QIMR Berghofer
Medical Research Institute

Introduction to Polygenic Risk Scores: concept and calculation

Baptiste Couvy-Duchesne & Lucia Colodro Conde
(+ credits to Sarah Medland)

3th October 2017

Outline

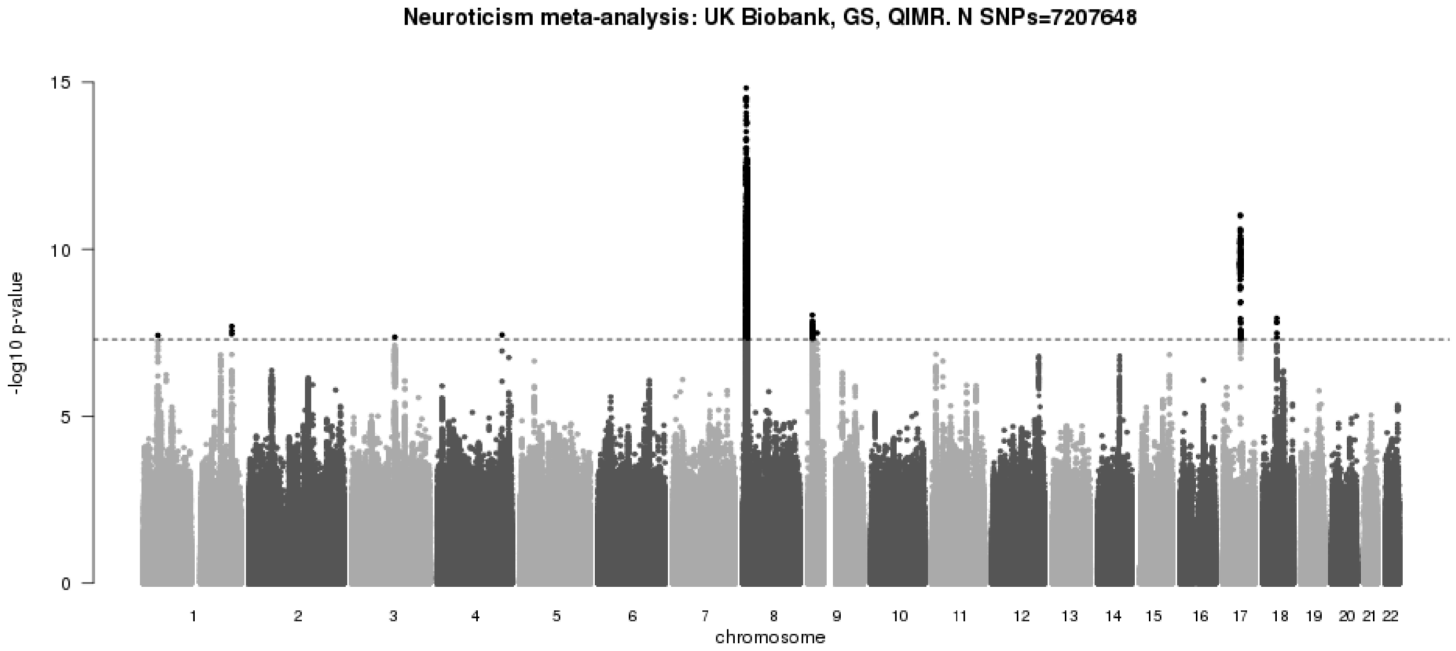
Concept

What are Polygenic risk scores (PRS)?

PRS are a quantitative measure of the cumulative genetic risk or vulnerability that an individual possesses for a trait.

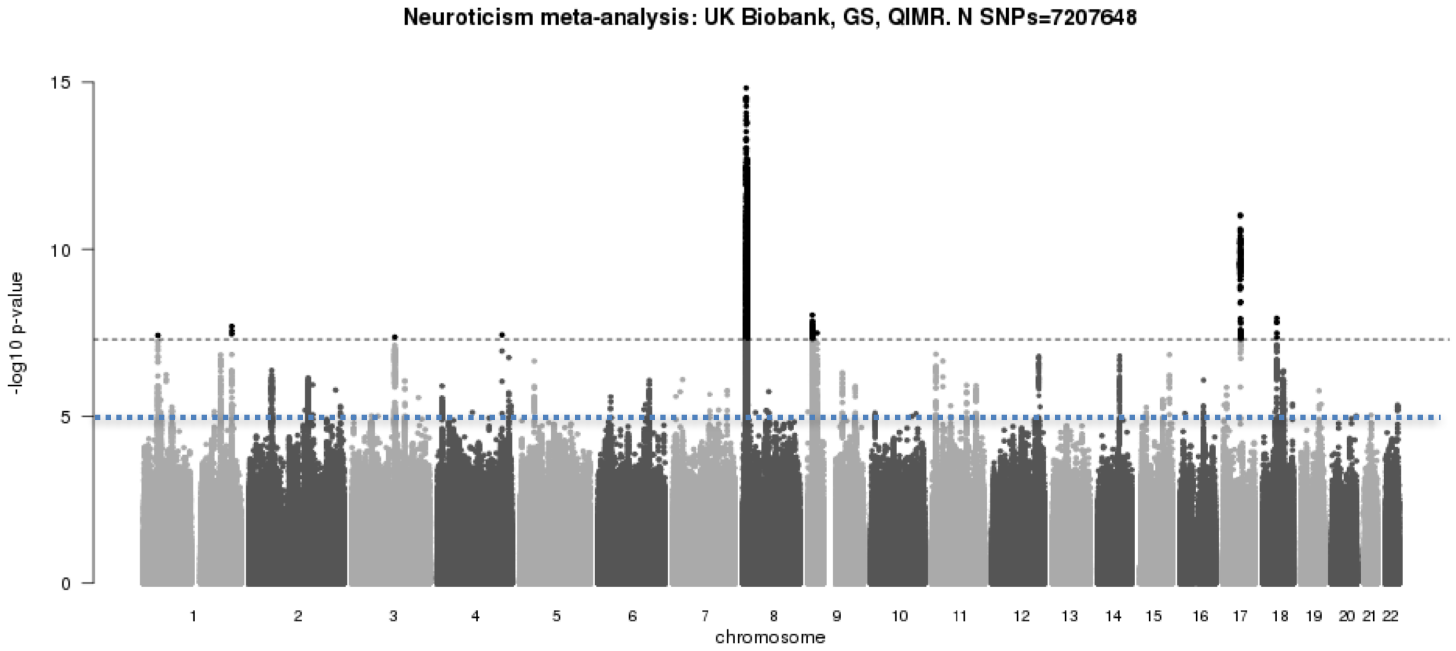
What are Polygenic risk scores (PRS)?

PRS are a quantitative measure of the cumulative genetic risk or vulnerability that an individual possesses for a trait.



What are Polygenic risk scores (PRS)?

PRS are a quantitative measure of the cumulative genetic risk or vulnerability that an individual possesses for a trait.



The classics

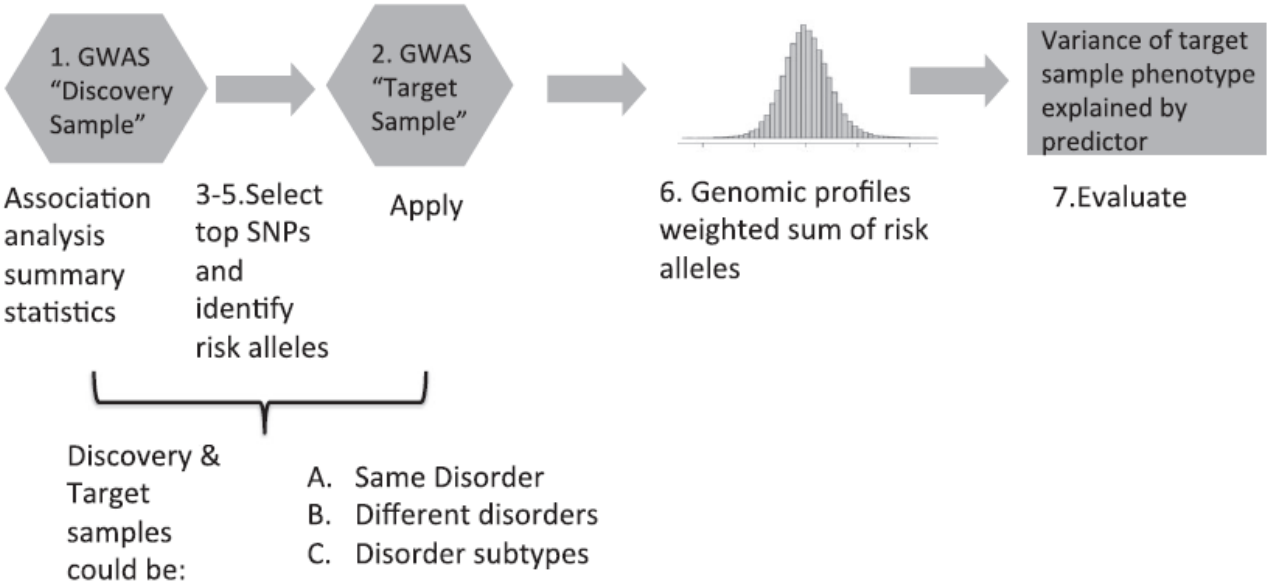
- Wray NR, Goddard, ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007; 7(10):1520-28.
- International Schizophrenia Consortium, Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, Sklar P . Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature. 2009; 460(7256):748-52
- Evans DM, Visscher PM., Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Human Molecular Genetics. 2009; 18(18): 3525-3531



Further reading

- Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013 Mar;9(3):e1003348. Epub 2013 Mar 21. Erratum in: PLoS Genet. 2013;9(4). (**Important discussion of power**)
- Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, Middeldorp CM. Research review: Polygenic methods and their application to psychiatric traits. J Child Psychol Psychiatry. 2014;55(10):1068-87. (**Very good concrete description of the traditional methods**).
- Wray NR, Yang J, Hayes BJ, Price AL, Goddard ME, Visscher PM. Pitfalls of predicting complex traits from SNPs. Nat Rev Genet. 2013;14(7):507-15. (**Very good discussion of the complexities of interpretation**).
- Witte JS, Visscher PM, Wray NR. The contribution of genetic variants to disease depends on the ruler. Nat Rev Genet. 2014;15(11):765-76. (**Important in the understanding of the effects of ascertainment on PRS work**).
- Shah S, Bonder MJ, Marioni RE, Zhu Z, McRae AF, Zhernakova A, Harris SE, Liewald D, Henders AK, Mendelson MM, Liu C, Joehanes R, Liang L; BIOS Consortium, Levy D, Martin NG, Starr JM, Wijmenga C, Wray NR, Yang J, Montgomery GW, Franke L, Deary IJ, Visscher PM. Improving Phenotypic Prediction by Combining Genetic and Epigenetic Associations. Am J Hum Genet. 2015; 97(1):75-85. (**Important for the conceptualization of polygenicity**)

Traditional approach

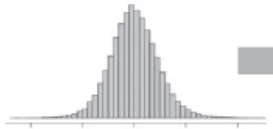


Wray et al (2014) *J Child Psychol Psychiatry*

Traditional approach

1. GWAS "Discovery Sample"

2. GWAS "Target Sample"



Variance of target sample phenotype explained by predictor

Association analysis summary statistics

3-5.Select top SNPs and identify risk alleles

Apply

6. Genomic profiles weighted sum of risk alleles

7.Evaluate

Discovery & Target samples could be:

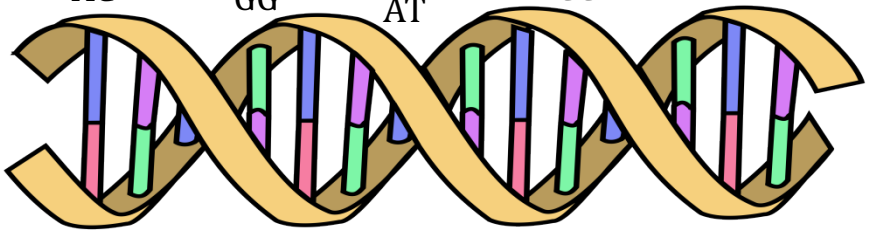
A. Same Disorder
B. Different disorders
C. Disorder subtypes

$1\times-.02 + 2\times.01 + 1\times.002 + 0\times.03 + 2\times.025$

Polygenic score:

.052

AC GG AT CC TT



$\beta_C=-.02$ $\beta_G=.01$ $\beta_A=.002$ $\beta_G=.03$ $\beta_T=.025$

Effect size from GWAS

Wray et al (2014) *J Child Psychol Psychiatry*

Usage

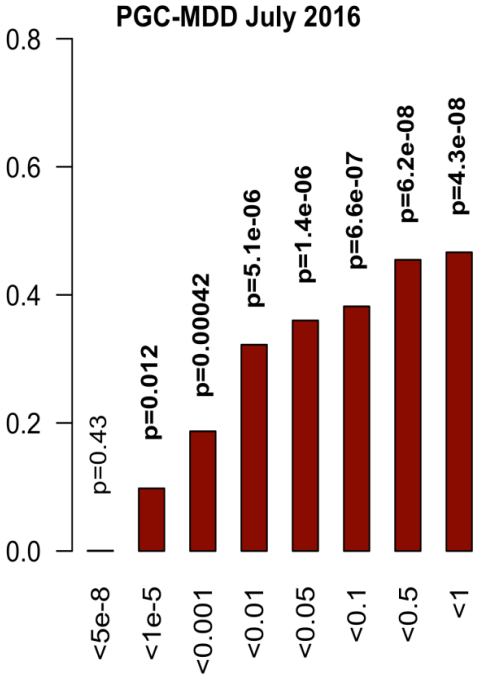
Main uses of PRS

1) Single disorder analyses

2) Cross-disorder analysis

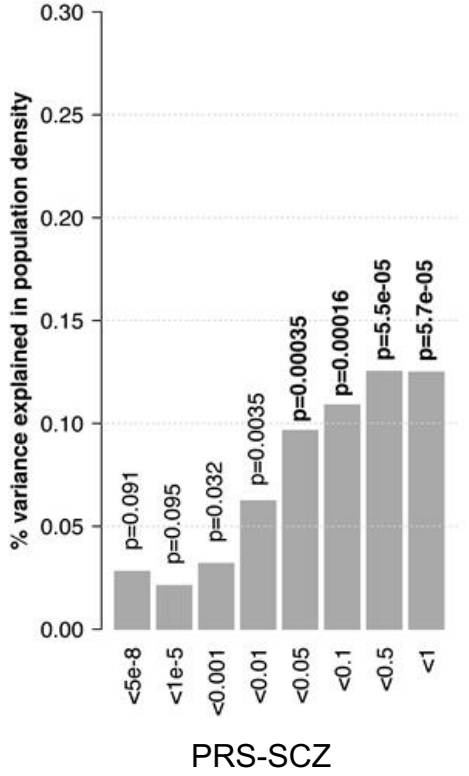
3) Sub-type analysis

Single trait analyses



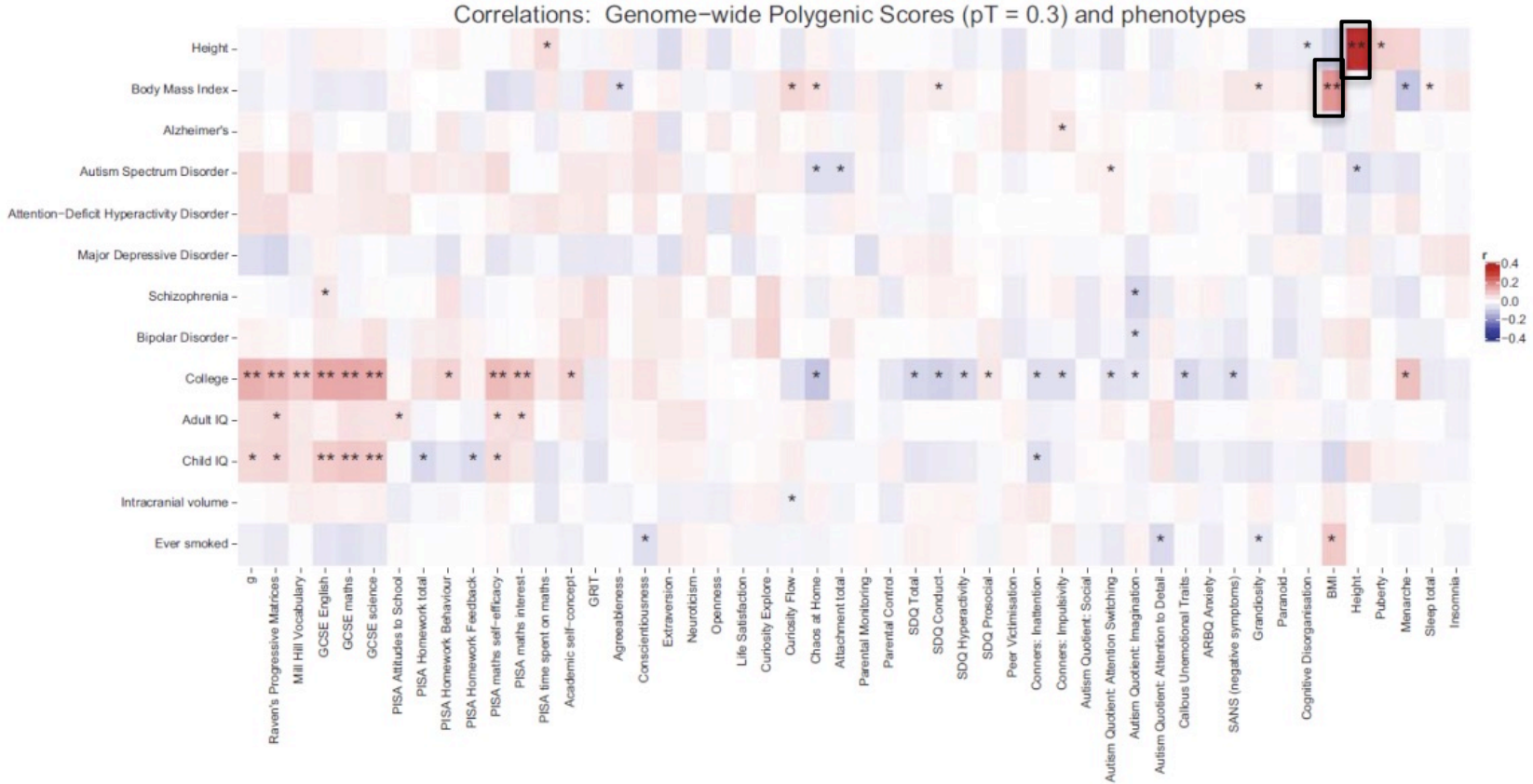
Colodro-Conde L,
Couvy-Duchesne B, et al, (2017)
Molecular Psychiatry

Cross-trait analysis



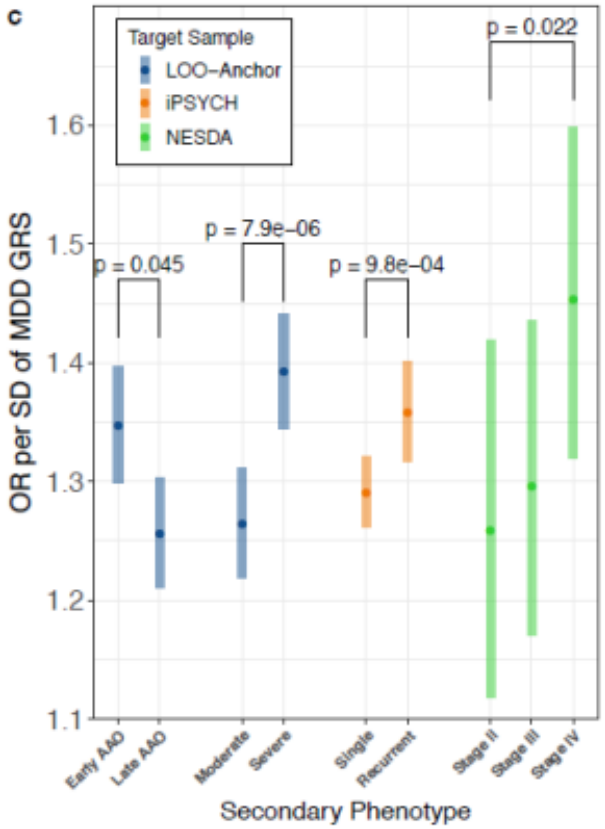
Colodro-Conde L, Couvy-Duchesne B, et al, 2017 *BioRxiv*

Single and cross-trait analyses



Krapohl et al (2016) *Molecular Psychiatry*

Sub-type analysis



Wray NM, et al, 2017 *BioRxiv*

PRS and power

The power of the predictor is a function of the power of the GWAS in the discovery sample (due to its impact on the accuracy of the estimation of the betas).

"I show that discouraging results in some previous studies were due to the low number of subjects studied, but a modest increase in study size would allow more successful analysis. However, I also show that, for genetics to become useful for predicting individual risk of disease, hundreds of thousands of subjects may be needed to estimate the gene effects."

(Dudbridge, 2013)

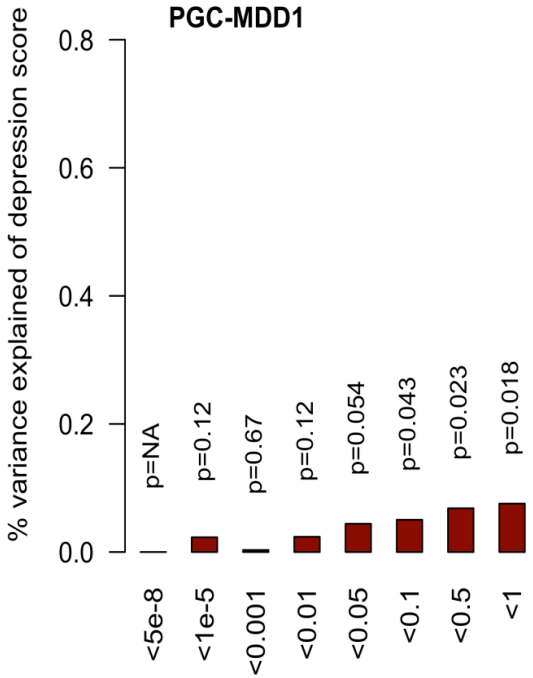
PRS and power

For simple power calculations you can use a simple regression power calculator (for r2 = 0.5% of the variance).
As a general rule of thumb you usually want 2,000+ people in the target dataset.

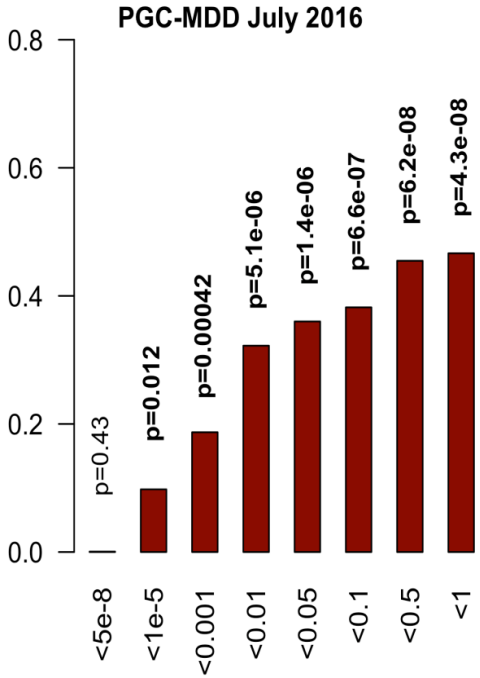
→ R AVENGEME (https://github.com/DudbridgeLab/avengeme)

```
sampleSizeForGeneScore(targetQuantity, targetValue, nsnp, n2 = NA, vg1 = 0,
  cov12 = vg1, pi0 = 0, weighted = TRUE, binary = FALSE,
  prevalence = 0.1, sampling = prevalence, lambdaS = NA,
  shrinkage = FALSE, logrisk = FALSE, alpha = 0.05, r2gx = 0,
  corgx = 0, r2xy = 0, adjustedEffects = FALSE)
```

PRS and power



PGC-MDD1: N=18k
max variance explained = 0.08%,
p=0.018



PGC-MDD2: N=163k
max variance explained =0.46%,
p= 5.01e-08

Colodro-Conde L,
Couvy-Duchesne B, et al, (2017)
Molecular Psychiatry

GENERAL STEPS OF PROCESSING

Overview of methods

Classic / OLS	BLUP	PRSice
dosage	Best guess	Dosage or best guess
clumping	BLUP effects summed over all SNPs	Clumping within windows
Multiple PRS by p-value thresholds	Unique PRS	All p-value threshold tested

All methods only require GWAS summary statistics and target sample

(1) GWAS summary statistics

SNP identifier

- rs999
- Chr:BP 2:2450
- Chr:BP:Alleles 2:2450:AAA_T
- Chr:BP: SNP/INDEL 2:2450:SNP or 2:2450:INDEL

Both Alleles (reference / alternative)

Pvalue

Effect

- Beta from association with continuous trait
- OR from an ordinal trait - convert to log(OR)

(2) Find SNPs in common with your local sample and QC

- Maximise overlap using HRC imputation (Michigan Imputation Server: https://imputationserver.sph.umich.edu/index.html)

/reference/genepi/GWAS_release/Release8/Release8_HRCr1/info/Metadata_allchr.txt

- QC
  - R^2 >=0.6
  - MAF>=0.01
  - No indels
  - Coherent strands

SNP name: Chromosome:bp
REF: reference allele
ALT: alternative allele
bp_Build37: SNP basepair number
SNP_dbSNP: SNP rs number
MAF: Minor Allele Frequency
Rsq_rederived: R squared (R2) corresponding to the quality of imputation across platforms

(3) Clumping

- Select most associated SNP per **LD region**
- Plink1.9 **--bfile** bfileReferencePanelForLD
  --extract QCedListofSNPs
  --clump gwasFileWithPvalue
  --clump-p1 1
  --clump-p2 1
  --clump-r2 0.1
  --clump-kb 10000
  --out OutputName

--bfile
/mnt/lustre/reference/genepi/public_reference_panels/1000G_20101123_v3_GIANT/derived_plinkformat/chr"$j".phase1_release_v3.20101123.snps_indels_svs.genotypes.refpanel.ALL

(4) Calculate risk scores

- Plink1.9 **--dosage** dosageData **format=1**
  --fam famFile.fam
  --score effectSizesClumpedSNP
  --q-score-range pvalueThresholds.txt
  --q-score-file pvaluesClumpedSNP
  --out Output

(4) Calculate risk scores

--dosage

/mnt/lustre/reference/genepi/GWAS_release/Release8/Release8_HRCr1/PLINK_dosage/BlockPLINK_chr"$j"."$k".dose.gz

--fam

/mnt/lustre/reference/genepi/GWAS_release/Release8/Release8_HRCr1/PLINK_dosage/GWAS.fam

--q-score-range

S1	0.00	0.00000005
S2	0.00	0.00001
S3	0.00	0.001
S4	0.00	0.01
S5	0.00	0.05
S6	0.00	0.1
S7	0.00	0.5
S8	0.00	1.0

(5) Run association analysis, controlling for relatedness

```
gcta    --reml
        --mgrm-bin GRM
        --pheno phenotypeToPredict.txt
        --covar discreteCovariates.txt
        --qcovar quantitativeCovariates.txt
        --out Output
        --reml-est-fix
        --reml-no-constrain
```

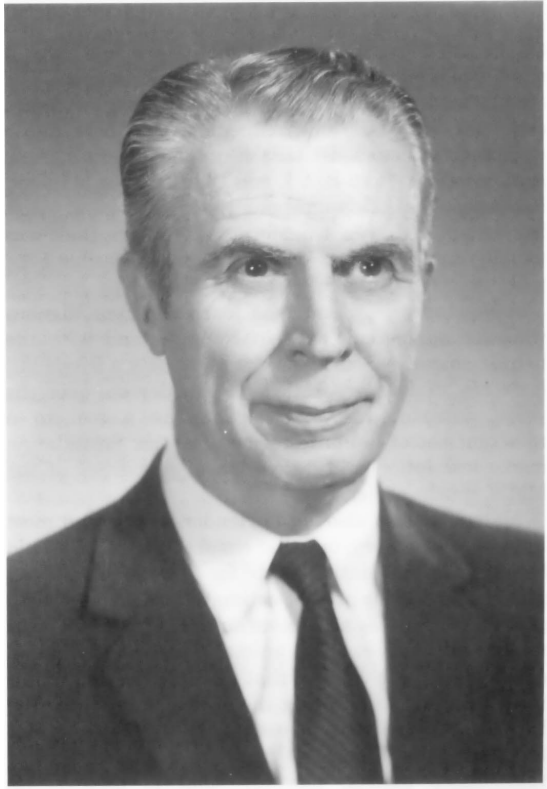
--mgrm-bin
/reference/genepi/GWAS_release/Release8/Release8_Observed/GRM/GRM_allindividuals_AncestryFiltered_autosomes

Other methods

Genetic Best Linear Unbiased Predictor

Application to genetic data (animal breeding) HENDERSON, C. R. (1950). Estimation of genetic parameters

Review of method and example:
Henderson, C. R. (1975). Best Linear Unbiased Estimation and Prediction under a Selection Model



Charles Roy Henderson
1911-1989

BLUP in context of linear models

GWAS estimates: marginal SNP effect

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

N individuals
Y_{Nx1} phenotype centered
X_{Nx1} SNP centered

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Joint and conditional SNP effect

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

N individuals
Y_{Nx1} phenotype centered
X_{Nxm} SNPs centered

Yang et al., 2012

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

BLUP effect

$$\mathbf{y} = \mathbf{Zs} + \mathbf{e},$$

N individuals
Y_{Nx1} phenotype centered
Z_{Nxm} SNPs centered
s_{mx1} vector of SNP effects assumed ~N(0, $\boldsymbol{\sigma}^2_s$)

Goddard et al., 2009

$$\hat{\mathbf{s}} = \mathbf{Z}'(\mathbf{ZZ}'\sigma_s^2 + \mathbf{I}\sigma_e^2)^{-1}\mathbf{y}$$

Calculating BLUP effect sizes

$$\mathbf{y} = \mathbf{Zs} + \mathbf{e},$$
$$\hat{\mathbf{s}} = \mathbf{Z}'(\mathbf{ZZ}'\sigma_s^2 + \mathbf{I}\sigma_e^2)^{-1}\mathbf{y}$$

Z'Z: nxn variance-covariance matrix of genotypes
Often not available from GWAS
Can be estimated from the GWAS allele frequencies and LD from a reference panel (assumed same population)
Yang et al., 2012

```
gcta64    --bfile ReferencePanelForLD
          --cojo-file GWAS_sumstat.ma
          --cojo-sblup 1.33e6
          --cojo-wind 1000
          --thread-num 20
```

BLUP limitations and perspective

$$\mathbf{y} = \mathbf{Z}\mathbf{s} + \mathbf{e},$$

$$\hat{\mathbf{s}} = \mathbf{Z}'(\mathbf{Z}\mathbf{Z}'\sigma_s^2 + \mathbf{I}\sigma_e^2)^{-1}\mathbf{y}$$

Requires to inverse $(\mathbf{Z}\mathbf{Z}'\sigma_s^2 + \mathbf{I}\sigma_e^2)$

Which can be computationally intensive for large sample sizes

Open field of prediction models

- BLUP "shrinks" the estimates: hypothesis of normally distributed effect sizes "infinitesimal model"
- Other shrinkage methods include LASSO: hypothesis of mixture of effect sizes (double exponential…)

- Non-additive models? That may include epistasis, dominance
- Semi-parametric models
  see Goddard et al., 2009 for review

LDpred

Bayesian estimation of the BLUP effect sizes: “posterior mean effect size of each marker by using a prior on effect sizes and LD information from an external reference panel”
Vilhjalmsson et al., 2015

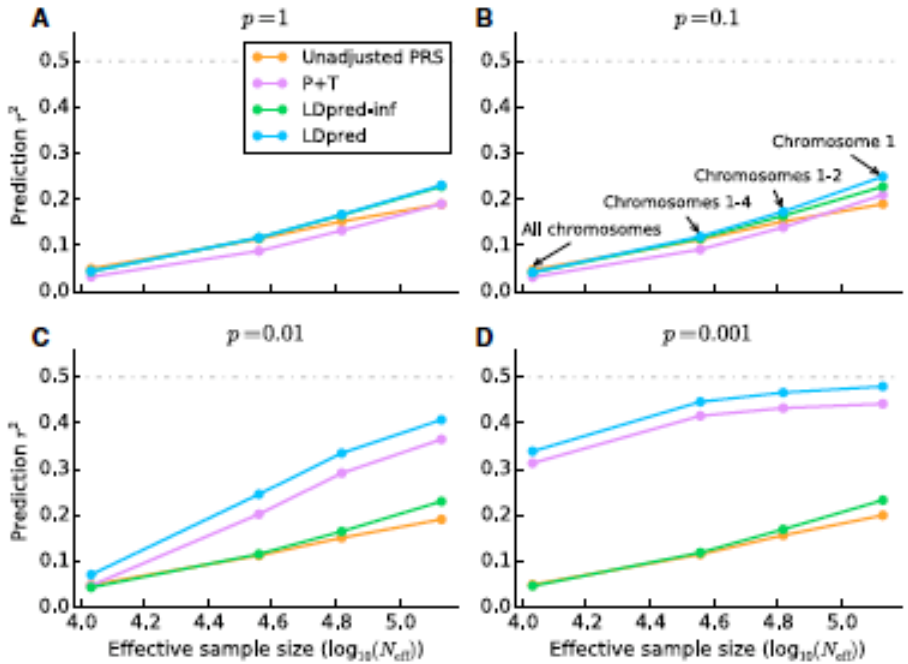
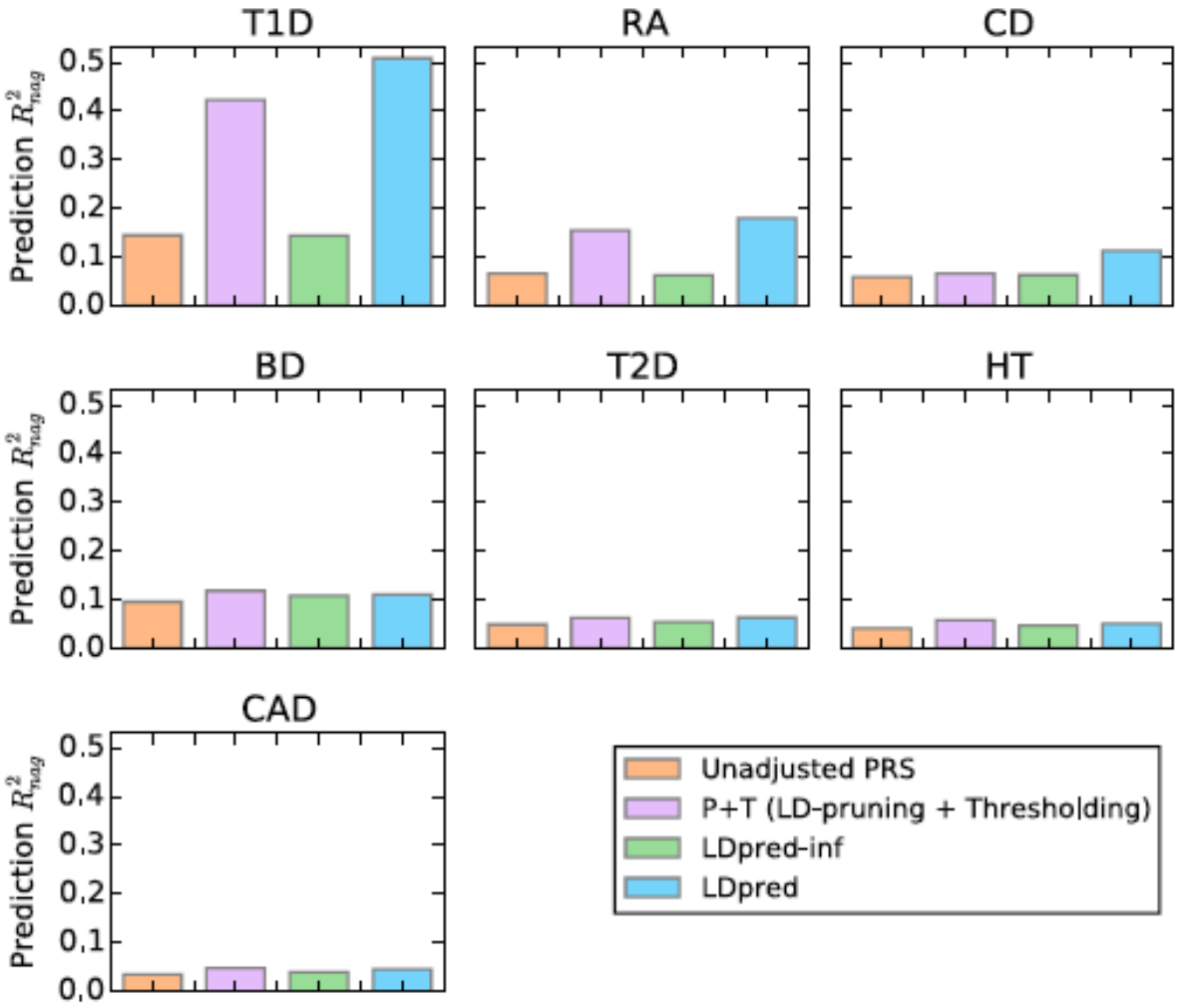


Figure 2. Comparison of Four Prediction Methods Applied to Simulated Traits
Prediction accuracy of the four different methods listed in Table S1 when applied to simulated traits with WTCCC genotypes. The four subfigures correspond to $p = 1$ (A), $p = 0.1$ (B), $p = 0.01$ (C), and $p = 0.001$ (D) for the fraction of simulated causal markers with (non-zero) effect sizes sampled from a Gaussian distribution. To aid interpretation of the results, we plot the accuracy against the effective sample size, defined as $N_{eff} = (N/M_{sim})M$, where $N = 10{,}786$ is the training sample size, $M = 376{,}901$ is the total number of SNPs, and M_{sim} is the actual number of SNPs used in each simulation: 376,901 (all chromosomes), 112,185 (chromosomes 1–4), 61,689 (chromosomes 1 and 2), and 30,004 (chromosome 1). The effective sample size is the sample size that maintains the same N/M ratio if all SNPs are used.

LDpred

Application to real data
Vilhjalmsson et al., 2015

BLUP marginally better than
Pruning + Thresholding



PRSice: Euesden et al., 2014

Multiple testing due to the high resolution in p-value threshold.

Authors suggest p<0.001 if using the best fit PRS.

Significance threshold dependent on LD in the target sample and distribution of the phenotype predicted.

Unclear if it holds for phenotypes with skewed distributions and for non UK samples.

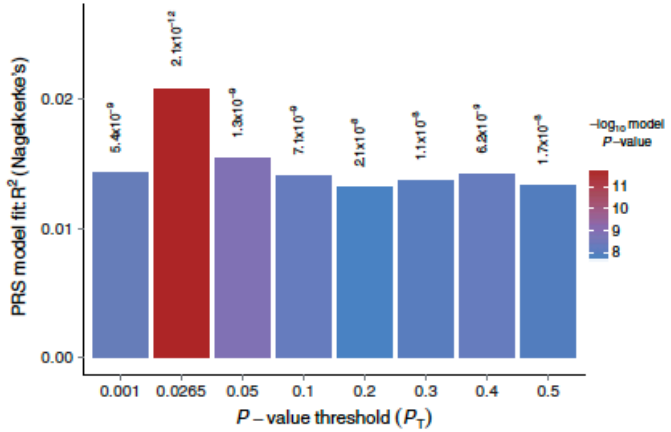


Fig. 1. Bar plot from PRSice showing results at broad *P*-value thresholds for Schizophrenia PRS predicting MDD status. A bar for the best-fit PRS from the high-resolution run is also included

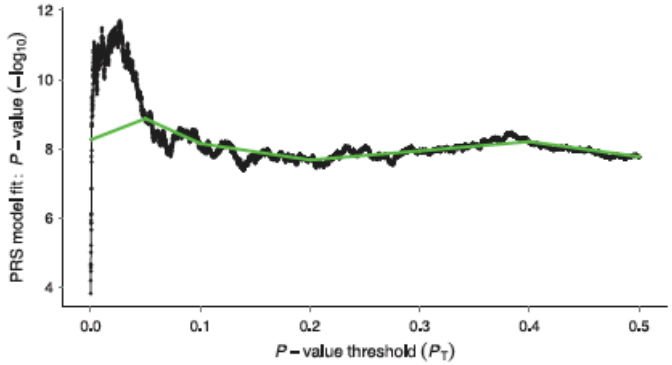


Fig. 2. High-resolution PRSice plot for SCZ predicting MDD status. The thick line connects points at the broad *P*-value thresholds of Fig.1

Classic / OLS	BLUP	PRSice
Dosage or best guess	Best guess	Dosage or best guess
clumping	BLUP effects summed over all SNPs	clumping
Multiple PRS by p-value thresholds	Unique PRS	All p-value threshold tested
Bonferroni correction		Unclear significance threshold for association
	Hypothesis: effect sizes of SNPs normally distributed	
Fast (can be parallelized)	Matrix inversion, can be long for large N	Slower and harder to parallelize (R package)
PLINK	GCTA, PLINK	R (PLINK)

Going further

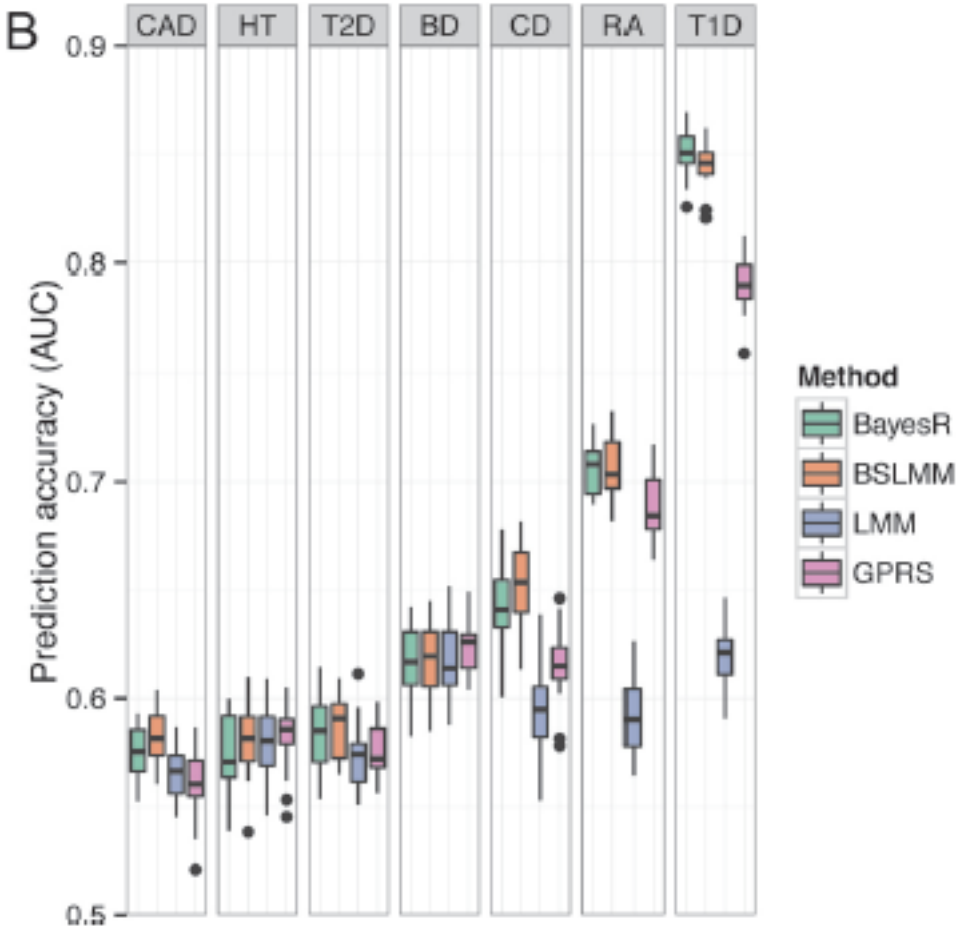
BayesR: Moser et al., 2015: bayesian mixture model that simultaneously

- variants discovery
- estimation of genetic variance (heritability)
- prediction in new sample

“All in one” tool but computationally demanding and limited to ~1M SNPs

Marginally better than LMM-BLUP or standard PRS (GPRS)

Require individual GWAS data

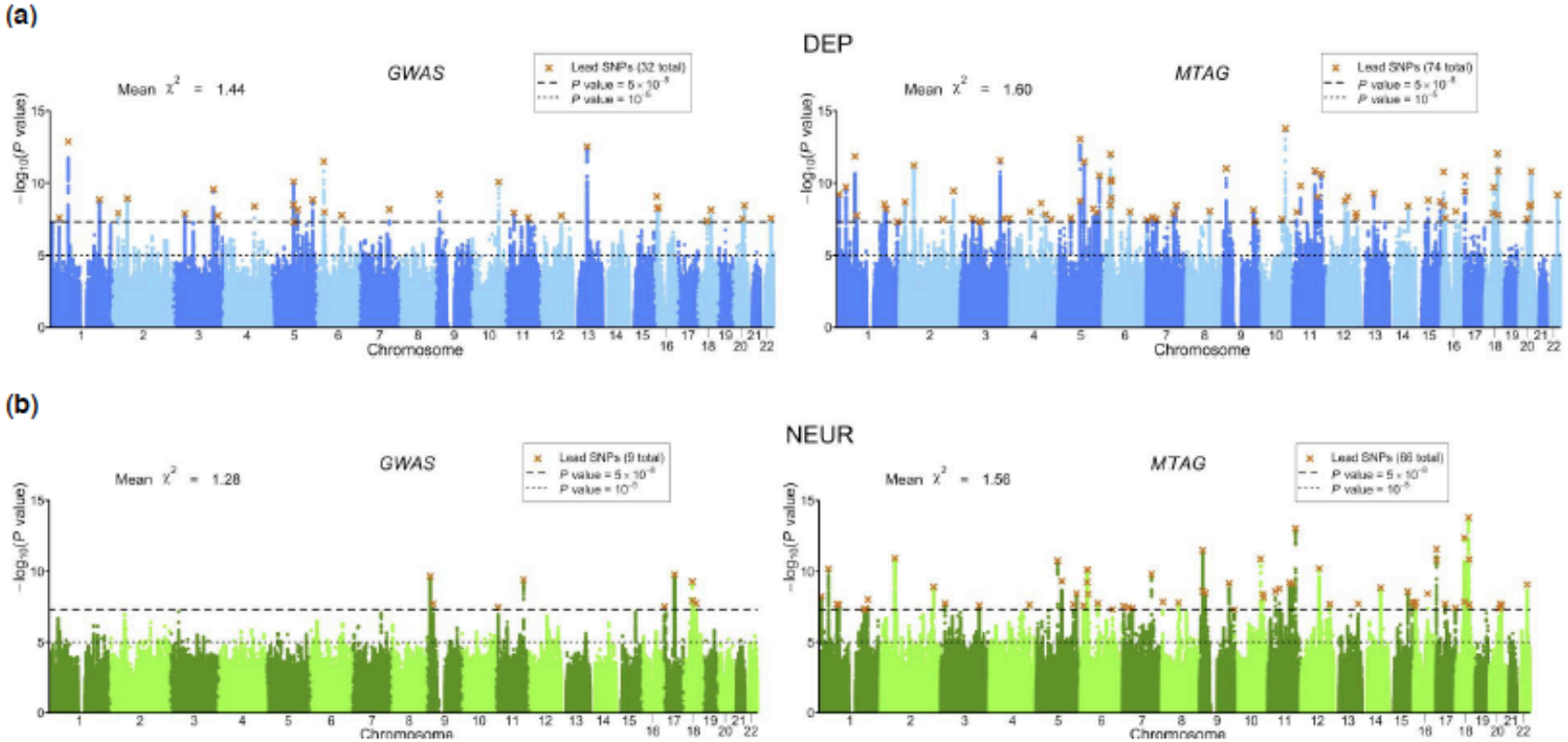


Multi-trait PRS

QIMR Berghofer
Medical Research Institute

MTAG: Multi-trait analysis of GWAS

Turley et al., 2017, BiorXiv



MTAG: Multi-trait analysis of GWAS

Turley et al., 2017, BiorXiv

Meta-analyse GWAS of genetically correlated traits that are not necessarily the same

Uses combined effect sizes to improve prediction of PRS

$$\hat{\beta}_{\mathrm{MTAG},j,t} = \frac{\frac{\boldsymbol{\omega}_t'}{\omega_{tt}}\left(\boldsymbol{\Omega} - \frac{\boldsymbol{\omega}_t\boldsymbol{\omega}_t'}{\omega_{tt}} + \boldsymbol{\Sigma}_j\right)^{-1}}{\frac{\boldsymbol{\omega}_t'}{\omega_{tt}}\left(\boldsymbol{\Omega} - \frac{\boldsymbol{\omega}_t\boldsymbol{\omega}_t'}{\omega_{tt}} + \boldsymbol{\Sigma}_j\right)^{-1}\frac{\boldsymbol{\omega}_t}{\omega_{tt}}}\hat{\boldsymbol{\beta}}_j,$$

SNP j

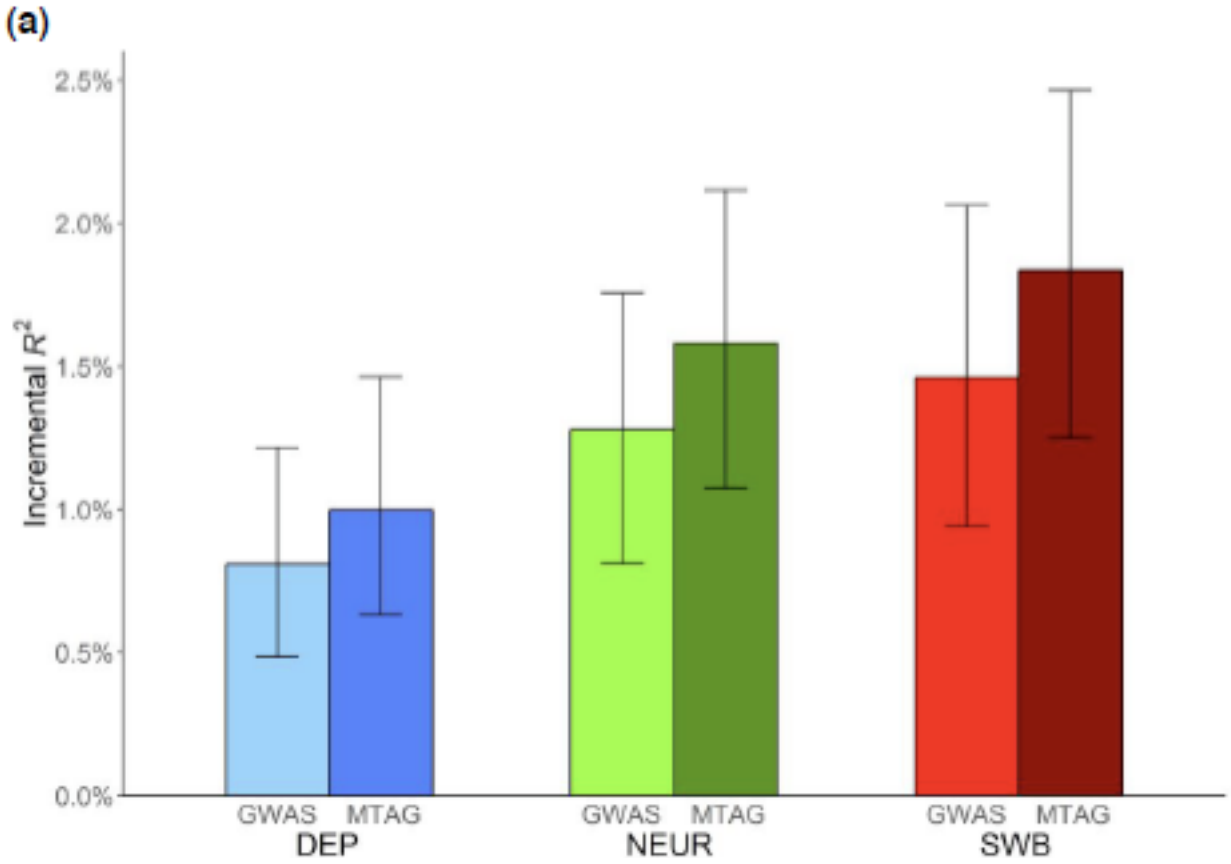
Trait t

Σj: TxT variance-covariance of estimation error: diagnonal are intercept of LDSCR, off diagonals are intercepts from bivariate LDSCR

Ω_{TxT}: genetic variance-covariance matrix of effect sizes across traits Estimated from data

ω: t th column of diagonal element of Ω

βj: GWAS effect sizes



DEP: MDD
Neff=354,862

NEUR: Neuroticism
N=168,105

SWB: subjective well being
N=388,538

Assumed overlap between MDD and SWB samples

Target: Health and Retirement Study (HRS), N>6,000

Thank you

www.qimrberghofer.edu.au

QIMR Berghofer
Medical Research Institute