THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Statistical Genetic Approaches in Eye Disease

Xikun Han

Bachelor of Medicine (BMed), Master of Science (MSc)

A thesis submitted for the degree of Doctor of Philosophy at
The University of Queensland in 2021
Faculty of Medicine

Abstract

Glaucoma and age-related macular degeneration (AMD), two leading causes of blindness and vision loss, are highly heritable and have important genetic basis. In this thesis, I aim to identify novel genes associated with the two major eye diseases and other related eye traits (*i.e.* glaucoma endophenotypes: intraocular pressure [IOP] and vertical cup-to-disc ratio [VCDR]), improve our understanding of their genetic architecture, provide new insight into the underlying biological mechanisms, uncover potential causal risk factors, as well as facilitate the development of personalized approaches for prevention and treatment of eye diseases.

In the introduction chapter, I reviewed some key concepts for complex traits. Understanding the genetic architecture of complex traits is one of the main themes of statistical genetics and lays the foundation for theoretical and practical studies. Following that, two important statistical genetic approaches were introduced. Firstly, the use of polygenic risk score (PRS) to construct a predictive tool based on individual genetic variants (single nucleotide polymorphism [SNP]), followed by Mendelian randomization for causal inference using SNPs as genetic instrumental variables. The subsequent sections provide an introduction of two main eye diseases investigated in the thesis, glaucoma and AMD.

In chapter 2, I evaluated the penetrance and risk effect of Myocilin gene p.Gln368Ter variant (rs74315329), one of the most common “monogenic” variants, for glaucoma risk in large population-based and registry-based studies. This study shows that approximately 50% of *MYOC* p.Gln368Ter carriers aged 65 and over had glaucoma or ocular hypertension, with an even higher prevalence observed in Australian registry-based studies. These findings provide evidence to support early detection and monitoring of p.Gln368Ter variant carriers, and to direct at-risk individuals to appropriate management.

In Chapter 3, I performed a large-scale multi-trait analysis of glaucoma and its endophenotypes (IOP and VCDR) to identify novel glaucoma risk loci, leveraging the high genetic correlation between the two traits with glaucoma. Based on the genetic discoveries, I built, validated, and evaluated the utility of the newly derived glaucoma PRS in various populations and clinical data sets. This study shows glaucoma PRS is predictive of

increased risk of glaucoma, earlier age of glaucoma diagnosis, equivalent risk of high penetrance variant, increased probability of progression, and requirement of treatment. These findings on glaucoma demonstrate how PRS enables effective risk stratification and would facilitate the development of personalized methods for prevention and treatment.

In Chapter 4, I conducted large-scale genome-wide association studies (GWAS) to enhance our understanding of the genetics of optic disc morphology. Specifically, in Chapter 4A, I carried out a large GWAS for vertical disc diameter, tripling the previously studied sample size. A subsequent study in Chapter 4B is to apply deep learning algorithms to optic nerve head photographs, which allows automated labelling of vertical disc diameter and VCDR, and enables systematic evaluation of the distribution of optic nerve head parameters and glaucoma risk across different ancestries. The automated labelling dramatically increases SNP-based heritability (indicating more accurate phenotyping than clinician gradings), and genetic discovery, with more than 200 loci for both VCDR and vertical disc diameter (doubled the number of loci from previous studies), and many of the novel VCDR loci also conferring risk for glaucoma.

In Chapter 5, I presented the work to map new AMD risk genes from a large meta-analysis including International AMD Genomics Consortium data, Genetic Epidemiology Research on Aging study, UK Biobank, and FinnGen study. In this study, 69 independent genome-wide significant SNPs were identified, 12 of which were novel. From further functional annotations, we found that most of the novel genes are expressed in the retina and potentially involved in the pathways of AMD pathogenesis.

Chapter 6 is a continuation of the work from Chapter 5 to evaluate the association between inflammatory and lipid biomarkers and AMD risk. Previous studies have shown the presence of complement, inflammatory factors, and lipids in drusen, the hallmark lesions of AMD. Genetic studies also indicated the importance of pathways involved in complement cascade, high-density lipoprotein particle remodelling, and cholesterol transporter activity. In this chapter, I investigated the potential causal associations between serum C-reactive protein levels and lipid biomarkers with AMD risk using Mendelian randomization approaches. The findings provide evidence to support associations of these biomarkers with the risk of AMD.

In the final discussion chapter, I reviewed a series of pertinent questions in the genetic risk profiling of glaucoma (Chapter 7A), provided an overview of recent advances about the genetics of glaucoma, as well as its endophenotypes, and discussed what the prospects are for glaucoma genetic risk predictions. Finally, I highlighted the key findings from the work in this thesis and discussed the challenges and futures of fine-mapping causal variants and translating genetic risk predictions into clinical care.

Declaration by author

This thesis is composed of my original work, and contains no material previously published or written by another person except where due reference has been made in the text. I have clearly stated the contribution by others to jointly-authored works that I have included in my thesis.

I have clearly stated the contribution of others to my thesis as a whole, including statistical assistance, survey design, data analysis, significant technical procedures, professional editorial advice, financial support and any other original research work used or reported in my thesis. The content of my thesis is the result of work I have carried out since the commencement of my higher degree by research candidature and does not include a substantial part of work that has been submitted to qualify for the award of any other degree or diploma in any university or other tertiary institution. I have clearly stated which parts of my thesis, if any, have been submitted to qualify for another award.

I acknowledge that an electronic copy of my thesis must be lodged with the University Library and, subject to the policy and procedures of The University of Queensland, the thesis be made available for research and study in accordance with the Copyright Act 1968 unless a period of embargo has been approved by the Dean of the Graduate School.

I acknowledge that copyright of all material contained in my thesis resides with the copyright holder(s) of that material. Where appropriate I have obtained copyright permission from the copyright holder to reproduce material in this thesis and have sought permission from co-authors for any jointly authored works included in the thesis.

Publications included in this thesis

First or joint-first authored publications that have been incorporated into my thesis are listed below:

1. **Han X.**[†], Souzeau E.[†], Ong J. S., An J., Siggs O. M., Burdon K. P., Best S., Goldberg I., Healey P. R., Graham S. L., Ruddle J. B., Mills R. A., Landers J., Galanopoulos A., White A. J. R., Casson R., Mackey D. A., Hewitt A. W., Gharahkhani P., Craig J. E., MacGregor S[*]. Myocilin Gene Gln368Ter Variant Penetrance and Association With Glaucoma in Population-Based and Registry-Based Studies. ***JAMA Ophthalmology***. 2019;137(1):28-35.
   -- Incorporated as Chapter 2

2. 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., Igo R. P., Jr., Graham S. L., Healey P. R., Ong J. S., Zhou T., Siggs O., Law M. H., Souzeau E., Ridge B., Hysi P. G., Burdon K. P., Mills R. A., Landers J., Ruddle J. B., Agar A., Galanopoulos A., White A. J. R., Willoughby C. E., Andrew N. H., Best S., Vincent A. L., Goldberg I., Radford-Smith G., Martin N. G., Montgomery G. W., Vitart V., Hoehn R., Wojciechowski R., Jonas J. B., Aung T., Pasquale L. R., Cree A. J., Sivaprasad S., Vallabh N. A., consortium Neighborhood, Eye U. K. Biobank, Vision Consortium, Viswanathan A. C., Pasutto F., Haines J. L., Klaver C. C. W., van Duijn C. M., Casson R. J., Foster P. J., Khaw P. T., Hammond C. J., Mackey D. A., Mitchell P., Lotery A. J., Wiggs J. L., Hewitt A. W., MacGregor S. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. ***Nature Genetics.*** 2020;52(2):160-166.
   -- Incorporated as Chapter 3

3. **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., Mackey D. A., Gharahkhani P., Khawaja A. P., Hewitt A. W., Craig J. E., MacGregor S. Genome-wide association analysis of 95 549 individuals identifies novel loci and genes influencing optic disc morphology. ***Human Molecular Genetics***. 2019;28(21):3680-3690.
   -- Incorporated as Chapter 4A

4. **Han X.**[†*], Steven K., Qassim A., Marshall H., Bean C., Tremeer M., An J., Siggs O., Gharahkhani P., Craig J. E., Hewitt A. W., Trzaskowski M., MacGregor S. Automated AI labelling of optic nerve head enables new insights into cross-ancestry glaucoma risk and genetic discovery in >280,000 images from UKB and CLSA. ***American Journal of Human Genetics.*** 2021; 108(7):1204-1216.
   -- Incorporated as Chapter 4B

5. **Han X.**[†*], Gharahkhani P., Mitchell P., Liew G., Hewitt A. W., MacGregor S. Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration. ***Journal of Human Genetics***. 2020; 65, 657–665.
   -- Incorporated as Chapter 5

6. **Han X.**[†*], Ong J. S., An J., Hewitt A. W., Gharahkhani P., MacGregor S. Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration. ***European Journal of Epidemiology***. 2020;35(2):139-146.
   -- Incorporated as Chapter 6A

7. **Han X.**[†*], Ong J. S., Hewitt A. W., Gharahkhani P., MacGregor S. The effects of eight serum lipid biomarkers on age-related macular degeneration risk: a Mendelian randomization study. ***International Journal of Epidemiology***. 2020; 50(1):325-336.
   -- Incorporated as Chapter 6B

8. **Han X.**[†*], Hewitt A. W., MacGregor S. Predicting the future of genetic risk profiling of glaucoma: a narrative review. ***JAMA Ophthalmology***. 2021;139(2):224-231.
   -- Incorporated as Chapter 7A

[†]First or joint-first author; [*]corresponding author.

Submitted manuscripts included in this thesis

No manuscripts submitted for publication.

Other publications during candidature

1. **Han X.**[†*], Ong J. S., An J., Craig J. E., Gharahkhani P., Hewitt A. W., MacGregor S. Association of Myopia and Intraocular Pressure With Retinal Detachment in European Descent Participants of the UK Biobank Cohort: A Mendelian Randomization Study. ***JAMA Ophthalmology***. 2020;138(6):671-8.

2. **Han X.**,[†*] Lee SS.,[*] Ingold N., McArdle N., Khawaja AP., MacGregor S., Mackey DA. Associations of sleep apnoea with glaucoma and age-related macular degeneration: an analysis in the United Kingdom Biobank and the Canadian Longitudinal Study on Aging. ***BMC Medicine.*** 2021; 19(1):104.

3. MacGregor S., Ong J. S., An J., **Han X.**, Zhou T., Siggs O. M., Law M. H., Souzeau E., Sharma S., Lynn D. J., Beesley J., Sheldrick B., Mills R. A., Landers J., Ruddle J. B., Graham S. L., Healey P. R., White A. J. R., Casson R. J., Best S., Grigg J. R., Goldberg I., Powell J. E., Whiteman D. C., Radford-Smith G. L., Martin N. G., Montgomery G. W., Burdon K. P., Mackey D. A., Gharahkhani P., Craig J. E., Hewitt A. W. Genome-wide association study of intraocular pressure uncovers new pathways to glaucoma. ***Nature Genetics.*** 2018;50(8):1067-1071.

4. Ong JS, An J, **Han X**, Law MH, Nandakumar P; 23andMe Research team; Esophageal cancer consortium, Schumacher J, Gockel I, Bohmer A, Jankowski J, Palles C, Olsen CM, Neale RE, Fitzgerald R, Thrift AP, Vaughan TL, Buas M, Hinds DA, Gharahkhani P, Kendall BJ, MacGregor S. Multitrait genetic association analysis identifies 50 new risk loci for gastro-oesophageal reflux, seven new loci for Barrett's oesophagus and provides insights into clinical heterogeneity in reflux diagnosis. ***Gut.*** gutjnl–2020 (2021) doi:10.1136/gutjnl-2020-323906.

5. Siggs OM, **Han X**, Qassim A, Souzeau E, Kuruvilla S, Marshall HN, Sean Mullany 1, Mackey DA, Hewitt AW, Gharahkhani P, MacGregor S, Craig JE. Association of monogenic and polygenic risk with the prevalence of open-angle glaucoma. ***JAMA Ophthalmology***. 2021; doi: 10.1001/jamaophthalmol.2021.2440.

6. Gharahkhani P., Jorgenson E., Hysi P., Khawaja A.P., Pendergrass S., **Han X.**, Ong J.S., Hewitt A. W., ..., Mackey D.A., Kubo M., Aung T., Craig J. E., MacGregor S., Wiggs J. L.. Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. ***Nature Communications.*** 2021;12(1):1258.

7. Ong JS, Dixon-Suen SC, **Han X**, An J; Esophageal Cancer Consortium; 23 and Me Research Team, Liyanage U, Dusingize JC, Schumacher J, Gockel I, Böhmer A, Jankowski J, Palles C, O'Mara T, Spurdle A, Law MH, Iles MM, Pharoah P, Berchuck A, Zheng W, Thrift AP, Olsen C, Neale RE, Gharahkhani P, Webb PM, MacGregor S. A comprehensive re-assessment of the association between vitamin D and cancer susceptibility using Mendelian randomization. ***Nature Communications.*** 2021;12(1):246.

8. Qassim A., Souzeau E., Siggs O. M., Hassall M. M., **Han X.**, Griffiths H. L., Frost N. A., Vallabh N. A., Kirwan J. F., Menon G., Cree A. J., Galanopoulos A., Agar A., Healey P. R., Graham S. L., Landers J., Casson R. J., Gharahkhani P., Willoughby C. E., Hewitt A. W., Lotery A. J., MacGregor S., Craig J. E. An Intraocular Pressure Polygenic Risk Score Stratifies Multiple Primary Open-Angle Glaucoma Parameters Including Treatment Intensity. ***Ophthalmology***. 2020;127(7):901-907.

9. Qassim A, Mullany S, Awadalla MS, Hassall MM, Nguyen T, Marshall H, Kolovos A, Schulz AM, **Han X**, Gharahkhani P, Galanopoulos A, Agar A, Healey PR, Hewitt AW, Landers J, Casson RJ, Graham SL, MacGregor S, Souzeau E, Siggs OM, Craig JE. A polygenic risk score predicts intraocular pressure readings outside office hours and early morning spikes as measured by home tonometry. **Ophthalmology Glaucoma.** 2020 Dec 11:S2589-4196(20)30320-3. doi: 10.1016/j.ogla.2020.12.002.

10. An J., Gharahkhani P., Law M. H., Ong J. S., **Han X.**, Olsen C. M., Neale R. E., Lai J., Vaughan T. L., Gockel I., Thieme R., Bohmer A. C., Jankowski J., Fitzgerald R. C., Schumacher J., Palles C., Beacon, andMe Research Team, Whiteman D. C., MacGregor S. Gastroesophageal reflux GWAS identifies risk loci that also associate with subsequent severe esophageal diseases. ***Nature Communications.*** 2019;10(1):4219.

11. Ong J. S., Law M. H., An J., **Han X.**, Gharahkhani P., Whiteman D. C., Neale R. E., MacGregor S. Association between coffee consumption and overall risk of being diagnosed with or dying from cancer among >300 000 UK Biobank participants in a large-scale Mendelian randomization study. ***International Journal of Epidemiology***. 2019;48(5):1447-1456.

12. Liyanage U. E., Law M. H., **Han X.**, An J., Ong J. S., Gharahkhani P., Gordon S., Neale R. E., Olsen C. M., andMe Research Team, MacGregor S., Whiteman D. C. Combined analysis of keratinocyte cancers identifies novel genome-wide loci. ***Human Molecular Genetics***. 2019;28(18):3148-3160.

13. Campos A.I., Ingold N., Huang Y., Kho P.K., **Han X.**, Ong J.S., Garcia-Marin L.M., 23andMe Research Team, Law M.H., Martin N.G., Dong X., Cuellar-Partida G., MacGregor S., Aslibekyan S., Renteria M.E.. Genome-wide analyses in 1,987,836 participants identify 39 genetic loci associated with sleep apnoea. medRxiv. 2020; doi: https://doi.org/10.1101/2020.09.29.20199893.

Contributions by others to the thesis

All of the research described in this thesis was primarily performed by the author, with contributions from Stuart MacGregor, Puya Gharahkhani, Jue-Sheng Ong, Alex W. Hewitt, and other co-authors listed on the included publications and manuscripts. The specific contributions are stated in each chapter.

Statement of parts of the thesis submitted to qualify for the award of another degree

No works submitted towards another degree have been included in this thesis.

Research Involving Human or Animal Subjects

No animal or human subjects were involved in this research.

Acknowledgements

I have been incredibly lucky through the PhD journey. Looking back over these years, it is an amazing endeavour that would not have been possible without the support from many people.

I can clearly remember the week in December 2016 when Stuart invited me to visit his Statistical Genetics Lab. Six months later, I started my PhD in Brisbane. I would like to express my deepest appreciation to my supervisor, Stuart, for providing me the opportunity to work with and learn from him in statistical genetics, and the excellent mentorship and guidance during my candidature. I would also like to extend my gratitude to my associate supervisor, Puya, for helping me along the way.

I thank my collaborators Maciej, Kaiah, Samantha, David, Emmanuelle, Ayub, Mark, Henry, and Jamie for their support in my research. Special thanks to Alex and Jue-Sheng for their constant support.

I would like to express my gratitude to all members from the Statistical Genetics Lab and GenEpi-QIMR family for providing me with the excellent research environment. Many thanks to Jue-Sheng, Matthew, Jiyuan, Upekha, Mathias, Nathan, Weixiong, Santiago, Adrian, Brittany, Jackson, Jose, Geoff, Chang, Daniel, Reece, Henry, Milly, Sami, etc., for being there for the wonderful coffee time and Schotts beer. I thank all members in my milestone review panel, Sarah, Eske, Ann-Marie, Manuel, Michelle, and John, for giving me constructive feedback on my work. I thank the reviewers for their comments and suggestions. I also thank Yang, Sifan, Xiaolin, Jiali, and Yeda, for cooking delicious foods to celebrate Chinese New Year in Brisbane.

Finally, I would like to thank my parents, to whom this thesis is dedicated, for continuing support of all kinds.

Financial support

This research was supported by the University of Queensland Research Training Scholarship and Queensland Institute of Medical Research Berghofer PhD Top Up Scholarship.

Keywords

Statistical genetics, genome-wide association study, polygenic risk score, Mendelian randomization, ophthalmology, glaucoma, age-related macular degeneration, UK Biobank.

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 060412, Quantitative Genetics, 60%

ANZSRC code: 111301, Ophthalmology, 20%

ANZSRC code: 111706, Epidemiology, 20%

Fields of Research (FoR) Classification

FoR code: 0604, Genetics, 60%

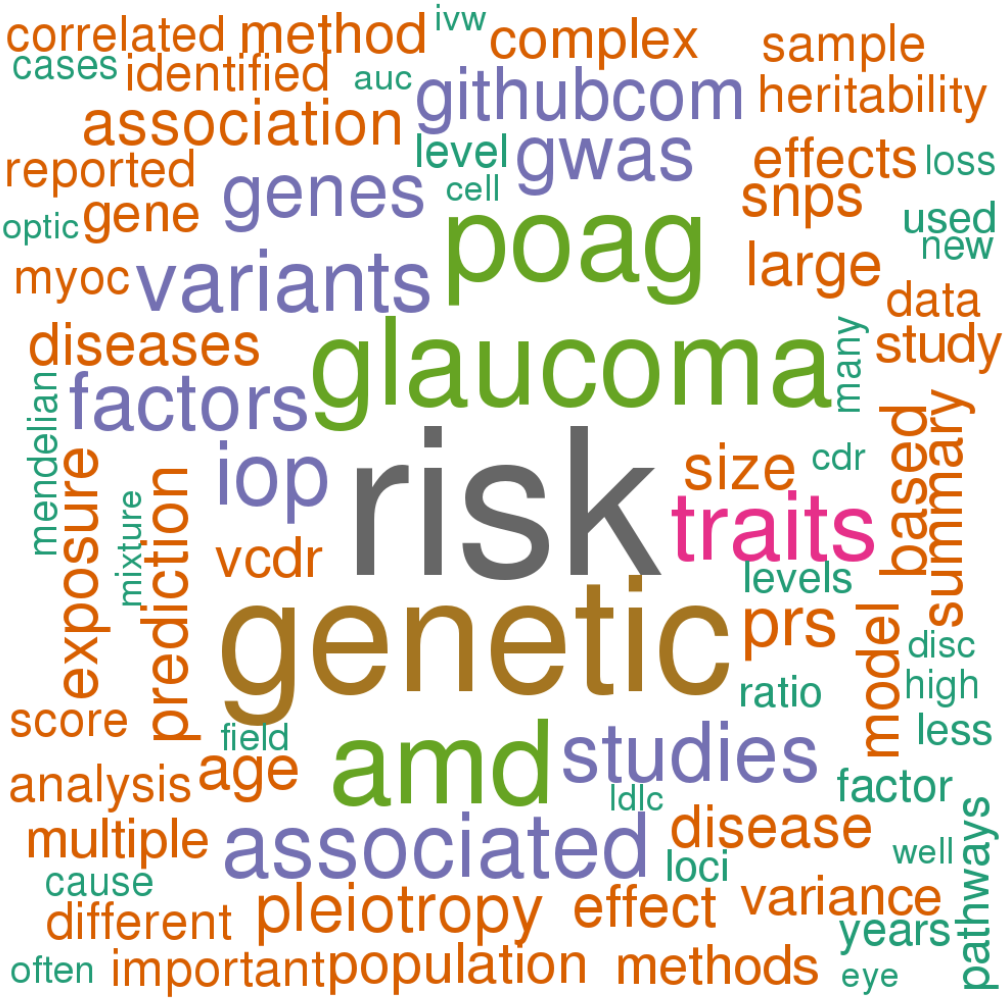
FoR code: 1113, Ophthalmology and Optometry, 20%

FoR code: 1117, Public Health and Health Services, 20%

For my parents

Table of contents

CHAPTER 1


correlated
method
ivw
complex
sample
cases
identified
auc
githubcom
heritability
association
level
gwas
effects
loss
reported
genes
cell
snps
optic
gene
poag
used
new
myoc
variants
large
data
diseases
glaucoma
study
factors
many
mendelian
risk
size
cdr
based
summary
exposure
iop
traits
mixture
prediction
vcdr
levels
genetic
prs
disc
model
high
score
ratio
field
amd
studies
less
analysis
age
ldlc
factor
multiple
associated
disease
well
loci
cause
pathways
different
pleiotropy
effect
variance
years
often
important
population
methods
eye

Chapter 1. Introduction

1.1 The view of complex traits

One major goal of genetic studies is to uncover the role of genes in diseases and traits. In the post-genomics era, it is known that many of the complex traits (or common diseases), from height, eye colour to depression and glaucoma, are influenced by genetics and environment. However, it has been a century-long effort to understand how genes contribute to diseases and traits. In the early 20th century, the Biometrician-Mendelian Debate involved the Mendelians geneticists who were motivated by Mendel's principles of inheritance and focused on individuals and discrete characters with discontinuity of variation, whereas the biometric school was interested in population and continuous traits with small gradual changes.[1] Sir Ronald Aylmer Fisher reconciled biometry with the Mendelian scheme in his 1918 landmark paper that showed variances can be decomposed into components and hypothesized characters such as stature were determined by a large number of "Mendelian factors" with the limiting model known as "infinitesimal model".[2–5]

In the past decades, with the advance of biotechnology to identify molecular markers and the flourishing of computational and statistical genetic methods, the field of quantitative genetics for complex traits has changed remarkably.[6] Around 2005, underlying the rationale "common disease and common variant" hypothesis, the theoretical and practical aspects of genome-wide association study (GWAS) were discussed, showing its advantages compared with the alternative methods, such as linkage mapping and candidate-gene association studies.[7,8] GWAS is a design to detect associations between single nucleotide polymorphisms (SNPs) and complex traits genome-widely rather than via a gene-by-gene candidate approach.[8] As of 4th June 2020, the GWAS Catalog (a collection of human genome-wide association studies) contains 4,566 publications and 186,829 unique SNP-trait associations.[9] These findings provide valuable insights into the genetic architecture of complex traits.

Another debate "nature versus nurture" is about the relative importance of genetic and environmental factors in trait variation. Heritability is an estimable population parameter of the ratio between genetic variance and phenotypic variance.[10] In the broad-sense of heritability (H^2), the genetic variance includes additive variance, dominance variance, and

epistatic variance, whereas the narrow-sense heritability (h^2) only estimates the proportion of additive genetic variance in total phenotypic variance.[11] With the large number of genetic variants identified from GWAS, they typically capture a small proportion of phenotypic variance, leading to the "missing" heritability question.[12] Recent studies have shown a large proportion of heritability can be recovered by including all common SNPs, with the remaining heritability can be explained by rare variants.[13,14] The GWAS signals generally spread across the genome, each SNP with a relatively small effect size.[15] Empirical analysis have also shown that the genetic variance contributed by each chromosome was approximately proportional to its physical length,[16–18] supporting a polygenic genetic architecture for complex traits.[19]

However, until now it remains unclear the underlying genetic architecture of complex traits, for instance, how many genes actually contribute to traits, and how these genes are interconnected to regulate gene expressions and phenotypes.[20,21] Understanding the genetic architecture of complex traits is one of the main themes of quantitative genetics. It would shed light on the biological mechanisms underlying disease pathogenesis, and further provide insights for druggable targets.[22,23] In parallel, it would also contribute to the development of personalized risk prediction models for risk stratification, and the identification of potential causal risk factors for disease prevention and treatment.

1.2 Polygenic risk score

Polygenic risk score (PRS), also called a genetic risk score, or allele score, is a method to calculate a score based on a selected set of genetic variants and their effect sizes.[24,25] Complex diseases typically have a polygenic basis, with genetic signals spread across the genome.[20,21] For example, Loh et al. showed more than 70% of 1-Mb genomic regions across the whole genome harbor at least one genetic variant contributes to schizophrenia risk, implying a large proportion of genes can increase disease risk.[26] Empirical and theoretical studies have shown that genome-wide markers, each with relatively small effects, can improve the prediction of genetic value.[27,28]

With the successful application of PRS in different diseases, PRS is a promising tool for risk stratification, genetic screening, and the development of personalized risk management strategies.[29–35] At the population level, the predictive accuracy of PRS is historically

assessed by area under the receiver-operator characteristic curve (AUC), an index to quantify the diagnostic ability of PRS to separate individuals with disease and those without disease. The theoretical predictive accuracy of PRS could be derived quantitatively based on the heritability of disease, study sample size, and the underlying effect size distributions of SNPs.[36,37] Currently, the AUCs of PRSs for most complex diseases are usually low or moderate (less than 0.7 or 0.8).[24,38,39] In the near future, with the available of large sample size from biobanks, the projected risk prediction ability of PRS will increase at a steady rate until plateauing (*e.g.* with sample sizes exceeding one million).[40] Even at that stage, PRS alone is unlikely to have a very high discriminatory ability (*e.g.* AUC > 0.9) for common diseases.[24,41] However, it should be noted that developing a diagnostic model based on PRS alone is an ambitious goal, since most common diseases show a low or moderate heritability and environmental factors also play an important role in diseases development. A more realistic and practical utility of PRS is for risk stratification, identifying a subgroup of individuals with higher risk. This has been well studied in clinical genetic testing to identify carriers of rare monogenic variants usually conferring several fold increased disease risk (high penetrate variants), such as *BRCA1* and *BRCA2* for breast cancer, and *MYOC* p.Gln368Ter for glaucoma.[42,43] Knowing such variants is important to identify individuals at elevated genetic risk for prevention and early treatment, however, at the population level, these variants are not very helpful to improve diagnostic models because the majority of patients are not monogenic variant carriers (high false negative). In the same scenario, the clinical utility of PRS should not be judged based on its diagnostic ability, but the risk stratification ability to stratify populations into subgroups for further screening and early therapeutic treatment. Recent studies have also reported PRS can identify individuals with genetic risk equivalent to “monogenic mutations”, showing its clinical utility in screening at-risk populations.[33,35]

Apart from a larger sample size, another angle to improve the predictive ability of PRS is to develop powerful and efficient statistical methods (Table 1). The classical approach to construct PRS is pruning+thresholding (P+T method): pruning SNPs to account for linkage disequilibrium (LD), and selecting SNPs with P values less than a particular threshold.[29] In addition to the conventional P+T approach, there are many new methods have been proposed to optimise the predictive algorithm and to improve the genetic prediction value. For instance, LDpred is a Bayesian shrinkage method to infer the posterior mean effect size for each SNP by using GWAS summary statistics and LD information from an external

reference panel;[44] various other Baysian-based methods can use both individual level data and summary statistics;[45–47] methods to leverage genetic functional annotation data in prediction model,[48,49] and to jointly model genetic correlated traits and trans-ethnic data to improve predictive power.[35,50–54] However, sophisticated statistical methods do not necessarily have superior performance.[55,56] Different approaches may rely on different assumptions of the underlying genetic architecture, and currently there is no optimal unified PRS approach that can outperform its counterparts for different traits.[55,56] Further empirical studies are needed to benchmark the performance of these methods in different scenarios.

Table 1. Summary of polygenic risk score methods.

Name	Description	Source	Refs
P+T	prune SNPs to account for LD, and then select SNPs with P values less than a particular threshold	https://www.cog-genomics.org/plink2/score http://www.prsice.info/	29,57
LDpred	derive PRS based on a LD reference panel and summary statistics and assume different fractions of causal SNPs	https://github.com/bvilhjal/ldpred https://github.com/privefl/bigsnpr	44,58
lassosum	penalized regression framework using summary statistics and a reference panel	https://github.com/tshmak/lassosum	59
SBLUP	best linear unbiased prediction using summary data	https://cnsgenomics.com/software/gcta/#SBLUP	60
SBayesR	Bayesian multiple regression on summary statistics with a finite mixture of normal distributions prior	https://cnsgenomics.com/software/gctb/#SummaryBayesianAlphabet	47
PRS-CS	Bayesian regression framework to use a continuous shrinkage (CS) prior on SNP effect sizes	https://github.com/getian107/PRScs	46
PRS-CSx	couples genetic effects across populations to improve polygenic prediction in diverse populations	https://github.com/getian107/PRScsx	61
NPS	non-parametric shrinkage method that allows for linkage disequilibrium in summary statistics	https://github.com/sgchun/nps	62
SDPR	robust Bayesian nonparametric method using summary statistics	https://github.com/eldronzhou/SDPR	63
DBSLMM	deterministic Bayesian sparse linear mixed model with a flexible assumption on the effect size distribution	https://biostat0903.github.io/DBSLMM/index.html	64
AnnoPred	leverage genomic and epigenomic functional annotations	https://github.com/yiminghu/AnnoPred	48
LDpred-funct	leverage trait specific functional enrichments	https://github.com/carlaml/LDpred-funct	49
PolyPred	leverages functionally informed fine-mapping to improve trans-ethnic polygenic prediction	https://github.com/omerwe/polyfun	65
PleioPred	joint modeling of genetically correlated traits and	https://github.com/yiminghu/PleioPre	51

	functional annotations	d	
MTGBLUP	multi-trait genomic best linear unbiased prediction	https://github.com/uqrmaie1/mtgblup	66
MTAG	multi-trait analysis of GWAS summary statistics	https://github.com/JonJala/mtag	53
CTPR	cross-trait penalized regression (CTPR) to incorporate shared genetic effects across multiple traits	https://github.com/wonilchung/CTPR	67
Genomic SEM	A multivariate method to jointly analyse multiple traits based on a structural equation modelling framework.	https://github.com/MichelNivard/GenomicSEM	68

1.3 Mendelian randomization

With the advent of large-scale genome-wide association studies and the availability of GWAS summary statistics, Mendelian randomization (MR) has become a very popular causal inference approach in the past decade. In traditional epidemiology designs, observational studies may be biased by confounding factors and/or reverse causality. Randomized clinical trials (RCTs) are often considered the gold standard for testing causality but are often very expensive and sometimes impossible. MR is an instrumental variable (IV, *e.g.* genetic variants) analysis, and is akin to RCT by distributing genotypes randomly at conception at a population level (analogous to random allocation of a treatment).

Genetic variants associated with interested exposure are usually selected as instrumental variables to investigate the potential causal relationship between the exposure and an outcome. In general, genetic instrumental analysis in MR is less susceptible to reverse causality. However, MR is subject to the following important IV assumptions that require careful evaluation:[69,70]

(1) Relevance assumption: that genetic variants are associated with the exposure (often evaluated by ensuring the strength of the IV-exposure association exceeding an F-statistics of 10 or using genome-wide significant SNPs);

(2) Independence assumption: that genetic variants are not associated with any confounder of the exposure-outcome association;

(3) Exclusion restriction: that genetic variants affect the outcome only through the exposure of interest;

(4) In general, MR estimates represent a *linear effect* of the risk factor on outcome over a

lifetime.

The relevance assumption can be tested by the instrument strength, and usually the statistical power can be boosted by a two-sample MR framework, which allows the exposure and the outcome derived from two separate studies.[71] The polygenicity of complex traits provides emerging evidence of the presence of pervasive pleiotropic effects,[72,73] which means GWAS signals can possibly be associated with multiple traits (including problematic confounding factors). In the MR framework, there are two types of pleiotropy to consider: vertical pleiotropy and horizontal pleiotropy.[74] In vertical pleiotropy (also known as mediation) genetic variants are associated with multiple traits on the same pathway, which does not invalidate the MR assumptions; while horizontal pleiotropy occurs when genetic variants affect multiple traits through separate pathways, which may violate assumption 2 or 3 and potentially leads to false positive MR findings. Thus, detecting and accounting for horizontal pleiotropy have become an important aspect of the methodology and applied MR studies. For instance, a recent study proposed a method CAUSE to account for both correlated pleiotropy (where horizontal pleiotropic effects on outcome are correlated with effects on exposure) and uncorrelated pleiotropy (where horizontal pleiotropic effects on outcome are uncorrelated with effects on exposure [Instrument Strength Independent of Direct Effect, InSIDE assumption]).[75] Uncorrelated pleiotropy occurs when variants influence exposure and outcome through separate pathways. It is assumed that variants affect the outcome with a random positive or negative direction. If on average the random effects have zero mean (balanced horizontal pleiotropy), the uncorrelated pleiotropy only adds random noise and leads to an unbiased estimation in inverse-variance weighted method (MR-IVW).[74] When there is non-zero mean random effect (directional pleiotropy), MR-Egger regression method can be used to model a non-zero intercept.[76] Several outlier removal approaches, such as MR-PRESSO and generalized summary MR (GSMR), were also developed to deal with directional pleiotropy.[77,78] Correlated pleiotropy occurs when variants influence outcome and exposure through shared factors or pathways.[75] For example, in the MR analysis for the association between low-density lipoprotein cholesterol (LDL-C) and coronary artery disease (CAD), a subset of LDL-C genetic variants are associated with high-density lipoprotein cholesterol (HDL-C) or triglycerides, which are phenotypically and genetically correlated with LDL-C and potentially affect CAD risk. For correlated pleiotropic effects in a shared pathway, a multivariable MR (MVMR) method can be used to model multiple traits together.[79] MVMR can estimate the direct effect of the exposure on the

outcome conditioning on other putative exposure traits. However, in MVMR it also requires that the set of SNPs used as genetic IVs are only associated with the included exposure variables but do not influence the outcome other than through these included exposure variables (a generalized assumption in univariable MR).[79] The authors of CAUSE demonstrated the new MR method can leverage information from all SNPs and account for both correlated and uncorrelated horizontal pleiotropic effects and avoid more false positives.[75] However, it only models a single unobserved shared factor and is unable to account for measured known shared factors like multivariable MR.[75]

In practice, MR studies often rely on multiple MR approaches that can present unique strengths and limitations to infer evidence for causality (Table 2). For the rest of the thesis, MR analysis conducted will mainly rely on the conventional inverse variance weighted (IVW) model to interpret estimates, whilst in sensitivity analysis various alternative MR approaches are applied to evaluate the robustness of MR findings.

Table 2. Summary of Mendelian randomization methods.

Name	Description and source[1]	refs
two-stage least squares	ratio estimate based on individual level data. https://github.com/sb452/mr-code	80
IVW	inverse-variance weighted combination of Wald ratio estimates based on summarized data, efficient but biased when one or more genetic variants are invalid IVs. Equivalent to a weighted linear regression with a zero intercept term.	81
weighted median estimate	allow up to 50% of the weights to come from invalid instrumental variables.	82
weighted mode-based estimate	the largest weights from a subset of variants are contributed by valid instruments.	83
Contamination mixture	two-component mixture distributions of valid and invalid IVs.	84
MR-Mix	model four components in the MR mixture models. https://github.com/gqi/MRMix	85
MR-TRYX	discover putative risk factors for diseases from horizontal pleiotropy effects https://github.com/explodecomputer/tryx	86
MR-Egger	detect and correct for the bias due to directional pleiotropy.	76
MR-PRESSO	remove candidate instruments based on a heterogeneity measure, similar in efficiency to the IVW method. https://github.com/rondolab/MR-PRESSO	77
GSMR	detect outlier variants using HEIDI test. https://cnsgenomics.com/software/gsmr/	78
CAUSE	account for correlated and uncorrelated pleiotropic effects. https://github.com/jean997/cause	75
multivariable MR (IVW, Egger)	estimate the effect of multiple exposure variables.	79,87
MR-BMA	multivariable MR method based on Bayesian model averaging that scales to high-throughput data. https://github.com/verena-zuber/demo_AMD	88

[1]Most of the MR methods are provided in two R packages: TwoSampleMR and MendelianRandomization.

1.4 Glaucoma

Glaucoma, an age related heterogeneous eye disease, is characterized by loss of retinal ganglion cells (RGC), thinning of the retinal nerve fiber layer (RNFL), and cupping of the optic disc.[89–91] It is also the leading cause of irreversible blindness globally. The two main forms of glaucoma are primary open angle glaucoma (POAG) and primary angle closure glaucoma (PACG).[92,93] A meta-analysis from 50 population-based studies has shown that the worldwide age-standardized prevalence of glaucoma in population aged 40 years or older is 3.54%, and the global prevalence of POAG and PACG is 3.05% and 0.50%, respectively.[94] The prevalence of glaucoma also varies across different geographic areas and ethnic groups. For instance, the prevalence of POAG and PACG is highest in Africa (4.20%) and Asia (1.09%), respectively, and worldwide Asia alone accounts for 53.4% of POAG cases and 76.7% PACG cases.[92–94] Both genetic and environmental risk factors may play an important role in the difference of prevalence. Moreover, the different definition and classification methods for glaucoma from different studies would also contribute to the difference.[93] For POAG, per decade increase of age is associated with higher risk (odds ratio, OR = 1.73), and men are more likely to have POAG than women (OR = 1.36).[94] Compared to people of European ancestry, African, Asian and Hispanic have increased risk, the OR is 2.80, 1.43, and 2.00, respectively. It has been projected that approximately 60 million people were affected by glaucoma worldwide in 2010, and the number of people with glaucoma increases to 76 million in 2020 and to 112 million in 2040.[94,95]

1.4.1 POAG risk factors

Primary open angle glaucoma, accounting for more than 80% of glaucoma cases globally (>85% in European and Afriacan ancestries, and >65% in Asian ancestry), is mostly asymptomatic until late in the disease progression when visual problems arise. It includes both adult-onset POAG (usually after the age of 40 years) and juvenile open angle glaucoma (JOAG, age of onset between 3 to 40 years old).[91] Epidemiological studies have elucidated several risk factors, including older age, ethnic background, family history, elevated intraocular pressure (IOP), myopia (short-sightedness), and central corneal thickness (CCT)[96,97]. However, the associations with myopia and CCT might be partly due to enlargement of optic disc and biased measurement of intraocular pressure. Cerebrospinal fluid (CSF) pressure and ocular perfusion pressure are suggested to be associated with glaucoma risk.[98–100] The evidence for several systemic diseases, such as hypertension,

diabetes mellitus, cardiovascular diseases, obstructive sleep apnoea, and migraine, still needs further investigations.[91]

Currently, IOP and vertical cup-to-disc ratio (VCDR) are well studied, and are considered as two key endophenotypes of POAG. However, their joint and specific contributions to POAG risk remain unclear.

IOP is the sole modifiable risk factor for POAG. Progression of POAG usually stops if the IOP level is lowered by 30% to 50% from baseline,[90] supporting the strategy of early screening and detection. While higher IOP confers greater risk for POAG, it can also develop at a normal IOP level as it was shown that not all patients with elevated IOP develop POAG. This is the concept of normal-tension glaucoma (NTG), defined as "visual field loss and optic nerve abnormalities consistent with glaucoma and IOP that does not exceed 21 mmHg".[101]

High VCDR, often a sign of glaucomatous visual field (VF) loss, is commonly used to define glaucoma in general population based prevalence surveys.[93] For instance, the category 1 diagnosis from structural and functional evidence is: "eyes with a cup disc ratio (CDR) or CDR asymmetry > 97.5th percentile for the normal population, or a neuroretinal rim width reduced to < 0.1 CDR"; the category 2 diagnosis for advanced structural damage is: "if the subject could not satisfactorily complete visual field testing but had a CDR or CDR asymmetry > 99.5th percentile for the normal population, glaucoma was diagnosed solely on the structural evidence".[93] Although VCDR is listed as a risk factor of glaucoma, it should be noted that VCDR may be used to define glaucoma cases. However, large VCDR does not necessarily mean glaucomatous damage because of the variation of optic disc size (*e.g.* physiologic cupping).[102]

1.4.2 Genetics of POAG

Genetic factors play an important role in the development of POAG.[103,104] In the general population, participants having a first degree relative with glaucoma have almost 10 times increased risk of developing POAG.[105] A recent large-scale study showed that the heritability of glaucoma is approximately 70% based on reconstructed family data.[106] Identifying POAG risk genes will improve our understanding of the underlying pathogenic mechanisms, increase risk prediction for POAG, as well as aid in fine-mapping of potential causal genes,

the development of therapeutics for treatment, and personalized approach for genetic screening.

In the past decades, genetic linkage analyses have identified many genes associated with the risk of POAG, such as myocilin (*MYOC*), *OPTN*, and *TBK1*.[107–109] The *MYOC* gene is considered one of the most important genes associated with POAG, and the pathogenic variants in *MYOC* gene have been found in 2% to 4% of POAG cases.[109,110] The p.Gln368Ter (rs74315329) variant is the most common glaucoma associated *MYOC* variant amongst populations of European ancestry.[107,111,112] Previous studies have shown that the association between *MYOC* p.Gln368Ter variant and POAG is very high (odds ratio, OR > 10). Moreover, *MYOC* p.Gln368Ter carriers are usually diagnosed at a younger age and have higher IOP levels.[113,114] However, the estimated penetrance of *MYOC* p.Gln368Ter in POAG and ocular hypertension has been inconsistent between family-based studies and general population-based studies.[107,111,112,115–117] In Chapter 2, I leveraged large scale datasets of UK Biobank (UKB) and registry-based datasets from Australia to investigate the penetrance and effect size of the *MYOC* p.Gln368Ter with respect to glaucoma and ocular hypertension.

The pace of gene discoveries for complex-traits have increased remarkably during the past decade.[6] With the aid of GWAS, at least 50 genes have been identified to be associated with POAG.[118–126] Two recent studies from our group have identified more than 100 loci from multi-trait and cross-ancestry meta-analysis.[35,127] As presented for many other complex traits or common diseases, current gene discoveries only account for a moderate fraction of heritability, and the biological mechanisms of many genes are largely unknown. The projected number of independent loci will increase in the future when larger sample size is available, *e.g.* several thousand SNPs will be reported when sample size reaches one million for many diseases and traits,[40] and each loci only has a very small effect on disease risk. Uncovering the biological functions for each identified loci is unrealistic, but if some genes could be used as drug targets would be a tremendous advance. Another application of the gene discoveries is to build a genetic risk prediction model, where sample size is a key limiting factor of prediction accuracy.[62] As shown in a “15 years of GWAS” paper,[128] GWAS has no signs of slowing down, and large sample size will continue to contribute to new gene discoveries and provide insight into the understanding of the genetics of POAG.

1.4.3 Genetics of IOP and VCDR

As key quantitative endophenotypes for POAG, the genetics of IOP and VCDR, can be very informative to dissect the genetic risk factors underlying POAG. Our recent study has shown the genetic correlation between POAG and IOP is 0.71 (standard error, SE = 0.04), and the genetic correlation between POAG and VCDR is 0.50 (SE = 0.05). Leveraging the high genetic correlation between IOP/VCDR and POAG, multi-trait GWAS analysis is an ideal approach to boost GWAS power for gene discoveries, uncover novel glaucoma risk genes and pathways, and improve the accuracy of genetic risk prediction models for glaucoma risk.

The availability of large biobanks have accelerated the gene discoveries for IOP dramatically in recent years.[124–126,129,130] For instance, in 2017, an IOP GWAS meta-analysis of 37,930 participants from the International Glaucoma Genetic Consortium (IGGC) reported 9 genomic regions passing genome-wide significance level; in 2018, including 103,914 IOP measures from UKB identified more than 100 loci, and more than half of them (85 of 101) are novel IOP genes.[125,126,131] More importantly, we found 53 of them were associated with glaucoma after Bonferroni correction, indicating these IOP genes contribute to glaucoma risk.[126] Another independent study based on UKB, IGGC and EPIC-Norfolk reported that 48 IOP loci were nominally associated with glaucoma ($P < 0.05$), and 14 of them were associated with glaucoma after Bonferroni correction ($P < 0.00042$).[125] Several new pathways were identified to be associated with IOP and glaucoma from the large scale IOP GWAS. From our IOP GWAS,[126] 11 pathways were highlighted after multiple corrections, and 9 pathways with a P value less than 0.05 in glaucoma pathway analysis. Vascular development pathway, positive regulation of locomotion, cell motility and cell migration showed the strongest evidence. Khawaja *et al.* found that angiopoietin-receptor tyrosine kinase signaling pathway, lipid metabolism, mitochondrial function, and developmental processes are potential pathways to regulate IOP levels.[125] For glaucoma risk prediction, Khawaja and colleagues built a polygenic risk score (PRS) prediction model based on 120 IOP variants, three known POAG SNPs, sex, and age. The area under the curves (AUC) was 0.76 for HTG in the National Eye Institute Glaucoma Human Genetics Collaboration Heritable Overall Operational Database (NEIGHBORHOOD) study, and 0.74 for glaucoma in UK Biobank participants.

Previously, the VCDR GWAS of 32,272 participants from the IGGC reported 30 genomic

regions, only six of them were associated with POAG (including *BMP2*, *CDKN1A*, *CDKN2B-AS1*, *FLNB*, *RERE*, and *SIX6*).[132,133] In Chapter 4, retinal fundus photographs from UKB and Canadian Longitudinal Study on Aging were used to increase sample size for VCDR and vertical disc diameter (VDD), boosting the sample size to more than 100,000 and revealing more than 200 independent genome-wide loci for both VCDR and VDD. The detailed results are presented in Chapter 4A and 4B.

1.5 Age-related macular degeneration

Age-related macular degeneration (AMD), a degenerative disorder of the central retina, is the leading cause of central vision loss in the elderly. The prevalence of AMD is 8.7% globally among individuals older than 45 years, with a higher prevalence of 12.3% in Europeans.[134] It has been estimated that 196 million people were affected by AMD in 2010, and the number of people with AMD will increase to 288 million in 2040.[134] The progression of AMD is classified as early-stage (medium drusen and no retinal pigmentary changes), intermediate-stage (large drusen or medium drusen with retinal pigmentary changes) and late-stage (two subtypes: geographic atrophy [GA] and choroidal neovascularization [CNV]).[135,136] The pathogenesis of AMD progression remains largely unclear. In recent years, anti-vascular endothelial growth factor (VEGF) therapies have become an effective treatment to reduce the progression of CNV subtype.[137] However, the treatment is not curative, and there are no effective medications for GA subtype, which accounts for 90% of late AMD cases. In addition, a better prevention strategy is to identify and to treat AMD at an earlier stage before serious vision loss occurs. Therefore, it is important to find new pathogenesis mechanisms and therapeutic targets for AMD.

1.5.1 AMD risk factors

Several risk factors have been identified for AMD, such as age, smoking, and genetic factors.[138] Advancing age is one of the most important risk factors for AMD. In European populations, the prevalence of AMD is 5.7% in the age group between 50 to 59 years old, whereas the prevalence is 22.5% and 33.6% in the 70-79 and 80-84 age groups, respectively,[134] with a similar age trend in Asian and African populations. Apart from age, smoking is consistently associated with AMD risk from different studies, and is considered

as the most important modifiable risk factor.[139] Current smokers have two times increased risk relative to non-smokers, and are more likely to develop AMD at an earlier age.[140,141] However, the association of past smoking and AMD risk is controversial. Some studies reported no association, and the incident risk of AMD may decrease after smoking cessation.[140,142] A variety of less robust risk factors have been reported for AMD. For instance, hyperopia was associated with an increased AMD risk in a meta-analysis[143], which was also supported by a Mendelian randomization study, showing a minimal influence of refractive error on AMD risk.[144] Some lipid biomarkers have been shown to be associated with AMD risk.[145–151] An observational meta-analysis showed that higher levels of high-density lipoprotein cholesterol (HDL-C) is associated with an increased risk of AMD, whereas higher levels of low-density lipoprotein cholesterol (LDL-C), total cholesterol (CHOL), and triglycerides (TG) are associated with a decreased AMD risk.[146] In addition, two previous MR studies also supported a potential causal role of higher levels of HDL-C on the risk of advanced AMD, but not other lipid biomarkers.[147,148] Other reported possible risk factors for AMD include sunlight exposure,[152,153] alcohol consumption[154], dietary patterns,[155,156] physical activity,[157] serum C-reactive protein levels,[158,159] circulating vitamin D levels,[160] obesity,[161,162] diabetes,[163] and cardiovascular diseases.[164]

1.5.2 Genetics of AMD

Genome-wide association studies have identified more than 50 genes association with AMD risk, such as complement pathway related genes, complement factor H (*CFH*), factor I (*CFI*), complement components *C2*, *C3*, *C9*, and lipid related genes *CETP*, *LIPC*, *ABCA1*.[165–167] These genetic findings have advanced our understanding of the biological mechanisms underlying AMD pathogenesis, as well as have provided therapeutic targets for clinical trials.[168,169] Leveraging these genetic findings, genetic risk prediction models for AMD have shown risk stratification and predictive capacity.[166,167] For example, Fritsche et al. reported that the risk to develop advanced AMD was increased by 44-fold for participants in the top decile of genetic risk score relative to those in the bottom decile.[166] In Chapter 5, from a meta-analysis of AMD GWAS, we identified 69 lead common risk variants, and in aggregate, a polygenic risk prediction model can reach an AUC value of 0.76 (95% CI: 0.72–0.80) in a population based study.[167]

Recent bulk and single-cell transcriptome-wide atlas of human retina have also provided

new insights into gene fine-mapping and potentially pathogenic cell types for AMD.[170,171] For example, the Eye Genotype Expression (EyeGEx) database, a large resource for retinal transcriptomes established from 453 AMD cases and controls, has presented a gene expression reference database for ocular traits complementing the GTEx project.[170] A transcriptome-wide association analysis (TWAS) based on the retinal eQTL data and AMD GWAS summary statistics have prioritized three potential target genes (*RLBP1*, *HIC1*, and *PARP12*).[170] Another study reported the first single-cell transcriptomic atlas of human retina from six postmortem retinas with a total of 20,091 and 3,248 cells respectively from two scRNA-seq platforms, and showed some cell types are preferentially associated with the risk of AMD.[171]

CHAPTER 2

Myocilin gene Gln368Ter variant penetrance and association with glaucoma in population-based and registry-based studies

Xikun Han, Emmanuelle Souzeau, Jue-Sheng Ong, Jiyuan An, Owen M Siggs, Kathryn P. Burdon, Stephen Best, Ivan Goldberg, Paul R. Healey, Stuart L. Graham, Jonathan B. Ruddle, Richard A. Mills, John Landers, Anna Galanopoulos, Andrew White, Robert Casson, David A. Mackey, Alex W. Hewitt, Puya Gharahkhani, Jamie E. Craig, Stuart MacGregor.

JAMA Ophthalmology. 2019;137(1):28-35.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Emmanuelle Souzeau assisted with data analysis of Australian registry-based studies and manuscript preparation. Stuart MacGregor, Puya Gharahkhani, and Jamie E. Craig obtained funding and designed the study. All authors contributed to interpretation of the results and the final version of the paper.

Chapter 2. Myocilin gene Gln368Ter variant penetrance and association with glaucoma in population-based and registry-based studies

The p.Gln368Ter (rs74315329) risk allele in the myocilin gene (*MYOC*) was initially reported to have high penetrance in glaucoma registry-based studies, but much lower estimates were recently obtained from population-based studies. This disparity was investigated using data from Australia and the United Kingdom. This cross-sectional study within the UK Biobank (UKB) included participants of white British ancestry. Glaucoma cases were defined by International Classification of Diseases, Ninth Revision (ICD-9) and Tenth Revision (ICD-10) diagnoses and self-reported questionnaires. Carriers of the *MYOC* p.Gln368Ter variant were identified using genotype imputation from arrays. In contrast, two Australian registry-based studies, the Australian and New Zealand Registry of Advanced Glaucoma and the Glaucoma Inheritance Study in Tasmania, ascertained glaucoma cases referred by eye care clinicians, with historic control participants recruited from other Australian studies. Samples were either directly sequenced or had genotypes determined by imputation (for the Australian registry and historic control participants). Recruitment to the UKB occurred between 2006 and 2010, and data analysis occurred from September 2017 to July 2018. A total of 411 337 UKB participants of white British ancestry (mean [SD] age, 56.6 [8.0] years) were included, plus 3071 Australian registry and 6750 historic control participants. In the UKB, the minor allele frequency of the *MYOC* p.Gln368Ter variant was 1 in 786 individuals (0.13%). The odds ratio of p.Gln368Ter in patients with primary open-angle glaucoma (POAG) was 6.76 (95% CI, 4.05-11.29); glaucoma (POAG, self-reported glaucoma, and unspecified glaucoma), 4.40 (95% CI, 3.38-5.71); OHT, 3.56 (95% CI, 2.53-4.92); and OHT and glaucoma combined, 4.18 (95% CI, 3.05-5.67). The penetrance of the *MYOC* p.Gln368Ter variant was 7.6% in patients with glaucoma, 24.3% in patients with OHT, and 30.8% in patients with OHT and glaucoma combined. In the Australian registry studies, the odds of *MYOC* p.Gln368Ter variant were 12.16 (95% CI, 6.34-24.97) in patients with advanced glaucoma and 3.97 (95% CI, 1.55-9.75) in those with nonadvanced glaucoma; the penetrance of glaucoma was 56.1%, and penetrance in those considered to have glaucoma or be glaucoma suspects was 69.5%. In conclusion, the *MYOC* p.Gln368Ter variant confers a very high-risk effect size for advanced glaucoma; the risk is lower in nonadvanced glaucoma and OHT. In the general population sample, approximately 50% of *MYOC* p.Gln368Ter carriers 65 years and older had glaucoma or OHT, with higher prevalence in the Australian registry studies.

2.1 Introduction

Glaucoma is the leading cause of irreversible blindness globally. The most common forms of glaucoma are primary open angle glaucoma (POAG) and primary angle closure glaucoma (PACG). For the population over 40 years old, the worldwide age-standardized prevalence of glaucoma, POAG and PACG is 3.54%, 3.05% and 0.50%, respectively.[94] It is estimated there were 60.5 million POAG and PACG patients worldwide in 2010 and that number will be 112 million by 2040.[94,95] Elevated intraocular pressure (IOP) is the major modifiable risk factor for POAG. Progression of POAG is arrested or reduced if the IOP is lowered by 30–50% from baseline levels.[90]

Genetic factors play an important role in glaucoma.[103,104,172] Having a first-degree relative with glaucoma raises the likelihood of developing glaucoma by 9.4 fold relative to the general population.[105] A recent large-scale study estimated the heritability of glaucoma to be 70% using reconstructed family data.[106] The myocilin gene (*MYOC*) at the *GLC1A* locus was the first gene discovered to be associated with POAG.[109,110] Pathogenic variants in *MYOC* have been found in 2-4% of POAG cases.[107,111] The exact pathogenic mechanisms by which disease-causing variants in *MYOC* cause glaucoma have not been elucidated completely, but evidence supports a dominant-negative mechanism.[173,174]

The p.Gln368Ter (rs74315329) is the most common *MYOC* variant amongst populations of European ancestry.[107,111,112] The association between p.Gln368Ter and POAG has a high odds ratio (OR>10), with p.Gln368Ter associated with younger age at onset and greater severity of IOP elevation.[113,114] The estimated penetrance of p.Gln368Ter in glaucoma and ocular hypertension (OHT) has been inconsistent between family studies and general population-based studies.[107,111,112,115–117] There are several potential explanations for this inconsistency. Estimates from family studies may be inflated due to ascertainment bias, aggregation of other genetic factors, and/or confounded by common environmental risk factors. Conversely, estimates from general population-based designs are likely to be low due to undersampling of cases (especially more severely affected cases) amongst volunteer-based studies.[117] Additionally, for both family studies and general population-based studies, statistical power is typically low due to relatively low numbers of p.Gln368Ter carriers.

In this study, we explore the penetrance and association of *MYOC* p.Gln368Ter with glaucoma and OHT in white Europeans enrolled in the UK Biobank (UKB) study, and compare the results with registry-based studies.

2.2 Methods

2.2.1 UK Biobank

The UKB project is a large-scale prospective cohort study of approximately 500,000 individuals across the United Kingdom, aged between 40 and 69 at the time of recruitment (2006-2010). Detailed information of the UKB study is available online (http://www.ukbiobank.ac.uk/resources/) and the genotype curation process is described in Bycroft et al[130]. The study was approved by the National Research Ethics Service Committee North West – Haydock, in accordance with the Declaration of Helsinki.

In the UKB study, the genotypic data were imputed with IMPUTE4 using the Haplotype Reference Consortium (HRC), UK10K and 1,000 Genomes Phase 3 reference panels. Among 487,409 participants passing genotyping quality control, 409,694 had white-British ancestry based on self-reported ethnicity. We identified 438,870 individuals who fell within the same genetic principal components-based clustering as those who self-reported white-British, based on the first two genetic principal components (eFigure 1 in the supplement).

Our previous study showed that p.Gln368Ter can be imputed with high accuracy from genotyping arrays.[114] In this study, the imputation posterior probability for each of the three genotypes (GG, AG, and homozygous for risk allele AA) was used to identify p.Gln368Ter carriers. We calculated the genotype dosages based on imputation posterior probability. As the dosage only ranged from 0 to 1.1, there were no p.Gln368Ter homozygous carriers in our study. For downstream analyses requiring best guess genotypes, we set the dosage threshold of heterozygous AG at 0.8.

A subset of 112,690 UKB participants underwent IOP measurements at the UKB Assessment Centre using Reichert Ocular Response Analyser,[175] the detailed procedure is available online (http://biobank.ctsu.ox.ac.uk/crystal/docs/Intraocularpressure.pdf). The

mean corneal-compensated IOP (IOPcc, UKB Field 5254 and 5262) and mean Goldmann-correlated IOP (IOPg, Field 5255 and 5263) for each participant was calculated at the initial assessment visit, with measurements < 5 or > 60 mmHg set as "missing". OHT or high IOP was defined as mean IOPcc > 21 mmHg (N= 6,827).

Among the 438,870 subjects with genetic data, we removed participants who withdrew consent (N=10). From the remaining 438,860 participants, 7,997 glaucoma cases were identified using the following criteria: 1) had International Classification of Diseases (ICD)-9 or ICD-10 diagnosis ("Primary Open Angle Glaucoma", "Other Glaucoma", "Glaucoma, unspecified"); or 2) responded "Glaucoma" in "eye problems/disorders" (Field 6148); or 3) responded "Glaucoma" in self-reported non-cancer illness (Field 20002). There were 1,111 POAG cases identified by ICD-9 or ICD-10 diagnosis "Primary Open-Angle Glaucoma". In participants with IOP measurements, we defined "OHT or glaucoma" cases as individuals who had mean IOPcc > 21 mmHg or were identified as glaucoma cases. The information for age at glaucoma onset was gathered from Field 4689 and 20009. Field 21022 was used as age at recruitment.

Finally, "healthy" controls for glaucoma, POAG, and OHT were selected as individuals who: 1) did not have other serious eye diseases (Field 6148, 26,576 individuals excluded); and 2) did not have other kinds of glaucoma diagnosed by ICD-9 or ICD-10 (glaucoma suspect, PACG, or secondary glaucoma, 947 individuals excluded). For POAG controls, 6,886 other glaucoma cases were set as not available (NA) status. In total 411,337 UKB participants were included in this study. The flow-chart illustrating the selection criteria is shown in the eFigure 2 in the supplement.

2.2.2 Australian registry-based studies

In addition to examining population-based data, we considered participants from two registry-based studies: the Australian and New Zealand Registry of Advanced Glaucoma (ANZRAG) and the Glaucoma Inheritance Study in Tasmania (GIST). Recruitment has been previously described. [176,177] In brief, glaucoma patients from Australia and New Zealand had been referred to the ANZRAG by their ophthalmologists. Participants in the GIST were recruited from surveys distributed to ophthalmology clinics and advertisements around Tasmania. Clinical information was collected by the patient's treating ophthalmologist. Participants from ANZRAG/GIST were considered to have glaucoma if they had

glaucomatous visual field defects on standard automated perimetry and neuroretinal rim thinning (cup-to-disc ratio, CDR ≥0.7 or CDR asymmetry ≥0.2). Glaucoma suspects had OHT as defined by IOP above 21 mmHg or had pre-perimetric glaucoma based on glaucomatous appearance of the optic disc or thinning or the retinal nerve fibre layer with no glaucomatous field changes.

There were two arms to the ANZRAG/GIST component: 1) a sequencing-based study within ANZRAG/GIST alone to estimate the penetrance of p.Gln368Ter and 2) an array-based genome-wide association study (GWAS) to allow estimation of OR of glaucoma in a large sample of cases and controls (with controls sourced from outside ANZRAG/GIST).

In the sequencing-based study, all glaucoma participants and their relatives in the ANZRAG/GIST underwent Sanger sequencing for *MYOC* exon 3 as previously described.[178] There were 174 participants with a Sanger-validated *MYOC* p.Gln368Ter carriers, including 164 with a known age at diagnosis.

In the array-based study, we selected a total of 3,071 unrelated glaucoma cases from ANZRAG/GIST and 6,750 unscreened controls from the Brisbane Adolescent Twin Study, the Australian Cancer Study, a study of inflammatory bowel diseases, and a study of endometriosis.[120,122,179] The samples were genotyped on Illumina Omni1M, OmniExpress, or HumanCoreExome arrays (approximately two thirds of cases were genotyped on HumanCoreExome arrays, with the remainder typed on arrays with higher SNP density [Omni1M/OmniExpress], with a similar proportion among the controls).[180] Genotype imputation was performed using Minimac3 through the Michigan Imputation Server, with the HRC r1.1 as the reference panel. We investigated the effect size of p.Gln368Ter for advanced (N=1,753) and non-advanced glaucoma (N=1,318) separately. Advanced and non-advanced glaucoma cases were defined as previously described.[181] Approval was obtained from QIMR Berghofer Institute of Medical Research, the Southern Adelaide Clinical Human Research Ethics Committee/Flinders University, the University of Tasmania, and the Royal Victorian Eye and Ear Hospital in accordance with the Declaration of Helsinki.

2.2.3 Statistical Analysis

Descriptive statistics are presented as mean (standard deviation) for continuous variables or as number (percentages) for categorical variables. Continuous variables were compared

between groups using analysis of variance, whereas Pearson chi-square or Fisher exact tests were used for categorical variables. We explored the frequency of glaucoma and OHT in different age groups (less than 50, 50-59, 60-65 and more than 65 years old) in *MYOC* p.Gln368Ter carriers. We also investigated the cumulative risk of glaucoma by age 50, 60 and 65 years old using a Cox model (adjusted for sex, and the first six genetic principal components, UKB) or the Kaplan-Meier method (ANZRAG). The association between p.Gln368Ter dosage and disease status was estimated using logistic regression adjusted for sex, age, and the first six PCs. To control the bias from familial relationships in association analysis, we used a relationship-based pruning strategy in PLINK (https://www.cog-genomics.org/plink/) to exclude one member of each pair of samples if the genomic relatedness greater than 0.2.[182] The R "survival" package was used in analyses (R version 3.4.1; http://www.r-project.org). We used two-sided *P* values and an alpha level of 0.05.

2.3 Results

Table 1 shows the baseline characteristics of the 411,337 UKB participants included in this study. Approximately 46% of participants were male. The mean age of participants was 56.59 years old, with a mean IOPcc of 16.06 mmHg and mean IOPg of 15.93 mmHg. We observed a trend that the average level of IOP increased with age.

Table 1. Characteristics of 411,337 UK Biobank study participants.

Variable		Age <50 years (N=94164)	Age 50-60 years (N=138395)	Age 60-65 years (N=101322)	Age > 65 years (N=77456)
Age	Years	45.0 ± 2.7	54.8 ± 2.9	61.9 ± 1.4	66.9 ± 1.5
Sex	Male	42529 (45.2%)	60806 (43.9%)	46539 (45.9%)	38851 (50.2%)
	Female	51635 (54.8%)	77589 (56.1%)	54783 (54.1%)	38605 (49.8%)
IOPcc	mmHg	15.20 ± 3.17	15.78 ± 3.39	16.49 ± 3.55	16.85 ± 3.72
IOPg	mmHg	15.38 ± 3.43	15.76 ± 3.57	16.23 ± 3.64	16.38 ± 3.80

Abbreviations: IOP, intraocular pressure; IOPcc, corneal-compensated IOP; IOPg, Goldmann-correlated IOP.

From 411,337 UKB participants, it was estimated that 1,046 participants carried the p.Gln368Ter AG genotype. The minor allele frequency (MAF) of risk allele A of p.Gln368Ter

was 1/786 (0.13%) and the observed MAFs were roughly the same across different age groups. As expected given the frequency, no AA homozygotes were observed. The *MYOC* p.Gln368Ter penetrance and its association with glaucoma and OHT are summarised in Table 2. The penetrance of p.Gln368Ter in glaucoma, POAG, OHT, and OHT or glaucoma was estimated to be 7.55%, 1.63%, 24.30%, and 30.84%, respectively. The ORs (95% confidence interval, CI) of p.Gln368Ter in glaucoma, POAG, OHT, and OHT or glaucoma were 4.40 (3.38, 5.71), 6.76 (4.05, 11.29), 3.56 (2.53, 4.92), and 4.18 (3.05, 5.67), respectively. For p.Gln368Ter carriers, their IOPcc was 2.04 mmHg (95% CI: 1.44 - 2.64 mmHg) higher than individuals of GG genotype.

Table 2. Disease frequency, penetrance and risk effect of *MYOC* p.Gln368Ter in the UK Biobank.

Phenotype	rs74315329 AG[a]	rs74315329 GG	OR (95% CI)[c]	*P* value
Glaucoma	79 (**7.55%**)[b]	7918 (1.93%)	4.40(3.38,5.71)	< 0.001
POAG	16 (**1.63%**)	1095 (0.27%)	6.76(4.05,11.29)	< 0.001
OHT	52 (**24.3%**)	6775 (8.04%)	3.56(2.53,4.92)	< 0.001
OHT or glaucoma	66 (**30.84%**)	8015 (9.51%)	4.18(3.05,5.67)	< 0.001
IOPcc, mmHg	18.10 ± 4.47	16.06 ± 3.50	-	< 0.001
IOPg, mmHg	17.74 ± 4.28	15.92 ± 3.62	-	< 0.001

Abbreviations: CI, confidence interval; IOP, intraocular pressure; IOPcc, corneal-compensated IOP; IOPg, Goldmann-correlated IOP; OHT, ocular hypertension; OR, odds ratio; POAG, primary open angle glaucoma.

[a] In the UK Biobank, there are 1,046 carriers of *MYOC* p.Gln368Ter and 410,291 non-carriers. In the subset of participants with IOP measurements, there are 214 carriers of *MYOC* p.Gln368Ter and 84,267 non-carriers.

[b] Number of cases (frequency). Penetrance showed in bold.

[c] In association analysis, relatives are removed if the genomic relatedness greater than 0.2 (~5% of the individuals are removed due to relatedness prior to the statistical test being applied).

P values are the association between p.Gln368Ter dosage and disease status using logistic regression adjusted for sex, age, and the first six genetic principal components (general linear model was used for IOP levels).

In the UKB, the age-related frequency of glaucoma, POAG, OHT, and OHT or glaucoma is summarised in Table 3 and Figure. The frequency of glaucoma, POAG, OHT, and OHT or glaucoma in p.Gln368Ter carriers older than 65 years old was 15.46%, 4.09%, 40.00%, and 48.00%, respectively. We gathered age at onset glaucoma information for 4,915 individuals:

the mean age and standard deviation at diagnosis was 53.49 ± 10.81 years old. The cumulative risk of glaucoma at 50, 60 and 65 years old was 2.27%, 8.14% and 15.60%, respectively (eTable 1 in the Supplement).

Table 3. Age-related prevalence of glaucoma and OHT in the UK Biobank.

Age Group	rs74315329 AG	rs74315329 GG	*P* value
Glaucoma			
<50 years	3 (**1.26%**)[a]	469 (0.5%)	0.12
50-59 years	19 (**5.21%**)	1817 (1.32%)	<0.001
60-65 years	27 (**10.84%**)	2555 (2.53%)	<0.001
>65 years	30 (**15.46%**)	3077 (3.98%)	<0.001
POAG			
<50 years	1 (**0.42%**)	36 (0.04%)	0.09
50-59 years	3 (**0.86%**)	216 (0.16%)	0.02
60-65 years	5 (**2.2%**)	326 (0.33%)	<0.001
>65 years	7 (**4.09%**)	517 (0.69%)	<0.001
OHT			
<50 years	9 (**21.95%**)	715 (3.98%)	<0.001
50-59 years	12 (**15.58%**)	1730 (6.38%)	<0.001
60-65 years	11 (**23.91%**)	2172 (9.82%)	<0.001
>65 years	20 (**40%**)	2158 (12.65%)	<0.001
OHT or glaucoma			
<50 years	9 (**21.95%**)	791 (4.4%)	<0.001
50-59 years	15 (**19.48%**)	1985 (7.32%)	<0.001
60-65 years	18 (**39.13%**)	2583 (11.68%)	<0.001
>65 years	24 (**48%**)	2656 (15.57%)	<0.001

Abbreviations: IOP, intraocular pressure; OHT, ocular hypertension; POAG, primary open angle glaucoma.

[a] Number of cases (frequency). Frequency in risk allele carriers showed in bold.

P values are from chi-square test or Fisher exact test comparing disease frequency based on p.Gln368Ter genotypes in different age groups (~5% of the individuals are removed due to relatedness prior to the statistical test being applied).

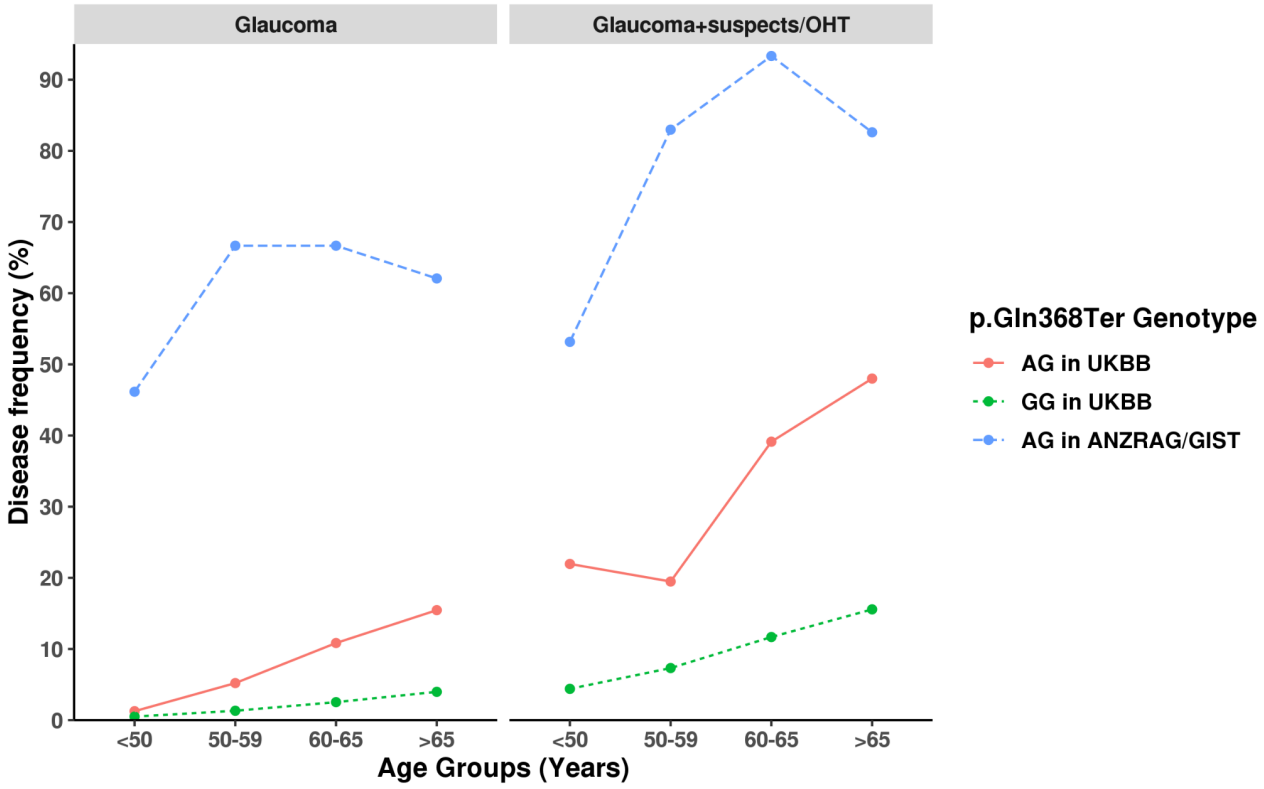


Figure. Age-associated frequency of glaucoma and ocular hypertension (OHT) or glaucoma suspects in p.Gln368Ter risk allele carriers.

ANZRAG indicates the Australian and New Zealand Registry of Advanced Glaucoma; GIST, the Glaucoma Inheritance Study in Tasmania; UKB, the UK Biobank.

Among the 164 individuals carrying the p.Gln368Ter variant in ANZRAG/GIST, 92 (56.1%) had glaucoma, 22 (13.4%) were glaucoma suspects and 50 (30.5%) were unaffected. There were 77 males and 87 females. The mean age at glaucoma diagnosis was 53.18 ± 12.93 years and their mean IOP at diagnosis was 32.47 ± 9.47 mmHg. The penetrance of p.Gln368Ter with respect to glaucoma was 56.10%; for a wider definition including both glaucoma and glaucoma suspects the penetrance was 69.51% (Table 4). Figure and eTable 2 present the age-related frequency of glaucoma and glaucoma/suspects in ANZRAG/GIST registry-based studies in p.Gln368Ter carriers. The cumulative risk of glaucoma in *MYOC* p.Gln368Ter carriers at 50, 60 and 65 years old was 55.88%, 80.49% and 87.06% respectively, and the cumulative risk of glaucoma/suspects at 50, 60 and 65 years old was 77.59%, 94.38% and 95.96%, respectively (eTable 3 in the Supplement). Based on imputed p.Gln368Ter status, the OR (95% CI) of p.Gln368Ter for advanced and non-advanced glaucoma was 12.16 (6.34, 24.97) and 3.97 (1.55, 9.75), respectively.

Table 4. Penetrance of p.Gln368Ter in ANZRAG and GIST registry-based studies.

Phenotype	Disease status	Rs74315329 AG, No.(%)	Mean age at last examination (y)	*P* value	Mean max recorded IOP	*P* value
Glaucoma	No	72 (43.90%)	48.11 ± 16.74	<0.001	19.32 ± 6.07	<0.001
	Yes	92 (**56.10%**)	69.13 ± 12.65		31.26 ± 9.55	
Glaucoma or glaucoma suspect	No	50 (30.49%)	42.50 ± 16.25	<0.001	16.48 ± 2.87	<0.001
	Yes	114 (**69.51%**)	67.14 ± 13.12		30.01 ± 9.43	

Abbreviations: IOP, intraocular pressure, mmHg. Penetrance showed in bold.

2.4 Discussion

To our knowledge, this is the largest study to examine the penetrance and association of the *MYOC* p.Gln368Ter on glaucoma and OHT in a cohort of European white-British individuals and compare it with data from two large registry-based studies. We found that p.Gln368Ter was robustly associated with glaucoma, POAG and OHT and that its penetrance increased with age.

The p.Gln368Ter variant was well imputed (imputation quality score of 93.8%) and the MAF was 0.13% in UKB. In our study, the MAF was similar to those reported from exome sequencing databases, i.e. 0.15% (192/126640) in Non-Finnish European in the Genome Aggregation Database (http://gnomad.broadinstitute.org/), but much higher than that recently reported in the TwinsUK cohort (MAF 0.07%, 8/12184).[117] The lower MAF seen in the TwinsUK cohort suggests that the set of volunteers ascertained was biased toward healthy individuals ("healthy individual" bias).

In the UKB study, POAG or glaucoma cases were identified by ICD-9, ICD-10 diagnosis or self-reported questionnaires and the frequency of POAG and glaucoma was 0.27% and 1.94%, respectively. A previous study estimated the prevalence of POAG and glaucoma in

Europe as 2.51% and 2.93%, respectively.[94] Previous studies also showed that 50% of glaucoma cases are undiagnosed.[183,184] Due to the lack of a comprehensive eye examination in the UKB, the proportion of glaucoma or POAG cases defined here were lower than expected. However, IOP is a key risk factor for POAG and the main mechanism of p.Gln368Ter is via elevation of IOP.[109,181,185] The penetrance and risk effect of p.Gln368Ter in OHT serves as a proxy for POAG.[117] The prevalence of OHT (defined as IOPcc >21 mmHg) reported here was 8.08%, which is similar to an earlier study.[186]

Family studies have shown that p.Gln368Ter had a high penetrance in POAG and OHT.[107,112,115] For instance, Craig et al.[112] reported the age-related penetrance of p.Gln368Ter for OHT or POAG as 72% (28/39) at age 40 years and 82% (14/17) at age 65 years. Another study by Allingham and colleagues observed that 100% (9/9) people with the p.Gln368Ter variant had elevated IOP, and 78% (7/9) had POAG by age 70.[115] In the current ANZRAG/GIST study, our data indicated that the cumulative risk for glaucoma and glaucoma/suspects was 87.06% and 95.96% respectively at 65 years old, which was consistent with findings from previous family-based studies.[115, 107,112]

From their population-based study, Nag et al.[117] reported that the penetrance of p.Gln368Ter in relation to OHT was 12.5% (1/8) and 19.4% (6/31) in the TwinsUK and the Rotterdam Study, respectively. The penetrance of p.Gln368Ter for POAG was 12.5% (1/8) and 9.7% (3/31) in the TwinsUK and the Rotterdam Study, respectively. For OHT, our study showed that the penetrance of p.Gln368Ter (24.3%, 52/214) was lower than the previous family studies but higher than in the population-based study. With approximately 100,000 participants having IOP measurements, our study provided a more robust estimation of p.Gln368Ter penetrance in OHT in population-based studies. However, our number of POAG cases was much lower than expected, given the typical prevalence of POAG in Europe (2.51%).[94] Hence, the “true” penetrance of p.Gln368Ter in POAG is likely to be larger than estimated in the UKB samples here. This again may reflect the bias of a volunteer-based study design.[187] We also proposed a method to calculate the penetrance of p.Gln368Ter based on its OR, MAF and disease frequency (eMethod in the Supplement). According to our proposed method, if the prevalence of glaucoma and POAG was 2.93% and 2.51%, respectively, in population over 40 years old, using the ORs and MAF of p.Gln368Ter from the UKB, the estimated overall penetrance of p.Gln368Ter for glaucoma and POAG was derived to be 10.7% and 15.1%, respectively.

The penetrance of p.Gln368Ter with respect to OHT and glaucoma combined is a more comprehensive indicator. Our study showed that in p.Gln368Ter carriers, the cumulative risk of OHT or glaucoma was 38.69% at age 65 years in the population-based study and 95.96% in glaucoma-based registries; p.Gln368Ter genotyping has great potential for early identification of individuals at risk for developing these eye diseases.[116,181,188]

The penetrance of p.Gln368Ter with respect to glaucoma in UKB was lower than family studies. There are several potential reasons. On the one hand, estimates from family studies may have been inflated by ascertainment bias. On the other hand, the penetrance in general population-based studies may be underestimated due to undersampling of cases. Furthermore, it remains possible that aggregated genetic and environmental risk factors in family studies may have led to increased penetrance in p.Gln368Ter carriers. Recruitment based on families with multiple affected individuals is likely to lead to an increase in the number of common variants of individually small effect (polygenes) in a family, potentially increasing the penetrance of variants such as p.Gln368Ter. This supports the use of cascade testing as close relatives share the same genetic background.

Limitations

This study has some limitations. The genotypes of p.Gln368Ter in UKB are based on best-guessed imputed genotypes. Reassuringly, our previous study presented evidence that the p.Gln368Ter variant could be imputed with high accuracy.[114] Thus, the imputed genotype is unlikely to make a meaningful difference in our results. Another limitation of the UKB study is that some glaucoma cases were defined by self-reported questionnaires, which could lead to recall bias. However, our study is one of the largest studies to investigate the penetrance and risk effect of p.Gln368Ter in OHT, which could serve as a proxy for glaucoma or POAG.[117] Furthermore, glaucoma cases with eye disorders may be less likely to participate the UKB project compared with healthy individuals,[187] which could lead to a lower estimated penetrance of p.Gln368Ter in glaucoma. Moreover, in ANZRAG/GIST, the controls were genotyped on different platforms. As a sensitivity analysis, we substituted the Australian controls for controls from UKB; our results were essentially unchanged. Finally, some individuals with high IOP present in the UKB cohort may be on medications or have undergone ophthalmic surgery to reduce their IOP levels. In our sensitivity analysis to adjust for the change in IOP post-medication, when we added 25% to the measured IOP levels for

individuals taking IOP-lowering medications,[119,189] the resultant OR and penetrance with respect to OHT only increased slightly.

Conclusions

Our study suggests that the *MYOC* p.Gln368Ter variant has a high penetrance in OHT and glaucoma. Genetic testing for p.Gln368Ter could help identify individuals who are at greater risk of developing glaucoma and direct them to early screening and appropriate management.

2.5 Acknowledgment

Funding/Support: This work was conducted using the UK Biobank Resource (application number 25331). This work was supported by grants from the National Health and Medical Research Council (NHMRC) of Australia (#1107098; 1116360, 1116495, 1023911), the Ophthalmic Research Institute of Australia, the BrightFocus Foundation. XH is supported by the University of Queensland Research Training Scholarship. JSO is supported by scholarships from the University of Queensland and QIMR Berghofer Medical Research Institute. KPB, JEC and AWH are supported by NHMRC Fellowships. SM is supported by an Australian Research Council Future Fellowship.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Additional Contributions: We gratefully thank all the participants or volunteers who participated in the studies. The authors would like to acknowledge David Whiteman PhD MD, Graham Radford-Smith MD and Nick Martin PhD from QIMR Berghofer Medical Research Institute and Grant Montgomery PhD from Institute for Molecular Bioscience, University of Queensland, for providing access to control samples - they were not financially compensated for their contributions this work.

Conflict of Interest Disclosures: All authors have completed and submitted the ICMJE Form for Disclosure of Potential Conflicts of Interest. Dr Goldberg reports serving on advisory board for Novartis and Allergan, Mundipharma, and Pfizer, and receiving speakers' fees from Mundipharma and Pfizer. No other disclosures were reported.

2.6 Supplement

Supplementary Tables, Figures and Methods are available at:
https://jamanetwork.com/journals/jamaophthalmology/fullarticle/2704062

CHAPTER 3

Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression

Jamie E. Craig[†], **Xikun Han**[†*], Ayub Qassim[†], Mark Hassall, Jessica N. Cooke Bailey, Tyler G. Kinzy, Anthony P. Khawaja, Jiyuan An, Henry Marshall, Puya Gharahkhani, Robert P. Igo Jr., Stuart L. Graham, Paul R. Healey, Jue-Sheng Ong, Tiger Zhou, Owen Siggs, Matthew H. Law, Emmanuelle Souzeau, Bronwyn Ridge, Pirro G. Hysi, Kathryn P. Burdon, Richard A. Mills, John Landers, Jonathan B. Ruddle, Ashish Agar, Anna Galanopoulos, Andrew J. R. White, Colin E. Willoughby, Nicholas H. Andrew, Stephen Best, Andrea L. Vincent, Ivan Goldberg, Graham Radford-Smith, Nicholas G. Martin, Grant W. Montgomery, Veronique Vitart, Rene Hoehn, Robert Wojciechowski, Jost B. Jonas, Tin Aung, Louis R. Pasquale, Angela Jane Cree, Sobha Sivaprasad, Neeru A. Vallabh, NEIGHBORHOOD consortium, UK Biobank Eye and Vision Consortium, Ananth C. Viswanathan, Francesca Pasutto, Jonathan L. Haines, Caroline C. W. Klaver, Cornelia M. van Duijn, Robert J. Casson, Paul J. Foster, Peng Tee Khaw, Christopher J. Hammond, David A. Mackey, Paul Mitchell, Andrew J. Lotery, Janey L. Wiggs, Alex W. Hewitt, Stuart MacGregor.

Nature Genetics. 2020;52(2):160-166.

Contribution of candidate:

In this study, I am listed as the corresponding and joint-first author. I contributed to study design, data analysis, and the first draft of the manuscript. S.M., J.E.C., A.W.H., X.H., P.G., J.L.W. and D.A.M. designed the study and obtained the funding. X.H., A.Q., M.H., J.N.C.B., T.G.K., A.P.K., P.G.H., J.A., H.M., P.G., R.P.I., J.-S.O., T.Z., O.S., M.H.L. and S.M. analyzed the data. J.E.C., X.H., A.Q., M.H., A.P.K., H.M., R.P.I., S.L.G., P.R.H., O.S., E.S., B.S., P.G.H., K.P.B., R.A.M., J.L., J.B.R., A.A., A.G., A.J.R.W., C.E.W., N.A., S.B., A.L.V., I.G., G.R.-S., N.G.M., G.W.M., V.V., R.H., R.W., J.B.J., T.A., L.R.P., A.J.C., S.S., N.A.V., A.C.V., F.P., J.L.H., C.C.W.K., C.M.v.D., R.J.C., P.J.F., P.T.K., C.J.H., D.A.M., P.M., A.J.L., J.L.W., A.W.H. and S.M. contributed to data collection and contributed to genotyping. X.H., J.E.C., A.Q., A.W.H. and S.M. wrote the first draft of the paper. All authors contributed to the final

version of the paper.

Chapter 3. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression

Glaucoma, a disease characterized by progressive optic nerve degeneration, can cause visual loss and blindness, which can be prevented through timely diagnosis and treatment. We characterize optic nerve photographs of 67,040 UK Biobank participants and use a multitrait genetic model to identify risk loci for glaucoma. A glaucoma polygenic risk score (PRS) enables effective risk stratification in unselected glaucoma cases and modifies penetrance of the *MYOC* variant encoding p.Gln368Ter, the most common glaucoma-associated myocilin variant. In the unselected glaucoma population, individuals in the top PRS decile reach an absolute risk for glaucoma 10 years earlier than the bottom decile and are at 15-fold increased risk of developing advanced glaucoma (top 10% versus remaining 90%, odds ratio = 4.20). The PRS predicts glaucoma progression in prospectively monitored, early manifest glaucoma cases ($P = 0.004$) and surgical intervention in advanced disease ($P = 3.6 \times 10^{-6}$). This glaucoma PRS will facilitate the development of a personalized approach for earlier treatment of high-risk individuals, with less intensive monitoring and treatment being possible for lower-risk groups.

3.1 Introduction

Glaucoma refers to a group of ocular conditions united by a clinically characteristic optic neuropathy associated with, but not dependent on, elevated intraocular pressure.[190] It is the leading cause of irreversible blindness worldwide and is predicted to affect 76 million in 2020.[94,95] There is no single definitive biomarker for glaucoma, and diagnosis involves assessing clinical features, with characterization of the optic nerve head carrying the strongest evidential weight. Primary open-angle glaucoma (POAG) is the most prevalent subtype of glaucoma in people of European and African ancestry.[94,186] POAG is asymptomatic in the early stages, and currently approximately half of all cases in the community are undiagnosed even in developed countries.[183] Early detection is paramount, as existing treatments are unable to restore vision that has been lost, and late presentation is a major risk factor for blindness.[191] Thus, better strategies to identify high-risk individuals are urgently needed,[192] and more refined approaches can capitalize on the fact that POAG is one of the most heritable of all common human diseases.[106,124,193] The lack of a currently cost effective screening strategy for glaucoma,[192] coupled with very high heritability make glaucoma an ideal candidate disease for the development and application of a polygenic risk score to facilitate risk stratification.

Overlap of features shared by healthy optic nerves with those in early stages glaucoma, makes it a difficult disease to diagnose early, necessitating costly ongoing monitoring of patients for progressive optic nerve degeneration.[190] Once a glaucoma diagnosis is established, rates of progression vary widely between individuals, and considerable time can elapse before surveillance techniques adequately differentiate slow from more rapidly progressing cases.[190] Progressive vision loss from glaucoma can be slowed, or in some cases halted, by timely intervention to reduce intraocular pressure using medical therapy, laser trabeculoplasty or incisional surgery.[190] The ability to predict progression is currently crude, with delays in treatment escalation for high-risk individuals an important and inevitable consequence, as well as substantial cost and morbidity associated with overtreatment of lower risk cases.

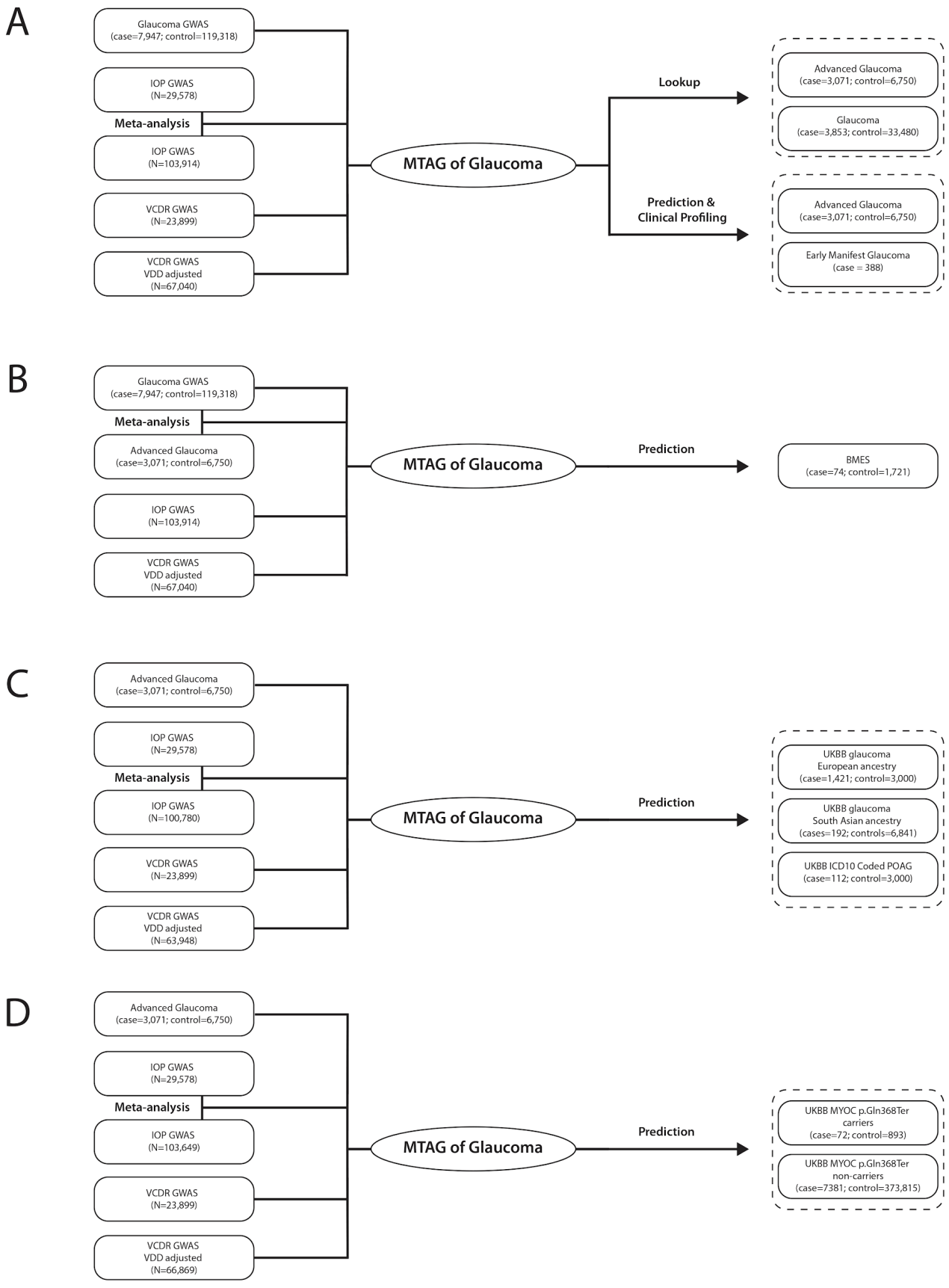
The chronicity, heritability, clinical heterogeneity and treatability of POAG make it an ideal candidate for genetic risk profiling.[194,195] In this study, we evaluated the optic nerve head in

67,040 UK Biobank participants (UKB), enabling the largest genome-wide association study (GWAS) on optic nerve morphology to date, using vertical cup-disc ratio (VCDR) as an endophenotype for glaucoma. We then incorporated additional genetic data from a second well established glaucoma endophenotype, intraocular pressure (IOP), and combined this with glaucoma disease status using a recently developed multiple trait analysis of GWAS (MTAG)[53] approach to firstly identify new risk loci for glaucoma, and secondly generate a comprehensive glaucoma polygenic risk score (PRS). We examined the impact of newly implicated glaucoma genes in independent case-control cohorts from Australia, the United States, and the United Kingdom, and then evaluated the utility of the PRS for predicting glaucoma risk, and important clinical outcomes in well-characterised cases across a range of disease severities.

3.2 Results

3.2.1 Study design

Our overall study design is illustrated in **Extended Data Fig. 1a**. We first conducted a GWAS on glaucoma (7,947 cases and 119,318 controls) and on the key endophenotypes for glaucoma: VCDR (including new data on 67,040 UKB participants, and International Glaucoma Genetics Consortium, IGGC, N = 23,899) and intraocular pressure (including data on 103,914 UKB participants and GWAS summary statistics from IGGC, N = 29,578, Supplementary Table 1). These data were then combined using MTAG[53] to identify new glaucoma risk loci and to construct a PRS. The clinical significance of the PRS was investigated in advanced glaucoma cases in two populations, and a separate prospectively monitored clinical cohort with early manifest glaucoma. The predictive ability of the PRS was also explored in other datasets; however, to ensure our results generalize to further cohorts, we selected mutually exclusive samples for inclusion in the discovery and testing datasets to ensure no sample overlap. When required, we re-derived the PRS to ensure no sample overlap (**Extended Data Fig. 1** panel b, c and d, and Supplementary Note).



Extended Data Fig. 1. Study Design

The multi-trait analysis of GWAS (MTAG) algorithm was applied to datasets of European descent (unless

otherwise specified). In Panel a, MTAG was applied to four datasets (glaucoma case-control GWAS from the UKB; GWAS meta-analysis of intraocular pressure (IOP) from the International Glaucoma Genetics Consortium [IGGC] and the UKB; Vertical cup-disc ratio [VCDR] GWAS data that was either adjusted for vertical disc diameter [VDD] in the UKB dataset; or not adjusted for VDD in the IGGC). Novel variants identified through this analysis were then confirmed in two independent data sets; an Australasian cohort of advanced glaucoma (ANZRAG) and a consortium of cohorts from the United States (NEIGHBORHOOD). The clinical significance of the PRS derived from the MTAG analysis was validated in independent samples; firstly, in advanced glaucoma cases (ANZRAG and samples from Southampton/Liverpool in the UK) and secondly, in a prospectively monitored clinical cohort with early manifest glaucoma (PROGRESSA). Panel b displays the prediction in the Blue Mountains Eye Study (BMES), where we removed the IGGC VCDR and IGGC IOP GWAS from the training datasets, given that they contain BMES data. Panel c shows the prediction in the UKB glaucoma and ICD-10 POAG cases. Here we removed all glaucoma cases and 3000 controls with IOP/VCDR measurements as well as their relatives from UKB VCDR/IOP GWAS. We also evaluated the performance of PRS in non-European ancestry (192 cases and 6,841 controls of South Asian ancestry in UKB). Panel d shows the cumulative risk of glaucoma in UKB. For the analysis of *MYOC* p.Gln368Ter carriers (N=965, case=72, control=893) participants were stratified into tertiles of PRS. We also examined cumulative risk of glaucoma in the general population (*i.e.* in *MYOC* p.Gln368Ter non-carriers, N=381,196, case=7381, control=373,815) stratifying by deciles of the PRS. The discovery and testing datasets were designed to derive the PRS with no sample overlap (Supplement Note).

3.2.2 Discovery of novel optic nerve morphology loci

GWAS of VCDR (adjusted for vertical disc diameter) identified 76 statistically independent, genome-wide significant SNPs (66 loci), of which 49 SNPs (43 loci) had not previously been associated with VCDR (Supplementary Figure 1, Supplementary Figure 2, and Supplementary Table 2). Using LD score regression, we found no evidence for genomic inflation (intercept=1.04, SE=0.01, Supplementary Figure 3). The genetic correlation between VCDR (adjusted for vertical disc diameter) and glaucoma in UKB was 0.50 (SE=0.05); the correlation in effect size estimates at the 76 SNPs was 0.60 ($P=9.0\times10^{-9}$, Supplementary Figure 4). We further combined UKB VCDR (adjusted for vertical disc diameter) GWAS and IGGC VCDR GWAS summary statistics using MTAG, and identified 107 independent genome-wide significant SNPs (across 90 loci, Supplementary Table 3) for VCDR (adjusted for vertical disc diameter). As previously reported, the genetic correlation between intraocular pressure and glaucoma was high (0.71),[126] but as expected the genetic correlation between VCDR (adjusted for vertical disc diameter) and intraocular pressure was substantially lower (0.22, SE=0.03).

3.2.3 Discovery of novel glaucoma loci via multitrait analysis

Given the high correlation between glaucoma and its endophenotypes, we then conducted a multivariate GWAS (with 8,002,429 SNPs after quality control) to identify 114 statistically independent SNPs (107 loci, $P < 5\times10^{-8}$) associated with glaucoma - this includes all previously published glaucoma loci as well as 49 novel loci (**Figure 1**, Supplementary Figure 5, Supplementary Figure 6, and Supplementary Table 4). At the more stringent multiple testing threshold ($P < 1 \times 10^{-8}$) suggested by a simulation study[196] 95 loci reach significance, 39 of which are novel (Supplementary Table 4). Half of top SNPs (27 of 49) at these novel loci were not associated individually with any of the individual input traits at the genome-wide significance level ($P=5\times10^{-8}$), and were only found to reach this threshold for glaucoma due to the MTAG method leveraging the strong correlation between the input traits. We then attempted to replicate the 49 novel SNPs in two independent glaucoma cohorts (ANZRAG and NEIGHBORHOOD). Given the much smaller effective sample size of these replication cohorts (versus the discovery datasets from the MTAG analysis), we did not expect all of the SNPs to be strongly associated - rather if they were genuine associations we would expect the ORs to be highly concordant, with some of the smaller ORs being individually non-significant. The concordance between the discovery cohort and our replication cohorts log ORs was excellent (correlation 0.88, $P=1.6\times10^{-36}$), indicating our multivariate model was successful in identifying genuine glaucoma risk loci (**Figure 2** and Supplementary Figure 7). Of the 49 novel SNPs, nine SNPs were replicated after Bonferroni correction ($P<0.05/49=0.001$, one-sided test, bold text in Supplementary Table 4), 26 SNPs were associated at a nominal significance level ($P<0.05$, one-sided test, italic text in Supplementary Table 4), and 46 (94%) were in the expected direction. Whilst the concordance between the multivariate and the glaucoma replication sample log ORs was high, only nine of the 49 loci were associated with glaucoma after correction for multiple comparisons, and further studies are required to replicate the remaining 40 loci for glaucoma.

We conducted a genome-wide gene-based association analysis and a gene set enrichment analysis to assess which predefined biological pathways were enriched in our multitrait glaucoma GWAS - we found 196 genes and 14 gene sets, respectively, that were significant after Bonferroni correction (Supplementary Table 5 and Supplementary Table 6). The most significant pathways were also previously implicated (i.e. extracellular matrix, collagen, and

circulatory system development).[126,197] Further studies are warranted to investigate the role of these pathways in the risk of glaucoma.

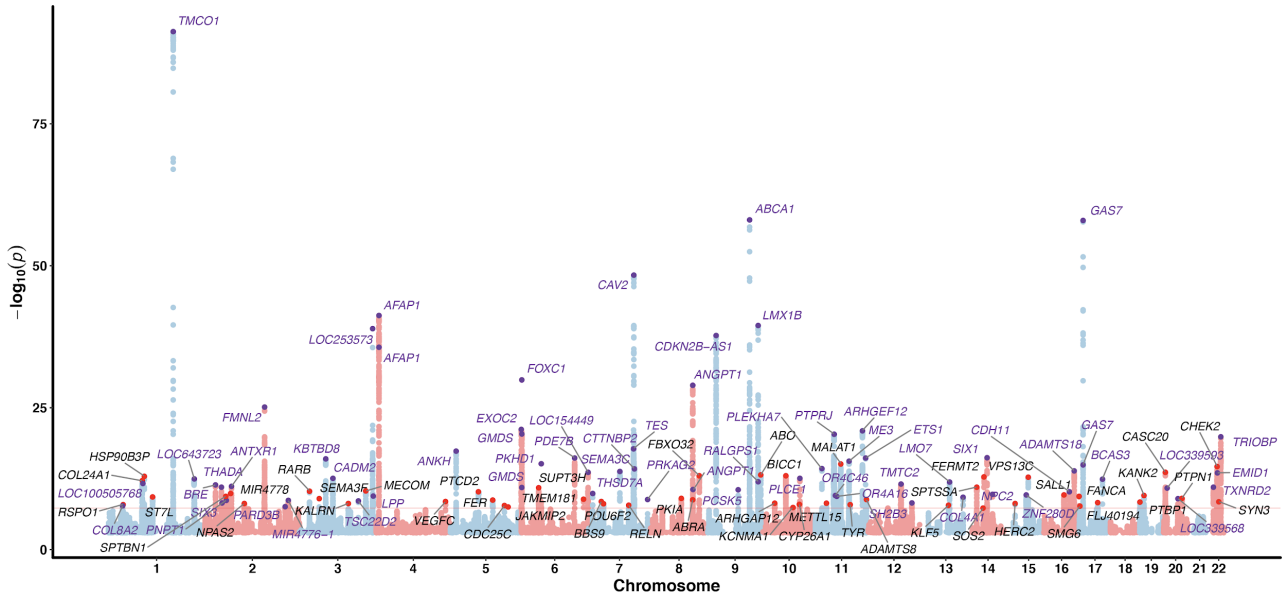


Figure 1. Manhattan plot displaying glaucoma-specific P values from the multi-trait GWAS (MTAG) analysis.

The samples used in multi-trait analysis are presented in Extended Data Fig.1a. Novel SNPs are highlighted in red dots, with the nearest gene names in black text. Known SNPs are highlighted in purple dots, with the nearest gene names in purple text. The red line is the genome-wide significance level at 5×10^{-8}.

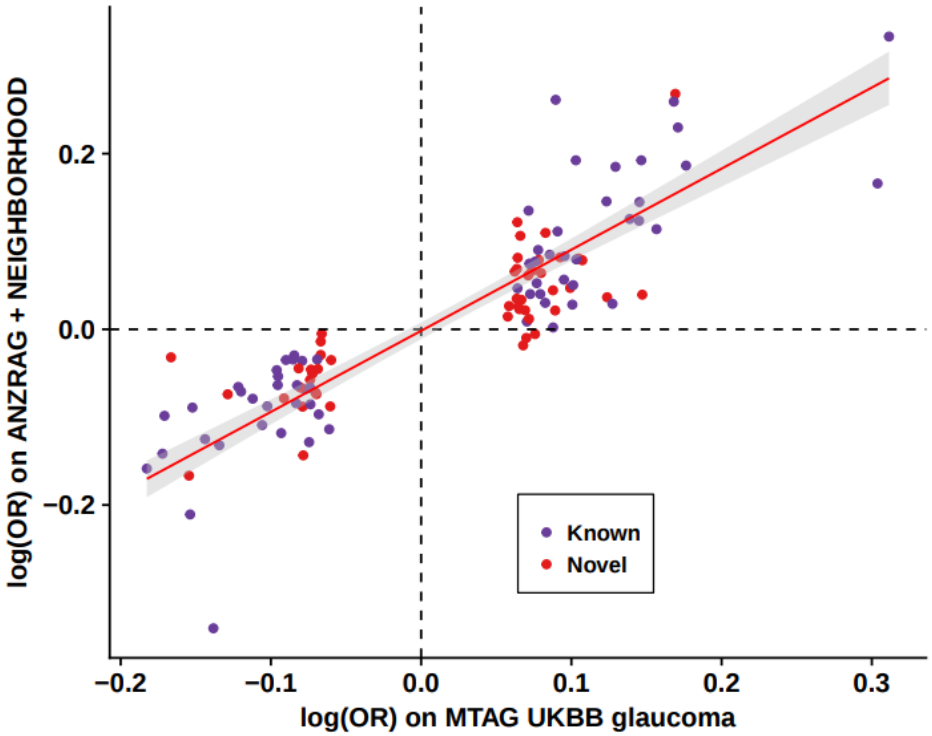


Figure 2. Comparison of the effect sizes (log odds ratio) for 114 genome-wide significant independent SNPs identified from the glaucoma multiple trait analysis of GWAS in the UKB versus those in independent glaucoma cohorts (meta-analysis of ANZRAG and NEIGHBORHOOD).

Pearson's correlation coefficient is 0.88 (P value=1.6×10^{-36}). The red line is the best fit line, with the 95% confidence interval region in grey. Novel glaucoma SNPs are highlighted in red and known SNPs in purple.

3.2.4 Optimizing prediction of glaucoma risk by combining correlated traits

We derived our PRS based on the MTAG of GWAS data from glaucoma and its endophenotypes. As well as increasing the number of SNPs that reach genome-wide significance (mean chi-squared statistic increased from 1.12 to 1.30, implying our effective sample size was 2.59 times larger than if we had used UKB glaucoma cases and controls alone), our multivariate model improved the power of risk prediction by reducing the error in the estimate of the effect size for every SNP (assuming the MTAG homogeneity assumption is true, see discussion).[53] We first tested the discriminatory power of the MTAG-derived PRS in the ANZRAG cohort of advanced glaucoma. We found SNPs with MTAG P values ≤0.001

(corresponding to 2,673 uncorrelated SNPs after LD-clumping at r^2 = 0.1 and P value threshold at 0.001) had the highest Nagelkerke R^2 (13.2%) and AUC (0.68, 95%CI: 0.67-0.70) (Supplementary Table 7). The MTAG PRS has better prediction ability than any of the input traits alone (Supplementary Table 8). Based on this we set the P value threshold at 0.001 for all the remaining prediction target sets (PROGRESSA, Blue Mountains Eye Study [BMES], UKB).

The MTAG-derived PRS was effective at separating advanced glaucoma individuals in terms of risk, with a clear dose-response over deciles (**Figure 3a**, Supplementary Figure 8). In ANZRAG, individuals in the top decile of the PRS had 14.9-fold higher risk (95%CI: 10.7-20.9) relative to the bottom decile, with even better discrimination for the more common high-tension glaucoma (OR=21.5, 95%CI: 12.5-37.0) than normal-tension glaucoma (Supplementary Figure 9). We replicated the dose-response of the PRS in a smaller UK advanced glaucoma dataset (Southampton and Liverpool); the top versus bottom PRS decile had OR=11.6 (95%CI: 6.0-25.3), with again better discrimination for high-tension glaucoma (OR=12.9, 95%CI: 6.2-31.3). While comparing the top and bottom deciles shows the dose-response across deciles, one can also consider the risk in the high PRS individuals versus all others; when this is done in ANZRAG, the OR is 4.2 and 8.5 in the top 10% and 1%, respectively, of individuals versus all remaining individuals (Supplementary Table 9).

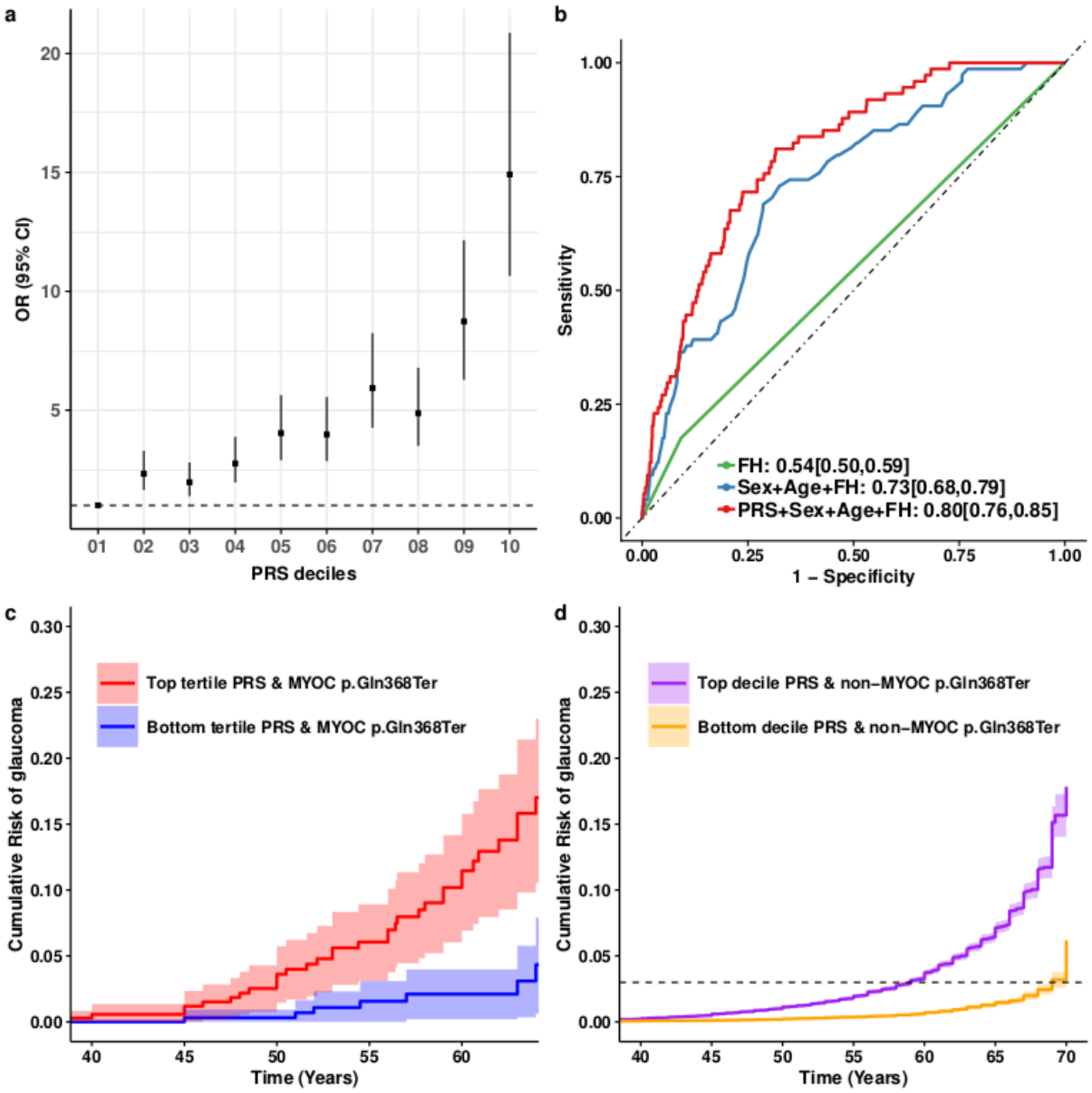


Figure 3. Multiple trait analysis of GWAS PRS prediction.

Panel a shows the odds ratio (OR) of developing advanced glaucoma in the ANZRAG cohort (with 1,734 advanced glaucoma cases and 2,938 controls) for each PRS decile. The square dots are the OR values (adjusted for sex and the first four principal components) and the error bars are 95% confidence interval. The dashed line is the reference at the bottom PRS decile (OR=1). Panel b shows the AUCs of PRS in BMES. The MTAG-derived PRS provided additional predictive ability on top of traditional risk factors (age, sex, and self-reported family history (FH), DeLong's test P value 0.002). The AUC is based on a logistic regression model with the coefficients for age, sex, FH and PRS estimated from the BMES data (Supplementary Table 10). Panel c displays the cumulative risk of glaucoma in UKB *MYOC* p.Gln368Ter carriers stratifying by the PRS (adjusted for sex and first six genetic principal components). Here the cumulative risk of tertiles (with 95% confidence intervals) of PRS are displayed given the relatively small number of *MYOC* p.Gln368Ter carriers (N=965). Panel d plots the cumulative risk of glaucoma for people in the top and bottom decile (with 95% confidence intervals) of PRS of the UKB who do not have the *MYOC* p.Gln368Ter variant (adjusted for sex and first six genetic principal components). The dashed line is the reference line of cumulative risk at 3%.

3.2.5 Glaucoma risk score performance in individuals carrying high penetrance variants

Previous studies indicated that PRS modifies the penetrance of rare BRCA1/2 mutation carriers for breast, ovarian, and prostate cancers.[31,198] Although the MTAG-derived PRS only contains common variants, given it indexes general glaucoma risk, we hypothesized that it could stratify individuals carrying known high-penetrance glaucoma variants. Pathogenic *MYOC* (Myocilin) gene variants account for 2-4% of POAG cases among most populations, the most common disease causing variant being p.Gln368Ter (rs74315329).[116] Penetrance is age-related and is lower in population-based than family based studies.[43,116] We speculated that this difference in penetrance could be due to enrichment of common glaucoma associated variants in families modifying age related penetrance. Within UKB, we identified 965 *MYOC* p.Gln368Ter carriers based on imputation (Supplementary Note).[114] **Figure 3c** shows the cumulative risk of glaucoma in p.Gln368Ter carriers, stratifying by PRS tertiles. For p.Gln368Ter carriers in the lowest tertile PRS, glaucoma risk remained very low (2%) up to age 60. In contrast, the highest tertile PRS group had substantially increased risk of early diagnosis, reaching a 6-fold increase in absolute risk of glaucoma by age 60, relative to the lowest PRS tertile (considering whole age range, hazard ratio=3.4, 95%CI: 1.7-6.6). This supports the utility of PRS in optimizing risk stratification and prediction, and early screening for patients carrying high penetrance *MYOC* variants in the presence of high PRS scores.

3.2.6 Potential for glaucoma risk score in screening in the general population

We considered a general population screening scenario using UKB (PRS was re-derived to ensure no sample overlap, **Extended Data Fig. 1d**), where we excluded the 965 *MYOC* p.Gln368Ter carriers. Over the 40-69 year old age range for individuals sampled in UKB, glaucoma prevalence increases from 0.1% at age 40, reaching 3% (95%CI: 2.9-3.1%) by age 64. The MTAG-derived PRS stratifies UKB participants very effectively; for those in the top PRS decile, 3% prevalence (prevalence in general population) is reached by age 59, whilst it takes an additional 10 years for this disease prevalence to be reached for people in the bottom PRS decile. Alternatively, the prevalence can be well stratified by PRS deciles

(**Figure 3d**).

To benchmark the performance of the MTAG-derived PRS with traditional risk factors, we computed the AUC in datasets for which this was possible: BMES, UKB glaucoma (broad glaucoma definition), and UKB POAG (ICD-10 definition) (**Figure 3b**, Supplementary Table 11, [PRS was re-derived to ensure no sample overlap], **Extended Data Fig. 1**). In the BMES our PRS provided additional predictive ability beyond that imparted by traditional risk factors (age, sex, and self-reported family history (FH)), with a statistical significant change in the AUC (from 0.73 to 0.80, P=0.002, **Figure 3b**). Clear improvement in prediction using this PRS is also observed in people of South Asian ancestry (Supplementary Table 11), though we were underpowered to explore this further across other groups.

A previous study examined the cost-effectiveness requirements for glaucoma screening and highlighted the key age 50-60 bracket.[192] In the BMES data (**Extended Data Fig. 1b**), screening only those with a top decile PRS identified 40% of all early onset cases in age 50-60 bracket (40% of the 10 cases, P=0.013). Such individuals represent a set of individuals likely to benefit from referral for immediate clinical assessment — with skilled clinical examination, retinal imaging, and visual fields. We replicated this result in the UKB POAG cohort (ICD10 cases in **Extended Data Fig. 1c**, top 10% PRS screening finds 29% of 24 cases aged 50-60, P=0.0075). In this way, PRS-based screening would satisfy the cost-effectiveness requirements of *Burr et al*[192], identify a meaningful proportion of cases, and capture those cases most at risk of severe disease.

3.2.7 Clinical implications of the glaucoma risk score

We evaluated the predictive power of the PRS in advanced glaucoma; in 1,336 ANZRAG advanced POAG cases with accurate age at diagnosis information available (Supplementary Table 12), the PRS was significantly associated with age at diagnosis of POAG ($P=1.8\times10^{-5}$). Individuals in the top 10% of the PRS distribution were on average diagnosed 7 years younger than people in the bottom 10% (**Figure 4a**). We also found ANZRAG individuals with higher PRS had more family members affected by glaucoma ($P=3.5\times10^{-9}$), with the highest decile having twice as many members affected (Supplementary Figure 10).

Retinal nerve fibre layer thinning is a major structural change evident in early stage glaucoma.[199] In the early manifest glaucoma (PROGRESSA) cohort the PRS predicted both the proportion lost, and rate of loss of peripapillary retinal nerve fibre layer. Given that glaucomatous loss of retinal ganglion cells generally progresses unequally between eyes, with some quadrants of the retina damaged more rapidly than others, we analysed the most affected quadrant of the most affected eye in individuals with early manifest glaucoma and greater than two years of longitudinal optical coherence tomography data. The PRS was significantly associated with the proportion of retinal nerve fibre layer lost from baseline to most recent review, even after adjustment for known risk factors; age, intraocular pressure and retinal nerve fibre layer thickness at presentation (P=0.004; **Figure 4b**, Supplementary Table 13). Expressed in terms of rate of loss, each decile change in PRS was associated with an accelerated progression rate of 0.05 μM/year, which was twice the rate of thinning per mmHg (approximately 1 decile change for intraocular pressure) of baseline intraocular pressure (0.022 μM/year).

Incisional surgery for glaucoma (trabeculectomy) is highly effective at reducing intraocular pressure, but has important complications which can adversely impact vision.[190] Trabeculectomy is performed either when intraocular pressure is unable to be controlled with medical or laser therapy, or when there is progressive visual field loss despite well controlled intraocular pressure. Patients with a high PRS were more likely to have undergone surgery for glaucoma (**Figure 4c**, Supplementary Figure 11). In the ANZRAG cohort of POAG cases, a higher PRS was associated with requiring trabeculectomy, even after adjustment for maximum recorded intraocular pressure and age ($P=3.6\times10^{-6}$), the OR of requiring trabeculectomy in either eye for people in the top PRS decile was 1.78 (95%CI: 1.07–3.00) compared to the bottom decile. We observed a very similar trend in our UK replication (Southampton/Liverpool) samples (Supplementary Figure 11).

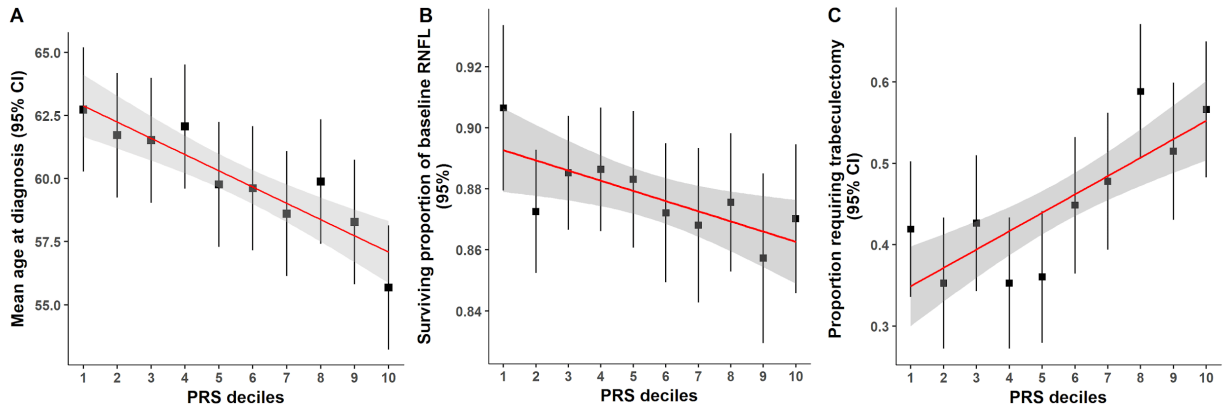


Figure 4. Clinical implications of the glaucoma PRS

Panel a shows the mean age at diagnosis (years) for each decile of PRS in the ANZRAG cohort (linear regression $P=1.8\times10^{-5}$). A total of 1,336 cases had accurate age at diagnosis information. We calculated the mean age at diagnosis for each decile of PRS, adjusted for sex and the first four principal components in a linear regression model. The square dots are the regression-based mean age at diagnosis, with error bars for 95% confidence intervals. The red line is the line of best fit, with 95% confidence intervals in grey. Panel b shows the proportion of preserved baseline retinal nerve fibre layer for PROGRESSA participants with early manifest glaucoma plotted against PRS decile (N=388; linear regression P=0.004). The square dots are the retinal nerve fibre layer proportions, with error bars showing 95% confidence intervals. The remaining retinal nerve fibre layer proportion is calculated for the most affected quadrant of the most affected eye of each patient — as determined on optical coherence tomography scans at baseline and latest follow-up scan. Panel c displays the proportion of patients requiring trabeculectomy in either eye in the ANZRAG POAG cohort (linear regression $P=3.6\times10^{-6}$). There were 1,360 cases with records of surgical treatment status. The square dots represent the observed average proportion of cases in each decile of PRS who required trabeculectomy, with 95% confidence interval bars. The line of best fit is shown in red, with 95% confidence interval shaded in grey.

3.3 Discussion

Through a large-scale multivariate GWAS we identified novel genes for glaucoma, the leading cause of irreversible blindness worldwide.[94] Despite a smaller replication cohort, many of these novel hits were replicated, and all but three SNPs showed a consistent direction of effect. We then expanded this analysis to derive a PRS and interrogated its utility across a wide spectrum of clinically relevant glaucoma outcomes.

From the multivariate GWAS, we identified 49 novel loci associated with glaucoma (nine of which replicated after correction for multiple comparisons in independent glaucoma case-

control cohorts; 26 were replicated with P<0.05). Interestingly, most of the loci replicated at P<0.001 are genes previously associated with glaucoma risk factors (myopia, CCT, IOP, VCDR). Specifically, *RSPO1* is associated with ocular axial length.[200] *BICC1* is associated with myopia and corneal astigmatism.[201,202,203] *POU6F2* modulates corneal thickness and increases glaucoma risk in animal experiments.[204] *FBXO32*, *PTPN1*, and *VPS13C* are associated with IOP,[125,126,131] whilst *CASC20* was identified in our VCDR (adjusted for vertical disc diameter) GWAS. These findings show our multivariate GWAS improves power to identify novel glaucoma genes and advance our understanding of the causes of glaucoma risk.

The MTAG-derived PRS was validated in independent samples, confirming its high predictive ability. Individuals in the top PRS decile were at 15-fold increased risk of advanced glaucoma, and at 21.5-fold increased risk of advanced high tension glaucoma, relative to the bottom decile; a substantial improvement on previously reported genetic profiling strategies (where, based on SNPs that were genome-wide significantly associated with intraocular pressure and SNPs previously associated with VCDR and glaucoma, top decile individuals had a 5.6-fold increased risk).[126] This new glaucoma PRS also outperforms those derived from other well-studied conditions; for example our OR comparing the top 1% PRS individuals versus the remaining individuals was 8.5 which is higher than that seen in a recent study which surveyed coronary artery, atrial fibrillation, type 2 diabetes, inflammatory bowel disease and breast cancer.[33] The aetiology of complex diseases depends on both environmental and genetic factors, thus PRS alone will never achieve the very high predictive power (*e.g.* AUC >0.99) required for accurate population screening.[36] Our glaucoma PRS will be primarily useful for stratifying individuals into risk groups; for example in the BMES data, screening the top decile of the PRS in individuals between 50-60 years old identifies 40% of cases. Moreover, as argued by Khera et al[33], individuals with a high PRS for glaucoma are likely to be at a similar risk to individuals carrying rare “high penetrance” *MYOC* mutations.[43] Finally, the PRS performance for glaucoma is particularly noteworthy given the clinical implications of identifying at-risk individuals and the prevention of irreversible blindness with readily available treatment proven to be effective at preventing visual loss.

Whilst current treatments are effective in preventing or reducing POAG progression,[195] many patients are not diagnosed before irreversible damage to visual function has already

occurred. Earlier diagnosis of glaucoma can reduce glaucoma blindness, and our work demonstrates that people with a higher PRS require earlier clinical assessment. In the UKB, individuals in the top PRS decile reach an equivalent absolute risk for glaucoma 10 years earlier than people in the bottom decile. In advanced glaucoma cases, individuals in the top decile were diagnosed 7 years earlier than those in the bottom decile. Similarly, the MTAG-derived PRS was associated with significantly earlier disease onset in UK Biobank *MYOC* p.Gln368Ter carriers who are at high disease risk. The MTAG-derived PRS can also identify people with early manifest glaucoma who are at higher probability of disease progression, as well as the likelihood of requiring surgical intervention, which is highly effective at reducing intraocular pressure, but carries substantial treatment morbidity meaning it should always be targeted specifically to those at higher risk of disease progression and blindness.

A concern in MTAG method is the homogeneous assumption which could be violated for some SNPs that have no effect on one trait but non-null for other traits (*i.e.* it is possible that a small number of the variants may be more specific for IOP or VCDR rather than glaucoma). The homogeneity assumption has been studied in detail by Turley et al.[53] We have evaluated the possible inflation using max False Discovery Rate (maxFDR) as recommended by Turley et al.[53] The baseline maxFDR for MTAG glaucoma-specific input GWAS summary statistics is 0.049, and the maxFDR for MTAG glaucoma-specific output summary statistics is 0.03. As these are similar, there is no evidence of inflation due to violation of the homogeneity assumption. As recommended by the MTAG authors, we also performed replication analysis to assess the credibility of novel SNPs in two independent data sets (an Australasian cohort of advanced glaucoma [ANZRAG] and a consortium of cohorts from the United States [NEIGHBORHOOD]); this analysis shows there is very good concordance between the MTAG based effect sizes and those from the glaucoma cohorts. Furthermore, using MTAG output instead of the individual input traits improves the predictions in independent cohorts (Supplementary Table 8), providing additional evidence that we are not merely identifying IOP or VCDR specific loci that have no effect on glaucoma. Further research needs to be undertaken to investigate the biological mechanisms of these novel genes on glaucoma risk.

A limitation of this work, is that in our 7,947 UKB glaucoma cases, only a small proportion had documented disease subtype; however, since the proportion of UK glaucoma cases that have POAG is high (87% in a recent study[186]), this is unlikely to have a large influence

on our results. A further limitation is that it is not yet clear how applicable our findings are to other populations. We showed that the PRS improved prediction accuracy over and above traditional risk factors in homogeneous groups (as defined by genetic principal components) of either European or South Asian ancestry. The performance of the PRS in other populations should be tested to investigate the generalizability of our findings. The performance of the PRS in aiding clinical decision making and guiding earlier treatment could be evaluated prospectively in a longitudinal intervention study, with participants randomized to have their PRS provided or withheld from their treating specialist.

In summary, we have applied a multivariate approach using weighted data on glaucoma, and endophenotypes intraocular pressure and VCDR, to identify novel glaucoma loci, and develop a polygenic risk score. This PRS was shown to be predictive of: 1) increasing risk of advanced glaucoma; 2) glaucoma status significantly beyond traditional risk factors; 3) earlier age of glaucoma diagnosis; 4) high levels of absolute risk in persons carrying high penetrance glaucoma variants; 5) increasing likelihood of disease progression in early stage disease, and 6) increasing likelihood of incisional glaucoma surgery in advanced disease. This glaucoma PRS has good predictive power across a range of clinical cohorts and its application will facilitate the rational allocation of resources through clinical screening and timely treatment in high-risk patients, with reduced clinical monitoring costs in lower risk groups.

3.4 Methods

3.4.1 Study Design And Overview

Our overall study design is illustrated in **Extended Data Fig. 1**. We first conducted a GWAS on glaucoma and on the key endophenotypes for glaucoma: VCDR and intraocular pressure. These data were then combined using MTAG,[53] a method for combining multiple genetically correlated traits to maximize power for identifying new loci and improving genetic risk prediction. Specifically, our MTAG analysis outputs *glaucoma-specific effect size estimates* and P-values for single nucleotide polymorphisms (SNPs) across the genome. Newly associated loci ($P<5\times10^{-8}$) were validated in two independent cohorts with well-characterised POAG. We created a PRS based on the MTAG GWAS summary statistics. The clinical significance of the PRS was investigated in advanced glaucoma cases in two populations, and a separate prospectively monitored clinical cohort with early manifest glaucoma. The predictive ability of the PRS was also explored in other datasets; however, to ensure our results generalize to further cohorts, we selected mutually exclusive samples for inclusion in the discovery and testing datasets to ensure no sample overlap. When required, we re-derived the PRS to avoid any sample overlap (**Extended Data Fig. 1**). Study procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research.

3.4.2 Study Populations

Detailed information of individual studies, phenotypic definitions, and genetic quality control procedures are provided in the Supplementary Note.

The UK Biobank (UKB) is a population-based study of half a million people living in the United Kingdom.[130] We measured VCDR and vertical disc diameter in all subjects with gradable retinal images (67,040 participants following exclusions, detailed in Supplementary Note) and undertook a GWAS to identify SNPs influencing optic nerve head morphology. Vertical disc diameter adjustment of the VCDR was used to account for optic cup and disc size covariation.[205,206] To improve power in the multi-trait analysis we combined the VCDR data with data on corneal-compensated intraocular pressure (103,914 participants), and glaucoma (7,947 cases, 119,318 controls) in the MTAG analysis.[126] We also used publicly

available VCDR and intraocular pressure GWAS summary results for individuals of European descent from the International Glaucoma Genetics Consortium (IGGC; N_{VCDR}=23,899, $N_{intraocular\ pressure}$=29,578).[133]

The Australian & New Zealand Registry of Advanced Glaucoma (ANZRAG) comprises 3,071 POAG cases of European descent, who were compared to 6,750 controls.[122,176] For sub-analyses restricted to advanced POAG, there were 1,734 advanced POAG cases and 2,938 controls, and of these cases 1,336 participants had accurate age at diagnosis information available. Replication of the ANZRAG findings was performed using 332 advanced glaucoma cases from Southampton and Liverpool in the United Kingdom; for case-control analysis, cases were matched to 3,000 randomly selected European ancestry individuals from the QSkin Sun and Health study.[207] The National Eye Institute Glaucoma Human Genetics Collaboration Heritable Overall Operational Database (NEIGHBORHOOD) GWAS results were generated through meta-analyzing summary data from eight independent datasets (3,853 POAG cases, 33,480 controls) of European ancestry from the United States.[208]

The Blue Mountains Eye Study (BMES) is a population-based cohort study investigating the etiology of common ocular diseases among suburban residents aged 49 years or older, in Australia.[183] Data from 74 POAG cases and 1,721 controls of European descent with genotype information were included.

The Progression Risk Of Glaucoma: RElevant SNPs with Significant Association (PROGRESSA) study is a prospective longitudinal study of the clinical and genetic risk factors, and course of early-stage glaucoma (N=388). Patients with confirmed early manifest POAG on sequential automated perimetry testing were consecutively recruited from ophthalmology clinics in South Australia (detailed criteria in Supplementary Note). Individuals underwent six-monthly evaluation of intraocular pressure, optic disc assessment, retinal nerve fibre layer analysis by optical coherence tomography, and achromatic Humphrey visual field perimetry. Longitudinal data were used from all visits since baseline presentation; participants were followed for one to eight years. The change in retinal nerve fibre layer was measured between the baseline optical coherence tomography and the most recent scan in the most-affected quadrant of the most-affected eye. Treating clinicians and graders were unaware of the patient's genetic risk for glaucoma or any PRS data.

POAG in the ANZRAG, NEIGHBORHOOD, BMES, and PROGRESSA cohorts was defined as outlined previously.[91] and in accordance with the consensus statement from the World Glaucoma Association.[209] Intraocular pressure was not used in the clinical case definition of POAG.[209]

3.4.3 Statistical Analysis

Detailed information on the statistical analysis is provided in the Supplementary Note.

For the VCDR (adjusted for vertical disc diameter) and intraocular pressure GWAS in UKB, we used linear mixed models (BOLT-LMM software) to account for cryptic relatedness and population stratification adjusting for sex, age and the first ten principal components.[210] We meta-analyzed UKB intraocular pressure GWAS results with those from the IGGC using the inverse variance weighted method (METAL software).[211] For the UKB glaucoma GWAS, we removed relatives (pi-hat>0.2 calculated using identity by descent determined based on autosomal markers) and used PLINK software for association analysis.[182]

We then conducted a multitrait GWAS using the MTAG (version 1.0.7) software to combine the European descent GWAS summary statistics from UKB glaucoma, UKB VCDR (adjusted for vertical disc diameter), IGGC VCDR and the intraocular pressure meta-analysis (**Extended Data Fig. 1**).[53] MTAG performs joint analysis of GWAS summary results from related traits to improve statistical power to identify new genes and to maximize the predictive ability of our polygenic risk scores.[53] In MTAG, GWAS summary results from related traits are used to construct the variance–covariance matrix of their SNP effects and estimation error; MTAG improves the accuracy of effect estimates by incorporating information from other genetic correlated traits. The MTAG method explicitly models sample overlap in the input studies and provides valid estimates even when sample overlap is present.[53] To benchmark the increase in effective sample size relative to just using UKB glaucoma, we calculated ($\underline{\chi 2}_{\text{MTAG}}$ -1) / ($\underline{\chi 2}_{\text{GWAS}}$-1), where $\underline{\chi 2}_{\text{MTAG}}$ and $\underline{\chi 2}_{\text{GWAS}}$ are the mean chi-squared statistics from MTAG and the UKB glaucoma analyses, respectively.[53]

We used a stepwise model selection procedure in the GCTA-COJO software to identify

independent genome-wide significant SNPs.[212] Gene-based and pathway analysis were conducted in MAGMA (v1.06), as implemented in FUMA (version 1.3.1).[213,214]

Prediction was based on the estimated glaucoma odds ratios (OR) from the MTAG analysis. To derive a PRS we considered a range of P-value thresholds (5×10-8, 1×10-5, 0.001, 0.05, 1) with LD-clumping r2=0.1 for inclusion of SNPs in the prediction model, applying each to our first prediction cohort (advanced glaucoma from ANZRAG). To avoid falsely inflating prediction accuracy, we applied the threshold with greatest predictive value in ANZRAG ($P \leq 0.001$) for the subsequent predictions into other target sets (rather than repeatedly taking the best P-value threshold for each of the datasets). We tested the LDpred[44] approach for PRS construction although the predictions were no better than those from the thresholding approach described above. There was no sample overlap between any of the training and target datasets (**Extended Data Fig. 1**).

Bivariate LD score regression was used to estimate the genetic correlation between pairs of traits.[73] The "pROC" package was used to calculate the area under the curve (AUC).[215] Analyses were performed with R software.[216]

3.5 Acknowledgements

This work was conducted using the UK Biobank Resource (application number 25331) and publicly available data from the International Glaucoma Genetics Consortium. The UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council (UK), Department of Health (UK), Scottish Government, and Northwest Regional Development Agency. It also had funding from the Welsh Assembly Government, British Heart Foundation, and Diabetes UK. The eye and vision dataset has been developed with additional funding from The NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology, Fight for Sight charity (UK), Moorfields Eye Charity (UK), The Macula Society (UK), The International Glaucoma Association (UK) and Alcon Research Institute (USA). This work was also supported by grants from the National Health and Medical Research Council (NHMRC) of Australia (#1107098; 1116360, 1116495, 1023911), the Ophthalmic Research Institute of Australia, the BrightFocus Foundation, UK and Eire Glaucoma Society and Charitable Funds from Royal Liverpool University Hospital. SM, JEC, KPB, and AWH are supported by NHMRC Fellowships. SM was supported by an Australian Research Council Future Fellowship. LRP is supported by NIH R01 EY015473. XH is supported by the University of Queensland Research Training Scholarship and QIMR Berghofer PhD Top Up Scholarship. We thank David Whiteman, Rachel Neale and Catherine Olson for providing access to QSKIN samples for use as controls as part of NHMRC Grant 1063061. We thank Scott Wood, John Pearson and Scott Gordon from QIMR Berghofer for support. The NEIGHBORHOOD consortium is supported by NIH grants P30 EY014104, R01 EY015473 and R01 EY022305.

Competing interests

D.A.M. is consultant/advisor to Allergan, Inc. J.E.C., A.W.H. and S.M. are listed as coinventors on a patent application for the use of genetic risk scores to determine risk and guide treatment.

Data sharing

The UKB data are available through the UK Biobank Access Management System https://www.ukbiobank.ac.uk/. The GWAS summary statistics from the glaucoma MTAG analysis is available for research use at https://xikunhan.github.io/site/publication/. We will

return the derived data fields following UKB policy; in due course, they will be available through the UK Biobank Access Management System.

3.6 Supplement

Supplementary Note, Figs. 1–13 and Tables 1–13 are available at:
https://www.nature.com/articles/s41588-019-0556-y

CHAPTER 4A

Genome-wide association analysis of 95,549 individuals identifies novel loci and genes influencing optic disc morphology

Xikun Han, Ayub Qassim, Jiyuan An, Henry Marshall, Tiger Zhou, Jue-Sheng Ong, Mark M Hassall, Pirro G Hysi, Paul J Foster, Peng T Khaw, David A Mackey, Puya Gharahkhani, Anthony P Khawaja, Alex W Hewitt, Jamie E Craig, Stuart MacGregor.

Human Molecular Genetics. 2019;28(21):3680-3690.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Ayub Qassim assisted with the data analysis of gene expression in human ocular tissues and manuscript preparation. Stuart MacGregor, Jamie E Craig, Alex W Hewitt, and Puya Gharahkhani obtained funding and designed the study. All authors contributed to interpretation of the results and the final version of the paper.

Chapter 4A. Genome-wide association analysis of 95,549 individuals identifies novel loci and genes influencing optic disc morphology

Optic nerve head morphology is affected by several retinal diseases. We measured the vertical optic disc diameter (DD) of the UK Biobank (UKB) cohort (N = 67 040) and performed the largest genome-wide association study (GWAS) of DD to date. We identified 81 loci (66 novel) for vertical DD. We then replicated the novel loci in International Glaucoma Genetic Consortium (IGGC, N = 22 504) and European Prospective Investigation into Cancer–Norfolk (N = 6005); in general the concordance in effect sizes was very high (correlation in effect size estimates 0.90): 44 of the 66 novel loci were significant at $P < 0.05$, with 19 remaining significant after Bonferroni correction. We identified another 26 novel loci in the meta-analysis of UKB and IGGC data. Gene-based analyses identified an additional 57 genes. Human ocular tissue gene expression analysis showed that most of the identified genes are enriched in optic nerve head tissue. Some of the identified loci exhibited pleiotropic effects with VCDR, intraocular pressure, glaucoma and myopia. The genome-wide genetic correlation between DD and vertical cup-to-disc ratio (VCDR) was very high (rg=0.50, $P=6.18\times10^{-21}$), whereas genetic correlation between DD and POAG was very small (rg=0.01, P=0.78). These results can enhance our understanding of the genetics of optic disc morphology and shed light on the genetic findings for other ophthalmic disorders such as glaucoma and other optic nerve diseases.

4A.1 Introduction

The optic disc is an oval structure representing the exit point of the retinal ganglion cell axons as they form the optic nerve responsible for transmitting vision to the brain. Anatomically, the optic disc can be divided into the neuroretinal rim where the nerves fibre layer turns outwards away from the retina, and the optic cup, which does not contain any nerve fibre layer and is located more central to the neuroretinal rim. The optic disc is clinically examined using fundoscopy and optical imaging technology (such as the Heidelberg Retinal Tomography [HRT] or the Optical Coherence Tomography [OCT]) for signs of retinal and optic nerve diseases.[217,218] Common metrics of the optic disc morphology include the disc size (measured as the vertical disc diameter [DD] or disc area) and the vertical cup-to-disc ratio (VCDR). Primary open angle glaucoma (POAG) is an optic neuropathy characterised by an accelerated loss of the retinal ganglion layer and is a leading cause of blindness globally.[219] The loss of the retinal ganglion cells manifests as an enlarged optic cup and VCDR and is considered a hallmark of POAG.[93]

A better understanding of factors that influence optic disc size is of high clinical relevance. Optic disc size affects the structural morphology of the optic nerve head and may influence the vulnerability of the nerve fibres.[220] Small optic nerves are associated with disorders such as optic nerve hypoplasia and increased risk of non-arteritic anterior ischaemic optic neuropathy and optic disc drusen.[221,222] There is a strong correlation between the optic disc size and the VCDR (clinically and genetically) and this should be taken into account in funduscopic examination.[223] Adjusting optic disc parameters such as VCDR for disc size improves their diagnostic power and clinical utility for glaucoma assessment.[224] For example, adjusting the VCDR to DD improves its sensitivity of identifying eyes with perimetric glaucoma from 67% to 76.6% (at 80% specificity).[224] Optic disc morphology traits are highly heritable with an estimated heritability of disc size and VCDR of 0.48 - 0.57 from family studies while the single nucleotide polymorphism (SNP) -based heritability are estimated to range from 0.27 to 0.31.[133,225] To date, however, less than 20 genes are shown to be implicated with disc size.[133,226–230] In this study, we conduct the largest genome-wide association study (GWAS) for disc size (triple the previously studied sample size) to investigate the development mechanisms of the optic nerve and shed light on the genetic understanding for some eye diseases such as POAG and other optic nerve diseases.

4A.2 Results

In the discovery stage, we conducted GWAS on vertical DD in 67,040 UK Biobank (UKB) samples, then we replicated the novel associated candidate loci in independent cohorts from the International Glaucoma Genetics Consortium (IGGC, N=22,504) and EPIC-Norfolk (N=6,005).

4A.2.1 UK Biobank disc diameter GWAS identifies 66 novel loci

From the vertical disc diameter GWAS of 67,040 UKB participants, we identified 91 lead genome-wide significant independent SNPs (81 loci), of which 67 SNPs (66 loci) had not previously been associated with disc diameter (Figure 1A, Supplementary Figure S1 and Table S1). Interestingly, we also identified two genes located in the X chromosome (*EFNB1* and *ZIC3*, Supplementary Table S2), which play an important role in eye development.[231,232] We conducted linkage disequilibrium (LD) score regression and observed no evidence for genomic inflation (intercept=1.05, SE=0.01, Supplementary Figure S2). As previously reported,[133] the genetic correlation between disc diameter and VCDR was very high (rg=0.50, $P=6.18\times10^{-21}$). The genetic correlation between DD and POAG was very small (rg=0.01, P=0.78). The strong association between DD and VCDR is due to the physiologically larger vertical cup diameter and optic disc rim area found in larger optic discs.[223] A higher count of optic nerve fibres is found histologically in eyes with larger optic discs,[233] representing the larger neuroretinal rim area seen on fundoscopy. The genetic correlation of disc size between UKB and IGGC was 0.83 ($P=1.31\times10^{-76}$).

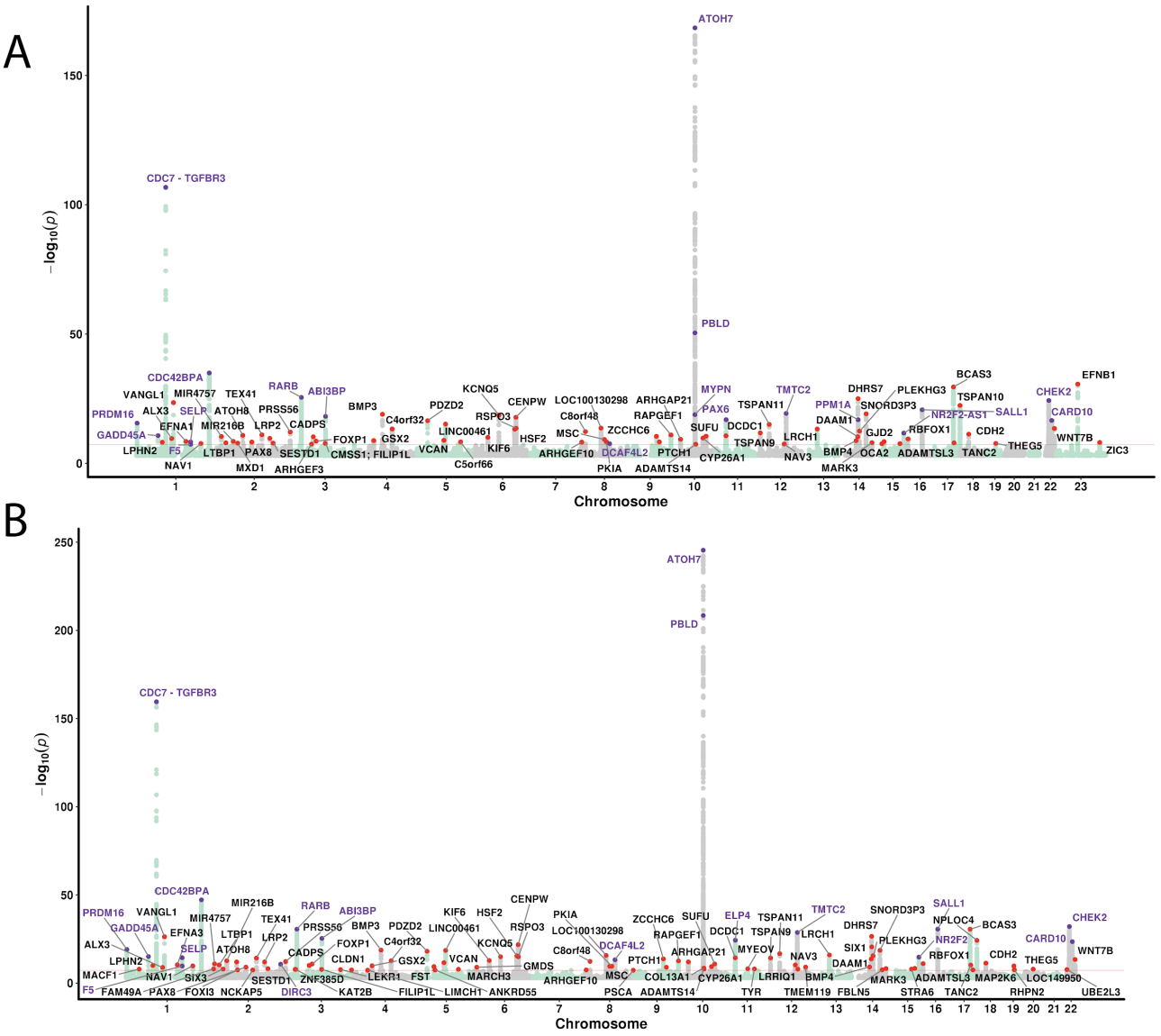


Figure 1. Manhattan plot of disc size genome-wide association studies.

A) UK Biobank dataset; **B)** UK Biobank and IGGC meta-analysis. Novel loci are highlighted in red dots, with the nearest gene names in black text. Known loci are highlighted in purple dots, with the nearest gene names in purple text. The red line is the genome-wide significance level ($P = 5\times10^{-8}$).

4A.2.2 Replication in IGGC and EPIC-Norfolk datasets

We then replicated the identified lead vertical DD loci in IGGC and EPIC-Norfolk datasets. The correlation in effect size estimates at the lead genome-wide significant SNPs were 0.90 ($P=2.85\times10^{-33}$, Figure 2), indicating the identified disc diameter SNPs from UKB could be well replicated (Supplementary Table S1). Of the 64 novel loci from autosomal chromosomes, 19 loci could be replicated in IGGC and EPIC-Norfolk after Bonferroni

correction (P=0.05/64=7.8×10^{-4}, Table 1), and 44 loci have nominal association (P=0.05, Supplementary Table S1). The X chromosome GWAS results are not available in the IGGC and EPIC-Norfolk cohorts. Therefore, further studies are needed to confirm the associations of these two X chromosome genes with DD.

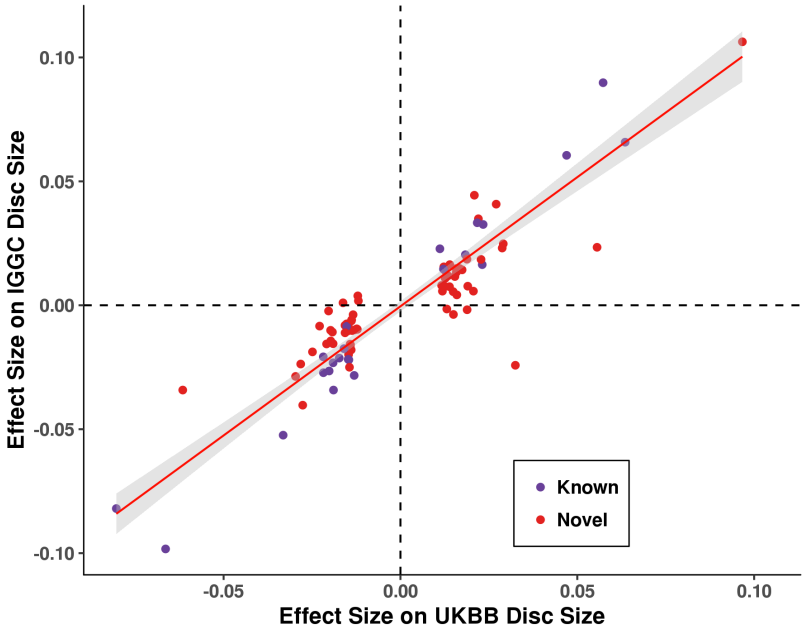


Figure 2. Comparison of the effect sizes for 91 genome-wide significant independent SNPs identified from UK Biobank disc size GWAS versus those in independent cohort of IGGC disc size GWAS.

Pearson's correlation coefficient is 0.90 (P value=2.85×10^{-33}). The red line is the best fit line, with the 95% confidence interval region in grey. Novel disc size SNPs are highlighted in red and known SNPs in purple.

Table 1: List of 19 novel disc size loci replicated in IGGC and EPIC-Norfolk datasets after Bonferroni correction.

SNP	CHR	BP	Nearest Gene	EA	NEA	FREQ	BETA (UKB)	P (UKB)	Z score (META)[1]	P (META)
rs12136690	1	116208944	*VANGL1*	C	T	0.76	-0.02	3.1E-24	5.40	6.7E-08
rs56412756	1	201605477	*NAV1*	C	T	0.92	0.02	2.0E-08	-3.96	7.6E-05
rs9967780	2	56234942	*MIR216B*	G	T	0.78	-0.01	3.0E-09	4.73	2.3E-06
rs4832012	2	86000500	*ATOH8*	G	C	0.49	-0.01	1.4E-11	3.69	2.3E-04
rs1365902	2	145470699	*TEX41*	T	C	0.33	-0.01	8.7E-12	-4.91	9.1E-07
rs3914468	2	170157400	*LRP2*	A	G	0.70	-0.01	2.0E-10	-3.58	3.4E-04
rs77877421	3	71182447	*FOXP1*	A	T	0.94	-0.03	2.6E-09	-3.47	5.1E-04
rs72759609	5	31952051	*PDZD2*	T	C	0.90	0.03	3.1E-17	4.02	5.9E-05
rs58531939	5	87823968	*LINC00461*	T	C	0.91	-0.03	6.3E-16	-4.55	5.3E-06
rs2092524	6	39529692	*KIF6*	G	A	0.66	-0.01	8.5E-11	4.18	3.0E-05
rs12661045	6	122682795	*HSF2*	C	T	0.70	0.02	6.1E-14	-3.46	5.5E-04
rs2152876	6	126761228	*CENPW*	G	A	0.54	-0.02	1.5E-18	5.36	8.2E-08
rs9401928	6	127298394	*RSPO3*	G	A	0.55	-0.02	2.7E-14	4.44	9.0E-06
rs6999835	8	78948855	*PKIA*	T	C	0.63	0.01	5.1E-09	3.74	1.8E-04
rs10512176	9	89252706	*ZCCHC6*	T	C	0.72	-0.01	3.5E-11	-4.07	4.8E-05
rs10764494	10	25058144	*ARHGAP21*	C	A	0.32	-0.01	4.7E-10	3.84	1.2E-04
rs76567987	12	31037655	*TSPAN11*	A	G	0.84	0.02	8.4E-16	4.01	6.1E-05
rs9534439	13	47192049	*LRCH1*	T	C	0.19	0.02	5.1E-14	4.36	1.3E-05
rs61985972	14	59550263	*DAAM1*	A	G	0.94	0.03	5.2E-11	5.02	5.2E-07

Abbreviations: BETA, beta coefficient; CHR, Chromosome; EA, effect allele; FREQ, allele frequency of effect allele; NEA, non-effect allele; SNP, single nucleotide polymorphism; P, P values.

UKB, UK biobank data; IGGC, International Glaucoma Genetic Consortium; META, meta-analysis results of IGGC and EPIC-Norfolk datasets.

Chromosomal position is based on the NCBI RefSeq hg19 human genome reference assembly.

[1] A sample size weighted meta-analysis was conducted.

4A.2.3 Meta-analysis of UKB and IGGC

We subsequently conducted a GWAS meta-analysis to combine UKB and IGGC disc size datasets, and identified 115 independent genome-wide significant SNPs (from 101 loci, and an additional 26 novel disc size loci, Supplementary Figure S3 and Table S3).

4A.2.4 Functional annotation

We further investigated the potential biological functions of the identified genome-wide significant variants ($P<5\times10^{-8}$). Figure 3A showed the functional categories of the genome-wide significant variants. We found 0.83% of the significant variants (77 variants from 39 genes, Supplementary Table S4) located in the exonic regions, and 85% of the significant variants were intronic or intergenic. A 15-core chromatin state was evaluated for 127 tissue or cell types,[214,234] and we found 5.48% (510 variants, Figure 3B and Supplementary Table S5) significant variants were active transcription start sites. The RegulomeDB score was used to identify regulatory elements, and Figure 3C indicated that 1.36% (113 variants, Figure 3C and Supplementary Table S6) significant variants were at least eQTL and transcription factor binding sites. Figure 3D presented the Combined Annotation Dependent Depletion (CADD) score distribution, which is an integrative metric to measure variant deleteriousness.[235,236] The higher the score, the more likely that a SNP is deleteriousness (suggested threshold 12.37),[235] which could be used to prioritize causal variations (Supplementary Table S7).

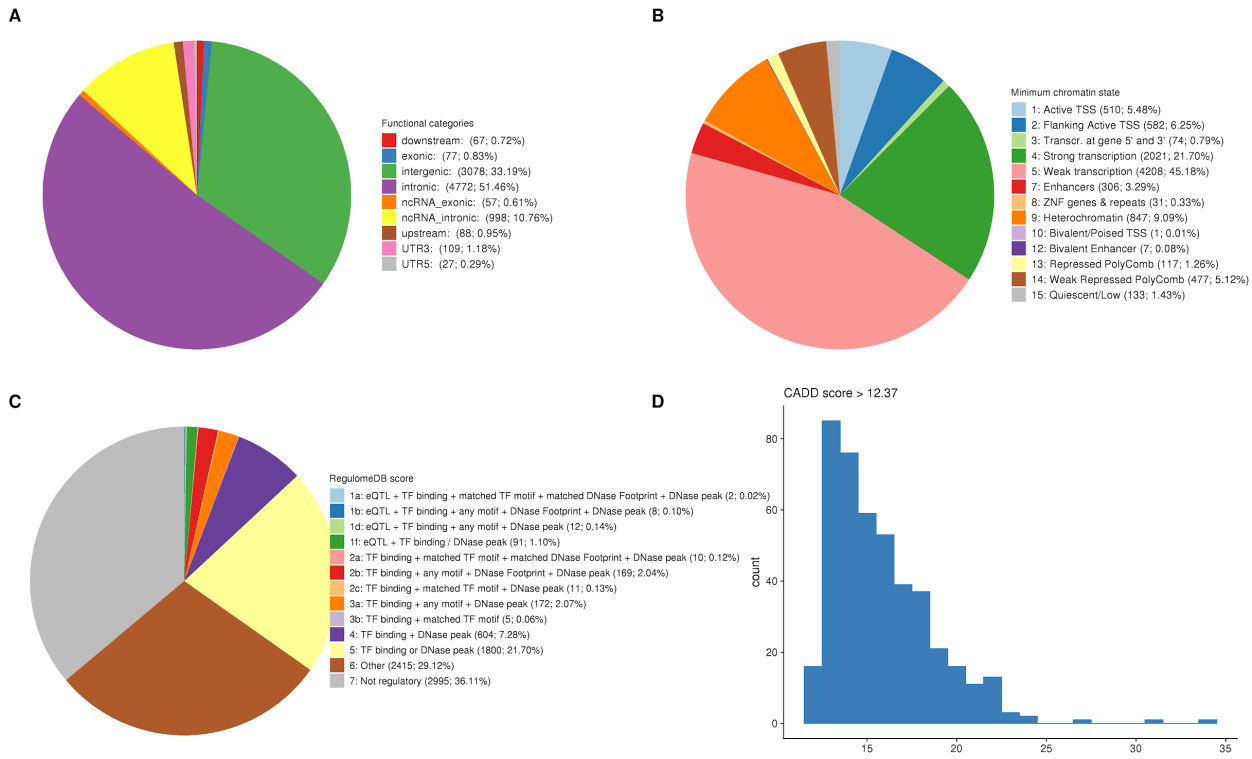


Figure 3. Functional annotation of genome-wide significant variants ($P<5\times10^{-8}$).

A) functional categories; **B)** the minimum chromatin state across 127 tissues; **C)** the Regulome

database score; **D)** the frequency of SNPs with high CADD score (more than the threshold 12.37). For **A**-**C**, the numbers and percentages in parentheses in the legends refer to the number of significant SNPs and the percentages in all identified significant variants. CADD, combined annotation dependent depletion. ncRNA, noncoding RNA; TF, transcription factor; TSS, transcription start site; UTR, untranslated region.

4A.2.5 Gene-based and pathway analysis

We then conducted a genome-wide gene-based association analysis and identified an additional 57 novel genes (without genome-wide significant SNPs in genes, Supplementary Figure S4 and Table S8). For instance, gene *THSD4* was associated with eye tail length and outercanthal width,[237] and genes *RNLS*, *DENND1A*, *RASGEF1B*, *FAM150B*, and *NCOA2* were associated with myopia. Tissue expression analysis of GTEx data (V7 30 general tissue types) indicated the gene expression profiles were enriched in nerve tissue (Supplementary Figure S5). Pathway analysis of 10,678 gene sets (MsigDB v6.2, curated gene sets: 4,761, Gene Ontology terms: 5,917) resulted in 29 significant gene sets after Bonferroni correction, which include sensory organ development, tissue development, and morphogenesis (Supplement Table S9). The top pathway is RAMJAUN_APOPTOSIS_BY_TGFB1_VIA_MAPK1_UP, which is a transforming growth factor-beta (TGFbeta) activated signalling pathway, involved in apoptosis and the regulation of cell growth and survival. [238,239]

4A.2.6 Genetic correlation with other traits

We estimated the genetic correlation between disc size and 832 traits in LD-Hub database (v1.9.0).[240] We only found significant genetic correlation between disc size and myopia (UKB data field 6147: Reason for glasses/contact lenses, rg=-0.24, $P = 5.94\times10^{-8}$, Supplementary Table S10) after Bonferroni correction (0.05/832).

We also investigated GWAS Catalog[9], a curated collection of published genome-wide association studies, for disc size genome-wide significant SNPs (Supplementary Table S11). Our results showed some of the lead disc size loci had pleiotropy effects. For instance, lead SNPs in genes *CDC42BPA* and *ANKRD55* were associated with macular thickness, and lead SNPs in *ANKRD55*, *PRSS56*, *KCNQ5*, *NPLOC4*, and *BMP4* were related to myopia.

4A.2.7 Gene expression in human ocular tissues

We also investigated the expression profile of the genes nearest to the identified SNPs in ocular tissue: optic nerve head, optic nerve, retina, trabecular meshwork, iris, ciliary body, sclera, cornea, as well as foetal retinal tissue (Supplementary Table S12, Table S13, Supplementary Figure S6 and S7).[241] The majority (94/106, 89%) displayed differential expression in the optic nerve head relative to all other ocular tissue (Supplementary Table S14). *BCAS3* (Microtubule Associated Cell Migration Factor), *DHRS7* (Dehydrogenase/Reductase 7) and *NPLOC4* (Nuclear Protein Localization Protein 4 Homolog) were the most significantly differentially expressed genes in the optic nerve head and were all novel discoveries. The SNP rs12147505 in *DHRS7* with no linkage disequilibrium with rs34935520 (in SIX6, $R^2 < 0.001$),[133] had the highest magnitude of effect on disc size; it is a protein coding gene functioning as a catalyst in oxidation and reduction of a wide range of substrates.[242] It is expressed in all ocular tissues, with highest expression in the corneal stroma followed by the optic nerve head, and suggested to be a risk locus for POAG.[180]

4A.2.8 eQTL and transcriptome-wide association analysis

We looked up the lead DD genome-wide significant SNPs in retina from the Eye Genotype Expression (EyeGEx) database to identify expression quantitative trait loci (eQTL).[170] We identified 38 SNP-gene pairs (cis-eQTLs) after gene-level multiple testing correction across the genome (Supplementary Table S15). We also conducted summary data-based Mendelian randomization (SMR) and heterogeneity in dependent instruments (HEIDI) analysis to test the effects of genetic variants on disc size that is mediated by gene expression levels.[243] From SMR approach, we identified 15 genes after multiple testing correction ($P_{SMR} < 0.05/5592 = 8.94 \times 10^{-6}$, Supplementary Table S16). The HEIDI tests ($P_{HEIDI} \geq 0.05$) suggested that genes *CTD-2292P10.4*, *CTNNAL1*, *MFSD13A*, TSPAN11 and *TANC2* (based on updated EyeGEx database) are associated with DD via the underlying GWAS hits.

4A.3 Discussion

We conducted the largest optic disc size GWAS to date and identified 101 loci including 159

genes using manual grading of the UKB fundus photographs. We identified for the first time two genes located in the X chromosome (*EFNB1* and *ZIC3*) associated with vertical disc diameter. Our results indicate that sensory organ development, tissue development, and morphogenesis are involved in the biological pathway for optic disc morphology. The discovery of genes and pathways involved in optic disc morphology is important in the understanding of the genetic architecture and development mechanisms of optic nerve head and would increase our knowledge of diseases related to this - POAG and other optic nerve diseases.

The identified optic disc size genes have important functions. For instance, the top two novel replicated genes are *VANGL1* and *CENPW*. VANGL Planar Cell Polarity Protein 1 (*VANGL1*) regulates the establishment of planar cell polarity, which plays a key role in tissue morphogenesis, embryonic development, and the development of eye tissues.[244–246] *CENPW* encodes Centromere Protein W, which is related to cell cycle, mitotic state, and chromosome maintenance.[247] The lead SNP rs2152876 in *CENPW* exhibits a pleiotropic effect, as its proxy SNPs ($R^2 > 0.8$) are associated with intraocular pressure[131], height[248], hip circumference[249], and the age onset of menarche[250]. The encoded protein by *STRA6* acts as a receptor for retinol-binding protein responsible for the cellular uptake of vitamin A, which is critical to the normal development of the eyes.[251] Indeed mutations in *STRA6* impairing this function lead to severe developmental abnormalities in the eyes such microphthalmia, anophthalmia and coloboma.[251,252] *SIX3*, *PRSS56* and *PAX6* are also involved in the eye development. *PAX6* has been labelled as the master control gene for the morphogenesis of the eye, and is regulated by the transcriptional regulator *SIX3*.[253] *BCAS3* and *RSPO3* are involved in angiogenesis, vascular support and cell migration.[254,255] *BMP4* antagonises transforming growth factor-beta 2 (TGF-β2) signalling, a cytoke involved in the synthesis and deposition of extracellular matrix in the optic nerve head.[256] This pathway is implicated in the pathological remodelling of the optic nerve head in glaucoma,[256] and deficiency of *BMP4* results in an abnormal optic nerve with loose connective tissue.[257] All together, these gene findings help us have a better understanding of the development of the eye and related traits.

Optic disc size is highly correlated with the vertical cup-to-disc ratio (VCDR)[133], one of the main glaucoma endophenotypes, and is important for the interpretation of a glaucomatous optic disc.[220,223] Clinically, adjusting VCDR to DD improves its utility as larger discs are more

likely to have physiologically larger cups.[224] In clinical genetics, genes are more likely to be involved in the pathogenesis of glaucoma if associated with larger VCDR but not disc size, or VCDR adjusted to disc size. For instance, variation in the *PDZD2* gene is associated with optic cup area and VCDR,[133] and our study identifies the same variation to be strongly associated with disc size. This would suggest that the observed association with VCDR is likely due to the disc size rather than a pathological enlargement of the optic cup. Similarly, the previously reported association between *F5* and VCDR is likely related to disc size due to its larger association with disc size in our study and previously.[230] When the disc size is adjusted for, Springelkamp *et al.* have reported the estimated effect size of the *F5* variant on VCDR is negligible.[133] Several of the identified disc size genes are correlated with intraocular pressure. For instance, genes *TMEM119*, *CENPW*, *LTBP1*, *TEX41*, and *PKIA* are reported to be associated with intraocular pressure,[125,126,129,131] which could represent pleiotropic effects of these genes. Correlating disc size loci with the genes for glaucoma and its endophenotypes would help to identify the role of these genes in glaucoma pathogenesis.

Understanding the genetics of optic disc size will also contribute towards the understanding of the etiology of myopia.[227,258] The clinical morphology of the optic disc in myopic eye is distinct and the disc size correlates with the magnitude of the refractive error especially in high myopes.[259] From the genetic correlation results using LD Hub,[240] we found disc size is negatively correlated with myopia (rg=-0.24, $P=5.94\times10^{-8}$), and positively correlated with hypermetropia (rg=0.24, $P=9.18\times10^{-5}$). Several novel disc size associated SNPs discovered in this study have been previously reported to be associated with myopia. Of these, variant rs74764079 located in the exonic region of bone morphogenetic protein 3 (*BMP3*) had a large CADD score (24.5). The Serine Protease 56 (*PRSS56*) and Bone Morphogenetic Protein 4 (*BMP4*) genes are involved in ocular growth as variations in these genes are associated with microphthalmia (abnormally small eyes).[260–262] Variations in *PRSS56* and *BMP4* are reported to be associated with myopia in previous GWAS.[202,263] Our results support the involvement of these genes in optic disc morphology. The tetraspanin 10 (*TSPAN10*) gene is involved in cellular protein trafficking and regulation and organ development.[264] Variations in *TSPAN10* and *NPLOC4* are associated with myopia and macular thickness (derived from ocular coherence tomography [OCT] scans).[201,265] The association with optic disc size further supports their involvement in eye development.

There are several limitations for this study. Firstly, we only evaluated individuals of European

ancestry in our disc size GWAS, hence the generalizability of the genetic findings to other populations remains unclear. However, the concordance in lead SNP effect sizes was also very high between UKB disc size GWAS and IGGC disc size GWAS in Asian population (Pearson's correlation coefficient 0.72, Supplementary Figure S8). Another limitation is that in UKB we measured the vertical disc diameter rather than disc area (measured in IGGC dataset). However, in our UKB GWAS, we used the rank-based inverse-normal transformed disc diameter, which we subsequently rescaled to disc area (Supplementary Methods).[266,267] In our sensitivity analysis, we used multiple trait analysis of GWAS (MTAG) to joint analyse UKB and IGGC disc size summary statistics; the results are essentially identical to traditional inverse-variance meta-analysis (Supplementary Table S3, Figure S9, Supplementary Methods). Nonetheless, the identified novel loci and genes from meta-analysis and gene-based analysis still need further studies to replicate these findings. A third limitation is that disc size was not available for both the left and right eyes in all cohorts. For instance due to the lengthy manual process of grading 67,040 UKB fundus photos, we graded the DD on the left eye where the image quality was good, otherwise the right eye was used; however, the DD between both eyes are expected to be very similar. In the EPIC-Norfolk sample set (N = 6,005), the measurement of the vertical disc diameter was 2.34 ± 0.26 mm in the right eyes, compared 2.33 ± 0.26 mm in the left eyes, which is also consistent with previous studies.[268] A fourth limitation is that although differences in ocular magnification and tilted appearance of some optic discs could affect disc size measurements, in practice the effects of these are expected to be small [269]. In our sensitivity analysis, we adjusted spherical equivalent refractive error in UKB disc size GWAS; the results were essentially unchanged.

In conclusion, we conducted a meta-analysis GWAS of 95,549 individuals for disc size and identified 101 genomic loci across 159 genes. This study enhanced our understanding of the genetics of disc size and would shed light on the genetic findings for other eye traits such as VCDR, intraocular pressure, POAG, and myopia.

4A.4 Materials and Methods

Study procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research.

4A.4.1 UK Biobank disc diameter phenotype data

UK Biobank is a large-scale cohort study that included over half a million participants aged between 40-69 years in 2006-2010 from across the United Kingdom. In the UKB, 87,685 left fundus retinal eye images were available (two assessment visits), covering 84,871 participants (UKB Field: 21015). The longest vertical DD was measured at the inner edges of the scleral ring from non-stereo fundus images obtained using a Topcon 3D OCT-1000 MKII (Topcon Corporation). The images had a 45^{o} field of view and were cropped and enlarged to facilitate grading using a custom Java program.

Two thousand images were randomly selected for quality control, and the Pearson's correlation coefficient of the DD measurements between the two examiners was 0.64 (95% confidence interval [CI]: 0.61-0.67; supplementary Figure S10). The second visit DD measurements were used if available, otherwise, we used the first visit measurements (N=52,199, proportion 76%). If the left eye images were ungradable, we used the right eye images instead (N=6,181, proportion 9%, UKB Field 21016). In DD GWAS, we excluded non-white British ancestry participants based on principal components (PCs). Finally, 67,040 participants were included in our analysis. For statistical analysis, we applied a rank-based inverse-normal transformation to DD, effectively rendering the results in terms of disc area (see statistical methods, below).

4A.4.2 UK Biobank genotype data

Detailed information of the genotype data and quality control procedures for UKB was reported by Bycroft and colleagues.[130] Briefly, approximately 488,000 participants were genotyped for 805,426 markers on Axiom arrays (Affymetrix Santa Clara, USA). After standard quality control procedures, ~96M genotypes were imputed using Haplotype Reference Consortium (HRC) and UK10K haplotype resources.[130,270,271] In the association analysis, we removed single nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) less than 0.01 or imputation quality score less than 0.3. Finally, 8,928,767 SNPs were kept for association analysis.

4A.4.3 IGGC disc size summary statistics

The phenotype and genotype data of disc size for IGGC have been previously described elsewhere.[133] We downloaded the publicly available disc area GWAS summary statistics for

22,504 individuals of European ancestry from IGGC. In IGGC, the genetic data was imputed to 1000 Genomes reference panel, and the disc area GWAS was adjusted for age, sex and the first five principal components.[133]

4A.4.4 EPIC-Norfolk Eye Study

The European Prospective Investigation into Cancer (EPIC) study is a pan-European prospective cohort study designed to investigate the aetiology of major chronic diseases.[272] EPIC-Norfolk, one of the UK arms of EPIC, recruited and examined 25,639 participants between 1993 and 1997 for the baseline examination.[273] 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. Ophthalmic assessment formed part of the third health examination and this has been termed the EPIC-Norfolk Eye Study.[274]

In total, 8,623 participants were seen for the Eye Study between 2004 and 2011. Digital photographs of the optic disc and macula were taken using a TRC-NW6S non- mydriatic retinal camera and IMAGEnet Telemedicine System (Topcon Corporation, Tokyo, Japan) with a 10-megapixel Nikon D80 camera (Nikon Corporation, Tokyo, Japan). Pupils were not dilated. Images were graded at the Moorfields Reading Centre. Measurement of the vertical diameter of the optic disc was made using adobe photoshop C55 software.

99.7% of EPIC-Norfolk are of European descent and we excluded participants of non-white European ancestries. 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.

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.[133] 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 6,005 of these participants (which includes no overlap with the 1,096 participants contributing to the IGGC meta-analysis[133]) that contributed to the EPIC-Norfolk disc size GWAS 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 implausible identity-by-descent values. We removed related individuals with pairwise relatedness corresponding to third-degree relatives or closer across all genotyped participants. Following these exclusions, there were no ethnic outliers. Data were pre-phased using SHAPEIT version 2 and imputed to the Phase 3 build of the 1000 Genomes project (October 2014) using IMPUTE (version 2.3.2).

We examined the relationship between allele dosage and mean of right and left vertical disc diameter using linear regression adjusted for age, sex and the first five principal components. Analyses were carried out using SNPTEST software (version 2.5.1).[275]

4A.4.5 Ocular gene expression analysis

Gene expression data was available from RNA extraction of 21 healthy donor eyes from 21 individuals. We analysed 63 tissues of cornea (epithelium, stroma and endothelium), trabecular meshwork, ciliary body, iris, retina, optic nerve and optic nerve head. RNA quality was assessed using Agilent Bioanalyzer 2100 RNA 6000 Nano Assay and samples were included for sequencing only if the RIN scores were greater than or equal to 3.8 and both 28S and 18S ribosomal RNA intensity peaks were prominent. RNA sequencing was done using Illumina NextSeq® 500 (San Diego, USA), followed by quality check (FASTQC v0.11.3). Trimgalore (v0.4.0) was used to trim low quality bases (Phred score < 28) and reads shorter than 20 bases after trimming were discarded. All reads which passed every quality control step were then aligned to the human genome (GRCh38 assembly) with ≤ 2 mismatches per read. Downstream analysis was done with edgeR (version 3.22.5).[276] We selected genes expressed 10 times (1.5 counts per million) in at least 5 tissue samples and normalised the libraries using trimmed mean of M-values (TMM).[277] Estimating dispersions was done via Cox-Reid profile-adjusted likelihood method,[278] and differential expression was compared between optic nerve head and all other tissues via negative binomial generalised linear model.[279] Genes were filtered to those nearest to the identified SNPs, and the differential expression P-values were adjusted using Bonferroni correction.

4A.4.6 Genome-wide association analysis and meta-analysis

For DD GWAS in UKB, we first applied a rank-based inverse-normal transformation.[267]

Since disc diameter and disc area are monotonically related, applying such a transformation makes the phenotype correlation between diameter and area effectively approach one, although to render them back to the same output scale, one should multiply by the standard deviation of the trait, which is approximately 0.4 mm^2 for disc area.[133] To ensure consistency with previously reported results, all our analyses are presented on the disc area scale.[133] For association analysis in UKB, we used a linear mixed model in BOLT-LMM software (version 2.3).[210] Analysis was performed under an additive genetic model, adjusted for the effect of sex, age, the first ten principal components, two indicator variables for examiners who performed the measurements, and fundus retinal image assessment visits. (84) A stepwise model selection procedure in the GCTA-COJO software (version 1.91.7beta) was used to identify independent lead genome-wide significant SNPs.[212] We reported independent SNPs with both unconditional P values and joint P values less than 5×10^{-8}. For genomic regions with multiple independent SNPs, we defined a 'locus' as a region at least 400 kilobases from the adjacent locus. Bivariate linkage disequilibrium (LD) score regression was used to estimate the genetic correlation between pairs of traits.[73] To replicate the lead SNPs from UKB, we conducted a sample size based meta-analysis in METAL (2011-03-25 release) for IGGC and EPIC-Norfolk datasets.[211] For the UKB and IGGC meta-analysis, we performed the inverse-variance weighted fixed-effect meta-analysis in METAL.[211] In our sensitivity analysis, rather than performing meta-analysis using the effect size estimates and standard errors, we also conducted the multiple trait analysis of GWAS (MTAG, software version 1.08) approach, a framework to generalize the standard inverse-variance meta-analysis method, with the approach able to joint analyse the same trait with different measures or even different traits with a high genetic correlation.[53] The general analyses were performed with R software (version 3.4.1).[216] Additional details are provided in supplementary Methods.

4A.4.7 Gene-based and pathway tests

We used MAGMA (v1.07) for gene-based and pathway analysis as implemented in FUMA (version 1.3.4).[213,214] In gene-based tests, GWAS summary statistics of SNPs were mapped to 18,619 genes, and the association P values for a set of SNPs were calculated. The default parameters in FUMA were used. Bonferroni method was used for multiple testing correction ($P < 0.05/18,619$). In pathway tests, 10,678 predefined gene sets (MsigDB v6.2, curated gene sets: 4,761, GO terms: 5,917) were tested for enrichment.

4A.4.8 Functional annotation

The functional annotation of SNPs was implemented in FUMA (version 1.3.4).[214] Firstly, the functional annotation of SNPs on the genes were obtained from ANNOVAR.[280] The detailed annotations included downstream, exonic, intergenic, and upstream. We then checked the chromatin states of the SNPs. The 15-core chromatin states were predicted by ChromHMM software based on 5 chromatin marks for 127 epigenomes.[234] We also looked up the Regulome database score and CADD score.[235,281] Briefly, the Regulome database score represents the evidence of regulatory function of SNPs based on eQTLs and chromatin marks. The highest score 1a means that those SNPs may affect regulatory elements while score 7 means not regulatory. CADD score is an integrative metric to measure variant deleteriousness of SNPs predicted by 63 functional annotations. The higher the score, the more likely that a SNP is deleteriousness (suggested threshold 12.37),[235] which could be used to prioritize causal variations.

4A.4.9 eQTL lookup and SMR method

The lead DD SNPs were looked up in the Eye Genotype Expression (EyeGEx) database of retinal tissue to identify retina specific expression quantitative trait loci (eQTL).[170] We also applied SMR (summary data-based Mendelian randomization) and HEIDI (heterogeneity in dependent instruments) tests based on DD meta-analysis summary statistics and the EyeGEx eQTL data.[243] The SMR approach uses both of GWAS summary statistics and eQTL data to test if the effect of a SNP on the phenotype is mediated by gene expression, which could be used to prioritize GWAS hits for further functional studies. The HEIDI method can test the null hypothesis that there is a single causal variant affecting both gene expression levels and phenotype risk.

4A.5 Acknowledgements

This work was conducted using the UK Biobank Resource (application number 25331) and publicly available data from the International Glaucoma Genetics Consortium. SM, JEC and AWH are supported by NHMRC Fellowships. SM was supported by an Australian Research Council Future Fellowship. XH is supported by the University of Queensland Research Training Scholarship and QIMR Berghofer PhD Top Up Scholarship. We thank Scott Wood, Xiaping Lin, John Pearson and Scott Gordon from QIMR Berghofer for support.

Conflict of interest disclosures

The authors declare no potential conflicts of interest.

4A.6 Supplement

Supplementary Figures and Tables are available at:
https://academic.oup.com/hmg/article-abstract/28/21/3680/5559949

CHAPTER 4B

Automated AI labelling of optic nerve head enables new insights into cross-ancestry glaucoma risk and genetic discovery in >280,000 images from UKB and CLSA

Xikun Han, Kaiah Steven, Ayub Qassim, Henry N Marshall, Cameron Bean, Michael Tremeer, Jiyuan An, Owen Siggs, Puya Gharahkhani, Jamie E Craig, Alex W Hewitt, Maciej Trzaskowski, Stuart MacGregor.

American Journal of Human Genetics. 2021; 108(7):1204-1216.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Kaiah Steven and Maciej Trzaskowski contributed to the training of convolutional neural network models and manuscript preparation. S.M., M.T., X.H. and K.S. designed the research. S.M., M.T., A.W.H., J.E.C. and P.G. obtained the funding. X.H., K.S., M.T. and S.M. executed the research and analysed the data. X.H., K.S., M.T. and S.M. wrote the first draft of the manuscript. A.Q., H.N.M., C.B., M.T., O.S., P.G., J.E.C. and A.W.H. interpreted the results. All authors contributed to interpretation of the results and the final version of the manuscript.

Chapter 4B. Automated AI labelling of optic nerve head enables new insights into cross-ancestry glaucoma risk and genetic discovery in over 280,000 images from the UK Biobank and Canadian Longitudinal Study on Aging

Cupping of the optic nerve head, a highly heritable trait, is a hallmark of glaucomatous optic neuropathy. Two key parameters are vertical cup-to-disc ratio (VCDR) and vertical disc diameter (VDD). However, manual assessment often suffers from poor accuracy and is time-intensive. Here, we show convolutional neural network models can accurately estimate VCDR and VDD for 282,100 images from both UK Biobank and an independent study (Canadian Longitudinal Study on Aging), enabling cross-ancestry epidemiological studies and new genetic discovery for these optic nerve head parameters. Using the AI approach we perform a systematic comparison of the distribution of VCDR and VDD, and compare these with intraocular pressure and glaucoma diagnoses across various genetically determined ancestries, which provides an explanation for the high rates of normal tension glaucoma in East Asia. We then used the large number of AI gradings to conduct a more powerful genome-wide association study (GWAS) of optic nerve head parameters. Using the AI based gradings increased estimates of heritability by ~50% for VCDR and VDD. Our GWAS identified more than 200 loci for both VCDR and VDD (double the number of loci from previous studies), uncovers dozens of novel biological pathways, with many of the novel loci also conferring risk for glaucoma.

4B.1 Introduction

The optic nerve head is the exit point of retinal ganglion cell axons from the eye to the brain.[282] It is commonly assessed during ophthalmic examinations using fundoscopy or optical imaging technology for multiple ocular diseases, such as glaucoma, which is the leading cause of irreversible blindness globally and is characterized by characteristic cupping of the optic disc as a result of retinal ganglion cell apoptosis.[90,91] Enlarged vertical cup-to-disc ratio (VCDR) is considered a hallmark of glaucomatous optic neuropathy and is often used to define glaucoma in general population based prevalence surveys.[93] However, VCDR alone is not adequate to assess glaucomatous damage in part because of the variation of optic disc size. For instance, a vertical cup:disc ratio of 0.5 in a small optic disc could be pathologic whereas a vertical cup:disc ratio of 0.8 in a large disc size may represent physiologic cupping. Adjusting for optic disc size is hence important to maximizing the clinical utility of VCDR in diagnosing glaucoma.

Family studies have shown that optic disc morphology traits are highly heritable with an estimated heritability of 0.48 and 0.57 for VCDR and optic disc diameter, respectively.[225] Large-scale genome-wide association studies (GWAS) for optic disc morphology have begun to shed light on the development and pathogenesis of glaucoma and other optic nerve diseases.[133,206,230] However, both large sample sizes and accurate phenotyping are critical in GWAS and further progress will be difficult under the existing manual phenotype paradigm. Manual assessment of optic disc photographs is time-intensive and often suffers from poor inter-observer concordance, even when performed by trained specialists and an alternative approach is required.[283,284] Clinical estimates of VCDR are more difficult from monoscopic photographs compared with stereoscopic viewing of the optic nerve head which can be achieved during slit-lamp biomicroscopy or from stereoscopic photographs.

Recent advances in artificial intelligence (AI) algorithms have shown exciting promise in healthcare[285], including the automated diagnosis of eye diseases.[286,287] With the high performance of AI technology, the U.S. Food and Drug Administration approved the first medical device to use AI technology to detect diabetic retinopathy in 2018.[288,289] The probabilistic nature and non-linear capabilities, as well as analytical capabilities to deal with single and multimodal, high-dimensional data, has seen application of AI experience lower resistance to adoption in the medical field when applied to computer vision applications. Two

fundamental properties have facilitated AI application to medical diagnostics. Firstly, the problem space (medical imaging) is, relative to other medical domains, well studied and very well understood. Secondly, an observation of the output can be quickly validated by a clinical practitioner, who by having access to additional clinical or historical data about that patient, may suggest alternative diagnosis. A motivating factor driving utilisation of AI on data such as fundus images is the large volume of images available for algorithms to be trained on. Furthermore, standardised imaging techniques can drastically reduce the dataset heterogeneity. This is highlighted by the collection of images as part of the UK Biobank (UKB) and the Canadian Longitudinal Study on Aging (CLSA) biobanks completed over a decade. Automated diagnosis from retinal fundus imaging has been approached through a number of different algorithms, ranging from multi-stage "classical" learning algorithms to end-to-end deep learning models.[290–293]

In this study, a convolutional neural network (CNN) model was utilised in a transfer learning approach, training on clinical assessments of the optic nerve head in ~70,000 photographs (Labelled Training Data) of UKB participants. Automatic labelling by the CNN model dramatically boosts the effective sample size (n=282,100 total images graded), presenting an opportunity to greatly expand the utility of the GWAS approach for VCDR and optic disc diameter. We also apply the AI labels systematically across the multiple different ancestries in UKB and CLSA and investigate how VCDR and other glaucoma risk factors, such as IOP, relate to glaucoma risk in different ancestries.

4B.2 Results

4B.2.1 Study Design And Overview

The overall study design is summarised in Figure 1. We use transfer learning to train three CNN models for image gradability, VCDR, and vertical disc diameter (VDD) values from ~70,000 UKB fundus images graded by clinicians. These models were then applied to all UKB fundus images (85,736 participants and 175,770 images in total) and another independent cohort - CLSA (29,635 participants and 106,330 images in total). We performed the largest AI-based GWAS for VCDR and VDD, and replicated novel genetic discoveries in clinician-graded fundus images from International Glaucoma Genetics Consortium (IGGC) and in glaucoma case-control studies (UKB and the Australian and New Zealand Registry

of Advanced Glaucoma; ANZRAG). The large scale biobank data for both VCDR and IOP also allow us to systematically compare the glaucoma risk and optic nerve head parameters across different ancestries.

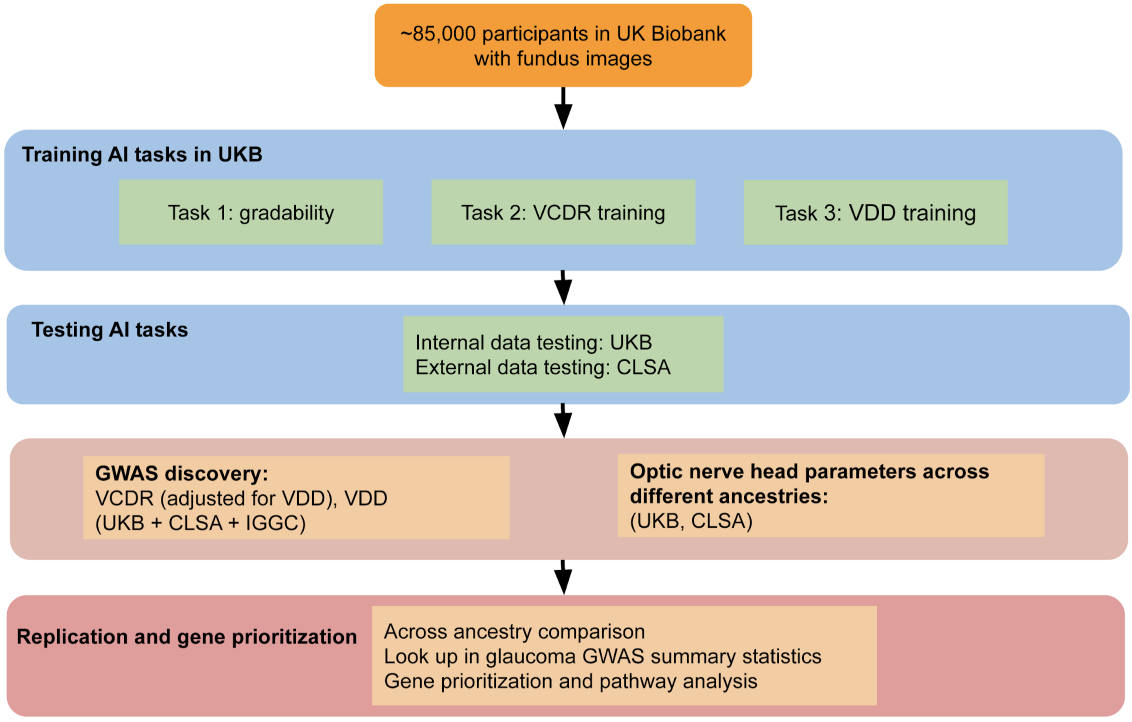


Figure 1. Flowchart of AI framework and datasets. In UK Biobank (UKB), the fundus retinal eye images were available for ~85,000 participants (~68,000 participants in the baseline visit and ~19,000 participants in the first repeat assessment visit). In our previous study, vertical cup-to-disc ratio (VCDR) and vertical disc diameter (VDD) were graded by two clinicians in ~70,000 photographs using a custom Java program. These clinical assessments were used as Training Data for three convolutional neural network (CNN) models for image gradability, VCDR, and VDD values. The learned models were then applied to all UKB fundus images (85,736 participants and 175,770 images in total) and another independent cohort - the Canadian Longitudinal Study on Aging (CLSA, 29,635 participants and 106,330 images in total). The AI labels were further used to systematically evaluate optic nerve head parameters across the multiple different ancestries in UKB and CLSA, and allowed us to perform the largest AI-based GWAS for VCDR and VDD.

4B.2.2 Study data and performance of the trained AI model

In the UKB, 85,736 participants had at least one fundus retinal image, with a total of 175,770 images available (Table 1). The mean age at baseline was 57.0 (SD: 8.1) years and 54% were women. In the CLSA cohort, 29,635 participants with 106,330 images were included

in the analysis, of whom 50% were women, and the mean age at recruitment was 62.6 (SD: 10.0) years.

We first trained a convoluted neural network to assess if each image was gradable in the UKB training sample. We found that most participants (> 95%) had gradable images in the UKB and the CLSA cohort (Supplementary Figure 3). We then predicted the measurements of both VCDR and VDD, and compared the AI-based measures with clinician gradings. The AI-based VCDR and VDD measurements exhibited a higher concordance to clinician gradings compared with previous gradings by two clinicians.[35,206,294,295] For instance, the Pearson's correlation coefficient of the VCDR measurements in the UKB samples was 0.81 (95% confidence interval [CI]: 0.80-0.81), and 0.84 (95% CI: 0.82-0.86) for an independent Canadian data set (CLSA) (Supplementary Figure 4). We therefore speculated that with the improved accuracy of VCDR and VDD measurements and the larger number of images graded, the optic nerve head assessment would increase the power for genetic discovery.

Table 1. Characteristics of retinal fundus images from the UK Biobank and Canadian Longitudinal Study on Aging participants.

Variable		UKB	CLSA
Number of images		175,770	106,330
Number of participants		85,736	29,635
% with at least one gradable image		95%	99%
Sex	Women (%)	44,017 (54%)	14,941 (51%)
Age at recruitment	Mean (SD), years	57 ± 8	63 ± 10
Vertical cup-disc-ratio	Unit in 0-1	0.37 ± 0.14	0.35 ± 0.15
Vertical disc diameter	Unit in pixel count	129.0 ± 10.5	121.4 ± 10.6

CLSA, Canadian Longitudinal Study on Aging cohort; SD, standard deviation; UKB, UK Biobank.

4B.2.3 Optic nerve head parameters and intraocular pressure across different ancestries

We compared AI model-derived VCDR and VDD measurements across different genetically-defined ancestry groups. VDD was similar across 3 ancestral groups (Europeans, East Asians and South Asians) and larger in Africans (Figure 2B, 2E). On average, after adjusting for age, sex, and VDD, VCDR was markedly higher in Asians and Africans than it was in Europeans (similar results in UKB Figure 2A and in CLSA Figure 2D). A different ancestry-based trend was also observed for intraocular pressure (IOP); relative to Europeans, South Asians had similar IOP, East Asians had lower IOP, and Africans had higher IOP (Figure 2C,F).

We then examined whether the systematically assessed VCDR, VDD and IOP can explain the observed prevalence of glaucoma seen across different ancestries in the UK and Canada. Figure 3 shows the glaucoma risk of Africans, East Asians and South Asians, with European ancestry (the most common ancestry in UKB and CLSA data sets) as the baseline. Consistent with previous epidemiological studies, Africans have the highest glaucoma risk (Figure 3 base model, correcting for only age and sex OR = 2.5 relative to the reference of Europeans). As seen in Figure 2, Africans have higher VCDR and higher IOP than Europeans and when these were corrected for, the glaucoma risk approached that of Europeans in both CLSA and UKB. East Asians had a similar base model risk to Europeans, although the contribution of IOP and VDR differs; on average their IOP is lower and their VCDR is larger (Figure 2), with the pattern of glaucoma risk changing as either IOP alone or VCDR alone were adjusted for in the regression model. Adjusting for both IOP and VCDR, the risk of glaucoma in East Asians was not significantly different to Europeans, suggesting that the higher VCDR and lower IOP in this group relative to Europeans counteract each other, explaining the similar glaucoma incidences between these ancestries. Interestingly, in South Asians, IOP is similar to Europeans, but VCDR is higher (Figure 2). This means that South Asian base model risk does not change when IOP is included in the model, but when VCDR is included the glaucoma risk decreases to become indistinguishable from the incidence in Europeans. In summary, by examining individuals of varying ancestry living in the UK and Canada, we show that relative to European ancestry, African ancestry glaucoma incidence is driven by both elevated VCDR and IOP, East Asian ancestry glaucoma is driven by elevated VCDR but ameliorated by lower IOP and finally that South Asian glaucoma is driven by elevated VCDR, but not by changes in IOP (relative to that in Europeans).

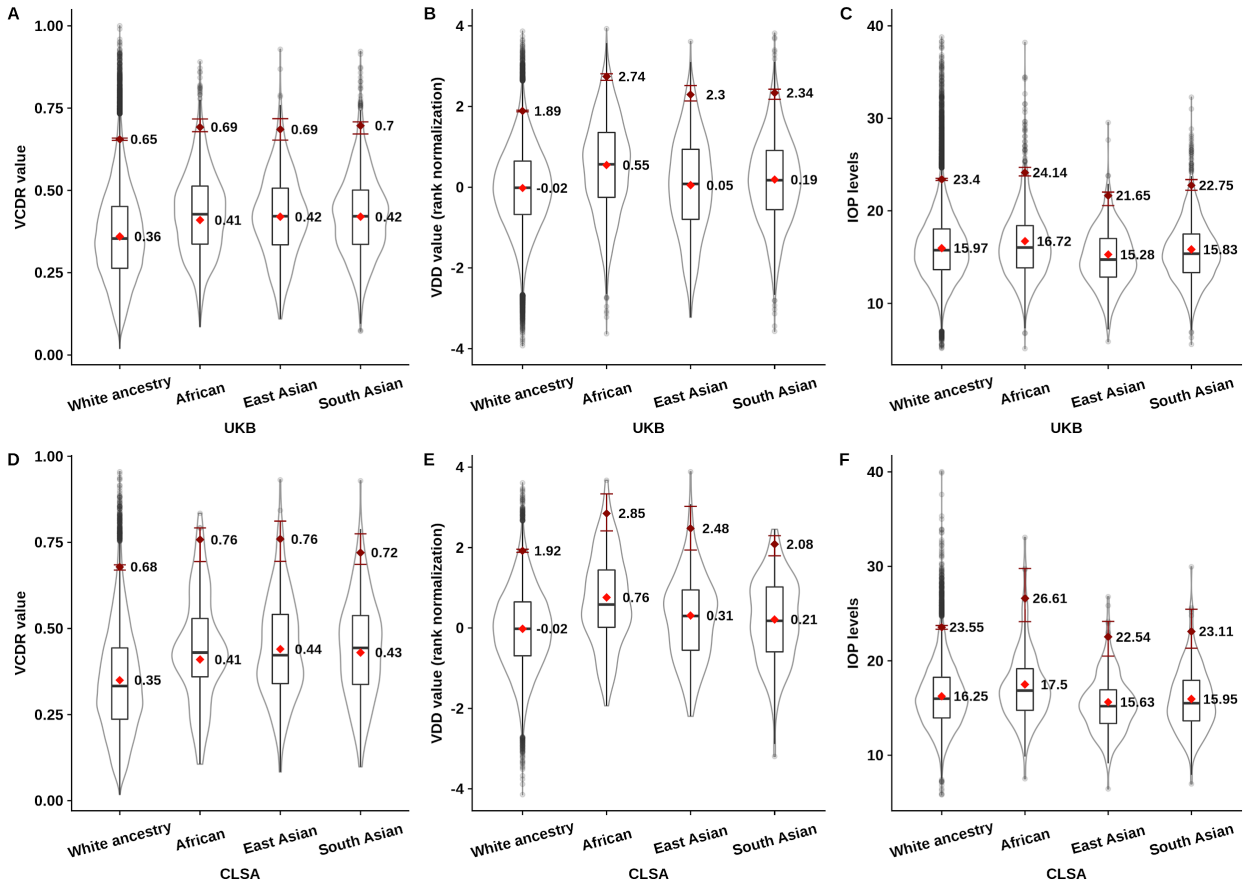


Figure 2. Optic nerve head measurements and intraocular pressure across different ancestry groups. Panel A shows the boxplot for VCDR values from different ancestry groups in UK Biobank. The box represents median value with first and third quartiles. The red diamond is the mean value of VCDR after accounting for age, sex, and VDD, where the mean value is annotated as text. The dark red diamond is the 97.5th percentile of VCDR value. The dark red error bar is the 95% confidence interval (2.5% to 97.5% quantiles) of the 97.5th percentile based on 1000 bootstrapped samples, which is essential for CLSA data, where the sample size for African, East Asian and South Asian was substantially smaller (N < 300). Panel B shows the boxplot for VDD values from different ancestry groups in UK Biobank. Due to the scale from fundus images, the VDD was rank normalized (mean = 0, SD = 1). The red diamond is the mean value of VDD after accounting for age and sex. Panel C shows the boxplot for IOP levels from different ancestry groups in the UK Biobank (truncated at 40 mm Hg, with 15 participants between 40 - 60 mm Hg). Panel D, E and F show the boxplots for VCDR, VDD and IOP in the CLSA cohort, respectively.

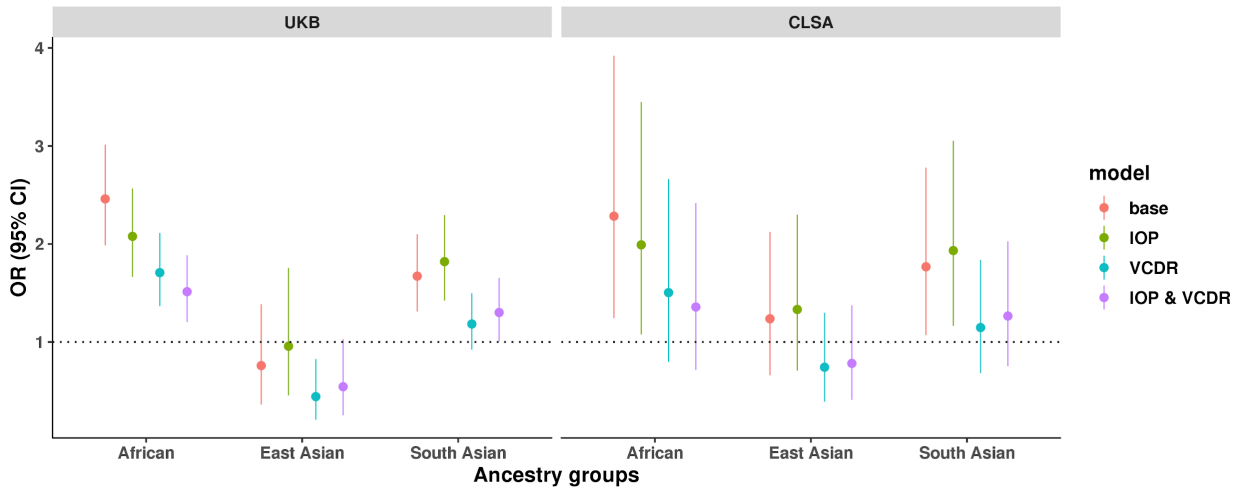


Figure 3. Glaucoma risk across different ancestry groups. The figure shows the risk of glaucoma in different ancestry groups. The horizontal line at OR = 1 is the reference for European ancestry. The Y-axis is the odds ratio (OR) and 95% confidence interval (CI) for three ethnic groups (African, South Asian, and East Asian). In each different model, different covariates were adjusted to evaluate the association of ethnic groups and glaucoma risk. In the base model, only sex and age were adjusted for; the other models also include either IOP, VCDR, or both (IOP & VCDR).

4B.2.4 AI-based phenotypes greatly increase SNP-based heritability and identify more loci

In the GWAS of VDD-adjusted VCDR, 145 and 19 statistically independent genome-wide significant SNPs were respectively identified in the UKB alone and CLSA alone (Supplementary Figure 5). The analogous numbers of SNPs for VDD were 142 and 17 for UKB and CLSA, respectively. We found weak evidence of genomic inflation from linkage disequilibrium score regression (Supplementary Table 3). From UKB, the AI-based GWAS of VDD-adjusted VCDR and VDD identified substantially more loci than our previous GWAS based on clinician gradings (76 for VDD-adjusted VCDR and 91 for VDD)[35,206]. Strikingly, the SNP-based heritability increased by ~50% for VCDR and VDD (Supplementary Figure 6). For instance, the SNP-based heritability for VCDR was 0.22 from clinician gradings (only single measure), whereas the heritability increased to 0.35 from AI-based GWAS (average of multiple measures). The increased heritability indicated that AI-based phenotyping was substantially cleaner than clinician gradings, which may be a result of two aspects: 1) higher accuracy of AI-based gradings; 2) improved accuracy from multiple measures per individual. We further tested the hypothesis in UKB and CLSA using only one measure per individual from AI-based gradings. The SNP-based heritability from a single measure (left or right eyes

in the baseline or first follow-up visit) was ~0.3, which is roughly in the middle of heritability estimation from clinician gradings and AI-based multiple measures (Supplementary Figure 6). These results indicate the higher accuracy of AI-based single measure per individual contributes to the increase of heritability estimation, and averaging of multiple measures per individual can further increase the heritability. Consistent with our previous study, correcting for VDD in VCDR GWAS also improved the relevance to glaucoma, with a higher genetic correlation with glaucoma in VDD-adjusted VCDR compared with unadjusted VCDR GWAS (genetic correlation rg = 0.502 vs 0.457 in UKB, and 0.543 vs 0.481 in CLSA).

4B.2.5 Validation AI-based GWAS

We then compared AI-based and clinician grading-based GWAS using independent samples from the IGGC. The concordance of SNP effect sizes of top SNPs between the AI-based and clinician gradings was essentially one (Panel A and D in Figure 4), and nearly all previously published loci using clinician ratings were replicated. The estimated effect sizes at the top SNPs from AI-based GWAS were also highly concordant between UKB and CLSA (Panel B and E in Figure 4). When combining UKB and CLSA AI-based GWAS we identified 193 and 188 loci for VDD-adjusted VCDR and VDD, respectively, again exhibiting very high concordance with IGGC (Panel C and F in Figure 4). The high concordance and more loci support the better-powered GWAS from AI-based measurements.

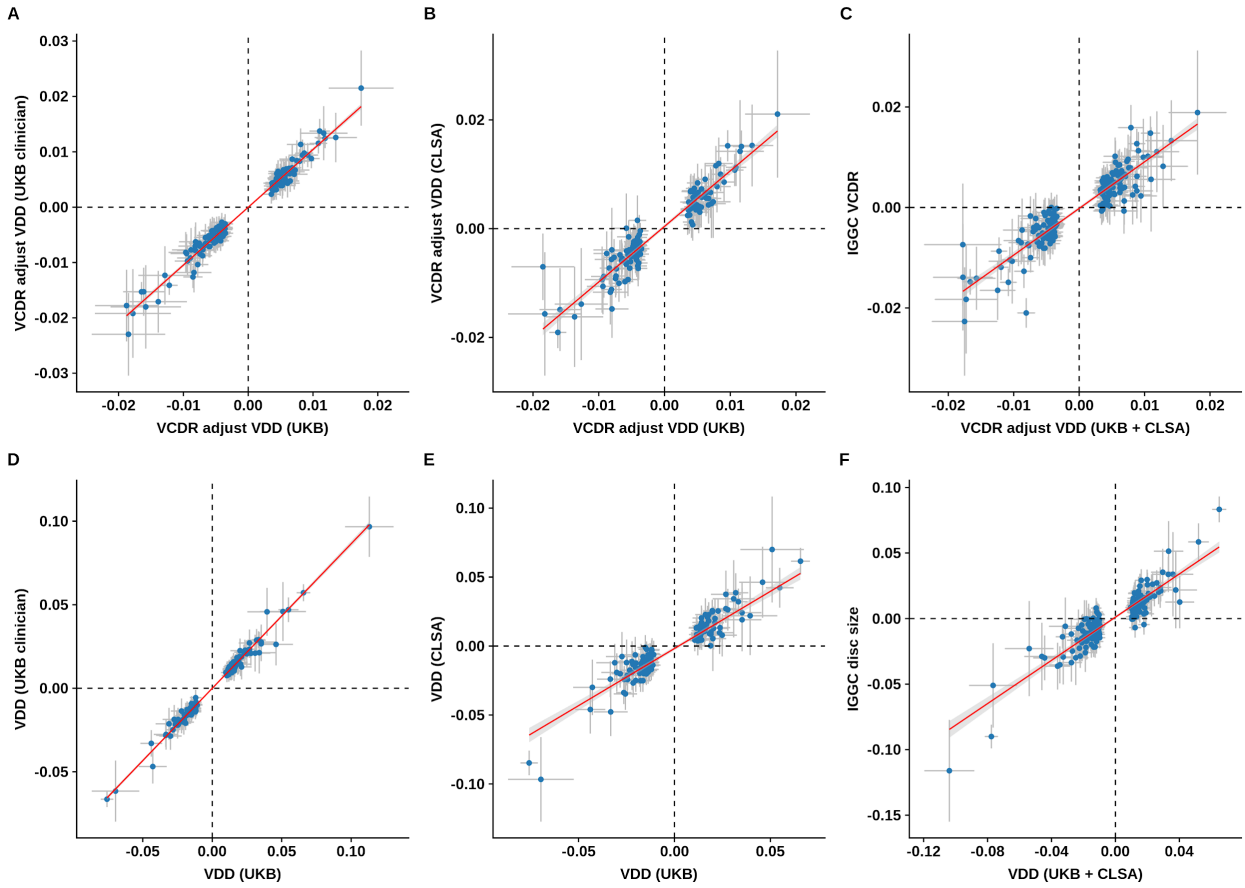


Figure 4. Validation AI-based GWAS. The figure shows the effect sizes for VDD-adjusted VCDR and VDD from different data sets. The vertical and horizontal error bars are the 95% confidence interval for SNP effect sizes. The red line is the best fit line with 95% confidence interval region in grey.

4B.2.6 New genetic discovery of optic nerve head measures, cross-ancestry comparison, and implications for glaucoma

To maximize power for locus discovery, we combined UKB, CLSA and IGGC GWAS (European ancestry), and identified 230 and 231 independent genome-wide significant SNPs for VDD-adjusted VCDR and VDD, respectively (Figure 5). Of them, we found 111 and 107 novel loci for VDD-adjusted VCDR and VDD, respectively (Supplementary Table 4 and 5). We then compared the effect sizes of top VDD-adjusted VCDR and VDD loci across different ancestries (Asian and African GWAS), due to the much smaller available sample sizes, their confidence intervals of effect estimations were very large, however the clear linear trend indicated the loci identified from European ancestry also had an effect on Asian populations (Figure 6A, B, for VCDR and VDD the Pearson's correlation coefficient is 0.65 [P value 3.6×10^{-27}] and 0.62 [P value 9.3×10^{-23}], respectively). The sample size of African

ancestry was much smaller than Asian ancestry (N = 2,245 versus 8,373 for VCDR) and showed a lower concordance (Supplementary Figure 7). The genetic correlations across the genome were essentially one based on the Popcorn approach for VCDR and VDD (Supplementary Table 6). We also compared the effect sizes of VDD-adjusted VCDR top loci with their effect sizes on glaucoma (Figure 6C), and found a relatively high concordance (Pearson's correlation coefficient 0.71, $P = 4.1 \times 10^{-37}$). Of the 230 VCDR (adjusted for VDD) loci, 205 (89%) were in the same direction, 131 were associated with glaucoma at a nominal significance level ($P<0.05$) and 68 were associated with glaucoma after Bonferroni correction ($P< 0.05/230= 2.2 \times 10^{-4}$, the nearest gene names are highlighted in Figure 6C, *e.g. RBPMS*, *AFAP1*, *LMX1B*, *ABCA1*, *CAV1*, and *GAS7*).

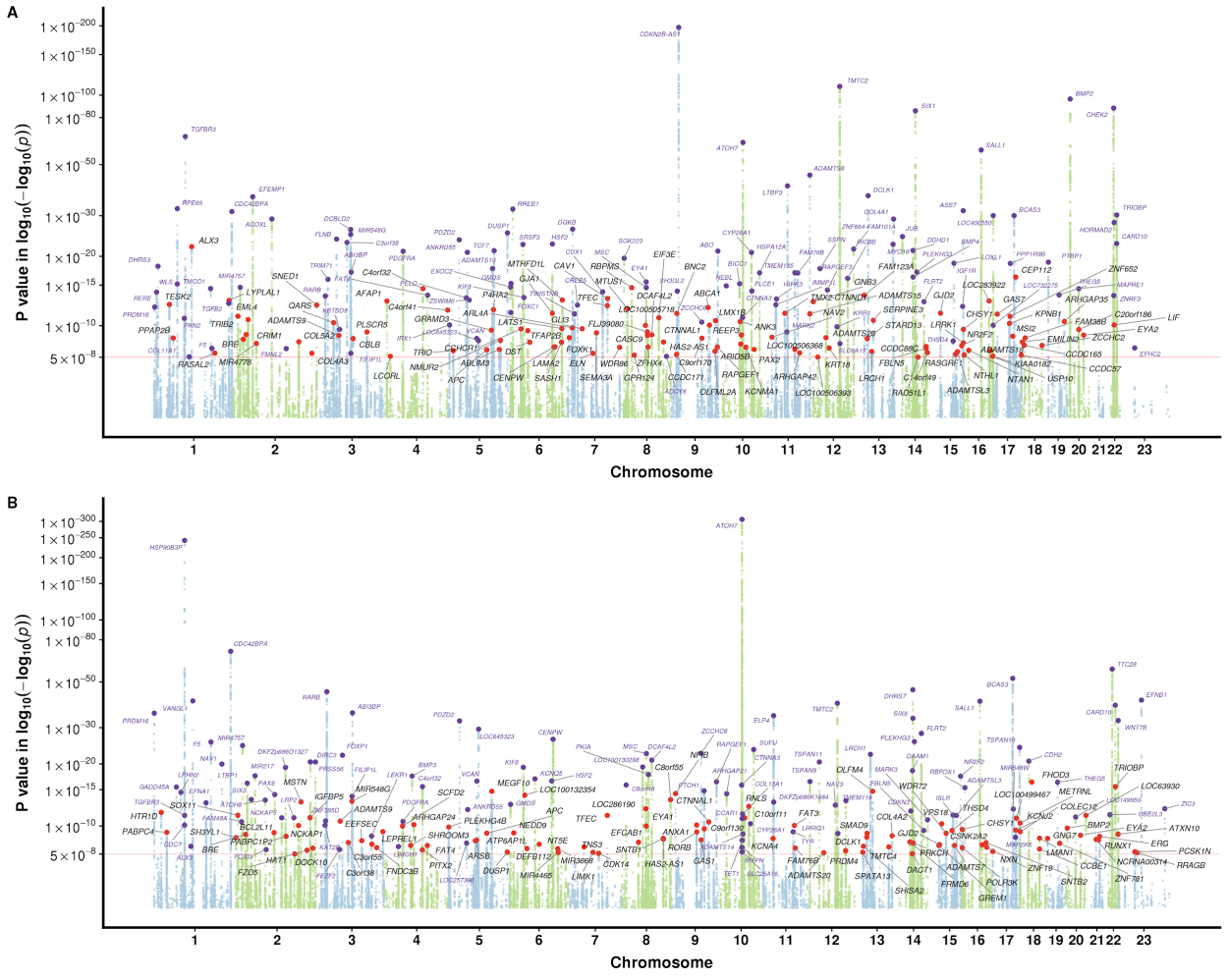


Figure 5. AI enables new genetic discovery for optic nerve head measures. Manhattan plot panel A shows P values for VDD-adjusted VCDR from the meta-analysis of UKB, CLSA, and IGGC (European ancestry). Panel B shows P values for VDD from the meta-analysis of UKB, CLSA, and IGGC

(European ancestry). The Y-axis is in log-log scale. The red horizontal line is the genome-wide significance level at P = 5×10^{-8}. SNPs with P value less than 1×10^{-4} are not shown in Manhattan plot. Previously unknown loci are highlighted with red dots, with the nearest gene names in black text. Known SNPs are highlighted with purple dots, with the nearest gene names in purple text.

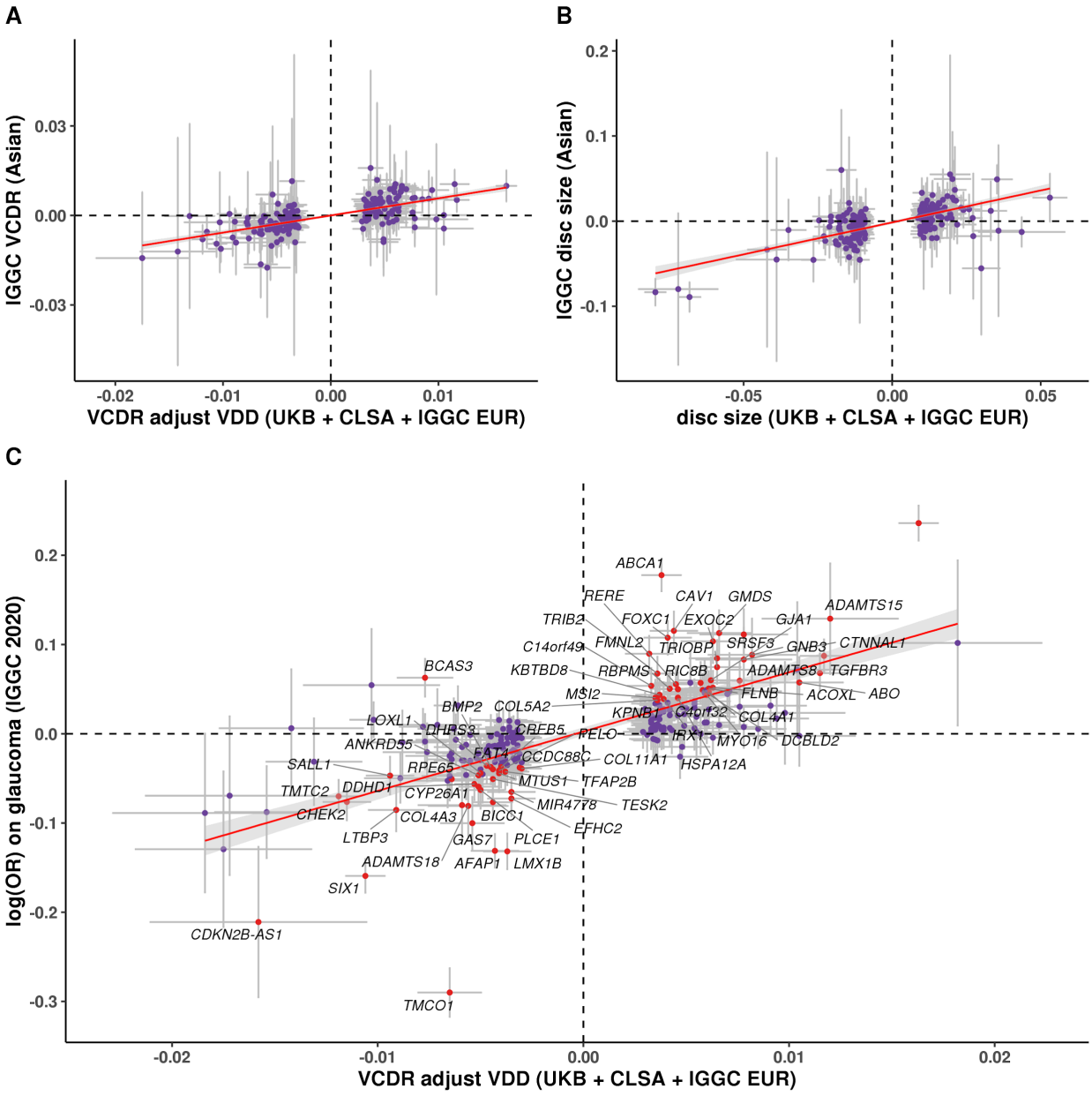


Figure 6. Comparison of the effect sizes for VCDR (adjusted for VDD) and VDD lead SNPs versus those observed in the Asian ancestry group and in independent glaucoma cohorts. Panel A and B show the effect sizes for lead VCDR (adjusted for VDD) and VDD loci (European versus Asian population). Panel C shows the effect sizes for VCDR (adjusted for VDD) lead SNPs versus log odds ratio in meta-analysis of UKB and ANZRAG glaucoma GWAS. The 24 SNPs associated with

glaucoma after Bonferroni correction ($P<0.05/227 = 2.2 \times 10^{-4}$) are highlighted with red dots, with the nearest gene names in black text.

4B.2.7 Gene prioritization and pathway analysis

We performed TWAS analysis in FUSION based on the VDD-adjusted VCDR and VDD GWAS summary statistics and retinal gene expression data. For VDD-adjusted VCDR we identified 101 genes that were significant after Bonferroni correction for multiple testing, nine of which were not genome-wide significant in the per-SNP analysis (Supplementary Figure 8A and 8B). For VDD we identified 64 genes that were significant after Bonferroni correction for multiple testing, 13 of which were not genome-wide significant in the per-SNP analysis. From SMR analysis, we identified 29 and 24 genes for VDD-adjusted VCDR and VDD, respectively, that were significant after multiple testing. We also compared the genes identified from both FUSION and SMR, 11 and 8 genes overlap from the two methods for VDD-adjusted VCDR and VDD, respectively (Supplementary Figure 8C and 8D). For instance, of the 11 genes that were associated with VDD-adjusted VCDR for the two approaches, 6 genes also passed the HEIDI tests (*P4HTM*, *SNX32, RASGRF*, *HAUS4*, *LRP11*, *AC012613.2*), suggesting the effects on VCDR may be mediated via these gene expression in retina tissue. The large increase in power resulting from the use of AI grading to improve accuracy and enable substantially larger datasets with multiple images per participant meant we were able to discover many new biological pathways influencing optic nerve head development and aging. Our pathway enrichment analysis uncovered 65 pathways for VCDR and 82 pathways for VDD after Bonferroni correction for multiple testing (Supplementary Table 7 and 8). As well as extracellular matrix pathways uncovered by our previous work, these new pathway analysis uncovered associations with telencephalon (forebrain) regionalization, embryo development, and anatomical tube development. There were several unexpected but statistically robust associations with kidney development (e.g. GO mesonephros development, $P_{raw} = 3.45 \times 10^{-8}$, P=0.00053 after correction for multiple comparisons). The genes driving the kidney development pathway enrichment included *BMP2*, *BMP4*, *EYA1*, *FAT4*, *FOXC1*, *GLI3*, *PAX2*, *RARB*, *SIX1*, and *SALL1*. Several kidney pathways were also significant in the pathway enrichment analysis applied to our VDD GWAS.

4B.3 Discussion

Our results show the promising application of AI algorithms in genetics studies. Large scale biobanks such as UKB and CLSA have accumulated hundreds of thousands of optic nerve images containing important information for glaucomatous optic neuropathy. However, the time-intensive and moderate agreement of manual assessment have impeded the usage of retinal fundus images. We trained a deep learning model using clinically estimated VCDR and VDD, and found the trained model has a high accuracy. The large scale biobank data for both VCDR and IOP allow us to systematically compare the glaucoma risk and optic nerve head parameters across different ancestries. Combining genetic and image data, we doubled the number of loci for both VCDR and VDD, with increased heritability.

The scope of available deep learning models for computer vision tasks is extensive and continuously developing. Various approaches to grade fundus images often utilise intricate data preprocessing methods[296–298] as well as computationally heavy models and training methods[292,299]. In the instance of statistically powered, large scale population study, fast inference and quick iterations are key, making heavy computational and design costs harder to justify. Here we demonstrate that a relatively lightweight, pretained CNN model is capable of producing highly accurate estimations of VCDR and VDD as evinced by high correlation with clinical grading, improved genetic discovery and further validations in independent samples.

Our AI approach has dramatically accelerated the pace of genetic discoveries. In our previous study, we laboriously manually assessed a subset of UKB images. With the deep learning model trained on clinical measurements, we were able to predict on a new subject within a fraction of a second, making time and effort of image labelling trivial, even when applied to large scale datasets (~1 hour for ~0.3 million images). Sample size is one of the most important limiting factors for genetic discovery. Leveraging the AI-based algorithm and large scale data, we were able to conduct the most powerful GWAS of optic nerve head parameters to date. We doubled the number of genome-wide significant loci for both VCDR and VDD. Interestingly, the estimated SNP-based heritability also increased by ~50% for VCDR and VDD (Supplementary Figure 6); the estimate for VCDR is not substantially lower than the heritability estimate from twin studies (~50%), although given more accurate (AI

based) phenotypes, the twin study based heritability estimate may increase. The increased heritability is a result of more accurate measurements, which arises in part due to the higher accuracy of AI-based predictions and in part to the AI approach allowing time-efficient grading of multiple measures per individual.

Many of the newly identified VCDR genes are associated with other eye traits (*e.g.* glaucoma, IOP, exfoliation syndrome, myopia). For some loci associated with IOP, it is likely that they have an effect on VCDR as a secondary effect of the locus first acting on IOP. Loci including genes such as *ABCA1*, *CAV1*, *AFAP1* and *LMX1B* were associated with VCDR for the first time; a likely explanation for this association is that the associated variant alters IOP and subsequently VCDR. Over 20 of the VCDR loci are also associated with refractive error, with multiple aspects of eye physiology likely involved (axial length, corneal thickness, retinal ganglion cell function). We also found a significant genome-wide genetic correlation between VCDR (adjusted for VDD) and myopia (rg = 0.3, $P = 1\times10^{-14}$), as well as with well studied traits which are associated with myopia such as years of education.[300]

In addition, several of the new VCDR genes provide possible links to retinal ganglion cell biology and they may constitute possible drug repositioning candidates. There are too many to discuss individually but one SNP of interest is rs17855988; this missense variant in the elastin gene (*ELN*) has been associated with facial ageing. Elastin in the sclera is most dense around the optic nerve head[301] and *ELN* expression has been shown to be high in exfoliation glaucoma lens[302]. A subset of the VDD loci have been found to be associated at genome-wide significance levels in previous glaucoma GWAS. However, in the majority of cases, the association with glaucoma appears to be driven by the lead SNP having a primary effect on VCDR (where the variance explained in VCDR for the peak SNP is larger than that for VDD: *e.g.* SNPs in or near *GMDS*, *CAV1*, *MYOF*, *SIX6*, *CHEK2*, *TMTC2*). Hence, the primary link between the disc parameters and glaucoma is via VCDR rather than via VDD. This is also shown in the lower genetic correlation between glaucoma and VDD (rg = 0.01) compared with glaucoma and VCDR (rg = 0.5).[35,206] With the high genetic correlation between VCDR and glaucoma, a multitrait analysis has recently shown that including VCDR can improve the power to identify glaucoma genes and to enable the development of polygenic risk score.[35] Future studies of glaucoma would benefit from incorporating these accurate AI derived VCDR estimates.

Previous studies have looked at the differences between VDD across different ancestries.[303,304] Our results were consistent with this, with Africans having the largest disc size, followed by those of Asian ancestry. For VCDR, an early study (100 black and 100 white) found that blacks had larger VCDR (mean values: blacks 0.35, white 0.24).[305] A subsequent larger study (1534 black and 1853 white) reported larger VCDR in blacks (mean values: blacks 0.56, whites 0.49).[306] A subsequent study in three different Asian ancestries, showed that VCDR values were similar between the studied ancestries (mean VCDR 0.40, 0.42 and 0.40, in Malay, Chinese, Indian, respectively).[307] It is striking that despite VCDR theoretically being a simple parameter to assess, the mean VCDR varies widely across studies, possibly due to differences in measurement protocol, sex, age and eye disease status. A further study[93] looked at the 97.5th percentile of VCDR instead of the mean and reported broadly similar values in the Netherlands (0.73), Bangladesh (0.7), Mongolia (0.70), Singapore (0.7), Tanzania (0.7). A major advantage of our study is that we use our AI derived gradings in two population-based cohort studies to systematically assess VCDR differences across ancestries in a consistent study design. By leveraging large sample sizes, we are able to clearly show both Asian and African ancestry individuals have larger VCDR values than Europeans. Our primary results in Figure 2 correct VCDR for VDD, given previous studies showing that correcting for VDD enhances the relevance to glaucoma.[223]

The raised VCDR in Asian and African ancestry individuals living in the UK and Canada is in keeping with elevated glaucoma rates in these ancestries.[94] When combined with data on IOP, a combination of VCDR and IOP explains the vast majority of the variation between glaucoma rates in Europeans relative to Africans, South Asians and East Asians. Although crucially, our data show (Figure 3) that the relative contributions of VCDR and IOP are clearly different between all 4 major populations groups that we consider. For individuals of European, South Asian or African ancestry, the vast majority of broadly defined glaucoma cases are open angle glaucoma (OAG). In East Asia, angle closure glaucoma (ACG) is common and a limitation of our analysis is that we cannot distinguish between ACG and OAG in all cases - where available we have removed known cases of ACG in the broad glaucoma definition, but some ACG cases will remain.

A strength of our study is that a large number of clinically assessed images were used to train the deep learning model for VCDR and VDD; this allowed us to generate accurate predictions. Our study has shown that the AI-based measurements have a high accuracy.

The AI-based optic nerve head assessment has also boosted the available sample size and dramatically expanded gene discovery for these key ocular phenotypes. We show that this deep learning model can also be used to assess future fundus images automatically and rapidly, especially in population-based studies with a large number of images. Moreover, the implementation of transfer-learning techniques shows that AI-aided labelling, with adequate sample size, has a potential to deliver equally successful genetic discoveries in other image based biological phenotypes. Our study has several limitations. Firstly, although our AI approach was able to grade a large proportion of images (particularly in the CLSA study), a small proportion remained ungradable due to poor picture quality. Future studies could explore adversarial architectures to improve clinical ratings of VCDR and VDD. However, a set of high quality truth labels would still be necessary for initial pre-training. Finally, although we were able to use genetic data to clearly identify the major ancestries within UKB and CLSA (European, African, South Asian, East Asian), there remained a group of uncategorized individuals with mixed ancestries that we did not include in our epidemiological or genetic analyses.

To conclude, we showed that AI-based optic nerve head assessment has a high accuracy and this greatly improves our power to discover new genes. These findings provide new insights into the pathogenesis of glaucomatous optic neuropathy. We also use the systematic assessment of VCDR across different ancestries to help explain how the pattern of IOP and VCDR measures underpin observed glaucoma risk; such findings in mixed ancestry groups living in the UK and Canada help explain the differing characteristics of glaucoma across ancestries. For example, relative to Europeans, individuals with East Asian ancestry are more likely to have lower IOP and increased VCDR. Given these East Asians are genetically similar to East Asians in countries such as China and Japan, this provides support for the assertion that normal tension forms of glaucoma predominate in East Asia due to genetic predisposition for high VCDR, despite low IOP.

4B.4 Methods

4B.4.1 Study populations

UK Biobank

The UK Biobank is a population-based cohort study with deep genetic and phenotypic data from ~500,000 participants aged between 40 to 69 years at the time of recruitment (2006-2010), living in the United Kingdom.[130] Retinal fundus images were available for both left and right eyes from two assessment visits, covering ~85,000 participants (~68,000 participants in the baseline visit and ~19,000 participants in the first repeat assessment visit [2012-2013]). In our previous study, vertical cup-to-disc ratio (VCDR) and vertical disc diameter (VDD) were graded by two clinicians using a custom Java program.[35] Detailed image processing and quality control methods were described previously.[35] Briefly, given the time-consuming nature of manual grading, we only graded the left eye images (if the left eye images were ungradable, the right eye images were used instead) and one visit (if the second visit measurements were unavailable, the first visit measurements were used instead) of white British ancestry participants. A total of 67,040 participants with both VCDR and VDD measurements were included in our previous GWAS. In this study, we used a CNN model to grade left and right eye images from two visits for all participants, irrespective of ancestry, with a total of 175,770 images.

In the UKB, ~488,000 participants were genotyped for 805,426 variants on Axiom arrays (Affymetrix Santa Clara, USA). The genetic data, quality control procedures and imputation methods have been described previously.[130] Briefly, ~96 million variants were imputed using Haplotype Reference Consortium (HRC) and UK10K haplotype resources[270,271,308], and 487,409 individuals passed genotyping quality control. Of them, 438,870 individuals were genetically similar to those of white-British ancestry.[126,130] For the GWAS in UKB, we retained SNPs with MAF > 0.01 and imputation quality score > 0.8. To verify self-reported diverse ancestry information (data field 21000 in UKB), we used a K-means clustering method based on genetic principal components (PCs). The genetic clusters were compared with self-reported ancestry. Participants within the same self-reported ancestry groups were largely in the same genetic clusters (*e.g.* African [N=9791], South Asian [N=2594], and East Asian [N=9941], detailed in Supplementary Figure 1), and on average ~20% of them have fundus retinal images.

The Canadian Longitudinal Study on Aging

The Canadian Longitudinal Study on Aging (CLSA) is a national, longitudinal cohort study of 51,338 participants from 10 Canadian provinces, aged 45 to 85 years at enrollment.[309,310] Recruitment and baseline data collection were completed in 2015, with participants followed-up every 3 years, and an initial follow-up visit completed in 2018. In this study the nerve

head photographs are available for a subset cohort "Comprehensive cohort" of 30,097 participants (for both left and right eyes, and the baseline and first follow-up visit). Retinal fundus imaging was performed using a Topcon (TRC-NW8) non-mydriatic retinal camera, with images saved in jpg format. A random sample of 1000 images was graded by a clinician for both VCDR and VDD using a custom Java program. The latest genome-wide genotype data (August 2019 release) are available for 19,669 participants of the Comprehensive cohort, comprising 794,409 genetic variants genotyped on the Affymetrix Axiom array, and ~40 million genetic variants imputed using the Haplotype Reference Consortium.[270] Variant- and sample- based quality control procedures were consistent with standards of the UK Biobank[130] with detailed steps presented in the CLSA support document (available at https://www.clsa-elcv.ca/researchers/data-support-documentation). For the GWAS analysis, we included 18,304 participants of European ancestry based on the K-means cluster method on genetic principal components, and the largest cluster also contains the majority of individuals that self-report European ancestry. SNPs with MAF > 0.01 and imputation quality score > 0.8 were retained in association analysis. From the K-means clustering method, the sample size for African South Asian, and East Asian is 135, 219, and 217, respectively (PC plot was shown in Supplementary Figure 2).

The International Glaucoma Genetic Consortium

The International Glaucoma Genetic Consortium (IGGC) is one of the largest international consortia established to identify glaucoma genetic risk variants through large-scale meta-analysis. The phenotype and genotype data of VCDR and optic disc area for IGGC have been previously described elsewhere.[133,311] It should be noted the optic disc area is not in the same scale as VDD from the AI gradings. When comparing and meta-analyzing the VDD and disc area data, we applied a rank-based inverse normal transformation to AI gradings and rendered them back to disc area scale, as detailed in our previous study.[206] Publicly available summary statistics were downloaded for individuals of European descent (N_{VCDR}= 25,180, N_{disc}= 24,509, from the latest HRC imputation), as well as Asian descent (N_{VCDR}= 8,373, N_{disc}= 7,307).[133,311]

Glaucoma GWAS dataset

The glaucoma datasets were described in our previous study, including 34,179 primary open-angle glaucoma cases and 349,321 controls from a large-scale multi-ethnic meta-

analysis (Gharahkhani, et al. 2020, in press)[312]. The detailed information of phenotype definition and genetic association analyses were presented in detail previously.[312] The GWAS summary statistics were used to look up each of the VCDR loci (adjusted for VDD), and replicate their effects on glaucoma.

4B.4.2 AI algorithm on retinal images

Three separate CNN models were used to make inferences about image gradability, VCDR, and VDD values of retinal fundus images in UKB. The image gradability (gradable or ungradable) was defined as a binary classification, while the latter two tasks were modelled as regression problems. Images with a higher likelihood of gradability (i.e. designated softmax probability more than 0.5) were assigned as gradable. While a variety of CNN model architectures were tested, the final architecture used for all CNN models was ResNet-34.[313] Pre-trained weights, initially trained on ImageNet[314] classification tasks, were utilised for each model as a form of transfer learning. Untrained layers specific to each model were additionally added, forming a custom regression (Relu) and classification (softmax) heads for each respective task. All fundus images were cropped and scaled to a pixel ratio of (1080, 800) before training or validation. We used the highest native resolution for the UKB training images as we found that using lower resolution negatively impacted inference metrics. The total dataset sizes used for the VCDR, VDD and gradability tasks were 71,950, 50,984, and 75,718, respectively. Each dataset was randomly split into 80% training and 20% validation. The model performance was validated by sample hold out, with final testing performed on images from the CLSA dataset. Model requirements for regression tasks were defined achieving a validation loss equal or lower than human inter-rater loss. The gradebillity task criteria was defined as accuracy above 95%. Both regression tasks utilised mean square error loss function, while the classification model optimised over the binary cross entropy loss function. Training of all models was completed using the FastAI framework[315], while utilising the in-built data augmentations functionality to improve accuracy and generalisability. The specifics of which augmentations were used can be found in Supplementary Table 1. It should be noted that the regression task for VDD was dependent on image scale, as such, augmentations which introduced scaling were omitted. Training was carried out in two stages: the first involved freezing the pretrained weights and only training the task head; the second, the ‘fine tuning’ stage, all model weights were unfrozen.

Each stage was trained with cyclical training rate as described elsewhere[316], and performed until the validation loss reached a plateau.

4B.4.3 Optic nerve head parameters, intraocular pressure and glaucoma risk across different ancestries

Previous studies have reported differences in VCDR and VDD values across different ancestry groups.[303,304] Taking advantage of the diverse ancestries available in UKB and CLSA, we compared our AI derived VCDR and VDD values, as well as intraocular pressure (IOP, corneal-compensated[126]) values across different ancestry groups. We used the K-means clustering method to define ancestry groups based on genetic data (detailed above). Boxplots were used to show the differences of optic nerve head measurements across different ancestry groups (*e.g.* median value, upper and lower quartiles). The mean values of VCDR across different ancestries were estimated after adjusting for age, sex, and VDD. The 97.5th percentile of optic nerve head measurements and its 95% confidence interval (2.5% to 97.5% quantiles) were also calculated based on 1,000 bootstrapped samples, on account of the substantially smaller sample size for individuals of African, East Asian and South Asian ancestry. We then investigated how VCDR and IOP relate to glaucoma risk in different ancestries. The definition of glaucoma cases and controls was detailed in our previous study.[35] Briefly, in UKB glaucoma cases were ascertained from International Classification of Diseases diagnosis, record-linkage data from general practitioners, and self-reported previous diagnosis. In the CLSA, participants were interviewed in-person with the question “Has a doctor ever told you that you have glaucoma?”. Logistic regression models were used to evaluate the association between genetically-defined ancestry groups and glaucoma risk. In each different model, different covariates were adjusted to evaluate the association of ethnic groups and glaucoma risk. In the base model, only sex and age were adjusted for; the other models also include either IOP, VCDR, or both (IOP & VCDR).

4B.4.4 Genome-wide association analysis and meta-analysis

For both UKB and CLSA, the VCDR and VDD GWAS association tests were carried out using a linear mixed model (using BOLT-LMM version 2.3)[210] to account for cryptic relatedness and population stratification, adjusting for sex and age. The first ten principal components were also included in the model to speed up the convergence of computations.[317] The average values of measurements from left and right eyes and multiple

visits (if available) were used, and were first transformed using a rank-based inverse-normal method before association tests.[267] To account for optic disc size covariation, VCDR grading was adjusted for VDD in GWAS analyses.[205,206] The VDD-adjusted VCDR and VDD GWAS results from UKB and CLSA were then meta-analysed with those from the IGGC based on the inverse variance-weighted method (METAL software 2011-03-25 release).[211] We also conducted association tests for VCDR and VDD in African and South Asian populations in UKB. Due to the relatively small size of each of these populations (Supplementary Table 2, less than the recommended sample size of 5000 in BOLT-LMM), PLINK was used instead, after removing related individuals.[182]

SNP-based heritability was calculated by LD score regression (LDSC) from GWAS summary statistics.[73,318] Bivariate LD score regression was used to estimate the genetic correlation between pairs of traits in European ancestry.[73] We selected independent SNPs based on the PLINK clumping method with P value < 5×10^{-8}, $r^2 < 0.01$, and a window of 1Mb from the index variant.[182] To define novel loci from the AI-based GWAS, we checked previous UKB VCDR and VDD GWAS based on clinician gradings[35,206], we also looked up the proxy SNPs ($r^2 > 0.8$) of top loci and their nearest genes in GWAS Catalog.[9]

4B.4.5 Cross population genetic effects on optic nerve head parameters

We evaluated the effects of genetics variants on VCDR and VDD cross different populations based on the following methods: 1) we first compared and replicated the AI-based top loci from European ancestry with the GWAS from African and South Asian samples. The effect sizes and standard errors of top loci were shown in a scatter plot for different ancestries; 2) we calculated the trans-ethnic genetic effect correlation for VCDR and VDD using the "Popcorn" package.[319] Specifically, the GWAS summary statistics for VCDR and VDD from European ancestry were compared with that in Asian and African ancestry.

4B.4.6 Transcriptome-wide association study and pathway analysis

To prioritize potential causal genes, transcriptome-wide association study analysis (TWAS) was performed in FUSION using GWAS summary statistics and retina gene expression data.[320] In FUSION, a reference data with both gene expression and genetic variants (SNPs) were used to train predictive models, which were used to impute the expression-trait association directly from large-scale GWAS summary statistics.[320] The weights of retina gene expression were obtained from 406 individuals from Eye Genotype Expression

database (EyeGEx).[170,320] We also used the EyeGEx to perform a summary data-based Mendelian randomization (SMR) to investigate the association of gene expression levels (exposure) and phenotype (outcome).[243] The heterogeneity in dependent instruments (HEIDI) tests were used to evaluate the null hypothesis that a single causal variant affecting both gene expression and outcome, and the significance threshold was set at 0.05 ($P_{HEIDI} \geq 0.05$ not reject the null hypothesis).[243] Pathway analysis were conducted in MAGMA as implemented in FUMA (version 1.3.6).[213,214] All other analyses were performed with R software.[216]

4B.5 Acknowledgements

This work was conducted using the UK Biobank Resource (application number 25331) and publicly available data from the International Glaucoma Genetics Consortium. The UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council (UK), Department of Health (UK), Scottish Government, and Northwest Regional Development Agency. It also had funding from the Welsh Assembly Government, British Heart Foundation, and Diabetes UK. The eye and vision dataset has been developed with additional funding from The NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology, Fight for Sight charity (UK), Moorfields Eye Charity (UK), The Macula Society (UK), The International Glaucoma Association (UK) and Alcon Research Institute (USA). This work was also supported by grants from the National Health and Medical Research Council (NHMRC) of Australia (#1107098; 1116360, 1116495, 1023911), the Ophthalmic Research Institute of Australia, the BrightFocus Foundation, UK and Eire Glaucoma Society and Charitable Funds from Royal Liverpool University Hospital. SM, JEC, and AWH are supported by NHMRC Fellowships. XH is supported by the University of Queensland Research Training Scholarship and QIMR Berghofer PhD Top Up Scholarship. We thank Scott Wood, John Pearson and Scott Gordon from QIMR Berghofer for support. The NEIGHBORHOOD consortium is supported by NIH grants P30 EY014104, R01 EY015473 and R01 EY022305. AI engineering was performed at and funded by Max Kelsen.

This research was made possible using the data/biospecimens collected by the Canadian Longitudinal Study on Aging (CLSA). Funding for the Canadian Longitudinal Study on Aging (CLSA) is provided by the Government of Canada through the Canadian Institutes of Health Research (CIHR) under grant reference: LSA 94473 and the Canada Foundation for Innovation. This research has been conducted using the CLSA dataset [Baseline Comprehensive Dataset version 4.0, Follow-up 1 Comprehensive Dataset version 1.0], under Application Number 190225. The CLSA is led by Drs. Parminder Raina, Christina Wolfson and Susan Kirkland.

Data availability

UK Biobank data are available through the UK Biobank Access Management System https://www.ukbiobank.ac.uk/. We will return the derived data fields following the UK biobank policy and in due course they will be available through the UK Biobank Access Management System.

Data are available from the Canadian Longitudinal Study on Aging (www.clsa-elcv.ca) for researchers who meet the criteria for access to de-identified CLSA data.

Disclaimer

The opinions expressed in this manuscript are the author's own and do not reflect the views of the Canadian Longitudinal Study on Aging.

Competing interests

K.S., C.B., M.T. and M.T. are employees of Max Kelsen.

4B.6 Supplement

Supplementary Figures and Tables are available at:
https://www.cell.com/ajhg/fulltext/S0002-9297(21)00189-0

CHAPTER 5

Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration

Xikun Han, Puya Gharahkhani, Paul Mitchell, Gerald Liew, Alex W. Hewitt, Stuart MacGregor.

Journal of Human Genetics. 2020; 65, 657–665.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Stuart MacGregor, Alex W Hewitt, and Puya Gharahkhani obtained funding and designed the study. All authors contributed to interpretation of the results and the final version of the paper.

Chapter 5. Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration

Age-related macular degeneration (AMD) is the leading cause of central vision loss among the elderly population in the Western world. To accelerate the understanding of the genetics of AMD, we conducted a meta-analysis of genome-wide association studies (GWAS) combining data from the International AMD Genomics Consortium AMD-2016 GWAS (16,144 advanced AMD cases and 17,832 controls), AMD-2013 GWAS (17,181 cases and 60,074 controls), and new data on 4017 AMD cases and 14,984 controls from Genetic Epidemiology Research on Aging study. We identified 12 novel AMD loci near or within *C4BPA—CD55*, *ZNF385B*, *ZBTB38*, *NFKB1*, *LINC00461*, *ADAM19*, *CPN1*, *ACSL5*, *CSK*, *RLBP1*, *CLUL1*, and *LBP*. We then replicated the associations of the novel loci in independent cohorts, UK Biobank (5860 cases and 126,726 controls) and FinnGen (1266 cases and 47,560 control). In general, the concordance in effect sizes was very high (correlation in effect size estimates 0.89), 11 of 12 novel loci were in the expected direction, 5 were associated with AMD at a nominal significance level, and rs3825991 (near gene *RLBP1*) after Bonferroni correction. We identified an additional 21 novel genes using a gene-based test. Most of the novel genes are expressed in retinal tissue and could be involved in the pathogenesis of AMD (i.e., complement, inflammation, and lipid pathways). These findings enhance our understanding of the genetic architecture of AMD and shed light on the biological process underlying AMD pathogenesis.

5.1 Introduction

Age-related macular degeneration (AMD), a degenerative disorder of the central retina, is the leading cause of central vision loss in the elderly population in the Western world.[134,321–323] AMD is classified as non-neovascular (dry AMD) and neovascular (wet) AMD. For the population aged over 45 years, the global prevalence of AMD is 8.69%, with a higher prevalence in Europeans (12.3%).[134] It is estimated that the number of AMD patients is 196 million in 2020, rising to 288 million in 2040.[134]

AMD is highly heritable with heritability estimates between 46% and 71%.[324] A recent genome-wide association study (GWAS) from the International AMD Genomics Consortium (IAMDGC) has identified 52 independent variants across 34 loci.[166] Understanding the genetic contributions for AMD is important to reveal insights into the biological mechanisms of AMD, and discover potential genetic variations for clinical diagnostic, predictive, and therapeutic targets.[138,166]

Recent statistical methodology and application studies have shown that multivariate GWAS can leverage multiple input summary statistics of the same trait or genetically correlated traits, and gain the power for identifying new genes.[53,325] Compared with the traditional meta-analysis that assumes the input GWAS summary statistics are from the same trait (a genetic correlation close to one) and is sensitive to sample overlap,[211] the multiple trait analysis of GWAS (MTAG) approach, a framework to generalize the standard inverse-variance meta-analysis method, can jointly analyse GWAS summary statistics from the same trait or multiple correlated phenotypes, with or without overlapping samples.[53] In this study, we identify novel AMD loci using the state-of-the-art multivariate GWAS method to combine several large AMD GWAS datasets.

5.2 Methods

5.2.1 Study overview

Our study design is displayed in Figure 1. We conducted a meta-analysis of GWASs based on the MTAG approach,[53] which generalizes the standard inverse-variance meta-analysis method to jointly analyse GWAS summary statistics with overlapping samples. We applied

MTAG to three input summary statistics: AMD-2016 GWAS[166] and AMD-2013 GWAS[165] from the IAMDGC, and AMD GWAS in Genetic Epidemiology Research on Aging (GERA) study.[326] We then replicated the novel AMD loci in independent datasets from the UKB and FinnGen Studies.

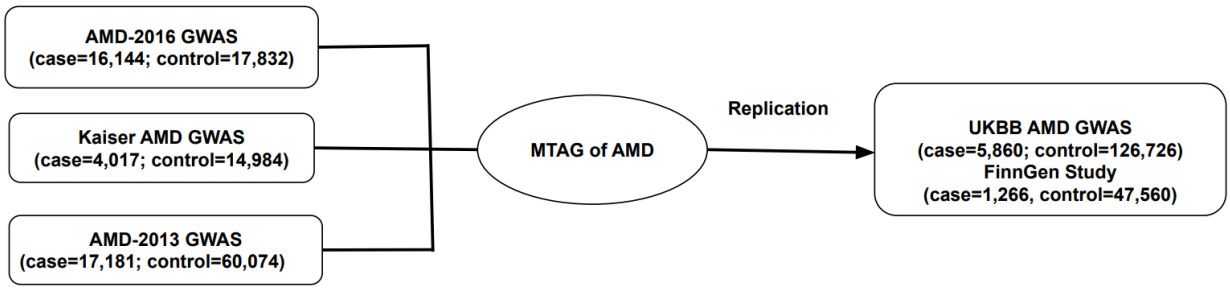


Figure 1. Study Design.

The multivariate analysis of GWAS (MTAG), a method to jointly analyze summary statistics, was applied to three input summary statistics of AMD GWAS: AMD-2016 GWAS[166] and AMD-2013 GWAS[165] from the International AMD Genomics Consortium (IAMDGC), and AMD GWAS in Genetic Epidemiology Research on Aging (GERA) study. The novel loci were then replicated in the UK Biobank and FinnGen studies.

5.2.2 International Age-Related Macular Degeneration Genomics Consortium

GWAS summary statistics: AMD-2016 GWAS and AMD-2013 GWAS

We downloaded two publicly available AMD summary statistics from the IAMDGC: AMD-2016 GWAS and AMD-2013 GWAS (web resources in the supplement).[165,166] In the AMD-2016 GWAS, there are 16,144 cases and 17,832 controls of European descent with P values and directions available in the summary statistics. We used the same method as mentioned in Burgess and colleagues' study to derive the beta-coefficients and standard errors (SEs) for all SNPs.[147] Briefly, the P values and directions of associations from the summary statistics were used to calculate z-scores. With the assumption that SE multiplied by sqrt(MAF × (1-MAF)) should be a constant, where "MAF" is the minor allele frequency, we estimated the constant using the average estimations from 34 genome-wide significant variants from Fritsche and colleagues' study.[147,166] The constant was further used to

calculate SEs and beta-coefficients for other variants. The validity of the method was also fully assessed in Burgess and colleagues' study.[147]

For the AMD-2013 GWAS, the GWAS summary statistics from 17,181 AMD cases and 60,074 controls were used. We used the same method above to calculate the beta-coefficients and SEs for all variants.

5.2.3 Genetic Epidemiology Research on Aging (GERA) study

The Genetic Epidemiology Research on Aging (GERA) cohort is a substudy of the longitudinal cohort enrolled in the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH). The detailed description of the study design can be found in the database of Genotypes and Phenotypes (dbGaP, study accession: phs000674.v1.p1).[326,327] In our authorized access data, 78,486 participants have both phenotype and genotype data. We only included self-reported whites for the following analysis.

We performed genotype quality control using PLINK software (version 1.90 beta).[182] For samples, we removed individuals with >3% missing genotypes. For markers, SNPs with call rate <95%, MAF <0.01, and Hardy-Weinberg equilibrium (HWE) P $<1\times10^{-6}$ were discarded. For relatives, we calculated identity by descent using autosomal SNPs and only kept one of any pair of individuals with pi-hat >0.2 for analysis. Michigan Imputation Server was used for imputation (parameters: HRC reference panel, version r1.1 2016; phasing, ShapeIT; population, EUR).[328] SNPs with imputation quality score >0.3 and MAF >0.01 were retained for association analysis.

Macular degeneration cases were recorded in the electronic health record (EHR) system as International Classification of Diseases, Ninth Revision (ICD-9) diagnosis codes (362.5, 362.50, 362.51, 362.52, and 362.57). Finally, we reported GWAS on 4,017 macular degeneration cases and 14,984 controls from the GERA cohort.

5.2.4 Replication in UK Biobank and FinnGen Study datasets

The UK Biobank project is a large-scale prospective cohort study of half a million participants across the United Kingdom, aged between 40 and 69 at the time of recruitment (2006-2010).[329] In our analysis, we only included participants with written consent and of white-British ancestry based on self-reported ethnicity and genetic principal components.[126,329] To control relatedness between samples, we used a pruning method in PLINK software (version

1.90 beta) to keep one of any pair of individuals with pi-hat >0.2. We identified 5,860 AMD cases using the following criteria: 1) ICD-9 or ICD-10 diagnosis codes (3625 and H353); 2) responded “Macular degeneration” in “eye problems/disorders” (Field 6148); 3) responded “macular degeneration” in self-reported non-cancer illness (Field 20002). We selected 126,726 “healthy” controls who did not have serious eye diseases (Field 6148). The UKB genotype data and quality control procedures were described previously.[329] In our association analysis, we only included SNPs with MAF >0.01 and imputation quality score >0.3.

The FinnGen study (https://www.finngen.fi/en) is a nation-wide study launched in Finland in 2017. The FinnGen study combines both genetic information and health care data to improve personalised healthcare. We downloaded the available summary statistics from the public release of FinnGen data freeze 2 results for 1,266 AMD cases (wet or dry macular degeneration) and 47,560 controls. The UKB and FinnGen AMD results were meta-analysed as the replication sample.

5.2.5 The Blue Mountains Eye Study

The Blue Mountains Eye Study (BMES) is a population-based cohort study investigating the etiology of common ocular diseases among suburban residents aged 49 years or older, in Australia.[323] The full description of the study design, phenotype definition, and genetic data were described previously.[166,323,330,331] In brief, retinal photographs were assessed for AMD lesions following the Wisconsin Age-Related Maculopathy Grading System for late AMD cases.[332] The late AMD cases were defined as presence of neovascular AMD or pure geographic atrophy. The controls were defined as no soft (distinct or indistinct) or intermediate drusen, any retinal pigment abnormalities (either depigmentation or increased pigment), and no signs of early or late AMD. DNA samples were obtained during the 5-year follow-up and ancillary surveys, which were performed between 1997 and 2000. Participants were genotyped with Human610-Quad arrays (Illumina, San Diego, CA, USA). Genotype data were imputed in Michigan Imputation Server. We included 100 late AMD cases and 2,136 controls of European descent with genetic information in our analysis.

5.2.6 Statistical analysis

For both of GERA and UKB AMD GWASs, we conducted logistic regression models under an additive genetic model adjusting for sex, age, and the first ten genetic principal

components in PLINK software (version 2.0).[182] Bivariate LD score regression was used to estimate the genetic correlation between pairs of AMD datasets.[73] We then used the MTAG software (version 1.0.8) to meta-analyse the GWAS summary statistics from AMD-2016 GWAS, AMD-2013 GWAS, and AMD GWAS in GERA study (Figure 1).[53] The default quality control procedures in MTAG were used to filter SNPs with MAF > 0.01. We then used a stepwise model selection procedure in the GCTA-COJO software (1.91.7beta) to identify lead independent genome-wide significant SNPs (both conditional and unconditional P value < 5×10^{-8}).[126,212] The lead SNPs were looked up in the Eye Genotype Expression (EyeGEx) database of retinal tissue to identify retina specific expression quantitative trait loci (eQTL) and expression-trait associations from transcriptome-wide association study (TWAS) summary results.[170] We applied SMR and HEIDI tests based on AMD meta-analysis summary statistics and the EyeGEx eQTL data.[243] We conducted gene-based and pathway analysis in MAGMA (v1.06) as implemented in FUMA platform (version 1.3.4).[213,214]

To derive a PRS, we selected the lead independent genome-wide significant SNPs, and the PRS was weighted based on the estimated AMD odds ratios (OR) from the MTAG analysis. The "pROC" package was used to calculate the area under the curve (AUC).[215] All general analyses were performed with R (version 3.4.1).[216]

5.3 Results

5.3.1 Meta-analysis of AMD GWAS identifies 12 novel loci

We conducted a meta-analysis based on MTAG method to combine three AMD GWAS summary statistics: AMD-2016 GWAS and AMD-2013 GWAS from the International AMD Genomics Consortium (IAMDGC), and AMD GWAS in the Genetic Epidemiology Research on Aging (GERA) cohort (Figure 1). The genetic correlations between the AMD input datasets were very high based on LD score regression method (Supplementary Table S1). We then investigated the MTAG output summary statistics, and found no evidence of genomic inflation (lambda genomic control 1.18, LD score regression intercept 1.03, Supplementary Figure S1). There is also no evidence of inflation due to violation of the homogeneity assumption in MTAG (max False Discovery Rate 0.0016). From the MTAG GWAS output, we identified 69 lead independent genome-wide significant SNPs (12 of them are novel loci, Figure 2, Supplementary Figure S2, and Supplementary Table S2).

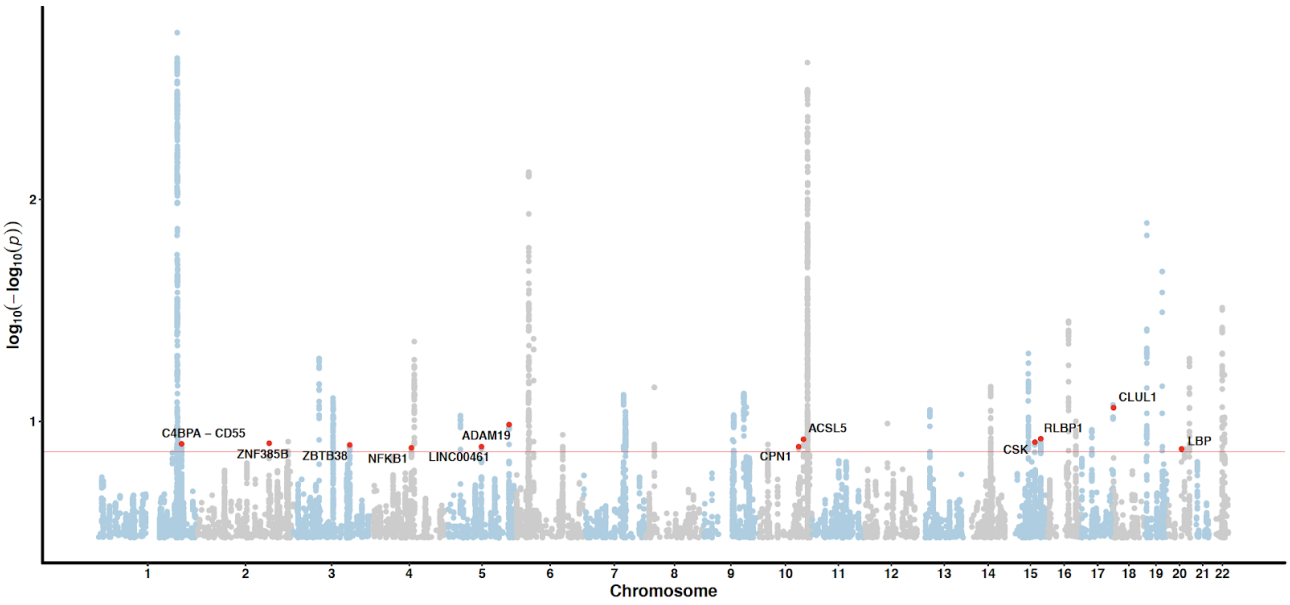


Figure 2. Manhattan plot of the meta-analysis of genome-wide association studies for AMD.

Novel loci are highlighted in red dots, with the nearest gene names in black text. The red line is the genome-wide significance level at 5×10^{-8}.

We then replicated the 12 novel AMD loci in the UKB and FinnGen AMD studies. The concordance of SNP effect sizes between the MTAG discovery cohorts and replication datasets (UKB and FinnGen) was high (Pearson's correlation coefficient 0.89, P value 1.2×10^{-24}, Figure 3). Of the 12 novel loci, 11 were in the expected direction (binomial test P value = 6.3×10^{-3}), five were associated with AMD at a nominal significance level (P value < 0.05), and one (rs3825991 in gene *RLBP1*, P value = 1.6×10^{-3}) after Bonferroni correction (Table 1). We also built a polygenic risk score using the 12 novel SNPs, and the score was strongly associated with AMD status in UKB (P = 2.4×10^{-4}).

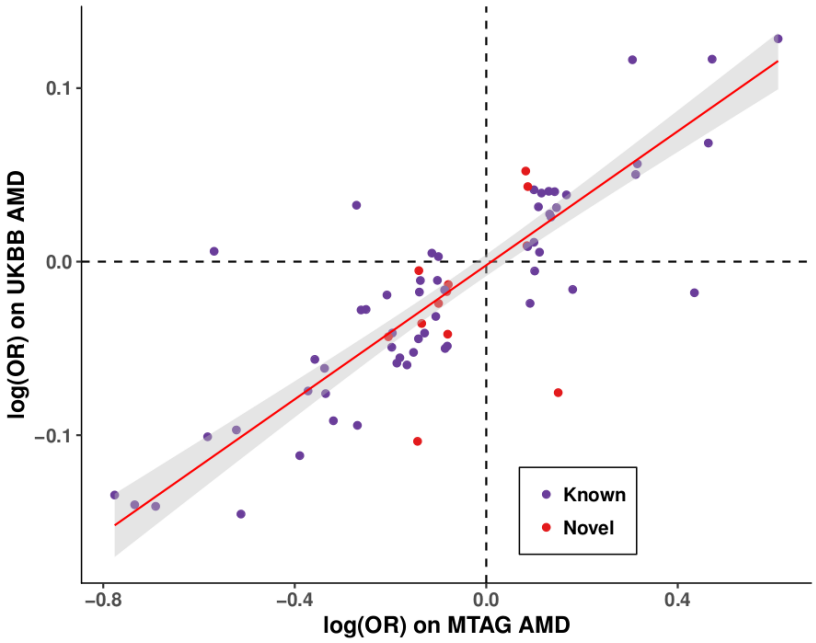


Figure 3. Comparison of the effect sizes for 69 genome-wide significant independent SNPs identified from meta-analysis of AMD GWASs versus those in UK Biobank and FinnGen AMD GWAS.

Pearson's correlation coefficient is 0.89 (P value=1.2×10^{-24}). The red line is the best fit line, with the 95% confidence interval region in grey. Novel AMD SNPs are highlighted in red and known SNPs in purple.

Most of the novel genes are expressed in retinal tissue and could be involved in the pathogenesis of AMD (Box 1). For instance, *C4BPA - CD55* loci is involved in the regulation of complement activation,[333,334] and *NFKB1* and *LBP* are important factors for inflammatory response pathways.[335,336] *LINC00461* was identified as the most significant loci associated with macular thickness.[265] *RLBP1* is associated with multiple Mendelian retinal dystrophy,[337,338] and also one of the strongest AMD-associated candidate genes from a recent transcriptome-wide association analysis.[339] These findings are important for our understanding of the pathogenesis of AMD development, and could potentially constitute therapeutic targets for AMD.[340]

Table 1. List of 12 novel AMD loci from the meta-analysis of genome-wide association studies.

SNP	CHR	BP	Nearest Gene	EA	NEA	FREQ	P (2016)	P (2013)	P (GERA)	BETA (MTAG)	P (MTAG)	P (replication)[1]	P (eQTL)[2]	P (TWAS)[3]
rs11120691	1	207486475	*C4BPA - CD55*	G	T	0.44	2.3×10^{-5}	8.2×10^{-5}	0.1	-0.08	1.2×10^{-8}	0.02	0.006	0.47
rs259842	2	180738840	*ZNF385B*	C	T	0.62	5.6×10^{-6}	2.4×10^{-3}	0.01	-0.08	1.1×10^{-8}	0.04	0.10	0.43
rs2011092	3	141124607	*ZBTB38*	C	T	0.35	4.4×10^{-6}	4.6×10^{-4}	0.22	0.09	1.4×10^{-8}	0.27	**6.2×10^{-15}**	9.4×10^{-5}
rs1005819	4	103504305	*NFKB1*	T	C	0.42	8.6×10^{-6}	4.0×10^{-4}	0.23	-0.08	2.4×10^{-8}	0.49	0.007	0.08
rs17421410	5	87836307	*LINC00461*	G	A	0.07	9.8×10^{-7}	3.4×10^{-3}	0.29	0.15	2.0×10^{-8}	0.08	0.004	0.48
rs6899205	5	156943285	*ADAM19*	A	G	0.28	5.8×10^{-7}	1.2×10^{-6}	0.98	-0.10	2.0×10^{-10}	0.13	6.7×10^{-4}	0.04
rs7896471	10	101788308	*CPN1*	T	G	0.04	1.5×10^{-7}	0.01	0.57	-0.20	1.9×10^{-8}	0.07	0.02	0.004
rs1926564	10	114139896	*ACSL5*	A	G	0.90	3.4×10^{-7}	5.1×10^{-4}	0.79	-0.13	4.9×10^{-9}	0.39	0.007	0.39
rs1378940	15	75083494	*CSK*	A	C	0.68	3.5×10^{-6}	1.8×10^{-4}	0.36	0.09	8.3×10^{-9}	9.6×10^{-3}	**1.6×10^{-9}**	0.25
rs3825991	15	89761664	*RLBP1*	A	C	0.48	1.9×10^{-7}	3.7×10^{-3}	0.26	0.08	4.3×10^{-9}	**1.6×10^{-3}**	**1.7×10^{-20}**	1.0×10^{-6}
rs9973159	18	597950	*CLUL1*	T	C	0.15	9.4×10^{-8}	4.3×10^{-6}	0.04	-0.14	2.9×10^{-12}	0.22	**5.9×10^{-18}**	0.001
rs2232613	20	36997655	*LBP*	T	C	0.08	4.3×10^{-7}	0.06	0.03	-0.14	3.0×10^{-8}	7.0×10^{-3}	-	0.06

Abbreviations: BETA, beta coefficient; CHR, Chromosome; EA, effect allele; eQTL, expression quantitative trait loci; FREQ, allele frequency of effect allele; MTAG, multiple trait analysis of GWAS; NEA, non-effect allele; P, P values; SNP, single nucleotide polymorphism; TWAS, transcriptome-wide association analysis.

UKB, UK biobank data; 2016, AMD-2016 GWAS; 2013, AMD-2013 GWAS; GERA, Genetic Epidemiology Research on Aging study.

Chromosomal position is based on the NCBI RefSeq hg19 human genome reference assembly.

[1] Novel genes passing multiple testing correction (P<0.05/12) in replication datasets (meta-analysis of UK biobank and FinnGen study) are highlighted in **bold font**.

[2] eQTL passing gene-level multiple testing correction are highlighted in **bold font**.

[3] TWAS P values look up from Ratnapriya et al 2019. For loci rs7896471 and rs2232613 are based on genes *DNMBP* and *KIAA1755*, respectively.

Box 1. Biology annotations of 12 novel AMD loci.

Nearest Genes	Gene function
C4BPA - CD55	Complement Component 4 Binding Protein Alpha (C4BPA) and Decay Accelerating Factor For Complement (CD55) are involved in the regulation of complement activation.[333,334] Previous functional studies proposed that the expression of CD55 in retinal pigment epithelium cells could be a potential therapeutic target for AMD.[340] *CD55* was also reported to be associated with myopia.[341]
ZNF385B	Zinc Finger Protein 385B (ZNF385B) is highly expressed in retinal tissue.[342] Patient with a 2q31.2-32.3 deletion presented microphthalmia and retinal coloboma.[343]
ZBTB38	*ZBTB38* encodes Zinc Finger And BTB Domain Containing 38, a zinc finger transcriptional activator that binds methylated DNA. Its function in eye is uncharacterized.
NFKB1	Nuclear Factor Kappa B Subunit 1 (NFKB1) is related to many biological processes such as inflammation, immunity, differentiation, cell growth, tumorigenesis, and apoptosis. The activation of NF-κB is an important pathway to the development of AMD.[336,344]
LINC00461	LINC00461 is a long non-coding RNA and expressed predominantly in the brain and visual cortex. It is the most significant loci associated with macular thickness.[265] It is also associated with retinal vascular caliber,[345,346] and macular telangiectasia type 2.[347]
ADAM19	Disintegrin And Metalloproteinase Domain-Containing Protein 19 (ADAM19) is a member of the ADAM (a disintegrin and metalloprotease domain) family. It is associated with Alzheimer's disease and could play an important role in retinal degeneration diseases.[348,349]
CPN1	Carboxypeptidase N Subunit 1 (CPN1) plays a central role in regulating the biologic activity of peptides such as kinins and anaphylatoxins. It could be involved in choroid development,[350] and a recent Bayesian functional association study also showed *CPN1* is associated with AMD.
ACSL5	ACSL5 plays a key role in lipid biosynthesis and fatty acid degradation.[351]
CSK	C-Terminal Src Kinase (CSK) plays an important role in T-cell activation and the phosphorylation of C-terminal tyrosine residues. It is expressed in retinal vascular endothelial cells.[352]
RLBP1	Retinaldehyde Binding Protein 1 (RLBP1) is related to multiple Mendelian retinal dystrophy.[337,338] A recent study showed this gene could increase AMD risk by the interaction effect between the nuclear and mitochondrial genome.[353] A transcriptome-wide association study also identified this gene associated with AMD.[339]
CLUL1	*CLUL1* encodes Retinal Clusterin-Like Protein. Clusterin is expressed in many eye tissues, such as retinal pigment epithelium, ganglion cells, and photoreceptor cells. Although candidate gene study found no pathogenic variants,[354] a recent study showed an interaction effect of gene-age for AMD risk.[355]
LBP	Lipopolysaccharide Binding Protein (LBP) is involved in inflammatory response through NF-kB and MAPK signaling. It protects human retinal pigment epithelial cells against oxidative stress-induced apoptosis, which contributes to the pathogenesis of AMD development.[335]

5.3.2 Gene-based and pathway analysis

We then conducted a genome-wide gene-based association analysis and identified an additional 21 novel genes (defined as no genome-wide significant SNPs within the region of a gene, Supplementary Figure S3 and Table S3). For example, the novel gene *PDGFB*, encodes Platelet Derived Growth Factor Subunit B, which is a member of the protein family comprised of both platelet-derived growth factor (PDGF) and vascular endothelial growth factor (VEGF), could provide genetic insight into the development of VEGF and PDGF inhibition for neovascular AMD.[356,357] Pathway analysis of 10,678 gene sets (MsigDB v6.2, curated gene sets: 4,761, Gene Ontology terms: 5,917) resulted in 19 significant gene sets after FDR correction, which include complement cascade, high density lipoprotein particle remodeling, cholesterol transporter activity, and negative regulation of macrophage derived foam cell differentiation (Supplement Table S4).

5.3.3 eQTL and transcriptome-wide association analysis

We also looked up the 69 genome-wide significant SNPs in retina from the Eye Genotype Expression (EyeGEx) database to identify expression quantitative trait loci (eQTL).[170] We found 12 genome-wide significant SNPs were significant eQTL for 25 SNP-gene pairs (cis-eQTLs) after gene-level multiple testing correction across the genome (Table 1 and Supplementary Table S5). Five SNP-gene cis-eQTLs were from our novel AMD SNPs: rs2011092 (cis-eQTL target gene *ZBTB38*), rs1378940 (*MPI*), rs3825991 (*RLBP1*), rs9973159 (*CLUL1*), and rs9973159 (*RP11-806L2.2*). Of the 12 genome-wide significant SNPs we identify here, one was study-wide significant in a previous transcriptome-wide association study (TWAS) based on earlier AMD GWAS summary statistics[170] (Table 1).

To test the effects of genetic variants on AMD risk that is mediated by gene expression levels, we also conducted summary data-based Mendelian randomization (SMR) and heterogeneity in dependent instruments (HEIDI) tests.[243] SMR investigates the relationships between gene expression levels (exposure) and phenotype (outcome) using genetic variants as instrumental variables. We identified 19 genes after multiple testing corrections ($P_{SMR} < 0.05/5075 = 9.85 \times 10^{-6}$, Supplementary Table S6). We further used the HEIDI method to test the null hypothesis that there is a single causal variant affecting both gene expression levels and AMD risk. We identified 12 genes that passed the HEIDI test ($P_{HEIDI} \geq 0.05$), *HLA-DQB2*, *PILRA*, *PILRB*, *STAG3L5P*, *PMS2P1*, *TSC22D4*, *BLOC1S1*, *B3GLCT*, *RLBP1*, *POLDIP2*, *CLUL1*, and *RP11-806L2.2* (based on updated EyeGEx database),

which are associated with AMD risk underlying the GWAS hits suggesting that these genes are good candidates of prioritizing for functional follow-up studies.

5.3.4 Prediction value of AMD polygenic risk score

We constructed an AMD polygenic risk score (PRS) from the 69 lead SNPs ($PRS_{69\text{-}SNP}$), and then evaluated the prediction performance in 100 late AMD cases and 2,136 controls from the Blue Mountains Eye Study (BMES). The area under the curve (AUC) of the $PRS_{69\text{-}SNP}$ was 0.76 (95% confidence interval [CI]: 0.72 - 0.80). To assess the improvement of our new PRS compared with previous AMD PRS, we derived a PRS from previously published 52 SNPs ($PRS_{52\text{-}SNP}$) for BMES. The prediction ability of our new $PRS_{69\text{-}SNP}$ was better than that based on previously published SNPs $PRS_{52\text{-}SNP}$ ($AUC_{52\text{-}SNP}$ = 0.74, 95% CI: 0.70 - 0.79), although the AUC improvement was not significant (P = 0.21).

5.4 Discussion

We have conducted a large meta-analysis of GWAS for AMD and identified 69 genome-wide significant SNPs (12 novel). We found most of the novel genes are expressed in the retina and could be involved in AMD pathogenesis. Through genome-wide gene-based association analysis, we identified an additional 21 novel genes. Pathway analysis indicated complement cascade, high density lipoprotein particle remodeling, cholesterol transporter activity, and negative regulation of macrophage derived foam cell differentiation are involved in the biological process underlying AMD risk.

In this study, we conducted a multivariate GWAS (based on MTAG method) rather than a traditional inverse-variance meta-analysis. Traditional meta-analysis assumes the input GWAS summary statistics are derived for the same trait (a genetic correlation close to one).[211] In practice, the heterogeneity of the case phenotype would lead to a lower genetic correlation (less than one) even for the same trait. Recent statistical methodology studies showed that multivariate GWAS can leverage multiple input summary statistics of the same trait with different measures or even different traits with a high genetic correlation.[53,325] Our input files AMD-2016 GWAS and AMD-2013 GWAS are summary statistics for advanced AMD, and the AMD cases in GERA are identified using electronic health records, which could include both of advanced and early or intermediate AMD cases. The MTAG approach

is able to handle this issue by leveraging the high genetic correlation between the input summary statistics and maximizing the statistical power to detect genetic associations for advanced AMD (our index input AMD-2016 GWAS, which has the highest power). More importantly, the MTAG approach can handle sample overlap between the input GWAS summary statistics.[53] In our multivariate GWAS, there is some sample overlap between the AMD-2016 GWAS and the AMD-2013 GWAS. In this scenario, MTAG framework is an ideal method for taking full advantage of the large public available GWAS summary statistics.

The gene discovery from our MTAG GWAS will contribute towards the understanding of the biology mechanisms and the etiology of AMD. As we presented in Box 1, most of the novel loci are potentially involved in the biological process of AMD. For instance, macular thickness is an important quantitative trait for AMD.[358] A recent first macular thickness GWAS identified 139 loci, and some of them are known AMD genes, such as *RDH5*, *NPLOC4*, *RAD51B*, and *SLC16A8*.[265] Our meta-analysis of GWASs identified a novel AMD loci *LINC00461*, which is the most significant signal from the macular thickness GWAS.[265] *LINC00461* is a long non-coding RNA and expressed predominantly in the brain and visual cortex.[359] Previous GWAS also indicated *LINC00461* is associated with retinal vascular caliber,[345,346] a risk factor of AMD, and macular telangiectasia type 2,[347] a rare neurovascular degenerative retinal disease. Our meta-analysis of GWASs also identified novel genes involved in the regulation of complement activation,[333,334] lipid biosynthesis,[351] inflammatory response,[335,336] and Mendelian retinal diseases.[337,338] All together, these gene findings help us have a better understanding of the pathogenesis of AMD.

In this study, we conducted a meta-analysis of GWASs for individuals of European ancestry, hence the generalizability of the novel AMD genes to other populations still needs further replication. Besides, our replication dataset of UKB has a relatively small sample size and young participants (40-69 years old), and the AMD cases were identified using both hospital health records and self-reported cases. Although the concordance of SNP effect sizes between the MTAG discovery cohorts and replication cohorts was high and most of them were in the expected direction, replication datasets with larger sample size of clinical diagnosis cases would improve the power to replicate our novel genes. Moreover, although we looked up the eQTL and TWAS results in retinal tissue and further literature search indicated most of these genes are probably involved in the pathogenesis of AMD, additional functional studies are warranted to investigate the underlying biological mechanisms of the

novel genes. Finally, using BMES samples we evaluated the prediction value of a PRS based on i) 69 lead SNPs identified here and ii) 52 previously published SNPs. Both PRSs were derived using the AMD consortium data which included a subset of the BMES samples used here to test the PRS and in theory this could induce slight over-fitting due to sample overlap. In practice this would have a negligible effect on our results because i) the sample overlap is very small (~0.5% of cases) and ii) we only used a small number of SNPs in our PRS. Although our new PRS improved the prediction AUC (from 0.74 previously, to 0.76 here), the increase was not statistically significant, possibly due to the limited number of advanced AMD cases in BMES or the small effect sizes of the additional GWAS signals. We performed an exploratory analysis using the PRS in the UKB cohort although the AUC values were substantially lower (data not shown), reflecting the fact that due to their relatively young age, most UKB cases did not have advanced AMD.

In conclusion, we conducted a meta-analysis GWAS for AMD and identified 12 novel loci. Most of the novel genes are expressed in retinal tissue and could be involved in the pathogenesis development of AMD. These findings enhance our understanding of the disease mechanisms of AMD.

5.5 Acknowledgements

This work was conducted using the UK Biobank Resource (application number 25331), the Genetic Epidemiology Research on Aging (GERA) cohort (dbGaP, study accession: phs000674.v3.p3), and publicly available data from the International AMD Genomics Consortium (IAMDGC). We want to acknowledge the participants and investigators of the FinnGen study. We thank Scott Wood, Xiaping Lin, John Pearson and Scott Gordon from QIMR Berghofer for support.

The GERA data came from a grant, the Resource for Genetic Epidemiology Research in Adult Health and Aging (RC2 AG033067; Schaefer and Risch, PIs) awarded to the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and the UCSF Institute for Human Genetics. The RPGEH was supported by grants from the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, Kaiser Permanente Northern California, and the Kaiser Permanente National and Northern California Community Benefit Programs. The RPGEH and the Resource for Genetic Epidemiology Research in Adult Health and Aging are described in the following publication, Schaefer C, et al., The Kaiser Permanente Research Program on Genes, Environment and Health: Development of a Research Resource in a Multi-Ethnic Health Plan with Electronic Medical Records, In preparation, 2013.

Funding

SM and AWH are supported by Australian National Health and Medical Research Council (NHMRC) Fellowships. We acknowledge funding from NHMRC grants 1116360, 1150144 and 1123248.

Conflict of interest disclosures

The authors declare no potential conflicts of interest.

5.6 Supplement

Supplementary Tables and Figures are available at:

https://www.nature.com/articles/s10038-020-0750-x

CHAPTER 6A

Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration

Xikun Han, Jue-Sheng Ong, Jiyuan An, Alex W. Hewitt, Puya Gharahkhani, Stuart MacGregor.

European Journal of Epidemiology. 2020;35(2):139-146.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Stuart MacGregor, Alex W Hewitt, and Puya Gharahkhani obtained funding and designed the study. All authors contributed to interpretation of the results and the final version of the paper.

Chapter 6A. Using Mendelian randomization to evaluate the causal relationship between serum C-reactive protein levels and age-related macular degeneration

Serum C-reactive protein (CRP), an important inflammatory marker, has been associated with age-related macular degeneration (AMD) in observational studies; however, the findings are inconsistent. It remains unclear whether the association between circulating CRP levels and AMD is causal. We used two-sample Mendelian randomization (MR) to evaluate the potential causal relationship between serum CRP levels and AMD risk. We derived genetic instruments for serum CRP levels in 418,642 participants of European ancestry from UK Biobank, and then conducted a genome-wide association study for 12,711 advanced AMD cases and 14,590 controls of European descent from the International AMD Genomics Consortium (IAMDGC). Genetic variants which predicted elevated serum CRP levels were associated with advanced AMD (odds ratio [OR] for per standard deviation [SD] increase in serum CRP levels: 1.31, 95% confidence interval [CI]: 1.19 to 1.44, $P = 5.2 \times 10^{-8}$). The OR for the increase in advanced AMD risk when moving from low (<3 mg/L) to high (>3 mg/L) CRP levels is 1.29 (95% CI: 1.17 - 1.41). Our results were unchanged in sensitivity analyses using MR models which make different modelling assumptions. Our findings were broadly similar across the different forms of AMD (intermediate AMD, choroidal neovascularization, and geographic atrophy). We used multivariable MR to adjust for the effects of other potential AMD risk factors including smoking, body mass index, blood pressure and cholesterol; this did not alter our findings. Our study provides strong genetic evidence that higher circulating CRP levels lead to increases in risk for all forms of AMD. These findings highlight the potential utility for using circulating CRP as a biomarker in future trials aimed at modulating AMD risk via systemic therapies.

6A.1 Introduction

Age-related macular degeneration (AMD) is the leading cause of irreversible central vision loss among the elderly population in the Western world.[134,138,321–323] The etiology of AMD is not yet well understood; however, several hypotheses focus on the pathogenic pathways related to genetic predisposition, inflammation, complement, lipid, and oxidative stress.[138,166,360,361] In support of this, genome-wide association studies (GWAS) have identified a variety of complement pathway related genes, such as complement factor H (*CFH*), factor I (*CFI*), and the complement components *C2*, *C3*, and *C9*.[165,166] The presence of complement and inflammatory reactions in drusen, the hallmark lesions of AMD, suggests the important role of inflammation in AMD pathogenesis.

C-reactive protein (CRP) is the most studied systemic marker of inflammation,[362] and could induce proinflammatory responses and the progression of AMD.[361] Drugs targeted to CRP that alleviate inflammatory responses have been postulated to prevent the progression of AMD.[360,361] However, observational studies have shown mixed conclusions on the association between circulating CRP levels and the risk of AMD.[159,363–369] Previous genetic studies have found no evidence of association between genetic variants in the *CRP* gene and AMD risk.[370–373] However, these genetic variants at the *CRP* locus only account for a relatively small proportion of the variability of circulating CRP levels, and more robust instruments for quantifying the genetic contribution to circulating CRP are needed.[373] Therefore, it remains unclear whether elevated circulating CRP levels are causally related to AMD risk.

Mendelian randomization (MR) is an instrumental-variable based approach to investigate the causal relationships between risk factors and outcomes via the use of genetic instruments (single nucleotide polymorphisms [SNPs] being most commonly used).[70,374] In MR analysis, as genetic instruments are distributed randomly at conception, the predicted circulating CRP levels are unlikely to be related to confounders of AMD risks or consequentially influenced by AMD disease status through reverse causality. Therefore, the study design of MR is akin to a natural analogue of traditional RCT where unmeasured confounding are randomized across both the genetically predisposed (circulating CRP-increasing allele carriers) and unaffected group (reference; non-effect allele carriers).[70] In

this study, we investigate the causal relationship between circulating CRP levels and AMD risk, which would provide therapeutic implications for the prevention and treatment of AMD.

6A.2 Methods

6A.2.1 Study design

To investigate the causal relationship between serum CRP levels and AMD risk, we applied the two-sample MR framework,[375] an approach to make causal inference using GWAS summary statistics for exposure and outcome from separate GWASs. We conducted a GWAS for serum CRP levels in the UK Biobank (UKB) cohort to obtain genetic instruments for measured circulating CRP levels. We then conducted a series of GWAS analyses for advanced AMD and other AMD subtypes using the individual level data from the International AMD Genomics Consortium (IAMDGC). To assess sample overlap between UKB and IAMDGC datasets, we ran LD score regression between CRP GWAS in UK Biobank and advanced AMD GWAS in IAMDGC dataset. The intercept of genetic covariance is 0.0058 (standard error 0.0104), which indicates that the intercept is approximately zero and there is little or no sample overlap between the two datasets.

The UK Biobank study was approved by the National Research Ethics Service Committee North West—Haydock, all participants provided informed written consent, and all study procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research. In the International AMD Genomics Consortium, all groups collected data according to the Declaration of Helsinki principles. Study participants provided informed consent, and protocols were reviewed and approved by the local ethics committees.

6A.2.2 Genetic instruments for serum C-reactive protein levels

The UKB is a large-scale population-based cohort study of half a million people aged between 40-69 years living in the United Kingdom.[130] The serum CRP levels were available for 469,881 individuals (UKB data field 30710, http://biobank.ctsu.ox.ac.uk/crystal/field.cgi?id=30710) as part of the recent UKB release (2019 release) for serum biochemistry data, and were measured using immunoturbidimetric method (high sensitivity analysis on a Beckman Coulter AU5800). The reportable range of

high sensitivity serum CRP is from 0.08 mg/L to 80 mg/L (mean and standard deviation: 2.60 ± 4.34 mg/L, Supplementary Table 1). We included 418,642 participants of white British ancestry in the following serum CRP GWAS analysis (Supplementary Figure 1). We calculated the average values of serum CRP levels for individuals that underwent two assessments. We applied a rank-based inverse-normal transformation to serum CRP levels.

For the serum CRP GWAS in UKB, we conducted a linear mixed model under an additive genetic model implemented via the BOLT-LMM software (version 2.3).[210] The model was adjusted for sex, age and the first ten principal components (PCs). We selected independent genome-wide significant variants as genetic instruments for serum CRP levels using the following criteria: 1) P value on serum CRP < 5×10^{-8}; 2) linkage disequilibrium (LD) between SNPs r^2 < 0.01; and 3) the SNPs being present in the AMD GWAS summary statistics (described below). The LD-clumping procedure was performed using PLINK (version 1.9).[182]

In our sensitivity analyses, we used the following methods to derive the genetic instruments: 1) in UKB, we removed 16,946 (4%) participants with circulating CRP levels > 10 mg/L (*e.g.* due to a serious infection in the participant) and adjusted for body mass index (BMI, data field 21001) in the association models; 2) we used previously reported circulating CRP variants (44 SNPs in our AMD GWAS summary statistics described below);[376] 3) to evaluate the potential pleiotropic effects of CRP genetic instruments, we also ran a series of GWASs for other potential AMD risk factors including smoking, BMI, systolic blood pressure (SBP), high-density lipoprotein cholesterol (HDL-C), and glycated haemoglobin (HbA1c) in UKB (Supplementary Table 1). In the causal inference of CRP levels on the risk of AMD, we adjusted for these risk factors by a multivariable MR analysis (see statistical analysis section, below).

6A.2.3 Age-related macular degeneration dataset

The International Age-related Macular Degeneration Genomics Consortium (IAMDGC) dataset is the largest European GWAS focusing on AMD susceptibility (16,144 advanced AMD cases and 17,832 controls).[166] The full description of the study design, phenotype definition, and genetic data were described previously.[166] Briefly, in IAMDGC, data were gathered from 26 studies with each including (a) advanced AMD cases with choroidal neovascularization (CNV) and/or geographic atrophy (GA) in at least one eye and age at

first diagnosis more than 50 years old; (b) intermediate AMD cases with pigmentary changes in the retinal pigment epithelium or more than five macular drusen greater than 63 μm in diameter and age at first diagnosis more than 50 years old; or (c) controls without known advanced or intermediate AMD.[166] The individual level AMD phenotype data and genetic data are available in the database of Genotypes and Phenotypes (dbGaP, study accession: phs001039.v1.p1).[166] We downloaded the imputation data for 35,358 participants. The imputation was based on the 1000 Genomes Project reference panel (1000 Genomes Project Phase I, version 3) using Minimac.[166,377] SNPs with imputation quality score >0.3 and MAF >0.01 were retained for association analysis. In our association analysis, we removed non-European ancestry participants based on the first two principal components inferred ancestry.[166] Finally, we included 12,711 advanced AMD cases (8,544 CNV, 2,656 GA, and 1,511 mixed AMD [both of CNV and GA] cases), 5,336 intermediate AMD cases, and 14,590 controls in our analysis (Supplementary Table 2). We ran GWAS analyses for 12,711 advanced AMD cases and other AMD subtypes with 14,590 controls in PLINK software (version v2.00a1LM) adjusting for sex, age, and the first ten PCs.

6A.2.4 Statistical Analysis

To assess the power of our MR analyses, we used the mRnd (http://cnsgenomics.com/shiny/mRnd/) method to evaluate power for different AMD subtypes.[378] We conducted two-sample MR for circulating CRP levels and AMD risk using inverse-variance weighted (IVW) method as the main analysis.[81,379] We verified the estimates using the MR weighted median and MR-Egger methods to allow violations of MR assumptions.[71,380] Specifically, the weighted median MR method allows genetic variants representing over 50% of the weight in the MR analysis are valid instruments, while MR-Egger method can detect and correct for the bias due to directional pleiotropy (pleiotropic effects of genetic instruments do not average to zero).[76,381] The intercept from MR-Egger method was used to assess directional pleiotropy (*i.e.* intercept P value < 0.05).[76] Although pleiotropy is concerning, if the pleiotropic effects are equally to be positive or negative (no directional pleiotropy), the overall estimate would be unbiased.[381] We also used the funnel plot and MR-PRESSO method to evaluate bias from outliers and assess the heterogeneity of genetic instruments.[76,77] To further assess potential pleiotropic effects of related risk factors, we conducted a multivariable MR analysis.[79,382,383] In univariate MR analysis, the causal effect of a risk factor (CRP level) on the outcome (AMD) was assessed via genetic

variants that are solely associated with that specific risk factor. The univariate IVW MR method is a weighted linear regression method to regress the effects of genetic instruments on AMD (outcome) against their effects on CRP level (exposure), with a forced intercept term at zero and weighted by inverse-variance. In multivariable MR analysis, we conducted GWAS for other potential AMD risk factors including smoking, BMI, SBP, HDL-C, and HbA1c. The analytic framework for multivariable MR-IVW method is similar to univariable MR-IVW except regressing on the effects of multiple risk factors in a single regression model.[79] In general, the univariate MR estimates the total effect of the circulating CRP on AMD risk, whereas multivariable MR could estimate the direct causal effect of circulating CRP on AMD risk when conditioned on the presence of other AMD risk factors.[79,384]

We performed MR analyses using R packages MendelianRandomization and TwoSampleMR.[76,385] All analyses were performed with R (version 3.4.1).[216]

6A.3 Results

6A.3.1 Genetic instruments and statistical power

In our UKB circulating CRP GWAS, we identified 526 independent genome-wide significant SNPs as genetic instruments (Supplementary Table 3), which explained 13% of the variance of circulating CRP levels (Supplementary Figure 2). Our MR analyses yield adequate power to detect moderate effect sizes (*e.g.* odds ratio [OR] 1.2 per standard deviation increase of circulating CRP levels); our power for advanced AMD, intermediate, GA, CNV, and mixed AMD is 100%, 99%, 91%, 100%, and 75%, respectively (Supplementary Table 4).

6A.3.2 Circulating CRP levels are associated with advanced AMD

The MR scatter plot indicates that higher serum CRP levels were associated with increased risk of advanced AMD (Figure 1). The overall MR-IVW OR of advanced AMD per standard deviation (SD, 4.34 mg/L) increase in circulating CRP levels was 1.31 (95% confidence interval [CI]: 1.19 to 1.44, $P = 5.2 \times 10^{-8}$, Table 1), which is 1.06 for each one mg/L increase in circulating CRP levels. Another way of interpreting these results is to consider a more clinically relevant change. For example we can consider a change in CRP for those with high (> 3 mg/L) versus low (< 3 mg/L) levels. The estimated odds ratio for the difference between these groups is 1.29 ($\exp(\log_e 1.31 / 4.34 * 4.09)$; where 4.09 mg/L is the change in CRP between the median level in the high group and the median level in the low group).

The estimation between circulating CRP levels and advanced AMD was similar to the results from MR-Egger method (OR = 1.41, 95% CI: 1.22 - 1.63, P = 1.9×10^{-6}) and MR weighted median method (OR = 1.17, 95% CI: 1.00 - 1.37, P = 0.046) with overlapping confidence intervals. We found no evidence of directional pleiotropy effects based on MR-Egger intercept test (intercept -0.003, P = 0.15). The MR-PRESSO outlier-corrected result was not meaningfully different from the MR-IVW estimate (OR = 1.17, 95% CI: 1.07 - 1.29, P = 8.2×10^{-4}) and the MR funnel plot showed no evidence of asymmetry (Supplementary Figure 3). To further investigate whether pleiotropy effects distorted our estimates, we also conducted a multivariable Mendelian randomization analysis to adjust for other potential AMD risk factors including: smoking; body BMI; SBP; HDL-C; and HbA1c. The association between circulating CRP levels and advanced AMD was essentially unchanged in multivariable MR analysis (OR=1.27, 95% CI: 1.14 to 1.40, P = 7.1×10^{-6}). The consistency of total effect and direct effect of CRP levels on AMD risk based on univariate and multivariable MR estimates supported an independent association between circulating CRP levels and the risk of AMD (Table 1).

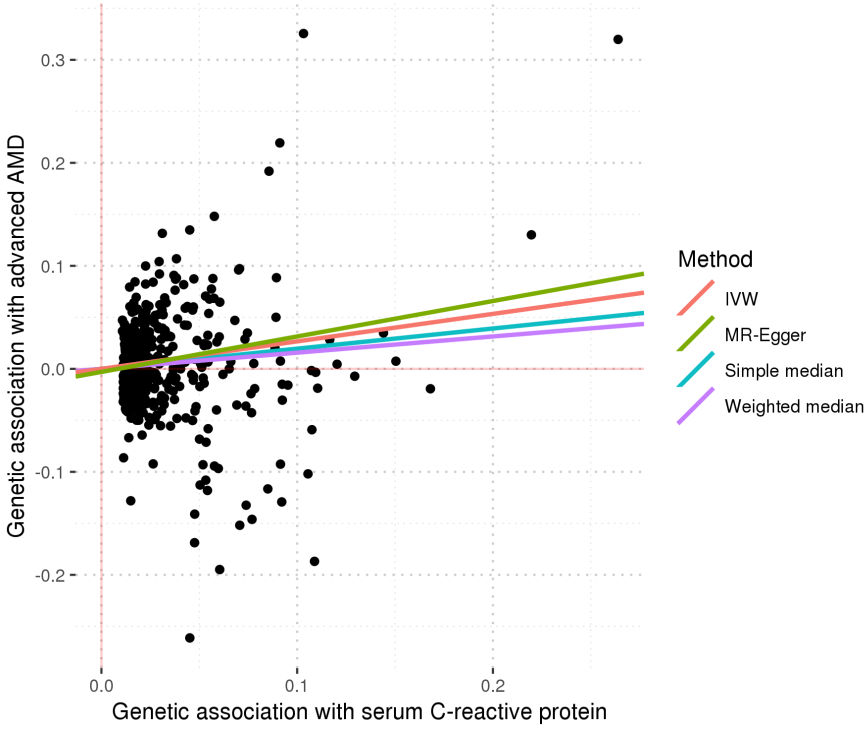


Figure 1. Serum C-reactive protein-increasing risk variants are associated with increased risk of advanced age-related macular degeneration.

The x-axis shows the estimates for the 526 genetic instruments for serum C-reactive protein levels, the y-axis shows the estimates (log odds ratios) of the effects of the same variants on advanced age-related macular degeneration. The Mendelian randomization (MR) inverse-weighted (IVW), MR-Egger, simple median and weighted median method lines are plotted with red, green, blue, and purple lines, respectively.

Table 1. Mendelian randomization estimates of the associations between serum C-reactive protein levels and age-related macular degeneration.

Trait[1]	Method	OR[2]	95% CI	P-value
Advanced AMD	IVW	1.31	[1.19, 1.44]	5.2×10^{-8}
	Multivariable MR[3]	1.27	[1.14, 1.40]	7.1×10^{-6}
	Weighted median	1.17	[1.00, 1.37]	0.047
	MR-Egger	1.41	[1.22, 1.63]	1.9×10^{-6}
	(intercept)	-0.003	[-0.007, 0.001]	0.15
Intermediate AMD	IVW	1.15	[1.04, 1.27]	7.1×10^{-3}
	Multivariable MR	1.12	[1.00, 1.24]	0.046
	Weighted median	1.1	[0.92, 1.32]	0.29
	MR-Egger	1.25	[1.08, 1.45]	2.7×10^{-3}
	(intercept)	-0.003	[-0.008, 0.001]	0.11
GA AMD	IVW	1.28	[1.11, 1.48]	7.3×10^{-4}
	Multivariable MR	1.19	[1.02, 1.38]	0.03
	Weighted median	1.07	[0.82, 1.38]	0.63
	MR-Egger	1.21	[0.98, 1.50]	0.08
	(intercept)	0.002	[-0.004, 0.008]	0.49
CNV AMD	IVW	1.28	[1.15, 1.43]	3.5×10^{-6}
	Multivariable MR	1.26	[1.13, 1.42]	5.4×10^{-5}
	Weighted median	1.25	[1.05, 1.48]	0.01
	MR-Egger	1.39	[1.19, 1.62]	3.1×10^{-5}
	(intercept)	-0.003	[-0.007, 0.001]	0.17
Mixed AMD	IVW	1.52	[1.28, 1.79]	1.3×10^{-6}
	Multivariable MR	1.48	[1.23, 1.77]	2.2×10^{-5}
	Weighted median	1.61	[1.17, 2.21]	3.3×10^{-3}
	MR-Egger	2.08	[1.63, 2.67]	6.7×10^{-9}
	(intercept)	-0.01	[-0.02, -0.005]	6.9×10^{-4}

All AMD	IVW	1.26	[1.16, 1.37]	1.1×10^{-7}
	Multivariable MR	1.23	[1.12, 1.34]	1.2×10^{-5}
	Weighted median	1.17	[1.02, 1.34]	0.03
	MR-Egger	1.36	[1.21, 1.55]	9.9×10^{-7}
	(intercept)	-0.003	[-0.007, 0.0004]	0.08

AMD, age-related macular degeneration; CI, confidence interval; IVW, inverse-variance weighted; MR, Mendelian randomization; OR, odds ratio.

[1] Different subtypes of age-related macular degeneration: advanced AMD, intermediate AMD, geographic atrophy (GA) AMD, choroidal neovascularization (CNV) AMD, mixed AMD (CNV and GA), and all AMD (both of intermediate AMD and advanced AMD).

[2] The intercepts for MR-Egger are shown on the raw scale rather than the exponential scale.

[3] Multivariable Mendelian randomization analysis is a regression-based MR method adjusting here for the effects of smoking, body mass index, systolic blood pressure, high-density lipoprotein cholesterol, and glycated haemoglobin (HbA1c).

6A.3.3 Sensitivity analysis

We constructed genetic instruments for circulating CRP by removing participants with serum CRP > 10 mg/L and adjusting for body mass index (BMI) in circulating CRP GWAS. The average MR-IVW OR of advanced AMD per SD (1.83 mg/L) increase in circulating CRP levels was 1.22 (95% CI: 1.09 - 1.37, $P = 6.4 \times 10^{-4}$). We also repeated our MR analysis using 44 previously reported circulating CRP variants (independent from the UKB cohort) as genetic instruments,[376] the estimation was similar to our main analysis (OR per unit change in the natural-log-transformed CRP (mg/L) was 1.40, 95% CI: 1.16 - 1.70, $P = 5.4 \times 10^{-4}$); this shows our results are robust to the particular SNP instruments used, although as expected our power is highest (and consequential our confidence intervals are narrowest) with the full set of genome-wide significant SNPs.

6A.3.4 Circulating CRP levels are associated with different AMD subtypes

We then evaluated the causal relationships between circulating CRP levels and different AMD subtypes (Table 1 and Supplementary Figure 4). The MR-IVW ORs of circulating CRP levels on different AMD subtypes were highly consistent: for intermediate AMD, GA, CNV, and mixed AMD types the ORs were 1.15 (95% CI: 1.04 - 1.27, $P = 7.1 \times 10^{-3}$), 1.28 (95% CI: 1.11 - 1.48, $P = 7.3 \times 10^{-4}$), 1.28 (95% CI: 1.15 - 1.43, $P = 3.5 \times 10^{-6}$), and 1.52 (95% CI: 1.28 - 1.79, $P = 1.3 \times 10^{-6}$), respectively. Our results indicated circulating CRP levels

were causally associated with each of intermediate AMD, CNV, GA, and mixed AMD types. These results show that the overall findings are unlikely to be driven by a very strong association on specific AMD subtypes, suggesting CRP may be involved in different stages and types of AMD progression.

6A.4 Discussion

In this study, we conducted comprehensive MR analyses to investigate the causal relationships between circulating CRP levels and the risk of different AMD subtypes. We found that higher circulating CRP levels were associated with increased risk of advanced AMD and other AMD subtypes. These findings enhance our understanding of the underlying pathological mechanism of AMD and could have clinical utility for identification of high-risk individuals.

Our study corroborates results from previous observational studies and meta-analysis showing elevated circulating CRP is a risk factor for AMD.[159,363,364,367,368] A meta-analysis showed that the OR for higher circulating CRP level (CRP >3 mg/L vs <1 mg/L) was 2.19 (95% CI: 1.38 - 3.47) for advanced AMD; the OR was 1.31 (95% CI: 1.04 - 1.65) for combined early and late AMD.[159] Another pooled analysis of five cohorts also indicated that elevated CRP levels (CRP >3 mg/L vs <1 mg/L) increased the risk of overall incident AMD (OR = 1.49; 95% CI: 1.06 - 2.08) and neovascular AMD (OR = 1.84; 95% CI: 1.14 - 2.98).[368] In our MR analysis, the OR is also higher for advanced AMD (OR = 1.31; 95% CI: 1.19 - 1.44) compared with only intermediate AMD (OR = 1.15; 95% CI: 1.04 - 1.27) albeit with overlapping CIs. These results may indicate circulating CRP levels have a larger effect on advanced AMD than early or intermediate AMD. However, some observational studies failed to obtain evidence for the association between circulating CRP levels and AMD risk.[365,366] The inconsistent results from observational studies may be due to selection bias of AMD subtype composition (small proportion of advanced AMD cases), small sample size, and sub-optimal study designs (i.e. susceptible to confounding for cross sectional or case-control designs).[365,366] The key advantage of Mendelian randomization analysis is that the causal inference drawn through genetic instruments is less likely to be susceptible to confounding and reverse causation. As an ancillary analysis we used reverse-direction MR to examine the effect of AMD on circulating CRP levels but found no effect (MR-IVW P value 0.56).[70]

Several studies investigated the association of genetic variants in *CRP* gene and AMD risk, but found no evidence for an association.[370–373] Although the variants in *CRP* gene that were used in these studies are associated with circulating CRP levels, these SNPs in aggregation only explain a relatively small proportion of the variance in circulating levels of CRP (r^2 < 2%), hindering power for a proper MR analysis.[373] Moreover, the sample sizes for AMD cases and controls in their studies were relatively small. In our MR analysis, we conducted the largest GWAS for circulating CRP levels to date, and the lead 526 circulating CRP levels related SNPs explained 13% of the variance. In our sensitivity analysis, we also used a 44 SNP set (explaining about 7% of the variance; our results were similar, although as expected confidence intervals were considerably wider.[376]

A concern in MR analysis is the possibility of pleiotropic effects of genetic instruments.[381] It is possible that a subset of our CRP variants might have been associated with AMD risk through measured or unmeasured confounders, which may violate one of the MR assumptions.[70] To address this concern, our sensitivity MR analysis performed using the MR-Egger and weighted median methods results in similar conclusion showing that our findings were robust.[82,374] We found no evidence of directional pleiotropy based on MR-Egger intercept test and the MR funnel plot showed no evidence of asymmetry. We also used a multivariable Mendelian randomization analysis to adjust for potential AMD risk factors including smoking, BMI, SBP, HDL-C, and HbA1c. The associations between circulating CRP levels and advanced AMD or other AMD subtypes in the multivariable model were similar to those estimated from the main analysis. The sensitivity analysis to construct genetic instruments for circulating CRP by removing participants with serum CRP > 10 mg/L and adjusting for BMI or using 44 previously reported circulating CRP variants also showed similar results to the main analysis. These results indicate the finding of an association between circulating CRP and AMD risk is unlikely to be driven by horizontal pleiotropy effects.

There are several limitations in our study. In the GWASs of circulating CRP and AMD, we only included European ancestry participants, thus it is unclear whether our results are also applicable to people not of European ancestry. The generalizability of the association between circulating CRP levels and AMD risk in other ethnic groups would require further investigation. Secondly, we estimated the overall population-averaged effect of elevated

circulating CRP levels and AMD risk assuming linearity, and did not attempt to dissect potential non-linear relationships between circulating CRP levels and AMD risk. Thirdly, the MR findings reflect the change in AMD risk due to a genetically predisposed (lifetime) change in circulating CRP levels, hence the short-term effect of increasing circulating CRP levels on AMD risk is unknown.

In conclusion, elevated circulating CRP levels were associated with increased risk of AMD. Our study provided strong evidence for a causal effect of inflammation as proxied via higher circulating CRP concentrations on AMD risk, regardless of AMD disease subtypes. Further studies are warranted to investigate the clinical utility of serum CRP levels in combination with the other AMD predictors for identification of high-risk individuals and therapeutic treatment in preventing AMD.

6A.5 Acknowledgements

This work was conducted using the UK Biobank Resource (application number 25331). For the AMD datasets, all contributing sites and additional funding information are acknowledged in this publication: Fritsche et al. (2016) Nature Genetics 48 134–143, (doi:10.1038/ng.3448); The International AMD Genomics consortium's web page is: http://eaglep.case.edu/iamdgc_web/, and additional information is available on: http://csg.sph.umich.edu/abecasis/public/amd2015/.

The AMD case-control datasets used for the analyses described in this manuscript were obtained from the NEI Study of Age-Related Macular Degeneration (NEI-AMD) Database found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001039.v1.p1 through dbGaP accession number 20740. Funding support for NEI-AMD was provided by the National Eye Institute. We would like to thank NEI-AMD participants and the NEI-AMD Research Group for their valuable contribution to this research.

SM and AWH are supported by Australian National Health and Medical Research Council (NHMRC) Fellowships. We acknowledge funding from NHMRC grants [1116360, 1150144 and 1123248].

The authors thank all of the participants who took part in the UK Biobank and the International AMD Genomics Consortium and support staff who made this study possible.

We thank Scott Wood, Xiaping Lin, and John Pearson from QIMR Berghofer for IT support.

Competing Interests

The authors have declared that no competing interests exist.

6A.6 Supplement

Supplementary Tables and Figures are available at:

https://link.springer.com/article/10.1007%2Fs10654-019-00598-z

CHAPTER 6B

The effects of eight serum lipid biomarkers on age-related macular degeneration risk: a Mendelian randomization study

Xikun Han, Jue-Sheng Ong, Alex W. Hewitt, Puya Gharahkhani, Stuart MacGregor.

International Journal of Epidemiology. 2020; 50(1):325-336

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Stuart MacGregor, Alex W Hewitt, and Puya Gharahkhani obtained funding and designed the study. All authors contributed to interpretation of the results and the final version of the paper.

Chapter 6B. The effects of eight serum lipid biomarkers on age-related macular degeneration risk: a Mendelian randomization study

Age-related macular degeneration (AMD) is a leading cause of vision loss. While lipids have been studied extensively to understand their effects on cardiovascular diseases, their relationship with AMD remains unclear. In this study, two-sample Mendelian randomization (MR) analyses were performed systematically to evaluate the causal relationships between eight serum lipid biomarkers, consisting of apolipoprotein A1 (ApoA1), apolipoprotein B (ApoB), total cholesterol (CHOL), high-density lipoprotein cholesterol (HDL-C), direct low-density lipoprotein cholesterol (LDL-C), lipoprotein A (Lp(a)), triglycerides (TG), and non-HDL cholesterol (non-HDL-C), and the risk of different AMD stages and subtypes. We derived 64 to 407 genetic instruments for eight serum lipid biomarkers in 419,649 participants of European descent from the UK Biobank cohort. We conducted genome-wide association studies (GWAS) for 12,711 advanced AMD cases (8,544 choroidal neovascularization [CNV] and 2,656 geographic atrophy [GA] specific AMD subtypes) and 5,336 intermediate AMD cases with 14,590 controls of European descent from the International AMD Genomics Consortium. Higher HDL-C and ApoA1 levels increased the risk of all AMD subtypes. LDL-C, ApoB, CHOL, and non-HDL-C levels were associated with decreased risk of intermediate and GA AMD but not with CNV. TG levels were associated with decreased risk of different AMD subtypes. Sensitivity analyses revealed no evidence for directional pleiotropy effects. In our multivariable MR analyses, adjusting for the effects of correlated lipid biomarkers yielded similar results. These results suggest the role of lipid metabolism in drusen formation and particularly in AMD development at both the early and intermediate stages. Mechanistic studies are warranted to investigate the utility of lipid pathways for therapeutic treatment in preventing AMD.

6B.1 Introduction

Age-related macular degeneration (AMD) is a leading cause of vision loss among elderly in western countries.[134,138,322] The global prevalence of AMD is 8.7% among individuals aged 45 years and over, with a higher prevalence of 12.3% in Europeans.[134] The progression of AMD is classified as early, intermediate, and late stages.[135,136] The clinical hallmark in the early stage of AMD is the presence of drusen, which are formed by deposits of extracellular debris between the retinal pigment epithelium and Bruch's membrane.[386] The initiation and formation of drusen are not yet well understood; histochemical studies support an “oil spill” model, indicating lipid-rich extracellular lesions in drusen.[387–389] Approximately 40% of druse content is comprised of lipids.[390] Intermediate AMD is characterized by extensive intermediate drusen or at least one large drusen of diameter ≥125 μm.[391] AMD has two advanced types: 1) geographic atrophy (GA, dry) AMD, accounting for 90% of AMD, is characterized by drusen and retinal pigment epithelium degeneration (focal hyperpigmentation or atrophy); and 2) choroidal neovascularization (CNV, wet) AMD, is characterized by abnormal vascular proliferation underneath the retina. Currently, anti-vascular endothelial growth factor therapies have been used to reduce the progression of CNV.[137] However, the treatment is not curative, and there are no effective medications for GA. Moreover, a better scenario is to treat AMD at an earlier stage before serious vision loss occurs. It is therefore important to find new pathogenesis pathways and intervention targets for AMD.

In recent years, epidemiological and genetic studies have shown the potential role of lipids in AMD risk.[145–151] For instance, an observational meta-analysis reported a higher level of high-density lipoprotein cholesterol (HDL-C) was associated with an increased risk of AMD, whereas higher levels of total cholesterol (CHOL), low-density lipoprotein cholesterol (LDL-C), and triglycerides (TG) were associated with a decreased risk of AMD.[146] However, observational studies have shown inconsistent results respecting the association between lipids and AMD risk,[146,149] and are susceptible to confounding factors or influenced by reverse causality.[145,146,150] Genome-wide association studies have identified more than 30 genes associated with AMD, and some of them are also associated with lipid traits, such as *ABCA1*, *APOE*, *CETP*, *LIPC*, and *VEGFA*.[166]

Mendelian randomization (MR) is an approach to investigate the causal relationships between risk factors and outcomes via the use of genetic variants as natural experiments. Compared with traditional observational studies, MR is less likely to be affected by confounding or reverse causation.[70,374] Two previous MR studies have shown a causal relationship between increased HDL-C levels and advanced AMD risk.[147,148] However, the relationship between HDL-C and different AMD subtypes remains unclear. More importantly, the associations between different lipid subfractions, such as apolipoprotein A1 (ApoA1), apolipoprotein B (ApoB), and lipoprotein A (Lp(a)), and AMD risk have not been well studied. Elucidating these causal relationships might help us identify lipid-modifying therapeutics targets for AMD.

In this study, we systematically investigate the association between eight major serum lipid biomarkers (ApoA1, ApoB, CHOL, HDL-C, LDL-C, Lp(a), TG, and non-HDL-C) available in UK Biobank and the risk of different AMD subtypes using large scale genetic data from the International AMD Genomics Consortium via a two-sample MR framework. To our knowledge, our study is the first to consider the effect of a wide range of lipid biomarkers (eight in total, including ApoA1, ApoB, and Lp(a)) on the risk of AMD and its subtypes. This study would help us glean a better understanding of the role of lipids in different AMD stages and subtypes, and provide therapeutic implications for AMD.

6B.2 Methods

We performed genome-wide association studies (GWAS) for each of the eight serum lipid biomarkers in the UK Biobank cohort to identify genetic instruments. We then conducted a series of GWAS analyses on AMD outcomes of interest (namely, for intermediate AMD, advanced AMD and its subtypes CNV and GA) using the individual level data from the International AMD Genomics Consortium (independent samples from UK Biobank). Causal inferences can then be drawn via two-sample MR analysis to evaluate the (genetic) causal relationships between each of the eight serum lipid biomarkers and different AMD subtypes using GWAS summary statistics.[375]

The UK Biobank study was approved by the National Research Ethics Service Committee North West—Haydock, all participants provided informed written consent, and all study

procedures were performed in accordance with the World Medical Association Declaration of Helsinki ethical principles for medical research. In the International AMD Genomics Consortium, all groups collected data according to the Declaration of Helsinki principles. All study participants provided informed consent, and protocols were reviewed and approved by the local ethics committees.[166]

6B.2.1 Serum lipid biomarkers in UK biobank

The UK Biobank is a prospective cohort study with deep genetic and phenotypic data collected on half a million people aged between 40-69 years across the United Kingdom.[130] The sample collection and quality control procedures for serum lipid biomarkers were described in detail elsewhere (see: http://biobank.ctsu.ox.ac.uk/crystal/crystal/docs/biomarker_issues.pdf). We identified 438,870 individuals who were genetically similar to those of white-British ancestry.[126] For lipid biomarker GWAS analyses, we only included participants of white British ancestry.[126] The serum lipid biomarkers ApoA1, ApoB, CHOL, HDL-C, direct LDL-C, Lp(a), and TG, were measured using standard procedures in Beckman Coulter AU5800. We calculated non-HDL-C by subtracting HDL-C from total cholesterol.[392] The sample size and characteristics for each of the serum lipid biomarkers were presented in Table 1. The distributions of some serum lipid biomarkers were right-skewed (such as Lp(a) and TG, Supplementary Figure S1). We applied a rank-based inverse-normal transformation to the concentration values for each lipid biomarker in order to interpret genetic estimates in standard deviation (SD) units.[267] We computed the phenotypic correlation between lipid biomarkers using the transformed concentration values (Supplementary Figure S2).

6B.2.2 Genetic instruments for serum lipid biomarker

For the GWAS of serum lipid biomarkers, we conducted linear mixed models using BOLT-LMM software (version 2.3).[210] The models were adjusted for sex and age. The first ten principal components were also included as covariates to speed up the convergence of BOLT-LMM's mixed model computations. The genetic instruments for each of the serum lipid biomarkers were selected based on the following criteria: 1) P-value from GWAS < 5×10^{-8}; 2) linkage disequilibrium (LD) between single nucleotide polymorphisms (SNPs) r^2 < 0.001 within a clumping window of 10,000 kilobase;[383] and 3) the SNPs being present in the AMD GWAS summary statistics (described below). We randomly selected 5,000 UK

Biobank white British ancestry individuals as the reference panel.[35] The LD-clumping procedure was performed using PLINK (version 1.9).[182]

6B.2.3 Age-related macular degeneration datasets

The International AMD Genomics Consortium has collected the largest European AMD samples (16,144 advanced AMD cases and 17,832 controls, Supplementary Table S1).[166] The detailed description of the study design, AMD subtype definitions, and genetic data were presented previously.[166] In brief, AMD samples were gathered from 26 studies with each including: 1) intermediate AMD cases with more than five macular drusen greater than 63 μm in diameter or pigmentary changes in the retinal pigment epithelium and age at first diagnosis more than 50 years old; 2) advanced AMD cases with CNV and/or GA in at least one eye and age at first diagnosis more than 50 years old; 3) controls without known advanced or intermediate AMD.[166] The individual level AMD phenotypic and genetic data were obtained from the database of Genotypes and Phenotypes (dbGaP, study accession: phs001039.v1.p1).[166] The genetic imputation was based on the 1000 Genomes Project reference panel (1KGP Phase I, version 3) using Minimac.[166] The SNPs were filtered by imputation quality score (> 0.3) and minor allele frequency (MAF > 0.01) for association analysis. In the association analysis, non-European ancestry participants were removed based on the first two principal components inferred ancestry.[166] For different AMD subtype GWASs, we included 5,336 intermediate AMD cases, 8,544 CNV cases, 2,656 GA cases, 12,711 advanced AMD cases (CNV, GA cases, and 1,511 mixed AMD cases with both CNV and GA), and 14,590 controls. The association analyses were implemented in PLINK software (version v2.00a1LM) adjusting for sex, age, and the first ten principal components.

6B.2.4 Statistical Analysis

The R packages MendelianRandomization and TwoSampleMR were used for MR analyses.[76,385] All general analyses were performed with R (version 3.4.1). We used a two-sided alpha at 0.00625 (0.05/8) to account for the multiple testing of eight lipid-related traits, although given the high genetic correlation between the lipid-related traits the Bonferroni correction can be considered overly conservative.

6B.2.5 Power calculation for MR analysis

We first assessed the power of the MR analyses for different lipid biomarkers with different

AMD subtypes. We calculated the phenotypic variance explained by genetic instruments for each biomarker using the formula (2 × MAF × (1 - MAF) × beta^2)/var(biomarker), where MAF refers to the minor allele frequency, beta is the estimated effect size of each SNP, and var(biomarker) is variance (typically very close to one) after the rank-based inverse-normal transformation.[393] We assumed different effect sizes of lipid biomarkers on AMD risk, and used the mRnd (http://cnsgenomics.com/shiny/mRnd/) method to calculate the power for MR analyses.[378]

6B.2.6 Univariable Mendelian randomization analysis

For the two-sample MR analysis between each lipid biomarker and AMD risk, the univariable inverse-variance weighted (MR-IVW) method was used in the main analysis.[81,379] MR-IVW is a weighted linear regression method to regress the effects of genetic instruments on AMD (outcome) against their effects on lipid biomarker (exposure), with a forced intercept term at zero and weighted by inverse-variance.[81]

6B.2.7 Sensitivity analysis

We then conducted various sensitivity analyses which allow violations of MR assumptions to assess the robustness of MR findings.[381] In particular, the weighted median MR method enables robust inference to be made providing more than 50% of the genetic variants are valid instruments.[381] The MR-Egger method models an intercept term to detect and correct for bias due to directional pleiotropy.[76,381] Although pleiotropy is concerning, if the pleiotropic effects of genetic instruments average to zero (equally to be positive or negative, no directional pleiotropy), the overall estimate would be unbiased.[381] The intercept term from MR-Egger method was used to assess evidence for directional pleiotropy (*i.e.* intercept close to zero and P value > 0.05).[76] We also applied the MR pleiotropy residual sum and outlier (MR-PRESSO) method to evaluate potential bias from outliers and assess the overall heterogeneity of our MR estimates.[77] MR-PRESSO method can identify outlier variants and correct for their effects via outlier removal (MR-PRESSO outlier test). We also implemented a leave-one-chromosome-out analysis by excluding genetic variants in each chromosome out in turn and re-computing the IVW MR estimates, as a means to assess the influence of particular genes from the same chromosome on the overall MR findings.

Bi-directional MR analyses were used to estimate the potential effects of different AMD subtypes on serum lipid biomarker levels. In the reverse-directional analysis, the genetic

instruments for different AMD subtypes were selected via similar criteria as was the case for serum lipid biomarkers as described earlier.

6B.2.8 Multivariable Mendelian Randomization analysis

We performed a regression-based multivariable MR (MVMR) analysis by selecting groups of exposures to avoid collinearity (Figure 1). In the multivariable MR-IVW analysis, the genetic instrumental variables associated with any of the included set of exposures were included.[79,394] The multivariable MR-Egger method is an extension of univariable MR-Egger method to account for multiple lipid biomarkers, and at the same time models an intercept term to correct for both measured and unmeasured pleiotropy.[87]

We also used a recently developed MVMR approach based on Bayesian model averaging (MR-BMA) that scales to high-throughput data to detect true causal risk factors even when the candidate risk factors are highly correlated.[88,395] In the MR-BMA analysis, we included all genetic variants that were genome-wide significant for any lipid biomarker and selected 807 independent genetic variants as instrumental variables. The genetic correlation between lipid biomarkers was computed using the effect sizes of the independent genetic variants. The MR-BMA used a shotgun stochastic search algorithm to evaluate the posterior probability of all combinations of risk factors and then computed for each risk factor its marginal inclusion probability. More details were given in the Supplementary Material.

6B.3 Results

6B.3.1 Serum lipid biomarkers, genetic instruments and statistical power

We included 419,649 participants with at least one lipid biomarker measured in UK Biobank. The proportion of females was 54% and the mean age was 56.83 (SD 8.01) years old (Table 1). We observed a high genetic correlation between ApoA1 and HDL-C, ApoB and LDL-C / CHOL / non-HDL-C levels (maximum genetic association of $|r| < 0.978$, Figure 1 and Supplementary Figure S3). For different serum lipid biomarkers, we identified 64 to 407 genome-wide significant independent variants as genetic instruments, and they collectively explained 9% to 16% of the phenotypic variance (Table 1). We calculated the MR analysis statistical power for different AMD subtypes; even with 9% variance explained, our power for intermediate, advanced AMD, GA, and CNV AMD subtypes was 95%, 100%, 79%, 98%, respectively, assuming moderate effect sizes (*e.g.* odds ratio [OR] 1.2, Supplementary Table

S2).

Table 1. Serum lipid biomarkers in UK Biobank.

Variables	N	Mean (SD)	Median (IQR)	N SNPs[2]	Variance explained
Sex	419649	227003 (54%)[1]	-	-	-
Age, years	419649	56.83 (8.01)	58 (51 to 63)	-	-
Apolipoprotein A1 (ApoA1), g/L	382867	1.54 (0.27)	1.52 (1.35 to 1.7)	407	0.11
Apolipoprotein B (ApoB), g/L	417522	1.03 (0.24)	1.02 (0.87 to 1.18)	241	0.13
Cholesterol (CHOL), mmol/L	419516	5.71 (1.14)	5.67 (4.93 to 6.44)	231	0.09
HDL cholesterol (HDL-C), mmol/L	384986	1.45 (0.38)	1.4 (1.18 to 1.68)	488	0.12
LDL direct (LDL-C), mmol/L	418780	3.57 (0.87)	3.53 (2.96 to 4.13)	215	0.10
Lipoprotein A (Lp(a)), nmol/L	334646	44.12 (49.49)	20.11 (9.33 to 60.1)	64	0.16
Triglycerides (TG), mmol/L	419185	1.76 (1.02)	1.49 (1.06 to 2.16)	394	0.11
non-HDL-C, mmol/L	384915	4.26 (1.08)	4.2 (3.49 to 4.94)	207	0.09

[1] The frequency and percentage of females are presented.

[2] Number of genetic instruments.

IQR, interquartile range; N, sample size; SD, standard deviation.

The biochemistry markers are described on http://biobank.ctsu.ox.ac.uk/crystal/label.cgi?id=17518.

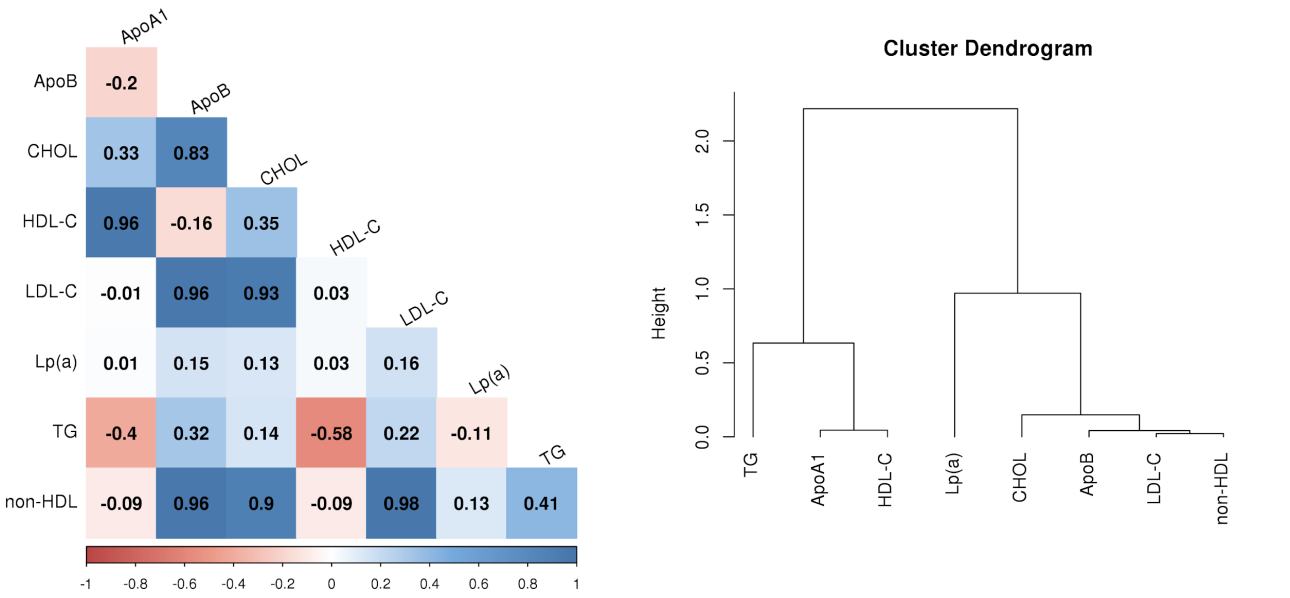


Figure 1. The genetic correlation and cluster of eight serum lipid biomarkers.

The left panel displays genetic correlation between each pair of serum lipid biomarkers based on N = 807 independent genetic variants that were genome-wide significant for any lipid biomarker. The right panel shows the cluster of the eight serum lipid biomarkers.

6B.3.2 The associations between eight serum lipid biomarkers and different AMD subtypes

In the univariable MR analysis, one SD higher HDL-C levels increased the risk of advanced AMD by 19% (MR-IVW OR 1.19, 95% CI 1.07 - 1.33, P value 1.2×10^{-3}). The association was consistent across different AMD subtypes and across different MR methods (weighted median, MR-Egger, Figure 2 and Supplementary Table S3). As expected given the high correlation with HDL-C, higher ApoA1 levels were also associated with increased risk of different AMD subtypes.

Raised LDL-C levels were nominally associated with decreased risk of advanced AMD (OR 0.87, 95% CI 0.76 - 1.00, P value 0.04). However, when split by AMD subtype, the association was primarily with GA (OR 0.70, 95% CI 0.59 - 0.83, P value 3.8×10^{-5}) and intermediate AMD (OR 0.77, 95% CI 0.67 - 0.87, P value 6.5×10^{-5}); there was no strong evidence for association with CNV AMD (OR 0.93, 95% CI 0.80 - 1.08, P value 0.34). Similarly, for the correlated traits ApoB, CHOL and non-HDL-C, all were not associated with CNV AMD, but were associated with GA and intermediate AMD.

Higher levels of TG were associated with decreased risk of different AMD subtypes, and the estimates were broadly consistent across different AMD subtypes (intermediate AMD OR 0.74, 95% CI 0.66 - 0.83, P value 2.5×10^{-7}; advanced AMD OR 0.81, 95% CI 0.72 - 0.90, P value 1.4×10^{-4}). Lp(a) levels were not associated with any of the AMD subtypes (intermediate AMD OR 0.96, 95% CI 0.85 - 1.09, P value 0.53; advanced AMD OR 1.00, 95% CI 0.89 - 1.12, P value 0.94) even though the variance explained by the genetic instruments for Lp(a) was higher than any other lipid biomarkers.

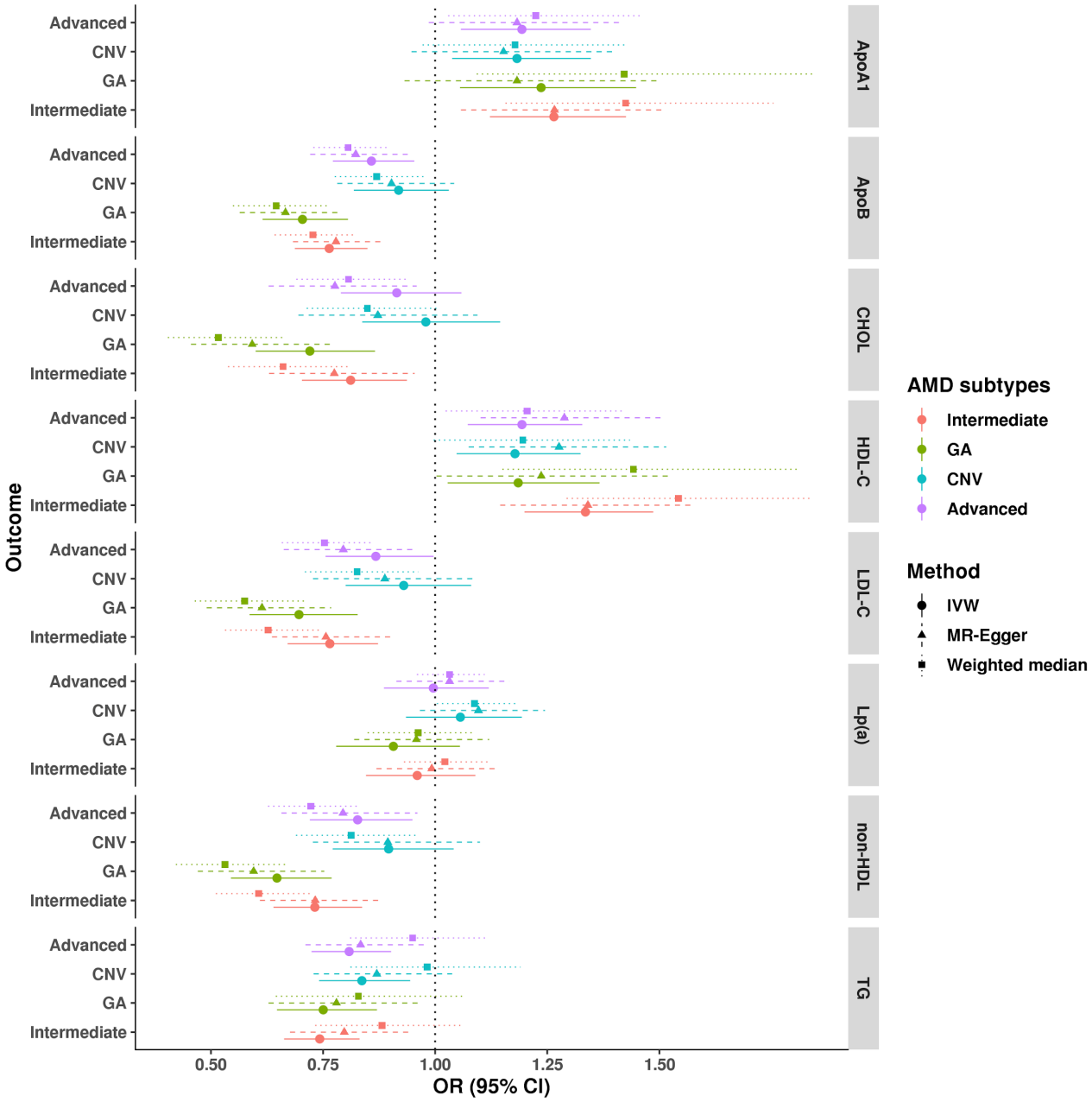


Figure 2. Univariable Mendelian randomization estimates of the associations between eight serum lipid biomarkers and different age-related macular degeneration

subtypes.

The x-axis is the odds ratio (OR) of the effects of lipid biomarkers on age-related macular degeneration (AMD) subtypes. The vertical dashed line is the reference at OR=1. The y-axis presents different AMD subtypes, highlighted in different colours. Different Mendelian randomization methods are displayed in different line types (MR-IVW, solid line; MR-Egger, dashed line; Weighted median, dotted line).

6B.3.3 Sensitivity analysis

We applied MR median-weighted and MR-Egger methods to validate the MR-IVW estimates (Figure 2, Supplementary Table S3); their estimates were broadly consistent with the MR-IVW method with overlapping confidence intervals. The MR-Egger intercepts showed no evidence of directional pleiotropy effects (intercepts were approximately 0, P > 0.05).

We conducted MR-PRESSO outlier-corrected tests, and found that most of the MR analyses were not meaningfully changed after removing outlier variants except the effects of HDL-C, ApoA1 and TG on CNV AMD risk (Supplementary Figure S4 and Supplementary Table S4). The removed outlier SNPs were mainly from genes *CETP*, *LIPC*, *APOE*, and *ABCA1*. Given the strong associations between variants in these genes and both lipid biomarkers and AMD risk, removing these variants would affect the estimated effect sizes in MR analyses.[166,396]

To further investigate the robustness of MR results, we applied a leave-one-chromosome-out analysis by leaving genetic variants in each chromosome out in turn for the MR analyses (Supplementary Figure S5). We found a striking difference in the results for Lp(a) depending on chromosome 6. However, most of the variance in Lp(a) is controlled by variants in *LPA* (96.9%, in chr6). We found no association between the SNP rs10455872 (the top SNP in *LPA* region associated with Lp(a) levels) and AMD risk (OR = 1.03, P = 0.40 for advanced AMD; OR = 0.97, P = 0.61 for GA AMD).

We found weak evidence of liability towards AMD on lipid traits via reverse-direction MR analyses (Supplementary Figure S6). To investigate the influence of lipid-related drugs on our MR results, we identified 87,904 participants taking statins (data coding C10AA) in UK Biobank.[397] We also found 6,030 participants with self-reported or medical electronic health records of macular degeneration. We removed both statin users and AMD cases in UK Biobank to re-select the genetic instruments for serum lipid biomarkers from GWAS. The MR results were unchanged (Supplementary Figure S7).

6B.3.4 Multivariable Mendelian randomization

We conducted multivariable MR analyses (MVMR-IVW method) to estimate the direct effects of serum lipid biomarkers on AMD risk conditional on other serum lipid biomarkers. We selected groups of exposures to avoid collinearity. In the classic trio (HDL-C, LDL-C and TG), we included N = 700 independent SNPs associated with any of the three biomarkers as instrumental variables. The associations of HDL-C and LDL-C with AMD risk were essentially unchanged in multivariable MR analyses compared with univariable MR analysis (first column in Figure 3). We further replaced HDL-C with ApoA1 in the trio (that is ApoA1, LDL-C and TG, second column in Figure 3), the results were similar to the trio HDL-C, LDL-C and TG. The MVMR results for Lp(a), CHOL and non-HDL-C were similar to univariable MR results (columns 3, 4, and 5 in Figure 3). The multivariable MR-Egger intercepts showed no evidence of directional pleiotropy effects (Supplementary Table S5).

We conducted multivariable MR-BMA analyses to select causal serum lipid biomarkers. When the prior probability was set at 0.125 or 0.25 (corresponding to a priori of one or two expected causal biomarkers), we found ApoA1 has relatively higher probabilities and causal effects for all AMD subtypes, and TG has the highest probability to be the causal risk factor for intermediate AMD (Supplementary Figure S8 and Supplementary Material).

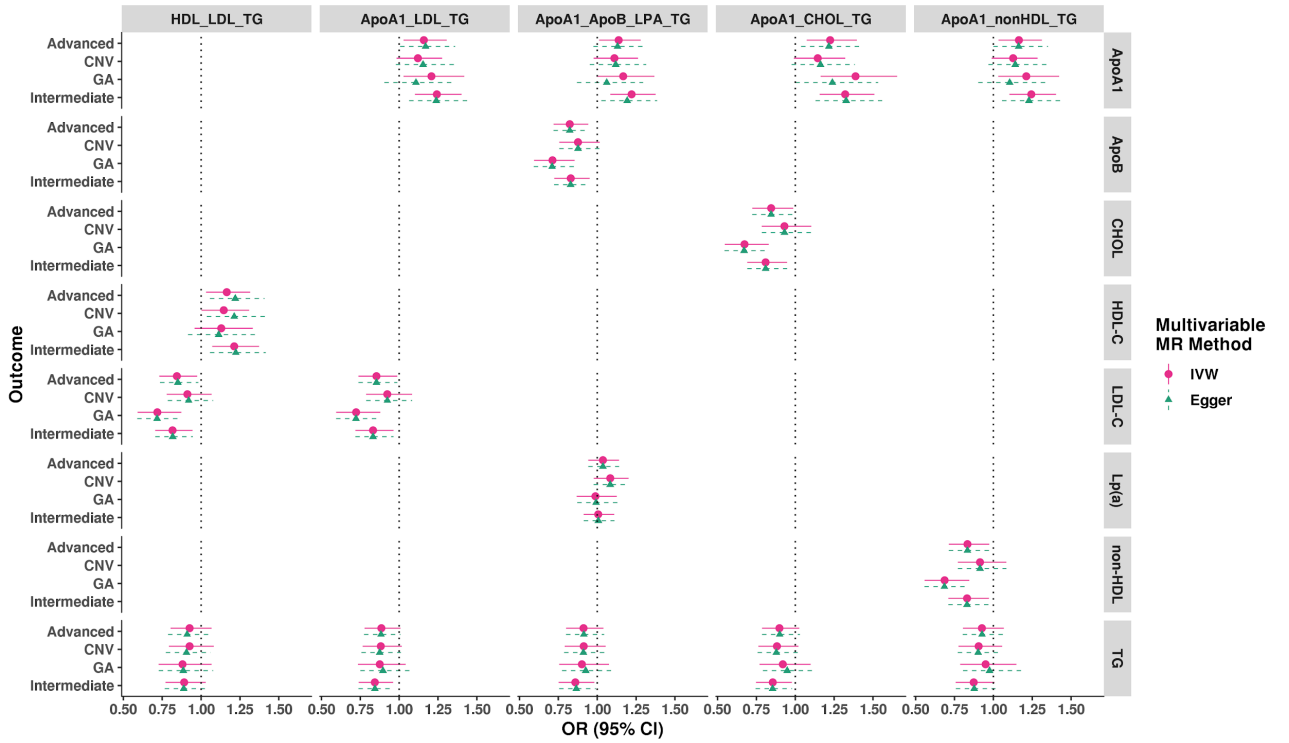


Figure 3. Multivariable Mendelian randomization estimates of the associations between eight serum lipid biomarkers and different age-related macular degeneration subtypes.

The x-axis is the estimated odds ratio (OR) on AMD subtypes per standard deviation (SD) increase in lipid concentration levels for each lipid biomarker evaluated. The vertical dashed line is the reference at OR=1. The y-axis lists the different AMD subtypes. The multivariable IVW estimates are shown in solid line, while the multivariable estimates adjusted for the MR-Egger intercept are shown in dashed line. Each column facet indicates the selected group of exposures in multivariable MR analysis, where all independent SNPs associated with any of the included exposures were fitted.

6B.4 Discussion

We systematically evaluated the effects of eight serum lipid biomarkers on the risk of different AMD subtypes. We found that higher HDL-C and ApoA1 levels increased the risk of all AMD subtypes, whereas LDL-C, ApoB, CHOL, and non-HDL-C levels appeared to be only associated with decreased risk of intermediate and GA AMD. TG levels were associated with decreased risk of different AMD subtypes. The role of lipids on cardiovascular diseases risk is well studied. Compared with cardiovascular disease risk, most of these serum lipid biomarkers showed the opposite direction effects on AMD risk.[398]

These findings suggest varying roles of lipids in different AMD stages and subtypes.

Previous observational studies have suggested a potential relationship between lipid biomarkers and AMD risk; however, the results were inconsistent.[145,146,150] We found that genetically elevated HDL-C levels increased the risk of AMD, consistent with findings from previous observational and MR studies.[146–148,150] Typically, HDL-C can mediate reverse cholesterol transport and have atheroprotective functions, such as anti-inflammatory, antioxidant, and endothelial cell maintenance.[399] However, dysfunctionally elevated HDL-C could have pro-inflammatory and pro-oxidant roles that impair cholesterol efflux and promote the accumulation of drusen.[400,401] Our results indicate that the effect of HDL-C levels on intermediate AMD (OR 1.34, 95% CI 1.20 - 1.49) appeared larger than advanced AMD, which was also highlighted in a recent observational study,[150] where the effect sizes of estimates were broadly consistent with observational studies. We do, however, find evidence that the effect predicted by these HDL-C genetic instruments are rather heterogeneous. For instance, removing genetic instruments from the gene *CETP* (chromosome 16) attenuated the effect of HDL-C on AMD risk towards the null (shown evidently for the CNV subtype); while excluding variants from the gene *LIPC* (chromosome 15) amplified the association (Supplementary Figure S4 and S5). These results suggest that serum HDL-C risk variants in *CETP* and *LIPC* might have counteracting effects on AMD risk, as discussed in previous literature.[147,150,402] We speculate that HDL-C related genes may affect AMD risk via different pathways. As the major apolipoproteins in HDL-C particles (genetic correlation 0.96), higher ApoA1 levels also increase the risk of AMD. In our MR-BMA analysis, serum ApoA1 levels have relatively higher probabilities and effects for AMD compared with other lipid biomarkers. A recent study also showed that extra-large and large HDL particles are putative risk factors for AMD.[88]

The relationship between LDL-C and AMD risk has proved controversial in previous observational and genetic studies. For instance, a meta-analysis study showed a protective tendency between LDL-C levels and AMD risk.[146] Further stratified analysis based on AMD subtypes revealed a protective effect on early stage, but not on late stage. A recent large-scale epidemiologic study also indicated that LDL-C levels were only associated with early AMD.[150] Previous MR studies, by contrast, showed no evidence of association between LDL-C levels and advanced AMD risk.[147,148] In this study, we observed a nominal association between higher LDL-C levels and decreased advanced AMD risk. Importantly, LDL-C levels

exhibit a clear protective effect on intermediate and GA AMD subtypes. The nominal association between LDL-C and advanced AMD was likely driven by GA AMD subtype even though only a smaller proportion of advanced AMD cases were GA in the data sets. For ApoB, CHOL, and non-HDL-C, all of them were associated with intermediate and GA AMD, but not CNV AMD. Previous observational studies have also shown that drusen are more likely to be involved in the development of GA AMD rather than CNV AMD.[403,404] Since drusen are a major pathological hallmark of early and intermediate stage AMD, these results suggest that LDL-C and ApoB may be involved in the formation of drusen in the early and intermediate stages of AMD, and the developing of GA AMD;[405] in contrast, their roles in CNV AMD appear limited.

Previous observational studies have shown that higher TG levels reduce the risk of early stage AMD but not late stage.[146,150] In our univariable MR analysis, raised TG levels were associated with decreased risk of different AMD subtypes; however, the effect size on CNV AMD subtype was smaller and was not that robust based on MR-PRESSO outlier-corrected tests. In this study, we find no evidence of the association between Lp(a) and AMD risk. Serum Lp(a) levels are mainly genetically determined by the genetic variations in *LPA* gene region,[406,407] none of which showed association with AMD risk. We found a SNP rs7412 in *APOE* that is both associated with Lp(a) concentrations and AMD risk. However, apoE proteins are thought to influence Lp(a) catabolism through lipoprotein receptor clearance pathways such as LDL receptor (maintains the plasma levels of LDL) rather than directly affect Lp(a) assembly or secretion.[408]

These findings aid us in the understanding of lipid metabolism in drusen formation and AMD development, as well as the clinical implications of modifying blood lipid concentrations in preventing AMD. The clinical hallmark of early stage AMD is the presence of drusen, with approximately 40% of druse content comprised of lipids. Lipids may be involved in the initiation and formation of drusen in the early and intermediate stages. This is supported by the associations between HDL-C / LDL-C / TG and intermediate AMD. Both CNV and GA AMD are subtypes of advanced AMD, the late stage of AMD that could cause vision loss. This study shows that LDL-C and TG are associated with GA AMD, and their roles in CNV AMD appear limited, suggesting different pathogenesis pathways for GA and CNV AMD subtypes. Currently, there is no effective medications for GA subtype, and the anti-vascular endothelial growth factor therapies for CNV are also not curative.[137] These Mendelian

randomization findings suggest the potential utility of lipid modifying therapies in AMD treatment, and shed light on the different roles of lipid subfractions on different AMD subtypes (Figure 2). A recent study also showed that high-dose statins may have a particular role in large drusenoid deposits AMD patients, and result in regression of large drusen and improvement of visual acuity.[409] Further clinical trials are warranted to investigate different lipid-modifying drugs in specific AMD subtypes rather than a broad range of AMD subtypes.

A strength of this study is that we used large-scale data sets with standard protocols to measure lipid biomarkers; this allowed us to systematically evaluate the effects of lipids on AMD risk. Compared to traditional observational studies, MR findings are less likely to be affected by confounding or bias from reverse causation. To the best of our knowledge to date, this is the first study to have comprehensively evaluated the causal relationships between lipid / lipoprotein biomarkers and different AMD stages and subtypes through a MR framework. In particular, unlike some previous studies we have considered a wide range of lipids and lipoproteins. Dyslipidemia has been involved in the formation of drusen, which are characterized in the early stage of AMD. This study based on different AMD stages and subtypes provides new insights for the role of lipids in AMD risk and development. At the same time, our results should be interpreted in light of its limitations. Firstly, this study is based on European ancestry participants, the generalizability of our findings in other ethnic groups needs further investigation. Moreover, in MR framework, the genetically predisposed biomarker changes are assumed to have a linear and lifetime effect on AMD risk. The potentially non-linear relationships and short-term effects of these biomarkers are unclear. This study indicates the role of circulating lipids on AMD risk; further studies are needed to investigate the effects of retina-specific lipid metabolism on AMD risk. Finally, in this study we used publicly available AMD samples and were unable to assess the potential selection bias due to competing risk, such as coronary artery disease (CAD). We performed an exploratory analysis computing the genetic correlation between CAD and AMD and found the correlations were close to zero (data not shown). We also conducted a MVMR analysis of CAD, LDL-C and TG (as three exposures) on AMD risk, and found no evidence of association between CAD and AMD risk (Supplementary Figure S9), the MVMR results for LDL-C were essentially unchanged relative to the previous MVMR results in Figure 3, suggesting that broadly speaking our typically elderly AMD samples were not enriched for cardioprotective genetic factors; these results suggest our MR findings are unlikely to be driven by competing risk of conditions with shared etiology.[410]

Conclusion

This study provides genetic evidence that elevated circulating HDL-C and ApoA1 levels increase the risk of all AMD subtypes, whereas LDL-C, ApoB, CHOL, and non-HDL-C levels are particularly associated with decreased risk of intermediate and GA AMD. The inconsistent results from previous studies could be partly explained by the large heterogeneity of AMD disease (different stages and subtypes) in these studies. This study provides new insights into the pathogenesis of AMD. Further studies are warranted to investigate the role of lipid metabolism in drusen formation and AMD development in the early and intermediate stages, and the utility of lipid pathways for therapeutic treatment in preventing AMD.

6B.5 Acknowledgements

The authors thank all of the participants who took part in the UK Biobank and the International AMD Genomics Consortium and support staff who made this study possible. We thank Scott Wood, Xiaping Lin, and John Pearson from QIMR Berghofer for IT support.

Funding

This work was conducted using the UK Biobank Resource (application number 25331). For the AMD datasets, all contributing sites and additional funding information are acknowledged in this publication: Fritsche et al. (2016) Nature Genetics 48 134–143, (doi:10.1038/ng.3448); The International AMD Genomics consortium's web page is: http://eaglep.case.edu/iamdgc_web/, and additional information is available on: http://csg.sph.umich.edu/abecasis/public/amd2015/.

The AMD case-control datasets used for the analyses described in this manuscript were obtained from the NEI Study of Age-Related Macular Degeneration (NEI-AMD) Database found at https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001039.v1.p1 through dbGaP accession number 20740. Funding support for NEI-AMD was provided by the National Eye Institute. We would like to thank NEI-AMD participants and the NEI-AMD Research Group for their valuable contribution to this research.

XH is supported by the University of Queensland Research Training Scholarship and QIMR Berghofer PhD Top Up Scholarship. SM and AWH are supported by Australian National Health and Medical Research Council (NHMRC) Fellowships. We acknowledge funding from NHMRC grants [1116360, 1150144 and 1123248].

Data Availability Statement

UK Biobank data are available through the UK Biobank Access Management System https://www.ukbiobank.ac.uk/. The International Age-related Macular Degeneration Genomics Consortium data are available from the database of Genotypes and Phenotypes (dbGaP, study accession: phs001039.v1.p1). Applications are assessed for meeting the required criteria for access.

Competing Interests

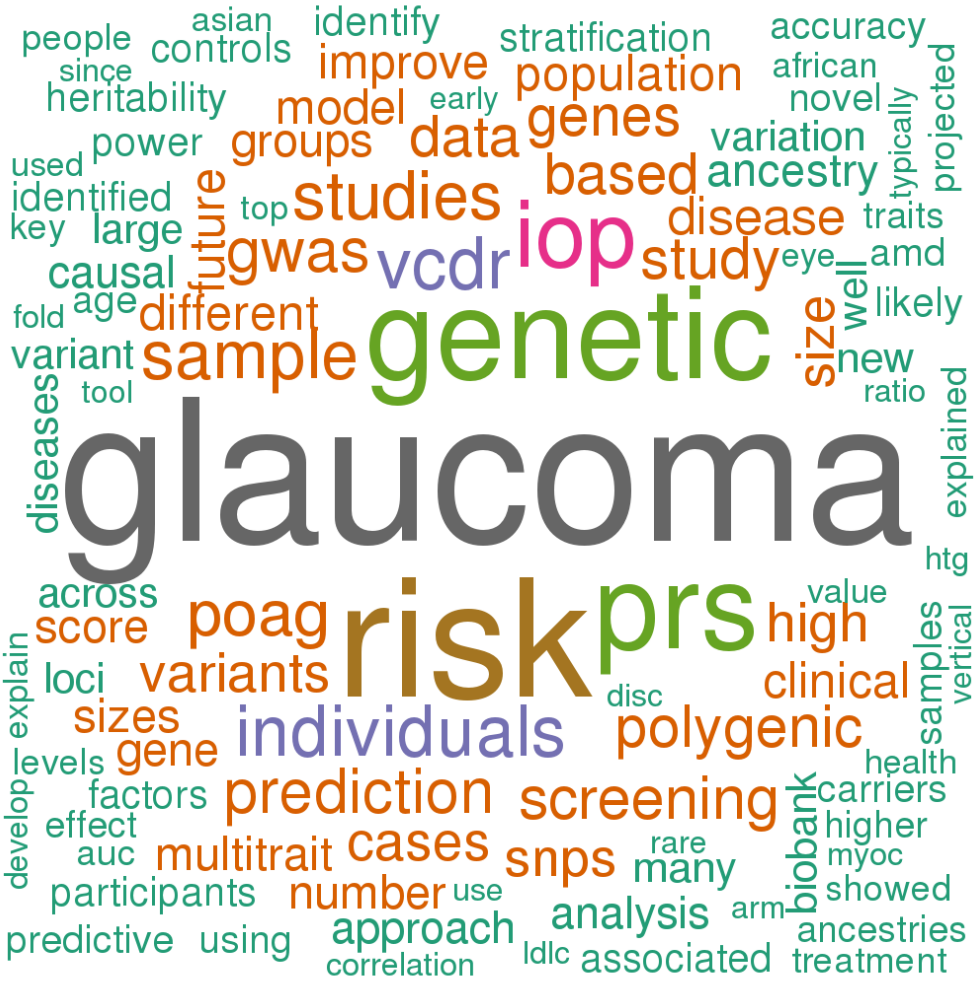
The authors have declared that no competing interests exist.

6B.6 Supplement

Supplementary Tables and Figures are available at:

https://doi.org/10.1093/ije/dyaa178

Chapter 7. Discussion



CHAPTER 7A

Predicting the future of genetic risk profiling of glaucoma: a narrative review

Xikun Han, Alex W Hewitt, Stuart MacGregor.

JAMA Ophthalmology. 2021;139(2):224-231.

Contribution of candidate:

In this study, I contributed to study design, data analysis, and the first draft of the manuscript. Stuart MacGregor contributed to study design and the first draft of the manuscript. Alex W Hewitt contributed to interpretation of the results and revision of the manuscript.

Chapter 7A. Predicting the Future of Predicting the Future: Pressing Questions in the Genetic Risk Profiling of Glaucoma

Glaucoma is the world's leading cause of irreversible blindness. Primary open angle glaucoma (POAG) is typically asymptomatic early in the disease process, and unfortunately many are diagnosed too late to prevent vision loss. Genome wide association studies, which evaluate the association between genetic variants and phenotype across the genome, have mapped many genes for POAG. As well as uncovering new biology, genetic information can be combined into a polygenic risk score (PRS), which aggregates an individual's disease risk over many genetic variants. In this non-systematic review performed from June 21 2019 to October 1 2020, we address a series of questions to explain the challenges and opportunities in translating recent genetic discoveries in POAG. We summarize what is known about POAG genetics and how its endophenotypes, such as intraocular pressure or cup-disc ratio, can help with prediction. We discuss the sample sizes available, and how increases in the future may have an effect on the utility of prediction approaches. We explore particular scenarios such as the use of PRS in risk stratification and applications for individuals who are particularly high-risk for POAG as a result of them carrying both a high penetrance mutation and an unfavourable PRS. Finally, we discuss the issue of equity in applying these tests and the prospects for prediction for people from various ancestry groups. The cost-effectiveness evaluation of glaucoma PRS in direct-to-consumer genetic testing and across different ancestry groups is warranted in future research. In conclusion, advances in glaucoma genetics have opened the door for risk stratification based on genetic risk predictions. The PRS approach has shown good promise in predicting who will be at highest risk of POAG, which could improve outcomes if these predictions can be acted upon to result in improved clinical outcomes.

7A.1 Introduction

Glaucoma, the world's leading cause of irreversible blindness, is a heterogeneous group of diseases characterized by progressive degeneration of retinal ganglion cells (RGC), thinning of the retinal nerve fiber layer (RNFL), and excavation of the optic disc.[90,91,411] This article will focus on the most common form of glaucoma, primary open angle glaucoma (POAG).[92,93] The global prevalence of glaucoma in the population 40 years or older is 3.54%, and the prevalence of POAG is approximately 3.05%.[94,95] The prevalence of glaucoma varies across the world and is highest in Africa ancestries (4.20%)[92,93]. POAG accounts for the vast majority of glaucoma cases of African and European ancestry and approximately half of Asians with the disease.[94,186]

The biological mechanisms underlying POAG are not well understood and the risk factors contributing to its progression have not been fully characterized.[411] Like other complex (multifactorial) diseases, both genetic and environmental factors play an important role in the development and progression of POAG.[103,104] Elevated intraocular pressure (IOP) is currently the sole modifiable risk factor for POAG. Given higher IOP confers greater risk for POAG, high-tension glaucoma (HTG) is a commonly used subcategory; HTG is typically defined as IOP > 21 mmHg, although the specific threshold is somewhat arbitrary.[186] POAG can develop and progress despite a IOP recordings in the "normal" range, normal-tension glaucoma (NTG).[412,413] Conversely, not all people with elevated IOP develop POAG. Apart from IOP, vertical cup-to-disc ratio (VCDR) is another key endophenotype of POAG. Larger VCDR, a sign of glaucomatous optic cupping and visual field loss, is generally used to define POAG in population-based prevalence surveys.[93]

Genetic factors play an important role in glaucoma.[103,104] During the past few decades, genetic linkage analysis has identified genes such as myocilin (*MYOC*), *OPTN* and *TBK1*.[107–109] Pathogenic variants in *MYOC* account for approximately 2-4% of POAG cases.[109,110,116] The p.Gln368Ter (rs74315329) variant is the most common *MYOC* variant amongst populations of European ancestry.[107,111,112] *MYOC* p.Gln368Ter carriers are generally diagnosed earlier than other cases and have elevated IOP.[43,113,114] *OPTN* or *TBK1* variant carriers typically manifest with NTG.[414,415]

The pace of gene discoveries for glaucoma accelerated during the past decade via genome-

wide association studies (GWAS), a design to detect associations between single nucleotide polymorphisms (SNPs) and complex traits genome-wide rather than via a gene-by-gene candidate approach.[6,8] Investigations into the genetics of POAG will improve our understanding of the allelic architecture, aid in molecular fine-mapping, and improve risk prediction and genetic screening for POAG.

Polygenic risk scores (PRS), also known as genetic risk scores or allele scores, are profiles based on aggregating multiple risk alleles and their effect sizes.[24,25] Complex traits and diseases, such as glaucoma, typically have a polygenic basis.[20,21] While the biology mechanisms of the discovered genes are largely unknown, this does not preclude their use in prediction. Previous studies have shown that using genome-wide markers can improve predictions,[27] and PRS are a promising tool for risk stratification, genetic screening, and the development of risk management strategies.[29–32,34] An illustrative schematic diagram to identify individuals at high-risk using polygenic risk score is shown in Figure 1. In the review, we address a series of pertinent questions, providing an overview of recent advances of genetics in our understanding of both the risk factors for glaucoma (IOP, VCDR), as well as the disease itself. We also discuss what the prospects are for improving upon recently reported glaucoma genetic risk predictions.[35]

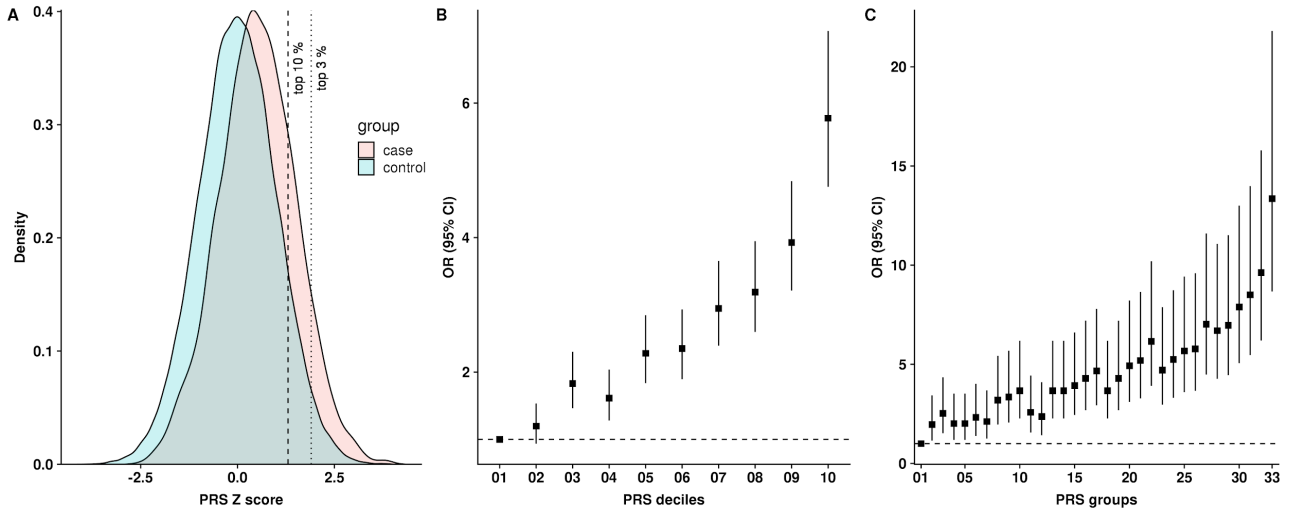


Figure 1. Illustrative diagram of identifying individuals at high-risk using polygenic risk score.

A simulation dataset was created with 100,000 individuals, with 3% of them having glaucoma (N = 3,000, in line with the glaucoma prevalence). A standardized glaucoma polygenic risk score was simulated for both

glaucoma cases and controls, and the PRS for glaucoma cases are on average 0.5 standard deviation higher than controls. Panel A shows the simulated density distribution of PRS for glaucoma cases and controls. The dashed vertical line is the cut-off point for individuals at the top 10% PRS. The dotted vertical line is the cut-off point for individuals at the top 3% PRS. These cut-off points could be potential thresholds to define individuals at high risk. Panel B shows the odds ratio for PRS split into 10 groups, with the first group set as the baseline. Panel C is similar to B, but with 33 groups. These figures give an illustration of how individuals can be stratified into high-risk or low-risk groups using a PRS.

7A.2 What is known about the genetics of glaucoma and its endophenotypes IOP and VCDR?

Studies have provided evidence for the importance of a genetic component in glaucoma. In the general population, participants with a first degree relative with glaucoma are at almost 10 times higher risk of glaucoma.[105,416,417] Heritability is a population parameter to describe the relative proportion of genetic and environmental factors in trait variation.[418] A recent large-scale study, using reconstructed family data, estimated the heritability of glaucoma to be 0.7.[106] The availability of large biobanks, such as UK Biobank (UKB), has dramatically accelerated the gene discoveries for glaucoma.[125,126,130] Nearly 100 genes are associated with POAG.[35,118–126] However, these genes only account for a small fraction of the disease heritability[35,125,126] and larger studies are warranted.

IOP and VCDR are key endophenotypes of glaucoma. Twin studies have estimated the heritability of IOP to range from 0.35 to 0.67.[193] Subsequent GWASs allowed estimation of array-based heritability - this measures the degree to which common variants on genotyping arrays explain trait variation. Since only common (and not rare) variants are included, the array-based heritability provides a lower bound on the overall heritability. The array-based heritability for IOP has been estimated to be 16% in UK Biobank participants.[126] However, the true value is likely higher, given there is substantial measurement error if only one IOP measurement is taken (*e.g.* the left eye IOP only explains 40% of the variance in right eye IOP in UK Biobank, with much of the remaining 60% likely due to measurement error). Recent gene discovery efforts using GWAS have identified more than 100 genes associated with IOP levels.[125,126,131,133] Collectively, these IOP genes explained 9% to 17% (variation is due chiefly to measurement error in different studies and to age specific effects) of the variance of IOP levels.[125]

For VCDR, a previous study from the International Glaucoma Genetic Consortium (IGGC) identified nearly 30 loci associated VCDR, with a SNP-based heritability estimate of 0.31.[133] Our recent study in UKB tripled the sample size and identified 76 independent SNPs, explaining 6% of the variance of VCDR.[35]

7A.3 To what extent will glaucoma endophenotypes improve risk prediction for glaucoma?

Our recent study has demonstrated that IOP and glaucoma have a large shared genetic component, with a genetic correlation of 0.71.[126] We also found a strong genetic correlation between VCDR and glaucoma (genetic correlation 0.5).[35] Leveraging the high genetic correlation between glaucoma and its endophenotypes, GWASs of IOP and VCDR can uncover novel glaucoma genes and pathways, and improve the prediction of POAG.[419,420] A previous study found 101 genome-wide significant IOP SNPs, 53 of which affected glaucoma.[126]

Recent studies have shown that multi-trait GWAS, a generalized meta-analysis method to incorporate genetic correlated traits, can improve power for identifying novel genes and improve the accuracy of genetic risk prediction.[53] With the high genetic correlation between POAG and its endophenotypes (IOP and VCDR), the multi-trait GWAS method boosts power to uncover POAG genes and improve genetic predictions. Our study modeling glaucoma and IOP/VCDR data in a multi-trait GWAS approach increased the effective sample size for glaucoma 2.6 fold, and doubled the variance explained (variance explained 6% by UKB glaucoma alone to 13% by multi-trait GWAS approach).[35] This multi-trait approach combined ~8000 glaucoma cases, ~119,000 controls, ~130,000 individuals with IOP measurements and ~100,000 individuals with VCDR measurements. Assuming the contributions of IOP and VCDR contribute to the effective sample size in proportion to the estimated genetic correlation with POAG (genetic correlations 0.7 and 0.5, respectively), we estimate that approximately 4 IOP samples or 7 VCDR samples contribute the same power as one sample in glaucoma GWAS (assuming a 1:1 ratio of case and control). For example, 100 glaucoma cases plus 100 controls have equivalent power to 800 individuals with IOP measured or 1,400 individuals with VCDR measured. Since glaucoma is relatively rare in the general population, biobanks will contribute more to glaucoma gene mapping efforts if

they have IOP or VCDR measured on their (largely glaucoma free) participants, than if such biobanks merely identify glaucoma cases/controls. Naturally, if both case-control and endophenotype data are available for use in a multi-trait model, this will maximize power.

7A.4 How many glaucoma samples are required for "good" prediction of risk?

Leveraging large datasets of glaucoma, IOP and VCDR, our recent study has shown that a PRS derived from multi-trait analysis provided additional predictive ability beyond traditional glaucoma risk factors, with a significant change in the AUC (from 0.73 to 0.80). In the general population, participants in the top PRS decile reach an absolute risk (3%) for glaucoma 10 years earlier than the bottom decile and are at 15-fold higher risk of developing advanced glaucoma. These findings demonstrate the prospect of PRS in identifying individuals in high risk groups, which could be an effective tool for risk stratification.

To predict what is expected in the future from GWAS on glaucoma and its endophenotypes given larger sample sizes, we applied a novel statistical method, "GENetic Effect-Size distribution Inference from Summary-level data" (GENESIS), to model the effect size distribution of common variants, characterize the polygenic architecture of the traits, and project the likely improvements in variance explained by future GWASs.[40] The detailed descriptions of the modelling data and methods are in the **supplement**. Based on the modelling, with a sample size of 40,000 (equivalent to 20,000 cases, 20,000 controls), the projected number of underlying susceptibility SNPs is 27, which are predicted to explain 15% of glaucoma phenotypic variation. Doubling the sample size to 80,000 (equivalent to 40,000 cases, 40,000 controls) is predicted to identify 90 susceptibility SNPs and explain 23% of glaucoma variation. From the GENESIS analysis, the predicted best AUC for the PRS alone is 0.59, 0.62, 0.67 for sample sizes 20,000, 40,0000, and 80,000, respectively.
We then projected the polygenic architecture of glaucoma endophenotypes (IOP and VCDR) using GENESIS. For IOP, with extant sample sizes of ~100,000, the projected number of underlying susceptibility SNPs is 67, which explains 3.5% of IOP variation. When doubling the sample size to 200,000, the projected number of underlying susceptibility SNPs is 200, which are predicted to explain 5.5% of IOP variation. Quadrupling the IOP GWAS sample size to 400,000 would identify approximately 655 susceptibility SNPs and explain 9.3% of

IOP variation. The explained variance of IOP measurements would depend on factors, such as diurnal variation, age and measurement errors.

To characterize the polygenic architecture of VCDR, we applied GENESIS to UKB VCDR GWAS summary statistics with a sample size of ~67,000. The projected number of underlying susceptibility SNPs is 64, which explains 5.5% of VCDR variation. When the sample size is 100,000, the projected number of underlying susceptibility SNPs is 101, which explains 6.5% of VCDR variation. A VCDR GWAS of 200,000 samples would identify 272 susceptibility SNPs and explain 9.2% of VCDR variation. For both IOP and VCDR, by combining these traits with glaucoma in a multitrait model, there is likely to be excellent scope to reveal novel glaucoma genes and to improve glaucoma risk predictions.

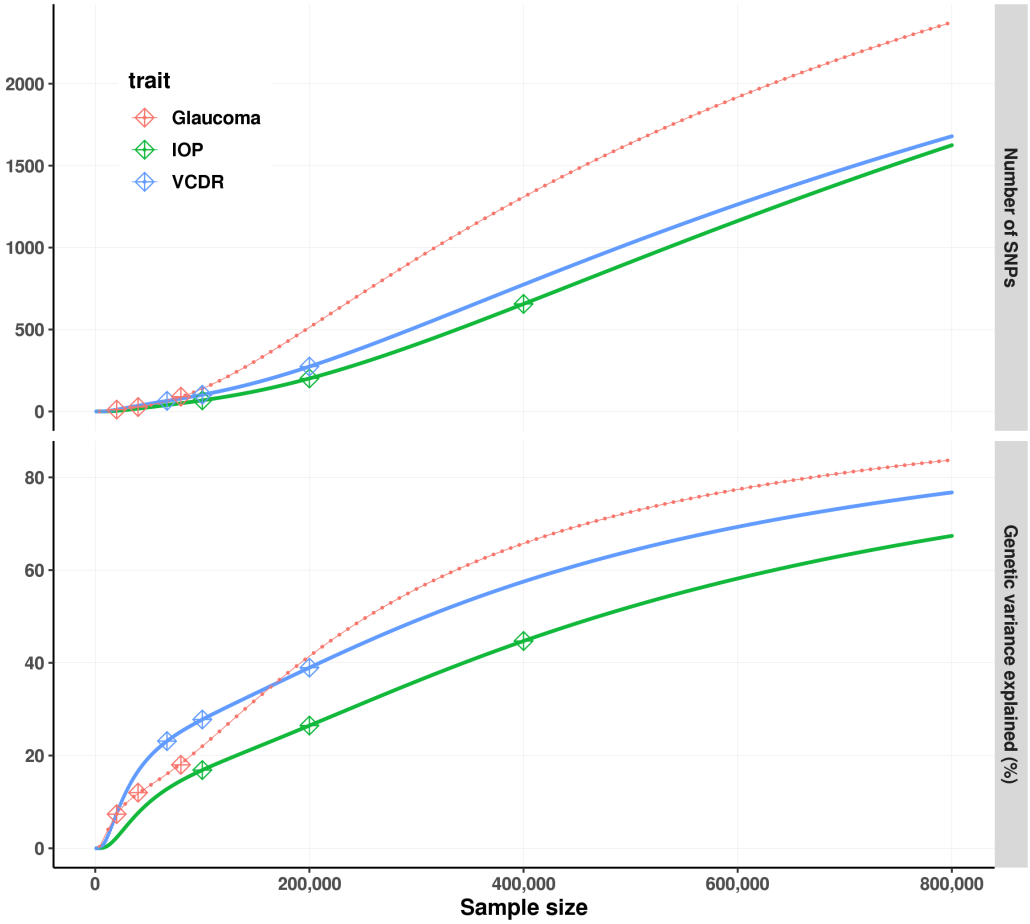


Figure 2. The projection of the number of discovered SNPs and genetic variance explained for glaucoma, IOP, and VCDR.

The X-axis is the sample size of GWAS summary statistics. For glaucoma, the sample size equals the total number of N cases and N controls, assuming a 1:1 ratio. Diamond plus symbols show the projection at different sample sizes (roughly current sample size, double, and quadruple). In Panel A, the Y-axis is the projected

number of independent SNVs. In panel B, the Y-axis is the genetic variance explained (%), which is equal to phenotypic variance explained multiply by heritability.

7A.5 What are the prospects for larger sample sizes? What is the limit in terms of improvement?

Sample sizes for glaucoma GWAS have steadily increased over the last decade, culminating in the International Glaucoma Genetics Consortium (IGGC) glaucoma meta-analysis.[312] The IGGC meta-analysis comprised 34,179 glaucoma cases and 349,321 controls. The primary determinant of power to identify new loci is the number of cases; the number of array genotyped glaucoma cases worldwide exceeds over 75,000 currently - for example 23andMe have data on 43,254 participants with self-reported POAG cases. Biobanks and other studies focusing on glaucoma are likely to take the number of cases over 100,000 in the not too distant future although the challenge will be efficiently collating these for meta-analysis.

As noted above, in addition to case-control samples, data on IOP and VCDR will also be important in increasing discovery power. The largest IOP GWAS comprised almost 140,000 individuals,[125] although there are >200,000 individuals with IOP and array genotypes worldwide - for example the GERA cohort[129] comprises almost 70,000 individuals (non-overlapping with Khawaja *et al*'s study[125]). For VCDR, the largest published GWAS comprises >90,000 individuals[35]; increasing this sample size is more difficult. Nonetheless, based on ongoing studies across the world, it is anticipated that 100,000 individuals will be exceeded in the near future.

For both glaucoma and the endophenotypes IOP and VCDR, increasing in the number of individuals who are phenotyped and genotyped is likely to yield improvements in prediction accuracy. For example, the predicted AUCs for glaucoma for the 34,000 and 75,000 cases scenario (assuming twice as many controls available) are 0.68 and 0.73, respectively. 75,000 glaucoma cases, combined in a multitrait analysis with N= 200,000 IOP and N=100,000 VCDR datasets (the endophenotypes add the equivalent of approximately additional 64,000 case samples) are expected to increase the AUC to 0.75. If hypothetically, the number of samples was doubled over the coming years, this would increase the AUC further, with the AUC beginning to plateau beyond this point.

In our modelling, AUC values are for a baseline model without age and sex included - in practice if age and sex are included, AUCs increase by 0.05-0.1 units.[35] Nonetheless, since glaucoma is not 100% heritable, stochastic environmental factors will prevent the AUC for a glaucoma PRS from exceeding 0.9, meaning that it will never be possible to develop genetic risk predictions which are diagnostic for individual people. Rather, the power of these PRSs in glaucoma lies in risk stratification - whilst risk estimates for individual people will be noisy and inaccurate, as a group those in high risk individuals are at greatly increased risk and will benefit from early screening and interventions.

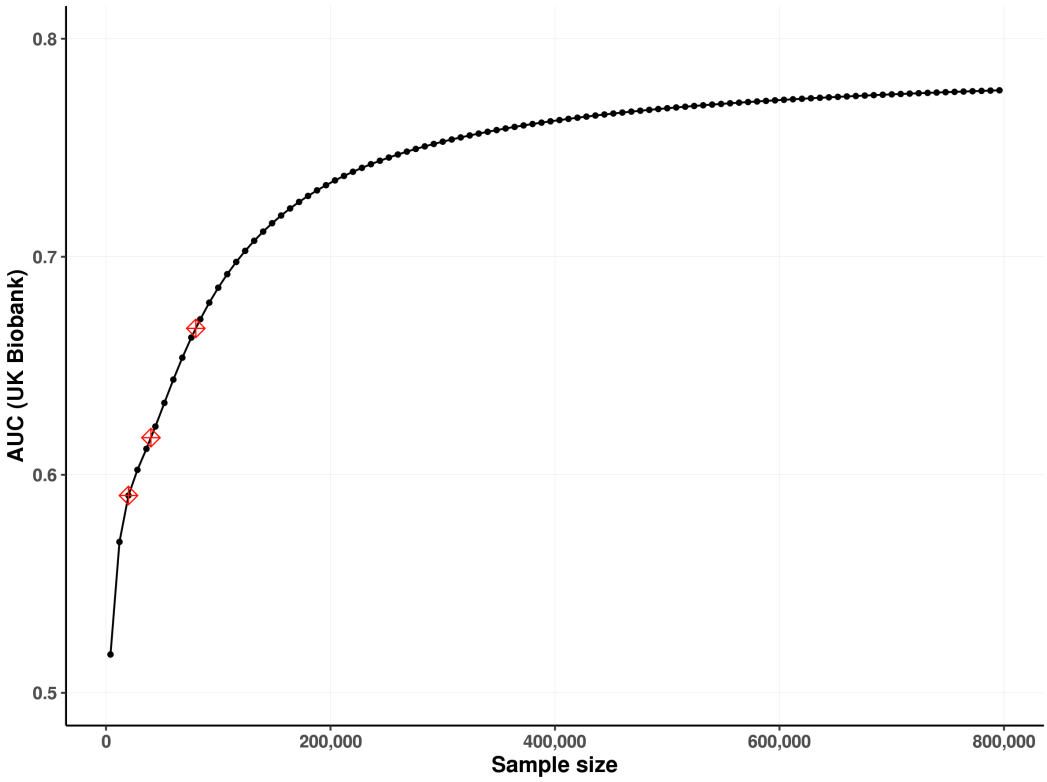


Figure 3. The projection of prediction value of polygenic risk score for glaucoma.

The X-axis is the sample size for glaucoma. The sample size equals the total number of N cases and N controls, assuming a 1:1 ratio. We note here that the AUC values are all for a baseline model without age included - in practice if age is included, all of these AUCs increase by between 0.05 and 0.1 units. Diamond plus symbols show the projection at different sample sizes (roughly current sample size, double, and quadruple).

7A.6 If IOP based screening is not currently recommended - what are the prospects for PRS based screening?

Raised IOP is the principal modifiable risk factor for glaucoma. In the past, IOP has been postulated as a screening tool for glaucoma.[92,93] However, IOP-based population screening is not currently recommended. Chan and colleagues using a community based cross-sectional study of a UK population showed 76% of POAG cases have IOP below 21 mmHg, and no specific IOP threshold can provide adequately sensitivity and specificity values for glaucoma.[186] Although there is no established evidence-based population screening for glaucoma, target screening of individuals at-risk may be cost-effective, *i.e.* sub-groups of older adults.[192]

Using a multi-trait PRS, our recent study considered a target population screening scenario in the key 50–60 age bracket, and showed the PRS can identify high-risk individuals.[35] The PRS can also improve the predictive ability beyond traditional risk factors (age, sex and family history). Participants in the top decile PRS were affected 10 years earlier than people in the bottom PRS decile - the age at which 3% prevalence reached was 59 and 69 in these respective groups. As shown in Figures 2 and 3, increased sample sizes in the foreseeable future will translate directly to improved prediction of glaucoma risk and in turn this will increase the degree of stratification by age that is possible. Since the PRS contains both SNPs which likely act via changes to IOP as well as SNPs which likely act via the nerve head (as measured by variation in VCDR), a PRS based approach is potentially more informative than an approach based solely on IOP. In practice, the utility of a genetic based approach will depend on both the accuracy of the PRS based predictions as well as more general health economic considerations.[421]

7A.7 What proportion of the population are at "high penetrance" risk (e.g. equivalent risk to Myocilin gene Gln368Ter variant)?

Traditionally, clinical genetic testing has primarily focused on identifying carriers of rare monogenic mutations conferring several fold increased disease risk (*e.g.* high penetrance disease causing variants).[42] For instance, the rare *BRCA1* and *BRCA2* mutation carriers are used in genetic screening for breast and ovarian cancers.[422,423] The ascertainment of monogenic mutations can be used in cascade genetic testing for carriers and their family

members and identifying at-risk unaffected relatives for early monitoring,[181] and has shown clear benefit in clinical care.[421] In European ancestry populations, the Myocilin gene Gln368Ter variant is by far the most common high penetrance glaucoma risk variant. Gln368Ter variant carriers have four fold increased risk of non-advanced glaucoma and have 12 fold increased risk for advanced glaucoma.[43] However, the proportion of Gln368Ter variant carriers is low (1 in 786 individuals, 0.13%) and the majority of glaucoma cases are not Gln368Ter carriers. In our recent study, the multi-trait PRS showed effective risk stratification in a case-control advanced glaucoma sample. Individuals in the top 1% of the PRS had a 8.5 fold higher risk relative to the remaining 99%, with even better discrimination value for high-tension glaucoma. Since this elevation of risk is similar to that for Gln368Ter variant carriers, currently the PRS based approach identifies 7 times more individuals at high risk than an approach screening using Gln368Ter variants alone.[117] Hence as shown in other diseases[33], in glaucoma identifying individuals with risk equivalent to monogenic mutations can have clinical utility for screening. As sample sizes increase and the PRS becomes more accurate, the proportion of individuals at “high penetrance” like risk will steadily increase. In addition, the two POAG subtypes (HTG and NTG) may have different genetic bases. The multi-trait PRS had a higher predictive value for HTG subtype - this may be due to 1) a larger proportion of glaucoma cases are HTG which were used to derive the PRS; 2) large IOP GWAS in the multi-trait PRS model was more predictive of HTG. However, currently there are no NTG-specific large-scale GWAS available to train a NTG-specific PRS model. In the future research, with large-scale well-defined glaucoma GWAS, the genetic heterogeneity of the two different glaucoma subtypes should be evaluated.

7A.8 What are the prospects for prediction in different ancestry groups?

During the past decades, genetic studies have predominantly included only European participants. The predictive accuracy of European ancestry derived PRS has been shown to be lower in non-European ancestries (e.g. Asian and African).[424,425] The different linkage disequilibrium patterns, allele frequencies, and genetic architecture may affect the transferability of PRS to people of different ancestries.[424] Nonetheless we showed a European ancestry based glaucoma PRS led to a statistically significant improvement in prediction accuracy in people of South Asian ancestry[35].

The prevalence of glaucoma is dramatically higher in individuals of African ancestry. A recent study identified the first glaucoma risk locus (*APBB2* gene) in individuals of African ancestry.[426] Given *APBB2* was not significant in European or Asian ancestry GWAS, one may be tempted to conclude there are genetic differences between ancestries. However, the key *APBB2* variants are monomorphic in non-African ancestry populations, making it difficult to directly assess the contribution of this locus. When the IGGC cross-ancestry meta-analysis considered the overlap on a genome-wide basis, most glaucoma loci showed a consistent effect across people of European, Asian and African ancestries.[427] It seems likely therefore that conducting large scale GWAS from diverse human populations would improve PRS prediction accuracy and contribute to the transferability of PRS across different ancestries. In the near term, since the majority of GWAS to date have been conducted in European or Asian ancestries, prediction accuracy is likely to be highest in these populations. In the longer term, incorporating a wider range of ancestries in future GWASs would improve prediction performance, particularly in African ancestries who are affected by glaucoma at high rates. Increasing the diversity of genomic research is also important to ensure health equity, and clinical use of PRS may exacerbate health disparities.[428] Four aspects have been proposed[429] to ensure everyone can benefit from genomics research, including increasing the diversity of populations in genetic studies, creating more diverse reference genomes, training more diverse scientists, and developing better methods for predicting across diverse ethnic groups and for separating gene and environment effects. These strategies would improve the generalizability of PRS to different ethnic groups and help health equity.

7A.9 Limitations of genetic risk profiling of glaucoma.

There are several limitations of PRS for glaucoma. First of all, glaucoma PRS studies to date have occurred in research settings and the cost-effectiveness of PRS based genetic screening program is warranted before adopting genetic testing in the general population. Secondly, particularly, in the direct-to-consumer setting, more research is needed on effective communication of PRS results to participants so that early and effective intervention can take place to prevent glaucoma. Finally, genetic studies to date predominantly include only European-descent samples and there is an urgent need to collect samples from different ethnic groups to increase diversity and reduce health

disparities. Recent initiatives to include diverse populations in genomics research include TOPMed and H3Africa consortia.[430,431]

7A.10 Conclusions

Recent advances in glaucoma genetics have mapped many genes implicated in disease pathogenesis and opened the door for risk stratification based on genetic risk predictions. Given the relatively strong predictive power of a POAG PRS and the increasing number of people with genomic data in clinical settings (over 60 million by 2025)[432,433], glaucoma genetic prediction is likely to steadily improve. There is good potential for the PRS based genetic screening program in glaucoma, although cost-effectiveness will need to be formally evaluated. Prospective studies validating the clinical utility for PRS profiling in POAG are also clearly needed. The next steps for implementing these advances into improvements in public health will depend on randomized trials to demonstrate efficacy in real world settings as well as health economics evaluations to guide practical implementation.

7A.11 Acknowledgements

Funding/Support: This work was conducted using the UK Biobank Resource (application number 25331). SM and AWH are supported by Australian National Health and Medical Research Council (NHMRC) Fellowships. We acknowledge funding from NHMRC grants [1116360, 1150144 and 1123248]. The authors thank all of the participants who took part in the UK Biobank and support staff who made this study possible.

Role of the Funder/Sponsor: The funding sources had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

Conflict of Interest Disclosures: A.W.H. and S.M. are listed as coinventors on a patent application for the use of genetic risk scores to determine risk and guide treatment for glaucoma.

7A.12 Supplement

Supplementary Methods and Table are available at:
https://doi.org/10.1001/jamaophthalmol.2020.5404

Chapter 7B. General discussion

The availability of large scale biobanks with deep phenotyping and genomic data[130,434] has given researchers many new opportunities to answer scientific questions about disease mechanisms, prevention, diagnosis, and treatment, which may eventually revolutionize healthcare.[434] One milestone is the release of approximately half a million of individuals with deep phenotyping and genomic data from the UK Biobank (UKB) resource in 2017.

In the work shown in this thesis, by leveraging large scale biobanks and data from international consortia I have identified many new genes associated with eye diseases and traits (Chapters 3, 4, and 5), revealing their polygenic basis. A summary of the key findings and implications from this thesis is shown in Table 1. As the GWAS sample size increases, more genetic discoveries and better insights into the complex genetic architecture of eye diseases will be revealed in the future.[40] One important use for these genetic findings is to develop a polygenic risk score for disease risk prediction. Chapter 3 demonstrates how glaucoma PRS is predictive of glaucoma risk, and enables risk stratification for individuals in high risk and low risk groups. The PRS tool is unlikely to be a diagnostic model, however, it can identify a subgroup of individuals at higher risk, with equivalent risk to high penetrate "rare mutations" (*e.g.* Myocilin gene Gln368Ter variant, as presented in Chapter 2), which has been well studied in clinical genetic testing. The prediction accuracy will increase steadily in the near future when more samples are available as projected in Chapter 7A. The accumulation of large scale genetic data also provides new opportunities to investigate causal inference based on genetic instruments. In Chapter 6, Mendelian randomization analyses were performed to investigate potential causal associations between circulating inflammatory and lipid biomarkers and the risk of AMD. These findings provide new insights into the biological mechanisms of AMD.

Table 1. Summary of key findings and implications from each chapter.

Chapters	Key findings & implications
1. Introduction	a. The concept of complex traits was reviewed. b. Two key statistical genetics approaches were introduced (polygenic risk score and Mendelian randomization). c. General induction of two important eye diseases: glaucoma and age-related macular degeneration.
2. Myocilin gene Gln368Ter variant penetrance and association with glaucoma in population-based and registry-based studies	a. Approximately 50% of *MYOC* p.Gln368Ter carriers older than 65 years had glaucoma or ocular hypertension in UK Biobank, with an even higher prevalence in Australian registry-based studies. b. This study provides evidence to support early screening and monitoring of p.Gln368Ter variant carriers.
3. Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression	a. From multi-trait analysis, 114 statistically independent SNPs were identified for glaucoma, 49 of them had not previously been associated with glaucoma. b. Multi-trait GWAS derived glaucoma polygenic risk score has good predictive ability across a variety of population and clinical datasets, including glaucoma status, age of glaucoma diagnosis, risk stratification in general population and persons carrying *MYOC* p.Gln368Ter variant, glaucoma disease progression, and requirement of glaucoma surgery. c. The multi-trait glaucoma PRS will help the development of a personalized approach for earlier treatment of individuals at a high-risk group, with less intensive monitoring and treatment for lower-risk groups.
4A. Genome-wide association analysis of 95,549 individuals identifies novel loci and genes influencing optic disc morphology 4B. Automated AI labelling of optic nerve head enables new insights into cross-ancestry glaucoma risk and genetic discovery in >280,000 images from UKB and CLSA	a. Deep learning models were trained from optic nerve head photographs, which enables automated labelling of vertical disc diameter and VCDR for further cross-ancestry epidemiological studies and genetic analysis. b. The distributions of VCDR, vertical disc diameter, intraocular pressure, as well as glaucoma risk were systematically evaluated across different ancestries, which provide a possible explanation for the high prevalence of normal tension glaucoma in the East Asian population. c. AI-based gradings dramatically increased SNP-based heritability, and identified more than 200 loci for both VCDR and disc diameter (doubled the number of loci from previous studies), and many of the novel VCDR loci also conferring risk for glaucoma. d. AI approaches provide accurate phenotyping and dramatically accelerate the pace of genetic discoveries, which give new insights into the pathogenesis of glaucomatous optic neuropathy.
5. Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration	a. From genome wide meta-analysis, 69 statistically independent SNPs were identified for AMD, 12 of them had not previously been associated with AMD. b. From further functional annotation, most of the novel genes are expressed in the retina and may be involved in pathways of AMD pathogenesis.
6A. Using Mendelian randomization to evaluate the	a. Elevated circulating C-reactive protein levels are associated with increased risk of AMD.

causal relationship between serum C-reactive protein levels and age-related macular degeneration 6B. The effects of eight serum lipid biomarkers on age-related macular degeneration risk: a Mendelian randomization study	b. Elevated circulating ApoA1 and HDL-C levels are associated with the risk of all AMD subtypes, whereas ApoB, LDL-C, CHOL, and non-HDL-C levels are particularly associated with decreased risk of intermediate and GA AMD. c. This study supports the important role of the inflammatory biomarker CRP and lipid metabolism in drusen formation and AMD development, suggesting the potential utility of disease prediction and targeting CRP and lipid pathways for therapeutic treatment in preventing AMD.
7A. Predicting the future of genetic risk profiling of glaucoma: a narrative review 7B. General discussion	a. A series of pertinent questions in the genetic risk profiling of glaucoma were reviewed. b. The genetic risk prediction performance for glaucoma was projected when larger sample size is available. c. The key findings from the work in this thesis were summarized, the challenges of fine-mapping causal variants or genes and application of polygenic risk score were discussed.

In the digital era, large biobanks and national registries with massive data have become available, such as UK Biobank[130], China Kadoorie biobank[435], BioBank Japan[436], All of Us[437], TOPMed Programme[438], FinnGen study[439], H3Africa[440], Million Veteran program[441], Estonian Biobank[442], and Canadian Longitudinal Study on Aging[310]. Whilst some of these biobank studies (such as UK Biobank) have already delivered many new findings, other studies are only beginning to reach their potential. Traits or variables from wearable-devices[443], new knowledge from deep learning (*e.g.* medical images)[444], longitudinal follow-ups, omics data from different levels (genomic, transcriptomic, epigenetic, proteomic, and metabolomic data, etc.)[445,446] will further expand the size and depth of the data. Integrating large amounts of data resources will uncover more biological mechanisms and improve medical care. For instance, a deep learning model was used to extract knowledge from retinal fundus images that can predict cardiovascular risk factors.[444] Another study showed that retinal fundus images coupled with genetic data can predict AMD progression.[287] Artificial intelligence (AI) algorithms are revolutionizing the field of medical images for disease screening, prevention, diagnosis, and treatment.[286,287,314,447] In Chapter 4B, we have shown that AI models offer the opportunity to conduct large-scale genetic studies based on the automated highly accurate imaging-derived phenotypes, and to identify novel genes associated with glaucoma risk.

Despite the remarkably large number of GWAS loci identified to date, it still remains a challenge to identify causal variants or genes, and to interpret the underlying biological processes.[448] Fine-mapping methods based on statistical genetic approaches, integrating

large-scale omics data, and advance of novel biotechnologies such as CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) technology[449] have provided many new opportunities.

7B.1 Fine-mapping of causal variants or genes

In recent years, many new statistical fine-mapping methods and biotechnology techniques have been developed and have been applied to uncover potential causal variants or genes. Traditionally, SNPs with P values less than a particular threshold (*e.g.* $P < 5 \times 10^{-8}$ after accounting for multiple testing) are regions of interest. However, neighbouring SNPs may pass the genome-wide significant level because of linkage disequilibrium (LD) with the index SNP. Clumping or stepwise conditional analysis methods have been developed to prioritize lead SNPs accounting for LD blocks.[212] However, the top SNPs may be just surrogates for the underlying causal variants, and more importantly, causal variants are not necessary to be SNPs with the smallest P value in a region.[450,451] Factors such as multiple causal SNPs in a region, local LD structure, sample size, SNP effect sizes, and whether causal variants can be genotyped or imputed can further affect the accuracy and power of fine-mapping.[451] Various statistical fine-mapping approaches have been adopted to prioritize potential causal variants or genes, such as Bayesian based approaches (*e.g.* CAVIAR,[452] PAINTOR,[453] FINEMAP,[454] and SuSiE[455]), leveraging functional annotations,[456,457] integrating GWAS with gene expression data (*e.g.* PrediXcan, FUSION, SMR, FOCUS),[243,320,458,459] and trans-ethnic analysis.[460] For instance, a breast cancer GWAS has identified variants among more than 150 regions, however, fewer than 20 regions have been well studied.[461] A recent fine-mapping study has been conducted based on stepwise regression models and a Bayesian approach (PAINTOR analysis to leverage genetic associations, LD structure, and genomic features), with further epigenetic expression and chromatin conformation data to infer 191 likely target genes.[461] Studies have also shown trans-ethnic diverse populations can increase power to identify novel loci and improve fine-mapping resolution.[460] Compared with European and Asian populations, African ancestry populations have low linkage disequilibrium structure and high genetic heterogeneity (the proposed “Out of Africa” model).[462] Simulation and empirical studies have also demonstrated that the inclusion of African samples leads to improvement in gene discoveries and fine-mapping.[460,463] Different approaches have been developed to perform trans-ethnic meta-analysis. For instance, in

the traditional fixed effects model (FE), the effect sizes for each SNP are assumed to be homogeneous across multiple studies (or multiple ancestry groups), and an inverse-variance-weighted method is used to calculate the meta effect sizes. However, because of the genetic heterogeneity between different ancestry groups, the effect sizes across different studies are unlikely to be homogeneous, and the random-effects model (RE) allows the effect sizes to be heterogeneous across studies, following a normal distribution. The limitation of this classical random-effects model is that accounting for additional variability leads to a conservative estimation and loses power in trans-ethnic meta-analysis. Han and Eskin have proposed two new random-effects models, 1) Han and Eskin's random effects model (RE-HE), and 2) binary effects model (BE), both of which are powerful and account for heterogeneity,[464,465] and have shown a comparable performance with another Bayesian-based approach (MANTRA).[466,467] However, statistical fine-mapping models alone are unlikely to determine causal variants or genes for all loci. The underlying biological mechanisms for hundreds of disease associated loci remain to be explored.[468]

With the advance of new biotechnologies, integrating large-scale population-based omics data is another promising angle to systematically identify druggable targets,[469] harnessing omics such as transcriptome,[320] methylome,[470] proteome,[471] and metabolome.[472] For instance, circulating proteins and metabolites represent intermediate phenotypes for disease outcomes, as well as one major source of drug targets.[469] A recent phenome-wide Mendelian randomization study has comprehensively evaluated potential causal effects of approximately 1,000 plasma proteins on 225 human phenotypes based on a two sample MR framework and colocalization analysis, and has provided valuable information in validating and prioritizing therapeutic targets from the pQTL MR analysis.[471] Another burgeoning and rapid developing field is single-cell RNA-sequencing (scRNA-seq)[473], which enables the quantification of gene expression levels at the single cell level.

Functional experiments can provide further insights into the identification of causal variants and validation of drug targets. CRISPR technology, a revolutionary discovery that has been awarded the 2020 Nobel Prize in Chemistry, has shown a promising future for functional validation. For instance, CRISPR-mediated gene editing approach has been applied to perturb gene functions (*e.g.* CRISPR activation and interference [CRISPRa/i])[474,475] that prioritized from schizophrenia GWAS risk variants, and the downstream gene expressions were evaluated in different neuronal cell types based on the human induced pluripotent stem

cell (hiPSC) approach.[476] The advancing of CRISPR-based genome-targeting technologies,[477] including base editing, gene editing, epigenome editing, chromatin imaging, and chromatin topology manipulations, also enables large-scale functional screenings that have shown exciting application in genetic screens, and will pave the way for a better understanding of disease molecular mechanisms and therapeutic targets identification.[478,479]

With the burgeoning of PRS from large-scale genetic data, how to translate PRS into clinical care remains to be answered.

7B.2 Genetic risk predictions in action

7B.2.1 Develop a polygenic risk score

Even though the specific effects of most variants and genes remain unknown, aggregating many risk variants into a polygenic risk score has exhibited increasing utility and promising future in precision medicine in recent years. Detailed guidelines and reporting standards to perform polygenic risk score analysis have been presented.[24,480,481] A 22-item Polygenic Risk Score Reporting Statement is recently published to provide recommendations for PRS studies, including study design, risk model development and evaluation, as well as reporting limitations and clinical implications.[481] In general, a good PRS study should follow the best practices of both GWAS and risk prediction model studies, where genetic risk score is included as additional information on top of traditional epidemiological risk factors. Typically, the model development procedures and evaluation metrics are the same as epidemiological risk prediction models (*e.g.* the well-known Framingham Risk Score). One main advantage of polygenic risk prediction models is genetic information could be available from birth, and different polygenic risk scores can be developed based on the same genetic data ("all-in-one") weighted by their effect sizes for each specific disease, such as coronary heart disease, breast cancer, colorectal cancer, and glaucoma.

7B.2.2 Cost-effectiveness evaluation of PRS

The cost-effectiveness analysis (CEA) of genetic tests for some rare diseases has been well-studied,[482] however, as far as I know, there are only several studies to evaluate the cost-effectiveness of PRS for common diseases. For colorectal cancer (CRC), a recent published study has demonstrated that a CRC screening based on PRS is unlikely to be

cost-effective compared with colonoscopy screening.[483] However, it is expected to become cost-effective if the predictive AUC increased by 0.05 (beyond 0.65), or there was a 30% reduction in the price of polygenic testing, or a 5% increase in screening participation. With the availability of large-scale population genetic data (*i.e.* UK Biobank, 23andMe,[484] and All of Us Research Program[437]) to boost the training sample size and to improve prediction accuracy of PRS, as well as the reduction in the cost of genetic testing, it is anticipated that PRS-based CRC screening will become cost-effective in the foreseeable future. Another modelling study for prostate cancer showed a polygenic risk-tailored screening programme can prevent 6.3% deaths from prostate cancer, as well as lead to one-third fewer overdiagnosis compared with current recommended age-based screening strategy (in men aged 55 to 69).[485] Given the high heritability of some eye diseases (*e.g.* glaucoma), there is good potential to apply PRS-based population screening programs. However, currently there is no cost-effectiveness evaluation of glaucoma population screening based on polygenic risk score, even though glaucoma PRS has shown a relatively high predictive accuracy.[35] In the future, before applying direct-to-consumer genetic tests for glaucoma and other complex diseases, high quality cost effectiveness studies are warranted to assess the utility and efficacy of PRS in population risk stratification and disease screening.

7B.2.3 Application of PRS in clinical decision making

To adopt PRS in clinical decision making, randomized clinical trials are needed to evaluate the efficiency of PRS for drug prescriptions and therapeutic responses. [63] A recent ODYSSEY OUTCOMES trial has demonstrated the benefits of PRS in guiding therapeutic decision making. [486] This study showed that the incidence of major adverse cardiovascular events (MACE) in the placebo arm was 17% and 11% for high and lower PRS groups, respectively. For participants in the high PRS group, alirocumab treatment can reduce the relative incidence of MACE by 37% compared with placebo arm, while in lower PRS group, alirocumab treatment only decreased MACE incidence by 13% compared with placebo arm. Further including baseline LDL-C levels with PRS to stratify participants into four subgroups (combinations of high vs lower PRS groups, LDL-C ≥ 100 mg/dL vs < 100 mg/dL), alirocumab treatment reached the highest relative MACE incidence reduction in participants with high PRS and LDL-C ≥ 100 mg/dL (HR = 0.5, P value = 0.015) compared to the placebo arm, and showed no evidence to reduce MACE incidence in participants with lower PRS and baseline LDL-C < 100 mg/dL (HR = 0.94, P value = 0.42) compared to placebo arm. In

the future, when PRS is widely available in clinical settings, incorporating and optimizing PRS-based therapeutic targeting in clinical pathways will be an important step towards precision medicine.

7B.3 Conclusion

In summary, the work in this thesis enhanced our understanding of the genetics in eye diseases and related quantitative traits. From large-scale genome-wide association studies, many novel genes were identified in this thesis, paving the way to uncover the underlying biological mechanisms. One important utility of genetic data is to develop a personalized disease risk prediction tool in aid of risk stratification, clinical screening, and therapeutic targeting. In my work, I have developed a glaucoma polygenic risk score and have comprehensively evaluated its predictive ability across a variety of clinical datasets and different populations. These discoveries will mark a great starting point upon which we can realise the clinical benefits in risk prediction, prevention, and treatment from genetic data, making it one of the exemplary fields for the success of translating genetic findings into clinical practise for many other complex traits and diseases.

Bibliography

1. Olby, R. The Dimensions of Scientific Controversy: The Biometric—Mendelian Debate. *Br. J. Hist. Sci.* **22**, 299–320 (1989).
2. Fisher, R. A. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* vol. 52 399–433 (1919).
3. Barton, N. H., Etheridge, A. M. & Véber, A. The infinitesimal model: Definition, derivation, and implications. *Theor. Popul. Biol.* **118**, 50–73 (2017).
4. Lynch, M. & Walsh, B. Genetics and analysis of quantitative traits. (1998).
5. Walsh, B. & Lynch, M. Evolution and Selection of Quantitative Traits. *Oxford Scholarship Online* (2018) doi:10.1093/oso/9780198830870.001.0001.
6. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
7. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.* **6**, 95–108 (2005).
8. Wang, W. Y. S., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* **6**, 109–118 (2005).
9. Morales, J. *et al.* A standardized framework for representation of ancestry data in genomics studies, with application to the NHGRI-EBI GWAS Catalog. *Genome Biol.* **19**, 21 (2018).
10. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era—concepts and misconceptions. *Nat. Rev. Genet.* **9**, 255–266 (2008).
11. Falconer, D. S., Mackay, T. F. C. & Frankham, R. Introduction to quantitative genetics (4th edn). *Trends Genet.* **12**, 280 (1996).
12. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).

13. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* **42**, 565–569 (2010).
14. Wainschtein, P., Jain, D. P., Yengo, L., Zheng, Z. & Cupples, L. A. Recovery of trait heritability from whole genome sequence data. *BioRxiv* (2019).
15. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99**, 139–153 (2016).
16. Visscher, P. M. *et al.* Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet.* **2**, e41 (2006).
17. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nat. Genet.* **43**, 519–525 (2011).
18. Lee, S. H. *et al.* Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs. *Nat. Genet.* **44**, 247–250 (2012).
19. Gottesman, I. I. & Shields, J. A polygenic theory of schizophrenia. *Proceedings of the National Academy of Sciences* vol. 58 199–205 (1967).
20. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177–1186 (2017).
21. Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* **173**, 1573–1580 (2018).
22. King, E. A., Davis, J. W. & Degner, J. F. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. *PLoS Genet.* **15**, e1008489 (2019).
23. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).
24. Chatterjee, N., Shi, J. & García-Closas, M. Developing and evaluating polygenic risk

prediction models for stratified disease prevention. *Nat. Rev. Genet.* **17**, 392–406 (2016).

25. Torkamani, A., Wineinger, N. E. & Topol, E. J. The personal and clinical utility of polygenic risk scores. *Nat. Rev. Genet.* (2018) doi:10.1038/s41576-018-0018-x.
26. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47**, 1385–1392 (2015).
27. Meuwissen, T. H., Hayes, B. J. & Goddard, M. E. Prediction of total genetic value using genome-wide dense marker maps. *Genetics* **157**, 1819–1829 (2001).
28. Wray, N. R., Goddard, M. E. & Visscher, P. M. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).
29. International Schizophrenia Consortium *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
30. Maas, P. *et al.* Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States. *JAMA Oncol* **2**, 1295–1302 (2016).
31. Kuchenbaecker, K. B. *et al.* Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *J. Natl. Cancer Inst.* **109**, (2017).
32. Desikan, R. S. *et al.* Genetic assessment of age-associated Alzheimer disease risk: Development and validation of a polygenic hazard score. *PLoS Med.* **14**, e1002258 (2017).
33. Khera, A. V. *et al.* Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet.* **50**, 1219–1224 (2018).
34. Seibert, T. M. *et al.* Polygenic hazard score to guide screening for aggressive prostate cancer: development and validation in large scale cohorts. *BMJ* **360**, j5757 (2018).

35. Craig, J. E. *et al.* Multitrait analysis of glaucoma identifies new risk loci and enables polygenic prediction of disease susceptibility and progression. *Nat. Genet.* **52**, 160–166 (2020).
36. Dudbridge, F. Power and predictive accuracy of polygenic risk scores. *PLoS Genet.* **9**, e1003348 (2013).
37. Chatterjee, N. *et al.* Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat. Genet.* **45**, 400–5, 405e1–3 (2013).
38. Mega, J. L. *et al.* Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet* **385**, 2264–2271 (2015).
39. Mavaddat, N. *et al.* Prediction of breast cancer risk based on profiling with common genetic variants. *J. Natl. Cancer Inst.* **107**, (2015).
40. Zhang, Y., Qi, G., Park, J.-H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat. Genet.* **50**, 1318–1326 (2018).
41. Zhang, Y. D. *et al.* Assessment of polygenic architecture and risk prediction based on common variants across fourteen cancers. *Nat. Commun.* **11**, 3353 (2020).
42. Wright, C. F. *et al.* Assessing the Pathogenicity, Penetrance, and Expressivity of Putative Disease-Causing Variants in a Population Setting. *Am. J. Hum. Genet.* **104**, 275–286 (2019).
43. Han, X. *et al.* Myocilin Gene Gln368Ter Variant Penetrance and Association With Glaucoma in Population-Based and Registry-Based Studies. *JAMA Ophthalmol.* **137**, 28–35 (2019).
44. Vilhjálmsson, B. J. *et al.* Modeling Linkage Disequilibrium Increases Accuracy of Polygenic Risk Scores. *Am. J. Hum. Genet.* **97**, 576–592 (2015).

45. Moser, G. *et al.* Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model. *PLoS Genet.* **11**, e1004969 (2015).
46. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.-C. A. & Smoller, J. W. Polygenic prediction via Bayesian regression and continuous shrinkage priors. *Nat. Commun.* **10**, 1776 (2019).
47. Lloyd-Jones, L. R. *et al.* Improved polygenic prediction by Bayesian multiple regression on summary statistics. *Nat. Commun.* **10**, 5086 (2019).
48. Hu, Y. *et al.* Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comput. Biol.* **13**, e1005589 (2017).
49. Auton, A., Price, A. L. & Research Team, 23andme. Modeling functional enrichment improves polygenic prediction accuracy in UK Biobank and 23andMe data sets. *bioRxiv* (2018).
50. Li, C., Yang, C., Gelernter, J. & Zhao, H. Improving genetic risk prediction by leveraging pleiotropy. *Hum. Genet.* **133**, 639–650 (2014).
51. Hu, Y. *et al.* Joint modeling of genetically correlated diseases and functional annotations increases accuracy of polygenic risk prediction. *PLoS Genet.* **13**, e1006836 (2017).
52. Zhu, X. *et al.* Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.* **96**, 21–36 (2015).
53. Turley, P. *et al.* Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat. Genet.* **50**, 229–237 (2018).
54. Márquez-Luna, C., Loh, P.-R., South Asian Type 2 Diabetes (SAT2D) Consortium, SIGMA Type 2 Diabetes Consortium & Price, A. L. Multiethnic polygenic risk scores improve risk prediction in diverse populations. *Genet. Epidemiol.* **41**, 811–823 (2017).
55. Kulm, S., Mezey, J. & Elemento, O. Benchmarking the Accuracy of Polygenic Risk Scores and their Generative Methods. *medRxiv* (2020).
56. Ni, G. *et al.* A comprehensive evaluation of polygenic score methods across cohorts in

psychiatric disorders. *medRxiv* (2020).

57. Euesden, J., Lewis, C. M. & O'Reilly, P. F. PRSice: Polygenic Risk Score software. *Bioinformatics* **31**, 1466–1468 (2015).
58. Privé, F., Arbel, J. & Vilhjálmsson, B. J. LDpred2: better, faster, stronger. *BioRxiv* (2020).
59. Mak, T. S. H., Porsch, R. M., Choi, S. W., Zhou, X. & Sham, P. C. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.* **41**, 469–480 (2017).
60. Robinson, M. R. *et al.* Genetic evidence of assortative mating in humans. *Nature Human Behaviour* **1**, 1–13 (2017).
61. Ruan, Y. *et al.* Improving Polygenic Prediction in Ancestrally Diverse Populations. doi:10.1101/2020.12.27.20248738.
62. Chun, S. *et al.* Non-parametric Polygenic Risk Prediction via Partitioned GWAS Summary Statistics. *Am. J. Hum. Genet.* **107**, 46–59 (2020).
63. Zhou, G. & Zhao, H. A fast and robust Bayesian nonparametric method for prediction of complex traits using summary statistics. *Cold Spring Harbor Laboratory* 2020.11.30.405241 (2020) doi:10.1101/2020.11.30.405241.
64. Yang, S. & Zhou, X. Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets. *Am. J. Hum. Genet.* **106**, 679–693 (2020).
65. Weissbrod, O. *et al.* Leveraging fine-mapping and non-European training data to improve trans-ethnic polygenic risk scores. *medRxiv* (2021).
66. Maier, R. *et al.* Joint analysis of psychiatric disorders increases accuracy of risk prediction for schizophrenia, bipolar disorder, and major depressive disorder. *Am. J. Hum. Genet.* **96**, 283–294 (2015).
67. Chung, W. *et al.* Efficient cross-trait penalized regression increases prediction accuracy in large cohorts using secondary phenotypes. *Nat. Commun.* **10**, 569 (2019).

68. Grotzinger, A. D. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* **3**, 513–525 (2019).
69. Haycock, P. C. *et al.* Best (but oft-forgotten) practices: the design, analysis, and interpretation of Mendelian randomization studies. *Am. J. Clin. Nutr.* **103**, 965–978 (2016).
70. Davies, N. M., Holmes, M. V. & Davey Smith, G. Reading Mendelian randomisation studies: a guide, glossary, and checklist for clinicians. *BMJ* **362**, k601 (2018).
71. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Stat. Med.* **36**, 1783–1802 (2017).
72. Sivakumaran, S. *et al.* Abundant pleiotropy in human complex diseases and traits. *Am. J. Hum. Genet.* **89**, 607–618 (2011).
73. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
74. Hemani, G., Bowden, J. & Davey Smith, G. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Hum. Mol. Genet.* **27**, R195–R208 (2018).
75. Morrison, J., Knoblauch, N., Marcus, J. H., Stephens, M. & He, X. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nat. Genet.* (2020) doi:10.1038/s41588-020-0631-4.
76. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *Int. J. Epidemiol.* **44**, 512–525 (2015).
77. Verbanck, M., Chen, C.-Y., Neale, B. & Do, R. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nat. Genet.* **50**, 693–698 (2018).
78. Zhu, Z. *et al.* Causal associations between risk factors and common diseases inferred

from GWAS summary data. *Nat. Commun.* **9**, 224 (2018).

79. Sanderson, E., Davey Smith, G., Windmeijer, F. & Bowden, J. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *Int. J. Epidemiol.* **48**, 713–727 (2019).
80. Angrist, J. D. & Imbens, G. W. Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity. *J. Am. Stat. Assoc.* **90**, 431–442 (1995).
81. Burgess, S., Butterworth, A. & Thompson, S. G. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genet. Epidemiol.* **37**, 658–665 (2013).
82. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genet. Epidemiol.* **40**, 304–314 (2016).
83. Hartwig, F. P., Davey Smith, G. & Bowden, J. Robust inference in summary data Mendelian randomization via the zero modal pleiotropy assumption. *Int. J. Epidemiol.* **46**, 1985–1998 (2017).
84. Burgess, S., Foley, C. N., Allara, E., Staley, J. R. & Howson, J. M. M. A robust and efficient method for Mendelian randomization with hundreds of genetic variants. *Nat. Commun.* **11**, 376 (2020).
85. Qi, G. & Chatterjee, N. Mendelian randomization analysis using mixture models for robust and efficient estimation of causal effects. *Nat. Commun.* **10**, 1941 (2019).
86. Cho, Y. *et al.* Exploiting horizontal pleiotropy to search for causal pathways within a Mendelian randomization framework. *Nat. Commun.* **11**, 1010 (2020).
87. Rees, J. M. B., Wood, A. M. & Burgess, S. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured

pleiotropy. *Stat. Med.* **36**, 4705–4718 (2017).

88. Zuber, V., Colijn, J. M., Klaver, C. & Burgess, S. Selecting likely causal risk factors from high-throughput experiments using multivariable Mendelian randomization. *Nat. Commun.* **11**, 29 (2020).
89. Agarwal, R., Gupta, S. K., Agarwal, P., Saxena, R. & Agrawal, S. S. Current concepts in the pathophysiology of glaucoma. *Indian J. Ophthalmol.* **57**, 257–266 (2009).
90. Jonas, J. B. *et al.* Glaucoma. *Lancet* (2017) doi:10.1016/s0140-6736(17)31469-1.
91. Kwon, Y. H., Fingert, J. H., Kuehn, M. H. & Alward, W. L. Primary open-angle glaucoma. *N. Engl. J. Med.* **360**, 1113–1124 (2009).
92. Cedrone, C., Mancino, R., Cerulli, A., Cesareo, M. & Nucci, C. Epidemiology of primary glaucoma: prevalence, incidence, and blinding effects. *Prog. Brain Res.* **173**, 3–14 (2008).
93. Foster, P. J., Buhrmann, R., Quigley, H. A. & Johnson, G. J. The definition and classification of glaucoma in prevalence surveys. *Br. J. Ophthalmol.* **86**, 238–242 (2002).
94. Tham, Y. C. *et al.* Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis. *Ophthalmology* **121**, 2081–2090 (2014).
95. Quigley, H. A. & Broman, A. T. The number of people with glaucoma worldwide in 2010 and 2020. *Br. J. Ophthalmol.* **90**, 262–267 (2006).
96. Kim, K. E. *et al.* Prevalence, Awareness, and Risk Factors of Primary Open-Angle Glaucoma: Korea National Health and Nutrition Examination Survey 2008-2011. *Ophthalmology* **123**, 532–541 (2016).
97. Boland, M. V. & Quigley, H. A. Risk factors and open-angle glaucoma: classification and application. *J. Glaucoma* **16**, 406–418 (2007).

98. Hayreh, S. S., Zimmerman, M. B., Podhajsky, P. & Alward, W. L. Nocturnal arterial hypotension and its role in optic nerve head and ocular ischemic disorders. *Am. J. Ophthalmol.* **117**, 603–624 (1994).

99. Ren, R. *et al.* Cerebrospinal fluid pressure in glaucoma: a prospective study. *Ophthalmology* **117**, 259–266 (2010).

100. Topouzis, F. *et al.* Association of open-angle glaucoma with perfusion pressure status in the Thessaloniki Eye Study. *Am. J. Ophthalmol.* **155**, 843–851 (2013).

101. Rand Allingham, R., Moroi, S., Bruce Shields, M. & Damji, K. *Shields' Textbook of Glaucoma*. (Lippincott Williams & Wilkins, 2020).

102. Jonas, J. B., Zäch, F. M., Gusek, G. C. & Naumann, G. O. Pseudoglaucomatous physiologic large cups. *Am. J. Ophthalmol.* **107**, 137–144 (1989).

103. Liu, Y. & Allingham, R. R. Major review: Molecular genetics of primary open-angle glaucoma. *Exp. Eye Res.* **160**, 62–84 (2017).

104. Wiggs, J. L. & Pasquale, L. R. Genetics of glaucoma. *Hum. Mol. Genet.* **26**, R21–R27 (2017).

105. Wolfs, R. C. *et al.* Genetic risk of primary open-angle glaucoma. Population-based familial aggregation study. *Arch. Ophthalmol.* **116**, 1640–1645 (1998).

106. Wang, K., Gaitsch, H., Poon, H., Cox, N. J. & Rzhetsky, A. Classification of common human diseases derived from shared genetic and environmental determinants. *Nat. Genet.* **49**, 1319–1325 (2017).

107. Fingert, J. H. *et al.* Analysis of myocilin mutations in 1703 glaucoma patients from five different populations. *Hum. Mol. Genet.* **8**, 899–905 (1999).

108. Rezaie, T. *et al.* Adult-onset primary open-angle glaucoma caused by mutations in optineurin. *Science* **295**, 1077–1079 (2002).

109. Stone, E. M. *et al.* Identification of a gene that causes primary open angle glaucoma.

Science **275**, 668–670 (1997).

110. Sheffield, V. C. *et al.* Genetic linkage of familial open angle glaucoma to chromosome 1q21-q31. *Nat. Genet.* **4**, 47–50 (1993).

111. Alward, W. L. *et al.* Clinical features associated with mutations in the chromosome 1 open-angle glaucoma gene (GLC1A). *N. Engl. J. Med.* **338**, 1022–1027 (1998).

112. Craig, J. E. *et al.* Evidence for genetic heterogeneity within eight glaucoma families, with the GLC1A Gln368STOP mutation being an important phenotypic modifier. *Ophthalmology* **108**, 1607–1620 (2001).

113. Fingert, J., Stone, E., Sheffield, V. & Alward, W. Myocilin Glaucoma. *Surv. Ophthalmol.* **47**, 547–561 (2002).

114. Gharahkhani, P. *et al.* Accurate Imputation-Based Screening of Gln368Ter Myocilin Variant in Primary Open-Angle Glaucoma. *Invest. Ophthalmol. Vis. Sci.* **56**, 5087–5093 (2015).

115. Allingham, R. R. *et al.* Gln368STOP myocilin mutation in families with late-onset primary open-angle glaucoma. *Invest. Ophthalmol. Vis. Sci.* **39**, 2288–2295 (1998).

116. Hewitt, A. W., Mackey, D. A. & Craig, J. E. Myocilin allele-specific glaucoma phenotype database. *Hum. Mutat.* **29**, 207–211 (2008).

117. Nag, A. *et al.* Evaluation of the Myocilin Mutation Gln368Stop Demonstrates Reduced Penetrance for Glaucoma in European Populations. *Ophthalmology* **124**, 547–553 (2017).

118. Bailey, J. N. C. *et al.* Genome-wide association analysis identifies TXNRD2, ATXN2 and FOXC1 as susceptibility loci for primary open-angle glaucoma. *Nat. Genet.* **48**, 189–194 (2016).

119. Hysi, P. G. *et al.* Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. *Nat. Genet.* **46**, 1126–

1130 (2014).

120. Burdon, K. P. *et al.* Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *Nat. Genet.* **43**, 574–578 (2011).

121. Chen, Y. *et al.* Common variants near ABCA1 and in PMM2 are associated with primary open-angle glaucoma. *Nat. Genet.* **46**, 1115–1119 (2014).

122. Gharahkhani, P. *et al.* Common variants near ABCA1, AFAP1 and GMDS confer risk of primary open-angle glaucoma. *Nat. Genet.* **46**, 1120–1125 (2014).

123. Thorleifsson, G. *et al.* Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma. *Nat. Genet.* **42**, 906–909 (2010).

124. Choquet, H. *et al.* A multiethnic genome-wide association study of primary open-angle glaucoma identifies novel risk loci. *Nat. Commun.* **9**, 2278 (2018).

125. Khawaja, A. P. *et al.* Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nat. Genet.* (2018) doi:10.1038/s41588-018-0126-8.

126. MacGregor, S. *et al.* Genome-wide association study of intraocular pressure uncovers new pathways to glaucoma. *Nat. Genet.* **50**, 1067–1071 (2018).

127. Gharahkhani, P. *et al.* Genome-wide meta-analysis identifies 127 open-angle glaucoma loci with consistent effect across ancestries. *Nat. Commun.* **12**, 1258 (2021).

128. Loos, R. J. F. 15 years of genome-wide association studies and no signs of slowing down. *Nat. Commun.* **11**, 5900 (2020).

129. Choquet, H. *et al.* A large multi-ethnic genome-wide association study identifies novel genetic loci for intraocular pressure. *Nat. Commun.* **8**, 2108 (2017).

130. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

131. Gao, X. R., Huang, H., Nannini, D. R., Fan, F. & Kim, H. Genome-Wide Association

Analyses Identify New Loci Influencing Intraocular Pressure. *Hum. Mol. Genet.* (2018) doi:10.1093/hmg/ddy111.

132. Springelkamp, H. *et al.* Meta-analysis of genome-wide association studies identifies novel loci that influence cupping and the glaucomatous process. *Nat. Commun.* **5**, 4883 (2014).

133. Springelkamp, H. *et al.* 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 (2017).

134. Wong, W. L. *et al.* Global prevalence of age-related macular degeneration and disease burden projection for 2020 and 2040: a systematic review and meta-analysis. *Lancet Glob Health* **2**, e106–16 (2014).

135. Lim, L. S., Mitchell, P., Seddon, J. M., Holz, F. G. & Wong, T. Y. Age-related macular degeneration. *Lancet* **379**, 1728–1738 (2012).

136. Ferris, F. L., 3rd *et al.* Clinical classification of age-related macular degeneration. *Ophthalmology* **120**, 844–851 (2013).

137. CATT Research Group *et al.* Ranibizumab and bevacizumab for neovascular age-related macular degeneration. *N. Engl. J. Med.* **364**, 1897–1908 (2011).

138. Mitchell, P., Liew, G., Gopinath, B. & Wong, T. Y. Age-related macular degeneration. *Lancet* **392**, 1147–1159 (2018).

139. Velilla, S. *et al.* Smoking and age-related macular degeneration: review and update. *J. Ophthalmol.* **2013**, 895147 (2013).

140. Mitchell, P., Wang, J. J., Smith, W. & Leeder, S. R. Smoking and the 5-year incidence of age-related maculopathy: the Blue Mountains Eye Study. *Arch. Ophthalmol.* **120**, 1357–1363 (2002).

141. Smith, W. *et al.* Risk factors for age-related macular degeneration: Pooled findings from

three continents. *Ophthalmology* **108**, 697–704 (2001).

142. Khan, J. C. *et al.* Smoking and age related macular degeneration: the number of pack years of cigarette smoking is a major determinant of risk for both geographic atrophy and choroidal neovascularisation. *Br. J. Ophthalmol.* **90**, 75–80 (2006).

143. Pan, C.-W. *et al.* Refractive errors and age-related macular degeneration: a systematic review and meta-analysis. *Ophthalmology* **120**, 2058–2065 (2013).

144. Wood, A. & Guggenheim, J. A. Refractive Error Has Minimal Influence on the Risk of Age-Related Macular Degeneration: A Mendelian Randomization Study. *Am. J. Ophthalmol.* **206**, 87–93 (2019).

145. Chakravarthy, U. *et al.* Clinical risk factors for age-related macular degeneration: a systematic review and meta-analysis. *BMC Ophthalmol.* **10**, 31 (2010).

146. Wang, Y. *et al.* The Association between the Lipids Levels in Blood and Risk of Age-Related Macular Degeneration. *Nutrients* **8**, (2016).

147. Burgess, S. & Davey Smith, G. Mendelian Randomization Implicates High-Density Lipoprotein Cholesterol-Associated Mechanisms in Etiology of Age-Related Macular Degeneration. *Ophthalmology* **124**, 1165–1174 (2017).

148. Fan, Q. *et al.* HDL-cholesterol levels and risk of age-related macular degeneration: a multiethnic genetic study using Mendelian randomization. *Int. J. Epidemiol.* **46**, 1891–1902 (2017).

149. van Leeuwen, E. M. *et al.* A new perspective on lipid research in age-related macular degeneration. *Prog. Retin. Eye Res.* **67**, 56–86 (2018).

150. Colijn, J. M. *et al.* Increased High-Density Lipoprotein Levels Associated with Age-Related Macular Degeneration: Evidence from the EYE-RISK and European Eye Epidemiology Consortia. *Ophthalmology* **126**, 393–406 (2019).

151. Klein, R. *et al.* Oxidized Low-density Lipoprotein and the Incidence of Age-related

Macular Degeneration. *Ophthalmology* **126**, 752–758 (2019).

152. Sui, G.-Y. *et al.* Is sunlight exposure a risk factor for age-related macular degeneration? A systematic review and meta-analysis. *Br. J. Ophthalmol.* **97**, 389–394 (2013).

153. Fletcher, A. E. *et al.* Sunlight exposure, antioxidants, and age-related macular degeneration. *Arch. Ophthalmol.* **126**, 1396–1403 (2008).

154. Chong, E. W.-T., Kreis, A. J., Wong, T. Y., Simpson, J. A. & Guymer, R. H. Alcohol consumption and the risk of age-related macular degeneration: a systematic review and meta-analysis. *Am. J. Ophthalmol.* **145**, 707–715 (2008).

155. Reynolds, R., Rosner, B. & Seddon, J. M. Dietary omega-3 fatty acids, other fat intake, genetic susceptibility, and progression to incident geographic atrophy. *Ophthalmology* **120**, 1020–1028 (2013).

156. Amirul Islam, F. M. *et al.* Dietary patterns and their associations with age-related macular degeneration: the Melbourne collaborative cohort study. *Ophthalmology* **121**, 1428–1434.e2 (2014).

157. McGuinness, M. B. *et al.* Physical Activity and Age-related Macular Degeneration: A Systematic Literature Review and Meta-analysis. *Am. J. Ophthalmol.* **180**, 29–38 (2017).

158. Colak, E., Majkic-Singh, N., Zoric, L., Radosavljevic, A. & Kosanovic-Jakovic, N. The role of CRP and inflammation in the pathogenesis of age-related macular degeneration. *Biochem. Med.* **22**, 39–48 (2012).

159. Hong, T., Tan, A. G., Mitchell, P. & Wang, J. J. A review and meta-analysis of the association between C-reactive protein and age-related macular degeneration. *Surv. Ophthalmol.* **56**, 184–194 (2011).

160. Annweiler, C. *et al.* Circulating vitamin D concentration and age-related macular degeneration: Systematic review and meta-analysis. *Maturitas* **88**, 101–112 (2016).

161. Adams, M. K. M. *et al.* Abdominal obesity and age-related macular degeneration. *Am. J. Epidemiol.* **173**, 1246–1255 (2011).

162. Zhang, Q.-Y. *et al.* Overweight, Obesity, and Risk of Age-Related Macular Degeneration. *Invest. Ophthalmol. Vis. Sci.* **57**, 1276–1283 (2016).

163. Choi, J. K., Lym, Y. L., Moon, J. W., Shin, H. J. & Cho, B. Diabetes mellitus and early age-related macular degeneration. *Arch. Ophthalmol.* **129**, 196–199 (2011).

164. Hyman, L., Schachat, A. P., He, Q. & Cristina Leske, M. Hypertension, Cardiovascular Disease, and Age-Related Macular Degeneration. *Arch. Ophthalmol.* **118**, 351–358 (2000).

165. Fritsche, L. G. *et al.* Seven new loci associated with age-related macular degeneration. *Nat. Genet.* **45**, 433–9, 439e1–2 (2013).

166. Fritsche, L. G. *et al.* A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants. *Nat. Genet.* **48**, 134–143 (2016).

167. Han, X. *et al.* Genome-wide meta-analysis identifies novel loci associated with age-related macular degeneration. *J. Hum. Genet.* **65**, 657–665 (2020).

168. Yehoshua, Z. *et al.* Systemic complement inhibition with eculizumab for geographic atrophy in age-related macular degeneration: the COMPLETE study. *Ophthalmology* **121**, 693–701 (2014).

169. Geerlings, M. J., de Jong, E. K. & den Hollander, A. I. The complement system in age-related macular degeneration: A review of rare genetic variants and implications for personalized treatment. *Mol. Immunol.* **84**, 65–76 (2017).

170. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* **51**, 606–610 (2019).

171. Menon, M. *et al.* Single-cell transcriptomic atlas of the human retina identifies cell types

associated with age-related macular degeneration. *Nat. Commun.* **10**, 4902 (2019).

172. Hewitt, A. W., Craig, J. E. & Mackey, D. A. Complex genetics of complex traits: the case of primary open-angle glaucoma. *Clin. Experiment. Ophthalmol.* **34**, 472–484 (2006).

173. Gobeil, S. *et al.* Intracellular sequestration of hetero-oligomers formed by wild-type and glaucoma-causing myocilin mutants. *Invest. Ophthalmol. Vis. Sci.* **45**, 3560–3567 (2004).

174. Jain, A. *et al.* CRISPR-Cas9-based treatment of myocilin-associated glaucoma. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 11199–11204 (2017).

175. Chan, M. P. *et al.* Associations with Intraocular Pressure in a Large Cohort: Results from the UK Biobank. *Ophthalmology* **123**, 771–782 (2016).

176. Souzeau, E. *et al.* Australian and New Zealand Registry of Advanced Glaucoma: methodology and recruitment. *Clin. Experiment. Ophthalmol.* **40**, 569–575 (2012).

177. Coote, M. A., McCartney, P. J., Wilkinson, R. M. & Mackey, D. A. The 'GIST' score: ranking glaucoma for genetic studies. Glaucoma Inheritance Study of Tasmania. *Ophthalmic Genet.* **17**, 199–208 (1996).

178. Souzeau, E. *et al.* Higher prevalence of myocilin mutations in advanced glaucoma in comparison with less advanced disease in an Australasian disease registry. *Ophthalmology* **120**, 1135–1143 (2013).

179. Nyholt, D. R. *et al.* Genome-wide association meta-analysis identifies new endometriosis risk loci. *Nat. Genet.* **44**, 1355–1359 (2012).

180. Gharahkhani, P. *et al.* Analysis combining correlated glaucoma traits identifies five new risk loci for open-angle glaucoma. *Sci. Rep.* **8**, 3124 (2018).

181. Souzeau, E. *et al.* Myocilin Predictive Genetic Testing for Primary Open-Angle Glaucoma Leads to Early Identification of At-Risk Individuals. *Ophthalmology* **124**, 303–309 (2017).

182. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

183. Mitchell, P., Smith, W., Attebo, K. & Healey, P. R. Prevalence of open-angle glaucoma in Australia. The Blue Mountains Eye Study. *Ophthalmology* **103**, 1661–1669 (1996).

184. Wensor, M. D., McCarty, C. A., Stanislavsky, Y. L., Livingston, P. M. & Taylor, H. R. The prevalence of glaucoma in the Melbourne Visual Impairment Project. *Ophthalmology* **105**, 733–739 (1998).

185. Shepard, A. R. *et al.* Glaucoma-causing myocilin mutants require the Peroxisomal targeting signal-1 receptor (PTS1R) to elevate intraocular pressure. *Hum. Mol. Genet.* **16**, 609–617 (2007).

186. Chan, M. P. Y. *et al.* Glaucoma and intraocular pressure in EPIC-Norfolk Eye Study: cross sectional study. *BMJ* **358**, j3889 (2017).

187. Fry, A. *et al.* Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank Participants With Those of the General Population. *Am. J. Epidemiol.* **186**, 1026–1034 (2017).

188. Souzeau, E. *et al.* Predictive genetic testing in minors for Myocilin juvenile onset open angle glaucoma. *Clin. Genet.* **88**, 584–588 (2015).

189. van der Valk, R. *et al.* Intraocular pressure--lowering effects of all commonly used glaucoma drugs: a meta-analysis of randomized clinical trials. *Ophthalmology* **112**, 1177–1185 (2005).

190. Weinreb, R. N. & Khaw, P. T. Primary open-angle glaucoma. *Lancet* **363**, 1711–1720 (2004).

191. Fraser, S., Bunce, C. & Wormald, R. Risk factors for late presentation in chronic glaucoma. *Invest. Ophthalmol. Vis. Sci.* **40**, 2251–2257 (1999).

192. Burr, J. M. *et al.* The clinical effectiveness and cost-effectiveness of screening for open

angle glaucoma: a systematic review and economic evaluation. *Health Technol. Assess.* **11**, iii–iv, ix–x, 1–190 (2007).

193. Sanfilippo, P. G., Hewitt, A. W., Hammond, C. J. & Mackey, D. A. The heritability of ocular traits. *Surv. Ophthalmol.* **55**, 561–583 (2010).

194. Leske, M. C., Heijl, A., Hyman, L., Bengtsson, B. & Komaroff, E. Factors for progression and glaucoma treatment: the Early Manifest Glaucoma Trial. *Curr. Opin. Ophthalmol.* **15**, 102–106 (2004).

195. Garway-Heath, D. F. *et al.* Latanoprost for open-angle glaucoma (UKGTS): a randomised, multicentre, placebo-controlled trial. *Lancet* **385**, 1295–1304 (2015).

196. Wu, Y., Zheng, Z., Visscher, P. M. & Yang, J. Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. *Genome Biol.* **18**, 86 (2017).

197. Huang, L. *et al.* Genome-wide analysis identified 17 new loci influencing intraocular pressure in Chinese population. *Sci. China Life Sci.* **62**, 153–164 (2019).

198. Lecarpentier, J. *et al.* Prediction of Breast and Prostate Cancer Risks in Male BRCA1 and BRCA2 Mutation Carriers Using Polygenic Risk Scores. *J. Clin. Oncol.* **35**, 2240–2250 (2017).

199. Na, J. H. *et al.* Detection of glaucoma progression by assessment of segmented macular thickness data obtained using spectral domain optical coherence tomography. *Invest. Ophthalmol. Vis. Sci.* **53**, 3817–3826 (2012).

200. Cheng, C.-Y. *et al.* Nine loci for ocular axial length identified through genome-wide association studies, including shared loci with refractive error. *Am. J. Hum. Genet.* **93**, 264–277 (2013).

201. Pickrell, J. K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat. Genet.* **48**, 709–717 (2016).

202. Verhoeven, V. J. M. *et al.* Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat. Genet.* **45**, 314–318 (2013).

203. Lopes, M. C. *et al.* Identification of a candidate gene for astigmatism. *Invest. Ophthalmol. Vis. Sci.* **54**, 1260–1267 (2013).

204. King, R. *et al.* Genomic locus modulating corneal thickness in the mouse identifies POU6F2 as a potential risk of developing glaucoma. *PLoS Genet.* **14**, e1007145 (2018).

205. Bengtsson, B. The variation and covariation of cup and disc diameters. *Acta Ophthalmol.* **54**, 804–818 (1976).

206. Han, X. *et al.* Genome-wide association analysis of 95,549 individuals identifies novel loci and genes influencing optic disc morphology. *Hum. Mol. Genet.* (2019) doi:10.1093/hmg/ddz193.

207. Olsen, C. M. *et al.* Cohort profile: the QSkin Sun and Health Study. *Int. J. Epidemiol.* **41**, 929–929i (2012).

208. Wiggs, J. L. *et al.* The NEIGHBOR consortium primary open-angle glaucoma genome-wide association study: rationale, study design, and clinical variables. *J. Glaucoma* **22**, 517–525 (2013).

209. Weinreb, R. N., Garway-Heath, D. F., Leung, C., Medeiros, F. A. & Liebmann, J. *Diagnosis of Primary Open Angle Glaucoma: WGA consensus series - 10*. (Kugler Publications, 2017).

210. Loh, P.-R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat. Genet.* **47**, 284–290 (2015).

211. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

212. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics

identifies additional variants influencing complex traits. *Nat. Genet.* **44**, 369–75, S1–3 (2012).

213. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: generalized gene-set analysis of GWAS data. *PLoS Comput. Biol.* **11**, e1004219 (2015).
214. Watanabe, K., Taskesen, E., van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **8**, 1826 (2017).
215. Robin, X. *et al.* pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* vol. 12 77 (2011).
216. R Core Team. R: A Language and Environment for Statistical Computing. (2017).
217. Quigley, H. A. Glaucoma. *Lancet* **377**, 1367–1377 (2011).
218. Cekić, S., Stanković-Babić, G., Visnjić, Z., Jovanović, I. & Risimić, D. Optic disc abnormalities - diagnosis, evolution and influence on visual acuity. *Bosn. J. Basic Med. Sci.* **10**, 125–132 (2010).
219. Casson, R. J., Chidlow, G., Wood, J. P. M., Crowston, J. G. & Goldberg, I. Definition of glaucoma: clinical and experimental concepts. *Clin. Experiment. Ophthalmol.* **40**, 341–349 (2012).
220. Hoffmann, E. M., Zangwill, L. M., Crowston, J. G. & Weinreb, R. N. Optic disk size and glaucoma. *Surv. Ophthalmol.* **52**, 32–49 (2007).
221. Li, A., Li, L., Li, M. & Shi, X. A new characterization for nonarteritic anterior ischemic optic neuropathy. *Int. J. Clin. Exp. Med.* **8**, 18681–18688 (2015).
222. Chang, M. Y. & Pineles, S. L. Optic disk drusen in children. *Surv. Ophthalmol.* **61**, 745–758 (2016).
223. Crowston, J. G. *et al.* The effect of optic disc diameter on vertical cup to disc ratio percentiles in a population based cohort: the Blue Mountains Eye Study. *Br. J. Ophthalmol.* **88**, 766–770 (2004).

224. Jonas, J. B., Bergua, A., Schmitz-Valckenberg, P., Papastathopoulos, K. I. & Budde, W. M. Ranking of optic disc variables for detection of glaucomatous optic nerve damage. *Invest. Ophthalmol. Vis. Sci.* **41**, 1764–1773 (2000).

225. Klein, B. E. K., Klein, R. & Lee, K. E. Heritability of risk factors for primary open-angle glaucoma: the Beaver Dam Eye Study. *Invest. Ophthalmol. Vis. Sci.* **45**, 59–62 (2004).

226. Macgregor, S. *et al.* Genome-wide association identifies ATOH7 as a major gene determining human optic disc size. *Hum. Mol. Genet.* **19**, 2716–2724 (2010).

227. Ramdas, W. D. *et al.* A genome-wide association study of optic disc parameters. *PLoS Genet.* **6**, e1000978 (2010).

228. Khor, C. C. *et al.* Genome-wide association studies in Asians confirm the involvement of ATOH7 and TGFBR3, and further identify CARD10 as a novel locus influencing optic disc area. *Hum. Mol. Genet.* **20**, 1864–1872 (2011).

229. Gasten, A. C. *et al.* A genetic epidemiologic study of candidate genes involved in the optic nerve head morphology. *Invest. Ophthalmol. Vis. Sci.* **53**, 1485–1491 (2012).

230. Springelkamp, H. *et al.* Meta-analysis of Genome-Wide Association Studies Identifies Novel Loci Associated With Optic Disc Morphology. *Genet. Epidemiol.* **39**, 207–216 (2015).

231. Nagai, T. *et al.* The expression of the mouse Zic1, Zic2, and Zic3 gene suggests an essential role for Zic genes in body pattern formation. *Dev. Biol.* **182**, 299–313 (1997).

232. Cavodeassi, F., Ivanovitch, K. & Wilson, S. W. Eph/Ephrin signalling maintains eye field segregation from adjacent neural plate territories during forebrain morphogenesis. *Development* **140**, 4193–4202 (2013).

233. Jonas, J. B., Schmidt, A. M., Müller-Bergh, J. A., Schlötzer-Schrehardt, U. M. & Naumann, G. O. Human optic nerve fiber count and optic disc size. *Invest. Ophthalmol. Vis. Sci.* **33**, 2012–2018 (1992).

234. Ernst, J. & Kellis, M. ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods* **9**, 215–216 (2012).

235. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).

236. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).

237. Cha, S. *et al.* Identification of five novel genetic loci related to facial morphology by genome-wide association studies. *BMC Genomics* **19**, 481 (2018).

238. Ramjaun, A. R., Tomlinson, S., Eddaoudi, A. & Downward, J. Upregulation of two BH3-only proteins, Bmf and Bim, during TGF beta-induced apoptosis. *Oncogene* **26**, 970–981 (2007).

239. Saika, S. TGFbeta pathobiology in the eye. *Lab. Invest.* **86**, 106–115 (2006).

240. Zheng, J. *et al.* LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* **33**, 272–279 (2017).

241. Hoshino, A. *et al.* Molecular Anatomy of the Developing Human Retina. *Dev. Cell* **43**, 763–779.e4 (2017).

242. Haeseleer, F. & Palczewski, K. [24] Short-chain dehydrogenases/reductases in retina. in *Methods in Enzymology* vol. 316 372–383 (Academic Press, 2000).

243. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

244. Iliescu, A., Gravel, M., Horth, C., Apuzzo, S. & Gros, P. Transmembrane topology of mammalian planar cell polarity protein Vangl1. *Biochemistry* **50**, 2274–2282 (2011).

245. Belotti, E. *et al.* Molecular characterisation of endogenous Vangl2/Vangl1 heteromeric

protein complexes. *PLoS One* **7**, e46213 (2012).

246. Wolff, T. & Rubin, G. M. Strabismus, a novel gene that regulates tissue polarity and cell fate decisions in Drosophila. *Development* **125**, 1149–1159 (1998).

247. Prendergast, L. *et al.* Premitotic assembly of human CENPs -T and -W switches centromeric chromatin to a mitotic state. *PLoS Biol.* **9**, e1001082 (2011).

248. N'Diaye, A. *et al.* Identification, replication, and fine-mapping of Loci associated with adult height in individuals of african ancestry. *PLoS Genet.* **7**, e1002298 (2011).

249. Shungin, D. *et al.* New genetic loci link adipose and insulin biology to body fat distribution. *Nature* **518**, 187–196 (2015).

250. Elks, C. E. *et al.* Thirty new loci for age at menarche identified by a meta-analysis of genome-wide association studies. *Nat. Genet.* **42**, 1077–1085 (2010).

251. Casey, J. *et al.* First implication of STRA6 mutations in isolated anophthalmia, microphthalmia, and coloboma: a new dimension to the STRA6 phenotype. *Hum. Mutat.* **32**, 1417–1426 (2011).

252. Pasutto, F. *et al.* Mutations in STRA6 cause a broad spectrum of malformations including anophthalmia, congenital heart defects, diaphragmatic hernia, alveolar capillary dysplasia, lung hypoplasia, and mental retardation. *Am. J. Hum. Genet.* **80**, 550–560 (2007).

253. Azuma, N. *et al.* Mutations of the PAX6 gene detected in patients with a variety of optic-nerve malformations. *Am. J. Hum. Genet.* **72**, 1565–1570 (2003).

254. Jain, M., Bhat, G. P., Vijayraghavan, K. & Inamdar, M. S. Rudhira/BCAS3 is a cytoskeletal protein that controls Cdc42 activation and directional cell migration during angiogenesis. *Exp. Cell Res.* **318**, 753–767 (2012).

255. Kazanskaya, O. *et al.* The Wnt signaling regulator R-spondin 3 promotes angioblast and vascular development. *Development* **135**, 3655–3664 (2008).

256. Zode, G. S., Clark, A. F. & Wordinger, R. J. Bone morphogenetic protein 4 inhibits TGF-beta2 stimulation of extracellular matrix proteins in optic nerve head cells: role of gremlin in ECM modulation. *Glia* **57**, 755–766 (2009).

257. Chang, B. *et al.* Haploinsufficient Bmp4 ocular phenotypes include anterior segment dysgenesis with elevated intraocular pressure. *BMC Genet.* **2**, 18 (2001).

258. Leung, C. K.-S. *et al.* Optic disc measurements in myopia with optical coherence tomography and confocal scanning laser ophthalmoscopy. *Invest. Ophthalmol. Vis. Sci.* **48**, 3178–3183 (2007).

259. Jonas, J. B., Gusek, G. C. & Naumann, G. O. Optic disk morphometry in high myopia. *Graefes Arch. Clin. Exp. Ophthalmol.* **226**, 587–590 (1988).

260. Orr, A. *et al.* Mutations in a novel serine protease PRSS56 in families with nanophthalmos. *Mol. Vis.* **17**, 1850–1861 (2011).

261. Paylakhi, S. *et al.* Müller glia-derived PRSS56 is required to sustain ocular axial growth and prevent refractive error. *PLoS Genet.* **14**, e1007244 (2018).

262. Reis, L. M. *et al.* BMP4 loss-of-function mutations in developmental eye disorders including SHORT syndrome. *Hum. Genet.* **130**, 495–504 (2011).

263. Kiefer, A. K. *et al.* Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet.* **9**, e1003299 (2013).

264. Charrin, S., Jouannet, S., Boucheix, C. & Rubinstein, E. Tetraspanins at a glance. *J. Cell Sci.* **127**, 3641–3648 (2014).

265. Gao, X. R., Huang, H. & Kim, H. Genome-wide association analyses identify 139 loci associated with macular thickness in the UK Biobank cohort. *Hum. Mol. Genet.* (2018) doi:10.1093/hmg/ddy422.

266. Bartling, H., Wanger, P. & Martin, L. Measurement of optic disc parameters on digital

fundus photographs: algorithm development and evaluation. *Acta Ophthalmol.* **86**, 837–841 (2008).

267. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

268. Quigley, H. A., Brown, A. E., Morrison, J. D. & Drance, S. M. The size and shape of the optic disc in normal human eyes. *Arch. Ophthalmol.* **108**, 51–57 (1990).

269. Garway-Heath, D. F. *et al.* Measurement of optic disc size: equivalence of methods to correct for ocular magnification. *Br. J. Ophthalmol.* **82**, 643–649 (1998).

270. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

271. UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).

272. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* **26 Suppl 1**, S6–14 (1997).

273. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95–103 (1999).

274. Khawaja, A. P. *et al.* The EPIC-Norfolk Eye Study: rationale, methods and a cross-sectional analysis of visual impairment in a population-based cohort. *BMJ Open* vol. 3 e002684 (2013).

275. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.* **39**, 906–913 (2007).

276. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–

140 (2010).

277. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).

278. Cox, D. R. & Reid, N. Parameter Orthogonality and Approximate Conditional Inference. *J. R. Stat. Soc. Series B Stat. Methodol.* **49**, 1–18 (1987).

279. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).

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

281. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).

282. Meyer, C. H., Rodrigues, E. B. & Schmidt, J. C. Congenital optic nerve head pit associated with reduced retinal nerve fibre thickness at the papillomacular bundle. *Br. J. Ophthalmol.* **87**, 1300–1301 (2003).

283. Tielsch, J. M., Katz, J., Quigley, H. A., Miller, N. R. & Sommer, A. Intraobserver and interobserver agreement in measurement of optic disc characteristics. *Ophthalmology* **95**, 350–356 (1988).

284. Varma, R., Steinmann, W. C. & Scott, I. U. Expert agreement in evaluating the optic disc for glaucoma. *Ophthalmology* **99**, 215–221 (1992).

285. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).

286. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).

287. Yan, Q. *et al.* Deep-learning-based Prediction of Late Age-Related Macular

Degeneration Progression. *Nat Mach Intell* **2**, 141–150 (2020).

288. Food, U. S., Administration, D. & Others. FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems. *News Release, April* (2018).

289. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N. & Folk, J. C. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digit Med* **1**, 39 (2018).

290. An, G. *et al.* Glaucoma Diagnosis with Machine Learning Based on Optical Coherence Tomography and Color Fundus Images. *J. Healthc. Eng.* **2019**, 4061313 (2019).

291. Katz, N., Goldbaum, M. & Nelson, M. An image processing system for automatic retina diagnosis. *Imaging and Remote …* (1988).

292. Sahlsten, J. *et al.* Deep Learning Fundus Image Analysis for Diabetic Retinopathy and Macular Edema Grading. *Sci. Rep.* **9**, 10750 (2019).

293. Sengupta, S., Singh, A., Leopold, H. A., Gulati, T. & Lakshminarayanan, V. Ophthalmic diagnosis using deep learning with fundus images - A critical review. *Artif. Intell. Med.* **102**, 101758 (2020).

294. Wolfs, R. C., Ramrattan, R. S., Hofman, A. & de Jong, P. T. Cup-to-disc ratio: ophthalmoscopy versus automated measurement in a general population: The Rotterdam Study. *Ophthalmology* **106**, 1597–1601 (1999).

295. Harper, R., Reeves, B. & Smith, G. Observer variability in optic disc assessment: implications for glaucoma shared care. *Ophthalmic Physiol. Opt.* **20**, 265–273 (2000).

296. Sisodia, D. S. & Nair, S. Diabetic retinal fundus images: Preprocessing and feature extraction for early detection of diabetic retinopathy. *Biomedical and* (2017).

297. Orlando, J. I., Prokofyeva, E. & del Fresno, M. Convolutional neural network transfer for automated glaucoma identification. *12th international* (2017).

298. Singh, A., Dutta, M. K., ParthaSarathi, M., Uher, V. & Burget, R. Image processing based automatic diagnosis of glaucoma using wavelet features of segmented optic disc from fundus image. *Comput. Methods Programs Biomed.* **124**, 108–120 (2016).

299. Li, Z. *et al.* Efficacy of a Deep Learning System for Detecting Glaucomatous Optic Neuropathy Based on Color Fundus Photographs. *Ophthalmology* vol. 125 1199–1206 (2018).

300. Cuellar-Partida, G. *et al.* Assessing the Genetic Predisposition of Education on Myopia: A Mendelian Randomization Study. *Genetic Epidemiology* vol. 40 66–72 (2016).

301. Gelman, S., Cone, F. E., Pease, M. E. & Nguyen, T. D. The presence and distribution of elastin in the posterior and retrobulbar regions of the mouse eye. *Exp. Eye Res.* (2010).

302. Rebecca, M. *et al.* Elastin modulation and modification by homocysteine: a key factor in the pathogenesis of Pseudoexfoliation syndrome? *Br. J. Ophthalmol.* **103**, 985–992 (2019).

303. Marsh, B. C. *et al.* Optic nerve head (ONH) topographic analysis by stratus OCT in normal subjects: correlation to disc size, age, and ethnicity. *J. Glaucoma* **19**, 310–318 (2010).

304. Lee, R. Y. *et al.* Ethnic variation in optic disc size by fundus photography. *Curr. Eye Res.* **38**, 1142–1147 (2013).

305. Beck, R. W., Messner, D. K., Musch, D. C., Martonyi, C. L. & Lichter, P. R. Is there a racial difference in physiologic cup size? *Ophthalmology* **92**, 873–876 (1985).

306. Varma, R. *et al.* Race-, age-, gender-, and refractive error-related differences in the normal optic disc. *Arch. Ophthalmol.* **112**, 1068–1076 (1994).

307. Soh, Z. D. *et al.* Asian-specific vertical cup-to-disc ratio cut-off for glaucoma screening: An evidence-based recommendation from a multi-ethnic Asian population. *Clin.*

Experiment. Ophthalmol. (2020) doi:10.1111/ceo.13836.

308. Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. (2017) doi:10.1101/166298.

309. Raina, P. S. *et al.* The Canadian longitudinal study on aging (CLSA). *Can. J. Aging* **28**, 221–229 (2009).

310. Raina, P. *et al.* Cohort Profile: The Canadian Longitudinal Study on Aging (CLSA). *Int. J. Epidemiol.* **48**, 1752–1753j (2019).

311. Bonnemaijer, P. W. M. *et al.* Multi-trait genome-wide association study identifies new loci associated with optic disc parameters. *Commun Biol* **2**, 435 (2019).

312. Gharahkhani, P. *et al.* 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* 2020.01.30.927822 (2020) doi:10.1101/2020.01.30.927822.

313. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 770–778 (2016).

314. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems 25* (eds. Pereira, F., Burges, C. J. C., Bottou, L. & Weinberger, K. Q.) 1097–1105 (Curran Associates, Inc., 2012).

315. Howard, J. & Gugger, S. Fastai: A Layered API for Deep Learning. *Information* **11**, 108 (2020).

316. Smith, L. N. Cyclical Learning Rates for Training Neural Networks. in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* 464–472 (2017).

317. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed-model

association for biobank-scale datasets. *Nat. Genet.* **50**, 906–908 (2018).

318. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).

319. Brown, B. C., Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, Ye, C. J., Price, A. L. & Zaitlen, N. Transethnic Genetic-Correlation Estimates from Summary Statistics. *Am. J. Hum. Genet.* **99**, 76–88 (2016).

320. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

321. Friedman, D. S. *et al.* Prevalence of age-related macular degeneration in the United States. *Arch. Ophthalmol.* **122**, 564–572 (2004).

322. Klein, R., Klein, B. E. & Cruickshanks, K. J. The prevalence of age-related maculopathy by geographic region and ethnicity. *Prog. Retin. Eye Res.* **18**, 371–389 (1999).

323. Mitchell, P., Smith, W., Attebo, K. & Wang, J. J. Prevalence of Age-related Maculopathy in Australia: The Blue Mountains Eye Study. *Ophthalmology* **102**, 1450–1460 (1995).

324. Seddon, J. M., Cote, J., Page, W. F., Aggen, S. H. & Neale, M. C. The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences. *Arch. Ophthalmol.* **123**, 321–327 (2005).

325. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).

326. Kvale, M. N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051–1060 (2015).

327. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285–1295 (2015).

328. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat. Genet.* **48**, 1284–1287 (2016).

329. Bycroft, C., Freeman, C., Petkova, D. & Band, G. Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017) doi:10.1101/166298.

330. Mitchell, P., Wang, J. J., Foran, S. & Smith, W. Five-year incidence of age-related maculopathy lesions: the Blue Mountains Eye Study. *Ophthalmology* **109**, 1092–1097 (2002).

331. Wang, J. J. *et al.* Ten-year incidence and progression of age-related maculopathy: the blue Mountains Eye Study. *Ophthalmology* **114**, 92–98 (2007).

332. Klein, R. *et al.* The Wisconsin age-related maculopathy grading system. *Ophthalmology* **98**, 1128–1134 (1991).

333. Blom, A. M., Kask, L. & Dahlbäck, B. CCP1–4 of the C4b-binding protein α-chain are required for factor I mediated cleavage of complement factor C3b. *Mol. Immunol.* **39**, 547–556 (2003).

334. Lambris, J. D., Ricklin, D. & Geisbrecht, B. V. Complement evasion by human pathogens. *Nat. Rev. Microbiol.* **6**, 132–142 (2008).

335. Liu, L. *et al.* Lycium barbarum polysaccharides protected human retinal pigment epithelial cells against oxidative stress-induced apoptosis. *Int. J. Ophthalmol.* **8**, 11–16 (2015).

336. Kaarniranta, K. & Salminen, A. NF-kappaB signaling as a putative target for omega-3 metabolites in the prevention of age-related macular degeneration (AMD). *Exp. Gerontol.* **44**, 685–688 (2009).

337. Burstedt, M. S. *et al.* Ocular phenotype of bothnia dystrophy, an autosomal recessive retinitis pigmentosa associated with an R234W mutation in the RLBP1 gene. *Arch. Ophthalmol.* **119**, 260–267 (2001).

338. Scimone, C. *et al.* A novel RLBP1 gene geographical area-related mutation present in a young patient with retinitis punctata albescens. *Hum. Genomics* **11**, 18 (2017).

339. Ratnapriya, R. *et al.* Author Correction: Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat. Genet.* (2019) doi:10.1038/s41588-019-0430-y.

340. Ma, K. N., Cashman, S. M., Sweigard, J. H. & Kumar-Singh, R. Decay accelerating factor (CD55)-mediated attenuation of complement: therapeutic implications for age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **51**, 6776–6783 (2010).

341. Fan, Q. *et al.* Meta-analysis of gene–environment-wide association scans accounting for education level identifies additional loci for refractive error. *Nat. Commun.* **7**, 11008 (2016).

342. Wagner, A. H. *et al.* Exon-level expression profiling of ocular tissues. *Exp. Eye Res.* **111**, 105–111 (2013).

343. Mencarelli, M. A. *et al.* Clinical and molecular characterization of a patient with a 2q31.2-32.3 deletion identified by array-CGH. *Am. J. Med. Genet. A* **143A**, 858–865 (2007).

344. Wu, T., Chen, Y., Chiang, S. K. S. & Tso, M. O. M. NF-κB Activation in Light-Induced Retinal Degeneration in a Mouse Model. *Invest. Ophthalmol. Vis. Sci.* **43**, 2834–2840 (2002).

345. Jeganathan, V. S. E. *et al.* Retinal vascular caliber and age-related macular degeneration: the Singapore Malay Eye Study. *Am. J. Ophthalmol.* **146**, 954–9.e1 (2008).

346. Sun, C., Wang, J. J., Mackey, D. A. & Wong, T. Y. Retinal vascular caliber: systemic, environmental, and genetic associations. *Surv. Ophthalmol.* **54**, 74–95 (2009).

347. Scerri, T. S. *et al.* Genome-wide analyses identify common variants associated with macular telangiectasia type 2. *Nat. Genet.* **49**, 559–567 (2017).

348. Tanabe, C. *et al.* ADAM19 is tightly associated with constitutive Alzheimer's disease APP alpha-secretase in A172 cells. *Biochem. Biophys. Res. Commun.* **352**, 111–117 (2007).

349. Masuzzo, A., Dinet, V., Cavanagh, C., Mascarelli, F. & Krantic, S. Amyloidosis in Retinal Neurodegenerative Diseases. *Front. Neurol.* **7**, 127 (2016).

350. Cohen-Tayar, Y. *et al.* Pax6 regulation of Sox9 in the mouse retinal pigmented epithelium controls its timely differentiation and choroid vasculature development. *Development* **145**, (2018).

351. Rajkumar, A. *et al.* ACSL5 genotype influence on fatty acid metabolism: a cellular, tissue, and whole-body study. *Metabolism* **83**, 271–279 (2018).

352. Bharadwaj, A. S. *et al.* Role of the retinal vascular endothelial cell in ocular disease. *Prog. Retin. Eye Res.* **32**, 102–180 (2013).

353. Persad, P. J. *et al.* Joint Analysis of Nuclear and Mitochondrial Variants in Age-Related Macular Degeneration Identifies Novel Loci TRPM1 and ABHD2/RLBP1. *Invest. Ophthalmol. Vis. Sci.* **58**, 4027–4038 (2017).

354. Sturgill, G. M. *et al.* Mutation screen of the cone-specific gene, CLUL1, in 376 patients with age-related macular degeneration. *Ophthalmic Genet.* **27**, 151–155 (2006).

355. Winkler, T. W. *et al.* Investigating the modulation of genetic effects on late AMD by age and sex: Lessons learned and two additional loci. *PLoS One* **13**, e0194321 (2018).

356. Jaffe, G. J. *et al.* Dual Antagonism of PDGF and VEGF in Neovascular Age-Related Macular Degeneration: A Phase IIb, Multicenter, Randomized Controlled Trial. *Ophthalmology* **124**, 224–234 (2017).

357. Siedlecki, J. *et al.* Combined VEGF and PDGF inhibition for neovascular AMD: anti-angiogenic properties of axitinib on human endothelial cells and pericytes in vitro. *Graefes Arch. Clin. Exp. Ophthalmol.* **255**, 963–972 (2017).

358. Farsiu, S. *et al.* Quantitative classification of eyes with and without intermediate age-related macular degeneration using optical coherence tomography. *Ophthalmology* **121**, 162–172 (2014).

359. Oliver, P. L. *et al.* Disruption of Visc-2, a Brain-Expressed Conserved Long Noncoding RNA, Does Not Elicit an Overt Anatomical or Behavioral Phenotype. *Cereb. Cortex* **25**, 3572–3585 (2015).

360. Kauppinen, A., Paterno, J. J., Blasiak, J., Salminen, A. & Kaarniranta, K. Inflammation and its role in age-related macular degeneration. *Cell. Mol. Life Sci.* **73**, 1765–1786 (2016).

361. Molins, B., Romero-Vázquez, S., Fuentes-Prior, P., Adan, A. & Dick, A. D. C-Reactive Protein as a Therapeutic Target in Age-Related Macular Degeneration. *Front. Immunol.* **9**, 808 (2018).

362. Pepys, M. B. & Hirschfield, G. M. C-reactive protein: a critical update. *J. Clin. Invest.* **111**, 1805–1812 (2003).

363. Seddon, J. M., Gensler, G., Milton, R. C., Klein, M. L. & Rifai, N. Association between C-reactive protein and age-related macular degeneration. *JAMA* **291**, 704–710 (2004).

364. Seddon, J. M., George, S., Rosner, B. & Rifai, N. Progression of age-related macular degeneration: prospective assessment of C-reactive protein, interleukin 6, and other cardiovascular biomarkers. *Arch. Ophthalmol.* **123**, 774–782 (2005).

365. McGwin, G., Hall, T. A., Xie, A. & Owsley, C. The relation between C reactive protein and age related macular degeneration in the Cardiovascular Health Study. *Br. J. Ophthalmol.* **89**, 1166–1170 (2005).

366. Wu, K. H. C. *et al.* Circulating inflammatory markers and hemostatic factors in age-related maculopathy: a population-based case-control study. *Invest. Ophthalmol. Vis. Sci.* **48**, 1983–1988 (2007).

367. Klein, R. *et al.* Inflammation, complement factor h, and age-related macular degeneration: the Multi-ethnic Study of Atherosclerosis. *Ophthalmology* **115**, 1742–1749 (2008).

368. Mitta, V. P. *et al.* C-reactive protein and the incidence of macular degeneration: pooled analysis of 5 cohorts. *JAMA Ophthalmol.* **131**, 507–513 (2013).

369. Yip, J. L. Y. *et al.* Cross Sectional and Longitudinal Associations between Cardiovascular Risk Factors and Age Related Macular Degeneration in the EPIC-Norfolk Eye Study. *PLoS One* **10**, e0132565 (2015).

370. Despriet, D. D. G. *et al.* Complement factor H polymorphism, complement activators, and risk of age-related macular degeneration. *JAMA* **296**, 301–309 (2006).

371. Schaumberg, D. A. *et al.* A prospective assessment of the Y402H variant in complement factor H, genetic variants in C-reactive protein, and risk of age-related macular degeneration. *Invest. Ophthalmol. Vis. Sci.* **47**, 2336–2340 (2006).

372. Kim, I. K. *et al.* Comprehensive analysis of CRP, CFH Y402H and environmental risk factors on risk of neovascular age-related macular degeneration. *Mol. Vis.* **14**, 1487–1495 (2008).

373. Cipriani, V. *et al.* Association of C-Reactive Protein Genetic Polymorphisms With Late Age-Related Macular Degeneration. *JAMA Ophthalmol.* **135**, 909–916 (2017).

374. Pingault, J.-B. *et al.* Using genetic data to strengthen causal inference in observational research. *Nat. Rev. Genet.* **19**, 566–580 (2018).

375. Davey Smith, G. & Hemani, G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23**, R89–98 (2014).

376. Ligthart, S. *et al.* Genome Analyses of >200,000 Individuals Identify 58 Loci for Chronic Inflammation and Highlight Pathways that Link Inflammation and Complex Disorders. *Am. J. Hum. Genet.* **103**, 691–706 (2018).

377. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

378. Brion, M.-J. A., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int. J. Epidemiol.* **42**, 1497–1501 (2013).

379. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).

380. Burgess, S. & Thompson, S. G. Interpreting findings from Mendelian randomization using the MR-Egger method. *Eur. J. Epidemiol.* **32**, 377–389 (2017).

381. Burgess, S., Bowden, J., Fall, T., Ingelsson, E. & Thompson, S. G. Sensitivity Analyses for Robust Causal Inference from Mendelian Randomization Analyses with Multiple Genetic Variants. *Epidemiology* **28**, 30–42 (2017).

382. Burgess, S. & Thompson, S. G. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *Am. J. Epidemiol.* **181**, 251–260 (2015).

383. Hemani, G. *et al.* The MR-Base platform supports systematic causal inference across the human phenome. *Elife* **7**, (2018).

384. Burgess, S. *et al.* Dissecting Causal Pathways Using Mendelian Randomization with Summarized Genetic Data: Application to Age at Menarche and Risk of Breast Cancer. *Genetics* **207**, 481–487 (2017).

385. Yavorska, O. O. & Burgess, S. MendelianRandomization: an R package for performing Mendelian randomization analyses using summarized data. *Int. J. Epidemiol.* **46**, 1734–1739 (2017).

386. de Jong, P. T. V. M. Age-related macular degeneration. *N. Engl. J. Med.* **355**, 1474–1485 (2006).

387. Pauleikhoff, D., Harper, C. A., Marshall, J. & Bird, A. C. Aging changes in Bruch's membrane. A histochemical and morphologic study. *Ophthalmology* **97**, 171–178 (1990).

388. Curcio, C. A., Johnson, M., Rudolf, M. & Huang, J.-D. The oil spill in ageing Bruch membrane. *Br. J. Ophthalmol.* **95**, 1638–1645 (2011).

389. Curcio, C. A. Soft Drusen in Age-Related Macular Degeneration: Biology and Targeting Via the Oil Spill Strategies. *Invest. Ophthalmol. Vis. Sci.* **59**, AMD160–AMD181 (2018).

390. Wang, L. *et al.* Abundant lipid and protein components of drusen. *PLoS One* **5**, e10329 (2010).

391. Age-Related Eye Disease Study Research Group. Risk factors associated with age-related macular degeneration. A case-control study in the age-related eye disease study: Age-Related Eye Disease Study Report Number 3. *Ophthalmology* **107**, 2224–2232 (2000).

392. Ridker, P. M., Rifai, N., Cook, N. R., Bradwin, G. & Buring, J. E. Non–HDL Cholesterol, Apolipoproteins A-I and B100, Standard Lipid Measures, Lipid Ratios, and CRP as Risk Factors for Cardiovascular Disease in Women. *JAMA* **294**, 326–333 (2005).

393. Park, J.-H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).

394. Sanderson, E., Davey Smith, G., Bowden, J. & Munafò, M. R. Mendelian randomisation analysis of the effect of educational attainment and cognitive ability on smoking behaviour. *Nat. Commun.* **10**, 2949 (2019).

395. Zuber, V. *et al.* High-throughput multivariable Mendelian randomization analysis prioritizes apolipoprotein B as key lipid risk factor for coronary artery disease. *medRxiv* (2020).

396. Liu, D. J. *et al.* Exome-wide association study of plasma lipids in >300,000 individuals.

Nat. Genet. **49**, 1758–1766 (2017).

397. Wu, Y. *et al.* Genome-wide association study of medication-use and associated disease in the UK Biobank. *Nat. Commun.* **10**, 1891 (2019).

398. Xu, Q., Cao, S., Rajapakse, S. & Matsubara, J. A. Understanding AMD by analogy: systematic review of lipid-related common pathogenic mechanisms in AMD, AD, AS and GN. *Lipids Health Dis.* **17**, 3 (2018).

399. Angelica, M. D. & Fong, Y. HDL function, dysfunction, and reverse cholesterol transport. *Atheroscler. Thromb. Vasc. Biol* **141**, 520–529 (2008).

400. Handa, J. T., Cano, M., Wang, L., Datta, S. & Liu, T. Lipids, oxidized lipids, oxidation-specific epitopes, and Age-related Macular Degeneration. *Biochim. Biophys. Acta Mol. Cell Biol. Lipids* **1862**, 430–440 (2017).

401. G, H. B., Rao, V. S. & Kakkar, V. V. Friend Turns Foe: Transformation of Anti-Inflammatory HDL to Proinflammatory HDL during Acute-Phase Response. *Cholesterol* **2011**, 274629 (2011).

402. Neale, B. M. *et al.* Genome-wide association study of advanced age-related macular degeneration identifies a role of the hepatic lipase gene (LIPC). *Proc. Natl. Acad. Sci. U. S. A.* **107**, 7395–7400 (2010).

403. Complications of Age-related Macular Degeneration Prevention Trial (CAPT) Research Group. Risk factors for choroidal neovascularization and geographic atrophy in the complications of age-related macular degeneration prevention trial. *Ophthalmology* **115**, 1474–9, 1479.e1–6 (2008).

404. Friberg, T. R., Bilonick, R. A. & Brennen, P. Is drusen area really so important? An assessment of risk of conversion to neovascular AMD based on computerized measurements of drusen. *Invest. Ophthalmol. Vis. Sci.* **53**, 1742–1751 (2012).

405. García-Layana, A., Cabrera-López, F., García-Arumí, J., Arias-Barquet, L. & Ruiz-

Moreno, J. M. Early and intermediate age-related macular degeneration: update and clinical review. *Clin. Interv. Aging* **12**, 1579–1587 (2017).

406. Berglund, L. & Ramakrishnan, R. Lipoprotein (a) an elusive cardiovascular risk factor. *Arterioscler. Thromb. Vasc. Biol.* **24**, 2219–2226 (2004).
407. Clarke, R. *et al.* Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N. Engl. J. Med.* **361**, 2518–2528 (2009).
408. Moriarty, P. M., Varvel, S. A., Gordts, P. L. S. M., McConnell, J. P. & Tsimikas, S. Lipoprotein(a) Mass Levels Increase Significantly According to APOE Genotype: An Analysis of 431 239 Patients. *Arterioscler. Thromb. Vasc. Biol.* **37**, 580–588 (2017).
409. Vavvas, D. G. *et al.* Regression of Some High-risk Features of Age-related Macular Degeneration (AMD) in Patients Receiving Intensive Statin Treatment. *EBioMedicine* **5**, 198–203 (2016).
410. Schooling, C. M., Lopez, P., Au Yeung, S. L. & Huang, J. V. Bias from competing risk before recruitment in Mendelian Randomization studies of conditions with shared etiology. *bioRxiv* 716621 (2019) doi:10.1101/716621.
411. Weinreb, R. N., Aung, T. & Medeiros, F. A. The Pathophysiology and Treatment of Glaucoma: A Review. *JAMA* **311**, 1901–1911 (2014).
412. Sommer, A. *et al.* Relationship between intraocular pressure and primary open angle glaucoma among white and black Americans. The Baltimore Eye Survey. *Arch. Ophthalmol.* **109**, 1090–1095 (1991).
413. Kamal, D. & Hitchings, R. Normal tension glaucoma—a practical approach. *Br. J. Ophthalmol.* **82**, 835–840 (1998).
414. Minegishi, Y., Nakayama, M., Iejima, D., Kawase, K. & Iwata, T. Significance of optineurin mutations in glaucoma and other diseases. *Prog. Retin. Eye Res.* **55**, 149–181 (2016).

415. Ritch, R. *et al.* TBK1 gene duplication and normal-tension glaucoma. *JAMA Ophthalmol.* **132**, 544–548 (2014).

416. Green, C. M. *et al.* How significant is a family history of glaucoma? Experience from the Glaucoma Inheritance Study in Tasmania. *Clin. Experiment. Ophthalmol.* **35**, 793–799 (2007).

417. Wu, J. *et al.* Disease severity of familial glaucoma compared with sporadic glaucoma. *Arch. Ophthalmol.* **124**, 950–954 (2006).

418. Visscher, P. M., Hill, W. G. & Wray, N. R. Heritability in the genomics era — concepts and misconceptions. *Nature Reviews Genetics* vol. 9 255–266 (2008).

419. Nannini, D. R., Kim, H., Fan, F. & Gao, X. Genetic Risk Score Is Associated with Vertical Cup-to-Disc Ratio and Improves Prediction of Primary Open-Angle Glaucoma in Latinos. *Ophthalmology* (2018) doi:10.1016/j.ophtha.2017.12.014.

420. Tham, Y. C. *et al.* Aggregate Effects of Intraocular Pressure and Cup-to-Disc Ratio Genetic Variants on Glaucoma in a Multiethnic Asian Population. *Ophthalmology* **122**, 1149–1157 (2015).

421. Khawaja, A. P. & Viswanathan, A. C. Are we ready for genetic testing for primary open-angle glaucoma? *Eye* **32**, 877–883 (2018).

422. Berry, D. A. *et al.* BRCAPRO validation, sensitivity of genetic testing of BRCA1/BRCA2, and prevalence of other breast cancer susceptibility genes. *J. Clin. Oncol.* **20**, 2701–2712 (2002).

423. King, M.-C., Marks, J. H., Mandell, J. B. & New York Breast Cancer Study Group. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643–646 (2003).

424. Martin, A. R. *et al.* Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations. *Am. J. Hum. Genet.* **100**, 635–649 (2017).

425. Duncan, L. *et al.* Analysis of polygenic risk score usage and performance in diverse human populations. *Nat. Commun.* **10**, 3328 (2019).
426. Genetics of Glaucoma in People of African Descent (GGLAD) Consortium *et al.* Association of Genetic Variants With Primary Open-Angle Glaucoma Among Individuals With African Ancestry. *JAMA* **322**, 1682–1691 (2019).
427. Gharahkhani, P., Jorgenson, E., Hysi, P. & Khawaja, A. P. 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 …. *BioRxiv* (2020).
428. Martin, A. R. *et al.* Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet.* **51**, 584–591 (2019).
429. McGuire, A. L. *et al.* The road ahead in genetics and genomics. *Nat. Rev. Genet.* **21**, 581–596 (2020).
430. Ramsay, M., Sankoh, O. & as members of the AWI-Gen study and the H3Africa Consortium. African partnerships through the H3Africa Consortium bring a genomic dimension to longitudinal population studies on the continent. *Int. J. Epidemiol.* **45**, 305–308 (2016).
431. Kowalski, M. H. *et al.* Use of >100,000 NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium whole genome sequences improves imputation quality and detection of rare variant associations in admixed African and Hispanic/Latino populations. *PLoS Genet.* **15**, e1008500 (2019).
432. Birney, E., Vamathevan, J. & Goodhand, P. Genomics in healthcare: GA4GH looks to 2022. *bioRxiv* 203554 (2017) doi:10.1101/203554.
433. Saunders, G. *et al.* Leveraging European infrastructures to access 1 million human genomes by 2022. *Nat. Rev. Genet.* **20**, 693–701 (2019).
434. Shilo, S., Rossman, H. & Segal, E. Axes of a revolution: challenges and promises of

big data in healthcare. *Nat. Med.* **26**, 29–38 (2020).

435. Chen, Z. *et al.* China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol.* **40**, 1652–1666 (2011).

436. Nagai, A. *et al.* Overview of the BioBank Japan Project: Study design and profile. *J. Epidemiol.* **27**, S2–S8 (2017).

437. All of Us Research Program Investigators *et al.* The 'All of Us' Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).

438. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Cold Spring Harbor Laboratory* 563866 (2019) doi:10.1101/563866.

439. Borodulin, K. *et al.* Cohort Profile: The National FINRISK Study. *Int. J. Epidemiol.* **47**, 696–696i (2018).

440. Mulder, N. *et al.* H3Africa: current perspectives. *Pharmgenomics. Pers. Med.* **11**, 59–66 (2018).

441. Gaziano, J. M. *et al.* Million Veteran Program: A mega-biobank to study genetic influences on health and disease. *J. Clin. Epidemiol.* **70**, 214–223 (2016).

442. Leitsalu, L. *et al.* Cohort Profile: Estonian Biobank of the Estonian Genome Center, University of Tartu. *Int. J. Epidemiol.* **44**, 1137–1147 (2015).

443. Strain, T. *et al.* Wearable-device-measured physical activity and future health risk. *Nat. Med.* **26**, 1385–1391 (2020).

444. Poplin, R. *et al.* Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nat Biomed Eng* **2**, 158–164 (2018).

445. Mason, C. E., Porter, S. G. & Smith, T. M. Characterizing multi-omic data in systems biology. *Adv. Exp. Med. Biol.* **799**, 15–38 (2014).

446. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**,

1177932219899051 (2020).

447. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).

448. Cannon, M. E. & Mohlke, K. L. Deciphering the Emerging Complexities of Molecular Mechanisms at GWAS Loci. *Am. J. Hum. Genet.* **103**, 637–653 (2018).

449. Cong, L. *et al.* Multiplex genome engineering using CRISPR/Cas systems. *Science* **339**, 819–823 (2013).

450. van de Bunt, M. *et al.* Evaluating the Performance of Fine-Mapping Strategies at Common Variant GWAS Loci. *PLoS Genet.* **11**, e1005535 (2015).

451. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).

452. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).

453. Kichaev, G. *et al.* Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).

454. Benner, C. *et al.* FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).

455. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J. R. Stat. Soc. Series B Stat. Methodol.* **82**, 1273–1300 (2020).

456. Kichaev, G. & Pasaniuc, B. Leveraging Functional-Annotation Data in Trans-ethnic Fine-Mapping Studies. *Am. J. Hum. Genet.* **97**, 260–271 (2015).

457. Weissbrod, O. *et al.* Functionally-informed fine-mapping and polygenic localization of complex trait heritability. *Cold Spring Harbor Laboratory* 807792 (2019) doi:10.1101/807792.

458. Gamazon, E. R. *et al.* A gene-based association method for mapping traits using

reference transcriptome data. *Nat. Genet.* **47**, 1091–1098 (2015).

459. Mancuso, N. *et al.* Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).

460. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).

461. Fachal, L. *et al.* Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. *Nat. Genet.* **52**, 56–73 (2020).

462. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).

463. Fernández-Rhodes, L. *et al.* Trans-ethnic fine-mapping of genetic loci for body mass index in the diverse ancestral populations of the Population Architecture using Genomics and Epidemiology (PAGE) Study reveals evidence for multiple signals at established loci. *Hum. Genet.* **136**, 771–800 (2017).

464. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

465. Han, B. & Eskin, E. Interpreting meta-analyses of genome-wide association studies. *PLoS Genet.* **8**, e1002555 (2012).

466. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).

467. Wang, X. *et al.* Comparing methods for performing trans-ethnic meta-analysis of genome-wide association studies. *Hum. Mol. Genet.* **22**, 2303–2311 (2013).

468. Sullivan, P. F. & Geschwind, D. H. Defining the Genetic, Genomic, Cellular, and Diagnostic Architectures of Psychiatric Disorders. *Cell* **177**, 162–183 (2019).

469. Suhre, K., McCarthy, M. I. & Schwenk, J. M. Genetics meets proteomics: perspectives for large population-based studies. *Nat. Rev. Genet.* (2020) doi:10.1038/s41576-020-

0268-2.

470. Richardson, T. G. *et al.* Systematic Mendelian randomization framework elucidates hundreds of CpG sites which may mediate the influence of genetic variants on disease. *Hum. Mol. Genet.* **27**, 3293–3304 (2018).

471. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *Nat. Genet.* **52**, 1122–1131 (2020).

472. Qin, Y. *et al.* Genome-wide association and Mendelian randomization analysis prioritizes bioactive metabolites with putative causal effects on common diseases. doi:10.1101/2020.08.01.20166413.

473. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nature Methods* vol. 6 377–382 (2009).

474. Qi, L. S. *et al.* Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. *Cell* **152**, 1173–1183 (2013).

475. Gilbert, L. A. *et al.* CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. *Cell* **154**, 442–451 (2013).

476. Schrode, N. *et al.* Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* **51**, 1475–1485 (2019).

477. Adli, M. The CRISPR tool kit for genome editing and beyond. *Nat. Commun.* **9**, 1911 (2018).

478. Doench, J. G. Am I ready for CRISPR? A user's guide to genetic screens. *Nat. Rev. Genet.* **19**, 67–80 (2018).

479. Fernando, M. B., Ahfeldt, T. & Brennand, K. J. Modeling the complex genetic architectures of brain disease. *Nat. Genet.* **52**, 363–369 (2020).

480. Choi, S. W., Mak, T. S.-H. & O'Reilly, P. F. Tutorial: a guide to performing polygenic risk score analyses. *Nat. Protoc.* **15**, 2759–2772 (2020).

481. Wand, H. *et al.* Improving reporting standards for polygenic scores in risk prediction studies. *medRxiv* (2020).

482. Payne, K., Gavan, S. P., Wright, S. J. & Thompson, A. J. Cost-effectiveness analyses of genetic and genomic diagnostic tests. *Nat. Rev. Genet.* **19**, 235–246 (2018).

483. Naber, S. K. *et al.* Cost-Effectiveness of Risk-Stratified Colorectal Cancer Screening Based on Polygenic Risk: Current Status and Future Potential. *JNCI Cancer Spectr* **4**, kz086 (2020).

484. Stoeklé, H.-C., Mamzer-Bruneel, M.-F., Vogt, G. & Hervé, C. 23andMe: a new two-sided data-banking market model. *BMC Med. Ethics* **17**, 19 (2016).

485. Callender, T. *et al.* Polygenic risk-tailored screening for prostate cancer: A benefit-harm and cost-effectiveness modelling study. *PLoS Med.* **16**, e1002998 (2019).

486. Damask, A. *et al.* Patients With High Genome-Wide Polygenic Risk Scores for Coronary Artery Disease May Receive Greater Clinical Benefit From Alirocumab Treatment in the ODYSSEY OUTCOMES Trial. *Circulation* **141**, 624–636 (2020).