Polygenic Risk Scores

Dr Paul O'Reilly
(www.pauloreilly.info @paul_f_oreilly)

Senior Lecturer in Statistical Genetics
SGDP Centre, IoPPN, KCL

MSc Genomic Medicine, IoPPN, 19th February 2016

Outline of Lecture

- Recap of GWAS

Outline of Lecture

- Recap of GWAS
- Background to polygenic score methods:
  - From discovery to prediction
  - Aggregated predictions for group-level inference

Outline of Lecture

- Recap of GWAS
- Background to polygenic score methods:
  - From discovery to prediction
  - Aggregated predictions for group-level inference
- The Polygenic Risk Score (PRS) method:
  - Step-by-step guide to the PRS method
  - Some applications of PRS
  - Reading List: Recent history of the PRS method in 4 papers

Outline of Lecture

- Recap of GWAS
- Background to polygenic score methods:
  - From discovery to prediction
  - Aggregated predictions for group-level inference
- The Polygenic Risk Score (PRS) method:
  - Step-by-step guide to the PRS method
  - Some applications of PRS
  - Reading List: Recent history of the PRS method in 4 papers
- *Appendix: Some work by my group*

Genes discovered that decide whether we will be tall or short

By DAILY MAIL REPORTER
UPDATED: 11:06, 30 September 2010

Comments (13) | Share

Scientists have pinpointed why some people shoot up to become superstar basketball players while others struggle to reach 5ft.

The genetic finding is a giant stride towards explaining why comedian John Cleese towers over his friend Ronnie Corbett.

While obesity is caused by a mix of genetic and environmental factors, about 80 per cent of variation in human height is determined by our genes.

Scientists Identify Gene Responsible For Tooth Growth

February 26, 2009

Print repost

Scientists have reported new insights gathered from a single gene that could one day be used to help adults grow a new set of teeth.

Scientists from the University of Rochester bred mice that lacked the oddskipped related-2 (Osr2) gene. They

Scientists Link Gene Mutation to Autism Risk

Kevin P. Casey for The New York Times

Members of a team conducting autism studies: from left, Evan E. Eichler, Dr. Jay A. Shendure and Brian O'Roak of the University of Washington.

By BENEDICT CAREY
Published: April 4, 2012 | 522 Comments

Teams of scientists working independently have for the first time identified several gene mutations that they agree sharply increase the chances that a child will develop autism. They have found further evidence that the risk increases with the age of the parents, particularly in fathers over age 35.

RECOMMEND
TWITTER
LINKEDIN
COMMENTS (522)

LIVE BBC NEWS CHANNEL

Page last updated at 18:35 GMT, Sunday, 6 December 2009

E-mail this to a friend Printable version

Obesity gene discovery 'may cut cases blamed on abuse'

Scientists have discovered what they believe is a genetic cause of severe obesity in children.

The team concluded that the loss of a key segment of DNA can be to blame.

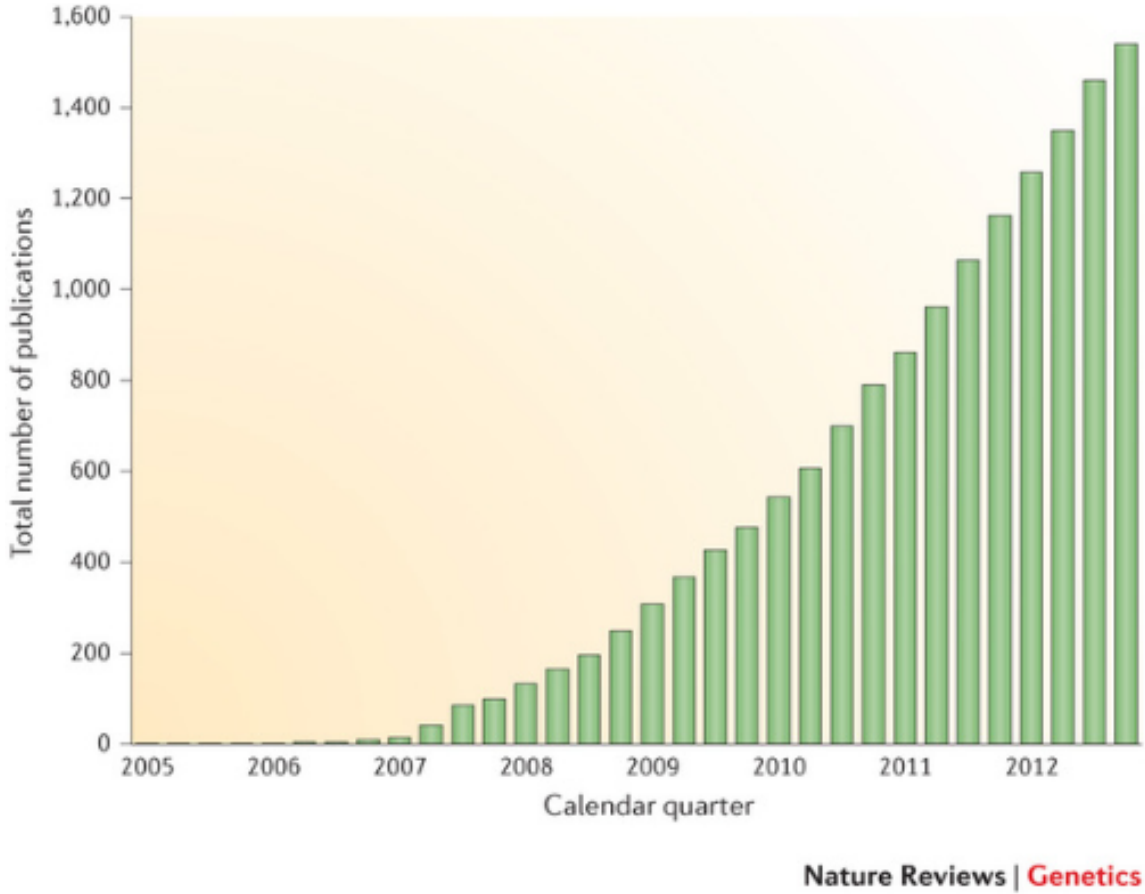
It said the findings might improve diagnosis of severe obesity - which on occasion has been wrongly attributed to abusive overfeeding.

Child obesity levels have been rising for decades

SEE ALSO
- Child obesity 'is levelling off'
  03 Nov 09 | Health
- GPs 'struggle with child obesity'
  01 Sep 09 | Health
- Child obesity 'may harm thyroid'
  04 Dec 08 | Health
- Obesity 'spreads among the young'
  31 Jul 09 | Health

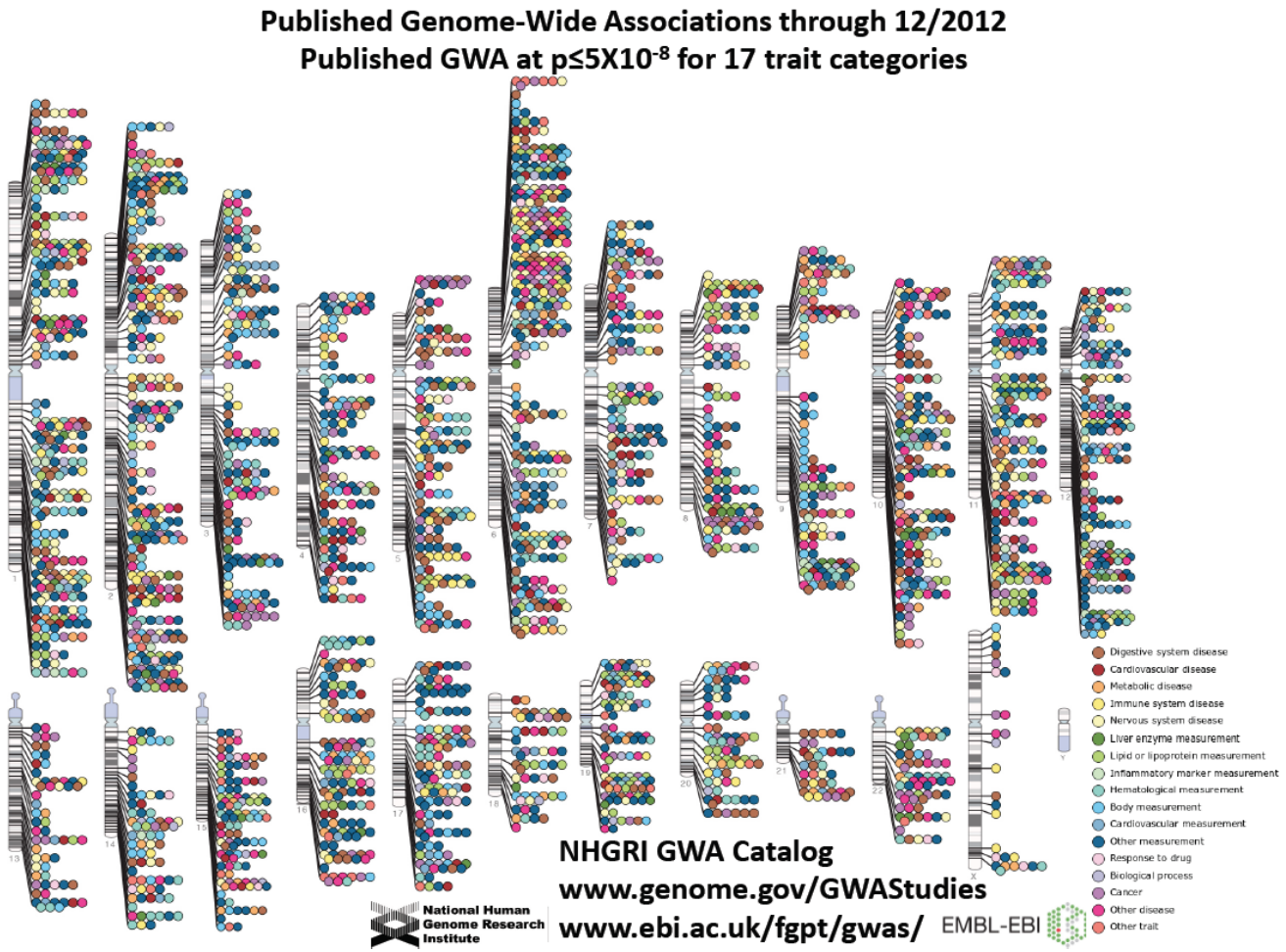
RELATED BBC LINKS
- BMI calculator

The GWAS era: Publications



- The number of GWAS publications has grown exponentially in the last decade, and as a result the field of genetics has become dominated by GWAS.

The GWAS era: Discovered variants

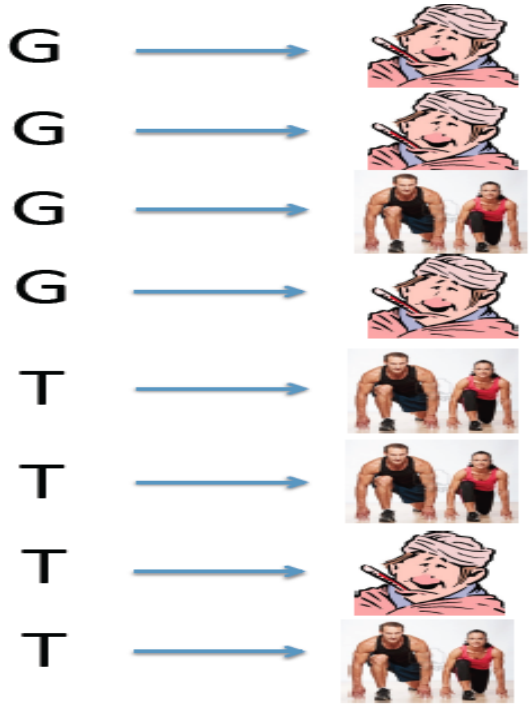


- The genetic variants discovered by GWAS, affecting numerous human diseases and traits, cover the entire human genome.

Association testing

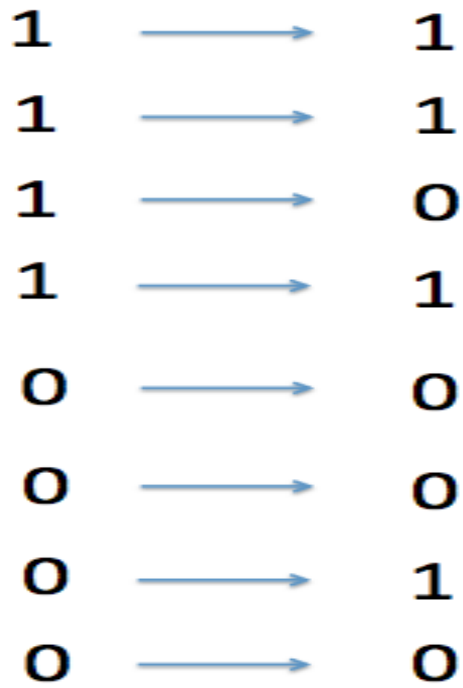
- Smoking has a strong *effect* on many disease outcomes
- The effect can be estimated by testing the association between smoking and the disease under study

Association testing



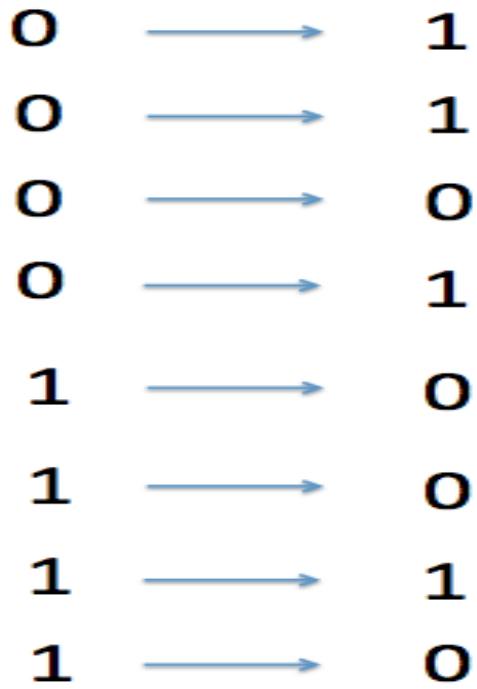
- Different alleles at a SNP can also have an effect on diseases/disorders
- Genetic Association Studies test the association between the alleles/genotypes of a SNP and a trait of interest

Association testing



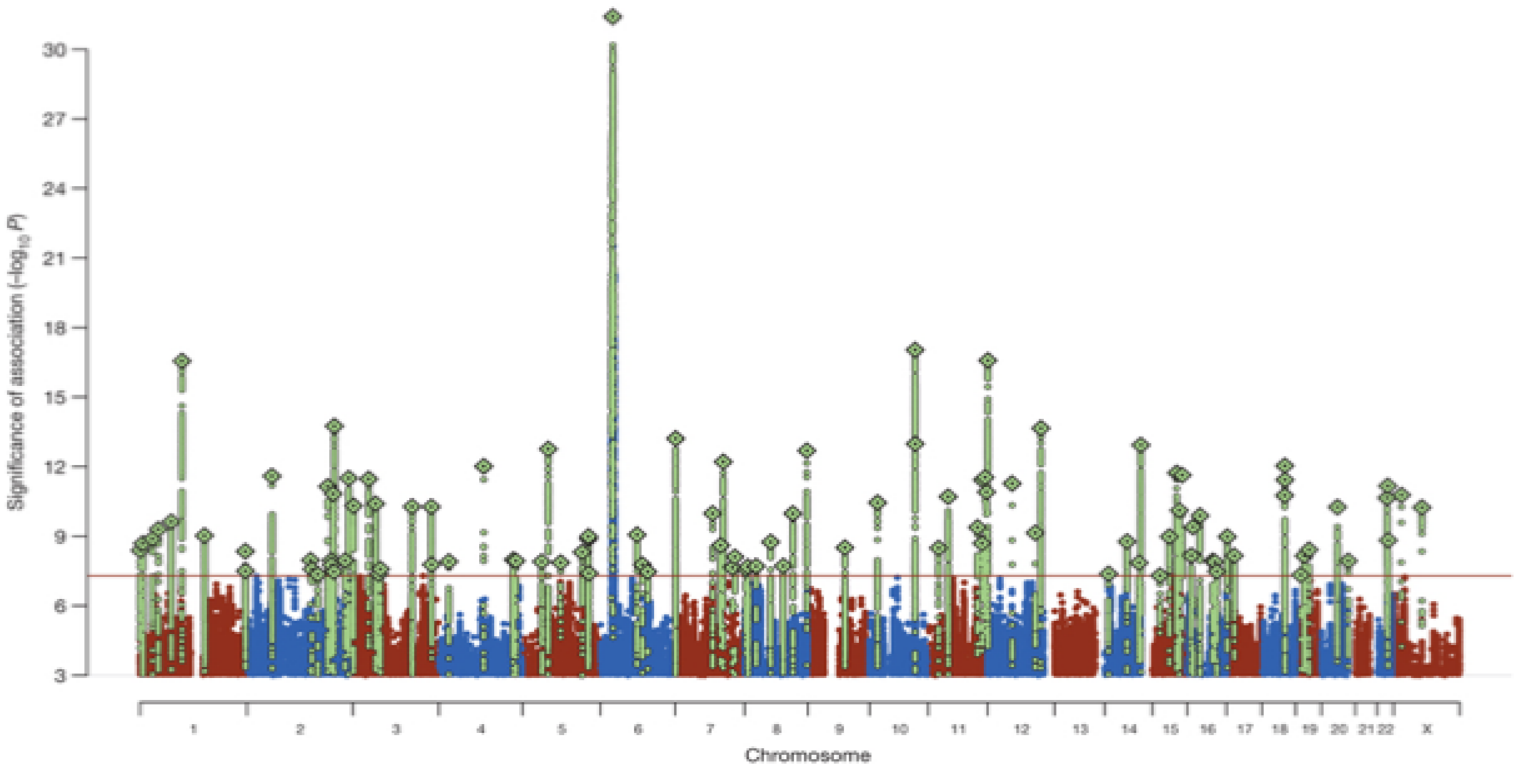
- To perform these association tests we need to first code the information of the variables numerically
- Generally alleles are coded as 0 and 1, and genotypes are coded 0, 1, 2.

Association testing



- It makes no difference how the coding is done here in terms of the association test result
- However, genotypes are coded in a certain way to reflect the *inheritance model*

GWAS results



- Manhattan plot illustrating GWAS results (on Schizophrenia) across all SNP across the genome

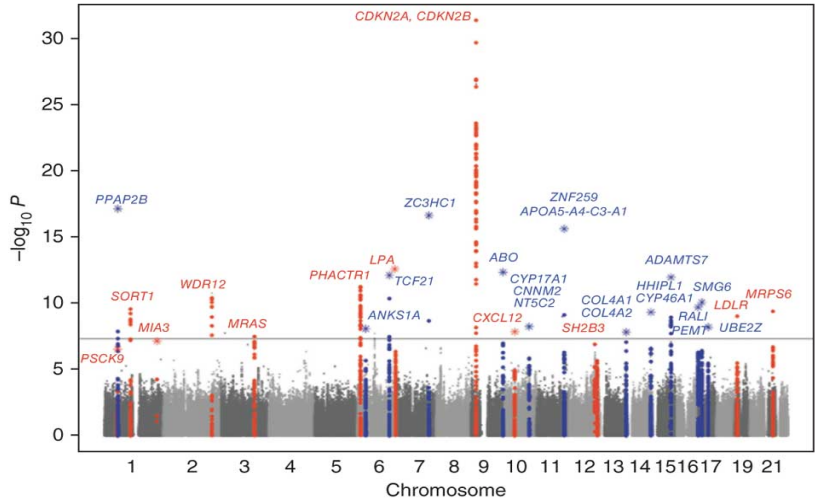
Going beyond GWAS

- A major ambition of the genomic era is to use genetics to predict what diseases and disorders different people may get later in life
- This could initiate *personalised/precision medicine*
- We can use genetics to predict human traits such as height, blood pressure and cholesterol levels, as well as IQ, alcohol consumption and extraversion - so it is not all about disease

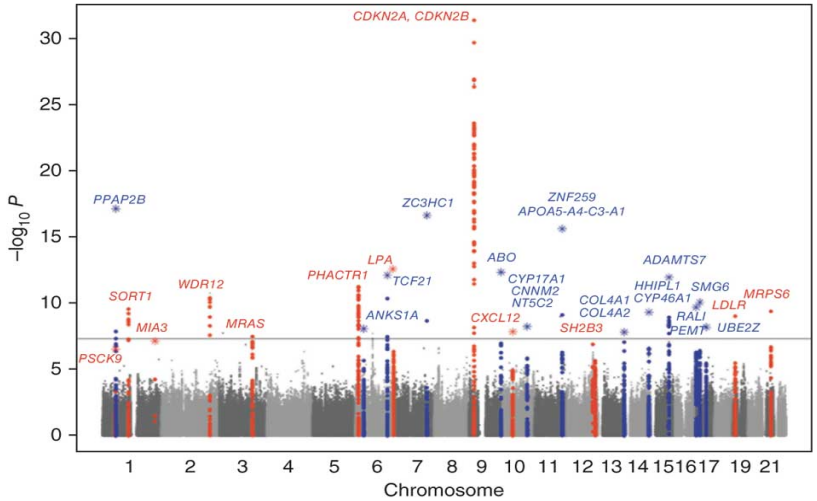
Moving from discovery to prediction

Moving from discovery to prediction

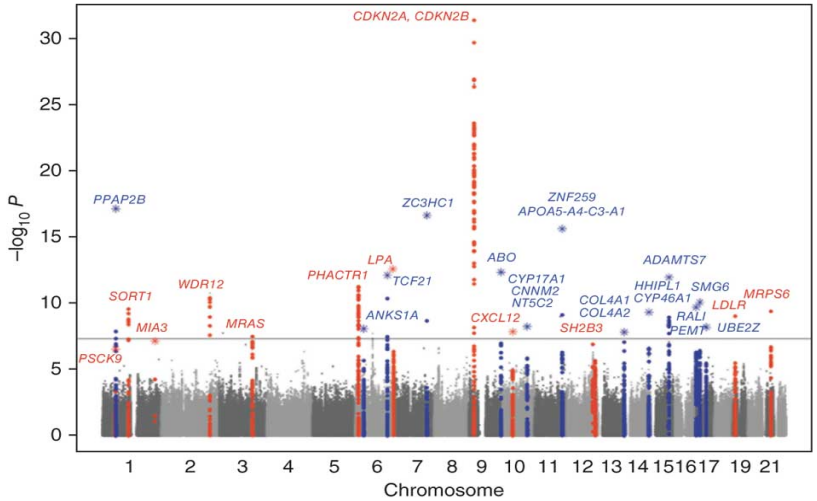
Moving from discovery to prediction



Moving from discovery to prediction

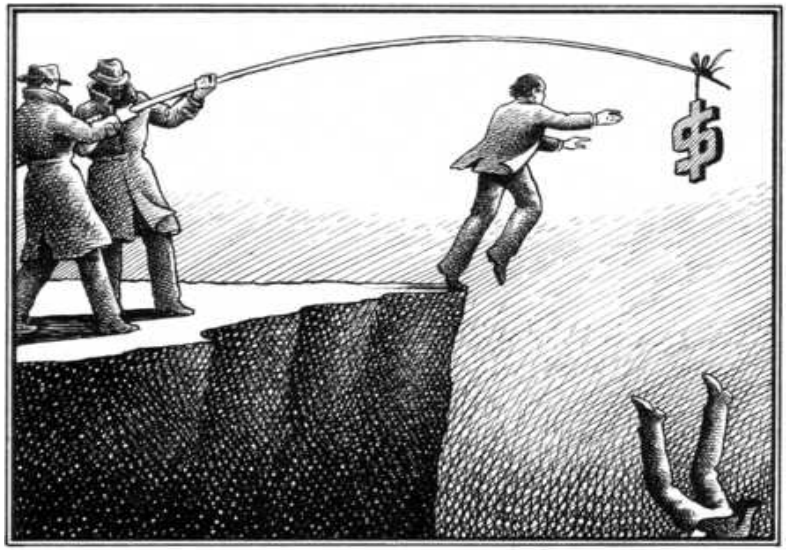


Moving from discovery to prediction

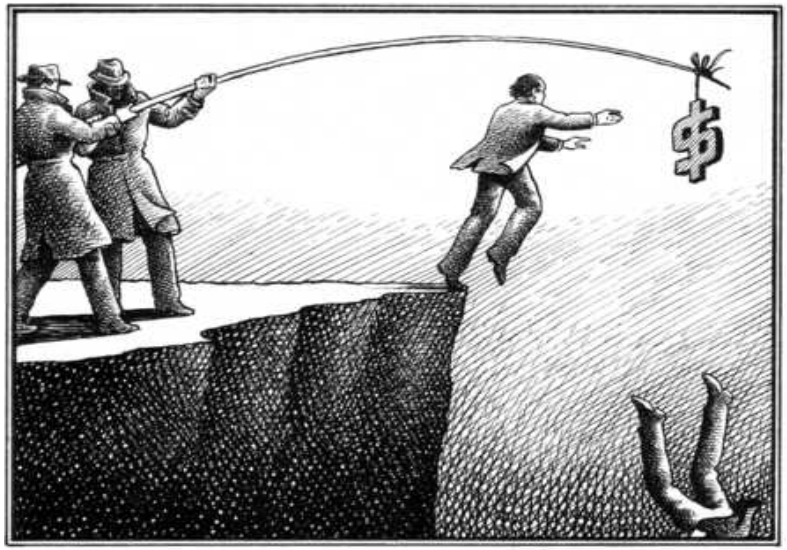


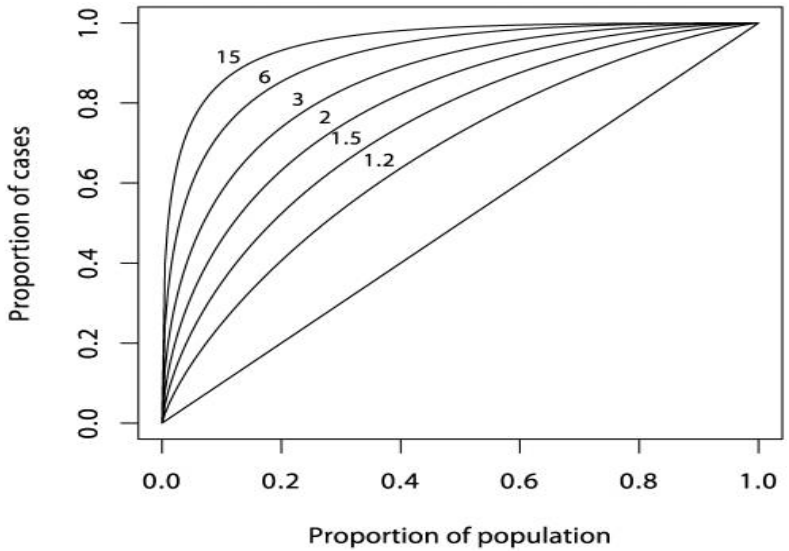
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Prediction - doomed?

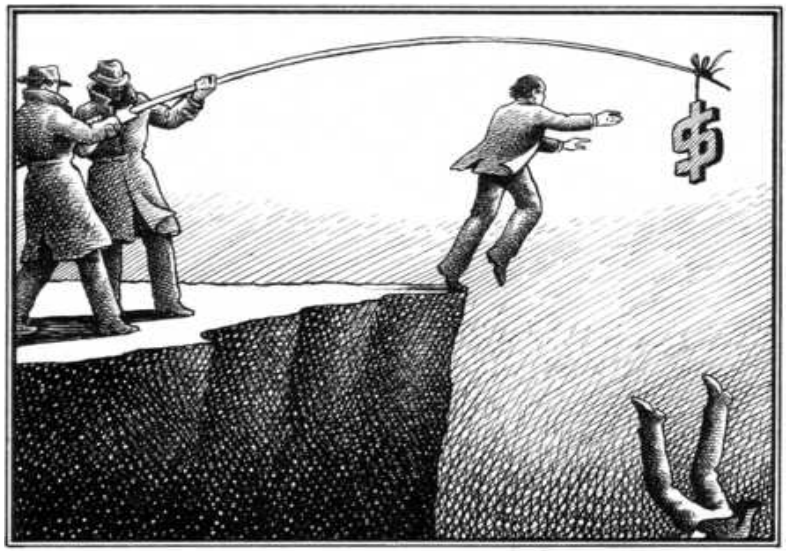


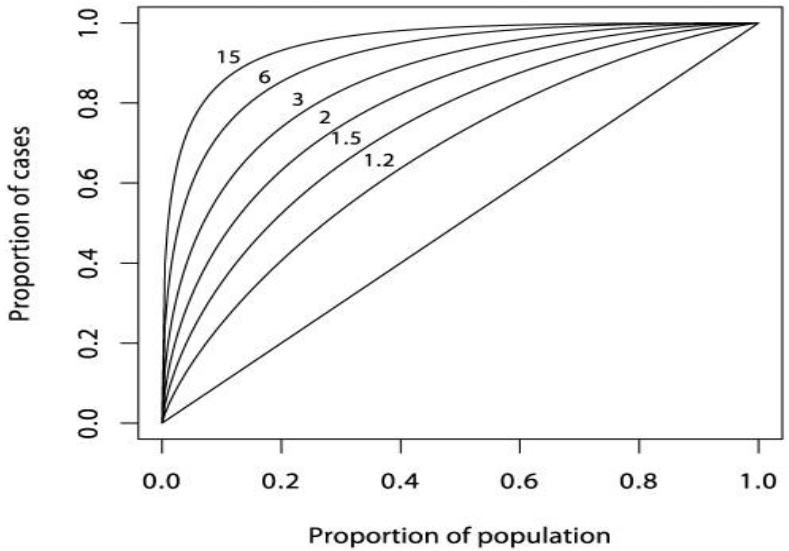
Prediction - doomed?





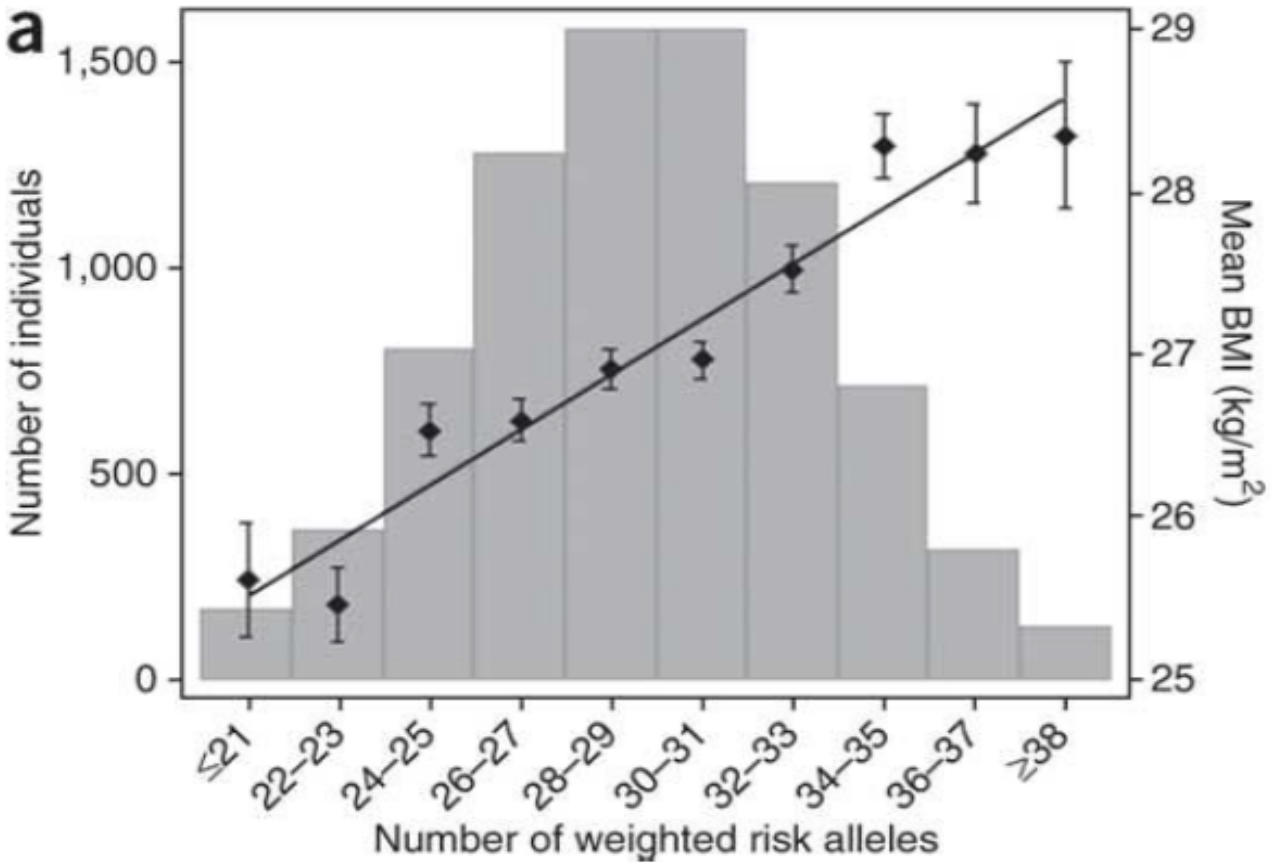
Prediction - doomed?





- Individual-level disease prediction from genetics is still very challenging (but this may change soon)
- However, aggregating predictions across samples allows powerful group-level inference and many different applications

Prediction in groups of people

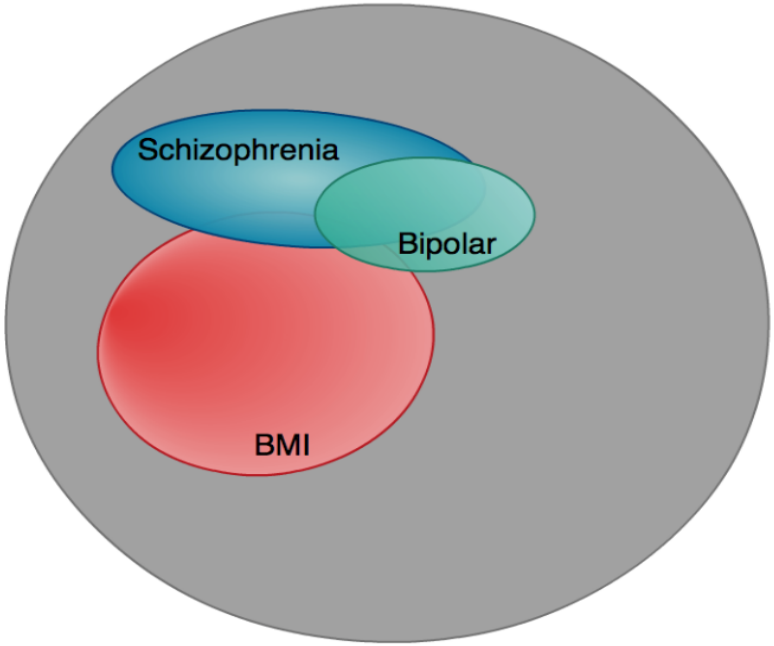


Allen et al. 2010

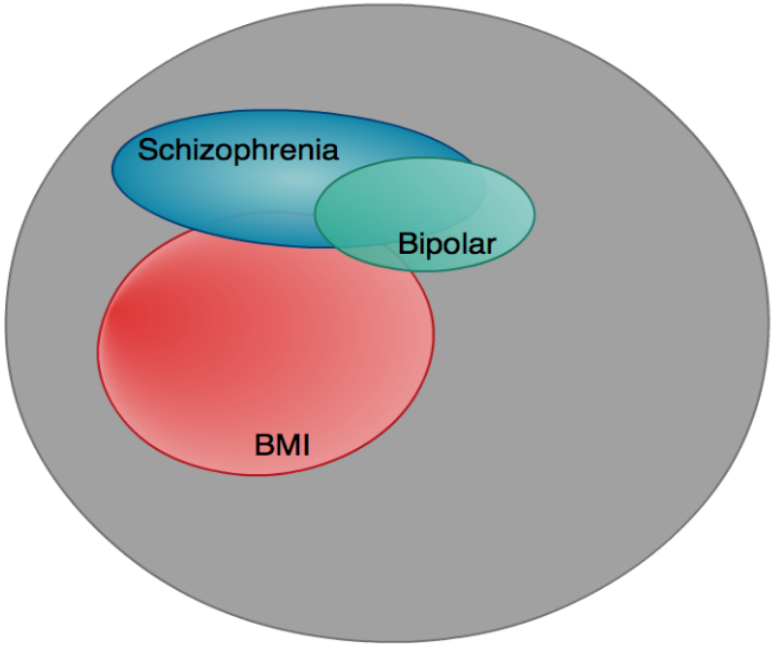
- Genome-wide significant SNPs for BMI are predictive of BMI

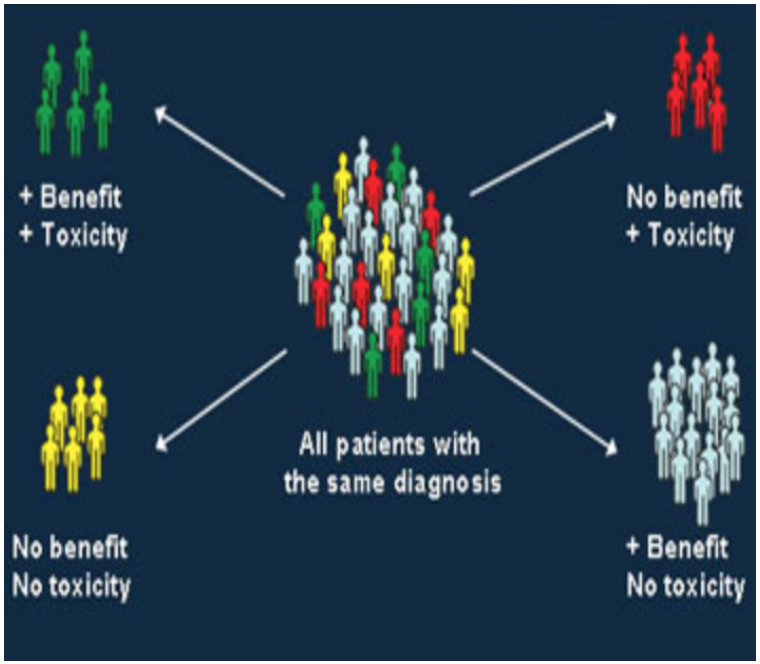
Why predict in groups of people?

Why predict in groups of people?

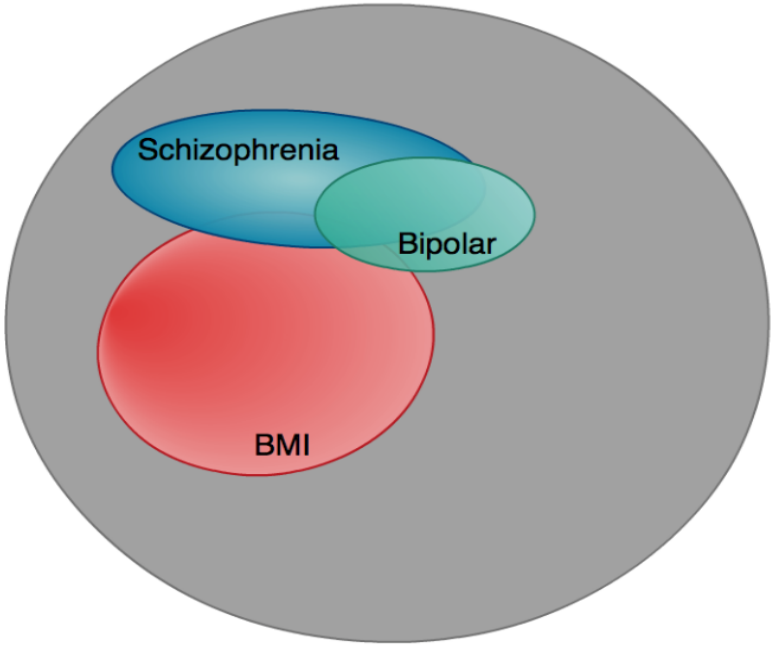


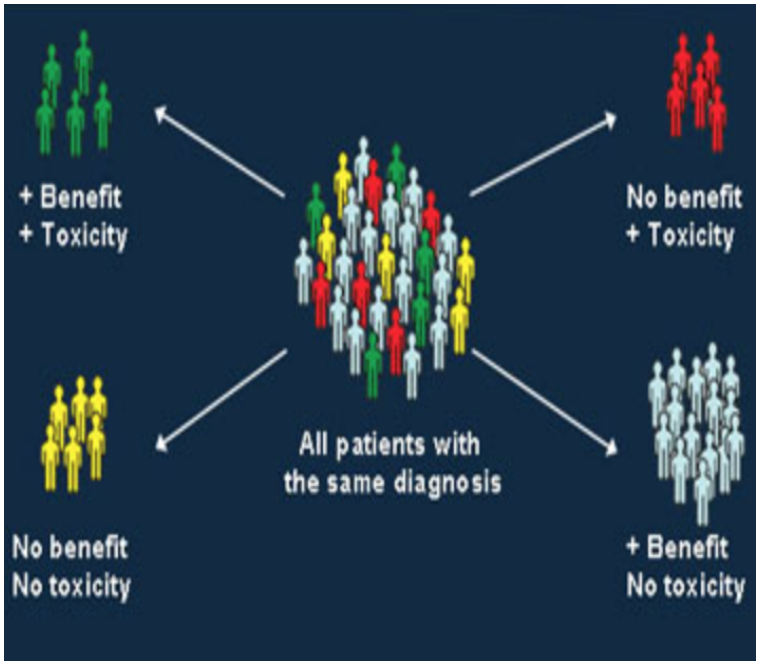
Why predict in groups of people?





Why predict in groups of people?





- There are a huge number of scientific questions that can be answered by comparing predictions in groups of people
- We will see some examples later in the lecture
- With larger GWAS samples, predictions at the individual-level will also become accurate soon

How can we optimise prediction from genetics?

- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)

How can we optimise prediction from genetics?

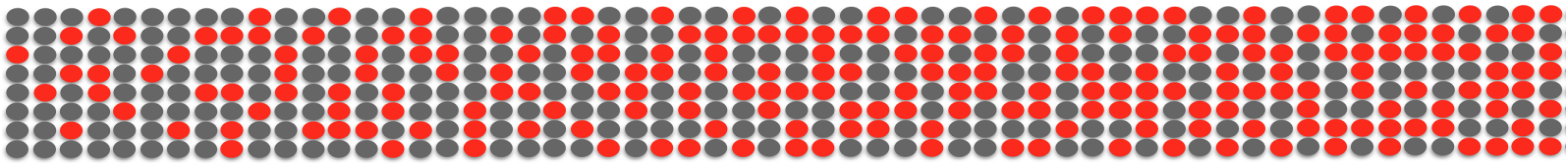
- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)
- But what if we want to optimise phenotype PREDICTION?

How can we optimise prediction from genetics?

- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)
- But what if we want to optimise phenotype PREDICTION?
- For polygenic complex traits, maybe better to include more variants $\rightarrow$ so long as they are enriched for real signal

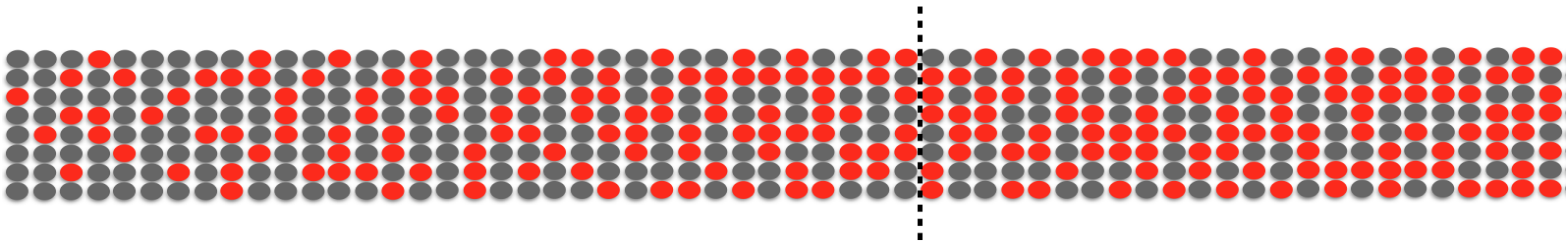
How can we optimise prediction from genetics?

- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)
- But what if we want to optimise phenotype PREDICTION?
- For polygenic complex traits, maybe better to include more variants $\rightarrow$ so long as they are enriched for real signal



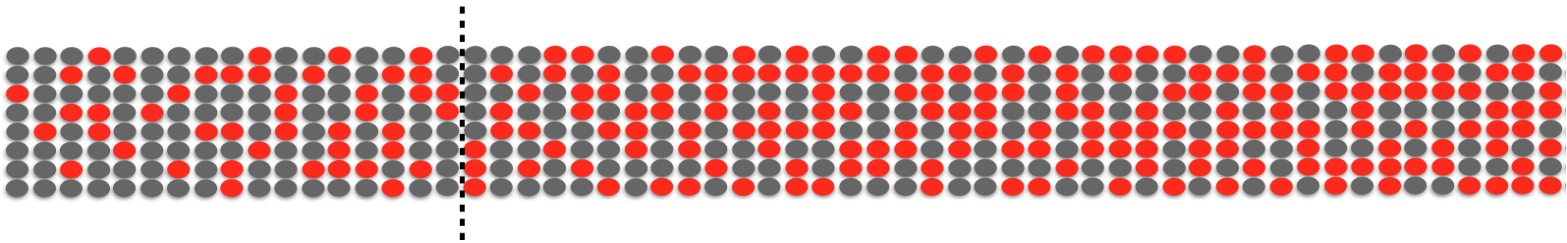
How can we optimise prediction from genetics?

- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)
- But what if we want to optimise phenotype PREDICTION?
- For polygenic complex traits, maybe better to include more variants $\rightarrow$ so long as they are enriched for real signal



How can we optimise prediction from genetics?

- The primary aim of GWAS is susceptibility locus DISCOVERY - with variants declared only if we are certain ($< 5e-8$)
- But what if we want to optimise phenotype PREDICTION?
- For polygenic complex traits, maybe better to include more variants $\rightarrow$ so long as they are enriched for real signal



Polygenic Risk Scoring

- GWAS results have shown that hundreds or thousands of variants affect complex traits - **polygenic effects**

Polygenic Risk Scoring

- GWAS results have shown that hundreds or thousands of variants affect complex traits - **polygenic effects**
- Polygenic Risk Scoring exploits these polygenic effects to predict human diseases/traits/behaviour

Polygenic Risk Scoring

- GWAS results have shown that hundreds or thousands of variants affect complex traits - **polygenic effects**
- Polygenic Risk Scoring exploits these polygenic effects to predict human diseases/traits/behaviour
- There has been widespread and increasing application of *polygenic risk scores* over the last $\approx$ 2 years

Polygenic Risk Scoring

- GWAS results have shown that hundreds or thousands of variants affect complex traits - **polygenic effects**
- Polygenic Risk Scoring exploits these polygenic effects to predict human diseases/traits/behaviour
- There has been widespread and increasing application of *polygenic risk scores* over the last $\approx$ 2 years
- Almost all aspects of genetic epidemiology and population genetics can be considered from a polygenic prospective

Polygenic Risk Scoring

- GWAS results have shown that hundreds or thousands of variants affect complex traits - **polygenic effects**
- Polygenic Risk Scoring exploits these polygenic effects to predict human diseases/traits/behaviour
- There has been widespread and increasing application of *polygenic risk scores* over the last $\approx$ 2 years
- Almost all aspects of genetic epidemiology and population genetics can be considered from a polygenic prospective
- Represents a significant shift from discovery to prediction in the GWAS era

How do you think polygenic risk scores are calculated?

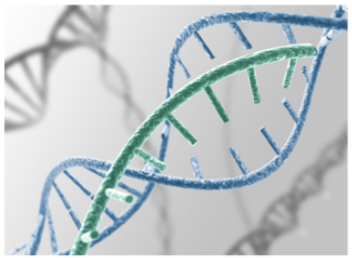
Polygenic Risk Scores

What are Polygenic Risk Scores?

- A score that predicts an individual's risk of disease, based on the combination of their genotypes and effect size estimates from GWAS results

Why might they prove so useful for medical research?

- They can assess shared genetic aetiology among phenotypes, act as a biomarker for disease, infer whether a biological factor is causally associated with a disorder, and screen subjects for clinical trials

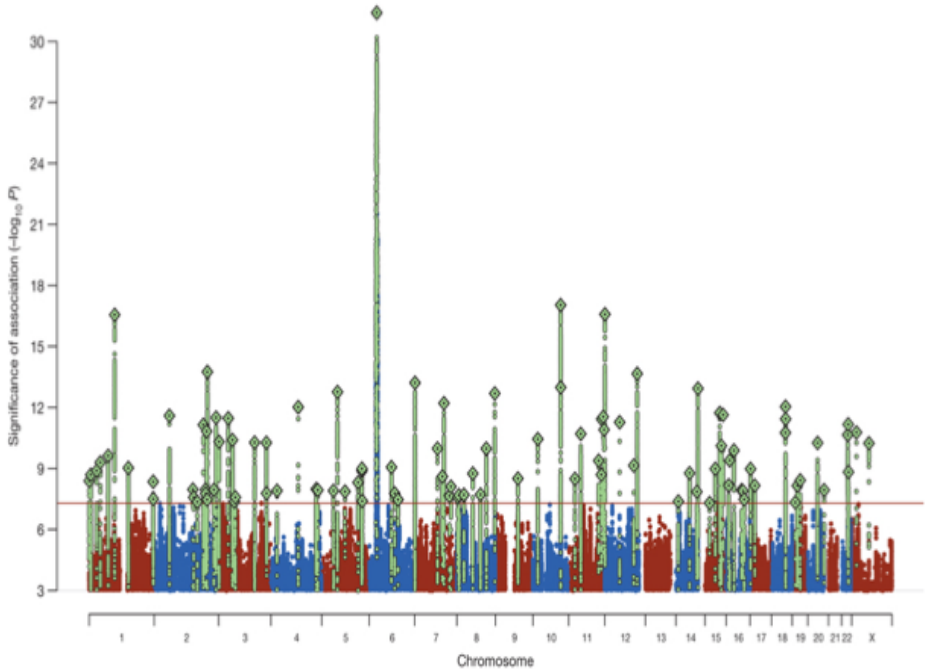


Polygenic Risk Scores

Schizophrenia

2014

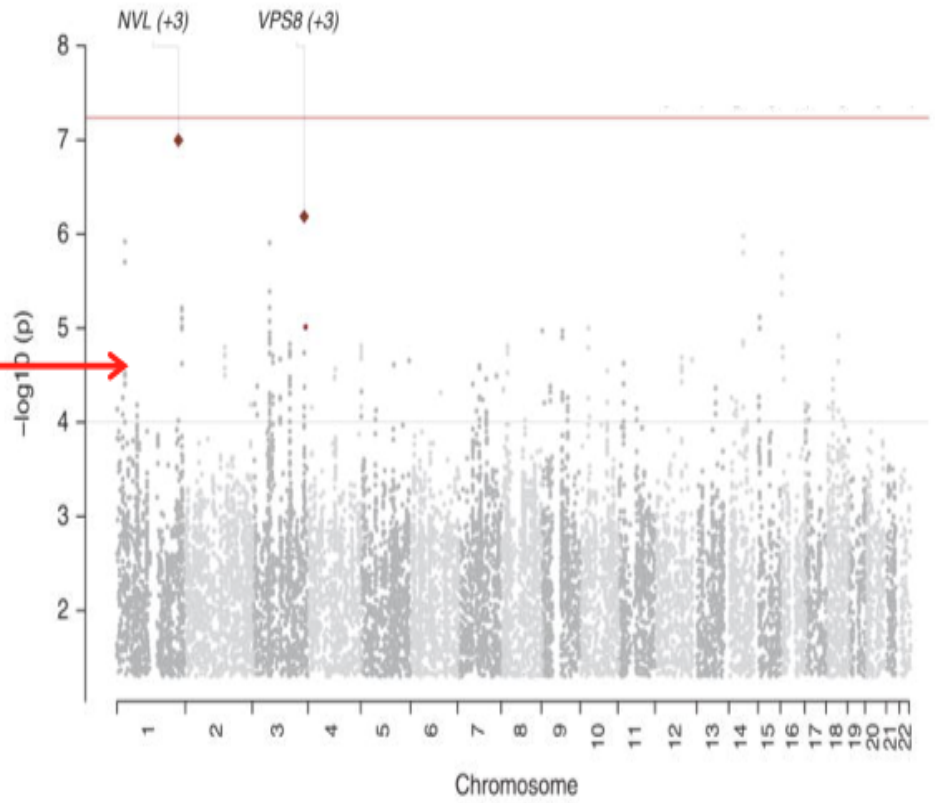
108 Genes



- There are many significant SNPs that could be used in a polygenic risk score from the recent Schizophrenia GWAS

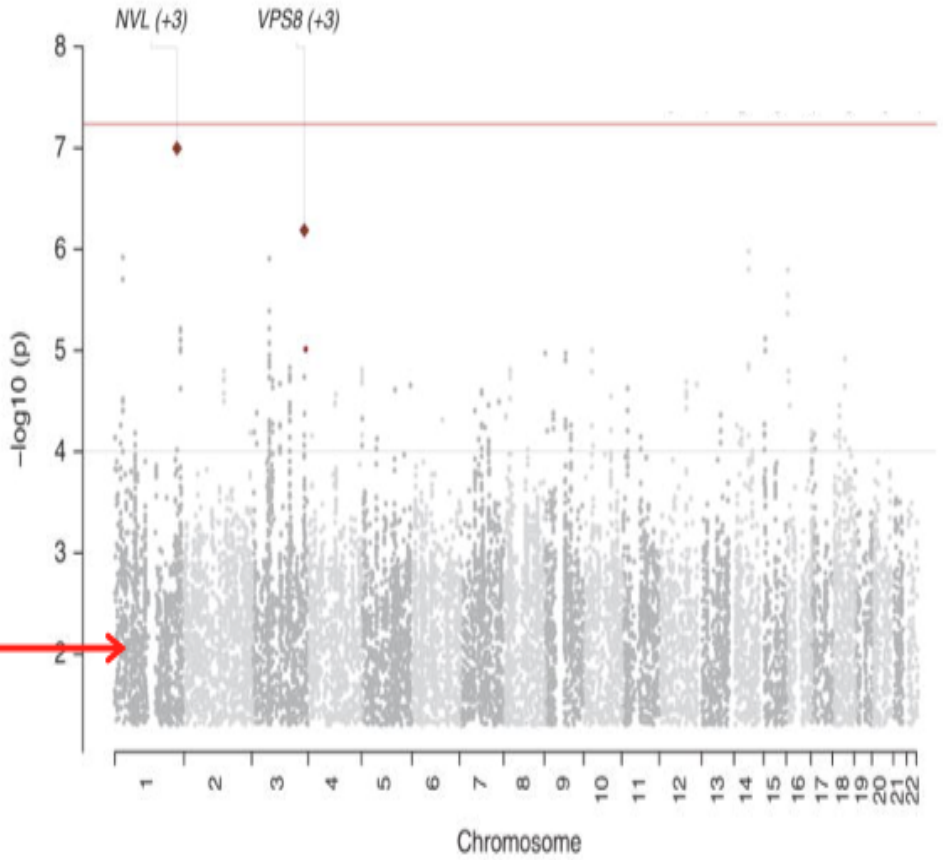
Polygenic Risk Scores

These SNPs are more strongly associated with the phenotype



Polygenic Risk Scores: using *all* SNPs

Than these SNPs



Polygenic Risk Scores (PRS): the method

- The PRS method predicts an individual's trait value from their genetic profile

Polygenic Risk Scores (PRS): the method

- The PRS method predicts an individual's trait value from their genetic profile
- It is a sum of risk alleles from genome-wide SNPs, weighted by their GWAS-derived effect size estimates

Polygenic Risk Scores (PRS): the method

- The PRS method predicts an individual's trait value from their genetic profile
- It is a sum of risk alleles from genome-wide SNPs, weighted by their GWAS-derived effect size estimates
- Only SNPs exceeding a P-value threshold, P_t, in the discovery GWAS, contribute to the score

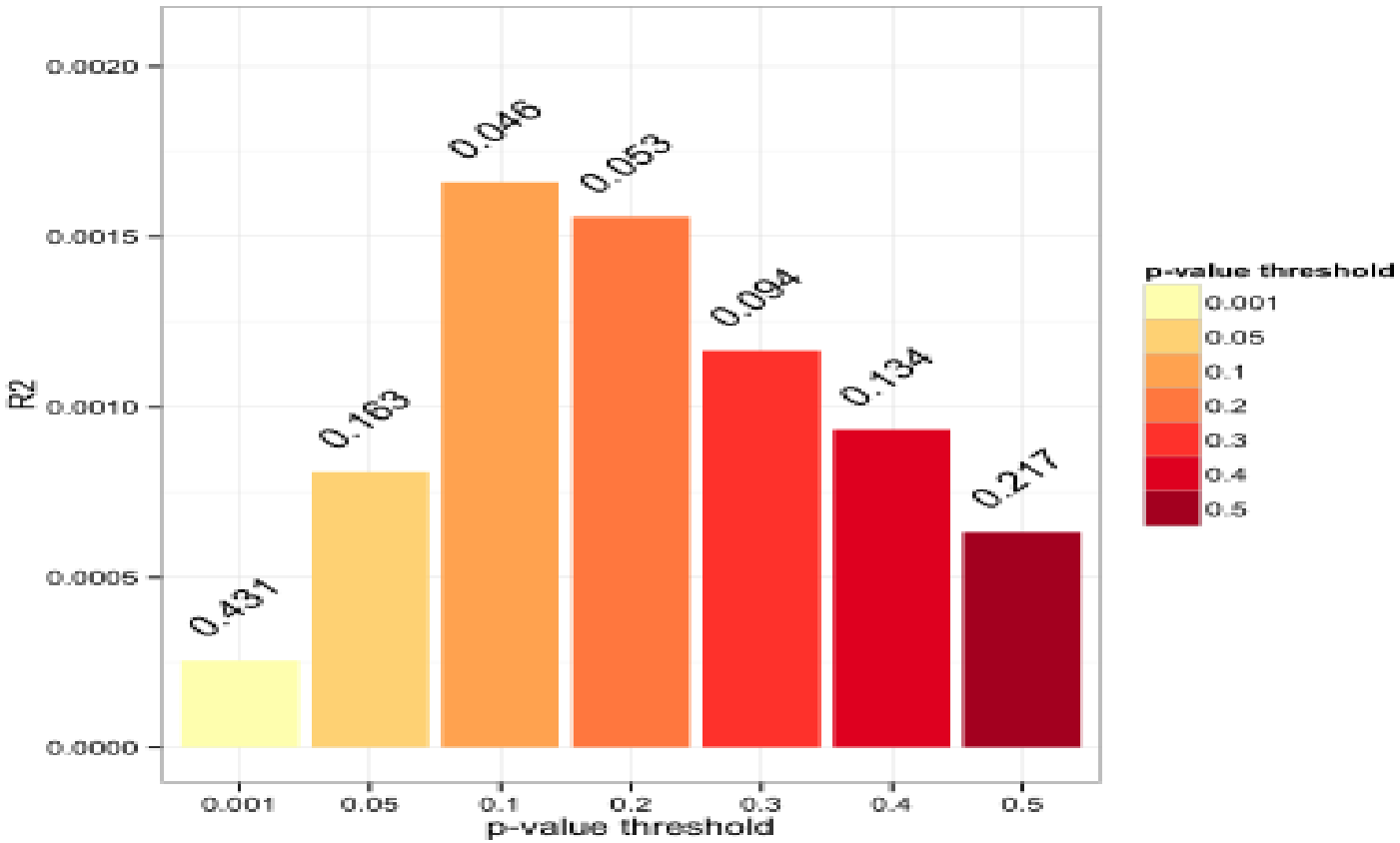
Polygenic Risk Scores (PRS): the method

- The PRS method predicts an individual's trait value from their genetic profile
- It is a sum of risk alleles from genome-wide SNPs, weighted by their GWAS-derived effect size estimates
- Only SNPs exceeding a P-value threshold, P_t, in the discovery GWAS, contribute to the score
- PRS are calculated for several P_t (...10^{-3}, 0.01, 0.05, 0.1...) in individuals of a **target data set**

Polygenic Risk Scores (PRS): the method

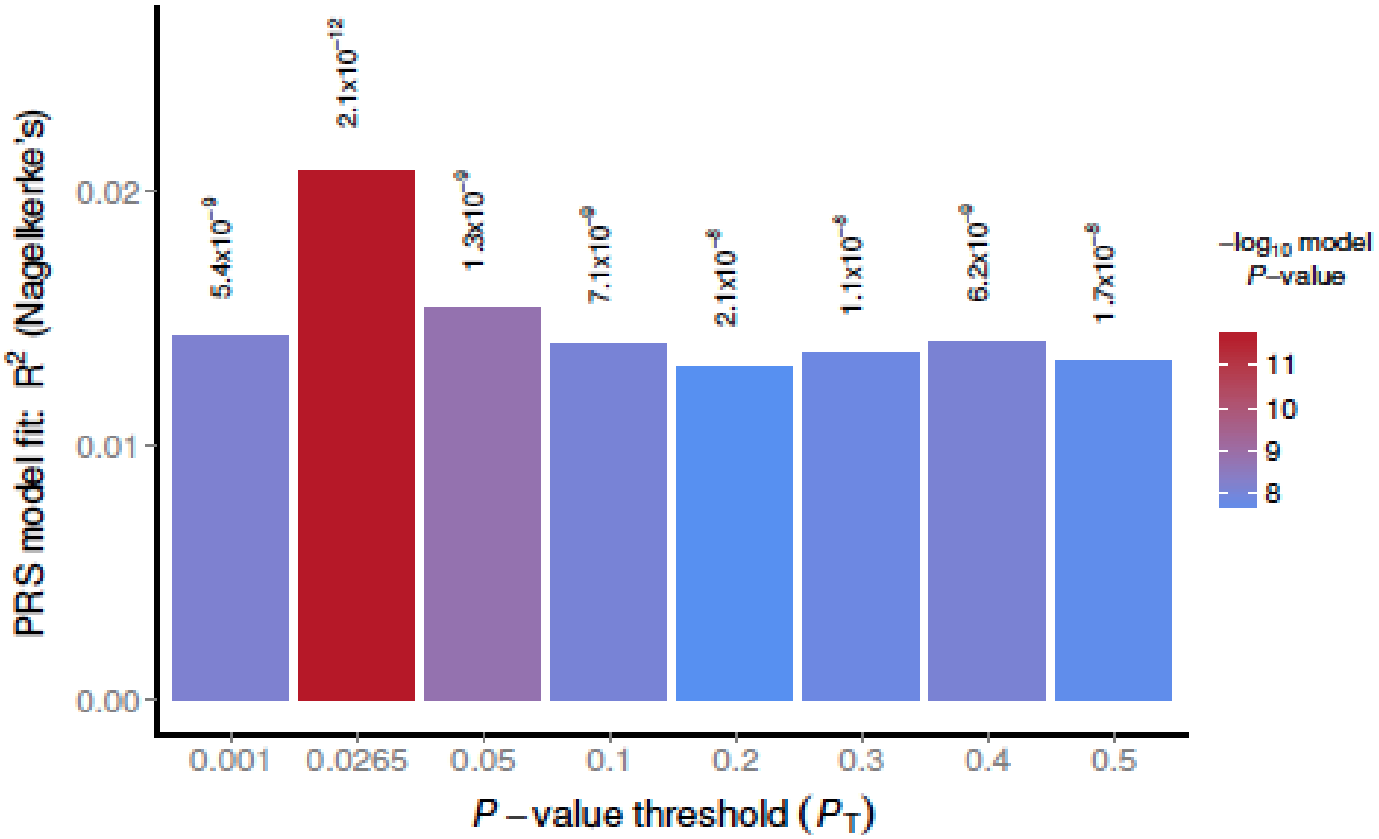
- The PRS method predicts an individual's trait value from their genetic profile
- It is a sum of risk alleles from genome-wide SNPs, weighted by their GWAS-derived effect size estimates
- Only SNPs exceeding a P-value threshold, P_t, in the discovery GWAS, contribute to the score
- PRS are calculated for several P_t (...10^{-3}, 0.01, 0.05, 0.1...) in individuals of a **target data set**
- Finally - regression is then performed to test for association between PRS and the trait in the target sample

Polygenic Risk Scores: at different thresholds

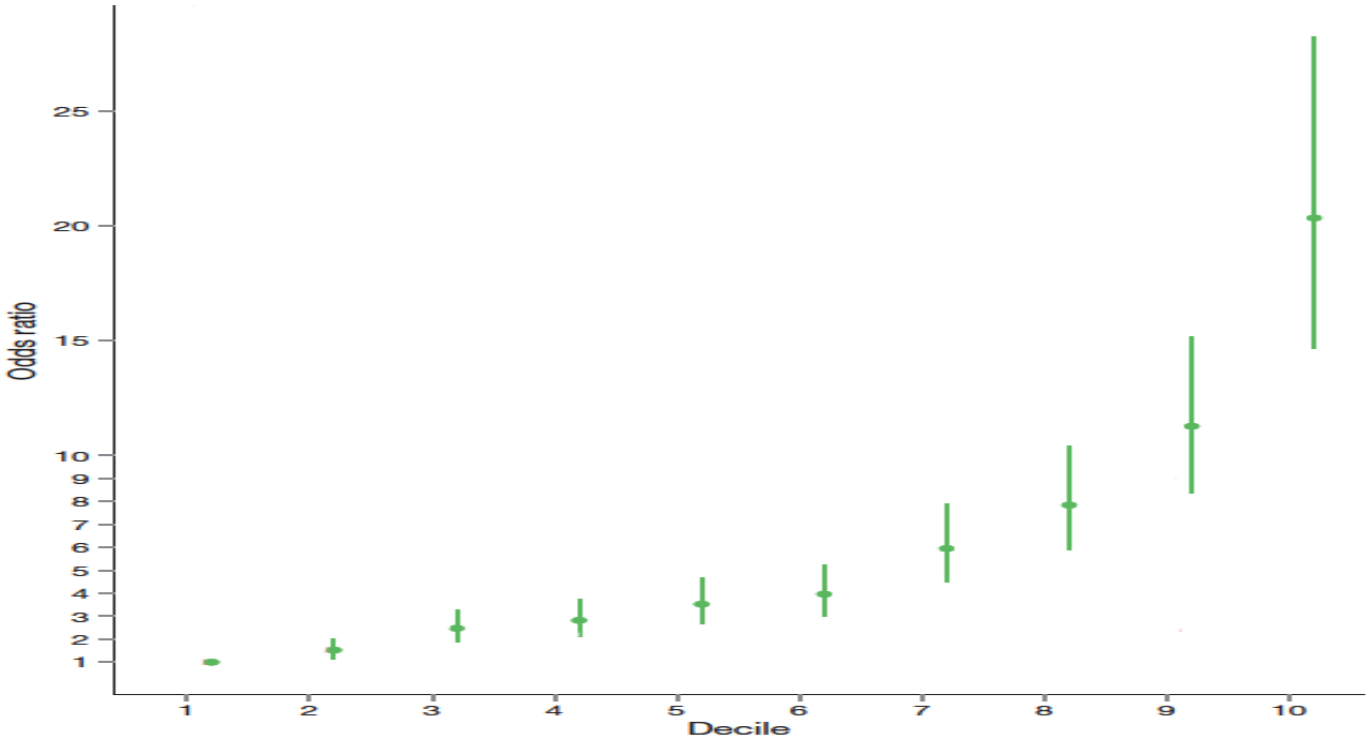


- Usually calculated at several broad thresholds

Polygenic Risk Scores (PRS)



Polygenic Risk Scores: on the spectrum



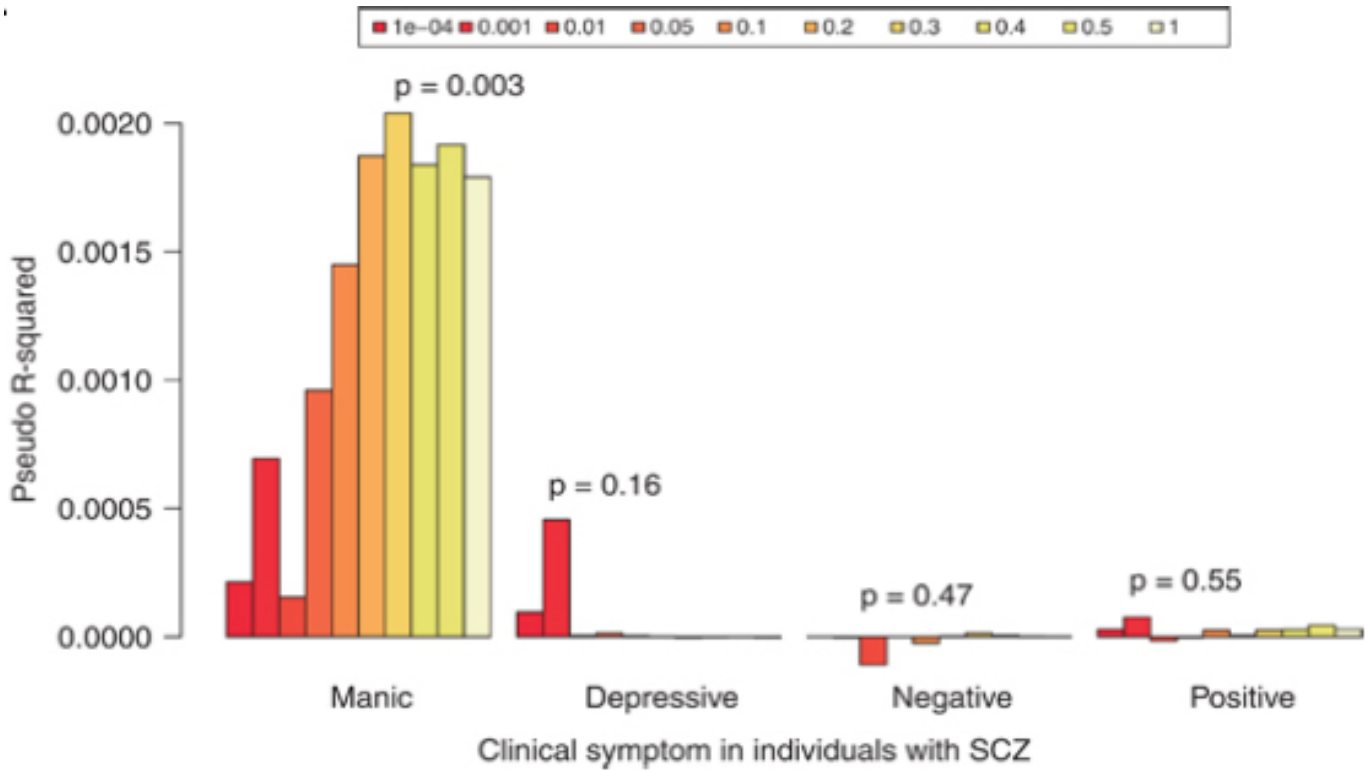
- Individuals can be placed on a *spectrum* of genetic burden to a disease/trait, mirroring eg. the 'autistic spectrum'

Polygenic Risk Scores (PRS)

ID	SCZ	Sex	BMI	$PRS_{(1e-7)}$	$PRS_{(1e-5)}$
1	1	F	22.1	3.14	2.88
2	0	M	24.6	0.26	0.74
3	1	M	33.2	0.81	0.83
4	0	F	25.3	0.02	0.08
5	0	M	26.8	2.64	2.91

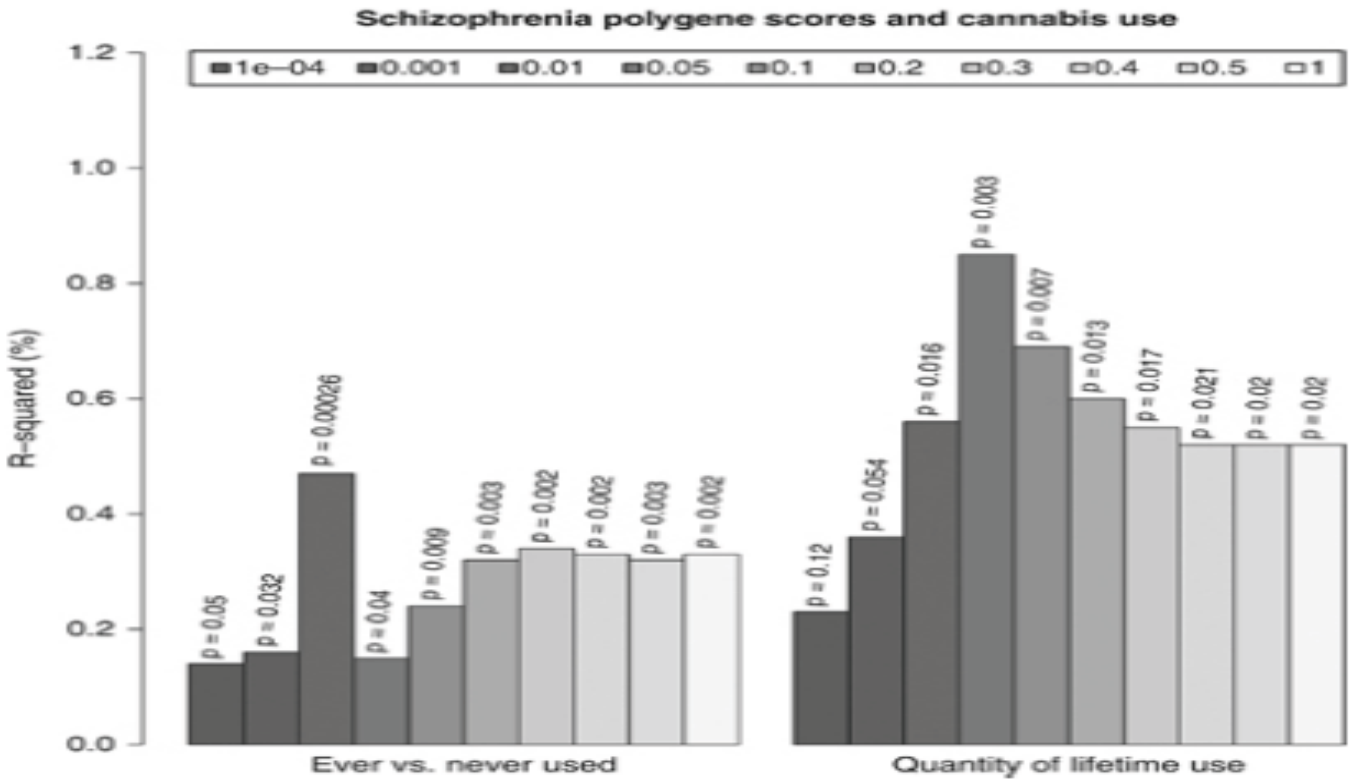
Testing for shared genetic aetiology

- Assessing genetic overlap between Schizophrenia and symptoms of Bi-polar disorder



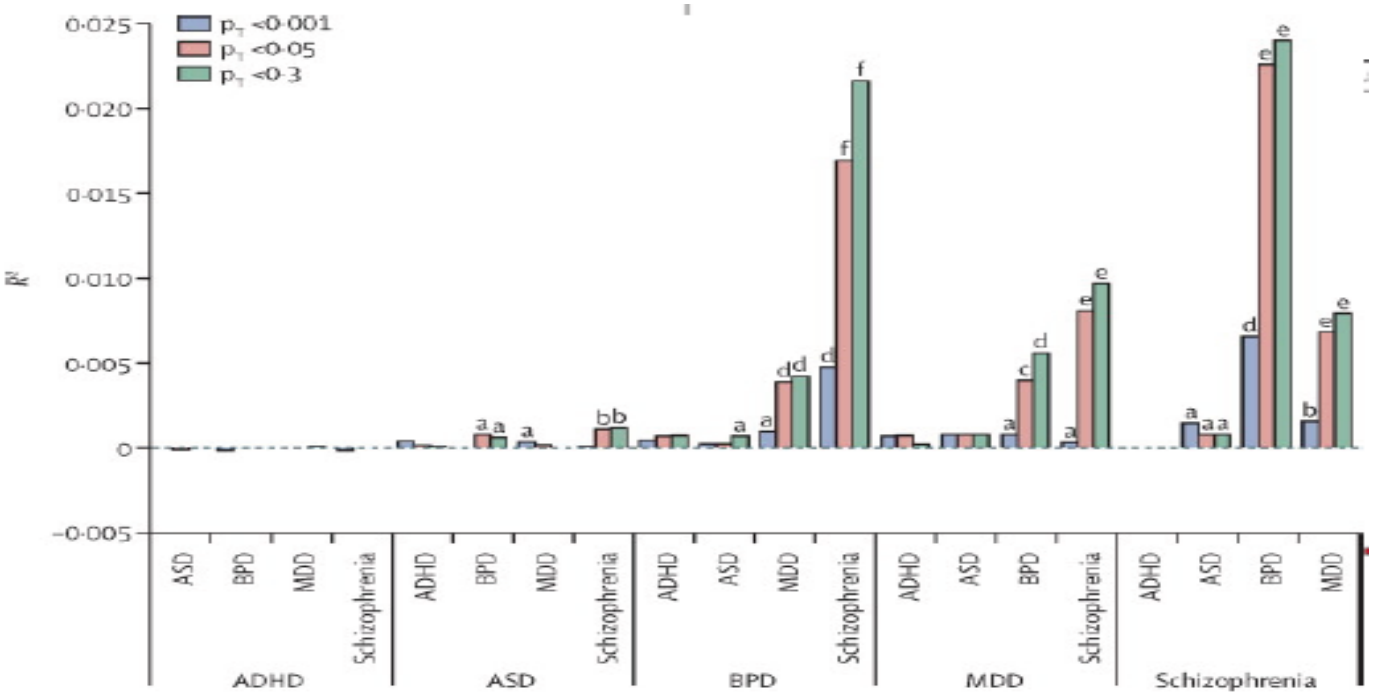
Testing for shared genetic aetiology

- Assessing shared genetic aetiology between Schizophrenia and Cannabis use



Testing for shared genetic aetiology

- Assessing the genetic overlap between the major psychiatric disorders



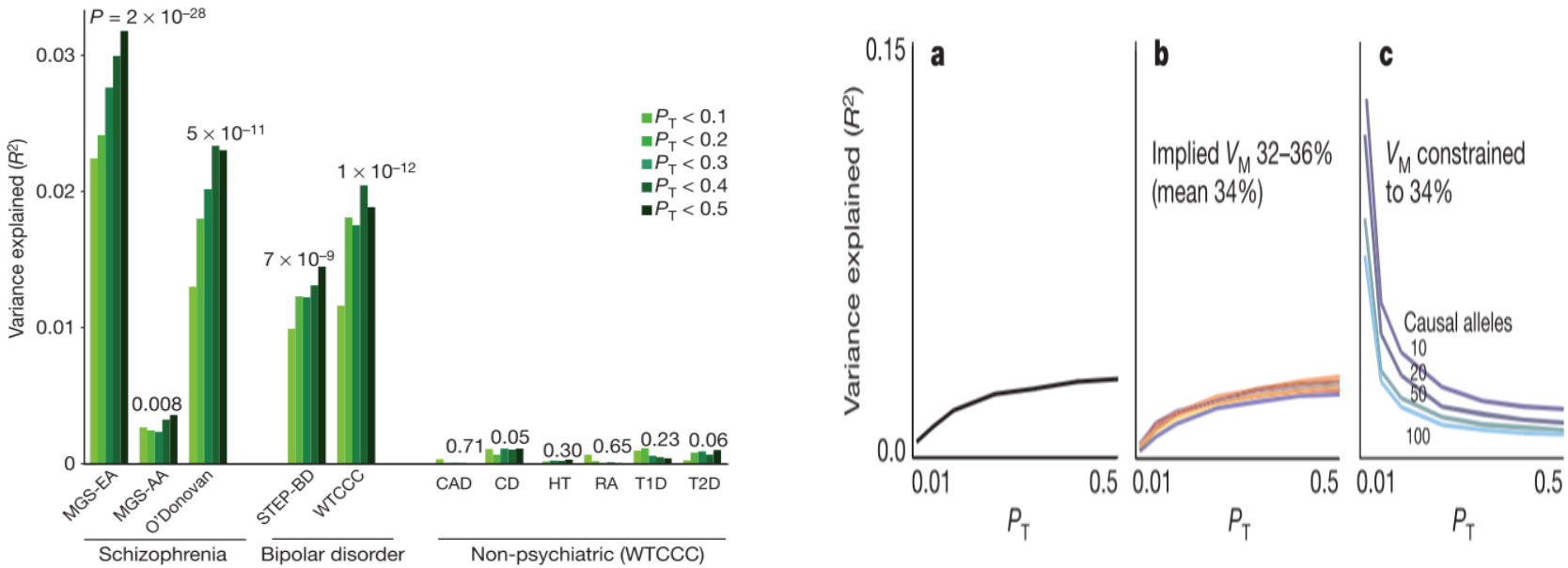
Any questions??

Problems with the PRS method?

Reading List:

Recent history of PRS in 4 papers

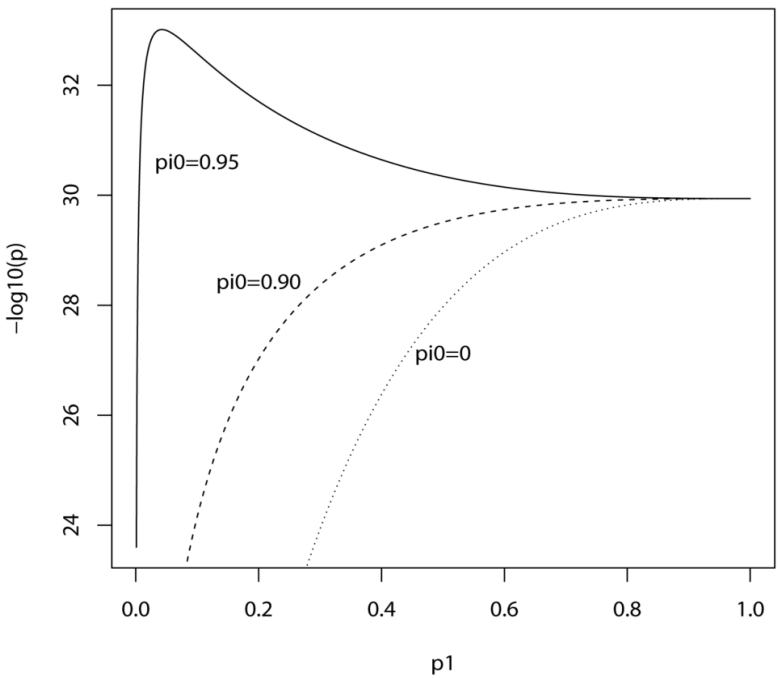
Paper (1): ISC GWAS 2009

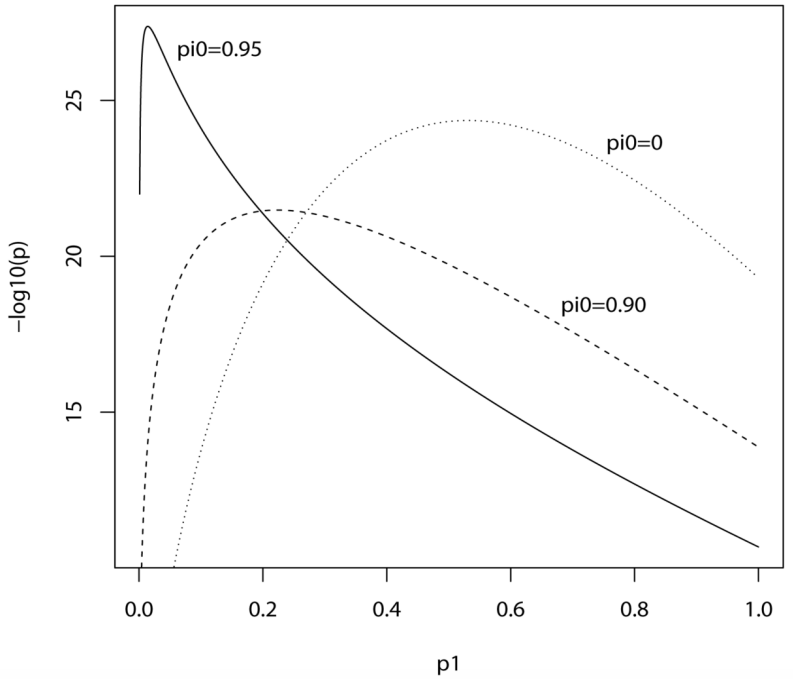


Purcell et al. 2009. *Nature*. 460:748-52. doi: 10.1038/nature08185

- SCZ PRS predict SCZ and BIP, but not WTCCC diseases
- Simulations showed that observed results consistent with 34% variance explained by markers

Paper (2): Frank Dudbridge 2013

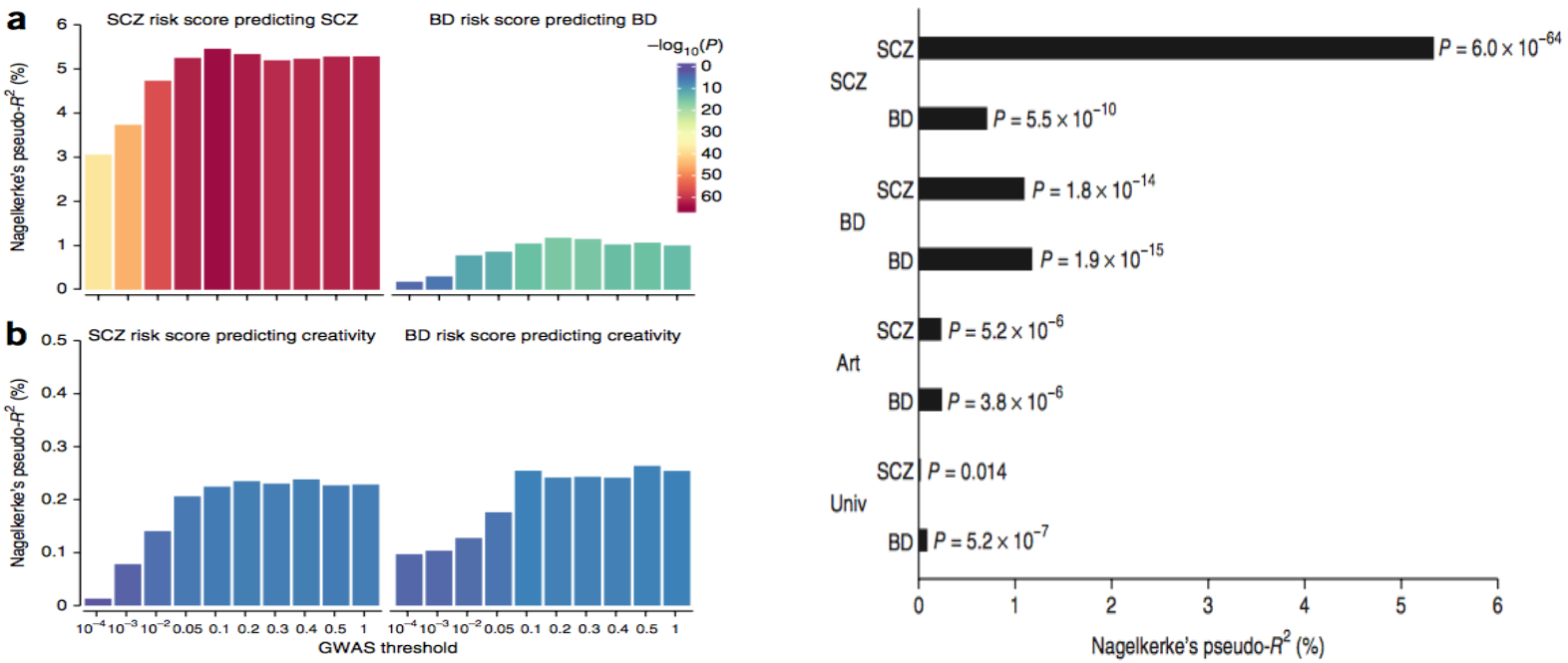




Dudbridge F. 2013. *PLoS Genet*. 9(3):e1003348. doi: 10.1371/journal.pgen.1003348

- Showed how fraction of causal SNPs can affect prediction across P_t
- Provided analytical solutions to describe how the power and accuracy of PRS for association and prediction are a function of power, discovery:target ratio, effect size distn. etc

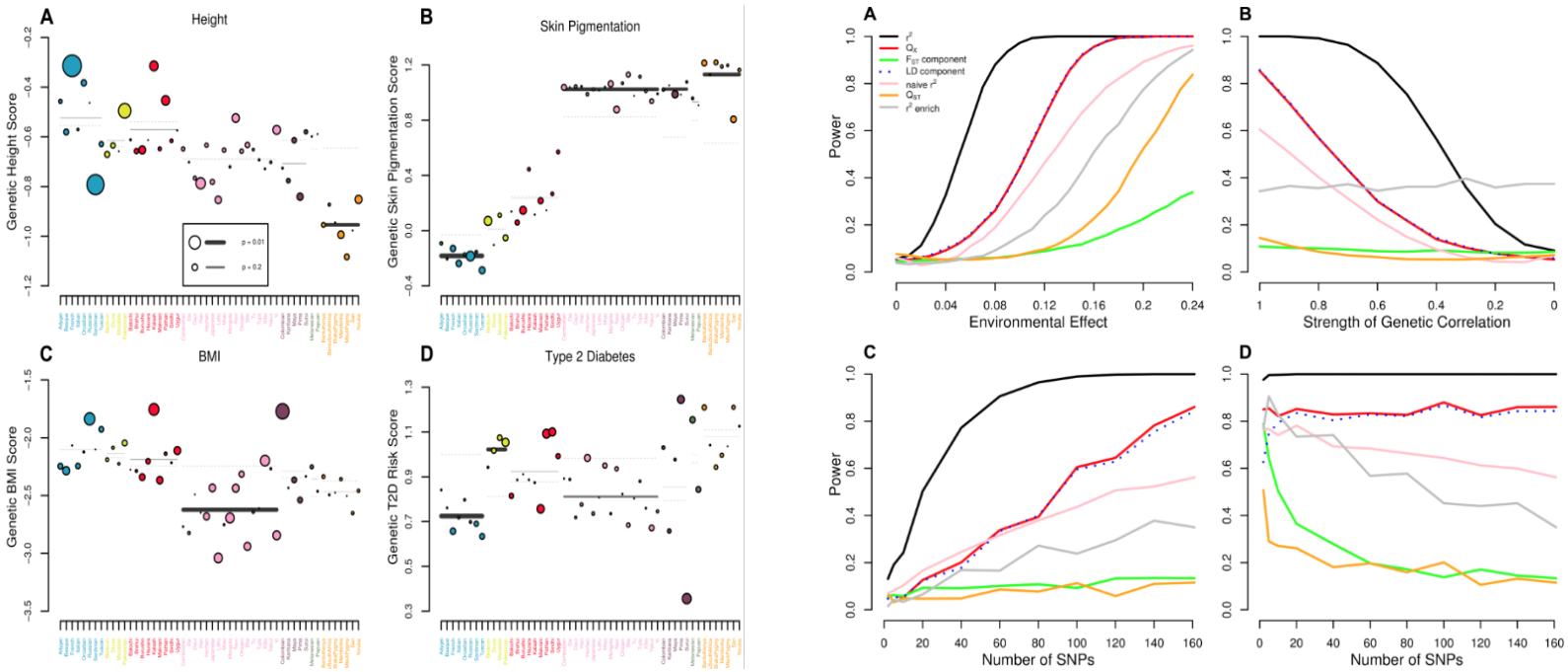
Paper (3): SCZ-Creativity PRS



Power et al. 2015. *Nat Neuro*. 18:953-5. doi: 10.1038/nn.4040

- SCZ and BIP PRS associated with creative/artistic measures in Icelandic pop
- Association is small but significant

Paper (4): Evolutionary insights from PRS



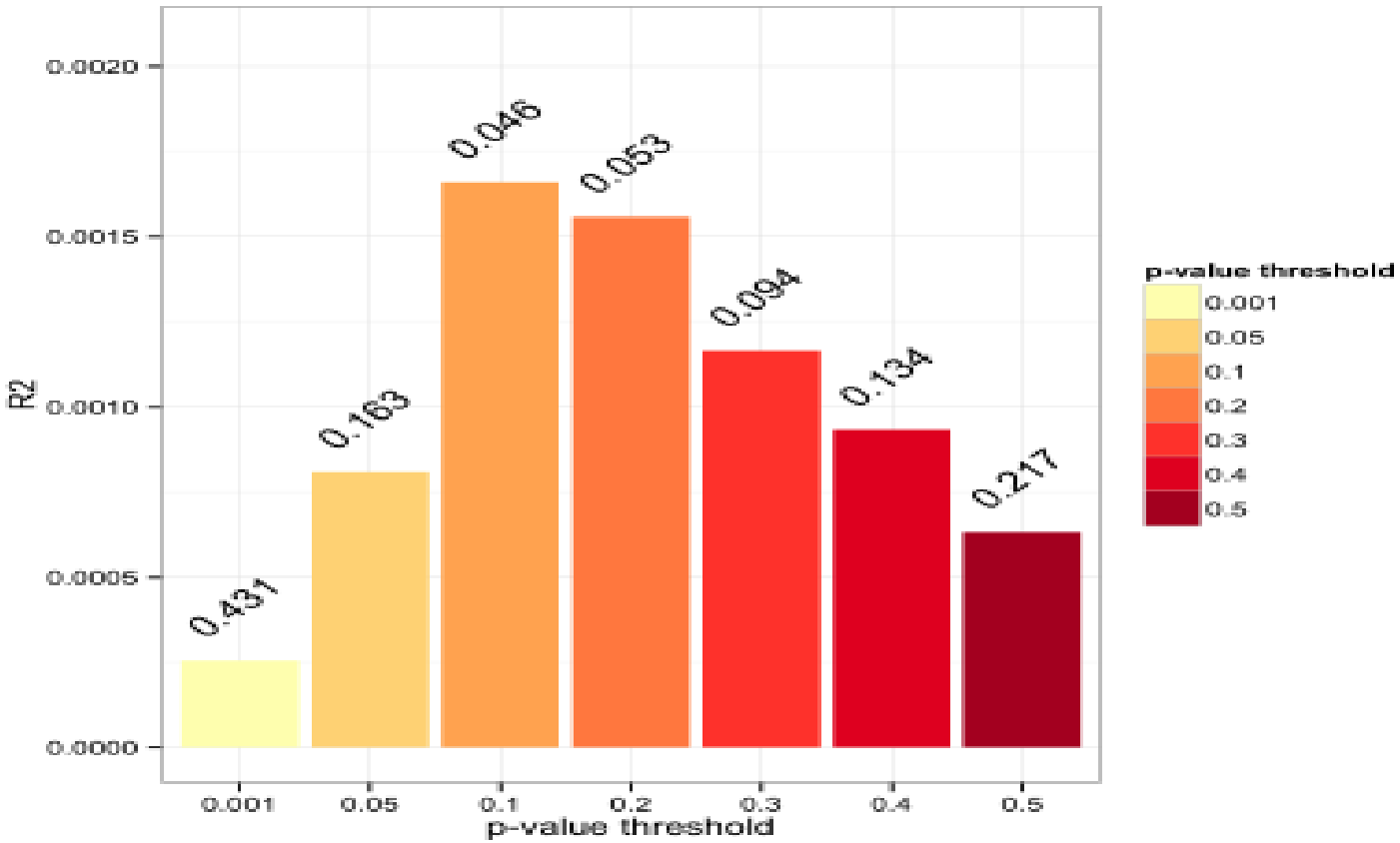
Berg and Coop. 2014. *PLoS Genet*. 10(8):e1004412. doi: 10.1371/journal.pgen.1004412

- Tests for detecting selection usually focus on single variants
- Here, a polygenic score was tested against environmental variables and across populations

Appendix: Some work by my group..

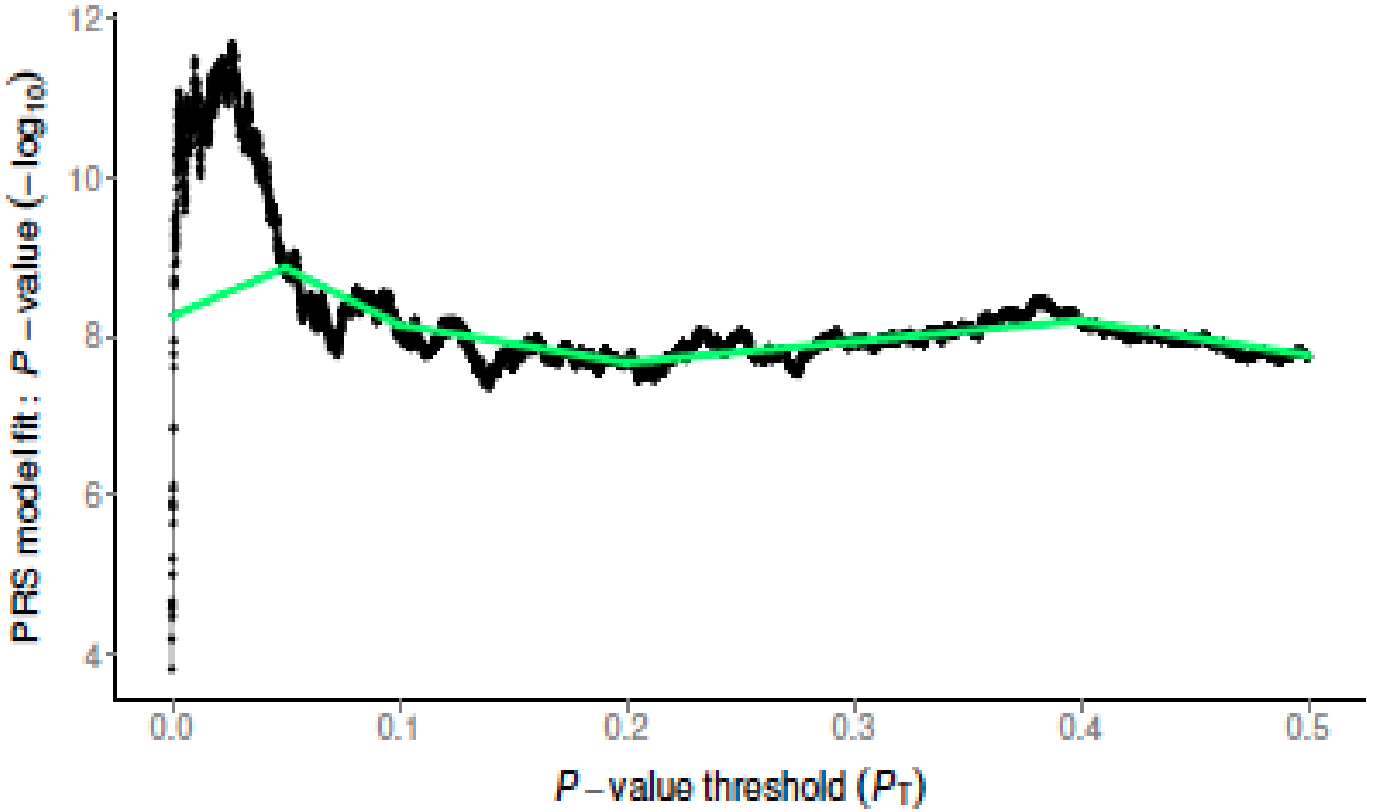
- We have produced the first dedicated PRS software: PRSice
- PRSice calculates, applies, evaluates and plots PRS results
- PRSice can calculate PRS at 'high-resolution' to give the best-fit PRS (rather than rely on broad P_t thresholds)
- Freely available for download from: www.PRSice.info
- Performed large cross-trait PRS study
- Tested for *polygenic adaptation*

Polygenic Risk Scores: using broad thresholds



- Usually calculated at several broad thresholds

PRSice: high-resolution scoring



PRSice webpage: www.PRSice.info

PRSice: Polygenic Risk Score software

by Jack Euesden, Cathryn Lewis & Paul O'Reilly

Statistical Genetics Unit
King's College London

PRSice (pronounced 'precise') is a software package for calculating, applying, evaluating and plotting the results of polygenic risk scores. PRSice can run at high-resolution to provide the best-fit PRS as well as provide results calculated at broad *P*-value thresholds, illustrating results corresponding to either (see below), can thin SNPs according to linkage disequilibrium and *P*-value ("clumping"), handles genotyped and imputed data, can calculate and incorporate ancestry-informative variables, and can be applied across multiple traits in a single run.

Based on a permutation study we estimate a significance threshold of $P = 0.001$ for high-resolution PRS analyses - the work on this is included in our *Bioinformatics* paper on PRSice.

PRSice is a software package written in R, including wrappers for bash data management scripts and PLINK2 (Chang et al. 2015) to minimise computational time; thus much of its functionality relies entirely on computations written originally by Shaun Purcell in PLINK. PRSice runs as a command-line program with a variety of user-options and is freely available for download below, compatible for Unix/Linux/Mac OS and in dockerised form also Windows.

For more details on the authors, see: Jack's homepage, Cathryn's homepage, Paul's homepage.

Downloads

PRSice v1.23 can be downloaded HERE - this includes toy data, a vignette using these data that guide users through the implementation of PRSice via several examples (including running on a cluster, illustration of output/plots etc), and a user manual describing all user-options. All versions previous to v1.2 should be considered beta.

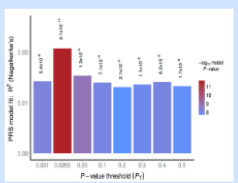
The PRSice user manual can also be obtained directly here: PRSice User Manual

The PRSice vignette can also be obtained directly here: PRSice Vignette

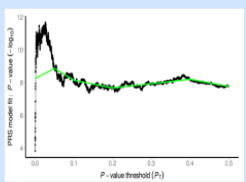
For Windows users, we suggest either running PRSice on a cluster or using the version of PRSice dockerised by Stephen Newhouse: Dockerised PRSice

If you have any questions about PRSice, or would like to be added to the mailing list to receive emails on software updates etc, then please email PRSice.info@gmail.com

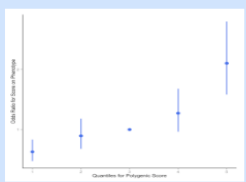
Example Output



Example bar plot produced by PRSice



Example high-resolution plot produced by PRSice



Example quantile plot produced by PRSice

The first two figures are based on a PRSice run over PGC Schizophrenia and RADIANT-UK Major Depressive Disorder data, as shown in our paper, while the quantile plot is produced from simulated data.

PRSice: Polygenic Risk Score software

- Example inputs:

```
R --file=./PRSICE_SOFTWARE_v0.1.R \
        --args  base REDUCED_GWAS_SCZ_CHR19.assoc \
                target RADIANT_CHR19 \
                fastscore F \
                plink ./PLINK_mac_2014-08-08 \
                covary F \
                report.individual.scores T
```

PRSice: Polygenic Risk Score software

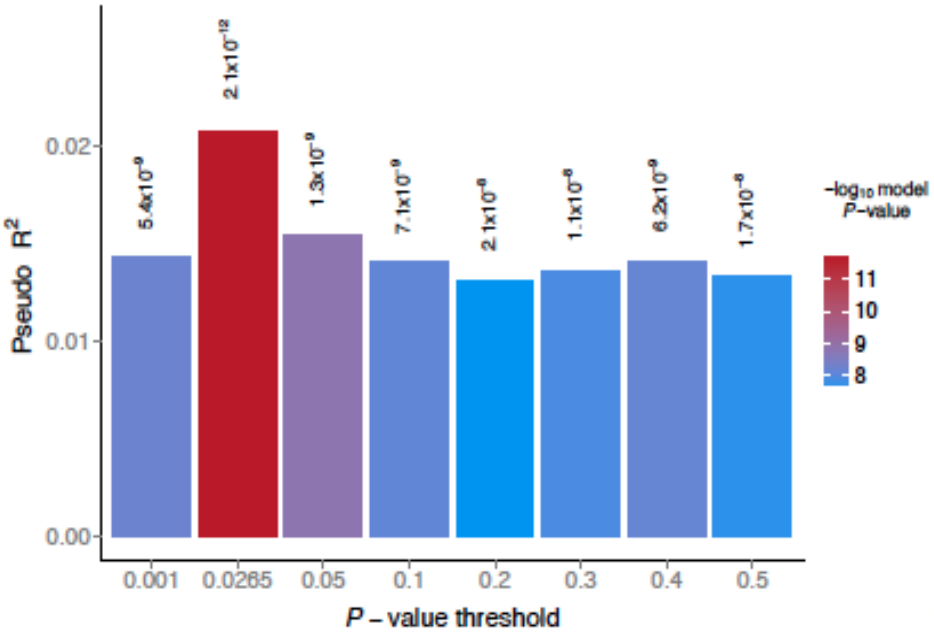
- Example outputs:

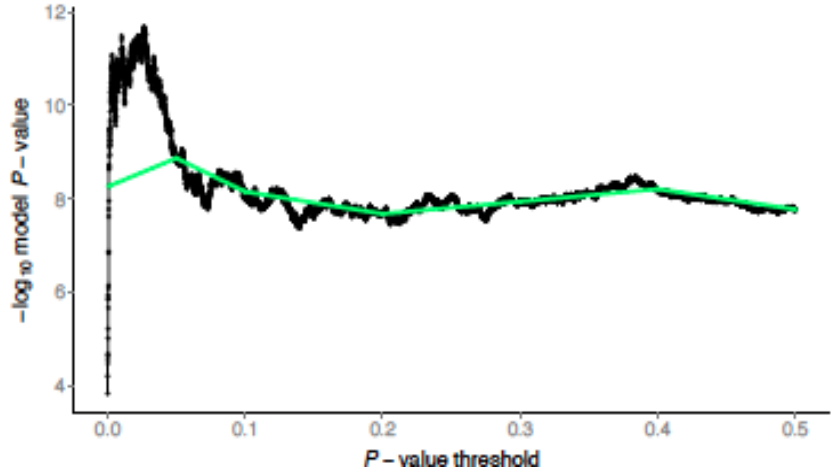
```
IID pT_0.1
476120 -0.00230224
476121 -0.00231043
476359 -0.001604
476360 -0.00165873
476361 -0.000793023
476363 -0.00197644
476364 0.00022314
476448 -0.00106749
477005 -0.00121219
477088 -0.00107726
477114 -0.00214565
477158 -0.0010211
477257 -0.0014548
477660 -0.00113063
477662 -0.00142114
477702 -0.0032354
477745 -0.00154415
```

```
thresh p.out r2.out nsnps
0.001 0.436268248939325 0.000251335434298746 14
0.05 0.171394400496722 0.000776073951446416 276
0.1 0.0476095496508912 0.00162876061437323 457
0.2 0.0554749622868169 0.00152239045280368 738
0.3 0.100139808111387 0.00112176524621458 950
0.4 0.14356627851544 0.000887576439545803 1142
0.5 0.232043924436186 0.000592453186939881 1331
PRSICE_RAW_RESULTS_DATA.txt (END)
```

PRSice: Polygenic Risk Score software

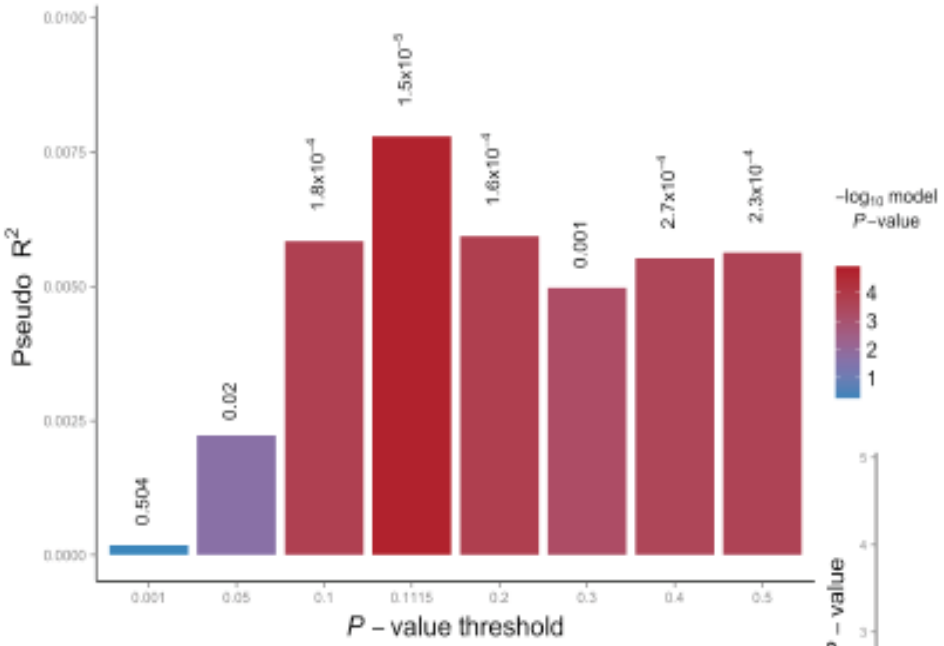
- Schizophrenia predicting MDD

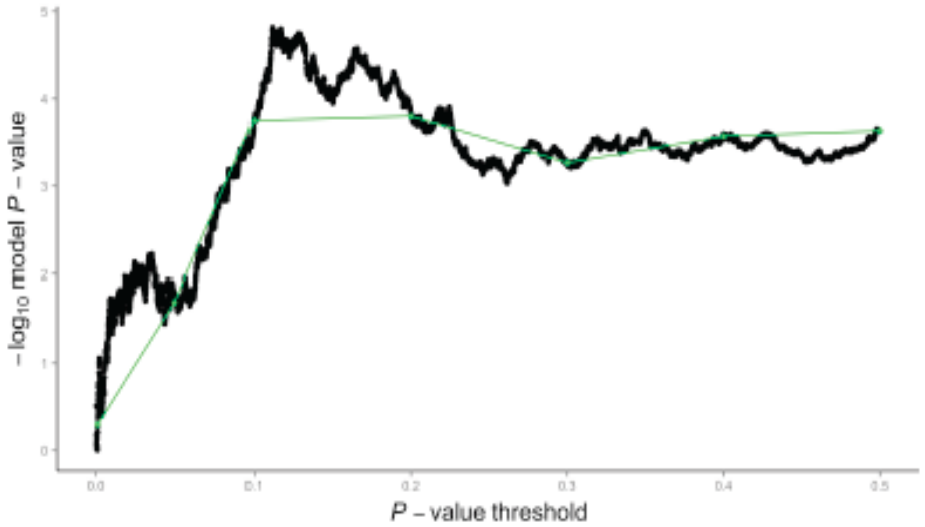




PRSice: Polygenic Risk Score software

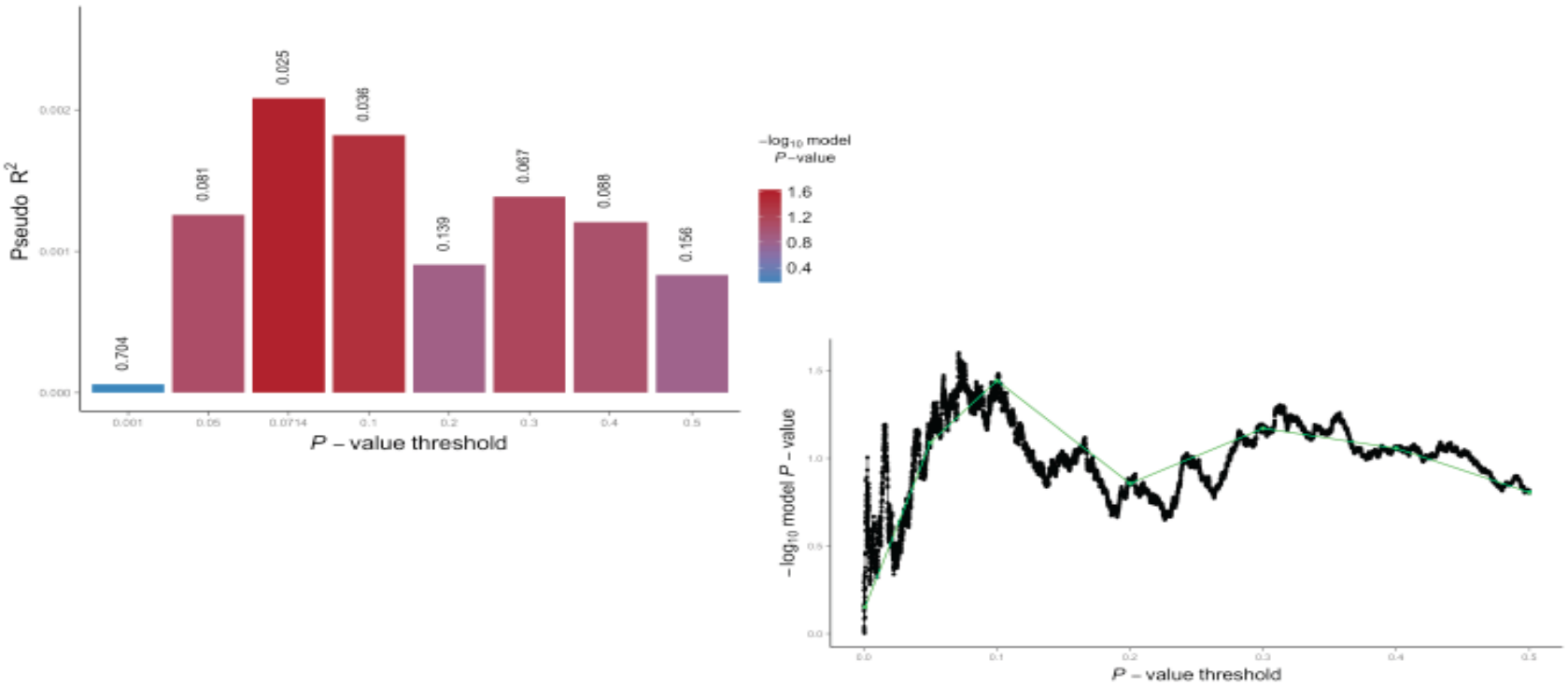
- Ever-smoked predicting MDD





PRSice: Polygenic Risk Score software

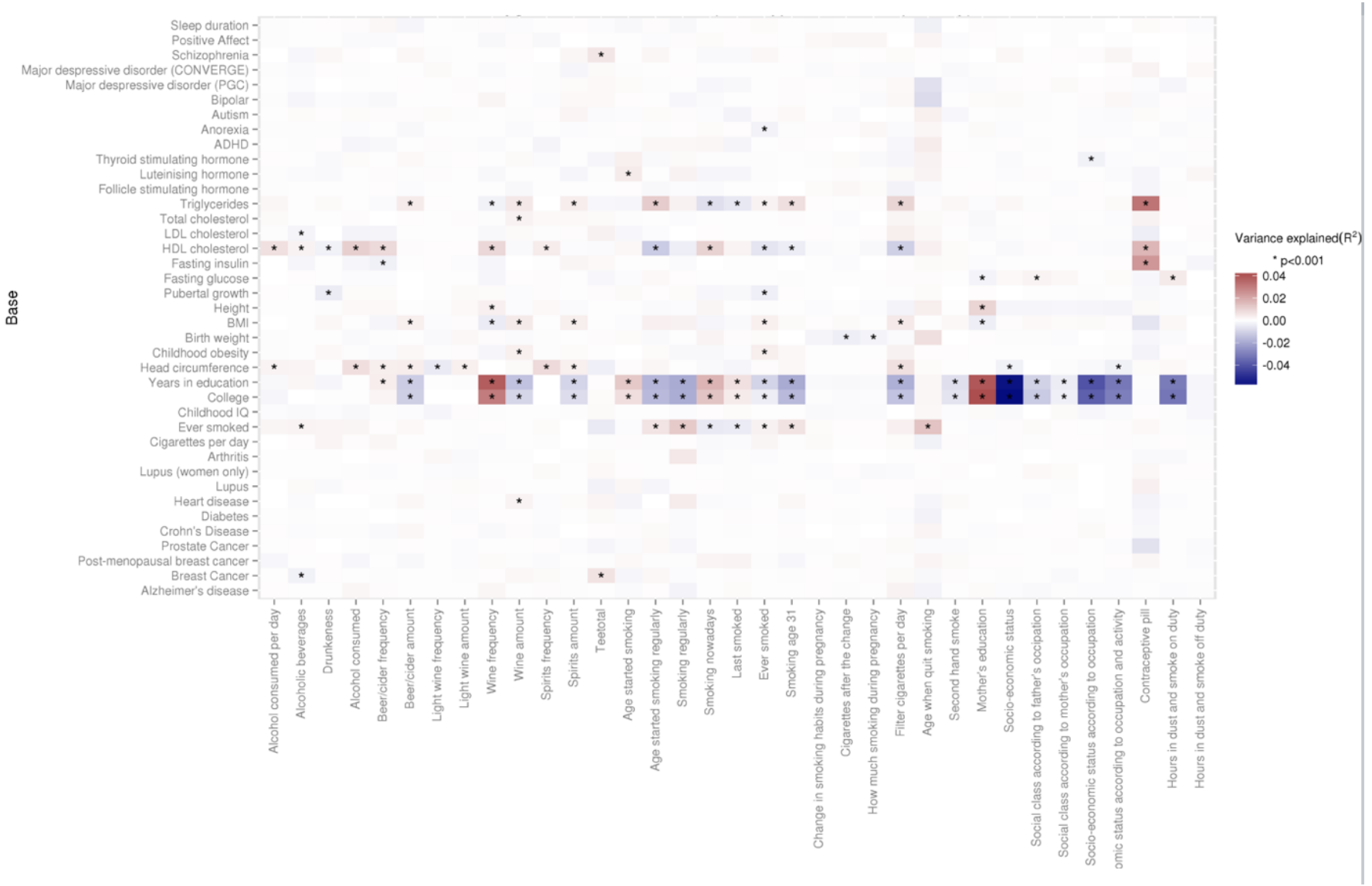
- Number of cigarettes smoked predicting MDD



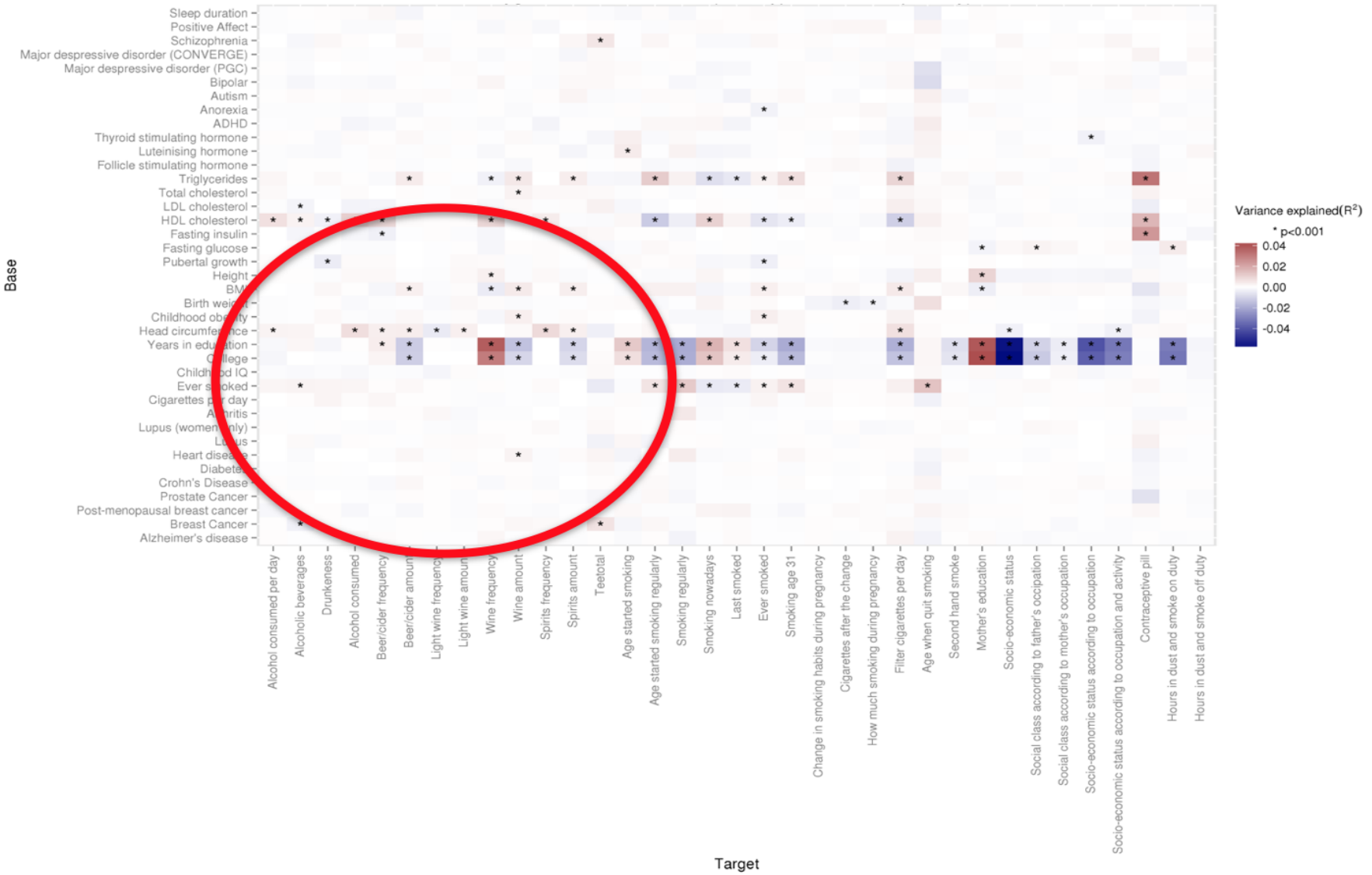
We suggest a significance threshold of 0.001 for high-resolution PRS (Euesden, Lewis, O'Reilly. *Bioinf*. 2015).

PRSice: Cross-trait PRS study

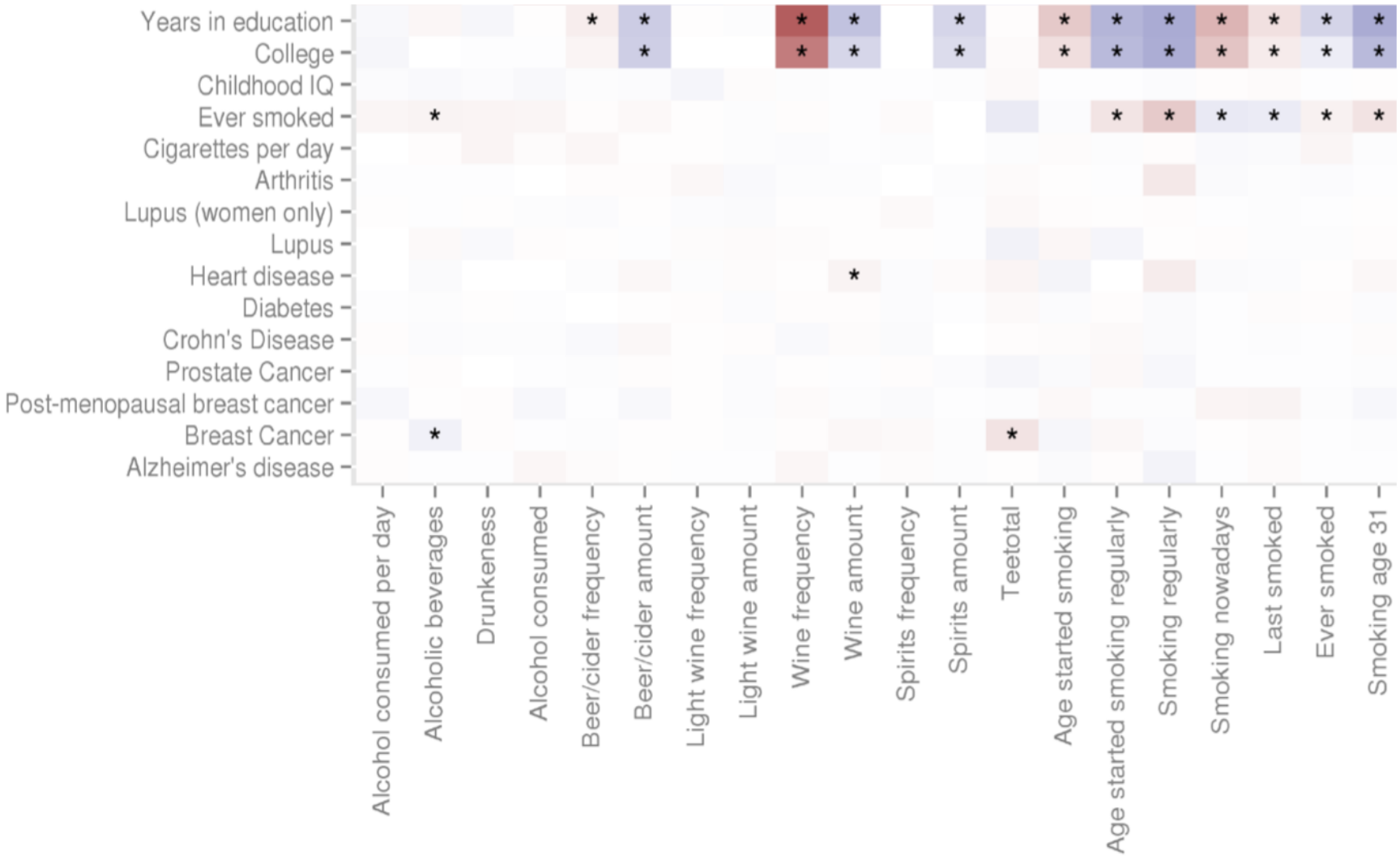
PRSice: Cross-trait PRS study



PRSice: Cross-trait PRS study



PRSice: Cross-trait PRS study



The End!

The End!

Thanks for listening!!