Implementation, improvement and application of Polygenic Risk Scores

Jack Euesden
PhD Student
Statistical Genetics Unit
King's College London

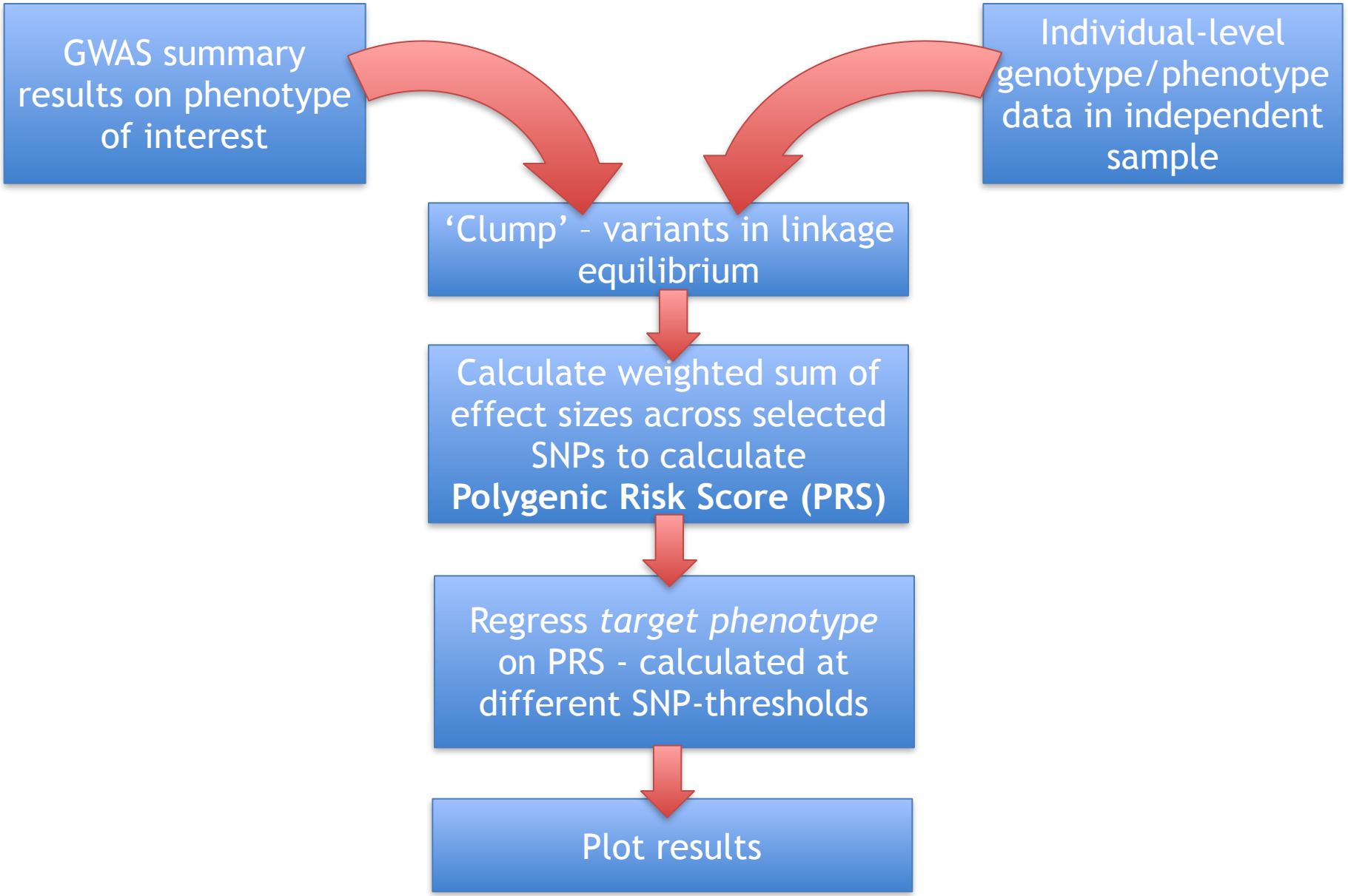
Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

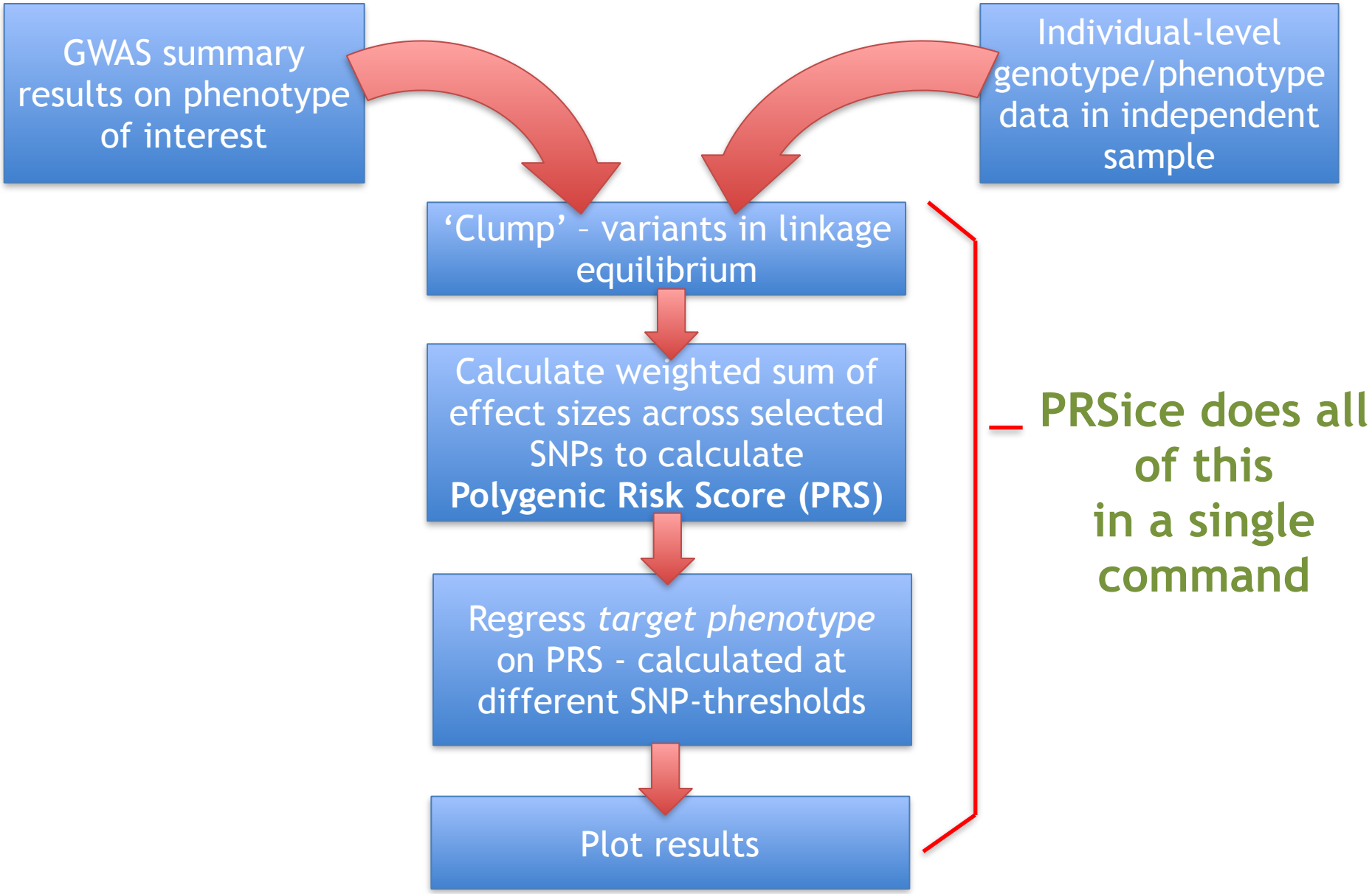
Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

Standard PRS 'Pipeline'



Standard PRS 'Pipeline'



PRSice software: 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

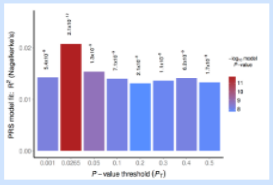
The PRSice vignette can also be obtained directly here: PRSice Vignette

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

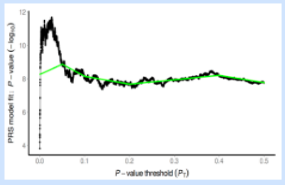
If you have any questions about PRSice, or would like to be added to the mailing list to receive emails on software updates etc, then please email PRSice.info@gmail.com

---

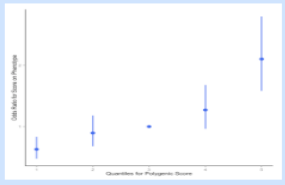
Example Output



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.

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T
```

Set of options supplied to command line

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ###############################
 #
 #   Remove Ambiguous SNPs
 #
 ###############################
 ###############################
 #
 #   Clump
 #
 ###############################
 ###############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ###############################
 ###############################
 #
 #   Polygenic scoring!
 #
 ###############################
 ###############################
 #
 #   Covary by generated dimensions
 #
 ###############################
 ###############################
 #
 #   Regression Models
 #
 ###############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ###############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ###############################
```

Set of options supplied to command line

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

Clean input data

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

Prepare SNPs in Linkage Equilibrium

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

Calculate Polygenic Scores

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

(optionally) generate covariates

www.PRSice.info

Running PRSice

```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

Regress score on phenotype, across thresholds

www.PRSice.info

Running PRSice

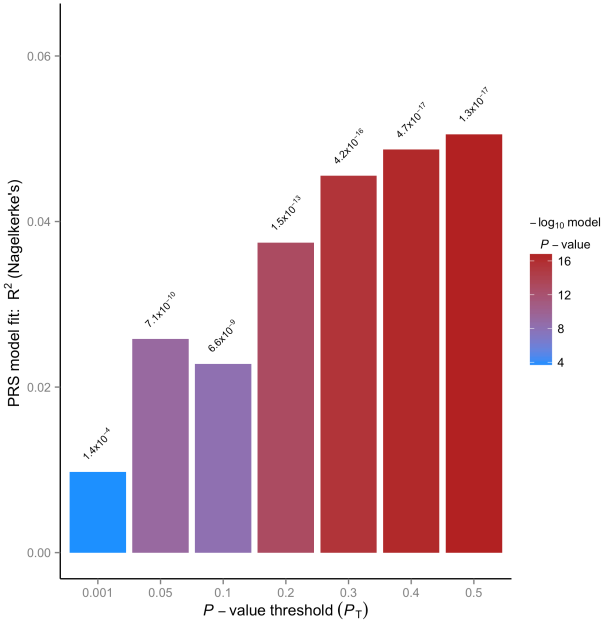
```
R --file=./PRSice_v1.23.R -q  --args base TOY_BASE_GWAS.assoc target TOY_TARGET_DATA plink ./plink_1.9_mac_160914 fastscore T

 ##############################
 #
 #   Remove Ambiguous SNPs
 #
 ##############################
 ##############################
 #
 #   Clump
 #
 ##############################
 ##############################
 #
 #   Deal with strand flips if target is in genotype format and produce input files for polygenic scoring
 #
 ##############################
 ##############################
 #
 #   Polygenic scoring!
 #
 ##############################
 ##############################
 #
 #   Covary by generated dimensions
 #
 ##############################
 ##############################
 #
 #   Regression Models
 #
 ##############################
Regression Models: 10% Complete
Regression Models: 30% Complete
Regression Models: 40% Complete
Regression Models: 60% Complete
Regression Models: 70% Complete
Regression Models: 90% Complete
Regression Models: 100% Complete
 ##############################
 #
 #   Barplots
 # Bars for inclusion can be changed using the barchart.levels option
 ##############################
```

Generate Plots

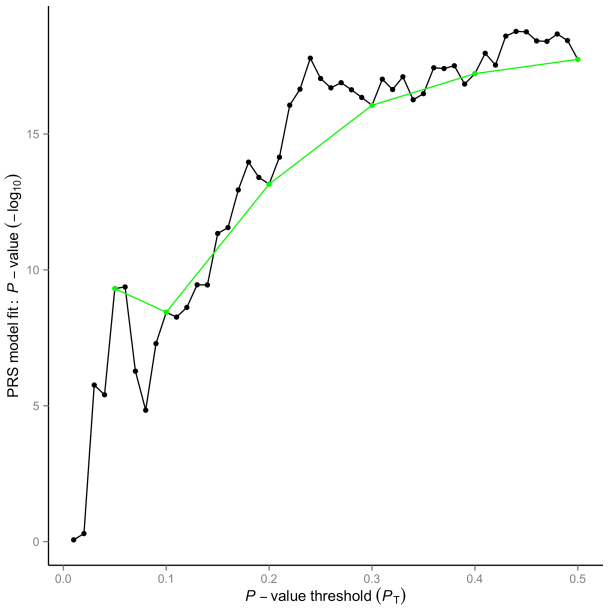
www.PRSice.info

PRSice plots

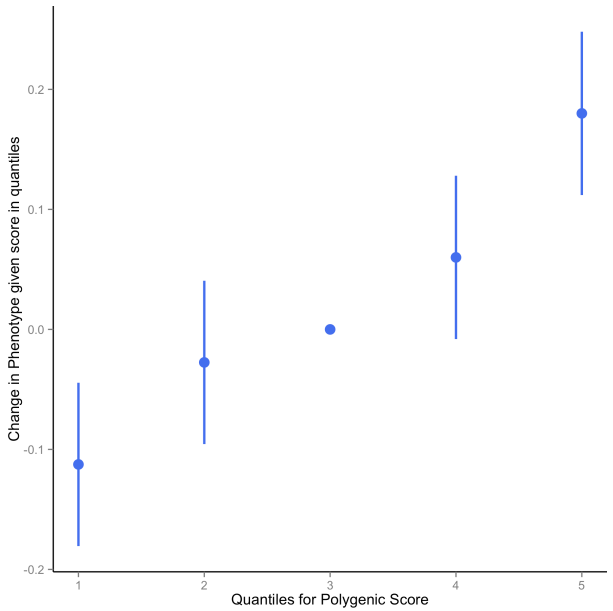
Bar plot



High-Resolution Plot



Quantiles Plot



Additional data outputs from PRSice:

- Polygenic Scores for each individual, at each threshold
- Model fit measures at each threshold

www.PRSice.info

Overview

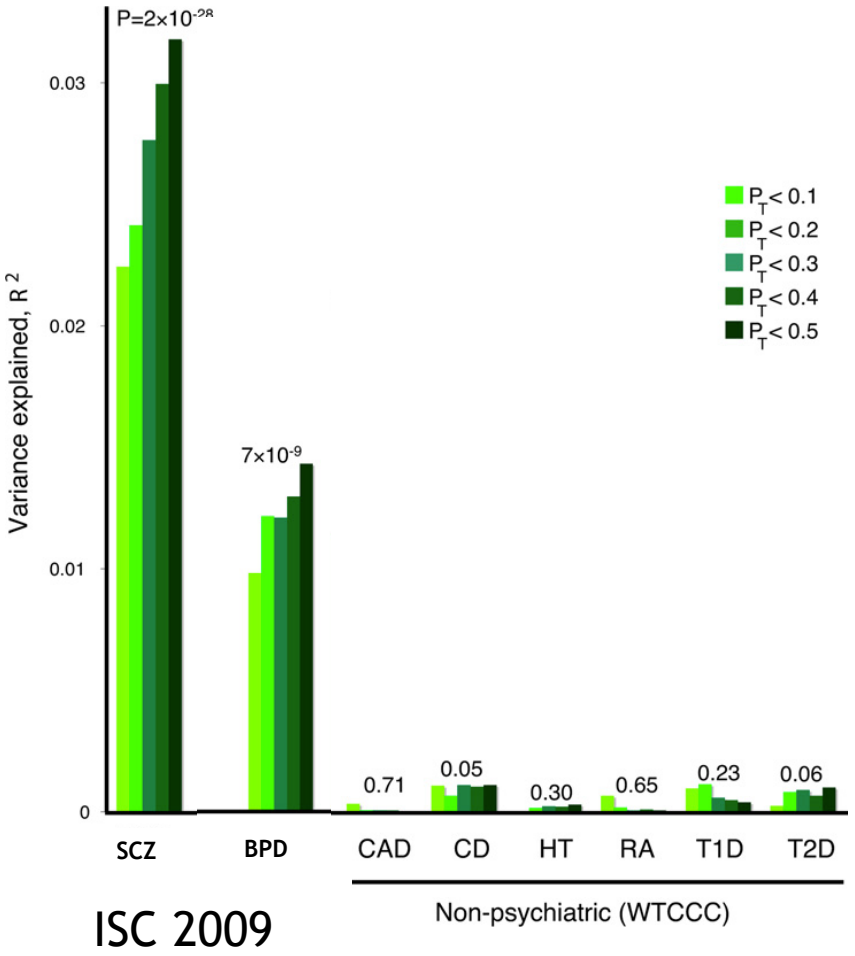
- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

High-resolution scoring in PRS

- It is standard to show the results from PRS regression testing at a small number of thresholds

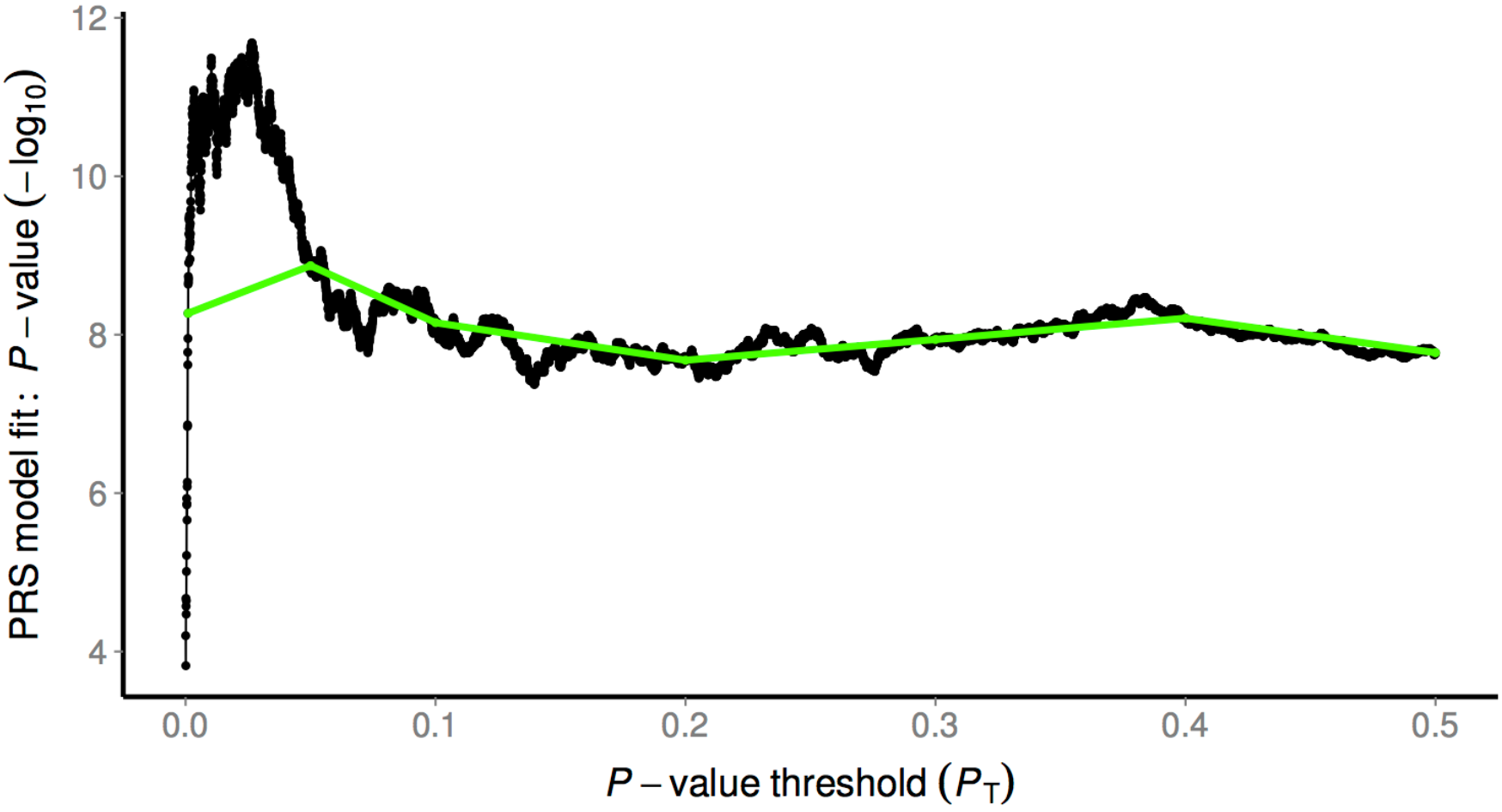


ISC 2009

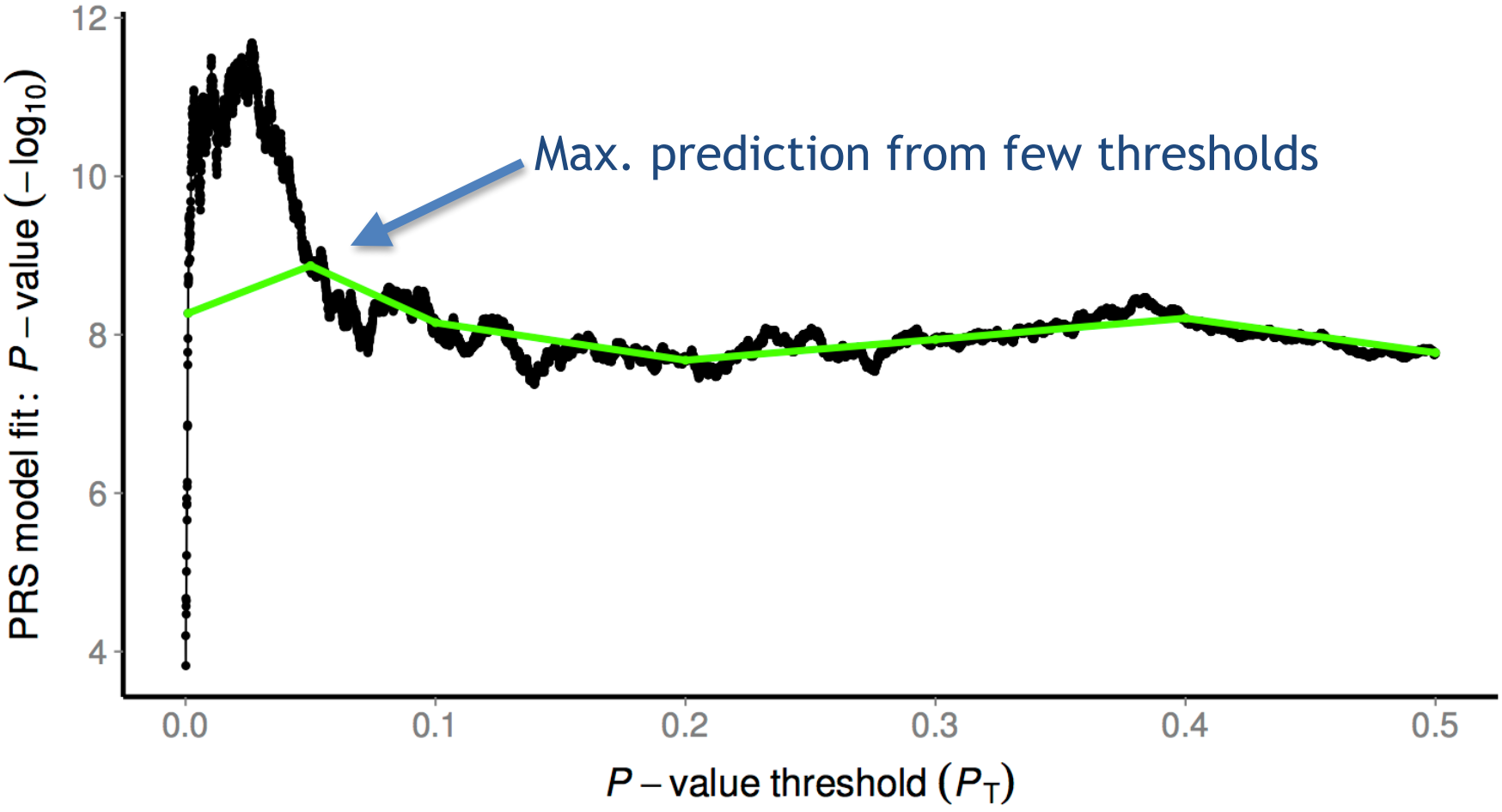
High-resolution scoring in PRS

- Improvement: find optimum threshold for maximum prediction
  - Standard in statistics to search for **most predictive model**
  - Most predictive model may be poorly captured by using a small number of thresholds
  - Performing a small number of tests is not the best solution to the multiple testing problem

High-resolution scoring in PRSice

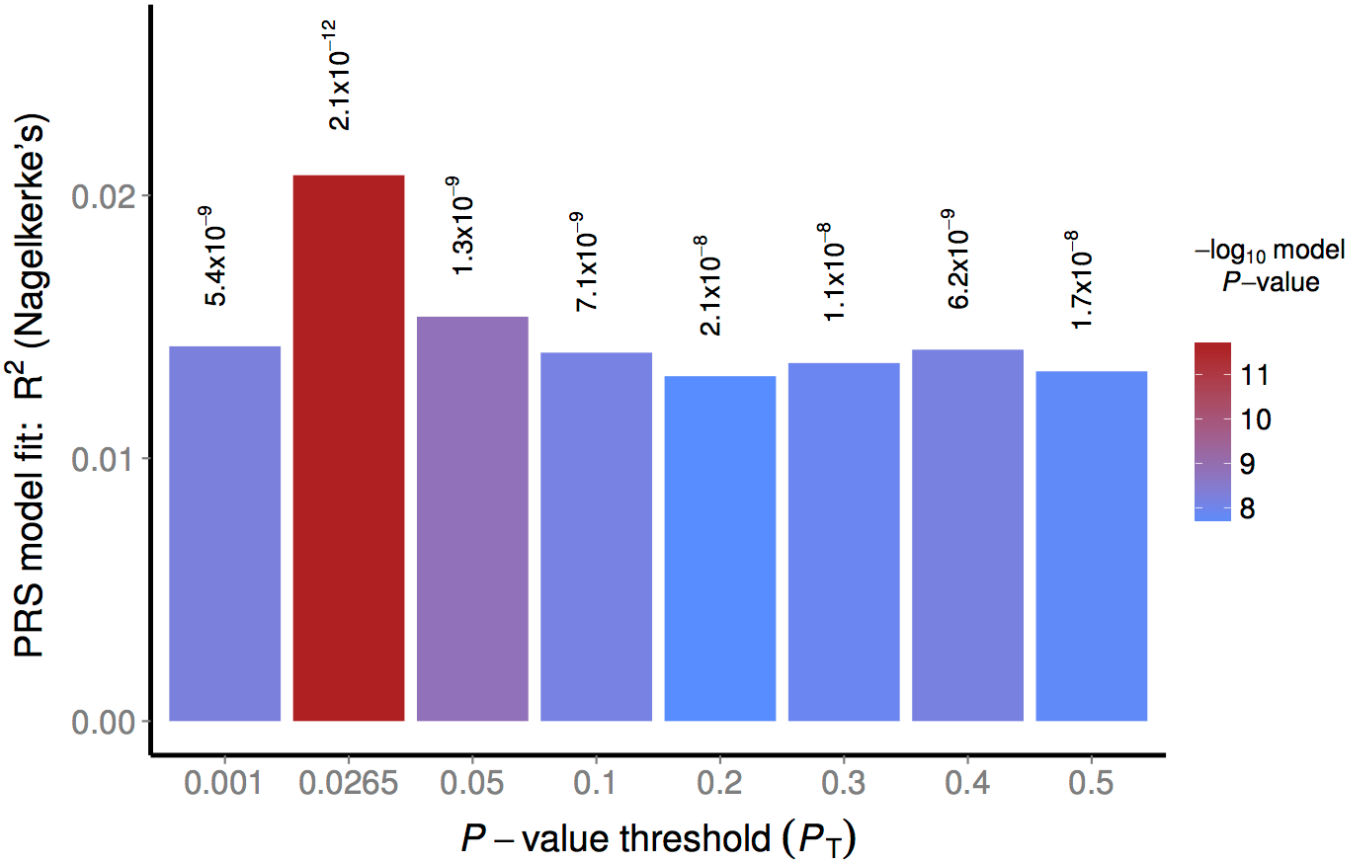


High-resolution scoring in PRSice



High-resolution scoring in PRS

- Most predictive (high-res) bar included



Adjusting multiple testing via permutation

- Testing thousands of thresholds – multiple testing problem?

- These tests are highly correlated – simulation study to find effective number of multiple tests

  - 10,000 permutations
  - 10,000 thresholds
  - Permuted alpha threshold = 0.004
  - Suggest alpha = 0.001

PRSice paper

Bioinformatics, 31(9), 2015, 1466–1468
doi: 10.1093/bioinformatics/btu848
Advance Access Publication Date: 24 December 2014
Applications Note

Genome analysis

PRSice: Polygenic Risk Score software

Jack Euesden*, Cathryn M. Lewis and Paul F. O'Reilly*

MRC Social, Genetic and Developmental Psychiatry Centre, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on September 16, 2014; revised on November 26, 2014; accepted on December 22, 2014

We suggest a significance threshold of $P < 0.001$ for high resolution scoring

www.PRSice.info

Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

To Clump or not to Clump?

- In summing genetic effects genome-wide we assume independence between SNPs
- However, we may want to include multiple nearby SNPs due to allelic heterogeneity
- Clumping seeks to solve this: SNPs are 'pruned' by taking the lowest *P*-value in an LD window – usually
  - Window of 250Kb
  - R^2 of 0.2

To Clump or not to Clump?

- Alternative: LASSO/ Elastic Net
  - Step 1: Multiply genotypes in TARGET data by BETA from BASE GWAS
  - Step 2: Regress all modified genotypes on phenotype using Penalised Regression
  - Step 3: Construct PRS from all SNPs retained in this new model

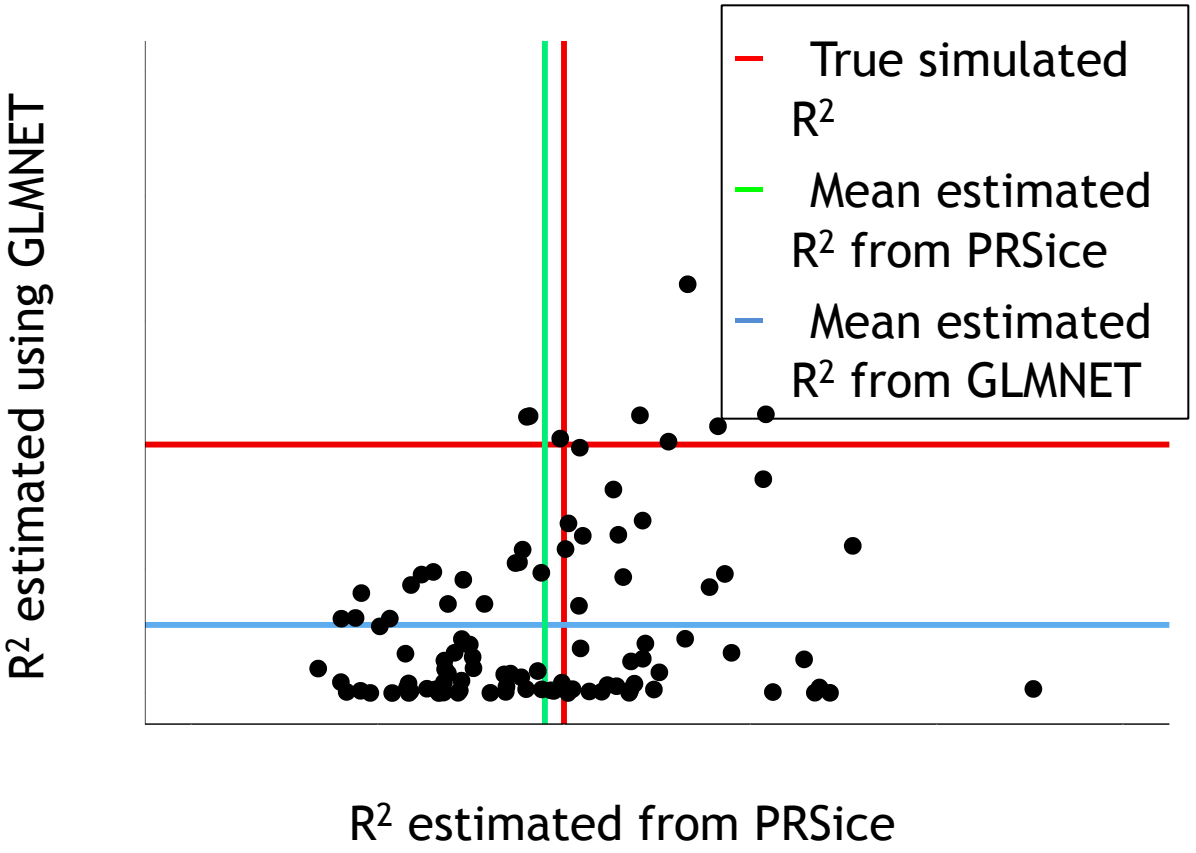
To Clump or not to Clump?

- Test in simulated data:
  - 5000 genotypes from WTCCC1
  - Randomly select 1000 SNPs in sequence
  - Simulate a proportion of causal SNPs
  - Simulate a quantitative trait with fixed h^2
  - Split into **base** and **target** 2:1
  - Test performance of method vs PRSice
  - 100 simulations per scenario

To Clump or not to Clump?

Simulate:

- 5 causal SNPs in 50 SNP window
- Effect sizes follow exponential distribution
- Select SNPs with:
  - Elastic Net
  - PRSice
- Compare performance



Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Alternative to clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

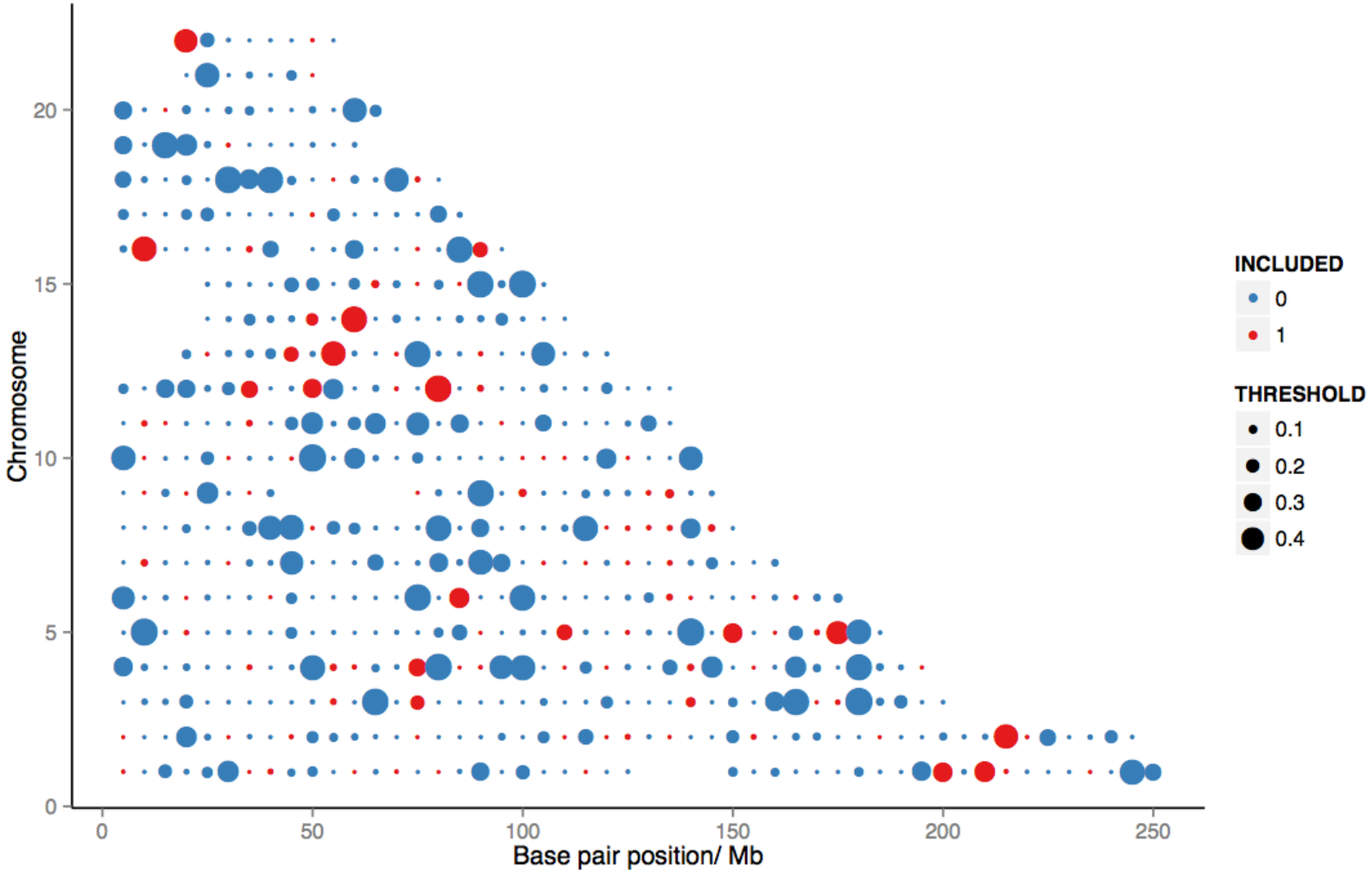
Tailoring PRS to scientific question

- The standard PRS method has been used in multiple applications, sometimes with different underlying hypotheses
- Power should be optimised by tailoring PRS methods to the corresponding scientific question
- We develop a new method to best score for use as a biomarker
- This is a specific scientific question, UNLIKE:
  - Assessing level of genetic overlap
  - Demonstrating a trait can predict itself

Tailoring PRS to a Scientific Question

- Method:
  - Split genome into chunks – e.g. 5Mb
  - At each chunk, regress lots of thresholds on phenotype and pick the best threshold
  - Retain chunks that predict phenotype
  - Sum these to make new score

Using PRS as a Biomarker



Overview

- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Replacing clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

Real data application – PRS as a Biomarker

- SCZ GWAS has higher power than MDD
- PGC SCZ predicting MDD
  - NB: performs best on different disorders
- Final model – $P = 4.47 \times 10^{-33}$

Real data application – PRS as a Biomarker

- Overfit?
- Run 100 permutations, calculate empirical *P*-value
- Empirical $P = 1.79 \times 10^{-28}$

Real data application – PRS as a Biomarker

- New method – 'PRS*l*ice'

- Utility?
  - Biomarker for high-risk individuals
  - Leverage shared component between two disorders to predict individual risk

Overview

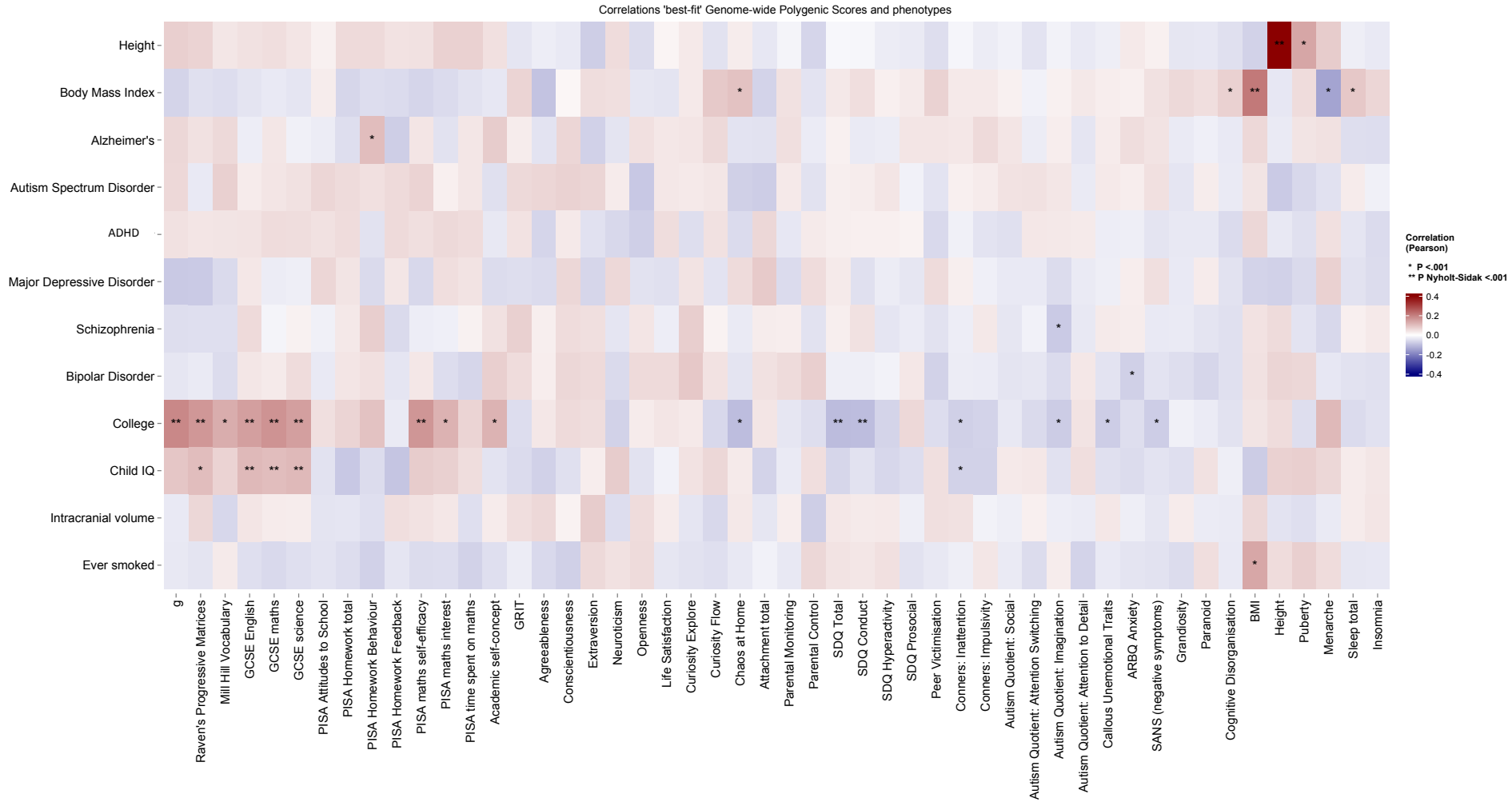
- Implementation via our PRSice software
- Improvements to PRS:
  - High-resolution PRS to increase power
  - Replacing clumping to capture more risk variants
  - PRS methods tailored to scientific question
- PRS applications:
  - PRS biomarker method applied to real data
  - 2 large cross-disorder analyses

Educational Phenotypes

- Use GWAS investigating a large number of physical and psychiatric traits
- Genotype data from Twins Early Development Study (TEDS)
- Test prediction on educational phenotypes, eg:
  - Maths age 16
  - Inattention
  - Imagination

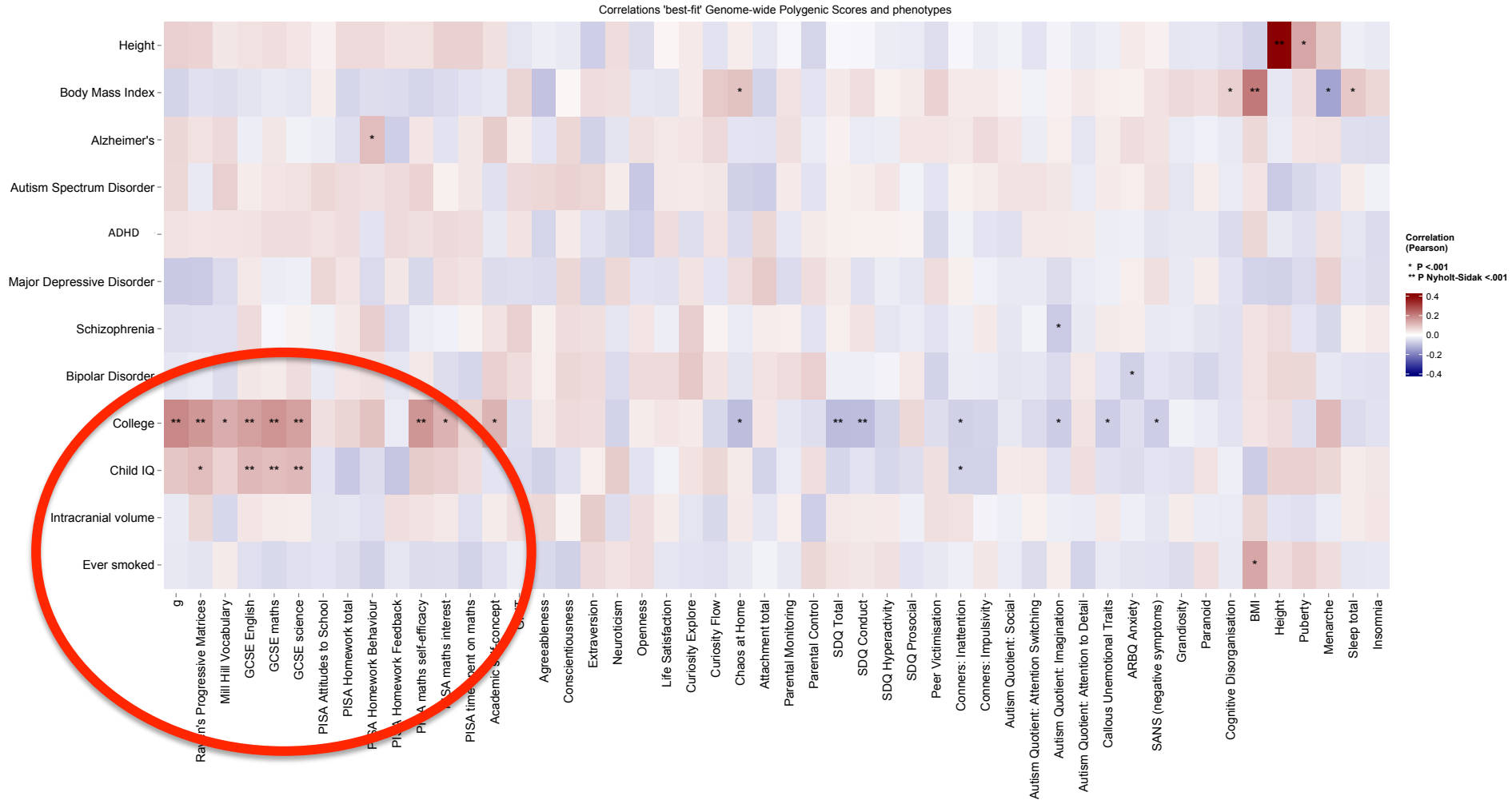
Krapohl et al 2015

Educational Phenotypes



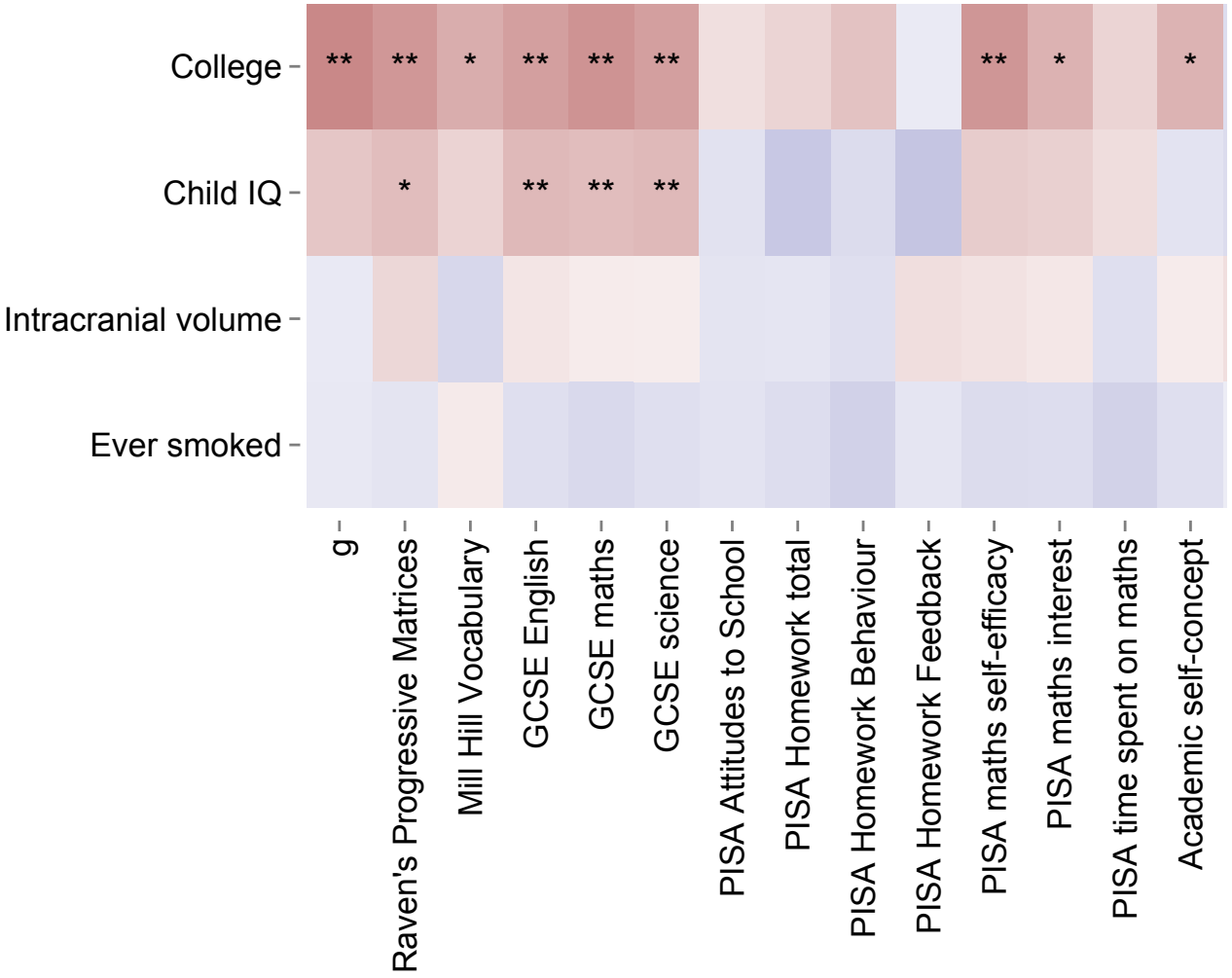
Krapohl et al 2015

Educational Phenotypes



Krapohl et al 2015

Educational Phenotypes



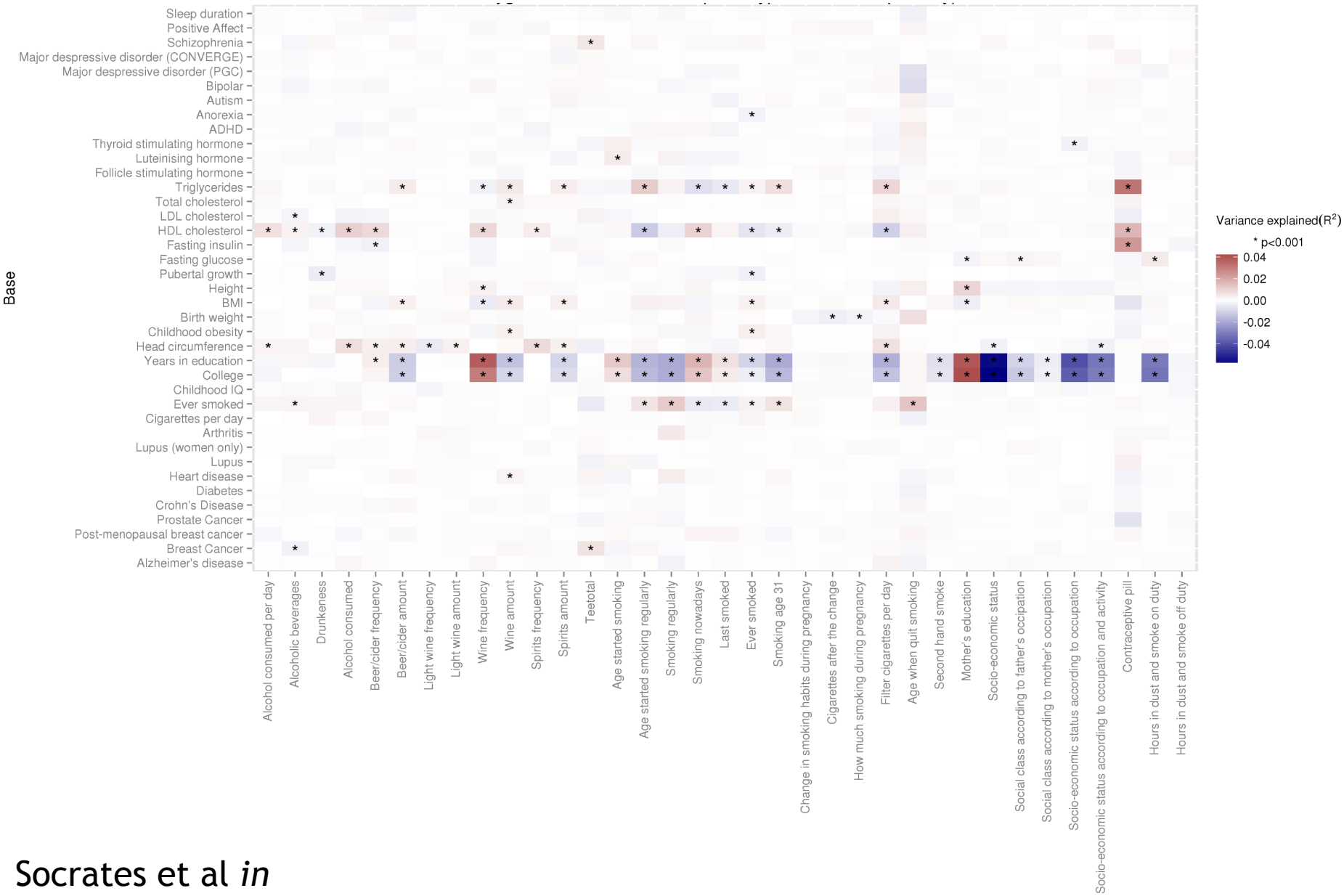
Krapohl et al 2015

Social Phenotypes

- Use GWAS investigating a large number of physical and psychiatric traits
- Genotype data from North Finland Birth Cohort
- Test prediction on ‘social’ phenotypes, eg:
  - Beer consumption
  - Wine consumption
  - Smoking behaviour

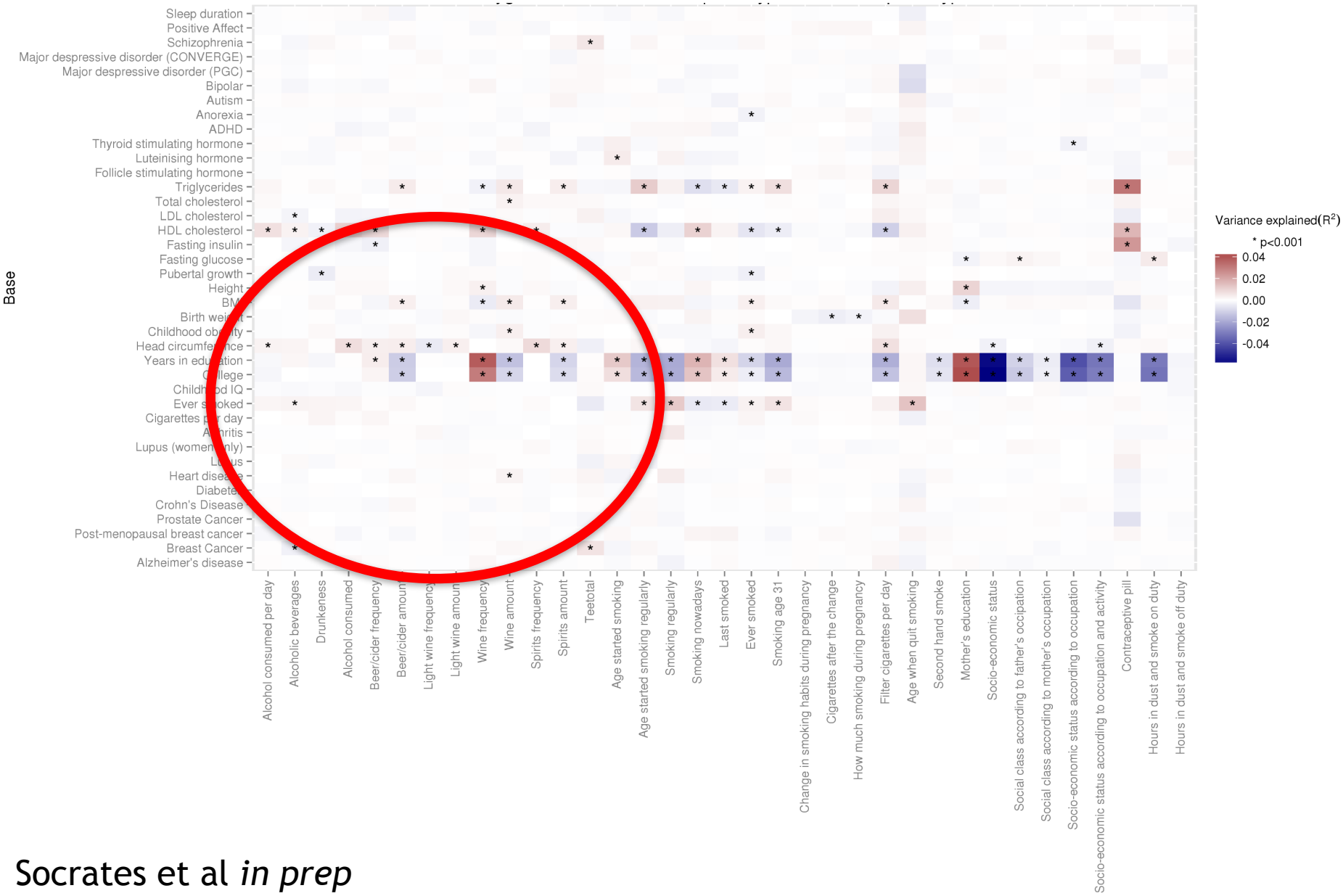
Socrates et al *in*

Social Phenotypes



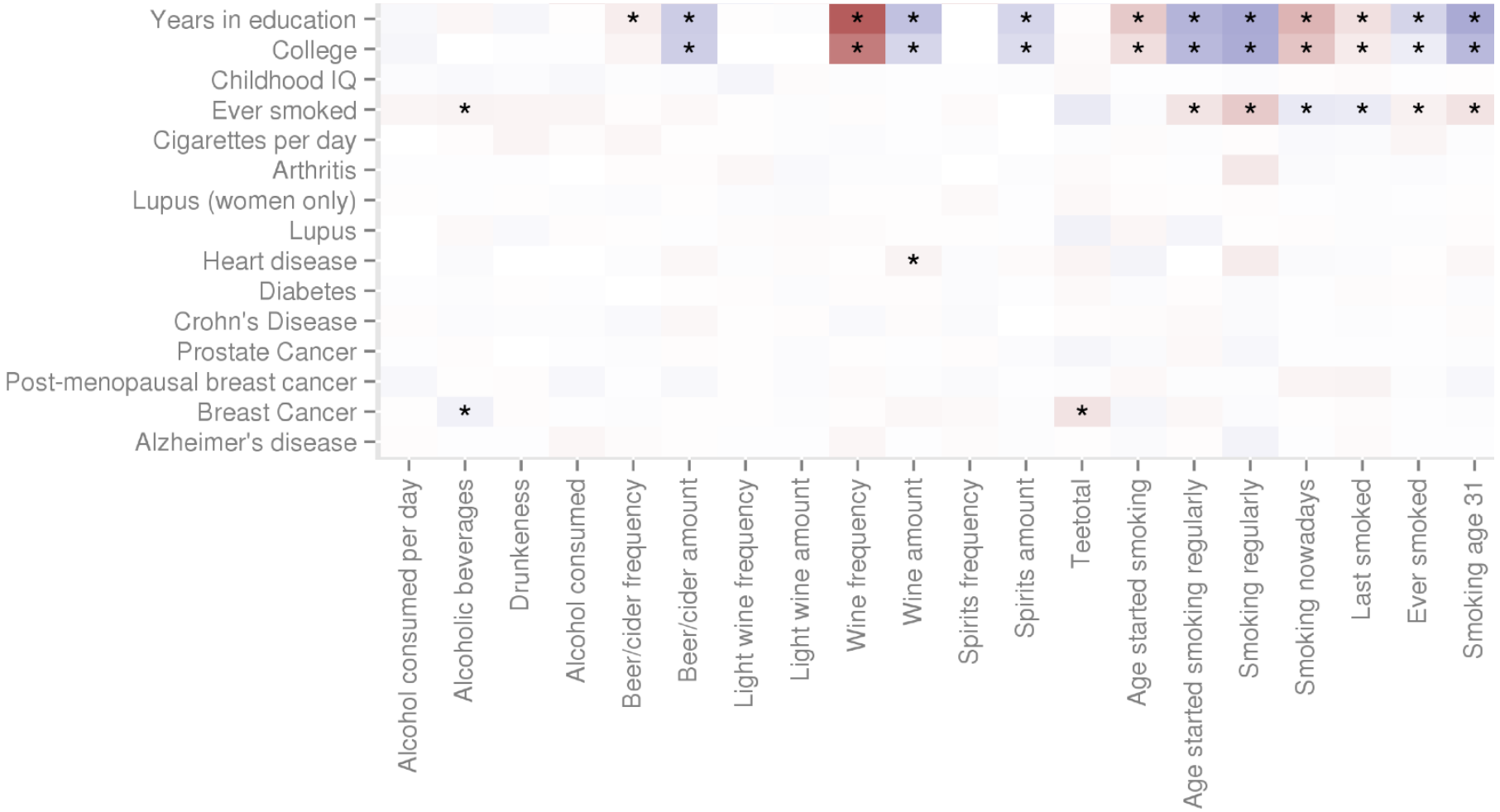
Socrates et al *in prep*

Social Phenotypes



Socrates et al *in prep*

Social Phenotypes



Socrates et al *in prep*

Future Directions

- Investigate biological pathways enriched within optimised threshold
- Consider Conditional and Joint models to improve SNPs in LD selection

Conclusions

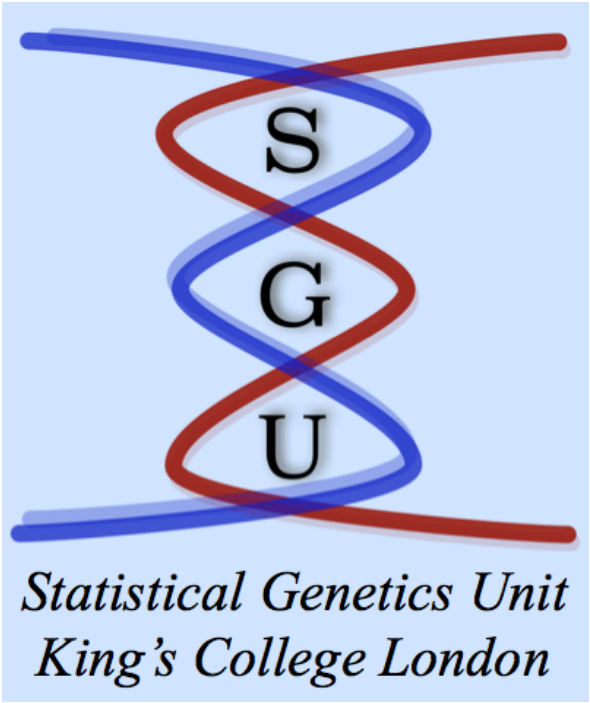
- Improvements to PRS:
  - Threshold selection by ‘high resolution’
  - Chunks optimise thresholds across genome, when using PRS as a ***biomarker***

- No improvement through using penalised regression
  - Consider other methods for achieving Linkage Equilibrium

Acknowledgements

- Paul O'Reilly
- Cathryn Lewis

- Eva Krapohl
- Adam Socrates



MRC | Medical Research Council

Guarantors of Brain

www.PRSice.info