THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

CREATE CHANGE

Head in the clouds Head in the sand

Source: My undergrad text book: Strachan & Read Human Molecular Genetics 3.

Polygenic Risk Scores: The Basics

Jian Zeng

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Slides credit: Naomi Wray

Acknowledgement of Country

- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

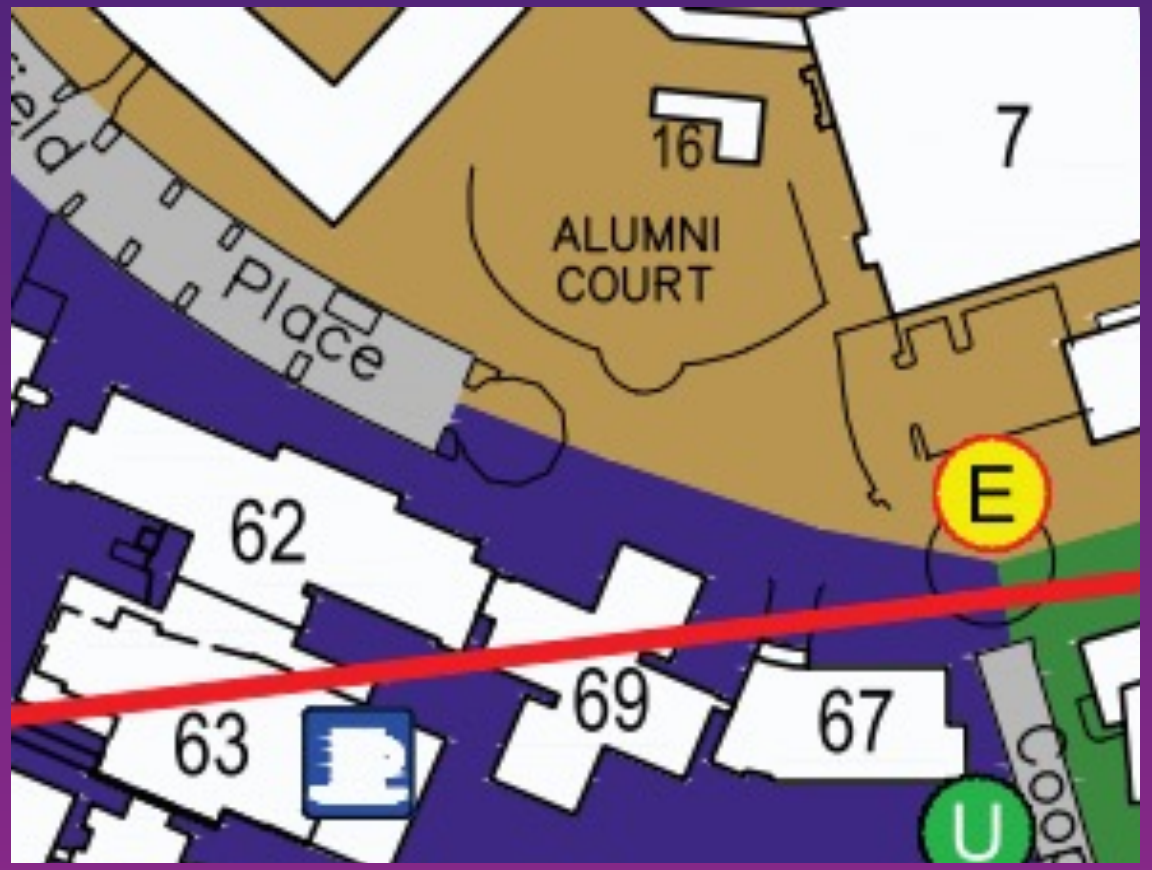
General Information:

- We are currently located in Building 69

Emergency evacuation point

- Food court and bathrooms are located in Building 63

- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module



Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

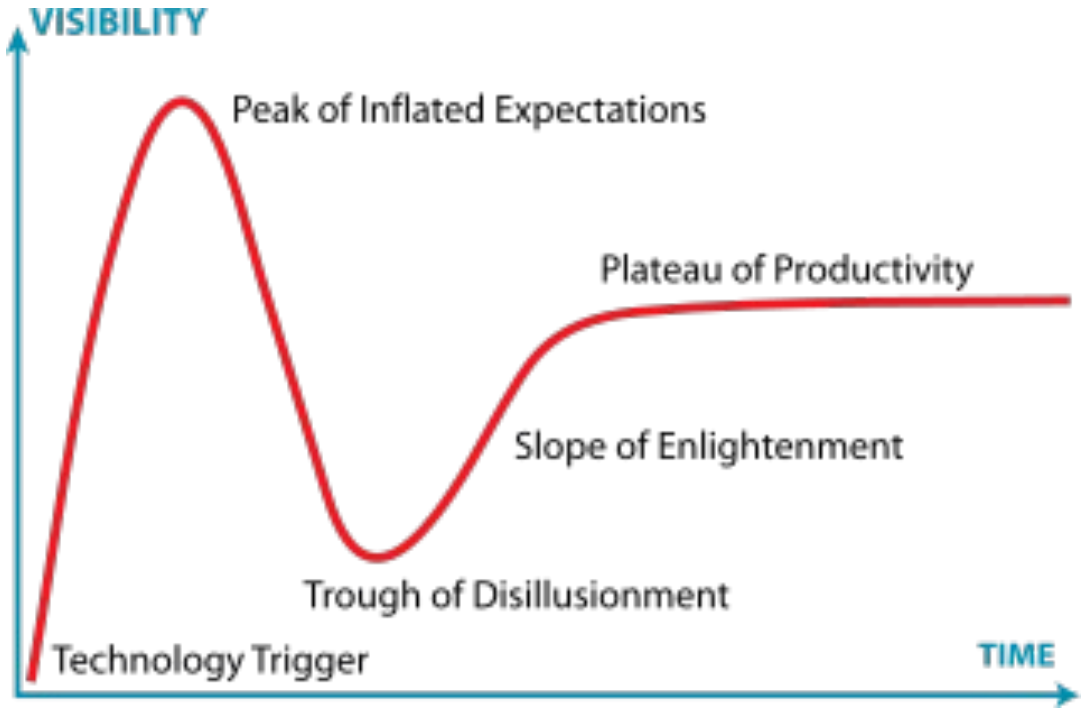
Please email pctgadmin@imb.com.au with your name and the below statement to confirm that you agree with the following:

“I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

Polygenic risk scores (PRS)

Polygenic risk scores (PRS) are predictors of the genetic susceptibilities of individuals to diseases.

Can be calculated for a wide range of diseases from a saliva or blood sample using genotyping technologies that are inexpensive.



Polygenic risk scores (PRS)

Polygenic risk scores (PRS) are predictors of the genetic susceptibilities of individuals to diseases.

Can be calculated for a wide range of diseases from a saliva or blood sample using genotyping technologies that are inexpensive.

Head in the clouds Head in the sand

Source: My undergrad text book: Strachan & Read Human Molecular Genetics 3.

Learning objectives

- Understand what PRS are and what they are not
- How to evaluate PRS and what the pitfalls are in application
- Understand the basic method to calculate PRS
- Get to know more advanced methods in common usage
- Discuss challenges, opportunities and future directions
- Know how to generate a PRS from start to end

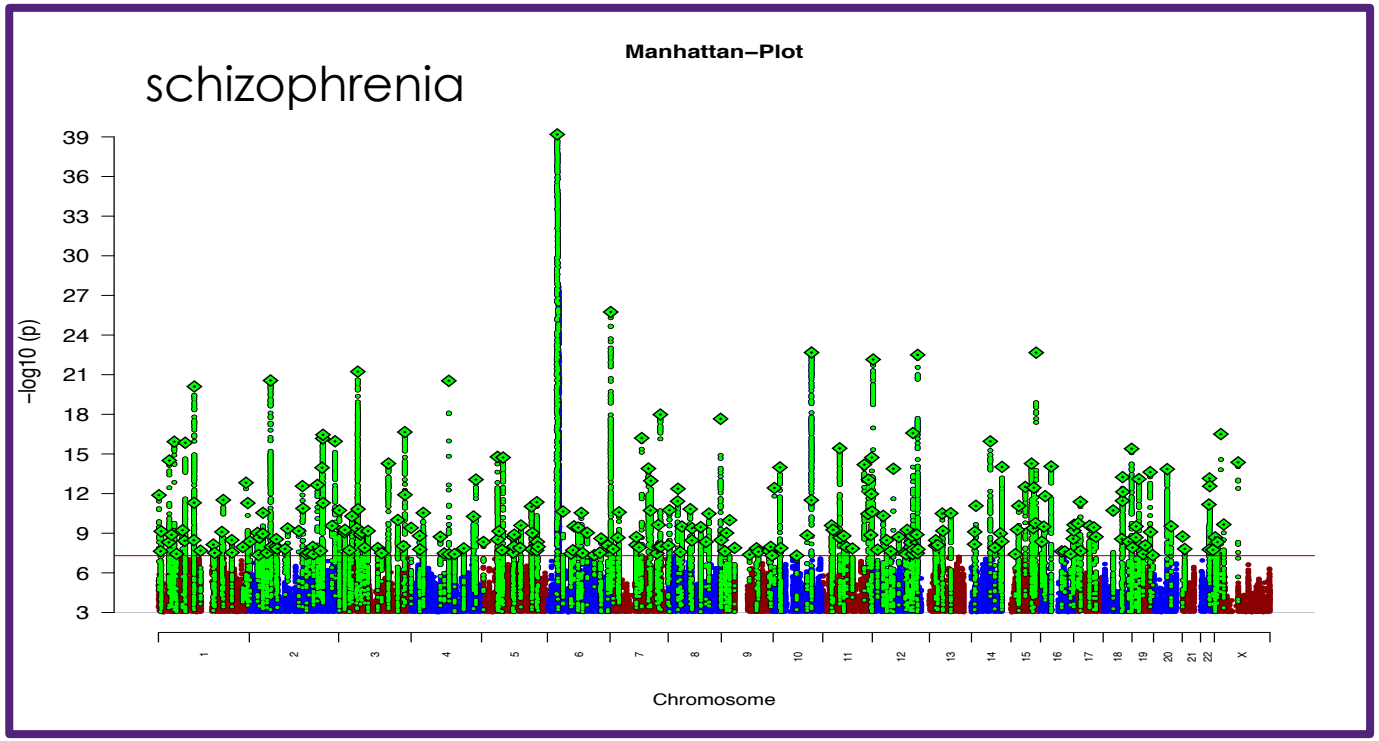
Module materials at

https://cnsgenomics.com/data/teaching/GNGWS23/module5/

Module Program

	Lecture	Practical
This afternoon	Basic science of PRS	Basic method to compute PRS (C+PT)
	Evaluations of PRS and pitfalls in application	Calculation of prediction accuracy; winner's curse
Tomorrow morning	Best Linear Unbiased Prediction (BLUP)	How to do BLUP using R and GCTA
	Bayesian methods to predict PRS	How to do BayesR using R and GCTB
Tomorrow afternoon	PRS prediction using summary data	How to do SBayesR using R and GCTB
	Trans-ancestry prediction; Improved PRS using functional annotations; Complete PRS pipeline	Our in-house PRS pipeline: from lab data to personal scores

Common diseases are polygenic



nature

Explore content | About the journal | Publish with us

nature > articles > article

Article | Published: 08 April 2022

Mapping genomic loci implicates genes and synaptic biology in schizophrenia

Vassily Trubetskoy, Antonio F. Pardiñas, Ting Qi, Georgia Panagiotaropoulou, Swapnil Awasthi, Tim B. Bigdeli, Julien Bryois, Chia-Yen Chen, Charlotte A. Dennison, Lynsey S. Hall, Max Lam, Kyoko Watanabe, Oleksandr Frei, Tian Ge, Janet C. Harwood, Frank Koopmans, Sigurdur Magnusson, Alexander L. Richards, Julia Sidorenko, Yang Wu, Jian Zeng, Jakob Grove, Minsoo Kim, Zhiqiang Li, Indonesia Schizophrenia Consortium, PsychENCODE, Psychosis Endophenotypes International Consortium, The SynGO Consortium, Schizophrenia Working Group of the Psychiatric Genomics Consortium

+ Show authors

Nature **604**, 502–508 (2022) | Cite this article

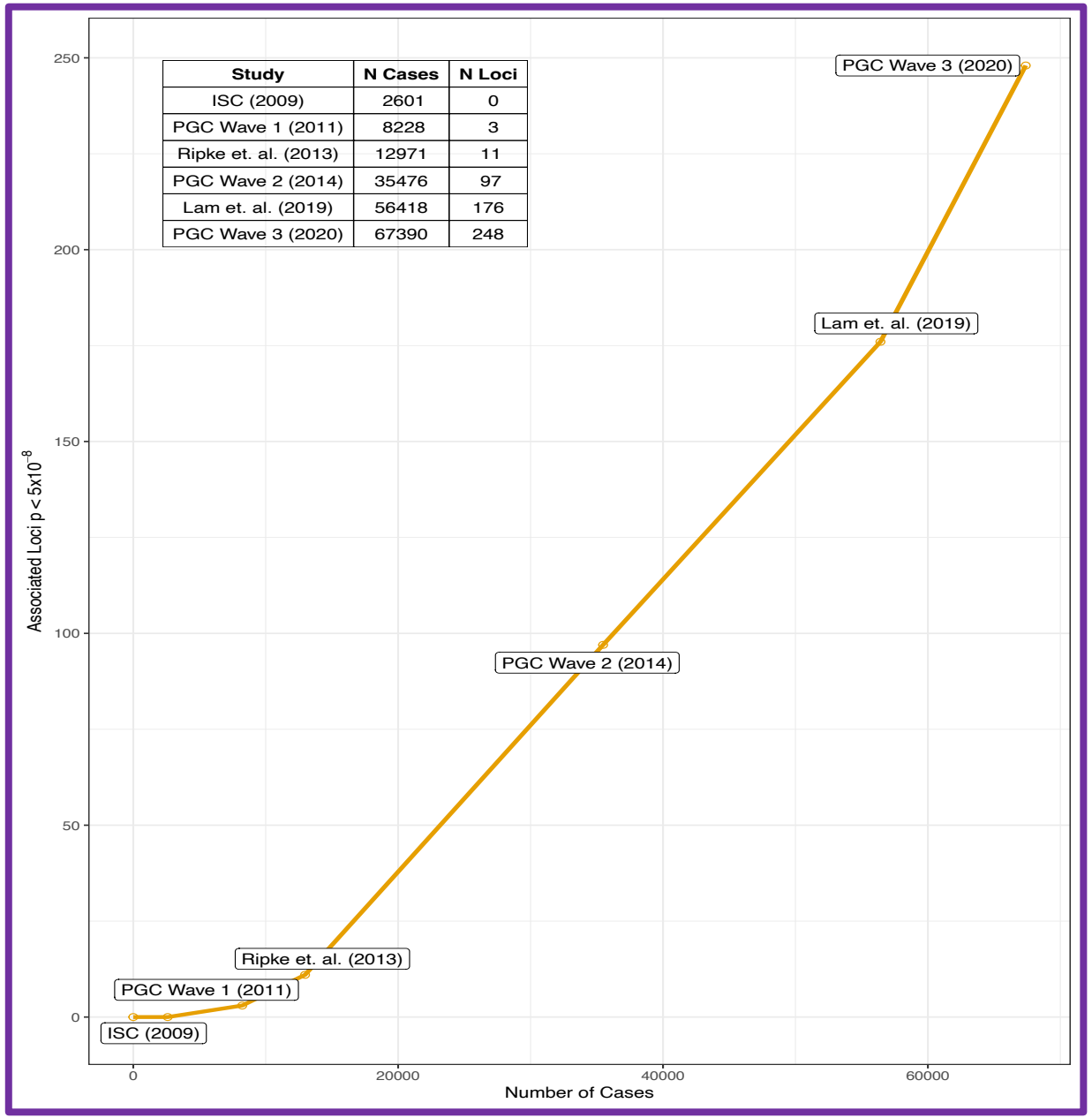
57k Accesses | **321** Citations | **463** Altmetric | Metrics

248 risk loci identified at genome-wide significance level.

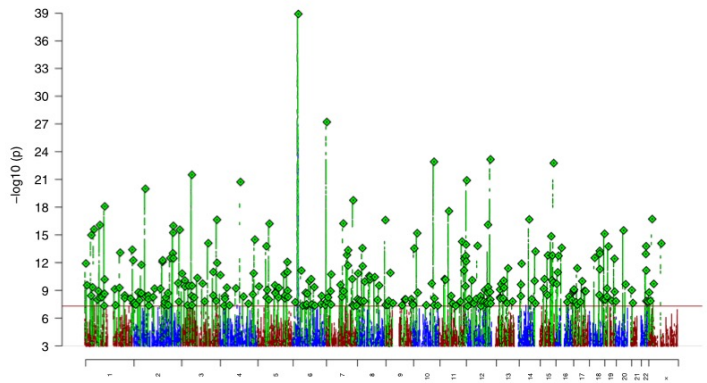
We predict thousands are associated with schizophrenia.

PGC

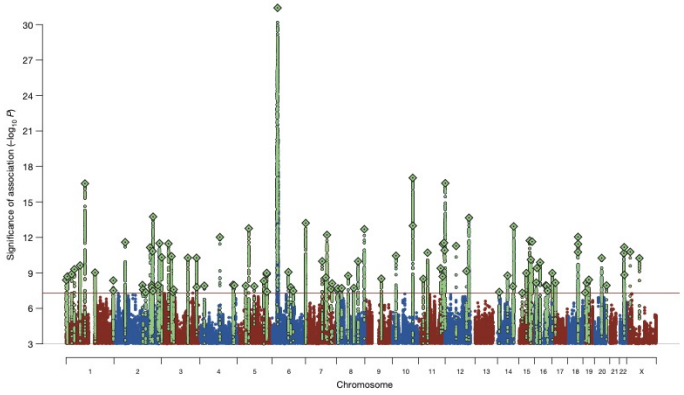
Common diseases are polygenic



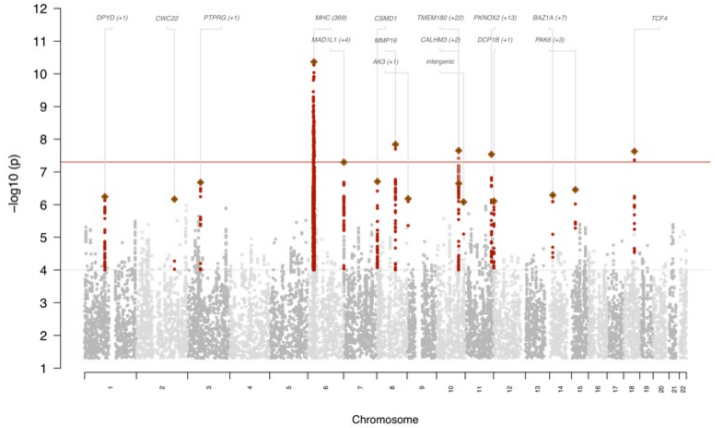
Study	N Cases	N Loci
ISC (2009)	2601	0
PGC Wave 1 (2011)	8228	3
Ripke et. al. (2013)	12971	11
PGC Wave 2 (2014)	35476	97
Lam et. al. (2019)	56418	176
PGC Wave 3 (2020)	67390	248



2022 PGC Wave 3

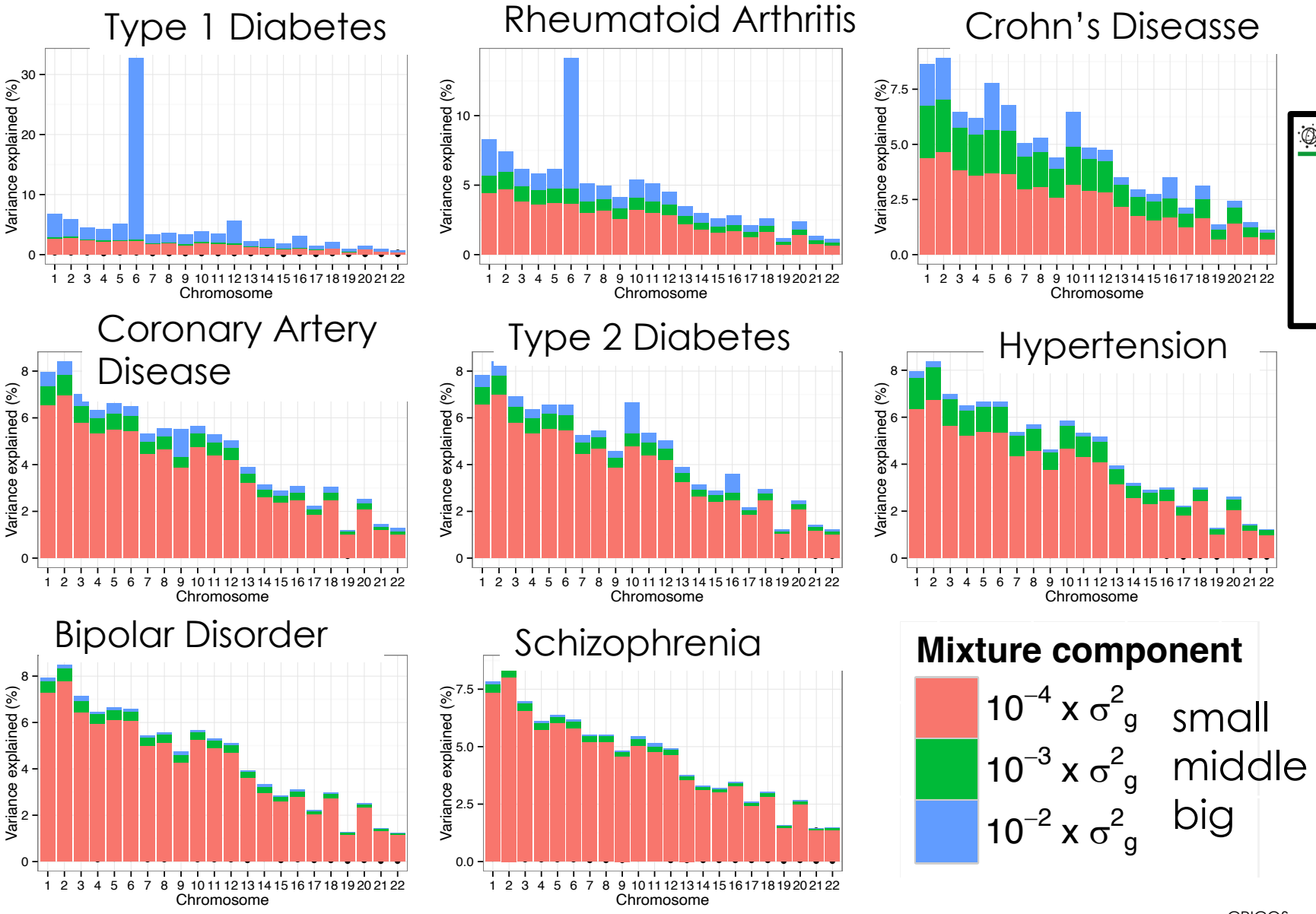


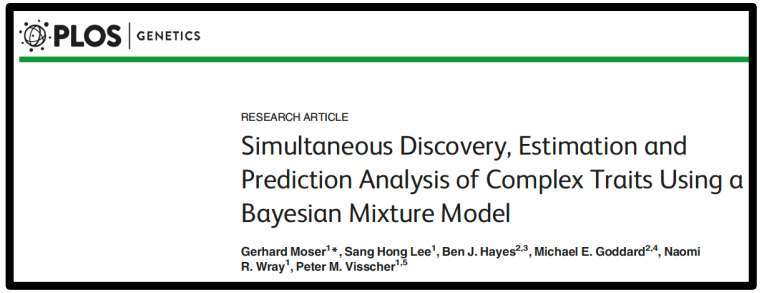
2014 PGC Wave 2



2011 PGC Wave 1

Many polygenic genetic architectures




PLOS | GENETICS

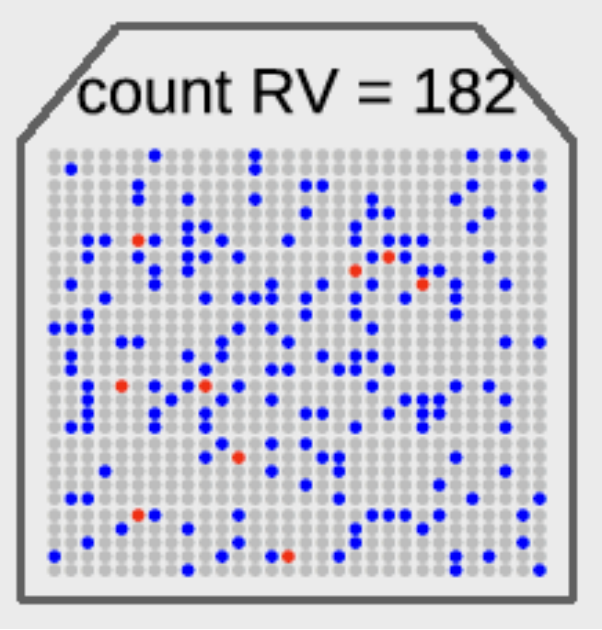
RESEARCH ARTICLE

Simultaneous Discovery, Estimation and Prediction Analysis of Complex Traits Using a Bayesian Mixture Model

Gerhard Moser[1]*, Sang Hong Lee[1], Ben J. Hayes[2,3], Michael E. Goddard[2,4], Naomi R. Wray[1], Peter M. Visscher[1,5]

Many DNA variants contribute to genetic risk, and most have very small effects.

Polygenic disease for an individual



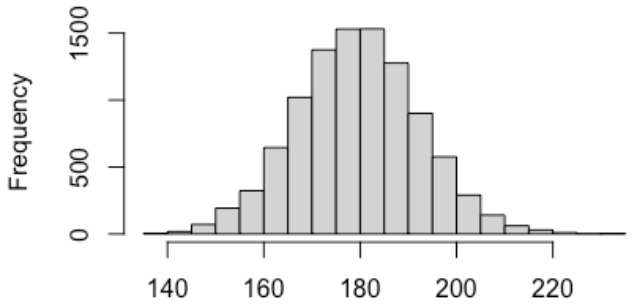
900 DNA polymorphic sites

RV =risk variant

Frequency of risk variant at each site: 0.1 (p)

Average person 900*2*0.1 = 180 risk variant

Mean +/- 3SD: 142 to 218



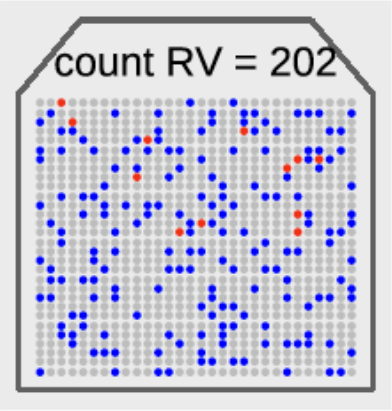
Count of RV in population

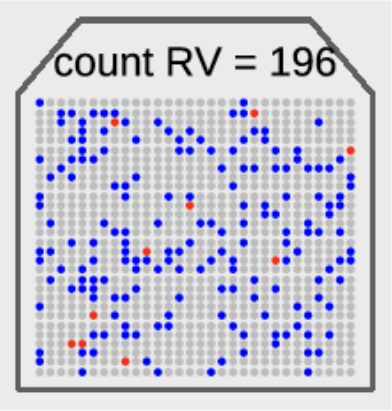
0 Grey: Homozygote no risk alleles (or equivalently 2 protective alleles)
1 Blue : Heterozygote one risk allele (and one non-risk/protective allele)
2 Red: Homozygote two risk alleles

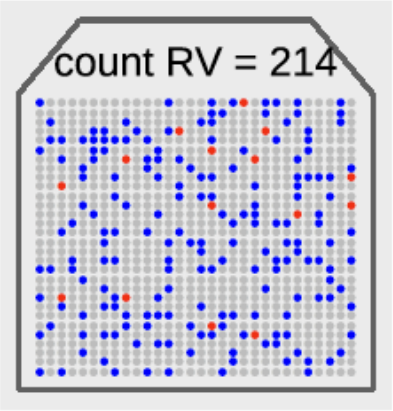
Toy example

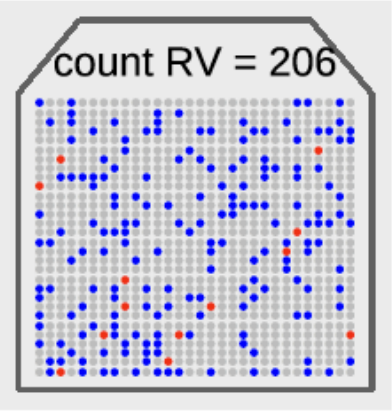
Polygenic disease for an individual

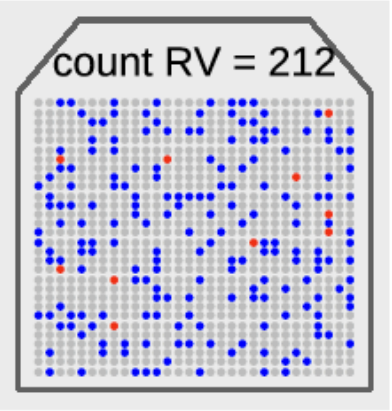
Affected over lifetime



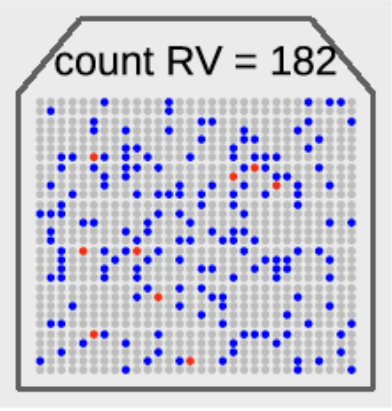


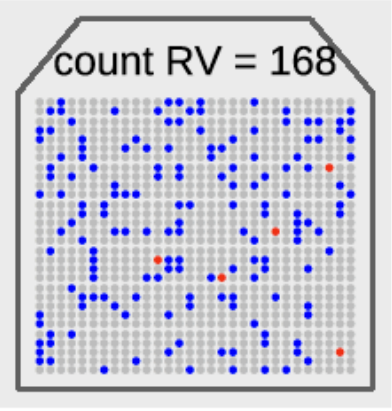


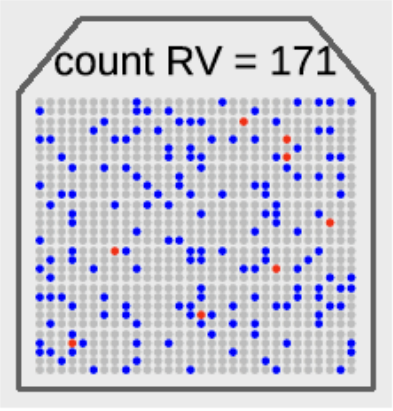


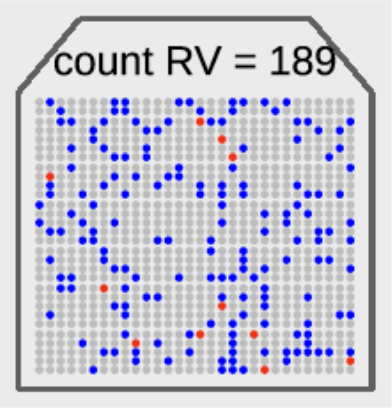


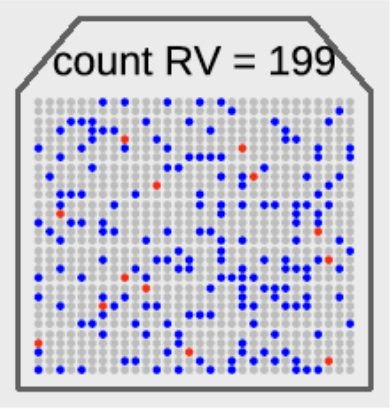
Not affected over lifetime





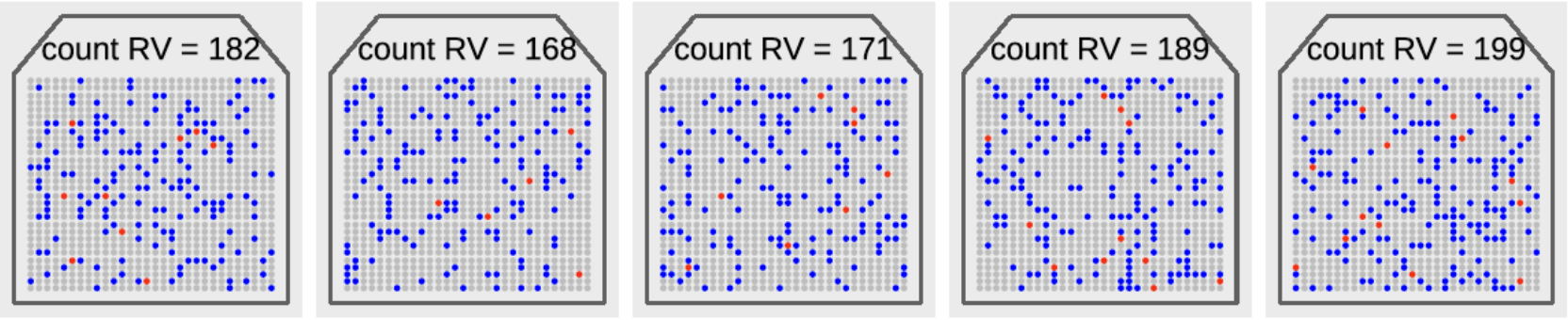






- We all carry risk variants for all diseases.
- Robustness
- Those affected carry a higher burden.
- Non-genetic factors contribute to risk too
- Each person carries a unique portfolio of risk alleles

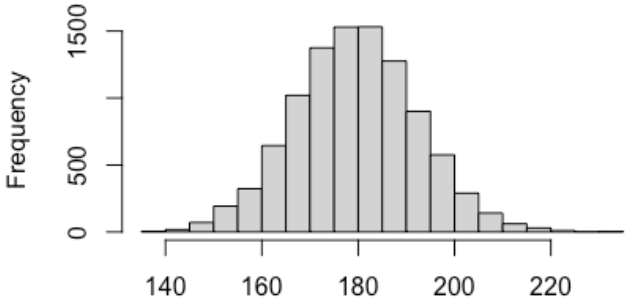
Polygenic score



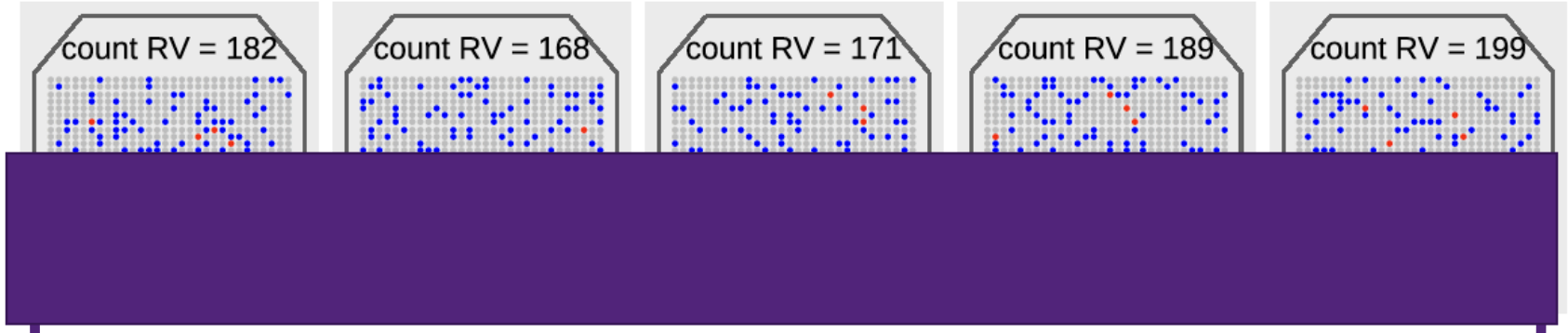
"True" polygenic score

Genetic variance between people attributed to all genetic factors V(A)

$$h^2 = \frac{V(A)}{V(P)}$$ heritability



Polygenic score



"True" polygenic score

Not all variants captured on genotyping arrays

Genetic variance between people attributed to all genetic factors V(A)

$$h^2 = \frac{V(A)}{V(P)}$$ heritability

$$h^2_{SNP} = h^2_g = \frac{V(A{:}SNP)}{V(P)}$$

SNP – based heritability

Genetic variance between people attributed to all genetic factors associated with SNPs on genotyping arrays

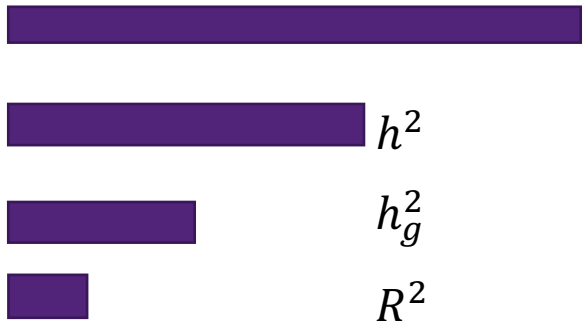
Limitations in prediction accuracy

- PRS have a **theoretical** upper limit dependent on the **heritability of the trait** (how much of the variance of trait values between people is attributed to genetic factors).

- PRS have a **technical** upper limit associated with the proportion of **variance tagged** by the DNA variants measured.

- PRS have a **practical** upper limit dependent on the **sample size of the discovery sample** used to estimate effect sizes of risk alleles, and the **quality** of the discovery sample.

- PRS can be pushed closer to the technical upper limit by the **statistical methodology** used to generate the optimal weighting given to the risk alleles, and new methods integrate new biological data.



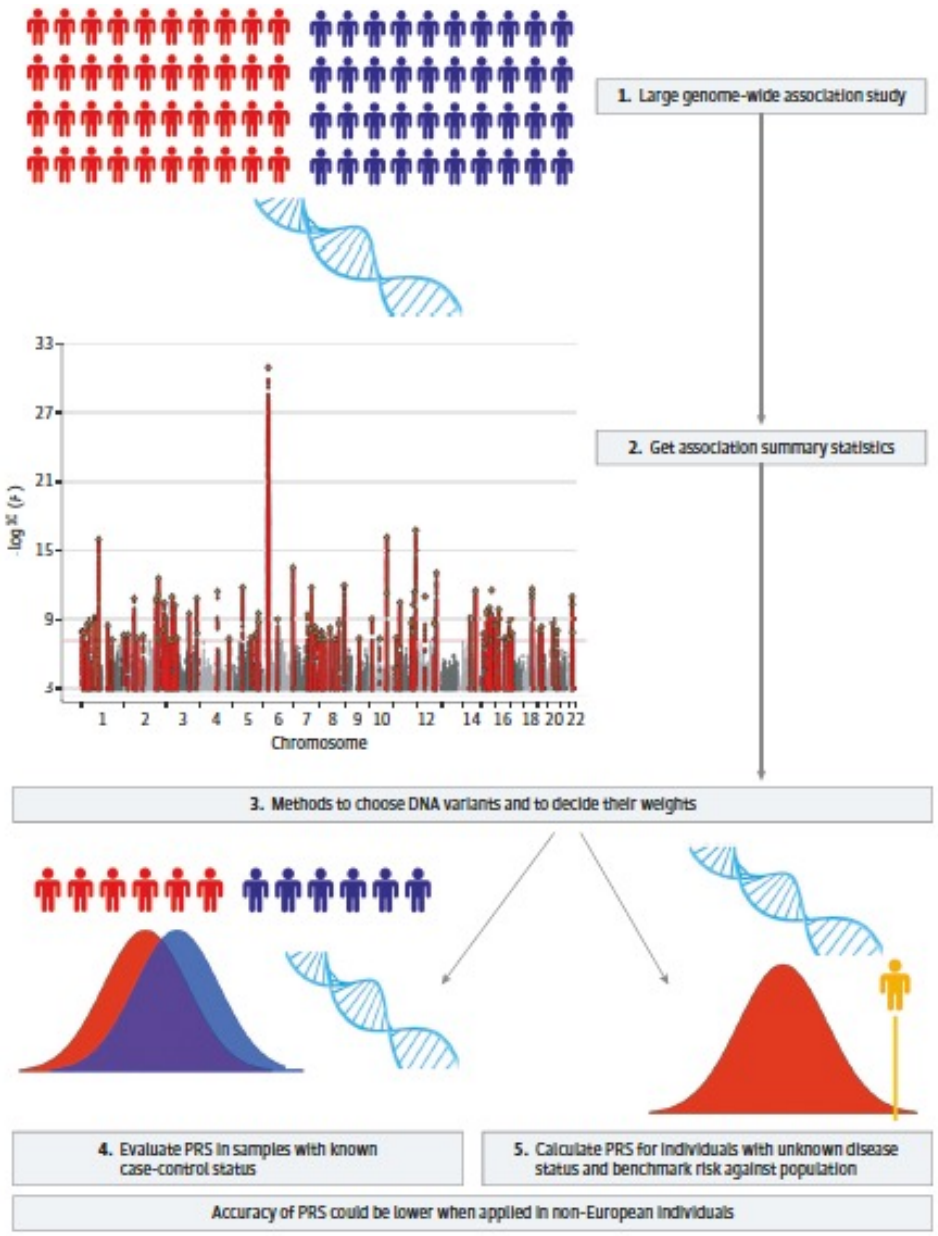
Schizophrenia

Max:
25% Liability
AUC 0.84

Current:
11% Liability
AUC 0.74

Polygenic scores cannot be highly accurate predictors of phenotypes

Polygenic risk scores



- A weighted count of risk alleles

$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

0, 1 or 2
Risk alleles

Which SNPs?

What weights?

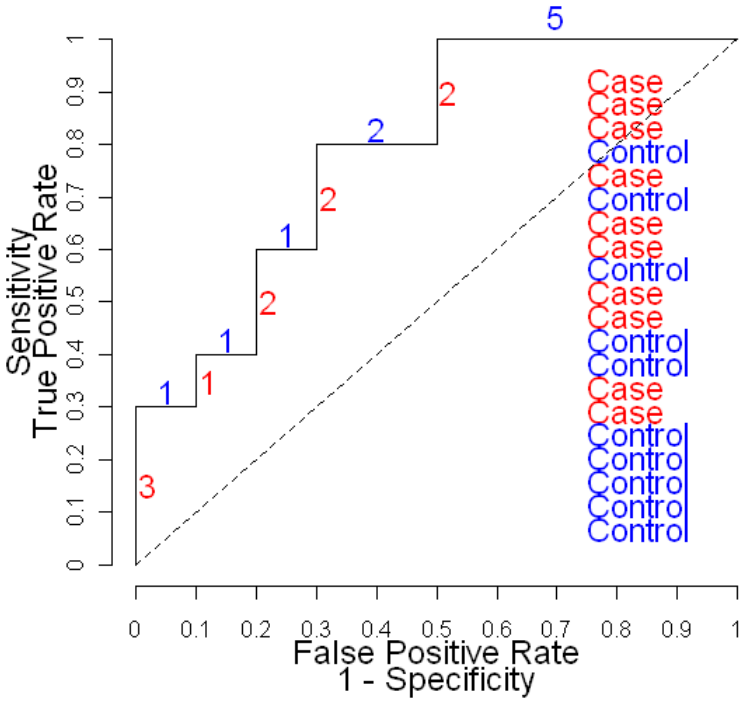
- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.
- Prediction robust to inclusion of false positives

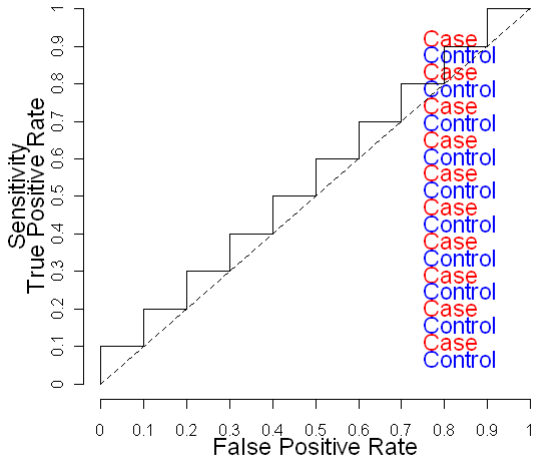
4. Evaluate
Y= b*PRS + e
R^2 = var(b*PRS)/Var(Y)

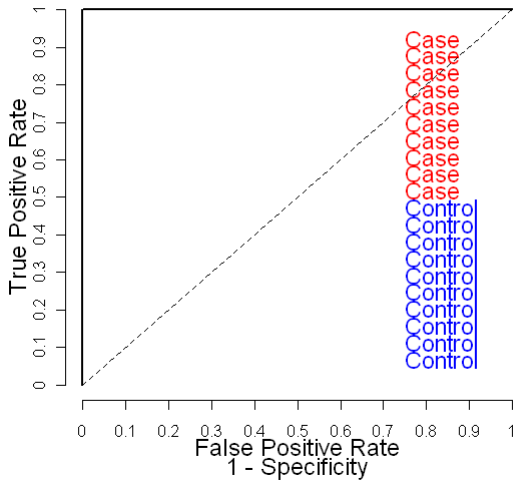
AUC statistic:
Probability that a case ranks higher than a control

Visualising Prediction Accuracy

Evaluation of disease risk prediction: Area Under the ROC Curve

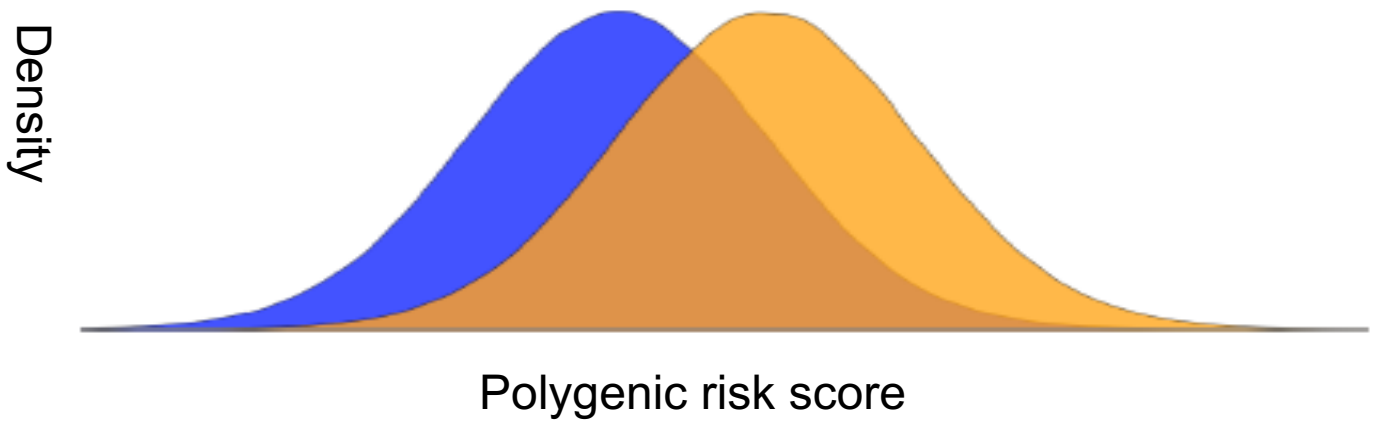


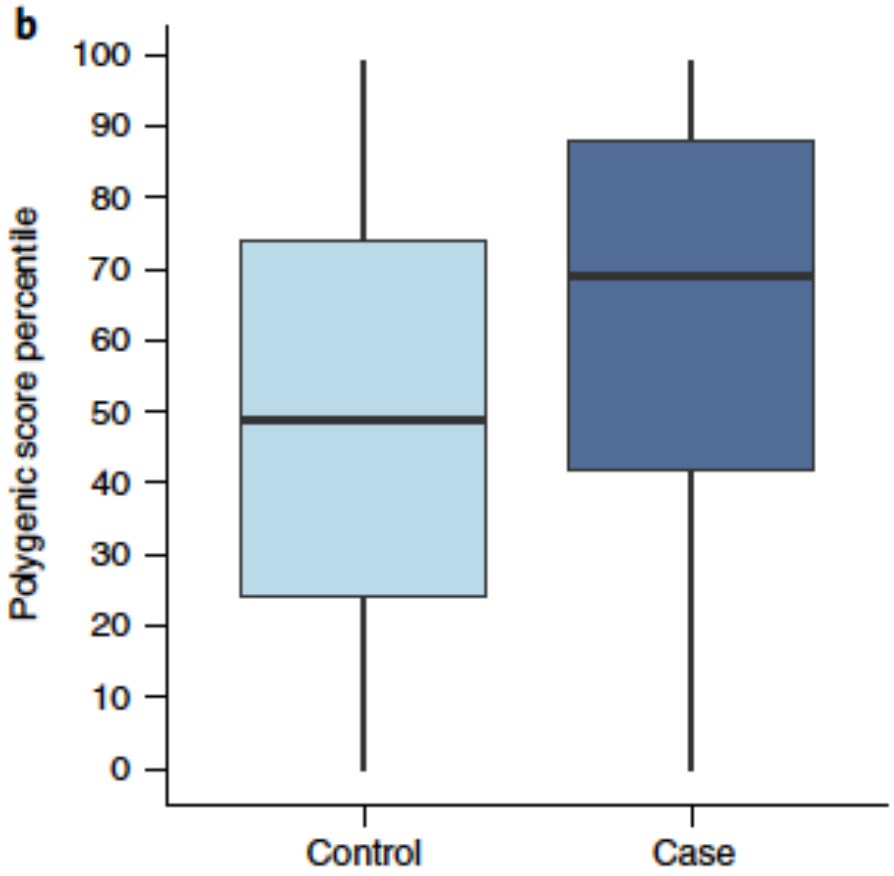




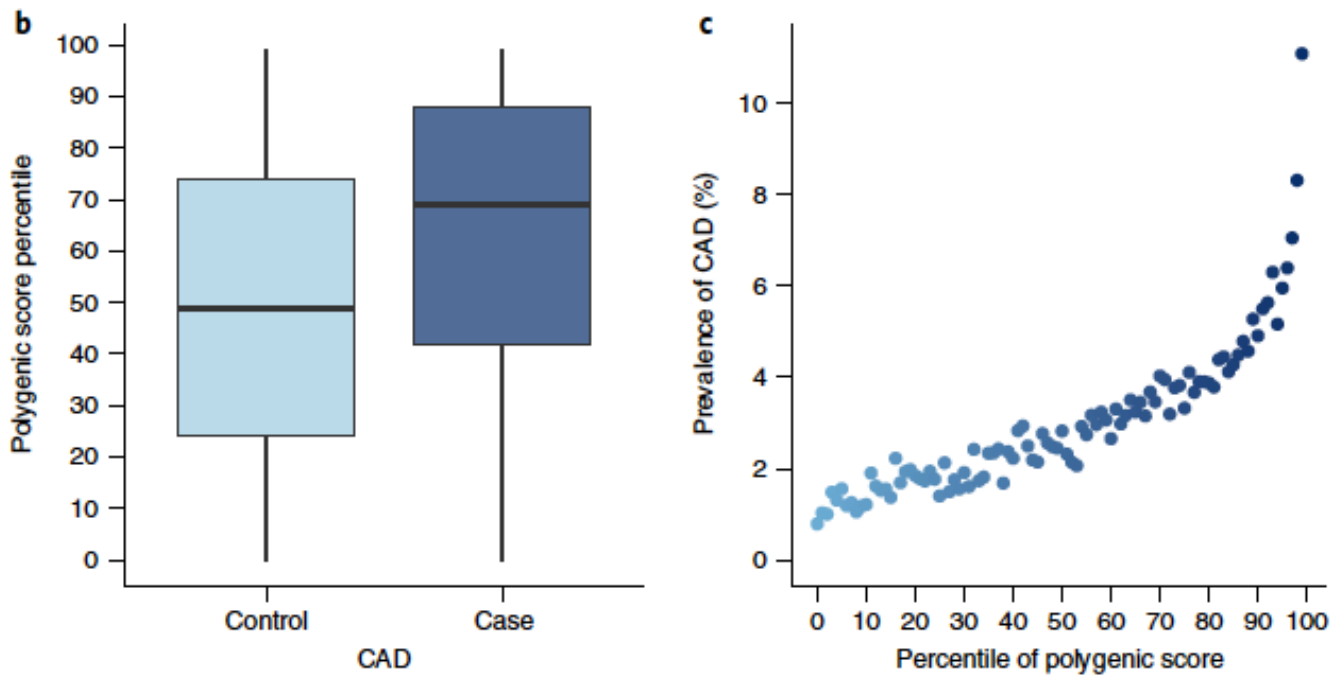
AUC = Probability that a randomly selected case has a higher test score than a randomly selected control

Different views of the same data





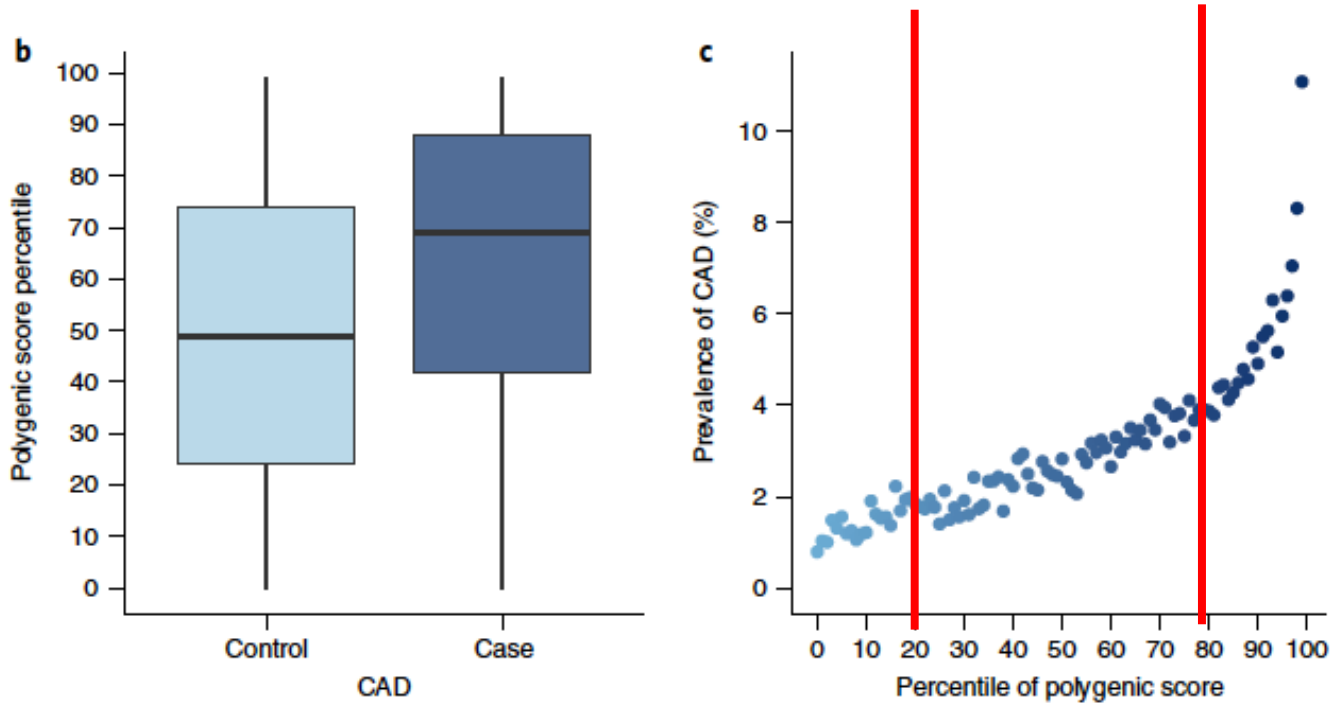
Different views of the same data

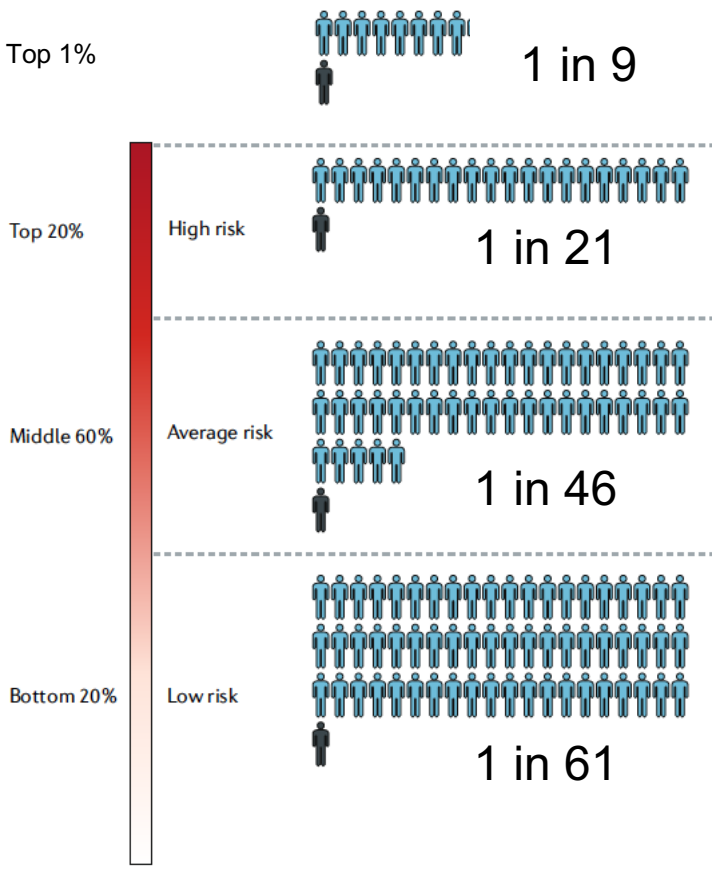


Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Evaluating PRS

Different views of the same data

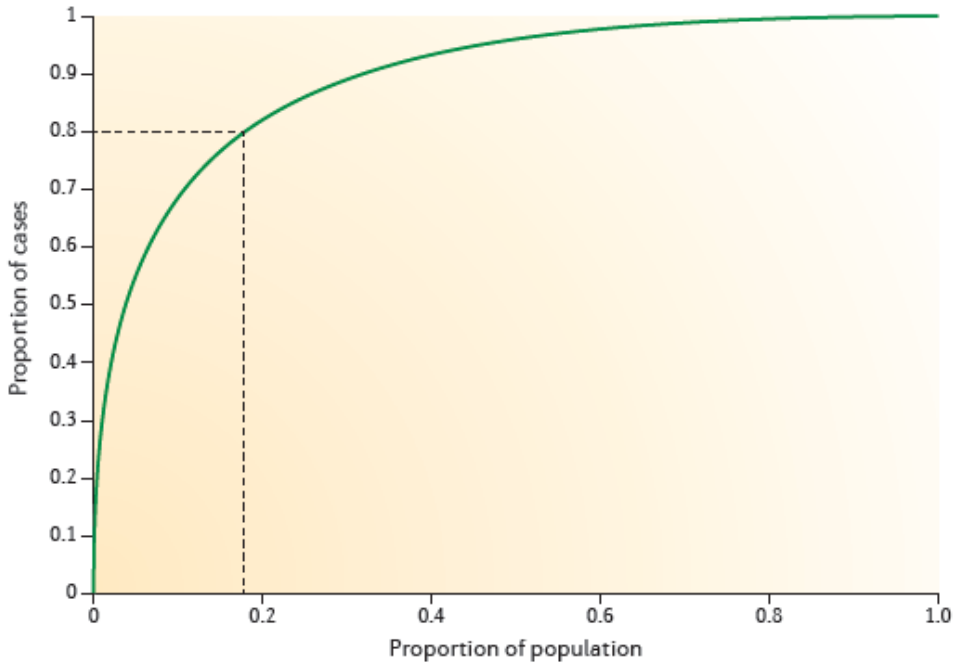




Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Torkamani et al, Nat Rev Genetics, 2018

Stratification & health economics



Population risk of 1%

80% of cases in
top 18% of genetic risk

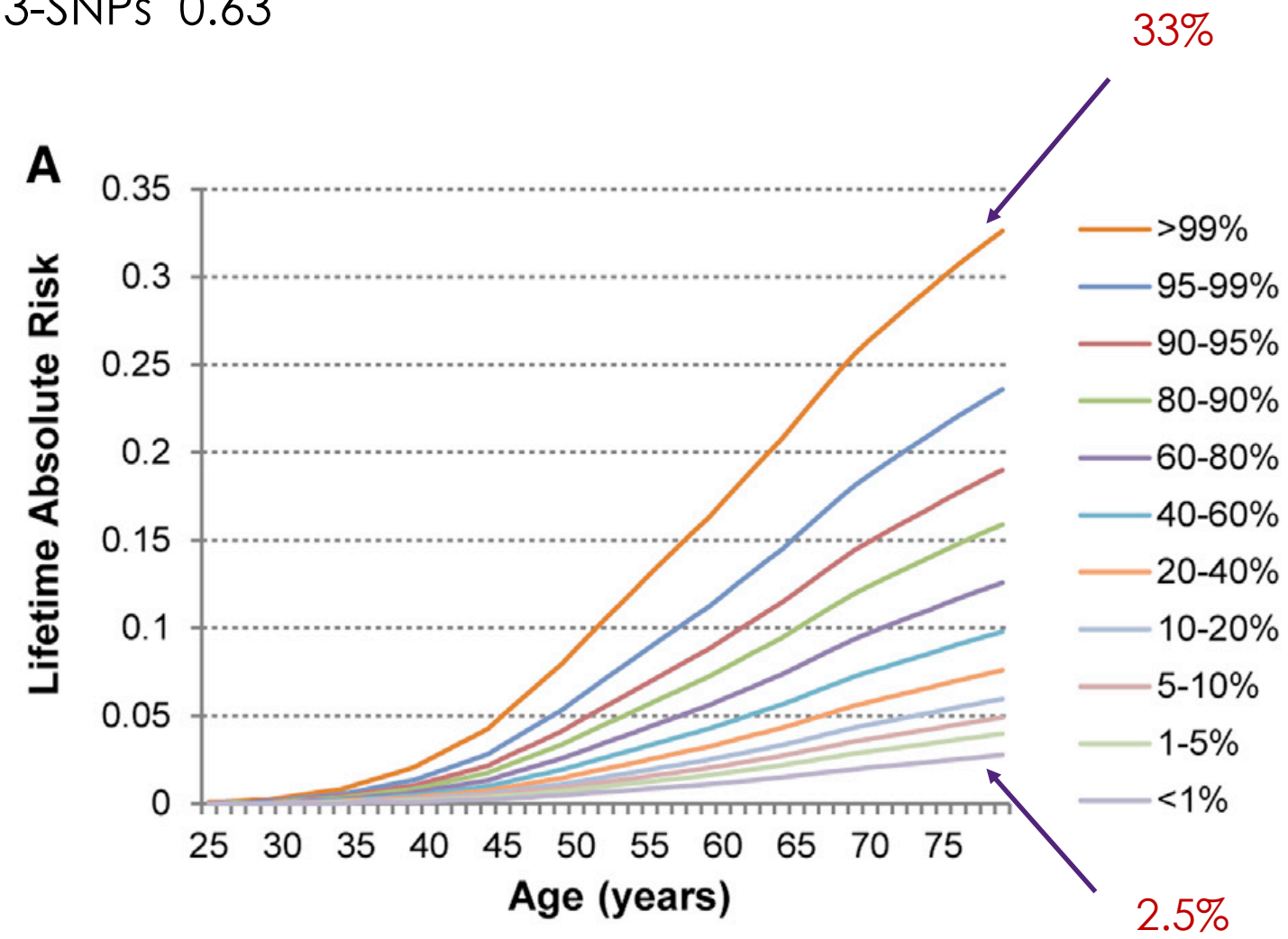
For every 1,000 people treated with intervention could "save" 10
Treat only 18% = 180 and "save" 8

Number of people treated to save 1 reduced from 100 to 22.5

Polychronakos & Li NRG (2011) Understanding Type I Diabetes through genetics. Nat Rev Genetics

Breast Cancer

AUC 313-SNPs 0.63



Mavaddat et al (2019) Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes. AJHG

Polygenic risk score applications

The University Of Queensland
AUSTRALIA

JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

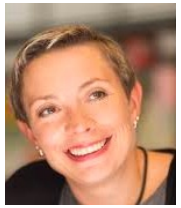
Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD

Goal:

- Understandable by interested clinician
- Technically accurate – backed up in Supplement & Rscript

Ian Hickie, UoSydney

Graham Murray, UoCambridge



Jehannine Austin, UoBritish Columbia

John McGrath, UQ

Tian Lin, UQ

Cohort where PRS applied:	Community Of 100 people in the population, 1 will get "the disease" in lifetime, assuming a disease of lifetime risk of 1%	Symptoms: help-seeking Of 100 people presenting at clinic with symptoms but without a clear diagnosis, a higher proportion than in a population sample will go on to get "the disease" in their lifetime	Established diagnosis 100 people with diagnosis of "the disease"
Utility of PRS:	PRS contribute to risk stratification Of 100 people in the top PRS stratum, a higher proportion will get "the disease" in their lifetime and hence are particularly encouraged to enter established disease screening	PRS contribute to clinical decisions Of 100 people presenting with symptoms AND in the top PRS stratum, a higher proportion than in the clinic-presenting cohort will go on to get diagnosis of "the disease" in their lifetime	PRS contribute to treatment choices Genetic information may contribute to more effective choice of treatment, with reduced adverse events
Likely applications:	Common diseases/ disorders for which there is already population screening	When there is no clear diagnosis based on presenting symptoms, guide monitoring of emergent symptoms	Potentially all common diseases/disorders but little data available to date
Likely first applications:	Cancers: breast and colorectal; common eye disorders: glaucoma, macular degeneration; heart disease	Differentiating between type 1 and type 2 diabetes	Inflammatory bowel disease is a flagship in the genetics of common disease; perhaps we will see first applications here?

Polygenic risk score applications

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

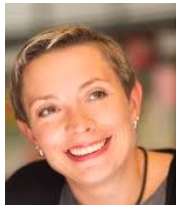
Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD

Goal:

- Understandable by interested clinician
- Technically accurate – backed up in Supplement & Rscript

Ian Hickie, UoSydney

Graham Murray, UoCambridge



Jehannine Austin, UoBritish Columbia

John McGrath, UQ

Tian Lin, UQ

Cohort where PRS applied:	Community	Of 100 people in the population, 1 will get "the disease" in lifetime, assuming a disease of lifetime risk of 1%
Utility of PRS:	PRS contribute to risk stratification	Of 100 people in the top PRS stratum, a higher proportion will get "the disease" in their lifetime and hence are particularly encouraged to enter established disease screening
Likely applications:	Common diseases/ disorders for which there is already population screening	
Likely first applications:	Cancers: breast and colorectal; common eye disorders: glaucoma, macular degeneration; heart disease	

PRS could have utility in community settings (stratification to better triage people into established screening programs)

Polygenic risk score applications

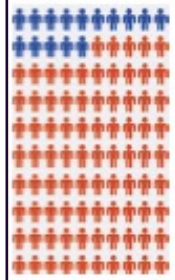
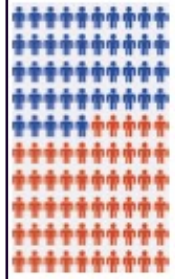
JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD

Goal:

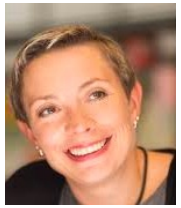
- Understandable by interested clinician
- Technically accurate – backed up in Supplement & Rscript

Cohort where PRS applied:	Symptoms: help-seeking
	 Of 100 people presenting at clinic with symptoms but without a clear diagnosis, a higher proportion than in a population sample will go on to get "the disease" in their lifetime
Utility of PRS:	PRS contribute to clinical decisions
	 Of 100 people presenting with symptoms AND in the top PRS stratum, a higher proportion than in the clinic-presenting cohort will go on to get diagnosis of "the disease" in their lifetime
Likely applications:	When there is no clear diagnosis based on presenting symptoms, guide monitoring of emergent symptoms
Likely first applications:	Differentiating between type 1 and type 2 diabetes

PRS could contribute to clinical decision-making for those presenting with symptoms but where formal diagnosis is unclear.

Ian Hickie, UoSydney

Graham Murray, UoCambridge



Jehannine Austin, UoBritish Columbia

John McGrath, UQ

Tian Lin, UQ

Polygenic risk score applications

JAMA Psychiatry | Review

From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

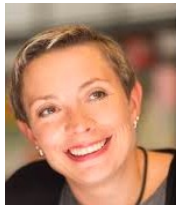
Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD

Goal:

- Understandable by interested clinician
- Technically accurate – backed up in Supplement & Rscript

Ian Hickie, UoSydney

Graham Murray, UoCambridge

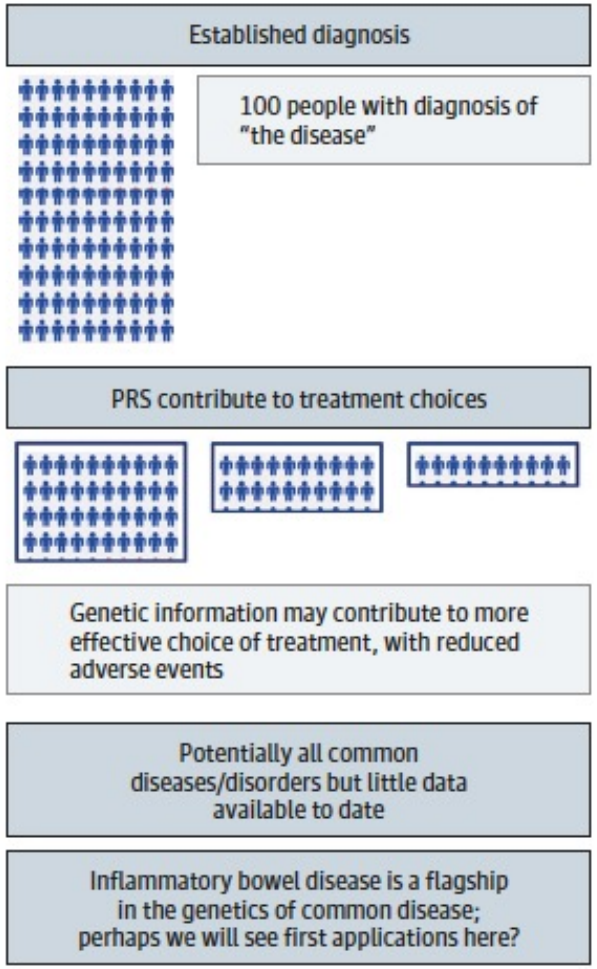


Jehannine Austin, UoBritish Columbia

John McGrath, UQ

Tian Lin, UQ

Cohort where PRS applied:	
Utility of PRS:	PRS could contribute to treatment choices, but more data are needed to allow development of PRS in this context.
Likely applications:	
Likely first applications:	



Australia vs other countries

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

ourfuturehealth.org.uk/participants/

Genome tools | Google Scholar | Cybrary | UQEmail | UniFi | Google | PubMed | NetBank - Logon | ABC News 24 (Au... | PGC | MND

+ Our Future Health

Home | Our programme | About us | Participants | Partners | News & events | Careers

Participants

Taking part in Our Future Health

Our aim is to recruit up to five million people over the age of 18, from all backgrounds and ethnic groups, and from all across the UK to take part. This will make Our Future Health the largest ever health research programme involving members of the public in the UK.

World-leading research to improve health

Our plan is to collect information from millions of people from across the UK, in a giant digital database. Researchers will use this resource to make new discoveries about human health and disease. This could transform the prevention, detection and treatment of conditions such as dementia, cancer, diabetes, heart disease and stroke. So future generations can live in good

NHS X

About us | Key tools and info | COVID-19 response | News

Accelerating Detection of Disease

Accelerating Detection of Disease (ADD) is the UK's largest ever health research programme. By building the most detailed picture we've ever had of the UK's health, it will help detect common diseases earlier and allow more people to live healthier lives for longer.

UK

All of Us RESEARCH PROGRAM

There's a gap in medical research that only you can fill.

The *All of Us* Research Program has a simple mission. We want to speed up health research breakthroughs. To do this, we're asking one million people to share health information. In the future, researchers can use this to conduct thousands of health studies.

US

Australian Government Department of Health

Search consultations

Consultation Hub | Find Consultations | We Asked, You Said, We Did

Medical Research Future Fund's Genomics Health Futures Mission - National Consultation on the Roadmap, Implementation Plan and Summary of Recommendations from the Scientific Strategy Committee Report

Overview

The Medical Research Future Fund's (MRFF) Genomics Health Futures Mission (the Mission) was announced as part of the 2018-19 budget to provide $500 million for research to deliver better testing, diagnosis and treatment.

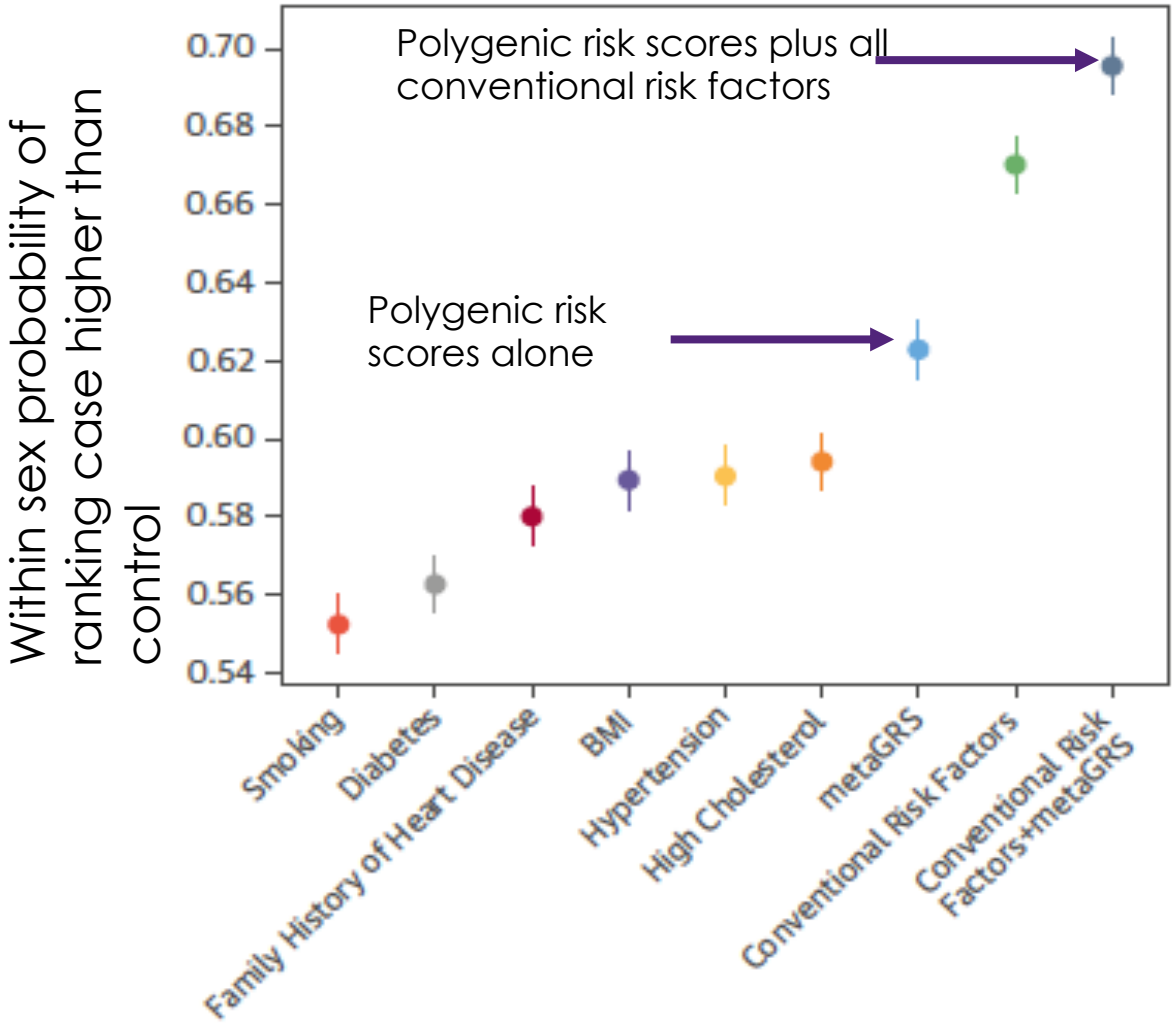
Closes 23 Apr 2021

Opened 14 Dec 2020

Finland, Estonia, ……

Increase prediction accuracy….

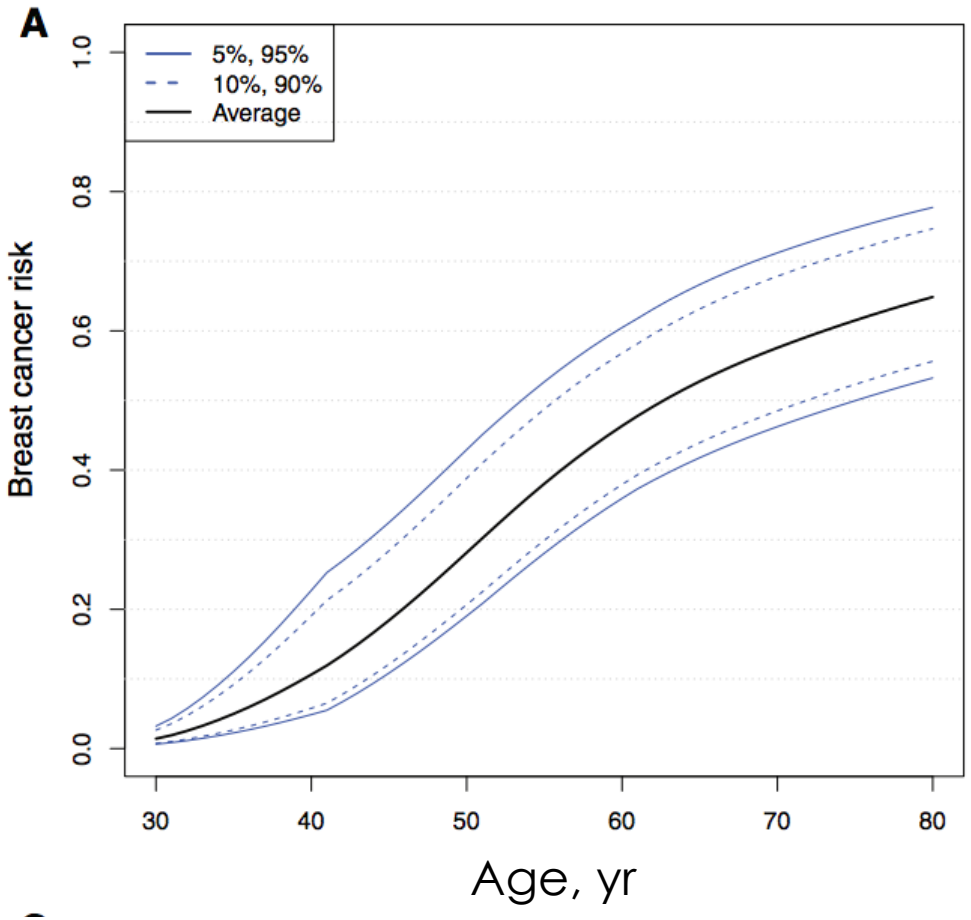
Combine PRS with conventional risk predictors
Coronary Artery Disease



Inouye et al (2018) Genomic risk prediction of CAD in 480K adults. JACC

Increase prediction accuracy….

Combine PRS with known risk mutations
Breast cancer



BRCA1 carriers

Kuchenbaecker et al: Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. J Natl Cancer Inst (2017)

Family history

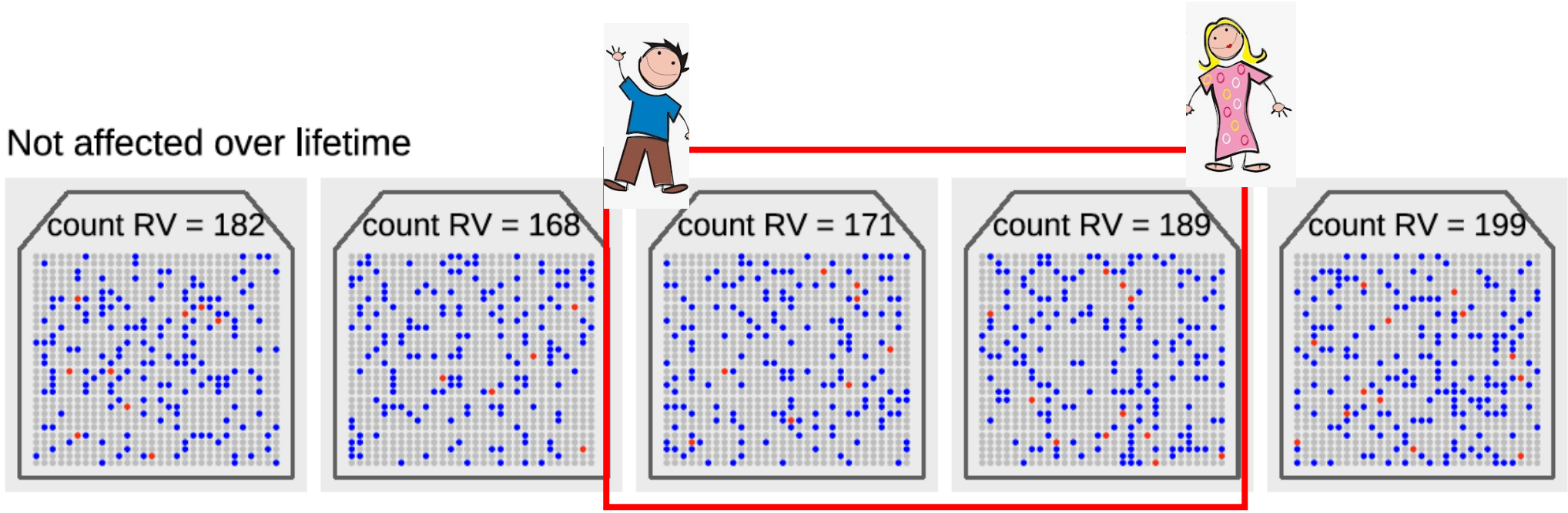
Will people withOUT known family history have high PRS?

Maybe, and that's important!

JAMA Psychiatry | Review

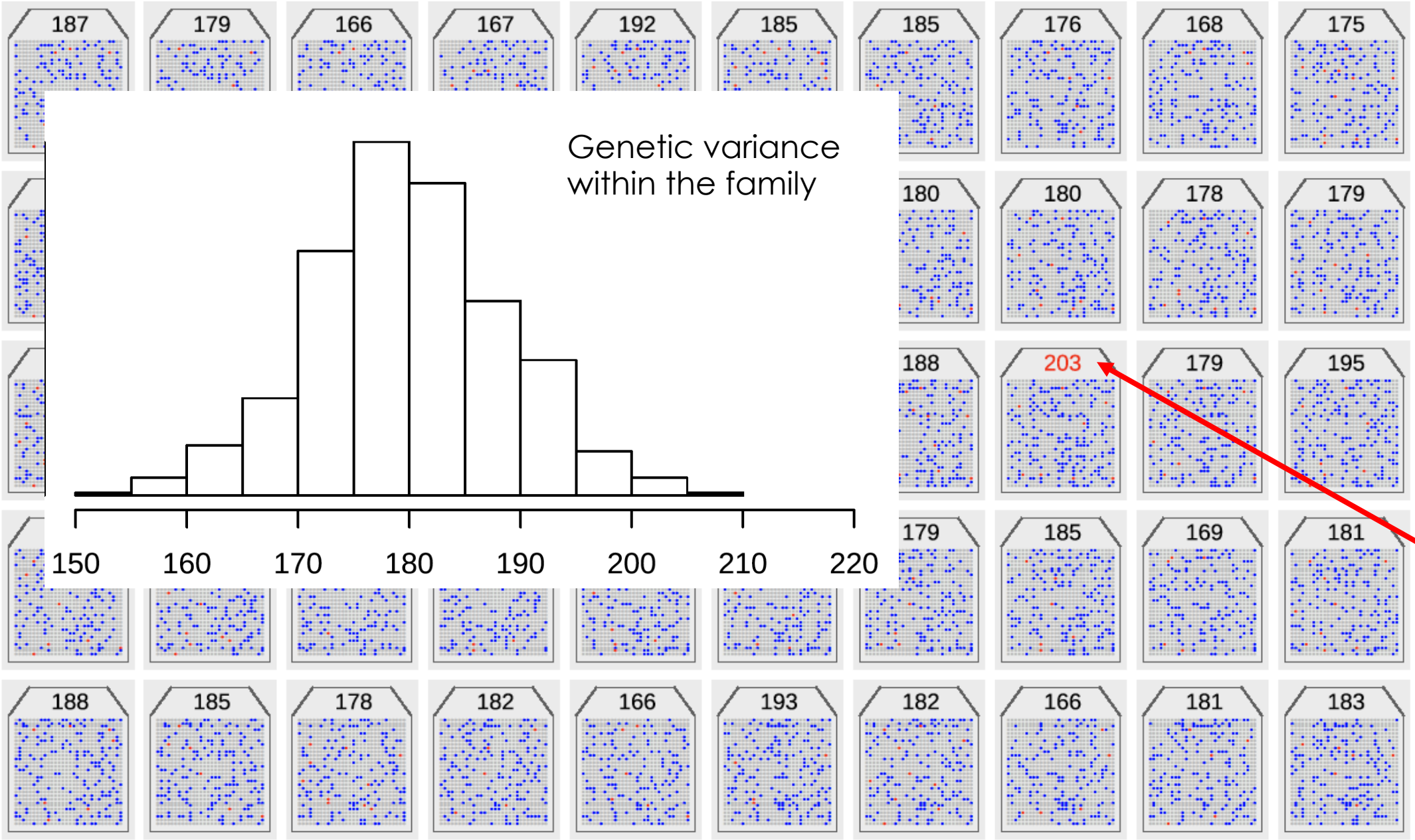
From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD



Grey: Homozygote: Two non-risk/protective alleles – always passes a non-risk allele to child at the locus
Red: Homozygote: Two risk alleles – always passes a risk allele to child at the locus
Blue: Heterozygotes: One risk allele & one non-risk allele –
passes a risk allele 50% of the time & a non-risk allele 50% of the time

Children (Parents: 171 & 189)



Children of these parents
Mean: 180
+/-3SD: 153-207

Population
Mean: 180
+/-3SD: 142-218

No family history, but by chance segregation of alleles has high genetic risk

Family history

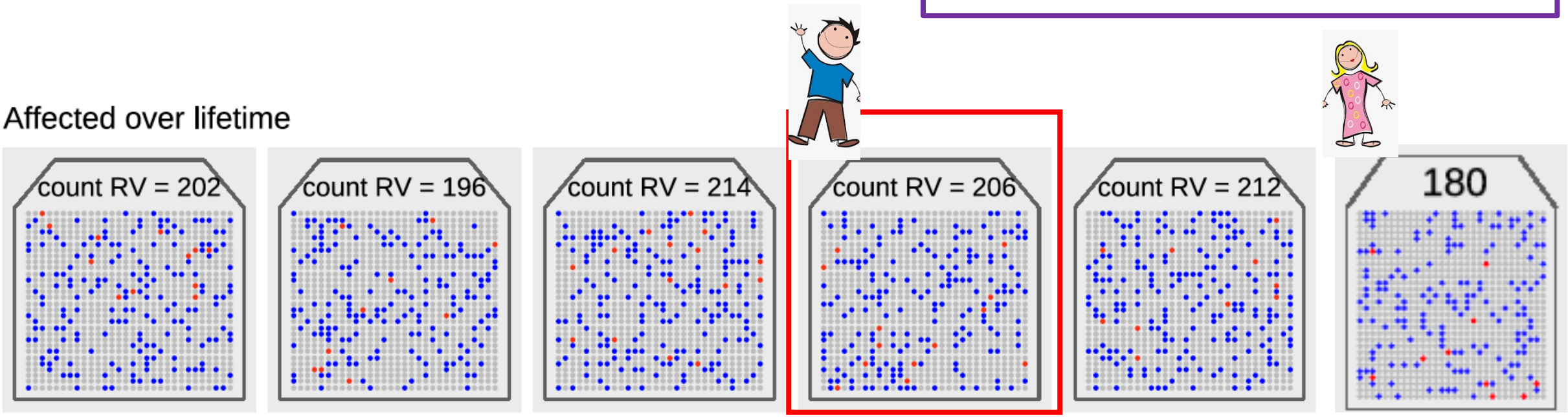
Will people with known family history have high PRS?

Maybe, maybe not!!

JAMA Psychiatry | Review

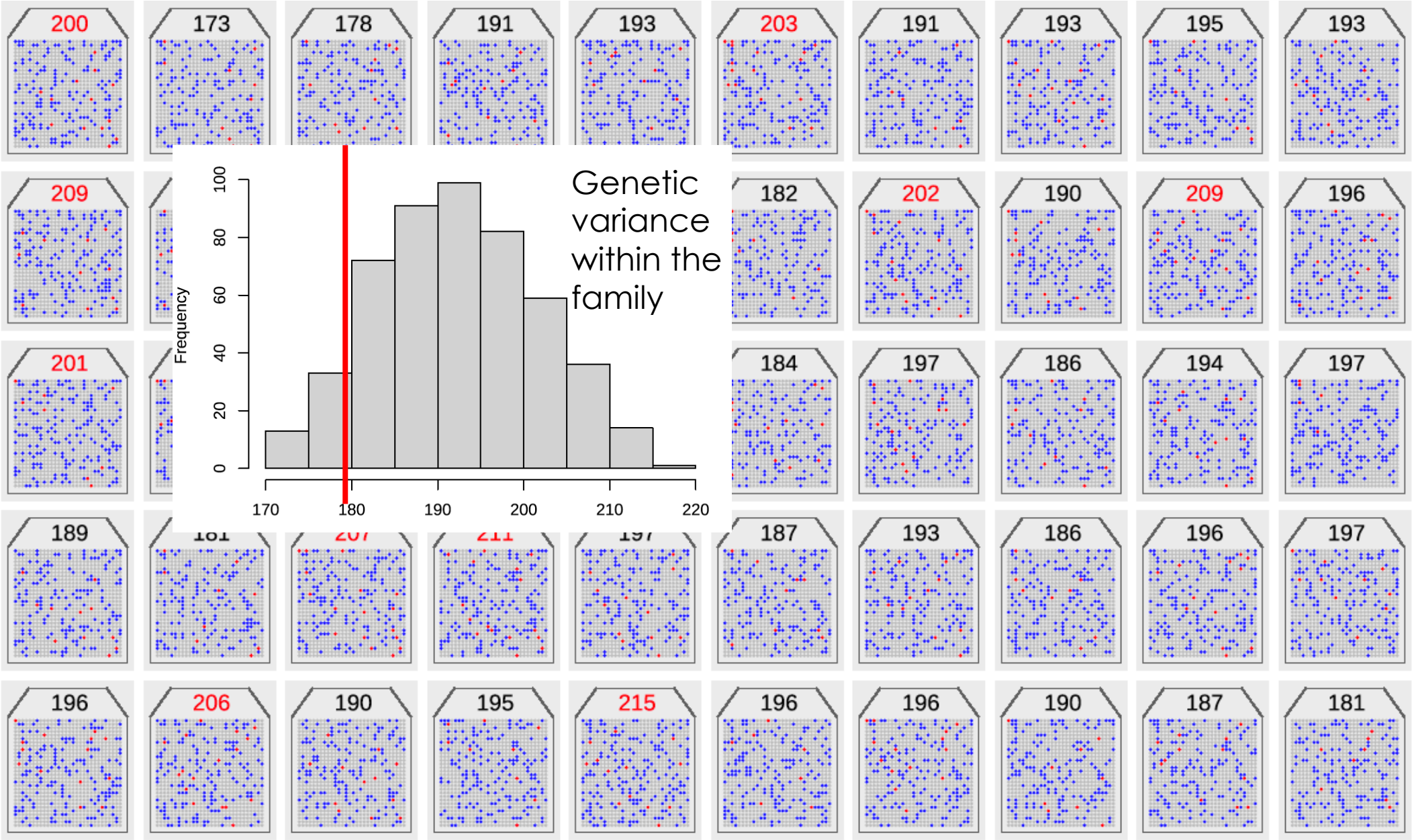
From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

Naomi R. Wray, PhD; Tian Lin, PhD; Jehannine Austin, PhD; John J. McGrath, MD, PhD; Ian B. Hickie, MD; Graham K. Murray, MD, PhD; Peter M. Visscher, PhD



Grey: Homozygote: Two non-risk/protective alleles – always passes a non-risk allele to child at the locus
Red: Homozygote: Two risk alleles – always passes a risk allele to child at the locus
Blue: Heterozygotes: One risk allele & one non-risk allele –
passes a risk allele 50% of the time & a non-risk allele 50% of the time

Children (Parents: 206 & 180)

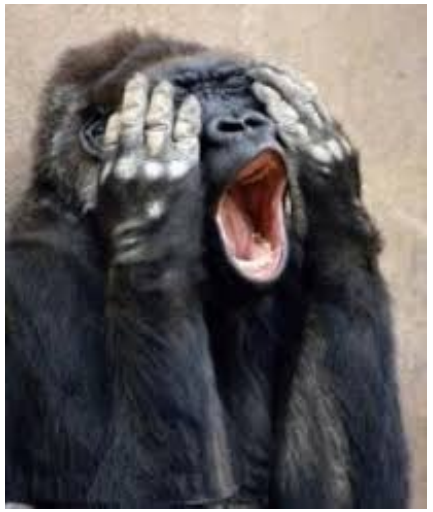


Children of these parents
Mean: 193
+/-3SD: 166-220

Population
Mean: 180
+/-3SD: 142-218

What's in a name?

- **PRS**- Polygenic risk score
- **GPRS**- Genomic or genetic profile risk score
- **PGS** -Polygenic score
- **GRS** - Genetic risk score
- **rsPS** – restricted to significant polygenic score
- **gePS** – global extended polygenic score
- **Multi-SNP score** (usually this uses only single nucleotide polymorphisms (SNPs) that are genome-wide significant, hence the same as gePS)
- **MetaGRS** – a PRS constructed from genetic data for the disease/trait of interest plus from other correlated traits
- **MTAG-GRS/PRS** a PRS constructed from GWAS data from multiple correlated traits
- **Genetic score**
- **Genotypic score**
- **Allele score**
- **Profile score**
- **Linear predictor** (this of course is a generic term, but has been used to describe PRS when risk alleles are the only predictors)



Polygenic risk score methods

A weighted sum of the count of risk alleles

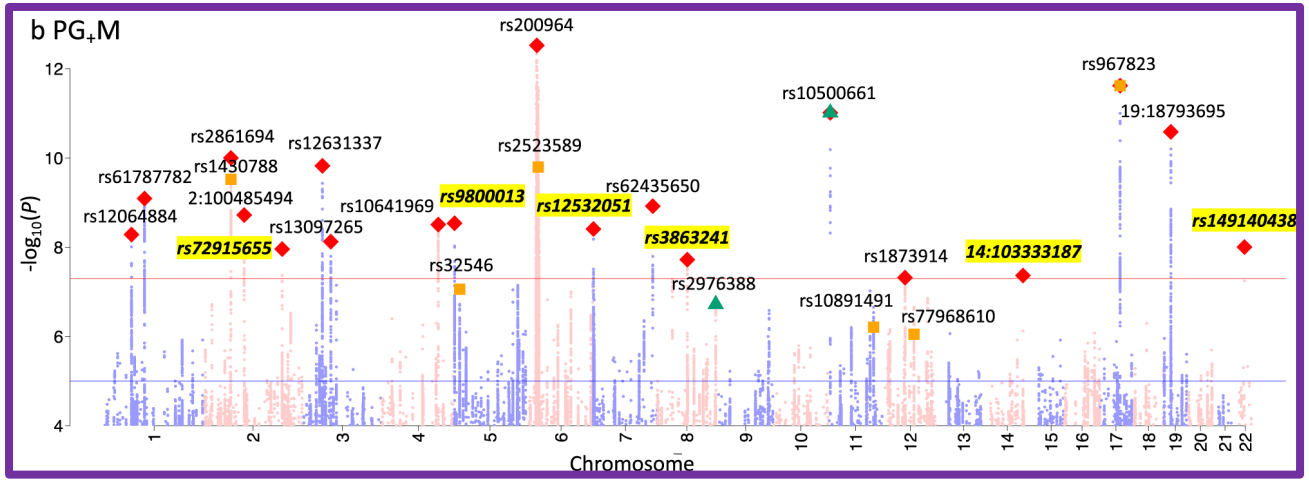
$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold



Polygenic risk score methods

A weighted sum of the count of risk alleles

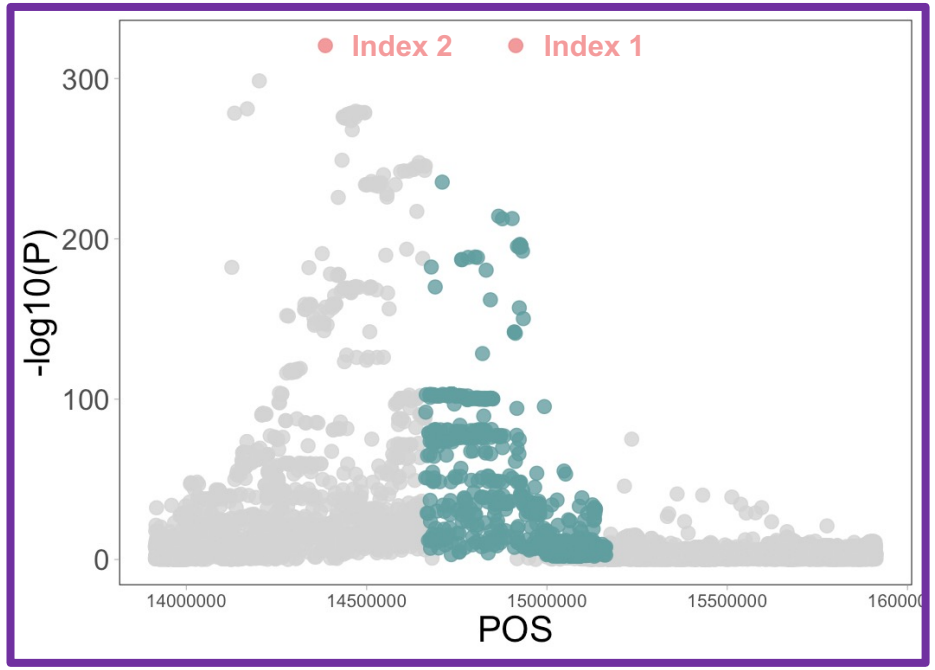
$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold



Polygenic risk score methods

A weighted sum of the count of risk alleles

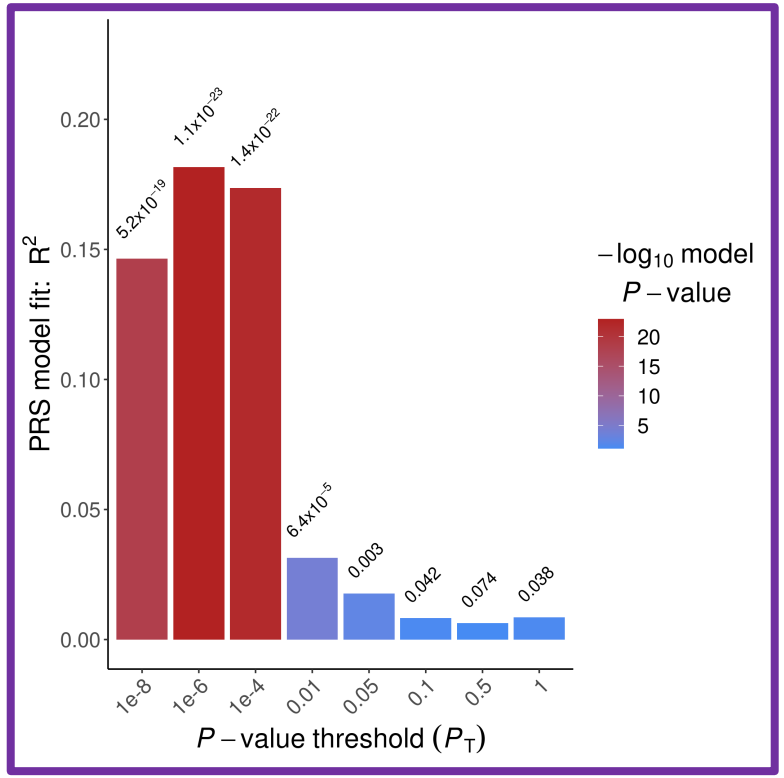
$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

How many SNPs?
Which SNPs?
What weights?

Basic method:

Clumping & P-value thresholding (C+PT):

- Select most associated SNP in tower – LD-based clumping
- Select on a p-value threshold



Polygenic risk score methods

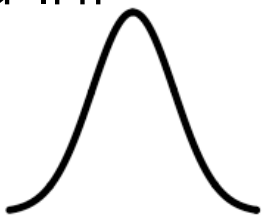
A weighted sum of the count of risk alleles

$$\text{PRS} = \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \widehat{\beta_3} x_{i3} + \cdots = \sum_{j=1}^{n_{SNP}} \widehat{\beta_j} x_{ij}$$

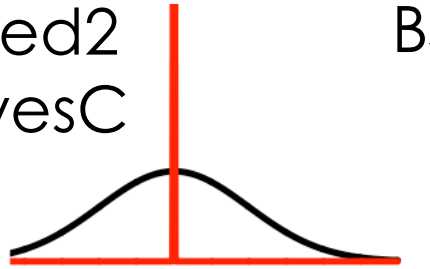
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture

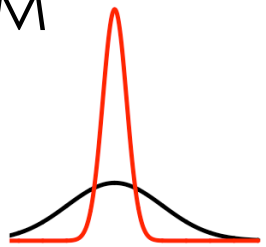
LDpred-Inf
SBLUP



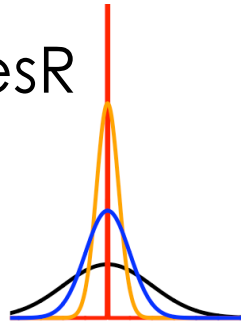
LDPred2
SBayesC



BSLMM



SBayesR



Polygenic risk score methods

Table 1. Summary of Methods Used to Generate Polygenic Scores

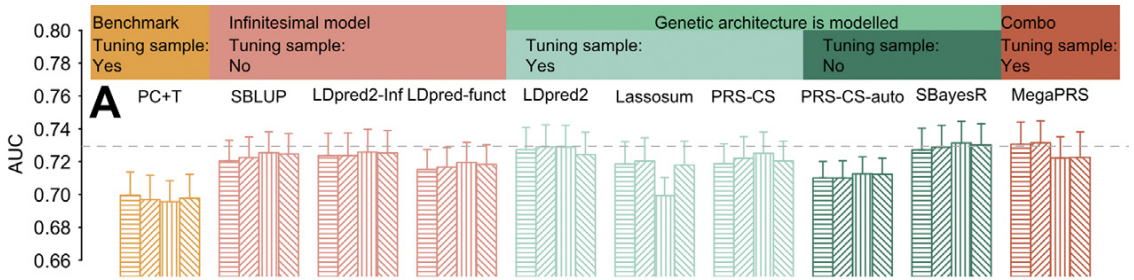
Method	Distribution of SNP Effects (β)	Tuning Sample	Predefined Parameters	Parameters Estimated in Tuning Sample
PC+T	None	Yes	–	*p*-value threshold
SBLUP	$\beta \sim N\left(0, \frac{h_g^2}{m}\right)$ h_g^2: SNP-based heritability, m: number of SNPs; $\lambda = m(1 - h_g^2)/h_g^2$	No	λ LD radius in kb	–
Ldpred2-Inf	Same as SBLUP	No	h_g^2 LD radius in cM or kb	–
LDpred-funct	$\beta_j \sim N(0, c\sigma_j^2)$ $\sum_{j=1}^{M} 1_{\sigma_j^2>0} c\sigma_j^2 = h_g^2$, c is a normalizing constant, σ_j^2 is the expected per SNP heritability under the baseline-LD annotation model estimated by stratified LDSC from the discovery GWAS within LDpred-funct software	No	h_g^2 LD radius in number of SNPs	–
LDpred2	$\beta_j \sim \begin{cases} N\left(0, \frac{h_g^2}{\pi m}\right), & \text{with probability of } \pi \\ 0, & \text{with probability of } 1-\pi \end{cases}$ When sparsity is "true," the β_j for SNPs in the $(1 - \pi)$ partition are all set to zero	Yes	h_g^2 π software default values, LD radius in cM or kb	π, sparsity
Lassosum	$f(\beta) = \mathbf{y}^T\mathbf{y} + (1 - s)\beta^T\mathbf{X}_r^T\mathbf{X}_r\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + s\beta^T\beta + 2\lambda\lVert\beta\rVert_1^1$ $\mathbf{X}_r$: $n \times m$ matrix of genotypes of LD reference sample, where n is sample size	Yes	LD blocks	λ, s
PRS-CS	$\beta_j \sim N\left(0, \frac{\sigma^2}{n}\psi_j\right)$ $\psi_j \sim G(a, \delta_j)$ $\delta_j \sim G(b, \phi)$, ϕ is a global scaling parameter	Yes	$a = 1$, $b = 0.5$ n LD blocks	ϕ
PRS-CS-auto	Same as PRS-CS, but estimates ϕ from the discovery GWAS	No	$a = 1$, $b = 0.5$ n LD blocks	–
SBayesR	$\beta_j \mid \pi, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2\sigma_\beta^2), & \text{with probability of } \pi_2 \\ \vdots & \\ N(0, \gamma_c\sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1}\pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2\ (d.f. = 4)$ $\pi_l \sim Dir(\mathbf{1})$, estimated from discovery GWAS in SBayesR software γ_l are scaling parameters	No	LD radius in cM or kb C = 4 γ software default values	–
MegaPRS	Lasso: $\beta_j \sim DE(\lambda/\sigma_j)$ Ridge regression: $\beta_j \sim N(0, v\sigma_j^2)$ BOLT-LMM: $\beta_j \sim \begin{cases} N\left(0, \frac{(1-f_2)\sigma_j^2}{\pi}\right), & \text{with probability of } \pi \\ N\left(0, \frac{f_2\sigma_j^2}{1-\pi}\right), & \text{with probability of } 1-\pi \end{cases}$ f_2 is the proportion of the total mixture variance in the second normal distribution BayesR: similar to SBayesR with C = 4, and π_l and γ_l estimated in the tuning sample σ_j^2 is the expected per SNP-heritability under BLD-LDAK model using SumHer	Yes	LD radius in cM or kb Parameters used in BLD-LDAK Grid search parameter values for each method	The tuning cohort is used to estimate the parameters that maximize prediction for each model, and from these the model that maximizes prediction is selected

Archival Report

Biological Psychiatry

A Comparison of Ten Polygenic Score Methods for Psychiatric Disorders Applied Across Multiple Cohorts

Guiyan Ni, Jian Zeng, Joana A. Revez, Ying Wang, Zhili Zheng, Tian Ge, Restuadi Restuadi, Jacqueline Kiewa, Dale R. Nyholt, Jonathan R.I. Coleman, Jordan W. Smoller, Schizophrenia Working Group of the Psychiatric Genomics Consortium, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium, Jian Yang, Peter M. Visscher, and Naomi R. Wray

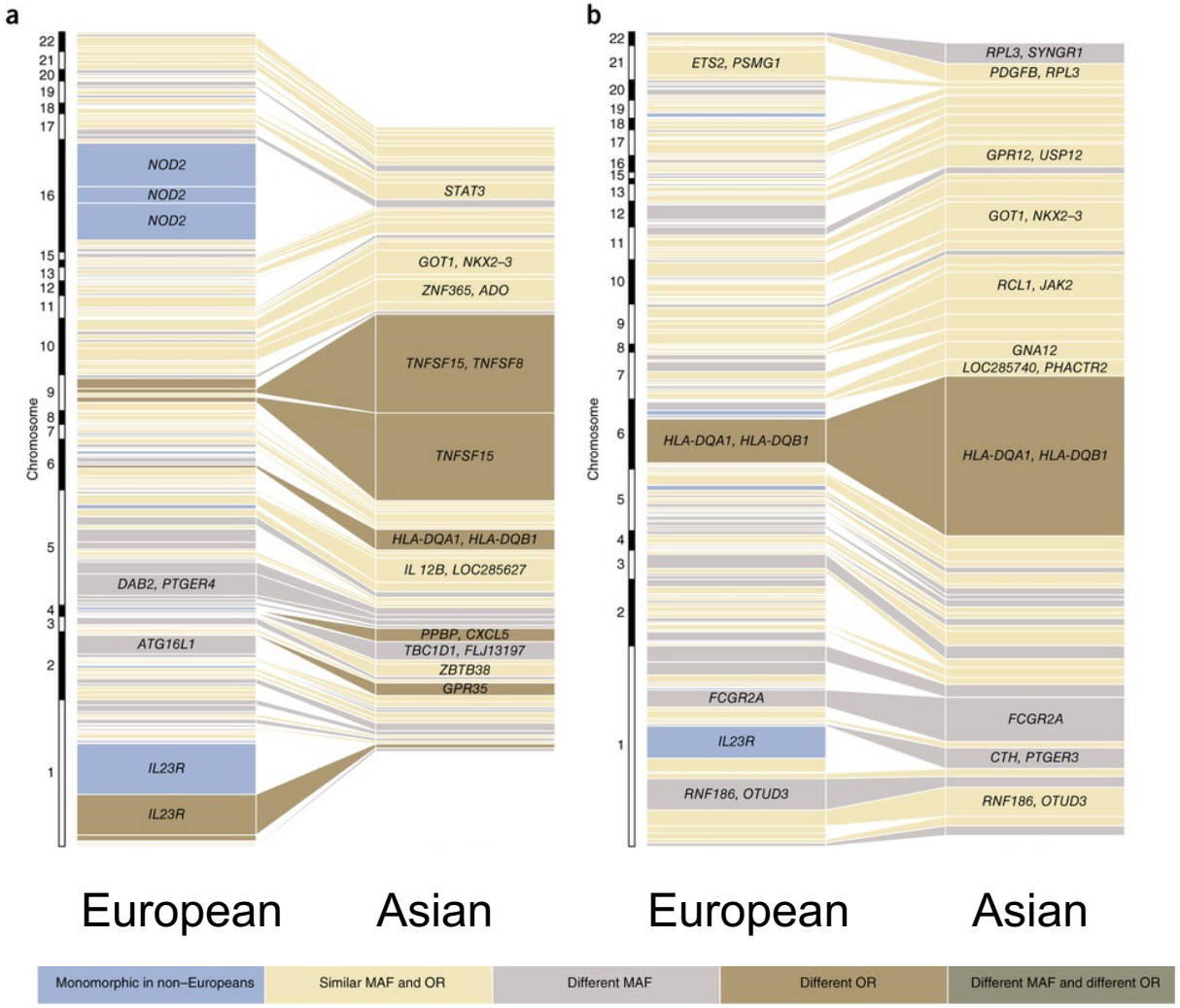


- Random effects models > fixed effects models
- Mixture models > non-mixture (infinitesimal) models

Ancestry

Crohn's Disease
r_g EUR-ASN 0.76

Ulcerative Colitis
r_g EUR-ASN 0.79

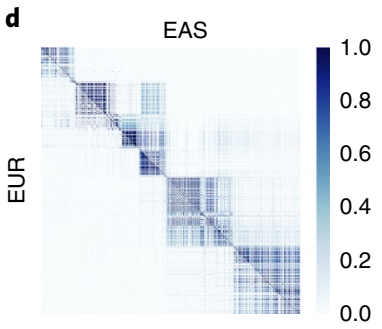


Issues

- Same causal variants
  - Different allele frequencies
  - LD differences
  - Different effect sizes
- Different causal variants
  - GxE
  - Different phenotype

In *general:*
We expect common causal variants to be shared across ancestries

But correlation structure differs



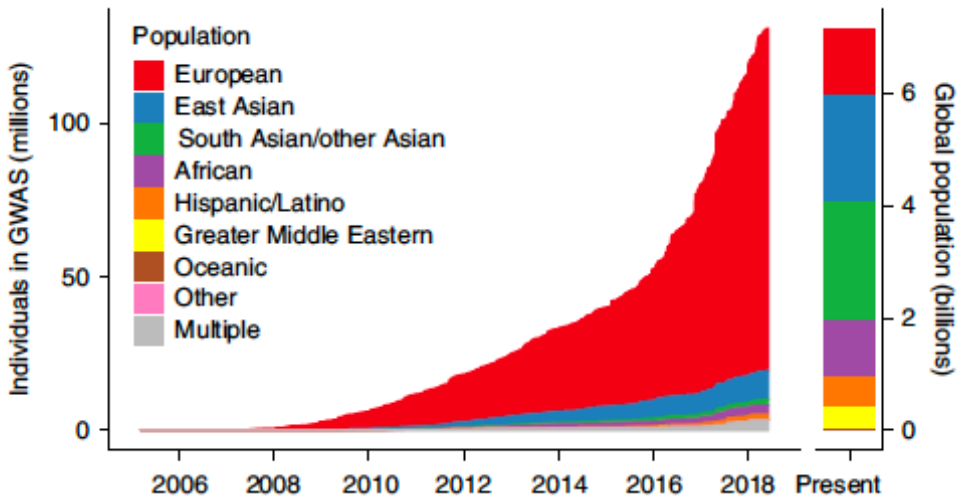
Liu et al (2015) Association analysis identifies 38 susceptibility loci for IBD and highlight shared genetic risk across populations. Nat Gen 2015

Ancestry

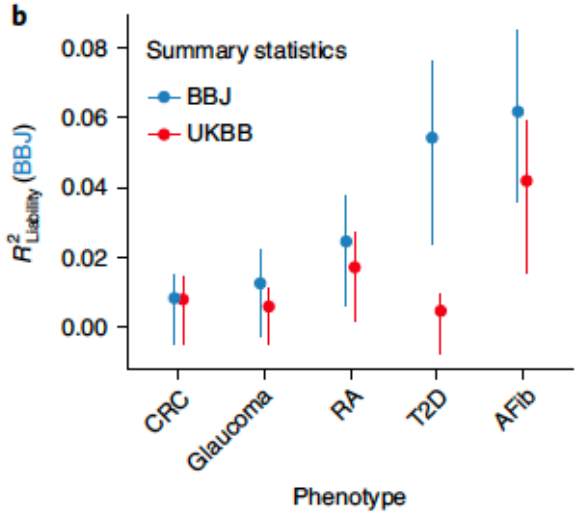
PERSPECTIVE
https://doi.org/10.1038/s41588-019-0379-x
nature genetics

Clinical use of current polygenic risk scores may exacerbate health disparities

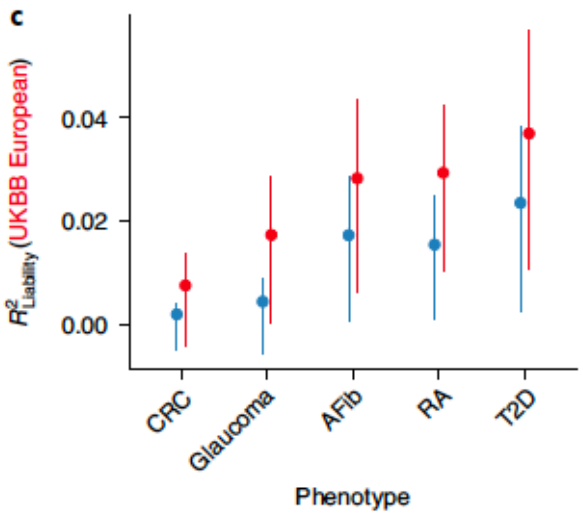
Alicia R. Martin[1,2,3*], Masahiro Kanai[1,2,3,4,5], Yoichiro Kamatani[5,6], Yukinori Okada[5,7,8], Benjamin M. Neale[1,2,3] and Mark J. Daly[1,2,3,9]

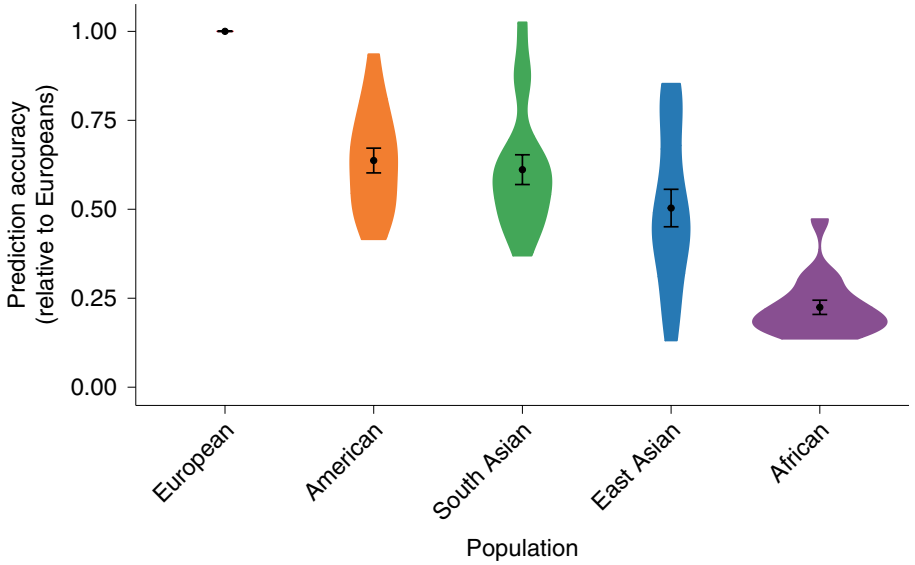


Predicted into Japanese



Predicted into European





Realistic expectations for PRS

- PRS are NOT diagnostic
- PRS will become more accurate as GWAS sample size increases…but still wont be diagnostic
- Very high PRS– immediate utility
- Combine PRS with other predictors
- The time is ripe for evaluation of PRS in clinical settings
- At the same time, more data and improved methods to ensure PRS have utility across ancestries
- Research designs: 44- fold difference in odds of having schizophrenia for lowest centile of PRS, the highest centile

History of PRS

GENETICS | HIGHLIGHTED ARTICLE
GENOMIC PREDICTION

GENOMIC PREDICTION

2019

Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans

Naomi R. Wray,*,†,1 Kathryn E. Kemper,* Benjamin J. Hayes,‡ Michael E. Goddard,§,** and Peter M. Visscher*,†

Shai Carmi @ShaiCarmi · Apr 6

Remember when the Broad Institute discovered polygenic scores? Now it seems as if they invented quantitative genetics.

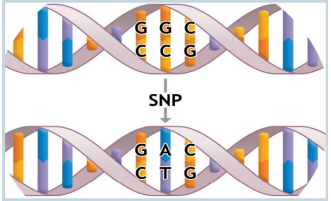
See below for a thread. (Happy if someone could send me the full text.)

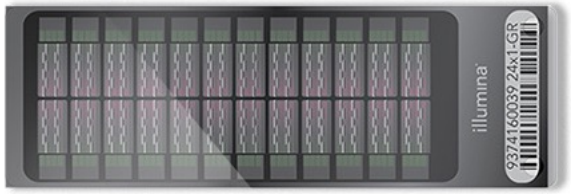
1/7

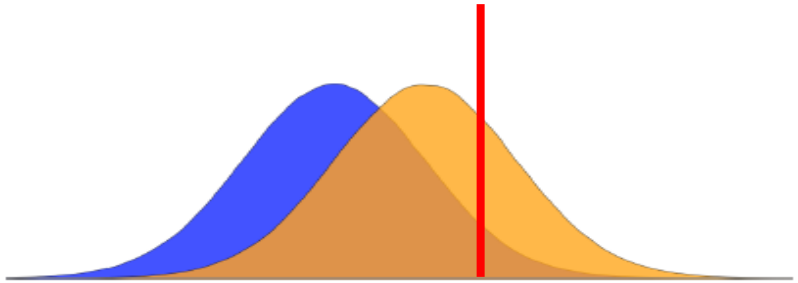
Concordance of a High Polygenic Score Among Relatives
ahajournals.org

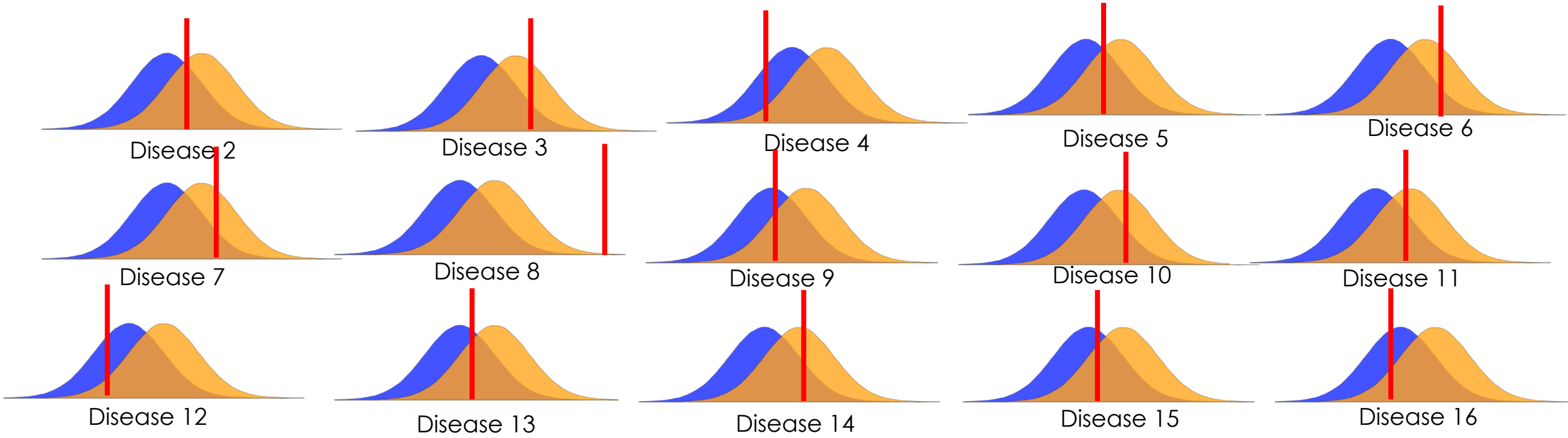
Justify for one disease and the rest come for free!

One disease









Summary

PRS are …

- PRS are imperfect genetic predictors with inherently limited accuracy.
- PRS are often combined with other predictive measures to predict the total disease risk.
- PRS are useful in risk stratification to better triage people into established screening programs.
- In principle, PRS are available for an individual for all common diseases from birth.

PRS are not …

- PRS are not diagnostic.
- PRS are not absolute risk and do not provide a baseline or timeframe for the progression of a disease.
- PRS accuracy will increase with GWAS sample size but are never going to be able to definitively predict complex conditions.
- PRS are not and never will be stand-alone predictors of common diseases.

Practical 1: Computation of PRS using C+PT

https://cnsgenomics.com/data/teaching/GNGWS23/model5/Practical1_PRS.html

Log into the cluster

cd to your working directory in scratch: cd /scratch/[your folder]

If you have not created a folder yet, you can do it by

cd /scratch/
mkdir [your folder]

What is the maximum prediction accuracy we can get?

h_m^2 : True variance explained by the predictor depends on the SNP set - subscript m.

$$R^2 = \frac{h_m^2}{1 + C}$$

Variance explained by the predictor

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

We want C to be as small as possible:

- C decreases as Discovery sample N increases
- C decreases as the number of SNPs in the SNP set m decreases

$$C \approx \frac{m}{N h_m^2}$$

Catch-22

As m gets smaller, h_m^2 also gets smaller

How to optimise m and h_m^2 to get max R^2 ?

How about whole genome sequencing?

Maximum depends on maximising h_m^2

We use GWAS data so the maximum h_m^2 is the SNP-based heritability

Theoretical maximum depends is the heritability of the trait

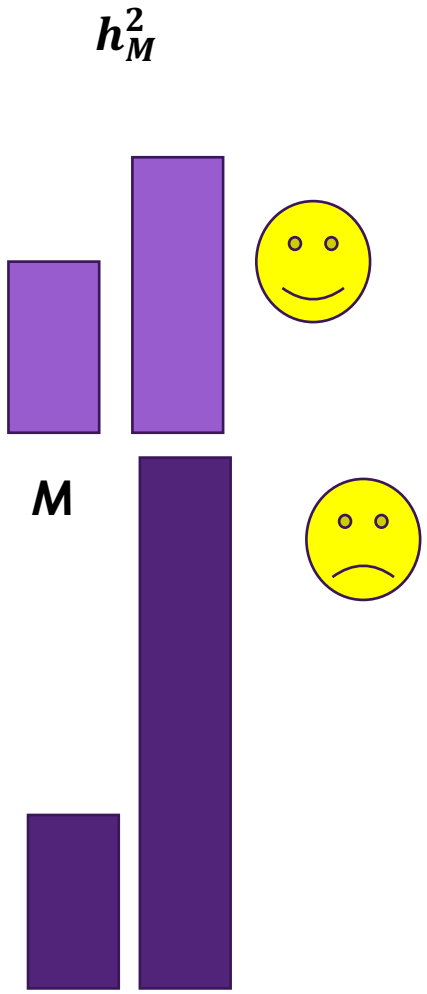
$$R^2 \approx \frac{h_m^2}{1 + \frac{m}{Nh_m^2}}$$

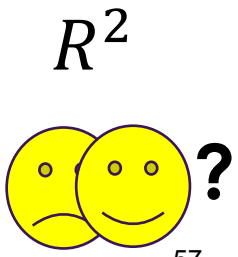
With whole genome sequencing the variance captured by all measured SNPs will increase

But the number of SNPs that we have estimate effect sizes for increases much more

Need MASSIVE discovery sample sizes for WGS association

Also..rare variants are less likely to be shared across populations





Wray et al (2019) Complex trait prediction from genome data. Genetics