Please cite this article in press as: Alipanahi et al., Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.05.004

ARTICLE

Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology

Babak Alipanahi,[1,5,*] Farhad Hormozdiari,[2,5] Babak Behsaz,[2,5] Justin Cosentino,[1,5] Zachary R. McCaw,[1,5] Emanuel Schorsch,[1] D. Sculley,[2] Elizabeth H. Dorfman,[1] Paul J. Foster,[3] Lily H. Peng,[1] Sonia Phene,[1] Naama Hammel,[1] Andrew Carroll,[1] Anthony P. Khawaja,[3,4,6] and Cory Y. McLean[2,6,*]

Summary

Genome-wide association studies (GWASs) require accurate cohort phenotyping, but expert labeling can be costly, time intensive, and variable. Here, we develop a machine learning (ML) model to predict glaucomatous optic nerve head features from color fundus photographs. We used the model to predict vertical cup-to-disc ratio (VCDR), a diagnostic parameter and cardinal endophenotype for glaucoma, in 65,680 Europeans in the UK Biobank (UKB). A GWAS of ML-based VCDR identified 299 independent genome-wide significant (GWS; $p \leq 5 \times 10^{-8}$) hits in 156 loci. The ML-based GWAS replicated 62 of 65 GWS loci from a recent VCDR GWAS in the UKB for which two ophthalmologists manually labeled images for 67,040 Europeans. The ML-based GWAS also identified 93 novel loci, significantly expanding our understanding of the genetic etiologies of glaucoma and VCDR. Pathway analyses support the biological significance of the novel hits to VCDR: select loci near genes involved in neuronal and synaptic biology or harboring variants are known to cause severe Mendelian ophthalmic disease. Finally, the ML-based GWAS results significantly improve polygenic prediction of VCDR and primary open-angle glaucoma in the independent EPIC-Norfolk cohort.

Introduction

Genome-wide association studies (GWASs) require accurate phenotyping of large cohorts, but expert phenotyping can be costly and time intensive. On the other hand, self-reported phenotyping, while cost-effective and often insightful,[1] can be inaccurate for nuanced phenotypes such as osteoarthritis[2] or infeasible to obtain for complex quantitative phenotypes. Population-scale biobanks, such as the UK Biobank (UKB)[3] and Biobank Japan,[4] that contain genomics, biomedical data, and health records for hundreds of thousands of individuals provide opportunities to study complex disorders and traits.[5] GWASs of individual blood- and urine-based biomarkers, which can be assayed accurately with high throughput, have shed light on disease etiology.[6,7]

Advances in deep learning have enabled the extraction of medically relevant features from high-dimensional data, such as using cardiac magnetic resonance imaging to infer cardiac and aortic dimensions,[8] color fundus photographs to detect glaucoma risk,[9] and optical coherence tomography images to predict age-related macular degeneration progression.[10] Using medically relevant features extracted from biobank data by machine learning (ML) models as GWAS phenotypes provides an opportunity to identify genetic signals influencing these traits. For example, Glastonbury et al. trained an ML model to predict mean adipocyte areas from histology images and used the predictions to perform a GWAS, doubling the cohort size in comparison to similar studies.[11]

Here, we propose training an ML model to automatically phenotype a large cohort for genomic discovery. The proposed paradigm has two phases: in the "model training" phase, a database of expert-labeled samples (for which genomics data are not required) is used to train and validate a phenotype prediction model (Figure 1A); in the "model application" phase, the model is applied to biobank data to predict phenotypes of interest, which are then analyzed for genomic associations (Figure 1B). This paradigm has several advantages. First, model application is scalable and efficient. Second, a single model can predict multiple phenotypes simultaneously. Third, the model can be applied retrospectively to existing data, resulting in new phenotypes or more accurate predictions for the existing phenotypes. Fourth, multiple lines of evidence can be integrated to predict a single phenotype, which would be prohibitively expensive if performed manually.

As a proof of concept, we investigate predicting glaucoma-related features from fundus images and performing genomic discovery on the predicted features. Glaucoma is an optic neuropathy that results from progressive retinal ganglion cell degeneration[12] and is the leading cause of irreversible blindness globally,[13] affecting more than 80 million people worldwide.[14] Moreover, glaucoma is one

[1]Google Health, Palo Alto, CA 94304, USA; [2]Google Health, Cambridge, MA 02142, USA; [3]NIHR Biomedical Research Centre at Moorfields Eye Hospital and UCL Institute of Ophthalmology, London EC1V 9EL, UK; [4]MRC Epidemiology Unit, University of Cambridge, Cambridge CB2 0SL, UK
[5]These authors contributed equally
[6]These authors contributed equally
*Correspondence: cym@google.com (C.Y.M.), babaka@google.com (B.A.)
https://doi.org/10.1016/j.ajhg.2021.05.004.

Please cite this article in press as: Alipanahi et al., Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.05.004

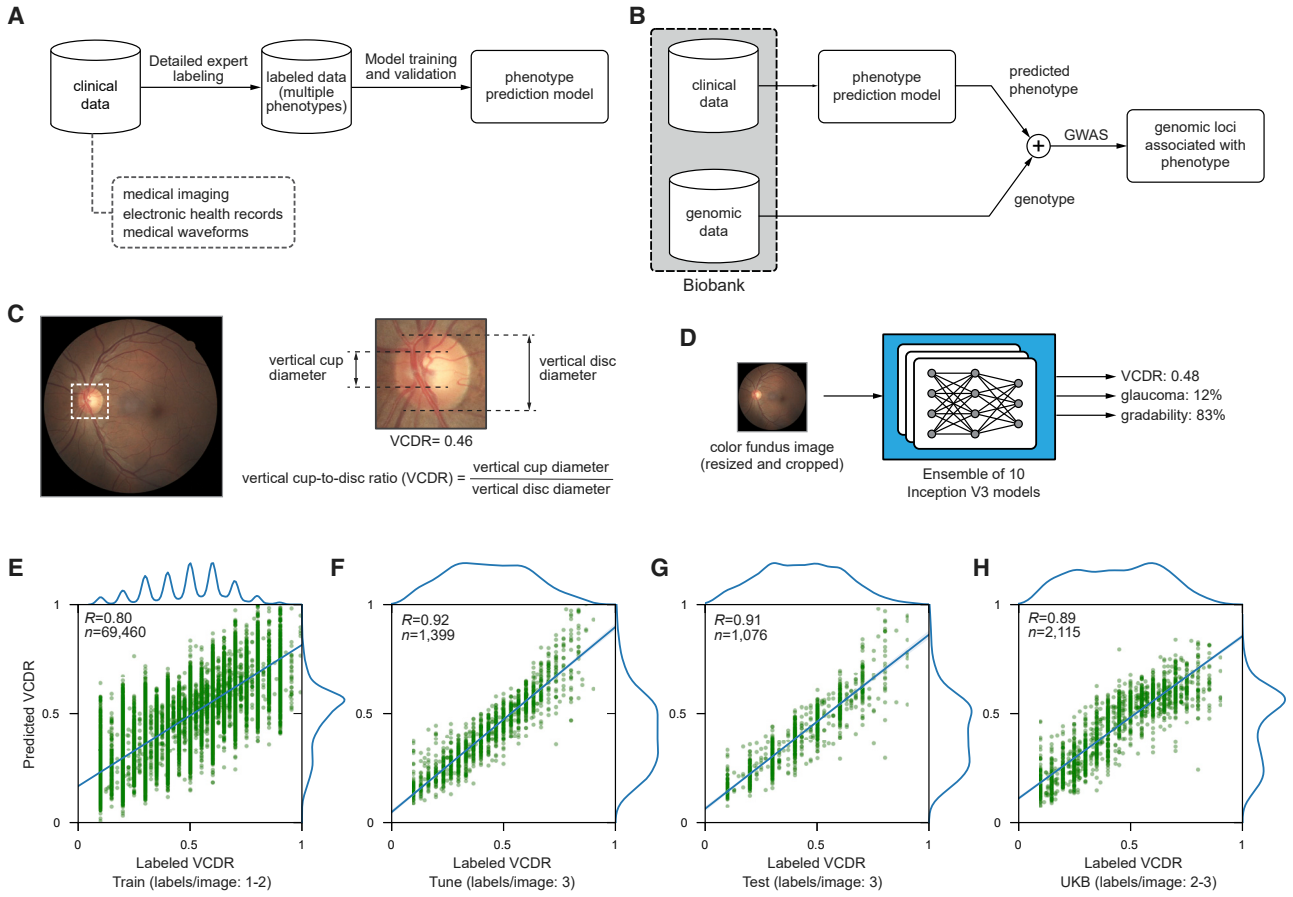

Figure 1. ML-based phenotyping concept and its application to VCDR
(A) "Model training" phase in which a phenotype prediction model is trained with expert-labeled data.
(B) "Model application" phase in which the validated phenotype prediction model is applied to new, unlabeled data followed by genomic discovery.
(C) Definition of vertical cup-to-disc ratio (VCDR) in a real fundus image.
(D) Schematic of the multi-task ensemble model used in phenotype prediction.
(E–H) Scatterplots of the ML-based VCDR versus expert-labeled VCDR values for the train (E), tune (F), test (G), and UK Biobank (H) datasets. Number of grades per image is shown in parentheses.

of the most heritable common human diseases, with heritability estimates of 70%,[15] and there is evidence for effective genomic risk prediction.[16,17]

The hallmark diagnostic feature of glaucoma is optic disc cupping.[12] The vertical cup-to-disc ratio (VCDR; Figure 1C), a quantitative indicator for optic nerve head morphology and a frequently reported quantitative measure of cupping, is an important endophenotype of glaucoma.[18–21] With the advent of very large biobank studies and routine retinal imaging in community optometric practices, there is huge potential for furthering our understanding of glaucoma through population-level analysis of VCDR. However, human grading of optic disc images to ascertain VCDR is costly and extremely resource intensive at large scale because it requires expert knowledge and deciphering the optic cup margin is challenging.

Here, we developed an ML model using 81,830 non-UKB, ophthalmologist-labeled fundus images to predict image gradability, VCDR, and referable glaucoma risk. We used the model to predict VCDR in 65,680 UKB participants of European ancestry from 175,337 fundus images. We then performed a GWAS on the ML-based VCDR phenotype (hereafter, "ML-based GWAS") and compared the results to prior VCDR GWASs, including a recent VCDR GWAS using phenotypes derived from expert-labeled UKB fundus images.[17] We show that ML-based phenotypes are accurate and substantially more efficient to obtain than expert-phenotyped VCDR measurements, identify novel genetic associations with plausible links to known VCDR biology, and produce more accurate polygenic risk scores for predicting VCDR in an independent population.

Methods

Model training and validation

We followed the procedure described previously by Phene et al.,[9] modifying only to remove all UKB images. Briefly, we used 81,830 color fundus images from AREDS,[22] EyePACS (see web

resources), Inoveon (see web resources) from the United States, and two eye hospitals in India (Narayana Nethralaya and Sankara Nethralaya). Ethics review and institutional review board exemption was obtained via Quorum Review Institutional Review Board. We trained ten independent multi-task Inception v3[23] deep convolutional neural networks on the fundus images, using weights learned from the Image Net dataset[24] as pre-trained weights for the convolutional layers. For each of the ten models, a different random seed, which randomly changes the ordering of the training data and selection of mini-batches, the random initialization of the last layers of the neural network, and random image augmentation and dropout patterns, was used. Furthermore, we performed image augmentation[25] and early stopping[26] based on mean squared error (MSE) for predicting VCDR on the tune dataset for picking the best model. The final prediction model is the average prediction of the ten models in the ensemble.

Phenotype calling in the UK Biobank cohort

We included UKB participants with color fundus images. After making predictions for 175,337 images, 21,400 were predicted to be ungradable and were removed. Individual-level VCDR values were computed as the average per-eye VCDR within a single visit, with preference for the initial visit (supplemental information).

Genome-wide association study

We used BOLT-LMM v.2.3.4[27,28] to examine associations between genotype and ML-based VCDR in European individuals in UKB by using the –lmm parameter to compute the Bayesian mixed model statistics. We used all genotyped variants with minor allele frequency > 0.001 to perform model fitting and heritability estimation. We performed rank-based inverse normal (INT) transformation to the ML-based VCDR phenotype to increase the power for association discovery.[29] Finally, in our association study, we used sex, age at visit, visit number (i.e., 1 or 2 to indicate visit 1 or visit 2), number of eyes used to compute VCDR, genotyping array indicator, refractive error, average gradability scores of all fundus images included for each participant, and the top 15 genetic principal components as covariates.

Detecting independent genome-wide significant loci

Genome-wide significant (GWS; $p \leq 5 \times 10^{-8}$) lead SNPs, independent at $R^2 = 0.1$, were identified via PLINK's –clump command (see web resources). The reference panel for linkage disequilibrium (LD) calculation contained 10,000 unrelated subjects of European ancestry from the UKB. Loci were formed around lead SNPs on the basis of the span of reference panel SNPs in LD with the lead SNPs at $R^2 \geq 0.1$. Loci separated by fewer than 250 kb were subsequently merged.

SNP-heritability estimates for ML-based VCDR

We computed the SNP heritability for ML-based VCDR by applying stratified LD score regression[30] on the VCDR GWAS summary statistics while using the 75 baseline LD annotations provided by S-LDSC authors (see web resources).

Replication of existing loci

Loci for ML-based VCDR and comparator studies were formed as described above, and the common reference panel of 10,000 randomly selected unrelated subjects from the UKB. Replication was assessed via the proportion of ML-based VCDR loci that overlapped with comparators and the proportion of comparator loci that overlapped with the ML-based VCDR loci. Thus, replication required that both studies had a GWS variant within a common genomic region, although not necessarily the same variant. Loci reaching GWS in the ML-based VCDR but not identified in any comparator GWASs of VCDR analyzed here are hereafter referred to as "novel loci."

Mendelian randomization and mediation analyses

We performed two sample Mendelian randomization analysis, implemented via TwoSampleMR (see web resources), to examine the causal association between intraocular pressure (IOP), as assessed by Khawaja et al.,[16] and ML-based VCDR. Per-SNP associations were meta-analyzed via Egger regression.[31]

We performed mediation analysis to estimate the association between ML-based VCDR and glaucoma, as assessed by Gharahkhani et al.[32] Mendelian randomization is in fact a special case of mediation analysis in which the instrumental variables (here, SNPs) have no effect on the outcome (here, glaucoma) other than through the mediator (here, ML-based VCDR). Our mediation analysis differs from Mendelian randomization in that, because limited availability of summary statistics from Gharahkhani et al., the SNP set was defined on the basis association with the mediator (ML-based VCDR) rather than the outcome (glaucoma). Among the 118 independent, significant glaucoma SNPs identified by Gharahkhani et al., 116 remained after harmonizing with VCDR. To account for probable direct effects of the candidate SNPs on glaucoma odds, for example via IOP, we again meta-analyzed the per-SNP associations via Egger regression.

VCDR polygenic risk score

We developed two polygenic risk scores (PRSs) by using the pruning and thresholding (P+T)[33] and elastic net[34] methods. The UKB test cohort was graded with the same guidelines used in grading other datasets used in this study. The HRT-derived VCDR was examined and, for participants with good quality scans in both eyes, the mean value of right and left eyes was considered, as previously described.[35] Genotyping was carried out on the Affymetrix UK Biobank Axiom array, as previously described.[36]

In the P+T model, we used a set of variants common to the UKB and EPIC-Norfolk cohorts. EPIC-Norfolk's imputation was performed with the HRC v.1 panel and excludes indels;[37] thus, to harmonize the variants, we filtered out variants from Craig et al. and our ML-based GWAS not present in EPIC-Norfolk. This resulted in 58 variants from the 76 reported variants from the Craig et al. GWAS (i.e., 18 variants were dropped) and 282 of the 299 variants from our ML-based GWAS (i.e., 17 fewer variants).

In the elastic model, we used the ML-predicted VCDR as the target label from the 62,969 UKB training samples to train the elastic model. For Craig et al., we used 76 variants that included the 58 variants from the P+T model and 18 additional proxy variants that are in high LD ($R^2 \geq 0.6$) with the 18 variants dropped from the Craig et al. P+T model. The same set of 282 variants used in P+T was used for the ML-based model. We performed 5-fold cross validation and used the L1-penalty ratios of [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0].

Glaucoma liability conditional analysis

We defined glaucoma risk liability as the logit transform of the highest-level of ML-based glaucoma probability ("likely glaucoma"; supplemental information) as

$$g = \log\left(\frac{p}{1-p}\right),$$

where p and g denote ML-based glaucoma risk probability and liability, respectively. We performed conditional analysis on ML-based glaucoma risk liabilities by using BOLT-LMM conditional on ML-based VCDR. In this conditional analysis, we additionally adjusted for the same covariates used in the primary ML-based VCDR GWAS.

Glaucoma subtypes prediction in the EPIC-Norfolk cohort

We analyzed 5,868 participants from the EPIC-Norfolk Eye Study cohort who were genotyped via the Affymetrix UK Biobank Axiom array, met inclusion criteria and quality control, and had scanning laser ophthalmoscopy VCDR measurements (supplemental information). Included participants had a mean age of 68 years (SD = 7.7, range 48–90), 55% were women, and the mean VCDR was 0.34 (SD = 0.23). Of the 5,868 samples, 175 were classified as primary open-angle glaucoma (POAG) cases (see supplemental information for detailed POAG criteria), of which 98 were classified as high tension glaucoma (HTG; IOP > 21 mmHg) and 77 as normal tension glaucoma (NTG; IOP ≤ 21 mmHg) on the basis of the corneal-compensated IOP at the Eye Study assessment. Pre-treatment IOP was imputed by dividing by 0.7 for participants using glaucoma medication at the time of assessment, as previously described.[16]

We extracted age, sex, POAG status, NTG status, and HTG status from all 5,868 samples. We fitted independent logistic regression models to predict POAG, HTG, and NTG statuses by using VCDR PRS, age, and sex as predictors. We considered both the ML-based elastic net VCDR PRS and the Craig et al. elastic net PRS described above.

Results

Overview of the ML-based phenotyping method

We used 81,830 fundus images graded by a panel of experts that passed our labeling guideline assessment (supplemental information) to train a phenotype prediction model that jointly predicts image gradability, VCDR, and referable glaucoma risk (Figure 1D). We split these images into "train," "tune," and "test" sets; training images were graded by one to two eye care providers with varied expertise, while images in the two latter sets were each graded by three glaucoma specialist experts. We benchmarked model performance on all data splits (Figures 1E–1G; Table S1). On the test set of 1,076 test images, the model achieved a Pearson's correlation of $R = 0.91$ between predicted and graded VCDR (95% confidence interval [CI] = 0.90–0.92) and root mean square error (RMSE) of 0.079 (95% CI = 0.074–0.085). Additionally, we validated model generalizability on 2,115 UKB fundus images each graded by two to three experts (hereafter, "UKB test set"), which achieved similar predictive performance to the test set (Figure 1H; $R = 0.89$, 95% CI = 0.88–0.90; RMSE = 0.092, 95% CI = 0.088–0.096; Table S1). We also validated that the model generalizes across ancestries in a larger set of 4,816 UKB fundus images with at least one manual grade (Figure S1).

ML-based GWAS replicates a manual phenotyping VCDR GWAS and discovers 93 additional novel loci

We applied the VCDR prediction model to the entire set of 175,337 UKB fundus images. Most images were either predicted to be easily gradable (predicted gradability > 0.9) or completely unusable (predicted gradability < 0.2) (Figure S2). We classified all 21,400 images with predicted gradability < 0.7 as "ungradable." Manual inspection of 100 randomly selected ungradable images showed they were typically completely dark, bleached white, or extremely out of focus. After removing the 21,400 ungradable images, aggregating predicted VCDR values across left and right eyes and the first and second visits for each individual, subsetting the cohort to individuals of European ancestry, and performing cohort quality control, a cohort of 65,680 individuals with VCDR phenotype remained for further analysis (supplemental information, Figures S3 and S4). To control for confounding factors (e.g., population structure) and increase power, we added age at the time of visit, sex, average image gradability, number of fundus images used in VCDR calculation, normalized refractive error, genotyping array type, and the top 15 genetic principal components as covariates.

We performed the ML-based GWAS by using BOLT-LMM (supplemental information). While genomic inflation Δ_{GC} was 1.20 (Figure S5), the stratified LD score regression-based (S-LDSC) intercept[30] was 1.06 (SEM = 0.02), indicating that most test statistic inflation can be attributed to polygenicity rather than population structure. The SNP-based heritability in the ML-based GWAS was 0.43 (SEM = 0.03), a majority of the 56% heritability estimated for VCDR by twin and family-based studies (Asefa et al., 2019)[38]. The ML-based GWAS identified 299 independent genome-wide significant (GWS) hits ($R^2 \leq 0.1$, $p \leq 5 \times 10^{-8}$) at 156 independent GWS loci after merging hits within 250 kb together (Figure 2A, Tables S2 and S3). Based on sum of single effects regression,[39] the number of causal variants within the 156 independent GWS loci was conservatively estimated at 813 (supplemental information; Tables S4 and S5).

To understand the influence of training dataset size on model performance and GWAS results, we retrained the ML model with as little as 10% of the full training set. Performance curves indicate that using fewer than 8,000 training images achieved a Pearson's correlation $R = 0.83$ (95% CI = 0.81–0.84) on the UKB test set, identified 131 GWS loci, and replicated 123 of the 156 loci identified in the full model (Figures S6 and S7). An analysis of the implications of phenotyping accuracy on genomic discovery suggested that the difference in power for the model trained with 10% of the training data and the model trained with all data would maximally reach 15% (Figure S8).

Next, we compared the ML-based GWAS results with those from the two largest existing VCDR GWASs. First, we compared with the VCDR meta-analysis from the International Glaucoma Genetics Consortium (IGGC) in

Please cite this article in press as: Alipanahi et al., Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.05.004

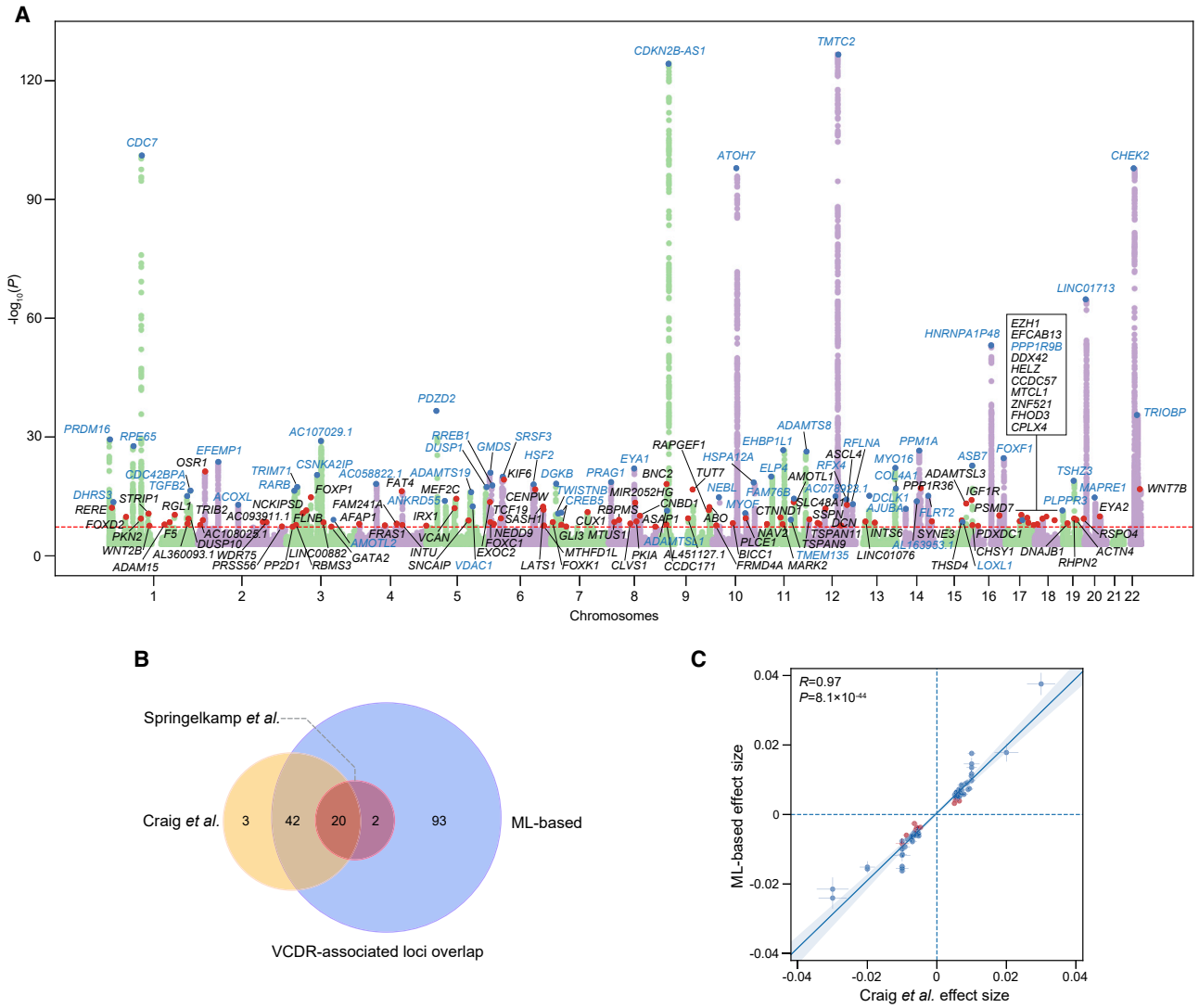

Figure 2. ML-based VCDR GWAS results and comparison to known associations

(A) Manhattan plot depicting ML-based VCDR-associated GWAS p values from the BOLT-LMM analysis. There are 156 GWS (genome-wide significant) loci, representing 299 independent ($R^2 = 0.1$) GWS hits. For each locus, the closest gene is shown. Blue gene names and dots indicate loci also identified in the Craig et al. study[17] and red dots and black gene names indicate novel loci. The dashed red line denotes the GWS p value, 5×10^{-8}.

(B) Venn diagram of loci overlap for three VCDR GWASs. ML-based GWAS replicates all 22 loci of the IGGC VCDR meta-analysis[20] and 62 of 65 loci identified by Craig et al.,[17] while discovering 93 novel loci (supplemental information).

(C) Effect sizes for the 73 GWS hits shared by the Craig et al.[17] and ML-based VCDR GWAS. The three Craig et al. hits not included failed the ML-based GWAS QC (rs61952219 for low imputation quality and rs7039467 and rs146055611 for violating Hardy-Weinberg equilibrium). Blue and red dots denote the SNP's being more significant in the ML-based and Craig et al. GWAS, respectively. Error bars depict standard errors. The banding in Craig et al. effect sizes is due to large effect sizes' being reported in multiples of 0.01. The blue line is the best fit line and the shaded area shows the 95% confidence interval.

23,899 Europeans[20] for which all summary statistics are publicly available (see web resources). The ML-based GWAS replicated all 22 GWS loci and exhibited strong genetic correlation (0.95, SEM = 0.03, $p = 2.1 \times 10^{-167}$) with the IGGC GWAS (Figure 2B, Table 1), and effect size regression analysis showed a slope significantly different from zero (slope = 0.983, SEM = 0.041, $p = 1 \times 10^{-61}$) and indistinguishable from one ($p = 0.67$; Figure S9; supplemental information). Second, we compared with a GWAS on 67,040 manually phenotyped UKB fundus images[17] for which only the independent genome-wide significant SNPs are publicly available. The ML-based GWAS replicated 62 out of 65 GWS loci with very similar estimated effect sizes (Figures 2B and 2C, Table 1) and more significant p values (Figure S10). The p values and effect sizes of the novel loci are shown in Figure S11. The three loci not replicated at the GWS level in the ML-based GWAS were all Bonferroni-replicated (adjusting for 65 tests), and p values ranged from 5.5×10^{-8} to 6.6×10^{-5}. Third, we compared our results with a meta-analysis of the Craig et al. and IGGC VCDR GWASs.[17] The ML-based GWAS replicated 82 of the 90 loci at GWS level, and the remaining eight

Please cite this article in press as: Alipanahi et al., Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.05.004

Table 1. Replicated loci of ML-based VCDR GWASs and meta-analysis at GWS level

Discovery GWAS details					
Study (phenotype)	Number of participants	Loci	Number of loci replicated in ML-basedVCDR GWAS	Number of loci replicated in ML-based + IGGCVCDR GWAS	S-LDSC-based genetic correlation with ML-based VCDR
ML-based (VCDR)	65,680	156	–	151	–
ML-based 10% (VCDR)	65,044	131	123	125	0.99 (2.1×10^{-3})
ML-based + IGGC[20] (VCDR)	89,579	189	151	–	0.97 (2.6×10^{-3})
IGGC[20] (VCDR)	23,899	22	22	22	0.95 (0.03)
Craig et al.[17] (VCDR)	67,040	65	62	63	N/A
Craig et al.[17] + IGGC[20] (VCDR)	90,939	90	82	85	N/A
Khawaja et al.[16] (IOP)	139,555	107	14	22	0.19 (0.02)
Gharahkhani et al.[32] (POAG)	383,500	118	32	40	N/A

"ML-based 10% (VCDR)" denotes the GWAS performed on VCDR predictions of the ML model trained with only 10% of the training data. "ML-based + IGGC (VCDR)" denotes meta-analysis of ML-based and IGGC VCDR GWAS. Likewise, "Craig et al. + IGGC (VCDR)" denotes meta-analysis of Craig et al. VCDR and IGGC VCDR GWAS. Genetic correlation was only computed when the full set of summary statistics were available.

loci were Bonferroni-replicated with and had p values ranging from 1.4×10^{-7} to 6.6×10^{-5} (Table 1).

Finally, we performed a meta-analysis of our ML-based GWAS with the IGGC VCDR GWAS, which resulted in 189 GWS loci (supplemental information; Table 1 and Tables S6 and S7). This ML-based meta-analysis replicated 63 out of 65 of Craig et al.'s discovery GWAS and 85 out of 90 Craig et al.'s meta-analysis at GWS level (Table 1). Taken together, these comparisons demonstrate that the ML-based GWAS accurately identifies known VCDR associations and additionally identifies over 90 novel loci (Figure 2B, Table S8), substantially increasing our understanding of the genetic underpinnings of this complex trait.

To assess the biological plausibility of the novel loci identified in the ML-based GWAS, we compared gene set enrichment analyses of the 156 ML-based loci to those of the 65 Craig et al. loci by using FUMA.[40] Nine eye-related gene sets were significantly enriched in both sets of loci. The enrichment odds ratios (ML-based enrichment over Craig et al. enrichment) of all nine gene sets were greater than one, suggesting improved identification of functionally relevant pathways in the ML-based loci (Figure S12). To assess effects of distal *cis*-regulatory interactions, we also performed enrichment analyses of the 156 ML-based loci and the 65 Craig et al. loci by using GREAT.[41] Consistent with the FUMA results, the ML-based loci were more significantly enriched than the Craig et al. loci across all tested ontologies (Figure S13). The ML-based loci were significantly enriched for 22 gene sets, the majority of which are developmental and seven of which are eye related (Table S9). In contrast, the Craig et al. loci were significantly enriched for only three gene sets; two of these are eye-related sets that were also enriched in the ML-based results (Table S9).

Lastly, we performed a phenome-wide association study (PheWAS) over all 299 independent GWS hits by using OpenTargets (web resources). OpenTargets reported 62,753 (variant, phenotype) pairs that were nominally significant ($p \leq 0.05$); after Bonferroni correction, 974 pairs were significant (supplemental information). We observed that 314 of the 974 significant pairs belonged to the "anthropometric measurement" trait category, while the "eye measurement" category had 101 pairs (Table S10).

Biological significance of select novel VCDR-associated loci

Several of the VCDR-associated loci discovered in this study are known to be associated with intraocular pressure (IOP), including rs1361108 near *CENPW*,[42] rs2570981 in *SNCAIP*,[42] rs6999835 near *PKIA*,[16] and rs351364 in *WNT2B*.[16] This suggests that a proportion of the genetic variation in VCDR is mediated via IOP and pathophysiological processes affecting the anterior segment of the eye, consistent with IOP's being a strong risk factor for glaucoma.[43] Indeed, we observed that 13% (14 of 107) of the GWS loci from the latest IOP meta-analysis[16] were GWS in the ML-based VCDR GWAS. In addition, the overall genetic correlation between our ML-based VCDR GWAS and the IOP GWAS meta-analysis is 0.19 (SEM = 0.02, $p = 5.5 \times 10^{-15}$), indicating that VCDR is partially explained by IOP. Moreover, a Mendelian randomization (MR) analysis followed by Egger regression[31] suggests that IOP has a strong directional association with ML-based VCDR: the regression intercept does not differ significantly from zero (intercept = 0.001, SE = 0.002, p = 0.7), but the slope does (slope = 0.072, SE = 0.020, $p = 4 \times 10^{-4}$). The reverse analysis provided no evidence for a directional association between ML-based VCDR and IOP (supplemental information; Figure S14).

VCDR is an objective quantification of the proportion of neuronal tissue at the head of the optic nerve (Figure 1C). Interestingly, several VCDR-associated loci discovered in this study encompass genes involved in neuronal and synaptic biology, and thus may influence VCDR via direct effects on the retina and optic nerve rather than via IOP.

NCKIPSD (rs7633840) is involved in the formation and maintenance of dendritic spines, and modulates synaptic activity in neurons.[44] *CPLX4* (rs77759734) is required for the maintenance of synaptic ultrastructure in the adult retina.[45] *MARK2* (rs199826712) has roles in neuronal cell polarity and the regulation of neuronal migration.[46] These loci complement additional neuronal loci also discovered by Craig et al.; some notable examples include *MYO16* (rs10162202), *TRIM71* (rs56131903), and *FLRT2* (rs1289426). An increase in VCDR may be due not only to loss of retinal ganglion cell neurons but also loss of neural supporting tissue, such as glial cells. One of our novel VCDR-associated loci is an indel on chromosome 8 (chr8: 131,606,303_CTGTT_C), near *ASAP1*; this locus has been associated with glioma,[47] suggesting glial cells as potential mediators of the VCDR association.

Several genes at the novel VCDR-associated loci harbor mutations that cause severe Mendelian ophthalmic disease. Here, for the first time, we report common variants at these genes that are associated with VCDR variation at a population level. Three of our novel loci are at *ADAMTSL3* (rs59199978), *PITX2* (rs2661764), and *FOXC1* (rs2745572), all of which are associated with syndromic ocular anterior segment dysgenesis, which in turn causes raised IOP and secondary glaucoma. *ADAMTSL3* is an important paralog of *ADAMTSL1*—which itself is also associated with VCDR in our GWAS. A mutation in *ADAMTSL1* has been reported to cause inherited anterior segment dysgenesis and secondary congenital glaucoma.[48] Mutations in *PITX2* and *FOXC1* cause Axenfeld-Rieger syndrome.[49] Common variants at these loci may mark more subtle effects on ocular anterior segment development, resulting in subclinical changes in IOP and VCDR that are apparent on a population level. While *FOXC1* variants have been previously associated with glaucoma,[50] this is the first time they have been associated with population variation in VCDR. Mutations in *PRSS56*, a gene at one of our novel VCDR-associated loci, cause microphthalmia in humans.[51] Another two of our VCDR-associated loci are at *EYA1* and *EYA2* (eyes absent homologs 1 and 2), genes that are important for eye development in *Drosophila*. *EYA1* has been implicated in ocular anterior segment anomalies and cataract.[52] We also replicate some of the loci identified by Craig et al., such as *ELP4*, which has been associated with aniridia,[53] a condition characterized by the absence of an iris and that can predispose patients to glaucoma.[53,54]

ML-based GWAS improves VCDR polygenic risk scores

We developed P+T and elastic net PRSs for both the ML-based VCDR GWAS and the Craig et al. GWAS (Tables S11–S14). These PRSs were evaluated in two test sets: a holdout set of 2,076 subjects from UKB with VCDR measured by two to three experts and a set of 5,868 subjects from the European Prospective Investigation into Cancer Norfolk (EPIC-Norfolk) cohort with VCDR measured by scanning laser ophthalmoscopy (HRT).[55] Because the EPIC-Norfolk imputation was done with the HRC v.1 (Haplotype Reference Consortium) panel, which excludes indels,[37] we subset the ML-based GWAS summary statistics to HRC v.1.

For the P+T model, subsetting to HRC v.1 results in 282 hits, down from 299 original hits. With the effect sizes from the ML-based GWAS (Table S11), this model achieves a Pearson's correlation $R = 0.37$ (95% CI = 0.33–40) in the UKB adjudicated cohort. The P+T model from the Craig et al. GWAS does not include 18 out of 76 SNPs (absent in HRC v.1) and achieves a Pearson's correlation $R = 0.29$ (95% CI = 0.25–0.33). The performance metrics of the ML-based Craig et al. P+T models when not subset to HRC v.1 are shown in Figure S15. Performance in the EPIC-Norfolk set was slightly lower, but the P+T model still explained 9.6% of the total variance (Figure 3A). In both sets, the ML-based P+T model outperformed the Craig et al. P+T model (UKB: $\Delta R = 0.079$, $p < 0.031$, $n = 2{,}076$; EPIC: $\Delta R = 0.082$, $p < 5.9 \times 10^{-4}$, $n = 5{,}868$, permutation test).

We then used the ML-based VCDR values from UKB to train elastic net models; after removing all images used in building the adjudicated test set, the training set contained 62,969 samples. In contrast to the P+T model in which GWAS marginal effect sizes are used as PRS weights, elastic net jointly learns all weights in a supervised manner. To make up for the 18 missing Craig et al. SNPs, we identified LD-based proxies for all of the missing hits in HRC v.1 and included them in training the elastic net model. The ML-based elastic net model (Table S12) numerically improved upon the P+T model in both UKB ($R = 0.38$, 95% CI = 0.34–0.41) and EPIC ($R = 0.33$, 95% CI = 0.30–0.35) sets (Figure 3B). The elastic net model explains 14.2% and 10.6% of total VCDR variation in the UKB and EPIC-Norfolk sets, respectively. The Craig et al. elastic net model has a more pronounced improvement—probably because of the addition of proxy SNPs—but the ML-based model still significantly outperforms it (UKB: $\Delta R = 0.064$, $p < 9.6 \times 10^{-3}$, $n = 2{,}076$; EPIC: $\Delta R = 0.053$, $p < 6.8 \times 10^{-4}$, $n = 5{,}868$, permutation test).

Relationship of primary open-angle glaucoma and VCDR

To study the relationship between primary open-angle glaucoma (POAG) and VCDR, we defined POAG status in UKB by using a combination of self-report and hospital episode International Classification of Diseases 9/10 codes (supplemental information). ML-based VCDR has moderate predictive power for POAG with an area under the ROC curve (AUC) of 0.76 (n = 65,193, 95% CI = 0.74–0.78, POAG prevalence = 1.9%) and area under the precision-recall curve (AUPRC) of 0.14 (95% CI = 0.12–0.16). After binning individuals by ML-based VCDR, we computed odds ratios (ORs) in each bin versus the bottom bin (Figure 4A). The most extreme bin (VCDR > 0.7, n = 385), which corresponds to a diagnostic criterion for glaucoma,[18] has an OR of 74.3 (95% CI = 57.0–94.3) versus the bottom bin (VCDR < 0.3, n = 30,752).

Please cite this article in press as: Alipanahi et al., Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology, The American Journal of Human Genetics (2021), https://doi.org/10.1016/j.ajhg.2021.05.004

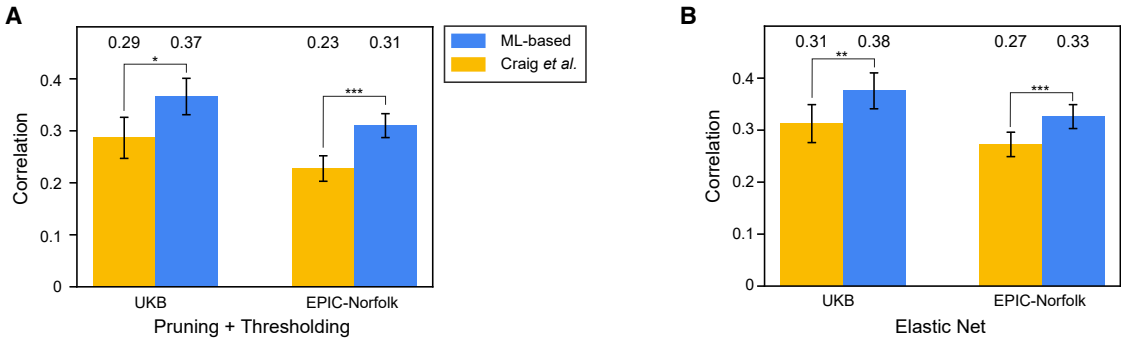

Figure 3. VCDR polygenic risk score performance metrics

(A and B) Pearson's correlations between measured VCDR values and predictions of the pruning and thresholding (P+T) (A) and the elastic net models (B) are shown for the PRS learned from ML-based and Craig et al.[17] hits. Error bars depict 95% confidence intervals. Numbers above bars are the observed Pearson's correlations. Indications of p value ranges (permutation test): *p ≤ 0.05, **p ≤ 0.01, ***p ≤ 0.001. The Craig et al. P+T model uses 58 out of 76 hits. Measured VCDR values were obtained from adjudicated expert labeling of fundus images (UKB, n = 2,076) and scanning laser ophthalmoscopy (HRT) (EPIC-Norfolk, n = 5,868).

We then performed mediation analysis (MA) to study the association of VCDR with glaucoma. Similar to MR, MA evaluates the association between an intermediary or mediating phenotype (here, VCDR) and an outcome phenotype (here, glaucoma). However, whereas in MR the SNP set is selected on the basis of association with the mediator, because of the limited availability of glaucoma summary statistics from the study by Gharahkhani et al.,[32] the SNP set for MA was selected on the basis of association with the outcome. Because, contrary to MR's exclusion restriction, the included SNPs may have affected glaucoma through a pathway other than VCDR (e.g., IOP), the per-SNP estimates of association were meta-analyzed with Egger regression (Egger et al., 1997[31]), which is robust to this assumption.[56] The Egger slope of 5.7 (SE = 1.8, $p = 3 \times 10^{-3}$) differs significantly from zero, providing evidence that VCDR, as ascertained by our ML-based models, is strongly associated with the odds of glaucoma (Figure S16). We note that the Egger intercept of 0.04 also differs significantly from zero ($p = 7 \times 10^{-7}$), indicating the presence of directional pleiotropy; that is, variants included in the analysis, on average, were associated with an increase in the odds of POAG through a pathway other than VCDR.

As shown above, VCDR is an informative endophenotype for glaucoma, and we hypothesize that its PRS should also be predictive of POAG. Indeed, 32 out of 118 loci previously associated with POAG[32] were significantly associated with ML-based VCDR in this study. We applied the ML-based elastic net model to the UKB individuals of European ancestry that do not have fundus images (n = 98,151) to estimate their genetic VCDR. As expected, this genetic model performs noticeably worse than the model using a direct measurement of the VCDR phenotype (AUC = 0.56, 95% CI = 0.55–0.57, AUPRC = 0.07, 95% CI = 0.066–0.073, n = 98,151, POAG prevalence = 5.5%). Nonetheless, when we binned samples by VCDR elastic net PRS, participants in the highest bin (PRS $Z > 2.5$, n = 567) had a considerably higher POAG prevalence (OR = 3.4, 95% CI = 2.6–4.3; Figure 4B) than those in the lowest bin (PRS $Z < -0.1$, n = 46,136).

In addition to VCDR, the ML model was trained to predict referable glaucoma risk;[9] this model output can be interpreted as the probability a specialist would refer an individual for detailed glaucoma evaluation. Because the model output is a continuous value, we can evaluate the contribution of features other than VCDR to referable glaucoma risk by regressing out the VCDR signal. We computed glaucoma risk liability as the logit transform of the ML-based glaucoma probability, which is highly correlated with ML-based VCDR (Figure 4C, Pearson's $R = 0.91$, n = 65,680, $p < 1 \times 10^{-300}$). While a large VCDR is the cardinal feature of a glaucomatous optic nerve, there are other features that suggest glaucoma that are difficult to quantify (e.g., bayoneting or baring of blood vessels and hemorrhages). To examine the genetic associations with glaucomatous optic disc features other than VCDR, we carried out a GWAS of ML-based glaucoma risk conditioned on ML-based VCDR by using BOLT-LMM. The observed SNP heritability was 0.062 (SEM = 0.013) with genomic inflation of 1.04 and S-LDSC-based intercept of 1.01 (SEM = 9.8×10^{-3}; Figure S17) and the GWAS identified eight GWS loci (Tables S15 and S16). Interestingly, two of these loci, *OCA2-HERC2* (Figure 4D; rs12913832, $p = 2.2 \times 10^{-66}$) and *TYR* (rs1126809, $p = 5.8 \times 10^{-13}$), have been previously associated with macular inner retinal thickness (retinal nerve fiber layer and ganglion cell inner plexiform layer) as derived from UKB optical coherence tomography images.[57] These inner retinal parameters have diagnostic utility for glaucoma that is considered complementary to VCDR and may be particularly efficacious at detecting early glaucoma.[58] Moreover, it is not currently possible to ascertain the thickness of the inner retina from fundus images, which are two-dimensional. Together, this suggests that ML-based phenotyping has the potential to identify glaucoma-related features from fundus images that are complementary to VCDR and not typically gradable by humans.

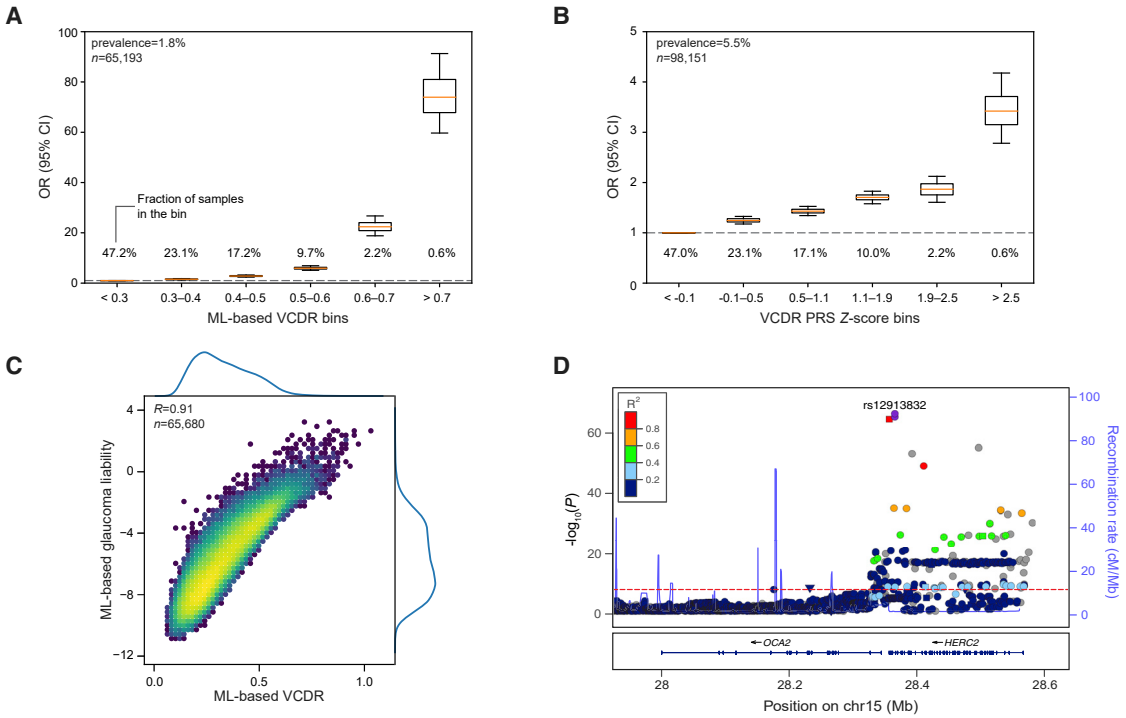

Figure 4. Relationship between glaucoma and VCDR

(A) Glaucoma odds ratios for each ML-based VCDR bin versus the bottom bin is shown. The fraction of individuals in each bin is shown (n = 65,193).

(B) Glaucoma odds ratios for different VCDR elastic net PRS bins versus the bottom bin for individuals with a glaucoma phenotype not used in the GWAS or developing the PRS (n = 98,151). The fractions are selected to match those from (A).

(C) A histogram of ML-based glaucoma liability versus ML-based VCDR (Pearson's correlation $R = 0.91$, n = 65,680, $p < 1 \times 10^{-300}$).

(D) LocusZoom for the strongest associated variant (rs12913832, $p = 2.2 \times 10^{-66}$) in the ML-based glaucoma liability GWAS conditioned on the ML-based VCDR.

Glaucoma prediction in the EPIC-Norfolk cohort

To further assess the utility of the ML-based elastic net VCDR PRS for prediction of glaucoma, we classified the status of EPIC-Norfolk participants (n = 5,868) for POAG (175 cases and 5,693 controls). We additionally sub-categorized POAG cases into HTG (98 cases) and NTG (77 cases). Given the enrichment of the VCDR PRS for variants associated with neuronal development and function, we hypothesized that the PRS would be particularly associated with NTG. We fit a logistic regression model to predict POAG status by using age, sex, and ML-based elastic net VCDR PRS as its three predictors.

The ML-based elastic net VCDR PRS was strikingly associated with POAG, and particularly NTG, in EPIC-Norfolk (Figure 5). The ORs (95% CI) comparing the top risk decile with the bottom decile were 9.7 (3.4–27.6) for POAG, 7.4 (2.2–25.2) for HTG, and 16.5 (2.2–125.9) for NTG (Figure 5). The overall prediction metrics were AUC = 0.74, 95% CI = 0.70–0.77, AUPRC = 0.08, 95% CI = 0.06–0.11, prevalence = 3.0% for POAG; AUC = 0.73, 95% CI = 0.68–0.78, AUPRC = 0.05, 95% CI = 0.03–0.08, prevalence = 1.7% for HTG; and AUC = 0.76, 95% CI = 0.71–0.80, AUPRC = 0.04, 95% CI = 0.03–0.06, prevalence = 1.3% for NTG. The AUC and AUPRC show nominally significant improvements over those from an analogous model using the Craig et al. elastic net VCDR PRS for POAG (ΔAUC = 0.014, 95% CI = 0.0–0.03, p = 0.03; ΔAUPRC = 0.008, 95% CI = 0.0–0.02, p = 0.03, paired bootstrap test) and HTG (ΔAUC = 0.014, 95% CI = 0.0–0.03, p = 0.04; ΔAUPRC = 0.006, 95% CI = 0.0–0.02, p = 0.04, paired bootstrap test).

Discussion

Large cohorts of genotyped and phenotyped individuals have enabled researchers to identify genetic influences of many traits. As methods to ascertain genetic variants in large cohorts continue to improve, we anticipate the major challenge for cohort generation to be accurate and deep phenotyping[59] at scale. Here, we demonstrated that ML-based phenotyping shows promise for improving both scalability to biobank-sized datasets and phenotyping accuracy. We predicted VCDR from all 175,337 UKB fundus images in less than 1 h on a distributed computing system. Multiple lines of evidence indicate that the model-based VCDR predictions improve accuracy over manual labeling, including the reproduction of known VCDR-related biology, identification of plausible novel genetic associations, and generation of polygenic risk scores that better predict VCDR in multiple held-out datasets. Additional

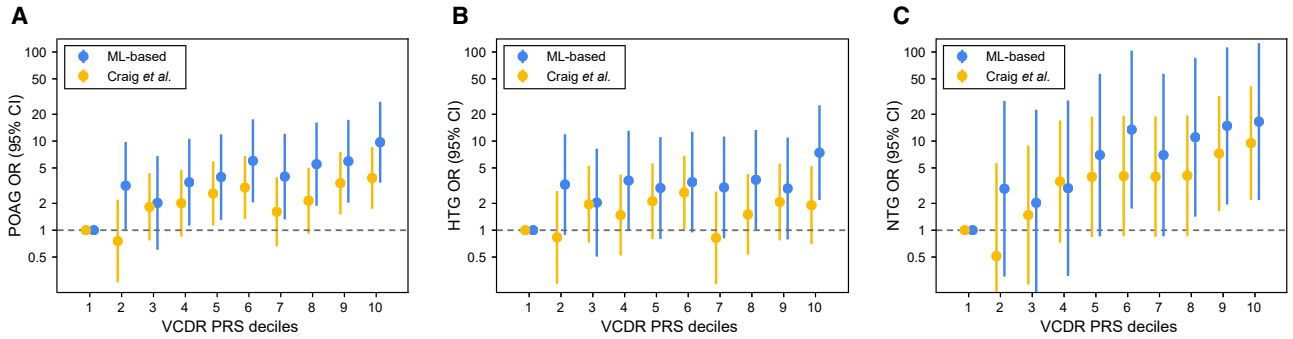

Figure 5. Primary open-angle glaucoma (POAG) prediction in the EPIC-Norfolk cohort
(A–C) Odds ratios and 95% CIs for POAG prevalence by decile of VCDR PRS; reference is decile 1. Results are from logistic regression models adjusted for age and sex for primary open-angle glaucoma (175 cases, 5,693 controls) (A), high-tension glaucoma (HTG; 98 cases, 5,693 controls) (B), and normal-tension glaucoma (NTG; 77 cases, 5,693 controls) (C). Results are presented for the ML-based elastic net VCDR PRS (blue) and the Craig et al.[17] elastic net VCDR PRS (yellow). Note the y axis log scale.

advantages of ML-based phenotyping over manual labeling are improved joint prediction accuracy for multiple correlated phenotypes and predicting liabilities instead of binary labels for binary phenotypes. By regressing out predicted VCDR from the predicted referrable glaucoma risk (i.e., whether the individual should seek further ophthalmologist care), we identified residual referrable risk not attributable to variation in VCDR.

The improvement of our model-based VCDR GWAS over the recent expert-labeled VCDR GWAS by Craig et al. is consistent with improved phenotyping accuracy by our model. The expert labels may include more noise or measurement error than the ML-based labels, as suggested by the inter-grader variability; the inter-grader Pearson's correlation between the two ophthalmologists as reported by Craig et al. for images graded multiple times was 0.75 (95% CI = 0.72–0.77), whereas the ML model achieves a Pearson's correlation of 0.89 between the model predictions and adjudicated expert labels (95% CI = 0.88–0.90). Noise or variability in human grading of VCDR can arise from difficulty in defining the cup-rim border of the optic disc. If the cup-rim border is sloping, rather than having vertical edges, defining it is challenging via two-dimensional images. In this situation, the average VCDR of multiple graders may be considered more accurate than a single grader's score. Our ML-based model was trained and tuned on images that were assessed by multiple graders and may therefore be expected to outperform a single human grader, on average.

The 93 novel VCDR-associated loci discovered by ML-based phenotyping substantially expand our knowledge of the biological processes underlying optic nerve head morphology. While elevated IOP is an established cause of glaucoma,[43] characterized by a pathologically enlarged VCDR, our results support the role of IOP's contributing to variation in VCDR within the healthy range as well. Of particular note were common VCDR-associated variants in genes harboring mutations that cause inherited anterior segment dysgenesis that is well characterized phenotypically. Our findings suggest these dysgenesis processes may also occur at subclinical levels and contribute to variation in the complex VCDR phenotype. Understanding the genotype-phenotype link in rare single-gene disorders can therefore improve our knowledge of some of the many contributory causes to complex traits. Our results also support an important role of neuronal development processes for VCDR. It remains uncertain whether these processes primarily influence VCDR during optic nerve development in early life, thereby reflecting population variation in baseline optic nerve head anatomy, or act later in life and reflect a pathological, glaucomatous change in VCDR over time. Interestingly, genes involved in developmental processes more broadly, including development of the cardiovascular and urogenital systems, were significantly enriched in our results (Table S8). This may suggest early life processes are a major determinant of VCDR variation in adult populations.

This study also showed that a substantial proportion of VCDR variation can be predicted with a polygenic risk score. Improving VCDR prediction produces a concomitant improvement in glaucoma prediction, as we demonstrated by stratifying glaucoma prevalence by using the VCDR PRS. While the UK National Screening Committee does not currently recommend population screening for glaucoma because tests lack sufficient positive predictive value,[60] using polygenic prediction to identify subsets of the general population that are at risk for glaucoma may enable effective screening. Notably, we identified a substantially higher POAG prevalence in the top decile of VCDR PRSs and it may be that current screening tests would have sufficient positive predictive value if applied to this enriched population subset. Earlier detection and treatment of glaucoma, a disease that causes progressive and irreversible vision loss, is a key strategy outlined by the World Health Organization for the prevention of blindness worldwide.[61]

While this study demonstrates the potential for ML-based phenotyping to expand our understanding of the genetic variation underlying complex traits, the method has important limitations that must be taken into

account. Application of this technique relies on the trained model's producing accurate predictions in the genomic discovery set. Here, we showed strong generalizability of the model trained on non-UKB fundus images to the UKB fundus images used for genomic discovery by manually labeling a small subset of UKB fundus images and validating model predictions against these ground truth labels. Application to other phenotypes derived from fundus images, or other data modalities such as optical coherence tomography or magnetic resonance imaging, would require similar demonstrations of model generalizability. Additionally, the initial model training can be costly and time intensive, as it requires manual labeling to be performed. While our ablation analysis showed that training on only 10% of the data still identified the majority of VCDR-associated loci, model performance did not appear to saturate even at the full training set size. Ongoing improvements to transfer learning may reduce future labeled data requirements,[62] although the ability to extrapolate consumer imaging improvements to biomedical imaging is unclear.[63]

Another limitation of our study was the absence of data for absolute vertical disc diameter (VDD), a commonly used proxy for disc size. While VDD is a heritable trait[64] that would be of interest given its correlation with VCDR, considerable challenges preclude extending ML-based phenotyping to VDD in our study. Because VDD is an absolute size measurement, it requires strict standardization of image acquisition. In particular, differences in absolute size measurements from images arise secondary to camera-related magnification and from ocular refraction, mostly determined by the length of the eye.[65] Since our training images were derived from multiple centers and multiple different cameras that were not standardized in terms of magnification and zoom, it is not possible to derive an accurate VDD on which to train an algorithm. Even within UKB, accurately measuring VDD from fundus images is not possible because there are no measurements of axial length. Correcting for magnification with spherical equivalent only corrects for about 30% of eye size-related magnification artifact, whereas axial length correction can account for nearly 100% of the variation.[65] Consequently, we cannot exclude the possibility that some loci discovered in this study would not reach genome-wide significance in a GWAS adjusted for VDD. However, the similar effect sizes estimated for loci significant both in our study and in Craig et al., and the increased number of loci discovered in an independent ML-based GWAS of VDD-adjusted VCDR in the UKB,[66] suggest that many of the loci discovered here influence VCDR independently of VDD.

In summary, we have proposed a method for performing genomic discovery on biobank-scale datasets by using machine learning algorithms for accurate phenotyping. A key benefit of the method is its ability to use a modest-sized biomedical dataset annotated with reasonable accuracy to train a model that identifies the underlying patterns and yields usable predictions. Extending the method to additional phenotypes and data modalities in large-scale biobanks could further expand our understanding of disease etiology and improve genetic risk modeling.

Data and code availability

Code and detailed instructions for model training, prediction, and analysis, as well as instructions for evaluating the trained model on fundus images, are available at https://github.com/Google-Health/genomics-research/tree/main/ml-based-vcdr. The UKB data are available for approved projects through the UK Biobank Access Management System (https://www.ukbiobank.ac.uk/). We have deposited the derived data fields and model predictions following UKB policy, which will be available through the UK Biobank Access Management System. Independent associated loci and polygenic risk score coefficients are available as supplemental tables (supplemental information). Full GWAS summary statistics are publicly available through the above GitHub link.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2021.05.004.

Acknowledgments

We acknowledge research participants from all datasets used for their contributions. We thank Roy Lee, Jonathan Krause, and Avinash Varadarajan, the eye experts who labeled fundus images, and other members of the genomics, ophthalmology, and data labeling teams in Google Health. This research has been conducted with the UK Biobank resource application 17643. We thank Inoveon, Aravind Eye Hospital, Sankara Nethralaya, and Narayana Nethralaya for providing de-identified data; Jorge Cuadros from EyePACS for data access and helpful conversations; and the National Eye Institute Study of Age-Related Macular Degeneration Research Group and study participants for their contributions to this research. The EPIC-Norfolk Eye Study (https://doi.org/10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1, MC-UU_12015/1; genetics MC_PC_13048), Research into Aging (262), and Cancer Research UK (C864/A14136). We appreciate the many study team members at the University of Cambridge who enabled this research. We acknowledge Michelle Chan and David Broadway who contributed to glaucoma status grading in EPIC-Norfolk. We acknowledge the UK Biobank Eye and Vision Consortium for design and oversight of eye and vision data in the UK Biobank and appreciate NIHR Biomedical Research Centre at Moorfields and UCL Institute of Ophthalmology for funding consortium activities. A.P.K. was supported by a Moorfields Eye Charity Career Development Fellowship and a UK Research and Innovation Future Leaders Fellowship. P.J.F. was supported by The Richard Desmond Charitable Trust. This study was funded by Google LLC.

Declaration of interests

P.J.F. and A.P.K. are employees of the UCL Institute of Ophthalmology, London, UK. The remaining authors are employees and shareholders of Google LLC.

Received: December 7, 2020
Accepted: May 10, 2021
Published: June 1, 2021

Web resources

1000 Genomes Project phase 3, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502
AREDS dataset, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000001.v3.p1
BaselineLD annotations, https://alkesgroup.broadinstitute.org/ldscore
BOLT-LMM software, https://alkesgroup.broadinstitute.org/boltlmm
EPIC-Norfolk Study, https://www.epic-norfolk.org.uk
EyePACS dataset, http://www.eyepacs.com/research
FUMA, https://fuma.ctglab.nl
GenomicRanges, https://bioconductor.org/packages/release/bioc/html/GenomicRanges.html
GREAT, http://great.stanford.edu/public/html
ImageNet dataset, https://www.image-net.org/
Inoveon, http://www.inoveon.com
LocusZoom, http://locuszoom.org
Meta-Soft, http://genetics.cs.ucla.edu/meta_jemdoc
OpenTargets, https://genetics.opentargets.org/
PLINK software, https://www.cog-genomics.org/plink/1.9
QCtools, http://www.well.ox.ac.uk/~gav/qctool_v1/
Scikit-learn, https://scikit-learn.org/stable
Springelkamp et al. summary statistics, https://academic.oup.com/hmg/article/26/2/438/2970289
TensorFlow, https://www.tensorflow.org
TwoSampleMR, https://github.com/mrcieu/twosamplemr
The UK Biobank Study, https://www.ukbiobank.ac.uk

References

1. Tung, J.Y., Do, C.B., Hinds, D.A., Kiefer, A.K., Macpherson, J.M., Chowdry, A.B., Francke, U., Naughton, B.T., Mountain, J.L., Wojcicki, A., and Eriksson, N. (2011). Efficient replication of over 180 genetic associations with self-reported medical data. PLoS ONE *6*, e23473.
2. Deveza, L.A., Melo, L., Yamato, T., Mills, K., and Hunter, D.J. (2017). Knee osteoarthritis phenotypes and their relevance for outcomes: a systematic review of the literature. Osteoarthritis Cartilage *25*, S57–S58.
3. Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M., et al. (2015). UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med. *12*, e1001779.
4. Nagai, A., Hirata, M., Kamatani, Y., Muto, K., Matsuda, K., Kiyohara, Y., Ninomiya, T., Tamakoshi, A., Yamagata, Z., Mushiroda, T., et al.; BioBank Japan Cooperative Hospital Group (2017). Overview of the BioBank Japan Project: Study design and profile. J. Epidemiol. *27* (3S), S2–S8.
5. DeBoever, C., Tanigawa, Y., Aguirre, M., McInnes, G., Lavertu, A., and Rivas, M.A. (2020). Assessing Digital Phenotyping to Enhance Genetic Studies of Human Diseases. Am. J. Hum. Genet. *106*, 611–622.
6. Wheeler, E., Leong, A., Liu, C.-T., Hivert, M.-F., Strawbridge, R.J., Podmore, C., Li, M., Yao, J., Sim, X., Hong, J., et al.; EPIC-CVD Consortium; EPIC-InterAct Consortium; and Lifelines Cohort Study (2017). Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. PLoS Med. *14*, e1002383.
7. Sinnott-Armstrong, N., Tanigawa, Y., Amar, D., Mars, N., Benner, C., Aguirre, M., Venkataraman, G.R., Wainberg, M., Ollila, H.M., Kiiskinen, T., et al.; FinnGen (2021). Genetics of 35 blood and urine biomarkers in the UK Biobank. Nat. Genet. *53*, 185–194.
8. Bai, W., Suzuki, H., Huang, J., Francis, C., Wang, S., Tarroni, G., Guitton, F., Aung, N., Fung, K., Petersen, S.E., et al. (2020). A population-based phenome-wide association study of cardiac and aortic structure and function. Nat. Med. *26*, 1654–1662.
9. Phene, S., Dunn, R.C., Hammel, N., Liu, Y., Krause, J., Kitade, N., Schaekermann, M., Sayres, R., Wu, D.J., Bora, A., et al. (2019). Deep Learning and Glaucoma Specialists: The Relative Importance of Optic Disc Features to Predict Glaucoma Referral in Fundus Photographs. Ophthalmology *126*, 1627–1639.
10. Yim, J., Chopra, R., Spitz, T., Winkens, J., Obika, A., Kelly, C., Askham, H., Lukic, M., Huemer, J., Fasler, K., et al. (2020). Predicting conversion to wet age-related macular degeneration using deep learning. Nat. Med. *26*, 892–899.
11. Glastonbury, C.A., Pulit, S.L., Honecker, J., Censin, J.C., Laber, S., Yaghootkar, H., Rahmioglu, N., Pastel, E., Kos, K., Pitt, A., et al. (2020). Machine Learning based histology phenotyping to investigate the epidemiologic and genetic basis of adipocyte morphology and cardiometabolic traits. PLoS Comput. Biol. *16*, e1008044.
12. Jonas, J.B., Aung, T., Bourne, R.R., Bron, A.M., Ritch, R., and Panda-Jonas, S. (2017). Glaucoma. Lancet *390*, 2183–2193.
13. Pascolini, D., and Mariotti, S.P. (2012). Global estimates of visual impairment: 2010. Br. J. Ophthalmol. *96*, 614–618.
14. Tham, Y.-C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., and Cheng, C.-Y. (2014). Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. Ophthalmology *121*, 2081–2090.
15. Wang, K., Gaitsch, H., Poon, H., Cox, N.J., and Rzhetsky, A. (2017). Classification of common human diseases derived from shared genetic and environmental determinants. Nat. Genet. *49*, 1319–1325.
16. Khawaja, A.P., Cooke Bailey, J.N., Wareham, N.J., Scott, R.A., Simcoe, M., Igo, R.P., Jr., Song, Y.E., Wojciechowski, R., Cheng, C.-Y., Khaw, P.T., et al.; UK Biobank Eye and Vision Consortium; and NEIGHBORHOOD Consortium (2018). Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. Nat. Genet. *50*, 778–782.
17. Craig, J.E., Han, X., Qassim, A., Hassall, M., Cooke Bailey, J.N., Kinzy, T.G., Khawaja, A.P., An, J., Marshall, H., Gharahkhani, P., et al.; NEIGHBORHOOD consortium; and UK Biobank Eye and Vision Consortium (2020). Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. Nat. Genet. *52*, 160–166.
18. Foster, P.J., Buhrmann, R., Quigley, H.A., and Johnson, G.J. (2002). The definition and classification of glaucoma in prevalence surveys. Br. J. Ophthalmol. *86*, 238–242.
19. Gordon, M.O., Beiser, J.A., Brandt, J.D., Heuer, D.K., Higginbotham, E.J., Johnson, C.A., Keltner, J.L., Miller, J.P., Parrish, R.K., 2nd, Wilson, M.R., and Kass, M.A. (2002). The Ocular

Hypertension Treatment Study: baseline factors that predict the onset of primary open-angle glaucoma. Arch. Ophthalmol. *120*, 714–720, discussion 829–830.

20. Springelkamp, H., Iglesias, A.I., Mishra, A., Höhn, R., Wojciechowski, R., Khawaja, A.P., Nag, A., Wang, Y.X., Wang, J.J., Cuellar-Partida, G., et al.; NEIGHBORHOOD Consortium (2017). New insights into the genetics of primary open-angle glaucoma based on meta-analyses of intraocular pressure and optic disc characteristics. Hum. Mol. Genet. *26*, 438–453.
21. Czudowska, M.A., Ramdas, W.D., Wolfs, R.C.W., Hofman, A., De Jong, P.T.V.M., Vingerling, J.R., and Jansonius, N.M. (2010). Incidence of glaucomatous visual field loss: a ten-year follow-up from the Rotterdam Study. Ophthalmology *117*, 1705–1712.
22. Age-Related Eye Disease Study Research Group (1999). The Age-Related Eye Disease Study (AREDS): design implications. AREDS report no. 1. Control. Clin. Trials *20*, 573–600.
23. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.308.
24. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2009.5206848.
25. Shorten, C., and Khoshgoftaar, T.M. (2019). A survey on Image Data Augmentation for Deep Learning. J. Big Data *6*, 60.
26. Prechelt, L. (1998). Early Stopping - But When? In Neural Networks: Tricks of the Trade, G.B. Orr and K.-R. Müller, eds. (Berlin, Heidelberg: Springer Berlin Heidelberg), pp. 55–69.
27. Loh, P.-R., Tucker, G., Bulik-Sullivan, B.K., Vilhjálmsson, B.J., Finucane, H.K., Salem, R.M., Chasman, D.I., Ridker, P.M., Neale, B.M., Berger, B., et al. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat. Genet. *47*, 284–290.
28. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P., and Price, A.L. (2018). Mixed-model association for biobank-scale datasets. Nat. Genet. *50*, 906–908.
29. McCaw, Z.R., Lane, J.M., Saxena, R., Redline, S., and Lin, X. (2020). Operating characteristics of the rank-based inverse normal transformation for quantitative trait analysis in genome-wide association studies. Biometrics *76*, 1262–1272.
30. Bulik-Sullivan, B.K., Loh, P.-R., Finucane, H.K., Ripke, S., Yang, J., Patterson, N., Daly, M.J., Price, A.L., Neale, B.M.; and Schizophrenia Working Group of the Psychiatric Genomics Consortium (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat. Genet. *47*, 291–295.
31. Egger, M., Davey Smith, G., Schneider, M., and Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. BMJ *315*, 629–634.
32. Gharahkhani, P., Jorgenson, E., Hysi, P., Khawaja, A.P., Pendergrass, S., Han, X., Ong, J.S., Hewitt, A.W., Segre, A., Igo, R.P., et al. (2020). A large cross-ancestry meta-analysis of genome-wide association studies identifies 69 novel risk loci for primary open-angle glaucoma and includes a genetic link with Alzheimer's disease. bioRxiv. https://doi.org/10.1101/2020.01.30.927822.
33. Chatterjee, N., Shi, J., and García-Closas, M. (2016). Developing and evaluating polygenic risk prediction models for stratified disease prevention. Nat. Rev. Genet. *17*, 392–406.
34. Zou, H., and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. J. R. Stat. Soc. Series B Stat. Methodol. *67*, 301–320.
35. Khawaja, A.P., Chan, M.P.Y., Broadway, D.C., Garway-Heath, D.F., Luben, R., Yip, J.L.Y., Hayat, S., Khaw, K.-T., and Foster, P.J. (2013). Laser scanning tomography in the EPIC-Norfolk Eye Study: principal components and associations. Invest. Ophthalmol. Vis. Sci. *54*, 6638–6645.
36. Khawaja, A.P., Rojas Lopez, K.E., Hardcastle, A.J., Hammond, C.J., Liskova, P., Davidson, A.E., Gore, D.M., Hafford Tear, N.J., Pontikos, N., Hayat, S., et al. (2019). Genetic Variants Associated With Corneal Biomechanical Properties and Potentially Conferring Susceptibility to Keratoconus in a Genome-Wide Association Study. JAMA Ophthalmol. *137*, 1005–1012.
37. McCarthy, S., Das, S., Kretzschmar, W., Delaneau, O., Wood, A.R., Teumer, A., Kang, H.M., Fuchsberger, C., Danecek, P., Sharp, K., et al.; Haplotype Reference Consortium (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat. Genet. *48*, 1279–1283.
38. Asefa, N.G., Neustaeter, A., Jansonius, N.M., and Snieder, H. (2019). Heritability of glaucoma and glaucoma-related endophenotypes: Systematic review and meta-analysis. Surv. Ophthalmol. *64*, 835–851.
39. Wang, G., Sarkar, A., Carbonetto, P., and Stephens, M. (2020). A simple new approach to variable selection in regression, with application to genetic fine mapping. J. R. Stat. Soc. Series B Stat. Methodol. *25*, 1.
40. Watanabe, K., Taskesen, E., van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with FUMA. Nat. Commun. *8*, 1826.
41. McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.
42. Gao, X.R., Huang, H., Nannini, D.R., Fan, F., and Kim, H. (2018). Genome-wide association analyses identify new loci influencing intraocular pressure. Hum. Mol. Genet. *27*, 2205–2213.
43. Chan, M.P.Y., Broadway, D.C., Khawaja, A.P., Yip, J.L.Y., Garway-Heath, D.F., Burr, J.M., Luben, R., Hayat, S., Dalzell, N., Khaw, K.-T., and Foster, P.J. (2017). Glaucoma and intraocular pressure in EPIC-Norfolk Eye Study: cross sectional study. BMJ *358*, j3889.
44. Lee, S., Lee, K., Hwang, S., Kim, S.H., Song, W.K., Park, Z.Y., and Chang, S. (2006). SPIN90/WISH interacts with PSD-95 and regulates dendritic spinogenesis via an N-WASP-independent mechanism. EMBO J. *25*, 4983–4995.
45. Reim, K., Regus-Leidig, H., Ammermüller, J., El-Kordi, A., Radyushkin, K., Ehrenreich, H., Brandstätter, J.H., and Brose, N. (2009). Aberrant function and structure of retinal ribbon synapses in the absence of complexin 3 and complexin 4. J. Cell Sci. *122*, 1352–1361.
46. Sapir, T., Sapoznik, S., Levy, T., Finkelshtein, D., Shmueli, A., Timm, T., Mandelkow, E.-M., and Reiner, O. (2008). Accurate balance of the polarity kinase MARK2/Par-1 is required for proper cortical neuronal migration. J. Neurosci. *28*, 5710–5720.
47. Melin, B.S., Barnholtz-Sloan, J.S., Wrensch, M.R., Johansen, C., Il'yasova, D., Kinnersley, B., Ostrom, Q.T., Labreche, K., Chen, Y., Armstrong, G., et al.; GliomaScan Consortium (2017). Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to

glioblastoma and non-glioblastoma tumors. Nat. Genet. *49*, 789–794.

48. Hendee, K., Wang, L.W., Reis, L.M., Rice, G.M., Apte, S.S., and Semina, E.V. (2017). Identification and functional analysis of an ADAMTSL1 variant associated with a complex phenotype including congenital glaucoma, craniofacial, and other systemic features in a three-generation human pedigree. Hum. Mutat. *38*, 1485–1490.
49. Seifi, M., and Walter, M.A. (2018). Axenfeld-Rieger syndrome. Clin. Genet. *93*, 1123–1130.
50. Bailey, J.N.C., Loomis, S.J., Kang, J.H., Allingham, R.R., Gharahkhani, P., Khor, C.C., Burdon, K.P., Aschard, H., Chasman, D.I., Igo, R.P., Jr., et al.; ANZRAG Consortium (2016). Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. Nat. Genet. *48*, 189–194.
51. Gal, A., Rau, I., El Matri, L., Kreienkamp, H.-J., Fehr, S., Baklouti, K., Chouchane, I., Li, Y., Rehbein, M., Fuchs, J., et al. (2011). Autosomal-recessive posterior microphthalmos is caused by mutations in PRSS56, a gene encoding a trypsin-like serine protease. Am. J. Hum. Genet. *88*, 382–390.
52. Azuma, N., Hirakiyama, A., Inoue, T., Asaka, A., and Yamada, M. (2000). Mutations of a human homologue of the Drosophila eyes absent gene (EYA1) detected in patients with congenital cataracts and ocular anterior segment anomalies. Hum. Mol. Genet. *9*, 363–366.
53. Wawrocka, A., and Krawczynski, M.R. (2018). The genetics of aniridia - simple things become complicated. J. Appl. Genet. *59*, 151–159.
54. D'Elia, A.V., Pellizzari, L., Fabbro, D., Pianta, A., Divizia, M.T., Rinaldi, R., Grammatico, B., Grammatico, P., Arduino, C., and Damante, G. (2007). A deletion 3¿ to the PAX6 gene in familial aniridia cases. Mol. Vis. *13*, 1245–1250.
55. Hayat, S.A., Luben, R., Keevil, V.L., Moore, S., Dalzell, N., Bhaniani, A., Khawaja, A.P., Foster, P., Brayne, C., Wareham, N.J., and Khaw, K.T. (2014). Cohort profile: A prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3). Int. J. Epidemiol. *43*, 1063–1072.
56. Bowden, J., Del Greco M, F., Minelli, C., Davey Smith, G., Sheehan, N., and Thompson, J. (2017). A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat. Med. *36*, 1783–1802.
57. Currant, H., Hysi, P., Fitzgerald, T.W., Gharahkhani, P., Bonnemaijer, P.W.M., Atan, D., Aung, T., Charng, J., Choquet, H., Craig, J., et al.; UK Biobank Eye and Vision Consortium; and International Glaucoma Genetics Consortium (2020). Genetic variation affects morphological retinal phenotypes extracted from UK Biobank Optical Coherence Tomography images. medRxiv. https://doi.org/10.1101/2020.07.20.20157180.
58. Khawaja, A.P., Chua, S., Hysi, P.G., Georgoulas, S., Currant, H., Fitzgerald, T.W., Birney, E., Ko, F., Yang, Q., Reisman, C., et al.; UK Biobank Eye and Vision Consortium (2020). Comparison of Associations with Different Macular Inner Retinal Thickness Parameters in a Large Cohort: The UK Biobank. Ophthalmology *127*, 62–71.
59. Delude, C.M. (2015). Deep phenotyping: The details of disease. Nature *527*, S14–S15.
60. UK National Screening Committee (2019). Screening for Glaucoma - External review against programme appraisal criteria for the UK National Screening Committee. https://legacyscreening.phe.org.uk/policydb_download.php?doc=1219.
61. World Health Organization (2019). World report on vision. ISBN: 9789241516570. https://www.who.int/publications/i/item/9789241516570.
62. Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., and Houlsby, N. (2019). Big Transfer (BiT): General Visual Representation Learning. arXiv, 1912.11370.
63. Raghu, M., Zhang, C., Kleinberg, J., and Bengio, S. (2019). Transfusion: Understanding Transfer Learning for Medical Imaging. arXiv, 1902.07208.
64. Han, X., Qassim, A., An, J., Marshall, H., Zhou, T., Ong, J.-S., Hassall, M.M., Hysi, P.G., Foster, P.J., Khaw, P.T., et al. (2019). Genome-wide association analysis of 95 549 individuals identifies novel loci and genes influencing optic disc morphology. Hum. Mol. Genet. *28*, 3680–3690.
65. Garway-Heath, D.F., Rudnicka, A.R., Lowe, T., Foster, P.J., Fitzke, F.W., and Hitchings, R.A. (1998). Measurement of optic disc size: equivalence of methods to correct for ocular magnification. Br. J. Ophthalmol. *82*, 643–649.
66. Han, X., Steven, K., Qassim, A., Marshall, H.N., Bean, C., Tremeer, M., An, J., Siggs, O., Gharahkhani, P., Craig, J.E., et al. (2020). Automated AI labeling of optic nerve head enables new insights into cross-ancestry glaucoma risk and genetic discovery in >280,000 images from the UKB and CLSA. Am. J. Hum. Genet. Published online June 1, 2021. https://doi.org/10.1016/j.ajhg.2021.05.005.

The American Journal of Human Genetics, Volume *108*

Supplemental information

Large-scale machine learning-based phenotyping significantly improves genomic discovery for optic nerve head morphology

Babak Alipanahi, Farhad Hormozdiari, Babak Behsaz, Justin Cosentino, Zachary R. McCaw, Emanuel Schorsch, D. Sculley, Elizabeth H. Dorfman, Paul J. Foster, Lily H. Peng, Sonia Phene, Naama Hammel, Andrew Carroll, Anthony P. Khawaja, and Cory Y. McLean

Table of Contents:

Figures

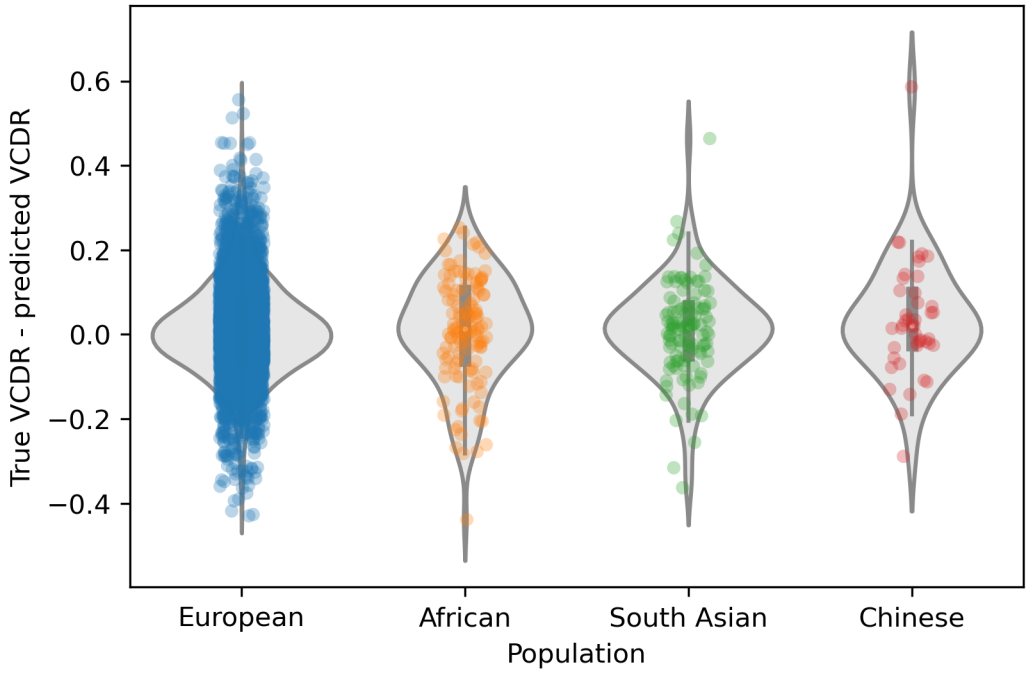

Figure S1. Accuracy of the VCDR prediction model across genetic ancestries. Of 4,816 total UK Biobank images with at least one manual VCDR grade, differences between the manual label ("True VCDR") and the model prediction ("predicted VCDR") are shown by ancestry (see "Genetic ancestry inference"; European *n*=4,538; African *n*=124; South Asian *n*=110, Chinese *n*=44). No significant differences were detected across ancestries (one-way ANOVA *P*=0.56 across the four groups; paired T-test *P*=0.78 when comparing European to non-European).

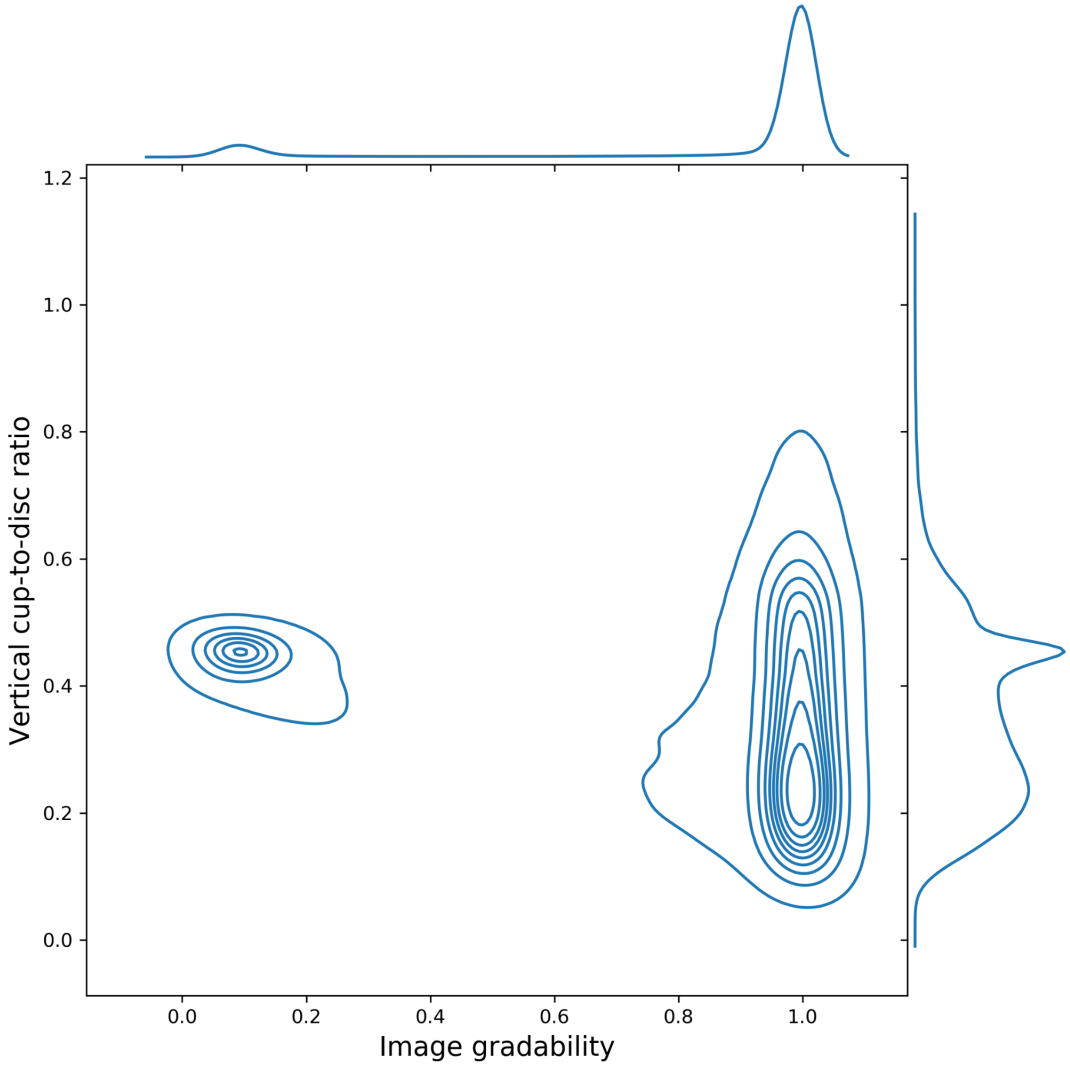

Figure S2. Kernel density estimate of the distribution of image gradability and VCDR predictions in UK Biobank images. The 21,400 images with gradability < 0.7 were omitted from further analysis.

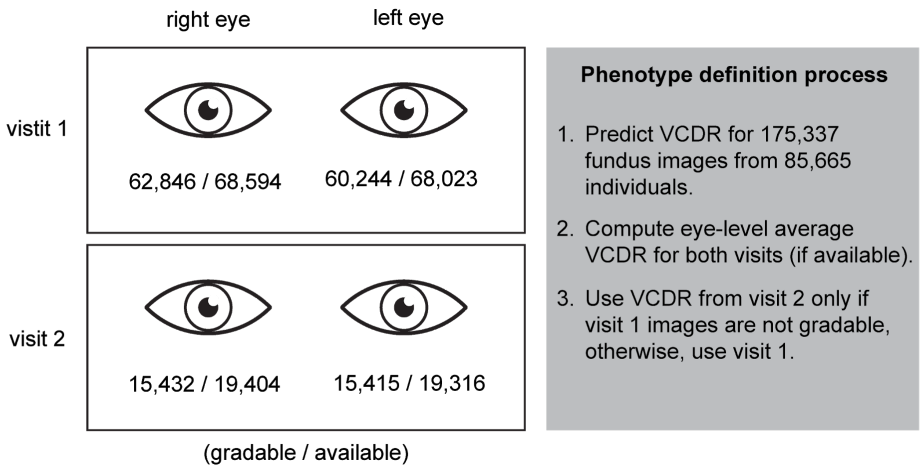

Figure S3. VCDR phenotype calling process. The numbers below each eye indicate gradable / available images.

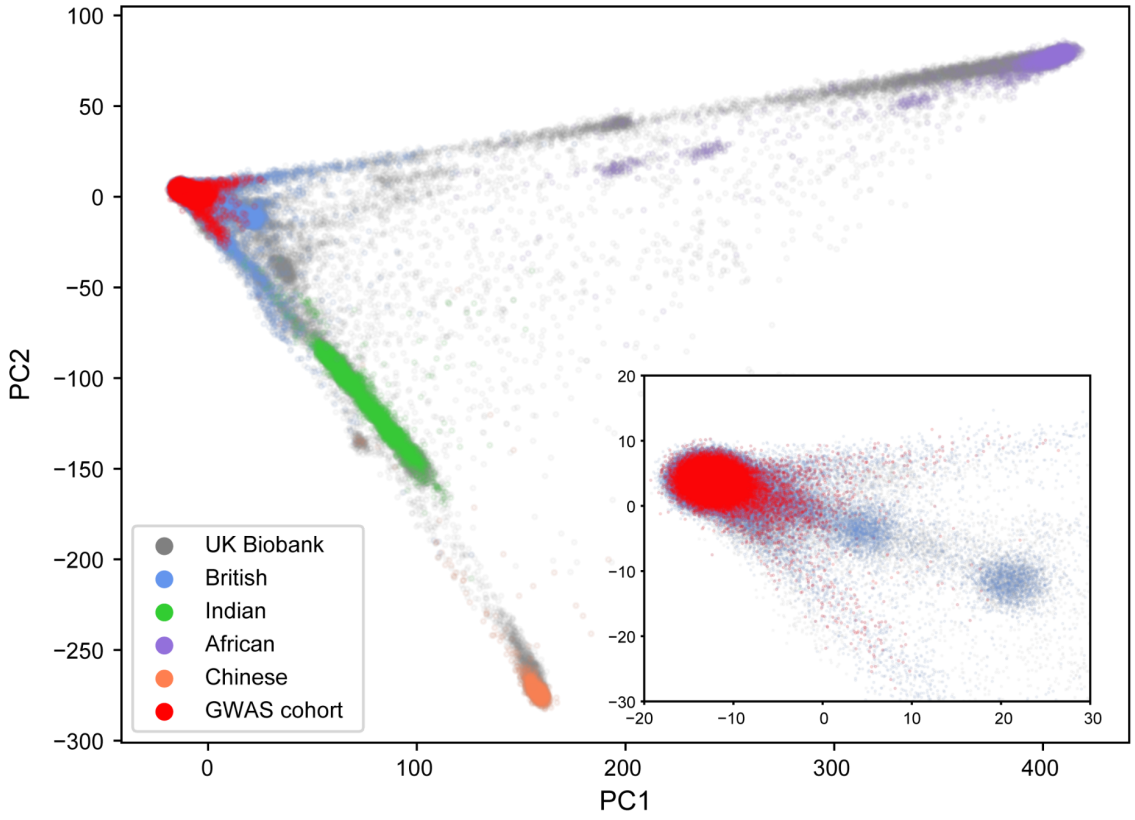

Figure S4. Genetic Principal Components. The first two PCs of all individuals in the UK Biobank and the GWAS cohort is shown. The individuals with self-reported "British", "Indian", "Afrcian" and "Chinese" ancestries are also shown for reference. The inset shows a zoomed version of the GWAS cohort.

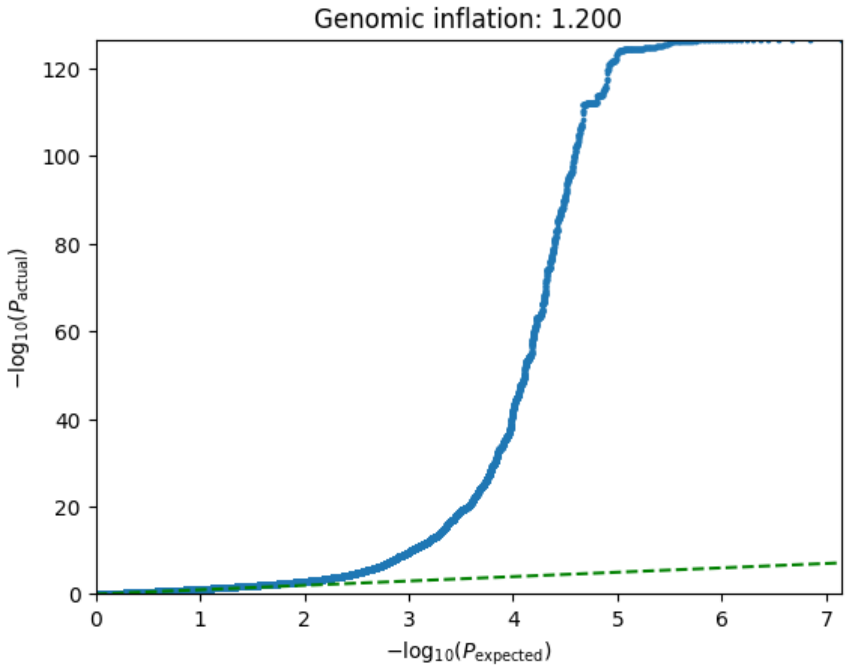

Figure S5. QQ-plot for the ML-based VCDR GWAS. The expected *P*-values are based on a uniform distribution.

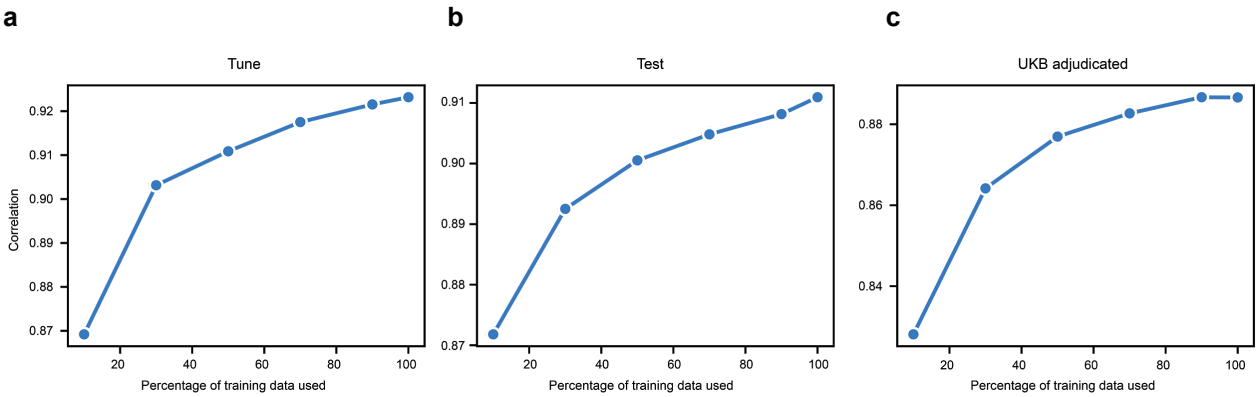

Figure S6. VCDR model performance as a function of the percentage of training data samples used to train the model. Pearson's correlation between the model-predicted VCDR and the expert-labeled VCDR at training data percentages from 10% to 100% for **a,** the tune dataset, **b,** the test dataset, and **c,** the UKB adjudicated dataset. See **Model Training and Evaluation** section for detailed dataset definitions.

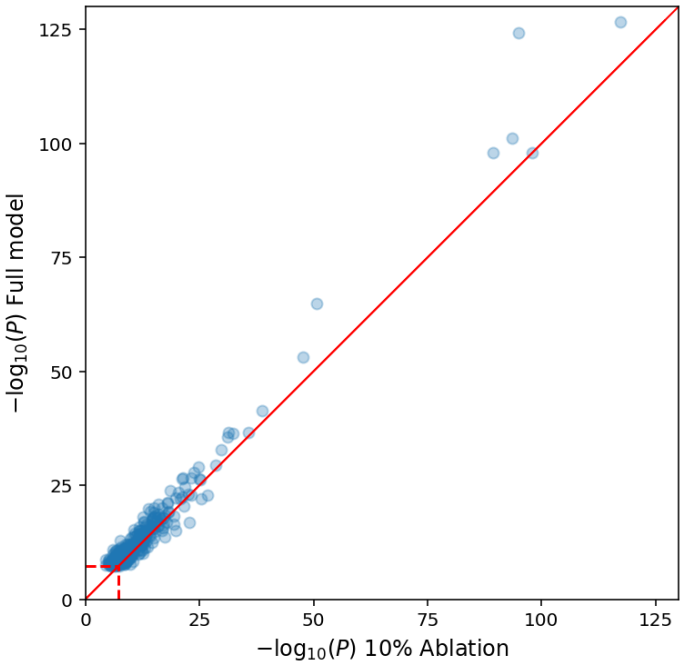

Figure S7. Comparison of the Full model *P*-values with the 10% Ablation *P*-values for 299 Full model hits. The dashed red horizontal and vertical lines indicate the GWS level ($P<5\times10^{-8}$).

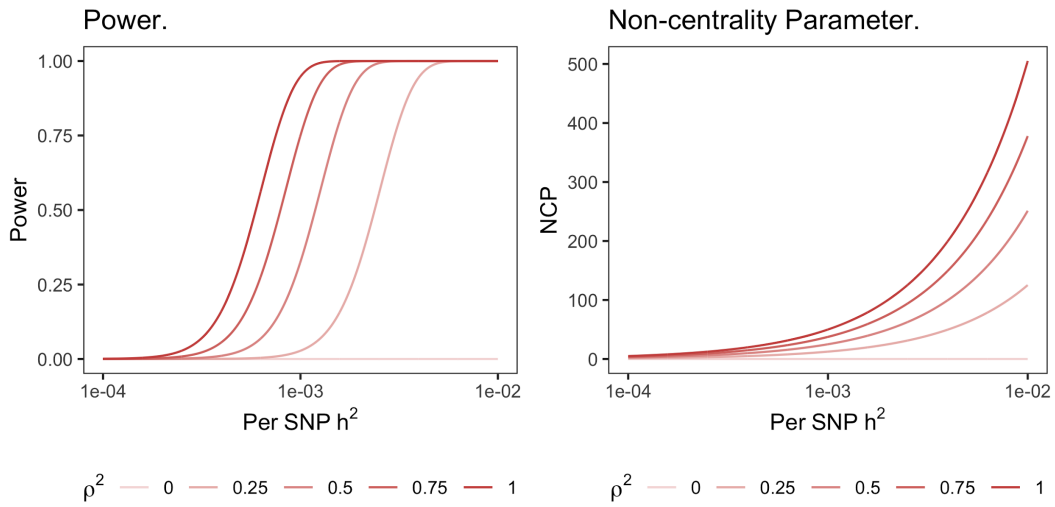

Figure S8. Power and non-centrality curves as a function of per-SNP heritability. The curves are stratified by the correlation between the mismeasured and true phenotypes. Sample size was set to *n*=50,000; changing the sample sizes amounts to horizontally shifting the power curves. The range of per-SNP heritabilities was selected to demonstrate the inflection points of the power curves.

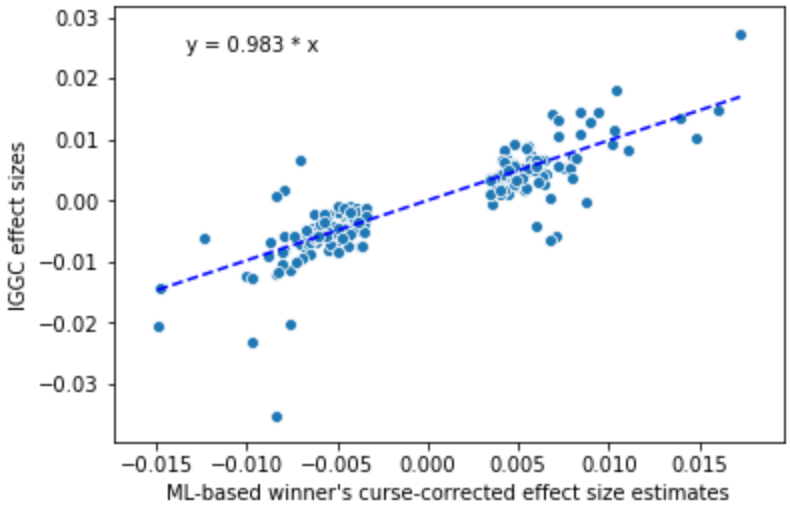

Figure S9. Regression of IGGC VCDR meta-analysis effect sizes on winner's curse-corrected ML-based VCDR effect size estimates. Of 299 independent GWS hits, 214 were present in IGGC. Regression slope was computed with intercept fixed to zero. The dotted blue line shows the line of best fit.

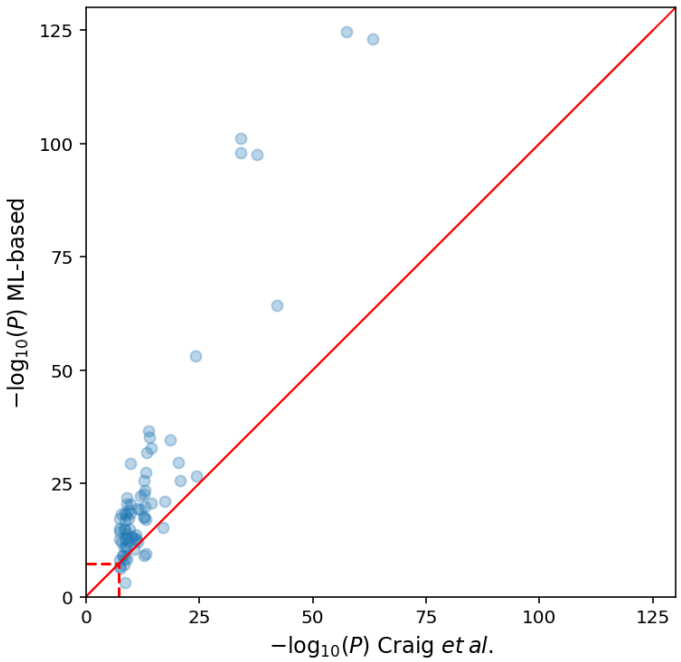

Figure S10. Comparison of ML-based VCDR *P*-values with the Craig *et al.* *P*-values for 73 Craig *et al.* hits. The dashed red horizontal and vertical lines indicate the GWS level ($P<5\times10^{-8}$).

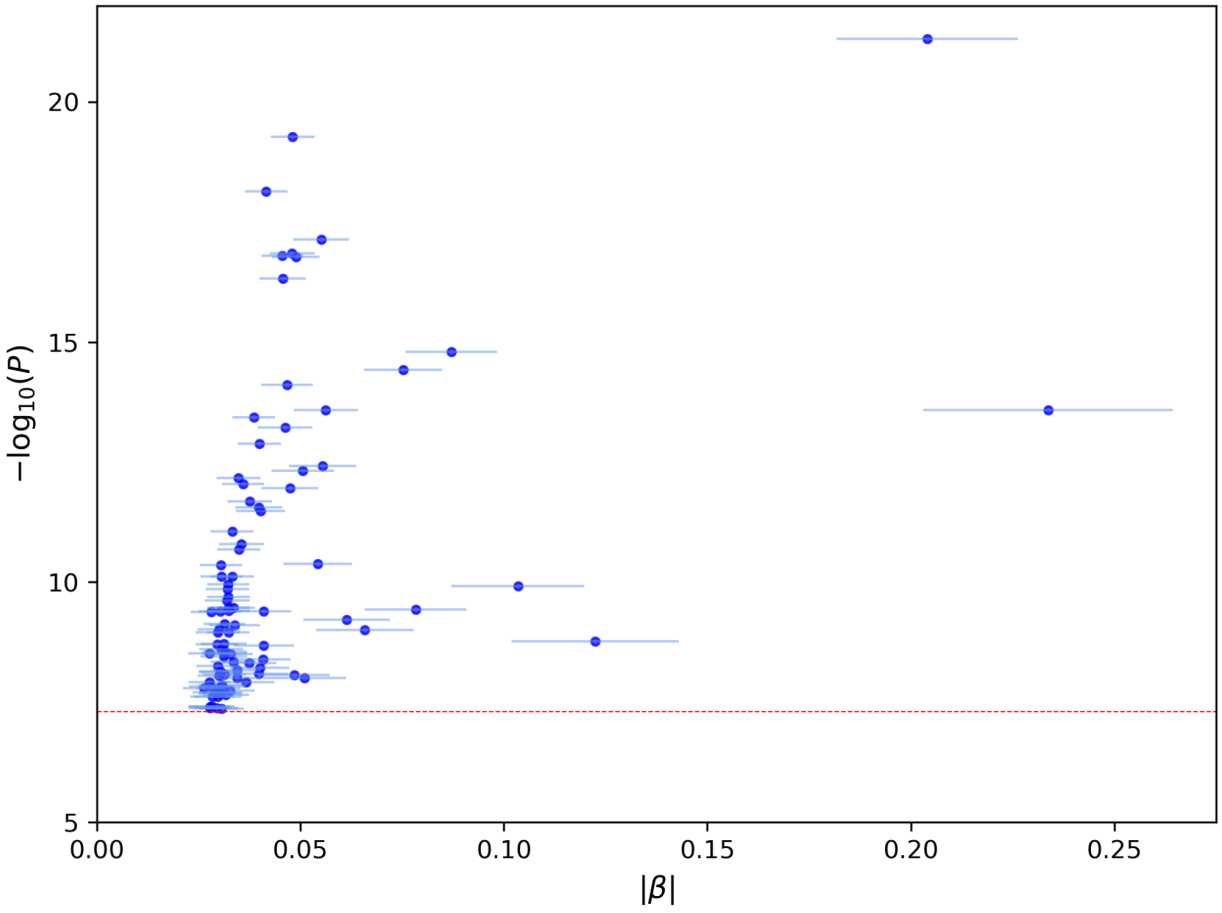

Figure S11. Novel ML-based VCDR loci. The -$\log_{10}(P)$ and the effect size magnitude (change in expected VCDR per risk allele) of the 93 novel loci not observed in Craig *et al.* or IGGC VCDR GWAS loci are shown. The error bars show standard errors of the effect sizes. The red dashed line denotes the genome-wide significance level.

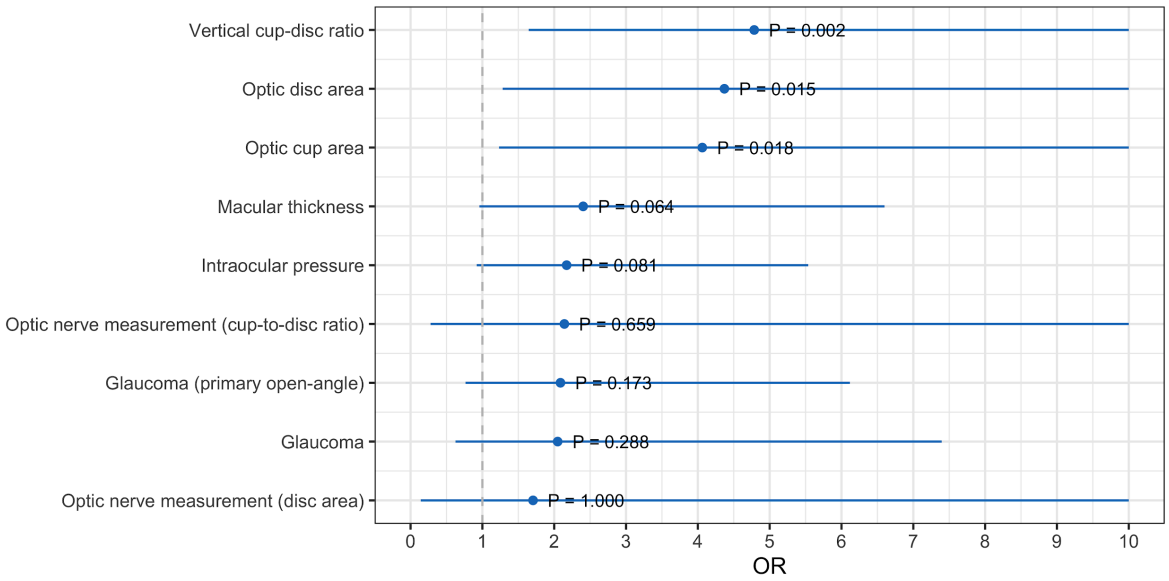

Figure S12. **FUMA enrichment of eye-related gene sets from the ML-based VCDR GWAS versus the VCDR GWAS of Craig *et al.*** Enrichment is quantified via the odds ratio, with confidence interval and *P*-value provided by Fisher's exact test.

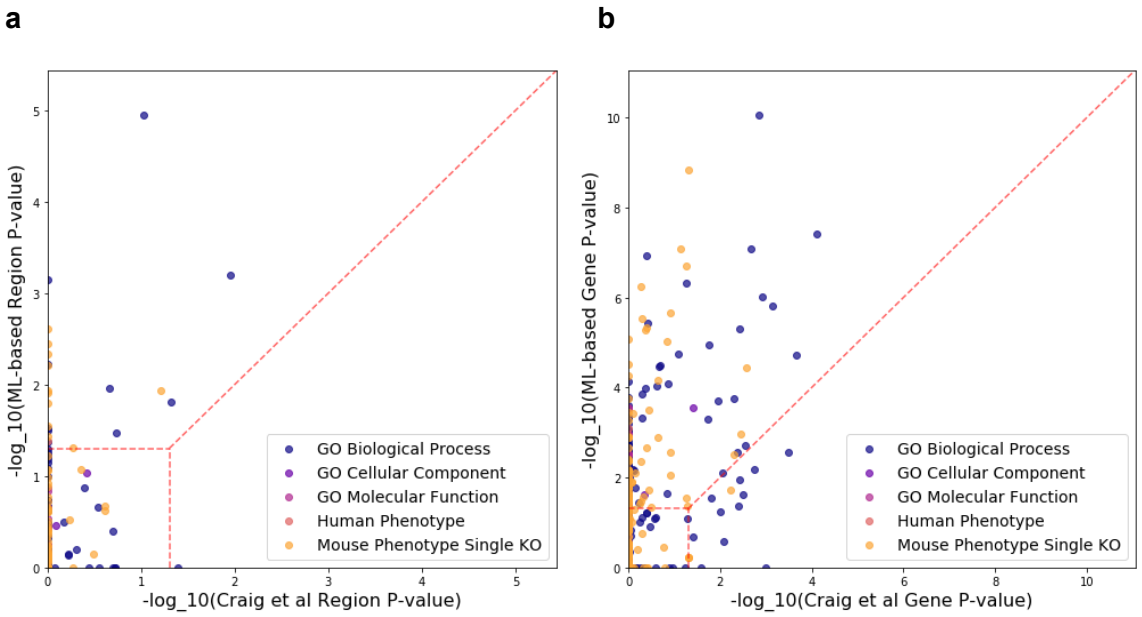

Figure S13. GREAT enrichment of loci from the ML-based VCDR GWAS vs the VCDR GWAS of Craig *et al.* a, Comparison of Bonferroni-corrected *P*-values for the region-based test reported by GREAT for the five listed ontologies. Dashed horizontal and vertical lines show the threshold for statistical significance at a Bonferroni-corrected $P \leq 0.05$, and the diagonal line indicates y=x. More ontology terms are statistically significant for the ML-based GWAS than Craig *et al.* **b,** Comparison analogous to that in **a**, for the gene-based test.

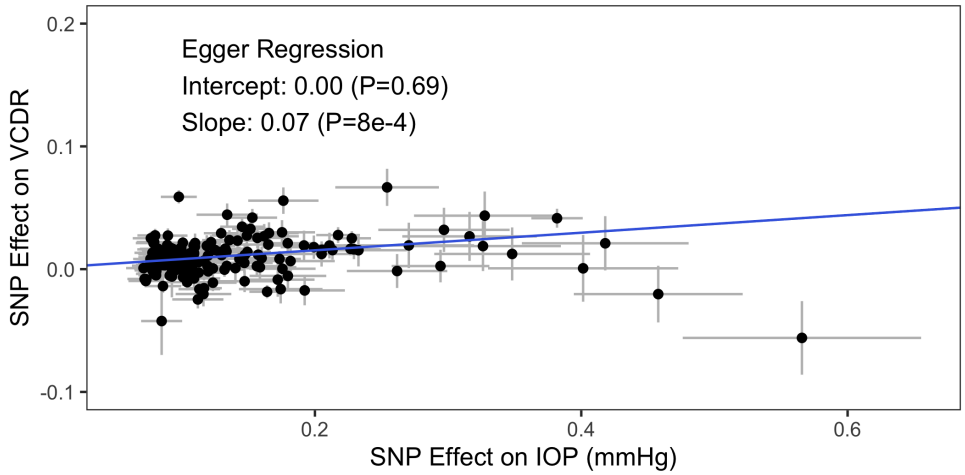

Figure S14. Egger Regression for Mendelian Randomization of the effect of IOP on ML-based VCDR. Independent, significant IOP-associated SNPs were ascertained from Khawaja *et al.* and harmonized with GWAS results for ML-based VCDR.

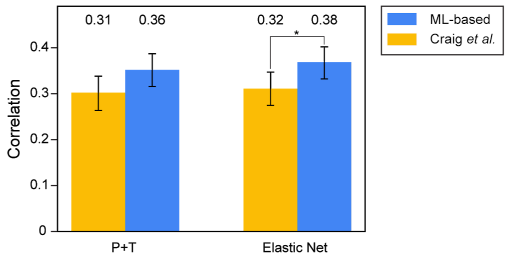

Figure S15. VCDR polygenic risk score performance metrics on the UKB imputation panel. Pearson's correlations between measured VCDR values and predictions of the pruning and thresholding (P+T) and the Elastic Net models are shown for the PRS learned from ML-based and Craig *et al.* hits. Error bars depict 95% confidence intervals. Numbers above bars are the observed Pearson's correlations. Indications of *P*-value ranges: * $P \leq 0.05$, ** $P \leq 0.01$, *** $P \leq 0.001$. Measured VCDR values were obtained from adjudicated expert labeling of fundus photographs (UKB, n=2,076).

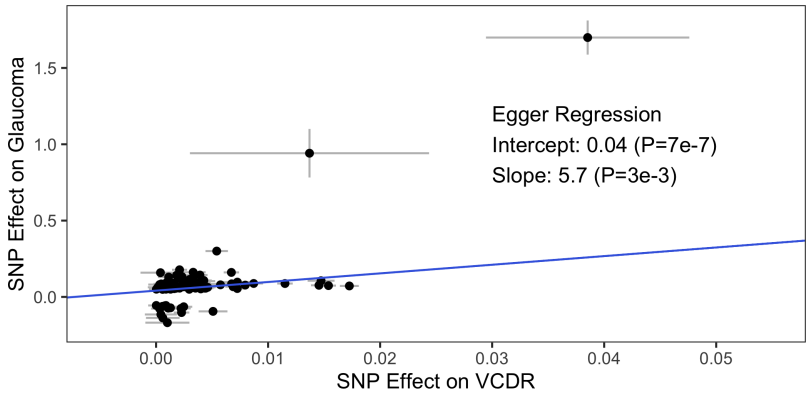

Figure S16. Egger Regression for Mendelian Randomization of the Effect of VCDR on Glaucoma log Odds. Independent, significant Glaucoma risk SNPs were ascertained from Gharahkhani *et al.* and harmonized with GWAS results from ML-based VCDR.

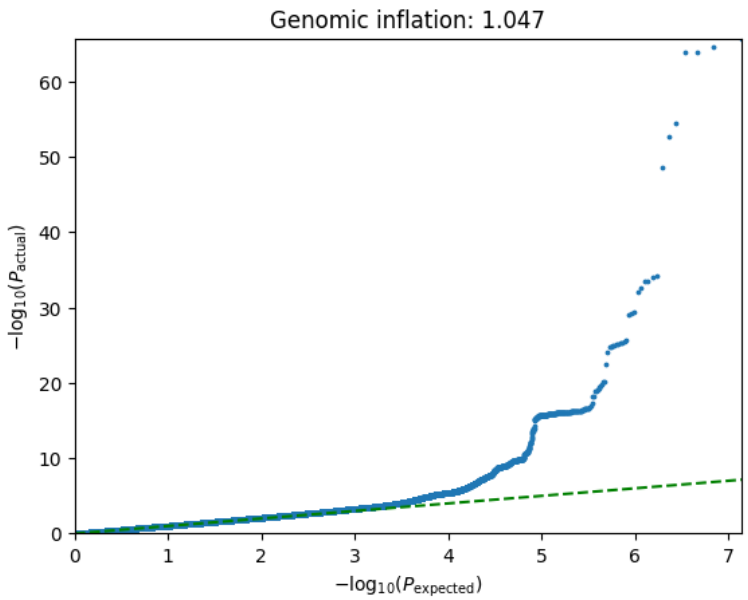

Figure S17. **QQ-plot for the ML-based glaucoma liability GWAS conditional on ML-based VCDR.** The expected *P*-values are based on a uniform distribution.

Tables

Table S1. Phenotype prediction model performance metrics. For VCDR Pearson's correlation is reported. AUC, area under ROC curve; AUPRC, area under precision-recall curve, RMSE, root mean square error. The numbers in parentheses are 95% confidence intervals.

Shared nomenclature for all GWAS results:

CHR, chromosome; POS, base-pair variant position; EA, effect allele; NEA, non-effect allele; EAF, effect allele frequency; BETA, estimated effect size; SE, standard error; P, GWAS *P*-value; NUM_INDV, sample size for the variant; SRC, imputed or genotyped variant; INFO, imputation INFO score (set to 1 for genotyped variants); CRAIG, locus replicated in Craig *et al.* GWAS, CRAIG_META, locus replicated in Craig *et al.* meta-analysis; GENE_CONTEXT: genomic context of the variant, as explained below.

- Overlapping gene(s)
  - `[A]`: variant overlaps gene A
  - `[A,B]`: variant overlaps genes A and B
- Downstream genes
  - `[]A`: variant position is $0 < p \leq 10^3$ bp upstream of closest downstream gene A
  - `[]-A`: variant position is $10^3 < p \leq 10^4$ bp upstream of closest downstream gene A
  - `[]--A`: variant position is $10^4 < p \leq 10^5$ bp upstream of closest downstream gene A
  - `[]---A`: variant position is $10^5 < p \leq 10^6$ bp upstream of closest downstream gene A
  - `[]`: closest downstream gene is further than 10^6 bp
- Upstream genes
  - The notation for upstream genes is similar, but gene A is on the left side, e.g., `B-[]` means variant position is $10^3 < p \leq 10^4$ bp downstream of closest gene B

For example, `FOXD2--[]---TRABD2B` indicates the variant is $10^4 < p \leq 10^5$ downstream of *FOXD2* and $10^5 < p \leq 10^6$ upstream of *TRABD2B*.

Table S2. ML-based VCDR GWAS independent GWS hits ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$).

Table S3. ML-based VCDR GWAS independent GWS loci ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$, distance between top hits > 250k).

Table S4. SuSiE per-SNP results for all fine-mapped SNPs. PIP is the posterior probability the SNP is causal, higher being more likely; and LOCUS_IDX is a locus identifying index (as defined in Table S3). SNPs with PIP = 0 are not shown.

Table S5. SuSiE results summarized per-locus. N_FINEMAPPED is the number of SNPs in loci with PIPs available. N_GWS is the number of genome-wide significant SNPs with MAF>0.05. N_CAUSAL is the sum of PIPs across SNPs in the locus. The estimated number of causal SNPs for a locus is min(N_GWS, N_CAUSAL).

Table S6. ML-based + IGGC VCDR meta-analysis independent GWS hits ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$).

Table S7. ML-based + IGGC VCDR meta-analysis independent GWS loci ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$, distance between top hits > 250k).

Table S8. Novel ML-based VCDR GWAS loci. One-hundred fifty-six (156) independent, genome-wide significant loci were identified by the ML-based VCDR GWAS, increasing to 189

after meta-analyzing with the IGGC's results. This table reports the numbers of loci that were not overlapped by any locus from each previous study, and hence were novel with respect to that study. For example, 135 of the 156 ML-based VCDR loci were not overlapped by any locus reported in the IGGC's original meta-analysis. In addition to comparing against individual studies, we compare against the unions of results from multiple studies. For example, 93 of the 156 ML-based VCDR loci were not overlapped by any locus from either the original IGGC meta-analysis or the Craig *et al.* VCDR GWAS.

Table S9. All GREAT ontology terms significant for at least one of the two sets of loci. All terms in the ontologies of Figure S13 were tested. Abbreviations: ML *P-val*, the Bonferroni-corrected *P*-value for the region-based test with the ML-based GWAS loci; Craig *P-val*, the Bonferroni-corrected *P*-value for the region-based test with the Craig *et al.* GWAS loci; GOBP, Gene Ontology Biological Process; MP1KO, Mouse Phenotype Single Knockout; HP, Human Phenotype.

Table S10. PheWAS results for ML-based VCDR hits. Results from queries to the OpenTargets website are presented for all 299 ML-based VCDR hits. Abbreviations: SNP ID: ML-based VCDR hit ID, P-value: *P*-value obtained from OpenTargets, Beta: estimated effect size obtained from OpenTargets, PASS: True if the *P*-value is smaller than the Bonferroni threshold and False otherwise.

Shared nomenclature for all PRS Tables:

CHR, chromosome; POS, base-pair variant position; REF, reference allele; ALT, alternative allele; EA, effect allele, COEFF, variants coefficient

Table S11. ML-based P+T PRS using EPIC panel (282 variants).

Table S12. ML-based ElasticNet PRS using EPIC panel (282 variants).

Table S13. ML-based P+T PRS using UKB panel (299 variants).

Table S14. ML-based ElasticNet PRS using UKB panel (299 variants).

Table S15. ML-based glaucoma risk conditioned on ML-based VCDR independent GWS hits ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$).

Table S16. ML-based glaucoma risk conditioned on ML-based VCDR independent GWS loci ($R^2 \leq 0.1$, $P \leq 5 \times 10^{-8}$, distance between top hits > 250k).

Methods

Phenotype Prediction Model

Data collection

Grading of images has been described in detail previously (Phene et al. 2019). In short, graders assessed each image for gradability, presence of various optic nerve head (ONH) features (including estimation of VCDR; the ratio between the vertical diameter of the cup and the vertical diameter of the disc) and referable glaucomatous optic neuropathy (GON). Gradability was measured based on image quality, blurring, media opacity, or any other confounding reason. If graders selected “ungradable” for a particular feature or referable GON, then no grade was collected for that aspect. To enable systematic training of graders, we developed grading guidelines and iterated on the guidelines with a panel of three fellowship-trained glaucoma specialists to increase inter-rater agreement; please refer to the Supplementary Table 1 in (Phene et al. 2019). Similar to clinical practice, for VCDR graders were asked to provide an estimate as a decimal between 0.0 and 1.0, with 0.1 increments (0.0 < VCDR < 1.0). For referable GON grading we developed guidelines for a four-point GON assessment ("non-glaucomatous", "low-risk glaucoma suspect", “high-risk glaucoma suspect”, and “likely glaucoma”) where the “high-risk glaucoma suspect” or “likely glaucoma” levels were considered referable, that is, the ONH appearance was worrisome enough to justify referral for comprehensive examination. Graders were asked to provide a referable GON grade after evaluating the image for the other ONH features.

Model training and validation

Data processing and model training has been described previously (Phene et al. 2019). In short, we first remove all UK Biobank (UKB) samples from the "train", "tune", and "test" sets used by (Phene et al. 2019). We use 81,830 color fundus images from AREDS (age-related eye disease study) (Age-Related Eye Disease Study Research Group 1999), EyePACS (https://www.eyepacs.org/), Inoveon (http://www.inoveon.com/) from United States and two eye hospitals in India (Narayana Nethralaya and Sankara Nethralaya). In total, 69,460 of the 79,355 training images were gradable. All color fundus images are cropped to center the retinal image and resized to 587×587 pixels. The prediction model consists of ten independently trained multi-task Inception V3 (Szegedy et al. 2016) deep convolutional neural networks. To accelerate model training, convolutional layers were initialized using the weights learned from the Image Net dataset (Deng et al. 2009). We used image augmentation (Shorten and Khoshgoftaar 2019) (randomly changing brightness, hue, contrast, saturation and flipping the image horizontally and vertically) to regularize model training in TensorFlow (Abadi et al. 2016). Full set of hyperparameters is given in the "Model hyper-parameters" section. We used early stopping (Prechelt 1998) based on root mean squared error (RMSE) for predicting VCDR in the tune set

for each model. The final prediction was the average prediction of the ten models in the ensemble. Model performance metrics are listed in Table S1.

Genomic Discovery

UK Biobank cohort

The UK Biobank is a very large multisite cohort study established by the Medical Research Council, Department of Health, Wellcome Trust medical charity, Scottish Government and Northwest Regional Development Agency. Detailed study protocols are available online (http://www.ukbiobank.ac.uk/resources/ and http://biobank.ctsu.ox.ac.uk/crystal/docs.cgi). A baseline questionnaire, physical measurements, and biological samples were undertaken in 22 assessment centers across the UK between 2006 and 2010. All UK residents aged 40 to 69 years who were registered with the National Health Service (NHS) and living up to 25 miles from a study center were invited to participate. The study was conducted with the approval of the North-West Research Ethics Committee (ref 06/MRE08/65), in accordance with the principles of the Declaration of Helsinki, and all participants gave written informed consent. This research has been conducted using the UK Biobank Resource under Application Number 17643.

Ophthalmic assessment was not part of the original baseline assessment and was introduced as an enhancement in 2009 for 6 assessment centers which are spread across the UK (Liverpool and Sheffield in North England, Birmingham in the Midlands, Swansea in Wales, and Croydon and Hounslow in Greater London). Imaging of both eyes was performed using the Topcon 3D OCT- 1000 Mark II in a dark room without pupil dilation. The instrument takes a color photograph of the retina as well as an optical coherence tomography scan; we used the color photographs in the current study. The right eye was imaged first. Refractive status of both eyes was measured by autorefraction (Tomey RC5000; Erlangen-Tennenlohe). Spherical equivalent was calculated as the sphere + 0.5 * cylinder and participant-level refractive error was taken as the mean of right and left values.

Genetic ancestry inference

To minimize the impact of population structure, we limited our GWAS cohort to individuals of European genetic ancestry, which was defined as follows:

1. Determine the set all individuals with self-reported "British" ancestry.
2. Compute the medioid of the British ancestry set in the 15-dimensional genetic principal component (PC) space.
3. Calculate the distance of each individual in the UK Biobank to the British medioid.
4. Form the "European" set by selecting all individuals with a distance from the British medioid less than 40 (based on the 99th percentile of distances of individuals who self-identify as British or Irish).

Using this scheme, approximately 99% of all individuals with self-reported British and Irish ancestries are included, and in total slightly over 8% of all individuals are filtered. It should be noted that our cohort selection is conservative and keeps only individuals very close to the core

British ancestry, e.g., it removes over 42% of individuals with self-reported "Any other white background" ancestries. The PC plot of all individuals in the UK Biobank and the GWAS cohort is shown in Figure S4.

Ancestry determination for African, Chinese, and South Asian samples was performed in an analogous manner, using self-reported ancestry and distance from the resulting medioid of (“African”, 35), (“Chinese”, 25), (“Indian”, 35), respectively. Distances were selected based on visual inspection of the top PCs. Samples with self-reported Indian and Pakistani ancestry overlap strongly in PC space and density of distance from the medioid value, so are merged and referred to as “South Asian” in Figure S1.

Phenotype calling

After predicting VCDR for all 175,337 fundus images from 85,665 individuals in UKB, we first remove the 21,400 images which are predicted as ungradable (gradability prediction < 0.7) for VCDR. Recall that there are two imaging visits, called visit 1 and 2. We define the phenotype only based on one of these visits, because there is an approximate 5 years difference between the two visits and many factors such as age, medications, eye operations can be materially different between the two visits.

If an individual has any gradable image(s) from visit 1, we define the phenotype based on these images; otherwise, we define it based on visit 2 (a.k.a. first repeat imaging visit). For a specific visit, we first average the VCDRs of each eye and then average these per eye VCDRs if both eyes have gradable images. Moreover, to account for the impact of image gradability on the phenotype, we computed the average gradability score of all images used in defining an individual's phenotype. For the details and statistics of phenotype calling, see Figure S3. To control for the small variations in phenotype calling, we add the visit number used (i.e., 1 or 2) and the number of eyes used in calling the phenotype (i.e., 1 or 2) as covariates. After subsetting to individuals of European ancestry and removing samples with excess heterozygosity or missingness, putative sex chromosome aneuplody, and missing refractive error report, we call the VCDR phenotype for 65,680 individuals.

Genome-wide association study

We use linear mixed models as implemented in BOLT-LMM v2.3.4 (Loh et al. 2015) to account for population structure and cryptic relationships in UK Biobank, and to increase association power. We applied BOLT-LMM to all individuals of European ancestry with available VCDR who passed our sample QC and had non-missing covariates *(n*=65,680). We used sex, age at visit, visit number (i.e., 1 or 2 to indicate visit 1 or visit 2), number of eyes used to compute VCDR (i.e., 1 or 2 to indicate one eye or both eyes are used), genotyping array indicator, refractive error, average gradability scores of all fundus images used in phenotype calling and the top 15 genetic principal components as covariates. To increase association power and make the normality assumption more plausible, ML-based VCDR was rank-based inverse normal (INT; (McCaw et al. 2019)) transformed. We considered the autosomal chromosomes for our GWAS and filtered out variants with minor allele frequency (MAF) < 0.001, imputation INFO score < 0.8,

or Hardy-Weinberg equilibrium (HWE) $P < 1\times10^{-10}$ in Europeans. Using these filters, 13,110,443 variants passed QC. To verify that our association results were not driven by population stratification, we applied LD score regression (Bulik-Sullivan et al. 2015).

Identification and comparison of loci

Genome-wide significant (GWS; $P \leq 5\times10^{-8}$) lead SNPs, independent at R^2=0.1, were identified using the `plink --clump` command (v1.90b4). The reference panel comprised a random sample of 10,000 unrelated subjects of white European ancestry from the UK Biobank. Around each lead SNP, a locus was defined as the span of reference panel SNPs in LD with the lead SNP at $R^2 \geq 0.1$. For consistency with locus formation as implemented by FUMA (Watanabe et al. 2017), loci separated by fewer than 250 kb were merged, and the most significant, independent SNP in the merged locus was retained as the lead SNP. Gene context annotations were added from the GRCh37 version of GenCode v34 "comprehensive gene annotations." Only protein-coding genes and level 1 long noncoding RNAs (lncRNA) were considered.

For comparing loci across studies, loci were formed within each using the common reference panel and procedure described above. Locus overlap metrics were calculated using the GenomicRanges package (Lawrence et al. 2013) in R (v3.2.3). In comparing loci from studies A and B, it is possible for a single locus from study A to overlap multiple loci from study B and conversely. Consequently, the “overlap” operation is asymmetric, and the number of loci from study A that overlap a locus from study B may differ from the number of loci from study B that overlap a locus from study A. To make this concrete, consider the 22 loci reported by the 2017 IGGC meta-analysis (Springelkamp et al. 2017). Twenty (20) of these loci were overlapped by, and thus replicated by, the subsequent Craig *et al.* GWAS (Craig et al. 2020). However, only 19 of the Craig *et al.* loci were overlapped by an IGGC locus. This occurs because a single, larger locus from Craig *et al.* on chromosome 22 spanning base pairs 28,175,232 to 30,620,360 (lead SNP: rs6005840, P=1.9x10^{-38}) overlapped with two smaller loci from Springelkamp *et al.*: one spanning 28,195,332 to 29,447,570 (lead SNP: rs5752773, P=4.6x10^{-21}) and the second spanning 29,888,485 to 30,620,360 (lead SNP: rs1003342, P=4.3x10^{-8}). From the perspective of Craig *et al.,* 65 - 19 = 46 of the reported loci were novel, even though 20 of the IGGC’s loci were replicated.

Similarly, of the 65 loci from Craig *et al.,* 62 were replicated by our ML-based GWAS, but only 61 of the loci from the ML-based GWAS overlapped with a locus reported by Craig *et al.* A single larger locus from the ML-based analysis on chromosome 3 spanning base pairs 98,486,551 to 100,810,114 (lead SNP: rs1871794, P=9.4x10^{-30}) overlapped two smaller loci from Craig *et al.*: one spanning 98,688,022 to 99,375,069 (lead SNP: rs4928176, P=4.5x10^{-15}), and the second spanning 100,593,266 to 100,869,589 (lead SNP: rs9827694, P=8.6x10^{-10}). In both cases, two nearby loci from an earlier study collapsed into a single larger locus in the later, better powered study.

We describe a locus from an earlier study as having been replicated by a later study if it overlaps at least 1 locus from the later study. Thus, 20 loci from IGGC were replicated by Craig *et al.,* and 62 loci from Craig *et al.* were replicated by the ML-based GWAS. We describe a locus

from a later study as novel if it is not overlapped by a locus from any previous study. Among the 156 loci from the ML-based GWAS, 93 were novel, not overlapping with any locus reported by either IGGC or Craig *et al.*

Fine-mapping

Fine-mapping of independent significant loci was performed via Sum of Single Effects Regression (SuSiE; v0.9.0) (Wang et al. 2020), as implemented in R. Briefly, SuSiE identifies the likely causal variants in a region using a variational approximation to Bayesian variable selection regression. A posterior inclusion probability (PIP) is assigned to each SNP in the locus, quantifying the probability that the SNP has a non-zero effect on the outcome. The sum of PIPs for SNPs in a locus is the posterior expectation of the number of causal variants in that locus. To estimate the total number of distinct genetic signals for ML-based VCDR detected in our analysis, PIPs were aggregated across all loci where SuSiE reported no more than the number of GWS variants in the locus. Loci where SuSiE reported more causal variants than GWS variants were considered potentially unreliable. The number of causal variants in such loci was conservatively estimated as the number of GWS variants in the locus, which is potentially an underestimate. Moreover, uncommon SNPs (those with minor allele frequencies below 5%) were removed from the fine-mapping analysis, some of which are likely causal. Nevertheless, the estimated number of genetic signals for VCDR detected by our analysis was 813.

Ablation analysis

In order to assess the dependence of model quality on the training data size, we analyzed model performance when trained on progressively smaller subsets of the full training data. Predicted VCDR vs adjudicated VCDR correlations for different sets are depicted in Figure S6. In particular, when training only on 10% of the data (~7,900 samples), the Pearson's correlations (ratio with regard to the original correlation) were 0.87 (94%), 0.87 (96%) and 0.83 (93%), for the Tune, Test, and UKB Adjudicated cohorts.

We also performed a GWAS using the "10% model" predictions, which identified 131 genome-wide significant loci, replicating 123 of the 156 loci identified by the full model. The scatter plot of *P*-values for the ML-based GWAS and the 10% ablation GWAS are presented in Figure S7.

Genomic discovery power analysis

To assess how the power for genomic discovery varied with phenotyping quality, we followed the "Noisy Measurement Model" (Hormozdiari et al. 2016). Specifically, consider the following:

$$Y = X\beta + \epsilon \quad (1)$$

where Y is the true VCDR, X is genotype, and ϵ is an environmental residual. Suppose Y and X have been standardized to mean zero and variance one. Let h^2 denote the per-SNP heritability, then the residual variance is $1-h^2$. We do not observe the true VCDR, but instead a mismeasured version Y^*, which is related to Y via

(2) $$Y^* = Y + \delta$$

where δ is mean-zero measurement error. Substituting (1) into (2) gives the variance component model

(3) $$Y^* = X\beta + \epsilon + \delta$$

From model (3) we can derive the asymptotic non-centrality parameter (NCP) of the standard Wald χ^2 test of association by considering a sequence of contiguous alternatives (Serfling 1980). The NCP for the χ^2 test based on *Y** takes the simple form

$$NCP = n \cdot \frac{\rho^2 h^2}{1 - \rho^2 h^2}$$

where *n* is the sample size, ρ^2 is the square of the correlation between the mismeasured *Y** and true *Y* phenotypes, and h^2 is the true heritability of *Y*. Power and Non-Centrality Curves as a function of per-SNP heritability, stratified by the correlation between the measured and true phenotypes are shown in Figure S8.

Applying the above model to compare the “10% model” and the model trained on the entire training set, at our GWAS sample size *n*=65,680, the difference in power between a GWAS where the correlation between the observed and true VCDR measurements is 0.89 and a GWAS where the correlation is 0.83 can reach as high as 15%.

Replication slope analysis

To jointly test the ML-based hits for replication of the IGGC VCDR meta-analysis, we first scaled the effect size estimates of the ML-based GWAS results to account for winner’s curse. Winner’s curse correction was performed by fitting a two-component Gaussian mixture model, as described in supplemental section 5.3 of (Turley et al. 2018):

$$f(\hat{\beta}_j|\pi, \tau^2) = \pi \cdot N(\hat{\beta}_j|0, \sigma_j^2) + (1 - \pi) \cdot N(\hat{\beta}_j|0, \sigma_j^2 + \tau^2).$$

Here $\hat{\beta}_j$ is the estimated effect size, π is the prior probability of belonging to the null component, σ_j^2 is the sampling variance (i.e., squared standard error) of $\hat{\beta}_j$, and τ^2 is the variance in effect sizes at non-null SNPs. Model parameters (π, τ^2) were estimated by maximum likelihood using the expectation maximization algorithm (McCaw, Julienne, and Aschard 2020; Meng and Rubin 1993). The observed effect sizes were shrunk to their posterior expectation via:

$$E[\beta_j|\hat{\beta}_j] = (1 - \gamma_j)\frac{\tau^2}{\tau^2 + \sigma_j^2} \cdot \hat{\beta}_j$$

where γ_j is the posterior responsibility of the null-component for SNP *j*:

$$\gamma_i = P[\beta_j = 0|\hat{\beta}_j] = \frac{\pi \cdot N(\hat{\beta}_j|0, \sigma_j^2)}{\pi \cdot N(\hat{\beta}_j|0, \sigma_j^2) + (1-\pi) \cdot N(\hat{\beta}_j|0, \sigma_j^2 + \tau^2)}.$$

Winner's curse correction was performed using all genotyped variants as input, with final model parameter estimates of $(\pi = 0.958, \tau^2 = 0.000427)$. We then identified 214 (of 299) ML-based hits additionally present in the IGGC VCDR meta-analysis and regressed the IGGC effect sizes on the winner's curse-corrected ML-based GWAS effect size estimates (Figure S9).

Meta-analysis

GWAS summary statistics for ML-based VCDR were combined with summary statistics from a previous meta-analysis of VCDR by the International Glaucoma Genetics Consortium (IGGC) using Meta-Soft (Han and Eskin 2011). The following strategy was adopted for selecting the final *P*-value. At each SNP, the I^2 statistic (Higgins and Thompson 2002) was calculated to quantify the proportion of total variation across studies that was attributable to effect size heterogeneity. For SNPs with $I^2 > 0$, a random effects meta-analysis was performed, using Han and Eskin's "RE2" model (see URLs), whereas for those SNPs with $I^2 = 0$, a fixed effects meta-analysis was performed. Selecting whether to perform random or fixed effects meta-analysis on the basis of I^2 is an effort to apply the most appropriate model for the observed effect sizes. Among the 8.6M SNPs present in both studies, 68% had $I^2 = 0$, while the remaining 32% had $I^2 > 0$. For the 4.5M SNPs present in our analysis but not in IGGC, the original *P*-value from the ML-based VCDR GWAS was retained. The S-LDSC intercept was 1.06 (s.e.m=0.01) and the SNP-heritability h2g was 0.37 (s.e.m=0.02).

Functional analyses with FUMA and GREAT

Functional analyses were performed in FUMA (Watanabe et al. 2017). We assigned each variant to the nearest gene within 10kb using FUMA's "SNP2GENE" functionality, and performed gene-set enrichment analysis using FUMA's "GENE2FUNC" functionality. In both cases, we adopted the default parameter settings. We compared the *relative enrichment* of gene sets that were significant according to both the ML-based and the Craig *et al.* GWAS of VCDR. Specifically, enrichment refers to the odds that a gene in the gene-set was detected in a given GWAS, and relative enrichment is the odds ratio comparing our GWAS with the Craig et al. GWAS. Fisher's exact test was applied to determine whether enrichment differed significantly between the two studies. Those sets where the relative enrichment (odds ratio) exceeds 1 represent biologically interesting gene-sets where the ML-based GWAS captured more of the constituent genes.

GREAT enrichment analyses were performed on the human GRCh37 assembly using GREAT v4.0.4 (McLean et al. 2010). The default "basal+extension" region-gene association rule was used with 5 kb upstream, 1 kb downstream, 1000 kb extension, and curated regulatory domains included. Analyses were performed using the same loci as in the FUMA analyses described above; 65 loci from Craig *et al.* and 156 loci from the ML-based GWAS. Terms were considered

statistically significant if the Bonferroni-corrected *P*-values for both the region-based and gene-based tests were ≤ 0.05.

Phenome-wide association study (PheWAS) using OpenTargets

PheWAS analyses were performed using the OpenTargets website (see URLs) for all 299 ML-based VCDR hits. For any given variant (e.g., GWAS hit), OpenTargets reports a set of phenotypes that are nominally significant ($P < 0.05$) for the variant. This analysis produced 62,753 (variant, phenotype) pairs, of which 974 pass a Bonferroni-corrected threshold ($P < 0.05/(299 * 4645)$) where 4,645 is the total number of phenotypes in OpenTargets. We observed that 314 out of 974 significant (variant, phenotype) pairs are classified in the "Anthropometric measurement" trait category and 101 of 974 are in the "Eye measurement" category. All nominally significant pairs are reported in Table S10.

VCDR-IOP Mendelian Randomization

Two sample Mendelian randomization (MR) for the association between intraocular pressure (IOP) and ML-based VCDR was performed using the TwoSampleMR (see **URLs**) package in R (4.0.2). Among the 187 independent significant SNPs for IOP from (Khawaja et al. 2018), 183 remained after harmonizing with ML-based VCDR. This provided 183 candidate instrumental variables for quantifying the association between IOP and ML-based VCDR. Based on Cochran's Q test, there was significant evidence of pleiotropy ($P<10^{-16}$). Therefore, per-SNP associations were meta-analyzed using Egger regression (Egger et al. 1997), which is robust to the exclusion restriction (Bowden et al. 2017). The Egger intercept did not differ from zero (intercept=0.001, P=0.69). The Egger slope of 0.07 ($P=4\times10^{-4}$) provided strong evidence of a directional association between IOP and ML-based VCDR. In a reversed analysis, regarding ML-based VCDR as the mediator and IOP as the outcome, the Egger slope was -0.03 (P=0.75), providing no significant evidence of association in the opposite direction.

Polygenic VCDR Model

Pruning and thresholding

Pruning and thresholding-based polygenic risk scores for VCDR were computed as the weighted sum of effect allele counts for independent genome-wide significant variants ($P\leq5\times10^{-8}$), where the weight of each variant was its estimated effect size from the GWAS (Chatterjee, Shi, and García-Closas 2016). To evaluate performance both within the UK Biobank and in the EPIC-Norfolk cohorts, index variants present in both cohorts were used in PRS creation, resulting in 58 of the 76 published variants from Craig *et al.* GWAS and 282 of the 299 index variants from the ML-based GWAS. The UK Biobank evaluation set consisted of adjudicated expert-annotated VCDR measurements in 2,076 individuals of European ancestry. The EPIC-Norfolk evaluation set consisted of scanning laser ophthalmoscopy (HRT)-measured VCDRs in 5,868 individuals.

Elastic net

Elastic net-based polygenic risk scores for VCDR were trained using the ML-predicted VCDR as the target label in 62,969 individuals using scikit-learn (Pedregosa et al. 2011). The Craig *et al.* model used 76 variants (the 58 described in the pruning and thresholding section above, plus 18 proxy variants present in both UK Biobank and EPIC-Norfolk that were in highest linkage disequilibrium ($R^2 \geq 0.6$) with the 18 dropped Craig *et al.* variants) and the ML-based model used the same 282 variants as described above. Each model was trained with 5-fold cross-validation and L1-penalty ratios of [0.1, 0.5, 0.7, 0.9, 0.95, 0.99, 1.0]. Model evaluation was performed in the same evaluation sets as described above. Both the UK Biobank and EPIC-Norfolk test sets were scored using the `plink --score` command and the correlations were computed using the scores in the resulting `*.profile` files.

Permutation *P*-values

A permutation test was applied to assess whether a polygenic risk score (PRS) trained using summary statistics from the ML-based GWAS significantly outperformed a PRS trained using summary statistics from the Craig *et al.* GWAS for predicting VCDR in the UK Biobank and EPIC-Norfolk cohorts. Phenotypic predictions were generated from both PRS. The test statistic was the difference in Pearson correlations between the observed and predicted phenotypes, comparing ML-based with Craig *et al.* A value exceeding zero indicates better performance by the ML-based PRS. Under the null hypothesis, the predictions from both PRS are exchangeable. To obtain a realization from the null distribution, for each subject, the predictions of the ML-based and Craig *et al.* PRS were randomly swapped, and the difference in correlations was recalculated. This procedure was repeated 10^5 times to obtain the null distribution. The one-sided *P*-value is given by the proportion of realizations from the null distribution that were as or more extreme than the observed difference in correlations.

Glaucoma Association

Mediation Analysis

A mediation analysis was performed to estimate the association between ML-based VCDR and glaucoma, as assessed by Gharahkhani *et al* (Gharahkhani et al. 2020). MR is a special case of mediation analysis in which the SNPs have no direct effect on the outcome; that is, the effect of genotype on the phenotype passes entirely through the mediator. Our mediation analysis differs from MR in that, due to limited availability of summary statistics from Gharahkhani *et al*, the SNP set was defined based on association with the mediator (ML-based VCDR) rather than the outcome (glaucoma). Among the 118 independent, significant glaucoma SNPs identified by Gharahkhani *et al*, 116 remained after harmonizing with the VCDR summary statistics available from our study. As expected, Cochran's Q test provided strong evidence of pleiotropy ($P<10^{-16}$), and the Egger intercept of 0.04 ($P=7\times10^{-7}$) suggested that variants with tended to increase VCDR also tended to increase the odds of glaucoma via an alternative pathway. The Egger

slope was 5.7 (P=3×10^{-3}; Figure S16), which is interpreted as a log odds ratio, provides substantial evidence that increased VCDR was associated with increased glaucoma odds. This estimate of the association between VCDR and glaucoma remains valid, despite the presence of pleiotropy, since Egger regression is robust to the exclusion restriction (Bowden et al. 2017). In a reversed analysis using the same set of candidate SNPs, but regarding glaucoma as the mediator and VCDR as the outcome, the Egger intercept was 0.00 (P=0.09), and the Egger slope was 0.02 (P=0.07), providing no strong evidence of association in the opposite direction.

Glaucoma liability conditional analysis

One of the main advantages of the ML-based model is that we can apply our ML-based model to different phenotypes without additional cost. We computed the glaucoma liability (ML-based glaucoma) for the same set of individuals in UK Biobank for whom we had calculated VCDR as described above. We performed GWAS on glaucoma liability (logit scale of glaucoma probability), using BOLT-LMM, conditional on ML-based VCDR and all covariates included in the ML-based VCDR GWAS. The LD score regression intercept was 1.00 (SE=0.001), with a SNP-heritability of 0.06 (0.01). Moreover, QQ-plot is depicted in Figure S17.

UK Biobank glaucoma phenotype

UK Biobank participants who underwent an ophthalmic examination also completed an ophthalmic touchscreen questionnaire and were considered to have POAG if they responded "Glaucoma" to the question "Has a doctor told you that you have any of the following problems with your eyes?". Participants were also considered to have POAG if they had a recorded hospital episode statistic ICD 10 code for POAG (H40.1). Controls were defined as participants who underwent the ophthalmic touchscreen questionnaire but did not meet the criteria to be a case. Additionally, we excluded participants with an ICD 9/10 hospital episode statistic code for types of glaucoma types other than POAG (ICD 9: 365.*; ICD 10: H40.0, H40.2, H40.3, H40.4, H40.5, H40.6, H40.8, H40.9, H42.*), participants meeting the case criteria but reporting an age of glaucoma onset prior to 30 years, and participants reporting glaucoma laser treatment or eye surgery but not reporting glaucoma on the touchscreen questionnaire. Applying these criteria, there were 7,654 cases and 182,726 controls.

EPIC-Norfolk cohort

The European Prospective Investigation into Cancer (EPIC) study is a pan-European prospective cohort study designed to investigate the etiology of major chronic diseases (Riboli and Kaaks 1997). EPIC-Norfolk, one of the UK arms of EPIC, recruited and examined 25,639 participants between 1993 and 1997 for the baseline examination (Day et al. 1999). Recruitment was via general practices in the city of Norwich and the surrounding small towns and rural areas, and methods have been described in detail previously (Hayat et al. 2014). Since virtually all residents in the UK are registered with a general practitioner through the National Health Service, general practice lists serve as population registers. Ophthalmic

assessment formed part of the third health examination and this has been termed the EPIC-Norfolk Eye Study (Khawaja et al. 2013).

In total, 8,623 participants were seen for the Eye Study between 2004 and 2011. Ophthalmic examination included tonometry (Ocular Response Analyzer; Reichert, New York, USA; software V.3.01), optic disc photography (Nikon D80 camera; Nikon Corporation, Tokyo, Japan), scanning laser ophthalmoscopy (Heidelberg Retinal Tomograph 3; Heidelberg Engineering, Heidelberg, Germany) and nerve fiber layer assessment (GDx-VCC; Zeiss, Dublin, California, USA). Participants meeting pre-defined criteria and an additional 1:10 participants underwent automated visual field testing (Humphrey 750i Visual Field Analyzer; Carl Zeiss Meditech Ltd, Welwyn Garden City, UK). 99.7% of EPIC-Norfolk are of European descent. The EPIC-Norfolk Eye Study was carried out following the principles of the Declaration of Helsinki and the Research Governance Framework for Health and Social Care. The study was approved by the Norfolk Local Research Ethics Committee (05/Q0101/191) and East Norfolk & Waveney NHS Research Governance Committee (2005EC07L). All participants gave written, informed consent.

Ascertainment of POAG in the EPIC Norfolk third health examination has been described previously (Chan et al. 2017). In brief, participants with study results suspicious of glaucoma (using pre-defined criteria) were referred for further examination by a glaucoma specialist at the regional University Hospital (Khawaja et al. 2013). Additionally, a diagnosis refinement process was undertaken by a second glaucoma specialist who independently reviewed the test results of all participants classified as glaucoma and a proportion of participants who were not classified as having glaucoma. POAG was defined as the presence of a glaucomatous optic disc together with either a corresponding visual field defect or otherwise unexplained non-specific visual field loss, open angles on gonioscopy, and absence of secondary causes of glaucoma. A glaucomatous disc was defined as one with focal or diffuse neuro-retinal rim thinning, and may possess, though not necessary for the definition, additional characteristic features such as bared circumlinear vessels, disc hemorrhages or nerve fiber layer defects. Pseudoexfoliative and pigmentary glaucoma were defined as secondary glaucoma in this study and therefore did not contribute to POAG cases. We defined controls as participants not meeting referral criteria for glaucoma on initial ophthalmic assessment and participants who attended the University Hospital for further examination and were not classified as having or being suspect for any type of glaucoma or ocular hypertension.

Initial genotyping on a small subset of EPIC-Norfolk was undertaken using the Affymetrix GeneChip Human Mapping 500K Array Set and 1,096 of these participants contributed to the IGGC meta-analysis (Springelkamp et al. 2017). Subsequently, the rest of the EPIC-Norfolk cohort were genotyped using the Affymetrix UK Biobank Axiom Array (the same array as used in UK Biobank); it is 5,868 of these participants (which includes no overlap with the 1,096 participants contributing to the IGGC meta-analysis) that contributed to the EPIC-Norfolk analyses in the current study. SNP exclusion criteria included: call rate < 95%, abnormal cluster pattern on visual inspection, plate batch effect evident by significant variation in minor allele frequency, and/or Hardy-Weinberg equilibrium $P < 10^{-7}$. Sample exclusion criteria included:

DishQC < 0.82 (poor fluorescence signal contrast), sex discordance, sample call rate < 97%, heterozygosity outliers (calculated separately for SNPs with minor allele frequency >1% and <1%), rare allele count outlier, and impossible identity-by-descent values. We removed individuals with relatedness corresponding to third-degree relatives or closer across all genotyped participants. Following these exclusions, there were no ethnic outliers. Imputation was carried out using the HRC v1.

Quality control for HRT3 images included requiring a topography SD > 40 µm and checking of the manually drawn optic disc margin contours by an ophthalmologist (with redrawing if necessary). The mean HRT3 VCDR of right and left eyes was considered as the participant's VCDR if good quality scans were available for both eyes. If a good quality scan was only available for one eye, the VCDR value for that eye was considered for the participant.

Model hyper-parameters

The hyper-parameters used very closely follow (Krause et al. 2018; Phene et al. 2019):

- Inception V3 architecture initialized with pretrained ImageNet weights
- Weight decay: an L2 kernel regularization penalty of 0.00004 applied to all 2D convolution and dense layers
- Input image resolution: 587 x 587
- Learning rate: 0.001 with an exponential decay rate of 0.99 every two epochs
- Optimizer: Adam optimizer with a 1st moment exponential decay rate, i.e., `beta_1`, of 0.9, a 2nd moment exponential decay rate, i.e., `beta_2`, of 0.999, and an `epsilon` of 0.1
- Model averaging: checkpoints were taken using a moving average of trainable parameters in the network, i.e., `tfa.optimizers.MovingAverage`, with an average decay of 0.9999
- Maximum training steps: The model was trained for 250,000 steps
- Early stopping checkpoint frequency: Model evaluation occurred every 500 steps and checkpoints monitored VCDR MSE
- Batch size: 16
- Data augmentation:
    - Random horizontal and vertical reflections
    - Random brightness changes (with a max delta of 0.1147528) [`tf.image.random_brightness`]
    - Random saturation changes between 0.5597273 and 1.2748845 [`tf.image.random_saturation`]
    - Random hue changes (with a max delta of 0.0251488) [`tf.image.random_hue`]
    - Random contrast changes between 0.9996807 and 1.7704824 [`tf.image.random_constrast`]

Full details are available in the `learning/` section of the associated open-source repository: https://github.com/Google-Health/genomics-research/tree/main/ml-based-vcdr

References

Abadi, Martín, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, et al. 2016. “TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems.” *arXiv [cs.DC]*. arXiv. http://arxiv.org/abs/1603.04467.

Age-Related Eye Disease Study Research Group. 1999. “The Age-Related Eye Disease Study (AREDS): Design Implications. AREDS Report No. 1.” *Controlled Clinical Trials* 20 (6): 573–600.

Bulik-Sullivan, Brendan K., Po-Ru Loh, Hilary K. Finucane, Stephan Ripke, Jian Yang, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Nick Patterson, Mark J. Daly, Alkes L. Price, and Benjamin M. Neale. 2015. “LD Score Regression Distinguishes Confounding from Polygenicity in Genome-Wide Association Studies.” *Nature Genetics* 47 (3): 291–95.

Chan, Michelle P. Y., David C. Broadway, Anthony P. Khawaja, Jennifer L. Y. Yip, David F. Garway-Heath, Jennifer M. Burr, Robert Luben, et al. 2017. “Glaucoma and Intraocular Pressure in EPIC-Norfolk Eye Study: Cross Sectional Study.” *BMJ* 358 (September): j3889.

Chatterjee, Nilanjan, Jianxin Shi, and Montserrat García-Closas. 2016. “Developing and Evaluating Polygenic Risk Prediction Models for Stratified Disease Prevention.” *Nature Reviews. Genetics* 17 (7): 392–406.

Day, N., S. Oakes, R. Luben, K. T. Khaw, S. Bingham, A. Welch, and N. Wareham. 1999. “EPIC-Norfolk: Study Design and Characteristics of the Cohort. European Prospective Investigation of Cancer.” *British Journal of Cancer* 80 Suppl 1 (July): 95–103.

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. “ImageNet: A Large-Scale Hierarchical Image Database.” *2009 IEEE Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.2009.5206848.

Gharahkhani, Puya, Eric Jorgenson, Pirro Hysi, Anthony P. Khawaja, Sarah Pendergrass, Xikun Han, Jue Sheng Ong, et al. 2020. “A Large Cross-Ancestry Meta-Analysis of Genome-Wide Association Studies Identifies 69 Novel Risk Loci for Primary Open-Angle Glaucoma and Includes a Genetic Link with Alzheimer’s Disease.” *bioRxiv*. https://doi.org/10.1101/2020.01.30.927822.

Han, Buhm, and Eleazar Eskin. 2011. “Random-Effects Model Aimed at Discovering Associations in Meta-Analysis of Genome-Wide Association Studies.” *American Journal of Human Genetics* 88 (5): 586–98.

Hayat, Shabina A., Robert Luben, Victoria L. Keevil, Stephanie Moore, Nichola Dalzell, Amit Bhaniani, Anthony P. Khawaja, et al. 2014. “Cohort Profile: A Prospective Cohort Study of Objective Physical and Cognitive Capability and Visual Health in an Ageing Population of Men and Women in Norfolk (EPIC-Norfolk 3).” *International Journal of Epidemiology* 43 (4): 1063–72.

Higgins, Julian P. T., and Simon G. Thompson. 2002. “Quantifying Heterogeneity in a Meta-Analysis.” *Statistics in Medicine* 21 (11): 1539–58.

Hormozdiari, Farhad, Eun Yong Kang, Michael Bilow, Eyal Ben-David, Chris Vulpe, Stela McLachlan, Aldons J. Lusis, Buhm Han, and Eleazar Eskin. 2016. “Imputing Phenotypes for Genome-Wide Association Studies.” *American Journal of Human Genetics* 99 (1): 89–103.

Khawaja, Anthony P., Michelle P. Y. Chan, Shabina Hayat, David C. Broadway, Robert Luben, David F. Garway-Heath, Justin C. Sherwin, et al. 2013. “The EPIC-Norfolk Eye Study: Rationale, Methods and a Cross-Sectional Analysis of Visual Impairment in a Population-Based Cohort.” *BMJ Open* 3 (3). https://doi.org/10.1136/bmjopen-2013-002684.

Krause, Jonathan, Varun Gulshan, Ehsan Rahimy, Peter Karth, Kasumi Widner, Greg S. Corrado, Lily Peng, and Dale R. Webster. 2018. “Grader Variability and the Importance of Reference Standards for Evaluating Machine Learning Models for Diabetic Retinopathy.” *Ophthalmology* 125 (8): 1264–72.

Lawrence, Michael, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. 2013. “Software for Computing and Annotating Genomic Ranges.” *PLoS Computational Biology* 9 (8): e1003118.

Loh, Po-Ru, George Tucker, Brendan K. Bulik-Sullivan, Bjarni J. Vilhjálmsson, Hilary K. Finucane, Rany M. Salem, Daniel I. Chasman, et al. 2015. “Efficient Bayesian Mixed-Model Analysis Increases Association Power in Large Cohorts.” *Nature Genetics* 47 (3): 284–90.

McCaw, Zachary R., Hanna Julienne, and Hugues Aschard. 2020. “MGMM: An R Package for Fitting Gaussian Mixture Models on Incomplete Data.” *Cold Spring Harbor Laboratory*. https://doi.org/10.1101/2019.12.20.884551.

McCaw, Zachary R., Jacqueline M. Lane, Richa Saxena, Susan Redline, and Xihong Lin. 2019. “Operating Characteristics of the Rank-Based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies.” *Biometrics*, December. https://doi.org/10.1111/biom.13214.

McLean, Cory Y., Dave Bristor, Michael Hiller, Shoa L. Clarke, Bruce T. Schaar, Craig B. Lowe, Aaron M. Wenger, and Gill Bejerano. 2010. “GREAT Improves Functional Interpretation of Cis-Regulatory Regions.” *Nature Biotechnology* 28 (5): 495–501.

Meng, Xiao-Li, and Donald B. Rubin. 1993. “Maximum Likelihood Estimation via the ECM Algorithm: A General Framework.” *Biometrika* 80 (2): 267–78.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. “Scikit-Learn: Machine Learning in Python.” *Journal of Machine Learning Research: JMLR* 12 (85): 2825–30.

Phene, Sonia, R. Carter Dunn, Naama Hammel, Yun Liu, Jonathan Krause, Naho Kitade, Mike Schaekermann, et al. 2019. “Deep Learning and Glaucoma Specialists.” *Ophthalmology* 126 (12): 1627–39.

Prechelt, Lutz. 1998. “Early Stopping - But When?” In *Neural Networks: Tricks of the Trade*, edited by Genevieve B. Orr and Klaus-Robert Müller, 55–69. Berlin, Heidelberg: Springer Berlin Heidelberg.

Riboli, E., and R. Kaaks. 1997. “The EPIC Project: Rationale and Study Design. European Prospective Investigation into Cancer and Nutrition.” *International Journal of Epidemiology* 26 Suppl 1: S6–14.

Serfling, Robert J. 1980. *Approximation Theorems of Mathematical Statistics*. Wiley.

Shorten, Connor, and Taghi M. Khoshgoftaar. 2019. “A Survey on Image Data Augmentation for Deep Learning.” *Journal of Big Data* 6 (1): 60.

Springelkamp, Henriët, Adriana I. Iglesias, Aniket Mishra, René Höhn, Robert Wojciechowski, Anthony P. Khawaja, Abhishek Nag, et al. 2017. “New Insights into the Genetics of Primary Open-Angle Glaucoma Based on Meta-Analyses of Intraocular Pressure and Optic Disc Characteristics.” *Human Molecular Genetics* 26 (2): 438–53.

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. “Rethinking the Inception Architecture for Computer Vision.” *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/cvpr.2016.308.

Turley, Patrick, Raymond K. Walters, Omeed Maghzian, Aysu Okbay, James J. Lee, Mark Alan Fontana, Tuan Anh Nguyen-Viet, et al. 2018. “Multi-Trait Analysis of Genome-Wide Association Summary Statistics Using MTAG.” *Nature Genetics* 50 (2): 229–37.

Wang, Gao, Abhishek Sarkar, Peter Carbonetto, and Matthew Stephens. 2020. “A Simple New

Approach to Variable Selection in Regression, with Application to Genetic Fine Mapping." *Journal of the Royal Statistical Society. Series B, Statistical Methodology* 25 (July): 1.
Watanabe, Kyoko, Erdogan Taskesen, Arjen van Bochoven, and Danielle Posthuma. 2017. "Functional Mapping and Annotation of Genetic Associations with FUMA." *Nature Communications* 8 (1): 1826.