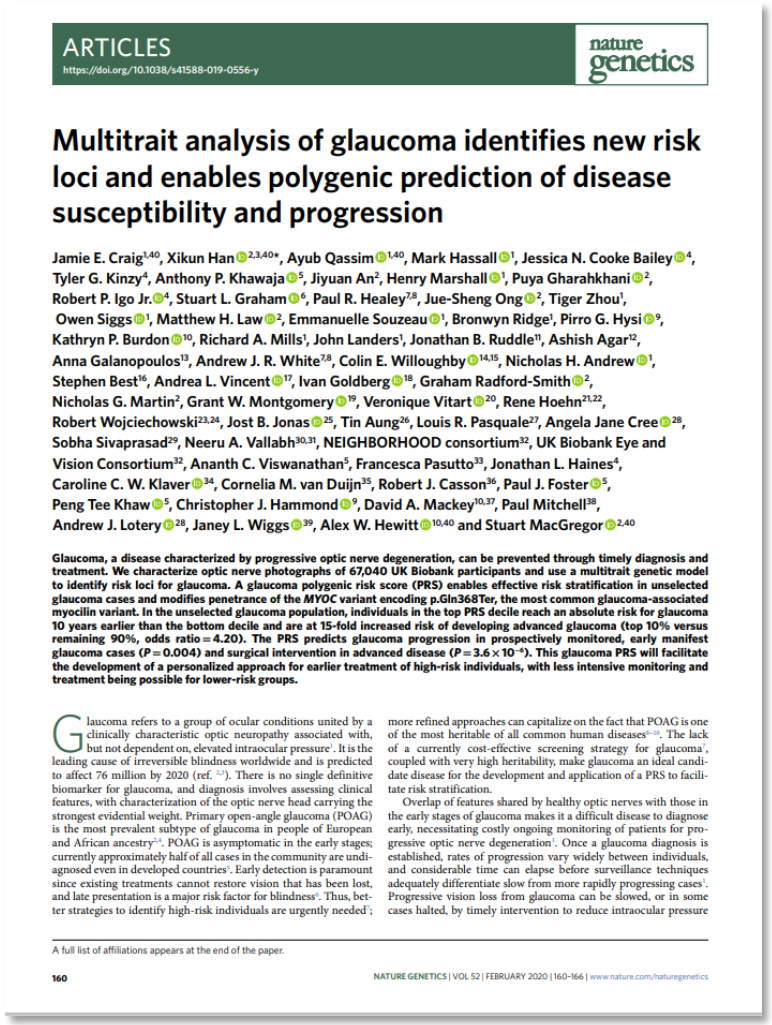
# Polygenetic Risk Score (PRS)

**UNIL BSc course: Solving Biological Problems that require Math 2020**

Alex Button

alexanderluke.button@unil.ch

# Motivation



ARTICLES
https://doi.org/10.1038/s41588-019-0556-y

nature genetics

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

Jamie E. Craig[1,40], Xikun Han[2,3,40*], Ayub Qassim[1,40], Mark Hassall[1], Jessica N. Cooke Bailey[4], Tyler G. Kinzy[4], Anthony P. Khawaja[5], Jiyuan An[2], Henry Marshall[1], Puya Gharahkhani[2], Robert P. Igo Jr.[4], Stuart L. Graham[6], Paul R. Healey[7,8], Jue-Sheng Ong[2], Tiger Zhou[1], Owen Siggs[1], Matthew H. Law[2], Emmanuelle Souzeau[1], Bronwyn Ridge[1], Pirro G. Hysi[9], Kathryn P. Burdon[10], Richard A. Mills[1], John Landers[1], Jonathan B. Ruddle[11], Ashish Agar[12], Anna Galanopoulos[13], Andrew J. R. White[7,8], Colin E. Willoughby[14,15], Nicholas H. Andrew[1], Stephen Best[16], Andrea L. Vincent[17], Ivan Goldberg[18], Graham Radford-Smith[2], Nicholas G. Martin[2], Grant W. Montgomery[19], Veronique Vitart[20], Rene Hoehn[21,22], Robert Wojciechowski[23,24], Jost B. Jonas[25], Tin Aung[26], Louis R. Pasquale[27], Angela Jane Cree[28], Sobha Sivaprasad[29], Neeru A. Vallabh[30,31], NEIGHBORHOOD consortium[32], UK Biobank Eye and Vision Consortium[32], Ananth C. Viswanathan[5], Francesca Pasutto[33], Jonathan L. Haines[4], Caroline C. W. Klaver[34], Cornelia M. van Duijn[35], Robert J. Casson[36], Paul J. Foster[5], Peng Tee Khaw[5], Christopher J. Hammond[9], David A. Mackey[10,37], Paul Mitchell[38], Andrew J. Lotery[28], Janey L. Wiggs[39], Alex W. Hewitt[10,40] and Stuart MacGregor[2,40]

**Glaucoma, a disease characterized by progressive optic nerve degeneration, can be prevented through timely diagnosis and treatment. We characterize optic nerve photographs of 67,040 UK Biobank participants and use a multitrait genetic model to identify risk loci for glaucoma. A glaucoma polygenic risk score (PRS) enables effective risk stratification in unselected glaucoma cases and modifies penetrance of the MYOC variant encoding p.Gln368Ter, the most common glaucoma-associated myocilin variant. In the unselected glaucoma population, individuals in the top PRS decile reach an absolute risk for glaucoma 10 years earlier than the bottom decile and are at 15-fold increased risk of developing advanced glaucoma (top 10% versus remaining 90%, odds ratio = 4.20). The PRS predicts glaucoma progression in prospectively monitored, early manifest glaucoma cases ($P = 0.004$) and surgical intervention in advanced disease ($P = 3.6 \times 10^{-4}$). This glaucoma PRS will facilitate the development of a personalized approach for earlier treatment of high-risk individuals, with less intensive monitoring and treatment being possible for lower-risk groups.**

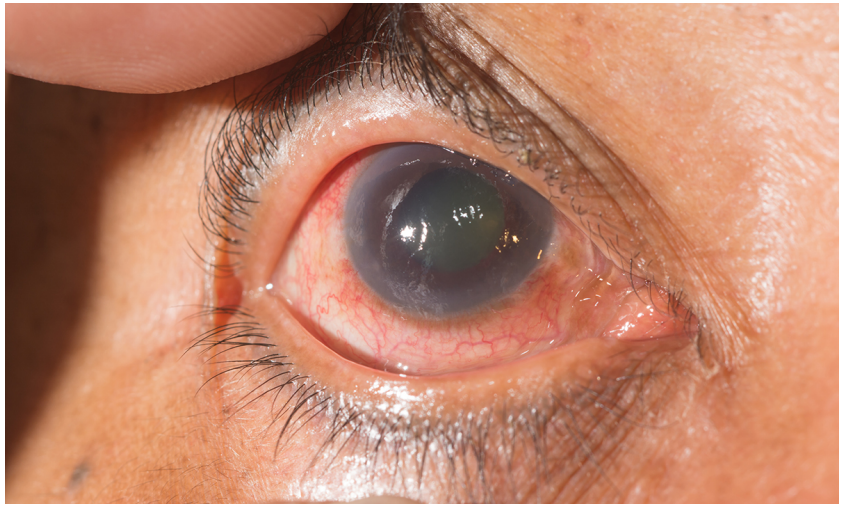
Glaucoma refers to a group of ocular conditions united by a clinically characteristic optic neuropathy associated with, but not dependent on, elevated intraocular pressure[1]. It is the leading cause of irreversible blindness worldwide and is predicted to affect 76 million by 2020 (ref. [2,3]). There is no single definitive biomarker for glaucoma, and diagnosis involves assessing clinical features, with characterization of the optic nerve head carrying the strongest evidential weight. Primary open-angle glaucoma (POAG) is the most prevalent subtype of glaucoma in people of European and African ancestry[2,4]. POAG is asymptomatic in the early stages; currently approximately half of all cases in the community are undiagnosed even in developed countries[5]. Early detection is paramount since existing treatments cannot restore vision that has been lost, and late presentation is a major risk factor for blindness[6]. Thus, better strategies to identify high-risk individuals are urgently needed[7]; more refined approaches can capitalize on the fact that POAG is one of the most heritable of all common human diseases[8–10]. The lack of a currently cost-effective screening strategy for glaucoma[7], coupled with very high heritability, make glaucoma an ideal candidate disease for the development and application of a PRS to facilitate risk stratification.

Overlap of features shared by healthy optic nerves with those in the early stages of glaucoma makes it a difficult disease to diagnose early, necessitating costly ongoing monitoring of patients for progressive optic nerve degeneration[1]. Once a glaucoma diagnosis is established, rates of progression vary widely between individuals, and considerable time can elapse before surveillance techniques adequately differentiate slow from more rapidly progressing cases[1]. Progressive vision loss from glaucoma can be slowed, or in some cases halted, by timely intervention to reduce intraocular pressure

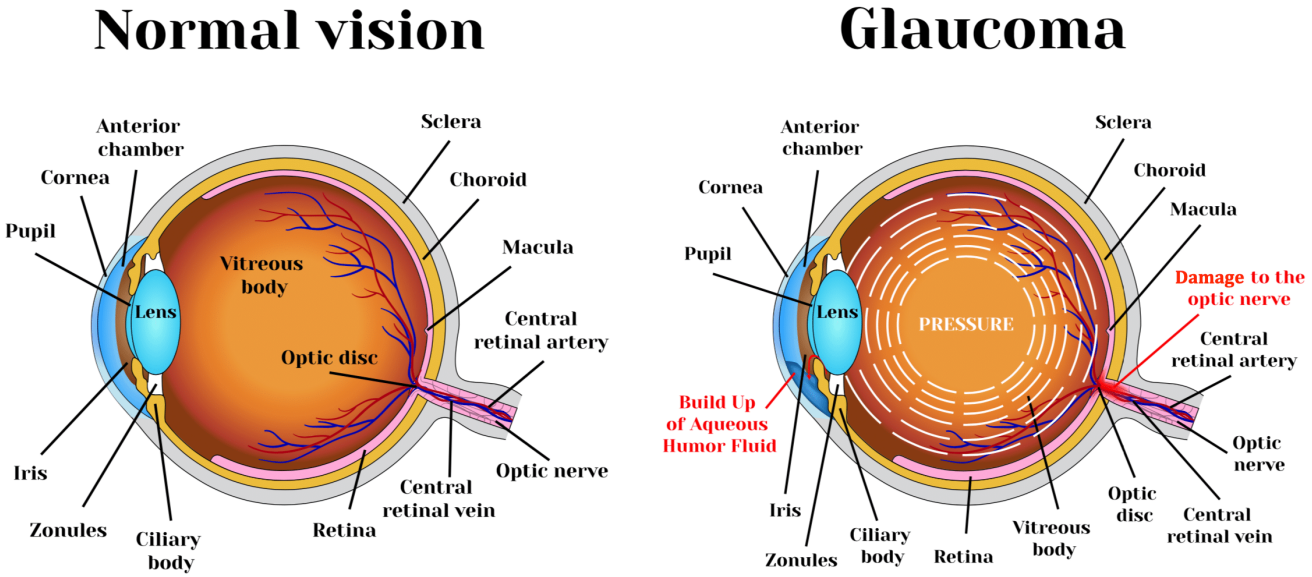
A full list of affiliations appears at the end of the paper.



Craig JE *et al.* Nat Genet. 2020 (2):160-166



https://insightplus.mja.com.au/2018/9/what-gps-can-do-to-boost-glaucoma-diagnosis-rates/




https://topdogtips.com/wp-content/uploads/2018/09/Primary-Open-Angle-Glaucoma-in-Dogs-POAG.jpg

# Motivation

$$PRS: Patient's\ genome \rightarrow Risk\ score$$

$$PRS = \sum_{i=0}^{N} \beta_i X_i$$

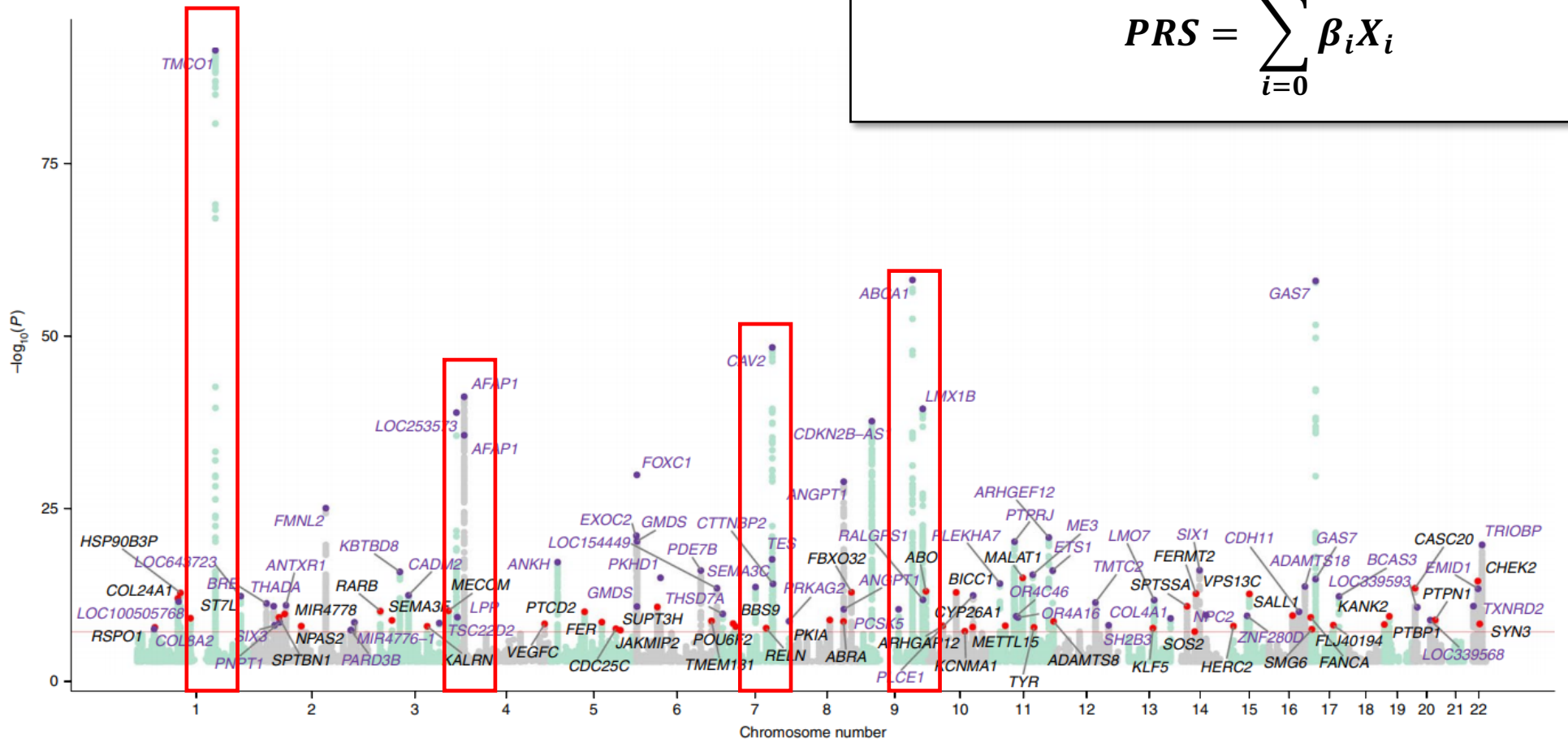
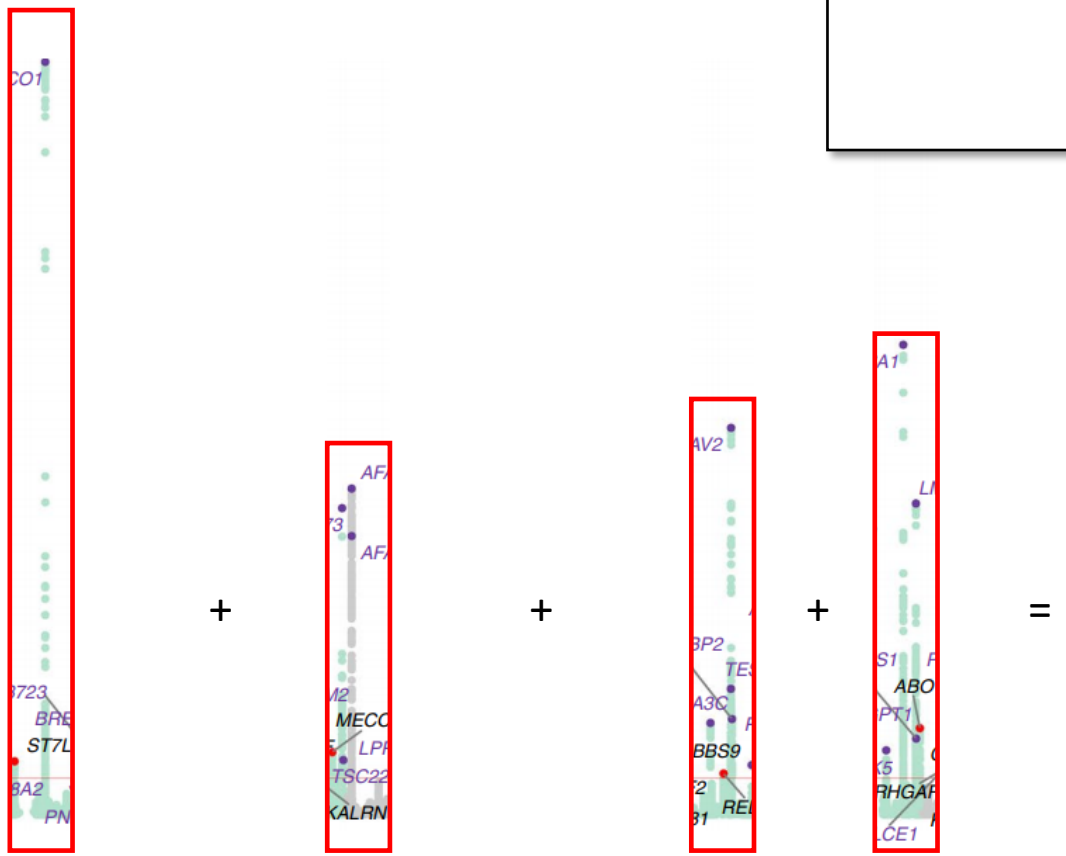



**Fig. 1 | Manhattan plot displaying glaucoma-specific *P* values from the MTAG analysis.** The samples used in the multitrait analysis are presented in Extended Data Fig. 1a. Previously unknown SNPs are highlighted with red dots, with the nearest gene names in black text. Known SNPs are highlighted with purple dots, with the nearest gene names in purple text. The red line is the genome-wide significance level at $5 \times 10^{-8}$.

# Motivation

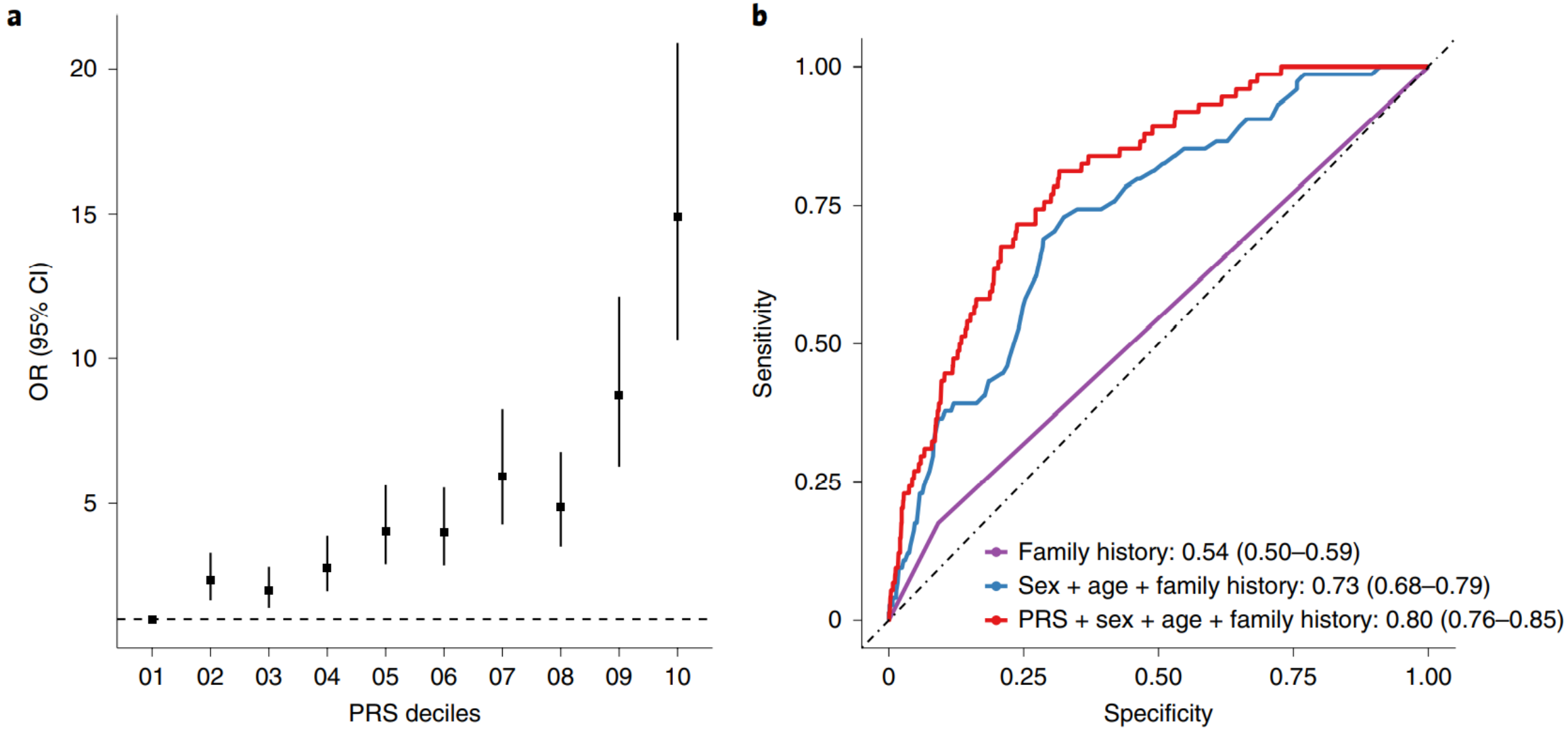
$$PRS : Patient's\ genome \rightarrow Risk\ score$$

$$PRS = \sum_{i=0}^{N} \beta_i X_i$$



Weighted sum of SNPs

Patient specific risk score


a
20
15
10
5
OR (95% CI)
01
02
03
04
05
06
07
08
09
10
PRS deciles
b
1.00
0.75
0.50
0.25
0
Sensitivity
0
0.25
0.50
0.75
1.00
Specificity
Family history: 0.54 (0.50–0.59)
Sex + age + family history: 0.73 (0.68–0.79)
PRS + sex + age + family history: 0.80 (0.76–0.85)

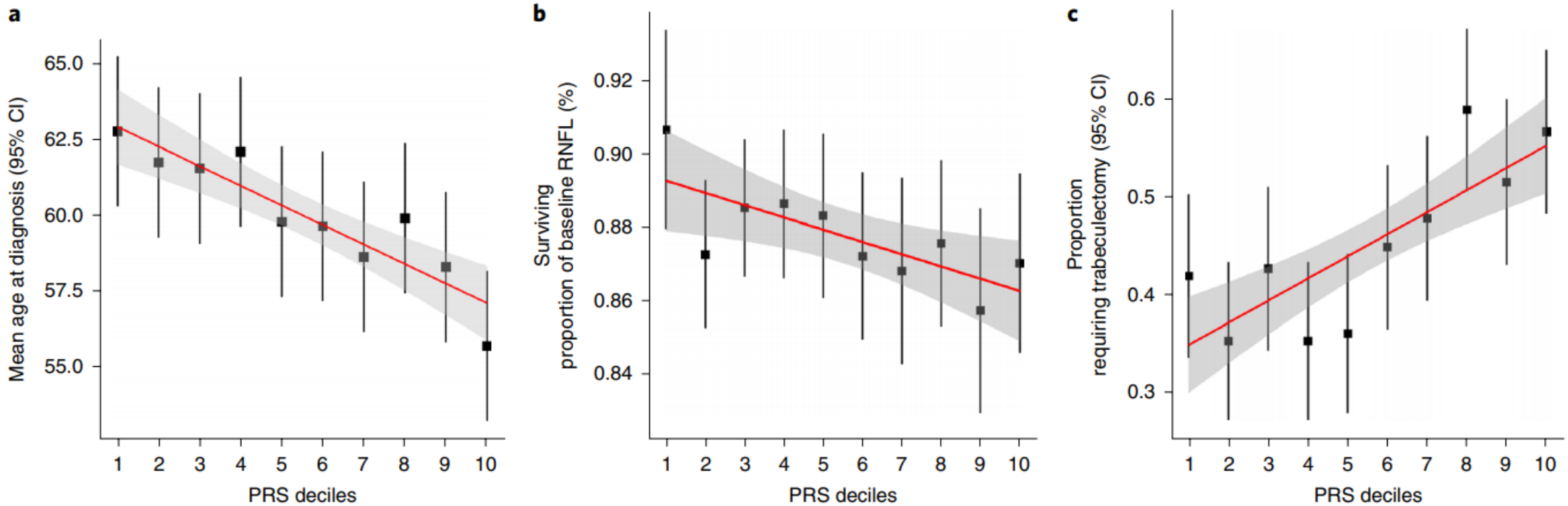



**Fig. 4 | Clinical implications of the glaucoma PRS. a**, Mean age at diagnosis (years) for each decile of the PRS in the ANZRAG cohort (linear regression $P = 1.8 \times 10^{-5}$). A total of 1,336 cases had accurate age at diagnosis information. We calculated the mean age at diagnosis for each decile of PRS, adjusted for sex and the first four principal components in a linear regression model. The black squares are the regression-based mean age at diagnosis, with the error bars for the 95% CIs. The red line is the line of best fit, with the 95% CIs in gray. **b**, Proportion of preserved baseline retinal nerve fiber layer for PROGRESSA participants with early manifest glaucoma plotted against PRS decile ($n = 388$; linear regression $P = 0.004$). The black squares are the retinal nerve fiber layer proportions, with the error bars showing the 95% CIs. The remaining retinal nerve fiber layer proportion is calculated for the most affected quadrant of the most affected eye of each patient, as determined on optical coherence tomography scans at baseline and latest follow-up scan. **c**, Proportion of patients requiring trabeculectomy in either eye in the ANZRAG POAG cohort (linear regression $P = 3.6 \times 10^{-6}$). There were 1,360 cases with records of surgical treatment status. The black squares represent the observed average proportion of cases in each decile of PRS who required trabeculectomy, with 95% CI bars. The line of best fit is shown in red, with the 95% CI shaded in gray.

# How do we calculate the PRS?

$$\boldsymbol{PRS : Patient's\ genome \rightarrow Risk\ score}$$

$$\boldsymbol{PRS = \sum_{i=0}^{N} \beta_i X_i}$$

- $X_i$ = allelic dosage of the $i^{th}$SNP position (0,1,2)

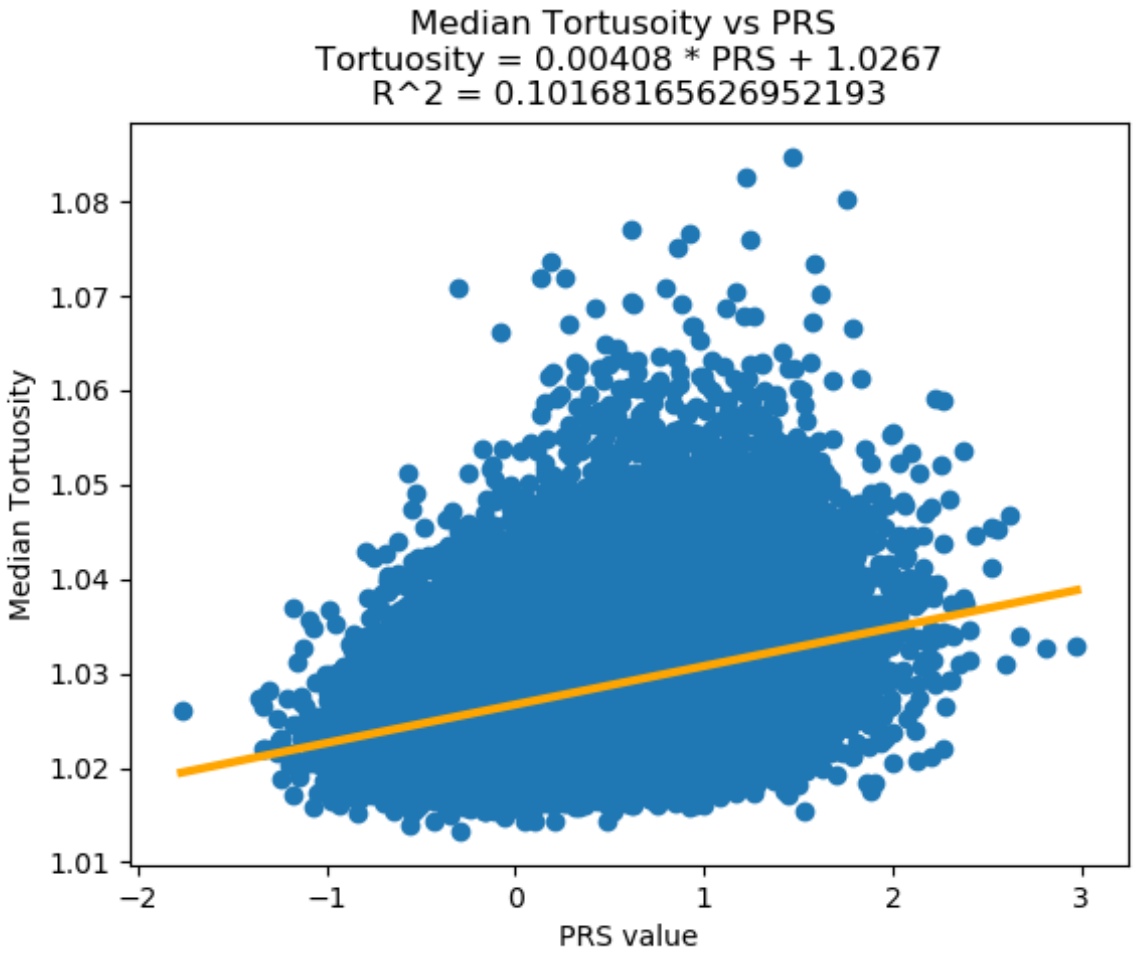
This value represents whether or not the patient possesses 0, 1, or 2 copies of the main effect allele in their chromosome

e.g for effect allele T this would be G/G(0), G/T(1), or T/T(2)

- $\beta_i$ = effect size of the $i^{th}$SNP position (float)

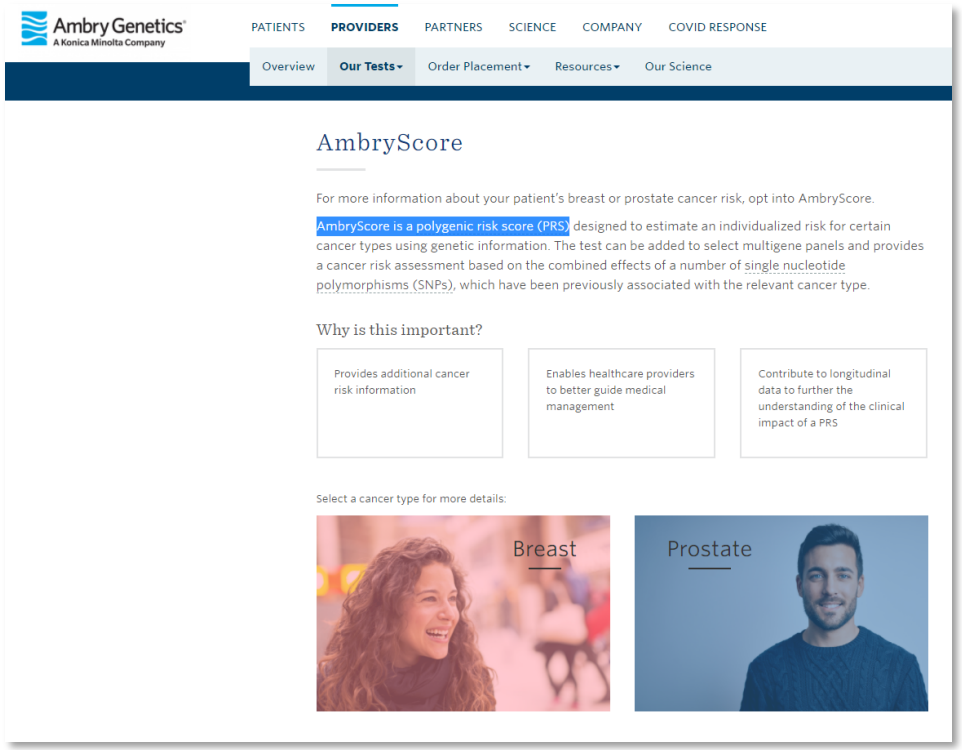
A measure of how much a trait varies with the allelic dosage at position $i$. (For continuous traits, this is often the slope of a linear fit between the trait and dosage)

# Real Results

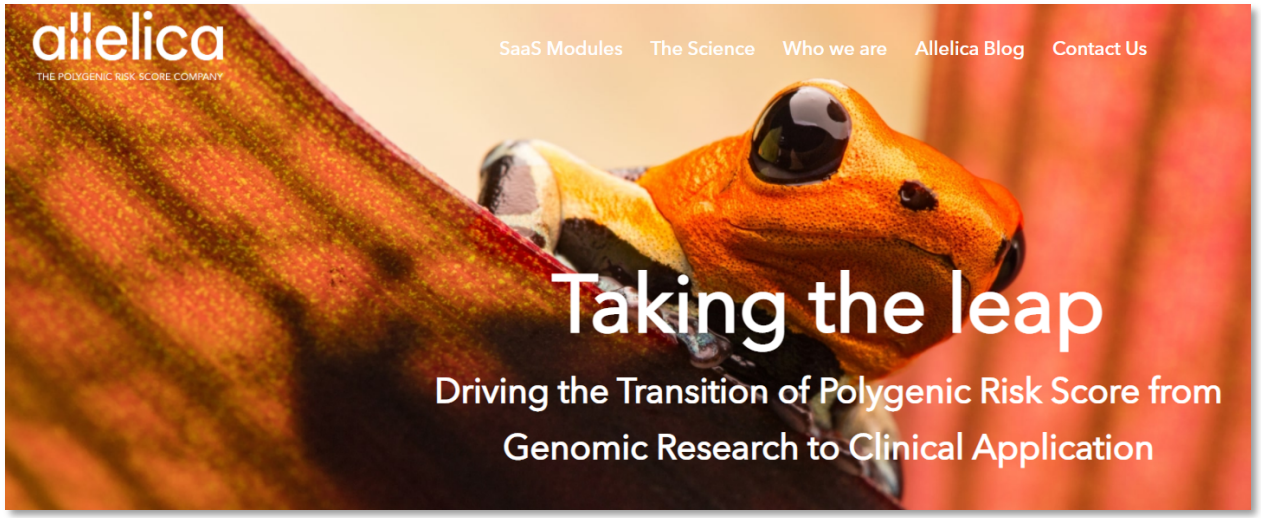



| Phenotype | Disease Indication | ROC-AUC |
|---|---|---|
| Median Tortuosity | Angina | 0.5448 |
| PRS | Angina | 0.5168 |
| Median Tortuosity | Diabetes | 0.5237 |
| PRS | Diabetes | 0.4996 |
| Median Tortuosity | DV Thrombosis | 0.5252 |
| PRS | DV Thrombosis | 0.4904 |
| Median Tortuosity | Heart Attack | 0.5338 |
| PRS | Heart Attack | 0.4968 |
| Median Tortuosity | Hypertension | 0.5644 |
| PRS | Hypertension | 0.5131 |
| Median Tortuosity | Stroke | 0.545 |
| PRS | Stroke | 0.5127 |


https://www.genome.gov/Health/Genomics-and-Medicine/Polygenic-risk-scores




https://www.ambrygen.com/providers/ambryscore



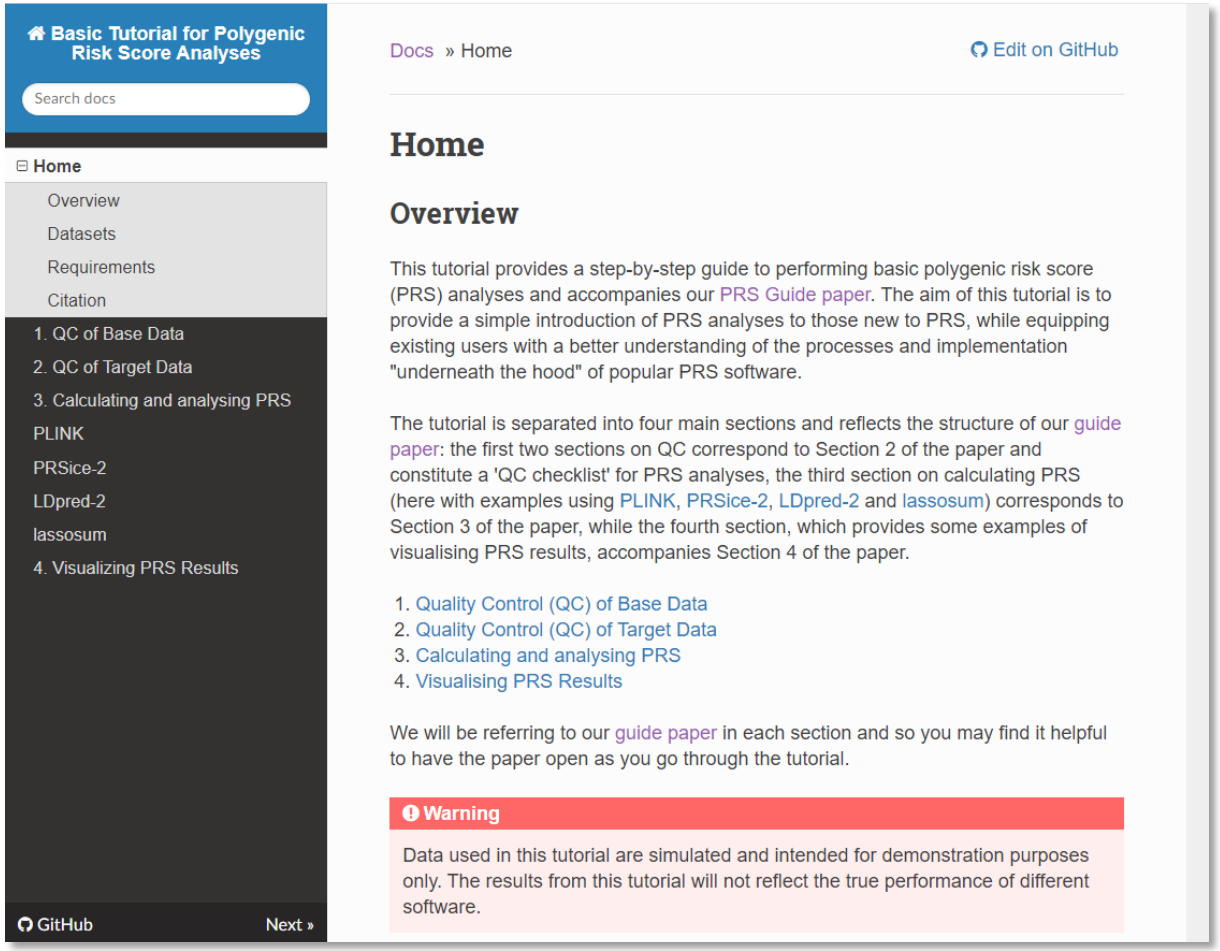

https://www.allelica.com/

**Project goal:**
Construct a polygenetic risk score to accurately evaluate a patient's disease risk.

Learning objectives:

- Learn to calculate the polygenetic risk score from GWAS summary statistics.
- Learn to manipulate and analyze genetic patient data (big data, high performance computing).
- Utilize the PRS for disease outcome prediction.
- Look into extensions to the basic PRS method (LASSO, linear mixed models).
- Investigate various phenotypes and their influence on the PRS and its predictive capabilities.

# Resources



Basic Tutorial for Polygenic Risk Score Analyses

Search docs

- Home
  - Overview
  - Datasets
  - Requirements
  - Citation
- 1. QC of Base Data
- 2. QC of Target Data
- 3. Calculating and analysing PRS
- PLINK
- PRSice-2
- LDpred-2
- lassosum
- 4. Visualizing PRS Results

GitHub Next »

Docs » Home — Edit on GitHub

## Home

### Overview

This tutorial provides a step-by-step guide to performing basic polygenic risk score (PRS) analyses and accompanies our PRS Guide paper. The aim of this tutorial is to provide a simple introduction of PRS analyses to those new to PRS, while equipping existing users with a better understanding of the processes and implementation "underneath the hood" of popular PRS software.

The tutorial is separated into four main sections and reflects the structure of our guide paper: the first two sections on QC correspond to Section 2 of the paper and constitute a 'QC checklist' for PRS analyses, the third section on calculating PRS (here with examples using PLINK, PRSice-2, LDpred-2 and lassosum) corresponds to Section 3 of the paper, while the fourth section, which provides some examples of visualising PRS results, accompanies Section 4 of the paper.

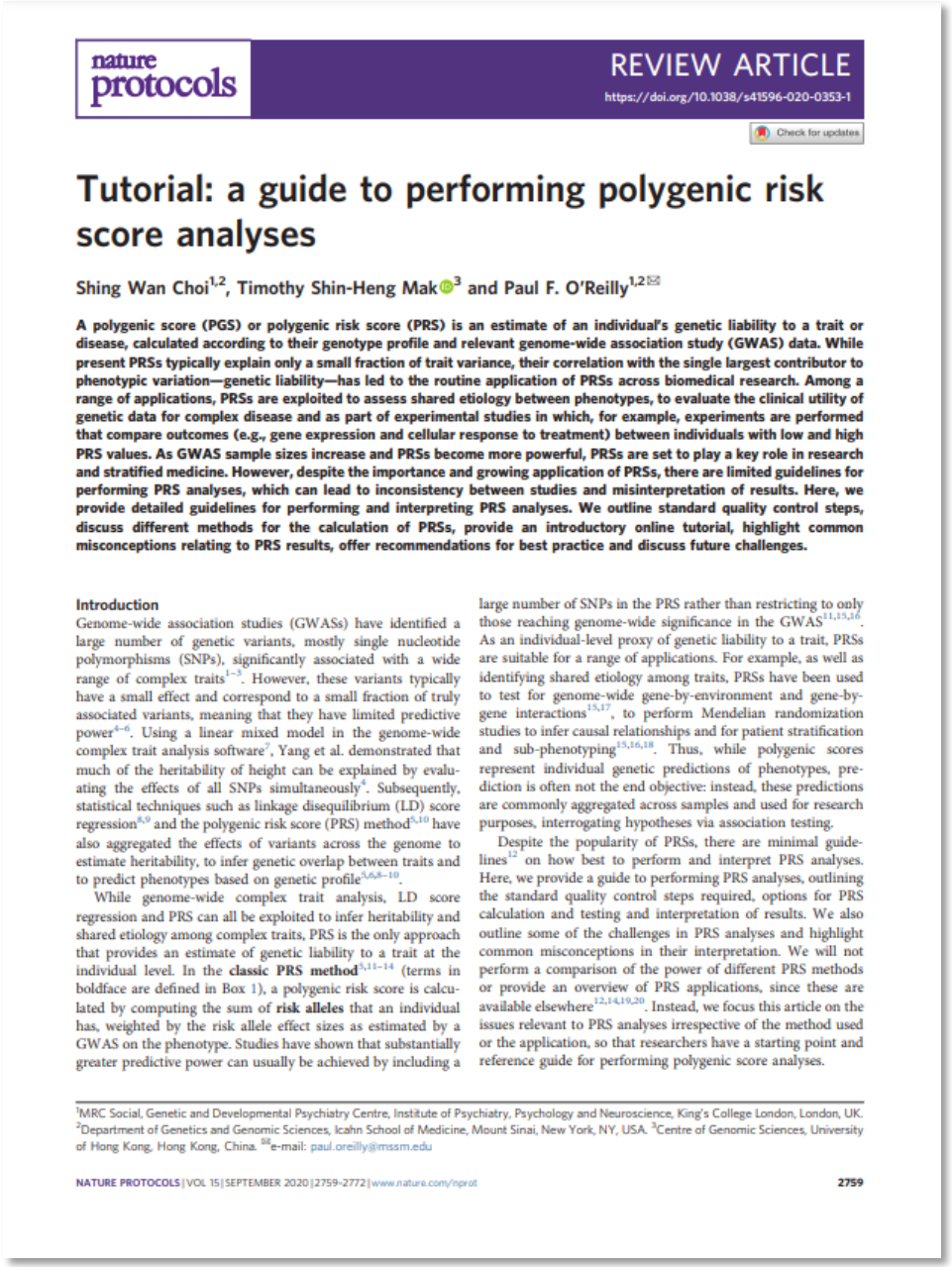
1. Quality Control (QC) of Base Data
2. Quality Control (QC) of Target Data
3. Calculating and analysing PRS
4. Visualising PRS Results

We will be referring to our guide paper in each section and so you may find it helpful to have the paper open as you go through the tutorial.

**Warning**

Data used in this tutorial are simulated and intended for demonstration purposes only. The results from this tutorial will not reflect the true performance of different software.

https://choishingwan.github.io/PRS-Tutorial/



nature protocols — REVIEW ARTICLE
https://doi.org/10.1038/s41596-020-0353-1
Check for updates

# Tutorial: a guide to performing polygenic risk score analyses

Shing Wan Choi[1,2], Timothy Shin-Heng Mak[3] and Paul F. O'Reilly[1,2]✉

**A polygenic score (PGS) or polygenic risk score (PRS) is an estimate of an individual's genetic liability to a trait or disease, calculated according to their genotype profile and relevant genome-wide association study (GWAS) data. While present PRSs typically explain only a small fraction of trait variance, their correlation with the single largest contributor to phenotypic variation—genetic liability—has led to the routine application of PRSs across biomedical research. Among a range of applications, PRSs are exploited to assess shared etiology between phenotypes, to evaluate the clinical utility of genetic data for complex disease and as part of experimental studies in which, for example, experiments are performed that compare outcomes (e.g., gene expression and cellular response to treatment) between individuals with low and high PRS values. As GWAS sample sizes increase and PRSs become more powerful, PRSs are set to play a key role in research and stratified medicine. However, despite the importance and growing application of PRSs, there are limited guidelines for performing PRS analyses, which can lead to inconsistency between studies and misinterpretation of results. Here, we provide detailed guidelines for performing and interpreting PRS analyses. We outline standard quality control steps, discuss different methods for the calculation of PRSs, provide an introductory online tutorial, highlight common misconceptions relating to PRS results, offer recommendations for best practice and discuss future challenges.**

**Introduction**

Genome-wide association studies (GWASs) have identified a large number of genetic variants, mostly single nucleotide polymorphisms (SNPs), significantly associated with a wide range of complex traits[1–3]. However, these variants typically have a small effect and correspond to a small fraction of truly associated variants, meaning that they have limited predictive power[4–6]. Using a linear mixed model in the genome-wide complex trait analysis software[7], Yang et al. demonstrated that much of the heritability of height can be explained by evaluating the effects of all SNPs simultaneously[4]. Subsequently, statistical techniques such as linkage disequilibrium (LD) score regression[8,9] and the polygenic risk score (PRS) method[5,10] have also aggregated the effects of variants across the genome to estimate heritability, to infer genetic overlap between traits and to predict phenotypes based on genetic profile[5,6,8–10].

While genome-wide complex trait analysis, LD score regression and PRS can all be exploited to infer heritability and shared etiology among complex traits, PRS is the only approach that provides an estimate of genetic liability to a trait at the individual level. In the **classic PRS method**[5,11–14] (terms in boldface are defined in Box 1), a polygenic risk score is calculated by computing the sum of **risk alleles** that an individual has, weighted by the risk allele effect sizes as estimated by a GWAS on the phenotype. Studies have shown that substantially greater predictive power can usually be achieved by including a large number of SNPs in the PRS rather than restricting to only those reaching genome-wide significance in the GWAS[11,15,16]. As an individual-level proxy of genetic liability to a trait, PRSs are suitable for a range of applications. For example, as well as identifying shared etiology among traits, PRSs have been used to test for genome-wide gene-by-environment and gene-by-gene interactions[15,17], to perform Mendelian randomization studies to infer causal relationships and for patient stratification and sub-phenotyping[15,16,18]. Thus, while polygenic scores represent individual genetic predictions of phenotypes, prediction is often not the end objective: instead, these predictions are commonly aggregated across samples and used for research purposes, interrogating hypotheses via association testing.

Despite the popularity of PRSs, there are minimal guidelines[12] on how best to perform and interpret PRS analyses. Here, we provide a guide to performing PRS analyses, outlining the standard quality control steps required, options for PRS calculation and testing and interpretation of results. We also outline some of the challenges in PRS analyses and highlight common misconceptions in their interpretation. We will not perform a comparison of the power of different PRS methods or provide an overview of PRS applications, since these are available elsewhere[12,14,19,20]. Instead, we focus this article on the issues relevant to PRS analyses irrespective of the method used or the application, so that researchers have a starting point and reference guide for performing polygenic score analyses.

[1]MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK. [2]Department of Genetics and Genomic Sciences, Icahn School of Medicine, Mount Sinai, New York, NY, USA. [3]Centre of Genomic Sciences, University of Hong Kong, Hong Kong, China. ✉e-mail: paul.oreilly@mssm.edu



Choi, S.W. *et al. Nat Protoc* **15,** 2759–2772 (2020)

# Appendix

# Project suggestions

- Compare different methods for calculating the PRS (LASSO, mixed linear models).
- Investigate different quality control methods (significance threshold, corrections for relatedness, linkage disequilibrium).
- Use different phenotypes for performing the GWAS analysis (Which phenotype leads to most accurate PRS?).
- Determine the correlation between a given phenotype and the corresponding PRS (How well do they correlate? Does combining them improve disease prediction?).
- Look into different diseases (Are some diseases more strongly influenced by genetic contributions? Are some more polygenetic?).