THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Complex trait genetics: mapping, correlation and causation

Gabriel Cuellar Partida

B.Sc.

A thesis submitted for the degree of Doctor of Philosophy at

The University of Queensland in 2016

School of Medicine

Abstract

Efforts to understand the genetic aetiology of complex traits have gained a lot of momentum in the last decade. Advancement of next generation sequencing and the ever-decreasing price of genotyping platforms have allowed us to carry out a vast number of genome wide association studies (GWAS). Until relatively recently, GWAS was guided by the common disease – common genetic variation paradigm. However, recent findings and developments have made us look at the bigger picture, including rare genetic variation. In addition, methodological developments are guiding the translation of GWAS findings. For instance, diverse statistical methods can be applied on genetically informative data to estimate the genetic correlation between complex diseases. The latter can have important medical implications, as genetically correlated diseases might be responsive to the same treatments. Also, approaches such as Mendelian randomization (MR) can help investigations of causal factors in disease when is unfeasible to carry out randomized control trials.

My focus is on the application of statistical methods in complex trait genetics – I cover a range of phenotypes, ranging from eye disease to cancer. I begin with a general introduction describing the methods used along this thesis as well as the important concepts. I make particular emphasis of methodological approaches and challenges during association studies of rare variants, as well as approaches used for the estimation of genetic correlation and for MR. In the latter part of the introduction, I present an overview of the traits examined and the gaps this thesis covers.

The second chapter displays a mapping study of exonic variants with central corneal thickness (CCT), an endophenotype of keratoconus. This work led to the identification of a missense mutation in *WNT10A*, associated to a 2-fold increase in risk of keratoconus.

The following three chapters interrogate epidemiological aspects of refractive error (RE) and myopia through genetic approaches. Many studies have observed a strong

correlation between myopia, time outdoors, and education level. One hypothesis to explain these associations is that less time outdoors and more education translates into more time performing near-work activities, which may promote eye elongation and the development of myopia. Another hypothesis is that light induces dopamine release, suppressing the eye elongation. Further, some studies suggested that vitamin D might play a role in the development of the condition. To investigate some of these hypotheses I carried out gene mapping and MR approaches. Chapter 3 introduces a GWAS study of conjunctival ultraviolet autofluorescence (CUVAF). CUVAF has excellent potential as a biomarker of sun exposure compared to survey data. Understanding the aetiology behind this biomarker is potentially helpful when assessing the hypotheses of sun exposure and myopia. In chapters 4 and 5, I present two MR studies assessing the causal relationship between RE (level of myopia), vitamin D and education levels. Using a sample of 37,382 individuals of European ancestry and 8,376 from Asian ancestry and SNPs in the *DHCR7* and *CYP2R1* genes as instrumental variables (IVs), we ruled out a causal association of vitamin D on RE. Chapter 5 describes the MR study of education and myopic RE. In this, using data from three different cohorts and an education level polygenic risk score derived from alleles effects from GWAS summary data, we estimated that approximately every 2 years of additional education result in an increase of myopia.

Following to chapter 6, I performed polygenic assessments of age-related macular degeneration (AMD) and primary open angle glaucoma (POAG). . Using genome-wide array data on Australian cases and controls, we estimated the array heritability of both diseases, and the variance explained by the genome-wide associated loci. Further, we assess whether there is some genetic overlap between AMD and POAG, beyond the signal seen at *ABCA1,* which at genome-wide significance level is associated to both. In addition, we investigated whether the difference in prevalence of POAG in males and females, can be due (at least in part) to genetics. Our analyses suggest that risk to POAG is conferred by many genetic variants of small effects and that the genetic overlap between POAG and AMD is not restricted to *ABCA1*. Moreover, we found evidence of genetic differences between genders in POAG.

In the last results' chapter I show the work investigating the genetic architecture of epithelial ovarian carcinomas (EOC) and its subtypes. I look into the array heritability of each subtype and their pairwise genetic correlations. Moreover, I examine their genetic overlap with risk factors including obesity, smoking behaviour, diabetes, age at menarche and height. Overall, this work shows that EOC and its subtypes do not have a large array heritability and that the genetic architecture of the subtypes is homogenous. Finally, I show evidence of a genetic overlap of EOC with obesity and diabetes.

In the last chapter, I discuss and propose future directions for the field of statistical genetics, particularly in the areas of mapping, correlation and causation.

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, 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 research higher degree 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.

Publications during candidature

Hysi, P.G., Cheng, C.Y., Springelkamp, H., Macgregor, S., Bailey, J.N., Wojciechowski, R., Vitart, V., Nag, A., Hewitt, A.W., Hohn, R., Venturini C., Mirshahi A., Ramdas W.D., Thorleifsson G., Vithana E., Khor C.C., Stefansson A.B., Liao J., Haines J.L., Amin N., Wang Y.X., Wild P.S., Ozel A.B., Li J.Z., Fleck B.W., Zeller T., Staffieri S.E., Teo Y.Y., **Cuellar-Partida G.**, *et al.* (2014) Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. *Nature genetics*, **46**, 1126-1130.

Cuellar-Partida, G., Lu, Y., Kho, P.F., Hewitt, A.W., Wichmann, H.E., Yazar, S., Stambolian, D., Bailey-Wilson, J.E., Wojciechowski, R., Wang, J.J. *et al.* (2016) Assessing the Genetic Predisposition of Education on Myopia: A Mendelian Randomization Study. *Genetic epidemiology*, 40 (1), 66-72.

Cuellar-Partida, G., Renteria, M.E. and MacGregor, S. (2015) LocusTrack: Integrated visualization of GWAS results and genomic annotation. *Source code for biology and medicine*, **10**, 1.

Cuellar-Partida, G., Springelkamp, H., Lucas, S.E., Yazar, S., Hewitt, A.W., Iglesias, A.I., Montgomery, G.W., Martin, N.G., Pennell, C.E., van Leeuwen, E.M. *et al.* (2015) WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Human molecular genetics*, 1;24(17), 5060-8..

Hibar, D.P., Stein, J.L., Renteria, M.E., Arias-Vasquez, A., Desrivieres, S., Jahanshad, N., Toro, R., Wittfeld, K., Abramovic, L., Andersson, M., Aribisala B.S., Armstrong N.J., Bernard M., Bohlken M.M., Boks M.P., Bralten J., Brown A.A., Chakravarty M.M., Chen Q., Ching C.R., **Cuellar-Partida G**,. *et al.* (2015) Common genetic variants influence human subcortical brain structures. *Nature*, **520**, 224-229.

Lu, Y., **Cuellar-Partida, G.**, Painter, J.N., Nyholt, D.R., Australian Ovarian Cancer, S., International Endogene, C., Morris, A.P., Fasching, P.A., Hein, A., Burghaus, S.

et al. (2015) Shared genetics underlying epidemiological association between endometriosis and ovarian cancer. *Human molecular genetics*, **24**, 5955-5964.

Miyake, M., Yamashiro, K., Tabara, Y., Suda, K., Morooka, S., Nakanishi, H., Khor, C.C., Chen, P., Qiao, F., Nakata, I., Akagi-Kurashige Y., Gotoh N., Tsujikawa A., Meguro A., Kusuhara S., Polasek O., Hayward C., Wright A.F., Campbell H., Richardson A.J., Schache M., Takeuchi M., Mackey D.A., Hewitt A.W., **Cuellar-Partida G.**, *et al.* (2015) Identification of myopia-associated WNT7B polymorphisms provides insights into the mechanism underlying the development of myopia. *Nature communications*, **6**, 6689.

Springelkamp, H., Iglesias, A.I., **Cuellar-Partida, G.**, Amin, N., Burdon, K.P., van Leeuwen, E.M., Gharahkhani, P., Mishra, A., van der Lee, S.J., Hewitt, A.W. *et al.* (2015) ARHGEF12 influences the risk of glaucoma by increasing intraocular pressure. *Human molecular genetics*, **24**, 2689-2699.

Strike, L.T., Couvy-Duchesne, B., Hansell, N.K., **Cuellar-Partida, G.**, Medland, S.E. and Wright, M.J. (2015) Genetics and brain morphology. *Neuropsychology review*, **25**, 63-96.

Yazar, S.*, **Cuellar-Partida, G**.*, McKnight, C.M., Quach-Thanissorn, P., Mountain, J.A., Coroneo, M.T., Pennell, C.E., Hewitt, A.W., MacGregor, S. and Mackey, D.A. (2015) Genetic and environmental factors in conjunctival UV autofluorescence. *JAMA ophthalmology*, **133**, 406-412. *Authors contributed equally.

Submitted papers during candidature

Cuellar-Partida G., Williams K., Yazar S., Guggenheim J.A., Hewitt A.W., Williams C., Wang J.J., Kho PF., CREAM consortium, Young T.L, Tideman W., Jonas J.B., Mitchell P., Wojciechowski R., Stambolian D., Hysi P.,Hammond C.J., Mackey D.A., Lucas R., MacGregor S. No evidence of a causal effect of vitamin D on degree of myopia: a Mendelian randomization study. *Under review in European Journal of Epidemiology. (Submitted in 2016) – incorporated as Chapter 5.*

Henriet Springelkamp, Adriana I Iglesias AI, Aniket Mishra, René Höhn, Robert Wojciechowski, Anthony P. Khawaja, Abhishek Nag, Ya Xing Wang, Jie Jin Wang, **Gabriel Cuellar-Partida***, et al.* New insights into genetics of primary open-angle glaucoma based on meta-analyses of intraocular 3 pressure and optic disc characteristics. *Under review in Human Molecular Genetics. (Submitted in 2016)*

Hieab HH Adams, Derrek P Hibar, Vincent Chouraki, Jason L Stein, Paul Nyquist, Miguel E Renteria, Stella Trompet, Alejandro Arias-Vasquez, Sudha Seshadri, Sylvane Desrivières, Ashley H Beecham, Neda Jahanshad, Katharina Wittfeld, Lucija Abramovic, Saud Alhusaini, Najaf Amin, Micael Andersson, Konstantinos A Arfanakis, Benjamin S Aribisala, Nicola J Armstrong, Lavinia Athanasiu, Tomas Axelsson, Alexa Beiser, Manon Bernard, Joshua C Bis, Laura ME Blanken, Susan H Blanton, Marc M Bohlken, Marco P Boks, Janita Bralten, Adam Brickman, Owen Carmichael, M Mallar Chakravarty, Ganesh Chauhan, Qiang Chen, Christopher RK Ching, **Gabriel Cuellar-Partida**, *et al.* Common genetic variation underlying human intracranial volume highlights developmental influences and continued relevance during late life. *Under review in Nature Neuroscience. (Submitted in 2015)*

Derrek P Hibar, Hieab HH Adams, Neda Jahanshad, Ganesh Chauhan, Jason L Stein, Edith Hofer, Miguel E Renteria, Joshua C Bis, Alejandro Arias-Vasquez, M Kamran Ikram, Sylvane Desrivières, Meike W Vernooij, Lucija Abramovic, Saud Alhusaini, Najaf Amin, Micael Andersson, Konstantinos Arfanakis, Benjamin S Aribisala, Nicola J Armstrong, Lavinia Athanasiu, Tomas Axelsson, Ashley H Beecham, Alexa Beiser, Manon Bernard, Susan H Blanton, Marc M Bohlken, Marco P Boks, Janita Bralten, Adam M Brickman, Owen Carmichael, M Mallar Chakravarty, Qiang Chen, Christopher RK Ching, Vincent Chouraki, Fabrice Crivello, **Gabriel Cuellar-Partida**, *et al.* Novel genetic loci associated with hippocampal volume are relevant to aging and dementia. *Under review in Nature Neuroscience. (Submitted in 2015)*

Jue-Sheng Ong, **Gabriel Cuellar-Partida**, Puya Gharahkhani, OCAC consortium, *et al*. Association Of Vitamin D Levels And Risk Of Ovarian Cancer: A Mendelian Randomization Study. Under review in *International Journal of Epidemiology. (Submitted in 2015)*

Daniel Hwang*, **Gabriel Cuellar-Partida***, Jue-Sheng Ong*, *et al*. Sweet taste perception is associated with body mass index at the phenotypic and genotypic level. *Under review in Twin Research and Human Genetics. *Authors contributed equally. (Submitted in 2016)*

Publications included in this thesis

Cuellar-Partida, G., Springelkamp, H., Lucas, S.E., Yazar, S., Hewitt, A.W., Iglesias, A.I., Montgomery, G.W., Martin, N.G., Pennell, C.E., van Leeuwen, E.M., Verhoeven V.J., Hofman A., Uitterlinden A.G., Ramdas W.D., Wolfs R.C., Vingerling J.R., Brown M.A., Mills R.A., Craig J.E., Klaver C.C., van Duijn C.M., Burdon K.P., MacGregor S., Mackey D.A. (2015) WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Human molecular genetics*, 1;24(17), 5060-8. – incorporated as Chapter 2.

Contributor	Statement of contribution
Cuellar-Partida G. (Candidate)	Designed experiments (50%) Wrote and edited the paper (70%) Statistical analyses (80%)
MacGregor S.	Designed experiments (50%) Wrote and edited paper (10%)
Springelkam H.	Statistical analyses (20%) Wrote and edited paper (10%)
Lucas, S.E., Yazar, S., Hewitt, A.W., Iglesias, A.I., Montgomery, G.W., Martin, N.G., Pennell, C.E., van	Provided samples and other materials (100%) Wrote and edited the paper (10%)

Leeuwen, E.M., Verhoeven V.J., Hofman A., Uitterlinden A.G., Ramdas W.D., Wolfs R.C., Vingerling J.R., Brown M.A., Mills R.A., Craig J.E., Klaver C.C., van Duijn C.M., Burdon K.P., Mackey D.A.	

Yazar S.*, **Cuellar-Partida G**.*, McKnight C.M., Quach-Thanissorn P., Mountain J.A., Coroneo M.T., Pennell C.E., Hewitt A.W., MacGregor S. and Mackey D.A. (2015) Genetic and environmental factors in conjunctival UV autofluorescence. *JAMA ophthalmology*, **133**, 406-412. *Authors contributed equally. – incorporated as Chapter 3.

Contributor	Statement of contribution
Cuellar-Partida G. (Candidate)	Wrote and edited the paper (40%) Statistical analyses (100%)
Yazar S.	Wrote and edited paper (40%) Collected data (40%)
MacGregor S.	Wrote and edited paper (10%) Designed experiments (50%)
Mackey D.A.	Provided samples and other materials (50%) Designed experiments (50%)
McKnight C.M., Quach-Thanissorn P., Mountain J.A., Coroneo M.T., Pennell C.E., Hewitt A.W.	Provided samples and other materials (50%) Wrote and edited the paper (10%)

Cuellar-Partida G., Lu Y., Kho P.F., Hewitt A.W., Wichmann H.E., Yazar S., Stambolian D., Bailey-Wilson J.E., Wojciechowski R., Wang J.J., Mitchell P., Mackey D.A., MacGregor S. (2016) Assessing the Genetic Predisposition of Education on Myopia: A Mendelian Randomization Study. *Genetic epidemiology*, 40 (1), 66-72. – incorporated as Chapter 4.

Contributor	Statement of contribution
Cuellar-Partida G. (Candidate)	Wrote and edited the paper (70%) Statistical analyses (70%) Designed experiments (50%)
Lu Y.	Statistical analyses (20%)
Kho P.F.	Wrote and edited paper (10%) Statistical analyses (10%)
MacGregor S.	Provided samples and other materials (50%) Wrote and edited paper (10%) Designed experiments (50%)
Hewitt A.W., Wichmann H.E., Yazar S., Stambolian D., Bailey-Wilson J.E., Wojciechowski R., Wang J.J., Mitchell P., Mackey D.A.	Provided samples and other materials (50%) Wrote and edited the paper (10%)

Cuellar-Partida G., Craig J.E, Burdon K.P., Wang J.J, Vote B.J, Souzeau E., McAllister I.L., Isaacs T., Lake S., Mackey D.A., Constable I.J., Mitchell P., Hewitt A.W., MacGregor S. (2016) Assessment of polygenic effects links primary open angle glaucoma and age-related macular degeneration. *Accepted for publication in Scientific Reports. – incorporated as Chapter 6.*

Contributor	Statement of contribution
Cuellar-Partida G. (Candidate)	Wrote and edited paper (70%) Statistical analyses (100%) Designed experiments (30%)
MacGregor S.	Wrote and edited paper (20%) Designed experiments (30%)
Craig J.E.	Designed experiments (20%) Provided samples or other materials (30%)
Hewitt A.W.	Designed experiments (20%)

	Provided samples or other materials (30%) Wrote and edited paper (5%)
All remaining authors	Provided samples and other materials (50%) Wrote and edited paper (5%)

Gabriel Cuellar-Partida, Yi Lu, Suzanne C Dixon, Australian Ovarian Cancer Study, Peter A. Fasching, Alexander Hein, Stefanie Burghaus, Matthias W. Beckmann, Diether Lambrechts, Els Van Nieuwenhuysen, Ignace Vergote, Adriaan Vanderstichele, Jennifer Anne Doherty, Mary Anne Rossing, Jenny Chang-Claude, Anja Rudolph, Shan Wang-Gohrke, Marc T. Goodman, Natalia Bogdanova, Thilo Dörk, Matthias Dürst, Peter Hillemanns, Ingo B. Runnebaum, Natalia Antonenkova, Ralf Butzow, Arto Leminen, Heli Nevanlinna, Liisa M. Pelttari, Robert P. Edwards, Joseph L. Kelley, Francesmary Modugno, Kirsten B. Moysich, Roberta B. Ness, Rikki Cannioto, Estrid Høgdall, Claus Høgdall, Allan Jensen, Graham G. Giles, Fiona Bruinsma, Susanne K. Kjaer, Michelle A.T. Hildebrandt, Dong Liang, Karen H. Lu, Xifeng Wu, Maria Bisogna, Fanny Dao, Douglas A. Levine, Daniel W. Cramer, Kathryn L. Terry, Shelley S. Tworoger, Meir Stampfer, Stacey Missmer, Line Bjorge, Helga B. Salvesen, Reidun K. Kopperud, Katharina Bischof, Katja K.H. Aben, Lambertus A. Kiemeney, Leon F.A.G. Massuger, Angela Brooks-Wilson, Sara H. Olson, Valerie McGuire, Joseph H. Rothstein, Weiva Sieh, Alice S. Whittemore, Linda S. Cook, Nhu D. Le, C. Blake Gilks, Jacek Gronwald, Anna Jakubowska, Jan Lubiński, Tomasz Kluz, Honglin Song, Jonathan P. Tyrer, Nicolas Wentzensen, Louise Brinton, Britton Trabert, Jolanta Lissowska, John R. McLaughlin, Steven A. Narod, Catherine Phelan, Hoda Anton-Culver, Argyrios Ziogas, Diana Eccles, Ian Campbell, Simon A. Gayther, Aleksandra Gentry-Maharaj, Usha Menon, Susan J. Ramus, Anna H. Wu, Agnieszka Dansonka-Mieszkowska, Jolanta Kupryjanczyk, Agnieszka Timorek, Lukasz Szafron, Julie M. Cunningham, Brooke L. Fridley, Stacey J. Winham, Elisa V. Bandera, Elizabeth M. Poole, Terry K. Morgan, Ellen L. Goode, Joellen M. Schildkraut, Celeste L. Pearce, Andrew Berchuck, Paul D. P. Pharoah, Penelope M. Webb, Georgia Chenevix-Trench, Harvey A. Risch, Stuart MacGregor. (2016) Assessing the Genetic Architecture of

Epithelial Ovarian Cancer Histological Subtypes. *Human Genetics, 1-16, doi: 10.1007/s00439-016-1663-9 – incorporated as Chapter 7.*

Contributor	Statement of contribution
Cuellar-Partida G. (Candidate)	Wrote and edited the paper (75%) Statistical analyses (100%) Designed experiments (50%)
Lu Y.	Designed experiments (20%)
MacGregor S.	Wrote and edited paper (10%) Designed experiments (30%)
Risch H.A.	Wrote and edited paper (10%) Provided samples or other materials (5%)
All remaining authors	Provided samples and other materials (95%) Wrote and edited the paper (5%)

Contributions by others to the thesis

Part of data used in Chapter 2 and 3, i.e., genotype data (including imputed data) of the BATS and TEST twin samples, were obtained or processed by Nicholas G Martin, Scott D. Gordon, Anjali K. Henders, Sarah E. Medland, Brian McEvoy, Dale R. Nyholt, Margaret J. Wright, Megan J. Campbell and Anthony Caracella.

Stuart Macgregor, provided comments on all the contents of this thesis

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

None.

Acknowledgements

The completion of this thesis would not have been possible without scholarship support from the University of Queensland (Australian Postgraduate Award, International Postgraduate Research Scholarship and UQ advantage Top-up scholarship) and QIMR Berghofer Medical Research Institute (Top-up scholarship).

I would like to thank my supervisor Stuart Macgregor for guiding me through all my PhD and for the many projects he exposed me to, allowing me to learn and explore many areas in the field, in addition to forge many collaborations. I also appreciate the degree of freedom and independence I was granted during the course of my PhD that allowed me to have fun and follow my own ideas. I also thank for the support given to attend workshops and conferences and to land on a good postdoctoral position. In brief, thank you so much for being an amazing mentor and to help me build my career.

In addition, many thanks to Sarah Medland and Miguel Renteria to permit me to be involved in projects from the ENIGMA consortium (among others) that greatly enhanced my technical skills and helped build my research experience. Also, I highly appreciate the advice given throughout this degree.

Many thanks as well for the tea-time group and friends Aniket, Puya, Yadav Sapkota, Jue Sheng, Daniel and Michael. Tea time played an essential role in the development and exchange of ideas that helped in every project and collaboration carried out, in addition to supply the needed caffeine for the day.

Thank you to Rocio, Qian, Sonia, Ivan, Nas, Virginia, Nadiah and Raidah, that have made life in Brisbane the best.

Special thanks to Pik Fang for all her support since my first research internship in Australia, followed by an internship in Singapore and finally, my PhD. Nothing of this would have happened without you. It makes me the happiest the fact you are here with me. I’m glad that you have also decide to pursue a career in research, and I’ll be there by your side helping you in whatever I can just as you have helped me.

Finally and most important, my infinite gratitude to my family for all their support during always. Thank you for the visits throughout my PhD for you advices and well, everything.

Keywords

genetics, genome-wide association studies, complex traits, heritability, genetic architecture, ophthalmology, ovarian cancer, genetic correlation, mendelian randomization, epidemiology

Australian and New Zealand Standard Research Classifications (ANZSRC)

ANZSRC code: 010401, Applied Statistics, 50%

ANZSRC code: 060412, Quantitative Genetics, 50%

Fields of Research (FoR) Classification

FoR code: 0104, Statistics, 50%

FoR code: 0606, Genetics, 50%

Table of Contents

List of Figures & Tables

List of Abbreviations used in the thesis

AMD	Age-related macular degeneration
ANZRAG	Australia and New Zealand Registry of Advanced Glaucoma
AREDS	Age-Related Eye Disease Study
BATS	Brisbane Adolescence Twin Study
BMES	Blue mountains eye study
BMI	Body mass index
CCT	Central corneal thickness
CNV	Copy number variant
CREAM	Consortium for Refractive Error and Myopia
CUVAF	Conjunctival ultraviolet autofluorescence
DZ	Dizygotic
EOC	Epithelial ovarian cancer
GCTA	Genome-wide Complex Trait Analysis
GIANT	Genetic Investigation of ANthropometric Traits
GO	Gene ontology
GRM	Genetic relationship matrix
GWAS	Genome wide association study
HWE	Hardy-Weinberg equilibrium
IBD	Identical be descent
IBS	Identical by state
IGGC	International Glaucoma Genetics Consortium
IOP	Intraocular pressure
IV	Instrumental variable
LD	Linkage disequilibrium
LMM	Linear mixed model
MAF	Minor allele frequency
MLE	Maximum likelihood estimation
MR	Mendelian randomization
MZ	Monozygotic
OCAC	Ovarian cancer association consortium

PC	Principal component
PCA	Principal component analysis
PGRS	Polygenic risk score
POAG	Primary open angle glaucoma
QC	Quality control
RCT	Randomized control trial
RE	Refractive error
REML	Restricted maximum likelihood
RS	Rotterdam study
RV	Rare Variant
SKAT	Sequence kernel association test
SNP	Single nucleotide polymorphism
SPHEQ	Spherical equivalent
SSGAC	Social Science Genetic Association Consortium
TEST	Twin eye study of Tasmania
TSLS	Two-step least squares
VEGAS	Versatile gene based association study

Chapter 1

Introduction

1.1 Complex traits and heritability

Complex traits and complex diseases are those that are influenced by the genetic makeup and the environmental exposures an individual is subjected to. In contrast to Mendelian traits that arise as result of specific changes in a particular position in the genome (hereafter referred as locus); complex traits are influenced by variations in several loci. These variations include single nucleotide polymorphisms (SNP), insertions and deletions (indel), copy number variation (CNV) and rearrangements [1-3]. The extent that these variations influence a complex trait is termed heritability (H^2), and it is defined as how much variation in a population's phenotype is due to the genetic variation among the members in that population[4, 5]:

$$H^2 = \frac{V_G}{V_P}$$

Where V_P(phenotypic variance) = V_G(genetic variance) + V_E(environmental variance).

As it can be observed in the formula, the heritability of a trait is relative and it varies between populations. For example, if we were estimating H^2 of the tail length in a population of mice in a laboratory setting, where we are able to minimize environmental differences, then, V_E would be closer to 0, and the resultant H^2 would be closer to 1. In contrast, a scenario where we were estimating H^2 of a disease in a human population with vast amount of variability in the environment, and low genetic variety between the disease status groups, H^2 would tend to 0.

Mechanisms on how the genetic component can influence a phenotype include additive genetic effects, dominance effects and gene-gene interactions (i.e. epistasis) [4]. Additive genetic effects (A), describe a fixed value contribution to a quantitative trait measure per each effect allele. For example, if each allele “A” in a specific locus contributes 1cm to height, then, a homozygote “AA” would translate to a gain of 2cm in height. Dominance effects (D) are the effects of those alleles “A” capable of masking the contribution of the recessive alleles “a” (i.e. the heterozygote “Aa” and the homozygote “AA” would exhibit the same phenotype). The gene-gene interactions also called epistasis (I) is where the effect of a gene depends on the presence of another gene. Altogether, these effects contribute to the total genetic variance, which we can expand as follow:

$$V_G = V_A(additive\ genetic\ variance) + V_D(dominance\ genetic\ variance) \\ + V_I(interaction\ genetic\ variance)$$

In order to estimate heritability in a population, variation of a trait must be partitioned into components that most of the times represent unmeasured genetic and environmental factors. One simple way to estimate the heritability would be to simply regress the children’s trait versus the mean of their parents, with whom they share half of their genome[6]. In an unrealistic scenario, where environmental factors are completely different between parents and children, the heritability would be the slope of this regression. However, this is not the case, and parents and offspring share a lot more beside half of their genome; they share a common environment (C) which confounds the estimate. In order to circumvent this, heritability is often estimated comparing the phenotypic concordance of monozygotic (MZ) twins versus dizygotic (DZ) twins [4, 6]. The advantage of this approach is that each of these pairs are expected to share all environmental factors, including those while in the womb, allowing to isolate the contribution of the shared genome to phenotypic concordance[4].

A caveat of this approach is that although DZ twins share on average 50% of their genome which translates into 50% sharing of additive effects, only around 25% of dominance effects and possible gene-gene interactions are shared between pairs. In order to accurately estimate H^2, we should be able to estimate the V_A, V_D and V_I from the predicted covariances among twins:

$$Cov_{MZ} = \begin{pmatrix} A+D+I+C+E & A+D+I+C \\ A+D+I+C & A+D+I+C+E \end{pmatrix}$$

$$Cov_{DZ} = \begin{pmatrix} A+D+I+C+E & 0.5A+0.25D+0.25I+C \\ 0.5A+0.25D+0.25I+C & A+D+I+C+E \end{pmatrix}$$

Nonetheless, is trivial to see that given five parameters and just three different equations, it is not possible to estimate all the parameters. As a result of this, and given that dominant effects cannot be passed on through generations (i.e. requires sharing both chromosomes)[6], heritability estimates are often calculated based solely in the additive genetic component $h^2 = \frac{V_A}{V_P}$. The h^2 estimate is termed the narrow-sense heritability, while $H^2$is called the broad-sense heritability[4].

During the last decade, where new technologies and massive reductions in costs have allowed us to type the genetic variation in big population samples, approaches that use this kind of data to estimate h^2have surged. These approaches are mainly based on the estimation of the proportion of identical-by-state (IBS) allele sharing between individuals using information of directly genotyped markers to compute a genetic relationship matrix (GRM). This GRM is then related to the phenotypic values by fitting the GRM as random effect in a linear-mixed model (LMM) using restricted maximum likelihood (REML) approach [7-9]. In order to get rid of potential confounding due to possible shared environment between individuals; this model should be fitted using unrelated individuals[8]. A caveat of this kind of approaches is that the GRM is not computed based on the whole genome but just typed variants, so the model is just able to estimate a proportion of h^2 which is usually called h_g^2 (i.e. the phenotypic variance explained by additive effects of genotyped markers). When estimating h_g^2in dichotomous traits (disease status) the estimated variance explained has to be transformed from the observed scale (0-1) to an unobserved continuous scale (liability) using a probit transformation. This is carried out so that h_g^2 is independent of the disease prevalence [7]. Moreover, as number of cases in case-control studies is usually higher than prevalence in the population, the estimates must be ascertainment-corrected through a further transformation.

I have very briefly shown the traditional approaches to estimate the proportion of phenotypic variance that is due to genetics, particularly additive genetic effects. The following question that arises is what do we do next? How does this estimate of heritability can help us understand complex traits and diseases?

This thesis focuses on answering these questions for a small number of traits. Here I apply a wide range of analytical approaches to interrogate parts of the genetic component of these traits. I take the space of this introduction to briefly describe the state of the approaches for this task. Further into the introduction, I describe the traits in question and the gaps of what is known about their genetic component and aetiology.

1.2 Genetic epidemiology and statistical genetics

Since centuries before Christ, where Hippocrates examined for the first time the relationship between the occurrence of disease and environmental factors[10], we have striven looking for patterns resulting on disease conditions. Modern epidemiologists collect data and design experiments in large populations in order to find correlations between exposures and disease, using different statistical approaches. In the best scenario, where the design of the experiment allows it (e.g. a randomized control trial (RCT)), it is possible to confirm or rule out a causal association between the exposure and the phenotype[11]. However, this, more than often, is not the case, and we can just observe the presence of a correlation, which at least helps us to narrow down the possible causes that trigger disease.

As described in the previous section, the genetic makeup of an individual is just another part in the big complex trait / disease puzzle. Genetic epidemiology utilizes epidemiologic methods to address the role of genetic markers in disease aetiology [5]. Once that it is known the extent of the genetic component in the trait of interest, statistical analyses are applied to diverse types of genetic data, such as genotype array data or next generation sequence in the quest to find the genetic variants underlying h^2.

The latest years have seen a fast paced development of technology, exponential growth of genetic data, and an ever growing knowledgebase about gene function. The statistical genetics field has arisen to cope with this, by focusing on the development and application of statistical methods on genetics. In comparison with genetic epidemiology, it is a heavily data-oriented discipline which relies on the exploration of big data sets to explore and answer question such as what are the genetic markers leading to disease[12].

1.3 Genome wide association studies

Genetic linkage is the propensity of alleles located physically close to each other to be inherited together [5, 13, 14]. This results from the fact that although recombination events are facilitated by chromosomal crossover during the meiosis, these events occur with small probability at any location along chromosome[14]. Moreover, the probability of recombination between two locations highly depends on the distance between them, and in humans is estimated to happen in about each 100 million base pairs[3, 14]. A linkage map is a genetic map that takes advantage of this phenomenon in order to co-localize genetic markers; for example, a greater frequency of recombination between two loci means that they are further apart, and vice versa.

Traditionally, this genetic map was necessary for identifying the location of genes that cause genetic diseases. Researchers used informative markers such as microsatellites from large pedigrees and assessed the probability that the co-segregation of the marker and the disease is due to the existence of linkage or to chance[5]. In the case of co-segregation due to linkage, it means that the causal gene is in the vicinity of the microsatellite. This kind of approach and the fast paced technology advancement gave rise to modern gene mapping.

In 2002 the international HapMap Project started developing a haplotype map of the human genome in order to describe common human genetic variation. After funding a large re-sequencing project to discover millions of additional SNPs to those already well characterized, they genotyped 269 individuals from diverse ancestry groups for all these SNPs[15].

The creation of this haplotype map was an important advancement for gene mapping studies. As previously mentioned, there is a strong correlation between alleles of nearby SNPs spanning for around 1Mb [3, 14]. This is because each SNP arose from single point mutations, and was then passed down on the chromosome surrounded by others which happened earlier. The latter, together with selection forces, rate of mutation, recombination, and genetic drift, have made SNPs to correlate with one another in segments of a chromosome at different extents. This non-random association between SNP alleles is called linkage disequilibrium (LD), which more formally is described as “the presence of statistical associations between alleles at different loci that are different from what would be expected if alleles were independently, randomly sampled based on their individual allele frequencies”[16]. The absence of LD is called linkage equilibrium.

Genome-wide association studies (GWAS) are designed based on the premise that SNPs within a certain distance (usually <500kb) are in LD (correlated) with each other. This allows scanning the whole genome for associations by genotyping just a fraction of the total number of SNPs [17]. However, this mainly applies when the study aims to look for common variation. As mentioned in the previous paragraph, LD arises when point mutations get surrounded by earlier ones, hence, rare variants may not be in high LD with others as these tend to be newer.

1.3.1 Association of common variants

GWAS is a hypothesis free approach, as its name suggests, instead of testing the association between a small number of SNPs close to candidate genes (e.g. genes with known relevant cellular functions) and a trait, its goal is to scan the whole genome. The latter causes a significant multiple testing burden, therefore, one should be cautious interpreting the significance of the results. Simulations and studies of the HapMap project have shown that even though there are tens of millions of SNPs in the genome, because of LD, the number of actual independent tests is around one million in the European population and around two million in the African population[18, 19]. Based on Bonferroni’s correction for multiple testing, an association with a p-value < $5x10^{-8}$ is deemed to be the genome-wide significance threshold[20].

To date GWAS have been performed more than 1300 different traits and diseases and around 20,000 SNPs have been associated to them with a p-value < $5x10^{-8}$ [21]. During a GWAS, single SNP association are mainly carried out through by regressing a continuous or a dichotomous trait with each of the genotyped SNPs in an additive scale (i.e. the number of minor alleles “A” 0, 1 or 2). Nonetheless, it is also possible to perform these regression by considering dominance effects (i.e. AA=Aa=1 and aa=0), or genotypic, where the homozygotes (for the major, or the minor alleles) and heterozygote are considered to have different effects.

Genetic associations can be easily confounded by population stratification where there exists differences in allele frequency between population groups in the study and where the disease prevalence or trait mean differ within these population groups. In GWAS, population stratification is usually evaluated by computing the Genomic Control λ (λ_{GC}), which corresponds to the median χ^2 (1 degree of freedom) of the association of all the SNPs, divided by the expected median χ^2 under the null distribution[22]. In practice, it is difficult to discern whether a $\lambda_{GC} > 1$ is due to population stratification or to a large signal of polygenic inheritance (i.e. many genes are associated to the trait). In most studies, a small λ_{GC} (e.g. < 1.05) is considered acceptable, suggesting no population stratification while a value above this has to be taken with caution[23]. However, is worth noting that λ_{GC} scales directly with sample size, so for bigger studies it is common practice to compute λ_{1000}, which is an extrapolation of the observed λ_{GC} to the one equivalent for a study of 1000 cases and 1000 controls [24].

Multiple approaches have been proposed to deal with population stratification. These approaches mainly differ regarding the type of stratification that they correct. In the simplest scenario, where population stratification is expected to arise as result of ancient divergence, dividing the test statistics by λ_{GC} is considered to provide enough correction [22, 23, 25]. Another commonly used and powerful approach to control for population structure is to perform principal component analysis (PCA) of the genotypes and use the top principal components (PCs) as covariates during the analyses. This approach is also capable to detect assay artefacts, which is useful when combining multiple data sets in the same study [22, 23, 25]. For bias arising due to cryptic relatedness between the individuals (e.g. families are included in the analysis) using PCs as covariates is not enough. A better approach in this scenario is to use linear mixed models (LMM) which can model family and population structure by fitting the SNP of interest and covariates as fixed effects and the generic relationship matrix (GRM) as a random effect [26].

To date, most GWAS findings agree that most complex traits and disease exhibit a large polygenic architecture, and the effect size of common SNPs are rather small which make them hard to detect. This has led GWAS to become a highly collaborative field [27] with the end of increase the number of samples by combining studies, then, increasing the power to detect associations. However, a problem with this is that it is not uncommon for the different studies to rely on different genotyping platforms, yielding different sets of SNPs. In order to circumvent this problem, genotype imputation is applied across studies. Genotype imputation is a process in which unobserved genotypes are inferred. This process takes advantage of the long range LD to infer haplotypes, and then these are compared to those haplotypes in reference samples such as those of HapMap or 1000 Genomes projects to fill the not genotyped SNPs [28]. This approach allows researchers to create a homogenous set of SNPs to perform meta-analyses with diverse studies as well as to narrow down the location of causal variants.

1.3.2 SNP set association analyses

1.3.2.1 Gene-based analyses

Although investigating association of single SNPs to disease have yielded great insight into disease aetiology, this approach has low power detecting trait associated genes when the SNP effects within a defined region or gene are small and just their cumulative effect is associated to the trait. Because of this, aside of performing associations between single SNPs and phenotypes during GWAS, SNP-set analyses has been established as complementary post-GWAS approaches [29, 30]. A key issue when performing SNP-set analysis is accounting for the correlation among SNPs (i.e. LD). In the simplest case, one could select the top associated SNP within the region of interest and apply multiple testing corrections through a Bonferroni's procedure to control the false positive rate[29]. However, this approach has low power as it does not combine the information from neighbouring SNPs which may be causal (although at a lesser extent). In order to analyze multiple SNPs, several methods have been developed. For example, if there is access to the raw phenotype and genotype data, it is possible to perform multivariable regressions (e.g. linear for quantitative traits or logistic for dichotomous) by fitting all the SNPs within a gene simultaneously (except for redundant SNPs). However, this may have low statistical power as result of the increase in degrees of freedom (which sometimes is evaded by cluster analysis of SNPs or PCA [31, 32]). Another method to combine information of multiple SNPs, is the Fisher's procedure to combine p-values; however given the unknown distribution and correlation between the SNPs, the significance must be estimated using permutations which are computationally expensive [29]. A variant of this method, which uses just the single SNP summary statistics and called versatile gene-based test (VEGAS), carries out the combination test and calculates the empirical p-value based on simulation of normal variables which are assigned values according to the LD structure (based on HapMap or 1000 Genomes references) between SNPs in the gene [30, 33]. Another, the gene-based association test that uses the extended Simes procedure to correct for multiple testing (GATES) can rapidly combine the SNP p-values, using only summary statistics and LD information. This method does not require of simulations or permutations, which makes it particularly fast; still, it is valid for SNPs in LD and is capable of weighting the SNPs based on functional information[34].

It is worth mentioning that although the approaches described above mainly focus on the integration of SNPs within a gene and flanking regions, it is possible to define functional or biological units in different ways. For example it is possible to perform these tests on fixed-width regions in the genome or focusing on regions with other functional annotations such as DNase I hypersensitivity sites or open chromatin.

1.3.2.2 Pathway-based analyses

The natural extension of these kinds of tests is pathway based analysis. A pathway describes a wide range of biological processes such a metabolism, cell cycle, development, etc. Analogous to gene-based analysis where all the SNPs within a gene are combined into a test statistic, here, all the gene-based statistics within a pathway are combined to test the relevance of a pathway in the trait of interest. Many different approaches have been developed for this endeavour, including enrichment analyses performed through rank comparison between a gene-set in the specified pathway and a permuted set (GenGen, MAGENTA)[35-37] or by a hypergeometric tests comparing significant versus non-significant genes in a pathway (INRICH, ALIGATOR)[38, 39]. These approaches tend to differ on their pathway definitions; with the vast amount of pathway databases, the most commonly used pathway definitions are those in the Gene Ontology (GO) database and the Kyoto Encyclopedia of Genes and Genomes (KEGG), although in some software like MAGENTA, other databases are used (e.g. the Molecular Signature Database and PANTHER which focus mostly in signaling pathways)[37].

A more recently developed approach: Data-driven Expression Prioritized Integration for Complex Traits (DEPICT) uses a different framework to test the association with a pathway. In addition to incorporate annotated gene sets (e.g. curated pathways, protein-protein interactions and phenotypic gene sets); it calculates for each gene a probability of belonging to each gene set based on expression patterns along thousands of microarray assays in different tissues/samples. In other words, instead of a binary setting where either a gene belongs or not to a pathway, each gene is assigned a probability of belonging to [40]. Once the gene-sets are created, DEPICT tests whether these are enriched for genes in the associated loci using precomputed GWAS on random phenotypes to estimate the background distribution [40].

1.3.3 The missing heritability problem

Even though all the approaches described in the previous sections produce invaluable knowledge on genes and genetic variants behind complex traits and disease aetiology; the variance explained by these is much smaller than the estimated additive genetic variance by twin and family estimates. This is due to many reasons, including 1) limited statistical power in the studies due to the high multiple-testing burden 2) the limited investigation carried out in structural genetic variation (e.g. CNVs and indels) and rare variants (RVs) and 3) the heritability estimates in twin studies may be inflated due to deviation of the assumptions [41, 42].

There is plenty of evidence showing that part of the missing heritability is due to our lack of power to detect causal variants of small effects. Studies have shown that fitting all the genotyped SNPs by computing the GRM and fitting this in a linear mixed model (as summarized in section 1.1) we are able to explain a much greater proportion of h^2(termed h_g^2) [8] of a trait than using just the genome-wide associated SNPs, indicating that the causal variants are there, but we just cannot discern true from false signals (i.e. hidden heritability). However, even then, this estimate fell short compared with the twins' estimate.

Till few years ago, the rationale for most GWAS was based on the common disease – common genetic variation hypothesis. Because of this, and technical limitations, most genotyping platforms just tagged variants present in more than 1% of the population. However, advances in high-throughput DNA sequencing technologies have driven the change of paradigm, allowing us to explore rare variation by either characterizing rare variants (RVs) (MAF<1%) in the genome through whole genome or exome sequencing [43]. Further, sequencing projects along with the 1000 Genomes Project identified several RVs [44] helping the creation of genotyping arrays tagging RVs.

Very recently, Jian Yang and colleagues found negligible missing heritability for height and BMI when using both common and rare variants (imputed), accounting properly for their LD and their MAF [45]. This highlights out the importance of not just considering common variants, but also investigating RVs.

1.3.4 Association of rare variants

There are many reasons to believe that RVs influence the expression and prevalence of complex traits and disease, and that these could account for a portion of the missing heritability [46]. For example, variation in exonic regions such as nonsynonymous mutations (i.e. missense and nonsense mutations) may be biologically significant but because they tend to have deleterious effects their frequencies never exceed low levels. Also, the recent expansion of the human population may have resulted in a great number of segregating, functionally relevant, rare variants that mediate a proportion of observed phenotypic variation [46]. Furthermore, the realization that there may be "synthetic associations" in some studies, where common variants appear to be associated due to linkage disequilibrium with several disease-associated rare variants, suggests that the latter might account for an important portion of phenotypic variance.

Although in recent years, sample sizes used in GWAS have increased dramatically, most studies have had a modest number of samples. This is a major problem when looking into association with RVs, given that sample size requirements may be much greater if the effect sizes are not bigger than those of common variants.

To date, very few studies have proven successful in the detection of associated rare variants to traits including insulin [47], lipids [48], blood cell counts [49] and liver disease [50]. Studies involving RVs present many statistical challenges ranging from variant calling to association testing. Statistical tests used for common variation would be underpowered to detect association of SNPs with a minor allele frequency (MAF) <1% even in studies with large sample sizes [51]. In addition, association analyses of RVs are more susceptible to genetic confounders, i.e. population structure, which is hard to control through traditional methods such as principal component analysis [52]. Moreover, independently from the technology used, genotype calling of rare variants is not as straightforward as is the calling of common variants. For example, if genotyping of RVs is carried out using a genotyping chip, only the common allele homozygote of the intensity cluster would be well populated, limiting the efficacy of calling algorithms[53].

1.3.4.1 Genotype calling considerations

Although new technologies allow the detection and genotyping of RVs through exome and whole-genome sequencing, genotyping arrays remain a more cost-effective way to interrogate SNPs identified in previous population studies[53]. However, using arrays to genotype RVs has its limitations. For example, standard calling algorithms like GenTrain, which uses a custom clustering algorithm to separate the two homozygotes and heterozygote genotypes, loses its efficacy when the frequency of one of the genotypes is low. Large sample sizes ease the problem by increasing the number of occurrences of the minor allele homozygote; however, this is not always feasible.

A now commonly used method to improve the calling of RVs, zCall [53] is implemented as a post-processing step after a default calling algorithm has been applied. Specifically, this algorithm separates the three clusters with a horizontal and a vertical line, which are defined by the mean and variance of the homozygote clusters and scaled by a z-score threshold. In the case of rare variants, where the minor allele homozygote threshold is hard to define, this is estimated by a linear regression model of common variant genotypes. To find the best z-score threshold, common sites are recalled using different values of z to find the best concordance. Finally, genotypes that were not previously called are assigned based on their position with respect to the horizontal and vertical lines.

Calling genotypes of rare SNPs is not only hard when genotyping through arrays. With sequencing technologies, individual samples are resequenced many times, and the combined reads are used to evaluate how likely a polymorphism truly exists at a particular locus [43, 54]. For most sequencing technologies, a large number of reads in rare variant loci are needed in order to avoid genotype misspecification. Genotype quality controls of sequencing technologies are dependent on the platform used, and so, are beyond the focus of this thesis where no sequencing studies were involved.

1.3.4.2 Population substructure

As mentioned in the previous sections, GWAS usually use PCA to account for population stratification by fitting into the model the first few PCs. However, to account for population stratification in analyses involving RVs, this approach may not be suitable; as the estimation of the variance-covariance matrix during PCA can become unstable for genetic loci with lower MAF, making principal components less reliable to identify population substructure [52, 55-57]. An alternative would be to use principal components generated from common variants; however, this would be problematic given that common variants are typically much older, thus, not accounting for the population substructure generated through rare variants.

Another approach proposed is to generate a score based on the deviation between the number of minor alleles in each subject and the expected number of minor alleles in the population across the genome [52]. The authors showed that this is a more sensitive approach to detect population outliers when considering just variants with a MAF < 5% than PCA. Moreover, an approach introduced by *Epstein et al* [58, 59] involves a permutation procedure during the association analysis that repeatedly shuffles the phenotype of study participants in a way that generates data sets with the same extent of confounding (in this case genetic) found in the original data. The genetic confounding is calculated in two steps. In the first step, each individual receives a stratification score based on the effect of informative SNPs (excluding the test SNP) on the phenotype. In the second step, based on its stratification score, each subject is assigned to a stratum where the association analysis is performed. The advantages of this approach, is that can be applied independently of the study and association test used. However, it may be computationally expensive.

1.3.4.3a Association tests

In the last few years, we have seen a tsunami of new approaches to circumvent the lack of power when analysing RVs. Most of these approaches are based on analysing (simultaneously) multiple SNPs (SNP-set) within a functional unit/region (e.g. gene), or within an annotated pathway. Independently of the definition of the SNP-set, the fundamental assumption of these association methods is that the frequency of observing at least one RV within each set is as high as observing a common variant, thus, improving the detection of association.

Researchers have come up with several strategies to combine the information of the different RVs within the set into a test statistic. One of the simplest approaches to achieve this involves testing the significance of a SNP-set by summarizing the test statistic of each SNP into one p-value. Similar approaches involve collapsing the multiple rare variants into a single number and performing a single univariate test. This collapsing can be achieved in several ways; for example we could use a dummy variable with values 0 or 1 depending on whether the individual has at least one rare allele, followed by the application of a univariate test [60].

Depending on the study design, several methods have been developed that involve collapsing of multiple rare variants. For quantitative traits, regression methods with collapsed variants, both assigning the same weight or a mixture of weights based on annotation or MAF to each RV and accounting for correlated RVs through permutation can be applied [60]. For case control studies, the weighted sum statistic [61], the cumulative minor allele test (CMAT) [62], and the combined multivariate and collapsing (CMC) [63] method which is an extension of the cohort allelic sum test (CAST) [64] can be applied.

An important drawback of collapsing methods is that are based on the assumption that every SNP within the tested unit confers an effect in the same direction, which is unlikely. A more probable scenario is that within the same locus, there are variants with protective and deleterious effects, which in addition, can interact with each other. Numerous methods have been designed that are able to accommodate these more realistic assumptions. Probably the method that has become most popular is the sequence kernel association test (SKAT) [65] and SKAT-O, which combines the burden and SKAT tests into a single statistics [66]. SKAT assumes a linear relationship between the phenotype and the set of rare variants. SKAT uses a variance-component score test in a linear mixed model (LMM) that assumes that the variant effects are drawn from a distribution with mean 0 and variance $w_j\tau$; where w_j is a user-specified weight for variant j, which usually is defined using the MAF just as in the weighted sum statistic and τ the variance component. The latter allows different variants to have different directions and magnitude effects, including no effects. SKAT-O realises that burden tests are more powerful when the variants within the defined functional unit have the same effect direction, while SKAT has a better performance when the variants present mixed effects; therefore, this approach maximizes power by adaptively using the data to combine optimally both tests [66]. Other approaches that combine multiple tests include the mixed effects score test (MiST) [67] and an approach based on Fisher's method for combining p-values [68]. The approach based on Fisher's method combines the p-values of the burden and variance-component test and then assess the significance through a permutation procedure, making it computationally intensive. MiST consists of a set of two score statistics, corresponding to grouping effects by variant characteristics (e.g. missense mutations) and effects of the individual variants. To evaluate the group effects, MiST include a modified version of the burden and SKAT tests in order to make these score statistics independent from the individual variant test score statistic under the null hypothesis. This modification facilitates the combination of the two score statistics by Fisher's and Tippett's methods.

Other approaches, such as the c-alpha test proposed by *Benjamin Neale et al* in 2011 [69] is equivalent to a SKAT in a case-control study where no covariates are included, and all w_j=1. However, one main difference is that it uses permutation to estimate the significance and control the presence of LD, which makes it computationally expensive for genome wide analyses. Another related general approach is the Estimated REgression Coefficients (EREC) method [70]. EREC reflects the ideas of the previously described collapsing methods and generalize them using a linear model framework; the approach combines information across the multiple variants within a gene by taking a weighted sum of minor alleles for each individual. It then relates the combined information and covariate(s) to the phenotype through an appropriate regression model. EREC derives optimal weights for each variant that theoretically would lead to the most powerful tests among all valid tests. EREC software also incorporates modified versions of the fixed-threshold and variable threshold (VT) methods [57], where the threshold is the frequency in which a rare variant is defined. It also includes the simplest collapsing method [64] (i.e. regression on number of rare mutations each subject carries) and the weighted sum statistic (WSS) [61].

Another kernel based association test for case-control studies is the kernel-based adaptive cluster (KBAC) method [71]. KBAC weights multi-site genotypes contrasting their frequencies between cases and controls. The rare admixture maximum likelihood test (RAML) [72] is a method which provides an omnibus test for joint effects of multiple variants on a phenotype. The backward support vector machine (BSVM)-based variant selection procedure is an approach that identifies informative disease-associated RVs and weight them into either risk or not risk categories [73]. Finally, the significance of the association between the disease and the informative variants remaining in the model is assessed by permutation tests.

1.3.4.3b Comparison of association tests

The power of RVs association tests tend to be low. It is clear that a method will perform well under the phenotype-causal variant model for which it was developed. In this section, I compare the approaches summarized above. Table 1 shows a summary of the attributes of the methods described above.

Table 1.1 Features of statistical tests for association between rare variants and complex traits. Variant based (V-b), Data adaptive (D-a), Combined/Optimal (C). Analytical and permutation refers to the way the significance of the association is calculated.

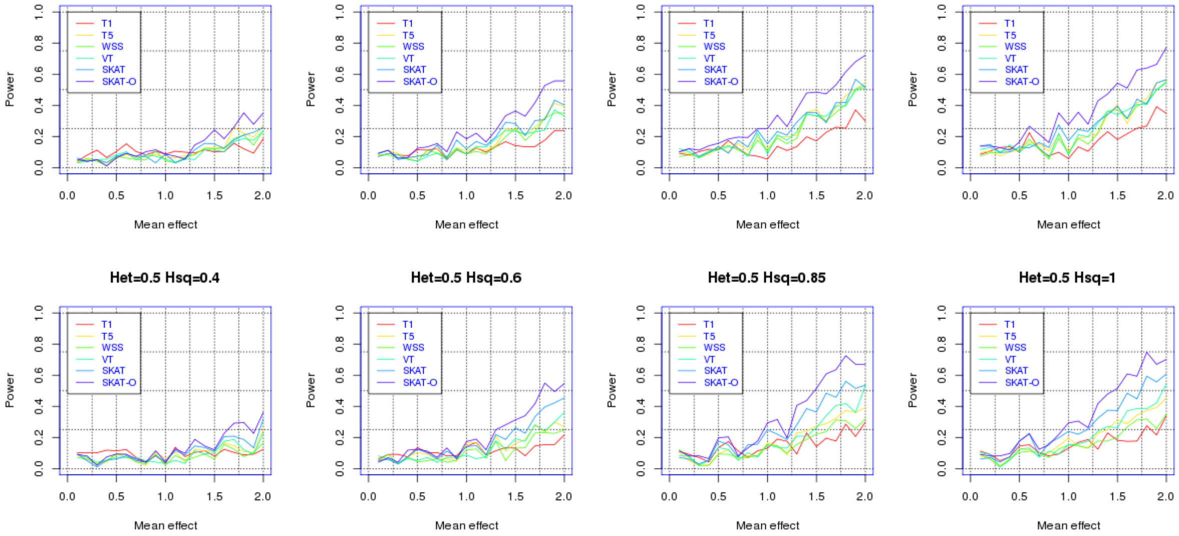
Method	C-α	BSVM	EREC	KBAC	RAML	SKAT	SKAT-O	MiST	Fisher[a]	WSS	CMC
Reference Year	*[22]* *2011*	*[26]* *2013*	*[23]* *2011*	*[25]* *2010*	*[28]* *2013*	*[21]* *2011*	*[24]* *2012*	*[23483651]* 2013	*[23032573]* 2013	*[19]* *2009*	[17] 2008
Type of method	*V-b*	*D-a*	*D-a*	*D-a*	*D-a*	*V-b*	*C*	*C*	*C*	*D-a*	*D-a*
Case/Control	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
Quantitative	*no*	*no*	*yes*	*no*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*no*
Covariates	*no*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*no*
Effect Heterogeneity	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*
Analytical	*no*	*no*	*no*	*no*	*no*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
Permutation	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*no*

In order to get a better sense of the behaviour and power of some of the most popular rare-variant association tests, I performed in-house simulations. I evaluated the approaches SKAT, SKAT-O, the weighted sum statistic (WSS), the variable (VT) and fixed threshold (T1 / T5) tests. I used exome genotype data from 4,000 individuals and randomly selected 100 different genes which had at least 4 SNPs. I simulated the effects of each SNP as drawn from a normal distribution $N(\delta,\delta^2)$. δ varied from 0.1 to 2 in steps of 0.1, each step assigning effects to 5 different genes. In other words, the SNPs within the first 5 genes had effects drawn from a $N(0.1,0.1^2)$, the SNPs of genes 6-10 had effects from a $N(0.2,0.2^2)$, and so forth. I generated phenotypes where the SNP effects explained 40%, 60%, 85% and 100% of the variability of the trait. Also, scenarios of heterogeneity (het=0.5) and not heterogeneity (het=0) were tested (het=0.5 means that I flipped the direction of the effect in 50% of the SNPs).

A good attribute of this simulation is the use of real genotype data to simulate the effects and phenotype. In addition, it does not make the assumption of effects proportional to the minor allele frequency of each variant.

Performance of each of the approaches tested is shown in Figure 1. In all the scenarios tested, SKAT-O showed the best performance. As expected, the fixed-threshold T1 and T5 approaches were the more affected since in the simulation the variants with MAF<0.01 or MAF<0.05 were not necessarily causal.

Figure 1.1 Evaluation results of different RVs association tests in different scenarios.



*Het (heterogeneity) describes the proportion of SNPs within a gene with different direction of effects. *Hsq is the variance explained by the “causal” SNPs.

1.4 Genetic correlation and pleiotropy

Pleiotropy has been noted for centuries; many Mendelian traits and diseases do not just exhibit a change in a single wild-type phenotype, but usually are accompanied by changes in other traits (i.e. they are syndromes, with multiple phenotypic effects). For example, albinism which results from the inheritance of recessive gene alleles results in a deficiency of skin, hair, and eye pigmentation but also causes defects in vision [74]. A more notable example of pleiotropy is the effect of the sex-determining region Y (*SRY*) gene that alters the expression of multiple genes that give rise to male-specific traits [75]. Formally, pleiotropy is defined as the phenomenon in which a mutation in a single locus affects at least two unrelated traits [76].

While many times the presence of pleiotropy is evident, it is challenging to discern if the genetic correlation between two traits is due to a single mutation affecting both traits (actual pleiotropy) or due to physical linkage between nearby loci. This is further hindered when the causal variant of a trait is not well mapped. However, regardless from the source of genetic correlation, estimating the proportion of variance that two traits share due to the genetic component can shed light into important aspects of the aetiology. For example, evidence of a genetic overlap between disorders can have an impact on the prospects for drug repositioning [77].

In order to estimate the genetic correlation between traits, it is necessary to have a genetic informative sample. Traditionally, the genetic overlap is estimated through a bivariate analysis in a sample of twins where the phenotypic covariance between the traits is decomposed into environmental and genetic components (similar to the estimation of heritability as described in section 1). The latter is achieved through a Cholesky decomposition of the variance components A (additive genetic effects), C (common environment), and E (unique environment), followed by estimation of the parameters through Maximum Likelihood Estimation (MLE) approach [78].

Now that it is common to have big population samples genotyped, it is no longer necessary to have pedigree data to estimate the genetic correlation. As described in previous sections, one can infer the genetic relationship between participants in one study by using the genotype data. A now commonly used approach to estimate the genetic correlation is to use the GRM of unrelated individuals (to avoid bias due to possible shared environment) in a bivariate mixed linear model and estimate the parameters using REML [79].

Another popular approach to investigate the genetic overlap between traits is the polygenic risk prediction approach. In contrast to the methods described above which require both traits to be measured in order to compute the genetic correlation, this approach can use GWAS summary results when raw data from one of the traits of interest is unavailable. Polygenic risk prediction involves the computation of a predicted trait value based on genotypes and often called “polygenic risk score” (PGRS). This PGRS is then used to examine its relationship with another trait [80]. It is computed by aggregating the estimated effects of many variants multiplied by observed genotypes into a single score for each individual. It is common practice to compute the PGRS selecting SNPs based on different significance thresholds, and removing those that are redundant through LD-clumping [80]. Subsequently, the association between the computed PGRS and the trait of interest is tested through regression analyses. If the association is significant, it means that there is a genetic correlation between the traits. Although this approach does not directly quantify the genetic correlation between traits, Dudbridge [81] proposed a procedure to estimate it which requires the estimates of parameters such as the heritability of each of the traits, number of independent SNPs, p-value of the association and sample sizes.

Since the beginning of large GWAS meta-analyses such as those from the Genetic Investigation of ANthropometric Traits (GIANT) consortium where the inclusion of hundreds of thousands of samples yielded a $\lambda_{GC} > 1.4$, it had been hard to discern whether this inflation was due to population structure bias or polygenicity of the trait. In a recently published paper, Bulik-Sullivan and colleagues [82] realize that under polygenic inheritance, variants in high LD with a causal variant will show inflation in the test statistics (i.e. the test statistic is correlated to LD), while inflation due to population structure is not correlated to LD. The latter holds as far as the allele frequency differences between subpopulations are not under strong selection. The authors showed in simulated and the PGC Schizophrenia GWAS results that it is possible to measure the amount o genetic variation tagged by the variants (h_g^2) and the level of confounding due to population structure by regressing the LD-score — defined as the sum of LD between the SNP and all SNPs within a region (usually 1 centiMorgan windows), against the χ^2 statistic. As heritability is a variance estimate, and the variance is just a special case of covariance where both random variables are the same, this approach can be extended to estimate genetic correlation. Bulik-Sullivan et al proposed the cross-trait LD score regression [83] where instead of regressing the LD score against the χ^2 statistic of a single GWAS, the regression is carried out against the product of the z scores from two studies. The latter gives an estimate of the genetic covariance (the slope of the regression) which after normalizing it by the h_g^2 gives the estimates of genetic correlation estimates.

1.5 Mendelian randomization

1.5.1 Definition and assumptions

Mendelian randomization (MR) has become a very popular approach in human genetics to try to answer questions about causality [84]. In observational studies, associations between an exposure and an outcome are subject to measurement error, confounders, reverse causation and many potential biases (e.g. recall bias, selection bias) that hinder distinguishing causal from non-causal relationships. Randomized control trials (RCTs) are often considered the gold standard in epidemiology to estimate causal relationships. An RCT is an experiment where the participants are randomly placed into the different treatment groups under study (e.g. placebo / not placebo). The randomization ensures that potential confounding factors are balanced between the groups, thus allowing unbiased estimation of the effect of the treatment in question. Although RCTs are evidently the ultimate means to estimate the causal effects of a modifiable factor on a trait, these are often not feasible to perform due to the high costs and long duration they involve or because the exposure in question can’t be applied (e.g. it would not be ethical to ask one of the groups to smoke or to gain weight). MR is an approach that can be applied to circumvent these limitations. An MR study makes use of genetic variants that are known to affect an exposure of interest (e.g. vitamin D levels, cholesterol, BMI, etc.) and thus functions as a “natural” RCT [85]. MR exploits the fact that alleles are segregated randomly during meiosis, thus ensuring that the genotypes are not related to any potential confounders. Some of the strengths of MR include that the use of genetic variants as proxies of exposures protect against reverse causation – provided that the genetic variants were adequately chosen. Also, in contrast to RCT, where the exposure of interest is administered during a relatively short time (i.e. during the trial period), MR allows to measure the long-term effects of lifetime exposure.

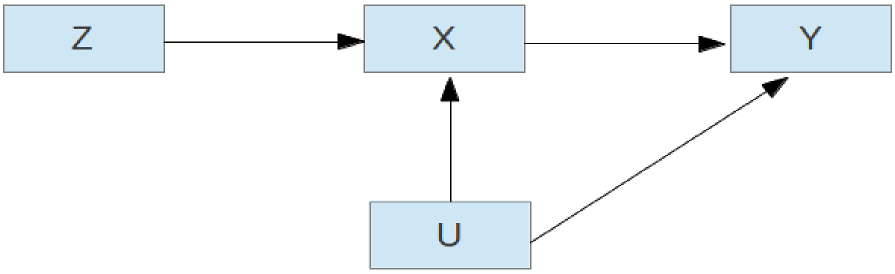
MR is tied to the three fundamental assumptions of instrumental variable analysis [86]. The first one is that the genetic variant acting as a proxy (i.e. the instrumental variable (IV)) must be robustly associated to the exposure of interest; generally with an F-statistic >10 (under the assumptions of a linear relationship between the IV and the exposure, the relative bias is approximately 1/*F*-statistic, meaning that the bias of the IV estimator is <10% of the bias of the observational estimator [87]). The second assumption states that the IV must not be associated to any of the confounders. In practice, this assumption is hard to test given the impossibility of gathering information on all the confounders. Publicly GWAS summary results can allow us to investigate the association of the IV in phenotypes not measured in the investigated sample increasing confidence that assumption 2 is met; however, is not possible to rule out a violation of the assumption completely. Finally, the third assumption stipulates that the IV is not directly associated to the outcome, but mediated by the exposure of interest. This assumption is also hard to prove for certain as is always possible that the variant affects the outcome via other pathway. In order to alleviate this possibility, ideally, the function of the genetic variants and genes affecting the exposure should be well characterized. Other strategies involve the use of multiple genetic variants with known effects on the exposure and fit different models [88]. In the case the models throw the same conclusion of causality, this will increase the confidence that assumption 3 is met and vice versa.

There are many ways the three assumptions summarized above can be violated. For example, genetic instruments might be weak. In this case, and assuming that increase of sample size is not possible, effect estimates from GWAS could be used to compute a PGRS and use this as an IV [89], as was done in the project displayed in chapter 4. Although a PGRS increases power as it explains a greater proportion of variance of the exposure, this IV is more susceptible to violations of the MR assumptions, as the more genetic variants used, the more likely one of those is going to be associated to other biological pathways in addition to the one of interest. Population stratification can also bias the estimate if the allele frequency of the genetic instrument differs between subpopulations in which the frequency or mean of their outcome also differs. In this case, just as in GWAS, restricting the analysis to a homogeneous populations and including the genotype-derived principal components can alleviate the problem.

1.5.2 Methods

The simplest way to estimate the causal effect of a continuous exposure on a continuous outcome (assuming linearity between the variables) in an MR setting is done by computing the Wald-type ratio estimate $\beta_{IV} = \frac{\beta_{zy}}{\beta_{zx}}$ [90]. Where β_{zy} is the regression coefficient of the outcome on the IV and β_{zx} the regression coefficient of the exposure on the IV [Figure 2]. The standard errors can be then approximated using the delta method [91]. This method is advantageous when β_{zy} and β_{zx} are computed in different samples and/or are extracted just from GWAS summary statistics.

Figure 1.2. Graph depicting the MR assumptions. The instrument variable Z is associated to the outcome Y through its effects on the exposure X. Z is not affected by confounders U.



Similarly, in scenarios where there is no access to individual level data and exists the possibility to combine multiple instruments, the causal estimate can be computed through an inverse variance weighted meta-analysis [92].

$$\hat{\beta}_{ivw} = \frac{\sum \hat{\beta}_{zx}\hat{\beta}_{zy}\sigma_{zy}^{-2}}{\sum \hat{\beta}_{zx}^{2}\sigma_{zy}^{-2}}$$

$$\sigma_{ivw} = \sqrt{\frac{1}{\sum \hat{\beta}_{zx}^{2}\sigma_{zy}^{-2}}}$$

A variety of methods exist to estimate the causal effect when individual data is available. One of the most commonly used is the two-stage least squares (2SLS) [90, 93]. In the first stage of this approach, an OLS regression between the IV and the exposure is performed followed by an OLS regression between the outcome of interest and the predicted values from the first stage regression. Standard errors from the second stage regression should then be corrected to account for the uncertainty of the predicted values of the exposure. Another approach that follows the same principle as 2SLS is the control function estimator. This method is also a two-stage estimator, but in contrast to 2SLS above, this one includes the residuals estimated from the first-stage regression in the second regression [94]. The idea behind this is that the first-stage residuals may be correlated with the confounders, thus including them in the second-stage regression will help control some confounding effect on the outcome. Other estimators for IV analyses include structural mean models [95] estimated through maximum likelihood and the generalized method of moments [90] which does not make strong assumptions about the relationship between the exposure and outcome so is more suitable for binary outcomes.

The advantages of using individual level data (i.e. when the genotypes, the exposure and the outcome are available) include the possibility to test more directly the three MR assumptions and having better precision, as the error in instrument variable estimate will be smaller when estimated from the same sample [96]. However, depending on the exposure of interest, getting sufficient individual level data may be problematic - MR generally requires very big sample sizes (tens of thousands) in light of the small fraction of variance explained by the IV. Publicly-available GWAS summary data from large consortia have proven to be a valuable resource to conduct well powered MR studies in a fast and cost-effective way [97]. In this setting, the causal estimate can be computed through the Wald-type ratio estimate or the inverse variance weighted meta-analysis as described above using the effect estimates and standard errors from the outcome GWAS and the known effect of the IV on the exposure. Although in this case the MR assumptions cannot be entirely tested, they can be investigated in a number of ways. These include obtaining information on the biology and function of the genetic variant, testing (wherever possible) the association of the IV with potential confounders using either individual level data or GWAS summary data, and ensuring that the effect on outcome and exposure were estimated in a population of the same ancestry (e.g. Asians, Europeans).

Although sample sizes in MR studies (and genetic studies in general) keep increasing, the use of multiple exposure-associated variants as IVs is attractive in scenarios where most of the single instruments are invalid (e.g. pleiotropic, weak). A variety of approaches to assess violation of the MR studies when using multiple IVs have been proposed. For example, a study investigating the causal relationship between triglycerides and risk for coronary artery disease (CAD) needed to rule out that the association seen was not confounded by LDL or HDL levels [98]. Doing this in an MR setting is not trivial in light that these factors are correlated to each other and many of the triglycerides-associated variants are also associated to LDL and HDL. In order to disentangle which was the causal exposure, they develop an approach which consisted on regressing the SNP-CAD effect estimates with those of the SNP-triglycerides, adjusting by the effect estimates of SNP-LDL and SNP-HDL. Doing this, they found evidence that an increase in triglycerides and LDL increases the risk of CAD, but a decrease of HDL is not causally related to CAD, but confounded by the other two factors[98]. Another related approach to assess bias in MR when using multiple invalid (weak) IVs is the Egger test [99]. Traditionally, this test is used as a tool for detecting small-study bias in meta-analysis. Given that MR of a single study with multiple IVs can be seen as analogous to a meta-analysis, this test can be applied by replacing the precision of a single study's estimate with the strength of the instrument [99]. It is reported that under certain conditions (the variants are not correlated with each other, and the direct effect of the variants on the outcome is 0) this approach can give protection against bias even when all the genetic variants violate the standard MR assumptions.

1.6 Aims and case studies

The work presented in this thesis is a medley of application of statistical genetics methods to diverse data sets available to answer questions about aetiology of some ocular traits and diseases including central corneal thickness (CCT), conjunctiva ultra violet autofluorescence (CUVAF), refractive error, keratoconus, glaucoma and age related macular degeneration. In addition to these, I also include work done investigating the genetic architecture of epithelial ovarian cancer (EOC) and its subtypes.

1.6.1 Central corneal thickness and keratoconus

In the following chapter I present a study involving an exome (Illumina Human Exome array) association study of CCT. As background, the cornea is the transparent dome-shaped surface of the eye that covers the iris, pupil and anterior chamber. One of its main functions is to allow the refraction of the light entering the eye as well as serving as a protective barrier. The cornea is comprised of many layers and has a thickness between 500 and 600 μm in its centre and 600-800μm in its periphery [100]. The three primary layers are the outer layer containing the epithelium, the stromal layer comprising 90% of the total corneal thickness built from an extracellular matrix rich in collagen fibrils and keratocytes and the inner layer containing endothelial cells [100]. A reduction in the thickness of the cornea can lead to the development of keratoconus. Keratoconus is a degenerative disorder of the eye where structural changes within the cornea along with the thinning cause the cornea to change into a conical shape, causing vision distortion. Different studies have shown that the breakage of the collagen cross-linkage in the stroma due to protease activity can lead the development of keratoconus by reducing corneal thickness [101]. Once Keratoconus initiates, it progressively dissolves the collagen fibrils in the Bowman's layer located between the stroma and epithelium layers[102].

As a complex disease, keratoconus risk is driven by both genetic and environmental factors [102, 103]. Historically, genetic studies of keratoconus (e.g. GWAS) have not been able to identify relevant disease loci as these have been hampered by the complex aetiology and low prevalence of the disease (1 in 2000 individuals) limiting the possibility to perform powered GWAS [104]. In order to dissect the genetics of keratoconus, researchers turned towards the endophenotype approach. Endophenotype is a term borrowed from psychiatric genetics to define the separation of psychiatric conditions into more stable phenotypes with clearer genetic connection. In the case of keratoconus, CCT acts as an endophenotype given that a reduction on CCT can lead to the disease. Genetic studies of CCT have been highly successful given that can be measure in anyone and is a trait in a continuous scale, thus providing increased power compared to a dichotomous disease status. Studies report that approximately 90% of the variance of CCT can be attributed to a genetic component. During the last years, GWAS of CCT successfully mapped 27 loci which together explain 8.3% of the additive variance [105]. Among these loci, several were found to be associated to keratoconus.

With the advent of the extension of paradigm in GWAS (i.e. from just investigating common variation to investigate also rare variation), we carried out an exome association analysis of CCT, followed by investigating the associated variants in a case-control sample of keratoconus. Details and results of this project form the chapter 2 of this thesis.

1.6.2 Conjunctival ultra violet autofluorescence

The conjunctiva is a thin translucent mucus membrane that covers the eye ball. It starts at the edge of the cornea and extends behind the eye where it folds and forms the inside surface of the eyelids. Its main function is to protect the eye from foreign particles and to keep it lubricated by producing mucus and tears[106]. There is evidence for an association between excess of ultraviolet radiation (UVR) and a number of diseases of the conjunctiva. For example, pterygia which refers to an abnormal growth of the conjunctiva is thought to arise as result of the sun's rays passing unobstructed through the lateral side of the eye causing degradation of the collagen fibres [107]. Pinguecula is a similar condition to pterygia also product of degeneration of the collagen fibres due to sun exposure and appears as a yellow-white deposit in the conjunctiva [108]. Photokeratitis is a painful eye condition also arising from UVR and is characterized by sunburn of the cornea and conjunctiva [108]. UVR is also associated with an increased risk of squamous cell carcinoma of the cornea and conjunctiva [109, 110], as well as other eye diseases outside the conjunctiva such as cortical cataract [111], iris melanoma [112] and macular degeneration [113].

Conjunctival ultra violet autofluorescence (CUVAF) has been developed as a way to measure the extent of UVR exposure of the eye [114]. This technique involves the use of black light emitted by a Wood lamp to examine the extent of actinic damage to the conjunctiva. Previous studies show that CUVAF can be an effective biomarker of the first stages of pinguecula and pterygium [115]. Also, this can be used as a marker of eye UV exposure. The latter is of great interest as measuring sun exposure accurately is challenging; this is usually carried out by questionnaires but these are subject to many biases and CUVAF may enable more precise estimates. CUVAF measures can aid the investigation of other ocular traits such as myopia. A considerable number of studies have shown that myopia is inversely associated to time spent outdoors, thus having an effective biomarker of sun exposure such as

CUVAF may help providing a frame to investigate what aspect of time spent outdoors can explain this apparent protective effect.

The utility of CUVAF to study ophthalmohelioses (sun-related eye diseases) and its potential use as sun biomarker encouraged the project depicted in chapter 3 of this thesis. In this, I show the results of genetic analyses carried out in three independent Australian cohorts. We report for the first time the heritability of CUVAF and perform a GWAS. Through the endophenotype approach GWAS of CUVAF can be potentially used as way to find risk genetic variants of pterygium and pinguecula. Further, we show the impact of geographic latitude and longitude on UVAF.

1.6.3 Refractive error and myopia

Refractive error arises when the length and/or curvature of the eye does not allow the direct focus of light on the retina. Myopia or short-sightedness is one of the most common refractive errors, where a more negative refractive error indicates a higher degree of myopia. Myopia is a condition where light does not focus on the retina but instead in front causing distant objects to appear blurry. Although myopia is in most cases a benign condition that can be corrected through the use of lenses or refractive surgery, having a high degree of myopia has been associated to cataract, glaucoma and retinal degeneration [116]. The importance of research of this condition has increased in the last decades due to dramatic increases in prevalence around the globe. In China, 90% of teenagers are short-sighted compared to 10-20% 60 years ago. The statistics are similar for South Korea and Singapore, and in the western world it is estimated that the prevalence has doubled over the same period.

There is compelling evidence that genetic factors and environmental inputs contribute to the development of myopia [117, 118]. Heritability estimates of myopia widely differ between studies; as reviewed in [119], heritability estimates have been reported to be between 11% and 87%. Although previous linkage studies had limited success on identifying genes involved in myopia, GWAS of myopia and refractive error from two large studies, one from the Consortium for Refractive Error and Myopia (CREAM) [120] and the other from 23andMe [121], successfully identified 22 loci associated to the trait. These genetic studies gave great insight into the biology of myopia by also identifying relevant pathways such as neurotransmission, retinoic acid metabolism and ion transport. However, even though genetic differences explains some proportion of myopia cases, genetic changes do not happen as fast to explain the soaring rates in myopia of the last decades. This dramatic increase of myopia points to an environmental effect. Educational attainment and time spent outdoors are the two factors more consistently associated to myopia. An emblematic study assessing the prevalence of myopia between generations of Alaskan Eskimos found a dramatic increase in the younger generation where education became compulsory [122]. After many independent population studies corroborating this finding, to date, educational attainment is considered the main risk factor for myopia. Multiple studies [123-128], including a very recent randomized trial [129] have found that time spent outdoors is inversely associated with myopia development. A protective mechanisms that may underlie this association is that time spent outdoors is accompanied by less time performing near work activities such as reading books or staring at a screen which promote eye elongation as a compensation mechanism to defocus [130-132]. Another hypothesis is that exposure to bright light enhances dopamine release in the retina suppressing axial elongation [133, 134]. Finally, recently few studies proposed that increased vitamin D concentrations may be behind the protective effect of time spent outdoors [135-137].

In chapters 4 and 5, I report the work where I assess the causal relationship of education and vitamin D on myopic refractive error. Specifically, chapter 4 displays the results of the application of an MR approach to assess the causal relationship of education on refractive error using individual level data in 3 independent cohorts. Chapter 5 illustrates an MR of vitamin D or refractive error using summary results of the large CREAM GWAS of refractive error.

1.6.4 Primary open angle glaucoma and age-related macular degeneration

Glaucoma comprises a group of age-related eye diseases characterized by an irreversible deterioration of the optic nerve resulting in visual field loss [138]. It is a progressive condition and one of the leading causes of blindness around the world [139, 140]. Primary open angle glaucoma (POAG) is the most common type of glaucoma in western countries [141] and can develop as a result of clogging of the drainage canals of the eye, resulting on an increase of intraocular pressure (IOP) which causes the progressive optic nerve damage [142]. POAG is relatively a rare disease and happens predominantly in older individuals; the lifetime risk at age 75% is approximately 2% [143]. Studies have identified many risk factors, including genetic variants that increase the risk of developing POAG. These factors include estrogen deficiency, taking corticosteroids, having certain eye conditions such as myopia, high blood pressure, diabetes, etc. [144-148]. Genetic studies of POAG have also identified multiple variants such as Gln368Ter in the *MYOC* gene which accounts for ~4% of all the POAG cases [149]. In the last years, SNPs in seven other loci (*CAV1* [150], *CDKN2BAS, TMCO1*[151], *SIX1*[152], *ABCA1, GMDS and AFAP1*[153, 154]) have been identified through GWAS of POAG and its endophenotypes IOP and optic disc parameters. However, the variance explained by these seven SNPs is relatively small and it is unclear how important other common genetic variants are in explaining trait variation.

Age-related macular degeneration (AMD) is an eye disease characterized by damage to the macula of the retina resulting in blurred or no vision in the center of the visual field [155]. This disease typically occurs in older people and the lifetime risk at age 75 is around 2.8% [143]. The pathogenesis of this disease is not completely understood; however, oxidative stress, mitochondrial dysfunction and inflammatory mechanisms are believed to lead to the accumulation of cellular damage resulting in the death of photoreceptors in the central visual field [155-157]. Environmental and life style factors that contribute to risk of this disease include smoking [158], high blood pressure [159] and obesity [160]. Identification of variants in the CFH and ARMS2-HTRA1 which may be responsible of as much as 50% of the risk of AMD was the first major success of GWAS [161]. To date, subsequent GWASs of AMD have also been very successful and a further 33 genetic loci have been identified [162, 163].

There were multiple aims for the work presented in chapter 6. First, there were many unknowns about the genetic architecture of POAG. The contribution of common (genome-wide) and rare variants (exome) to risk of POAG (i.e. POAG heritability) was still unknown. Further, epidemiological studies indicate that higher estrogen levels may help prevent POAG [164-166] and that the prevalence of POAG differs between sexes (the prevalence in men is higher than in women). Therefore, using genotype data of Australian cases and controls, we investigate the contribution of common and rare variants to POAG and we assess whether there are differences between POAG sexes (generated by genetic variants acting in men but not women, or vice versa). To date, at the genome-wide significance level, only the ABCA1 locus is associated to both AMD and POAG. In the second part of chapter 6 I report the genetic correlation between these age-related eye diseases.

1.6.5 Epithelial ovarian cancer

Ovarian cancer is the leading gynaecological malignancy in developed countries. Approximately 90% of ovarian cancer tumours are of epithelial origin. Epithelial ovarian cancer (EOC) is a heterogeneous disease and can be divided into various histological subtypes based on different morphological, molecular and genetic features. High-grade serous carcinomas are the most common subtype of EOC followed by endometrioid, mucinous, clear cell, Brenner and other minor types [167, 168]. So far epidemiological and genetic studies have identified many lifestyle, environmental and genetic factors associated to an increase in risk of EOC. Among these, smoking [169, 170], obesity [171-173], and type 2 diabetes [174, 175] are associated with an increased risk, while [later] age at menarche appears to confer some protection [176]. Studies have identified multiple subtype specific and non-specific genetic variants (e.g. mutations in *KRAS* increase risk of mucinous EOC [177] while somatic mutations in TP53 are present in most tumours [178]) that increase the risk of EOC.

Although genetic studies have shown evidence of overlap between specific genetic variants underlying risk to the different EOC subtypes, the extent of genetic correlation beyond the known risk markers has not been examined. Also, it is not known the proportion of heritability these known loci explain. Therefore, in chapter 7, using genotype data from the Ovarian Cancer Association Consortium, I examine the genetic architecture of EOC and its subtypes. I report the array heritability, contribution of the known loci to the heritability and the genetic correlation between the different EOC subtypes. I continue the chapter investigating the genetic overlap between EOC (and subtypes) and risk factors where summary GWAS statistics are available, namely BMI, height, obesity, diabetes type 2, age at Menarche and smoking.

Chapter 2

WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness.

This chapter is published as:

Cuellar-Partida, G., Springelkamp, H., Lucas, S.E., Yazar, S., Hewitt, A.W., Iglesias, A.I., Montgomery, G.W., Martin, N.G., Pennell, C.E., van Leeuwen, E.M., Verhoeven V.J., Hofman A., Uitterlinden A.G., Ramdas W.D., Wolfs R.C., Vingerling J.R., Brown M.A., Mills R.A., Craig J.E., Klaver C.C., van Duijn C.M., Burdon K.P., MacGregor S., Mackey D.A. (2015) WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Human molecular genetics*, 1;24(17), 5060-8.

Abstract

Keratoconus is a degenerative eye condition which results from thinning of the cornea and causes vision distortion. Treatments such as ultraviolet (UV) cross-linking have proved effective for management of keratoconus when performed in early stages of the disease. The central corneal thickness (CCT) is a highly heritable endophenotype of keratoconus, and it is estimated that up to 95% of its phenotypic variance is due to genetics. Genome-wide association efforts of CCT have identified common variants (i.e. minor allele frequency (MAF) >5%). However, these studies typically ignore the large set of exonic variants whose MAF is usually low. In this study, we performed a CCT exome-wide association analysis in a sample of 1029 individuals from a population-based study in Western Australia. We identified a genome-wide significant exonic variant rs121908120 (P = 6.63 × 10(-10)) in WNT10A. This gene is 437 kb from a gene previously associated with CCT (USP37). We showed in a conditional analysis that theWNT10A variant completely accounts for the signal previously seen at USP37. We replicated our finding in independent samples from the Brisbane Adolescent Twin Study, Twin Eye Study in Tasmania and the Rotterdam Study. Further, we genotyped rs121908120 in 621

keratoconus cases and compared the frequency to a sample of 1680 unscreened controls from the Queensland Twin Registry. We found that rs121908120 increases the risk of keratoconus two times (odds ratio 2.03, P = 5.41 × 10(-5)).

Introduction

Keratoconus is a degenerative eye disease with an incidence of around 1 in 2000 in the general population[179]. It is characterized by thinning and weakening of the cornea, and its symptoms range from mild astigmatism and myopia to severe vision distortion. Corneal collagen ultraviolet (UV) cross-linking is a minimally invasive and effective option for management of keratoconus at early stages[180] achieving biomechanical stabilization of the cornea and reducing (or in some cases halting) the disease progression rate. However, it is not uncommon for patients with mild or early stages of keratoconus to be misdiagnosed as cases of astigmatism or myopia and undiagnosed keratoconus can lead to corneal ectasia following laser refractive surgery (LASIK) [181].This makes it particularly important to find biomarkers that can point to keratoconus in its earliest stage.

Previous work has shown that keratoconus risk is affected by both genetic and environmental factors[102, 103]. Several strategies have been pursued to identify the genetic risk factors of keratoconus; however, given the low prevalence of the disease, it has been difficult to perform well powered genomic studies[104]. In contrast, genome-wide association studies (GWAS) of central corneal thickness (CCT), a highly heritable biometric trait which functions as endophenotype of keratoconus, have successfully identified 27 associated loci[105]. Lu et al[105] found that several of these CCT loci were also associated with keratoconus in a case-control analysis.

The identified CCT variants only explain around 8% of the variability of the trait[105]. CCT is highly heritable (~90%)[182] and hence there is substantial missing heritability. One possible component of the missing heritability is low frequency variants. The published CCT GWASs to date focused primarily on common variants (i.e. minor allele frequency (MAF) > 5%). This approach ignores a large number of coding exome variants, where the MAF is usually lower. Therefore, to determine the role of low-frequency coding variants in CCT, we evaluated putative functional coding variants from the Illumina Human Exome array. We performed the association using genotype data from 1029 individuals from the Raine cohort[183]. We replicated our results in two independent samples from i) the Rotterdam Study[184] and ii) Brisbane Adolescent Twin Study[185, 186]. Further, we investigated the significant associations with CCT in a sample of 621 Australian keratoconus cases and 1680 unscreened controls.

Results

We performed an exome-wide association analysis of CCT using data from the Western Australian Pregnancy (Raine) Cohort[183]. A sample of 1029 unrelated individuals of European descent and with CCT measures were used to test the association of the 43,435 exonic variants with a MAF>0.25% passing quality control. Sample characteristics are summarized in Table 1. We performed the association analysis of each variant through linear regression analysis adjusting for sex, age and the first 3 genetic principal components (PCs).The *genomic inflation* factor (λ) with respect to the median of χ2-statistics was 1.006 suggesting no inflation in the test statistics due to population structure[187] (Figure 1a.).

Figure 2 shows the results of the analysis. One single nucleotide polymorphism (SNP) reached the threshold of genome wide significance (P = $5.0\text{x}10^{-8}$): rs121908120 in *WNT10* on chromosome 2 (β=-23.84 ± 3.92,*P* = $6.63\text{x}10^{-10}$). *WNT10A* is expressed in all the ocular tissues reported in the ocular tissue database[188]. The SNP rs121908120 causes a missense mutation in *WNT10A* which results in a change in the amino acid 228 from phenylalanine to Isoleucine. According to SIFT[189] and PolyPhen[190], this missense mutation is deleterious (score 0) and probably damaging (score 0.994), respectively. This variant is 437kb upstream of rs10189064 (*P* = $3.11\text{x}10^{-4}$) in the *USP37* gene, which was previously associated with CCT[105]. These two variants are in moderate linkage disequilibrium (R^2 = 0.369). In order to assess the extent of independent effect of these two SNPs, we performed conditional analysis (i.e. using one as covariate and testing the other and vice versa). The results are summarized in Table 2. Our results show that conditioning rs121908120 by rs10189064 does not reduce the effect (β=-23.75 ± 5.33) of the SNP; however the *p-value* goes down to $9.28\text{x}10^{-6}$ probably due to a reduction in sample size, as just 938 individuals had information on the SNP rs10189064. On the other hand, conditioning rs10189064 on the variant in *WNT10A removes the effect completely - the effect changes from*β=-14.57 ± 4.02 (*P* = 3.11x10^{-4}) to β=0.44 ± 5.21 (*P* = 0.93). This suggests that the previously identified associated SNP in *USP37* is likely due to linkage disequilibrium (LD) confounding with the variant in *WNT10A*.

We replicated these results using data from the Rotterdam Study[184] (Table 2). The total sample size of this replication cohort was n=4,479. Although the effects were moderately smaller in these samples, the association signal for rs121908120 was clearly replicated (β = -12.68 ± 2.75, *P* = 3.87x10^{-6}). The results remained similar after conditioning on rs10189064 (β = -10.92 ± 3.7, *P* = 3.21x10^{-3}). In addition, we used available exome data from 147 participants from the Brisbane adolescent twin study (BATS) and Twin Eye Study from Tasmania (TEST)[185, 186]. However, this sample only had rs121908120 genotyped. We found that the effect in this sample replicates our finding (β = -28.73 ± 14.05, *P* = 0.04).

Although rs121908120 is the strongest candidate SNP in the region, we used the online tool LocusTrack [191] to look for additional SNPs in high LD with rs121908120, based on the1000 Genomes phase 3 European ancestry reference set. The only variant (rs146199923) with r^2=1 was 100kb downstream of rs121908120,close to the *FEV* gene (Supplementary Figure 3). Examining, GeneCards[192] and dbSNP, rs146199923 is not a strong candidate SNP as lies in an intergenic region outside conserved transcription factor binding sites and DNaseI hotspots.

We also investigated the effects of exome variants in previous associated loci[105]. Figure 1a displays the *p-value* distribution of SNPs within CCT known genes along with the distribution of all SNPs assessed in this study. We did not find any evidence of associated exome variants in these genes.

In addition to per SNP testing, gene-based analysis was done using the optimal unified approach SKAT-O[193]. However, the approach did not alter our conclusions, as *WNT10A* (*P* = 1.65x10^{-10}) was the only significant association after correction for multiple testing [Figure 1b]. The top 10 results for the gene-based test are summarized in Table 3. *WNT10A* was not associated in the gene-based analysis if rs121908120 was omitted (*P*=0.29). Following a similar approach, we performed pathway analysis. However, in order to avoid capturing the same signal as previous experiments we removed all SNPs within the *WNT10A* gene before the analysis. Although no pathways passed the significance threshold, interestingly, the top pathway (GO:2000096, P = $2.57\text{x}10^{-4}$) was the one described as the positive regulation of the *Wnt* receptor signaling pathway [Table 4]. Analyses from Lu et al indicated that extracellular matrix and collagen pathways are associated with CCT[105]. Analogous to inspecting variants within known associated genes, we examined the distribution of p-values in the collagen and extracellular matrix pathways and found an enrichment of small p-values for the extracellular matrix (λ=2.14) and collagen pathways (λ=2.32) [Figure 1c].

Further, we genotyped rs121908120 in 621 keratoconus cases and used data from 1680 individuals from the Queensland twin registry, genotyped on the Illumina HumanCoreExome array, as unscreened controls. The rs121908120 MAF in keratoconus cases was 0.05 while in controls it was 0.024 translating to a 2.03 fold increase in risk (Fisher exact test p-value = $5.41\text{x}10^{-5}$).

Discussion

Our study identified a missense mutation (rs121908120) in *WNT10A* associated with CCT and keratoconus. Previous GWAS of CCT identified rs10189064 in the *USP37* gene[105], which is in moderate LD with this newly found variant rs121908120. However, the SNP in *WNT10A* has a bigger effect on CCT, and completely accounts for the signal previously seen at *USP37* in the analyzed samples.

Aside from the *USP37*/*WNT10A* region, we did not detect association of exonic variants in or near CCT associated loci from a previous study focusing on common variation[105]. This indicates that the tagged SNPs are not in LD with the previously identified common variants in those genes, or the sample size is too small to detect any association.

WNT10A belongs to the WNT gene family. This family consists of structurally related genes encoding secreted signaling molecules that have been implicated in oncogenesis and in several developmental processes, including regulation of cell fate and patterning during embryogenesis[194, 195].Studies have shown that corneal endothelial cell fate is maintained by the hedgehog and WNT pathways[196, 197]. The corneal endothelium is responsible for maintaining the transport of fluids and solutes to the corneal stroma (which accounts for up to 90% of the total corneal thickness). A reduced endothelial cell density can have an impact on this fluid regulation leading to stromal swelling and scarring due to excess fluid[198, 199], which has also been described as a complication of keratoconus[200].

A handful of studies have described structural changes in the corneal epithelium in keratoconic eyes [201-203]. The corneal epithelium is an extremely thin layer composed of epithelial tissue covering the front of the cornea. Cornea epithelial cells renew continuously from limbal stem cells (LSCs) in order to maintain transparency for light transmission. A deficiency in LSCs can lead the cornea into a non-transparent or keratinized skin like epithelium[204]. Molecular analysis of the Wnt signaling pathway in limbal stem cells have shown that *WNT2*, *WNT6, WNT11, WNT16B* are over-expressed in the limbal region, while the expression of *WNT3, WNT7A, WNT7B* and *WNT10A* is upregulated in the central cornea (mature corneal epithelium)[194]. Based on this, we examined the p-value distribution of *WNT* genes and SNPs within them. We observed that in aggregate these genes tend to have small p-values, although non-significant (Figure 1a, Figure 1b and Table 5). The latter suggests that might be a matter of power or fine mapping of causal variants to see significant associations within these genes.

The strong association of *WNT10A* found in keratoconus adds evidence to its possible role in cornea stability. In addition, studies indicate that mutations in *WNT10A* are also a risk factor for ectodermal dysplasia including odonto-onycho-dermal dysplasia[205, 206]. This syndrome is associated with abnormalities of skin as well as epidermally derived structures including, hair, teeth, nails, tongue, and sweat glands. Hair including eyelashes is typically thin and sparse. Ocular features include chronic tearing, photophobia, and keratitis[207].

Other diseases that show corneal thinning as clinical feature include connective tissue disorders and osteogenesis imperfecta[208, 209]. Wnt signalling is essential for maintaining bone density and the homeostasis in connective tissue[210-212]. Our finding adds evidence on the link of these disorders to corneal thinning.
Pathway analyses performed by Lu et al[105] associated collagen and extracellular matrix pathways to CCT. Collagen fibrils are a major component of the cornea's extracellular matrix [213] and are the building blocks of the corneal stroma and Bowman's layer[214, 215].Our study was underpowered to detect significant pathway associations. However, we observed smaller p-value in these pathways than the expected from a uniform distribution.

In conclusion, our findings indicate that *WNT10A* plays a role in the corneal thickness homeostasis and that the mutation rs121908120 is a risk factor for keratoconus. Also, this finding adds evidence to the association of *WNT10A* to odonto-onycho-dermal dysplasia and the link of connective tissue disorders with corneal thinning. Furthermore, suggestive results on the association of WNT genes, and the fact that they are expressed in the different ocular tissues, indicate that may be a matter of extending the sample size or a finer mapping of variants to detect their association.

Methods

Our study consisted of two phases: in the first phase, we performed exome-wide association with CCT using the Raine study sample as discovery and the Rotterdam Study, the Brisbane adolescent twin study (BATS) and Twin Eye Study of Tasmania (TEST) for replication. In the second phase, we investigated the associated variant in Australian keratoconus patients and unscreened controls from the Queensland Twin Registry (QTwin).

Raine

Sample

Recruitment of the Western Australian Pregnancy (Raine) cohort has previously been described elsewhere in detail[216]. In brief, between 1989 and 1991 2,900 pregnant women were recruited prior to 18-weeks gestation into a randomized controlled trial to evaluate the effects of repeated ultrasound in pregnancy. Children have been comprehensively phenotyped from birth to 21 years of age (average ages of one, two, three, six, eight, ten, fourteen, seventeen and twenty-one) by trained members in the Raine research team. Most of the children are of Caucasian ethnicity. Data collection included questionnaires completed by the child's primary carer and by the adolescent from age 14, physical assessments by trained assessors at all follow up years, DNA collection from year 14 follow-up. The study was conducted with appropriate institutional ethics approval, and written informed consent was obtained from all mothers.

Phenotypes

At age 21, participants were invited for an eye study. CCT was obtained from the Pupil Center Pachymetry readout obtained by anterior segment tomography of each dilated eye taken with an Oculus Pentacam (Optikgerate GmbH, Wetzlar, Germany)[183].

Exome array

A total of 1825 participants were genotyped using the Illumina HumanExome-12v1_A array includes 247,870 markers. Approximately 90% of the markers are coding variants selected from >12,000 exome and genome sequences representing multiple ethnicities and complex traits. The remaining 10% comprises variants that have been associated with complex traits in previous Genome Wide Association Studies (GWAS), ancestry-informative markers, markers for identity-by-descent estimation, random synonymous SNPs and HLA tags[47]. Genotype calling was carried out in two steps. First, we called genotypes using Illumina GenomeStudio GenTrain clustering algorithm, together with the Illumina HumanExome-12v1_A product files. Quality control in the initial genotypes was done by excluding samples with a calling rate below 95%, and variants with a GeneTrain score <0.15, calling rate <0.95 or heterogeneity excess <-0.3 or >0.2.We performed principal component (PC) analysis, and excluded samples that were above 6s.d from the centroid of the 1000 Genomes[44] European population (GBR+CEU+FIN) PC1 and PC2. In the second step we used zCall[53] with the default parameters to improve calling of rare variants on the remaining samples. We excluded variants with calling rate < 99%, or which deviate from Hardy-Weinberg equilibrium (HWE) P < 10−6, resulting in 235,619 variants and 1563 individuals passing quality control (QC).

Statistical analysis:

Among the genotyped individuals passing QC, 1029 unrelated individuals (proportion of identity by descent <0.2) counted with CCT phenotype data. To ensure that the variants tested had at least 5 copies of the minor allele, we restricted the analyses to just those variants with a MAF > 0.25% (i.e. 43,435 SNPs). Single SNP based analysis was carried out using linear regression in plink[217, 218]and adjusting by sex, age and the first 3 PCs. The genotype cluster plot for the top associated variant rs121908120 is displayed in Supplementary Figure 1.

Gene and pathway based association analyses were performed using SKAT-O[193], which performs association test of SNP sets and optimally combines the burden test and the nonburden sequence kernel association test (SKAT).Gene-based SNP sets were created using the SNP-gene annotation file from the Illumina Human-Exome bead-chip. Pathways were based on the Gene Ontology database[219]. Pathway-based SNP sets were composed by the SNPs within the genes involved in each particular pathway.

We performed conditional analysis of rs10189064 and rs121908120 using Plink[217]. Given that the Illumina HumanExome-12v1_A does not contain rs10189064 among the tagged SNPs, we extracted the rs10189064 genotype from 938 individuals that were also genotyped in the Human660W-Quad bead chip for previous experiments. Genotyping and quality control details for the Human660W-Quad bead chip in the Raine sample are described elsewhere[105]. In brief 1593 individuals were genotyped in 2009 using the Human660W-Quad bead chip, as part of quality control (QC), the data were filtered by single nucleotide polymorphism (SNP) call rate <0.95, a Hardy-Weinberg equilibrium (HWE) p-value< 10^{-6} and a minor allele frequency (MAF) > 1%.To exclude population outliers, a principal component analysis (PCA) was carried out using SNPs with genotyping rate >0.98.

The Rotterdam Study

Samples

The Rotterdam Study is a population-based study held in Rotterdam, the Netherlands[184]. It consists of three cohorts. The original cohort, RS-I, started in 1990 and includes 7,983 subjects aged 55 years and older. The second cohort, RS-II, was added in 2000 and includes 3,011 subjects aged 55 years and older. The last cohort, RS-III, includes 3,932 subjects of 45 years of age and older and started in 2006. The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the "Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study)". All participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

Phenotype

CCT was measured using ultrasound pachymetry (Allergan Humphrey 850, Carl Zeiss Meditec, Dublin, CA, USA; subset of RS-I), and using a non-contact biometer (Lenstar LS900, Haag-Streit, Köniz, Switzerland; subset of RS-I, RS-II, and RS-III).

Genotyping and association

Genotyping of SNPs was performed using the Illumina Infinium II HumanHap550 array (RS-I), the Illumina Infinium HumanHap 550-Duo array (RS-I, RS-II), and the Illumina Infinium Human 610-Quad array (RS-I, RS-III). Samples with low call rate (<97.5%), with excess autosomal heterozygosity (>0.336), or with sex-mismatch were excluded, as were outliers identified by the identity-by-state clustering analysis (outliers were defined as being >3 s.d. from population mean or having identity-by-state probabilities >97%). A set of genotyped input SNPs with call rate >98%, MAF >0. 1% and Hardy-Weinberg P-value >10^{-6} was used for imputation. The Markov Chain Haplotyping (MACH) package version 1.0 software [220](Rotterdam, The Netherlands; imputed to plus strand of NCBI build 37, 1000 Genomes phase I version 3) and minimac version 2012.8.6 was used for the imputation. Number of samples remained for RS are summarized in Table 1. Imputation quality (r^2) for rs121908120 was RSI=0.61, RSII=0.57 and RSIII=0.63 and for rs10189064 r^2 was above 0.99 in the three studies. Association analyses of rs121908120 and rs10189064 variants with CCT were performed using the ProbABEL package[221] using age, sex, the first 5 PCs and the technique of measurement (the latter only for RS-I) as covariates.

Brisbane adolescent twin study and twin eye study in Tasmania

Samples

Methodologies and recruitment of participants from the Brisbane adolescent twin study (BATS) and twin eye study in Tasmania (TEST) are described elsewhere[185, 186].

Phenotype

CCT was measured in this cohort using ultrasound pachymetry and recorded for both eyes. Measurements were performed using a Tomey SP 2000 (Tomey Corp., Nagoya, Japan).

Genotyping and association

Genotyping of 147 individuals from BATS and TEST with eye phenotype was performed using the Illumina HumanCoreExome array. Samples and SNPs with low call rate (<98%) were excluded, as well as variants with MAF <0.1% and Hardy-Weinberg P-value >10^{-6}. Association was performed using Merlin which effectively accounts for family structure[222]. Age, sex and the first 3 principal components were used as covariates.

Queensland twin registry

Samples

Methodologies and recruitment of the Queensland Twin registry are described elsewhere (REF above). The unscreened controls for the keratoconus samples were a subset of the BATS and TEST projects. These controls were family members from BATS and TEST projects, selected to be unrelated to the 147 BATS and TEST individuals included in the CCT scan. The unscreened controls from were pruned to remove related individuals (typically both parents of a twin pair were used as controls).

Genotyping

Genotyping was carried out as described for the BATS and TEST CCT samples.

Keratoconus cases

Sample

Australian participants with keratoconus (n=621) were ascertained through the Department of Ophthalmology of Flinders Medical Centre, Adelaide, Australia; private optometry practices in Adelaide and Melbourne, Australia; the Royal Victorian Eye and Ear Hospital, Melbourne, Australia; and by Australia-wide mail out to members of Keratoconus Australia, a community-based support group for patients.

Phenotypes

The diagnosis of keratoconus was based on clinical examination and videokeratography pattern analysis. Clinical examination included slit lamp biomicroscopy, cycloplegic retinoscopy, and fundus evaluations. Slit lamp biomicroscopy was used to identify stromal corneal thinning, Vogt’s striae, or a Fleischer ring. A retinoscopic examination was performed with a fully dilated pupil to determine the presence or absence of retroillumination signs of keratoconus, such as the oil droplet sign and scissoring of the red reflex. Videokeratography evaluation was performed on each eye by topographic modeling. Patients were considered as having keratoconus if they had at least one clinical sign of the disease and by confirmatory videokeratography map with an asymmetric bowtie pattern with skewed radial axis above and below the horizontal meridian (AB/SRAX). A history of penetrating keratoplasty performed because of keratoconus was also sufficient for inclusion.

Genotyping

Genotyping of rs121908120 in 621 individuals was completed using a pre-designed Taqman assay (Life Technologies), amplified in SensiFAST Probe No-ROX master mix (Bioline) on a LightCycler480 Real-time PCR Machine (Roche), according to manufacturer’s protocol. The genotyping cluster plot for rs121908120 genotyping is displayed in Supplementary Figure 2.

Figure 2.1. Quantile-Quantile plots of single SNP association (a), gene-based association (b) and pathway-based (c) results in the discovery cohort (Raine study). Each dot represents an observed statistic (-log10P) versus the corresponding expected statistic. The black line corresponds to the null distribution. Dotted lines show the significance threshold based on a Bonferroni correction for multiple testing.

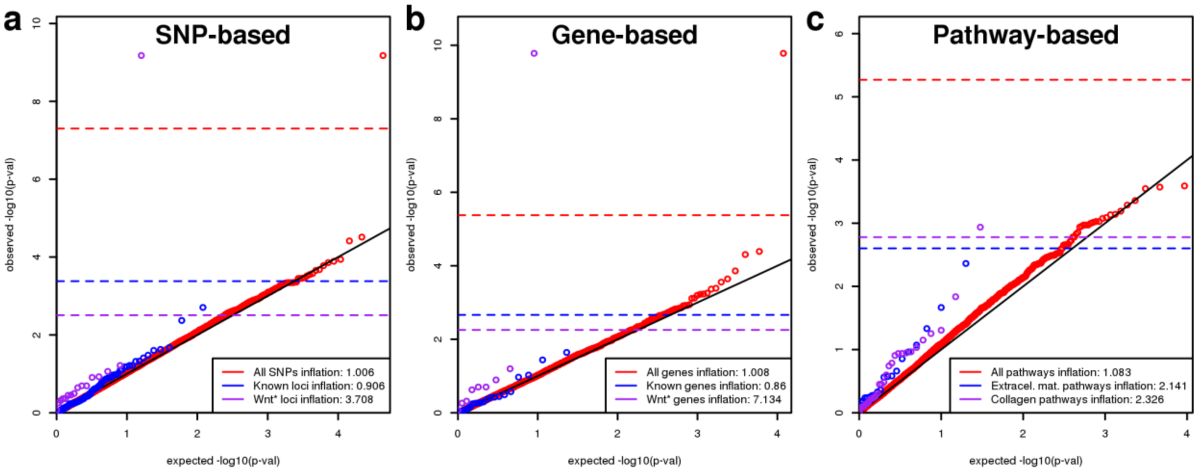


Figure 2.2. Manhattan plot of association results for central corneal thickness in the discovery cohort (Raine study). The plot shows -log10 –transformed p-values for all single nucleotide polymorphisms. The dotted horizontal line represents the threshold of genome-wide significance (p-value < 5.0 x 10^{-8}).

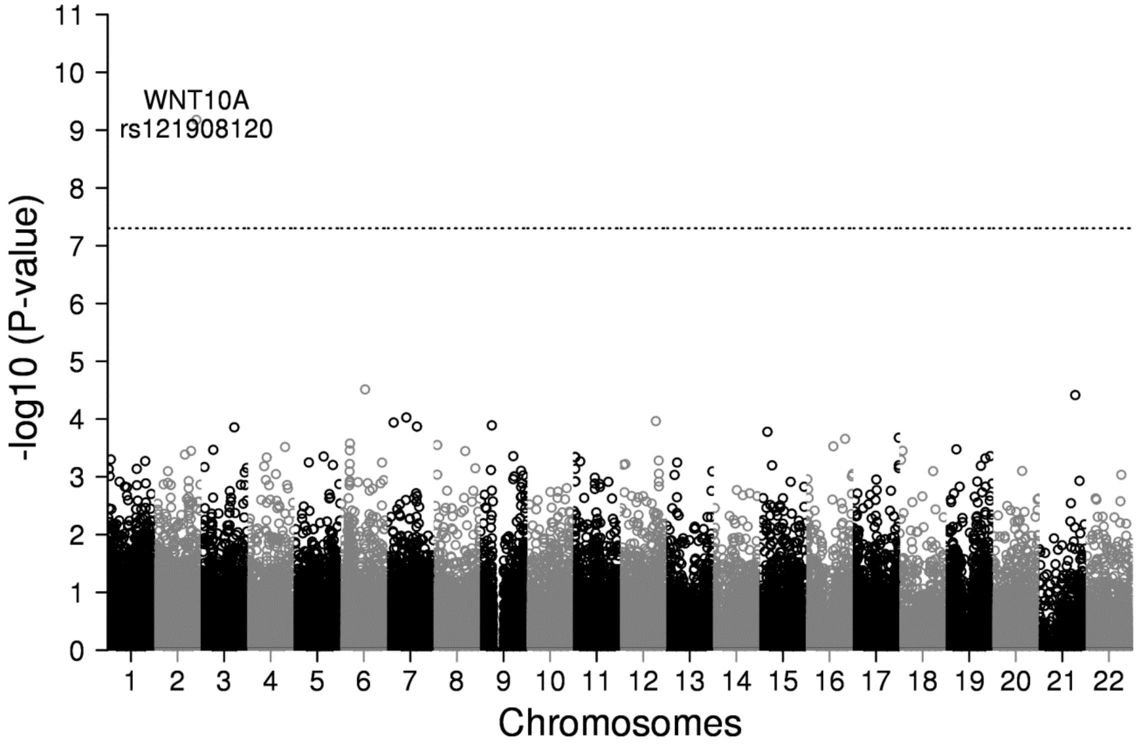


Table 2.1. Descriptive statistics of the samples.

	N	CCT (SD) (µm)	Age (SD) (years)	Males (%)
Raine	1,029	538.2 (32.3)	20.1 (0.4)	48%
BATS/TEST	147	554.5 (33.8)	22.6 (12.2)	54%
RS-I	873	544.4 (33.9)	76.3 (6.7)	48%
RS-II	1,215	547.7 (34.2)	72.6 (5.3)	47%
RS-III	2,391	550.3 (33.9)	62.3 (5.8)	43%

Abbreviations: CCT: Central corneal thickness; BATS/TEST: Brisbane Adolescent Twin Study / Twin Eye Study in Tasmania; RS-I, RS-II, RS-III: Rotterdam Study cohorts; SD = standard deviation.

Table 2.2. The results of the genome-wide association study with central corneal thickness (CCT) as outcome. Only the genome-wide significant single nucleotide polymorphism (SNP) is showed (rs121908120), together with a previously known associated SNP with CCT (rs10189064). Both SNPs were conditioned for the other SNP. Beta = effect size on central corneal thickness (µm) based on the minor allele.

SNP Minor / Major allele	Discovery			Replication										
	Raine			**BATS/TEST**			**RS-I**		**RS-II**		**RS-III**		**RS-Meta**	
	Beta ± s.e	**P-value**	**MAF**	**Beta ± s.e**	**P-value**	**MAF**	**Beta ± s.e**	**MAF**	**Beta ± s.e**	**MAF**	**Beta ± s.e**	**MAF**	**Beta ± s.e**	**P-value**
rs121908120 A/T	-23.8 ± 3.9	6.63E-10	0.030	-28.73 ± 14.1	4.10E-02	0.02	-8.76 ± 5.8	0.031	-18.46 ± 6.1	0.025	-12.18 ± 3.6	0.028	-12.68 ± 2.8	3.87E-06
rs10189064 A/G	-14.6 ± 4.0	3.11E-04	0.033	--	--	--	-4.04 ± 4.3	0.035	-12.8 ± 4.2	0.031	-6.71 ± 2.7	0.032	-7.52 ± 2.0	1.94E-04
rs121908120 a.f. rs10189064	-23.8 ± 5.3	9.28E-06	--	--	--	--	-9.27 ± 7.8	--	-11.27 ± 7.8	--	-11.46 ± 5.0	--	-10.92 ± 3.7	3.21E-03
rs10189064 a.f. rs121908120	0.44 ± 5.2	9.31E-02	--	--	--	--	0.56 ± 5.8	--	-7.99 ± 5.4	--	-0.78 ± 3.8	--	-2.34 ± 2.7	3.89E-01

Abreviations: a.f.. = adjusted for; BATS/TEST: Brisbane Adolescent Twin Study / Twin Eye Study in Tasmania; MAF = minor allele frequency; RS-I, RS-II, RS-III: Rotterdam Study cohorts; RS-Meta: Meta-analysed estimates from the 3 Rotterdam Study cohorts; s.e. = standard error of the beta.

Table 2.3. Top 10 results from the gene-based association with central corneal thickness (CCT) performed using the SKAT-O approach in the Raine cohort.

Gene	P-value	#SNP
WNT10A	1.65E-10	3
SH3BGR	4.07E-05	4
ANKRD6	4.94E-05	6
STEAP1B	1.38E-04	1
ATPBD4	2.28E-04	1
TAF11	2.78E-04	1
EFCAB7	4.10E-04	6
PRRG2	4.32E-04	3
CROCC	5.56E-04	1
C6orf1	5.88E-04	1

#SNP = number of single nucleotide polymorphisms used for the gene-based test.

Table 2.4. Top 10 results from the pathway-based association with central corneal thickness (CCT) performed using the SKAT-O approach in the Raine cohort.

Pathway	**P-value**	**#SNP**	**Definition**
GO:2000096	2.57E-04	11	Positive regulation of Wnt receptor signaling pathway
GO:0030145	2.68E-04	76	Manganese ion binding
GO:0038180	2.84E-04	19	Nerve growth factor signaling pathway
GO:1990090	4.39E-04	18	Cellular response to nerve growth factor stimulus
GO:0035249	5.18E-04	46	Synaptic transmission, glutamatergic
GO:0048406	6.45E-04	16	Nerve growth factor binding
GO:0007608	7.25E-04	175	Sensory perception of smell
GO:0003730	7.36E-04	20	mRNA 3'-UTR binding
GO:0007520	8.15E-04	53	Myoblast fusion
GO:0048172	8.41E-04	17	Regulation of short-term neuronal synaptic plasticity

#SNP = number of single nucleotide polymorphisms used for the pathway-based test.

Table 2.5. Results of WNT* genes available from the gene-based association with central corneal thickness (CCT) performed using the SKAT-O approach in the Raine cohort.

Chromosome	Position	Gene	P-value	#SNP
2	219745254	*WNT10A*	1.65E-10	3
10	102222811	*WNT8B*	0.06	1
11	75897369	*WNT11*	0.13	1
7	120969089	*WNT16*	0.2	2
7	116916685	*WNT2*	0.2	3
1	228109164	*WNT9A*	0.24	1
17	44928967	*WNT9B*	0.48	1
17	44839871	*WNT3*	0. 54	1
12	49359122	*WNT10B*	0.85	1

#SNP = number of single nucleotide polymorphisms used for the gene-based test.

Supplementary Material:

WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness

Figure 2.3 Supplementary Figure 1. Genotype cluster plot of rs121908120 (exome chip id: exm266718) in the discovery cohort (Raine). Black dots indicate not calling; blue and red indicate the homozygotes; purple indicates the heterozygotes.

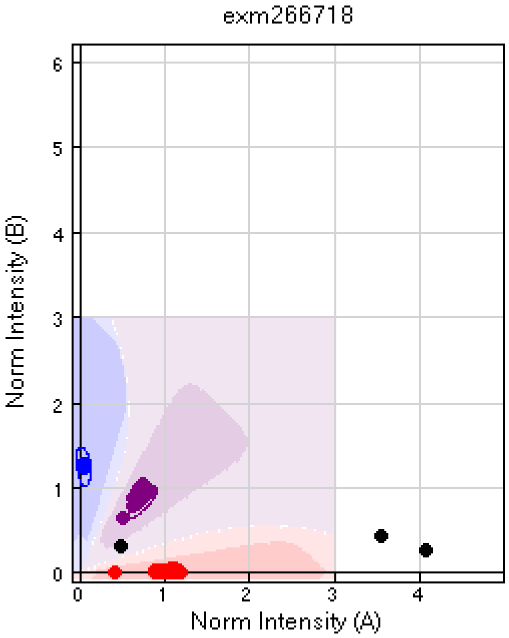


Figure 2.4 Supplementary Figure 2. Genotype cluster from a Taqman assay for rs121908120 (exome chip id: exm266718) in the Keratoconus cases. NTC refers to the no template control (i.e. no DNA added).

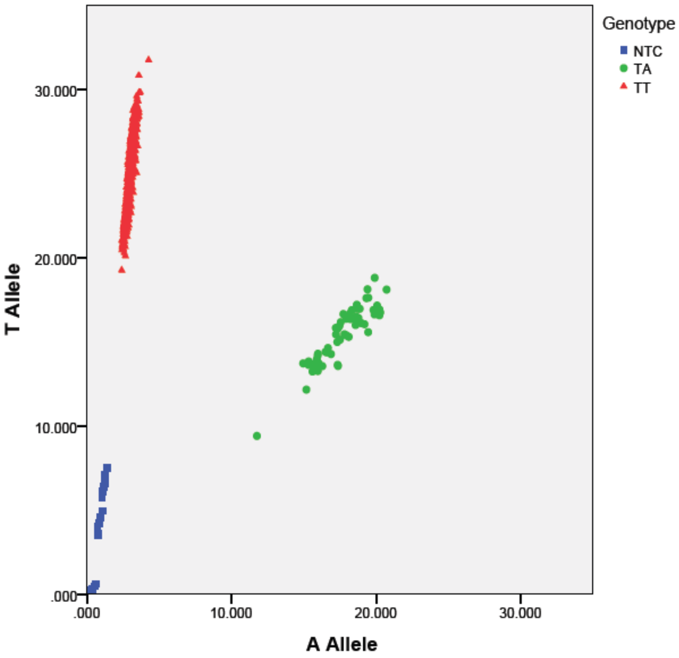
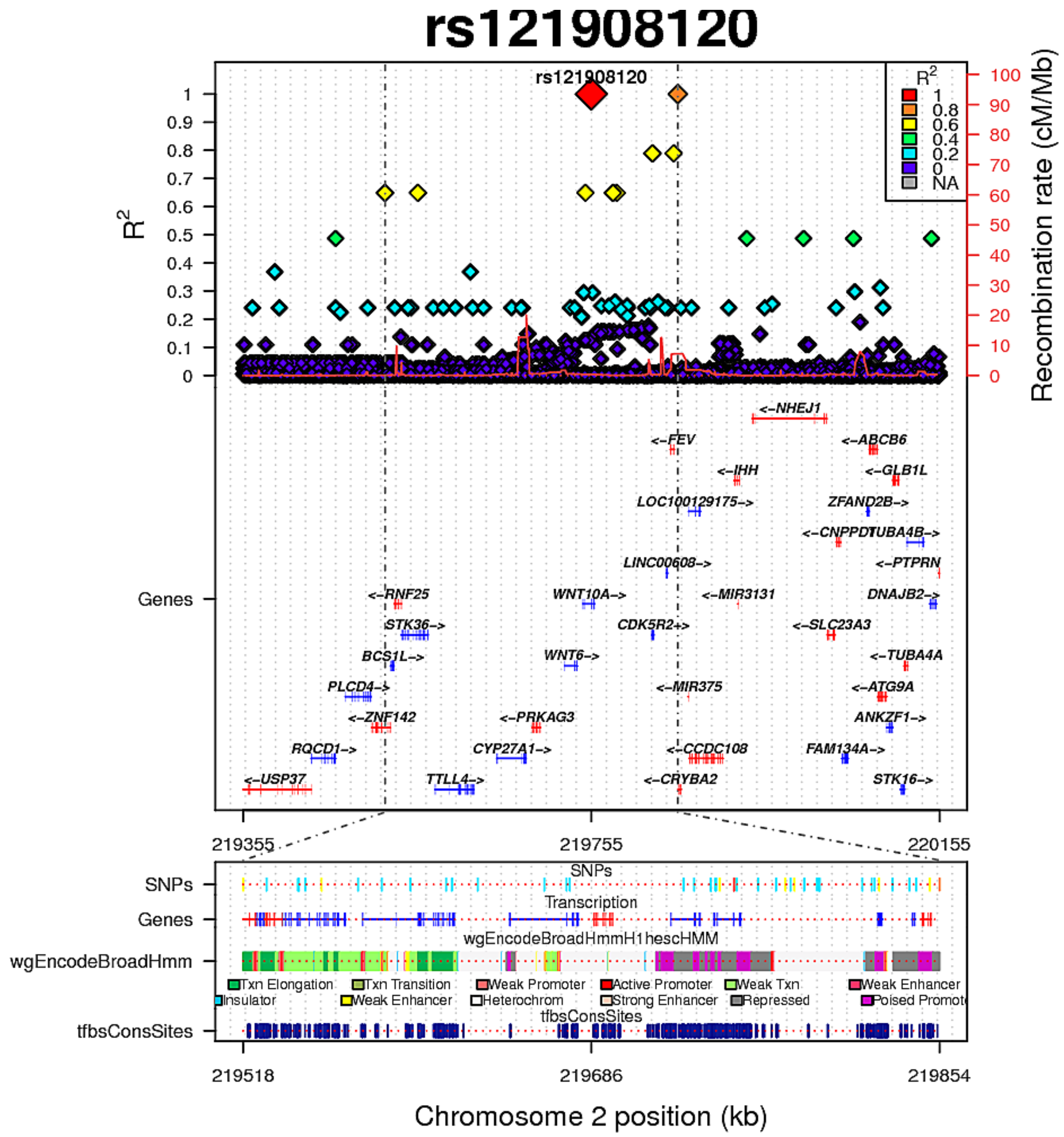


Figure 2.5 Supplementary Figure 3. LocusTrack plot displaying variants within the rs121908120 region (±400Kb). Left axis displays the extent of LD (r^2) of each variant with rs121908120. Dotted lines represent the region of the bottom panel. Bottom panel show SNPs in high LD with the same color code as the upper panel; Genes show the genes transcribed regions; wgEncodeBroadHMM track displays the chromatin state segmentation for Human Stem cells; tfbsConsSites show regions with a conserved transcription factor binding site.



Chapter 3

Genetic and environmental factors in conjunctival UV autofluorescence.

This chapter is published as:

Yazar, S.*, **Cuellar-Partida, G**.*, McKnight, C.M., Quach-Thanissorn, P., Mountain, J.A., Coroneo, M.T., Pennell, C.E., Hewitt, A.W., MacGregor, S. and Mackey, D.A. (2015) Genetic and environmental factors in conjunctival UV autofluorescence. *JAMA ophthalmology*, **133**, 406-412.

ABSTRACT

Importance: Conjunctival ultraviolet autofluorescence (CUVAF) has excellent potential as an objective biomarker of sun exposure. However, much variation in CUVAF is observed and the relative contribution of genes and environment to this variation has not yet been identified.

Objective: CUVAF photography was developed to detect and characterise pre-clinical sunlight-induced ocular damage. Ocular sun exposure has been related to cases of pterygia and also recently negatively correlated with myopia. We investigated sources of variation in CUVAF in relation to its potential clinical relevance.

Design: Cross-sectional analysis of three population-based cohort studies: Twins Eye Study in Tasmania, Brisbane Adolescent Twin Study and Western Australian Pregnancy Cohort (Raine) Study.

Setting: General community.

Participants: 295 Australian families from the Tasmanian and Brisbane twin studies and 661 participants from the 20-year follow-up of the Raine Study. Only individuals with available genotype data were included.

Methods: We compared the CUVAF levels in three cohorts and performed a classical twin study to partition variation in CUVAF. We also conducted a genome-wide association analysis to identify specific genetic variants associated with CUVAF.

Main Outcome Measure(s): The total area of CUVAF**,** heritability of CUVAF and single nucleotide polymorphisms (SNPs) associated with CUVAF from genome-wide association study.

Results: Within twin cohorts, individuals living closer to the equator (27.47° S) had higher levels of CUVAF compared to individuals from southern regions (42.88° S) (median of 45.2vs 28.7 mm^2) ($p<0.001$). The additive genetic component explained 37% (95% confidence interval [CI], 22%-50%) of the variation in CUVAF while 50% (95%CI; 29%-71%) was due to the common environment. The SNP rs1060043 located approximately 800bp away from the *SLC1A5* gene, a member of the solute carrier family 1, had a genome-wide significant association with a p-value of 3.2 x 10^{-8}**.** Gene-based analysis did not improve our power to detect association with other genes.

Conclusion: Our findings confirm that while there is a large environmental component to CUVAF (= sun exposure), genes also play a significant role. We identified a SNP (rs1060043) as being significantly associated with CUVAF; replication of this finding in future studies is warranted.

Introduction

Excessive sun exposure particularly ultraviolet-light (UV) increases the risk of many ocular diseases including pterygium [111], cortical cataract [107], ocular surface squamous neoplasia [109], climatic droplet keratopathy[223] and eyelid malignancy[110]. Despite early work suggesting sun exposure has a role in the pathogenesis of age-related macular degeneration [113] and ocular melanoma [112], these associations remain inconclusive. In recent years, a considerable number of epidemiological studies have reported that increased time spent outdoors is associated with lower rates of myopia in children, suggesting that sunlight brightness or UV-light may have a beneficial effect [128]. These conflicting reports on effects of sun exposure require a better understanding of mechanisms underlying ocular sun damage and related eye diseases.

A challenge of studying ophthalmohelioses [224] (sun-related ocular diseases) is the difficulty of assessing sun exposure. The usual method of determining an individual's sun exposure is byself-reported questionnaire which is subject to recall errors. Often questions are designed to assess whole-body sun exposure rather than ocular sun exposure, thus accuracy of these measures in ocular diseases is arbitrary. Conjunctival ultraviolet autofluorescence (CUVAF) photography was developed to detect precursors of ocular sun damage using a technique similar to UV fluorescence in the detection of UV exposure-related dermatologic diseases [225]. Previous studies have reported an association of CUVAF with the presence of pterygia[226] and shown increasing total area of CUVAF is associated with increasing prevalence of pterygium[227]. Time spent outdoors correlates highly with the level of CUVAF [128]. This suggests CUVAF could be regarded as an objective measure of sun damage corresponding to amount of time spent outdoors and could help characterize local sun exposure.

Multiple biological mechanisms have been proposed to explain the cause of detected CUVAF in other tissues. These include alterations of collagen cross-linking or changes in cell metabolites such as reduced nicotinamide adenine dinucleotide (NADH) or derivatives of amino acids like tryptophan[228].

CUVAF can be an ideal biomarker of ophthalmohelioses once its characteristics are defined better. In this current study, our main aim was to determine whether there is a genetic predisposition to variation in CUVAF identified in the three Australian cohorts. However, given that sun exposure is highly dependent on geographical location, the effect of latitudinal differences on CUVAF distribution was investigated. Following this analysis the contribution of genes to CUVAF variation was explored through a classical twin study and a genome-wide association study (GWAS).

Methodology

Participants

This study included two twin and one singleton cohorts each with Northern European ancestry from Australia. Twin pairs were identified from two existing cohorts, the Twin Eye Study in Tasmania (TEST) and the Brisbane Adolescent Twin Study (BATS). Methodologies of these studies were described in detail previously [186]. In brief, a total of 487 twin pairs (200 monozygotic [MZ], 287 dizygotic [DZ]) were recruited in the TEST through several overlapping methods, including utilization of national twin registry and existing state-wide studies. A total of 2443 individuals who were enrolled into BATS were invited to participate into the twin eye study. Among the 1199 individuals agreed to participate, there were 185 MZ and 278 DZ twin pairs. The Western Australian Pregnancy (Raine) Cohort is an ongoing longitudinal birth cohort of 2868 individuals whose mothers were initially recruited to evaluate prenatal ultrasound [216, 229]. Their offspring were subsequently assessed in detail during childhood (1, 2,3,5,8 and 10 years) and adolescence (14 and 17 years). At the 20-year cohort follow-up, 1344 participants underwent an ocular examination [183]. Comparison between the individuals who did and did not participate in the 20-year follow-up has been presented previously [135].

Ethics Approval

This study was conducted in accordance with the Declaration of Helsinki and informed consent was obtained from all adult participants and parents of minors. Approval for this study was obtained from the Human Research Ethics Committees of the University of Tasmania, Royal Victorian Eye and Ear Hospital, QIMR

Berghofer Medical Research Institute, Princess Margaret Hospital and the University of Western Australia.

Quantitative analysis of CUVAF

A camera system developed by Coroneo and colleagues [226, 230] was used to take CUVAF images for each participant. The camera system included a height adjustable table equipped with subject head-rest, camera positioning assembly, digital single-lens reflex camera (Nikon D100 (Nikon, Melville, New York, USA)), 105 mm f/2.8 Micro Nikkor (Nikkor, Melville, New York, USA) lens, and filtered electronic flash. Both nasal and temporal regions of both eyes were photographed at 0.94 magnification in total darkness. All images were saved in RGB format at the D100 settings of JPEG Fine (1:4 compression) and large resolution (3,000 2,000 pixels). The area of fluorescence in millimetres squared (mm^2) for each photograph was determined using Adobe Photoshop CS4 Extend (Adobe Systems Inc., San Jose, California, USA). Reliability of CUVAF as a biomarker of sunlight exposure has been validated previously [231].

Questionnaire

As part of the Raine Study 20-year examination, participants were asked to complete questionnaires regarding their socio-economic status, medical history and sun exposure. In relation to sun exposure, participants were asked to estimate time spent outdoors, with four possible responses to the question “In the summer, when not working at your job or at school, what part of the day do you spend outside?” Responses were ‘none’, ‘< ¼ of the day, approximately half of the day’ and ‘> ¾ of the day’. ‘None’ and ‘< ¼ of the day’ groups were combined due to low numbers in the ‘none’ category. Only socio-economic status and medical history questionnaires were available for TEST and BATS cohorts.

Study analysis was divided into three main components. These included: (1) comparison of CUVAF levels between TEST and BATS cohorts to identify effect of latitude; (2) a classical twin study using TEST and BATS cohorts to estimate heritability of CUVAF; (3) a meta GWAS study of CUVAF to identify common variants associated with this measurement by pooling data from all three cohorts.

Analytical Approach for Classical Twin Study

The classical twin model based on the multivariable linear structural equation was applied using OpenMx package in the statistical software R version 2.15.1 (R Foundation for Statistical Computing; http://www.r-project.org/). This model assumes the phenotypic variation observed between the MZ and DZ twins are due to variation in additive genetic (A), common environmental (C), and unique environmental (E) effects.

To determine the heritability of CUVAF, deterioration in the model fit was assessed by dropping each component in a hierarchical order from the full model. Each of the nested sub-models was then compared to the full model by chi-squared tests. The Akaike information criterion (AIC) was used to determine the best fitting model in which variation was explained by as a few parameters as possible. Before model fitting analyses, CUVAF was adjusted for age and gender.

Genotyping and quality control

TEST and BATS participants were genotyped using the Illumina Human 660W-Quad bead chip. A total of 1903 individuals from the Raine Study (some did not participate in the eye study) were genotyped in two different batches: 1593 individuals were genotyped in 2009 using the Human 660W-Quad bead chip and a further 310 individuals were genotyped in 2012 using the Illumina Human-OmniExpress bead chip.

As part of quality control (QC), the data were filtered by single nucleotide polymorphism (SNP) call rate <0.95, a Hardy-Weinberg equilibrium (HWE) p-value< 10^{-6} and a minor allele frequency (MAF) >0.01. To exclude population outliers, a principal component analysis (PCA) was carried out using SNPs with genotyping rate >0.98. Identical SNPs with the 1000 Genome panel were identified for the PCA analysis. All the samples beyond six standard deviations from PC1 and PC2 of 1000 Genomes British population were excluded. Individuals with identity-by-descent (IBD) estimate > 0.24 with another participants were also removed from the analysis. Genotype imputation.

TEST and BATS cohorts were imputed against the August 4, 2010 version of the publicly released 1000 Genomes Project European genotyping using MACH [220]. Likewise, Raine Study was imputed against the November 23, 2010 version of the 1000 Genome Project European genotyping using MACH. We applied a minimum passing threshold of 0.3 on the Rsq metric for each SNP as the recommended practice with MACH and a MAF>0.01.

Genome-wide Association (GWA) Studies of CUVAF

GWAS of twin cohorts and the Raine Study were conducted separately. 7,773,124 SNPs (439,454 genotyped) associations of 295 families from the TEST and BATS cohorts were carried out using MERLIN [222] with addition of age, sex and latitude as covariates in a linear model. For the Raine Study, a linear regression model in R with a PLINK interface [217] was used to determine associations between 9,131,795 SNPs (561,216 genotyped) and CUVAF. In this cohort, reported time spent outdoors had a correlation with CUVAF (r=0.19 p<0.001). Hence, it was included as a covariate along with age and gender for 661 individuals who remained in the analysis. Inverse variance weighted meta-analysis with common SNPs imputed in both cohorts (n = 5,003,381) was conducted using METAL [232]. Gene-based analysis was performed using Versatile Gene-based Association Study (VEGAS) [30] with the combined SNP p-values of the RAINE and TEST/BATS analyses as input along and the default parameters.

Results

After QC, 590 participants of 295 families from TEST/BATS and a total of 661 unrelated participants from the Raine Study had complete data available and were included in this current study. Characteristics of these three groups are displayed in Table 1. The age range varied between the cohorts, with the mean (range) age being 12 (5-51), 19 (13-28) and 20 (18-22) years in the TEST, BATS and Raine Study respectively. While there were more female (55% and 57%) participants in the TEST and BATS, more male participants (52%) participated in the Raine Study. Gender and age were correlated to CUVAF, correlation coefficient (r) being -0.09 (p=0.001) and 0.07(p=0.013) respectively in the pool of three cohorts.

Effect of latitude in distribution of CUVAF

CUVAF levels of two twin cohorts were compared based on their geographical locations. Of the 590 individuals, 146 were from Tasmania (Hobart latitude =42.88° S) and 444 from Queensland (Brisbane latitude = 27.47° S). The median CUVAF was higher in individuals from Queensland (45.41 mm^2, interquartile range [IQR]: 26.77, 68.50) compared to individuals from Tasmania (28.74mm^2, IQR: 15.01, 42.34)($p<0.001$). To ensure that this difference was not present due to confounding effect of a difference in age and gender distribution within the two twin cohorts, we adjusted CUVAF for age and gender prior to comparison. The difference remained, with median CUVAF being 43.36 mm^2 (IQR: 26.54, 66.69) in individuals from Queensland and 30.90 mm^2 (IQR: 18.96, 47.31) in individuals from Tasmania ($p<0.001$). Moreover, a similar difference was present when the analysis restricted to younger twin pairs (10-20 years old) (BATS: 47.43 mm^2 [IQR: 27.92, 66.4] vs TEST: 37.53 mm^2 [IQR: 23.64, 48.53]; $p=0.006$).

CUVAF heritability

Of the 295 twins pairs included in the analysis, 150 (50.8%) were MZ twins. The pairwise correlation coefficient of CUVAF was 0.88 for MZ twins and 0.70 for DZ twins. The slightly higher correlation of MZ twins suggests a stronger common environmental contribution for the phenotype variance, compared with the genetic contribution under a classical twin model. This observation was confirmed by univariate model fitting. The best-fit model was an additive genetic, common environment and unique environment (ACE) model adjusted by age and gender. With this model, we estimated the variation explained by the additive genetic component to be 0.37 (95% confidence interval [CI], 0.22-0.56) while the common environment component explained 0.5 (95%CI, 0.29-0.71) of the variability of the trait.

Genome-wide association (GWA)

A genome-wide significant locus rs1060043 at ($p=3.193\text{x}10^{-8}$) and suggestive loci are shown in Figure 1 and summarized in Table 2. The effect size of the CUVAF increasing allele was 11.34 mm^2 per copy. Figure 2 shows the region around the rs1060043 locus. The top ten CUVAF-associated genes obtained from the gene-based test using VEGAS and SNP meta-analysed p-value estimates are displayed in Table 3.

Discussion

A strong relationship between CUVAF and sun-related ocular damage has been reported previously [227, 231] suggesting that it could serve as a useful biomarker of ophthalmohelioses. In this study, we investigated the genetic characteristics of CUVAF. Given the possible confounding effect of geographical location of CUVAF, we initially explored the levels of CUVAF over two geographical regions defined by latitude in two ethnically homogeneous, European ancestry twin cohorts and identified lower amounts of CUVAF in individuals from lower ambient UVR region (Tasmania). Although previous studies report individuals from a higher ambient UVR region (Brisbane) spent less time outdoors compared to other regions of Australia including Tasmania, it must be noted that the intensity of UV exposure in Tasmania is lower [233]. The finding of higher CUVAF levels in Brisbane is consistent with previous work by Wlodarczyk et al. who reported Queensland as having double the pterygium surgical rate per 100,000 when compared to Tasmania [234]. Thus, pterygium may well be a sensitive indicator of UV exposure, since the cornea focuses peripheral incident light approximately twenty fold onto the usual limbal location of pterygia [224].

We assessed heritability of CUVAF and have shown additive genetic effect is responsible for up to 37% of the variance of detected CUVAF amounts indicating genes are a significant contributor to variation in CUVAF. This present finding corroborates earlier evidence showing that the tendency to develop pterygium may be inherited [235-237]. Interestingly, Hecht [235] identified eleven early onset pterygium cases in two generations resident mainly in the Midwestern USA, without known extreme environmental insult, and suggested a genetic-environmental model for pterygium two decades ago. There is also increased susceptibility to pterygium development in genetic conditions in which there are abnormal DNA repair mechanisms [238] such as xeroderma pigmentosum [239], porphyria cutanea tarda [240], polymorphous light eruption and possibly Cockayne syndrome [241].

To further understand the genetic contribution to development of CUVAF, we conducted a GWAS in both twin cohorts and the Raine Study. The meta-analysis of GWAS allowed the identification of a significant association of rs1060043, which is located 800bp upstream of the solute carrier Family 1 (Neutral Amino Acid Transporter), Member 5 (*SLC1A5*) gene on 19q13. *SLC1A5* is a peptide transporter gene expressed in retinal Muller cells and also serves as an effluxer of D-serine agonist in NMDA receptor sites [242]. Many of the genes that belong to SLC1 gene family and SLC families have been detected in human cornea, rabbit cornea and corneal epithelium cells (*SLC1A4, SLC6A14, SLC7A5*) [243-245]. Variants in *SLC45A2* and *SLC24A4* influence pigmentation traits including iris color [246]. The particular SNP identified in this study gives rise to a synonymous codon that is highly conserved in zebrafish and among multiple mammalian species including rhesus monkeys, chimpanzees, cattle and dogs suggesting that this gene has a critical function in mammals. The only locus in the best VEGAS pathway result was *C3orf58.* This gene and none of the other genes identified in the gene-based analysis had an ocular function.

The present study was designed to investigate whether genetic and environmental factors play a role in the development of CUVAF. This investigation had three important results. Firstly, individuals living in areas with higher UV radiation are more likely to have increased CUVAF. Secondly, although CUVAF was primarily caused by environmental factors, genetic factors also play a role in its development. Finally, a susceptibility locus related to CUVAF was detected. Although the study successfully demonstrated these findings, certain limitations in terms of its design and sample size must be acknowledged. For example, the environment of older twins varies, possibly due to relocation, compared to young twin pairs growing up together. Therefore inclusion of older adult twin pairs may have caused a selection bias when comparing the role of environment in presentation of CUVAF. On the other hand, when the analysis was restricted to younger twins, the effect of latitude on CUVAF remained the same. Thus this indicated that the effect of older individuals was minimal on representation of the young twin pairs in the current study. Moreover, a common limitation of single GWASs is being underpowered. Both our twins and singleton discovery cohorts were very limited in sample size that resulted in detection of inconsistent signals in individual cohort analysis. This issue was overcome by performing a meta-analysis which resulted more reliable outcomes. Overall, the current findings add to a growing body of literature contributing tothe understanding of CUVAF development. Further research investigating the role of genetics and the environment would assist in identifying individuals who are predisposed to ocular sun damage to recommend personalised health messages.

Acknowledgement

Author Contributions: Professor David Mackey and Dr Stuart MacGregor had full access to all of the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Study concept and design: Yazar, Cuellar-Partida, McKnight, Hewitt, MacGregor, Mackey

Acquisition, analysis, or interpretation of data: Yazar, Cuellar-Partida, McKnight, Quach-Thanissorn, Mountain, Pennell, Coroneo, Hewitt, MacGregor, Mackey

Drafting of the manuscript: Yazar, Cuellar-Partida, Stuart MacGregor, Mackey

Critical revision of the manuscript for important intellectual content: Yazar, Cuellar-Partida, McKnight, Quach-Thanissorn, Mountain, Hewitt, Pennell, Coroneo, MacGregor, Mackey

Statistical analysis: Cuellar-Partida, Stuart MacGregor

Obtained funding: Hewitt, Pennell, MacGregor, Pennell

Administrative, technical, or material support: Yazar, McKnight, Quach-Thanissorn , Mountain, Coroneo, Mackey

Study supervision: Mackey, Hewitt, MacGregor

Conflict of Interest: The authors have no proprietary or commercial interest in any materials or methods discussed in this article.

Funding/Support: The cohorts in this study were supported by the University of Western Australia (UWA), The Telethon Kids Institute, Raine Medical Research Foundation, Women’s and Infants’ Research Foundation, The Faculty of Medicine UWA, Curtin University, Edith Cowan University, Australian National Health and Medical Research Council Project Grants 1021105 and 350415, LEI, the Australian Foundation for the Prevention of Blindness, Alcon Research Institute, R01EY018246, Centre for Inherited Diseases Research, the Clifford Craig Medical Research Trust, Ophthalmic Research Institute of Australia (ORIA), American Health Assistance Foundation, Peggy and Leslie Cranbourne Foundation, Foundation for Children, Jack Brockhoff Foundation, the Pfizer Australia Senior Research Fellowship, and the Victorian Government of Australia. SM acknowledges Australian Research Council Future Fellowship and Australian National Health and Medical Research Council Career Development Fellowship support.

Figure 3.1. Manhattan plot of the meta-analysis association p-values for conjunctival UV autofluorescence (CUVAF). SNPs based on chromosomal position vs logarithm of the p-values. Red line denotes the genome-wide significance ($p<5x10^{-8}$). SNPs above the blue line represent the suggestive loci.

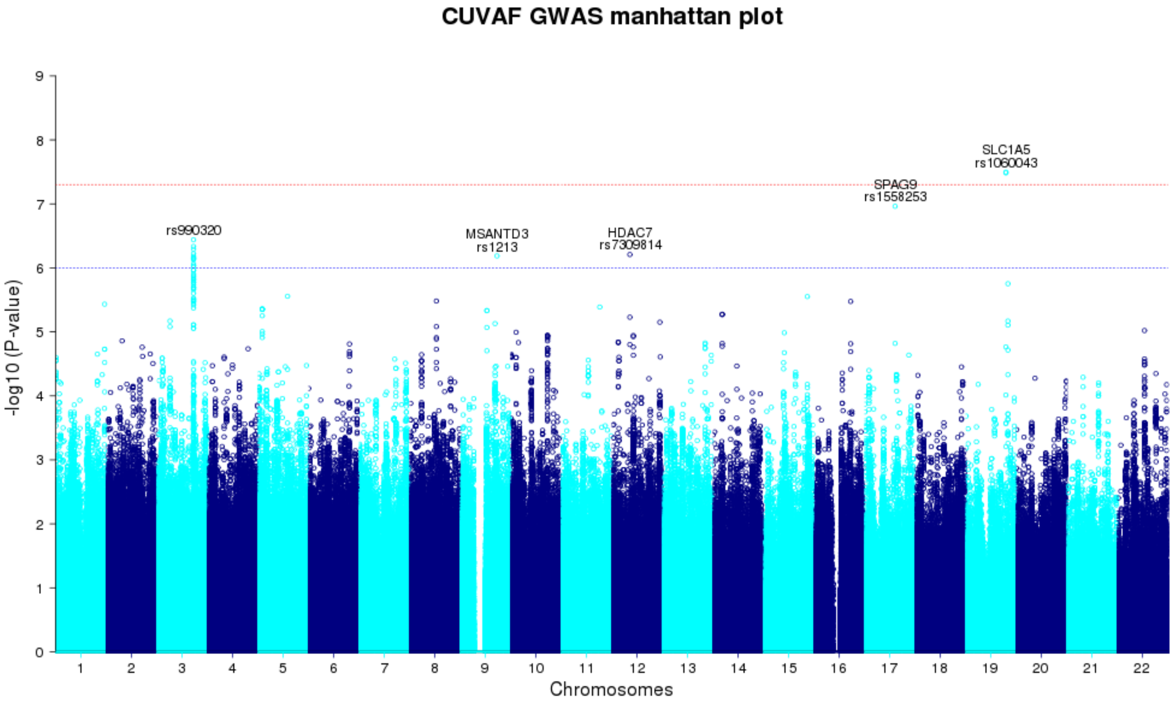


Figure 3.2. Association of variants at the SLC1A5 locus. P values (-log10) of SNP association with conjunctival UV autofluoresnce in the meta-analysis are plotted against their positions at the SLC1A5 locus.SNPs are colored to display their linkage disequilibrium (LD) with rs1060043.

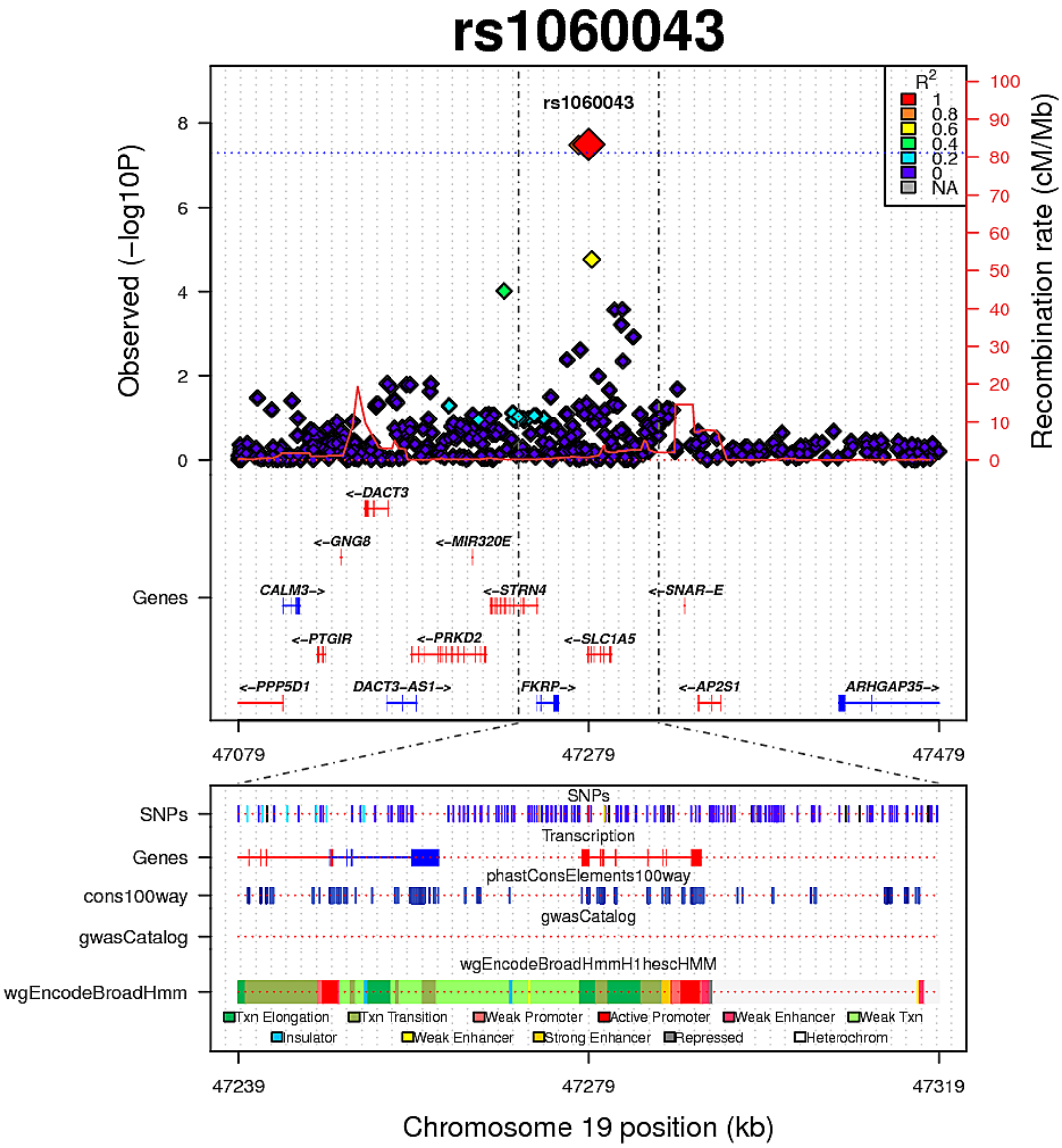


Table 3.1. Demographic characteristics of Conjunctival UV autofluorescence (CUVAF) study participants.

	TEST	BATS	Raine Study
Number of participants	146	444	661
Number of families	73	222	661
Mean age in years (range)	12 (5-51)	19 (13-28)	20 (18-22)
Number of MZ vs DZ twins	26/47	124/98	-
Gender (%females)	55%	57%	48%
Median CUVAF (IQR)	28.7 (15.0,42.3)	45.4 (26.7,68.5)	44.2 (20.3,69.8)

Table 3.2. Top five loci associated with conjunctival UV autofluorescence (CUVAF).

				TEST/BATS			Raine Study			Meta-analysis		
SNP	**CHR**	**Closest Locus**	**A1/A2**	**Effect**	**SE**	**p-value**	**Effect**	**SE**	**p-value**	**Effect**	**SE**	**p-value**
rs1060043	19	SLC1A5	A/G	7.32	2.70	0.006	16.71	3.13	$1.37x10^{-7}$	11.34	2.05	$3.19x10^{-8}$
rs1558253	17	SPAG9/NME1	T/G	-20.89	3.89	$8.47x10^{-8}$	-8.92	5.38	0.097	-16.78	3.16	$1.09x10^{-7}$
rs990320	3	C3orf58	T/C	-6.97	2.02	0.00058	-7.13	1.90	0.00019	-7.06	1.39	$3.64x10^{-7}$
rs7309814	12	HDAC7	C/G	16.89	4.56	0.00021	12.91	3.54	0.00062	13.97	2.80	$6.19x10^{-7}$
rs1213	9	MSANTD3	T/C	-34.68	10.91	0.0014	-35.53	9.27	0.00014	-35.18	7.07	$6.51x10^{-7}$

Table 3.3. VEGAS pathway analysis results for the ten most significant genes associated with conjunctival UV autofluorescence (CUVAF).

Chromosome	**Gene**	**Number of SNPs**	**Start Position**	**Stop Position**	**Test Statistic**	**p-value**	**Best-SNP**	**SNP p-value**
3	*IQCF3*	32	51837608	51839916	260.975	7.80×10^{-5}	rs9836804	6.77×10^{-6}
8	*PXMP3*	110	78055048	78075079	518.863	7.90×10^{-5}	rs7008266	8.26×10^{-6}
10	*ARMETL1*	81	14901256	14919989	354.384	1.37×10^{-4}	rs2688849	1.02×10^{-5}
14	*TRMT5*	62	60507919	60517535	285.471	1.77×10^{-4}	rs10129952	5.68×10^{-3}
9	*FANCC*	152	96901156	97119812	943.041	1.78×10^{-4}	rs4647558	3.57×10^{-5}
10	*HSPA14*	73	14920266	14953746	339.968	1.79×10^{-4}	rs2688849	1.02×10^{-5}
3	*C3orf58*	105	145173602	145193895	1003.223	2.18×10^{-4}	rs1075113	3.37×10^{-6}
16	*SNX20*	72	49264386	49272667	455.891	2.60×10^{-4}	rs6500327	4.40×10^{-5}
10	*SLIT1*	277	98747784	98935673	1073.202	2.65×10^{-4}	rs2636813	1.13×10^{-5}
1	*CD1A*	63	156490550	156494682	433.266	4.28×10^{-4}	rs614164	2.91×10^{-4}

Chapter 4

No evidence of a causal effect of vitamin D on myopic refractive error: a Mendelian randomization study.

This chapter is under review in *European Journal of Epidemiology.*

Gabriel Cuellar-Partida[1], Katie M Williams[2,3], Seyhan Yazar[4], Jeremy A. Guggenheim[5], Alex W Hewitt[6], Cathy Williams[7], Jie Jin Wang[8], Pik-Fang Kho[9], Saw Seang Mei[10,11,12], Cheng Ching-Yu[10,11,12], Wong Tien Yin[10,11,12], Aung Tin[10,11,12], Terri L. Young[13], Willem Tideman[14], Jost B Jonas[15,16], Consortium for Refractive Error and Myopia (CREAM); Paul Mitchell[8], Robert Wojciechowski[17], Dwight Stambolian[18], Pirro Hysi[3], Chris J Hammond[2,3], David A Mackey[4], Robyn Lucas[19], Stuart MacGregor[1].

Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
Department of Ophthalmology, King's College London, St. Thomas' Hospital, London, United Kingdom.
Department of Twin Research and Genetic Epidemiology, King's College London, St. Thomas' Hospital, London, United Kingdom.
Centre for Ophthalmology and Visual Science, Lions Eye Institute, University of Western Australia, Perth, Australia.
School of Optometry & Vision Sciences, Cardiff University, Cardiff, United Kingdom.
School of Medicine, Menzies Research Institute Tasmania, University of Tasmania, Hobart, Australia.
School of Social and Community Medicine, University of Bristol, Bristol, United Kingdom.
Centre for Vision Research, Department of Ophthalmology and Westmead Institute for Medical Research, University of Sydney, Sydney, Australia.

Department of Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.
Ophthalmology and Visual Sciences Academic Clinical Programme, Duke-NUS Graduate Medical School, National University of Singapore, Singapore.
Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
The Department of Ophthalmology and Visual Sciences, University of Wisconsin-Madison, Madison, Wisconsin, United States.
Department of Ophthalmology and Epidemiology, Erasmus Medical Center, Rotterdam, Netherlands.
Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Capital University of Medical Science, Beijing Ophthalmology & Visual Sciences Key Laboratory, Beijing, China.
Department of Ophthalmology, Medical Faculty Mannheim of the Ruprecht-Karls-University Heidelberg, Seegartenklinik Heidelberg, Germany.
Wilmer Eye Institute, Johns Hopkins Medical Institutions, Baltimore, MD USA.
Department of Ophthalmology, University of Pennsylvania, Philadelphia, PA, USA.
National Centre for Epidemiology and Population Health, Research School of Population Health, Australian National University, Canberra, Australian Capital Territory, Australia.

Corresponding author:
Mr Gabriel Cuellar-Partida
QIMR Berghofer Medical Research Institute
300 Herston Rd,
Brisbane, Queensland Australia 4006
Telephone: +61 8 9381 0707
Fax: +61 8 9381 0700
E-mail: *Gabriel.Cuellar@qimrberghofer.edu.au*

Word count: 2994

Abstract

The prevalence of myopia has reached alarming levels. Numerous studies have shown a strong association between less time spent outdoors and incidence of myopia. Recently, some studies showed a negative association between vitamin D (25(OH)D) levels and myopia. However, correlation does not imply causation. In this work we assess the causal role of 25(OH)D levels on the degree of myopia. To this end, we performed a Mendelian Randomization (MR) analysis using results from a meta-analysis of refractive error (RE) genome-wide association study (GWAS) that included 37,382 and 8,376 adult participants of European and Asian ancestry respectively, published by the CREAM consortium. Individual level data were available from the TwinsUK study (N=484). We used four single nucleotide polymorphisms (SNPs) in the *DHCR7* and *CYP2R1* genes with known effects on 25(OH)D concentration as instrument variables (IV). We estimated the causal effect of 25(OH)D on myopia level using a Wald-type ratio estimator based on the effect estimates from the CREAM GWAS. The estimated combined effect attributed to the 4 SNPs was 0.02 (95% CI: -0.06, 0.1) diopters (D) per 10 nmol/L increase in 25(OH)D concentration in Caucasians and -0.1 (95% CI: -0.26, 0.05) D per 10 nmol/L increase in Asians. The IVs were associated with 25(OH)D but not with the potential confounders, education, socio economic status, smoking and body mass index (P>0.05) in TwinsUK. Our study adds evidence that vitamin D is not directly involved with myopic refractive error as individuals genetically predisposed to lower 25(OH)D levels were not more myopic than expected.

Introduction

Myopia is the most common type of refractive error (RE). Its prevalence has notably increased worldwide in the past two decades, particularly in East Asian populations [247, 248].

Despite many international efforts, because of its complex nature the causes of myopia are not yet well understood [117, 118]. Numerous studies [123-129] have found that time spent outdoors is inversely associated with myopia development and a number of mechanisms have been proposed to explain this potential protective effect. One hypothesis is that time spent outdoors translates into less time performing near work activities, which may promote eye elongation as a compensatory mechanism to accommodation-induced defocus [130-132]. However, some studies have shown that the effect of time outdoors in the development of myopia is independent of the effect of near work activities [249-251]. Another hypothesis suggests that bright light enhances retinal dopamine release, which may suppress axial elongation [133, 134]. More recently, a few studies have proposed that higher vitamin D level (measured as the concentration of 25(OH)D) in serum or plasma reduces the risk of myopia [135-137]. However, a recent study from the British Avon Longitudinal Study of Parents and Children (ALSPAC) cohort found no evidence that 25(OH)D levels at age 10 years mediated the association between less time spent outdoors and higher incidence of myopia, and that the previously documented association between serum 25(OH)D and myopia was potentially confounded by time spent outdoors and the degree of sun exposure [134, 252].

In this work we aimed to clarify the role of 25(OH)D levels on myopia development. To this end, we carried out a Mendelian Randomization (MR) analysis. MR is an approach used to test and estimate the causal effect between an exposure and an outcome[253]. It uses an instrumental variable (IV) built from genetic variants with known effect on the risk factor, to make a causal inference. This approach is considered equivalent to a “natural” randomized controlled trial (RCT), as genotypes are segregated randomly from parent to offspring. Because of this random transmission of alleles, the genotypes are not related to any of the confounders (e.g. sex, age, or environmental factors such as time outdoors), which usually confound traditional epidemiological studies[93]. Here, we used genetic variants that are known to affect 25(OH)D concentrations as an IV to estimate the causal effect of 25(OH)D on RE.

Methods

Instrument variables

As IVs we used SNPs in *DHCR7* (rs7944926 and rs11234027) and *CYP2R1* (rs10741657 and rs12794714) which have been consistently reported to influence 25(OH)D levels[254-257]. Afzal *et al.* reported precise effect estimates (i.e. with low standard errors) for these 4 variants on 25(OH)D concentration in a sample of 35,334 individuals. We used these variants in preference to others in the same genes (e.g. *DHCR7: rs3829251, rs12785878 and CYP2R1: rs10766197, rs1562902, rs2060793)* [257, 258] since the latter are in linkage disequilibrium (LD) with our chosen variants and their effects were estimated in samples of lower size. We did not include variants in the *GC* gene, which encodes a vitamin D binding protein that affects 25(OH)D bioavailability though not synthesis, since its effects on 25(OH)D are reportedly unpredictable[259, 260]. Nevertheless, in Supplementary Table 1 we show the association between RE and genotype for SNPs in all 3 genes (*DHCR7*, *CYP2R1* and *GC*).

Beta coefficients and standard errors quantifying the association between RE and genotype for the SNPs rs7944926, rs10741657, rs12794714 and rs11234027 were obtained from a published genome-wide association study (GWAS) meta-analysis from the Consortium for Refractive Error and Myopia (CREAM). Full details of this meta-analysis are described elsewhere [120]. In brief, the meta-analysis included 37,382 participants from 27 studies of European ancestry and 8,376 from 5 Asian studies. All participants were aged 25 or older; mean age and RE (measured as spherical equivalent) in the European ancestry population were 55.7 (*s.d.*=12.3) and -0.1 (s.d.=0.76) respectively. Individuals of Asian ancestry had mean age of 55.8 (*s.d.*=5.54) and a mean RE of -0.34 (s.d.=1.52). Descriptions of each study cohort included in the CREAM GWAS are in Supplementary Table 2. Analysis of associations between whole genome imputed SNPs based on the HapMap 2 reference and RE was performed using age, sex and principal components as covariates (number of principal components included varied between studies). Each of the relevant SNPs was present in 25 or more of the European ancestry studies and in all the Asian studies.

Due to hyperopic shifts in adults above 50 years old, and to be able to compare more fairly our results to those from Yazar *et al*, we also performed this analysis using data from three younger cohorts: the Brisbane Adolescent Twin Study (BATS), the Twin Eye Study in Tasmania (TEST) and ALSPAC. We included 3,732 individuals from the BATS and TEST cohorts (mean age=16.90) and 3791 individuals from ALSPAC whose RE were measured at age 15. Details of the genotyping and phenotyping procedures are detailed elsewhere for ALSPAC[134, 261], BATS[185] and TEST[186].

The TwinsUK adult twin registry, based at St. Thomas' Hospital in London, compromises over 12,000 predominantly female European ancestry twins, from throughout the United Kingdom [262]. Twins who volunteered were largely unaware of the eye studies at the time of enrolment and gave fully informed consent under a protocol reviewed by the local research ethics committee (EC04/015), in accordance with the Helsinki Declaration. RE was measured using non-cycloplegic. Spherical equivalent was calculated for both eyes and the mean of the two eyes was considered. Other phenotypes were not necessarily measured at the same time as spherical equivalent. The concentration of 25(OH)D was measured in serum (units=nanomoles/litre). Smoking status (never=0, ex-smoker=1, current smoker=2), years of education, and vitamin D supplementation were assessed through questionnaire. Body mass index (BMI) was measured during clinical assessment. Socioeconomic status was graded from 1 to 5 using the Index of Multiple Deprivation score, which is based on the individual’s place of residence in the UK. Genotyping was carried out using two genotyping platforms: the HumanHap300k-Duo for part of the TwinsUK Cohort and the HumanHap610-Quad for the rest of the TwinsUK Cohort. Imputation was conducted with reference to HapMap 2 CEU population using IMPUTE2.

Statistical analysis

A recent study showed that each of the two SNPs in *DHCR7* reduced 25(OH)D concentrations by 2 nmol/L per allele [254]. Similarly, the variant rs10741657 and rs12794714 in *CYP2R1* reduced 25(OH)D concentrations by 2.5 nmol/L and 3 nmol/L respectively (Table 1) [254]. Each of the variants explained between 0.3% and 0.6% of the total variance in 25(OH)D concentrations[254]. The former study also showed that an allele score computed by summing each of the lowering alleles across the 4 genotypes explained 1% of the variance and that carrying the 8 lowering alleles resulted in a reduction of around 8nmol/L of 25(OH)D. This number is smaller than summing the effect of each allele given that the SNPs are not entirely independent from each other. Based on these effect parameters and those from the myopia GWAS meta-analysis, we estimated the causal effect of 25(OH)D on RE using the Wald-type ratio estimator:$\hat{\beta}_{iv} = \hat{\beta}_{zy}/\hat{\beta}_{zx}$,[93] where $\hat{\beta}_{iv}$ is the causal effect of vitamin D on RE, $\hat{\beta}_{zy}$ refers to the effect of the IV *z* (the SNP) on the outcome *y* (RE) and $\hat{\beta}_{zx}$ is the effect of the IV *z* on the exposure *x* (25(OH)D concentration). The standard error from this ratio estimate was approximated using the delta method[91] $\sigma_{zy}/\hat{\beta}_{zx}$. We also estimated the causal effect combining the ratio estimates of each variant using an inverse variance weighted model as described by Burgess *et al*.[92].

$$\hat{\beta}_{ivw} = \frac{\sum \hat{\beta}_{zx}\hat{\beta}_{zy}\sigma_{zy}^{-2}}{\sum \hat{\beta}_{zx}^{2}\sigma_{zy}^{-2}}$$

$$\sigma_{ivw} = \sqrt{\frac{1}{\sum \hat{\beta}_{zx}^{2}\sigma_{zy}^{-2}}}$$

In the TwinsUK data we had genotype data as well as sex, age, RE, 25(OH)D levels, BMI, smoking, vitamin D supplementation and socioeconomic status for 484 individuals (rather than relying on summary data from the larger CREAM data set). We tested the three fundamental MR assumptions to ensure the validity of the IV [93, 253]: 1) the IV must be strongly associated with the exposure variable; 2) the IV is not associated with potential confounders; 3) the IV is only associated with the outcome variable (RE) *via* the exposure (25(OH)D levels) [Figure 1]. For assumption 1, there is very strong evidence that the 4 SNPs we selected are robustly associated with 25(OH)D levels [254-256]. Additionally, in the TwinsUK Study, we showed a clearer association between an allele score containing these 4 SNPs and the 25(OH)D level. To test assumption 2, we performed a series of linear regressions between the aggregated allele score and smoking, BMI, education and socioeconomic status. Assumption 3 is difficult to test directly – however, the SNPs chosen play clear roles in vitamin D synthesis in the skin and metabolism in the liver and are unlikely to influence RE through other mechanisms.

Results

Using an MR approach, we investigated the causal association between 25(OH)D concentrations and RE, where a more negative RE indicates a higher degree of myopia.

We first computed the causal estimate using the RE GWAS summary results (N= 37,382 for Europeans and N= 8,376 from Asians) from CREAM [120] for the SNPs of interest. Based on the effects of these SNPs on 25(OH)D concentrations reported by Afzal *et al.*[254] and those reported in the RE GWAS, we estimated the causal effect of 25(OH)D concentration on RE to be not significantly different from 0 (i.e. the causal estimates β_{zy} varied from -0.7 to 0.6 diopters (D) per 10 nmol/L increase of 25(OH)D depending on the IV and the 95% CI overlapped with 0) [Table 1]. The causal estimates in three younger cohorts of European descent (TEST, BATS and ALSPAC (N= 7,523)) ranged from -0.12 to 0.19 with wider standard errors [Table 1].

Given that each of the SNPs in *DHCR7* and *CYP2R1* explain just a small fraction of 25(OH)D levels (0.3%-0.6%)[254], we investigated if their aggregated effect (which is reported to explain ~1% of the variance) had an effect on RE. We computed an inverse-variance weighted estimate of the causal effect combining the ratio estimate of each variant in a fixed-effect meta-analysis model [92]. As the SNPs within each gene are not independent of each other, we first computed the causal effect by combining the strongest SNP in each of the genes (i.e. rs7944926 and rs12794714) which yielded an estimate of $\hat{\beta}_{ivw}$=0.05 ± 0.05 (*P*>0.05) for Europeans, $\hat{\beta}_{ivw}$=-0.05 ± 0.06 (*P* > 0.05) for the young Europeans, and $\hat{\beta}_{ivw}$=-0.06 ± 0.12 (*P*>0.05) for Asians. By combining all SNPs, the estimate was again not statistically different from 0 in either Europeans ($\hat{\beta}_{ivw}$=0.02 ± 0.04; $P > 0.05$), young Europeans ($\hat{\beta}_{ivw}$=-0.06 ± 0.05; P>0.05) or Asians ($\hat{\beta}_{ivw}$=-0.1 ± 0.08; P>0.05).

The effect estimate for young Europeans ($e.g. \hat{\beta}_{ivw}$=-0.06 ± 0.05 D per 10nmol/L) was not significantly different from 0; however, it was different from the previously reported in Australian young adults by Yazar *et al.* (β=0.06 ± 0.02 D per 10nmol/L)[135], (P_{diff}=0.033 for the test of the null hypothesis of no difference between MR and observational estimates). Similarly, our estimates in Asians ($\hat{\beta}_{ivw}$=-0.1 ± 0.08) rule out the large effects reported in observational studies in Koreans by Choi *et al.* (β=0.12 ± 0.06)[136] (P_{diff}=0.035) and in Asian ancestry Australians reported by Yazar *et al.* (β=0.35 ± 0.11), (P_{diff}=0.001).

Finally, using individual level data from the TwinsUK study (N=484) (individual level data were not available within the wider CREAM study) we tested the MR assumptions [Figure 1]. Table 2 shows the effect estimates and partial correlations between the relevant SNPs and 25(OH)D level, using age, sex, vitamin D supplementation, smoking, BMI, education and socioeconomic status as covariates. Overall, the SNPs were clearly associated with 25(OH)D concentrations; the weakest association was observed for rs7944962 (R^2=0.37%; P=0.18) and greatest for rs10741657 (R^2=2.13%; P=1.2x10^{-3}). The aggregated allele score was strongly associated with 25(OH)D levels (R^2=2.38%; P=6.2x10^{-4}). The SNPs and allele score were not associated with any of the potential confounders (P>0.05). Multivariable linear regression showed no significant association between 25(OH)D levels (β_{vitD}=0.03 ± 0.03; P=0.35) and RE – however, this estimate was not significantly different from Yazar *et al.* (P_{diff}=0.28). As expected, education was negatively associated with RE (P=0.001). Neither smoking nor BMI was associated with RE.

Discussion

Some observational studies have reported that individuals with lower 25(OH)D levels are more myopic (i.e. have a more negative RE)[135-137]. However, whether this association is causal or not is unclear. Here, we hypothesized that this relationship is not causal and is more likely to be due to the confounding effects of increased time outdoors. In order to test this, we used an instrumental variable approach in an MR analysis, by using SNPs as an IV for 25(OH)D levels. Our results showed no evidence of a causal association between 25(OH)D levels and degree of RE. However, as it is difficult to prove that an exposure has no effect at all (e.g. the effect is very small to be detected), we show that if it exists it should be in the range of $\hat{\beta}_{ivw}$=0.02 (95% CI:-0.06, 0.1) diopters (D) per 10 nmol/L increase in 25(OH)D concentrations in Europeans, $\hat{\beta}_{ivw}$=-0.06 (95% CI:-0.15, 0.04) in younger Europeans and $\hat{\beta}_{ivw}$=-0.1 (95% CI:-0.26, 0.05) D per 10 nmol/L increase in Asians.

A study from Choi *et al*[136] involving 2038 adolescents from South Korea showed a significant association between 25(OH)D concentrations and RE; however, time spent outdoors was not entirely accounted for (i.e. physical exercise and area of residence were investigated but not outdoor time). A following study[135] involving young adults from Western Australia also reported that participants with low 25(OH)D_3 levels were more likely to be myopic even after accounting for the effect of time spent outdoors and conjunctival UV autofluorescence[135]; nonetheless, 25(OH)D_3 concentration is particularly affected by the amount of sun exposure, and given that time outdoors is hard to measure accurately it is possible that there was residual confounding. Further, a study from The Avon Longitudinal Study of Parents and Children (ALSPAC)[134, 252] investigated the association between 25(OH)D_2 and 25(OH)D_3 concentrations and myopia risk in 3677 participants. After an extensive analysis, they showed that 25(OH)D concentrations at age 10 years did not mediate the association between time spent outdoors and myopia measured at 8-9 years and at various time points between 7 and 15 years, respectively.

Whether 25(OH)D_3 concentrations cause myopia could be investigated via a RCT. However, this is costly and not always feasible. Instead, here we use MR, which is considered as a natural “RCT” in 34,000 individuals to test whether vitamin D has a causal role on RE. A strength of MR is that it allow us to measure differences in life-time exposure, while an RCT just describes the effect during the time of the study. Since we demonstrate no causal relationship over the lifetime, it is unlikely that an RCT over a shorter period would draw different conclusions (for this to happen, an unlikely series of events is required e.g. the effect of vitamin D increases cause X units *increase* in myopia for ages 5-9, followed by the effect of vitamin D increases causing exactly X units *decrease* in myopia for some later time period, such that over the lifetime any causal events exactly cancel out).

One of the strengths of our study is that the genetic variants we used are a robust proxy for 25(OH)D levels and have well understood roles in the vitamin D synthesis and metabolic pathway. Although these SNPs have a small effect on 25(OH)D levels (~8nmol/L), the underlying principle with instrumental variable analysis is that one intentionally carves out a small component of the overall trait variation that is not affected by confounding. Here, we found no evidence for these variants being associated with measured confounding variables. Instrumental variable estimates in our first analysis took advantage of the large samples sizes (37,382 Europeans and 8,376 Asians) of the CREAM GWAS meta-analyses and an extra sample of 7,523 young Europeans allowing us to estimate causal effect sizes with tight confidence intervals. These confidence intervals are evidence that our study is well powered to detect very small effects[263]. Also, to the extent possible, we have tested the MR assumptions using data from the TwinsUK study and followed the MR reporting checklist [264]. In the unlikely event there is confounding due to 25(OH)D associated SNPs being associated with an unmeasured confounding variable (MR assumption 2), for us to reach the conclusion we did (zero causal effect), there would have to be a true causal effect of 25(OH)D levels on RE which was exactly cancelled out by a confounding variable acting in the opposite direction and of the same magnitude [265]. Nevertheless, a number of limitations must be acknowledged. In the first part of our analysis, we did not have the actual effect estimates of the relevant genetic variants on the 25(OH)D levels of CREAM participants. Instead, the causal effects were computed using an approximation based on the effects estimated by Afzal *et al* [254] in a sample of 30,792 individuals. It must be noted any inaccuracy in estimates of the effect of each SNP may have caused a variation in the magnitude of our MR estimates. However our conclusions regarding the significance of the causal effect remain valid providing that the SNPs for 25(OH)D level constitute a strong instrument (i.e. if the SNPs are unambiguously associated with 25(OH)D level).

Although the estimates of the effect size for each of our SNPs on 25(OH)D levels vary across different published studies (across different ancestries), the SNPs are clearly significantly associated with 25(OH)D level, with the main determinant of the variation in effect size estimates being the small sample size in most studies (compared to the highly precise estimates from the 30,792 individuals in Afzal *et al* [254]). Another limitation is that it is not possible to completely rule out a very small but genuine effect of 25(OH)D level on myopia, particularly in Asian ancestry group. In summary, previous studies have shown that 25(OH)D level is a good biomarker for sun exposure[135, 136]. Careful analyses of 25(OH)D_2 and 25(OH)D_3 concentrations in the ALSPAC study suggested that the association seen between myopia and 25(OH)D levels is likely to reflect an association with sun exposure and/or time outdoors. Our study adds evidence that vitamin D is not directly involved with myopic RE as individuals genetically predisposed to lower 25(OH)D levels were not more myopic than expected.

Acknowledgements

Authors’ contributions:

Stuart MacGregor had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis
Study concept and design: SM, GCP.
Acquisition, analysis, or interpretation of data: GCP, KMW, SY, JAG, AWH, CW, JJW, PFK, SSM,CCY, WTY, AT, TLY, WT, JBJ, PM, RW, DS, PH, CJH, DAM, SM.
Drafting of the manuscript: GCP, SM, RL
Critical revision of the manuscript for important intellectual content: JAG, AWH, CW, JJW, PFK, SSM,CCY, WTY, AT, TLY, WT, JBJ, PM, RW, DS, PH, CJH, DAM, RL, SM
Statistical analysis: GCP, KMW, JAG

Data on refractive error GWAS was provided by the CREAM consortium. The list of members is included in the supplementary material. G.C.-P. is supported by scholarship from the University of Queensland and QIMR Berghofer Medical Research Institute. SM is supported by an Australian Research Council Future Fellowship. RL is supported by a National Health and Medical Research Career Development Fellowship. DAM acknowledges support from NHMRC grant APP1021105. KMW acknowledges support from a Medical Research Council (UK)

Clinical Research Training Fellowship and PGH a Fight for Sight Early Career Investigator Award. TwinsUK: The study was funded by the Wellcome Trust; European Community's Seventh Framework Programme (FP7/2007-2013). The study also receives support from the National Institute for Health Research (NIHR)-funded BioResource, Clinical Research Facility and Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust in partnership with King's College London. SNP Genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR. Funding support for SNP genotyping was also provided by the National Institutes of Health 1R01EY018246.

Table 4.1. Effect estimates (**β**) and standard errors (**SE**) of the instrumental variables (SNPs) on 25(OH)D concentrations based on [254] and refractive error based on the RE GWAS meta-analysis from CREAM [120]. **β_{zx}** and **β_{zy}** refers to the effect of the SNP on the exposure and outcome respectively and **β_{iv}** shows the causal effect estimates (diopters (D) per 10 nmol/L 25(OH)D increase).

SNP (IV)	Vitamin D*			RE Europeans *N*=37,382[a]				RE Europeans < 50 years *N*=7,523[b]				RE Asians *N*=8,376[a]			
	β_{zx}	SE_{zx}	R^2	β_{zy}	SE_{zy}	β_{iv}	σ_{iv}	β_{zy}	SE_{zy}	β_{iv}	σ_{iv}	β_{zy}	SE_{zy}	β_{iv}	SE_{iv}
rs11234027 A/G	-0.2	0.021	0.30%	0.014	0.022	-0.07	0.11	-0.024	0.030	0.12	0.15	0.040	0.036	-0.2	-0.18
rs7944926 A/G	-0.2	0.019	0.40%	-0.011	0.018	0.06	0.09	-0.038	0.025	0.19	0.13	0.020	0.042	-0.1	-0.21
rs10741657 G/A	-0.25	0.015	0.50%	0.007	0.017	-0.03	0.07	0.034	0.021	-0.14	0.08	0.024	0.05	-0.1	-0.20
rs12794714 A/G	-0.3	0.022	0.60%	-0.015	0.016	0.05	0.05	0.036	0.021	-0.12	0.07	0.015	0.042	-0.05	-0.14
rs7944926+rs12794714	-	-	1%	-	-	0.05	0.05	-	-	-0.05	0.06	-	-	-0.06	0.12
All combined	-	-	1%	-	-	0.02	0.04	-	-	-0.06	0.05	-	-	-0.1	0.08

*****Effect estimates were extracted from Azfal *et al*[254]

[a] Effect estimates are based on those from the large RE GWAS meta-analysis from CREAM[120].

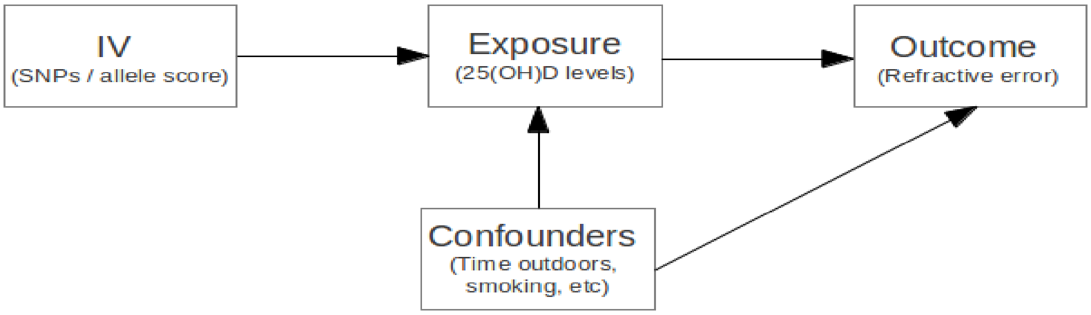
[b] Effect estimates were computed using data from individuals below 50 year old from the TEST, BATS and ALSPAC cohorts.

Table 4.2. Association between the IVs and 25(OH)D concentrations after adjusting for potential confounders*.

SNP (IV)	TwinsUK (N=484) 25(OH)D		
	β_{zx}	**SE_{zx}**	**R2**
rs11234027 A/G	-6.35	3.46	0.70%
rs7944926 A/G	-3.81	2.85	0.40%
rs10741657 G/A	-8.47	2.59	2.10%
rs12794714 A/G	-5.4	2.49	0.90%
Allele score	-2.8	0.81	2.40%

***** Vitamin D supplementation, sex, age, smoking, BMI, education and socioeconomic status.

Figure 4.1. Mendelian randomization assumptions. 1) SNPs (instrumental variable) are robustly associated with 25(OH)D concentrations (exposure variable); 2) SNPs are not correlated with the confounders; 3) The SNPs are associated to refractive error (outcome variable) through their effect on 25(OH)D concentrations.



Supplementary Material: No evidence of a causal effect of vitamin D on myopic refractive error: a Mendelian randomization study.

Table 4.3. **Supplementary Table 1.** Association between vitamin D SNPs and refractive error obtained from the genome-wide association study (GWAS) meta-analysis carried out by the Consortium for Refractive Error and Myopia (CREAM) [120].

				European Descent				Asian Descent			
SNP	Gene	Reference Allele	Other Allele	Beta	s.e.	P-value	Number of studies	Beta	s.e.	P-value	Number of studies
rs10741657	CYP2R1	G	A	0.007	0.017	0.691	25	0.024	0.050	0.626	5
rs10766197	CYP2R1	A	G	-0.005	0.016	0.734	27	0.019	0.037	0.605	4
rs12794714	CYP2R1	G	A	0.015	0.016	0.365	26	-0.015	0.042	0.717	5
rs1562902	CYP2R1	T	C	-0.011	0.016	0.480	27	-0.008	0.034	0.804	5
rs2060793	CYP2R1	G	A	0.008	0.016	0.643	27	0.023	0.050	0.638	5
rs2282679	GC	T	G	-0.022	0.018	0.217	28	0.039	0.039	0.324	5
rs7041	GC	C	A	-0.009	0.017	0.591	29	-0.007	0.035	0.840	5
rs705117	GC	T	C	0.027	0.028	0.350	28	-0.051	0.036	0.158	5
rs7944926	DHCR7	G	A	0.011	0.018	0.537	27	-0.020	0.042	0.630	5
rs11234027	DHCR7	G	A	-0.014	0.022	0.513	25	-0.040	0.036	0.260	5
rs12785878	DHCR7	T	G	0.010	0.019	0.600	25	-0.025	0.042	0.554	5
rs3829251	DHCR7	G	A	-0.012	0.022	0.589	27	-0.037	0.036	0.296	5

Table 4.4. Supplementary Table 2. Description of cohorts included in the GWAS meta-analysis carried out by CREAM [119].

Study	N	Mean age	age s.d.	Mean SPHEQ	SPHEQ s.d.
European Cohorts					
1985 British Birth Cohort	1658	42	0	-0.96	2
ALSPAC (Mothers)	1865	45	4.5	-0.76	2.16
ANZRAG	402	79.9	12	-0.26	2.36
AREDS1a1b	485	73.4	24.2	0.73	1.9
AREDS1c	1877	67.9	4.7	0.56	2.15
BMES	1550	73.8	7.76	0.62	2.12
CROATIA Korcula	822	56.3	13.3	-0.15	1.6
CROATIA Split	344	51.2	13	-1.68	1.61
CROATIA Vis	527	56.3	13.3	-0.13	1.75
DCCT	791	31.4	4.1	-1.47	1.8
EGCUT	782	57.2	17.8	0.48	3.18
ERF	2028	48.5	14.3	0.08	2.14
FECD	412	71.5	9.2	0.14	2.49
FITSA	98	68.1	3.7	1.54	1.7
Framingham	1497	55.6	8.9	0.03	2.41
Gutenberg Health Study 1	2750	55.6	10.8	-0.38	2.44
Gutenberg Health Study 2	1143	54.8	10.8	-0.41	2.58
KORA	1860	55.6	11.8	-0.82	7.33
OGP Talana	627	52.6	16.3	-0.2	2.04
ORCADES	504	57.6	13.5	0.03	2.08
RS1	5328	68.5	8.6	0.86	2.44
RS2	2009	64.2	7.4	0.48	2.51
RS3	1970	60.8	5.5	-0.35	2.62
TEST/BATS	403	38.7	13.7	-0.28	1.05
TwinsUK	3865	53.8	11	-0.4	2.73
WESDR	306	34.7	8.2	-1.5	2.02
Young Finns Study	1479	41.9	5	-1.04	2.01
Asian Cohorts					
Beijing Eye Study	578	62.1	8.8	2.45	15.73
SCES	1723	57.5	9	-0.77	2.65
SIMES	2273	58	10.8	-0.05	1.86
SINDI	2108	55.8	8.8	0.01	2.14
SP2	1694	47.7	10.8	-1.66	2.93

Chapter 5

Assessing the Genetic Predisposition of Education on Myopia: a Mendelian Randomization Study

This chapter is published as:

Cuellar-Partida, G., Lu, Y., Kho, P.F., Hewitt, A.W., Wichmann, H.E., Yazar, S., Stambolian, D., Bailey-Wilson, J.E., Wojciechowski, R., Wang, J.J. *et al.* (2016) Assessing the Genetic Predisposition of Education on Myopia: A Mendelian Randomization Study. *Genetic epidemiology*, 40 (1), 66-72.

Abstract

Myopia is the largest cause of uncorrected visual impairments globally and its recent dramatic increase in the population has made it a major public health problem. In observational studies, educational attainment has been consistently reported to be correlated to myopia. Nonetheless, correlation does not imply causation. Observational studies do not tell us if education causes myopia or if instead there are confounding factors underlying the association. In this work, we use a two-step least squares instrumental-variable (IV) approach to estimate the causal effect of education on refractive error, specifically myopia. We used the results from the educational attainment GWAS from the Social Science Genetic Association Consortium to define a polygenic risk score (PGRS) in three cohorts of late middle age and elderly Caucasian individuals (*N*=5,649). In a meta-analysis of the three cohorts, using the PGRS as an IV, we estimated that each z-score increase in education (approximately 2 years of education) results in a reduction of 0.92 ± 0.29 diopters (P=1.04x10^{-3}). Our estimate of the effect of education on myopia was higher (P=0.01) than the observed estimate (0.25 ± 0.03 diopters reduction per education z-score [~2 years] increase). This suggests that observational studies may actually underestimate the true effect. Our Mendelian Randomization (MR) analysis provides new evidence for a causal role of educational attainment on refractive error.

Introduction

The global prevalence of individuals with visual impairment in 2010 was estimated to be 285 million, of which 15% suffer blindness and 85% low vision [266]. Uncorrected refractive errors accounts for 43% of the 285 million visually impaired [266]. Myopia is the most common refractive error and occurs when the eye cannot clearly focus distant objects. More severe myopia has been associated with an increased risk of sight-threatening conditions including retinal detachment, subretinal neovascularization, macular haemorrhage, dense cataract, and glaucoma [267].

Myopia can often be corrected with optical aids such as spectacles, contact lenses, and, more recently, surgical intervention such as refractive surgery [267-270]. However, the high prevalence of myopia and cost of refractive care make this condition a significant public health concern worldwide [271]. The global economic cost in productivity loss from visual impairment due to uncorrected refractive errors is calculated to be USD$91.3 billion [272]. Also, it has been estimated that the money needed to educate personnel, establish and maintain refractive care facilities is around USD$20 billion globally [271].

Despite extensive international efforts, the causes of myopia are not yet well understood [117, 118]. Environmental factors related to socioeconomic status, time spent outdoors, near work activities and education have been consistently reported as being associated with myopia [117, 267, 273, 274]. Also, a growing body of evidence on the biological mechanisms underlying myopia suggests it results from complex interactions between the genetic makeup of an individual and the environmental exposures [268, 275-277].

Educational attainment is the most consistent environmental risk factor for myopia [274, 277]. Onset of myopia usually occurs during childhood, particularly during school years [247]. People with university-level education are 4x more likely to develop myopia than people with just primary education [247]. From the perspective of myopia epidemiology, level of education has been widely considered as a proxy measure for near work activity during the first three decades of life [117]. Near work activities, such as spending long hours in front of a computer, reading and writing, are considered important environmental risk factors for the development of myopia [130]. Performing near work activities requires the eye to generate extra optical power to focus the image on the retina, causing retinal defocus and degradation, which could then promote eye elongation as a compensatory mechanism [130, 131]. An alternate hypothesis suggests that individuals with higher education spend less time outdoors and this is the reason for an elevated risk of myopia [267, 278, 279]. Additional studies have found conflicting evidence regarding the near work hypothesis depending on the unit used to measure near work [249, 280, 281].

Recent studies have reported a gene-environment interaction between myopia genes and education [277, 282, 283]. It also has been proposed that a part of the association between education and myopia is due to pleiotropic effects (genes affecting both education and myopia, possibly as a result of education affecting subsequent myopia) [284]. A bivariate twin study has shown some evidence for a proportion of genetic factors influencing educational attainment and refractive error [274]. A large genome-wide association study (GWAS) meta-analysis estimated that genetic factors contribute to 40% of the variance in educational attainment [285]. The heritability of refractive error has been estimated to be as high as 90% [286].

In this report we investigate the effects of the genetic predisposition of education on refractive error (where a more negative refractive error indicates more myopia) in three independent cohorts of European descent. We hypothesize that the genetic correlation between refractive error and level of education is due to a causal association. We apply a Mendelian randomization (MR) approach using polygenic risk scores (PGRS) of educational attainment as an instrumental variable to establish the causal effect of education on refractive error. MR is considered to be equivalent to a randomized trial in which randomization is achieved with respect to predisposing genotypes. As genotypes are passed-on randomly from parental to offspring generations, they are immune to the confounding factors frequently present in observational studies [287].

Methods

Data

We analysed data from three different cohorts. Samples descriptors are summarized in Table I.

KORA

KORA ("Kooperative Gesundheitsforschung in der Region Augsburg" which translates as “Cooperative Health Research in the Region of Augsburg”) was accessed through dbGaP (dbGaP Study Accession: phs000303.v1.p1). The phenotyping and genotyping information are described in more detail elsewhere [288-291]. In brief, between 1984 and 2001, adults from 430,000 inhabitants living in Augsburg and 16 surrounding counties in Germany were randomly selected and separated in 4 different groups (S1-S4). One of the groups (S3/F3) was utilized for this study as was the only group with refractive error measured as spherical equivalent (SPHEQ). This study includes 1,981 subjects without medical conditions predisposing myopia, with education and genotype data along with refractive error measurements.

For each subject, eyeglass prescriptions were measured in addition to an evaluation by the Nikon Retinomax. Educational attainment was recorded as number of years of education (range 8 to 17, table I). Genotyping was done using the Illumina 2.5M chip or the Illumina Omni Express chip. Samples and SNPs were excluded if they had a low a genotype rate (<0.98). In addition, SNPs were removed if they had low minor allele frequency (<0.01) or Hardy-Weinberg P-value < 10^{-6}. The study was approved by the local ethics committee. Written informed consent was obtained from all participants before enrolment in accordance with the Declaration of Helsinki.

AREDS

The Age-Related Eye Disease Study (AREDS) was accessed through dbGaP (dbGaP Study Accession: phs000429.v1.p1). Detailed description of genotyping and phenotyping can be found elsewhere [292, 293]. In brief, AREDS participants were 55 to 80 years of age at enrolment and had to be free of any illness or condition that would make long-term follow-up or compliance with study medications unlikely or difficult. Based on ophthalmologic evaluations, 4,757 participants were enrolled in one of several categories, including a control group (AREDS 1c). Individuals included in this GWAS study are all Caucasians, who do not have age-related macular degeneration (AMD) and were further screened to also exclude individuals with cataracts, retinitis pigmentosa or other retinal degenerations, colour blindness, other congenital eye problems, LASIK, artificial lenses, and other eye surgery. For this work, we included 1842 participants from the control group (AREDS 1c), with refractive error measurements, education survey and genotype data. Refractive error was measured as SPHEQ, plus baseline measures of axis, sphere and cylinder are available for each eye. Educational attainment was recorded on a five point scale (table I). Genotyping was performed using the Illumina 2.5M chip. We applied the same quality control for samples and SNPs as for the KORA cohort. Written informed consent was obtained from all participants before enrollment in accordance with the Declaration of Helsinki.

BMES

The Blue Mountains Eye Study (BMES) is a population-based eye disease survey in individuals living in the Blue Mountains region, west of Sydney, Australia. Genotyping and phenotyping information is found elsewhere [294, 295]. In brief, 3,654 permanent residents aged 49 years or older participated (participation rate of 82.4%). During 1997-99 (BMES II A), 2,335 participants (75.1% of survivors) returned for examinations after 5 years. During 1999-2000, 1,174 (85.2%) new participants took part in an Extension Study of the BMES (BMES IIB). BMES cross-section II thus includes BMES IIA (66.5%) and BMES IIB (33.5%) participants (n=3,509). Participants underwent an eye examination including best-corrected visual acuity, objective and subjective refraction, slit-lamp examination. A Humphrey autorefractor was used to obtain an objective refraction. SPHEQ was calculated using the standard formula: SPHEQ = sphere + (cylinder/2). Educational attainment was recorded on a six point scale (table I). From the BMES cross section II who had blood samples collected, DNA was extracted for 3,189 (90.1 %) participants. Genotyping was performed on the Illumina Infinium platform using the Human660W-Quad, a WTCCC2 designed custom chip containing Human550 probes with 60,000 additional probes to capture common CNVs from the Structural Variation Consortium[296]. We applied the same QC for the SNPs as for the other two cohorts. Samples with call rate less than 95% were excluded from analysis. After initial QC, 2412 individuals had genotype and SPHEQ data; however, from these, just 1209 had education recorded. All BMES examinations were approved by the Human Ethics Committees of the Western Sydney Area Health Service and University of Sydney.

Genotype data from the remaining samples in the 3 cohorts were merged to perform relatedness filtering so that no pair of individuals had a probability of sharing an identity by descent allele (IBD) of more than 20% (~ first cousins). Further, principal component analysis was performed together with genotype data from the 1000 Genomes project. We removed all individuals that lay beyond >6 standard deviations from the 1000 Genomes northern European ancestry PC1 and PC2 centroid. The plot for the first two principal components after individuals removed is displayed in Supplementary Figure 1. Table 1 summarizes the sample sizes of each cohort after QC. Finally, we performed identity by state (IBS) clustering using PLINK –cluster which produced a single cluster, suggesting a homogeneous sample. We forced PLINK to generate 6 clusters but these were correlated to PC1 and adding the clusters as covariates did not alter our conclusions.

Statistical analysis

Given that educational attainment was coded differently in the three cohorts, it was transformed to z scores. Spearman correlations between educational attainment and refractive error were performed adjusting by sex and age.

MR is a method that permits the testing of a causal effect from observational data in the presence of confounding factors by using genetic information with a known effect on the exposure as an instrumental variable (IV) [93]. There are three fundamental assumptions to ensure the validity of the IV estimate in MR studies [93, 253]: 1) the IV must be strongly associated with the exposure variable (generally an F statistic > 10 is sufficient to ensure the validity of the IV); 2) the IV is not associated with potential confounders; 3) the IV is only associated with refractive error (outcome variable) *via* educational attainment (exposure variable) Figure 1.

Regression coefficients summarizing the results from Genome-wide association studies (GWAS) are an important source of data for MR studies. Multiple variants from these GWAS can be combined to create a powerful IV [92]. Here, we computed polygenic risk scores (PGRS) of education per individual based on the educational attainment GWAS summary results from the Social Science Genetic Association Consortium (SSGAC) [285]. These GWAS summary results were recomputed from the original SSGAC results [285] to exclude the KORA sample which was also involved in that study. The PGRS [80, 297] were estimated by summing each allele's estimated effect size multiplied by the number of risk alleles carried by each participant. We used SNPs across 12 different *P*-value thresholds (i.e. <1e-7, <1e-5, <1e-3, <1e-2, <5e-2, <1e-1, <2e-1, <3e-1, <4e-1, <5e-1), using the –score option in PLINK 1.9 [217]. Also, the PGRS were computed using the remaining SNPs after clumping for high linkage disequilibrium (clumping threshold: LD r^2=0.2 at a distance of <1Mb from the index SNP). In order to choose the PGRS with the best fit to the recorded educational attainment, we performed Spearman correlation after adjusting education by sex, age and the first 3 principal components (derived from the genome-wide genotypes) through linear regression (Figure 2).

We carried out the MR using a two-stage least squares (TSLS) approach with the *ivreg* function of the *AER* R package. In the first-stage, we predict education from the PGRS. In the second stage, we use the predicted values of education in a linear model with SPHEQ (refractive error). The *ivreg* function adjusts the second stage with the estimated residuals from the first stage to correctly account for the uncertainty of the predicted values of educational attainment. Age and sex were used as covariates. We used the Wu-Hausman test to test whether the TSLS estimates differed from the estimates obtained from a conventional linear regression between education and SPHEQ. A rejection of the null hypothesis (estimates do not differ) may indicate some inconsistency between conventional linear regression (i.e. the conventional observational study) and the TSLS which could be due to confounding or measurement errors. All the analyses were performed adjusting by sex, age and 3 principal components. Meta-analyses were performed using a weighted fixed-effect meta-analysis using the *RMETA* R package.

A study investigating genetic correlations showed a significant negative genetic correlation between attending college, obesity and smoking behavior, and a suggestive positive correlation with height [298]. Also, epidemiological studies have shown association between refractive error and anthropometric traits and smoking [136, 299]. In order to investigate potential pleiotropic effects, we performed a series of regressions between the educational attainment PGRS and BMI, height and smoking in the BMES cohort.

Results

Descriptions of the cohorts are displayed in Table I. Phenotypic correlation between educational attainment and refractive error (measured as the mean spherical equivalent, SPHEQ) for the AREDS, BMES and KORA cohorts after correcting by sex and age are summarized in Table II. Consistent with epidemiological studies, a strong negative correlation was observed in the three cohorts (ρ=-0.15 in AREDS; ρ=-0.06 in BMES; ρ=-0.10 in KORA) demonstrated by increased education resulting in more myopia.

We used data from the educational attainment GWAS from SSGAC to compute multiple PGRS of educational attainment based on different p-value thresholds of the genetic association between candidate SNPs and education. Correlation estimates between the PGRS and educational attainment are displayed in Figure 2. The PGRS computed from the top 10% SNPs (17,749 SNPs) of the educational attainment GWAS showed the most consistent and best fit to education in the three cohorts (*F*=35.5 in AREDS, *F*=9.1 in BMES and *F*=26.8 in KORA) and hence was used as IV for the MR analysis (formally, the 10% of SNPs PGRS was a strong instrument, clearly satisfying the first MR assumption). Further, we inspected the association between the PGRS and SPHEQ. The PGRS was significantly associated to SPHEQ in the AREDS (*R*=-0.09*; P*=1.4x10^{-3}) and BMES (*R*=-0.05*; P*=2.5x10^{-2}) cohorts, but not in KORA, where we observed a smaller effect size (*R*=-0.03*; P*=0.16) (Table 3).

We proceeded to perform a two-stage least squares IV analysis for the MR estimate. We found that each standard deviation from the mean of educational attainment (equals a 1 unit increase since we are working on a standardized scale, and corresponds to approximately 2 years of education) decreases SPHEQ by 0.64 – 1.33 diopters (Table IV). The IV estimates were statistically significant for the AREDS, but not for KORA and BMES cohorts. This is probably due to the smaller effect sizes seen for these cohorts and the fact that in BMES just 1209 out of 2344 participants had education measures. Further, in order to derive the most precise estimate, we meta-analysed the estimates of the three cohorts to yield the more precise estimate of 0.92 ±0.29 diopters reduction for approximately 2 years of education.

We observed that the causal effect estimate for AREDS was significantly higher than that estimated through standard observational methods (P_{diff}=7.3x10^{-3}). Provided the assumptions of the MR are satisfied, the MR estimate should reflect the true (i.e., unconfounded) effect of education on myopia. The fact that the observation study estimates are lower may be attributed to confounding (e.g. education in observational studies may be correlated with many other traits which modify myopia risk). Alternatively education SNPs (or SNPs in LD) may be associated with other traits underlying the association with refractive error. To test the latter, we investigated potential confounding effects using the BMES cohort where we had available data on smoking, height and BMI. We found no significant association between the PGRS or SPHEQ with height or BMI (*P*>0.05) (Supplementary Table 1). Smoking was nominally associated to the PGRS (*P*=0.039) but not to SPHEQ (*P*=0.129), thus it is unlikely that smoking mediates the association between the PGRS and SPHEQ. Further, the fact that educational attainment is difficult to assess accurately across studies may also have impacted results.

Discussion

In this Mendelian randomization study, we have estimated the causal effect of education on refractive error (measured as spherical equivalent) by using the genetic predisposition to education as an instrumental variable. As reported in epidemiological (observational) studies, we found a strong negative correlation between educational attainment and refractive error in three different cohorts. We also found that a genetic predictor of higher education was associated with refractive error. We assumed that any effect of education-associated genetic variants on other traits (e.g. obesity) was only via their effect on education (IV assumption 3). We also assumed that the genetic risk score for education was not associated with other traits that may confound the association between education and myopia (MR assumption 2). Our MR analysis showed a significant estimate of the causal effect of education on SPHEQ. Nevertheless, the observed MR estimate was significantly higher than the ones in the phenotypic (observational) association, suggesting some bias in either the instrumental variable or observational analysis (or both). We used principal components derived from genome-wide genotypes to control for potential population bias in all the analyses. Also, IBS clustering did not show evidence of population stratification. We believe that the observed bias may be result of an inaccurate or noisy measure of educational attainment or confounding in the observational studies. It is also possible that education SNPs (or SNPs in LD) may be associated with other traits (pleiotropic effects) underlying the association with refractive error. A recent paper from Bulik Sullivan et al [298], showed a polygenic risk score for attending college (yes/no) was correlated with the genetic risk score for a range of other traits: Alzheimer's disease, bipolar disorder, obesity, smoking and serum triglyceride levels. In the case of, say Alzheimer's, it is unlikely that Alzheimer's acts as a relevant mediator in the relationship between education-associated genes and refractive errors. For smoking, there was a nominal association between the PGRS and smoking although since we found no association between smoking and refractive error, it is unlikely that our results here are confounded via effects on smoking. Two other variables that are possible mediators [299], obesity and height, were not associated to the education PGRS Also, we note that genetic correlations between education and the other traits described in the Bulik Sullivan paper are weak [298], hence are unlikely to cause a meaningful violation of assumption 2 (although this is difficult to test). Bulik Sullivan use college yes/no, which is similar to the years of education variable we use here.

A possible source of bias in our estimates is the potential existence of an actual (unknown and/or unmeasured) pleiotropic effect of education-associated markers which cause myopia via pathways other than education. For this to violate MR assumption 3, any pleiotropic effects must not simply exist due to genetic effects influencing refractive error via their effect on education. One scenario would be a gene (or genes) with effects on both brain size and axial length, with bigger brain size being associated with a greater intelligence [300, 301] and higher education, leading to myopia. However, this scenario is unlikely given the propagation rate of genetic variants and the recent dramatic increase in myopia prevalence around the globe. Another scenario could be that education-associated genes could be inversely associated with e.g., athletic prowess, which, in turn, would be associated with increased outdoor exposure and a reduced risk of myopia. Future research should account for outdoor exposure.

A strength of our study is that it includes cohorts from Europe, Australia and the United States. However, due to modest individual sample size our results are strongest when combining the estimates. Our samples were all of European ancestry, as were the data used by the Social Science Genetic Association Consortium to derive the estimated SNP effects for educational attainment. However, as the current myopia epidemic is most marked in East Asian populations, it would be interesting in the future to perform this study in samples of Asian ancestry. Since our results indicate that observational studies may underestimate the true effect of education on myopia, for future studies of myopia where a correction for education is desired, it may be feasible to correct for a genetically derived education variable (particularly in scenarios where the education variable is missing or poorly measured). A practical limitation of this would that currently the education PGRS only explains 2% of the variance in the trait.

In conclusion we have shown that the genetic predisposition of higher education is negatively associated with refractive error. The results of our MR analysis are amongst the strongest to date in support of the notion that educational attainment is causally related to refractive error. Moreover, in the European ancestry samples studied here, the true causal effect of education on refractive error may be larger than predicted from the observational studies conducted to date.

Acknowledgements

GCP thanks the University of Queensland and QIMR Berghofer Medical Research Institute for scholarship support. PFK thanks QIMR Berghofer Medical Research

Institute for scholarship support. SM is supported by an Australian Research Council Future Fellowship. RW is supported by grant K08EY022943/EY/NEI NIH HHS/United States. DS is supported by grant R01 EY020483/EY/NEI NIH HHS/United States. JEBW is supported by the Intramural Research Program of the National Human Genome Research Institute, National Institutes of Health.

We thank Niels Rietveld (Social Science Genetic Association Consortium) for modifying and making available the summary GWAS results with the KORA samples excluded.

The dataset(s) used for the analyses described in this manuscript were obtained from the Age-Related Eye Disease Study (AREDS) Database found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000001.v3.p1. Funding support for AREDS was provided by the National Eye Institute (N01-EY-0-2127). We would like to thank the AREDS participants and the AREDS Research Group for their valuable contribution to this research.

The dataset(s) used for the analyses described in this manuscript were obtained from the NEI Refractive Error Collaboration (NEIREC) Database found at http://www.ncbi.nlm.nih.gov/gap through dbGaP accession number phs000303.v1.p1. Funding support for NEIREC was provided by the National Eye Institute. We would like to thank NEIREC participants and the NEIREC Research Group for their valuable contribution to this research.

CONFLICTS OF INTERESTS

The authors declare no conflicts of interest exist.

Table 5.1. Characteristics of the cohorts. AREDS and BMES educational attainment is coded as the higher level awarded. KORA education level is showed as education years completed. Mean and standard deviation is shown for Age and spherical equivalent (SPHEQ).

	AREDS	**BMES**	**KORA**
N	1459	2344	1846
Age (s.d.)	68.15 (4.80)	66.73 (8.96)	55.58 (11.77)
Male / Females	588 / 871	1327 / 1017	934 / 312
SPHEQ (s.d)	0.51 (2.13)	0.57 (2.02)	-0.28 (2.25)
Height (s.d)	-	M=1.72 (0.06) F=1.59 (0.07)	-
BMI (s.d)	-	M=27 (4), F=28 (5)	-
Smoking*	-	Yes=217; No=2040	-

Educational* attainment	1. Grade 11: 83;
	2. High school: 344;
	3. College: 472;
	4. Bachelors :248;
	5. Postgraduate: 312;
		1. Certificate-other: 134;
		2. Certificate-trade: 219;
		3. Diploma: 596;
		4. Bachelors: 191;
		5. Graduate diploma: 31;
		6. Higher degree: 38;
			Years of education:
			8: 168;
			10: 789;
			11: 246;
			12: 146;
			13: 223;
			15: 12;
			17: 262;

* Numbers may not add-up due to missing data.

Table 5.2. Phenotypic association (i.e. observational study estimates) of education with spherical equivalent after adjusting by sex and age. B+K+A represents the estimate of a weighted fixed-effect meta-analysis between the three cohorts.

	N	**Education level *ρ* (s.e)**	***P*-value**	**Education level β (s.e)**
AREDS	1459	-0.15 (0.013)	$1.9\text{x}10^{-9}$	-0.29 (0.06)
BMES	1209	-0.06 (0.028)	$1.9\text{x}10^{-2}$	-0.10 (0.06)
KORA	1846	-0.10 (0.023)	$2.1\text{x}10^{-6}$	-0.32 (0.05)
B+K+A		-0.11 (0.012)	$<2.2\text{x}10^{-16}$	-0.25 (0.03)

Abreviations: K+B+A: KORA + BMES + AREDS meta-analysis.

Table 5.3. Association of PGRS of education with spherical equivalent after adjusting by sex and age. B+K+A represents the estimate of a weighted fixed-effect meta-analysis between the three cohorts.

	N	**PGRS *R (s.e).***	***P*-value**
AREDS	1459	-0.09 *(0.015)*	4.1×10^{-4}
BMES(all)	2344*	-0.05 *(0.020)*	2.5×10^{-2}
KORA	1846	-0.03 *(0.022)*	1.6×10^{-1}
B+K+A		-0.05 *(0.013)*	1.4×10^{-4}

*Data available with genotype and spherical equivalent.

Abreviations: K+B+A: KORA + BMES + AREDS meta-analysis.

Table 5.4. Effect estimates from the two-stage least squares analyses using PGRS as an instrumental variable for education and spherical equivalent as outcome. P-value is for the test of whether beta is significantly different from zero. P-value$_{\text{diff}}$ corresponds to the significance of the endogeneity test (Wu-Hausman test), a rejection of the null hypothesis means that the beta effect estimates here are different from the observed (phenotypic) estimates in table II. B+K+A represents the estimate of a weighted fixed-effect meta-analysis between the three cohorts.

	N	**β *(s.e)***	***P*-value**	***P*-value$_{\text{diff}}$**
AREDS	1459	-1.33 *(0.42)*	1.41×10^{-3}	7.3×10^{-3}
BMES	1209*	-0.87 *(0.71)*	2.21×10^{-1}	2.1×10^{-1}
KORA	1846	-0.64 *(0.45)*	1.58×10^{-1}	2.2×10^{-1}
B+K+A		-0.92 *(0.29)*	1.04×10^{-3}	1.0×10^{-2}

*Data available with genotype, spherical equivalent and observed education.

Abreviations: K+B+A: KORA + BMES + AREDS meta-analysis.

Figure 5.1. Mendelian Randomization assumptions. 1) Educational attainment polygenic risk score (instrumental variable) is robustly associated with educational attainment (exposure variable); 2) IV is only associated with refractive error (outcome variable) via educational attainment (exposure variable); 3) IV is not associated to the confounders**.**

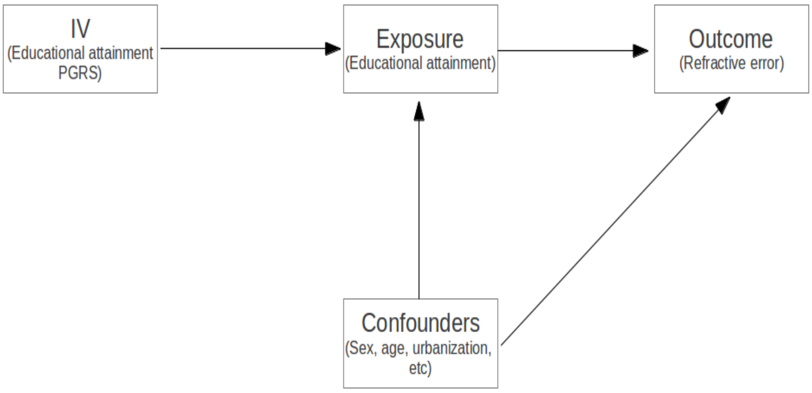
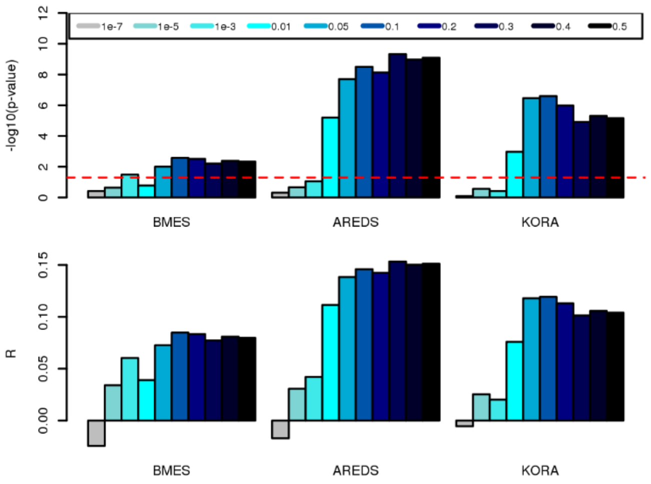


Figure 5.2. Polygenic risk scores (PGRS) of education predict educational attainment. Each bar represents the p-value threshold used to compute the PGRS of education. The upper panel shows the significance level of the association between PGRS of education and educational attainment; red dotted line represents –log10 (0.05). Lower panel indicates the Spearman correlation estimate.



Supplementary Material:

Assessing the Genetic Predisposition of Education on Myopia: a Mendelian Randomization Study

Figure 5.3. Supplementary Figure 1. Principal Component Analysis plot of the BMES, KORA and AREDS together with 1000 Genomes populations.

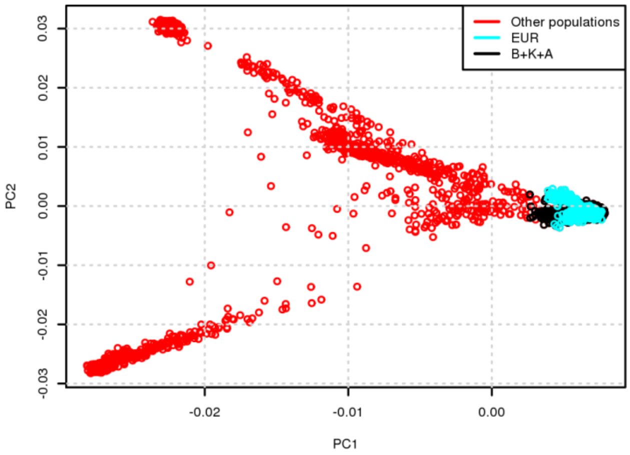


Table 5.5. Supplementary Table 1. Examination of potential confounders: Height, Smoking and BMI in the BMES cohort. All the estimate are based on linear regressions except for Smoking where we apply a logistic regression.

Formula: SPHEQ ~ Sex + Age + Height + Smoking + BMI + PGRS + PC1 + PC2 + PC3			
	Estimate	**s.e.**	**Pr(>|t|)**
Sex	-0.10	0.12	4.43E-01
Age	0.04	0.01	1.04E-12
Height	-0.57	0.69	4.12E-01
Smoking	0.21	0.15	1.57E-01
BMI	0.00	0.01	8.44E-01
PGRS	-3727.00	1529.00	1.49E-02
PC1	2.27	13.40	8.65E-01
PC2	0.51	4.09	9.00E-01
PC3	-5.88	5.22	2.60E-01
Formula: SPHEQ ~ Sex + Age + Height + Smoking + BMI			
Sex	-0.09	0.12	4.94E-01
Age	0.04	0.01	2.68E-12
Height	-0.63	0.69	3.61E-01
Smoking	0.22	0.14	1.29E-01
BMI	0.00	0.01	7.98E-01
Formula: Smoking ~ Sex + Age + PGRS + PC1 + PC2 + PC3			
Sex	0.39	0.14	5.85E-03
Age	-0.05	0.01	6.85E-09
PGRS	**-5322.00**	**2585.00**	**3.95E-02**
PC1	-9.02	9.28	3.31E-01
PC2	-5.24	6.39	4.12E-01
PC3	5.41	7.42	4.66E-01
Formula: Height ~ Sex + Age + PGRS + PC1 + PC2 + PC3			
Sex	0.13	0.00	2.00E-16
Age	0.00	0.00	2.00E-16

PGRS	84.21	47.80	7.83E-02
PC1	-1.20	0.42	4.36E-03
PC2	0.06	0.13	6.15E-01
PC3	0.29	0.16	8.06E-02
Formula: BMI ~ Sex + Age + PGRS + PC1 + PC2 + PC3			
Sex	-0.25	0.20	2.25E-01
Age	-0.07	0.01	1.26E-10
PGRS	-3725.00	3640.00	3.06E-01
PC1	23.39	31.93	4.64E-01
PC2	-2.57	9.80	7.93E-01
PC3	2.21	12.50	8.60E-01

Chapter 6

Assessment of polygenic effects links primary open-angle glaucoma and age-related macular degeneration

This chapter is under review in *Scientific Reports*.

Gabriel Cuellar-Partida,[1,*] Jamie E Craig,[2,8] Kathryn P Burdon,[3] Jie Jin Wang,[4] Brendan J. Vote,[5] Emmanuelle Souzeau,[2] Ian L. McAllister,[6] Timothy Isaacs,[6] Stewart Lake,[2] David A Mackey,[3,6] Ian J. Constable,[6] Paul Mitchell,[4] Alex W. Hewitt,[3,7,8] Stuart MacGregor[1,8,*].

1. Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, 4006, Australia.
2. Department of Ophthalmology, Flinders University, Adelaide, South Australia, 5001, Australia.
3. Menzies Institute for Medical Research, University of Tasmania, Hobart, 7001, Australia.
4. Centre for Vision Research, Department of Ophthalmology and Westmead Millennium Institute of Medical Research, University of Sydney, Sydney, New South Wales, 2145, Australia.
5. Launceston Eye Institute, Launceston, Tasmania, 7249, Australia.
6. Centre for Ophthalmology and Visual Science, Lions Eye Institute, University of Western Australia, Perth, Western Australia, 6150, Australia.
7. Centre for Eye Research Australia, University of Melbourne, Royal Victorian Eye and Ear Hospital, East Melbourne, Victoria, 3000, Australia
8. These authors jointly supervised this work

Corresponding Author:

* Gabriel Cuellar-Partida, e-mail: Gabriel.Cuellar@qimrberghofer.edu.au

** Stuart MacGregor, e-mail: Stuart MacGregor@qimrberghofer.edu.au

Abstract

Primary open-angle glaucoma (POAG) and age-related macular degeneration (AMD) are leading causes of irreversible blindness. Several loci have been mapped using genome-wide association studies. Until very recently, there was no recognized overlap in the genetic contribution to AMD and POAG. At genome-wide significance level, only ABCA1 harbors associations to both diseases. Here, we investigated the genetic architecture of POAG and AMD using genome-wide array data. We estimated the heritability for POAG (h2g= 0.42±0.09) and AMD (h2g= 0.71±0.08). Removing known loci for POAG and AMD decreased the h2g estimates to 0.36 and 0.24 respectively. There was evidence for a positive genetic correlation between POAG and AMD (rg= 0.47±0.25) which remained after removing known loci (rg = 0.64±0.31). We also found that the genetic correlation between sexes for POAG was likely to be less than 1 (rg= 0.33±0.24), suggesting that differences of prevalence among genders may be partly due to heritable factors.

Introduction

Primary open angle glaucoma (POAG) and age related macular degeneration (AMD) both show strong familial aggregation [302, 303]. AMD is a progressive disease characterized by retinal neovascularisation or atrophic degeneration in the macular and accounts for more than half of all blindness worldwide [303]. POAG is the most common type of glaucoma in populations of European ancestry, [141] and is characterized by a progressive destruction of the optic nerve leading to permanent visual loss. Genome-wide association studies (GWAS) meta-analyses have identified 35 genetic loci associated to AMD [162, 302, 304, 305] that explain a major part of the heritability of AMD [305]. In contrast, 7 loci of relatively small effect have been implicated in POAG [150-154]. At genome-wide significant level only the ATP-binding cassette transporter *ABCA1* is associated with both diseases [154, 305].

Quantification of the genetic contribution to disease can be estimated through the heritability (h^2), defined as the proportion of total phenotypic variation due to additive genetic factors. Traditionally this was performed using known family history (pedigree data). With a pedigree-based method, phenotypic similarity is related to the expected allele sharing across the genome among family members. An alternative way of measuring the genetic variance explained in ‘unrelated’ individuals (array heritability, $h^2{}_g$) is to use information from genetic markers typed on arrays instead of ‘expected sharing’ among family members [8]. Estimating $h^2{}_g$ in ‘unrelated’ individuals by including all single nucleotide polymorphisms (SNPs) with even a small effect on disease risk can give insight into the genetic architecture of a disease. Yang, *et al.* showed one can estimate the realized genetic relationship between distantly related individuals from genotype data [8].

In this work we estimate the proportion of variation explained, $h^2{}_g$, by all markers for POAG and AMD in a case-control setting using a restricted maximum likelihood (REML) approach implemented in GCTA [306] software. We also investigated whether POAG and AMD share a common genetic background beyond their overlap with the *ABCA1* locus using two methods. In the first, the genetic correlation is estimated from unrelated individuals without sample overlap, through a bivariate mixed-effects linear model [79, 306]; the second method, the cross-trait LD score regression, uses solely GWAS summary statistics and permits sample overlap [82, 83]. Furthermore, using these approaches, we also investigated whether there are significant genetic differences between genders in POAG and AMD. Even after accounting for confounders such as age, men are more likely to develop POAG than women [141, 143]. AMD has similar prevalence in each sex although prevalence is higher at later ages, due at least in part to differences in life expectancy. We try to address whether these differences can be attributed in part to genetic factors.

Results

Using genome-wide array data of 1382 AMD and 1105 POAG cases, and 1150 screened controls, we found that both AMD and POAG have a statistically significant ‘polygenic’ component underlying disease risk, $h^2{}_g = 0.71 \pm 0.08$ and $h^2{}_g = 0.42 \pm 0.09$ respectively. Autosomal $h^2{}_g$ estimates for the traits are shown in Table 1. After removing the effect of the known associated loci, the estimates of residual $h^2{}_g$ decreased from 0.42 to 0.36 for POAG. For POAG this implies the existence of many more common variants of small effect that collectively make a large contribution to genetic risk of POAG. In contrast, the $h^2{}_g$ explained by SNPs in AMD greatly decreased from 0.71 to 0.24 after removing the known loci. We found no evidence for genetic contribution of rare variants (MAF < 0.01) and X-chromosome (data not shown).

The bivariate linear mixed-model analysis showed a positive but non-significant genetic correlation between POAG and AMD $r_g = 0.17 \pm 0.19$ [Table 2]. However, using the cross-trait LD score regression approach (which allows the inclusion on more control individuals in the analysis), we observed a suggestive genetic correlation between both diseases $r_g = 0.47 \pm 0.25$ and a significant overlap between AMD and advanced POAG $r_g = 0.58 \pm 0.30$ [Table 3]. The genetic overlap became more apparent once we removed the known loci $r_g = 0.63 \pm 0.31$ for all POAG vs AMD and $r_g = 0.80 \pm 0.33$ for advanced POAG vs AMD.

Interestingly, we found that the genetic correlation between female and male POAG using the bivariate linear mixed model analysis was significantly lower than 1 ($r_g = 0.33 \pm 0.24$), indicating a significant difference between genetic architecture [Table 2]. In contrast there was not a significant difference in the genetic background between female and male AMD ($r_g = 0.71 \pm 0.23$) implying little or no sex specific genetic effects [Table 2]. Previous studies of POAG have shown that the effect sizes at some of the known loci are larger for advanced cases than for non-advanced cases. Here we found that h^2_g was larger for advanced than for non-advanced POAG although the increase was not significant and the estimated genetic correlation was 1 [Table 2 and Table 3].

Discussion

The estimated array heritabilities reported here differ from those of twin studies in two ways. First, when 'expected sharing' is used among (close) family members, both common and rare genetic variants contribute to the estimate of h^2. In contrast, estimating the genetic relationship between 'unrelated' individuals only uses information on the portion of the genome tagged by SNPs present on the array used[7]. This means h^2_g is a lower bound for h^2. Second, twin studies typically sample from the general population and hence provide heritability estimates for the disease up to the age at which the twins were ascertained. Twin studies generally ascertain individuals who are <75 years of age. In contrast, here we assume lifetime risk (up to age 75) to be of primary interest in our h^2_g calculations.

Given that rare variants are not highly correlated with common ones, their contribution to the GRM is limited. Because of this, we investigated the variance explained by creating the GRM using just rare variants on the array (MAF<0.01). The estimates were around 0, albeit with large standard errors. This suggests that the aggregate effects of the exonic variants on the arrays are not large.

The bivariate linear-mixed model allows the estimation of the SNP-correlation r_g between two traits. This correlation reflects the mean genetic correlation, meaning that small estimates can be the result of positive and negative correlations in different loci. The r_g estimate between POAG and AMD was positive but non-significant using the bivariate model, possibly due to sample size, as in this approach we had to randomly split the controls. However, using the cross-trait LD score regression approach where we could use all the controls for both diseases, we showed a significant genetic overlap between the diseases which extends beyond the known *ABCA1* locus. The overlap observed between AMD and POAG was even greater with advanced cases; however, this could be product of our limited sample size of non-advanced cases. We were unable to estimate reliable genetic correlations using the LD-score for non-advanced POAG (due to small sample size).

A potential mechanism mediating the correlation between AMD and POAG could be through heritable inflammatory mechanisms. Inflammatory events have been implicated in the development of AMD[307-310] and could affect IOP; elevated IOP is a major risk factor for POAG[311-313]. Our findings suggest that the observed overlap between POAG and AMD genetic associations at *ABCA1* represents just the tip of the iceberg in terms of genetic overlap. Larger studies of both diseases are likely to uncover more polygenes, with a subset of these polygenes expected to be common. Characterization of these common loci may offer new insights into molecular pathogenesis of both diseases.

The difference in genetic architecture between genders in POAG suggests a role of hormonal mechanisms in the patho-etiology of the disease. Several studies have reported that estrogen plays a protective role[147, 166, 314]. In addition, our results come in line with a recent study in a sample from the United States in which they found that SNPs in the estrogen pathway were associated to POAG in women but not in men[164].

Our findings regarding gender differences should be confirmed in subsequent studies as our analyses could have been hampered by our limited sample size once we stratified by sex. Also, although most of the test carried out in this study are highly correlated, some degree of multiple testing has to be acknowledged. This may impact our conclusions for the genders differences.

In summary, we have shown for the first time the important role of common variant polygenes in POAG risk. We also reveal a hitherto unappreciated genetic overlap between AMD and POAG. These results suggest that AMD GWAS could be used to prioritize POAG findings below the standard genome-wide significant threshold. We did not find significant differences between non-advanced and advanced POAG or between genders in AMD. Our results showing significant genetic differences between the POAG male and POAG female samples could explain the difference in prevalence between male and female POAG. Future work on the genetics of POAG should contemplate sex-stratified approaches.

Methods

Data

AMD cases were drawn from patients presenting to ophthalmology clinics across Australia (in particular the Lions Eye Institute, Western Australia; the Launceston Eye Institute, Tasmania; and the Flinders Medical Centre, South Australia) as well as from the population-based Blue Mountains Eye Study (BMES) [305, 315]. Advanced AMD was defined as geographic atrophy and/or choroidal neovascularisation in at least one eye and age at first diagnosis ≥ 50 years, and intermediate AMD was defined as pigmentary changes in the retinal pigmented epithelium or more than five macular drusen greater than 63µm and age at first diagnosis ≥ 50 years[305].

POAG cases were drawn from the Australia and New Zealand Registry of Advanced Glaucoma (ANZRAG) as previously described [316]. Advanced POAG was defined as a reliable 24-2 Humphrey visual field visual with a mean deviation of worse than -22dB or at least 2 out of 4 central fixation squares affected with a Pattern Standard Deviation of <0.5% and a cup:disc ratio of >0.95. Non-advanced POAG was defined as POAG related visual field loss with a corresponding optic disc appearance and cup:disc ratio of >0.7. Worst recorded intraocular pressure (IOP) was noted, but was not part of the inclusion criteria.

Controls for both diseases comprised 204 healthy controls from Flinders University, Australia and 955 healthy individuals from the BMES. The BMES is a population-based cohort study investigating the etiology of common ocular diseases among suburban residents aged 49 years or older, living in the Blue Mountains region, west of Sydney, Australia, during one of four surveys between 1992 and 2004[317]. All controls underwent a thorough ophthalmic evaluation and were confirmed to have no clinical signs of AMD or POAG.

All individuals were genotyped on the AMD consortium custom genotyping array[305]. This array includes 569,645 SNPs, approximately half of which tag common variation across the genome whilst the remainder are primarily non-synonymous coding SNPs (similar to those on the Illumina Exome arrays).

Approval for this work was obtained from the relevant Human Research Ethics Committees of the University of Sydney, the Royal Victorian Eye and Ear Hospital, the University of Tasmania, the University of Western Australia, as well as from the Southern Adelaide Clinical Human Research Ethics Committee. The study was carried out in accordance to the Declaration of Helsinki and informed consent was obtained from all participants.

Individuals were excluded to ensure that no pairs had an estimated genetic relationship > 0.05 (approximately a first cousin relationship). These individuals were excluded to minimize the chance that the phenotypic resemblance between close relatives could be because of non-genetic effects (for example, shared environment). We also excluded individuals who were beyond 6 s.d. from the genotype principal components (PCs) 1 and 2 from the 1000 Genomes[44] European population centroid.

Statistical analysis

Estimates of variance explained by all SNPs can be biased by genotyping errors and we therefore applied a stricter quality control than for typical GWAS analyses (99% calling rate, deviation from Hardy Weinberg Equilibrium ($P<1e-5$) and $MAF>0.0025$). Variance explained for the X-chromosome was estimated separately from the autosomes. Ten PCs were calculated using GCTA –pca flag and included as covariates to capture variance due to population stratification.

We used GCTA to calculate genetic relationship matrices (GRM): one for all variants in autosomes with a MAF > 0.01 (272,807 SNPs), another for all autosomes and variants MAF < 0.01 (32,299 SNPs) and one for the X-chromosome (6,902 SNPs). Both diseases were coded as binary traits (case-control status). The estimated variance explained was transformed from the observed scale to an unobserved continuous “liability” scale using a probit transformation[79]. The continuous scale is independent of the incidence of each category, enabling comparisons across traits or populations[7]. Phenotypes were modeled as a linear function of the sum of the additive effects due to all SNPs associated with trait-associated variants and residual effects. Variance components were estimated using residual maximum likelihood. For tests for whether a variance component is zero or not, the test is one-sided and under the null hypothesis the test statistic follows a 50:50 mixture of a point mass at zero and the χ_1 distribution. One-sided tests were performed for the significance of the autosomal and the sex chromosome specific variance explained (h^2_g) estimates.

Case-control studies usually have a much larger proportion of cases than do general populations and we hence correct for disease prevalence/lifetime risk. For late onset diseases such as POAG and AMD, lifetime risk increases as a person ages. To estimate variance explained we therefore need to specify the age which we are interested in. Here we assume lifetime risk to age 75 is of interest, resulting in lifetime risk estimates of 0.02[143] for POAG and 0.028 for AMD[143]. If older ages are of interest then prevalence in both cases is higher (e.g. for AMD lifetime risk to age 80 is 0.056), resulting in higher estimates of h^2_g. Similarly, if younger ages are of interest then the resultant h^2_g are lower.

To estimate the proportion of h^2_g that is explained by the SNPs already identified at genome-wide significance levels, we re-computed the GRM with the SNPs close to the genome-wide significant SNPs (+/- 1 megabase either side) removed. Since linkage disequilibrium very rarely extends beyond this, the resultant corrected h^2_g will not include the effect of the established risk loci. For POAG, 7 of the known loci in European ancestry populations (*CAV1*[150], *CDKN2BAS*[151], *TMCO1*[151], *SIX1*[152], *ABCA1*[153, 154], *GMDS*[154], *AFAP1*[154]) were removed. For AMD, we removed loci from 35 loci summarized in Fritsche *et al.[305]*.

Genetic correlation measures the proportion of genetic variance that two traits share. To minimize confounding by shared environmental factors, we estimated the genetic correlation (r_g) between traits of unrelated individuals using a bivariate mixed-effect linear model implemented in GCTA[79]. The genetic correlation is the estimated additive genetic covariance between traits, normalized by the geometric mean of the individual trait genetic variances (yielding values from -1 to +1). The additive genetic covariance was estimated by relating trait covariances between 'unrelated' individuals to genetic relationship estimates from genotype data. That is, information comes from the covariance between individuals from different sample sets (here POAG and AMD cases which although genotyped together, are independently ascertained samples). Increased covariance between traits with high genetic relationship values implies a positive genetic correlation between traits. To ensure no bias due to shared controls in the per disease analysis, controls were divided evenly and randomly (ensuring no overlap) between diseases.

We also estimated the genetic correlations using the recently developed cross-trait LD score regression approach[83] which requires only GWAS summary statistics and is not affected by sample overlap (e.g. overlap of controls). To this end, we first ran the genome-wide association analyses using the same samples as when computing h^2_g per each phenotype (i.e. we make use of all the controls for each phenotype) with SNPs with MAF >0.01, using the 10 first PCs as covariates. Genomic inflation factor for these GWAS ranged from 0.99 to 1.01. We used the LD-scores estimated by Bulik-Sullivan, *et al.*[82, 83] available at *http://www.broadinstitute.org/~bulik/eur_ldscores/* that are based on the 1000 Genomes European population and estimated by 1-cM windows. We then estimated the genetic correlation using the software available at *https://github.com/bulik/ldsc* with the default parameters.

To investigate differences between sexes in variance of liability captured by SNPs, we also estimated genetic correlation between sex where male cases and male controls were used as the first trait, female cases and female controls as the second trait. Finally, in the same manner, we investigated whether there was any difference in the genetic component between advanced and non-advanced POAG cases.

Acknowledgements

We thank the International AMD Genomics Consortium for carrying out the genotyping of the participants in this study.

GCP thanks the University of Queensland and QIMR Berghofer Medical Research Institute for scholarship support. SM is supported by an Australian Research Council Future Fellowship.

The UWA, LEI & Flinders group acknowledges financial support for participant recruitment and sample processing provided by the National Health and Medical Research Council (NHMRC) of Australia (#1023911), the Ophthalmic Research Institute of Australia, the BrightFocus Foundation and a Ramaciotti Establishment

Grant. KPB, JEC and AWH are supported by NHMRC Fellowships. The authors acknowledge the support of B. Usher-Ridge, L. Palmer, L Ma and DL Lim in patient recruitment and data collection.

The Westmead/Sydney samples were collected in three studies that were supported by the National Health and Medical Research Council (NHMRC), Australia: Grant IDs 974159, 211069, 457349 and 512423 supported the Blue Mountains Eye Study that provided population-based controls; Grant ID 302010 supported the Cataract Surgery and Risk of Age-related Macular Degeneration study that provided clinic-based early and late AMD cases and controls; and Grant ID 571013 supported the Genes and environment in late AMD study that provided clinic-based late AMD cases. The NHMRC had no role in the design or conduct of these studies.

Additional Information

Author contributions

GCP performed the analyses. GCP and SM designed the experiments and wrote the manuscript. AWH, JEC and SM supervised the project. AWH, JEC, DAM, KB, JJW, PM, BJV, ES, ILM, TI, SL and IJS collected the data. All authors reviewed the manuscript.

Competing financial interests

The authors declare no competing financial interests.

Table 6.1. Estimates of proportion of variation due to common genetic variants for POAG and AMD.

Trait	N_{Cases} / $N_{Controls}$	K (%)	h^2_g (s.e.)	P	h^2_g* (s.e.)	P
AMD	1382 / 1150	2.8	0.71 (0.08)	2.20E-16	0.24 (0.09)	2.49E-03
POAG	1105 / 1150	2.0	0.42 (0.09)	2.27E-06	0.36 (0.09)	4.04E-05
Advanced POAG	703 / 1150	2.0	0.42 (0.12)	1.35E-04	0.36 (0.12)	8.21E-04
Non-adv. POAG	402 / 1150	2.0	0.35 (0.18)	2.52E-02	0.29 (0.18)	5.03E-02
Female POAG	589 / 622	2.0	0.52 (0.16)	4.27E-04	0.49 (0.16)	1.00E-03
Male POAG	516 / 528	2.0	0.66 (0.19)	1.31E-04	0.62 (0.19)	4.82E-04
Male AMD	547 / 528	2.8	0.72 (0.20)	2.45E-05	0.34 (0.21)	4.50E-02
Female AMD	835 / 622	2.8	0.73 (0.15)	7.24E-11	0.42 (0.15)	2.11E-03

*Variance explained due to common genetic variance once we removed known associated loci for POAG[150-154] and AMD[305].

Table 6.2. Estimates of genetic correlations using a bivariate restricted maximum likelihood approach (REML) implemented in GCTA [78]. Controls were split evenly and randomly between sets in order to avoid bias in the estimate.

Set 1	Set 2	r_g (s.e.)	P	r_g^a (s.e)	P
AMD	**POAG**	0.17 (0.19)	1.79E-01	0.16 (0.30)	2.94E-01
AMD	**Adv. POAG**	0.20 (0.21)	1.62E-01	0.32 (0.36)	1.78E-01
AMD	**Not Adv. POAG**	0.12 (0.25)	3.17E-01	-0.08 (0.41)	4.25E-01
Not Adv. POAG	**Adv. POAG**	1.00 (0.46)	5.00E-01^b	1.00 (0.54)	5.00E-01^b
POAG male	**POAG female**	0.33 (0.24)	4.72E-02^b	0.25 (0.25)	9.97E-03^b
AMD male	**AMD female**	0.71 (0.23)	5.00E-01^b	0.14 (0.35)	5.13E-02^b

^{a} Estimated genetic correlation after removing known loci. For experiments involving only AMD or POAG only AMD or POAG loci were removed. For experiments involving POAG and AMD, both POAG and AMD loci were removed.

^{b} Significance estimate on whether r_g is different from 1 (H_0=1).

Table 6.3. Estimates of genetic correlations using cross-trait LD-score regression [82]. Controls for each set were the same as in Table 1, as this approach is not biased due to overlapping samples.

Set 1	Set 2	r_g	P	r_g[a]	P
AMD	**POAG**	0.47 (0.25)	6.20E-02	0.64 (0.31)	3.90E-02
AMD	**Adv. POAG**	0.58 (0.30)	5.00E-02	0.80 (0.33)	1.60E-02
AMD	**Non-adv. POAG**	0.39 (0.46)	3.96E-01	NA	NA
Non-adv. POAG	**Adv. POAG**	NA	NA	NA	NA
POAG male	**POAG female**	0.40 (0.36)	9.80E-02 [b]	0.58 (0.98)	5.00E-01 [b]
AMD male	**AMD female**	0.86 (1.08)	5.00E-01 [b]	NA	NA

[a] Estimated genetic correlation after removing known loci. For experiments involving only AMD or POAG only AMD or POAG loci were removed. For experiments involving POAG and AMD, both POAG and AMD loci were removed.

[b] Significance estimate on whether r_g is different from 1 (H_0=1).

Chapter 7

Assessing the Genetic Architecture of Epithelial Ovarian Cancer Histological Subtypes

This chapter in under review in *Human Genetics*.

Gabriel Cuellar[1; 2], Yi Lu[1], Suzanne C Dixon[3], Australian Ovarian Cancer Study[4; 5] , Peter A. Fasching[7; 8], Alexander Hein[8], Stefanie Burghaus[8], Matthias W. Beckmann[8], Diether Lambrechts[9; 10], Els Van Nieuwenhuysen[11], Ignace Vergote[11], Adriaan Vanderstichele[11], Jennifer Anne Doherty[12], Mary Anne Rossing[13; 14], Jenny Chang-Claude[15], Anja Rudolph[15], Shan Wang-Gohrke[16], Marc T. Goodman[17; 18], Natalia Bogdanova[19], Thilo Dörk[20], Matthias Dürst[21], Peter Hillemanns[22], Ingo B. Runnebaum[21], Natalia Antonenkova[23], Ralf Butzow[24], Arto Leminen[25], Heli Nevanlinna[25], Liisa M. Pelttari[25], Robert P. Edwards[26], Joseph L. Kelley[26], Francesmary Modugno[26-28], Kirsten B. Moysich[29], Roberta B. Ness[30], Rikki Cannioto[29], Estrid Høgdall[31; 32], Claus Høgdall[31; 32], Allan Jensen[31], Graham G. Giles[33-35], Fiona Bruinsma[35], Susanne K. Kjaer[31; 36], Michelle A.T. Hildebrandt[37], Dong Liang[38], Karen H. Lu[39], Xifeng Wu[37], Maria Bisogna[40], Fanny Dao[40], Douglas A. Levine[40], Daniel W. Cramer[41], Kathryn L. Terry[41], Shelley S. Tworoger[42; 43], Meir Stampfer[42; 43], Stacey Missmer[42-44], Line Bjorge[45; 46], Helga B. Salvesen[45; 46], Reidun K. Kopperud[45; 46], Katharina Bischof[45; 46], Katja K.H. Aben[47; 48], Lambertus A. Kiemeney[47], Leon F.A.G. Massuger[49], Angela Brooks-Wilson[50; 51], Sara H. Olson[52], Valerie McGuire[53], Joseph H. Rothstein[53], Weiva Sieh[53], Alice S. Whittemore[53], Linda S. Cook[54], Nhu D. Le[55], C. Blake Gilks[56], Jacek Gronwald[57], Anna Jakubowska[57], Jan Lubiński[57], Tomasz Kluz[58], Honglin Song[59], Jonathan P. Tyrer[59], Nicolas Wentzensen[60], Louise Brinton[60], Britton Trabert[60], Jolanta Lissowska[61], John R. McLaughlin[62], Steven A. Narod[63], Catherine Phelan[64], Hoda Anton-Culver[65; 66], Argyrios Ziogas[65], Diana Eccles[67], Ian Campbell[5], Simon A. Gayther[68], Aleksandra Gentry-Maharaj[69], Usha Menon[69], Susan J. Ramus[70], Anna H. Wu[70], Agnieszka Dansonka-Mieszkowska[71], Jolanta Kupryjanczyk[71], Agnieszka Timorek[72], Lukasz Szafron[71], Julie M. Cunningham[73], Brooke L. Fridley[74], Stacey J. Winham[75], Elisa V. Bandera[76], Elizabeth M. Poole[42; 43], Terry K. Morgan[77], Ellen L. Goode[78], Joellen M.

Schildkraut[79], Celeste L. Pearce[70; 80], Andrew Berchuck[81], Paul D. P. Pharoah[6; 59], Penelope M. Webb[3], Georgia Chenevix-Trench[4], Harvey A. Risch[82], Stuart MacGregor[1†]

1. Statistical Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston, QLD 4006, Australia.
2. School of Medicine, University of Queensland, St Lucia, QLD 4072, Australia.
3. Population Health Department, QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston, QLD 4006, Australia.
4. Cancer Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Road, Herston, QLD 4006, Australia.
5. Research Division, Peter MacCallum Cancer Centre, St Andrews Place, East Melbourne, Australia.
6. The Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
7. University of California at Los Angeles, David Geffen School of Medicine, Department of Medicine, Division of Hematology and Oncology.
8. University Hospital Erlangen, Department of Gynecology and Obstetrics, Friedrich-Alexander-University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen - EMN, Universitaetsstrasse 21-23, 91054 Erlangen, Germany.
9. Laboratory for Translational Genetics, Department of Oncology, University of Leuven, Belgium.
10.Vesalius Research Center, VIB, Leuven, Belgium.
11. Division of Gynecologic Oncology, Department of Obstetrics and Gynaecology and Leuven Cancer Institute, University Hospitals Leuven, Leuven, Belgium.
12. Department of Community and Family Medicine, Section of Biostatistics & Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, New Hampshire, USA
13. Department of Epidemiology, University of Washington, Seattle, WA, USA.
14. Program in Epidemiology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA.
15. German Cancer Research Center, Division of Cancer Epidemiology, Heidelberg, Germany.

16. Department of Obstetrics and Gynecology, University of Ulm, Ulm, Germany.
17. Cancer Prevention and Control, Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles, California, USA.
18. Community and Population Health Research Institute, Department of Biomedical Sciences, Cedars-Sinai Medical Center, Los Angeles, California, USA.
19. Radiation Oncology Research Unit, Hannover Medical School, Hannover, Germany.
20. Gynaecology Research Unit, Hannover Medical School, Hannover, Germany.
21. Department of Gynecology, Jena-University Hospital-Friedrich Schiller University, Jena, Germany.
22. Clinics of Obstetrics and Gynaecology, Hannover Medical School, Hannover, Germany.
23. N.N. Alexandrov National Cancer Centre of Belarus, Minsk, Belarus.
24. Department of Pathology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
25. Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
26. Department of Obstetrics, Gynecology and Reproductive Sciences, Division of Gynecologic Oncology, University of Pittsburgh School of Medicine, Pittsburgh, PA, USA.
27. Womens Cancer Research Program, Magee-Womens Research Institute and University of Pittsburgh Cancer Institute, Pittsburgh, PA, USA.
28. Department of Epidemiology, University of Pittsburgh Graduate School of Public Health, Pittsburgh, PA, USA.
29. Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY, USA.
30. The University of Texas School of Public Health, Houston, TX, USA.
31. Department of Virus, Lifestyle and Genes, Danish Cancer Society Research Center, Copenhagen, Denmark
32. Molecular Unit, Department of Pathology, Herlev Hospital, University of Copenhagen, Copenhagen, Denmark.
33. Department of Epidemiology and Preventive Medicine, Monash University, Melbourne, Victoria, Australia.

34. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Victoria, Australia.
35. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia.
36. Department of Gynaecology, Rigshospitalet, University of Copenhagen, Copenhagen, Denmark.
37. Department of Epidemiology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
38. College of Pharmacy and Health Sciences, Texas Southern University, Houston, Texas, USA.
39. Department of Gynecologic Oncology, The University of Texas MD Anderson Cancer Center, Houston, Texas, USA.
40. Gynecology Service, Department of Surgery, Memorial Sloan Kettering Cancer Center, New York, NY, USA.
41. Obstetrics and Gynecology Epidemiology Center, Brigham and Women's Hospital, Boston, Massachusetts, USA.
42. Department of Epidemiology, Harvard School of Public Health, Boston, Massachusetts, USA.
43. Channing Division of Network Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
44. Department of Obstetrics and Gynecology, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA.
45. Department of Gynecology and Obstetrics, Haukeland University Horpital, Bergen, Norway.
46. Centre for Cancer Biomarkers, Department of Clinical Science, University of Bergen, Bergen, Norway.
47. Radboud University Medical Centre, Radbond Institute for Health Sciences, Nijmegen, Netherlands
48. Netherlands Comprehensive Cancer Organisation, Utrecht, The Netherlands.
49. Radboud University Medical Center, Radboud Institute for Molecular Life Sciences, Department of Obstetrics and Gynaecology, Nijmegen, The Netherlands.
50. Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, BC, Canada.

51. Department of Biomedical Physiology and Kinesiology, Simon Fraser University, Burnaby, BC Canada.
52. Memorial Sloan Kettering Cancer Center, Department of Epidemiology and Biostatistics, New York, NY, USA.
53. Department of Health Research and Policy - Epidemiology, Stanford University School of Medicine, Stanford CA, USA.
54. Division of Epidemiology and Biostatistics, Department of Internal Medicine, University of New Mexico, Albuquerque, New Mexico, USA.
55. Cancer Control Research, BC Cancer Agency, Vancouver, BC, Canada.
56. Pathology and Laboratory Medicine, University of British Columbia, Vancouver BC, Canada.
57. International Hereditary Cancer Center, Department of Genetics and Pathology, Pomeranian Medical University, Szczecin, Poland.
58. Institute of Midwifery and Emergency Medicine, Clinic of Obstetrics and Gynecology, Frederick Chopin Clinical Provincial Hospital No 1, Faculty of Medicine, University of Rzeszów, Poland.
59. The Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK.
60. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda MD, USA.
61. M. Sklodowska-Curie Memorial Cancer Center, Warsaw, Poland.
62. Public Health Ontario, Toronto, ON, Canada.
63. Women's College Research Institute, University of Toronto, Toronto, Ontario, Canada.
64. Department of Cancer Epidemiology, Moffitt Cancer Center, Tampa, FL, USA.
65. Department of Epidemiology, University of California Irvine, Irvine, California, USA.
66. Center for Cancer Genetics Research & Prevention, School of Medicine, University of California Irvine, Irvine, California, USA.
67. Faculty of Medicine, University of Southampton, Southampton, UK.
68. Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, California, USA,.

69. Women's Cancer, Institute for Women's Health, University College London, London, United Kingdom.
70. Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, California, USA.
71. Department of Pathology and Laboratory Diagnostics, the Maria Sklodowska-Curie Memorial Cancer Center and Institute of Oncology, Warsaw, Poland.
72. Department of Obstetrics, Gynaecology and Oncology, IInd Faculty of Medicine, Warsaw Medical University and Brodnowski Hospital, Warsaw, Poland.
73. Department of Laboratory Medicine and Pathology, Division of Experimental Pathology, Mayo Clinic, Rochester, MN, USA.
74. Department of Biostatistics, University of Kansas, Kansas City, Kansas, USA.
75. Department of Health Sciences Research, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, MN, USA.
76. Rutgers Cancer Institute of New Jersey, New Brunswick, New Jersey, USA.
77. Departments of Pathology and Obstetrics & Gynaecology, OHSU, Portland, OR, USA.
78. Department of Health Science Research, Division of Epidemiology, Mayo Clinic, Rochester, Minnesota, USA
79. Department of Public Health Sciences, University of Virginia, Virginia, USA.
80. Department of Epidemiology, University of Michigan School of Public Health, Ann Arbor, Michigan, USA.
81. Duke Cancer Institute, Duke University Medical Center, Durham, North Carolina, USA
82. Department of Chronic Disease Epidemiology, Yale School of Public Health, New Haven, Connecticut, USA.

Abstract

Epithelial ovarian cancer (EOC) is one of the deadliest common cancers. The five most common types of disease are high-grade and low-grade serous, endometrioid, mucinous and clear–cell carcinoma. Each of these subtypes presents distinct molecular pathogeneses and sensitivities to treatments. Recent studies show that certain genetic variants confer susceptibility to all subtypes whilst other variants are subtype-specific. Here we perform an extensive analysis of the genetic architecture of EOC subtypes. To this end, we used data of 10,014 invasive EOC patients and 21,233 controls from the Ovarian Cancer Association Consortium genotyped in the iCOGS array (211,155 SNPs). We estimate the array heritability (attributable to variants tagged on arrays) of each subtype and their genetic correlations. We also look for genetic overlaps with factors such as obesity, smoking behaviours, diabetes, age at menarche, and height. We estimated the array heritabilities of high-grade serous disease (h^2_g= 8.8 ± 1.1%), endometrioid (h^2_g= 3.2 ± 1.6%), clear-cell (h^2_g = 6.7 ± 3.3%) and all EOC (h^2_g= 5.6 ± 0.6%). Known associated loci contributed approximately 40% of the total array heritability for each subtype. The contribution of each chromosome to the total heritability was not proportional to chromosome size. Through bivariate and cross-trait LD score regression, we found evidence of shared genetic backgrounds between the three high-grade subtypes: serous, endometrioid and undifferentiated. Finally, we found significant genetic correlations of all EOC with diabetes and obesity using a polygenic prediction approach.

Introduction

In developed countries, epithelial ovarian cancer (EOC) is the leading gynaecological malignancy with an estimated annual incidence rate of 12 per 100,000 and a poor 5 year survival between 20% and 50% [168, 318, 319]. About 90% of invasive tumours in the ovary are of epithelial origin [320]. These tumours are divided into various histological subtypes that include: serous, mucinous, endometrioid, clear cell, Brenner, other minor types, as well as undifferentiated, mixed and unclassified carcinomas [167, 168]. Serous carcinomas can be subdivided into high-grade (90%) and low-grade disease (10%) [321-323].

Each epithelial ovarian cancer histologic subtype exhibits a distinct etiologic and molecular pathogenesis and sensitivity to treatment (e.g., chemotherapeutic agents) [322, 324-327]. It has been suggested that serous carcinomas arise from the epithelial mucosal lining of the fallopian tube fimbriae or from endosalpingiotic deposits on the ovarian or peritoneal surfaces. Clear-cell and endometrioid subtypes may arise from endometriotic lesions [320, 328], while mucinous tumours do not yet have a clear origin, though metaplastic transformation of the epithelial lining of ovarian inclusion cysts has been suggested. Serous carcinoma is by far the most deadly type of EOC, with 5-year survival of less than 20% for patients suffering from high-grade disease and 50% for those with low-grade disease [323]. In contrast, women with mucinous, endometrioid or clear-cell carcinomas tend to have better prognosis, with estimated 5-year survivals of 50%-60% [323, 329]. These differences in survival are due at least in part to the fact that high-grade serous carcinomas are usually detected at advanced stages of disease but the other subtypes at earlier stages [323, 329, 330].

Genetic studies have shown that around 20% of patients with high-grade serous cancers carry germ-line and somatic mutations in *BRCA1* or *BRCA2* [331, 332] along with somatic mutations in *TP53* that are present in most tumours [178]. Alterations in *KRAS* and *BRAF* but not *TP53* have been associated with low-grade serous carcinomas [325, 333, 334]. Mucinous carcinomas also frequently have somatic mutations in *KRAS* [177] in addition to mutations in *HER2* [324]. Endometrioid and clear cell carcinomas often carry somatic mutations in *AR1D1A* and *PIK3CA* [335]. In addition, genome-wide association studies (GWAS) have found 20 common polymorphisms associated with risk of EOC [336-341].

Specific germ-line SNPs are commonly found in the different EOC subtypes. However, these variants explain only a fraction of the cases, thus it is not known whether or not other genetic components are shared among the subtypes. One of our previous studies [342] estimated the array heritability (i.e., heritability explained by about 200,000 genotyped SNPs but not all the genome) of all EOC to be 5.6%, and 8.8% for the most common EOC subtype, high-grade serous.

Beside genetic factors predisposing to these diseases, some environmental factors such as smoking [169, 170] and obesity [171-173] may be associated with increases in risk of some subtypes of EOC. In addition, traits including achieved height [171, 343] and diabetes mellitus [174, 175] have been positively associated to EOC. In contrast, some studies have shown that age at menarche [176] is inversely associated with risk of EOC. Evidence suggests that all these traits have heritable components. Genetic variation may explain as much as 80% of the total variance of height [8] or even 40% for smoking behaviour [344, 345]. It is possible that part of the heritability of EOC may be explained by the heritability of these traits, if they are associated with EOC risk.

In this work, we investigate three aspects of the genetic architecture of EOC and its subtypes: (i) the total genetic contribution of all array-genotyped SNPs (genome-wide, per chromosome and after accounting for known EOC associated loci); (ii) the genetic correlations between EOC subtypes; and (iii) the genetic correlations between EOC subtypes and risk factors such as obesity and smoking. To this end, we use genotype and risk-factor data from studies participating in the Ovarian Cancer Association Consortium (OCAC). We quantify genetic contributions to disease using genome-wide complex trait analysis (GCTA) [7, 8, 306]. Then, we evaluate shared genetic backgrounds between EOC subtypes and candidate risk factors using complementary approaches: bivariate linear mixed models [79], cross-trait LD score regression [83] and polygenic risk prediction [297].

Methods

Data

We used data from the Ovarian Cancer Association Consortium (OCAC). This dataset consists of custom Illumina iCOGS array genotyping of 47,630 cases and controls in 43 OCAC studies. Detailed description of the content of the array can be found elsewhere [339]. In brief, the array consists of 211,155 variants within breast, ovarian and prostate cancer susceptibility loci as well as candidate SNPs, SNPs associated with other cancers and SNPs associated with relevant quantitative traits such as body mass index (BMI) and the onset of menarche.

We applied standard quality control (QC) for the genotype data. First, we selected only samples from European ancestry studies and that were within 6 s.d. from the genotype-derived PC1 and PC2 from the 1000 Genomes European population [Supplementary Figure 1]. We excluded individuals with missing genotypes in 5% or more of the SNPs. Likewise, we removed SNPs with call rates below 99%, minor allele frequencies (MAF) below 1% and SNPs that deviated from Hardy-Weinberg equilibrium at $P<0.0001$ [346]. Further, given that our analytic methods are sensitive to relatedness (e.g., results may be biased by common environmental factors in relatives) we removed individuals such that no sample pairs had identity by descent (IBD) > 10% (i.e., less than second cousins), giving more priority to keeping cases than controls. In concordance with one of our previous work [342], we focused only on those with invasive EOC tumours. In total, 10,014 EOC cases and 21,233 controls met these criteria and were genotyped for 195,183 SNPs. The number of cases according to histologic subtype are displayed in Table 1. The numbers of initial cases and controls per study are summarized in Supplementary Table 1.

Analysis

We estimated the variance explained by all SNPs in the array (h^2_g) [7], the variance after removing known loci, and the variance explained by each chromosome for each of the EOC subtypes. We used GCTA to calculate one genetic relationship matrix (GRM) for all autosomes.

The estimated variance explained was transformed from the observed scale to an unobserved continuous “liability” scale using a probit transformation [7] taking into account the disease prevalence. The lifetime risk of the various EOC subtypes were calculated as the lifetime risk of ovarian cancer (~1% according to the Surveillance, Epidemiology and End Results (SEER), http://seer.cancer.gov/statfacts) multiplied by the relative proportion of each subtype according to SEER program DevCan database (http://surveillance.cancer.gov/devcan/canques.html) in all ovarian cancer. Given that around 90% of ovarian cancers are of epithelial origin, we used 0.9% as the prevalence for all EOC. As h^2_g, is derived solely from the SNPs tagged on the genotyping array instead of the whole genome, it provides a lower bound on heritability estimates [346]. Phenotypes were modeled as a linear function of the sum of the additive effects due to all SNPs associated with trait-associated variants and residual effects. Variance components were estimated using residual maximum likelihood (REML) [8]. For tests of whether a variance component is zero or not, the test is one-sided and under the null hypothesis that the test statistic follows a 50:50 mixture of a point mass at zero and the χ_1 distribution [8, 306]. One sided p-values were calculated to estimate statistical significance. Likewise, To estimate the proportion of h^2_g that is explained by the known loci (WNT4, RSPO1, SYNPO2, GPX6, ABO, ATAD5, C19orf62, CMYC, TIPARP, BNC2, ARHGAP27, TERT, RAD51B/C/D, BRIP1, BARD1, PALB2, NDN, CHMP4C, MLLT10, HNF1B, *BRCA1*, *BRCA2, KRAS, TP53, HER2, AR1D1A* and *PIK3CA* [336-341]), we re-computed the GRM with the SNPs (6,391 SNPs) close to the known loci SNPs (+/- 1 megabase either side) removed.

Similarly, in order to investigate the genetic contributions within of each of the chromosomes, we computed one GRM per chromosome and performed analyses using REML fitting the 22 genetic variance components in the model as implemented in GCTA with the flag –*mgrm* (multiple GRMs) [347]. Given that loading 22 GRMs with the 21,051 controls and the cases of the various histotypes was computationally intractable, we assigned to each case just one control of the same study, yielding smaller GRMs (e.g., for high-grade Serous cancer there were 3,705 cases and 3,705 controls). We then normalized the contribution of each chromosome by the number of independent SNPs (percentage) in the iCOGs array per chromosome. This number of independent SNPs was estimated through LD pruning using the PLINK command –*indep* 50 5 1.2, where 50 is the window size (#SNPs), 5 is the number of

SNPs the window can shift, and 1.2 is $1/(1-R^2)$, where R^2 is the multiple correlation coefficient for a SNP regressed on all other SNPs simultaneously [348]. In order to approximate the s.e. of the variance explained by each chromosome, we performed a jackknifing procedure up to 1000 times, taking 80% of the cases and 80% of the controls each time. Given the complexity of the sample, around 20% of the jackknifing repetitions did not converge within 1000 iterations so the standard errors were computed from just the 800 successful jackknifings.

To investigate the genetic correlations between the subtypes, in order to remove potential biases from overlapping control samples from the different studies, we matched each case to 1 control of the same study, and distributed controls in such a way that each EOC subtype had separate sets of controls. For example, all of the controls for mucinous EOC were different from the endometrioid EOC controls.

Genetic correlation (r_g) represents the proportion of the total genetic variance that two traits share. In order to investigate the r_g between EOC subtypes, we used two distinct approaches that can be applied to population-based samples. We first used the GRM in a bivariate mixed-effects linear model implemented in GCTA [349] to compute the genetic correlations between the various EOC subtypes. The estimated genetic correlation is the additive genetic covariance between traits, normalized by the geometric mean of the individual trait genetic variances (producing values from -1 to +1). The additive genetic covariance was estimated by relating trait covariances between unrelated individuals to genetic relationship estimates from marker data. Increased covariance between traits with high genetic relationship values implies a positive genetic correlation between traits. In order to control for any potential effects of population stratification, all the analyses were performed using the first 10 principal components (PCs) of the genotypes as covariates. Estimates are reported as genetic correlation ± standard error.

We also used cross-trait LD score regression [83], a recently developed approach that is able to estimate genetic correlations using solely GWAS summary statistics and is not affected by sample overlap. We first ran genome-wide association analyses using the same samples as when computing h^2_g per each EOC subtype (i.e., we repeatedly made use of all of the controls for analysis of each subtype) and with the 10 first PCs and study site as covariates. Genomic inflation factors for these GWAS analyses ranged from 0.99 for mucinous cancer to 1.07 for all EOC. We used the LD-scores estimated by Bulik-Sullivan, *et al.*[82, 83] available at *http://www.broadinstitute.org/~bulik/eur_ldscores/* which are based on the 1000 Genomes European population and estimated within 1-cM windows. We then estimated the genetic correlation using software available at *https://github.com/bulik/ldsc* with the default parameters.

Genetic correlations between EOC subtypes and risk factors

Using cross-trait LD score regression, we estimated genetic correlations between risk factors and EOC histotypes. To this end, we used publicly available GWAS summary results from the latest GWAS meta-analyses of BMI and height from the Genetic Investigation of Anthropometric Traits (GIANT) consortium. These analyses included 339,225 [350] and 253,288 [351] individuals, respectively. We also estimated genetic correlations using the GIANT extreme anthropometric traits GWAS which used obesity class 1 (BMI>30), class 2 (BMI>35) and class 3 (BMI>40) groups as cases, and individuals with BMI<=25 as controls, in a sample of 263,407 individuals [352]. Genetic overlaps with age at menarche was carried out based on the GWAS of the Reproductive Genetics Consortium which involved 182,416 women [353]. Smoking behaviour genetic predisposition was approximated based on the Tobacco and Genetics Consortium GWAS which involved 74,053 participants [354]. Finally, for diabetes, we used the summary results for type 2 diabetes GWAS of the DIAGRAM (DIAbetes Genetics Replication And Meta-analysis) consortium, which involved 34,840 cases and 114,981 controls [355].

We also carried out a polygenic risk-prediction approach. This method involves the computation of polygenic risk scores (PGRS) of each of the risk factors and uses these scores to predict disease status [297]. The PGRS describes a predicted phenotypic value based on the genetic component and is computed by aggregating the magnitude of associations of many variants. These associations are estimated using a discovery set of subjects (e.g., for height or BMI) to identify relevant SNPs and estimate the magnitude of association of each, and these magnitudes or the number of “high-risk” alleles in each SNP are then summed to create a score. Subsequently, we examine the association of this score within a target subject set (e.g., EOC cases and controls). If the score association is significant, it implies a genetic correlation between the two traits. In this study, we selected variants to compute the PGRS based on 11 p-value thresholds (<0.00001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.75, 1). Given the nature of the iCOGS array in which many loci have high densities of tagged SNPs, we performed linkage disequilibrium (LD) clumping in order to remove correlated variants (r^2>0.2) within 500kb windows for each component of the PGRS. The computations for PGRS and LD clumping were performed with PLINK [348]. Finally, we standardized each of the PGRS to have mean 0 and variance 1 and examined their associations with the various EOC subtypes through logistic regression, adjusted for the first 10 PCs.

Multiple testing correction

The polygenic risk prediction approach carries a high multiple testing burden, as does consideration of the various histologic groups and risk factors. However, given that we computed 11 PGRS for each trait based on sequential p-value thresholds, our statistics are not independent. In order to estimate the real number of independent hypotheses, we computed the correlation matrix of all the PGRS used in this study and fed this into a Matrix Spectral Decomposition (matSpD) algorithm [356], to estimate the number of independent variables. This algorithm provides an equivalent number of independent variables in a correlation matrix, by examining the ratio of the observed eigenvalue variance to its theoretical maximum. We estimated the number of independent PGRS to be 35 out of the 88 PGRS. As we examined these 35 independent PGRS in five separate EOC subtypes (high-grade serous, endometrioid, clear cell, mucinous and unknown), our significance threshold for the polygenic risk prediction analyses was 0.05/(35*5)=.00029.

Results

Genetic contribution of each chromosome and known loci

Fitting a GRM computed after removing known EOC-associated loci in univariate mixed-effect linear models implemented in GCTA [8, 306], we found that the known loci contributed about 40% of the total heritability of EOC and each of the subtypes **[Table 1]**. The estimated heritability of all EOC dropped from 5.6% to 3.6% once we removed known EOC-associated loci from the GRM. We observed a similar reduction of variance explained by the polygenic component for the EOC subtypes high-grade serous (8.8% to 4.7%), endometrioid (3.2% to 2.0%) and clear cell (6.7% to 4.6%) **[Table 1]**. Interestingly, in contrast to grade 1 and grade 2 (G1/G2) endometrioid where the heritability did not drop substantially (4.4% to 3.7%), grade 3 (G3) endometrioid h^2_g dropped from 4.9% to 0.9%. As shown previously [342], the heritability of mucinous cancer was not detectably different from 0. We were unable to perform any analyses for low-grade serous cancer given the small sample size (N_{cases}=350). We also had a set of cases with unknown EOC subtype classification; we expect that a high portion of these are individuals with undifferentiated high-grade serous, endometrioid or mixed serous EOC subtypes. For these, the heritability dropped from 7.0% to 4.1% after removing known loci.

In order to inspect the contributions of heritability per chromosome, we computed one GRM per chromosome, and fitted the multiple genetic variance components into linear mixed models as above. We found that the chromosomal contributions were not proportional to the number of independent SNPs in each of the chromosomes **[Figure 1]**. For example, the contribution of chromosomes 9, 11, 17 and 19 to high-grade serous EOC were larger than expected the 95% confidence interval (approximated through jackknifing 1000 times) did not overlap with 1. In contrast chromosomes 4, 10, 12, 14, 18 and 20 contributed less than expected.

Genetic correlation between EOC subtypes

We used the GRM as a random effect in a bivariate mixed-effects linear model implemented in GCTA to assess genetic heterogeneity across EOC histologic subtypes. **Table 2** summarizes the genetic correlations between the various EOC subtypes. We found significant genetic overlap between high-grade serous EOC and endometrioid EOC (r_g = 0.63 ± 0.27 ; *P*=.0029). Given that high-grade serous disease is not infrequently misclassified as endometrioid EOC [357], we also estimated the genetic correlations separating (G1/G2) endometrioid disease from (G3). Here we found that the genetic correlation between high-grade serous and G1/G2 endometrioid cancer was lower (r_g = 0.33 ± 0.23; P=.062) than between G3 endometrioid and high-grade serous cancer (r_g = 1.00 ± 0.83; P=.00078), suggesting that potential misclassification may have inflated the genetic correlation estimate when using all endometrioid EOC. Interestingly, we observed an appreciable but non-significant genetic overlap of about r_g = 0.5 between low-grade endometrioid and clear-cell EOC. We also found that the genetic correlations between “unknown/unclassified” EOC and high-grade serous and high-grade endometrioid disease were significant and essentially 1 (r_g = 1.0 ± 0.30; P=10^{-7} and r_g = 1.0 ± 0.96 *P*=.0049, respectively). The REML bivariate analyses involving Mucinous did not converge so did not yield any meaningful estimates. Further, removing known associated loci from the analyses affected the genetic correlation between endometrioid EOC (high and low grade) in a way that this was no longer significant [**Table 2]**.

Given that splitting the controls during the bivariate analyses to avoid sample overlap could have resulted in decreased power to detect genetic correlations; we complemented the genetic correlation analysis with the cross-trait LD score regression method, which is not biased by overlapping samples. In line with our results above, we found a statistically significant genetic correlation between high-grade serous EOC and endometrioid EOC (r_g =0.67 ± 0.25; *P*=7.4E-03), high-grade serous EOC and unknown EOC (r_g = 0.63 ± 0.25; *P*=.013) and endometrioid EOC and unknown EOC (r_g =1.00 ± 0.30; *P*=5.7E-04) **[Table 3]**.

Genetic overlap of EOC subtypes and associated environmental factors

In order to investigate the genetic overlap between all EOC and age at menarche, BMI, obesity, smoking, height and diabetes we used the cross-trait LD score regression method as well as a polygenic risk-prediction approach. We did not detect any significant genetic correlations using cross-trait LD score regression **[Table 4]**.

However, through the polygenic risk prediction approach, we found significant genetic overlap (at Bonferroni P-value threshold = .00029) of all EOC with obesity and with diabetes **[Table 5]**. The genetic overlap with diabetes appeared mainly in association with mucinous EOC. Overall, the directions of association are consistent with what has been reported in observational studies [170-173], although most of these associations are not significant.

Discussion

In this work, we have investigated the genetic architecture of EOC and its different subtypes. Our univariate analyses show an extent of hidden heritability inherent in the iCOGS array, with known associated loci accounting for about 40% of the total array heritability for most EOC histotypes, except for high-grade endometrioid, where they account for most of h_g^2. Is important to note that to reach these estimates we removed 2Mb per locus, which was done to ensure that no effect of these loci remained; however, this could also have inflated the estimates. We also showed that the hidden heritability is not spread proportionally across the chromosomes, with some contributing very little to the array heritability and others up to 5 times more than expected given their iCOGS SNP compositions. A limitation in our univariate experiments was that it was underpowered to compute meaningful estimates for low-grade serous and mucinous EOC. Although we had a bigger sample size for mucinous EOC than clear cell EOC, the analyses could have been affected by how each individual study deal with mucin-producing peritoneal tumours.

Using bivariate linear mixed-model and cross-trait LD score-regression approaches, we investigated genetic correlations between the various EOC subtypes. The bivariate linear mixed model provides unbiased estimates of genetic correlation and it requires individual genotype data in order to compute the GRM. Cross-trait LD score regression only requires summary results from the discovery set, and in contrast to the bivariate mixed-model approach, it allows sample overlap (in this case, overlapping controls) [83]. Whilst studies have shown shared germ-line risk mutations across the various EOC subtypes, these account for only a small fraction of general heritability [336-341]. We found a very high genetic correlation between high-grade serous EOC and poorly differentiated (G3, high-grade) endometrioid disease, and with unknown/unclassified EOC, which represents undifferentiated epithelial carcinoma. These correlations seem entirely reasonable, because high-grade endometrioid disease is sometimes misdiagnosed as high-grade serous, or may constitute a version of high-grade serous with slightly different differentiation. Undifferentiated ovarian carcinoma clinically resembles high-grade serous in response to treatment and in mortality. Low-grade serous, low-grade endometrioid and clear-cell carcinoma (which is relatively low grade) are heritability-distinct from the high-grade diseases and behave that way. Mucinous ovarian cancer seems to be a largely separate disease and has its own set of risk factors [327]. It does not appear to be related heritably to the other ovarian cancer histotypes.

We also considered whether the heritability of EOC and its subtypes could be explained (at least partly) via factors such as obesity, height, diabetes, smoking and age at menarche. As these factors have genetic components, it is plausible that the heritability of EOC could reflect the heritability of a causal factor. Using cross-trait LD score regression, we had insufficient power to detect genetic correlations, as this approach is greatly affected by small numbers of SNPs and by small sample sizes. However, through a polygenic risk prediction approach – which, although it does not directly quantify genetic overlap, is powerful for detecting genetic correlations between traits when the discovery and target sets are well powered [81], we found a significant positive genetic overlap between diabetes, obesity and all EOC. This genetic overlap appeared to be concentrated within mucinous disease and may not reflect other EOC histotypes. Genetic correlation in this analysis is estimated based on a large number of SNPs, so it is possible that the correlations seen between diabetes and obesity and EOC may be mediated by an upstream phenotype (e.g. hormonal changes). Genetic overlap analyses between EOC and the other risk factors did not reveal any other significant associations. Potential reasons for this include small sample sizes for some of the EOC subtypes, and incomplete mapping of relevant variants of the risk factors (i.e., variants in the iCOGS array explain only a limited amount of variance of the risk factors).

Is important to note that our results were derived from SNPs tagged in the iCOGS array. Hence the numbers of SNPs included in the analyses (195,183 SNPs) are smaller than in a typical GWAS array. Additional analyses could be performed on imputed genotypes from the iCOGS data; however, the iCOGS array is not designed to tag the whole genome, so imputation would likely still be limited to the existing tagged regions. Nevertheless, this array, which included several SNPs associated with other cancer types as well as with relevant quantitative traits such as BMI and the onset of menarche [339], allowed us to establish reasonably accurate estimates where the target sample sizes were well powered (e.g., high-grade serous, endometrioid, unknown/undifferentiated, and all EOC).

In summary, our results show that the major important EOC subtypes are genetically very homogeneous, and likely arise from a combination of known risk factors plus genetic contributions (beyond the known genetic predisposition mutations). This commonality highlights that high-grade disease could be considered a single clinical entity, with perhaps only minor variation between the serous, endometrioid and undifferentiated types. Low-grade histotypes, as well as mucinous ovarian cancer, likely represent more distinct pathologic variation. We also found that a great proportion of heritability is “missing”. Our analyses will be complemented once data of individuals genotyped in the OncoArray, which integrates a GWAS backbone, becomes available.

Acknowledgements

The Ovarian Cancer Association Consortium is supported by a grant from the Ovarian Cancer Research Fund thanks to donations by the family and friends of Kathryn Sladek Smith (PPD/RPCI.07). Funding of the constituent studies was provided by the California Cancer Research Program (00-01389V-20170, N01-CN25403, 2II0200); the Canadian Institutes of Health Research (MOP-86727); Cancer Australia; Cancer Council Victoria; Cancer Council Queensland; Cancer Council New South Wales; Cancer Council South Australia; Cancer Council Tasmania; Cancer Foundation of Western Australia; the Cancer Institute of New

Jersey; Cancer Research UK (C490/A6187, C490/A10119, C490/A10124); the Danish Cancer Society (94-222-52); the ELAN Program of the University of Erlangen-Nuremberg; the Eve Appeal; the Helsinki University Central Hospital Research Fund; Helse Vest; the Norwegian Cancer Society; the Norwegian Research Council; the Ovarian Cancer Research Fund; Nationaal Kankerplan of Belgium; the L & S Milken Foundation; the Polish Ministry of Science and Higher Education (4 PO5C 028 14, 2 PO5A 068 27); the Roswell Park Cancer Institute Alliance Foundation; the US National Cancer Institute (K07-CA095666, K07-CA80668, K07-CA143047, K22-CA138563, N01-CN55424, N01-PC67001, N01-PC067010, N01-PC035137, P01-CA017054, P01-CA087696, P30-CA072720, P30-CA15083, P30-CA008748, P50-CA159981, P50-CA105009, P50-CA136393, R01-CA149429, R01-CA014089, R01-CA016056, R01-CA017054, R01-CA049449, R01-CA050385, R01-CA054419, R01-CA058598, R01-CA058860, R01-CA061107, R01-CA061132, R01-CA063678, R01-CA063682, R01-CA067262, R01-CA071766, R01-CA074850, R01-CA080978, R01-CA083918, R01-CA087538, R01-CA092044, R01-CA095023, R01-CA122443, R01-CA112523, R01-CA114343, R01-CA126841, R01-CA136924, R03-CA113148, R03-CA115195, U01-CA069417, U01-CA071966, UM1-CA186107, UM1-CA176726 and Intramural research funds); the NIH/National Center for Research Resources/General Clinical Research Center (MO1-RR000056); the US Army Medical Research and Material Command (DAMD17-01-1-0729, DAMD17-02-1-0666, DAMD17-02-1-0669, W81XWH-07-0449, W81XWH-10-1-02802); the US Public Health Service (PSA-042205); the National Health and Medical Research Council of Australia (199600 and 400281); the German Federal Ministry of Education and Research of Germany Programme of Clinical Biomedical Research (01GB 9401); the State of Baden-Wurttemberg through Medical Faculty of the University of Ulm (P.685); the German Cancer Research Center; the Minnesota Ovarian Cancer Alliance; the Mayo Foundation; the Fred C. and Katherine B. Andersen Foundation; the Lon V. Smith Foundation (LVS-39420); the Oak Foundation; Eve Appeal; the OHSU Foundation; the Mermaid I project; the Rudolf-Bartling Foundation; the UK National Institute for Health Research Biomedical Research Centres at the University of Cambridge, Imperial College London, University College Hospital ‘Womens Health Theme’ and the Royal Marsden Hospital; and WorkSafeBC 14. Investigator-specific funding: G.C.P receives scholarship support from the University of Queensland and

QIMR Berghofer. Y.L. was supported by the NHMRC Early Career Fellowship. G.C.T. is supported by the National Health and Medical Research Council. S.M. was supported by an ARC Future Fellowship.

Table 7.1. Array heritabilities (h^2_g) and standard errors (s.e.) for invasive EOC according to histological subtype. Results for all iCOGS SNPs and after removing known associated loci. Disease prevalence of EOC subtypes is calculated as the lifetime risk of ovarian cancer multiplied by the relative proportion of the corresponding EOC subtype. See Methods section. Bolded estimates are statistically significantly different from 0.

Subtype	Cases	Controls	Life-time	All SNPs			Removing Known Loci*		
			risk	h^2_g	s.e.	P-value	h^2_g	s.e.	P-value
High-grade Serous	4098	21233	0.0055	**0.088**	**0.010**	**2.2E-16**	**0.047**	**0.009**	**1.83E-09**
Clear cell	620	21233	0.0005	**0.067**	**0.033**	**0.017**	**0.046**	**0.029**	**0.058**
Endometrioid (all)	1342	21233	0.001	**0.032**	**0.016**	**0.016**	**0.020**	**0.014**	**0.077**
Endometrioid G1/G2	906	21233	0.001	**0.044**	**0.024**	**0.025**	**0.037**	**0.021**	**0.037**
Endometrioid G3	436	21233	0.001	0.049	0.046	0.127	0.009	0.041	0.417
Mucinous	658	21233	0.0005	0.000	0.028	0.5	0.000	0.025	0.5
Unknown	2934	21233	0.009	**0.070**	**0.015**	**1.1E-10**	**0.041**	**0.012**	**1.1E-04**
All	10014	21233	0.009	**0.056**	**0.006**	**2.2E-16**	**0.036**	**0.005**	**2.2E-16**

*Loci removed: WNT4, RSPO1, SYNPO2, GPX6, ABO, ATAD5, C19orf62, CMYC, TIPARP, BNC2, ARHGAP27, TERT, RAD51B/C/D, BRIP1, BARD1, PALB2, NDN, CHMP4C, MLLT10, HNF1B, *BRCA1*, *BRCA2, KRAS, TP53, HER2, AR1D1A* and *PIK3CA.*

Table 7.2. Genetic correlations between major EOC subtypes as estimated from iCOGS array. Lower triangular matrix shows the genetic correlation using all the SNPs in the iCOGS array, while the upper triangular matrix shows the genetic correlation after removing known associated loci. For these calculations, each case was matched to one control in a way that none of the subtypes share any controls. Analyses for mucinous and low-grade serous EOC subtypes were underpowered to yield reliable estimates.

Subtype	**High-grade Serous**	**Endometrioid (all)**	**Endometrioid G1/G2**	**Endometrioid G3**	**Clear Cell**	**Unknown**
High-grade Serous	-	0.48 (0.35) P=0.072	0.24 (0.30) P=0.21	1.0 (2.66) P=0.5	0.29 (0.42) P=0.24	**1.0 (0.510)** **P=5.1E-04**
Endometrioid (all)	**0.63 (0.27)** **P=0.0029**	-	-	-	0.73 (0.64) P=0.088	0.50 (0.47) P=0.12
Endometrioid G1/G2	0.33 (0.23) P=0.062	-	-	0.36 (1.25) P=0.30*	0.42 (0.53) P=0.20	0.37 (0.41) P=0.18
Endometrioid G3	**1.0 (0.83)** **P=7.8E-04**	-	0.42 (0.56) P=0.2*	-	1.00 (1.68) P=0.5	1.0 (4.44) P=0.5
Clear Cell	0.28 (0.33) P=0.18	0.69 (0.56) P=0.074	0.52 (0.54) P=0.14	0.99 (0.87) P=0.073	-	0.09 **(**0.55**)** P=0.43
Unknown	**1.0 (0.30)** **P=1.0E-07**	**0.68 (0.33)** **P=0.0082**	0.42 (0.29) P=0.057	**1.0 (0.96)** **P=0.0049**	0.15 **(**0.39**)** P=3.5E-01	-

Bolded estimates are significantly different from 0.
* Significance (P-value) where the null hypothesis rG=1.

Table 7.3. Cross-trait LD score regression between EOC subtypes. Analyses for mucinous and low-grade serous EOC subtypes were underpowered to yield reliable estimates.

	HG Serous	Endometrioid	Endometrioid G1/G2	Endometrioid G3	Clear Cell	Unknown
HG Serous	-	0.82 (0.49) P=0.095	0.35 (0.41) P=0.41	1.0 (1.17) P=0.20	-	0.46 (0.46) P=0.31
Endometrioid	**0.67 (0.25)** **P=0.0074**	-	-	-	-	**1.0 (0.41)** **P=0.01**
Endometrioid G1/G2	0.35 (0.25) P=0.15	-	-	0.49 (0.70) P=0.47*	-	**0.85 (0.40)** **P=0.035**
Endometrioid G3	1.0 (0.79) P=0.15	-	0.53 (0.67) P=0.48*	-	-	1.0 (0.73) P=0.15
Clear Cell	0.53 (0.57) P=0.35	0.91 (0.80) P=0.26	0.71 (0.59) P=0.23	1.00 (1.06) P=0.29	-	-
Unknown	**0.63 (0.25)** **P=1.3E-02**	**1.0 (0.30)** **P=5.7E-04**	**0.77 (0.33)** **P=0.02**	1.00 (0.79) P=0.14	0.38 (0.53) P=0.47	-

Bolded estimates are significantly different from 0.

Table 7.4. Genetic correlation between risk factors and EOC subtypes using cross-trait LD score regression. Analyses for mucinous and low-grade serous EOC subtypes were underpowered to yield reliable estimates.

	All	HG Serous	Endometrioid	Clear Cell	Unknown
BMI	0.045 (0.07)	-0.04 (0.08)	0.18 (0.11)	-0.01 (0.16)	0.07 (0.08)
	P=0.52	P=0.63	P=0.10	P=0.96	P=0.38
Smoking	-0.34 (0.29)	-0.43 (0.33)	-0.37 (0.43)	-0.44 (0.66)	-0.17 (0.31)
	P=0.23	P=0.20	P=0.39	P=0.51	P=0.58
Height	0.081 (0.062)	0.13 (0.09)	0.03 (0.09)	0.24 (0.17)	0.00 (0.08)
	P=0.19	P=0.15	P=0.69	P=0.17	P=0.98
Menarche	-0.07 (0.08)	-0.23 (0.13)	-0.04 (0.12)	0.32 (0.36)	0.05 (0.09)
	P=0.38	P=0.06	P=0.75	P=0.36	P=0.59
Obesity*	0.05 (0.09)	-0.02 (0.09)	0.26 (0.17)	-0.18 (0.26)	0.12 (0.11)
>30 BMI	P=0.58	P=0.86	P=0.13	P=0.50	P=0.27
Obesity*	0.019 (0.087)	-0.03 (0.11)	0.02 (0.18)	-0.23 (0.37)	0.17 (0.12)
>35 BMI	P=0.83	P=0.80	P=0.90	P=0.54	P=0.17
Obesity*	-0.02 (0.15)	-0.02 (0.17)	-0.06 (0.30)	NA	0.03 (0.19)
>40 BMI	P=0.88	P=0.92	P=0.84		P=0.89
Diabetes	0.04 (0.12)	-0.04 (0.14)	0.04 (0.19)	-0.29 (0.38)	0.21 (0.14)
	P=0.75	P=0.74	P=0.84	P=0.45	P=0.15

*Reference group was individuals with BMI <=25

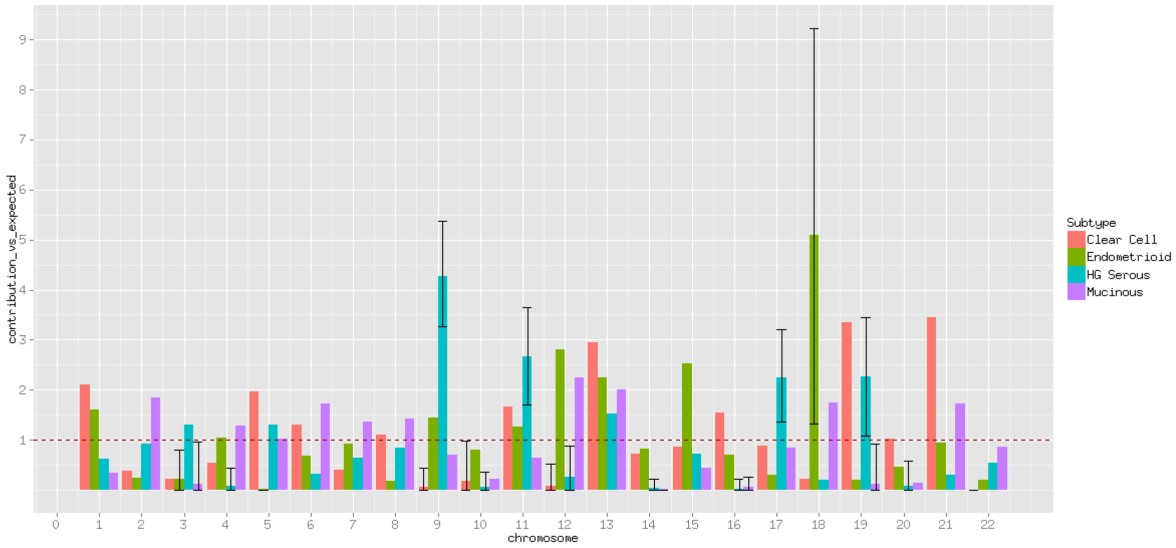
Table 7.5. Odds Ratios corresponding to 1 standard deviation increase in the PGRS and significance estimates (P-values) from the polygenic risk prediction approach between "environmental factors" PGRS and EOC subtypes. The displayed numbers correspond to the best association p-value out of the 11 different PGRS which were derived using different p-value thresholds. In this part we used the total set of controls with each of the EOC subtypes.

	HG Serous	Mucinous	Clear Cell	Endometrioid	Unknown	ALL
Menarche	0.99 (0.54)	1.09 (0.036)	1.05 (0.2)	1.04 (0.12)	1.04 (0.086)	1.02 (0.17)
BMI	1.04 (0.028)	1.05 (0.26)	1.06 (0.17)	1.07 (0.011)	1.04 (0.068)	1.04 (0.003)
Smoking	1.03 (0.11)	0.93 (0.067)	0.92 (0.049)	1.04 (0.18)	0.95 (0.0071)	0.97 (0.019)
Height	1.03 (0.14)	1.1 (0.015)	1.1 (0.025)	1.04 (0.17)	0.96 (0.06)	1.03 (0.022)
Diabetes	1.04 (0.021)	**1.18 (1.1e-05)**	1.08 (0.067)	1.07 (0.011)	1.04 (0.034)	**1.05 (4.1e-04)**
Obesity >30BMI	1.05 (0.0051)	1.06 (0.15)	1.06 (0.14)	1.04 (0.19)	1.04 (0.032)	**1.05 (2.6e-04)**
Obesity >35BMI	1.03 (0.08)	1.05 (0.21)	0.9 (0.012)	1.02 (0.42)	1.05 (0.028)	1.04 (0.0053)
Obesity >40BMI	1.03 (0.15)	1.06 (0.14)	0.87 (0.0015)	0.96 (0.13)	1.03 (0.19)	0.98 (0.21)

Bolded estimates are statistically significant (Bonferroni P-value threshold 2.9x10^{-4}).

*Reference group was individuals with BMI <=25

Figure 7.1. Contribution to the heritability by chromosome versus expected. Black vertical lines show the 95% confidence intervals approximated through jackknifing up to 1000 times. These are only shown for those instances that do not overlap with 1 to facilitate visualization.



Supplementary Material: Assessing the Genetic Architecture of Epithelial Ovarian Cancer Histological Subtypes.

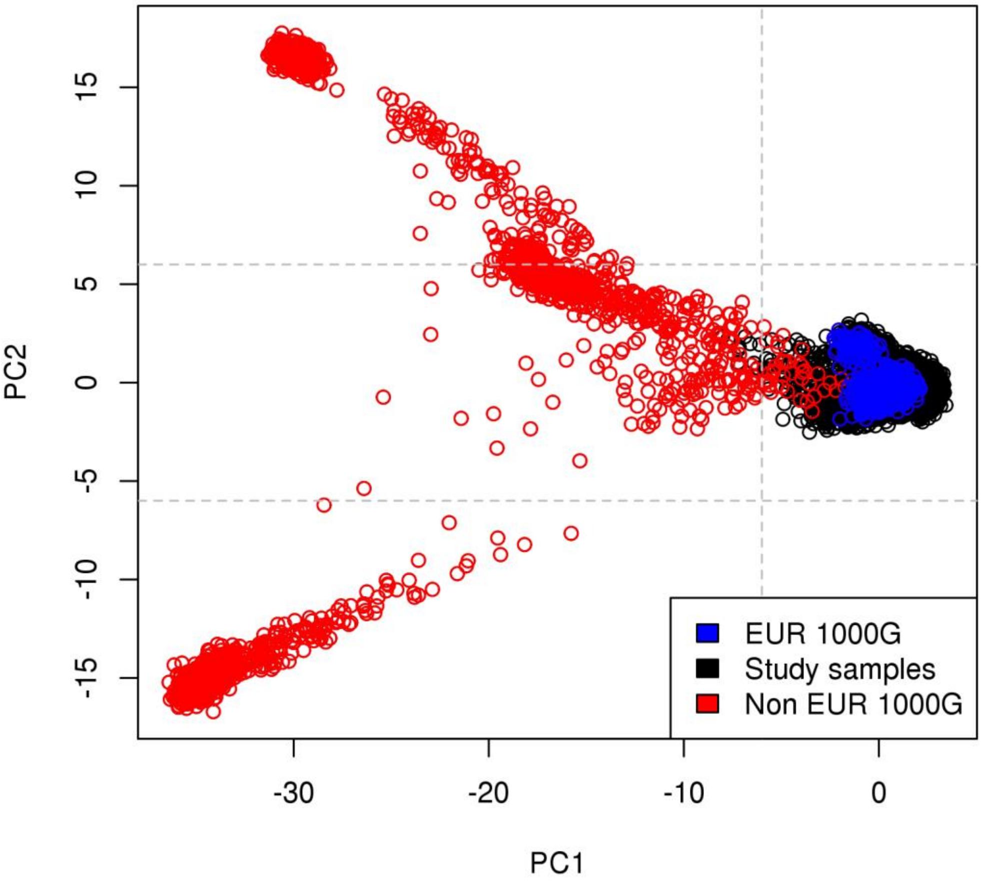


Figure 7.2. Supplementary Figure 1. Genotype principal component analysis of OCAC samples and 1000 Genomes. X and Y axes display the number of standard deviations from 1000 Genomes EUR populations. Dotted lines enclose the samples used in this study.

Table 7.6. Supplementary Table 1. Description of individual OCAC studies and case-control sample size. *Numbers differ from Table 1 in main manuscript, as these ones reflect the total number before Identity by descent (IBD) <0.10 filtering.

Study Name	Country	Code	Controls	Serous	Mucinous	Endometrioid	Clear Cell	HG Serous	Other	All invasive
Australian Cancer Study	Australia	ACS	175	104	7	22	9	89	32	166
Australian Ovarian Cancer Sutidy	Australia	AOC	802	448	35	84	43	409	118	714
Bavarian Ovarian Cancer Cases and Controls	Germany	BAV	142	56	8	13	6	42	10	93
Belgium Ovarian Cancer Study	Belgium	BEL	1348	194	23	22	23	182	17	274
Diseases of the Ovary and their Evaluation	USA	DOV	1119	293	18	84	29	235	136	515
Diseases of the Ovary and their Evaluation	USA	DVE	368	233	8	64	36	200	78	389
Germany Ovarian Cancer Study	Germany	GER	413	95	21	21	6	68	59	189
Hawaii Ovarian Cancer Study	USA	HAW	156	38	3	12	5	36	2	60
Hannover-Jena Ovarian Cancer Study	Germany	HJO	273	140	9	26	4	107	116	266
Hannover-Minsk Ovarian Cancer Study	Germany	HMO	138	50	7	12	1	1	121	142
Helsinki Ovarian Cancer Study	Finland	HOC	447	113	45	28	13	0	135	221
Hormones and Ovarian Cancer Prediction	USA	HOP	1464	377	30	84	42	333	145	654
Danish Malignant Ovarian Tumor Study	Denmark	MAL	828	272	42	54	33	183	53	440
Mayo Clinic Ovarian Cancer Case Control Study	USA	MAY	10	9	0	1	0	9	0	10
Melbourne Collaborative Cohort Study	Australia	MCC	65	34	7	7	6	19	21	63
MD Anderson Ovarian Cancer Study	USA	MDA	384	190	27	28	4	135	179	373
Memorial Sloan Kettering Cancer Center	USA	MSK	593	382	0	20	18	343	73	467
North Carolina Ovarian Cancer Study	USA	NCO	172	147	18	35	24	132	50	269
New England Case-Control Study	USA	NEC	979	371	41	140	33	331	60	634

Nurses' Health Study I and II	USA	NHS	425	68	7	14	6	0	100	127
New Jersey Ovarian Cancer Study	USA	NJO	180	100	7	27	20	80	27	169
University of Bergen, Haukeland University Hospital, Norway	Norway	NOR	370	135	15	27	11	85	87	237
Nijmegen Ovarian Cancer Study	Netherlands	NTH	323	116	33	64	20	64	52	255
Ovarian Cancer in Alberta and British Columbia	Canada	OVA	748	344	26	103	57	0	445	631
Polish Ovarian Cancer Study	Poland	POC	417	199	33	39	9	0	341	422
Polish Ovarian cancer Case Control Study (NCI)	Poland	POL	186	21	4	10	2	15	11	42
UK Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH) Ovarian Cancer Study	UK	SEA	1196	162	38	24	28	104	71	271
UK Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH) Ovarian Cancer Study	UK	SEB	4826	11	5	5	0	4	15	29
Southampton Ovarian Cancer Study	UK	SOC	0	102	33	62	11	72	79	267
Family Registry for Ovarian Cancer AND Genetic Epidemiology of Ovarian Cancer	USA	STA	313	154	16	32	20	135	35	251
Familial Ovarian Tumor Study	Canada	TOR	74	8	1	7	2	0	11	21
UC Irvine Ovarian Cancer Study	USA	UCI	367	166	19	48	23	143	32	277
UK Ovarian Cancer Population Study	UK	UKO	1103	117	24	32	25	93	51	236
Los Angeles County Case-Control Studies of Ovarian Cancer	USA	USC	1047	447	44	79	35	341	161	689
Warsaw Ovarian Cancer Study	Poland	WOC	203	132	8	20	17	131	25	202
Total*			21654	5828	662	1350	621	4121	2948	10065

Chapter 8

Discussion

8.1 General discussion

The series of studies that comprise this thesis are examples of how examination of germ-line variation can produce important insights into the biology of complex traits and diseases. For each specific study presented, the summary of the findings, their meaning, study limitation and future directions were discussed in their respective chapters. Therefore, I take the space of this section to discuss the strengths, general limitations and future directions of the approaches illustrated in this thesis.

8.2 Gene mapping studies

The first GWAS of age-related macular degeneration in 2005 marked the first milestone in the GWAS era by successfully identifying two variants using only 96 cases and 50 healthy controls [358]. Since then, progress in this area has been marked by advances in sequencing and genotyping technologies leading to faster, bigger and cheaper acquisition of genotype data. To date some of the biggest GWAS such as those for anthropometric traits (e.g. height and BMI) carried out by the GIANT consortium have included hundreds of thousands of samples, thus securing very precise estimates for common variants. These big studies, although well powered for lower frequency variants, still focus only in common variation given that they are tied to a genotype imputation procedure using (generally) 1000 Genomes Project reference panels.

The exome-chip association study of CCT described in chapter 2 served as an example of one of the limitations of the current imputation reference panels. In this project, we showed that an exonic variant in the gene *WNT10A* with a MAF of around 3% effectively accounted for all the variance previously seen in *USP37* in a big CCT GWAS meta-analysis. Although our analyses couldn't definitely prove that the missense variant in *WNT10A* is the causal variant, it is interesting to note that we were able to detect this variant using significantly less samples than the big CCT GWAS meta-analysis carried out by the IGGC [105]. One of the reasons for this, beside chance or the samples used, could be that the variant tagged is or is closer to the causal variant and that imputation for it is difficult. This points out that in addition to increase sample sizes in GWAS studies, it is imperative to improve and widen the imputation reference panels. In this direction, very recently, the haplotype reference consortium (HRC) has made available its first release which includes close to 65,000 haplotypes and around 40 million SNPs, all with a minor allele count of at least 5 (MAF > ~0.01%). This new reference panel guarantees an improved and broader imputation, particularly in the lower end of the MAF spectrum.

As to whether imputation using these large reference panels will be comparable (or better) to the use of current rare variants rare variants genotyping arrays (e.g. Illumina Human-Exome) is yet to be seen. As mentioned earlier in the introduction there have not been many successful studies using the Human-Exome array, possibly because (contrary from what was believed) these variants have modest-to-weak effect sizes or the array-tagged variants are not good proxies of causal variants. The Illumina Human-Exome array was designed to tag rare variants believed to have important effects (e.g. missense and nonsense mutations); however the overwhelming majority of variants in the genome are rare and numerous studies have shown that variants in regulatory regions are the major contributors to the heritability [359-361]. The latter suggests that using genotyping arrays to capture rare variation may not be cost effective as the arrays would require tagging many more variants than what they currently do (e.g. Human-Exome includes ~250,000 SNPs, with each rare variant typically only tagging a relatively small number of nearby SNPs). This reality argues that it may be better to stick to imputation when the study aim is to scan rare variants genome-wide. The latest imputation tests using the HRC and the UK10k reference panel report accurate imputation of variants with MAF >0.1% [362, 363]. Given current sample sizes, most GWAS would be underpowered to detect variants below that MAF anyway. In studies where the goal is inspecting rarer variants, different experimental designs may be useful such as selecting extreme phenotypes for sequencing.

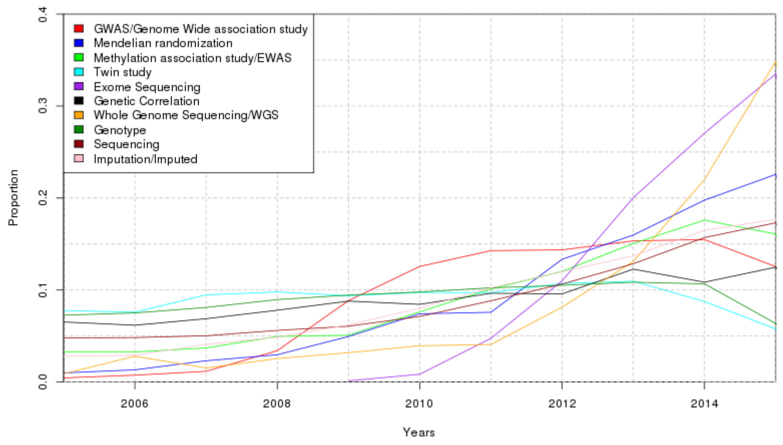
With the advent of studies with massive sample sizes such as the forthcoming ones using the UK biobank data resource with up to 500,000 genotyped (currently

150,000) and phenotyped individuals, those from large consortiums such as GIANT or those performed by direct to consumer genetic testing company 23andMe with up to a million genotyped individuals with diverse surveyed phenotypes, an interesting question is “when should GWAS stop?”. It is clear that we will reach a point where we are able to discern between not associated variants and variants with very tiny effects for some traits (certainly, this may not be true for rare diseases). Whether the variants with tiny effects would be somehow meaningful depends on the application. One possible answer is that we should stop doing GWAS once we stop detecting new pathways involved in the phenotype, as I heard said by Professor Matthew Brown during a seminar when this same question was raised. Although I agree with this, I believe that it just applies if the goal is to understand the biology of the trait (also, probably the catalogue of pathways to date is not complete). However, there are other applications of GWAS besides understanding the biology. For example, having accurate estimates of many variants with tiny effects is likely to aid the prediction of phenotypic values (e.g. through the computation of allelic scores). Nevertheless, in this direction, a limitation that should be acknowledged is that in contrast to Mendelian conditions, there will always be some degree of uncertainty when predicting complex traits, particularly in diseases with low heritability (e.g. epithelial ovarian cancer). In addition, although predicting differences in prevalence of a complex disease at a population level can greatly benefit from big GWAS of common variants, at the individual level looking for rare variants with high penetrance such as mutations in *MYOC* for glaucoma or BRCA1/2 for breast cancer so far has higher relevance for clinical applications. Another application of big GWAS is its use as a means to estimate genetic correlations through (for example) cross-trait LD score regression or the possibility to assess causal relations through MR.

8.2 Mendelian randomization studies

Although it is undisputable that GWAS discoveries have produced invaluable insights into the biology of many complex traits and diseases, translation of GWAS findings is happening slowly. To date performing a PubMed query of “GWAS” retrieve about 22700 different studies that have either performed GWAS or benefitted from the insights these have provided. Until the end of 2014, the GWAS catalogue [21] reports 14844 variants associated to 610 different traits. Although these numbers are likely to keep increasing, is interesting to note that we might have already passed the peak of GWAS studies. Investigating trends of studies in a very crude way, such as partitioning the number of publications with the term “GWAS or Genome-wide association study” by year, we can see that the proportion of studies has not increased meaningfully in the last 5 years [Figure 1]. In contrast, the term “Mendelian Randomization” has experience a steady increase in the same period. The latter is expected as Mendelian randomization may be one of the most simple and effective ways to reap what GWAS has sown in the last decade. Doing a quick inspection in the GWAS catalogue, there are 74 traits with the word “levels” on them (e.g. calcium levels, estradiol levels, folate levels, etc.) and presumably with at least one associated variant at the genome-wide significance level. Assuming that these variants are robust to the MR assumptions, these potentially could be used to assess the causality of these factors on many traits and diseases in studies where sample size allows.

Figure 8.1. Trends of research topics in the last 10 years approximated by PubMed queries.



Although MR can be vastly applied to many traits, so far, approximately only 607 studies with the term are reported in the literature. As detailed in the introduction chapter, MR studies are hindered by many requirements such as big samples and hard to test (at least completely) assumptions, which make these kind of studies challenging to carry out. In chapters 4 and 5 of this thesis, I presented MR studies to test the causal relationship of education and vitamin D levels on myopic refractive error. The MR of vitamin D levels detailed in chapter 4 made use of GWAS summary statistics of refractive error from the CREAM consortium. The greatest limitation of this study was the impossibility to test the MR assumptions in the complete CREAM sample, thus these tests were performed in a smaller sample. In this study, I found no support for the hypothesis that vitamin D levels have a causal relationship with refractive error. Although a negative result (no support of a causal relationship) is also subject to all the potential biases arising from violations in the assumption as positive results, the biases were probably very small; any real biases would have had to align perfectly in a way to move the effect estimate towards 0 (if there is really a true effect). The null association and narrow confidence intervals obtained in our study using SNPs with precise effects in vitamin D levels arguably provide robust evidence of no or very small effect of vitamin D on refractive error.

The MR of education level on refractive error required to build an allelic score using the top 10% independent variants in order to increase the variance explained of education level, with the goal of increasing power for the study. The latter increased the probability of the instrument being related to the confounders or directly associated to refractive error, thus violating the MR assumptions. Although in this study we assessed the MR assumptions to the extent possible by examining the association of the allele score with potential confounders, namely smoking, BMI and height, the estimated causal effect appeared to be higher than the observed. After conditioning on education level, the allelic score remained significantly associated to refractive error thus raising questions on whether it was operating through education or other pathways. However, is worth noting that the mediator being evaluated was education level measured as degree attained or years of education. These measures do not capture all aspects of education (e.g. does not give information on actual amount of time spent in school, reading, etc.) that may be affecting refractive error hence the results presented in chapter 5 are still consistent with the possibility that education if taken as a whole mediates the association of the allelic score.

The above is an example of how potentially inaccurate measurements of the exposure of interest can hamper the interpretation of MR studies. This indicates that, in addition to assess the fundamental assumptions for MR, it is important to examine whether the exposure used accurately captures the phenotype that may mediate the association between the genetic variant(s) and the outcome. A similar problem can arise when the exposure varies with time (e.g. carrying out an MR using estrogen levels in elderly women as exposure, or growth hormone in adults), making it difficult for the genotype-exposure relationship to be captured entirely.

The authors in [265] suggest that whenever the exposure measurement is unsuitable for an MR study, it may be better to just evaluate the association between the genetic variant and the outcome as once proposed in a letter by Katan [364]. Not pursuing defining the magnitude of effect by including the “faulty” exposure variable may defeat part of the purpose of MR; however, in order to obtain valid estimates, the MR assumptions must hold for the measured exposure available which many times is not completely characterized, like education level in our study.

8.2 Genetic correlation studies

In the forthcoming years, MR studies will keep producing valuable information about disease aetiology with good prospects of direct translation into the clinic. However, its reach has to be acknowledged. There are countless of modifiable traits for which suitable instruments are hard to find, and diseases which prevalence is so low that gathering enough data to perform a powerful MR study may not be possible. In these cases a possible way forward is to estimate the genetic correlation between traits.

In order for two traits to present genetic correlation the direction of SNP effects must be consistently aligned. Multiple mechanisms can give rise to genetic correlations. For example, the existence of a causal relationship between the traits. Although a genetic correlation is not *per se* prove of causality, the presence of it can certainly give some support to such a relation. Assuming a causal relationship between a heritable exposure and outcome, part of the heritability of the outcome will be mediated through the exposure. However, without prior knowledge, this value does not tell anything about direction or magnitude of effect, thus in many cases the conservative interpretation is that the same SNPs consistently affect in the same (positive genetic correlation) or opposite (negative) direction both traits. Another mechanism is that common genes control diverse pathways and that these have consistent effects in both traits. In some extreme instances, a genetic correlation could be present through a parent-offspring relation. For example, a genetic correlation seen between cognition and height (beside metabolic processes and nutrients intake mechanisms) can represent the situation where parents with a higher IQ have a better income and thus are able to provide more nutrients to their children, translating into them being taller. Given that parents and offspring share half of their genotype, estimating the genetic correlation between height and IQ will be confounded.

The common theme running through chapters 6 and 7 was the use of approaches to estimate the genetic correlation in diverse diseases. In particular, in chapter 6 we interrogate aspects of the genetic architecture of two age-related eye diseases, namely age-related macular degeneration (AMD) and primary open angle glaucoma (POAG). Our results suggest that variation in these diseases is partly underpinned by shared genetic factors. Although there is no comorbidity reported between these two diseases and their pathophysiology is different, there are diverse mechanisms that could be driving the observed genetic overlap. For example, diverse studies have reported that oxidative stress, inflammation and mitochondrial dysfunction are risk factors for both diseases [365-370]. These mechanisms are partly driven by genetics so it is possible to think that the observed genetic correlation can be mediated by them. Another interesting result in this study was the one supporting genetic differences between female and male POAG. Research has shown evidence that estrogen has protective effects against POAG potentially by promoting higher production of collagen fibers that increase flexibility of the eye, thus reducing IOP [164-166]. This could explain our findings given that genetic variation in the estrogen pathways may have a higher impact in women than in men.

In chapter 7 I made use of the diverse approaches available to estimate the genetic correlation between epithelial ovarian cancer, its subtypes and potential risk factors (associations from observational studies). The aim of this study was to add support (or not) for causality of some of the risk factors and to examine whether the different EOC subtypes were or not genetically homogeneous. For the EOC subtypes where power was enough to detect meaningful genetic correlations (high-grade serous, endometrioid and undifferentiated) we found that the subtypes were genetically highly homogenous. The estimation of genetic correlations between EOC and potential risk factors was consistent with causal roles for type 2 diabetes and obesity. Again, although the latter is not definitive, research shows that people with diabetes or obesity are more prone to develop EOC (among other types of cancer) but not the other way around, so reverse causality is improbable. However, we still cannot rule out the possibility of an upstream pathway affecting these conditions in the same direction.

So overall, these studies serve as good examples of how estimating the genetic correlation can lead to the development of hypotheses about the risk factors involved in the development of disease. I believe that one of the most important strengths of genetic correlation analyses is that (depending on the approach used) they do not require both phenotypes to be measured in the same sample. In the last years many studies have made publicly available their GWAS summary statistics, allowing us to check correlation of all these with our trait or disease of interest without the need to measure them. Arguably, these correlations are often more meaningful than direct phenotypic correlations as they are susceptible to less confounders. Increasing the number of publicly available GWAS summary results may allow us at some point to connect the dots of ‘what triggers what’ and causes disease. Among the interesting insights provided in these two chapters, I found the significant genetic differences between POAG genders an important one. Not much because of the finding, as is been long since the association between estrogen and IOP/POAG is believed to mediate the differences of prevalence between sexes, but because it suggests that would be a good idea to perform this kind of analysis for other traits and diseases where it is believe that hormones play a role.

To conclude, I have investigated aspects of different traits and diseases using genotype data and cutting edge statistical genetics approaches. Although the insights gained during the make of this thesis are far from translation into the clinic, they are an important contribution to the literature by adding or removing support to current hypotheses for the multiple traits and diseases investigated. The studies reported here also serve as background for future research featuring mapping, correlation and causation in these phenotypes.

Bibliography or List of References

1. Barreiro, L.B., et al., *Natural selection has driven population differentiation in modern humans.* Nat Genet, 2008. **40**(3): p. 340-5.
2. Feuk, L., A.R. Carson, and S.W. Scherer, *Structural variation in the human genome.* Nat Rev Genet, 2006. **7**(2): p. 85-97.
3. Nachman, M.W., *Single nucleotide polymorphisms and recombination rate in humans.* Trends Genet, 2001. **17**(9): p. 481-5.
4. Lynch, M. and B. Walsh, *Genetics and Analysis of Quantitative Traits.* 1998. **1**: p. 4.
5. Palmer, L.J., P.R. Burton, and G.D. Smith, *An introduction to genetic epidemiology.* 2011.
6. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era--concepts and misconceptions.* Nat Rev Genet, 2008. **9**(4): p. 255-66.
7. Lee, S.H., et al., *Estimating missing heritability for disease from genome-wide association studies.* Am J Hum Genet, 2011. **88**(3): p. 294-305.
8. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nat Genet, 2010. **42**(7): p. 565-9.
9. Zhou, X. and M. Stephens, *Genome-wide efficient mixed-model analysis for association studies.* Nat Genet, 2012. **44**(7): p. 821-4.
10. Merril, R.M., *An Introduction to Epidemiology.* 2012.
11. Moher, D., et al., *CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials.* BMJ, 2010. **340**: p. c869.
12. Montana, G., *Statistical methods in genetics.* Birefing in Bioinformatics, 2006.
13. Ewens, W.J., M. Li, and R.S. Spielman, *A review of family-based tests for linkage disequilibrium between a quantitative trait and a genetic marker.* PLoS Genet, 2008. **4**(9): p. e1000180.
14. Hunter, N., *Meiotic Recombination: The Essence of Heredity.* Cold Spring Harb Perspect Biol, 2015.
15. International HapMap, C., *A haplotype map of the human genome.* Nature, 2005. **437**(7063): p. 1299-320.
16. Slatkin, M., *Linkage disequilibrium--understanding the evolutionary past and mapping the medical future.* Nat Rev Genet, 2008. **9**(6): p. 477-85.
17. Bush, W.S. and J.H. Moore, *Chapter 11: Genome-wide association studies.* PLoS Comput Biol, 2012. **8**(12): p. e1002822.
18. Panagiotou, O.A., J.P. Ioannidis, and P. Genome-Wide Significance, *What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations.* Int J Epidemiol, 2012. **41**(1): p. 273-86.
19. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genet Epidemiol, 2008. **32**(4): p. 381-5.
20. Barsh, G.S., et al., *Guidelines for genome-wide association studies.* PLoS Genet, 2012. **8**(7): p. e1002812.
21. Welter, D., et al., *The NHGRI GWAS Catalog, a curated resource of SNP-trait associations.* Nucleic Acids Res, 2014. **42**(Database issue): p. D1001-6.

22. Pritchard, J.K. and N.A. Rosenberg, *Use of unlinked genetic markers to detect population stratification in association studies.* Am J Hum Genet, 1999. **65**(1): p. 220-8.
23. Price, A.L., et al., *New approaches to population stratification in genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 459-63.
24. Freedman, M.L., et al., *Assessing the impact of population stratification on genetic association studies.* Nat Genet, 2004. **36**(4): p. 388-93.
25. Price, A.L., et al., *The impact of divergence time on the nature of population structure: an example from Iceland.* PLoS Genet, 2009. **5**(6): p. e1000505.
26. Widmer, C., et al., *Further improvements to linear mixed models for genome-wide association studies.* Sci Rep, 2014. **4**: p. 6874.
27. Bulik-Sullivan, B.K. and P.F. Sullivan, *The authorship network of genome-wide association studies.* Nat Genet, 2012. **44**(2): p. 113.
28. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
29. Kang, G., B. Jiang, and Y. Cui, *Gene-based Genomewide Association Analysis: A Comparison Study.* Curr Genomics, 2013. **14**(4): p. 250-5.
30. Liu, J.Z., et al., *A versatile gene-based test for genome-wide association studies.* Am J Hum Genet, 2010. **87**(1): p. 139-45.
31. Gauderman, W.J., et al., *Testing association between disease and multiple SNPs in a candidate gene.* Genet Epidemiol, 2007. **31**(5): p. 383-95.
32. Wang, K. and D. Abbott, *A principal components regression approach to multilocus genetic association studies.* Genet Epidemiol, 2008. **32**(2): p. 108-18.
33. Mishra, A. and S. Macgregor, *VEGAS2: Software for More Flexible Gene-Based Testing.* Twin Res Hum Genet, 2015. **18**(1): p. 86-91.
34. Li, M.X., et al., *GATES: a rapid and powerful gene-based association test using extended Simes procedure.* Am J Hum Genet, 2011. **88**(3): p. 283-93.
35. Wang, K., M. Li, and M. Bucan, *Pathway-based approaches for analysis of genomewide association studies.* Am J Hum Genet, 2007. **81**(6): p. 1278-83.
36. Wang, K., M. Li, and H. Hakonarson, *Analysing biological pathways in genome-wide association studies.* Nat Rev Genet, 2010. **11**(12): p. 843-54.
37. Segre, A.V., et al., *Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits.* PLoS Genet, 2010. **6**(8).
38. Holmans, P., et al., *Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder.* Am J Hum Genet, 2009. **85**(1): p. 13-24.
39. Lee, P.H., et al., *INRICH: interval-based enrichment analysis for genome-wide association studies.* Bioinformatics, 2012. **28**(13): p. 1797-9.
40. Pers, T.H., et al., *Biological interpretation of genome-wide association studies using predicted gene functions.* Nat Commun, 2015. **6**: p. 5890.
41. Manolio, T.A., et al., *Finding the missing heritability of complex diseases.* Nature, 2009. **461**(7265): p. 747-53.
42. Zuk, O., et al., *The mystery of missing heritability: Genetic interactions create phantom heritability.* Proc Natl Acad Sci U S A, 2012. **109**(4): p. 1193-8.
43. Metzker, M.L., *Sequencing technologies - the next generation.* Nat Rev Genet, 2010. **11**(1): p. 31-46.
44. Genomes Project, C., et al., *An integrated map of genetic variation from 1,092 human genomes.* Nature, 2012. **491**(7422): p. 56-65.

45. Yang, J., et al., *Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index.* Nat Genet, 2015. **47**(10): p. 1114-20.
46. Pritchard, J.K., *Are rare variants responsible for susceptibility to complex diseases?* Am J Hum Genet, 2001. **69**(1): p. 124-37.
47. Huyghe, J.R., et al., *Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion.* Nat Genet, 2013. **45**(2): p. 197-201.
48. Peloso, G.M., et al., *Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks.* Am J Hum Genet, 2014. **94**(2): p. 223-32.
49. Auer, P.L., et al., *Rare and low-frequency coding variants in CXCR2 and other genes are associated with hematological traits.* Nat Genet, 2014. **46**(6): p. 629-34.
50. Kozlitina, J., et al., *Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease.* Nat Genet, 2014. **46**(4): p. 352-6.
51. Gorlov, I.P., et al., *Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms.* Am J Hum Genet, 2008. **82**(1): p. 100-12.
52. Qiao, D., M. Mattheisen, and C. Lange, *On association analysis of rare variants under population substructure: an approach for the detection of subjects that can cause bias in the analysis--T opt: an outlier detection method.* Genet Epidemiol, 2013. **37**(5): p. 431-9.
53. Goldstein, J.I., et al., *zCall: a rare variant caller for array-based genotyping: genetics and population analysis.* Bioinformatics, 2012. **28**(19): p. 2543-5.
54. Burkett, K. and C. Greenwood, *A sequence of methodological changes due to sequencing.* Curr Opin Allergy Clin Immunol, 2013. **13**(5): p. 470-7.
55. Price, A.L., et al., *Principal components analysis corrects for stratification in genome-wide association studies.* Nat Genet, 2006. **38**(8): p. 904-9.
56. Patterson, N., A.L. Price, and D. Reich, *Population structure and eigenanalysis.* PLoS Genet, 2006. **2**(12): p. e190.
57. Price, A.L., et al., *Pooled association tests for rare variants in exon-resequencing studies.* Am J Hum Genet, 2010. **86**(6): p. 832-8.
58. Epstein, M.P., A.S. Allen, and G.A. Satten, *A simple and improved correction for population stratification in case-control studies.* Am J Hum Genet, 2007. **80**(5): p. 921-30.
59. Epstein, M.P., et al., *A permutation procedure to correct for confounders in case-control studies, including tests of rare variation.* Am J Hum Genet, 2012. **91**(2): p. 215-23.
60. Morris, A.P. and E. Zeggini, *An evaluation of statistical approaches to rare variant analysis in genetic association studies.* Genet Epidemiol, 2010. **34**(2): p. 188-93.
61. Madsen, B.E. and S.R. Browning, *A groupwise association test for rare mutations using a weighted sum statistic.* PLoS Genet, 2009. **5**(2): p. e1000384.
62. Zawistowski, M., et al., *Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes.* Am J Hum Genet, 2010. **87**(5): p. 604-17.
63. Basu, S. and W. Pan, *Comparison of statistical tests for disease association with rare variants.* Genet Epidemiol, 2011. **35**(7): p. 606-19.

64. Li, B. and S.M. Leal, *Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.* Am J Hum Genet, 2008. **83**(3): p. 311-21.

65. Wu, M.C., et al., *Rare-variant association testing for sequencing data with the sequence kernel association test.* Am J Hum Genet, 2011. **89**(1): p. 82-93.

66. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* Am J Hum Genet, 2012. **91**(2): p. 224-37.

67. Sun, J., Y. Zheng, and L. Hsu, *A unified mixed-effects model for rare-variant association in sequencing studies.* Genet Epidemiol, 2013. **37**(4): p. 334-44.

68. Derkach, A., J.F. Lawless, and L. Sun, *Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests.* Genet Epidemiol, 2013. **37**(1): p. 110-21.

69. Neale, B.M., et al., *Testing for an unusual distribution of rare variants.* PLoS Genet, 2011. **7**(3): p. e1001322.

70. Lin, D.Y. and Z.Z. Tang, *A general framework for detecting disease associations with rare variants in sequencing studies.* Am J Hum Genet, 2011. **89**(3): p. 354-67.

71. Liu, D.J. and S.M. Leal, *A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions.* PLoS Genet, 2010. **6**(10): p. e1001156.

72. !!! INVALID CITATION !!!

73. Fang, Y.H. and Y.F. Chiu, *A novel support vector machine-based approach for rare variant detection.* PLoS One, 2013. **8**(8): p. e71114.

74. Britannica, E., *Albinism.* Encyclopædia Britannica, 2015.

75. Haqq, C.M., et al., *Molecular basis of mammalian sexual determination: activation of Mullerian inhibiting substance gene expression by SRY.* Science, 1994. **266**(5190): p. 1494-500.

76. Stearns, F.W., *One hundred years of pleiotropy: a retrospective.* Genetics, 2010. **186**(3): p. 767-73.

77. Emig, D., et al., *Drug target prediction and repositioning using an integrated network-based approach.* PLoS One, 2013. **8**(4): p. e60618.

78. Silventoinen, K., et al., *Heritability of adult body height: a comparative study of twin cohorts in eight countries.* Twin Res, 2003. **6**(5): p. 399-408.

79. Lee, S.H., et al., *Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood.* Bioinformatics, 2012. **28**(19): p. 2540-2.

80. Wray, N.R., et al., *Research review: Polygenic methods and their application to psychiatric traits.* J Child Psychol Psychiatry, 2014. **55**(10): p. 1068-87.

81. Dudbridge, F., *Power and predictive accuracy of polygenic risk scores.* PLoS Genet, 2013. **9**(3): p. e1003348.

82. Bulik-Sullivan, B.K., et al., *LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.* Nat Genet, 2015. **47**(3): p. 291-5.

83. Bulik-Sullivan, B., et al., *An atlas of genetic correlations across human diseases and traits.* Nat Genet, 2015.

84. Ebrahim., G.D.S.a.S., *Mendelian Randomization: Genetic Variants as Instruments for Strengthening Causal Inference in Observational Studies.* National Academies Press (US), 2008.

85. Smith, G.D. and S. Ebrahim, *'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?* Int J Epidemiol, 2003. **32**(1): p. 1-22.
86. Didelez, V. and N. Sheehan, *Mendelian randomization as an instrumental variable approach to causal inference.* Stat Methods Med Res, 2007. **16**(4): p. 309-30.
87. Burgess, S., S.G. Thompson, and C.C.G. Collaboration, *Avoiding bias from weak instruments in Mendelian randomization studies.* Int J Epidemiol, 2011. **40**(3): p. 755-64.
88. Palmer, T.M., et al., *Using multiple genetic variants as instrumental variables for modifiable risk factors.* Stat Methods Med Res, 2012. **21**(3): p. 223-42.
89. Evans, D.M., et al., *Mining the human phenome using allelic scores that index biological intermediates.* PLoS Genet, 2013. **9**(10): p. e1003919.
90. Palmer, T.M., et al., *Instrumental variable estimation of causal risk ratios and causal odds ratios in Mendelian randomization analyses.* Am J Epidemiol, 2011. **173**(12): p. 1392-403.
91. Thomas, D.C., D.A. Lawlor, and J.R. Thompson, *Re: Estimation of bias in nongenetic observational studies using "Mendelian triangulation" by Bautista et al.* Ann Epidemiol, 2007. **17**(7): p. 511-3.
92. Burgess, S., A. Butterworth, and S.G. Thompson, *Mendelian randomization analysis with multiple genetic variants using summarized data.* Genet Epidemiol, 2013. **37**(7): p. 658-65.
93. Lawlor, D.A., et al., *Mendelian randomization: using genes as instruments for making causal inferences in epidemiology.* Stat Med, 2008. **27**(8): p. 1133-63.
94. Terza, J.V., A. Basu, and P.J. Rathouz, *Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling.* J Health Econ, 2008. **27**(3): p. 531-43.
95. Goetghebeur, E. and V. Stijn, *Structural mean models for compliance analysis in randomized clinical trials and the impact of errors on measures of exposure.* Stat Methods Med Res, 2005. **14**(4): p. 397-415.
96. Bautista, L.E., et al., *Estimation of bias in nongenetic observational studies using "mendelian triangulation".* Ann Epidemiol, 2006. **16**(9): p. 675-80.
97. Burgess, S., et al., *Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors.* Eur J Epidemiol, 2015. **30**(7): p. 543-52.
98. Do, R., et al., *Common variants associated with plasma triglycerides and risk for coronary artery disease.* Nat Genet, 2013. **45**(11): p. 1345-52.
99. Bowden, J., G. Davey Smith, and S. Burgess, *Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression.* Int J Epidemiol, 2015. **44**(2): p. 512-25.
100. Hassell, J.R. and D.E. Birk, *The molecular basis of corneal transparency.* Exp Eye Res, 2010. **91**(3): p. 326-35.
101. Spoerl, E., G. Wollensak, and T. Seiler, *Increased resistance of crosslinked cornea against enzymatic digestion.* Curr Eye Res, 2004. **29**(1): p. 35-40.
102. Davidson, A.E., et al., *The pathogenesis of keratoconus.* Eye (Lond), 2014. **28**(2): p. 189-95.
103. Burdon, K.P. and A.L. Vincent, *Insights into keratoconus from a genetic perspective.* Clin Exp Optom, 2013. **96**(2): p. 146-54.

104. Li, X., et al., *A genome-wide association study identifies a potential novel gene locus for keratoconus, one of the commonest causes for corneal transplantation in developed countries.* Hum Mol Genet, 2012. **21**(2): p. 421-9.
105. Lu, Y., et al., *Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus.* Nat Genet, 2013. **45**(2): p. 155-63.
106. Kanski., J.J., *Clinical diagnosis in ophthalmology.* Elsevier Mosby, 2006.
107. Threlfall, T.J. and D.R. English, *Sun exposure and pterygium of the eye: a dose-response curve.* Am J Ophthalmol, 1999. **128**(3): p. 280-7.
108. Yam, J.C. and A.K. Kwok, *Ultraviolet light and ocular diseases.* Int Ophthalmol, 2014. **34**(2): p. 383-400.
109. Lee, G.A. and L.W. Hirst, *Ocular surface squamous neoplasia.* Surv Ophthalmol, 1995. **39**(6): p. 429-50.
110. Lindgren, G., B.L. Diffey, and O. Larko, *Basal cell carcinoma of the eyelids and solar ultraviolet radiation exposure.* Br J Ophthalmol, 1998. **82**(12): p. 1412-5.
111. Taylor, H.R., et al., *Effect of ultraviolet radiation on cataract formation.* N Engl J Med, 1988. **319**(22): p. 1429-33.
112. Shah, C.P., et al., *Intermittent and chronic ultraviolet light exposure and uveal melanoma: a meta-analysis.* Ophthalmology, 2005. **112**(9): p. 1599-607.
113. Klein, R., et al., *The epidemiology of age-related macular degeneration.* Am J Ophthalmol, 2004. **137**(3): p. 486-95.
114. Sherwin, J.C., et al., *Distribution of conjunctival ultraviolet autofluorescence in a population-based study: the Norfolk Island Eye Study.* Eye (Lond), 2011. **25**(7): p. 893-900.
115. McKnight, C.M., et al., *Pterygium and conjunctival ultraviolet autofluorescence in young Australian adults: the Raine study.* Clin Experiment Ophthalmol, 2015. **43**(4): p. 300-7.
116. Flitcroft, D.I., *The complex interactions of retinal, optical and environmental factors in myopia aetiology.* Prog Retin Eye Res, 2012. **31**(6): p. 622-60.
117. Sivak, J., *The cause(s) of myopia and the efforts that have been made to prevent it.* Clin Exp Optom, 2012. **95**(6): p. 572-82.
118. Wojciechowski, R., *Nature and nurture: the complex genetics of myopia and refractive error.* Clin Genet, 2011. **79**(4): p. 301-20.
119. Guggenheim, J.A., G. Kirov, and S.A. Hodson, *The heritability of high myopia: a reanalysis of Goldschmidt's data.* J Med Genet, 2000. **37**(3): p. 227-31.
120. Verhoeven, V.J., et al., *Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia.* Nat Genet, 2013. **45**(3): p. 314-8.
121. 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, 2013. **9**(2): p. e1003299.
122. Sorsby, A. and F.A. Young, *Transmission of refractive errors within Eskimo families.* Am J Optom Arch Am Acad Optom, 1970. **47**(3): p. 244-9.
123. Dirani, M., et al., *Outdoor activity and myopia in Singapore teenage children.* Br J Ophthalmol, 2009. **93**(8): p. 997-1000.
124. Rose, K.A., et al., *Outdoor activity reduces the prevalence of myopia in children.* Ophthalmology, 2008. **115**(8): p. 1279-85.
125. Jones, L.A., et al., *Parental history of myopia, sports and outdoor activities, and future myopia.* Invest Ophthalmol Vis Sci, 2007. **48**(8): p. 3524-32.

126. Low, W., et al., *Family history, near work, outdoor activity, and myopia in Singapore Chinese preschool children.* Br J Ophthalmol, 2010. **94**(8): p. 1012-6.

127. Guo, Y., et al., *Myopic shift and outdoor activity among primary school children: one-year follow-up study in Beijing.* PLoS One, 2013. **8**(9): p. e75260.

128. Sherwin, J.C., et al., *The association between time spent outdoors and myopia in children and adolescents: a systematic review and meta-analysis.* Ophthalmology, 2012. **119**(10): p. 2141-51.

129. He, M., et al., *Effect of Time Spent Outdoors at School on the Development of Myopia Among Children in China: A Randomized Clinical Trial.* JAMA, 2015. **314**(11): p. 1142-8.

130. Czepita, D.A. and M. Zejmo, *Environmental factors and myopia.* Ann Acad Med Stetin, 2011. **57**(3): p. 88-92; discussion 92.

131. Drexler, W., et al., *Eye elongation during accommodation in humans: differences between emmetropes and myopes.* Invest Ophthalmol Vis Sci, 1998. **39**(11): p. 2140-7.

132. Berntsen, D.A., D.O. Mutti, and K. Zadnik, *The effect of bifocal add on accommodative lag in myopic children with high accommodative lag.* Invest Ophthalmol Vis Sci, 2010. **51**(12): p. 6104-10.

133. Ashby, R.S. and F. Schaeffel, *The effect of bright light on lens compensation in chicks.* Invest Ophthalmol Vis Sci, 2010. **51**(10): p. 5247-53.

134. Guggenheim, J.A., et al., *Does vitamin D mediate the protective effects of time outdoors on myopia? Findings from a prospective birth cohort.* Invest Ophthalmol Vis Sci, 2014. **55**(12): p. 8550-8.

135. Yazar, S., et al., *Myopia is associated with lower vitamin D status in young adults.* Invest Ophthalmol Vis Sci, 2014. **55**(7): p. 4552-9.

136. Choi, J.A., et al., *Low serum 25-hydroxyvitamin D is associated with myopia in Korean adolescents.* Invest Ophthalmol Vis Sci, 2014. **55**(4): p. 2041-7.

137. Mutti, D.O. and A.R. Marks, *Blood levels of vitamin D in teens and young adults with myopia.* Optom Vis Sci, 2011. **88**(3): p. 377-82.

138. Franz Grehn, R.S., *Glaucoma.* Springer, 2009.

139. Resnikoff, S., et al., *Global data on visual impairment in the year 2002.* Bull World Health Organ, 2004. **82**(11): p. 844-51.

140. Kingman, S., *Glaucoma is second leading cause of blindness globally.* Bull World Health Organ, 2004. **82**(11): p. 887-8.

141. Tham, Y.C., et al., *Global prevalence of glaucoma and projections of glaucoma burden through 2040: a systematic review and meta-analysis.* Ophthalmology, 2014. **121**(11): p. 2081-90.

142. Mantravadi, A.V. and N. Vadhar, *Glaucoma.* Prim Care, 2015. **42**(3): p. 437-49.

143. Rudnicka, A.R., et al., *Variations in primary open-angle glaucoma prevalence by age, gender, and race: a Bayesian meta-analysis.* Invest Ophthalmol Vis Sci, 2006. **47**(10): p. 4254-61.

144. Jeppesen, P. and S. Krag, *[Steroid treatment and risk of glaucoma].* Ugeskr Laeger, 2015. **177**(34): p. 1620-3.

145. Zhao, D., et al., *Diabetes, fasting glucose, and the risk of glaucoma: a meta-analysis.* Ophthalmology, 2015. **122**(1): p. 72-8.

146. Cho, B.J., J.Y. Shin, and H.G. Yu, *Complications of Pathologic Myopia.* Eye Contact Lens, 2016. **42**(1): p. 9-15.

147. Newman-Casey, P.A., et al., *The potential association between postmenopausal hormone use and primary open-angle glaucoma.* JAMA Ophthalmol, 2014. **132**(3): p. 298-303.

148. Choi, J. and M.S. Kook, *Systemic and Ocular Hemodynamic Risk Factors in Glaucoma.* Biomed Res Int, 2015. **2015**: p. 141905.

149. Pang, C.P., et al., *TIGR/MYOC gene sequence alterations in individuals with and without primary open-angle glaucoma.* Invest Ophthalmol Vis Sci, 2002. **43**(10): p. 3231-5.

150. Thorleifsson, G., et al., *Common variants near CAV1 and CAV2 are associated with primary open-angle glaucoma.* Nat Genet, 2010. **42**(10): p. 906-9.

151. Burdon*, K.P., et al., *Genome-wide association study identifies susceptibility loci for open angle glaucoma at TMCO1 and CDKN2B-AS1. *JOINT FIRST AUTHORS.* Nat Genet, 2011. **43**(6): p. 574-8.

152. Wiggs, J.L., et al., *Common variants at 9p21 and 8q22 are associated with increased susceptibility to optic nerve degeneration in glaucoma.* PLoS Genet, 2012. **8**(4): p. e1002654.

153. Chen, Y., et al., *Common variants near ABCA1 and in PMM2 are associated with primary open-angle glaucoma.* Nat Genet, 2014. **46**(10): p. 1115-9.

154. Gharahkhani, P., et al., *Common variants near ABCA1, AFAP1 and GMDS confer risk of primary open-angle glaucoma.* Nat Genet, 2014. **46**(10): p. 1120-5.

155. Mehta, S., *Age-Related Macular Degeneration.* Prim Care, 2015. **42**(3): p. 377-91.

156. Marazita, M.C., et al., *Oxidative stress-induced premature senescence dysregulates VEGF and CFH expression in retinal pigment epithelial cells: Implications for Age-related Macular Degeneration.* Redox Biol, 2015. **7**: p. 78-87.

157. Haas, P., et al., *Impact of visceral fat and pro-inflammatory factors on the pathogenesis of age-related macular degeneration.* Acta Ophthalmol, 2015. **93**(6): p. 533-8.

158. Piermarocchi, S., et al., *Risk Factors and Age-Related Macular Degeneration in a Mediterranean-Basin Population: The PAMDI (Prevalence of Age-Related Macular Degeneration in Italy) Study - Report 2.* Ophthalmic Res, 2015. **55**(3): p. 111-118.

159. Yip, J.L., et al., *Cross Sectional and Longitudinal Associations between Cardiovascular Risk Factors and Age Related Macular Degeneration in the EPIC-Norfolk Eye Study.* PLoS One, 2015. **10**(7): p. e0132565.

160. Ghaem Maralani, H., et al., *Metabolic syndrome and risk of age-related macular degeneration.* Retina, 2015. **35**(3): p. 459-66.

161. Edwards, A.O., et al., *Complement factor H polymorphism and age-related macular degeneration.* Science, 2005. **308**(5720): p. 421-4.

162. Fritsche, L.G., et al., *Seven new loci associated with age-related macular degeneration.* Nat Genet, 2013. **45**(4): p. 433-9, 439e1-2.

163. 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, 2015.

164. Pasquale, L.R., et al., *Estrogen pathway polymorphisms in relation to primary open angle glaucoma: an analysis accounting for gender from the United States.* Mol Vis, 2013. **19**: p. 1471-81.

165. Vajaranant, T.S. and L.R. Pasquale, *Estrogen deficiency accelerates aging of the optic nerve.* Menopause, 2012. **19**(8): p. 942-7.
166. Wei, X., et al., *Is low dose of estrogen beneficial for prevention of glaucoma?* Med Hypotheses, 2012. **79**(3): p. 377-80.
167. Prat, J., *New insights into ovarian cancer pathology.* Ann Oncol, 2012. **23 Suppl 10**: p. x111-7.
168. Sung, P.L., et al., *Global distribution pattern of histological subtypes of epithelial ovarian cancer: a database analysis and systematic review.* Gynecol Oncol, 2014. **133**(2): p. 147-54.
169. Collaborative Group on Epidemiological Studies of Ovarian, C., et al., *Ovarian cancer and smoking: individual participant meta-analysis including 28,114 women with ovarian cancer from 51 epidemiological studies.* Lancet Oncol, 2012. **13**(9): p. 946-56.
170. Faber, M.T., et al., *Cigarette smoking and risk of ovarian cancer: a pooled analysis of 21 case-control studies.* Cancer Causes Control, 2013. **24**(5): p. 989-1004.
171. Aune, D., et al., *Anthropometric factors and ovarian cancer risk: a systematic review and nonlinear dose-response meta-analysis of prospective studies.* Int J Cancer, 2015. **136**(8): p. 1888-98.
172. Collaborative Group on Epidemiological Studies of Ovarian, C., *Ovarian cancer and body size: individual participant meta-analysis including 25,157 women with ovarian cancer from 47 epidemiological studies.* PLoS Med, 2012. **9**(4): p. e1001200.
173. Olsen, C.M., et al., *Obesity and risk of ovarian cancer subtypes: evidence from the Ovarian Cancer Association Consortium.* Endocr Relat Cancer, 2013. **20**(2): p. 251-62.
174. Gapstur, S.M., et al., *Type II diabetes mellitus and the incidence of epithelial ovarian cancer in the cancer prevention study-II nutrition cohort.* Cancer Epidemiol Biomarkers Prev, 2012. **21**(11): p. 2000-5.
175. Lee, J.Y., et al., *Diabetes mellitus and ovarian cancer risk: a systematic review and meta-analysis of observational studies.* Int J Gynecol Cancer, 2013. **23**(3): p. 402-12.
176. Gong, T.T., et al., *Age at menarche and risk of ovarian cancer: a meta-analysis of epidemiological studies.* Int J Cancer, 2013. **132**(12): p. 2894-900.
177. Cuatrecasas, M., et al., *K-ras mutations in mucinous ovarian tumors: a clinicopathologic and molecular study of 95 cases.* Cancer, 1997. **79**(8): p. 1581-6.
178. Cancer Genome Atlas Research, N., *Integrated genomic analyses of ovarian carcinoma.* Nature, 2011. **474**(7353): p. 609-15.
179. Rabinowitz, Y.S., *Keratoconus.* Surv Ophthalmol, 1998. **42**(4): p. 297-319.
180. Bikbov, M.M., G.M. Bikbova, and A.F. Khabibullin, *[Corneal collagen cross-linking in keratoconus management].* Vestn Oftalmol, 2011. **127**(5): p. 21-5.
181. Randleman, J.B., et al., *Risk factors and prognosis for corneal ectasia after LASIK.* Ophthalmology, 2003. **110**(2): p. 267-75.
182. Dimasi, D.P., K.P. Burdon, and J.E. Craig, *The genetics of central corneal thickness.* Br J Ophthalmol, 2010. **94**(8): p. 971-6.
183. Yazar, S., et al., *Raine eye health study: design, methodology and baseline prevalence of ophthalmic disease in a birth-cohort study of young adults.* Ophthalmic Genet, 2013. **34**(4): p. 199-208.

184. Hofman, A., et al., *The Rotterdam Study: 2014 objectives and design update.* Eur J Epidemiol, 2013. **28**(11): p. 889-926.
185. Wright MJ, M.N., *Brisbane Adolescent Twin Study: Outline of study methods and research projects.* Aust J Psychol, 2011(56): p. 65–78.
186. Mackey, D.A., et al., *Twins eye study in Tasmania (TEST): rationale and methodology to recruit and examine twins.* Twin Res Hum Genet, 2009. **12**(5): p. 441-54.
187. Yang, J., et al., *Genomic inflation factors under polygenic inheritance.* Eur J Hum Genet, 2011. **19**(7): p. 807-12.
188. Wagner, A.H., et al., *Exon-level expression profiling of ocular tissues.* Exp Eye Res, 2013. **111**: p. 105-11.
189. Ng, P.C. and S. Henikoff, *SIFT: Predicting amino acid changes that affect protein function.* Nucleic Acids Res, 2003. **31**(13): p. 3812-4.
190. Adzhubei, I.A., et al., *A method and server for predicting damaging missense mutations.* Nat Methods, 2010. **7**(4): p. 248-9.
191. Cuellar-Partida, G., M.E. Renteria, and S. MacGregor, *LocusTrack: Integrated visualization of GWAS results and genomic annotation.* Source Code Biol Med, 2015. **10**: p. 1.
192. Safran, M., et al., *GeneCards Version 3: the human gene integrator.* Database (Oxford), 2010. **2010**: p. baq020.
193. Lee, S., et al., *Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies.* Am J Hum Genet, 2012. **91**(2): p. 224-37.
194. Nakatsu, M.N., et al., *Wnt/beta-catenin signaling regulates proliferation of human cornea epithelial stem/progenitor cells.* Invest Ophthalmol Vis Sci, 2011. **52**(7): p. 4734-41.
195. Nusse, R., *Wnt signaling and stem cell control.* Cell Res, 2008. **18**(5): p. 523-7.
196. Ouyang, H., et al., *WNT7A and PAX6 define corneal epithelium homeostasis and pathogenesis.* Nature, 2014. **511**(7509): p. 358-61.
197. Hirata-Tominaga, K., et al., *Corneal endothelial cell fate is maintained by LGR5 through the regulation of hedgehog and Wnt pathway.* Stem Cells, 2013. **31**(7): p. 1396-407.
198. Meek, K.M., et al., *Transparency, swelling and scarring in the corneal stroma.* Eye (Lond), 2003. **17**(8): p. 927-36.
199. Srinivas, S.P., *Dynamic regulation of barrier integrity of the corneal endothelium.* Optom Vis Sci, 2010. **87**(4): p. E239-54.
200. Fan Gaskin, J.C., D.V. Patel, and C.N. McGhee, *Acute corneal hydrops in keratoconus - new perspectives.* Am J Ophthalmol, 2014. **157**(5): p. 921-8.
201. Hiratsuka, Y., K. Nakayasu, and A. Kanai, *Secondary keratoconus with corneal epithelial iron ring similar to Fleischer's ring.* Jpn J Ophthalmol, 2000. **44**(4): p. 381-6.
202. Reinstein, D.Z., T.J. Archer, and M. Gobbe, *Corneal epithelial thickness profile in the diagnosis of keratoconus.* J Refract Surg, 2009. **25**(7): p. 604-10.
203. Zhou, W. and A. Stojanovic, *Comparison of corneal epithelial and stromal thickness distributions between eyes with keratoconus and healthy eyes with corneal astigmatism >/= 2.0 D.* PLoS One, 2014. **9**(1): p. e85994.
204. Sejpal, K., P. Bakhtiari, and S.X. Deng, *Presentation, diagnosis and management of limbal stem cell deficiency.* Middle East Afr J Ophthalmol, 2013. **20**(1): p. 5-10.

205. Kantaputra, P., et al., *Tricho-odonto-onycho-dermal dysplasia and WNT10A mutations.* Am J Med Genet A, 2014. **164A**(4): p. 1041-8.

206. Nawaz, S., et al., *WNT10A missense mutation associated with a complete odonto-onycho-dermal dysplasia syndrome.* Eur J Hum Genet, 2009. **17**(12): p. 1600-5.

207. Zirbel, G.M., et al., *Odonto-onycho-dermal dysplasia.* Br J Dermatol, 1995. **133**(5): p. 797-800.

208. Evereklioglu, C., et al., *Central corneal thickness is lower in osteogenesis imperfecta and negatively correlates with the presence of blue sclera.* Ophthalmic Physiol Opt, 2002. **22**(6): p. 511-5.

209. Pedersen, U. and T. Bramsen, *Central corneal thickness in osteogenesis imperfecta and otosclerosis.* ORL J Otorhinolaryngol Relat Spec, 1984. **46**(1): p. 38-41.

210. Cisternas, P., C.P. Vio, and N.C. Inestrosa, *Role of Wnt signaling in tissue fibrosis, lessons from skeletal muscle and kidney.* Curr Mol Med, 2014. **14**(4): p. 510-22.

211. Lloyd, S.A., et al., *Shifting paradigms on the role of connexin43 in the skeletal response to mechanical load.* J Bone Miner Res, 2014. **29**(2): p. 275-86.

212. Xie, J., P.J. Tong, and L.W. Xiao, *[Progress on Wnt/beta-catenin signal pathway regulating the cartilage metabolism in osteonecrosis].* Zhongguo Gu Shang, 2013. **26**(7): p. 613-6.

213. Michelacci, Y.M., *Collagens and proteoglycans of the corneal extracellular matrix.* Braz J Med Biol Res, 2003. **36**(8): p. 1037-46.

214. Dua, H.S., et al., *Human corneal anatomy redefined: a novel pre-Descemet's layer (Dua's layer).* Ophthalmology, 2013. **120**(9): p. 1778-85.

215. Hayashi, S., T. Osawa, and K. Tohyama, *Comparative observations on corneas, with special reference to Bowman's layer and Descemet's membrane in mammals and amphibians.* J Morphol, 2002. **254**(3): p. 247-58.

216. McKnight, C.M., et al., *Birth of a cohort--the first 20 years of the Raine study.* Med J Aust, 2012. **197**(11): p. 608-10.

217. Purcell, S., et al., *PLINK: a tool set for whole-genome association and population-based linkage analyses.* Am J Hum Genet, 2007. **81**(3): p. 559-75.

218. Shaun Purcell, C.C., *PLINK 1.90*. 2007: https://www.cog-genomics.org/plink2.

219. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

220. Li, Y., et al., *MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.* Genet Epidemiol, 2010. **34**(8): p. 816-34.

221. Aulchenko, Y.S., M.V. Struchalin, and C.M. van Duijn, *ProbABEL package for genome-wide association analysis of imputed data.* BMC Bioinformatics, 2010. **11**: p. 134.

222. Abecasis, G.R., et al., *Merlin--rapid analysis of dense genetic maps using sparse gene flow trees.* Nat Genet, 2002. **30**(1): p. 97-101.

223. Gray, R.H., G.J. Johnson, and A. Freedman, *Climatic droplet keratopathy.* Surv Ophthalmol, 1992. **36**(4): p. 241-53.

224. Coroneo, M., *Ultraviolet radiation and the anterior eye.* Eye Contact Lens, 2011. **37**(4): p. 214-24.

225. Asawanonda, P. and C.R. Taylor, *Wood's light in dermatology.* Int J Dermatol, 1999. **38**(11): p. 801-7.

226. Ooi, J.L., et al., *Ultraviolet fluorescence photography: patterns in established pterygia.* Am J Ophthalmol, 2007. **143**(1): p. 97-101.
227. Sherwin, J.C., et al., *The association between pterygium and conjunctival ultraviolet autofluorescence: the Norfolk Island Eye Study.* Acta Ophthalmol, 2013. **91**(4): p. 363-70.
228. Svistun, E., et al., *Vision enhancement system for detection of oral cavity neoplasia based on autofluorescence.* Head Neck, 2004. **26**(3): p. 205-15.
229. Newnham, J.P., et al., *Effects of frequent ultrasound during pregnancy: a randomised controlled trial.* Lancet, 1993. **342**(8876): p. 887-91.
230. Ooi, J.L., et al., *Ultraviolet fluorescence photography to detect early sun damage in the eyes of school-aged children.* Am J Ophthalmol, 2006. **141**(2): p. 294-8.
231. Sherwin, J.C., et al., *Reliability and validity of conjunctival ultraviolet autofluorescence measurement.* Br J Ophthalmol, 2012. **96**(6): p. 801-5.
232. Willer, C.J., Y. Li, and G.R. Abecasis, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.
233. Lucas, R.M., et al., *Sun exposure over a lifetime in Australian adults from latitudinally diverse regions.* Photochem Photobiol, 2013. **89**(3): p. 737-44.
234. Wlodarczyk, J., et al., *Pterygium in Australia: a cost of illness study.* Clin Experiment Ophthalmol, 2001. **29**(6): p. 370-5.
235. Hecht, F. and M.G. Shoptaugh, *Winglets of the eye: dominant transmission of early adult pterygium of the conjunctiva.* J Med Genet, 1990. **27**(6): p. 392-4.
236. Elliott, R., *The aetiology and pathology of pterygium.* Trans Ophthalmol Soc Aust, 1966. **25**: p. 71-4.
237. Jacklin, H.N., *Familial Predisposition to Pterygium Formation; Report of a Family.* Am J Ophthalmol, 1964. **57**: p. 481-2.
238. Horkay, I., et al., *Repair of DNA damage in light sensitive human skin diseases.* Arch Dermatol Res, 1978. **263**(3): p. 307-15.
239. Goyal, J.L., et al., *Oculocutaneous manifestations in xeroderma pigmentosa.* Br J Ophthalmol, 1994. **78**(4): p. 295-7.
240. Hammer, H. and I. Korom, *Photodamage of the conjunctiva in patients with porphyria cutanea tarda.* Br J Ophthalmol, 1992. **76**(10): p. 592-3.
241. MacKenzie, F., L.W. Hirst, and A. Hilton, *Pterygia and retinitis pigmentosa.* Aust N Z J Ophthalmol, 1994. **22**(2): p. 145-6.
242. Dun, Y., et al., *Functional and molecular analysis of D-serine transport in retinal Muller cells.* Exp Eye Res, 2007. **84**(1): p. 191-9.
243. Katragadda, S., et al., *Identification and characterization of a Na+-dependent neutral amino acid transporter, ASCT1, in rabbit corneal epithelial cell culture and rabbit cornea.* Curr Eye Res, 2005. **30**(11): p. 989-1002.
244. Jain-Vakkalagadda, B., et al., *Identification of a Na+-dependent cationic and neutral amino acid transporter, B(0,+), in human and rabbit cornea.* Mol Pharm, 2004. **1**(5): p. 338-46.
245. Hosoya, K., et al., *Na(+)-dependent L-arginine transport in the pigmented rabbit conjunctiva.* Exp Eye Res, 1997. **65**(4): p. 547-53.
246. Chaitanya, L., et al., *Collaborative EDNAP exercise on the IrisPlex system for DNA-based prediction of human eye colour.* Forensic Sci Int Genet, 2014. **11**: p. 241-51.
247. Morgan, I. and K. Rose, *How genetic is school myopia?* Prog Retin Eye Res, 2005. **24**(1): p. 1-38.

248. Kempen, J.H., et al., *The prevalence of refractive errors among adults in the United States, Western Europe, and Australia.* Arch Ophthalmol, 2004. **122**(4): p. 495-505.
249. Ip, J.M., et al., *Role of near work in myopia: findings in a sample of Australian school children.* Invest Ophthalmol Vis Sci, 2008. **49**(7): p. 2903-10.
250. Lin, Z., et al., *Near work, outdoor activity, and their association with refractive error.* Optom Vis Sci, 2014. **91**(4): p. 376-82.
251. Mutti, D.O., et al., *Parental myopia, near work, school achievement, and children's refractive error.* Invest Ophthalmol Vis Sci, 2002. **43**(12): p. 3633-40.
252. Morgan, I.G. and K.A. Rose, *ALSPAC study does not support a role for vitamin D in the prevention of myopia.* Invest Ophthalmol Vis Sci, 2014. **55**(12): p. 8559.
253. Greenland, S., *An introduction To instrumental variables for epidemiologists.* Int J Epidemiol, 2000. **29**(6): p. 1102.
254. Afzal, S., et al., *Genetically low vitamin D concentrations and increased mortality: Mendelian randomisation analysis in three large cohorts.* BMJ, 2014. **349**: p. g6330.
255. Ahn, J., et al., *Genome-wide association study of circulating vitamin D levels.* Hum Mol Genet, 2010. **19**(13): p. 2739-45.
256. Wang, T.J., et al., *Common genetic determinants of vitamin D insufficiency: a genome-wide association study.* Lancet, 2010. **376**(9736): p. 180-8.
257. Zhang, Y., et al., *The GC, CYP2R1 and DHCR7 genes are associated with vitamin D levels in northeastern Han Chinese children.* Swiss Med Wkly, 2012. **142**: p. w13636.
258. Zhang, Z., et al., *An analysis of the association between the vitamin D pathway and serum 25-hydroxyvitamin D levels in a healthy Chinese population.* J Bone Miner Res, 2013. **28**(8): p. 1784-92.
259. Brondum-Jacobsen, P., et al., *No evidence that genetically reduced 25-hydroxyvitamin D is associated with increased risk of ischaemic heart disease or myocardial infarction: a Mendelian randomization study.* Int J Epidemiol, 2015. **44**(2): p. 651-61.
260. Theodoratou, E., et al., *Instrumental variable estimation of the causal effect of plasma 25-hydroxy-vitamin D on colorectal cancer risk: a mendelian randomization analysis.* PLoS One, 2012. **7**(6): p. e37662.
261. Boyd, A., et al., *Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children.* Int J Epidemiol, 2013. **42**(1): p. 111-27.
262. Moayyeri, A., et al., *The UK Adult Twin Registry (TwinsUK Resource).* Twin Res Hum Genet, 2013. **16**(1): p. 144-9.
263. Ruxton, N.C.G.D., *Confidence intervals are a more useful complement to nonsignificant tests than are power calculations. .* Behavioral Ecology, 2002: p. 14 (3): 446-447. .
264. Boef, A.G., O.M. Dekkers, and S. le Cessie, *Mendelian randomization studies: a review of the approaches used and the quality of reporting.* Int J Epidemiol, 2015.
265. VanderWeele, T.J., et al., *Methodological challenges in mendelian randomization.* Epidemiology, 2014. **25**(3): p. 427-35.
266. Pascolini, D. and S.P. Mariotti, *Global estimates of visual impairment: 2010.* Br J Ophthalmol, 2012. **96**(5): p. 614-8.

267. Foster, P.J. and Y. Jiang, *Epidemiology of myopia.* Eye (Lond), 2014. **28**(2): p. 202-8.
268. Young, T.L., *Molecular genetics of human myopia: an update.* Optom Vis Sci, 2009. **86**(1): p. E8-E22.
269. Javitt, J.C. and Y.P. Chiang, *The socioeconomic aspects of laser refractive surgery.* Arch Ophthalmol, 1994. **112**(12): p. 1526-30.
270. Jung, S.K., et al., *Prevalence of myopia and its association with body stature and educational level in 19-year-old male conscripts in seoul, South Korea.* Invest Ophthalmol Vis Sci, 2012. **53**(9): p. 5579-83.
271. Fricke, T.R., et al., *Global cost of correcting vision impairment from uncorrected refractive error.* Bull World Health Organ, 2012. **90**(10): p. 728-38.
272. Smith, T.S., et al., *Potential lost productivity resulting from the global burden of uncorrected refractive error.* Bull World Health Organ, 2009. **87**(6): p. 431-7.
273. Pan, C.W., D. Ramamurthy, and S.M. Saw, *Worldwide prevalence and risk factors for myopia.* Ophthalmic Physiol Opt, 2012. **32**(1): p. 3-16.
274. Dirani, M., S.N. Shekar, and P.N. Baird, *The role of educational attainment in refraction: the Genes in Myopia (GEM) twin study.* Invest Ophthalmol Vis Sci, 2008. **49**(2): p. 534-8.
275. Goldschmidt, E. and N. Jacobsen, *Genetic and environmental effects on myopia development and progression.* Eye (Lond), 2014. **28**(2): p. 126-33.
276. Mackey, D.A. and A.W. Hewitt, *Genome-wide association study success in ophthalmology.* Curr Opin Ophthalmol, 2014. **25**(5): p. 386-93.
277. Verhoeven, V.J., et al., *Education influences the role of genetics in myopia.* Eur J Epidemiol, 2013. **28**(12): p. 973-80.
278. French, A.N., et al., *Time outdoors and the prevention of myopia.* Exp Eye Res, 2013. **114**: p. 58-68.
279. Ngo, C.S., et al., *A cluster randomised controlled trial evaluating an incentive-based outdoor physical activity programme to increase outdoor time and prevent myopia in children.* Ophthalmic Physiol Opt, 2014. **34**(3): p. 362-8.
280. Jones-Jordan, L.A., et al., *Visual activity before and after the onset of juvenile myopia.* Invest Ophthalmol Vis Sci, 2011. **52**(3): p. 1841-50.
281. Yi, J.H. and R.R. Li, *[Influence of near-work and outdoor activities on myopia progression in school children].* Zhongguo Dang Dai Er Ke Za Zhi, 2011. **13**(1): p. 32-5.
282. Fan, Q., et al., *Education influences the association between genetic variants and refractive error: a meta-analysis of five Singapore studies.* Hum Mol Genet, 2014. **23**(2): p. 546-54.
283. Wojciechowski, R., et al., *Matrix metalloproteinases and educational attainment in refractive error: evidence of gene-environment interactions in the Age-Related Eye Disease Study.* Ophthalmology, 2013. **120**(2): p. 298-305.
284. Cohn, S.J., C.M. Cohn, and A.R. Jensen, *Myopia and intelligence: a pleiotropic relationship?* Hum Genet, 1988. **80**(1): p. 53-8.
285. Rietveld, C.A., et al., *GWAS of 126,559 individuals identifies genetic variants associated with educational attainment.* Science, 2013. **340**(6139): p. 1467-71.
286. Sanfilippo, P.G., et al., *The heritability of ocular traits.* Surv Ophthalmol, 2010. **55**(6): p. 561-83.

287. Davey Smith, G. and G. Hemani, *Mendelian randomization: genetic anchors for causal inference in epidemiological studies.* Hum Mol Genet, 2014. **23**(R1): p. R89-98.
288. Wichmann, H.E., et al., *KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes.* Gesundheitswesen, 2005. **67 Suppl 1**: p. S26-30.
289. Holle, R., et al., *KORA--a research platform for population based health research.* Gesundheitswesen, 2005. **67 Suppl 1**: p. S19-25.
290. Oexle, K., et al., *Novel association to the proprotein convertase PCSK7 gene locus revealed by analysing soluble transferrin receptor (sTfR) levels.* Hum Mol Genet, 2011. **20**(5): p. 1042-7.
291. Steffens, M., et al., *SNP-based analysis of genetic substructure in the German population.* Hum Hered, 2006. **62**(1): p. 20-9.
292. Age-Related Eye Disease Study Research, G., *A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E and beta carotene for age-related cataract and vision loss: AREDS report no. 9.* Arch Ophthalmol, 2001. **119**(10): p. 1439-52.
293. Age-Related Eye Disease Study Research, G., *A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss: AREDS report no. 8.* Arch Ophthalmol, 2001. **119**(10): p. 1417-36.
294. Foran, S., J.J. Wang, and P. Mitchell, *Causes of visual impairment in two older population cross-sections: the Blue Mountains Eye Study.* Ophthalmic Epidemiol, 2003. **10**(4): p. 215-25.
295. Schache, M., et al., *Genetic association of refractive error and axial length with 15q14 but not 15q25 in the Blue Mountains Eye Study cohort.* Ophthalmology, 2013. **120**(2): p. 292-7.
296. Conrad, D.F., et al., *Origins and functional impact of copy number variation in the human genome.* Nature, 2010. **464**(7289): p. 704-12.
297. International Schizophrenia, C., et al., *Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.* Nature, 2009. **460**(7256): p. 748-52.
298. Brendan Bulik-Sullivan , H.K.F., Verneri Anttila , Alexander Gusev , Felix R Day , ReproGen Consortium , Psychiatric Genomics Consortium , Genetic Consortium for Anorexia Nervosa of the Wellcome Trust Consortium 3 , John R.B. Perry , Nick Patterson , Elise Robinson , Mark J Daly , Alkes L Price , Benjamin M Neale *An Atlas of Genetic Correlations across Human Diseases and Traits.* biorxiv, 2015.
299. Roy, A., et al., *Variation of Axial Ocular Dimensions with Age, Sex, Height, BMI-and Their Relation to Refractive Status.* J Clin Diagn Res, 2015. **9**(1): p. AC01-4.
300. Kaup, A.R., et al., *A review of the brain structure correlates of successful cognitive aging.* J Neuropsychiatry Clin Neurosci, 2011. **23**(1): p. 6-15.
301. Goh, S., et al., *Neuroanatomical correlates of intellectual ability across the life span.* Dev Cogn Neurosci, 2011. **1**(3): p. 305-12.
302. Cooke Bailey, J.N., et al., *Advances in the genomics of common eye diseases.* Hum Mol Genet, 2013. **22**(R1): p. R59-65.
303. Fritsche, L.G., et al., *Age-related macular degeneration: genetics and biology coming together.* Annu Rev Genomics Hum Genet, 2014. **15**: p. 151-71.

304. Black, J.R. and S.J. Clark, *Age-related macular degeneration: genome-wide association studies to translation.* Genet Med, 2015.

305. Fritsche, L.G., et al., *Insights into Rare Genetic Variation From a Large Study of Age-Related Macular Degeneration.* Nature Genetics, 2015. **in press**.

306. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* Am J Hum Genet, 2011. **88**(1): p. 76-82.

307. Telander, D.G., *Inflammation and age-related macular degeneration (AMD).* Semin Ophthalmol, 2011. **26**(3): p. 192-7.

308. Stanton, C.M. and A.F. Wright, *Inflammatory biomarkers for AMD.* Adv Exp Med Biol, 2014. **801**: p. 251-7.

309. Wang, Y., V.M. Wang, and C.C. Chan, *The role of anti-inflammatory agents in age-related macular degeneration (AMD) treatment.* Eye (Lond), 2011. **25**(2): p. 127-39.

310. Hall, J.B., et al., *Estimating cumulative pathway effects on risk for age-related macular degeneration using mixed linear models.* BMC Bioinformatics, 2015. **16**: p. 329.

311. Cohen, L.P. and L.R. Pasquale, *Clinical characteristics and current treatment of glaucoma.* Cold Spring Harb Perspect Med, 2014(6).

312. Bodh, S.A., et al., *Inflammatory glaucoma.* Oman J Ophthalmol, 2011. **4**(1): p. 3-9.

313. Vohra, R., J.C. Tsai, and M. Kolko, *The role of inflammation in the pathogenesis of glaucoma.* Surv Ophthalmol, 2013. **58**(4): p. 311-20.

314. Cascio, C., et al., *The estrogenic retina: The potential contribution to healthy aging and age-related neurodegenerative diseases of the retina.* Steroids, 2015.

315. Craig, J.E., et al., *Rapid inexpensive genome-wide association using pooled whole blood.* Genome Res, 2009. **19**(11): p. 2075-80.

316. Souzeau, E., et al., *Australian and New Zealand Registry of Advanced Glaucoma: methodology and recruitment.* Clin Experiment Ophthalmol, 2012. **40**(6): p. 569-75.

317. Joachim, N., et al., *The Incidence and Progression of Age-Related Macular Degeneration over 15 Years: The Blue Mountains Eye Study.* Ophthalmology, 2015.

318. Chornokur, G., et al., *Common Genetic Variation In Cellular Transport Genes and Epithelial Ovarian Cancer (EOC) Risk.* PLoS One, 2015. **10**(6): p. e0128106.

319. Sopik, V., et al., *Why have ovarian cancer mortality rates declined? Part I. Incidence.* Gynecol Oncol, 2015.

320. Kurman R, C.M., Herrington C, Young R, *WHO classification of tumours of female reproductive organs. .* 2014(World Health Organization Classification of Tumours. France, IARC.).

321. Kurman, R.J. and M. Shih Ie, *Pathogenesis of ovarian cancer: lessons from morphology and molecular biology and their clinical implications.* Int J Gynecol Pathol, 2008. **27**(2): p. 151-60.

322. Shih Ie, M. and R.J. Kurman, *Ovarian tumorigenesis: a proposed model based on morphological and molecular genetic analysis.* Am J Pathol, 2004. **164**(5): p. 1511-8.

323. Malpica, A., et al., *Grading ovarian serous carcinoma using a two-tier system.* Am J Surg Pathol, 2004. **28**(4): p. 496-504.

324. Anglesio, M.S., et al., *Molecular characterization of mucinous ovarian tumours supports a stratified treatment approach with HER2 targeting in 19% of carcinomas.* J Pathol, 2013. **229**(1): p. 111-20.

325. Della Pepa, C., et al., *Low Grade Serous Ovarian Carcinoma: from the molecular characterization to the best therapeutic strategy.* Cancer Treat Rev, 2015. **41**(2): p. 136-43.

326. Soslow, R.A., *Histologic subtypes of ovarian carcinoma: an overview.* Int J Gynecol Pathol, 2008. **27**(2): p. 161-74.

327. Risch, H.A., et al., *Differences in risk factors for epithelial ovarian cancer by histologic type. Results of a case-control study.* Am J Epidemiol, 1996. **144**(4): p. 363-72.

328. Wiegand, K.C., et al., *ARID1A mutations in endometriosis-associated ovarian carcinomas.* N Engl J Med, 2010. **363**(16): p. 1532-43.

329. Simons, M., et al., *Survival of Patients With Mucinous Ovarian Carcinoma and Ovarian Metastases: A Population-Based Cancer Registry Study.* Int J Gynecol Cancer, 2015.

330. Devouassoux-Shisheboran, M. and C. Genestie, *Pathobiology of ovarian carcinomas.* Chin J Cancer, 2015. **34**(1): p. 50-5.

331. Alsop, K., et al., *BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group.* J Clin Oncol, 2012. **30**(21): p. 2654-63.

332. Berchuck, A., et al., *Frequency of germline and somatic BRCA1 mutations in ovarian cancer.* Clin Cancer Res, 1998. **4**(10): p. 2433-7.

333. Grisham, R.N., et al., *BRAF mutation is associated with early stage disease and improved outcome in patients with low-grade serous ovarian cancer.* Cancer, 2013. **119**(3): p. 548-54.

334. Jones, S., et al., *Low-grade serous carcinomas of the ovary contain very few point mutations.* J Pathol, 2012. **226**(3): p. 413-20.

335. Jones, S., et al., *Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma.* Science, 2010. **330**(6001): p. 228-31.

336. Bolton, K.L., et al., *Common variants at 19p13 are associated with susceptibility to ovarian cancer.* Nat Genet, 2010. **42**(10): p. 880-4.

337. Goode, E.L., et al., *A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24.* Nat Genet, 2010. **42**(10): p. 874-9.

338. Song, H., et al., *A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2.* Nat Genet, 2009. **41**(9): p. 996-1000.

339. Pharoah, P.D., et al., *GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer.* Nat Genet, 2013. **45**(4): p. 362-70, 370e1-2.

340. Permuth-Wey, J., et al., *Identification and molecular characterization of a new ovarian cancer susceptibility locus at 17q21.31.* Nat Commun, 2013. **4**: p. 1627.

341. Bojesen, S.E., et al., *Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer.* Nat Genet, 2013. **45**(4): p. 371-84, 384e1-2.

342. Lu, Y., et al., *Shared genetics underlying epidemiological association between endometriosis and ovarian cancer.* Hum Mol Genet, 2015. **24**(20): p. 5955-64.

343. Wiren, S., et al., *Pooled cohort study on height and risk of cancer and cancer death.* Cancer Causes Control, 2014. **25**(2): p. 151-9.

344. Vink, J.M. and D.I. Boomsma, *Interplay between heritability of smoking and environmental conditions? A comparison of two birth cohorts.* BMC Public Health, 2011. **11**: p. 316.
345. Vink, J.M., G. Willemsen, and D.I. Boomsma, *Heritability of smoking initiation and nicotine dependence.* Behav Genet, 2005. **35**(4): p. 397-406.
346. Lu, Y., et al., *Most common 'sporadic' cancers have a significant germline genetic component.* Hum Mol Genet, 2014. **23**(22): p. 6112-8.
347. Yang, J., et al., *Genome partitioning of genetic variation for complex traits using common SNPs.* Nat Genet, 2011. **43**(6): p. 519-25.
348. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets.* Gigascience, 2015. **4**: p. 7.
349. Cross-Disorder Group of the Psychiatric Genomics, C., et al., *Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.* Nat Genet, 2013. **45**(9): p. 984-94.
350. Locke, A.E., et al., *Genetic studies of body mass index yield new insights for obesity biology.* Nature, 2015. **518**(7538): p. 197-206.
351. Wood, A.R., et al., *Defining the role of common variation in the genomic and biological architecture of adult human height.* Nat Genet, 2014. **46**(11): p. 1173-86.
352. Berndt, S.I., et al., *Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture.* Nat Genet, 2013. **45**(5): p. 501-12.
353. Perry, J.R., et al., *Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche.* Nature, 2014. **514**(7520): p. 92-7.
354. Tobacco and C. Genetics, *Genome-wide meta-analyses identify multiple loci associated with smoking behavior.* Nat Genet, 2010. **42**(5): p. 441-7.
355. Morris, A.P., et al., *Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes.* Nat Genet, 2012. **44**(9): p. 981-90.
356. Nyholt, D.R., *A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other.* Am J Hum Genet, 2004. **74**(4): p. 765-9.
357. Gilks, C.B., et al., *Tumor cell type can be reproducibly diagnosed and is of independent prognostic significance in patients with maximally debulked ovarian carcinoma.* Hum Pathol, 2008. **39**(8): p. 1239-51.
358. Klein, R.J., et al., *Complement factor H polymorphism in age-related macular degeneration.* Science, 2005. **308**(5720): p. 385-9.
359. Finucane, H.K., et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nat Genet, 2015. **47**(11): p. 1228-35.
360. Gusev, A., et al., *Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases.* Am J Hum Genet, 2014. **95**(5): p. 535-52.
361. Schork, A.J., et al., *All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs.* PLoS Genet, 2013. **9**(4): p. e1003449.
362. Huang, J., et al., *Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.* Nat Commun, 2015. **6**: p. 8111.

363. Shane McCarthy, S.D., Warren Kretzschmar, Olivier Delaneau, and A.R. Wood, *A reference panel of 64,976 haplotypes for genotype imputation.* BioRxiv, 2016.

364. Katan, M.B., *Apolipoprotein E isoforms, serum cholesterol, and cancer.* Lancet, 1986. **1**(8479): p. 507-8.

365. Jorgenson, E., et al., *Common coding variants in the HLA-DQB1 region confer susceptibility to age-related macular degeneration.* Eur J Hum Genet, 2016.

366. Pinazo-Duran, M.D., et al., *Oxidative stress and mitochondrial failure in the pathogenesis of glaucoma neurodegeneration.* Prog Brain Res, 2015. **220**: p. 127-53.

367. Lascaratos, G., et al., *Resistance to the most common optic neuropathy is associated with systemic mitochondrial efficiency.* Neurobiol Dis, 2015. **82**: p. 78-85.

368. Itakura, T., D.M. Peters, and M.E. Fini, *Glaucomatous MYOC mutations activate the IL-1/NF-kappaB inflammatory stress response and the glaucoma marker SELE in trabecular meshwork cells.* Mol Vis, 2015. **21**: p. 1071-84.

369. Tanito, M., et al., *Correlation between Systemic Oxidative Stress and Intraocular Pressure Level.* PLoS One, 2015. **10**(7): p. e0133582.

370. Lambros, M.L. and S.M. Plafker, *Oxidative Stress and the Nrf2 Anti-Oxidant Transcription Factor in Age-Related Macular Degeneration.* Adv Exp Med Biol, 2016. **854**: p. 67-72.