

Polygenic risk scores

Sarah Medland and Lucía Colodro Conde

sarah/2020/thursday

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 traditional approach to calculating PRS is to **construct a weighted sum of the betas** (or other effect size measure) for a set of **independent loci thresholded at different significance levels**.
 - Typically the independence is LD based ($LD\ r^2 \leq .2$) via clumping.

The classics

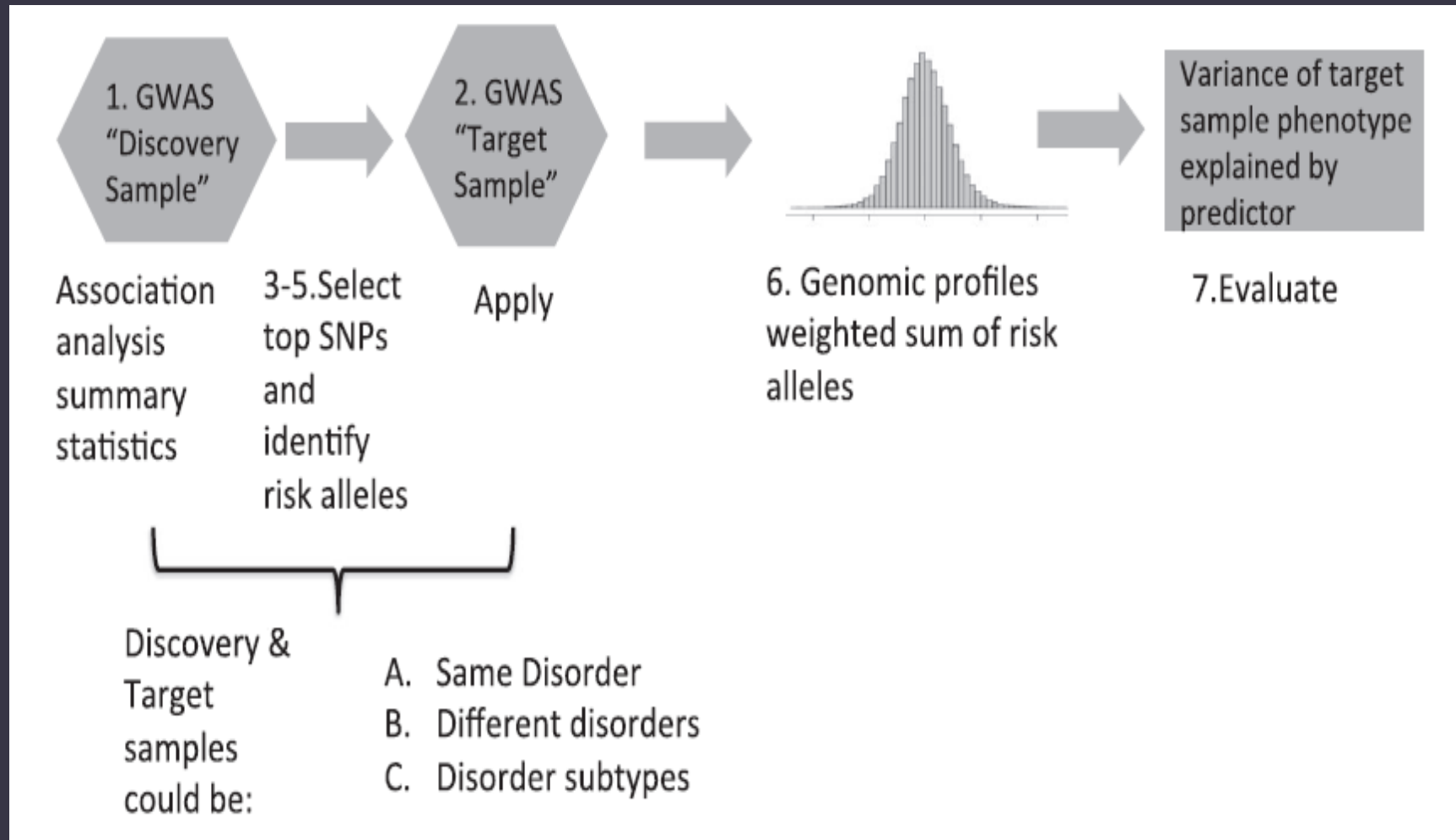
- Wray NR, Goddard, ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007; 7(10):1520-28.
- Evans DM, Visscher PM., Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. Human Molecular Genetics. 2009; 18(18): 3525-3531.
- 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, Brion MJ, Paternoster L, Kemp JP, McMahon G, Munafò M, Whitfield JB, Medland SE, Montgomery GW; GIANT Consortium; CRP Consortium; TAG Consortium, Timpson NJ, St Pourcain B, Lawlor DA, Martin NG, Dehghan A, Hirschhorn J, Smith GD. Mining the human phenome using allelic scores that index biological intermediates. PLoS Genet. 2013,9(10):e1003919.



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, Zernakova 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



<https://sites.google.com/broadinstitute.org/ukbbgwasresults/>

The banner features a blue-tinted background with a DNA double helix structure. On the left, there is a chemical structure of Thymine with the label 'Thymine' below it. On the right, there is a chemical structure of Guanine with the label 'Guanine' above it. The text 'UK Biobank GWAS Results' is prominently displayed in the center in a large, white, sans-serif font. In the top left corner, there is a logo for the Stanley Center for Genomic Medicine at the Broad Institute, with the text 'UK Biobank GWAS Results' next to it.

STANLEY CENTER
FOR GENOMIC MEDICINE
AT BROAD INSTITUTE

UK Biobank GWAS Results

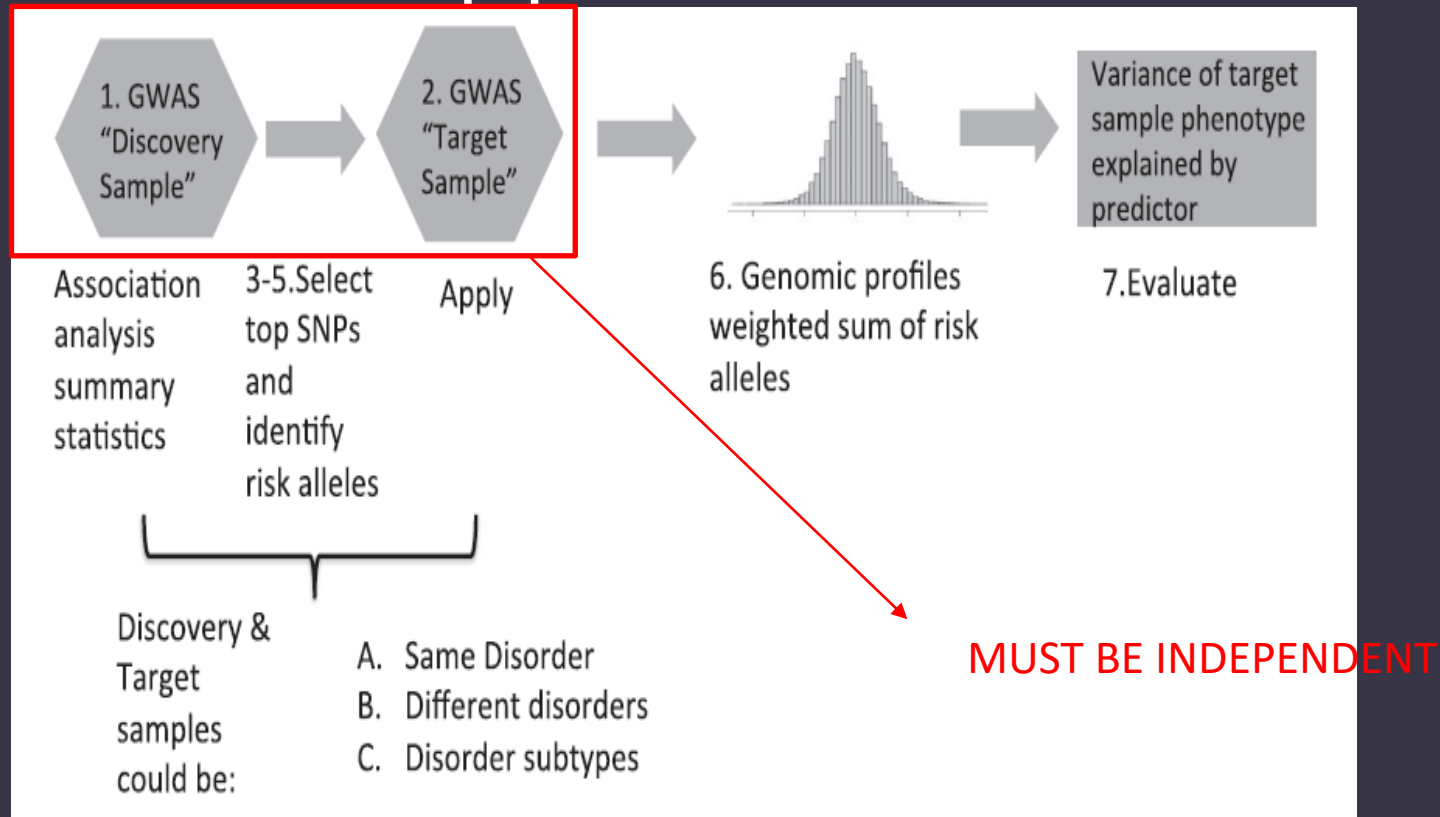
Thymine

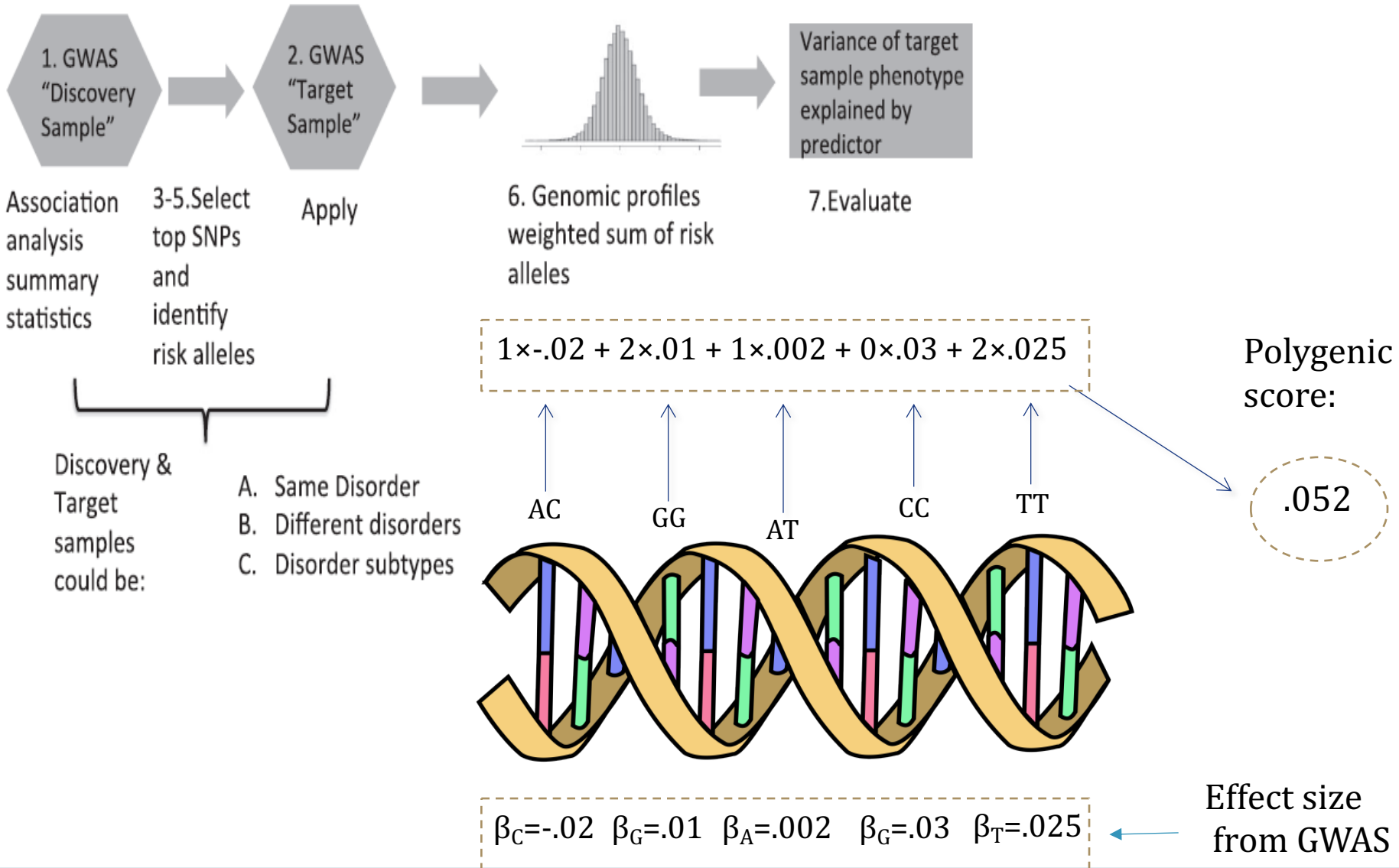
Guanine

UK Biobank GWAS Results

This site contains the results of the GWAS and heritability analyses conducted by the [Neale Lab](#). Please refer to the description of these analyses [here](#) for details.

Traditional approach





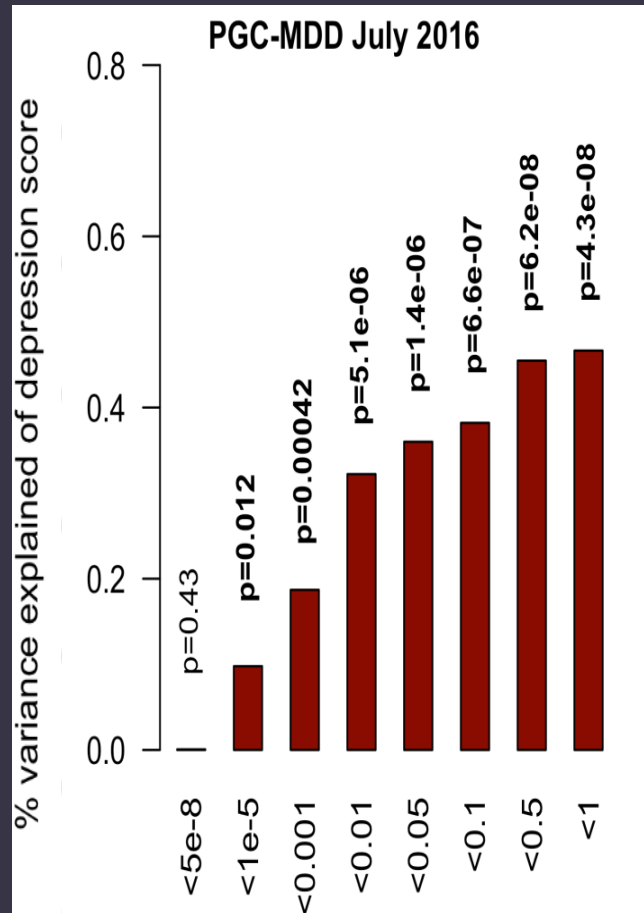
Main uses of PRS

1) Single disorder analyses

2) Cross-disorder analysis

3) Sub-type analysis

Single trait analyses



OPEN

Molecular Psychiatry (2017) 00, 1–7

www.nature.com/mp

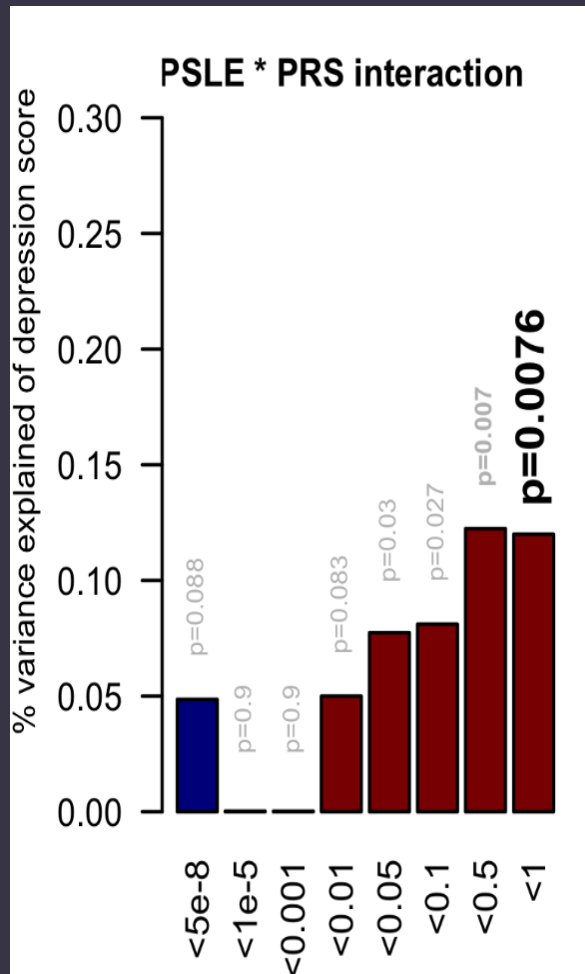
ORIGINAL ARTICLE

A direct test of the diathesis–stress model for depression

L Colodro-Conde^{1,2,12}, B Couvy-Duchesne^{1,3,12}, G Zhu¹, WL Coventry⁴, EM Byrne⁵, S Gordon¹, MJ Wright^{3,6}, GW Montgomery⁵, PAF Madden⁷, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium¹³, S Ripke^{8,9,10}, LJ Eaves¹¹, AC Heath⁷, NR Wray^{3,5}, SE Medland¹ and NG Martin¹

The diathesis–stress theory for depression states that the effects of stress on the depression risk are dependent on the diathesis or vulnerability, implying multiplicative interactive effects on the liability scale. We used polygenic risk scores for major depressive disorder (MDD) calculated from the results of the most recent analysis from the Psychiatric Genomics Consortium as a direct measure of the vulnerability for depression in a sample of 5221 individuals from 3083 families. In the same we also had measures of

Moderated single trait analyses



OPEN

Molecular Psychiatry (2017) 00, 1–7

www.nature.com/mp

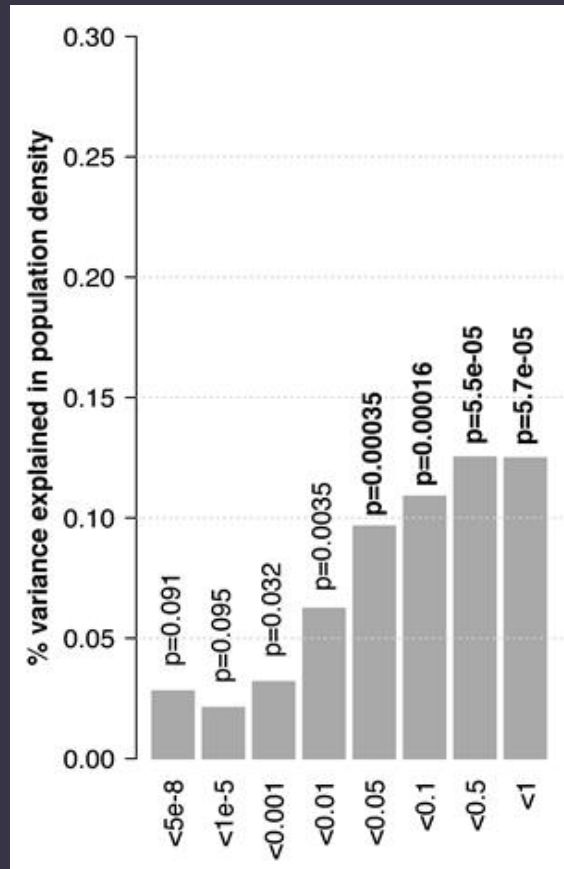
ORIGINAL ARTICLE

A direct test of the diathesis–stress model for depression

L Colodro-Conde^{1,2,12}, B Couvy-Duchesne^{1,3,12}, G Zhu¹, WL Coventry⁴, EM Byrne⁵, S Gordon¹, MJ Wright^{3,6}, GW Montgomery⁵, PAF Madden⁷, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium¹³, S Ripke^{8,9,10}, LJ Eaves¹¹, AC Heath⁷, NR Wray^{3,5}, SE Medland¹ and NG Martin¹

The diathesis–stress theory for depression states that the effects of stress on the depression risk are dependent on the diathesis or vulnerability, implying multiplicative interactive effects on the liability scale. We used polygenic risk scores for major depressive disorder (MDD) calculated from the results of the most recent analysis from the Psychiatric Genomics Consortium as a direct measure of the vulnerability for depression in a sample of 5221 individuals from 3083 families. In the same we also had measures of

Cross-trait analysis



PRS-SCZ

This Issue Views **4,143** | Citations **2** | Altmetric **119**

Original Investigation

September 2018

Association Between Population Density and Genetic Risk for Schizophrenia

Lucía Colodro-Conde, PhD¹; Baptiste Couvy-Duchesne, PhD^{1,2,3}; John B. Whitfield, PhD¹; [et al](#)

[» Author Affiliations](#)

JAMA Psychiatry. 2018;75(9):901-910. doi:10.1001/jamapsychiatry.2018.1581

Sub-type analysis

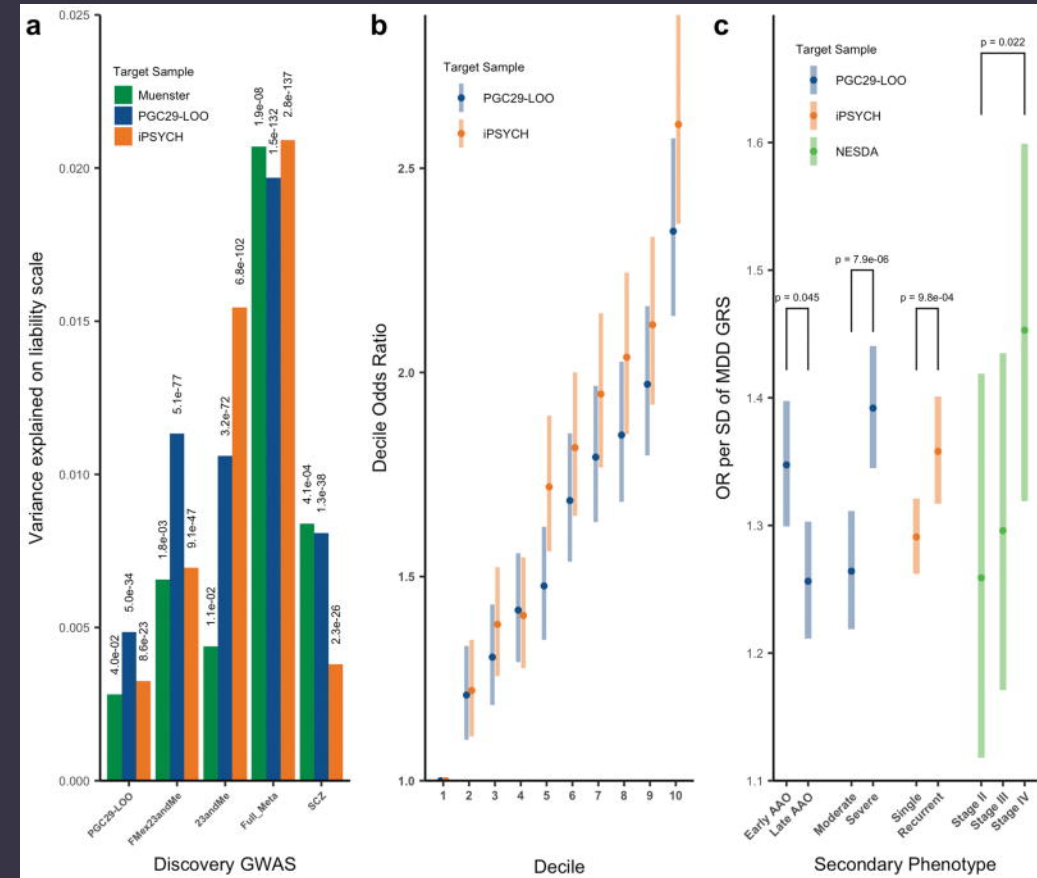
nature
genetics

Article | Published: 26 April 2018

Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression

Naomi R. Wray , Stephan Ripke, [...] the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium

Nature Genetics **50**, 668–681 (2018) | [Download Citation](#) 



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 regression power calculator (for r^2 of up to 0.5%).

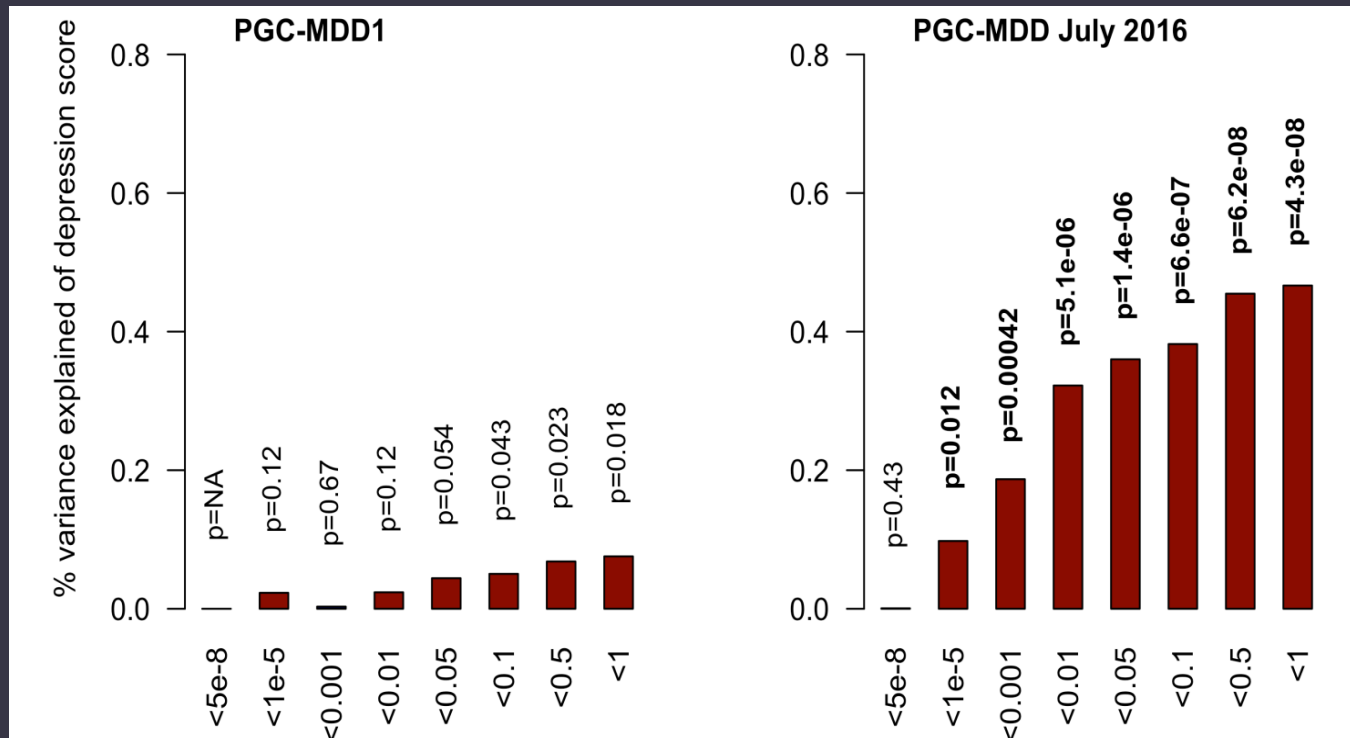
As a general rule of thumb you usually want 2,000+ people in the target dataset.

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

Power calculator for discovery (GWAS) sample needed to achieve prediction of r^2 in target sample

```
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)
```

Power of PRS analysis increases with GWAS sample size



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,
Couvry-Duchesne B, et al, (2017)
Molecular Psychiatry

Making a PRS

(1) GWAS summary statistics

→ From PGC results, other public domain GWAS, unpublished GWAS

SNP identifier (rs number, Chr:BP)

Both Alleles (effect/reference, A1/A2)

Effect

- Beta from association with continuous trait
- OR from an ordinal trait - convert to $\log(\text{OR})$
- Z-score, MAF and N (from an N weighted meta-analysis)

p-value

(frequency of A1)

(1) GWAS summary statistics

→ From PGC results, other public domain GWAS, unpublished GWAS

SNP identifier (rs number, Chr:BP)

Both Alleles (effect/reference, A1/A2)

Effect

- Beta from association with continuous trait
- OR from an ordinal trait - convert to $\log(\text{OR})$
- Z-score, MAF and N (from an N weighted meta-analysis)

p-value

(frequency of A1)

Make sure that your target genotypes are named the same way as your discovery data!

→ imputation reference and genomic build

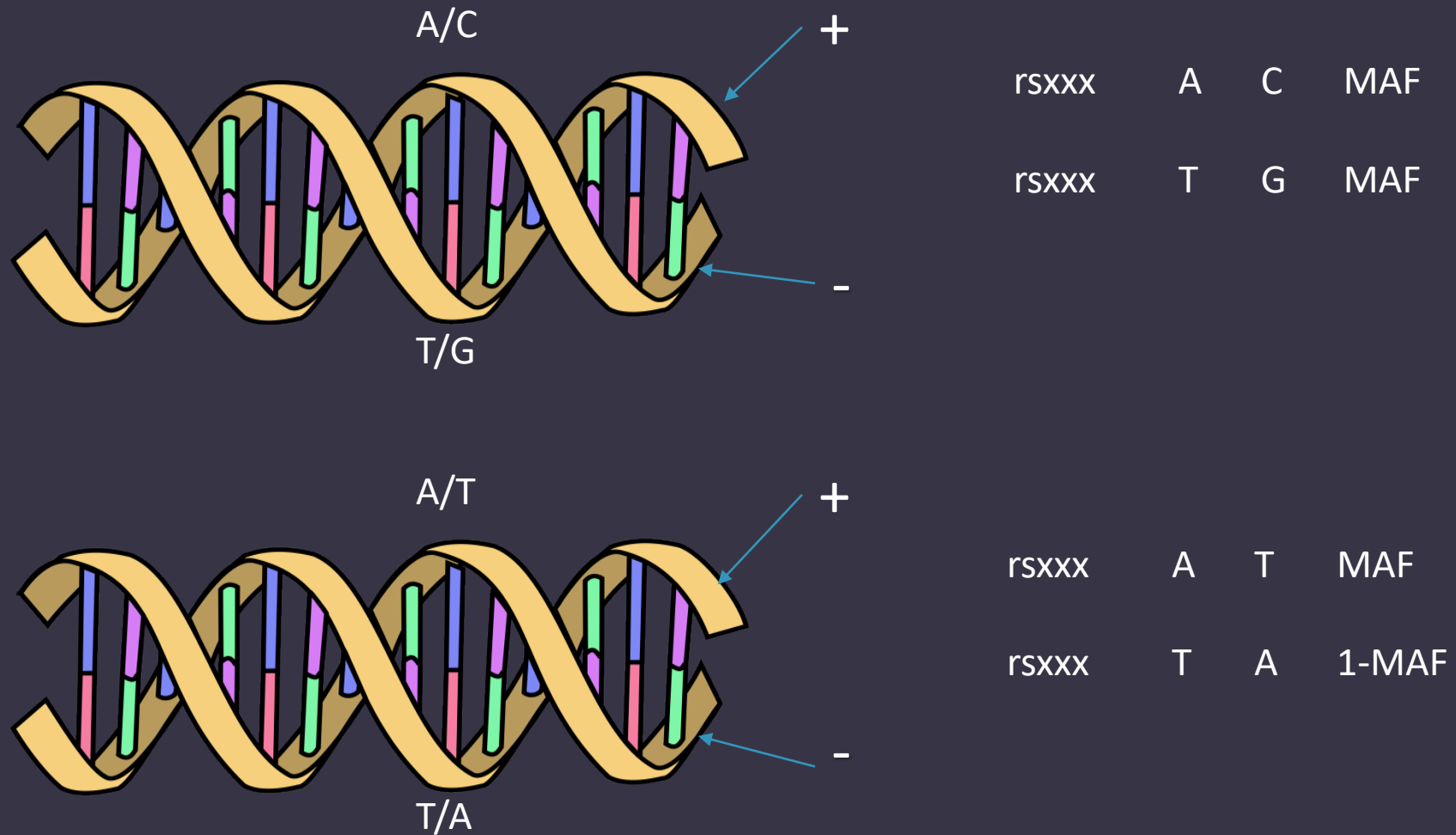
(2) Find SNPs in common with your local sample and QC

- Imputed data
- QC
 - $R^2 \geq 0.6$
 - $MAF \geq 0.01$
 - No indels
 - No ambiguous strands (*) - A/T or T/A or G/C or C/G

```
for ((i=1;i<=22;i++))  
do  
awk '{ if ($5<=.01 & $5<=.99 & $6>=.6) print $1}' file"$i".info >> available.snps  
done
```

(*) On ambiguous strands

GWAS chip results are expressed relative to the + or - strand of the genome reference



(3) Clumping

- Select most associated SNP per LD region (pruning)
- Plink1.9
 - bfile ReferencePanelForLD
 - extract QCedListofSNPs
 - clump gwasFileWithPvalue
 - clump-p1 (*#Significance threshold for index SNPs*)
 - clump-p2 (*#Secondary significance threshold for clumped SNPs*)
 - clump-r2 (*#LD threshold for clumping*)
 - clump-kb (*#Physical distance threshold for clumping*)
 - out OutputName

(4) Calculate risk scores

The traitX"\$i".selected files will contain the lists of top independent snps. Merge the alleles, effect & P values from the discovery data onto these files.

To do a final strand check merge the alleles of the target set onto these files. If any SNPs are flagged as mismatched you will have to manual update the merged file - flip the strands (ie an A/G snp would become a T/C snp) but leave the effect as is.

Create Score files (SNP EffectAllele Effect) and P files contain (SNP Pvalue).

```
for ((i=1;i<=22;i++))
do
awk '{ if ($6==$8 || $6==$9 ) print $0, "match" ; if ($6!=$8 && $6!=$9 ) print $0, "mismatch"}'
traitX."$j".merged > strandcheck.traitX."$i"
grep mismatch strandcheck.traitX*
done
```

(4) Calculate risk scores

```
for ((i=1;i<=22;i++))
do
plink --noweb --dosage Your_chr"$i".plink.dosage.gz format=1 Z --fam
Your_chr"$i".plink.fam --score traitX."$i".score --q-score-file traitX."$i".P --q-score-
range p.ranges --out Your_chr"$i".PRS
done
```

p.ranges

S1 0.00 0.000001

S2 0.00 0.01

S3 0.00 0.10

S4 0.00 0.50

S5 0.00 1.00

(5) Run PRS analysis –unrelated individuals

```
base <- lm (ICV ~ age + sex + PC1 + PC2 +PC3 +PC4 + other-covariates, data =mydata)
score1 <- lm (ICV ~ S1 + age + sex + PC1 + PC2 +PC3 +PC4 + other-covariates, data =mydata)
score2 <- lm (ICV ~ S2 + age + sex + PC1 + PC2 +PC3 +PC4 + other-covariates, data =mydata)
model_base <- summary(base)
model_score1 <- summary(score1)
model_score2 <- summary(score2)
model_base$r.squared
model_score1$r.squared
model_score2$r.squared
anova(base,score1)
anova(base,score2)
```

(5) Run PRS analysis, controlling for relatedness – twin pairs or small families

- You can add the PRS as a covariate on the means model in an open Mx script
- Allows you to do multivariate PRS analyses
- Or look at variance explained over time in longitudinal data
 - Test if the betas are equal across time points

(5) Run PRS analysis, controlling for relatedness in large/complex cohorts

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

Could run this analysis in a multilevel OpenMx model

Other Methods

Classic / Clump and Threshold	BLUP (LDpred)	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 thresholds 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)

Overlap and Overfitting

Q: How important is independence with Biobank size samples?

- Perceptions that this may not matter with biobank type discovery samples when the overlap is very small
- Impact of relatedness across the discovery and target samples is usually ignored

Q: How important is independence with Biobank size samples?

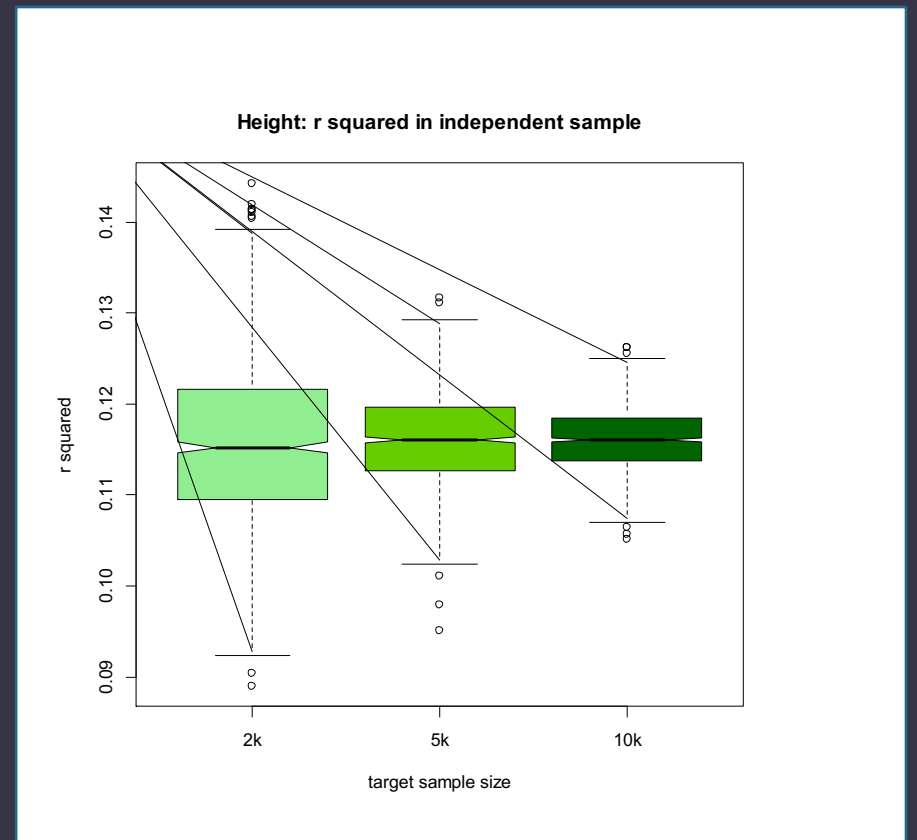
- To examine this
 - GWAS were conducted for a continuous (height)
 - ~340,000 individuals were extracted from the UK Biobank (app. 25331)
 - European Ancestry & Unrelated (less than 3rd degree relatedness)
 - Age, Sex and 10 PCs included as covariates
 - A set of 35,000 individuals held out to ensure independence of the target sample

Q: How important is independence with Biobank size samples?

- Discovery GWAS were clumped and PRS were calculated
- PRS analyses were conducted using target samples
 - of 2,000, 5,000 or 10,000 individuals randomly drawn from the hold-out sample (of 35,000)
 - 1,000 replicates
 - 4 PRS thresholds:
 - 0.00 0.0001
 - Age, Sex and 10 PCs included as covariates
- To examine overfitting the target samples were spiked with
 - 5, 10, 50, 100 or 200 overlapping individuals
 - 5, 10, 50, 100 or 200 1st degree relatives

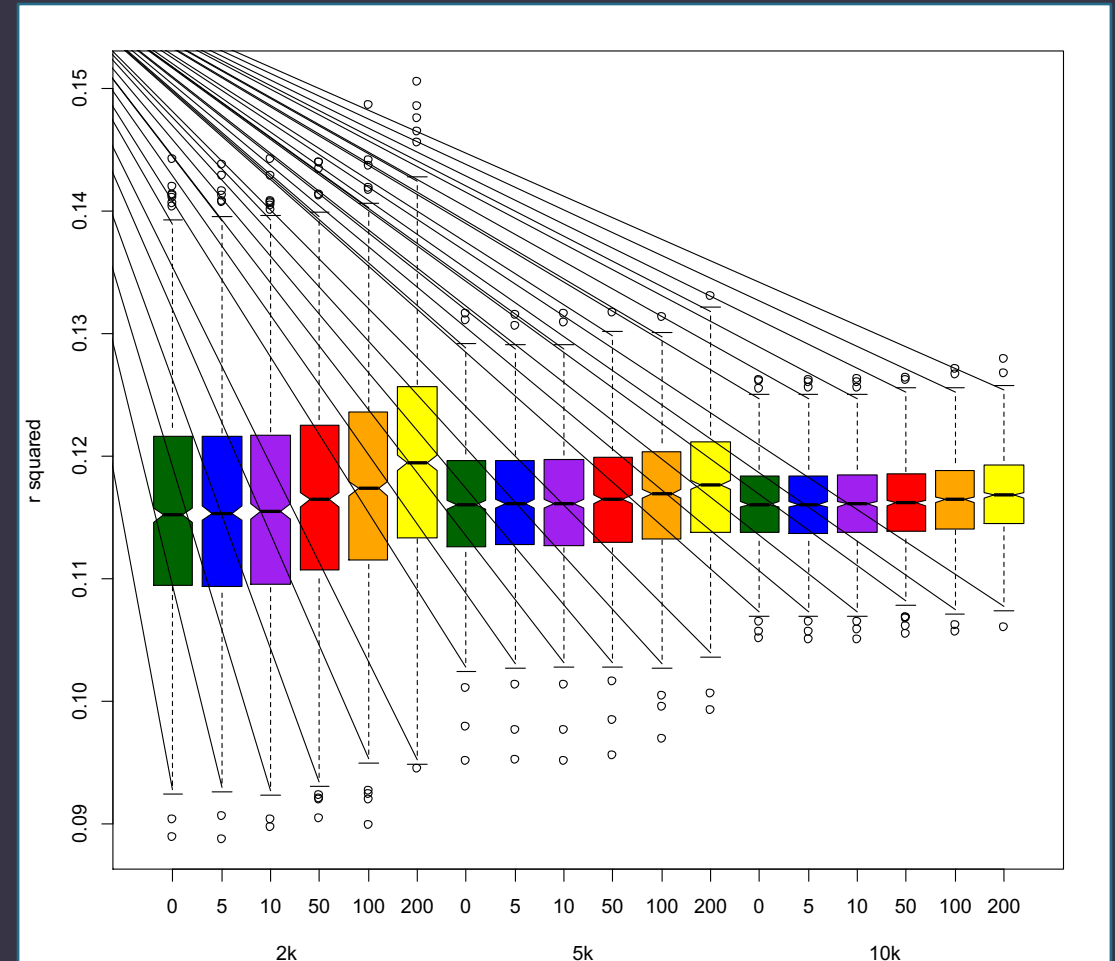
A: Variance explained

- PRS analyses in independent samples explained a median of 11.6% of variance



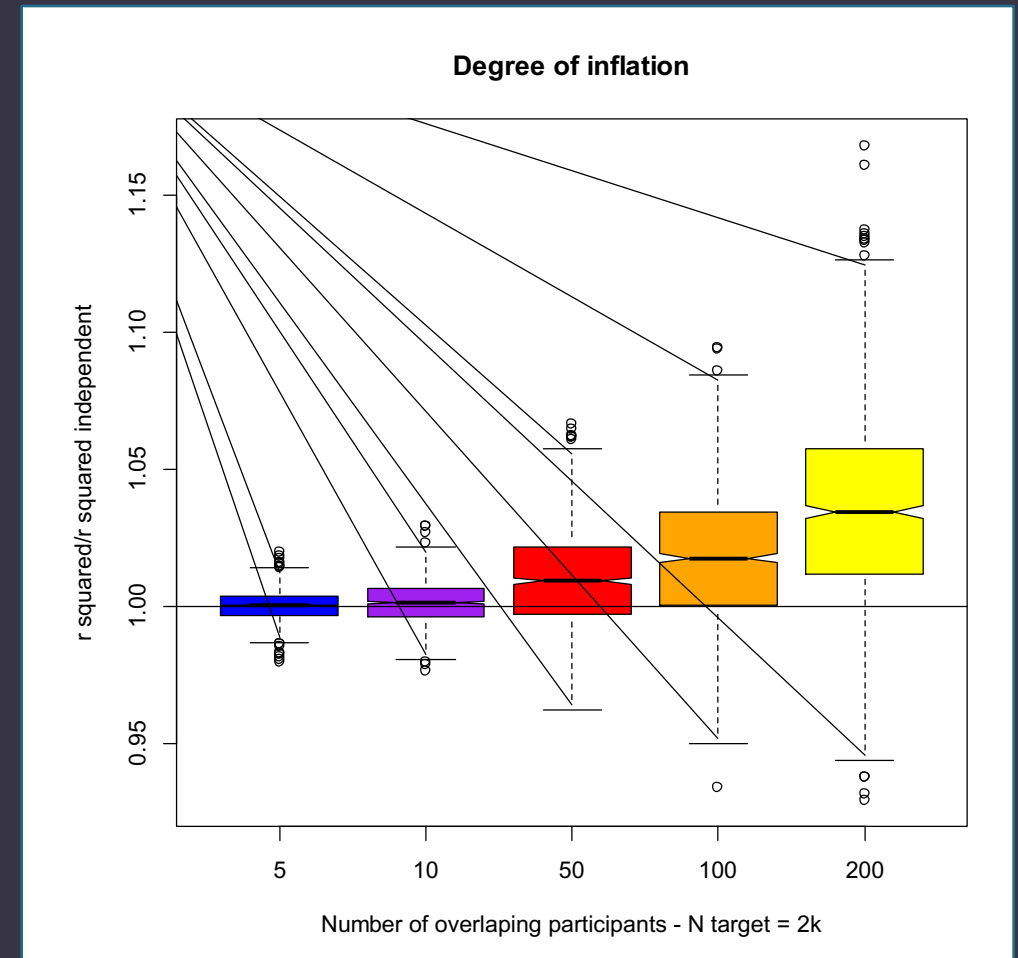
A: Impact of non-independent samples

- Yes – as expected there is bias in the estimate of variance explained and the p values
- Pattern of results the same across all N_s



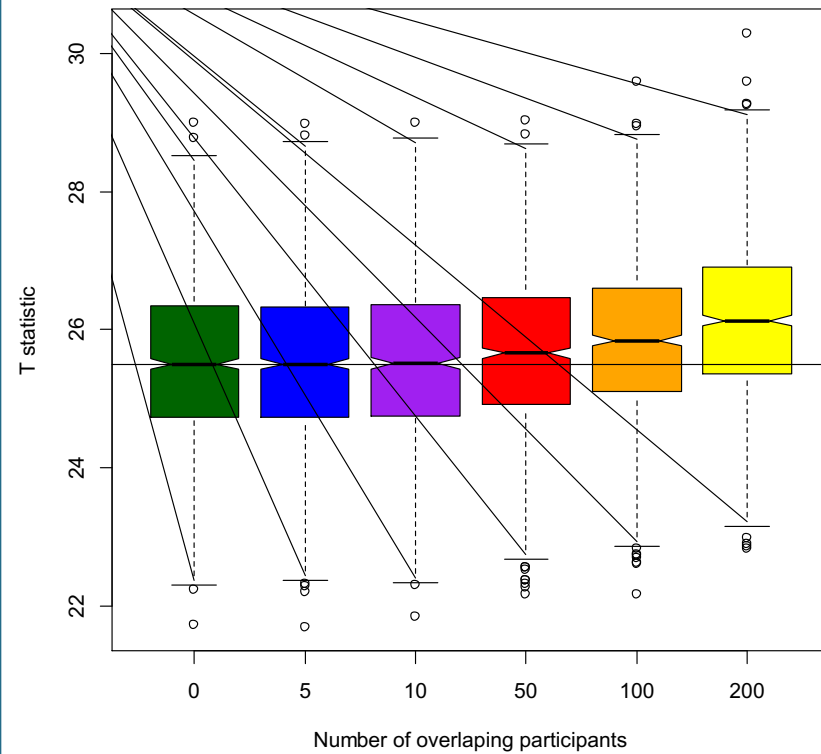
A: Impact of non-independent samples

- Inflation present
 - Extent is a function of the % overlap in the target sample
 - Confirms the cautions of Wray et al 2013 apply to biobank sized discovery samples
 - With 5 overlapping people in a target sample of 10k there was significant inflation
 - Median CIs did not include 1

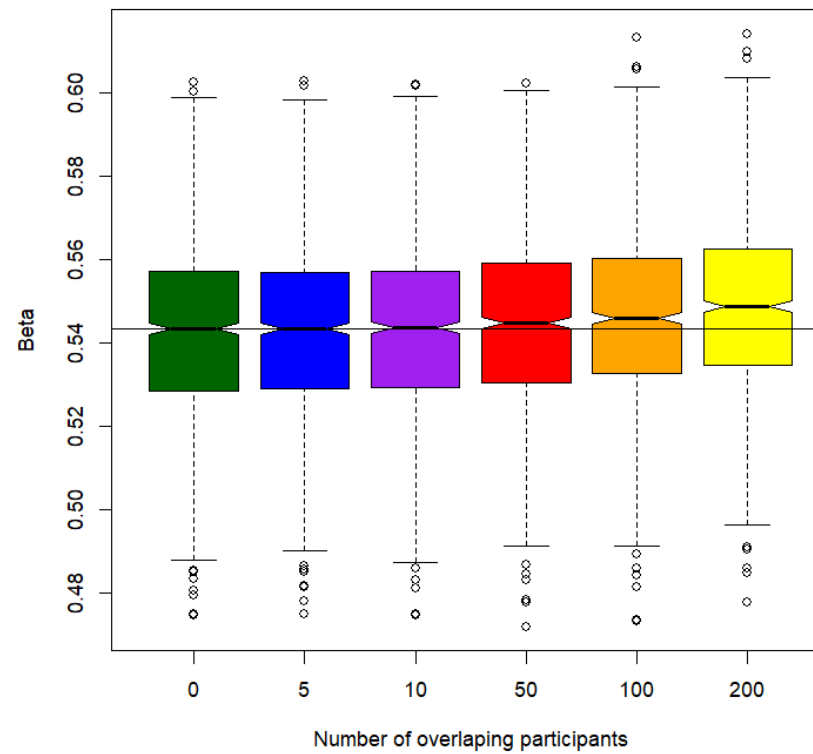


A: Impact of non-independent samples

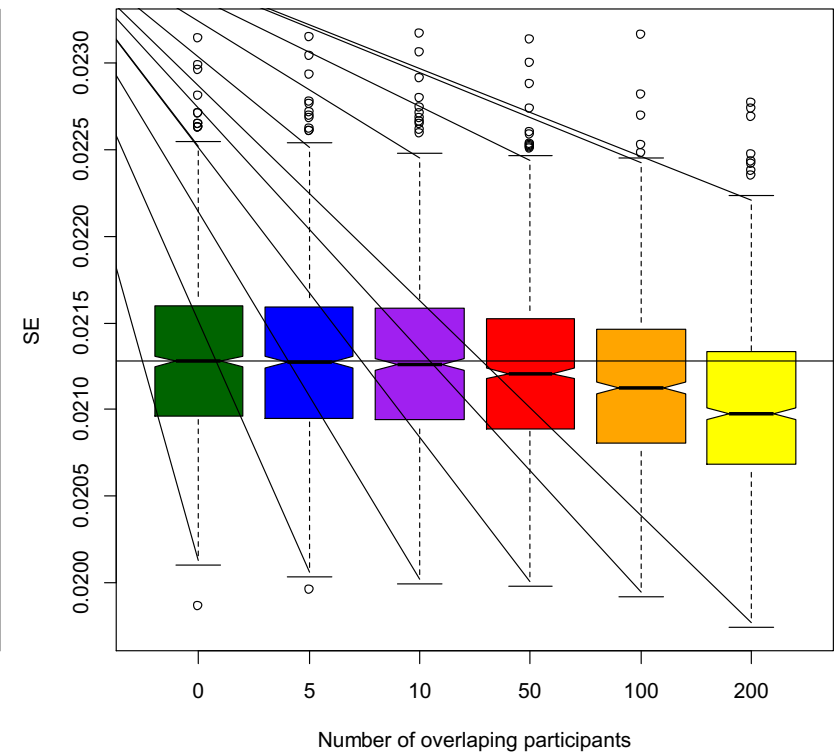
T statistic, total N=2k



Beta, total N=2k



SE, total N=2k

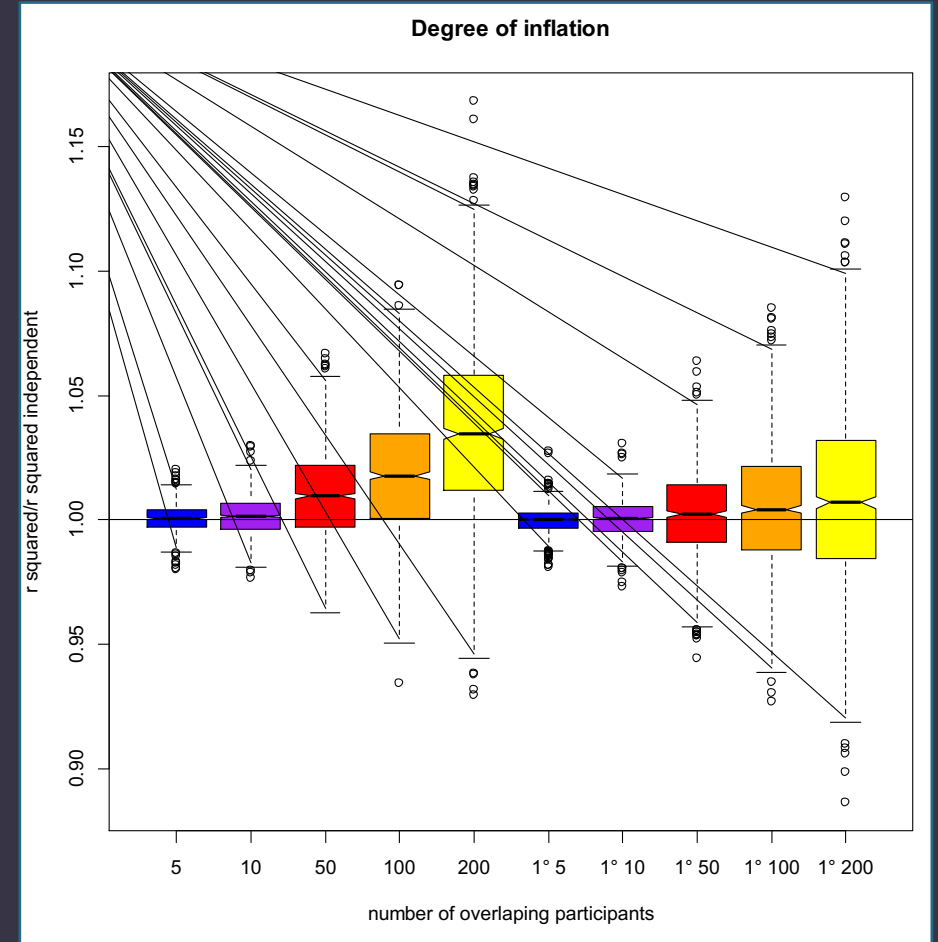


A: Impact of non-independent samples

- Inflation also present
 - In binary phenotypes
 - Even if the overlap is limited to only controls or only cases
- Expect that inflation will be worse for quantitative traits if overlap is restricted to the tails of the distribution
 - (Not tested)

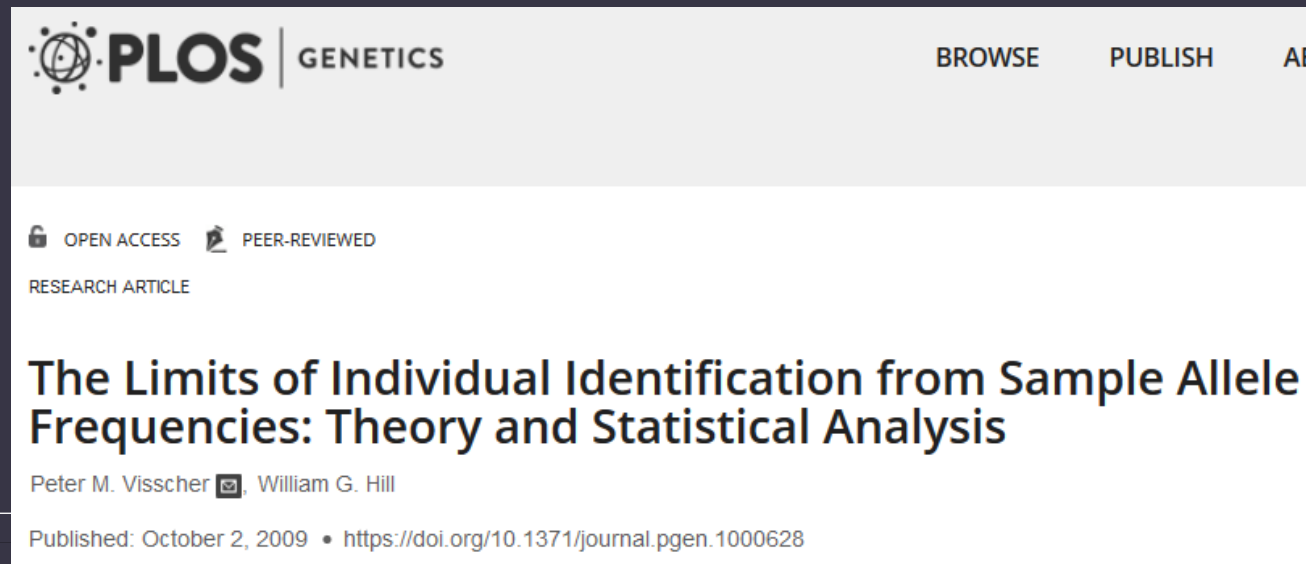
A: Impact of First Degree Relatives

- Inflation present
 - Proportional to the h^2 and the extent of overlap in the target sample (% of N)



Q: How to Identify non-independence?

- Homer et al method
 - Visscher and Hill 2009 more powerful
 - However, many cohorts do not provide true MAF, violates data access, not clear how well this really works with a realistic meta-analysis



The screenshot shows the top portion of a PLOS Genetics research article page. At the top left is the PLOS logo, followed by the journal name "PLOS GENETICS". To the right are navigation links for "BROWSE", "PUBLISH", and "ABOUT". Below the header, there are two icons: an open lock for "OPEN ACCESS" and a checkmark for "PEER-REVIEWED". Underneath these is the text "RESEARCH ARTICLE". The main title of the article is "The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis". Below the title, the authors are listed as "Peter M. Visscher" and "William G. Hill". At the bottom of the page, the publication date is "October 2, 2009" and the DOI link is "https://doi.org/10.1371/journal.pgen.1000628".

PLOS | GENETICS

BROWSE PUBLISH ABOUT

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

The Limits of Individual Identification from Sample Allele Frequencies: Theory and Statistical Analysis

Peter M. Visscher, William G. Hill

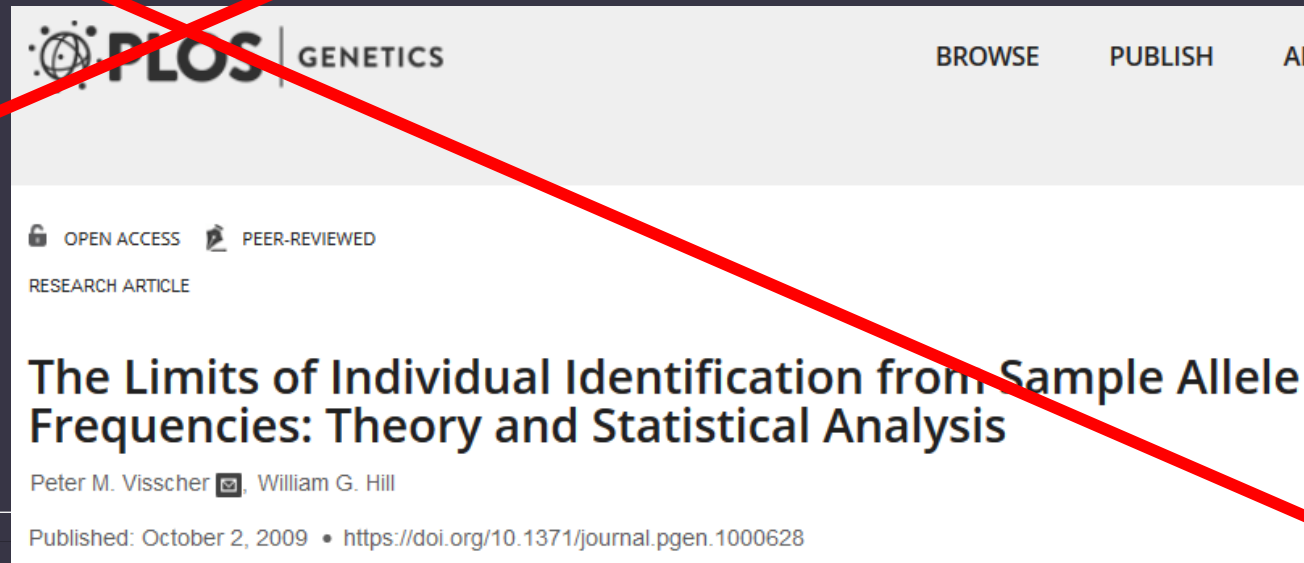
Published: October 2, 2009 • <https://doi.org/10.1371/journal.pgen.1000628>

Q: How to Identify non-independence?

- LD Score – (Maybe, more work needed...)
 - Using the Height data from the PRS analyses ran GWAS for 20 permutations
 - Sample 1 340,000 individuals
 - Sample 2 30,000 individuals
 - Overlap of 200 individuals
 - Covariance “Intercept” ranged from .067 (.017) to .075 (.017) indicating non-independence
 - Overlap of 5 individuals
 - Covariance “Intercept” ranged from .062 (.016) to .072 (.017) indicating non-independence

WHAT Are the Solutions if you find non-independence

- Homer et al method
 - Visscher and Hill 2009 more powerful
 - However, many cohorts do not provide true MAF, violates data access, not clear how well this really works with a realistic meta-analysis



WHAT Are the Solutions if you find non-independence

SCZ

The PGC has made the full results from all published PGC studies available for download. If you download these data, you and your immediate collaborators (“investigators”) acknowledge and agree to all of the following conditions:

7. Investigators will never attempt to identify any participant.

ata access, not
is

DOWNLOAD
SCZ2 README

TERMS AND CONDITIONS

Check here to indicate that you (and your immediate collaborators [“investigators”]) have read and agree to all of the terms and conditions of the PGC listed above. The Password is:

I Agree

scz2.snp.results.txt.gz (header row and 9,444,230 SNPs)

hg19chrc	hg19 chromosome as character string (chr1-chr22, chrX)
snpid	rs ID of SNP
a1	reference allele for OR (may not be minor allele)
a2	alternate allele
bp	hg19 base pair position of SNP
info	imputation quality score
or	odds ratio in PGC GWAS data
se	standard error of ln(OR) in PGC GWAS data
p	p-value in PGC GWAS data
ngt	number of samples in which SNP directly genotyped

WHAT Are the Solutions if you find non-independence

- Leave-one-out...
- If both groups have raw data access collaborate & exchange checksums
 - Make list of common non-ambiguous SNPs passing QC in discovery and target
 - Make n SNP set lists each with m SNPs
 - Export hardcall data from each SNP set (1 line per person but no IDs)
 - Parse the data obtaining a checksum for each line of data
 - Exchange and look at % of identical checksums




Google: checksum ripke


https://personal.broadinstitute.org/sripke/share_links/checksums_download/

WHAT Are the Solutions if you find non-independence

- Mak et al (2018) proposed using all available data in the discovery and use of cross-prediction with split-validation to reduce inflation
 - Focus is on situations where you have raw data for both discovery and target
 - They do not consider the more typical where you have discovery and raw target data






Cold Spring Harbor Laboratory



bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

New Results

Polygenic scores for UK Biobank scale data

Timothy Shin Heng Mak,  Robert Milan Porsch,  Shing Wan Choi,  Pak Chung Sham
doi: <https://doi.org/10.1101/252270>

WHAT Are the Solutions if you find non-independence

- Do you really need prediction
 - Are you trying to show polygenicity?
 - If not can you answer your question with LDSC, GWAS-SEM, MR, SECA or another approach?

Questions?

