_G628del) and p.(E3781K), were identified in both a case and a control. All other variants were observed in a single individual. Two variants, p.(S2242Y) and p.(P3372L), were previously observed in the study by Lechner et al.[17] and p.(E935G) is located at the same amino acid as a variant identified in the study by Davidson and colleagues.[15] In total, 18 variants were identified 20 times in cases and 12 variants were observed 13 times in the screened controls. For the population control comparison, 10 variants were observed in 12 cases and five variants were identified in the controls. Similar to the results for rare variants, very rare variants were not enriched in cases compared to either the screened controls ($P = 0.96$) or the population controls ($P = 0.15$; Table 3). Furthermore, the SKAT analyses demonstrated no significant enrichment of variants predicted to be damaging ($P = 0.06$; Table 4).

For the power calculations, the high-risk allele frequency was assumed to be the frequency of variants identified by the filtering strategies in the 230 screened controls. This was determined to be 0.104 (48/460) for strategy 1 and 0.028 (13/460) for strategy 2. Using these values, our study had 80% power to detect a relative risk of 1.5 and 2.0, respectively.

For each cohort, the allele frequencies were plotted against the variant position and mapped to a schematic of the gene (Fig. 2). The first exon showed a similar pattern of variation

TABLE 2. Variants Included in Analysis Under Filtering Strategy 1 (Rare Potentially Pathogenic Variants) Including the Position, Nucleotide and Protein Changes, Population Frequency, the 100-Way Vertebrate PhastCons Score, Cohort Frequency (Freq) and Allele Count (AC)

Position (hg19)	Nucleotide Change	Protein Change	MAF in ExAC NFE	PhastCons Score	Case Freq (AC)	Screened Controls Freq (AC)	Population Controls Freq (AC)	Included in FS2	Observed in Other Published Studies
chr16:88494603	c.725G>T	p.(S242I)	0.0024	0.9930	0.0013 (1)	0.0022 (1)	*	N	
chr16:88494809	c.931G>A	p.(G311R)	0.0000	0.0000	0.0013 (1)	0.0000	*	Y	
chr16:88495361	c.1483C>T	p.(P495S)	0.0015	0.0040	0.0013 (1)	0.0022 (1)	*	N	
chr16:88495385	c.1507C>T	p.(R503W)	0.0003	0.9980	0.0013 (1)	0.0000	*	Y	
chr16:88495400	c.1522G>A	p.(A508T)	0.0003	0.0000	0.0013 (1)	0.0000	*	Y	
chr16:88495443	c.1565G>T	p.(G522V)	0.0000	0.0010	0.0000	0.0022 (1)	*	Y	
chr16:88495461	c.1583C>G	p.(P528R)	0.0000	0.0000	0.0013 (1)	0.0000	*	Y	
chr16:88495487	c.1609G>A	p.(V537M)	0.0010	0.0000	0.0038 (3)	0.0022 (1)	*	N	[17]
chr16:88495568	c.1690G>C	p.(G564R)	0.0002	0.3100	0.0013 (1)	0.0000	*	Y	
chr16:88495575	c.1697C>T	p.(A566V)	0.0098	0.0000	0.0115 (9)	0.0283 (13)	*	N	[15,18]
chr16:88495753-88495761	c.1875_1883del	p.(P626_P628del)	0.0000	0.0000	0.0013 (1)	0.0022 (1)	*	Y	
chr16:88495872	c.1994C>T	p.(P665L)	0.0062	0.0420	0.0102 (8)	0.0109 (5)	*	N	[17]
chr16:88495913	c.2035G>A	p.(E679K)	0.0064	0.6550	0.0013 (1)	0.0022 (1)	*	N	[15]
chr16:88496430	c.2552T>C	p.(M851T)	0.0000	1.0000	0.0000	0.0022 (1)	*	Y	
chr16:88496682	c.2804A>G	p.(E935G)	0.0000	0.0040	0.0013 (1)	0.0000	*	Y	[15†]
chr16:88497443	c.3481C>A	p.(P1161T)	0.0000	0.0000	0.0000	0.0022 (1)	*	Y	
chr16:88497830	c.3868G>A	p.(D1290N)	0.0000	0.0020	0.0000	0.0022 (1)	*	Y	
chr16:88498350	c.4388C>T	p.(T1463M)	0.0035	0.0020	0.0000	0.0022 (1)	0.0058 (4)	N	[17]
chr16:88499040	c.5078G>T	p.(C1693F)	0.0002	0.0000	0.0026 (2)	0.0000	0.0000	Y	
chr16:88500687	c.6725C>A	p.(S2242Y)	0.0001	0.0000	0.0000	0.0000	0.0014 (1)	Y	[17]
chr16:88500713	c.6751C>T	p.(P2251S)	0.0000	0.0040	0.0000	0.0022 (1)	0.0000	Y	
chr16:88500822	c.6860C>G	p.(P2287R)	0.0002	0.0010	0.0013 (1)	0.0000	0.0000	Y	
chr16:88500957	c.6995C>T	p.(P2332L)	0.0033	0.0000	0.0064 (5)	0.0000	0.0000	N	
chr16:88501145	c.7183C>A	p.(P2395T)	0.0044	0.0000	0.0038 (3)	0.0022 (1)	0.0000	N	
chr16:88501224	c.7262G>A	p.(R2421H)	0.0002	0.0000	0.0013 (1)	0.0000	0.0000	Y	
chr16:88501344	c.7382G>A	p.(R2461Q)	0.0002	1.0000	0.0000	0.0022 (1)	*	Y	
chr16:88501431	c.7469C>A	p.(P2490H)	0.0034	0.0000	0.0013 (1)	0.0000	0.0000	N	[17]
chr16:88501605	c.7643C>T	p.(P2548L)	0.0000	0.0000	0.0026 (2)	0.0000	0.0000	Y	
chr16:88501749	c.7787C>T	p.(P2596L)	0.0010	0.0000	0.0000	0.0000	0.0058 (4)	N	
chr16:88501995	c.8033C>T	p.(A2678V)	0.0000	0.0000	0.0013 (1)	0.0000	0.0000	Y	
chr16:88502222	c.8260C>T	p.(H2754Y)	0.0009	0.0000	0.0000	0.0022 (1)	0.0000	Y	
chr16:88502276	c.8314C>A	p.(L2772M)	0.0000	0.0000	0.0000	0.0000	0.0014 (1)	Y	
chr16:88502304	c.8342C>T	p.(P2781L)	0.0000	0.0000	0.0013 (1)	0.0000	0.0000	Y	
chr16:88502583	c.8621C>T	p.(T2874M)	0.0011	0.0000	0.0026 (2)	0.0000	0.0000	N	
chr16:88502666	c.8704G>T	p.(D2902Y)	0.0010	0.9700	0.0038 (3)	0.0000	0.0029 (2)	N	[17]
chr16:88502723	c.8761C>G	p.(P2921A)	0.0000	0.0030	0.0000	0.0000	0.0014 (1)	Y	
chr16:88502862	c.8900C>T	p.(A2967V)	0.0000	0.0030	0.0013 (1)	0.0000	0.0000	Y	
chr16:88502972-88502986	c.9010_9024del	p.(L3004_T3008del)	0.0043	0.0000	0.0051 (4)	0.0022 (1)	0.0014 (1)	N	[15†,17†]
chr16:88503039	c.9077A>C	p.(E3026A)	0.0000	0.9690	0.0013 (1)	0.0000	0.0000	Y	
chr16:88503372	c.9410A>G	p.(E3137G)	0.0000	1.0000	0.0013 (1)	0.0000	*	Y	
chr16:88503596	c.9634A>T	p.(R3212W)	0.0000	0.0000	0.0000	0.0022 (1)	*	Y	
chr16:88503838	c.9876G>T	p.(E3292D)	0.0000	0.0010	0.0000	0.0022 (1)	*	Y	
chr16:88504077	c.10115C>T	p.(P3372L)	0.0000	1.0000	0.0000	0.0043 (2)	*	Y	[17]
chr16:88504239	c.10277G>A	p.(R3426Q)	0.0078	0.3970	0.0140 (11)	0.0196 (9)	*	N	[15,17]
chr16:88504492	c.10530delC	p.(I3510fs)	0.0000	0.0000	0.0000	0.0000	0.0014 (1)	Y	
chr16:88504766	c.10804C>T	p.(R3602C)	0.0031	0.0000	0.0013 (1)	0.0022 (1)	0.0087 (6)	N	[17]
chr16:88504775	c.10813T>G	p.(C3605G)	0.0000	0.0000	0.0000	0.0000	0.0014 (1)	Y	
chr16:88505303	c.11341G>A	p.(E3781K)	0.0005	1.0000	0.0013 (1)	0.0022 (1)	0.0000	Y	
chr16:88505654	c.11692G>A	p.(E3898K)	0.0000	1.0000	0.0013 (1)	0.0000	0.0000	Y	

FS2, Filtering Strategy 2 (very rare potentially pathogenic variants).
* Variant with insufficient coverage.
† Variant previously reported at the same amino acid, but a different nucleotide.

between the cases and screened controls (Fig. 2). Due to insufficient coverage of this region in the population controls, variation in this group could not be assessed. Of the 13 variants in cases in exon 1, two variants, p.(S242I) and p.(R503W), were located at highly conserved nucleotides (PhastCons score >0.99) and p.(E679K) was at a relatively conserved position with a PhastCons score of 0.655. The p.(R503W) variant was observed in a single case (frequency of 0.0013), whereas both p.(S242I) and p.(E679K) were identified in one case and one screened control (frequency of 0.0022).

TABLE 3. Association Analyses Using χ^2 or Fisher's Exact Test Under Each Filtering Strategy

	Cohort	Alternate Alleles	Wild Type Alleles	Alternate Allele Frequency	P Value	OR [95% CI]
Filtering strategy 1						
Comparison 1	Cases	73	697	0.09		
	Screened controls	48	412	0.1	0.66	0.90 [0.60–1.34]
Comparison 2	Cases	31	739	0.04		
	Population controls	22	670	0.03	0.47	1.28 [0.71–2.31]
Filtering strategy 2						
Comparison 1	Cases	20	750	0.03		
	Screened controls	13	447	0.03	0.96	0.92 [0.43–1.97]
Comparison 2	Cases	12	758	0.02		
	Population controls	5	687	0.01	0.15*	2.18 [0.71–7.11]

For each cohort, the number of alternate and wild type alleles and the frequency of the alternate allele are reported. For each comparison, the *P* value, the OR, and the 95% CI are shown.

* Fisher's exact *P* value.

Few variants were located in the proximal region of exon 2, with only one variant identified in cases. In contrast, a cluster of variants was observed in cases in the distal half of exon 2 (Fig. 2). Specifically, variants were observed in the region spanning just proximal of the first zinc finger motif to midway between the second and third zinc finger motif. Noticeably fewer variants were observed in the control cohorts within this region (Fig. 2). Three variants identified in cases within this cluster, p.(D2902Y), p.(E3026A) and p.(E3137G), were highly conserved (PhastCons scores > 0.96). The p.(D2902Y) variant was present at a similar frequency in cases (0.0038) and the population controls (0.0029), while p.(E3026A) and p.(E3137G) were present in one case and absent in controls. Two additional variants, p.(E3781K) and p.(E3898K), were identified at highly conserved residues at the very distal end of exon two (PhastCons score = 1). These variants were both identified a single case and neither of these variants were observed in the population controls despite good coverage in this region; however, the p.(E3781K) variant was observed in a one screened control.

Across *ZNF469*, three variants—p.(A566V), p.(P665L), and p.(R3426Q)—were identified at a frequency >0.01 in both the cases and the screened control cohort, despite a MAF < 0.01 in the public ExAC NFE database. The frequency of these variants could not be assessed in the population controls due to insufficient coverage; however, all three variants were observed in cases at a similar or lower frequency than the screened controls.

DISCUSSION

In our cohort of Australians of European descent, the overall frequency of rare potentially pathogenic variants in *ZNF469* was not different between keratoconus cases and two independent control cohorts using two filtering strategies. Despite a much larger cohort than the earlier studies, our findings were unable to replicate previous reports that show

TABLE 4. The Results of the SKAT Analyses

Comparison	Alternate Alleles	MAC	Carrier Individuals	P Value
Cases vs. screened controls	38	415	271	0.06
Cases vs. population controls	20	473	337	0.06

For each comparison, the number of alternate alleles included in the analysis, the total minor allele count (MAC), the number of individuals with these variants (carrier individuals), and the *P* value.

an association of potentially pathogenic variants in *ZNF469* and keratoconus development with relative risks of up to 12.[17,18] Our study had 80% power to detect a relative risk of 1.5 for rare potentially pathogenic variants, and 2.0 for very rare potentially pathogenic variants, commensurate with expected effect sizes for rare variants in a complex disease. In our study, we show that the reported large effect sizes do not exist when the whole gene is considered in both cases and controls and therefore if this gene is contributing to keratoconus risk, then effect sizes are likely to be quite small. The main finding of this paper is that the reported large effects of rare variants are likely spurious and are brought about by biased reporting of variants detected in fully sequenced cases not being present in controls who are not fully sequenced or due to the exclusion of variants that are observed in control individuals. Our work shows that controls have just as many variants meeting the potentially pathogenic criteria as keratoconus patients.
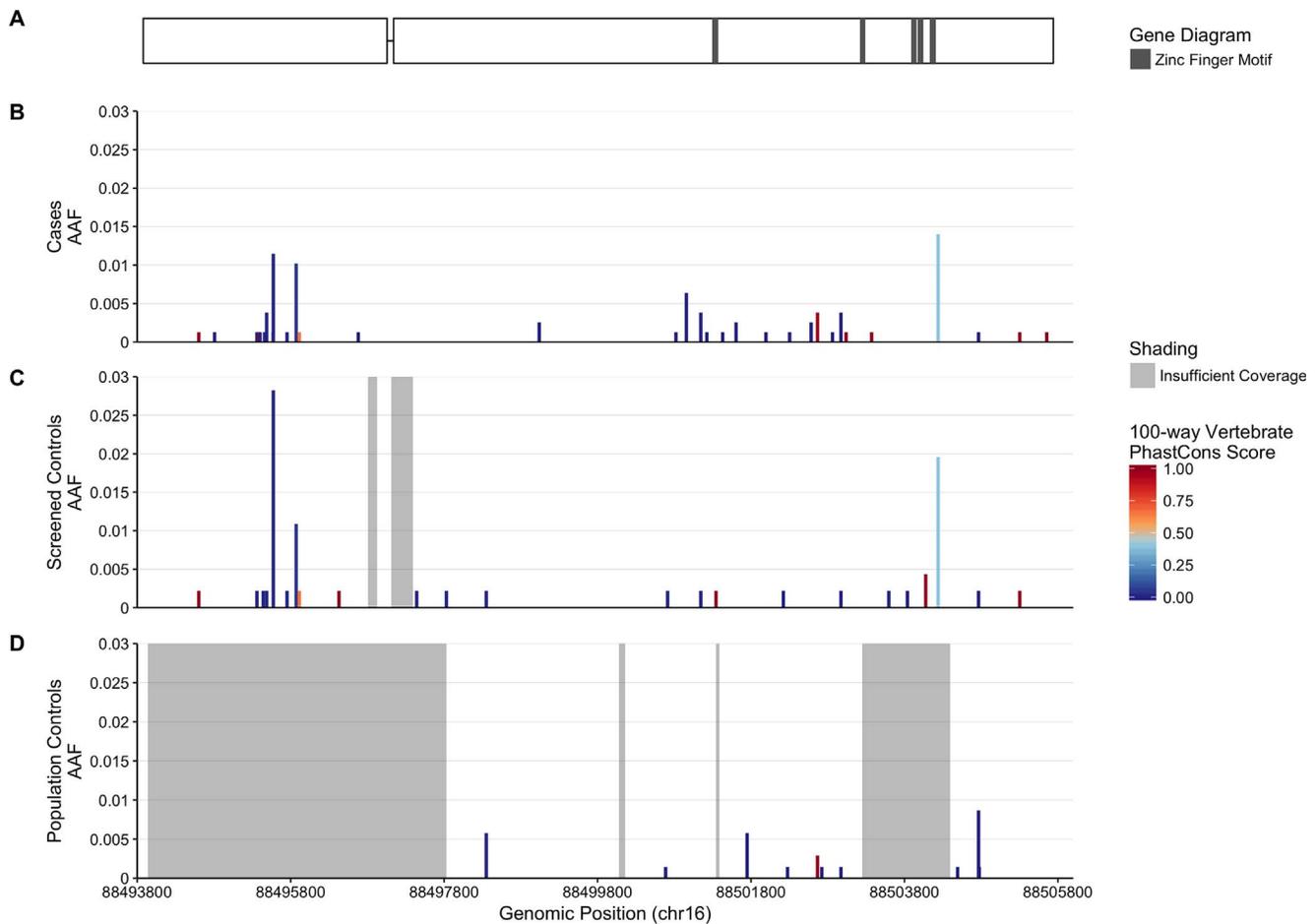
Due to the size of our study, we were also able to map the location of the rare potentially pathogenic variants identified in all three cohorts and show that they span the whole gene, with particular aggregation in the first exon and the distal half of the second exon. Only eight variants observed in cases were located at highly conserved nucleotides. Four of these variants were identified at similar frequencies in both cases and controls and four were only observed in a single case. While it is possible that these rare variants may contribute to keratoconus susceptibility in these few cases, on the whole, the evidence indicates that rare, potentially pathogenic variants in *ZNF469* do not make a substantial contribution to keratoconus risk. This finding is consistent with the work by Davidson and colleagues,[15] which showed that uncommon variants (MAF <0.025) did not segregate with disease in families with keratoconus and therefore, at least in isolation, do not contribute to keratoconus susceptibility. Furthermore, the results of the Polish study[16] indicate that potentially pathogenic variants are not enriched in keratoconus, and that *ZNF469* is highly allelic in the general population.

The first reports to assess coding variants in *ZNF469* were published by Lechner et al.[17] and Vincent et al.[18] Subsequent studies, including this one, have used similar filtering strategies to allow for direct comparison. Our filtering strategy 1 was based on the criteria used by Vincent et al.,[18] while filtering strategy 2 was based on the method used by Lechner et al.[17] with two key changes. Firstly, Lechner et al.[17] removed any variants from analysis that were present in both their cases and controls. We did not do this as keratoconus is a complex disease and therefore it is likely (and expected) that unaffected individuals will carry risk associated variants without ever

FIGURE 2. Summary of identified variants. (**A**) Schematic of *ZNF469* with zinc finger motifs shaded *dark gray*, (**B–D**) *bar plots* indicating the position and alternate allele frequency (AAF) of rare potentially pathogenic variants in 784 case alleles, 460 screened control alleles, and 692 population control alleles, respectively. The *bars* are colored according to the 100-way vertebrate PhastCons Score for the corresponding position, where *dark blue* is a score of 0 and *red* is a score of 1. *Light gray* shading on the plots for the control cohorts indicate regions with insufficient coverage for variant calling that were excluded from analysis when using these data. The gene schematic and all graphs are aligned vertically to share the same *x*-axis to allow for comparison.

developing disease. Secondly, Lechner et al.[17] only used SIFT predictions to classify variants as potentially pathogenic. SIFT uses protein conservation to calculate pathogenicity by comparing a query sequence to similar sequences with similar function.[33] As *ZNF469* is a highly variable gene with low conservation in lower mammals and vertrates,[41] SIFT may misclassify deleterious variants in regions of low conservation. In contrast, PolyPhen2 uses the properties of the substituted amino acids and the proximity to functional domains or structural features, as well as protein conservation to predict pathogenicity.[34] As the structure and function of the ZNF469 protein remains largely unknown,[41] PolyPhen2 is likely to better assess regions with poor conservation, particularly the regions that flank the zinc finger domains. Therefore, our study used both SIFT and PolyPhen2 to better capture the pathogenicity of nonsynonymous variants identified in *ZNF469*. Our study used these robust and complementary methods to replicate the analytical strategies of the previous studies; however, our findings do not support any enrichment of rare potentially pathogenic variants in *ZNF469* in keratoconus cases.

We assessed rare potentially pathogenic variants in *ZNF469* in the largest cohort of keratoconus patients to date by combining WES data and sequencing data from a targeted gene screen using pooled DNA samples. Following extensive

validation experiments, we demonstrated a high level of consistency of variant calls for individuals sequenced by both methods, validating the utility of pooling DNA samples to maximize cost-effectiveness. Furthermore, these validation experiments were used to develop stringent thresholds that were then applied to all data to ensure only high quality variants were included in analyses. While variants were not assessed for validation in the control cohorts, the inclusion of additional control variants that were filtered out due to these stringent thresholds would only strengthen our findings of no association. Furthermore, to minimize bias due to poor coverage of *ZNF469* in the population controls, a strict coverage threshold was applied such that only regions with sufficient coverage were assessed when comparing to this control group. While this may have resulted in the exclusion of important variants, the vast majority of these regions had sufficient coverage in the screened controls and therefore were still assessed.

As suggested by Davidson et al.,[15] it is likely that variation in *ZNF469* is underrepresented in public databases as a result of the poor coverage of the gene by the older WES capture techniques. According to the ExAC Browser, the mean coverage for *ZNF469* (ENSG00000225614) is 7.3 reads and the proportion of individuals with at least 10× coverage is less than 20%. Our experience with the poor coverage of *ZNF469*

in the population control data set, as well as the occurrence of three variants that were identified in our cases and screened controls at a frequency of > 0.01, despite being annotated with a MAF < 0.01 in the ExAC NFE database, supports this. As one might expect, these variants could not be assessed in the population controls due to insufficient coverage. All three of these variants were located at relatively nonconserved nucleotides (PhastCons scores <0.4). Two of these variants, p.(P665L) and p.(R3426Q), were observed at similar frequencies in both our cases and controls and therefore were hypothesized to be benign polymorphisms. These variants were similarly classified in the study by Lechner and colleagues, but were not identified in other reports.[15-18] The third variant, p.(A566V), was identified at more than twice the frequency in the screened controls (0.028) than the cases (0.012). Vincent et al.[18] reported this variant in one Indian and two Caucasian keratoconus cases, while Davidson et al.[15] identified the variant in two cases and two unaffected individuals from two separate consanguineous families of Middle Eastern origin. In addition, the work by Lechner et al.[17] excluded the variant from analysis due to a MAF > 0.01 in their control cohort. As our screened control cohort largely consists of advanced glaucoma patients, it is possible that these variants may be involved in glaucoma susceptibility. However, in the original GWAS that identified the association between the SNP upstream of *ZNF469* (rs9938149) and CCT as well as keratoconus, glaucoma cases were also assessed and no association was identified.[13] Therefore, it is more likely that these variants are benign, uncommon polymorphisms.

CONCLUSIONS

Our study included more keratoconus cases than the total number previously studied and demonstrated no significant difference in the overall frequency of potentially pathogenic variants in *ZNF469* compared to controls. In addition, our findings confirm the highly variable nature of *ZNF469* and the poor capture of the gene by previous generation WES capture methods, which is likely to have resulted in underestimates of alternate allele frequencies in public databases. Overall, this study indicates that rare potentially pathogenic variants are unlikely to contribute to keratoconus pathogenesis and do not account for the association at the distant upstream SNP, rs9938149.

References

1. Kymes SM, Walline JJ, Zadnik K, Gordon MO. Quality of life in keratoconus. *Am J Ophthalmol*. 2004;138:527–535.

2. Kennedy RH, Bourne WM, Dyer JA. A 48-year clinical and epidemiologic study of keratoconus. *Am J Ophthalmol*. 1986; 101:267–273.

3. Tuft SJ, Moodaley LC, Gregory WM, Davison CR, Buckley RJ. Prognostic factors for the progression of keratoconus. *Ophthalmology*. 1994;101:439–447.

4. Pearson AR, Soneji B, Sarvananthan N, Sandford-Smith JH. Does ethnic origin influence the incidence or severity of keratoconus? *Eye (Lond)*. 2000;14:625–628.

5. Kok YO, Tan GF, Loon SC. Review: keratoconus in Asia. *Cornea*. 2012;31:581–593.

6. Tanabe U, Fujiki K, Ogawa A, Ueda S, Kanai A. Prevalence of keratoconus patients in Japan [in Japanese]. *Nihon Ganka Gakkai Zasshi*. 1985;89:407–411.

7. Jonas JB, Nangia V, Matin A, Kulkarni M, Bhojwani K. Prevalence and associations of keratoconus in rural Maharashtra in central India: the central India eye and medical study. *Am J Ophthalmol*. 2009;148:760–765.

8. Nielsen K, Hjortdal J, Aagaard Nohr E, Ehlers N. Incidence and prevalence of keratoconus in Denmark. *Acta Ophthalmol Scand*. 2007;85:890–892.

9. Owens H, Gamble GD, Bjornholdt MC, Boyce NK, Keung L. Topographic indications of emerging keratoconus in teenage New Zealanders. *Cornea*. 2007;26:312–318.

10. Rabinowitz YS. Keratoconus. *Surv Ophthalmol*. 1998;42: 297–319.

11. Burdon KP, Vincent AL. Insights into keratoconus from a genetic perspective. *Clin Exp Optom*. 2013;96:146–154.

12. Nielsen K, Hjortdal J, Pihlmann M, Corydon TJ. Update on the keratoconus genetics. *Acta Ophthalmol*. 2013;91:106–113.

13. Lu Y, Vitart V, Burdon KP, et al. Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat Genet*. 2013;45:155–163.

14. Sahebjada S, Schache M, Richardson AJ, et al. Evaluating the association between keratoconus and the corneal thickness genes in an independent Australian population. *Invest Ophthalmol Vis Sci*. 2013;54:8224–8228.

15. Davidson AE, Borasio E, Liskova P, et al. Brittle Cornea Syndrome ZNF469 mutation carrier phenotype and segregation analysis of rare ZNF469 variants in familial Keratoconus. *Invest Ophthalmol Vis Sci*. 2015;56:578–586.

16. Karolak JA, Gambin T, Rydzanicz M, et al. Evidence against ZNF469 being causative for keratoconus in Polish patients. *Acta Ophthalmol*. 2016;94:289–294.

17. Lechner J, Porter LF, Rice A, et al. Enrichment of pathogenic alleles in the brittle cornea gene, ZNF469, in keratoconus. *Hum Mol Genet*. 2014;23:5527–5535.

18. Vincent AL, Jordan CA, Cadzow MJ, Merriman TR, McGhee CN. Mutations in the zinc finger protein gene, ZNF469, contribute to the pathogenesis of keratoconus. *Invest Ophthalmol Vis Sci*. 2014;55:5629–5635.

19. Kelly BJ, Fitch JR, Hu Y, et al. Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol*. 2015; 16:6.

20. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*. 2009; 25:2078–2079.

21. 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.

22. DePristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*. 2011;43:491–498.

23. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis

Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.0.1–11.0.33.

24. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. *Bioinformatics*. 2011;27:2156–2158.

25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available at: https://www.R-project.org/. Accessed November 30, 2016.

26. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. New York, NY: Springer-Verlag; 2009.

27. Wilke CO. cowplot: Streamlined Plot Theme and Plot Annotations for 'ggplot2', R package version 0.7.0; 2016. Available at: https://CRAN.R-project.org/package=cowplot. Accessed June 21, 2016.

28. Burdon KP, Macgregor S, Bykhovskaya Y, et al. Association of polymorphisms in the hepatocyte growth factor gene promoter with keratoconus. *Invest Ophthalmol Vis Sci*. 2011;52:8514–8519.

29. Untergasser A, Nijveen H, Rao X, et al. Primer3Plus, an enhanced web interface to Primer3. *Nucleic Acids Res*. 2007; 35(Web Server issue):W71–W74.

30. Rohrbach M, Spencer HL, Porter LF, et al. ZNF469 frequently mutated in the brittle cornea syndrome (BCS) is a single exon gene possibly regulating the expression of several extracellular matrix components. *Mol Genet Metab*. 2013;109:289–295.

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

32. Lek M, Karczewski KJ, Minikel EV, et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature*. 2016; 536:285–291.

33. Ng PC, Henikoff S. Predicting deleterious amino acid substitutions. *Genome Res*. 2001;11:863–874.

34. Adzhubei IA, Schmidt S, Peshkin L, et al. A method and server for predicting damaging missense mutations. *Nat Methods*. 2010;7:248–249.

35. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89:82–93.

36. Purcell S, Cherny SS, Sham PC. Genetic Power Calculator: design of linkage and association genetic mapping studies of complex traits. *Bioinformatics*. 2003;19:149–150.

37. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15:1034–1050.

38. Siepel A, Haussler D. Phylogenetic hidden Markov models. In: Nielson R, ed. *Statistical Methods in Molecular Evolution*. New York, NY: Springer-Verlag; 2005:325–351.

39. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12:996–1006.

40. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*. 2017;45:D158–D169.

41. Abu A, Frydman M, Marek D, et al. Deleterious mutations in the Zinc-Finger 469 gene cause brittle cornea syndrome. *Am J Hum Genet*. 2008;82:1217–1222.