Rare variant analyses across multiethnic cohorts identify novel genes for refractive error

Rare variant analyses across multiethnic cohorts identify novel genes for refractive error

Anthony M. Musolf[1*], Annechien E.G Haarman[2,3*], Robert N. Luben[4,5], Jue-Sheng Ong[6], Karina Patasova[7], Rolando Hernandez Trapero[8], Joseph Marsh[8], Ishika Jain[1], Riya Jain[1], Paul Zhiping Wang[9], Deyana D. Lewis[1], Milly S. Tedja[2], Adriana I. Iglesias[2], Hengtong Li[10], Cameron S. Cowan[11], Consortium for Refractive Error and Myopia (CREAM)[†], Ginevra Biino[12], Alison P. Klein[13], Priya Duggal[14], David A. Mackey[15], Caroline Hayward[8], Toomas Haller[16], Andres Metspalu[16], Juho Wedenoja[17,18], Olavi Pärssinen[19,20], Ching-Yu Cheng[21,22], Seang-Mei Saw[23,24], Dwight Stambolian[25], Pirro G. Hysi[7], Anthony P. Khawaja[4,5], Veronique Vitart[8], Christopher J. Hammond[7], Cornelia M. van Duijn[26#], Virginie J. M. Verhoeven[2,3,27#], Caroline C. W. Klaver[2,3,11,28#], and Joan E. Bailey-Wilson[1#]

*These authors contributed equally

#These authors jointly supervised this work

†Full list and affiliations of CREAM authors are provided in the Supplement "CREAM Authors and Affiliations"

1. Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Baltimore, Maryland, USA
2. Department of Ophthalmology, Erasmus Medical Center, Rotterdam, The Netherlands
3. Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
4. MRC Epidemiology, University of Cambridge School of Clinical Medicine, Cambridge, England, UK
5. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, England, UK
6. Statistical Genetics Laboratory, Department of Genetics and Computational Biology, QIMR Berghofer Medical Research Institute, Brisbane, QLD, Australia
7. Department of Twin Research and Genetic Epidemiology, King's College London, London, England, UK
8. MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Western General Hospital, Edinburgh, Scotland, UK
9. Institute for Biomedical Sciences, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
10. Data Science Unit, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
11. Institute for Molecular and Clinical Ophthalmology Basel, Basel, Switzerland
12. Institute of Molecular Genetics, National Research Council of Italy, Pavia, Italy
13. Department of Epidemiology, Johns Hopkins University Bloomberg School of Public Health, Baltimore, Maryland, USA
14. The Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, USA
15. Centre for Ophthalmology and Visual Science, Lions Eye Institute, University of Western Australia, Perth, Western Australia, Australia

16. Estonian Genome Center, Institute of Genomics, University of Tartu, Tartu, Estonia
17. Department of Ophthalmology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland
18. Department of Public Health, University of Helsinki, Helsinki, Finland
19. Department of Ophthalmology, Central Hospital of Central Finland, Jyväskylä, Finland.
20. Gerontology Research Center, Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland.
21. Centre for Quantitative Medicine, DUKE-National University of Singapore, Singapore, Singapore
22. Ocular Epidemiology Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
23. Saw Swee Hock School of Public Health, National University Health Systems, National University of Singapore, Singapore, Singapore
24. Myopia Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore
25. Department of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, USA
26. Nuffield Department of Population Health, University of Oxford, Oxford, England, UK
27. Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands
28. Department of Ophthalmology, Radboud University Medical Centre, Nijmegen, The Netherlands

Abstract

Refractive error is a complex eye condition caused by both genetic and environmental factors. Common genetic risk factors have been identified by genome-wide association studies (GWAS), but a great part of the refractive error heritability is still missing. Some of this heritability may be explained by rare variants (minor allele frequency [MAF] ≤ 0.01.). We performed multiple gene-based association tests for rare variants on exome array data from the Consortium for Refractive Error and Myopia (CREAM). The dataset consisted of over 27,000 total subjects from five cohorts of Indo-European and Eastern Asian ethnicity. We identified 129 unique genes associated with refractive error, many of which were replicated in multiple cohorts. Our best novel candidates included the retina expressed *PDCD6IP*, the circadian rhythm gene *PER3*, and *P4HTM*, which affects eye morphology. Future work will include functional studies and validation.

Introduction

Refractive error has become a major worldwide health concern, with the prevalence of the disease, particularly myopia (nearsightedness), becoming more frequent in both the United States[1] and Europe[2] and reaching epidemic proportions in parts of East Asia[3,4]. Refractive error is caused when the optics of the eye fail to project the focal point of light on the retina, causing a blurred image. Myopia is the refractive error mostly resulting from eye elongation, which can lead to serious ocular complications like myopic macular degeneration, glaucoma and retinal detachment[5-8], and is the second most common cause of blindness[9-11].

Refractive error is a highly complex trait that is known to have both an environmental and genetic etiology. Established environmental factors include prolonged near work, education, and little outdoor exposure[12]. Genome-wide association studies (GWAS) and genetic linkage studies have identified multiple associated variants for refractive error[13-18]. The Consortium for Refractive Error and Myopia (CREAM) has reported numerous risk variants using large-scale, multiethnic datasets[19-22], explaining approximately 18% of phenotypic variance[22].

Despite estimates that 50% to 80% of refractive error variance is determined by genetic factors[23-26], much of the refractive error heritability remains unaccounted for[19,21]. Since GWAS are particularly designed to identify common variants, some of the missing heritability may lie with rare variants (minor allele frequency [MAF] ≤ 0.01), which may be highly penetrant and exert a large effect on the phenotype[27]. Gene-based association tests, such as burden-style tests[28,29], offer increased power to find rare variants not identified by GWAS.

This study performs the first large-scale rare variant analysis on refractive error using multiethnic cohorts from CREAM. We used an initial discovery dataset consisting of over 13,000 Indo-Europeans and four replication datasets consisting of European ancestry Americans, European ancestry Australians, European ancestry Britons, and Eastern Asian ancestry Singaporeans. Gene-based tests were performed on each of the five cohorts and meta-analysis was performed subsequently. Pathway analysis was conducted on genome-wide significant genes and genes were prioritized based on annotation and biologic relevance to the trait.

Methods

Cohort Details, Genotyping and Joint Recalling of Exome Array Data

Fourteen population-based CREAM cohorts that had exome chip genotypes on individuals with refractive error measurements were used in this study. These 14 cohorts were: Singapore Chinese Eye Study (SCES), Singapore Malay Eye Study (SiMES), Singapore Indian Eye Study (SINDI), Age Related Eye Study (AREDS), Rotterdam Study I (RSI), Erasmus Rucphen Family (ERF), Raine Eye Health Study (REHS) of the Raine Study, Beaver Dam Eye Study (BDES), Estonian Genome Center for the University of Tartu (EGCUT), Finnish Twin Study on Aging (FITSA), Ogliastra, Croatia-Korcula, TwinsUK, and EPIC-Norfolk. Each individual cohort is described in further detail in the Supplementary Methods. All studies were performed in accordance with the Declaration of Helsinki and approved by the institutional review boards of the participating institutions. All participants provided written informed consent. The Institutional Review Board of the National Institutes of Health (NIH) determined that the analyses of

deidentified data performed in the current study and the meta-analysis qualified as "not human subjects research" and did not require specific protocol approval. The study was performed under guidelines agreed to in Data Use Agreements between the individual participating studies and the NIH and the Erasmus Medical Center where these analyses took place.

Thirteen cohorts had been genotyped on the Illumina HumanExome-12 v 1.0 or v 1.1, or the Illumina HumanCoreExome-12 v1.0; EPIC-Norfolk was genotyped on Affymetrix UK BioBank Axiom Array. The 13 cohorts on the Illumina arrays were jointly recalled to obtain a larger sample size of rare variants (here defined as variants with a MAF ≤ 0.01), as recalling genotypes simultaneously across all samples increases the ability to call rare variants with more discrete distinction between allele calls and sensitivity for low-frequency (high-intensity) loci. All data were recalled using GenomeStudio® v2011.1 (Illumina Inc., San Diego, CA, USA) per microarray platform and PLINK[30]. We note that these exome-based genotyping arrays consist of previously validated, high confidence rare variants, reducing the likelihood that findings might be the result of artifacts or genotyping errors that might affect sequencing studies. Further, since the imputation of very rare variants is difficult, only genotyped rare variants were used in this study; there were no imputed variants.

Combination of Cohorts for Mega-analysis

To increase power on rare variants, we sought to combine as many cohorts as possible into a mega-analysis. We thus performed principal components analysis (PCA) on all our cohorts after pruning the datasets for linkage disequilibrium using the pcair, part of the R package GENESIS. Pcair is designed to perform PCA in samples with cryptic relatedness and provides accurate ancestry inference that is not confounded by family structure[31]. For reference, we included individuals from all 11 HapMap reference panels in the PCA.

PCA showed two major groupings based on known ethnicity. The first consisted of the Han Chinese SCES and Malaysian SiMES cohorts, which were combined into the Eastern Asian combined cohort (EACC); we realize that technically the Malaysian population are Southeast Asians, but for simplicity will refer to this cohort as Eastern Asian. The second dataset consisted of the eight European cohorts (RSI, Croatia-Korcula, FITSA, EGCUT, TwinsUK, ERF, AREDS, and Ogliastra) and the one Indian cohort (SINDI). These cohorts were combined into the Indo-European combined cohort (IECC).

Analysis was performed on five discrete cohorts – IECC, EACC, EPIC-Norfolk, BDES, and REHS. The IECC analysis was performed in the Netherlands, while the EACC was performed in the United States as well as in the Netherlands. The BDES, EPIC-Norfolk, and REHS analyses were performed in their countries of origin (the United States, the United Kingdom, and Australia, respectively) as was legally required; these studies served on a per study basis as replication cohorts. A breakdown of all cohorts and the combined cohort with which they are grouped is provided in Supplementary Table 1.

Statistics and Reproducibility

Quality Control

For the combined cohorts, all raw cohort data were merged into a single file. All five cohorts then underwent identical quality control using PLINK[30]. Any individual not genotyped at 99% of all variants was removed and any variant not genotyped at 99% was also removed. Variants with a HWE p-value less

than a Bonferroni-corrected p-value (defined as 0.05 / total number of variants in the dataset) were also excluded. We also checked for batch effects and calculated the identity-by-descent (IBD) value of all individuals in the cohort, removing duplicates and twins. Many of the datasets exhibited cryptic relatedness amongst subjects (especially the Ogliastra study, which collected on the Italian island of Sardinia). Related individuals were not removed from the cohorts, as our analysis methods corrected for relatedness.

Final Sample Sizes

After QC, IECC had 13,037 individuals with 150,619 variants, EACC had 4,867 individuals with 98,750 variants, BDES had 1,740 individuals with 105,671 variants, REHS had 1,020 individuals with 92,313 variants, and EPIC-NORFOLK had 6,282 individuals with 637,160 variants.

Refractive Error Phenotype

Refractive error was defined as the quantitative phenotype spherical equivalent (SER), measured in diopters (D). Refractive error measurements in both eyes were taken from all participants and SER was calculated by adding the spherical refractive error + half the cylindrical refractive error in each eye, then taking the mean of both eyes. Individuals who had undergone procedures that could alter refraction, e.g., cataract surgery, laser refractive error procedures, retinal detachment surgery, and other ophthalmic conditions that may influence refraction were excluded from these analyses. The average spherical equivalents and standard deviations of each cohort are provided in Supplementary Table 1.

Gene-based Analysis using EMMAX-VT and EMMAX-CMC

Gene-based analysis was performed using a gene-based version of EMMAX.[32,33] EMMAX uses a kinship matrix to correct for population stratification and cryptic relatedness, which are present in these cohorts. EMMAX has been modified to perform gene-based burden-style tests, including the variable threshold (VT)[29] and the combined multivariate and collapsing (CMC)[28] methods through the software EPACTS (https://genome.sph.umich.edu/wiki/EPACTS), which we will term EMMAX-VT and EMMAX-CMC, respectively.[34].

We analyzed all five cohorts with EMMAX-VT and EMMAX-CMC using a maximum MAF = 0.01. We only included variants that were in an exon of a gene (as defined by RefSeq), including both nonsynonymous and synonymous variants. Common variants (MAF > 0.01) and variants with a MAF ≤ 0.01 that mapped to an intergenic region were excluded from the analysis. Any gene with a minor allele count (MAC) of less than three for the cohort was dropped from the analysis.

Initial analyses were performed without any covariates. We performed two follow-up analyses using age, sex, and education level (low, intermediate, and high). One covariate analysis included all three covariates, while the second used age and sex only (education level removed). We note that the inclusion of covariates resulted in no significant difference between significant genes; for brevity we only discuss the results without covariates. In addition, the Ogliastra cohort did not have data on age and education, thus approximately 3,000 individuals were removed from the IECC covariate analyses. Hence, the covariate analyses are underpowered with respect to the analyses without covariates

Gene-based Analysis using ACAT

The Aggregated Cauchy Association Test (ACAT)[35] is a novel method that allows individual p-values to be combined into a gene-based p-value that is particularly useful for rare variants. To take advantage of this method, we analyzed all variants with a MAF ≤ 0.01 (with a minimum allele count of 3) using the original, single variant-based version of EMMAX.[32,33] We then combined the EMMAX p-values for each gene using the ACAT package implemented through R. Only nonsynonymous and synonymous exonic variants were included in the analysis.

Meta-Analysis and Replication

The burden-style tests that created a single p-value for a gene precluded the use of popular meta-analysis programs such as METAL, which require the input of reference and alternative alleles. Instead the gene-based p-values from the EMMAX-VT, EMMAX-CMC and ACAT were combined across studies using the classic method described by Fisher[36]. Fisher's method was implemented through the R package metap[37]. We defined genome-wide significant as 1×10^{-5}, based on the standard for gene-based studies. Replication was defined as a having a P ≤ 0.05 in one cohort after being found to be genome-wide significant in one of the other four cohorts. We note that this replication value is liberal and may lead to an inflation in false positives. However, as this is a discovery analysis, we were willing to allow some extra false positives in order to capture as many true positives as possible. A more stringent replication p-value of 3.9e-04 was also used to adjust for 129 attempted replications and these more stringently replicated genes were also reported.

We performed two separate meta-analyses. The first combined all five cohorts (IECC, EACC, BDES, EPIC-Norfolk, and REHS), which will be referred to as the multiethnic meta-analysis. The second combined the four ethnically Indo-European cohorts (IECC, BDES, REHS, and EPIC-Norfolk), which will be referred to as the Indo-European meta-analysis. The Indo-European meta-analysis was designed to identify any genes that might be significant in Indo-European-derived individuals but not significant in Eastern Asians; thus, we also report the Eastern Asian analyses p-values.

To investigate whether signals identified by the rare variant analysis were being partially driven by common variants, we calculated polygenic risk scores (PRS) for all cohorts using common variants identified in previous GWAS[22]. PRS were calculated for each subject using PLINK (Supplementary Table 2). All rare variant analyses were then repeated using the PRS values for each subject as a covariate. We compared the explained variance (R^2) of our top individual genes between the analysis with and without including PRS (Supplementary Table 3-4).

Independent replication of the genome-wide significant genes was performed in the UK Biobank (UKBB) via extraction of all rare variants comprising the genome-wide significant genes and repeating the same analyses.

Pathway and Expression Analysis

All genome-wide significant genes in the four meta-analyses and the EACC analyses were analyzed using Ingenuity Pathway Analysis (IPA) (QIAGEN Inc., https://digitalinsights.qiagen.com/products-overview/discovery-insights-portfolio/analysis-and-visualization/qiagen-ipa/)[38]. We performed various analyses through IPA, including canonical pathway analysis (identifying which genes are in known pathways), upstream regulator analysis (which identifies genes, RNAs, and proteins that regulate the genes in the dataset), and causal network analysis (which expands the pathway analysis to include the upstream regulators in the pathway analysis). IPA also identified disease phenotypes, cellular/molecular

functions, and physiological networks associated with the genes in the dataset. Additional pathway and expression analysis were also performed with Functional Mapping and Annotation of GWAS[39,40] (FUMA), which provided tissue-enrichment information from GTEx and gene-group information from MsigDB. We repeated the IPA and FUMA analyses for our top prioritized genes from the schema proposed below.

Gene Prioritization based on Biological Function

To prioritize genes according to biological background, we evaluated genes following a modified schedule proposed by Fritsche et al.[41] and further adapted by Tedja et al.[21] Genes were ranked based on points equally assigned for the presence of replication, expression and biological plausibility. Evidence for ocular expression was based on single-cell expression data from adult human retina and developed organoids[42]. Biological plausibility was based on the presence of an ocular phenotype in OMIM and/or DisGeNET[43] as well as an ocular phenotype in a knock-out mouse model of this gene (Mouse Genome Informatics and International Mouse Phenotyping Consortium databases). The prioritization score ranged from zero to seven. In addition, we performed a look-up of the top-genes to screen for drugs that had these genes as target using SuperTarget[44], PharmGkb,[45] STITCH v5.0[46] and DrugBank v5.0.[47]

Variant Annotation for Potential Causal Variants

We performed annotation to identify potential causal variants within the significant genes. Therefore, we annotated all exonic variants from genome-wide significant genes using wANNOVAR[48-50], which collates functional predictions from popular prediction algorithms like SIFT[51], PolyPhen2[52], MutationTaster[53], CADD[54], and FATHMM[55]. We initially looked at the top-ranked genes in the prioritization approach described above, giving preference to variants that appeared to either be driving the gene-based association analysis or variants that the five annotation algorithms agreed upon as being damaging. We further expanded this approach to all significant genes identified in the meta-analyses.

Structural Analysis of Variants

We also performed structural analysis of all coding variants within our top prioritized genes, as well as all mutations predicted to be deleterious in all genome-wide significant genes. Crystal structures were obtained from the Protein Data Bank[56]; when crystal structures were not available, homology models were used for visualization and energy calculations. We used both FoldX RepairPDB and Position Scan [57] to predict differences in free energy between the wildtype and mutant proteins (ΔΔG, measured in kcal/mol). ChimeraX[58] was used to visualize affected proteins. We also incorporated prior information from publicly available databases (OMIM, PFam, ClinVar, gnomAD, UniProt, RCSB PDB) and predicted functional effects (Missense3D[59]).

Results

Overview of all Analyses

Across the three (i.e., VT, CMC and ACAT) multiethnic meta-analyses, the three Indo-European meta-analyses and the three EACC analyses, we identified a total of 129 unique genes that were significantly associated with the refractive error phenotype (Supplementary Tables 3-5). We found no statistically significant difference in p-value or the number of unique genome-wide significant genes when adding the PRS as covariates.

Multiethnic Meta-analyses

Forty-three genome-wide significant genes were found using EMMAX-VT (Figure 1A), 11 genome-wide significant genes using the EMMAX-CMC (Figure 1B), and 28 genome-wide significant genes using ACAT (Figure 1C).

Sixty-eight unique genes were identified across the three tests (Figure 2). Four genes were significant across all three tests - *GDF15* (19p13.11), *PDCD6IP* (3p22.3), *RRM2* (2p25.1), and *ST6GALNAC5* (1p31.1). *GDF15* (19p13.11) was one of the top two significant genes in all three approaches (EMMAX-VT P = 5.12×10^{-9}, EMMAX-CMC P = 1.12×10^{-9}, ACAT P = 1.95×10^{-9}). *GDF15*, *PDCD6IP*, and *RRM2* all replicated in at least one cohort; *ST6GALNAC5* only appeared in IECC and thus could not be replicated.

Overall, using a replication p-value of 0.05, 25 genes were replicated using the EMMAX-VT approach: 11 in the ACAT approach and 4 in the EMMAX-CMC approach. Three genes — *HCAR1*, *CCDC9*, and *NINJ2* — were replicated in more than one replication cohort, all in the EMMAX-VT approach. *MRPS27* in EMMAX-VT (REHS and EPIC-Norfolk) and *GDF15* in ACAT (IECC and REHS) had genome-wide significant p-values in two cohorts. If we use the more stringent replication threshold of 3.87×10^{-4}, then replications are observed for *GDF15* (VT, CMC, ACAT) and *MRPS27* (VT) with *PDCD6IP* (VT), *NDC80* (VT) and *LOXHD1* (ACAT) all having replication p-values very close to these thresholds. The list of all genome-wide significant genes for each test can be found in Supplementary Tables 6-8, while the full results of all p-values can be found in Supplementary Tables 9-11. Note that beta is provided for the individual CMC analyses and a direction for the individual VT analyses, as VT does not output a beta.

Indo-European Meta-analyses

As it is possible that Eastern Asians differ in genetic risk factor profile from Indo-Europeans, we performed meta-analyses on the four Indo-European ancestry cohorts. Forty-nine genes were genome-wide significant in the EMMAX-VT approach (Figure 3A), 13 genes in the EMMAX-CMC approach (Figure 3B), and 29 genes in the ACAT approach (Figure 3C). Four genes overlapped between all three tests — *GDF15*, *PIK3CA*, *RRM2*, and *ST6GALNAC5* (Figure 4). The signal at *PIK3CA* was unique to the Indo-European meta-analysis. *GDF15* and *RRM2* were both replicated in one cohort, while *PIK3CA* and *ST6GALNAC5* only appeared in IECC.

Overall, 24 genes were replicated at p=0.05 in EMMAX-VT, 8 genes in ACAT, and 4 genes in EMMAX-CMC. *NINJ2* in the EMMAX-VT and *STON1* and *SND1* in EMMAX-CMC were replicated in multiple cohorts. The list of all genome-wide significant genes for each test can be found in Supplementary Tables 12-14, while the full results of all p-values can be found in Supplementary Tables 15-17.

EACC Analysis

We also report the standalone results of EACC analysis. Thirty-one genome-wide significant genes were found in EACC using EMMAX-VT (Figure 5A), 5 genome-wide significant genes using EMMAX-CMC (Figure 5B), and 22 genome-wide significant genes using ACAT (Figure 5C). *GSTM5* (1p13.3) and *WEE1* (11p15.4) overlapped in all three tests (Figure 6). *SERTAD3* (chromosome 19) and *ZNF25* (chromosome 10) were genome-wide significant and only appeared in EACC, i.e., rare variants in these two genes did not exist in the other cohorts. 51 unique genome-wide significant genes were identified, 39 novel to the EACC analyses. The list of all genome-wide significant genes for each test can be found in Supplementary Tables 18-20.

Cohort Unique Genes

In addition to the two genes in the EACC EMMAX-VT analysis, there were 6 significantly associated genes that only had rare variants within a single cohort; no other rare variants existed in the other cohorts for these genes. *EDN3* and *CHMP1B* in the IECC EMMAX-VT analysis and *PRLH* in the IECC ACAT analysis. *KLF1* appeared only in the EPIC-Norfolk cohort, in both the EMMAX-VT and EMMAX-CMC analyses. The list of cohort unique genes appears in Supplementary Table 21.

Independent Replication in UK Biobank

We extracted the variants from the 129 significant unique genes and performed replication analyses in the UK Biobank. There were 7 genes with a $P < 0.05$ in EMMAX-CMC and 9 genes with a $P < 0.05$ in EMMAX-VT (Supplementary Table 22). *P4HTM*, *CCDC170*, and *CPB1* were found in both analyses. *STON1* was also replicated in the UK Biobank analyses; this gene had a significant meta-analysis p-value in the EMMAX-CMC analysis. Interestingly, the p-value in all cohorts was < 0.053.

Pathway and Expression Analysis on all Significant Genes

We performed IPA pathway analysis on the 129 unique genes. While this did not result in any genome-wide significant canonical pathways, the upstream regulators analysis identified over 172 associated transcription factors. The two highest were the cytokine *CSF2*, which is known to regulate neuroglia after retinal injuries[60], and the Transcription factor (TF) *MEF2C*, which is known to be expressed in the retina and controls photoreceptor gene expression[61] (Supplementary Table 23). The fourth ranked p-value was the Raf kinases, which are known to be involved in retinal development[62] and cell survival[63]; the fifth ranked p-value was *TBX5*, which is expressed in the retina and involved in eye morphogenesis[64,65]. Causal network analysis identified 288 associated pathways (Supplementary Table 24), including the *TRPC5* pathway, which regulates axonal outgrowth in developing ganglion cells[66].

The top overall associated physiological system functions were organ morphology, organismal development and embryonic development, while the top molecular/cellular functions were cell cycle and cellular assembly/organization. Cancer and organismal injuries/abnormalities were the top overall associated phenotypes (Supplementary Table 25). Six genes were associated with ophthalmic phenotypes: *CHST6*, *GCNT2*, *P4HTHM*, *USH2A*, *GRHL2*, and *MAPT*.

FUMA analysis found that the top enriched tissues were heart, brain, muscle, and adipose tissue (Supplementary Figure 1A). The top functional categories were cytoskeleton organization, cell cycle processes, mitotic nuclear division, and organelle organization (Supplementary Figure 1B.

Biological Plausibility and Prioritization of Genes

Of the 129 genome-wide significant genes from the six meta-analyses, 27.9% (36/129) have a known expression in human ocular tissue. 51.2% (66/129) of these genes showed evidence for a human ocular phenotype.

Seven genes had a biological plausibility score higher than 3 — *PER3* (internally replicated, expressed in ocular tissue and associated ocular phenotype, i.e., score of 5) and *PDCD6IP, MAPT, CHST6, GRHL2, USH2A,* and *P4HTM* (all with a score of 4). An additional 11 genes had a score of 3 — *GDF15*, *RRM2*, *HSPH1*, *TPR*, *KRT81*, *SPHK1, GSTM5, THSD7A, WEE1,* and *BUB1B* (Figure 7). Detailed background for the prioritization of the genes can be found in Supplementary Tables 26A-F. Table 1 provides the p-values and effect sizes (when available) for each gene. Supplementary Table 27 provides the average SER for minor allele carriers versus non-carriers for each variant in these prioritized genes; please note that this

table uses the single variant results which is restricted to MAC >= 3; some variants with MAC < 3 were used in the gene-based tests but will not be present in Supplementary Table 27. P-values and betas for each of the individual rare variants are also provided. In general, *PDCD6IP*, *MAPT*, and *USH2A* variants had the most negative average SER for carriers of the given rare variant (cases in the table), although genes *GRL2*, *CHST6*, *PDCD6IP*, and *USH2A* all had variants with high positive SER for rare variant carriers as well. Betas tended to conform with difference between rare variant carrier SER and SER in noncarriers (controls in the table), with many of the top variants having large betas. Perhaps the most interesting fact with respect to the betas is that most of the single variant betas tended to be positive and led to increased myopization (negative SER). However, there were still negative betas for some variants with more hyperopic mean SERs in carriers versus non-carriers, particularly in the IECC and the genes *PDCD6IP* and *USH2A* across cohorts.

The highest overall biological plausibility score belonged to the circadian rhythm gene *PER3* (1p36). It was genome-wide significant in both the all cohorts ACAT and Indo-European only meta-analyses (P = 1.08×10^{-6} and 1.15×10^{-6}, respectively); it was genome-wide significant in REHS and replicated in IECC. CMC betas were 0.1666, 0.1574, -0.1976, -0.1102, -0.518 for IECC, EACC, BDES, EPIC-Norfolk, and REHS respectively; none of the CMC p-values were significant, however (Supplementary Table 28). Circadian rhythm genes have been shown to be associated with refractive error[22] and *PER3* is located near the site of a known myopia locus (MYP14) at which the causal gene has not been identified[67-69]. PER3 was expressed in ON and OFF bipolar cells. Defects in this gene are associated with familial advanced sleep phase syndrome (OMIM 616882) and may contribute to other circadian phenotypes by altering the sensitivity to light[70]. In defocus experiments in chicks using -15D lenses, PER3 expression decreased by -1.26-fold in the retina[71]. Further chick defocusing experiments, showed that PER3 expression in the retina varies under altered visual conditions[72]. Recently published data from the Raine Study suggest that falling asleep later was associated with a higher risk of myopia progression[73].

Five genes had a score just below *PER3*, including the apoptosis gene *PDCD6IP* (3p22.3). This gene was found to be genome-wide significant in all-cohorts meta-analyses using all three tests (P = 1.07×10^{-7}, 1.45×10^{-7}, and 4.88×10^{-6}, respectively). Further *PDCD6IP* had a P of < 0.006 in both the EACC and IECC cohorts and did not appear in the other cohorts. Both betas in the CMC test were negative and with a large effect size for IECC (beta = -2.5) (Supplementary Table 28). Most rare variants in this gene in the EACC and IECC samples result in mean SER's in carriers that were smaller (more negative) than in non-carriers, which meant that the CMC test would be powerful to detect this association (Supplementary Table 27). It is particularly interesting because *PDCD6IP* has two low single variant p-values in both IECC and EACC (0.00556 and 0.00548, respectively) and there are no rare variants in this gene in any of the other cohorts. *PDCD6IP* is expressed in ganglion cells of peripheral retina and plays a role in programmed cell death in uveal melanoma[74] and may play a role in cornea lymphangiogenesis and vascular responses.[75]

MAPT (17q21.32) encodes tau proteins responsible for stabilizing microtubules; it was found to be genome-wide significant in the all cohorts EMMAX-VT analysis (P = 8.57×10^{-7}). It was genome-wide significant in REHS and replicated in EPIC-Norfolk. Betas from the CMC test were -0.4342, 0.3137, -0.4965, -0.171, and -0.8015 for IECC, EACC, BDES, EPIC-Norfolk, and REHS respectively (Supplementary Table 28). Again, none of the CMC test p-values were significant. Abnormal MAPT was present in human glaucoma patients with uncontrolled intraocular pressure[76] Cowan et al. showed that *MAPT* was expressed in several cell types in both the peripheral and foveal human retina: horizontal cells, rod

bipolar cells, ON and OFF bipolar cells GLY and GABA amacrine cells and ganglion cells[42]. A knock-out mouse model showed decreased total retina thickness.

CHST6 (16q23.1) was genome-wide significant in both the all cohorts and Indo-European only EMMAX-VT meta-analyses ($P = 8.99 \times 10^{-7}$ and 2.42×10^{-7}, respectively). The gene was genome-wide significant in IECC and replicated in BDES; it was also nearly replicated in EPIC-Norfolk. Though the CMC p-values were not significant, the beta for BDES was particularly large (0.95) (Supplementary Table 28). *CHST6* plays a role in maintaining corneal transparency. Mutations in this gene may result in macular corneal dystrophy (OMIM 217800), which is characterized by bilateral, progressive corneal opacification and a reduction of corneal sensitivity.[77] The mouse phenotype of a knock-out model corresponded to that of human, i.e. abnormal cornea morphology and decreased corneal (stroma) thickness. Since our reference expression database did not contain any corneal tissue, we couldn't score this category.

The transcription factor *GRHL2* (8q22.3) was genome-wide significant in the all cohorts EMMAX-VT meta-analysis ($P = 1.42 \times 10^{-6}$). It was genome-wide significant in REHS and replicated in IECC. Though the p-values for EMMAX-CMC were not significant, REHS had a large beta value of 0.87 (Supplementary Table 28). Mutations in *GRHL2* may lead to posterior polymorphous corneal dystrophy[78] (OMIM 618031), characterized by a variable phenotype ranging from an irregular posterior corneal surface with occasional opacities, corneal edema, reduced visual acuity, secondary glaucoma, and corectopia.

The transmembrane prolyl hydroxylase *P4HTM* (3p21.31) was only genome-wide significant in EACC using EMMAX-VT ($P = 1.00 \times 10^{-7}$). However, this gene was replicated independently in the UKBB analysis. Betas for the non-significant EMMAX-CMC test were -0.1769, -2.025, 0.6106, -0.1632, -0.1177 for IECC, EACC, BDES, EPIC-Norfolk, and REHS respectively (Supplementary Table 28). *P4HTM* has been shown to be expressed in different ocular cells (including horizontal cells and bipolar cells). It is associated with HIDEA, a severe autosomal recessive disorder that is characterized by multiple symptoms, including eye abnormalities[79] (OMIM 618493) and knock-out mice models have shown abnormal eye morphology[80].

The membrane gene *USH2A* (1q41) was genome-wide significant in the EACC ACAT analysis ($P = 7.55 \times 10^{-9}$). The EMMAX-CMC tests were not significant which is reflected in the betas which were all quite small except for 0.82 in the BDES sample (Supplementary Table 28). This reflects the wide variation in effect on SER exhibited by different rare variants in this gene, with some individual variants leading to much more myopic mean SER's in carriers compared to non-carriers while other rare variants led to more hyperopic mean SERs in carriers compared to non-carriers. (Supplementary Table 27). *USH2A* is well known to cause both Usher syndrome, which includes retinitis pigmentosa (RP) and mild to moderate hearing loss, as well as RP without hearing loss[81]. It is known to be expressed in the retina[82] and has been recently shown to be associated with high myopia[83]

Pathway and Expression Analysis on Top Prioritized Genes

We ran the IPA and FUMA analyses on the seven top prioritized genes. IPA did not identify any canonical pathways as significant; the only pathway shared across the genes was the 14-3-3-mediated signaling pathway (*MAPT* and *PDCD6IP*). The 14-3-3 proteins are a diverse group of signaling proteins.

Upstream regulator analysis found several transcription regulators of at least two genes include *NKX2-1* (*GRHL2* and *MAPT*), *PSEN1* (*MAPT* and *PER3*), and *SIRT1* (*MAPT* and *PDCD6IP*) (Supplementary Table 29). In the causal network analysis, the master regulator with the highest p-value covering multiple genes

was the cytokine macrophage migration inhibitory factor (*MIF*) (Supplementary Table 29), which covered five genes. Interestingly, *MIF* is an essential factor in the development of zebrafish eyes[84] and has been found to be a potential regulator of diabetic retinopathy[85]. *MIF* inhibitors may also be protective to photoreceptors[86]. The top functional analysis for disease result was hereditary eye disease (Supplementary Table 31). FUMA showed the top tissue expression occurred in the small intestinal terminal ileum, skeletal muscle, and the brain cortex; the latter being probably the best proxy for eye tissue (Supplementary Figure 2A). A heat map of the expression of the seven genes across all GTEx tissues is given in Supplementary Figure 2B).

Potential Causal Variants in the Prioritized Genes

We used annotation from wANNOVAR to identify potential causal variants within the top genes identified by the prioritization method (Table 2). For the two prioritized genes that were significant in the ACAT analyses, we were able to look at single variant p-values in addition to annotation to determine potential causal variants. There were three good candidate variants in *PDCD6IP*, which was genome-wide significant in IECC and replicated in EACC. rs199990824 (3:3879764) appeared in the EACC only, was predicted to be damaging by SIFT and MutationTaster, and had a CADD score of 26. The minor allele of rs199990824 appeared in 37 carriers (all heterozygotes) with an average SER of -2.04 D (SD = 3.29) compared to the non-carrier average of -0.44 D (SD = 2.27) and the overall cohort average of -0.45 D (SD=2.28); the single variant P was 0.000183. In the IECC, the best potential causal variant was rs62620697 (3: 33905532), which was predicted damaging by MutationTaster, had a CADD score of 23.8, and had a low single variant p-value of 0.002632. Carriers (N=9) of rs62620697 had an average SER of -2.17D (SD = 6.87) compared to that of non-carriers with an average SER of 0.20 (SD = 2.27). rs145293758 also had a low p-value (0.000311) but was not predicted damaging.

Potential candidate variants were also identified in *PER3*, which was genome-wide significant in REHS and replicated in IECC. The REHS signal was primarily driven by two variants - rs147327372 and rs144178755, which had single variant p-values of 1.72×10^{-8} and 0.004953, respectively. However, neither variant was predicted to be damaging by the prediction algorithms nor appeared in the other European cohorts and were not significant individually, although rs147327372 did have a p-value of 0.046 in EPIC-Norfolk in the single variant tests.

The signals in the other four genes, identified primarily by the two burden-style tests, were driven by a cumulative effect of several variants. In this case, we relied primarily on annotation and reported variants that were generally agreed upon by multiple prediction programs. Five good candidate variants were located in *MAPT*: rs139796158 (17:44055786) , rs76375268 (17:44060807), rs63750072 (17:44060859), rs143956882 (17:44067341) and rs63750191 (17:44101481). All these variants were nonsynonymous variants and predicted damaging by three of the four databases (except for rs76375268, which was predicted damaging by two). rs139796158, rs143956882, and rs63750191 all had CADD scores > 26. In *CHST6*, the best candidate variant was the missense variant rs140699573 (16:75512734). It was predicted damaging by SIFT, PolyPhen2, MutationTaster, and FATHMM and has a CADD score of 27.4. In *GRHL2*, the best candidate variant was rs142411476 (8:102570910). It was predicted damaging by two databases and had a CADD score of 22. In *P4HTM*, two variants of interest were identified: rs140290144 (3:49002551) and rs144279528 (3:49043292). These variants were predicted damaging by MutationTaster and had CADD scores of 22.1 and 27.3, respectively. Finally, in

USH2A, three variants (rs554957414 (1:216138793), rs148135241 (1:216373416), and rs201527662 (1:216419934) were all predicted damaging by the five prediction algorithms and had CADD scores above 22.

Structural Analysis of Prioritized Candidate Proteins

In addition to the annotation, we also performed protein structural modeling of all coding variants within the prioritized genes (98 variants across 6 genes/proteins) and calculated free energy difference (ΔΔG) between wildtype and mutant proteins (Supplementary Table 32); positive ΔΔG indicates a shift from a more stable to a less stable isoform. More detailed information on the structural analysis can be found in the Supplemental Methods.

In *PDCD6IP*, both rs145293758 (3:33905587) and rs200697599 (3: 33840234) were predicted to be highly destabilizing to protein structure (Supplementary Figure 3A). The variant rs145293758 leads to replacement of a proline (Pro737) for an asparagine near phosphorylation sites in the protein's self-associating domain, which could disrupt phosphorylation. rs200697599 (Ile5) and rs199990824 (Asp376; 3:33879764) result in changes to the protein's BRO1 domain, which is involved in endosomal targeting. The isoleucine to serine mutation at rs200697599 could introduce a phosphorylation site at the N-terminus while the asparagine to aspartic acid mutation at rs199990824 could disrupt hydrogen bonds. Recall that both rs145293758 and rs199990824 were identified as potential causal variants for refractive error in IECC and EACC, respectively, based on their annotation, and single variant p-values (Supplementary Table 27).

For *PER3*, several variants may affect structure, including rs140974114, which results a serine (Ser751) to aspartic acid substitution at the protein's nuclear localization signal and could disrupt hydrogen bonds and rs200140283, which results in an alanine (Ala681) to glycine substitution in the CSNK1E binding domain. Further potential disruptions occur at rs139315125 (His416), which takes place in the nuclear export signal 3 and rs77418803 (Ser919), which occurs near the nuclear export signal 2. The model is provided in Supplementary Figure 3B).

Of the variants in *MAPT*, two were predicted to be destabilizing (rs76375268 at Gly213 and rs63750191 at Gln741) (Supplementary Figure 3C). Further, rs73314997 (Ser318) and rs143956882 (Ser427) are located near known pathogenic mutations for frontotemporal dementia and Pick disease of the brain, respectively.

Three variants on the luminal domain of CHST6 were found to have a mild effect on protein stability. Two of these variants (rs201349198 at Ala326 and rs140699573 at Gln331) are positioned near variants known to cause macular corneal dystrophy (MCD) near the C-terminus. This suggests the C-terminus is sensitive to mutations enabling interference with keratan sulfation, which could cause a loss of function that can lead to a milder disease phenotype such as refractive error. The model can be found in Supplementary Figure 4A.

In *GRHL*2, variants were only predicted to have a mild effect on protein structure and were not located near known pathogenic variants (Supplementary Figure 4B).

For *P4HTM*, rs140290144 is predicted to be moderately destabilizing (Supplementary Figure 4C). It substitutes a valine for a buried isoleucine (Ile227) between two calcium binding sites; potential disruption of these calcium binding sites can result in loss of function. Similarly, rs144279528 occurs in the Fe-dependent 2-OG dioxygenase domain close to an iron binding residue. Substitution of asparagine

from the wildtype aspartic acid (Asp386) could have an impact on iron binding by introducing a glycosylation (due to location on protein surface) or disruption of hydrogen bonding.

Of particular interest in the protein modeling was that of usherin (*USH2A*), the known retinitis pigmentosa gene. Five variants were predicted to be highly destabilizing, particularly rs554957414 with a ΔΔG value of 99.19 kcal/mol). Three of these variants, including rs554957414 (Pro2329), result in the loss of proline and the loss of that ring structure could cause an increase in conformational flexibility and account for such high destabilization predictions (Supplementary Figure 5). Further, a mutation at rs201527662 (Cys934) results in the replacement of cysteine with tryptophan and will disrupt a disulfide bond between two cysteines.

We also compared the ΔΔG of these five candidate variants with the ΔΔG of all *USH2A* ClinVar (n = 63) and gnomAD (n = 1870) variants using the Wilcoxon rank sum test. A significant difference between the ClinVar variants and gnomAD variants was found (P = 0.0008) and the ΔΔG values of our candidate variants was much more similar to the known pathogenic variants than the putatively benign GnomAD variants (Supplementary Figure 6).

Potential Causal Variants in Other Genome-wide Significant Genes

We also identified variants within the other 122 genome-wide significant genes that had a high potential to be damaging. This included 25 variants across the five cohorts; the results are found in Supplementary Table 33. Like our prioritized genes, we also performed protein modeling on these variants (Supplementary Table 34).

Notable findings from the structural analysis include a valine to phenylalanine substitution (Val105) that would disrupt a helix in ALG3, which has been implicated in congenital disorders of glycosylation that have ocular phenotypes[87] (Supplementary Figure 7A). We also identified multiple glycine substitutions in TNFRSF13B in areas associated with heparan sulfate – glycosaminoglycan biosynthesis; heparan sulfate has been shown to play a role in eye pathologies[88] (Supplementary Figure 7B).

Discussion

In this large scale, gene-based analysis of rare variants in refractive error, 129 associated genes were identified. Though many of the genes were associated with eye conditions or ocular development, only ten genes had previously been identified with refractive error or myopia: six with myopia including two with high myopia — USH2A and GDF15[83,89] — and ten with refractive error. Pathway analysis revealed that 59 of these genes were involved in cell cycle, organ morphology, and embryonic development and 21 of these genes had upstream regulators that were directly involved in retinal development or eye morphogenesis. Given the substantial level of missing heritability still present within the refractive error, it is likely that at least some of this heritability is explained by rare variants within these genes. The fact that the significance of these genes and the explained variance of refractive error due to these genes did not significantly change after inclusion of GRS in the analysis, suggests that these association signals are independent from the effects of known common refractive error risk variants.

This is the first large scale meta-analysis using gene-based tests for rare variants in refractive error, which was undertaken to identify rare variants that may be partially responsible for missing heritability, particularly within the CREAM data set[21]. The CREAM data set is well-suited for this type of rare variant analysis. First, we were able to combine many smaller cohorts into two mega-analyses – IECC (N =

11,505) and EACC (N = 4,867). These meta-analyses greatly boosted power to detect variants with a MAF ≤ 0.01 and allowed more rare variants to be combined into a single, gene-based marker. In addition, we had three cohorts > 1000 subjects to observe replication and perform the combined meta-analyses. Genes identified in this study were done so across a very large pool of subjects, lowering the potential for type I error.

The multiethnic composition of this dataset also allowed for observation both across and within ethnicities. We have delineated how rare variants in some genes were found only in Indo-Europeans and others in Eastern Asians, as well as some that cut across the ethnic divide. Thus, we were able to identify risk genes that might contain rare variants that affect SER within a particular population (such as *ST6GALNAC5* in IECC), or more universally, like *PDCD6IP*.

PER3, PDCD6IP, MAPT, CHST6, *P4HTM, USH2A,* and *GRHL2* are good candidate genes, all known to be associated with ocular abnormalities. *PER3* is a circadian rhythm gene; circadian rhythm is associated with refractive error[22]. *PDCD6IP* and *MAPT* are both expressed in the retina while *CHST6*, and *GRHL2* are both involved in corneal dystrophy[78,90]. *P4HTM* affects eye morphology in mice knockouts[84]; it is also notable for being replicated in the UKBB analysis. *USH2A* is expressed in the retina and is a known RP gene[81,82].

Five of these prioritized genes were found to be regulated by the cytokine *MIF*, which has been shown to regulate zebrafish eye development[84] and have protective effects for photoreceptors[86]. More work on the *MIF* network with respect to refractive error is needed. We were further able to identify potential causal variants in these prioritized genes and, using structural analysis, were even able to determine the effect on protein stability.

STON1, *C5AR1*, and *WDFY3* were all replicated in UKBB. *C5AR1* is expressed in retinal Müller cells, which are known to play a role in retinal disease[91]. *STON1* is associated with AMD[92] while *WDFY3* is associated with inherited retinal dystrophies[93]. Other potential interesting candidates include *GDF15*, which was a top significant gene across all four meta-analyses, and has been found to be significantly overexpressed in highly myopic eyes[89] and patients with vitreoretinal disorders[94] and may also be a potential molecular marker of neurodegeneration in glaucoma[95], and *MRPS27*. This gene was genome-wide significant in the meta-analysis and in two individual cohorts, REHS and EPIC-Norfolk. While *MRPS27* is not known to be associated with eye disease, a common variant in this gene was found to be genome-wide significant in the GWAS meta-analysis of refractive error conducted by Hysi et al.[22]. Other candidate genes with known links to eye disease/functions include *HCAR1* with glaucoma[96,97] and *EPB41L2* with a potential role in phototransduction[98].

One final interesting set of genes was those that were genome-wide significant within a single cohort. This implies that there may be rare risk variants unique to a certain population that are fixed in other populations. This includes *ST6GALNAC5*, which was genome-wide significant in IECC in both EMMAX-VT and ACAT ($P = 5.84 \times 10^{-7}$, 9.03×10^{-10}). This gene catalyzes the transfer of sialic acid; polysialic acid has been shown to prevent vascular damage in retina[99] and to stimulate the generation of new rods in the retinas of developing zebrafish[100]. Other interesting significant genes unique to a single cohort included *SERTAD3* in EACC, which is overexpressed in retinoblastoma[101] and *KLF1* in EPIC-Norfolk, which may be expressed in the eye[102]. We also note that gene-based analyses for refractive error had been previously performed in BDES[103]. Of the five significant genes from that analysis, two were replicated at P≤0.05 —

PTCHD2 and *CRISP3*. *PTCHD2* is located near the known myopia locus *MYP14* on 1p36.22[69,104] and *CRISP3* is expressed in the retina[103,105].

This study used multiple tests (EMMAX-VT, EMMAX-CMC and ACAT) to identify significant genes and looked at overlap to find more robust signals. By using multiple tests that differ slightly in design, we were able to cast a wider net in our search. The ACAT test was particularly useful for identifying potential causal variants within a candidate gene, as it allowed us to observe which variants had significant single variant p-values. This enabled us to zero in on potential causal variants in genes like *PDCD6IP* and *PER3*, though we note that highlighting any potential causal variants are speculative at this point. We also felt it prudent to not give more weight to the result of one test over another and instead take the largest number of unique, significant genes since this was a discovery study, though we did try to give more weight to the genes that were identified by all three tests, such as *PDCD6IP*.

We note that the three tests did not always agree, though the two burden-style tests agreed more often than ACAT. This is not surprising given the different nature of the tests. Both EMMAX-VT and EMMAX-CMC were burden-style tests that create a new, gene-based marker on which the p-value is calculated. The ACAT test was an aggregation-style test created from single variant p-values that does not create a new gene-based marker[35]. This is a critical distinction; it means that the markers analyzed in the burden-style tests and the ACAT tests are different. The ACAT analyses may have been slightly underpowered with respect to the burden-style tests, as we used a minimum allele count of three in our analyses. For EMMAX-VT and EMMAX-CMC this was calculated across all variants within a gene and for ACAT at each individual variant, which resulted in certain variants being removed from the ACAT analysis that were present in the burden style analyses. Therefore, genes present in all three analyses indicate a more robust association with refractive error.

Since this is an exome microarray study, there were still large portions of the genome that would not have been covered in this work. Thus, there are almost certainly additional rare risk variants for refractive error in these cohorts that were not genotyped in this study. The goal of this discovery study was to provide an initial starting point for further analysis; we plan whole genome sequencing on high-risk individuals identified by this study. These non-genotyped variants could explain why we did not see replication with previous refractive error GWAS findings[21,22]. Some of the genes identified in the common variant GWAS may have included rare risk variants that were specific to a particular population that was not used in this study.

Another challenge is that due to the gene-based nature of this work, it is critical to remember that the gene-based markers across the cohorts are often made up of different variants. This means that the gene-based marker for gene A in IECC might be made up of three variants, and in REHS might be made up of seven variants, two of which are shared across the two cohorts. This means that it was possible that some cohorts may have had association tests that were less significant because of inclusion of non-significant rare variants that did not appear in other cohorts.

We also note that this was an exploratory analysis to determine candidate genes, and one of our goals was to cast a wide net to capture potential candidates. Therefore, we chose a more liberal replication significance threshold, which may allow for potential type I errors but would also ensure that a good candidate gene would not be missed or because functional rare variants did not appear in that cohort.

We also note that while we did utilize eye expression data in this study, we were limited to expression from retinal tissue only. We are actively seeking expression data from additional eye tissue, particularly corneal and scleral tissue, to further prioritize these genes.

This work identified 129 genome-wide significant genes for refractive error using the gene-based rare variant approach. Most of these genes are novel for association with refractive error but many have associations with other ocular abnormalities. This is the largest gene-based study of rare variants performed on refractive error. The fact that we found over 100 significant genes shows that rare variants (MAF ≤ 0.01) do account for some of the missing refractive error heritability not identified in the common variant GWAS. We were able to prioritize seven of these genes as our best candidate genes for causality based on biological function – *PDCD6IP*, *MAPT*, *CHST6*, *GRHL2*, *USH2A, P4HTM*, and *PER3* – as well as *GDF15* and *MRPS27* based on the strength of association. Validation studies, including replication within additional cohorts, are planned to identify the best candidates for functional studies to unravel the pathophysiology of refractive error and myopia. We also plan further analysis with the conversion of our quantitative refractive error phenotype to binary phenotypes to test for association with myopia, hyperopia, and astigmatism.

Acknowledgments

The authors gratefully acknowledge Sana Wajid of the Bioinformatics Core of the University of Pennsylvania for her quality control work on these data. This work was funded in part by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health. The acknowledgments for each individual study cohort are given alphabetically by study below. APK is supported by a UKRI Future Leaders Fellowship. Molecular graphics and analyses were performed with UCSF ChimeraX, developed by the Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco, with support from National Institutes of Health R01-GM129325 and the Office of Cyber Infrastructure and Computational Biology, National Institute of Allergy and Infectious Diseases

AREDS: AREDS was supported by the National Eye Institute (grants R01EY16482, R21EY015145, and P30EY11373) and by Research to Prevent Blindness and the Ohio Lions Eye Research Foundation. AREDS was also supported by contracts from National Eye Institute/National Institutes of Health, Bethesda, MD, with additional support from Bausch & Lomb Inc, Rochester, NY. The genotyping costs were supported by the National Eye Institute (R01EY020483 to D.S.) and some of the analyses were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, USA. AREDS acknowledges Frederick Ferris, National Eye Institute, National Institutes of Health, Bethesda, MD; and the Center for Inherited Disease Research, Baltimore, MD where SNP genotyping was carried out. The investigators gratefully acknowledge the advice and guidance of Hemin Chin of the National Eye Institute.

BDES: BDES was supported by the National Eye Institute of the National Institutes of Health under award numbers EY06594 (R. Klein and B. E. K. Klein), EY10605 (B. E. K. Klein) and R01EY021531 (A.P.K and P.D.) and some of the analyses were supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health, USA.

Croatia-Korcula: The Croatia-Korcula study was funded by the Medical Research Council (UK) "QTL in health and disease" programme core grants, currently MC_UU_00007/10, as well as grants from the

Republic of Croatia Ministry of Science, Education and Sports (108-1080315-0302; 216-1080315-0302) and the Croatian Science Foundation (8875). The study acknowledge Dr. Biljana Andrijević Derk, Valentina Lacmanović Lončar, Krešimir Mandić, Antonija Mandić, Ivan Škegro, Jasna Pavičić Astaloš, Ivana Merc, Miljenka Martinović, Petra Kralj, Tamara Knežević and Katja Barać-Juretić as well as the recruitment team from the Croatian Centre for Global Health, University of Split and the Institute of Anthropological Research in Zagreb for the ophthalmological data collection; the Wellcome Trust Clinical facility (Edinburgh, United Kingdom) for Exome array genotyping.

EGCUT: EGCUT was supported by the European Union H2020 grant 692145, Est.RC grant IUT20-60 and the European Regional Development Fund, in the frame of Centre of Excellence in Genomics and Estonian Research Infrastructure's Roadmap and the University of Tartu (SP1GVARENG). This research was supported by NIH grant 5R01 DK07 57 87 -13, under subward-agreement GENFDOOO1B52751; the European Union through Horizon 2020 research and innovation programme under grant 633589 and the European Regional Development Fund (Project No. 2014-2020.4.01.16-0125). This research was also supported by the European Union through the European Regional Development Fund (Project No. 2014-2020.4.01.16-0125) and the Estonian Research Council grant PUT (PRG687) European Union H2020 grant 654248 (Corbel). EGCUT acknowledges the High Performance Computing Center of the University of Tartu.

EPIC-Norfolk: The EPIC-Norfolk study (DOI 10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 and MC-UU_12015/1) and Cancer Research UK (C864/A14136). The genetics work in the EPIC-Norfolk study was funded by the Medical Research Council (MC_PC_13048).We are grateful to all the participants who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research.

FITSA: FITSA was supported by ENGAGE (FP7-HEALTH-F4-2007, 201413);
European Union through the GENOMEUTWIN project (QLG2-CT-2002-01254); the Academy of Finland Center of Excellence in Complex Disease Genetics (213506, 129680); the Academy of Finland Ageing Programme; and the Finnish Ministry of Culture and Education and University of Jyväskylä. FITSA acknowledges the contributions of Emmi Tikkanen, Samuli Ripatti, Markku Kauppinen, Taina Rantanen and Jaakko Kaprio.

Ogliastra: The Ogliastra Study gratefully acknowledges the population of Ogliastra, Sardinia, Italy. The Ogliastra study was funded by a grant from the Italian Ministry of Education, University and Research (MIUR) n°: 5571/DSPAR/2002.

RSI, ERF: The Rotterdam Study and ERF were supported by European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant 648268), Netherlands Organisation for Scientific Research (NWO, grant 91815655 to CCWK and NWO Veni 91617076 to VJMV), Ammodo Award (to CCWK), Erasmus Medical Center and Erasmus University, Rotterdam, The Netherlands; Netherlands Organization for Health Research and Development (ZonMw); the Research Institute for Diseases in the Elderly; the Ministry of Education, Culture and Science; the Ministry for Health, Welfare and Sports; the European Commission (DG XII); the Municipality of Rotterdam; the Netherlands Genomics Initiative/NWO; Center for Medical Systems Biology of NGI; Jacoba Breen Fonds, Topcon Europe; Ada Hooghart, Corina Brussee, Riet Bernaerts-Biskop, Amal Hamimida, Patricia van Hilten, Pascal Arp, Jeanette Vergeer, Sander Bervoets. The generation and management of the Illumina

exome chip v1.0 array data for the Rotterdam Study (RS-I) was executed by the Human Genotyping Facility of the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, Rotterdam, The Netherlands. The Exome chip array data set was funded by the Genetic Laboratory of the Department of Internal Medicine, Erasmus MC, from the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Research (NWO)-sponsored Netherlands Consortium for Healthy Aging (NCHA; project nr. 050-060-810); the Netherlands Organization for Scientific Research (NWO; project number 184021007) and by the Rainbow Project (RP10; Netherlands Exome Chip Project) of the Biobanking and Biomolecular Research Infrastructure Netherlands (BBMRI-NL; www.bbmri.nl). We thank Ms. Mila Jhamai, Ms. Sarah Higgins, and Mr. Marijn Verkerk for their help in creating the exome chip database. The authors are grateful to the study participants, the staff from the Rotterdam Study and the participating general practitioners and pharmacists.

REHS: The core management of the Raine Study is funded by the University of Western Australia, Australia; the Telethon Institute for Child Health Research, Australia; Raine Medical Research Foundation, Australia; Women's and Infant's Research Foundation, Australia; Curtin University, Australia; Murdoch University, Australia; Edith Cowan University, Australia; and the University of Notre Dame, Australia. The Generation-2 20-year follow-up of the Raine Study was funded by the National Health and Medical Research Council (NHMRC), Australia: project grant no.: 1 021 105. The Generation-2 28-year follow-up of the Raine Study was funded by the NHMRC, Australia: project grants 1 121 979 and 1 126 494

SCES, SiMES, SINDI: The Singapore studies (SCES, SiMES, SINDI) were supported by the National Medical Research Council, Singapore (NMRC 0796/2003, NMRC 1176/2008, STaR/0003/2008; CG/SERI/2010), Biomedical Research Council, Singapore (06/1/21/19/466, 09/1/35/19/616 and 08/1/35/19/550). The Singapore Tissue Network and the Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore provided services.

TwinsUK: TwinsUK received funding from the Wellcome Trust; the European Union MyEuropia Marie Curie Research Training Network; Guide Dogs for the Blind Association; the European 18 Community's FP7 (HEALTHF22008201865GEFOS); ENGAGE (HEALTHF42007201413); the FP-5 GenomEUtwin Project (QLG2CT200201254); US National Institutes of Health/National Eye Institute (1RO1EY018246); NIH Center for Inherited Disease Research; the National Institute for Health Research comprehensive Biomedical Research Centre award to Guy's and St. Thomas' National Health Service Foundation Trust partnering with King's College London. P.G.H. is the recipient of a Fight for Sight ECI award. We acknowledge the contribution of Drs Toby Andrew, Margarida Lopes, Samantha Fahy and Diana Kozareva.

Data Availability

The data that support the findings of this study are not publicly available due to information that could compromise research participant privacy and/or consent. European Union data privacy rulings currently forbid sharing of genomic data outside the EU and several of the participating studies have additional restrictions to protect the privacy of the study participants. Deidentified data were used here under data use agreements with each participating study. Data may be available by request from the individual participating studies if all regulatory conditions are met.

Code Availability

The R scripts used to run the EPACTS software for the association study are available upon request from the first authors or corresponding author.

Competing Interests

The authors state that they have no competing interests.

References

1. Vitale, S., Sperduto, R.D. & Ferris, F.L., 3rd. Increased prevalence of myopia in the United States between 1971-1972 and 1999-2004. *Arch Ophthalmol* **127**, 1632-9 (2009).
2. Williams, K.M. *et al.* Increasing Prevalence of Myopia in Europe and the Impact of Education. *Ophthalmology* **122**, 1489-97 (2015).
3. Morgan, I.G., Ohno-Matsui, K. & Saw, S.M. Myopia. *Lancet* **379**, 1739-48 (2012).
4. Wang, J. *et al.* Prevalence of myopia and vision impairment in school students in Eastern China. *BMC Ophthalmol* **20**, 2 (2020).
5. Verhoeven, V.J. *et al.* Visual consequences of refractive errors in the general population. *Ophthalmology* **122**, 101-9 (2015).
6. Tideman, J.W. *et al.* Association of Axial Length With Risk of Uncorrectable Visual Impairment for Europeans With Myopia. *JAMA Ophthalmol* **134**, 1355-1363 (2016).
7. Flitcroft, D.I. The complex interactions of retinal, optical and environmental factors in myopia aetiology. *Prog Retin Eye Res* **31**, 622-60 (2012).
8. Fricke, T.R. *et al.* Global prevalence of visual impairment associated with myopic macular degeneration and temporal trends from 2000 through 2050: systematic review, meta-analysis and modelling. *Br J Ophthalmol* **102**, 855-862 (2018).
9. Holden, B.A. *et al.* Global Prevalence of Myopia and High Myopia and Temporal Trends from 2000 through 2050. *Ophthalmology* **123**, 1036-42 (2016).
10. Bourne, R.R. *et al.* Causes of vision loss worldwide, 1990-2010: a systematic analysis. *Lancet Glob Health* **1**, e339-49 (2013).
11. Dolgin, E. The myopia boom. *Nature* **519**, 276-8 (2015).
12. Stambolian, D. Genetic susceptibility and mechanisms for refractive error. *Clin Genet* **84**, 102-8 (2013).
13. Stambolian, D. *et al.* Meta-analysis of genome-wide association studies in five cohorts reveals common variants in RBFOX1, a regulator of tissue-specific splicing, associated with refractive error. *Hum Mol Genet* **22**, 2754-64 (2013).
14. Fan, Q. *et al.* Meta-analysis of gene-environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nat Commun* **7**, 11008 (2016).
15. Kiefer, A.K. *et al.* Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet* **9**, e1003299 (2013).
16. Shi, Y. *et al.* Genetic variants at 13q12.12 are associated with high myopia in the Han Chinese population. *Am J Hum Genet* **88**, 805-813 (2011).

17. Nakanishi, H. *et al.* A genome-wide association analysis identified a novel susceptible locus for pathological myopia at 11q24.1. *PLoS Genet* **5**, e1000660 (2009).
18. Li, Y.J. *et al.* Genome-wide association studies reveal genetic variants in CTNND2 for high myopia in Singapore Chinese. *Ophthalmology* **118**, 368-75 (2011).
19. Verhoeven, V.J. *et al.* Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* **45**, 314-8 (2013).
20. Verhoeven, V.J. *et al.* Large scale international replication and meta-analysis study confirms association of the 15q14 locus with myopia. The CREAM consortium. *Hum Genet* **131**, 1467-80 (2012).
21. Tedja, M.S. *et al.* Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet* **50**, 834-848 (2018).
22. Hysi, P.G. *et al.* Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia. *Nat Genet* **52**, 401-407 (2020).
23. Lopes, M.C., Andrew, T., Carbonaro, F., Spector, T.D. & Hammond, C.J. Estimating heritability and shared environmental effects for refractive error in twin and family studies. *Invest Ophthalmol Vis Sci* **50**, 126-31 (2009).
24. Hysi, P.G., Wojciechowski, R., Rahi, J.S. & Hammond, C.J. Genome-wide association studies of refractive error and myopia, lessons learned, and implications for the future. *Invest Ophthalmol Vis Sci* **55**, 3344-51 (2014).
25. Pärssinen, O., Kauppinen, M., Kaprio, J., Koskenvuo, M. & Rantanen, T. Heritability of refractive astigmatism: a population-based twin study among 63- to 75-year-old female twins. *Invest Ophthalmol Vis Sci* **54**, 6063-7 (2013).
26. Pärssinen, O. *et al.* Heritability of spherical equivalent: a population-based twin study among 63-to 76-year-old female twins. *Ophthalmology* **117**, 1908-11 (2010).
27. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-53 (2009).
28. Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* **83**, 311-21 (2008).
29. Price, A.L. *et al.* Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* **86**, 832-8 (2010).
30. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559-75 (2007).
31. Conomos, M.P., Miller, M.B. & Thornton, T.A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet Epidemiol* **39**, 276-93 (2015).
32. Kang, H.M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* **42**, 348-54 (2010).
33. Price, A.L., Zaitlen, N.A., Reich, D. & Patterson, N. New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* **11**, 459-63 (2010).
34. Moutsianas, L. *et al.* The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. *PLoS Genet* **11**, e1005165 (2015).
35. Liu, Y. *et al.* ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet* **104**, 410-421 (2019).
36. Fisher, R.A. *Statistical methods for research workers*, (Oliver and Boyd, Edinburgh, 1925).
37. Dewey, M. metap: meta-analysis of significance values. R package version 1.4 edn (2020).
38. Krämer, A., Green, J., Pollard, J., Jr. & Tugendreich, S. Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics* **30**, 523-30 (2014).

39. Watanabe, K., Umićević Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P. & Posthuma, D. Genetic mapping of cell type specificity for complex traits. *Nat Commun* **10**, 3222 (2019).
40. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun* **8**, 1826 (2017).
41. Fritsche, L.G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat Genet* **48**, 134-43 (2016).
42. Cowan, C.S. *et al.* Cell Types of the Human Retina and Its Organoids at Single-Cell Resolution. *Cell* **182**, 1623-1640 e34 (2020).
43. Bauer-Mehren, A., Rautschka, M., Sanz, F. & Furlong, L.I. DisGeNET: a Cytoscape plugin to visualize, integrate, search and analyze gene-disease networks. *Bioinformatics* **26**, 2924-6 (2010).
44. Günther, S. *et al.* SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res* **36**, D919-22 (2008).
45. Whirl-Carrillo, M. *et al.* Pharmacogenomics knowledge for personalized medicine. *Clin Pharmacol Ther* **92**, 414-7 (2012).
46. Szklarczyk, D. *et al.* STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data. *Nucleic Acids Res* **44**, D380-4 (2016).
47. Wishart, D.S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res* **46**, D1074-D1082 (2018).
48. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**, e164 (2010).
49. Yang, H. & Wang, K. Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. *Nat Protoc* **10**, 1556-66 (2015).
50. Chang, X. & Wang, K. wANNOVAR: annotating genetic variants for personal genomes via the web. *J Med Genet* **49**, 433-6 (2012).
51. Sim, N.L. *et al.* SIFT web server: predicting effects of amino acid substitutions on proteins. *Nucleic Acids Res* **40**, W452-7 (2012).
52. Adzhubei, I., Jordan, D.M. & Sunyaev, S.R. Predicting functional effect of human missense mutations using PolyPhen-2. *Curr Protoc Hum Genet* **Chapter 7**, Unit7 20 (2013).
53. Schwarz, J.M., Cooper, D.N., Schuelke, M. & Seelow, D. MutationTaster2: mutation prediction for the deep-sequencing age. *Nat Methods* **11**, 361-2 (2014).
54. Rentzsch, P., Witten, D., Cooper, G.M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res* **47**, D886-D894 (2019).
55. Shihab, H.A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Hum Mutat* **34**, 57-65 (2013).
56. Berman, H.M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-42 (2000).
57. Schymkowitz, J. *et al.* The FoldX web server: an online force field. *Nucleic Acids Res* **33**, W382-8 (2005).
58. Pettersen, E.F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators, and developers. *Protein Sci* **30**, 70-82 (2021).
59. Khanna, T., Hanna, G., Sternberg, M.J.E. & David, A. Missense3D-DB web catalogue: an atom-based analysis and repository of 4M human protein-coding genetic variants. *Hum Genet* **140**, 805-812 (2021).
60. Paschalis, E.I. *et al.* Microglia Regulate Neuroglia Remodeling in Various Ocular and Retinal Injuries. *J Immunol* **202**, 539-549 (2019).

61. Wolf, A., Aslanidis, A. & Langmann, T. Retinal expression and localization of Mef2c support its important role in photoreceptor gene expression. *Biochem Biophys Res Commun* **483**, 346-351 (2017).
62. Sun, J., Yoon, J., Lee, M., Hwang, Y.S. & Daar, I.O. Sprouty2 regulates positioning of retinal progenitors through suppressing the Ras/Raf/MAPK pathway. *Sci Rep* **10**, 13752 (2020).
63. Wei, J., Jiang, H., Gao, H. & Wang, G. Raf-1 Kinase Inhibitory Protein (RKIP) Promotes Retinal Ganglion Cell Survival and Axonal Regeneration Following Optic Nerve Crush. *J Mol Neurosci* **57**, 243-8 (2015).
64. Sowden, J.C., Holt, J.K., Meins, M., Smith, H.K. & Bhattacharya, S.S. Expression of Drosophila omb-related T-box genes in the developing human and mouse neural retina. *Invest Ophthalmol Vis Sci* **42**, 3095-102 (2001).
65. Koshiba-Takeuchi, K. *et al.* Tbx5 and the retinotectum projection. *Science* **287**, 134-7 (2000).
66. Oda, M., Yamamoto, H., Matsumoto, H., Ishizaki, Y. & Shibasaki, K. TRPC5 regulates axonal outgrowth in developing retinal ganglion cells. *Lab Invest* **100**, 297-310 (2020).
67. Simpson, C.L. *et al.* Exome genotyping and linkage analysis identifies two novel linked regions and replicates two others for myopia in Ashkenazi Jewish families. *BMC Med Genet* **20**, 27 (2019).
68. Musolf, A.M. *et al.* Genome-wide scans of myopia in Pennsylvania Amish families reveal significant linkage to 12q15, 8q21.3 and 5p15.33. *Hum Genet* **138**, 339-354 (2019).
69. Wojciechowski, R. *et al.* Genomewide scan in Ashkenazi Jewish families demonstrates evidence of linkage of ocular refraction to a QTL on chromosome 1p36. *Hum Genet* **119**, 389-99 (2006).
70. Archer, S.N., Schmidt, C., Vandewalle, G. & Dijk, D.J. Phenotyping of PER3 variants reveals widespread effects on circadian preference, sleep regulation, and health. *Sleep Med Rev* **40**, 109-126 (2018).
71. Stone, R.A. *et al.* Image defocus and altered retinal gene expression in chick: clues to the pathogenesis of ametropia. *Invest Ophthalmol Vis Sci* **52**, 5765-77 (2011).
72. Stone, R.A. *et al.* Visual Image Quality Impacts Circadian Rhythm-Related Gene Expression in Retina and in Choroid: A Potential Mechanism for Ametropias. *Invest Ophthalmol Vis Sci* **61**, 13 (2020).
73. Lee, S.S. & Mackey, D.A. Prevalence and Risk Factors of Myopia in Young Adults: Review of Findings From the Raine Study. *Front Public Health* **10**, 861044 (2022).
74. Subramanian, L. *et al.* Ca2+ binding to EF hands 1 and 3 is essential for the interaction of apoptosis-linked gene-2 with Alix/AIP1 in ocular melanoma. *Biochemistry* **43**, 11175-86 (2004).
75. Zhou, H.J. *et al.* AIP1 mediates vascular endothelial cell growth factor receptor-3-dependent angiogenic and lymphangiogenic responses. *Arterioscler Thromb Vasc Biol* **34**, 603-15 (2014).
76. Gupta, N., Fong, J., Ang, L.C. & Yücel, Y.H. Retinal tau pathology in human glaucomas. *Can J Ophthalmol* **43**, 53-60 (2008).
77. Nakazawa, K. *et al.* Defective processing of keratan sulfate in macular corneal dystrophy. *J Biol Chem* **259**, 13751-7 (1984).
78. Liskova, P. *et al.* Ectopic GRHL2 Expression Due to Non-coding Mutations Promotes Cell State Transition and Causes Posterior Polymorphous Corneal Dystrophy 4. *Am J Hum Genet* **102**, 447-459 (2018).
79. Rahikkala, E. *et al.* Biallelic loss-of-function P4HTM gene variants cause hypotonia, hypoventilation, intellectual disability, dysautonomia, epilepsy, and eye abnormalities (HIDEA syndrome). *Genet Med* **21**, 2355-2363 (2019).
80. Leinonen, H. *et al.* Lack of P4H-TM in mice results in age-related retinal and renal alterations. *Hum Mol Genet* **25**, 3810-3823 (2016).

81. McGee, T.L., Seyedahmadi, B.J., Sweeney, M.O., Dryja, T.P. & Berson, E.L. Novel mutations in the long isoform of the USH2A gene in patients with Usher syndrome type II or non-syndromic retinitis pigmentosa. *J Med Genet* **47**, 499-506 (2010).
82. Fu, J. *et al.* Novel compound heterozygous nonsense variants, p.L150* and p.Y3565*, of the USH2A gene in a Chinese pedigree are associated with Usher syndrome type IIA. *Mol Med Rep* **22**, 3464-3472 (2020).
83. Wan, L., Deng, B., Wu, Z. & Chen, X. Exome sequencing study of 20 patients with high myopia. *PeerJ* **6**, e5552 (2018).
84. Ito, K., Yoshiura, Y., Ototake, M. & Nakanishi, T. Macrophage migration inhibitory factor (MIF) is essential for development of zebrafish, Danio rerio. *Dev Comp Immunol* **32**, 664-72 (2008).
85. Abu El-Asrar, A.M. *et al.* The Proinflammatory and Proangiogenic Macrophage Migration Inhibitory Factor Is a Potential Regulator in Proliferative Diabetic Retinopathy. *Front Immunol* **10**, 2752 (2019).
86. Kim, B. *et al.* MIF Inhibitor ISO-1 Protects Photoreceptors and Reduces Gliosis in Experimental Retinal Detachment. *Sci Rep* **7**, 14336 (2017).
87. Morava, E. *et al.* Ophthalmological abnormalities in children with congenital disorders of glycosylation type I. *Br J Ophthalmol* **93**, 350-4 (2009).
88. Park, P.J. & Shukla, D. Role of heparan sulfate in ocular diseases. *Exp Eye Res* **110**, 1-9 (2013).
89. Zhu, X. *et al.* Profiling and Bioinformatic Analysis of Differentially Expressed Cytokines in Aqueous Humor of High Myopic Eyes - Clues for Anti-VEGF Injections. *Curr Eye Res* **45**, 97-103 (2020).
90. Aldave, A.J. *et al.* Novel mutations in the carbohydrate sulfotransferase gene (CHST6) in American patients with macular corneal dystrophy. *Am J Ophthalmol* **137**, 465-73 (2004).
91. Cheng, L. *et al.* Modulation of retinal Müller cells by complement receptor C5aR. *Invest Ophthalmol Vis Sci* **54**, 8191-8 (2013).
92. Kawashima-Kumagai, K. *et al.* A genome-wide association study identified a novel genetic loci STON1-GTF2A1L/LHCGR/FSHR for bilaterality of neovascular age-related macular degeneration. *Sci Rep* **7**, 7173 (2017).
93. Martín-Sánchez, M. *et al.* A Multi-Strategy Sequencing Workflow in Inherited Retinal Dystrophies: Routine Diagnosis, Addressing Unsolved Cases and Candidate Genes Identification. *Int J Mol Sci* **21**(2020).
94. Ilhan, H.D., Bilgin, A.B., Toylu, A., Dogan, M.E. & Apaydin, K.C. The Expression of GDF-15 in the Human Vitreous in the Presence of Retinal Pathologies with an Inflammatory Component. *Ocul Immunol Inflamm* **24**, 178-83 (2016).
95. Ban, N., Siegfried, C.J. & Apte, R.S. Monitoring Neurodegeneration in Glaucoma: Therapeutic Implications. *Trends Mol Med* **24**, 7-17 (2018).
96. Kolko, M. *et al.* Lactate Transport and Receptor Actions in Retina: Potential Roles in Retinal Function and Disease. *Neurochem Res* **41**, 1229-36 (2016).
97. Harun-Or-Rashid, M. & Inman, D.M. Reduced AMPK activation and increased HCAR activation drive anti-inflammatory response and neuroprotection in glaucoma. *J Neuroinflammation* **15**, 313 (2018).
98. Cheng, C.L. & Molday, R.S. Interaction of 4.1G and cGMP-gated channels in rod photoreceptor outer segments. *J Cell Sci* **126**, 5725-34 (2013).
99. Karlstetter, M. *et al.* Polysialic acid blocks mononuclear phagocyte reactivity, inhibits complement activation, and protects from vascular damage in the retina. *EMBO Mol Med* **9**, 154-166 (2017).

100. Kustermann, S., Hildebrandt, H., Bolz, S., Dengler, K. & Kohler, K. Genesis of rods in the zebrafish retina occurs in a microenvironment provided by polysialic acid-expressing Muller glia. *J Comp Neurol* **518**, 636-46 (2010).
101. Jansen, R.W. *et al.* MR Imaging Features of Retinoblastoma: Association with Gene Expression Profiles. *Radiology* **288**, 506-515 (2018).
102. Chiambaretta, F. *et al.* Cell and tissue specific expression of human Kruppel-like transcription factors in human ocular surface. *Mol Vis* **10**, 901-9 (2004).
103. Chen, F. *et al.* Variation in PTCHD2, CRISP3, NAP1L4, FSCB, and AP3B2 associated with spherical equivalent. *Mol Vis* **22**, 783-96 (2016).
104. Li, Y.J. *et al.* An international collaborative family-based whole-genome linkage scan for high-grade myopia. *Invest Ophthalmol Vis Sci* **50**, 3116-27 (2009).
105. Wu, C. *et al.* BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biol* **10**, R130 (2009).

Figure Legends

Figure 1: P-values of the multiethnic meta-analysis. The gene-based p-values of the meta-analysis association study combining all five cohorts (N=26,946) using the A) EMMAX-VT test, B) EMMAX-CMC test, and C) ACAT. The line represents the genome-wide significant threshold of 1×10^{-5}.

Figure 2: Overlap between three tests in the multiethnic meta-analysis. A Venn diagram showing the overlap and unique significant genes in the multiethnic meta-analysis using the three different tests: EMMAX-VT (green), EMMAX-CMC (red), and ACAT (blue).

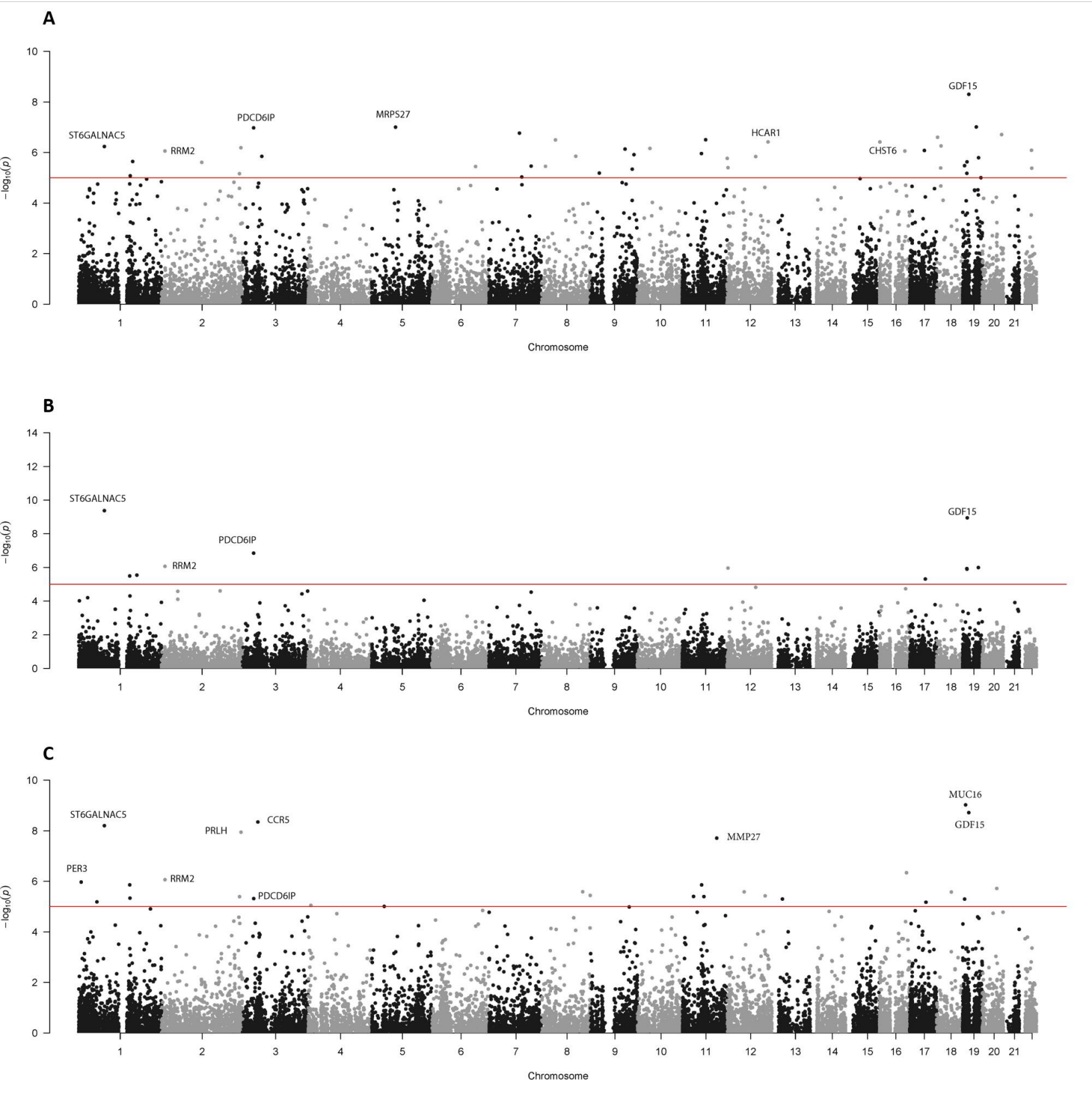
Figure 3: P-values of the Indo-European meta-analysis. The gene-based p-values of the meta-analysis association study (N=22,079) combining the four Indo-European derived cohorts using the A) EMMAX-VT test, B) EMMAX-CMC test, and C) ACAT. The line represents the genome-wide significant threshold of 1×10^{-5}.

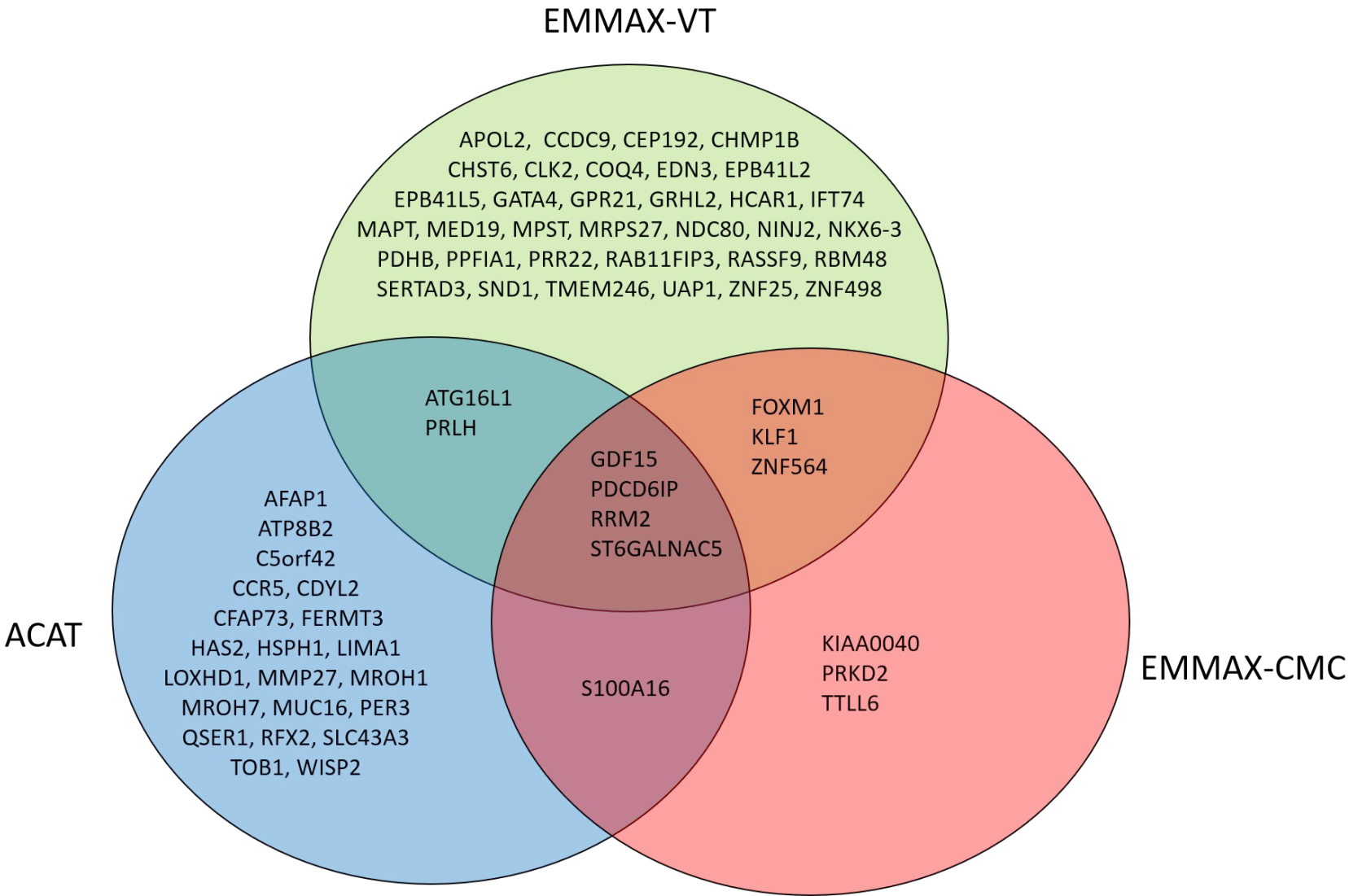
Figure 4: Overlap between three tests in the Indo-European meta-analysis. A Venn diagram showing the overlap and unique significant genes in the Indo-European cohorts meta-analysis using the three different tests: EMMAX-VT (green), EMMAX-CMC (red), and ACAT (blue).

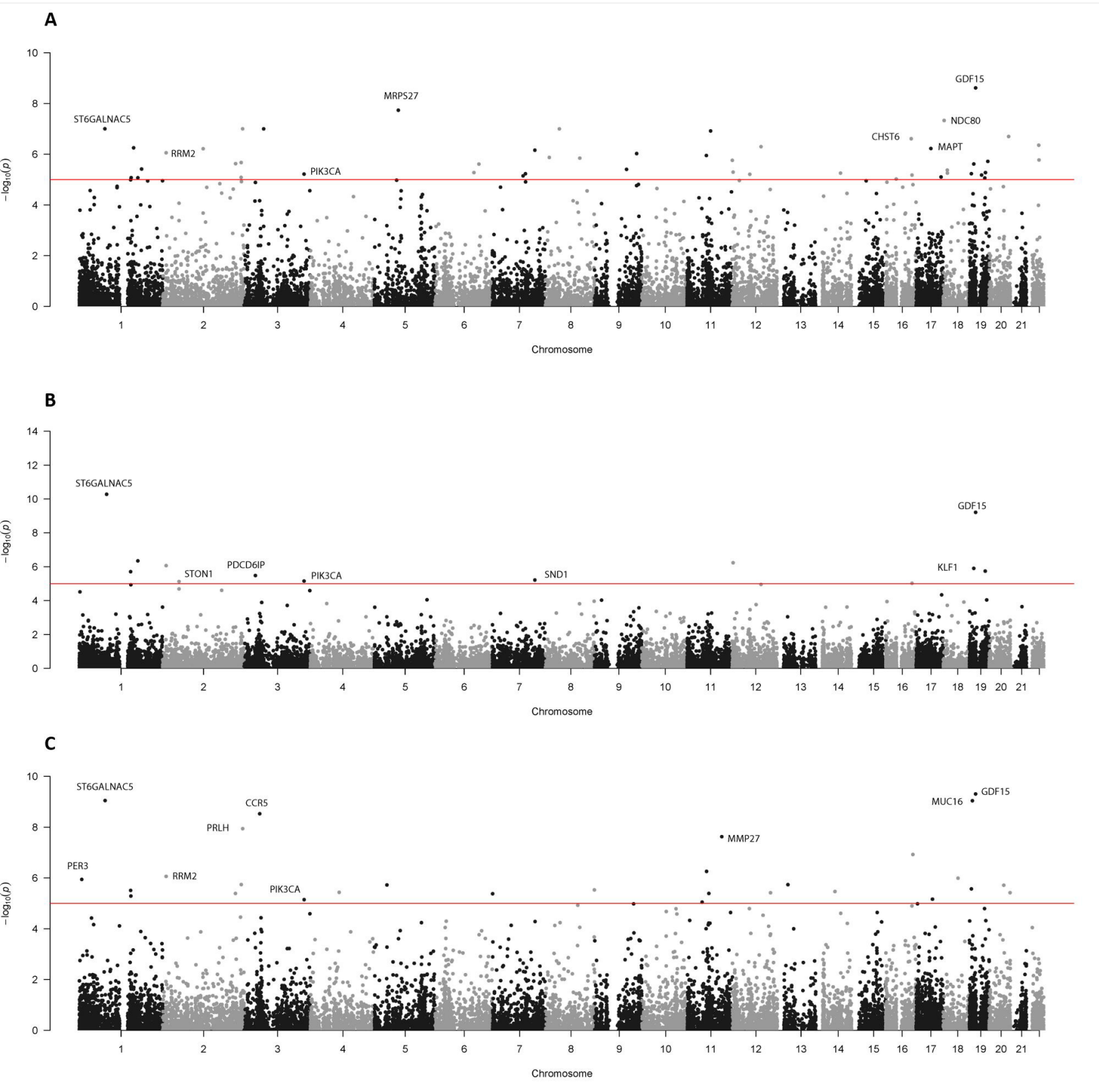
Figure 5: P-values of the analysis using the Eastern Asian EACC only. The gene-based p-values of the EACC association analysis (N=4,867) using the A) EMMAX-VT test, B) EMMAX-CMC test, and C) ACAT. The line represents the genome-wide significant threshold of 1×10^{-5}.

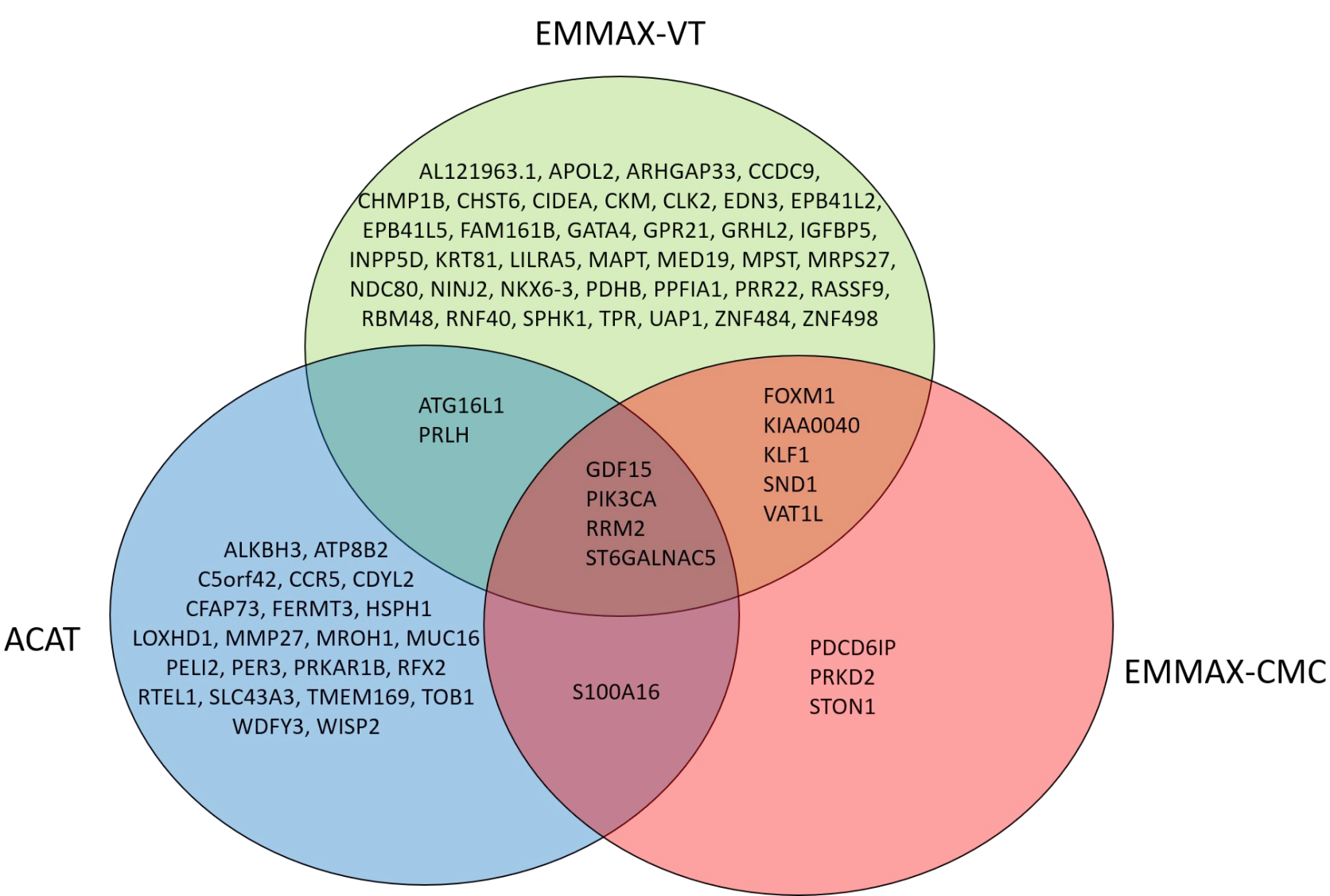
Figure 6: Overlap between three tests in the Eastern Asian EACC analysis. A Venn diagram showing the overlap and unique significant genes in the EACC analysis using the three different tests: EMMAX-VT (green), EMMAX-CMC (red), and ACAT (blue).

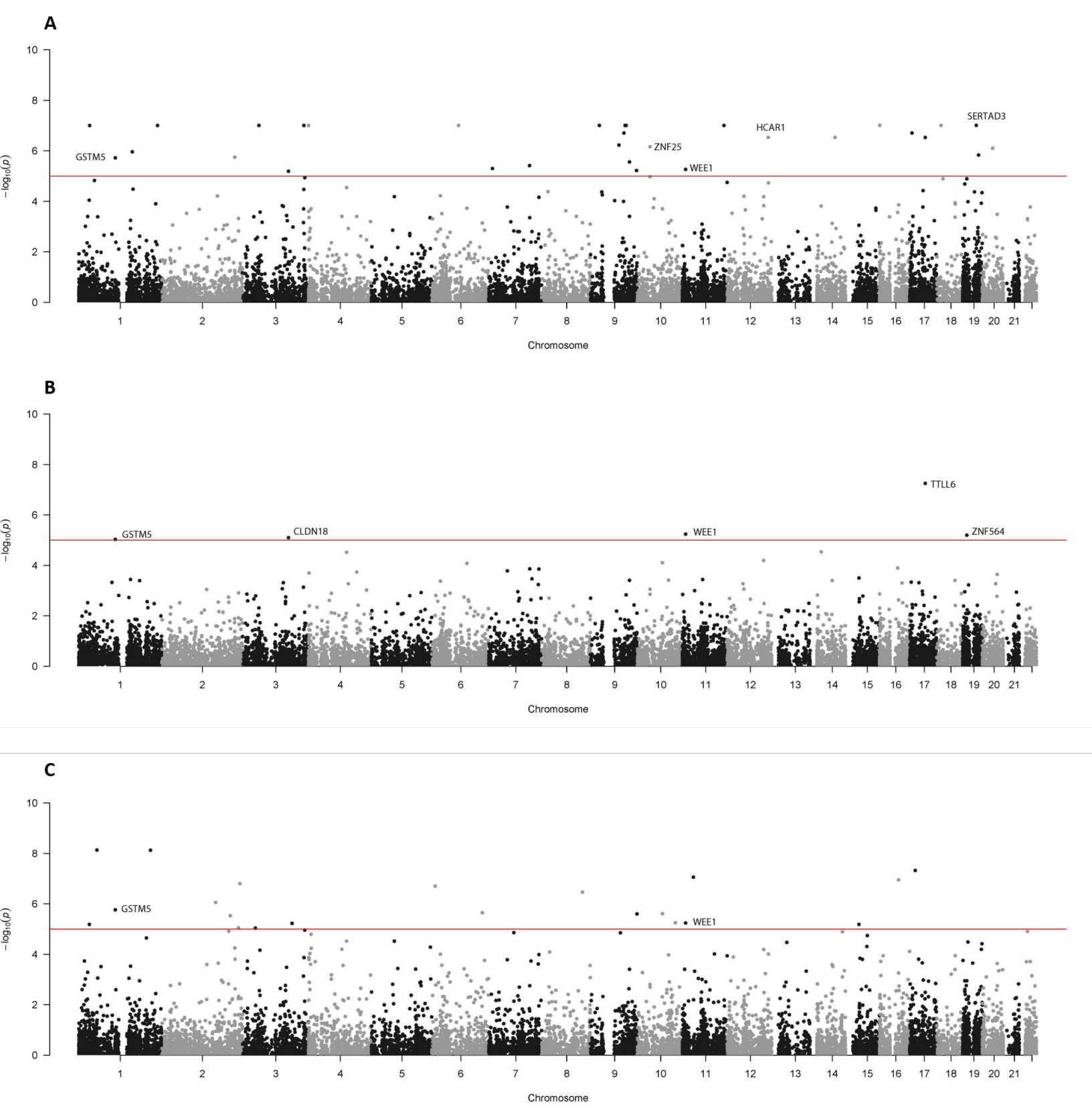
Figure 7: Prioritization of top genes from all 129 genome-wide significant genes. The top genes ranked by our prioritization schema. The figure contains the chromosome, basepair position, gene name, as well as the meta-analysis p-value and the individual cohort p-values for each gene. It also contains which test the given significant meta-analysis p-value refers to, and how many times the gene replicated in our internal analyses. Finally, it contains information regarding gene expression, whether the gene has a known ocular phenotype in mice or humans, overlap with the GWAS performed by Hysi et al., and the final overall prioritization score.

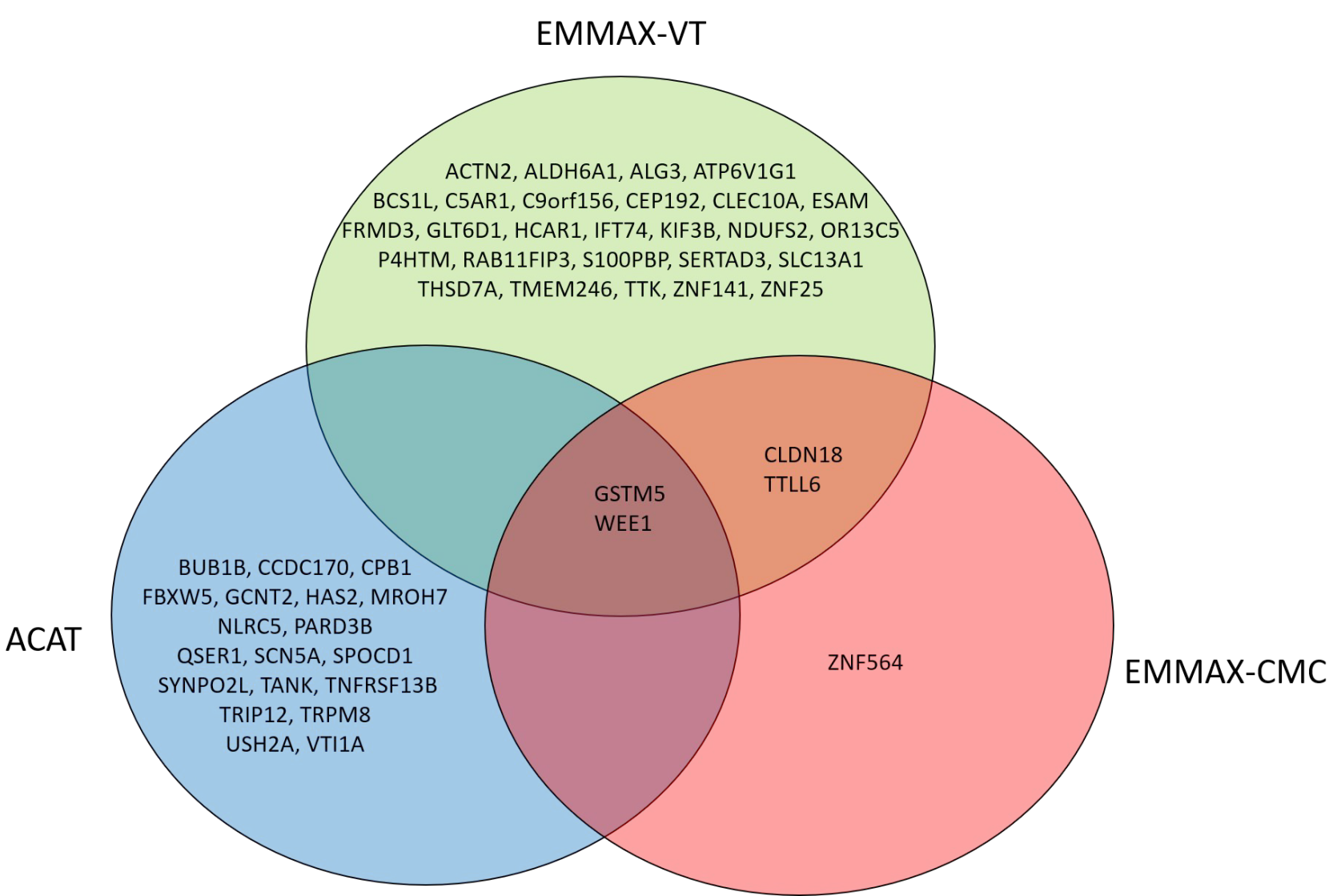

A
ST6GALNAC5
RRM2
PDCD6IP
MRPS27
HCAR1
CHST6
GDF15
−log10(p)
Chromosome
B
ST6GALNAC5
PDCD6IP
RRM2
GDF15
−log10(p)
Chromosome
C
ST6GALNAC5
PRLH
CCR5
PER3
RRM2
PDCD6IP
MMP27
MUC16
GDF15
−log10(p)
Chromosome


EMMAX-VT
APOL2, CCDC9, CEP192, CHMP1B
CHST6, CLK2, COQ4, EDN3, EPB41L2
EPB41L5, GATA4, GPR21, GRHL2, HCAR1, IFT74
MAPT, MED19, MPST, MRPS27, NDC80, NINJ2, NKX6-3
PDHB, PPFIA1, PRR22, RAB11FIP3, RASSF9, RBM48
SERTAD3, SND1, TMEM246, UAP1, ZNF25, ZNF498
ATG16L1
PRLH
FOXM1
KLF1
ZNF564
GDF15
PDCD6IP
RRM2
ST6GALNAC5
ACAT
AFAP1
ATP8B2
C5orf42
CCR5, CDYL2
CFAP73, FERMT3
HAS2, HSPH1, LIMA1
LOXHD1, MMP27, MROH1
MROH7, MUC16, PER3
QSER1, RFX2, SLC43A3
TOB1, WISP2
S100A16
KIAA0040
PRKD2
TTLL6
EMMAX-CMC


A
ST6GALNAC5
RRM2
PIK3CA
MRPS27
GDF15
NDC80
CHST6
MAPT
$-\log_{10}(p)$
Chromosome
B
ST6GALNAC5
STON1
PDCD6IP
PIK3CA
SND1
GDF15
KLF1
$-\log_{10}(p)$
Chromosome
C
ST6GALNAC5
CCR5
PRLH
PER3
RRM2
PIK3CA
MMP27
MUC16
GDF15
$-\log_{10}(p)$
Chromosome


EMMAX-VT
AL121963.1, APOL2, ARHGAP33, CCDC9, CHMP1B, CHST6, CIDEA, CKM, CLK2, EDN3, EPB41L2, EPB41L5, FAM161B, GATA4, GPR21, GRHL2, IGFBP5, INPP5D, KRT81, LILRA5, MAPT, MED19, MPST, MRPS27, NDC80, NINJ2, NKX6-3, PDHB, PPFIA1, PRR22, RASSF9, RBM48, RNF40, SPHK1, TPR, UAP1, ZNF484, ZNF498
ATG16L1
PRLH
FOXM1
KIAA0040
KLF1
SND1
VAT1L
GDF15
PIK3CA
RRM2
ST6GALNAC5
ACAT
ALKBH3, ATP8B2
C5orf42, CCR5, CDYL2
CFAP73, FERMT3, HSPH1
LOXHD1, MMP27, MROH1, MUC16
PELI2, PER3, PRKAR1B, RFX2
RTEL1, SLC43A3, TMEM169, TOB1
WDFY3, WISP2
S100A16
PDCD6IP
PRKD2
STON1
EMMAX-CMC


A
B
C
−log10(p)
Chromosome
GSTM5
ZNF25
WEE1
HCAR1
SERTAD3
CLDN18
TTLL6
ZNF564


EMMAX-VT
ACTN2, ALDH6A1, ALG3, ATP6V1G1
BCS1L, C5AR1, C9orf156, CEP192, CLEC10A, ESAM
FRMD3, GLT6D1, HCAR1, IFT74, KIF3B, NDUFS2, OR13C5
P4HTM, RAB11FIP3, S100PBP, SERTAD3, SLC13A1
THSD7A, TMEM246, TTK, ZNF141, ZNF25
ACAT
BUB1B, CCDC170, CPB1
FBXW5, GCNT2, HAS2, MROH7
NLRC5, PARD3B
QSER1, SCN5A, SPOCD1
SYNPO2L, TANK, TNFRSF13B
TRIP12, TRPM8
USH2A, VTI1A
GSTM5
WEE1
CLDN18
TTLL6
ZNF564
EMMAX-CMC

	Gene			Meta-analysis	Individual populations					Test	Internal replication		External replication UKBB		Expression	Biology		GWAS	Total	Drug
	Chr	Pos	Gene	P-value	EACC	EPIC	BDES	IECC	REHS	Analysis and cohort	One cohort <=10-5 and the other p<0.05	Overlap with other test	VT	CMC	Total (any of 4 models)	Ocular phenotype in mice	Ocular phenotype in human	GWAS overlap	Sum	4 databases
VT-ALL results	3	33877626	PDCD6IP	1.07E-07	4.10E-04	NA	NA	1.30E-05	NA	VT-all	1	1	2.60E-01	8.87E-02	1	0	1	0	4	X
	17	44039717	MAPT	8.57E-07	1.90E-01	3.90E-02	4.40E-01	1.70E-01	1.00E-07	VT-all	0	1	1.60E-01	1.03E-01	1	1	1	0	4	X
	16	75512672	CHST6	8.99E-07	5.60E-01	9.20E-02	9.50E-03	2.00E-07	6.00E-01	VT-all	0	1	7.20E-01	7.23E-01	0	1	2	0	4	X
	8	102555474	GRHL2	1.42E-06	NA	3.80E-01	8.60E-01	8.20E-03	3.00E-07	VT-all	0	1	9.60E-01	6.24E-01	0	1	2	0	4	X
	19	18497141	GDF15	5.12E-09	2.00E-01	9.90E-01	NA	2.00E-07	3.40E-05	VT-all	1	1	2.20E-01	3.92E-01	0	0	1	0	3	X
	2	10262920	RRM2	8.81E-07	NA	NA	2.70E-01	8.80E-03	1.80E-06	VT-all	1	1	5.20E-01	4.40E-01	0	0	1	0	3	X
Other tests	1	7845014	PER3	1.08E-06	1.16E-01	2.19E-01	4.81E-01	4.98E-02	1.20E-07	ACAT-all	1	0	5.90E-01	2.99E-01	1	1	2	0	5	X
	13	31712572	HSPH1	5.04E-06	4.37E-01	5.79E-01	2.36E-03	2.46E-01	3.19E-06	ACAT-all	1	0	2.20E-01	5.08E-01	1	0	1	0	3	X
	1	186313129	TPR	3.85E-06	NA	7.60E-01	1.40E-01	1.60E-03	1.50E-05	VT-IECC	1	0	4.30E-01	8.70E-01	1	1	0	0	3	X
	12	52681460	KRT81	6.17E-06	NA	3.20E-01	7.70E-01	1.80E-01	1.00E-07	VT-IECC	1	0	1.60E-01	9.79E-01	1	0	1	0	3	0
	17	74381555	SPHK1	7.84E-06	NA	3.00E-01	7.10E-01	9.20E-03	3.00E-06	VT-IECC	1	0	7.10E-01	2.18E-01	1	0	1	0	3	X
EACC only	3	49039984	P4HTM	1.65E-05	1.00E-07	5.10E-01	2.90E-01	2.40E-01	5.60E-01	VT-all	0	0	1.60E-02	1.09E-03	1	1	2	0	4	X
	1	215802301	USH2A	1.25E-05	7.55E-09	9.84E-01	3.51E-01	6.71E-01	8.11E-01	ACAT-all	0	0	7.20E-01	5.58E-01	1	1	2	0	4	X
	1	110257814	GSTM5	1.07E-04	1.90E-06	2.00E-01	3.70E-01	9.50E-01	NA	VT-all	0	1	8.70E-01	9.72E-01	1	0	1	0	3	X
	7	11500346	THSD7A	6.07E-04	5.10E-06	3.90E-01	1.40E-01	8.40E-01	8.30E-01	VT-all	0	0	6.50E-01	9.93E-01	1	1	1	0	3	0
	11	9606879	WEE1	2.55E-04	5.50E-06	8.30E-01	NA	5.80E-01	NA	VT-all	0	1	7.40E-01	4.50E-01	1	0	1	0	3	X
	15	40462771	BUB1B	1.82E-03	6.52E-06	9.80E-01	2.37E-01	7.50E-01	7.41E-01	ACAT-all	0	0	3.10E-01	1.82E-01	0	1	2	0	3	X

Table 1: P-values and Effect Sizes of Prioritized Genes

	Meta-analysis			Discovery Set			Replication Sets											
	Multiethnic P-values			IECC P-values			EACC P-values			BDES P-values			EPIC-Norfolk P-values			REHS P-values		
Gene	**CMC**	**VT**	**ACAT**	**CMC (beta)**	**VT**	**ACAT**	**CMC (beta)**	**VT**	**ACAT**	**CMC (beta)**	**VT**	**ACAT**	**CMC (beta)**	**VT**	**ACAT**	**CMC (beta)**	**VT**	**ACAT**
PDCD6IP	2.45e-7	1e-7	4.9e-6	3.4e-6 (0.0095)	1.3e-5	5.6e-4	0.002 (-0.76)	4.1e-4	5.5e-4	NA	NA	NA	NA	NA	NA	NA	NA	NA
PER3	0.08	0.03	1e-6	0.14 (0.17)	0.04	0.05	0.27 (0.16)	0.55	0.12	0.51 (-0.2)	0.42	0.48	0.43 (-0.11)	0.11	0.22	0.02 (-0.52)	0.04	1.2e-7
USH2A	0.44	0.90	1.3e-5	0.82 (0.01)	0.96	0.67	0.38 (-0.08)	0.75	7.6e-9	0.78 (-0.05)	0.72	0.35	0.33 (0.08)	0.90	0.98	0.08 (0.22)	0.19	0.81
MAPT	0.02	8.6e-7	0.57	0.02 (-0.43)	0.17	0.18	0.15 (0.31)	0.19	0.37	0.43 (-0.5)	0.44	0.24	0.48 (-0.17)	0.04	0.87	0.01 (-0.8)	1e-7	0.94
GRHL2	0.47	1.4e-6	0.06	0.31 (0.33)	0.008	0.08	NA	NA	NA	0.87 (-0.19)	0.86	0.08	0.39 (-0.55)	0.38	0.39	0.21 (0.87)	3e-7	0.21
CHST6	0.19	9e-7	0.06	0.58 (-0.09)	2e-7	0.40	0.38 (0.21)	0.56	0.58	0.05 (0.96)	0.009	0.01	0.27 (0.27)	0.09	0.16	0.35 (0.42)	0.6	0.42
P4HTM	0.09	1.7e-5	0.16	0.39 (0.01)	0.24	0.10	0.005 (-2.02)	1e-7	0.14	0.34 (0.61)	0.29	0.19	0.49 (-0.16)	0.51	0.54	0.83 (-0.12)	0.56	0.57

Legend: The summary statistics from our prioritized genes. The p-values for the CMC, VT, and ACAT analyses and betas for the CMC analyses for the meta-analysis and the five individual cohorts. Note that for the CMC meta-analysis, no beta is provided because Fisher's method does not provide an effect size.

Table 2: Potential Missense Causal Variants in Prioritized Genes

CHR	BP	rs ID	Gene	AA Change	MAF	SIFT	PolyPhen2	MT	FATHMM	CADD
1	7879401	rs147327372	PER3	Thr519Ala	0.002	T	B	N	T	0.01
1	7890153	rs144178755	PER3	Thr1040Asn	0.001	D	B	N	T	0.962
1	216138793	rs554957414	USH2A	Pro2329Leu	2e-6	D	D	D	D	29.1
1	216373416	rs148135241	USH2A	Ser1122Pro	0.004	D	D	D	D	22.8
1	216419934	rs201527662	USH2A	Cys934Trp	0.0002	D	D	D	D	36
3	33840234	rs200697599	PDCD6IP	Ile5Ser	0.0007	D	D	D	T	32
3	33879764	rs199990824	PDCD6IP	Asp376Asn	4e-6	D	B	D	T	26
3	33905532	rs62620697	PDCD6IP	Ala719Thr	4e-6	T	B	D	T	23.8
3	33905587	rs145293758	PDCD6IP	Pro737Arg	0.001	T	B	N	T	20.2
3	49039984	rs140290144	P4HTM	Ile227Val	0.006	T	B	D	T	22.1
3	49043292	rs144279528	P4HTM	Asp386Asn	8e-5	T	B	D	T	27.3
8	102570910	rs142411476	GRHL2	Arg183Gln	0.0002	T	D	D	T	22
16	75512734	rs140699573	CHST6	Gln331His	4e-6	D	D	D	D	27.4
17	44055786	rs139796158	MAPT	Ala118Gly	6e-5	D	D	D	T	26.4
17	44060807	rs76375268	MAPT	Gly213Arg	0.004	D	D	N	T	11.71
17	44060859	rs63750072	MAPT	Gln230Arg	0.04	D	D	D	T	4.652
17	44067341	rs143956882	MAPT	Ser427Phe	0.001	D	D	D	T	28.5
17	44101481	rs63750191	MAPT	Gln741Lys	3e-5	D	D	D	T	27.5

Legend: The best potential missense causal variants in our top prioritized genes. The headers represent: CHR = chromosome, BP = physical position in basepairs (hg19), Gene = gene location, AA change = amino acid change caused by mutation, MAF = minor allele frequency of the variant obtained from gnomAD, SIFT = pathogenicity prediction from SIFT (where T = tolerated and D = damaging), PolyPhen2 = pathogenicity prediction from PolyPhen2 (where B = benign and D = damaging), MT = pathogenicity prediction from MutationTaster (where N = neutral and D = damaging), FATHMM = pathogenicity prediction from FATHMM (where T = tolerated and D = damaging), CADD = CADD phred score