THE UNIVERSITY OF QUEENSLAND AUSTRALIA | CREATE CHANGE

Introduction to Polygenic Prediction

History, Theory, Applications & Methodology

Jian Zeng

j.zeng@uq.edu.au

THE UNIVERSITY OF QUEENSLAND AUSTRALIA | Institute for Molecular Bioscience

Program in Complex Trait Genomics

Slides credit: Naomi Wray

Acknowledgement of Country

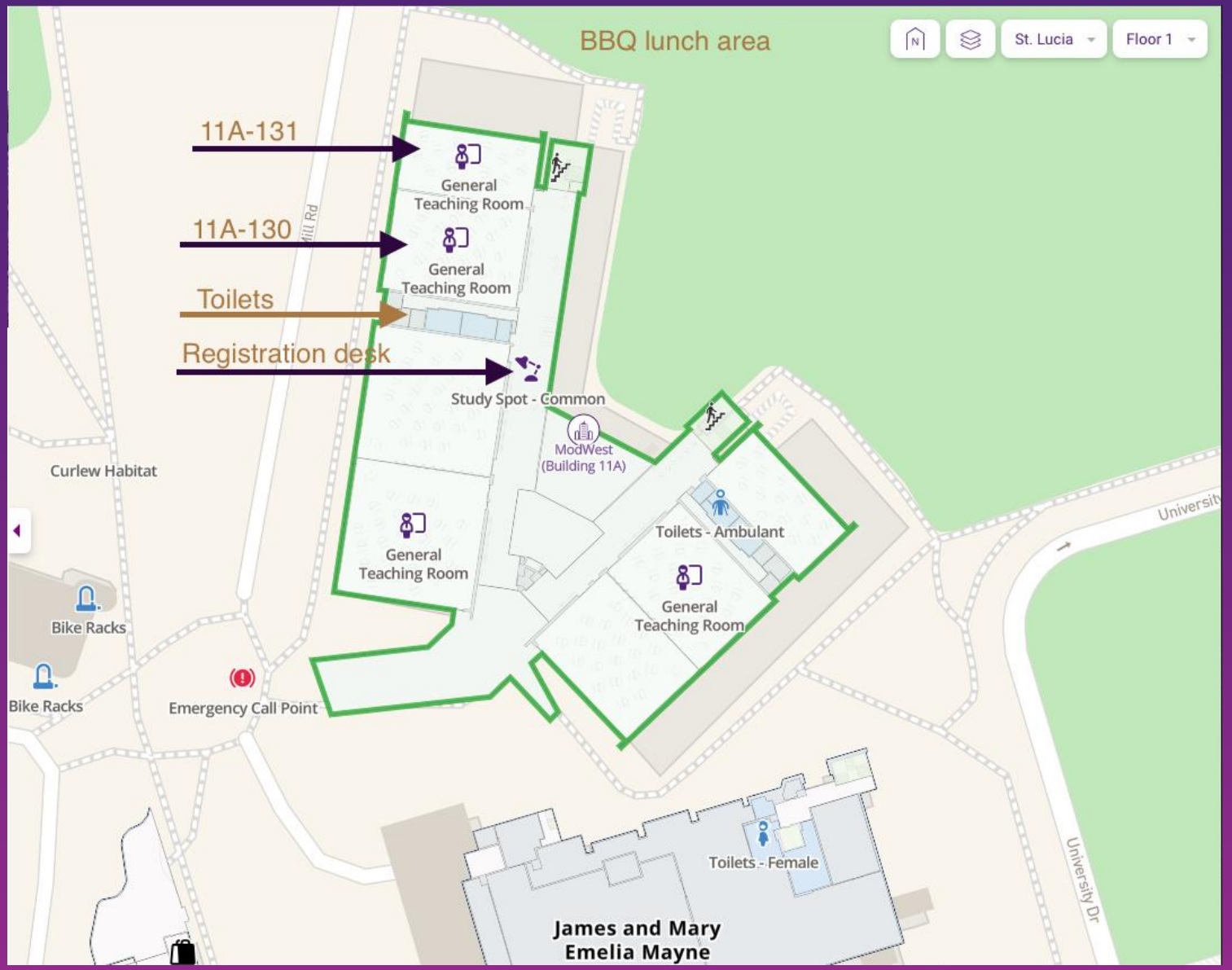
- The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.
- We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.
- We recognise their valuable contributions to Australian 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 11A MODWEST
  - Bathrooms
  - Vending machines
- Food court and other bathrooms are located in Building 63 or Building 21B
- 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

EMERGENCY CONTACT/UQ SECURITY: 3365 3333



Learning materials

Instructions to access WiFi/desktop/server:

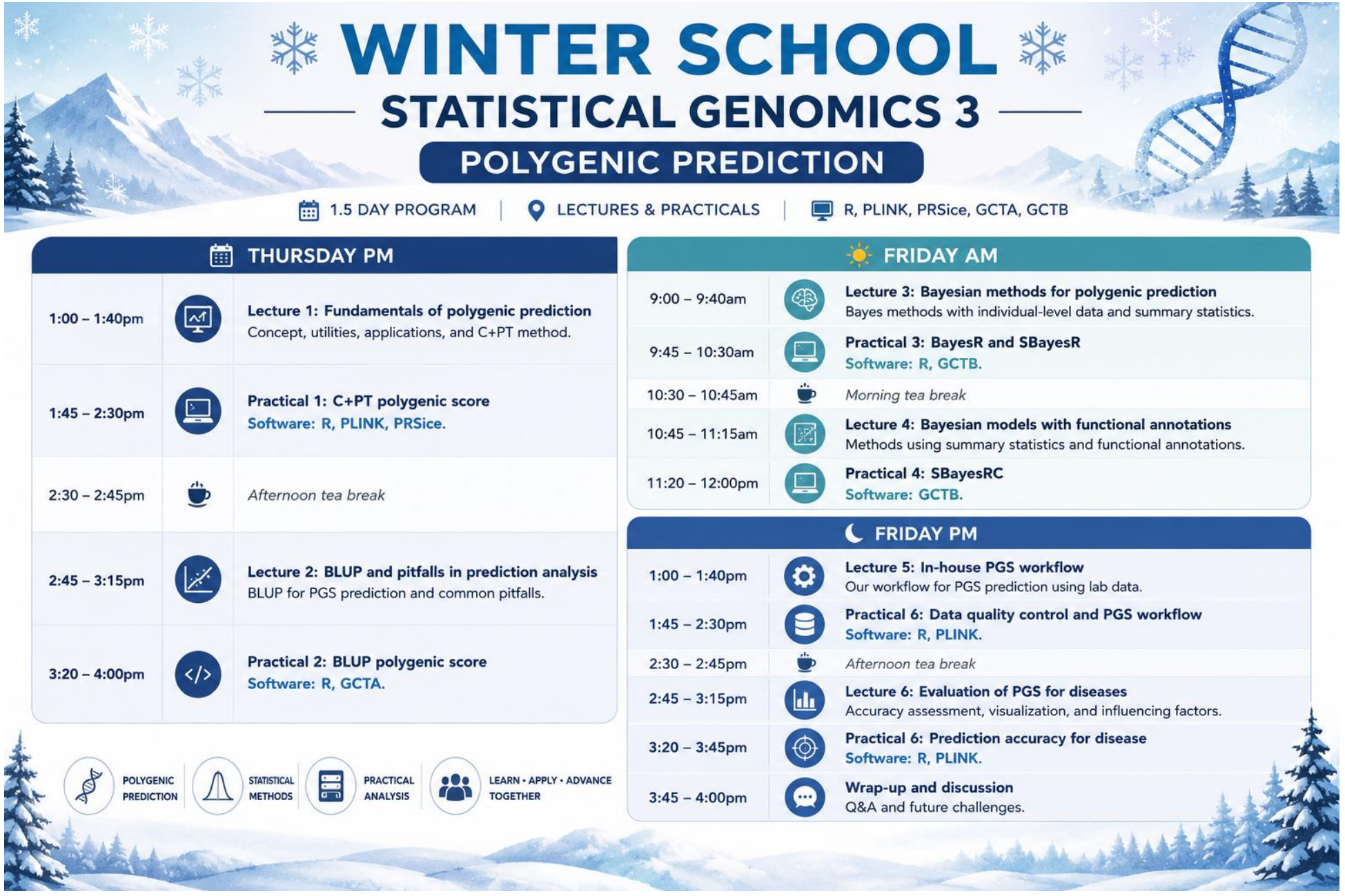
https://cnsgenomics.com/data/teaching/GNGWS26/module0/

The winter school server is available until **24th July 2026** (2 weeks after the course)

Slides and practical notes for this module:

https://cnsgenomics.com/data/teaching/GNGWS26/module5/

Module structure



Tutors

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

Jian Zeng

Tian Lin

Xuemin Wang

Moksedul Momin

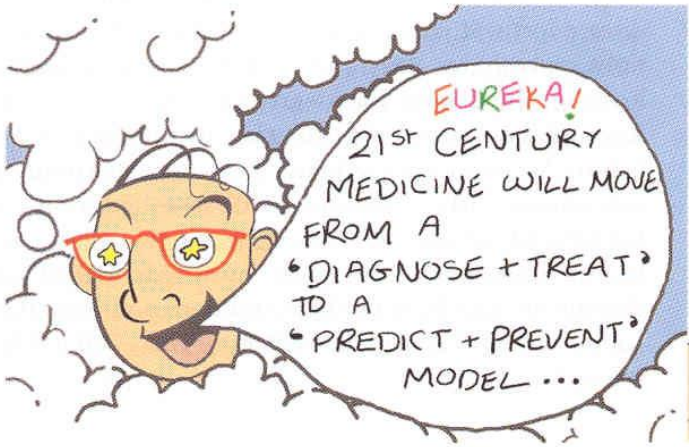
What are we predicting?

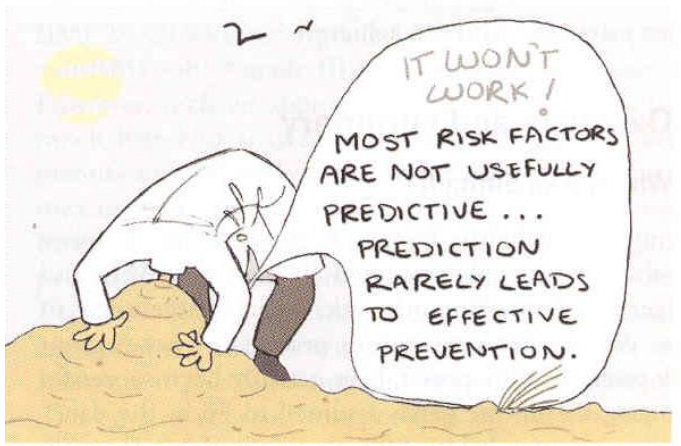
Polygenic scores (PGS) predict individual genetic values of complex traits using **genome-wide** variations.

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

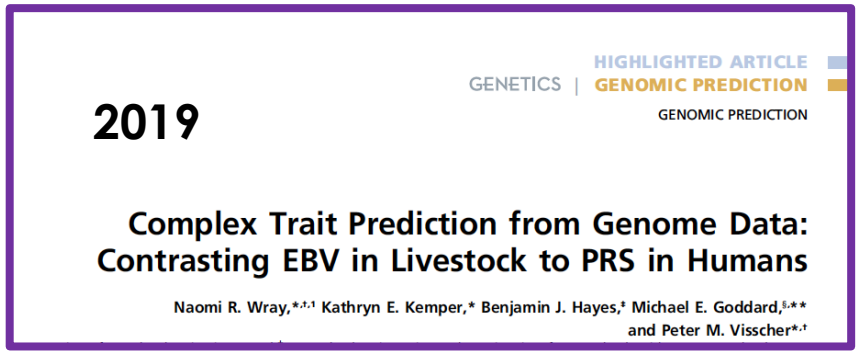
Applications

- Precision medicine (humans)
- Genomic selection (animals/plants)





Source: Strachan & Read Human Molecular Genetics 3.


HIGHLIGHTED ARTICLE
GENETICS | GENOMIC PREDICTION
GENOMIC PREDICTION

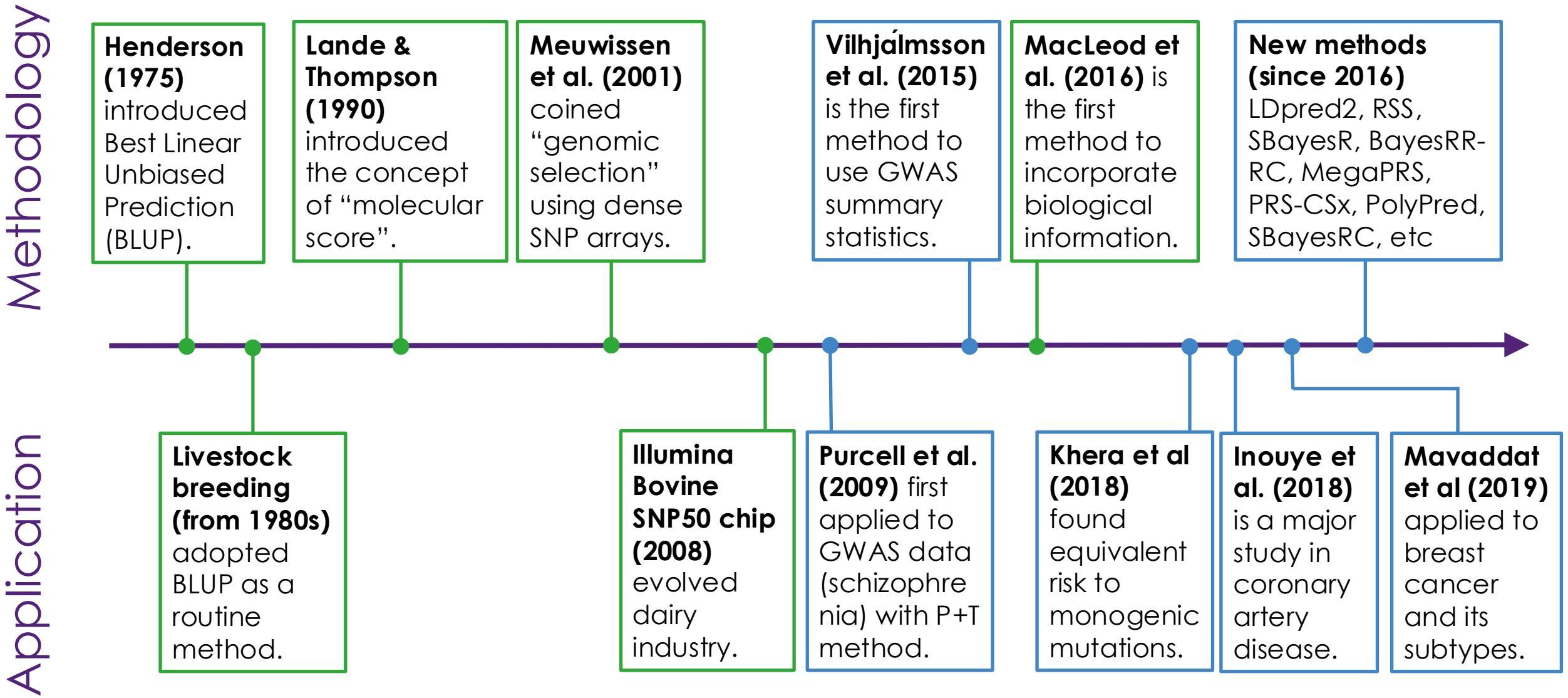
2019

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*,†

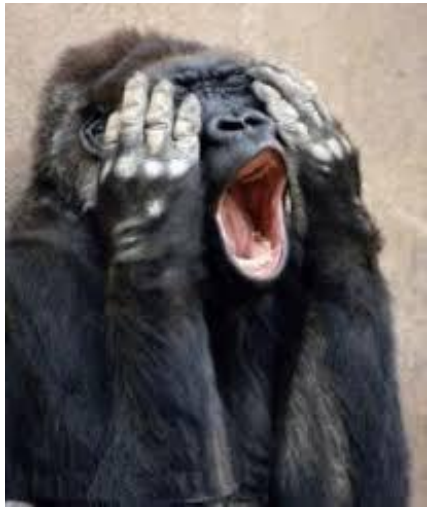
A brief history of PGS in humans & agriculture

The University Of Queensland AUSTRALIA



What's in a name?

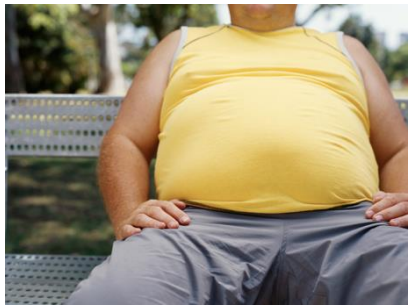
- **PRS**- Polygenic risk score
- **GPRS**- Genomic or genetic profile risk score
- **PGS** -Polygenic score
- **GRS** - Genetic risk score
- **rsPS** – restricted to significant polygenic score
- **gePS** – global extended polygenic score
- **Multi-SNP score** (usually this uses only single nucleotide 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)



Theory and methodology
of polygenic scores (PGS)
are built on
our understanding of
"polygenicity"
in complex traits.

Height

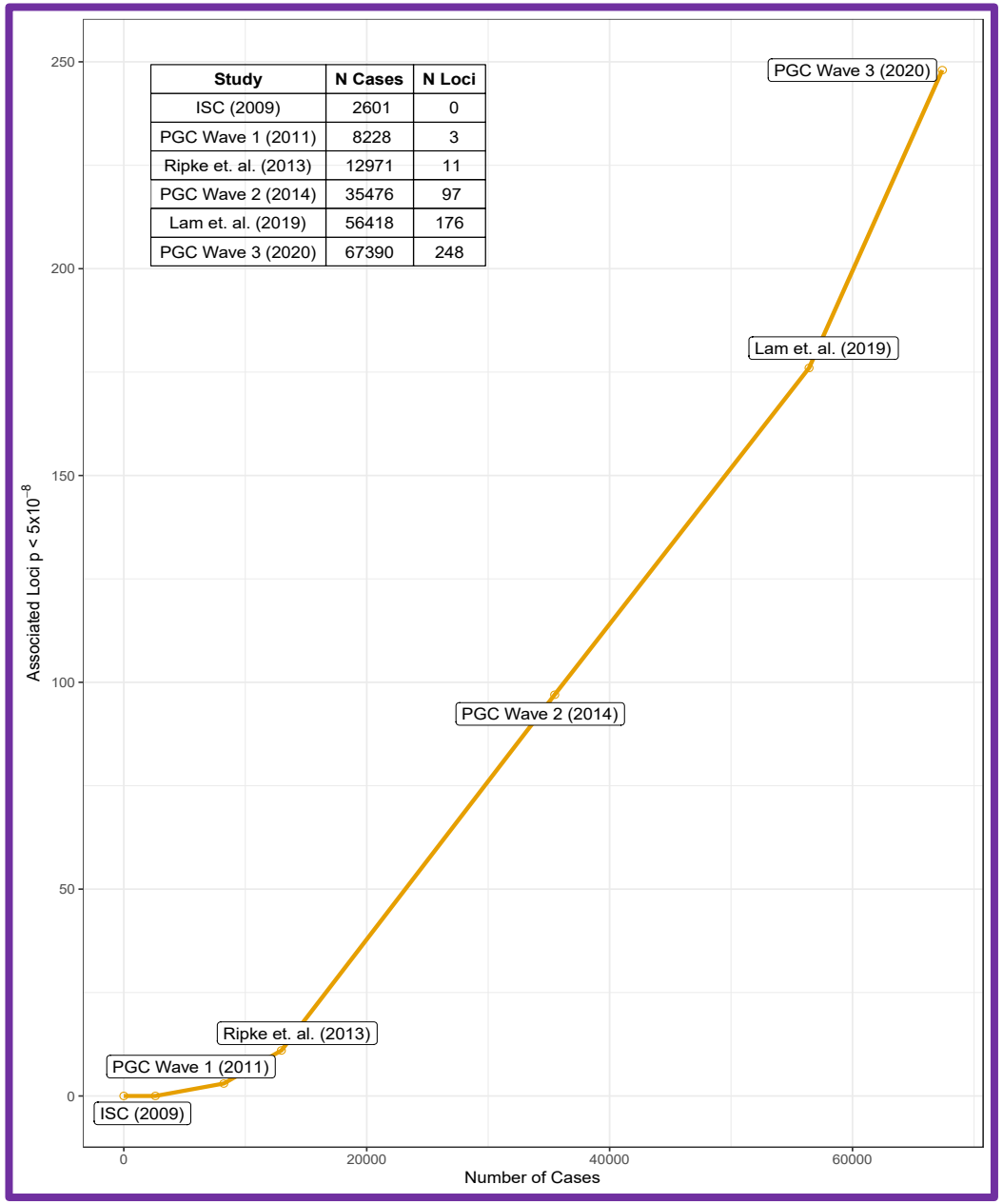
Schizophrenia



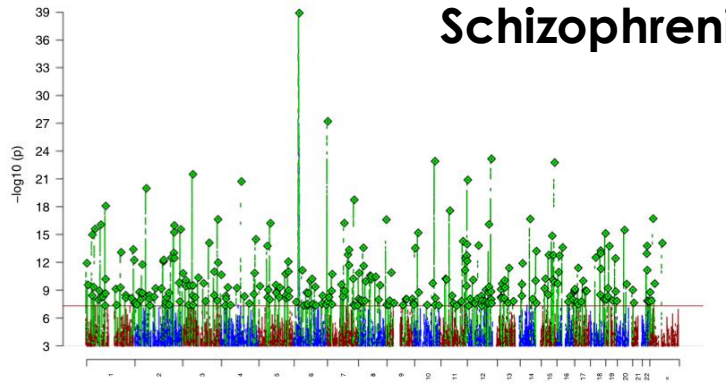
Obesity

Common diseases are polygenic

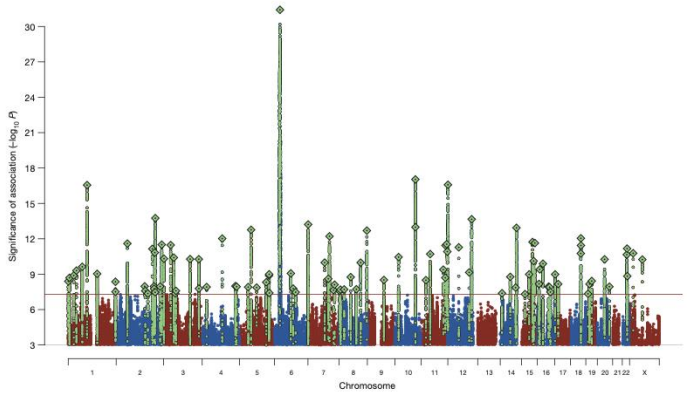
Study	N Cases	N Loci
ISC (2009)	2601	0
PGC Wave 1 (2011)	8228	3
Ripke et. al. (2013)	12971	11
PGC Wave 2 (2014)	35476	97
Lam et. al. (2019)	56418	176
PGC Wave 3 (2020)	67390	248



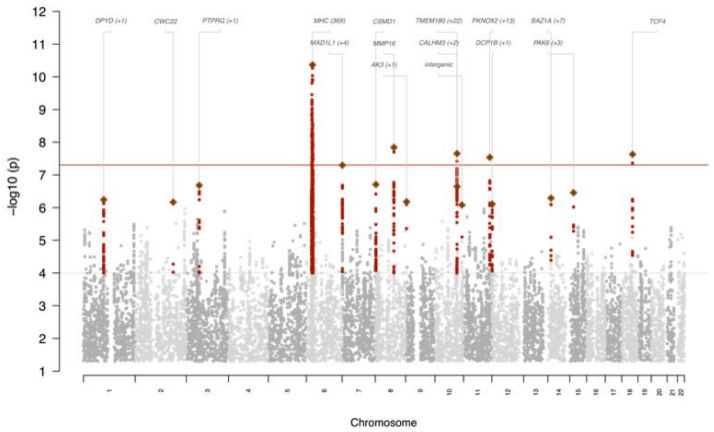
Schizophrenia



2022 PGC Wave 3

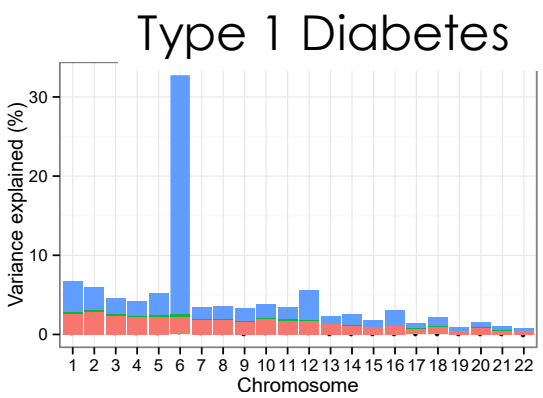


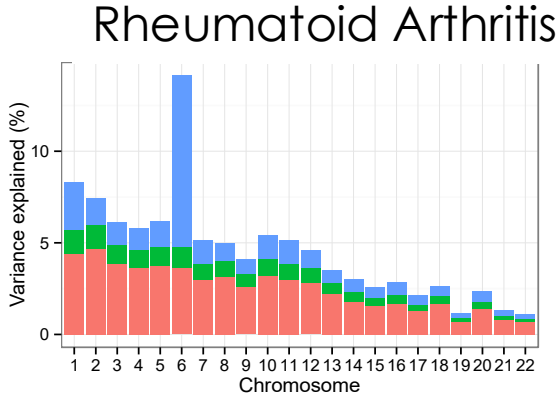
2014 PGC Wave 2

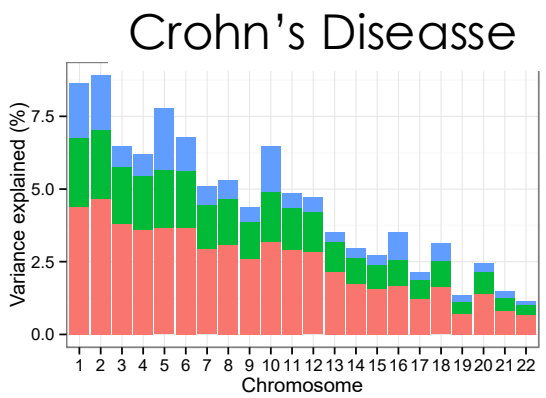


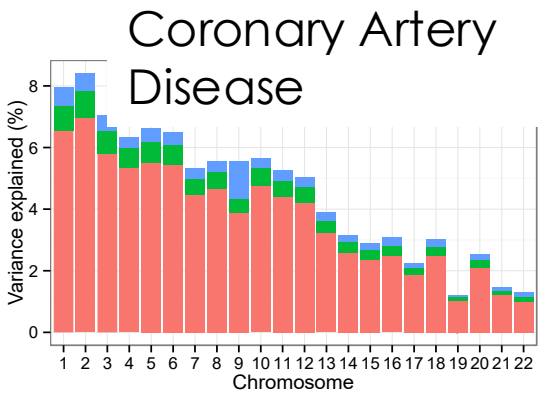
2011 PGC Wave 1

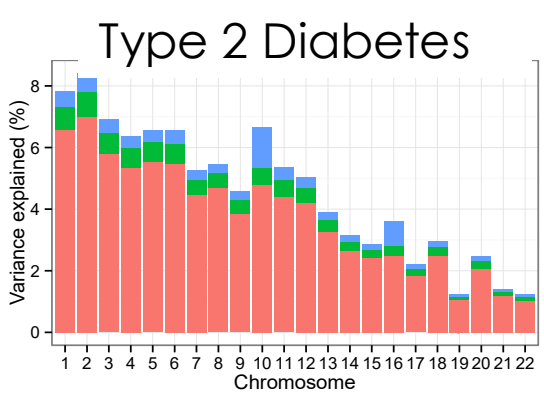
Many polygenic genetic architectures

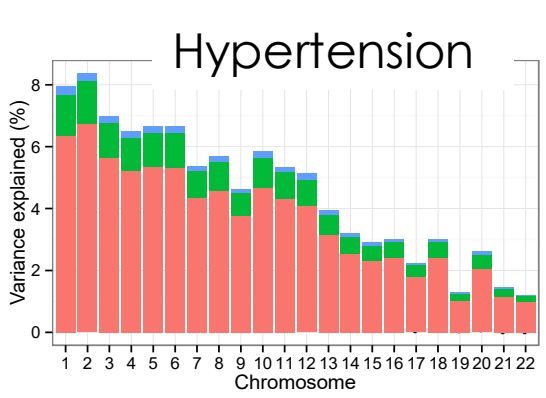


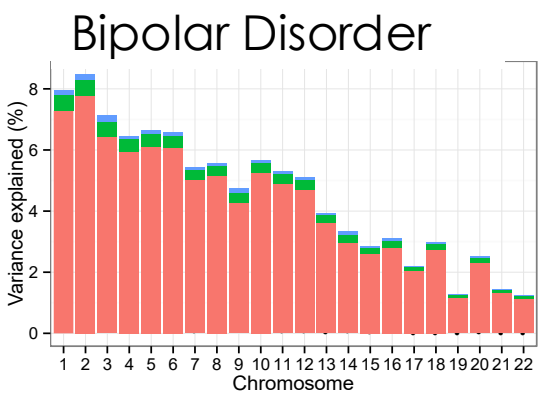


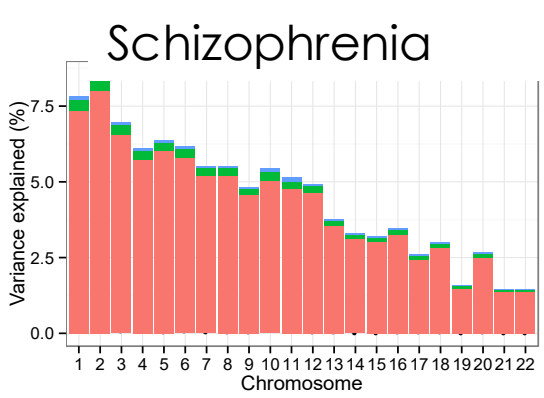


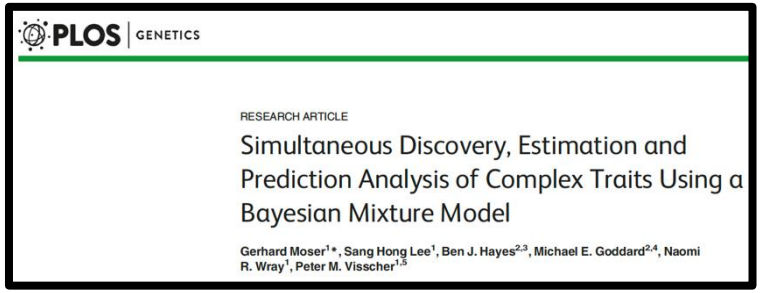











PLOS GENETICS

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]

Mixture component

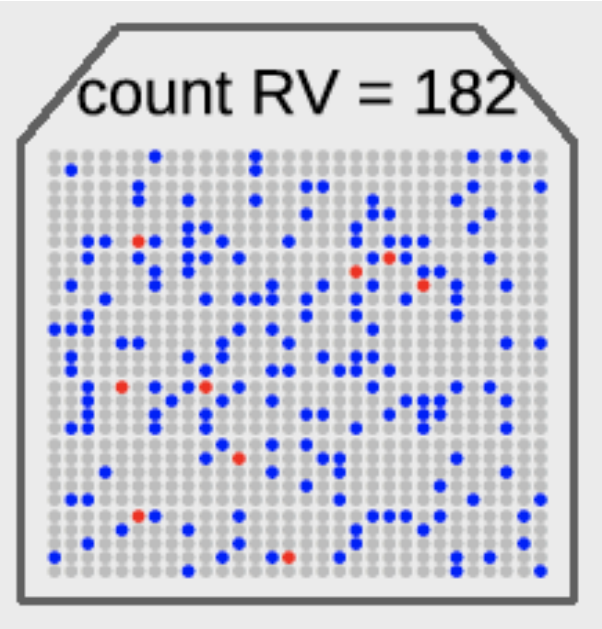
$10^{-4} \times \sigma^2_g$ small

$10^{-3} \times \sigma^2_g$ middle

$10^{-2} \times \sigma^2_g$ big

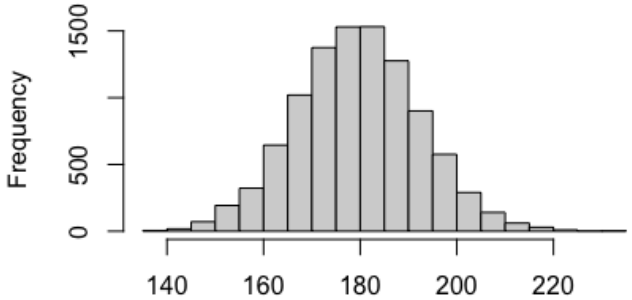
Many DNA variants contribute to genetic risk, and most have very small effects.

Polygenic disease for an individual



900 DNA polymorphic sites

RV =risk variant



Count of RV in population

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

Average person 900*2*0.1 = 180 risk variant

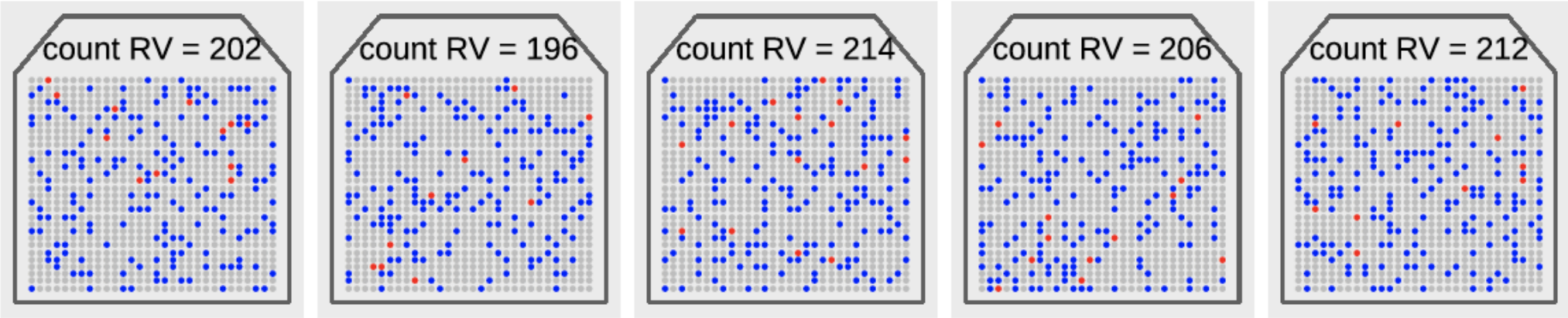
Mean +/- 3SD: 142 to 218

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

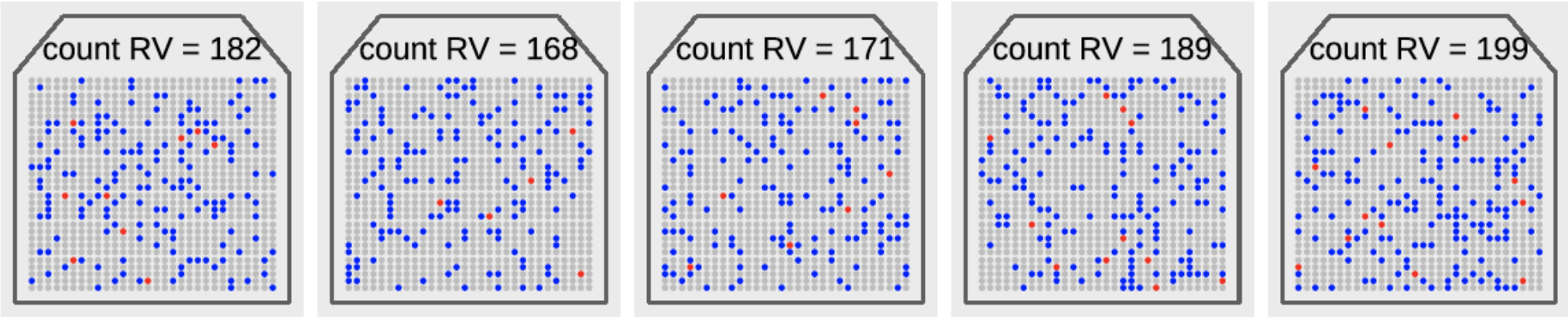
Toy example

Polygenic disease for an individual

Affected over lifetime

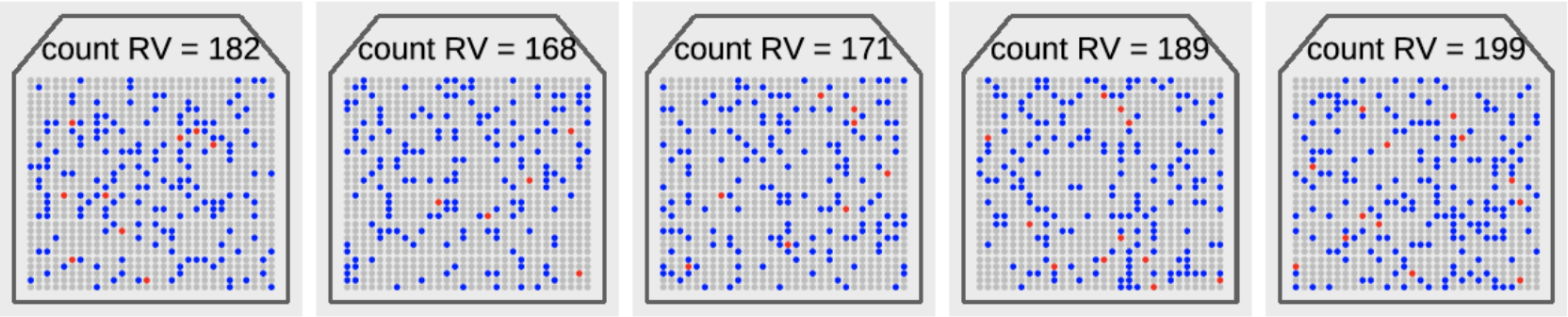


Not affected over lifetime



- 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

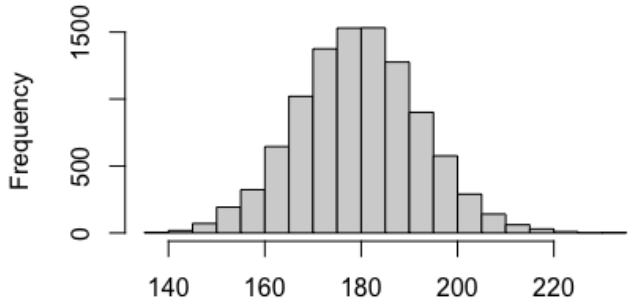
Polygenic score



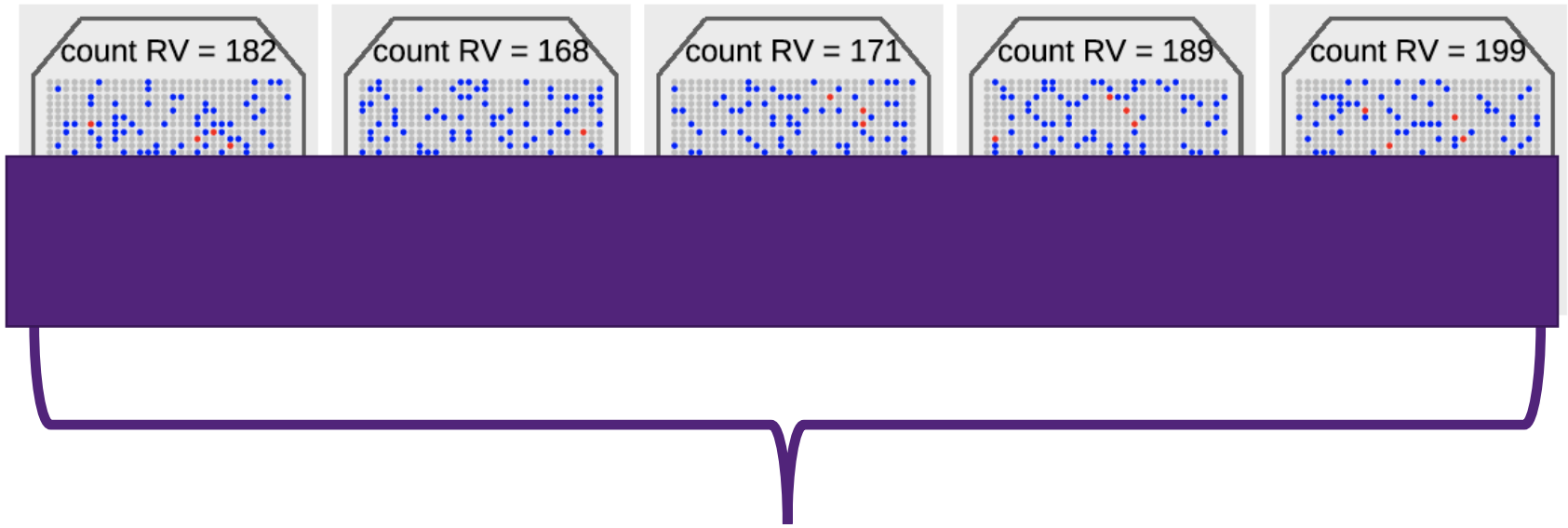
→ "True" polygenic score

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

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

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

$\mathrm{SNP-based\ heritability}$

Polygenic score

In reality, risk variants have different effect sizes.

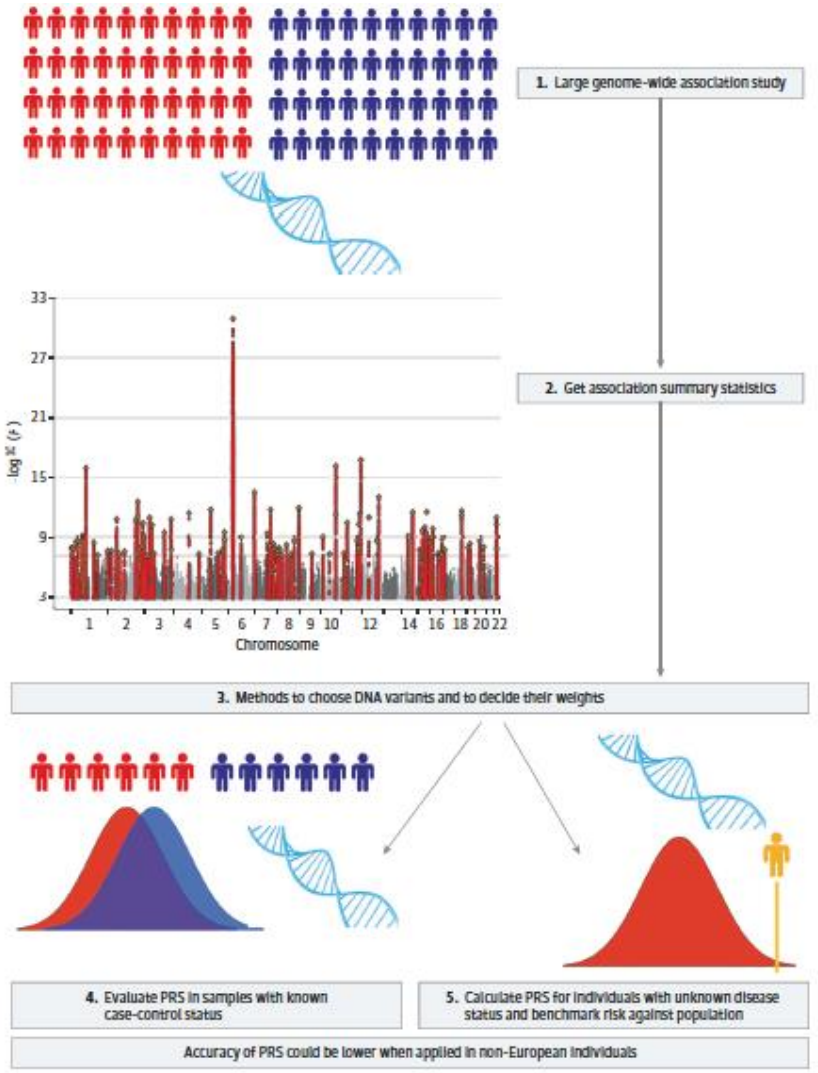
Therefore, PGS is a weighted count of risk alleles:

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

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.

Evaluate
Y= b*PGS + e
R^2 = var(b*PGS)/Var(Y)

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

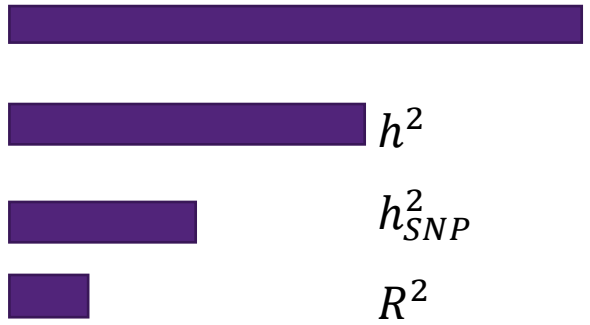
Limitations in prediction accuracy

- PGS 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).

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

- PGS 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.

- PGS 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

Parameters determining the prediction accuracy

The expected value of prediction accuracy:

Variance explained by the predictor

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

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

C: captures the error in estimation

As C→ 0, R^2 → h_m^2

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

- *N*: discovery sample size
- *m*: the number of SNPs (assume LD-independent)
- h_m^2: the SNP-heritability captured by *m* SNPs

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

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.

Variance explained by the predictor

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

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 ?

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

Family history

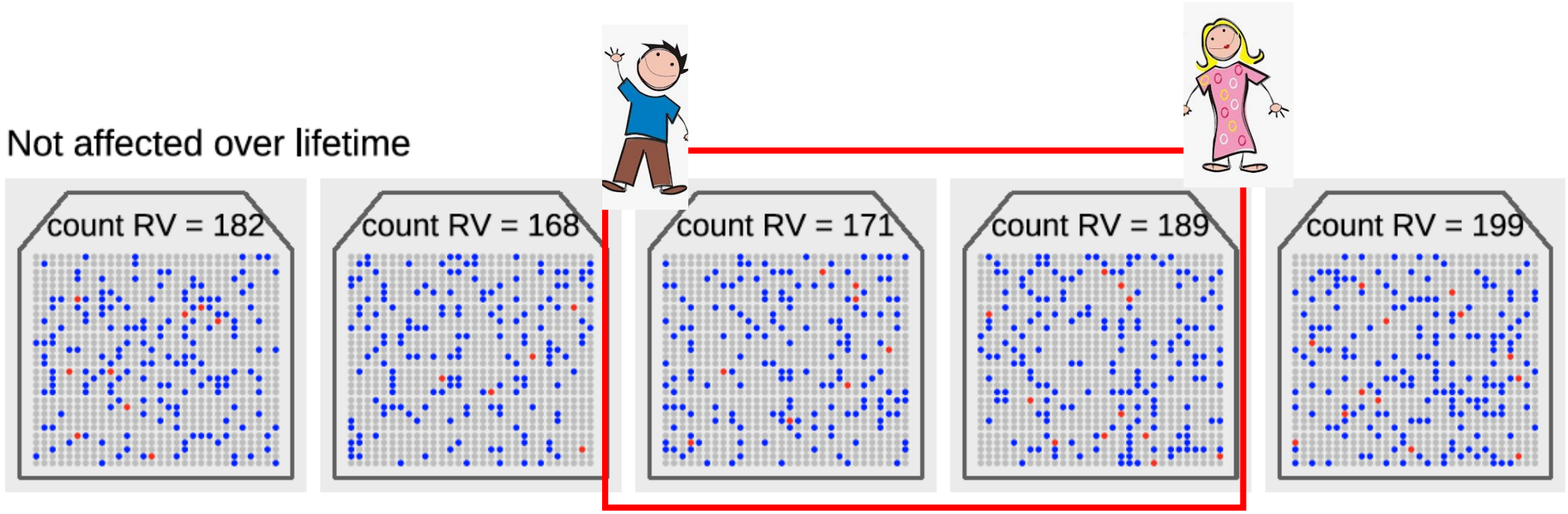
Will people withOUT known family history have high PGS?

Maybe, and that's important!

JAMA Psychiatry | Review

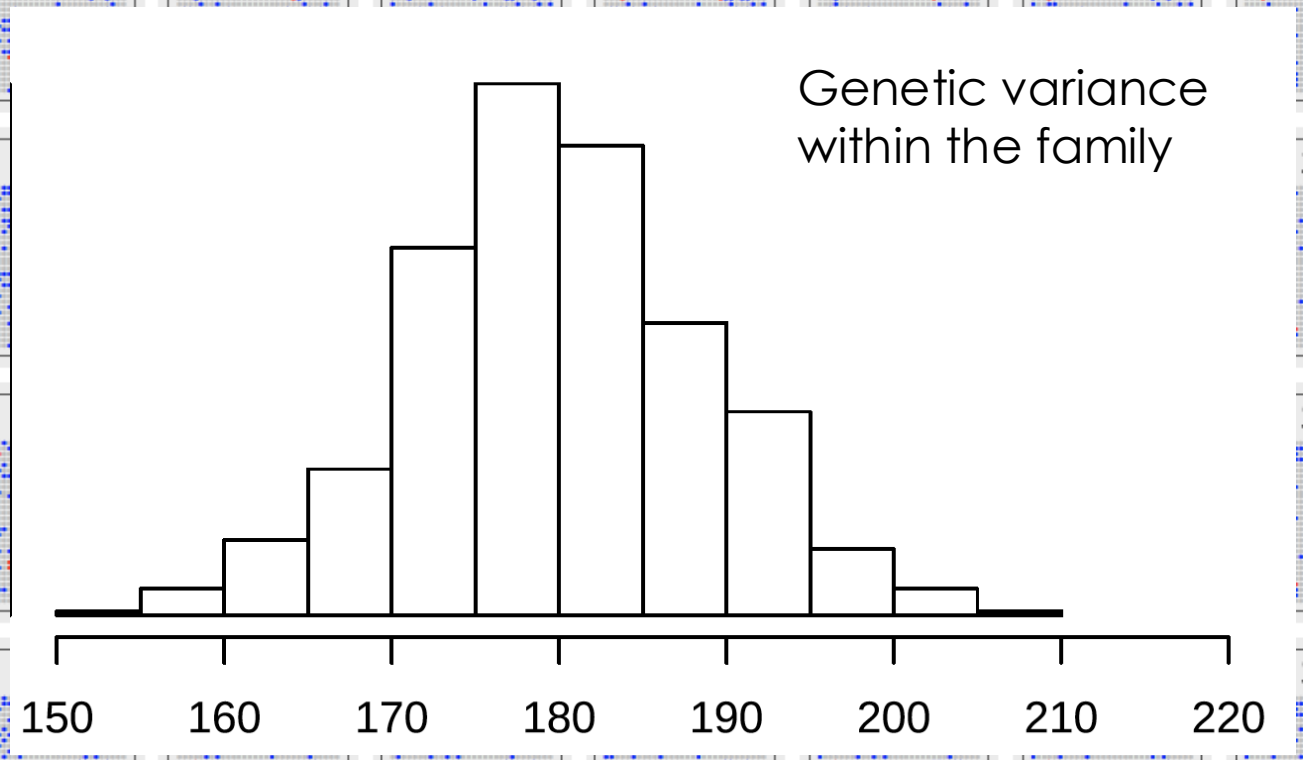
From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

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

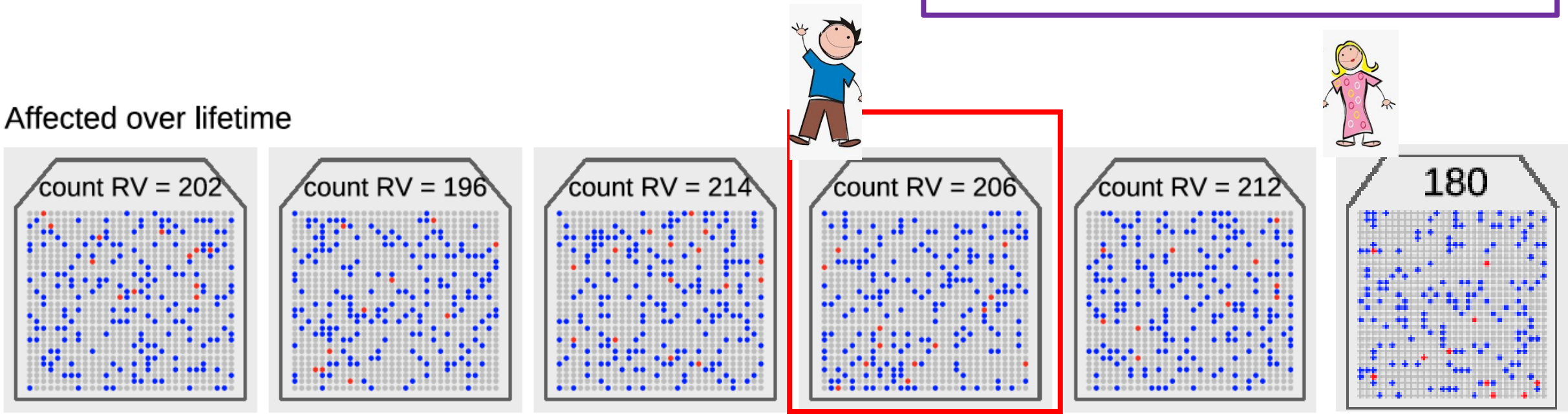
Will people with known family history have high PGS?

Maybe, maybe not!!

JAMA Psychiatry | Review

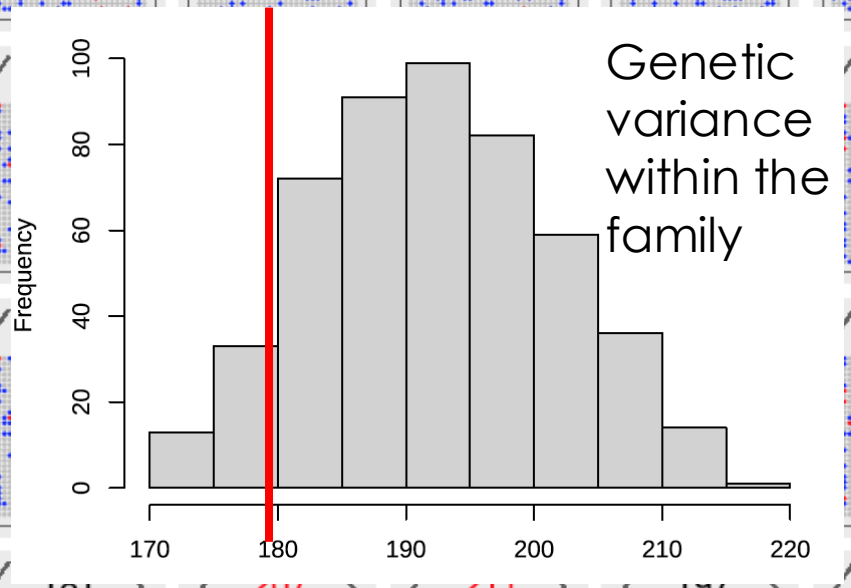
From Basic Science to Clinical Application of Polygenic Risk Scores
A Primer

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)



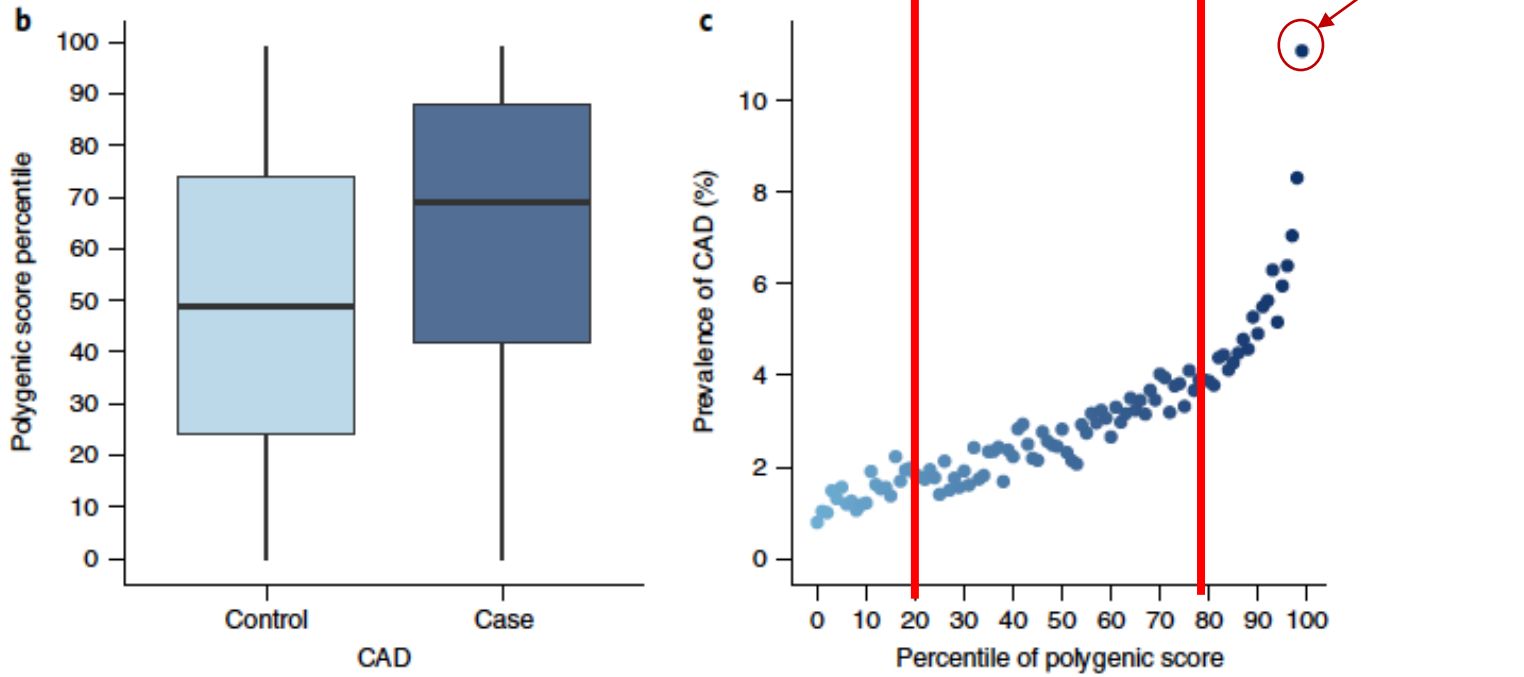
Children of these parents
Mean: 193
+/-3SD: 166-220

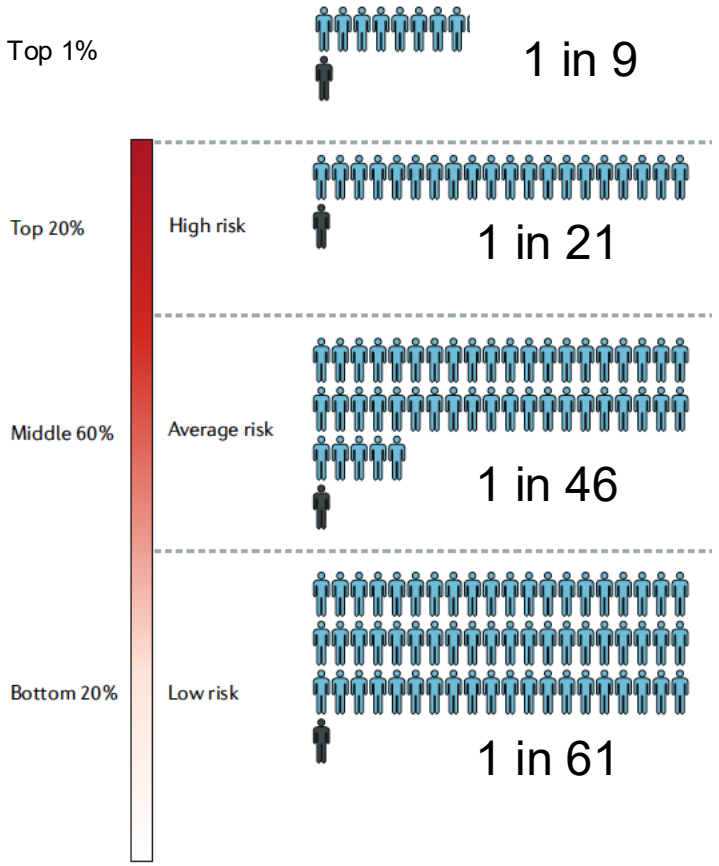
Population
Mean: 180
+/-3SD: 142-218

Utility and applications

How useful is PGS in practice?

Different views of the same data



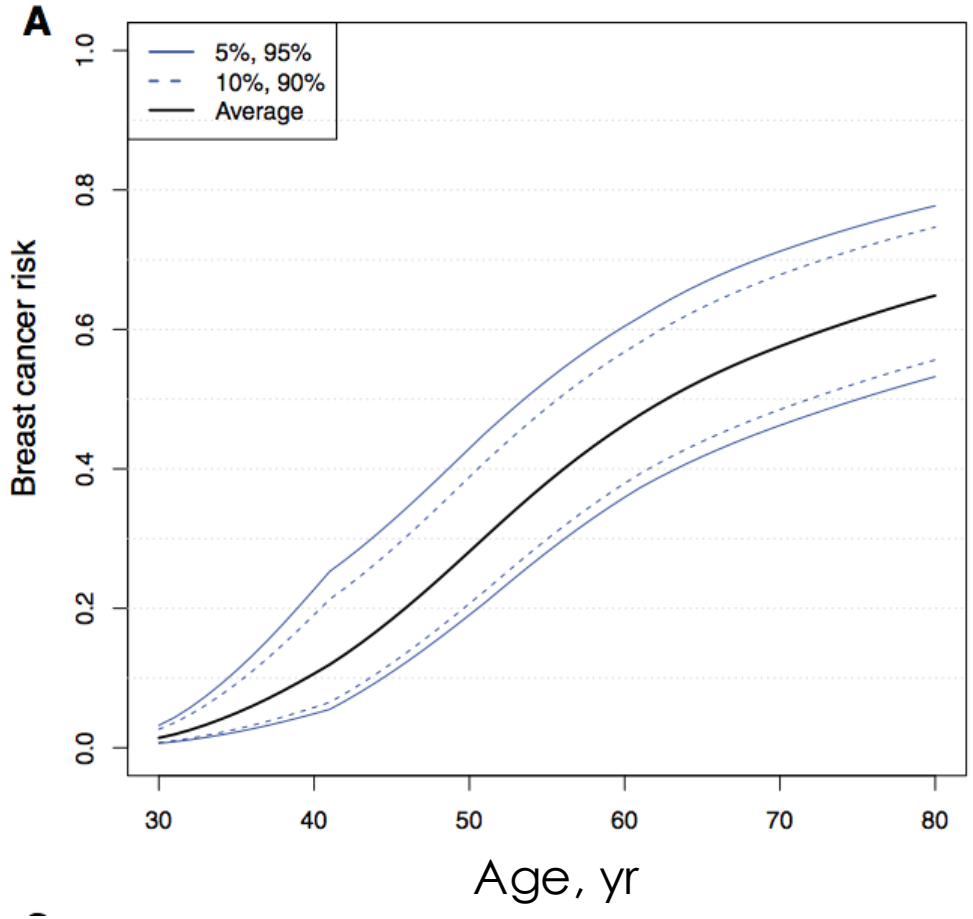


Khera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Torkamani et al, Nat Rev Genetics, 2018

Disease heterogeneity within patients

Combine PRS with known risk mutations
Breast cancer

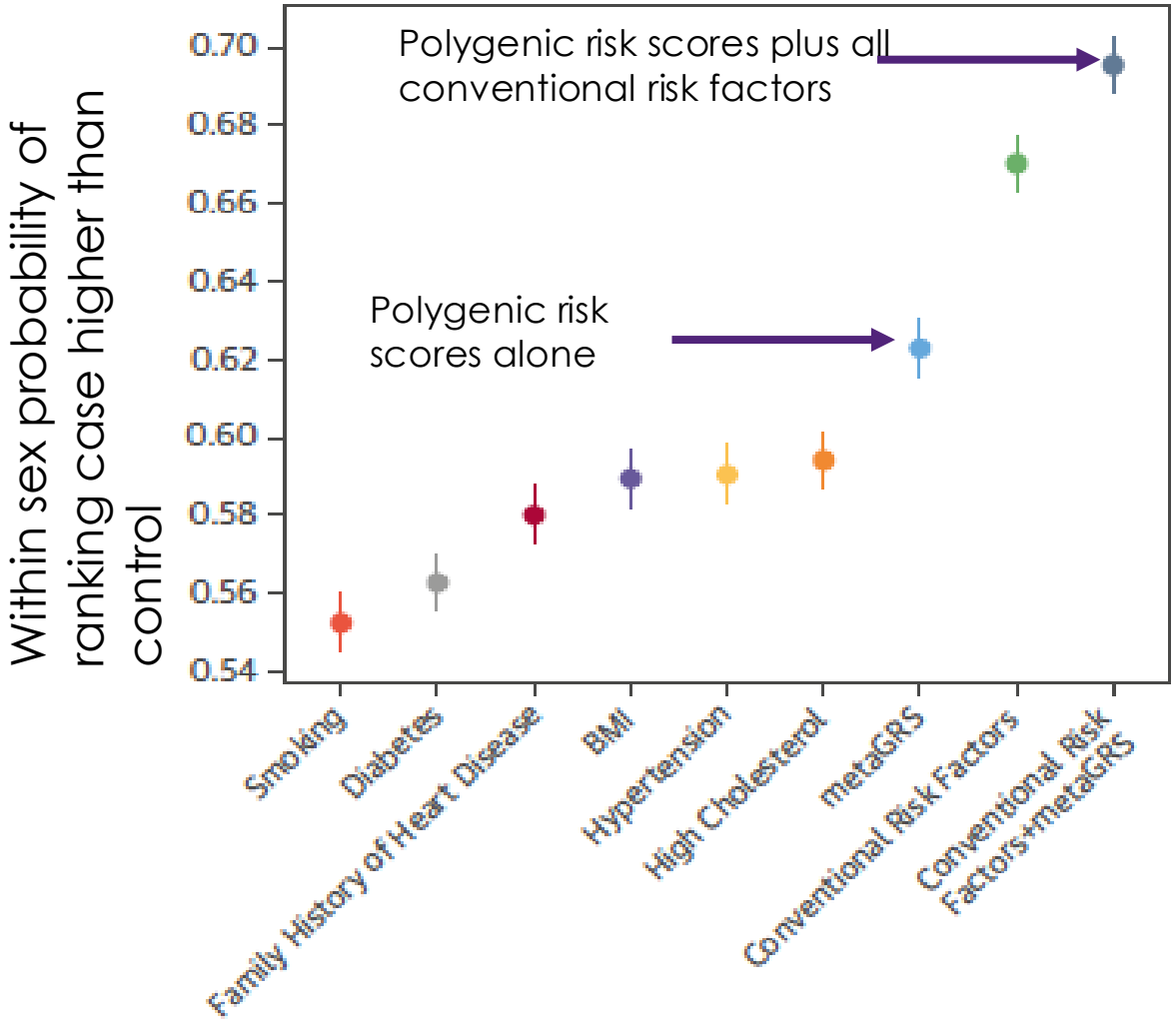


BRCA1 carriers

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)

Increase prediction accuracy

Combine PRS with conventional risk predictors
Coronary Artery Disease



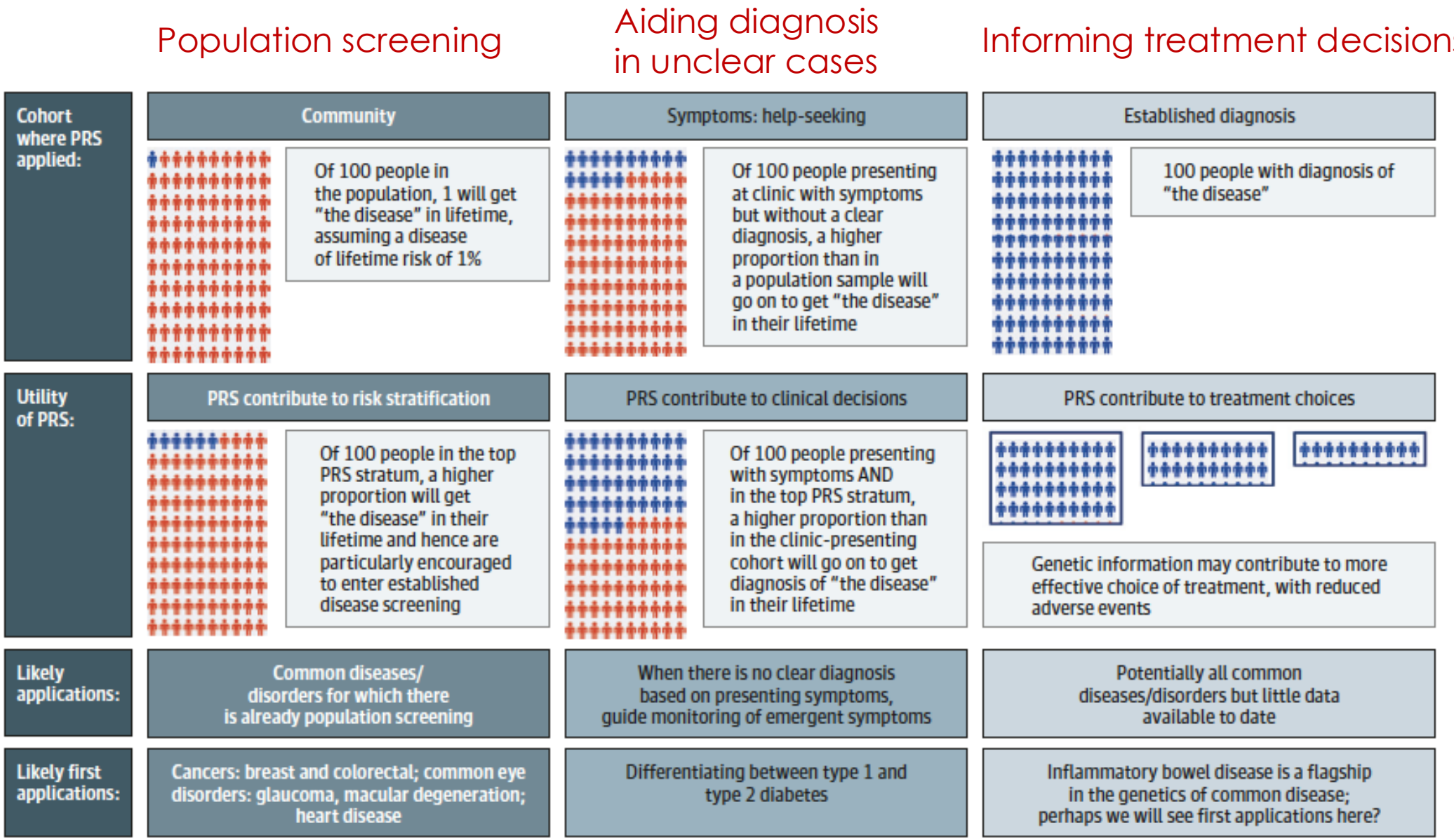
Inouye et al (2018) Genomic risk prediction of CAD in 480K adults. JACC

Clinical 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



Other applications of PGS

Identify correlates of genetic factors

e.g. Educational attainment PGS predicts early speech acquisition and is mediated by cognitive ability (Belsky et al., 2016).

Identify causal effects of genetic factors

Sibling data and family fixed effects → causal effect of PGS

Study GxE

e.g. Increase of compulsory schooling age in U.K. reduces BMI only among those with a high-BMI PGS (Barcellos, Carvalho, and Turley 2016)

Use as control variable

To control for confounding genetic factors or to increase statistical power for estimating the effect of a randomized treatment. If incremental R^2_{PGI} is 15%, then power increase is equivalent to 17% increase in sample size (Rietveld, 2013)

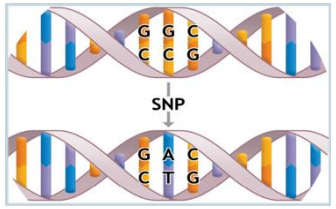
Genomic selection in livestock and crops

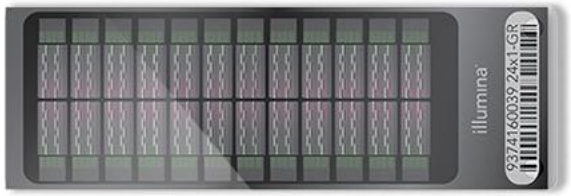
From Aysu Okbay

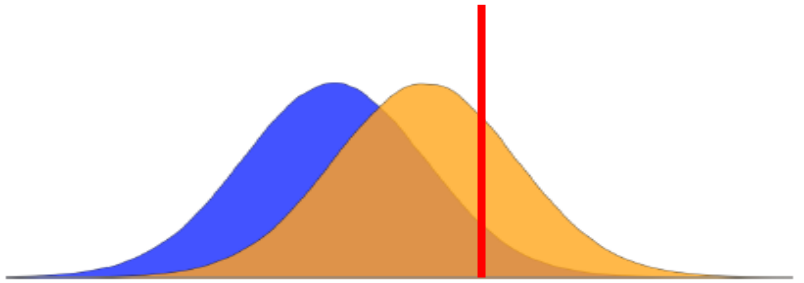
Justify for one disease and the rest come for free!

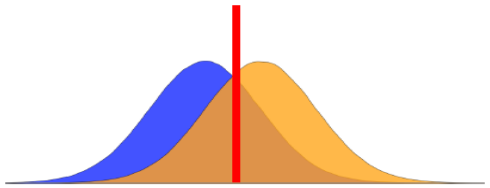
The University Of Queensland AUSTRALIA

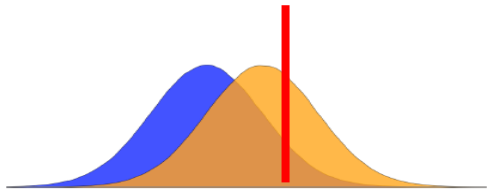
One disease

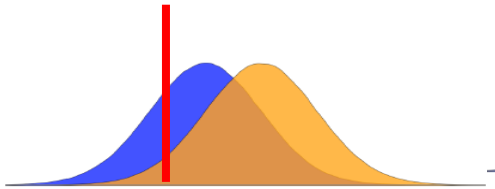


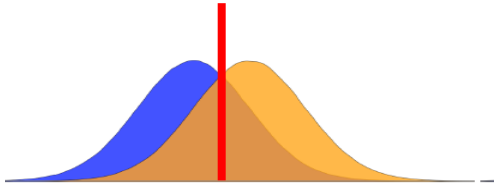


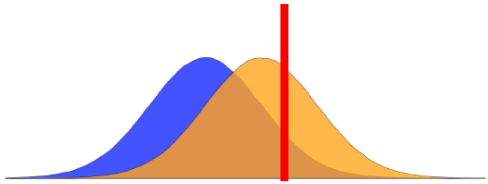


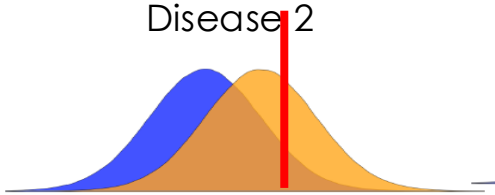

Disease 2

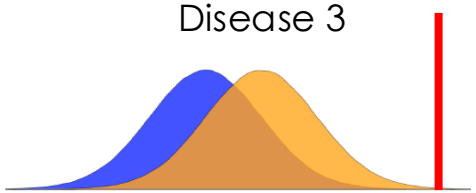

Disease 3

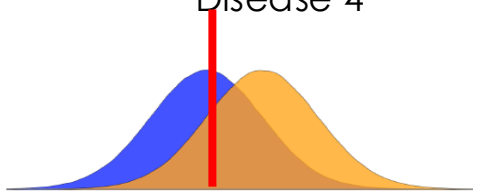

Disease 4

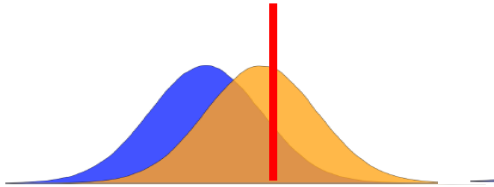

Disease 5

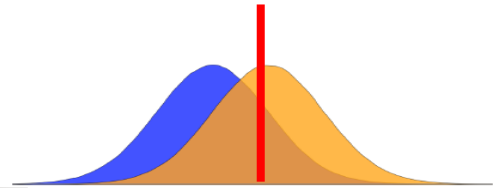

Disease 6

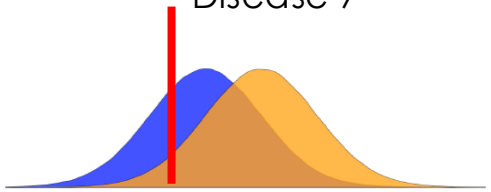

Disease 7

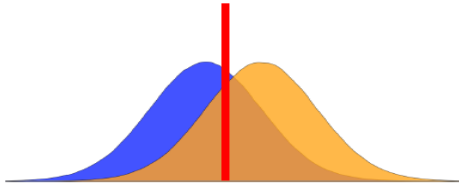

Disease 8

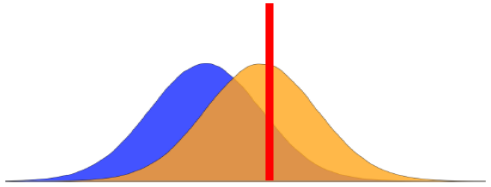

Disease 9

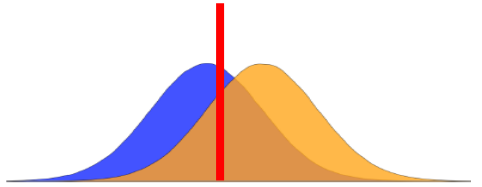

Disease 10

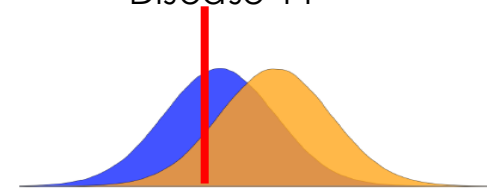

Disease 11


Disease 12


Disease 13


Disease 14


Disease 15


Disease 16

Methodology

Polygenic scores

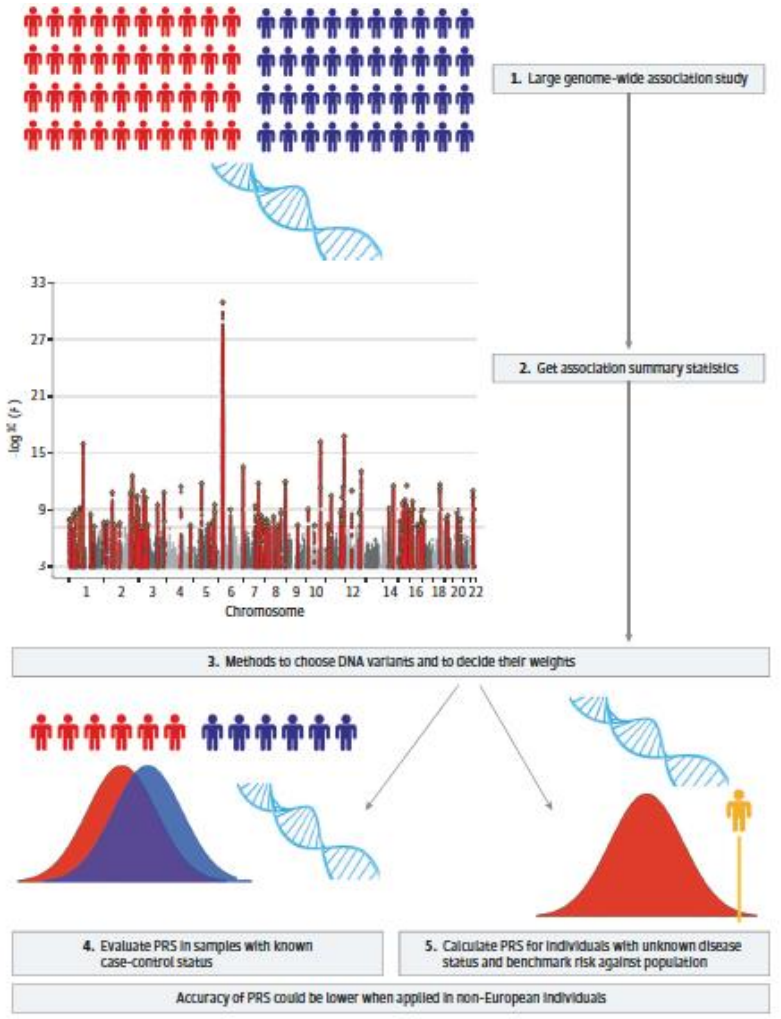
Polygenic score (PGS) is a weighted count of risk alleles

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

0, 1 or 2
Risk alleles

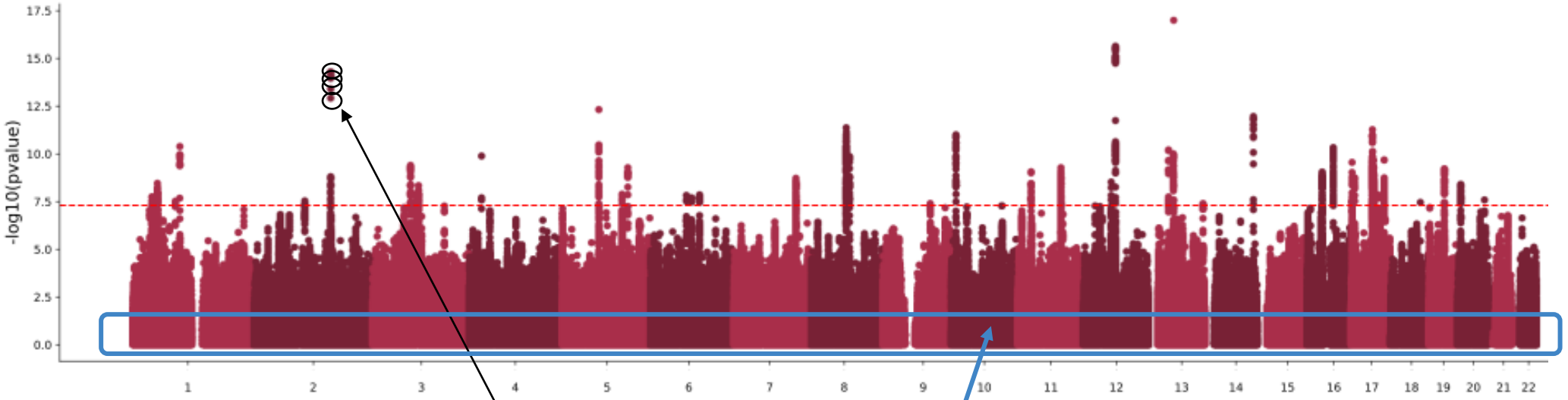
Which SNPs?

What weights?



- Don't need to know causal variants for prediction!
- Prediction can be based on correlated variants.

SNP Weights



GWAS results give us $\hat{\beta}_j^{GWAS}$, not β_j. Two issues to consider when constructing $\sum_{j=1}^{n_{SNP}} \hat{\beta}_j^{GWAS} x_{ij}$:

1. For some SNPs, $\hat{\beta}_j^{GWAS}$ may be a very noisy estimate of β_j and/or β_j may be close to 0, so adding those SNPs will add more noise than signal

2. If we include all SNPs, we will overweight ("double-count") SNPs with high LD scores

From Aysu Okbay

Two solutions

$$\sum_{j=1}^{n_{SNP}} \hat{\beta}_j x_{ij}$$

Clumping and P-value thresholding (C+PT)

Include only the most strongly associated SNP from each LD block (Purcell et al., 2009)

Weights: Set equal to GWAS coefficients.

Loci: Selected by

1. using a **clumping** algorithm that ensures the included SNPs are all approximately independent of each other
2. omitting SNPs whose *P* value for association with the phenotype is above a certain **threshold**

Whole-genome regression approaches

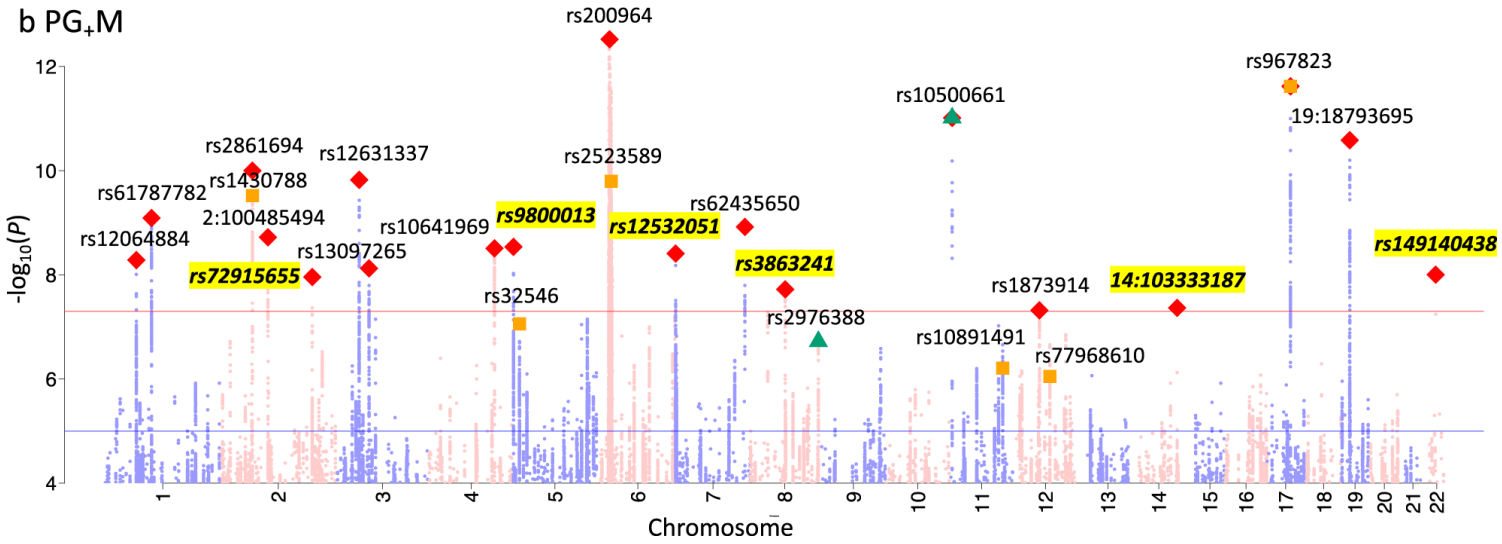
Include all SNPs but adjust the effect sizes for LD

Weights: Set to GWAS coefficients **adjusted for LD** → from a random-effect model regressing the phenotype on all SNPs

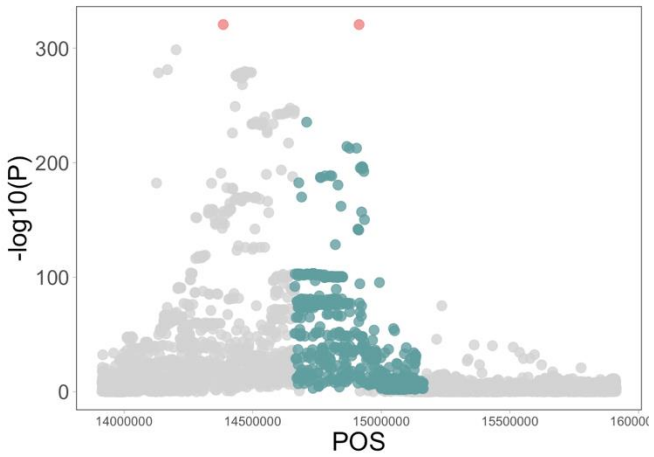
Loci: Include **all SNPs**, no LD-based pruning

Examples: BLUP (Meuwissen et al. 2001), LDpred (Vilhjalmsson et al. 2015, Prive et al. 2020), PRS-CS (Ge et al. 2019), SBayesR (Lloyd-Jones et al. 2019)

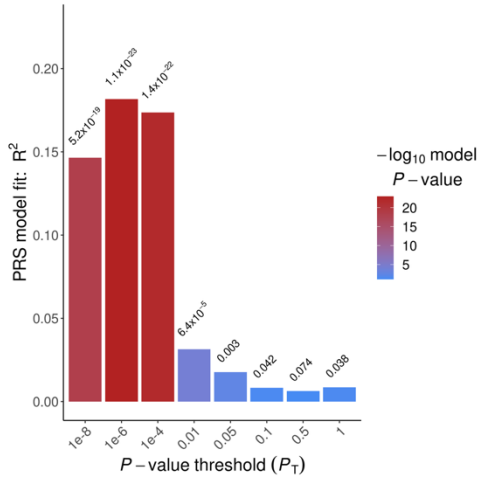
Clumping & P-value thresholding (C+PT, or P+T, C+T)



Step 1. Select most associated SNP in tower (LD-based clumping)



Step 2. Select on a p-value threshold in an independent tuning sample

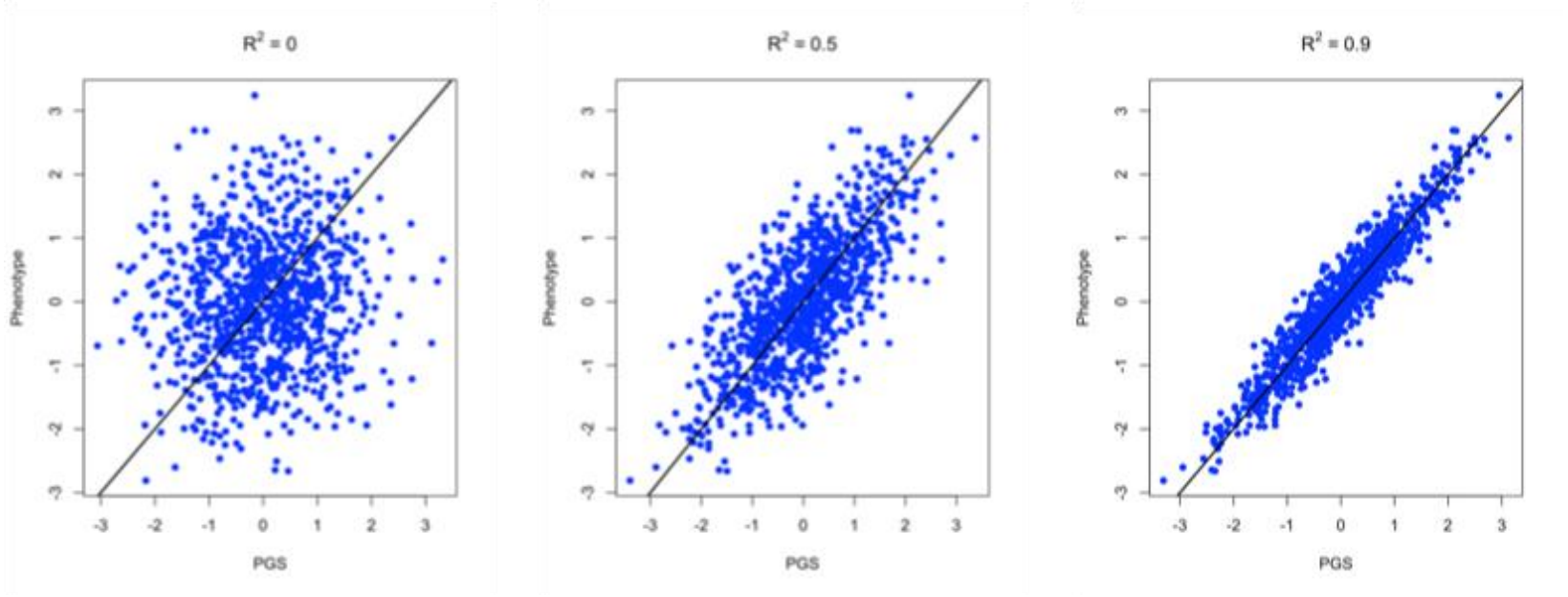


Prediction R^2

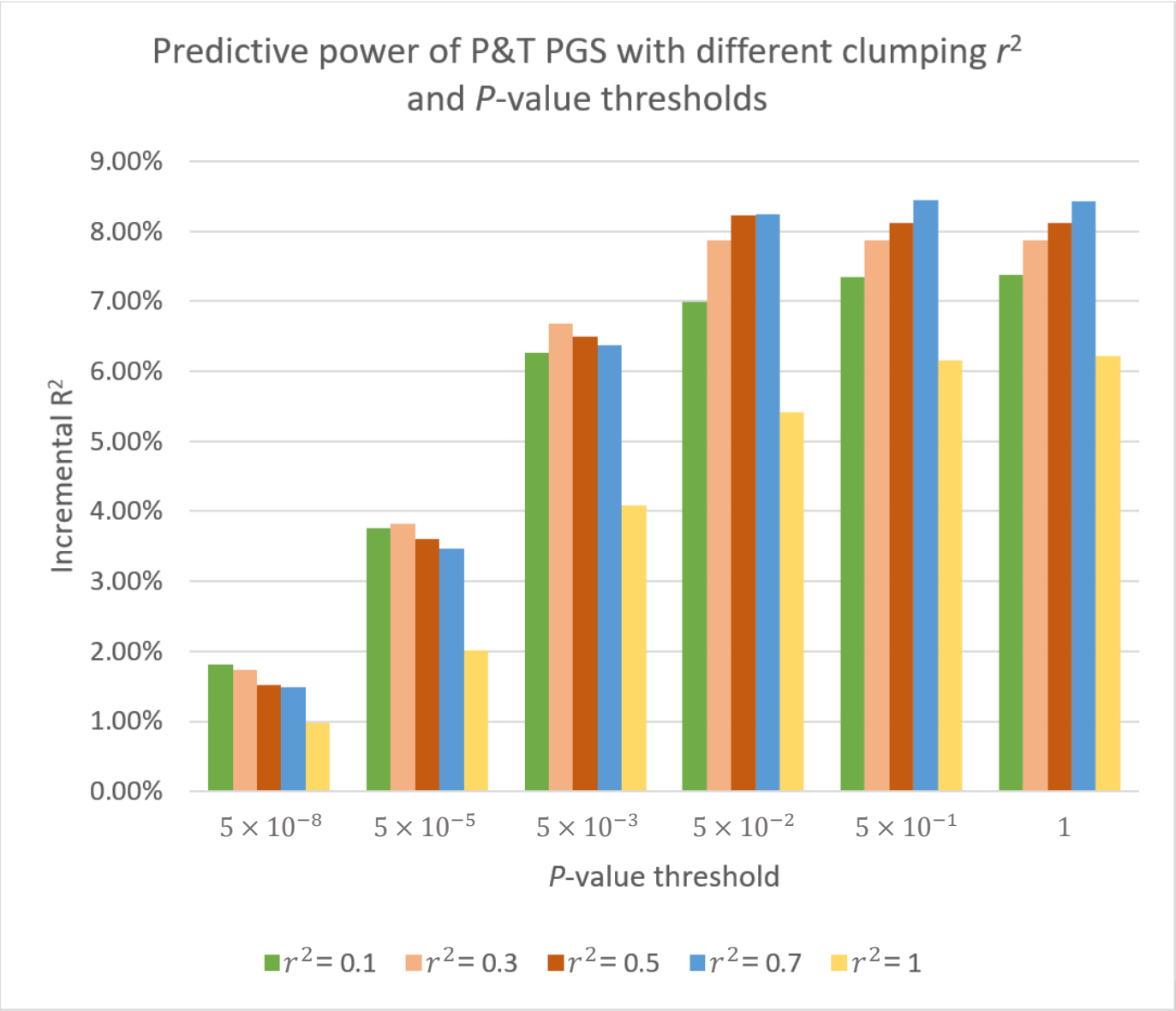
The proportion of phenotypic variance explained by PGS in the validation sample.

Often adjust for covariates (sex, age, and top 10 PCs) to account for differences between groups:

- Null model: y = covariates + e
- Full model: y = covariates + PGS + e
- Incremental R^2: $R^2_{Full} - R^2_{Null}$



Clumping & P-value thresholding (C+PT, or P+T, C+T)



- **Cohort:** Health and Retirement Study
- **Phenotype:** Educational attainment

From Aysu Okbay

More advanced methods

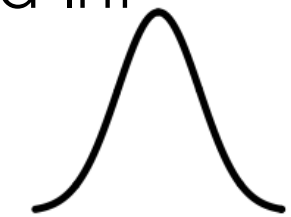
A weighted sum of the count of risk alleles

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

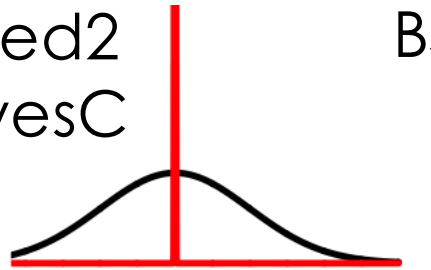
How many SNPs?
Which SNPs?
What weights?

New methods model genetic architecture

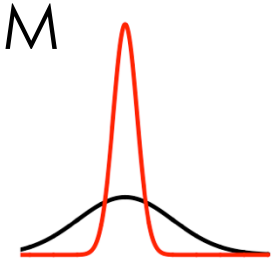
LDpred-Inf
SBLUP



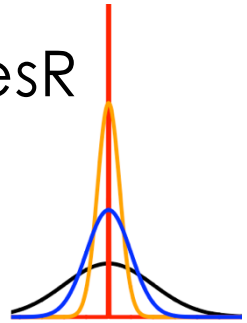
LDPred2
SBayesC



BSLMM



SBayesR



Polygenic risk score methods

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

Table 1. Summary of Methods Used to Generate Polygenic Scores

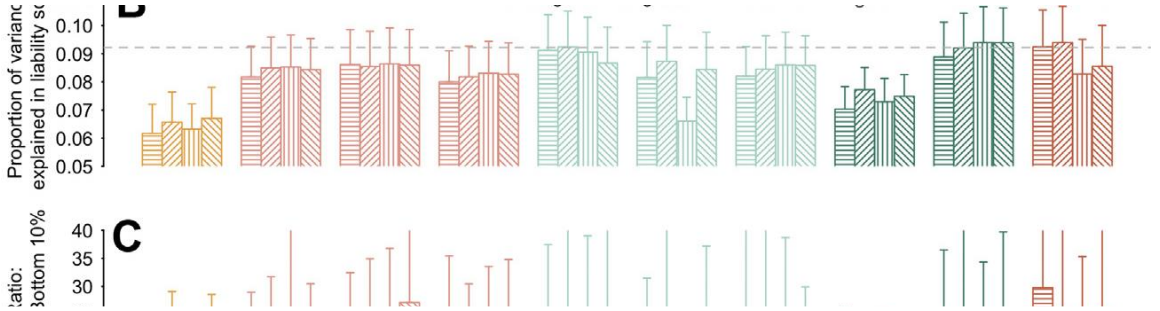
Method	Distribution of SNP Effects (β)	Tuning Sample	Predefined Parameters	Parameters Estimated in Tuning Sample
PC+T	None	Yes	–	*p*-value threshold
SBLUP	$\beta \sim N\left(0, \frac{h_g^2}{m}\right)$ h_g^2: SNP-based heritability, m: number of SNPs; $\lambda = m(1 - h_g^2)/h_g^2$	No	λ LD radius in kb	–
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	$f(\beta) = \mathbf{y}^T\mathbf{y} + (1-s)\beta^T\mathbf{X}_r^T\mathbf{X}_r\beta - 2\beta^T\mathbf{X}^T\mathbf{y} + s\beta^T\beta + 2\lambda\|\beta\|_1^1$ $\mathbf{X}_r$: $n \times m$ 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	$\beta_j \mid \pi, \sigma_\beta^2 \sim \begin{cases} 0, & \text{with probability of } \pi_1 \\ N(0, \gamma_2\sigma_\beta^2), & \text{with probability of } \pi_2 \\ \vdots \\ N(0, \gamma_C\sigma_\beta^2), & \text{with probability of } 1 - \sum_{c=1}^{C-1}\pi_c \end{cases}$ $\sigma_\beta^2 \sim Inv - \chi^2(d.f. = 4)$ $\pi_i \sim Dir(\mathbf{1})$, estimated from discovery GWAS in SBayesR software γ_i 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_i and γ_i 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

Summary

- Polygenic scores are imperfect but useful genetic predictors.
- Their accuracy is fundamentally limited by heritability, SNP set, and sample size.
- A high PRS is mostly a consequence of genetic sampling.
- PRS have the potential to differentiate risk between family members who have the same family history information.
- Being evaluated in clinical settings and are often combined with other predictive measures to predict the total disease risk.
- C+PT is a simple but commonly used method to calculate PGS, which involves SNP selection and requires an independent tuning sample.

PGS are not …

- Not diagnostic and never will be.
- Not absolute risk and do not provide a baseline or timeframe for the progression of a disease.
- Not and never will be stand-alone predictors of common diseases.
- Not equally applicable across populations – at least not yet.

Recommended reading

1. Wray NR, *et al*. Complex Trait Prediction from Genome Data: Contrasting EBV in Livestock to PRS in Humans: Genomic Prediction. *Genetics*. 2019 Apr;211(4):1131-1141. (**Review of polygenic prediction in livestock and humans**)

2. Wray NR, *et al*. From Basic Science to Clinical Application of Polygenic Risk Scores: A Primer. *JAMA Psychiatry*. 2021 Jan 1;78(1):101-109. (**Review of clinical application**)

3. Khera AV, *et al*. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet*. 2018 Sep;50(9):1219-1224. (**Demonstration of utility**)

4. Euesden J, *et al*. PRSice: Polygenic Risk Score software. *Bioinformatics*. 2015 May 1;31(9):1466-8. (**Popular tool implements the C+PT method**)

5. Martin AR, *et al*. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat Genet*. 2019 Apr;51(4):584-591. (**Poor cross-ancestry portability and consequences**)

Questions?

~ Gentle reminder to sign the signing sheet if you haven't ~

5 min break

Practical 1: Computation of PRS using C+PT

https://cnsgenomics.com/data/teaching/GNGWS26/module5/Practical1_Basic_method.html

To log into your server, type command below in **Terminal** for Mac/Linux users or in **Command Prompt** or **PowerShell** for Windows users.

```
ssh username@hostname
```

And then key in the provided password.