Genome-Wide Association Studies (GWAS) – #4

Allan McRae
a.mcrae@uq.edu.au

Meta-Analysis

We need very large sample sizes to detect associations with variants of small effect in GWAS

It is rare to have a large enough cohort to detect a large number of variants

- particularly for disease case control analysis

We can use a meta-analysis to combine results from a number of studies to effectively increase our sample size

Common approach for international consortia

Why Not One Big Analysis?

- Privacy
- Ethics
- Population stratification

Meta-Analysis

Done by pooling:

- Genetic effect of a SNP on a phenotype
- P-value of the association test

Many available approaches to perform a meta-analysis

Inverse Variance Weighted Method

Assumptions:

- There is one underlying ‘true’ effect
- All deviations of sample effects from the ‘true’ effect are due to chance

Prerequisites:

- Same scale must be used across samples
- Same reference allele on same strand

Inverse Variance Weighted Method

Pooled effect estimate

$$\beta_{pooled} = \frac{\sum_{i=1}^{N} (w_i * \beta_i)}{\sum_{i=1}^{N} (w_i)}$$

$$w_i = \frac{1}{\mathrm{var}(\beta_i)}$$

Standard error

$$\mathrm{var}(\beta_i) = se\ (\beta_i)^2$$

$$se_{pooled} = \sqrt{\frac{1}{\sum_{i=1}^{N} (w_i)}}$$

Inverse Variance Weighted Method

Does the assumption of equal effect size hold up?

Cochran's Q statistic → Test of homogeneity

$$Q = \sum_{i=1}^{N} w_i (\beta_i - \beta_{pooled})^2$$

χ^2-distributed with df=k-1 (k = number of samples)

Inverse Variance Weighted Method

Possible causes of heterogeneity related to bias in samples:

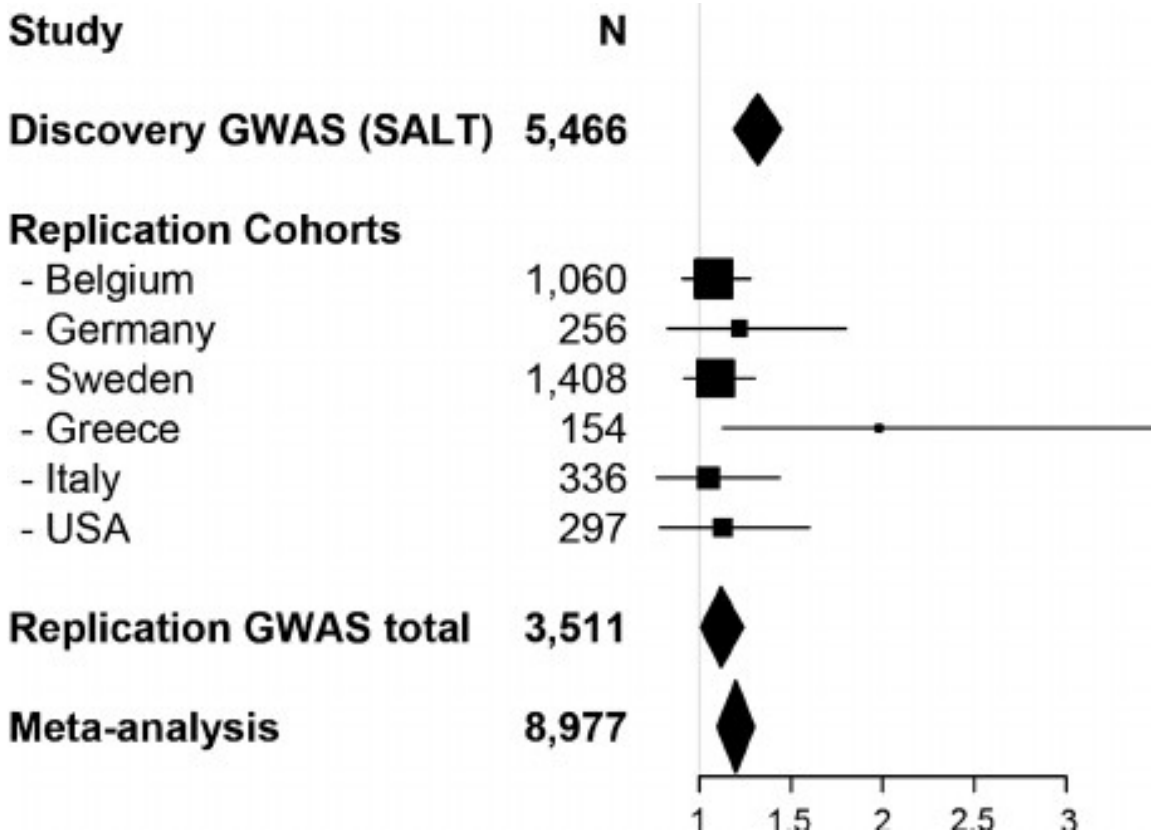
- Differential selection of cases and controls
- Poor genotyping
- Poor genotype data cleaning
- Different SNP platforms (imputed vs. observed SNPs)
- Poor/differential phenotyping
- Population stratification

Possible causes from genuine differences across samples:

- Different LD structure across populations
  (truly associated SNP vs. tested SNP)
- Different correlations of phenotypes across populations
  (truly associated phenotype vs. tested phenotype)

Meta-Analysis Example

Forrest plot of association between irritable bowel syndrome and marker rs12702514



Polygenic Risk Prediction

Marker loci associated with highly significant additive effects on the character can be included in a net molecular score, m, which for any individual is the sum of the additive effects on the character associated with these markers. Use of the net molecular score,

Genetics, 1990

Efficiency of Marker-Assisted Selection in the Improvement of Quantitative Traits

Russell Lande* and Robin Thompson†

Many Names...

PRS	Polygenic Risk Score
PGS	PolyGenic Score
GEBV	Genomic Estimated Breeding Values

Toy Example

G= risk allele
A = reference allele

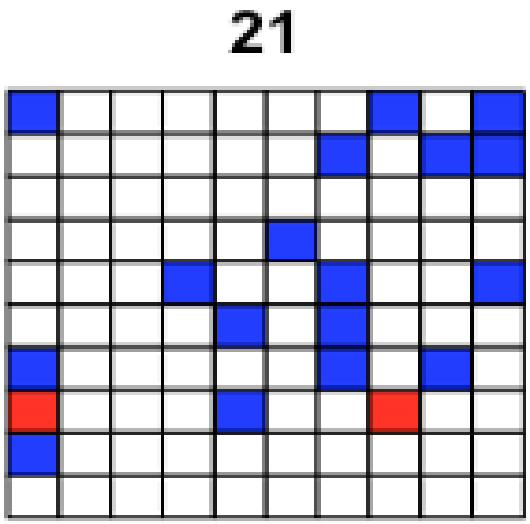
p = frequency of risk allele in the population
10%

0 risk alleles
Homozygote AA
p^2
81%

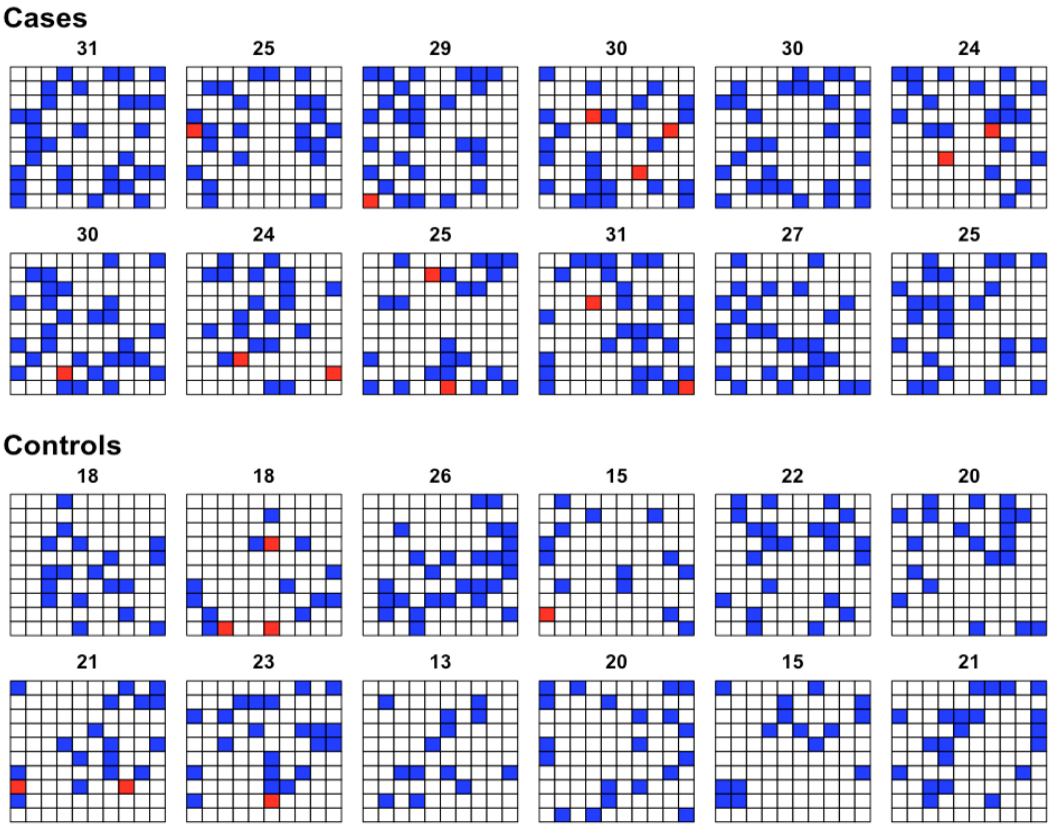
1 risk allele
Heterozygote AG
$2p(1-p)$
18%

2 risk alleles
Homozygote GG
$(1-p)^2$
1%

Imagine…
a disease with 100 contributing risk alleles



Visualising Polygenic Risk



Not all affected individuals carry the risk allele at any particular locus

Unaffected individuals carry many risk loci

Each affected person carries a unique portfolio

Cases carry more risk variants on average

PRS Schematic

Case-control Discovery

GWAS association results

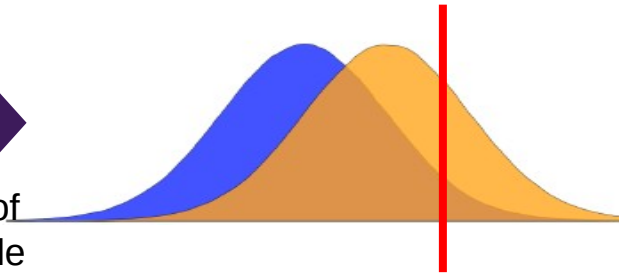
Identify risk loci and estimate effect sizes

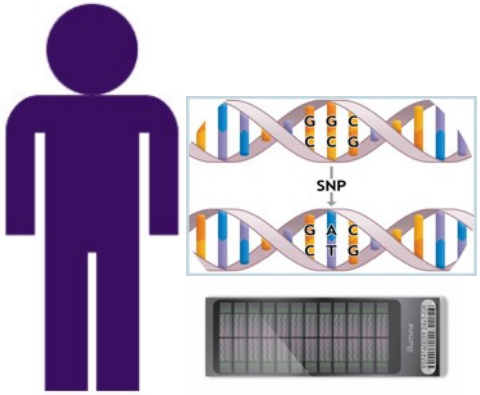
Case-control Target

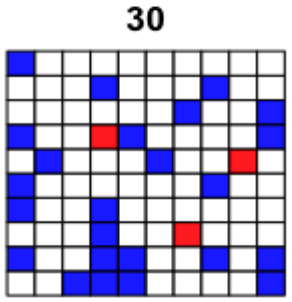
Genome-wide genotypes

Weighted sum of discovery-sample identified risk alleles

Evaluate







Calculating PRS

The PRS is a weighted sum of SNP effects:

$$PRS = \sum w_i * \beta_i x_i$$

w_i = Weights

β_i = Effect from GWAS

x_i = SNP genotype (0, 1, 2)

PRS Weights

Weighting based on p-value:

$$w_i = \begin{cases} 1; p < threshold \\ 0; p > threshold \end{cases}$$

Using a threshold well below GWAS significance ($5 * 10^{-8}$) generally results in better predictors

Could also include “functional” information

- up-weight regions of the genome that are annotated to have a biological function

Handling Linkage Disequilibrium

Ignoring LD in PRS calculations results in worse predictors

Effects in regions of high LD get included in PRS multiple times, while those in low LD do not

“Clumping” SNPs to get semi-independant dataset

- take most significant SNP
- remove all SNPs with LD $r^2 > 0.1$
- take next most significant SNP
- ....

Quantifying PRS Accuracy

Quantitative traits:

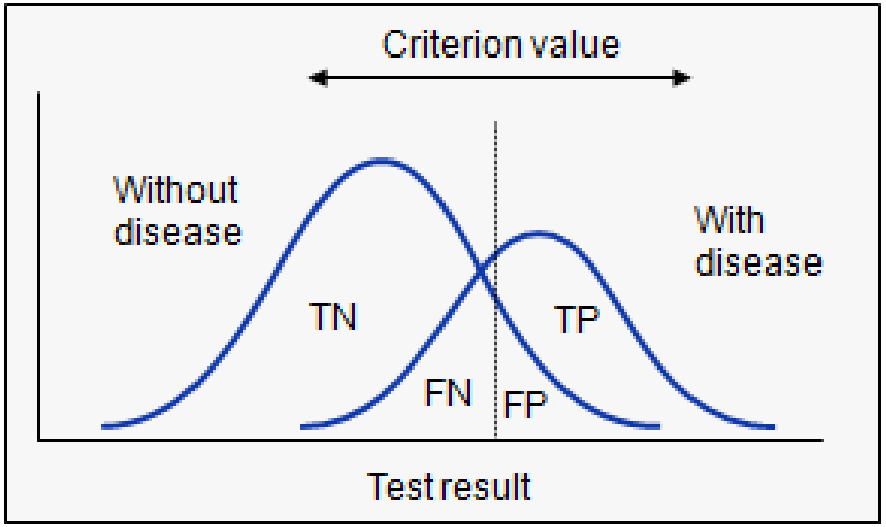
- Proportion of variance of trait PRS explains in target population
- Correlation between PRS and trait squared (R^2)

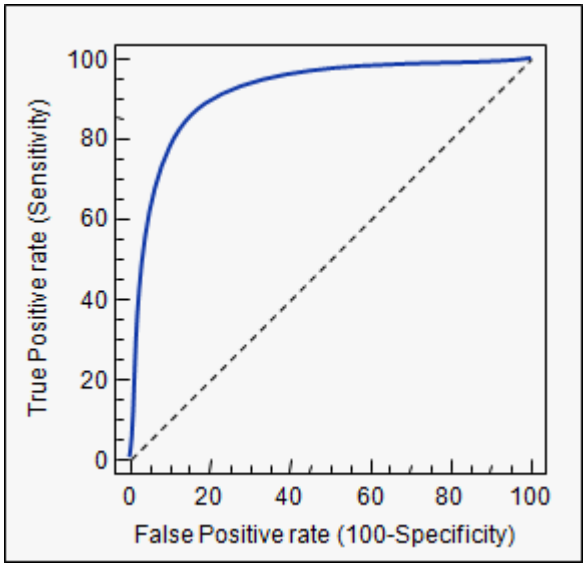
Binary (case/control) traits:

- Nagelkerke R^2
  - Affected by case/control ratio in test cohort
- Area Under the Curve (AUC)

Area Under the Curve

Receiver Operating Characteristics (ROC) Curve





For disease traits, can be interpreted as the probability that a randomly chosen case has a higher score than a randomly chosen control

PRS Accuracy Depends On...

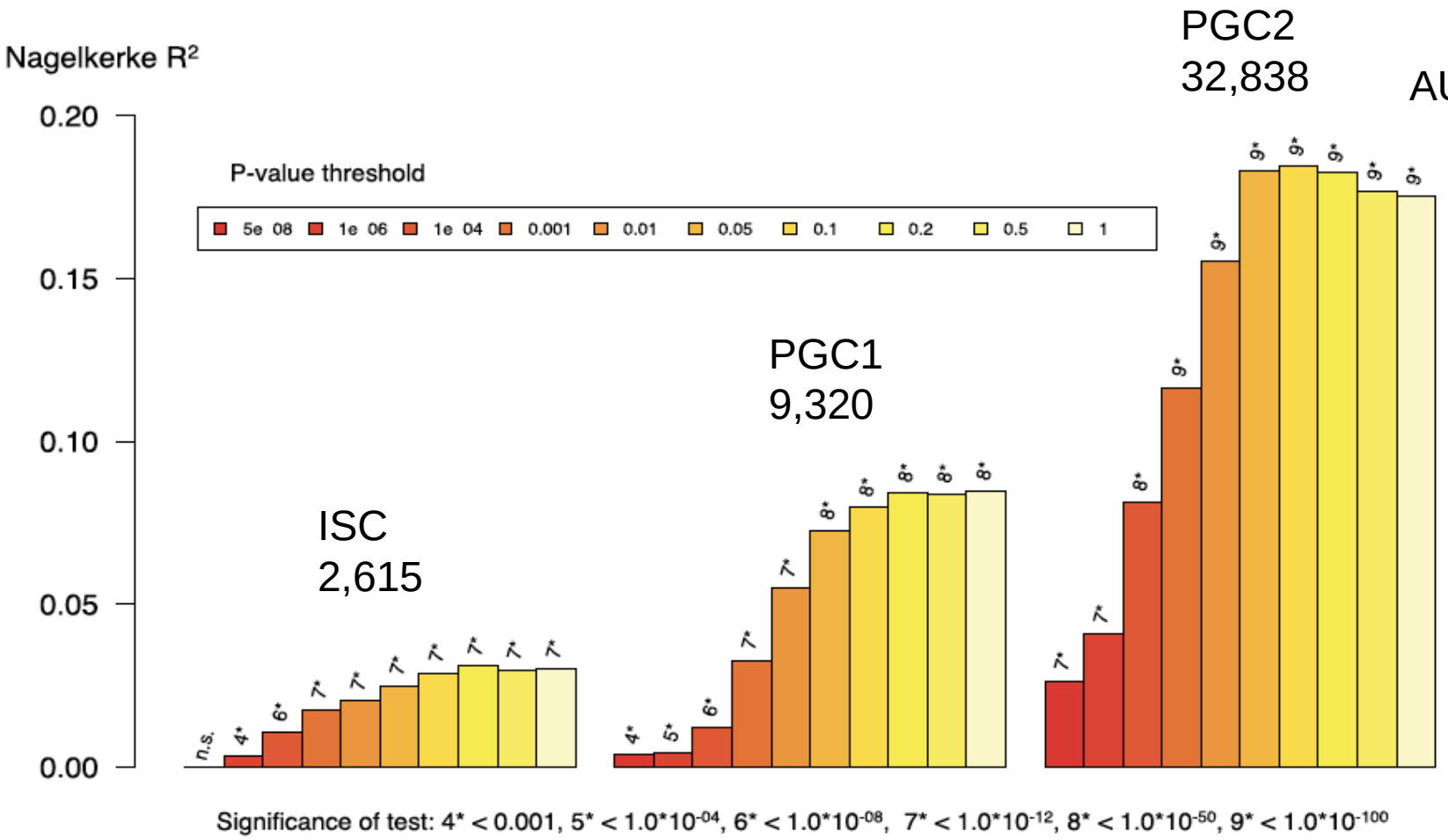
The genetic architecture of a disease
- How many risk loci
- How big the effect sizes
- Relative contribution of genetic factors to risk compared to non-genetic factors

The sample size for detecting risk loci

How well our technology tracks genetic risk factors

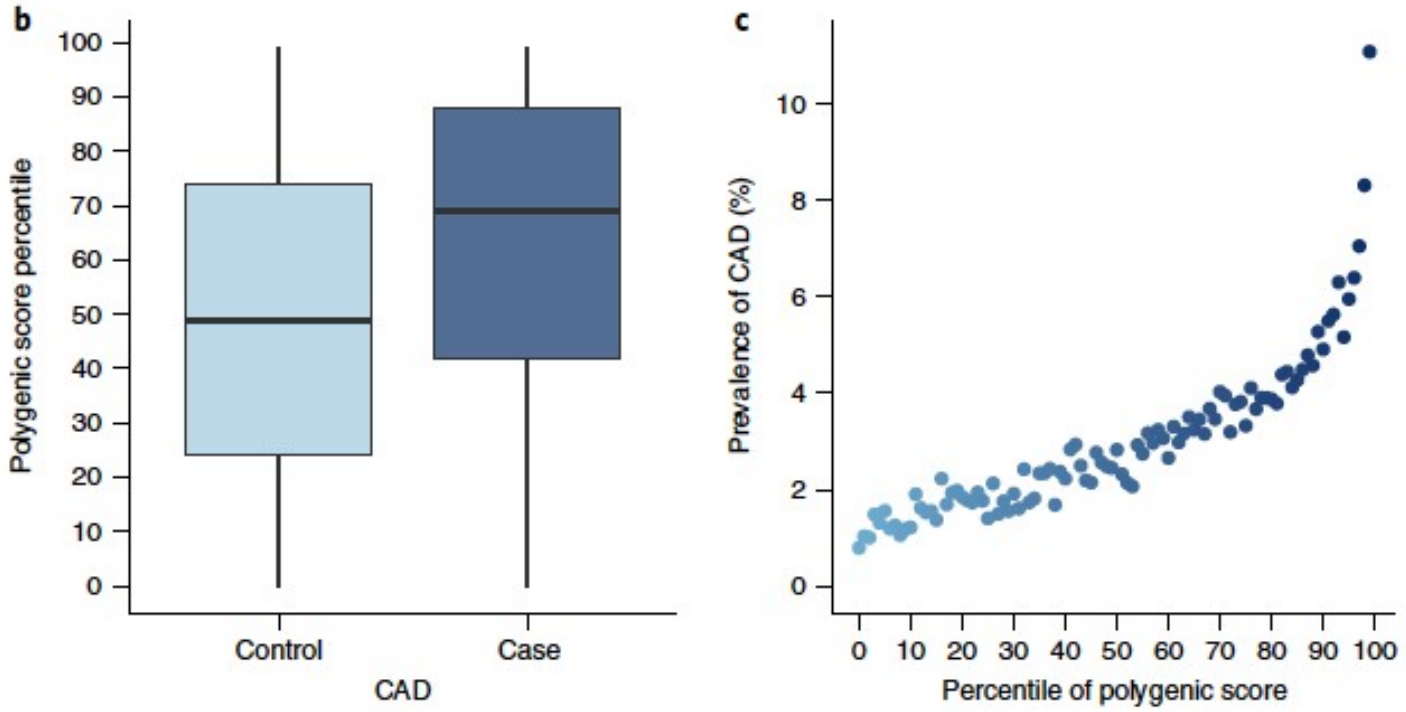
The methodology that optimises risk prediction likely depends on genetic architecture, which is different for different diseases.

Schizophrenia



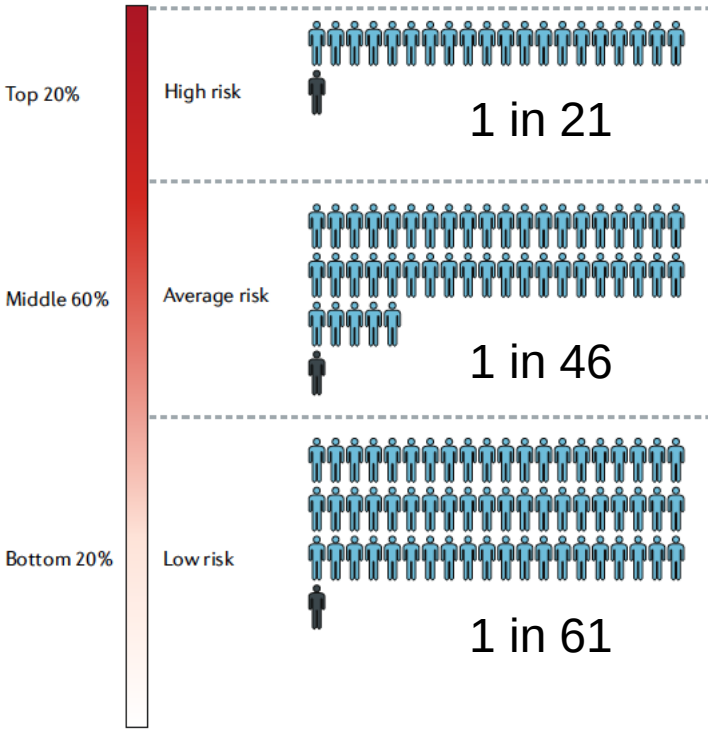
PGC-SCZ 2014 Biological insights from 108 schizophrenia-associated genetic loci

Coronary Artery Disease



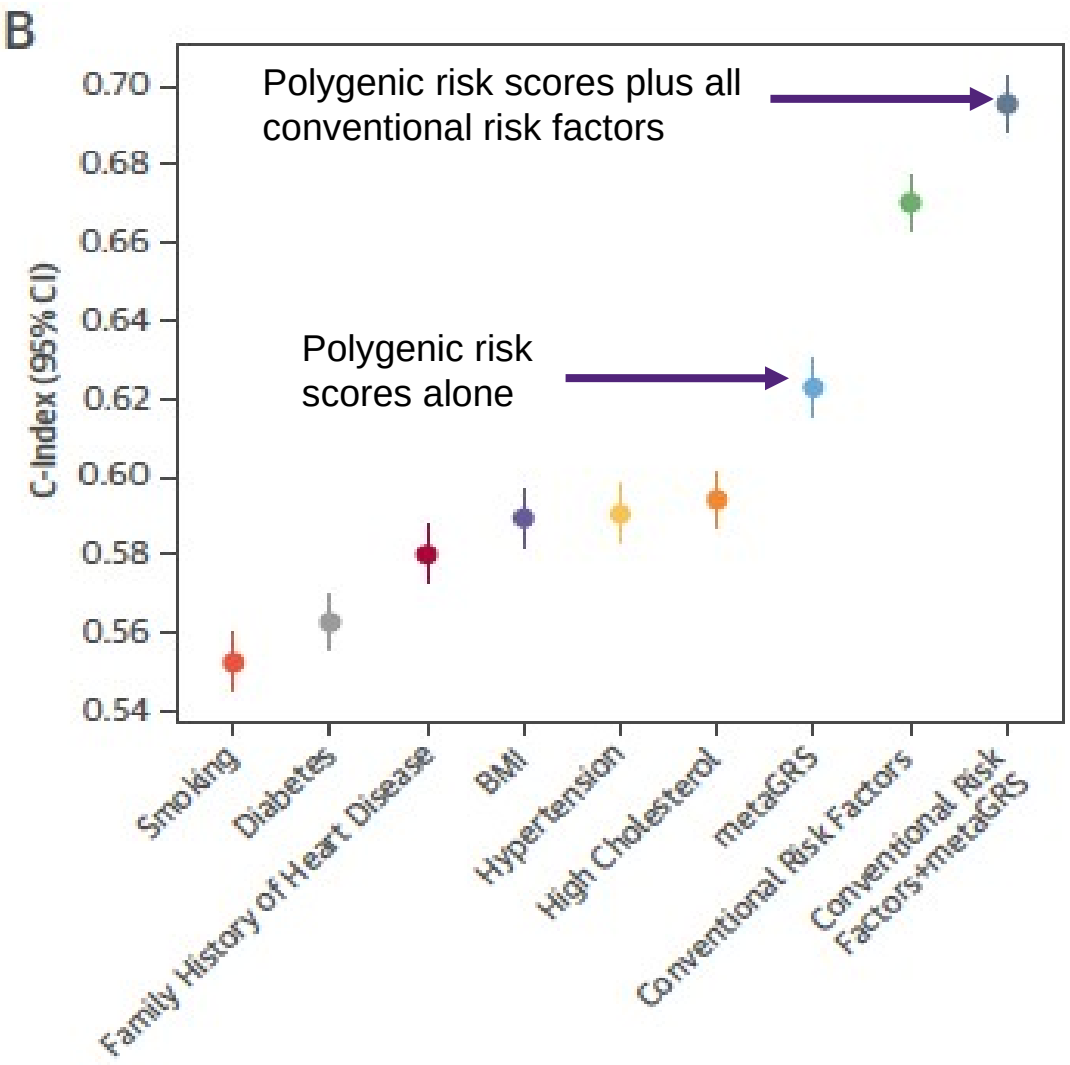
Kheera et al (2018) Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. Nature Genetics

Coronary Artery Disease



Torkamani *et al.*,
Nat Rev Genetics, 2018

Coronary Artery Disease



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

PRS Across Ancestries?

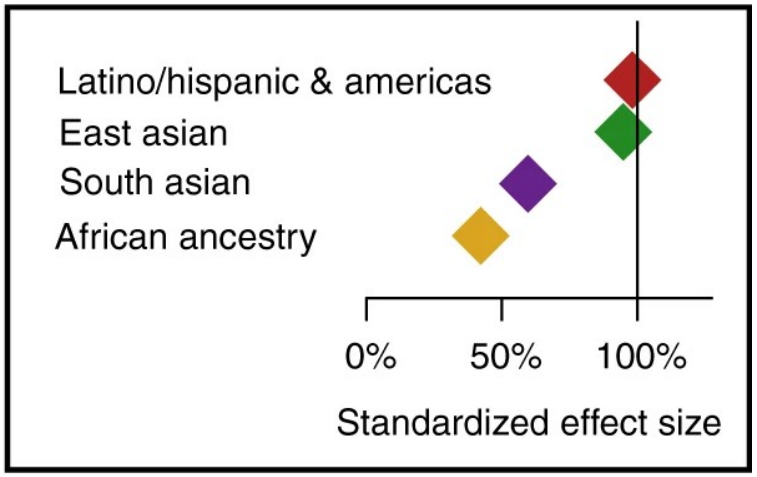
Most (~2/3) GWAS studies are performed in European populations, followed by East Asians (~1/4).

PRS developed in European populations tend to have poor performance in other ancestral groups

- Different LD structure (clumping)

- Different allele frequencies

- Ancestry specific genetic effects?



Duncan et al., Nature Communications, 2019