Methods

GERA. The Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort contains genome-wide genotype, clinical and demographic data of over 110,000 adult members from mainly 4 ethnic groups (non-Hispanic white, Hispanic/Latino, East Asian, and African American) of the Kaiser Permanente Northern California (KPNC) Medical Care Plan[1, 2]. The Institutional Review Board of the Kaiser Foundation Research Institute has approved all study procedures. Patients with pseudophakia were diagnosed by a Kaiser Permanente ophthalmologist and were identified in the KPNC electronic health record system based on the following International Classification of Disease, Ninth (ICD9) or Tenth Revision (ICD10) diagnosis codes: V43.1 (ICD-9 code) and Z96.1 (ICD-10 code). Cataract cases were also identified if they had a history of having a cataract surgery at KPNC. Our control group included all the non-cases. In total, 33,145 patients who have undergone cataract surgery and 64,777 controls from GERA were included in this study.

Protocols for participant genotyping data collection and previous quality control have been described in detail[2]. Briefly GERA participants' DNA samples were extracted from Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada) at KPNC and genotyped at the Genomics Core Facility of UCSF. DNA samples were genotyped at over 665,000 genetic markers on four ethnic-specific Affymetrix Axiom arrays (Affymetrix, Santa Clara, CA, USA) optimized for European, Latino, East Asian, and African American individuals[3, 4]. Genotype quality control (QC) procedures and imputation were conducted on an array-wise basis[2], after an updated genotyping algorithm with an advanced normalization step specifically for SNPs in batches not recommended or flagged by the outlier plate detector than has previously been done. Subsequently, variants were excluded if: >3 clusters were identified; the number of batches was <38/42 (EUR array), <3/5 (AFR), <3/6 (EAS), or <7/9 (LAT); and the ratio of expected allele frequency variance across packages was <100 (EUR), <50 (AFR), <100 (EAS), <200 (LAT). On the EUR array, variants were additionally excluded if heterozygosity >.52 or <.02, and if an association test between Reagent kit v1.0 and

v2.0 had $P<10^{-4}$. Imputation was done by array, and we additionally removed variants with call rates <90%. Genotypes were then pre-phased with Eagle[5] v2.3.2, and then imputed with Minimac3[6] v2.0.1, using two reference panels. Variants were preferred if present in the EGA release of the Haplotype Reference Consortium (HRC; n=27,165) reference panel[7], and from the 1000 Genomes Project Phase III release if not (n=2,504; e.g., indels)[8].

We first analyzed each ethnic group (non-Hispanic white, Hispanic/Latino, East Asian, and African American) separately. We ran a logistic regression of cataract and each SNP using PLINK[9] v1.9 (www.cog-genomics.org/plink/1.9/) adjusting for age, sex, and ancestry principal components (PCs), which were previously[1] assessed within each ethnic group using Eigenstrat[10] v4.2. We included as covariates the top ten ancestry PCs for the non-Hispanic whites, whereas we included the top six ancestry PCs for the three other ethnic groups. To adjust for genetic ancestry, we also included the percentage of Ashkenazi (ASHK) ancestry as a covariate for the non-Hispanic white sample analyses[1].

UK Biobank. The UK Biobank(UKB) is a large prospective study following the health of approximately 500,000 participants from 5 ethnic groups (European, East Asian, South Asian, African British, and mixed ancestries) resident in the UK aged between 40 and 69 years-old at the baseline recruitment visit[11, 12]. Demographic information and medical history were ascertained through touch-screen questionnaires. Participants also underwent a wide range of physical and cognitive assessments, including blood sampling. Cataract cases (N=34,699) were defined as participants with a self-reported cataract operation (f20004 code 1435) or/and a hospital record including a diagnosis code (ICD-10: H25 or H26). Controls (N=452,622) were participants who were not cases. Phenotyping, genotyping and imputation were carried out by members of the UK Biobank team. Imputation to the Haplotype Reference Consortium reference panel plus the 1000 Genomes Project and UK10K reference panels has been described (www.ukbiobank.ac.uk).

Following QC, over 10 million variants in 487,622 individuals were tested for cataract adjusting for age, sex, and genetic ancestry principal components.

GWAS analysis was performed by ethnic group. Ethnic groups were formed by those who reported any white group and with global ancestry $PC_1 \leq 50$ and $PC_2 \leq 50$ (where global PC_1 and PC_2 were calculated from the entire cohort), and by those reporting East Asian, South Asian, African British, and mixed/other ancestries. Ancestry PCs were then calculated within each ethnic group as done in GERA[1], using 50,000 random individuals and the rest projected just for Europeans, and GWAS analysis adjusted for 10 PCs in all ethnic groups.

The analyses presented in this paper were carried out under UK Biobank Resource project #14105.

GWAS meta-analyses. First, a meta-analysis of cataract was conducted in GERA to combine the results of the 4 ethnic groups using the R[13] (https://www.R-project.org) package ''meta''. Similarly, a meta-analysis was conducted in UKB to combine the results of the 5 ethnic groups. Three ethnic-specific meta-analyses were also performed: 1) combining European-specific samples (i.e. GERA non-Hispanic whites and UKB Europeans); 2) combining East Asian-specific samples (i.e. GERA and UKB East Asians); and 3) combining African-specific samples (i.e. GERA African Americans and UKB Africans). A meta-analysis was then conducted to combine the results from GERA and UKB. Two sex-specific meta-analyses were also performed: 1) combining women from GERA and UKB; and 2) combining men from GERA and UKB. Fixed effects summary estimates were calculated for an additive model. We also estimated heterogeneity index, I^2 (0–100%) and *P*-value for Cochrane's Q statistic among different groups, and studies. For each locus, we defined the top SNP as the most significant variant within a 2 Mb window. Novel loci were defined as those that were located over 1 Mb apart from any previously reported locus[14].

***Conditional & joint (COJO) analysis*.** A multi-SNP-based conditional & joint association analysis (COJO)[15] was performed on the combined European-specific (GERA non-Hispanic whites + UKB Europeans) meta-analysis results to potentially identify independent signals within the 44 identified genomic regions. To calculate linkage disequilibrium (LD) patterns, we used 10,000 randomly selected samples from GERA non-Hispanic white ethnic group as a reference panel. A *P*-value less than 5.0×10^{-8} was considered as the significance threshold for this COJO analysis.

Replication in 23andMe. Replication analysis of 54 loci identified in the combined (GERA+UKB) meta-analysis was conducted using self-reported data from a GWAS including 347,209 self-reported cataract cases and 2,887,246 controls (close relatives removed) of 5 ethnic groups (i.e. European, Latino, East Asian, South Asian, and African American) determined through an analysis of local ancestry[16], from 23andMe, Inc., research cohort. Participants provided informed consent and participated in the research online, under a protocol approved by the external AAHRPP-accredited IRB, Ethical & Independent Review Services (E&I Review). The self-reported phenotype was derived from survey questions. Cases were defined as those individuals that reported having cataract whereas controls were defined as individuals that reported not having cataract. Individuals that preferred not to/did not answer the cataract questions were excluded from the analysis. In 23andMe replication analysis, a maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments IBD. When selecting individuals for case/control phenotype analyses, the selection process is designed to maximize case sample size by preferentially retaining cases over controls. Specifically, if both an individual case and an individual control are found to be related, then the case is retained in the analysis. Variant QC is applied independently to genotyped and imputed GWAS results. The SNPs failing QC are flagged based on multiple criteria, such as Hardy-Weinberg P-value, call

rate, imputation R-square and test statistics of batch effects. Analyses were carried out through logistic regression assuming an additive model for allele effects and adjusting for age, sex, indicator variables to represent the genotyping platforms and the first five genotype principal components.

Variants prioritization. To prioritize variants within the 54 identified genomic regions for follow-up functional evaluation, a Bayesian approach (CAVIARBF)[17] was used, which is available publicly at https://bitbucket.org/Wenan/caviarbf. Each variant's capacity to explain the identified signal within a 2 Mb window (±1.0 Mb with respect to the original top variant) was computed for each identified genomic region. Then, the smallest set of variants that included the causal variant with 95% probability (95% credible set) was derived. Out of the 1,359 total variants, 43 variants had > 20% probability of being causal.

VEGAS2 prioritization. To prioritize genes and biological pathways, we conducted a gene-based and pathways analyses using the Versatile Gene-based Association Study - 2 version 2 (VEGAS2v02) web platform[18]. We first performed a gene-based association analysis on the combined (GERA+UKB) meta-analysis results using the default '-top 100' test that uses all (100%) variants assigned to a gene to compute gene-based *P*-value. Gene-based analyses were conducted on each of the individual ethnic groups (European-specific samples (GERA and UKB individuals), GERA Hispanic/Latinos, East Asian-specific samples (GERA and UKB individuals), UKB South Asians, and GERA African Americans) using the appropriate reference panel: 1000 Genomes phase 3 European population, 1000 Genomes phase 3 American population, 1000 Genomes phase 3 East and South Asian populations, and 1000 Genomes phase 3 African population, respectively. We then meta-analyzed the 5 ethnic groups gene-based results using Fisher's method for combining the *P*-values. As 22,673 genes were tested, the P-value adjusted for Bonferroni correction was set as $P<2.21 \times 10^{-6}$ (0.05/22,673).

Second, we performed pathways analyses based on VEGAS2 gene-based P-values. We tested enrichment of the genes defined by VEGAS2 in 9,732 pathways or gene-sets (with 17,701 unique genes) derived from the Biosystem's database (https://vegas2.qimrberghofer.edu.au/biosystems20160324.vegas2pathSYM). We adopted the resampling approach to perform pathway analyses using VEGAS2 derived gene-based P-values considering the default '−10 kbloc' parameter as previously decribed[19]. We then meta-analyzed the 5 ethnic groups gene-based results using Fisher's method for combining the *P*-values. As 9,732 pathways or gene-sets from the Biosystem's database were tested, the P-value adjusted for Bonferroni correction was set as $P<5.14x10^{-6}$ (0.05/9,732).

iSyTE analyses for lens gene expression. The iSyTE database was used to analyze mouse orthologs of the human candidate genes in the 54 loci linked to cataract. iSyTE contains genome-wide transcript expression information on mouse lens obtained from microarrays and RNA-sequencing (RNA-seq) studies[20, 21]. The Affymetrix 430 2.0 platform (GeneChip Mouse Genome 430 2.0 Array and/or 430A 2.0 Array) data used in this analysis was obtained on mouse whole lens tissue at embryonic day (E) stages E10.5, E11.5, E12.5, E16.5, E17.5, E19.5, as well as postnatal (P) day stages P0, P2, and P56, in addition to isolated lens epithelium at P28. The Illumina platform (BeadChip MouseWG-6 v2.0 Expression arrays) data used in this analysis was obtained on mouse whole lens tissue at P4, P8, P12, P20, P30, P42, P52, and P60. Because previously we have shown that lens-enriched expression of a candidate gene can be used as indicative of its potential function in the lens[20, 22], we also examined the lens-enrichment of the candidate genes. This was evaluated as elevated expression in the lens compared to that in mouse whole embryonic body (WB)-based on a previously described WB-*in silico* subtraction approach[20, 22, 23]. In brief, microarray files were imported in the R statistical environment (http://www.r-project.org), and processed using relevant packages implemented in Bioconductor (https://www.bioconductor.org). Probe sets were further processed to derive present/absent calls

and further by *limma* to collapse into genes, based on significant *p*-values and highest median expression. Comparative analysis was performed in *limma* to identify differential expression of genes in the lens datasets compared to WB datasets. Expression of candidate genes was also examined in RNA-seq data from wild-type mouse whole lenses at stages E10.5, E12.5, E14.5 and E16.5 obtained in a previous study[21].

Expression analyses in specific gene-perturbation mouse models of lens defects/cataract. The iSyTE database was also used to examine expression of mouse orthologs of the candidate genes in the context of nine different gene perturbation conditions in transgenic, mutant, or targeted knockout mouse models that exhibit lens defects and/or cataract. The following mouse lens gene expression microarray data were analyzed: *Brg1* dominant negative dnBrg1 transgenic mice at E15.5 (GSE22322), *E2f1:E2f2:E2f3* conditional lens-specific triple targeted knockout mice at E17.5 and P0 (GSE16533), *Foxe3 Cryaa*-promoter-driven lens over-expression transgenic mice at P2 (GSE9711), *Hsf4* germline targeted knockout mice at P0 (GSE22362), *Klf4* conditional lens-specific targeted knockout mice at E16.5 and P56 (GSE47694), *Mafg-/-:Mafk+/-* compound germline targeted knockout mice at P60 (GSE65500), *Notch2* conditional lens-specific targeted knockout mice at E19.5 (GSE31643), *Pax6* germline heterozygous targeted knockout mice at P0 (GSE13244), *Tdrd7* germline null (*Tdrd7*^{Grm5}) mice at P30 (GSE25776), *Sparc* germline targeted knockout mice (isolated lens epithelium) at P28 (GSE13402). Candidate genes were analyzed for significant differential expression in the lens (p-value ≤ 0.05) in one or more of the above gene-perturbation conditions and plotted in the graphs.

Genetic correlations. To estimate the genetic correlation of cataract with more than 700 diseases/traits, including vision disorders, from different publicly available resources/consortia, we used the LD Hub web interface[24], which performs automated LD score regression. In the LD Score regressions, we included only HapMap3 SNPs with MAF>0.01. Genetic correlations were

considered significant after Bonferroni adjustment for multiple testing (P<6.48x10^{-5} which corresponds to 0.05/772 phenotypes tested).

PheWAS analyses. PheWAS was carried out for the 54 lead SNPs in our loci of interest identified in the combined (GERA+UKB) multiethnic analysis. SNPs were queried against 776 traits ascertained for UKB participants and reported in the Roslin Gene Atlas[25], including disorders of the lens, anthropometric traits, hematologic laboratory values, ICD-10 clinical diagnoses and self-reported conditions. Among the 54 lead SNPs, 43 were available in Gene Atlas database. We reported SNPs showing genome-wide significant association with at least one trait (in addition to cataract).

Data availability. The GERA genotype data are available upon application to the KP Research Bank (https://researchbank.kaiserpermanente.org/). The combined (GERA+UKB) meta-analysis GWAS summary statistics are available from the NHGRI-EBI GWAS Catalog (https://www.ebi.ac.uk/gwas/downloads/summary-statistics). The variant-level data for the 23andMe replication dataset are fully disclosed in the manuscript. Individual-level data are not publicly available due participant confidentiality, and in accordance with the IRB-approved protocol under which the study was conducted.

REFERENCES

1. Banda Y, Kvale MN, Hoffmann TJ, Hesselson SE, Ranatunga D, Tang H, Sabatti C, Croen LA, Dispensa BP, Henderson M, et al. Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*. 2015;200:1285-95.
2. Kvale MN, Hesselson S, Hoffmann TJ, Cao Y, Chan D, Connell S, Croen LA, Dispensa BP, Eshragh J, Finn A, et al. Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics*. 2015;200:1051-60.
3. Hoffmann TJ, Kvale MN, Hesselson SE, Zhan Y, Aquino C, Cao Y, Cawley S, Chung E, Connell S, Eshragh J, et al. Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics*. 2011;98:79-89.
4. Hoffmann TJ, Zhan Y, Kvale MN, Hesselson SE, Gollub J, Iribarren C, Lu Y, Mei G, Purdy MM, Quesenberry C, et al. Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics*. 2011;98:422-30.

5. Loh PR, Danecek P, Palamara PF, Fuchsberger C, Y AR, H KF, Schoenherr S, Forer L, McCarthy S, Abecasis GR, et al. Reference-based phasing using the Haplotype Reference Consortium panel. *Nat Genet*. 2016;48:1443-1448.
6. Das S, Forer L, Schonherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M, et al. Next-generation genotype imputation service and methods. *Nat Genet*. 2016;48:1284-1287.
7. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, Kang HM, Fuchsberger C, Danecek P, Sharp K, et al. A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet*. 2016;48:1279-83.
8. Birney E, Soranzo N. Human genomics: The end of the start for population sequencing. *Nature*. 2015;526:52-3.
9. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience*. 2015;4:7.
10. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*. 2006;38:904-9.
11. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, et al. UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12:e1001779.
12. Allen NE, Sudlow C, Peakman T, Collins R, Biobank UK. UK biobank data: come and get it. *Sci Transl Med*. 2014;6:224ed4.
13. R: A Language and Environment for Statistical Computing. *The R Foundation for Statistical Computing*. 2014.
14. Boutin TS, Charteris DG, Chandra A, Campbell S, Hayward C, Campbell A, Eye UKB, Vision C, Nandakumar P, Hinds D, et al. Insights into the genetic basis of retinal detachment. *Hum Mol Genet*. 2020;29:689-702.
15. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, Replication DIG, Meta-analysis C, Madden PA, Heath AC, Martin NG, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet*. 2012;44:369-75, S1-3.
16. Durand EY, Do CB, Mountain JL, Macpherson JM. Ancestry Composition: A Novel, Efficient Pipeline for Ancestry Deconvolution. *bioRxiv*. 2014:010512.
17. Chen W, Larrabee BR, Ovsyannikova IG, Kennedy RB, Haralambieva IH, Poland GA, Schaid DJ. Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics. *Genetics*. 2015;200:719-36.
18. Mishra A, Macgregor S. VEGAS2: Software for More Flexible Gene-Based Testing. *Twin Res Hum Genet*. 2015;18:86-91.
19. Iglesias AI, Mishra A, Vitart V, Bykhovskaya Y, Hohn R, Springelkamp H, Cuellar-Partida G, Gharahkhani P, Bailey JNC, Willoughby CE, et al. Cross-ancestry genome-wide association analysis of corneal thickness strengthens link between complex and Mendelian eye diseases. *Nat Commun*. 2018;9:1864.
20. Kakrana A, Yang A, Anand D, Djordjevic D, Ramachandruni D, Singh A, Huang H, Ho JWK, Lachke SA. iSyTE 2.0: a database for expression-based gene discovery in the eye. *Nucleic Acids Res*. 2018;46:D875-D885.
21. Anand D, Kakrana A, Siddam AD, Huang H, Saadi I, Lachke SA. RNA sequencing-based transcriptomic profiles of embryonic lens development for cataract gene discovery. *Hum Genet*. 2018;137:941-954.
22. Lachke SA, Ho JW, Kryukov GV, O'Connell DJ, Aboukhalil A, Bulyk ML, Park PJ, Maas RL. iSyTE: integrated Systems Tool for Eye gene discovery. *Invest Ophthalmol Vis Sci*. 2012;53:1617-27.

23. Anand D, Agrawal S, Siddam A, Motohashi H, Yamamoto M, Lachke SA. An integrative approach to analyze microarray datasets for prioritization of genes relevant to lens biology and disease. *Genom Data*. 2015;5:223-227.
24. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, Hemani G, Tansey K, Laurin C, Early G, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*. 2017;33:272-279.
25. Canela-Xandri O, Rawlik K, Tenesa A. An atlas of genetic associations in UK Biobank. *Nat Genet*. 2018;50:1593-1599.