

# Polygenic risk scores

---



Sarah Medland

with thanks to Lucia Colodro Conde &  
Baptiste Couvy Duchesne

# 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

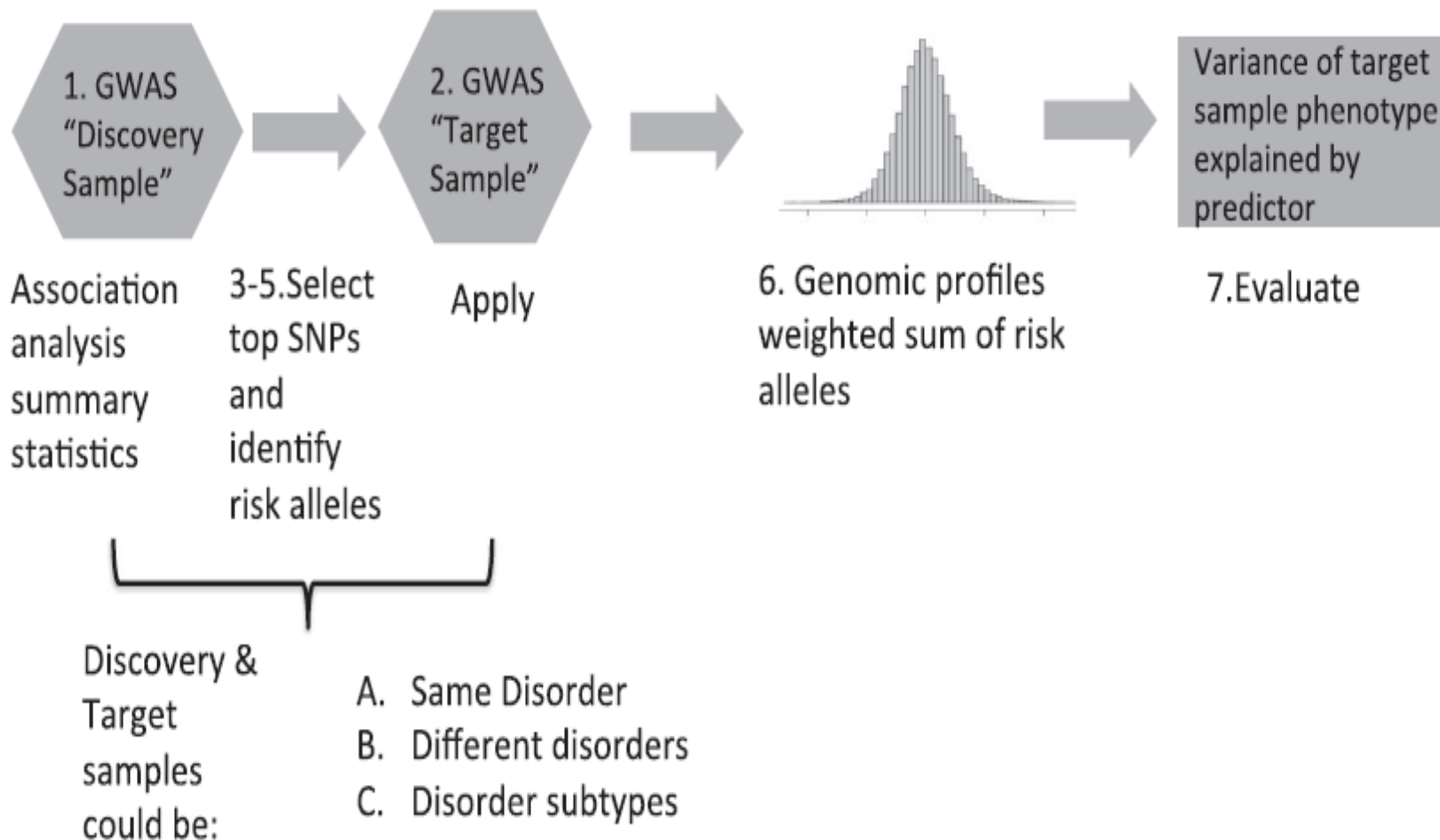
- Wray NR, Goddard, ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Research. 2007; 7(10):1520-28.
- Evans DM, Visscher PM., Wray NR. Harnessing the information contained within genome-wide association studies to improve individual prediction 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.



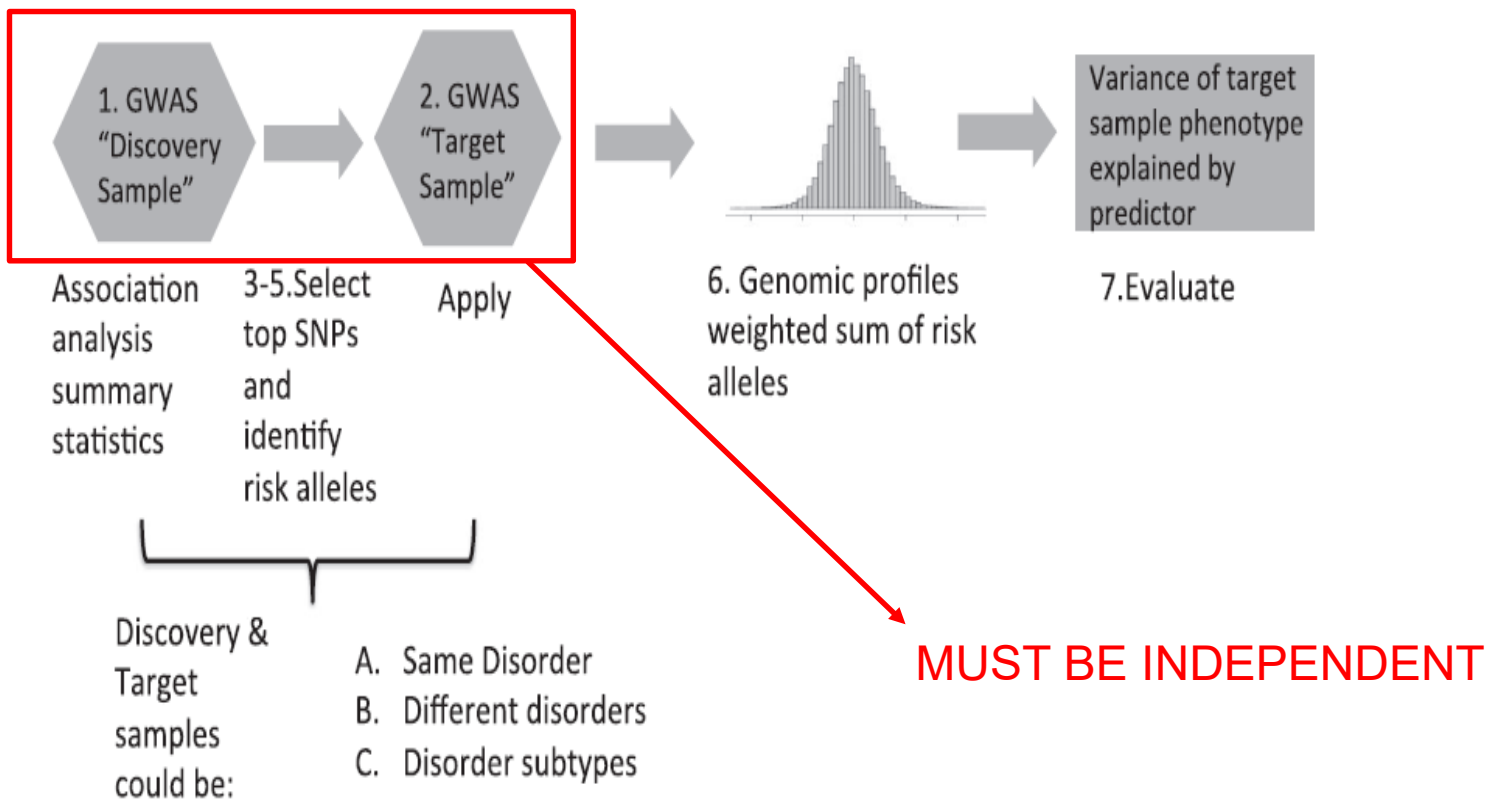
# Further reading

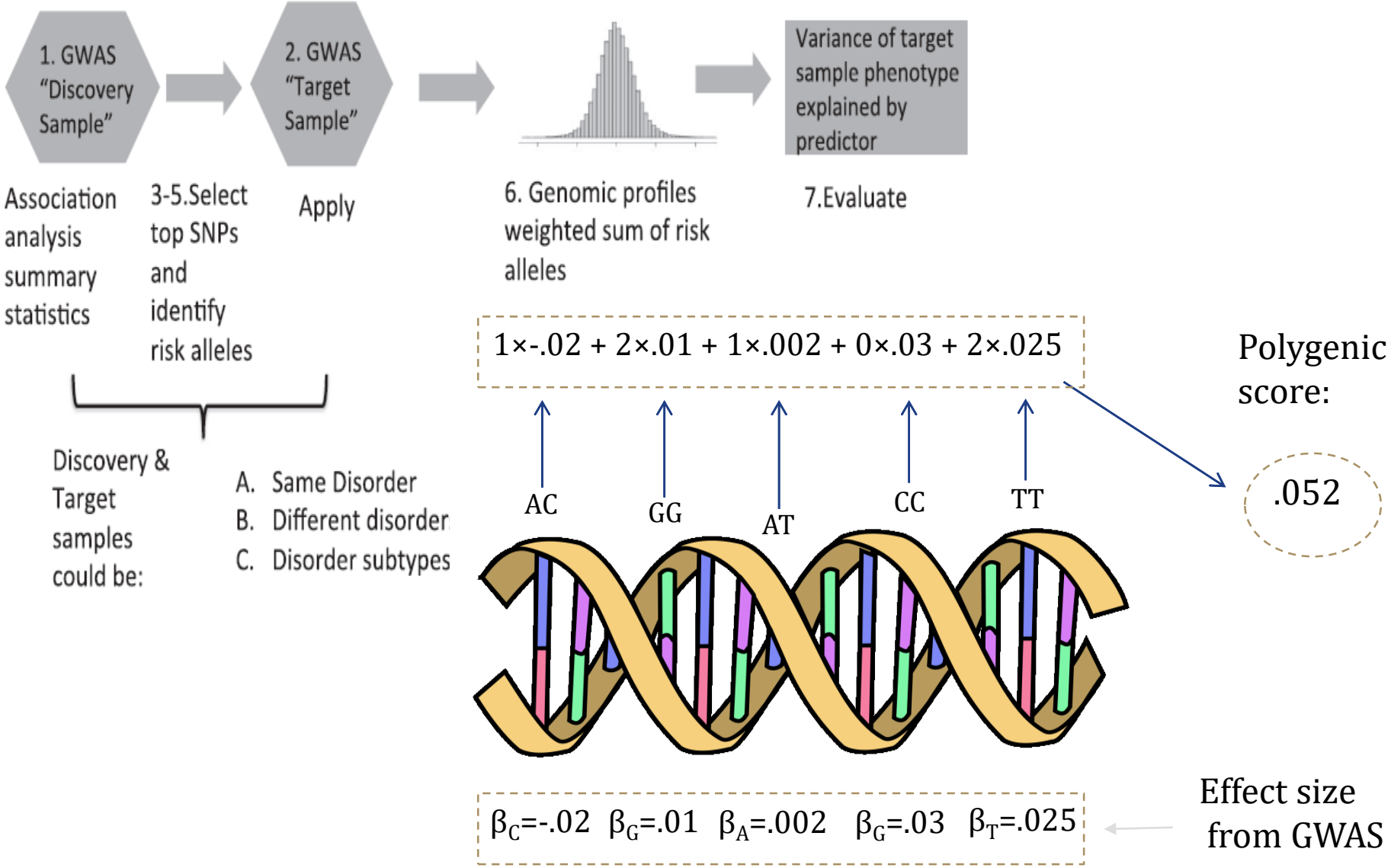
- Dudbridge F. Power and predictive accuracy of polygenic risk scores. PLoS Genet. 2013 Mar;9(3):e1003348. Epub 2013 Mar 21. Erratum in: PLoS Genet. 2013;9(4). **(Important discussion of power)**
- Wray NR, Lee SH, Mehta D, Vinkhuyzen AA, Dudbridge F, Middeldorp CM. Research review: Polygenic methods 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, 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



# Traditional approach



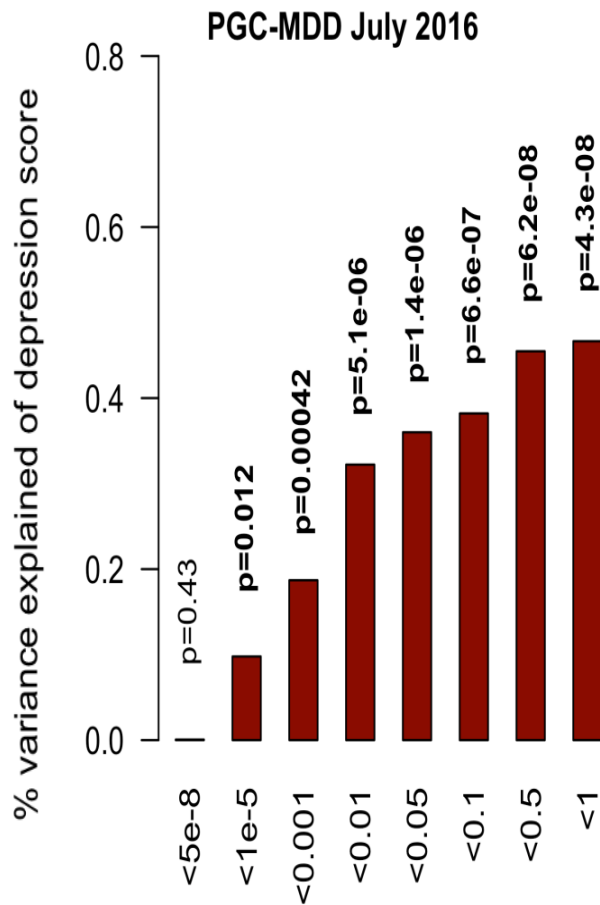


# Main uses of PRS

- 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](http://www.nature.com/mp)

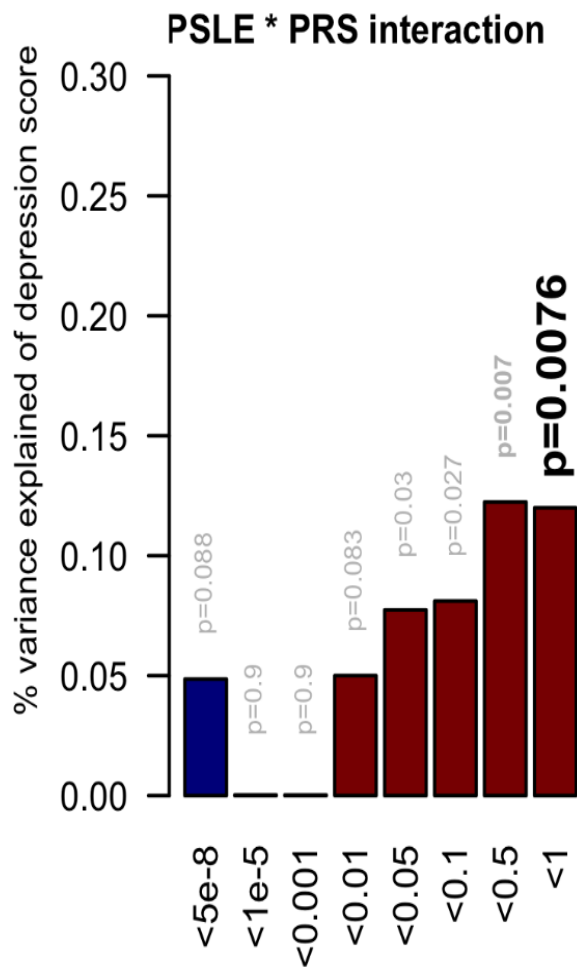
## ORIGINAL ARTICLE

### A direct test of the diathesis–stress model for depression

L Colodro-Conde<sup>1,2,12</sup>, B Couvy-Duchesne<sup>1,3,12</sup>, G Zhu<sup>1</sup>, WL Coventry<sup>4</sup>, EM Byrne<sup>5</sup>, S Gordon<sup>1</sup>, MJ Wright<sup>3,6</sup>, GW Montgomery<sup>5</sup>, PAF Madden<sup>7</sup>, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium<sup>13</sup>, S Ripke<sup>8,9,10</sup>, LJ Eaves<sup>11</sup>, AC Heath<sup>7</sup>, NR Wray<sup>3,5</sup>, SE Medland<sup>1</sup> and NG Martin<sup>1</sup>

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](http://www.nature.com/mp)

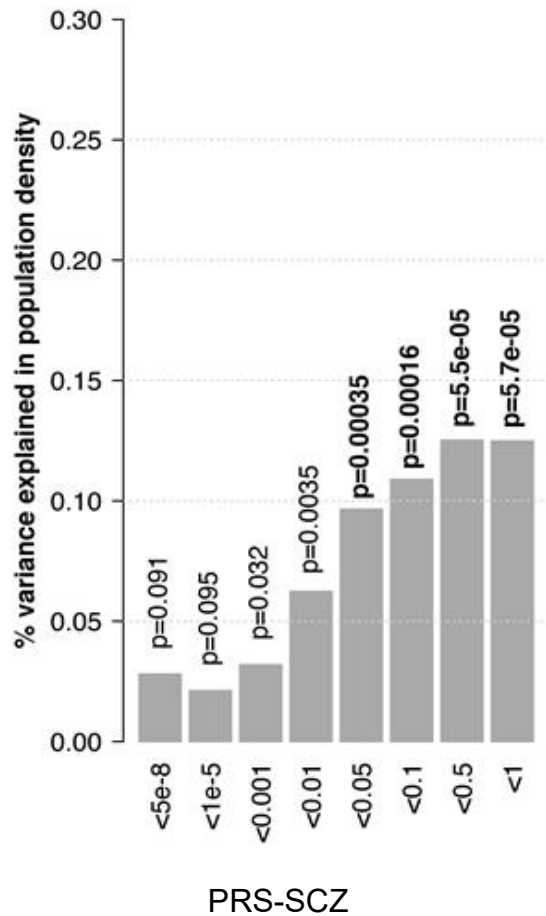
## ORIGINAL ARTICLE

### A direct test of the diathesis–stress model for depression

L Colodro-Conde<sup>1,2,12</sup>, B Couvy-Duchesne<sup>1,3,12</sup>, G Zhu<sup>1</sup>, WL Coventry<sup>4</sup>, EM Byrne<sup>5</sup>, S Gordon<sup>1</sup>, MJ Wright<sup>3,6</sup>, GW Montgomery<sup>5</sup>, PAF Madden<sup>7</sup>, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium<sup>13</sup>, S Ripke<sup>8,9,10</sup>, LJ Eaves<sup>11</sup>, AC Heath<sup>7</sup>, NR Wray<sup>3,5</sup>, SE Medland<sup>1</sup> and NG Martin<sup>1</sup>

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



**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, PhD<sup>1</sup>; Baptiste Couvy-Duchesne, PhD<sup>1,2,3</sup>; John B. Whitfield, PhD<sup>1</sup>; [et al](#)

[» Author Affiliations](#)

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




# Sub-type analysis

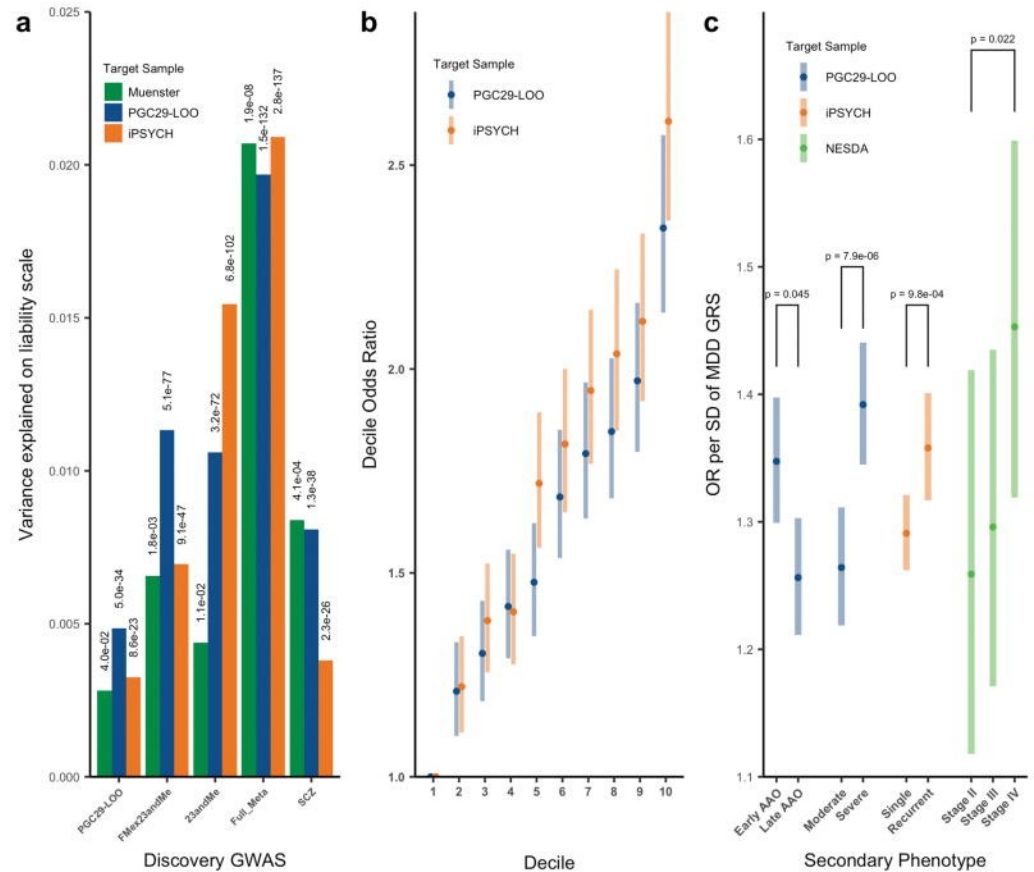
nature  
genetics

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%).

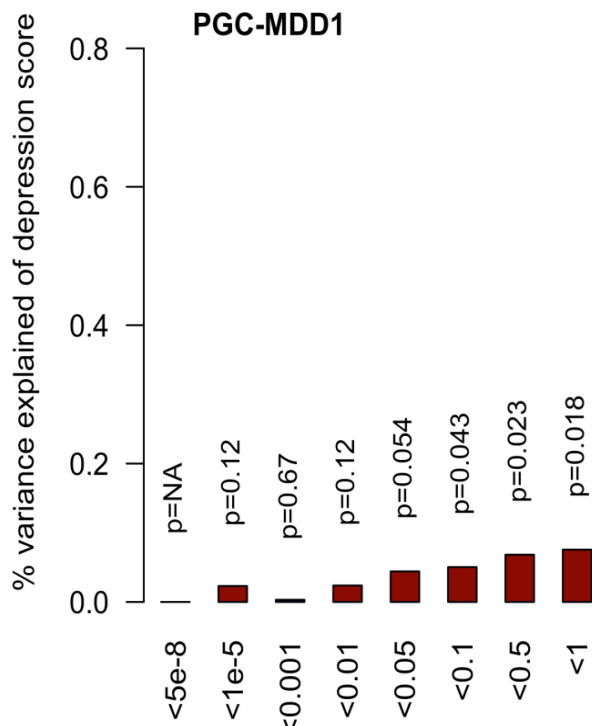
As a general rule of thumb you usually want 2,000+ people in the target dataset.

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

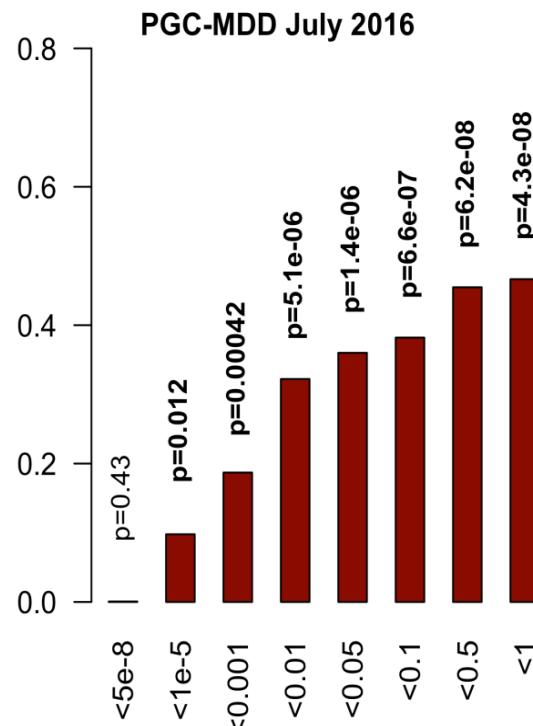
Power calculator for discovery (GWAS) sample needed to achieve prediction of  $r^2$  in target sample

```
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*



---

# General steps of processing

# (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

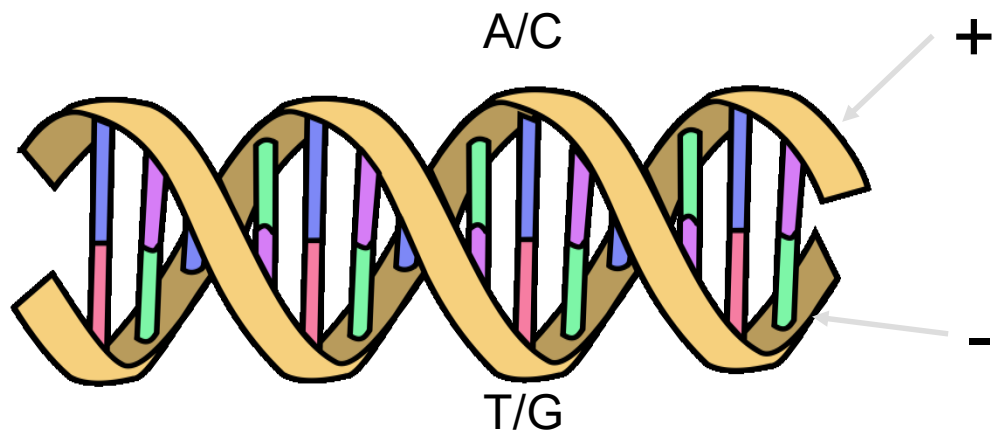
## (2) Find SNPs in common with your local sample and QC

- 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
```

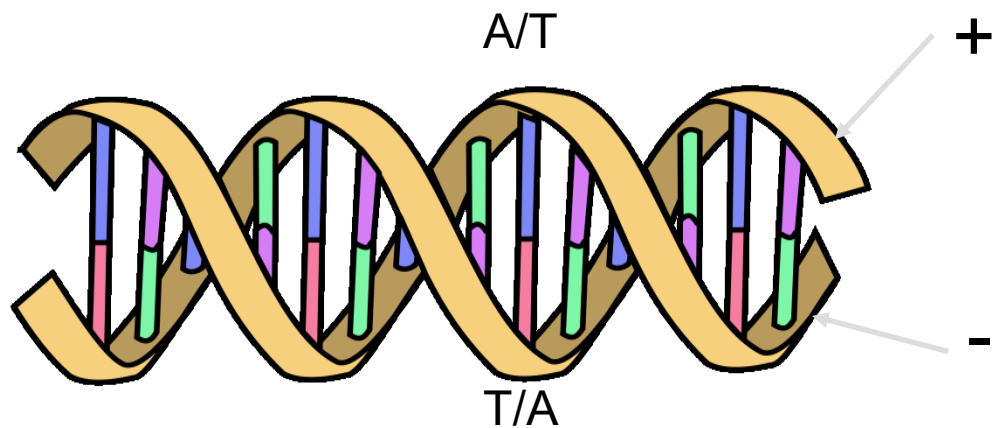
# (\* ) On ambiguous strands

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



rsXXX    A    C  
          MAF

rsXXX    T    G  
          MAF



rsXXX    A    T  
          MAF

rsXXX    T    A  
          1-MAF

# (3) Clumping

- Select most associated SNP per LD region (pruning)
- Plink1.9                    --bfile bfileReferencePanelForLD  
                                  --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

```
#Clump data in 2 rounds using plink2
```

```
#1st clumping & extract tops snps for 2nd round
```

```
for ((i=1;i<=22;i++))
```

```
do
```

```
plink2 --bfile reference --chr "$i" --extract available.snps --clump GWAS.noambig
```

```
--clump-p1 1 --clump-p2 1 --clump-r2 .5 --clump-kb 250 --out traitX"$i".round1
```

```
awk '{print $3, $5}' traitX"$i".round1.clumped > traitX"$i".round2.input
```

```
awk '{print $3}' traitX"$i".round1.clumped > traitX"$i".extract2
```

```
done
```

```
#2nd clumping & extract tops snps for profile
```

```
for ((i=1;i<=22;i++))
```

```
do
```

```
plink2 --bfile reference --chr "$i" --extract traitX"$i".extract2 --clump traitX"$i".round2.input --
```

```
clump-p1 1 --clump-p2 1 --clump-r2 .2 --clump-kb 5000 --out traitX"$i".round2
```

```
awk '{print $3}' traitX"$i".round2.clumped > traitX"$i".selected
```

```
done
```

## (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 association 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 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
```

---

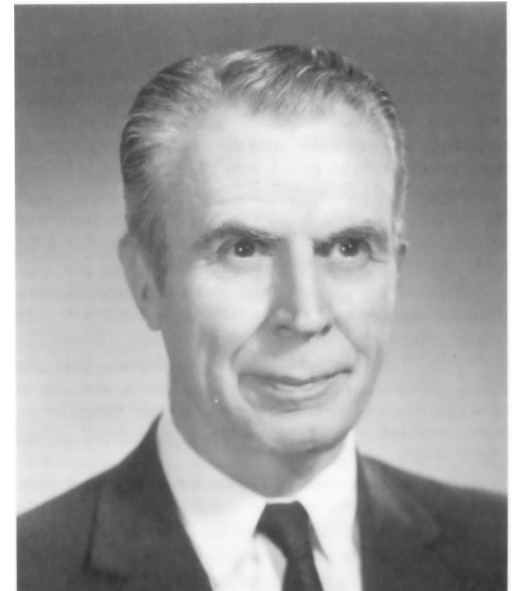
Other Methods

# 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

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

N individuals

$Y_{N \times 1}$  phenotype centered

$X_{N \times 1}$  SNP centered

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

Joint and conditional  
SNP effect

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

N individuals

$Y_{N \times 1}$  phenotype centered

$X_{N \times m}$  SNPs centered

Yang et al., 2012

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

BLUP effect

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

N individuals

$Y_{N \times 1}$  phenotype centered

$Z_{N \times m}$  SNPs centered

$S_{m \times 1}$  vector of SNP effects  
assumed  $\sim N(0, \boldsymbol{\sigma}_s^2)$

Goddard et al., 2009

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

# Calculating BLUP effect sizes

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

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

$\mathbf{Z}'\mathbf{Z}$ :  $n \times n$  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
```

$$\text{--cojo-sblup} = m * (1 / h_{\text{SNP}}^2 - 1)$$

With  $m$  the number of SNPs

# 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

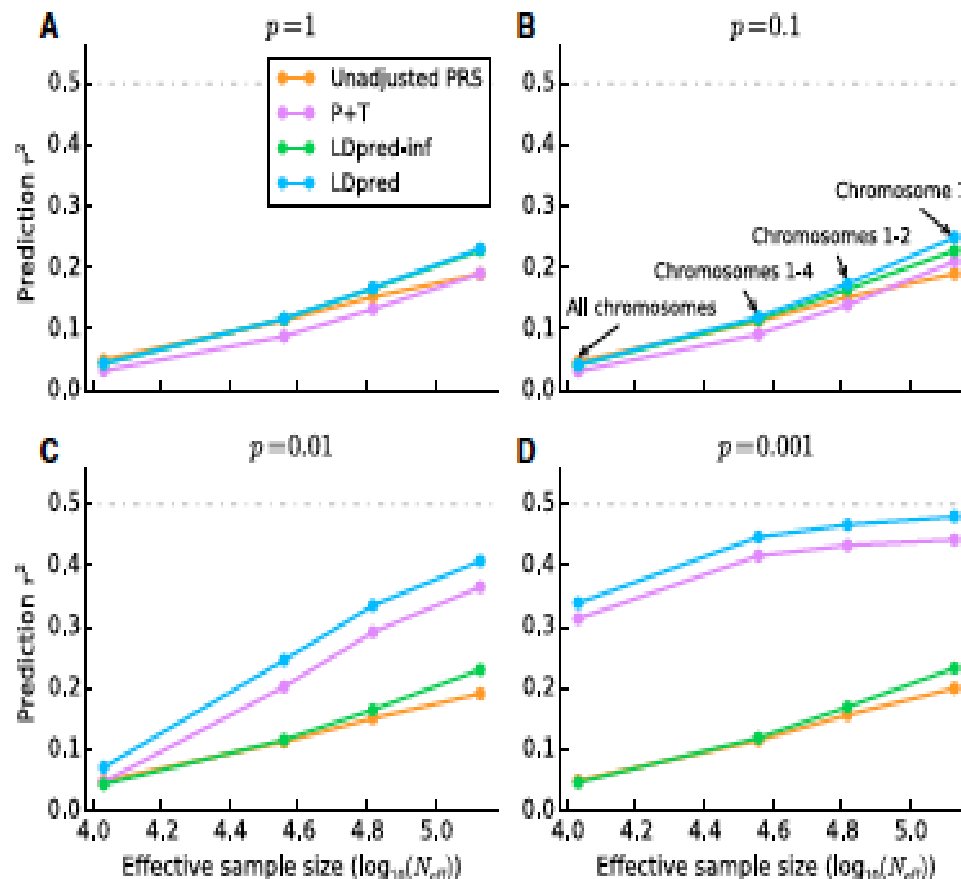
- BLUP “shrinks” the estimates: hypothesis of normally distributed effect sizes “infinitesimal model”
- Other shrinkage methods include LASSO: hypothesis of mixture of effect sizes (double exponential...)
- Non-additive models? That may include epistasis, dominance
- Semi-parametric models  
see **Goddard 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



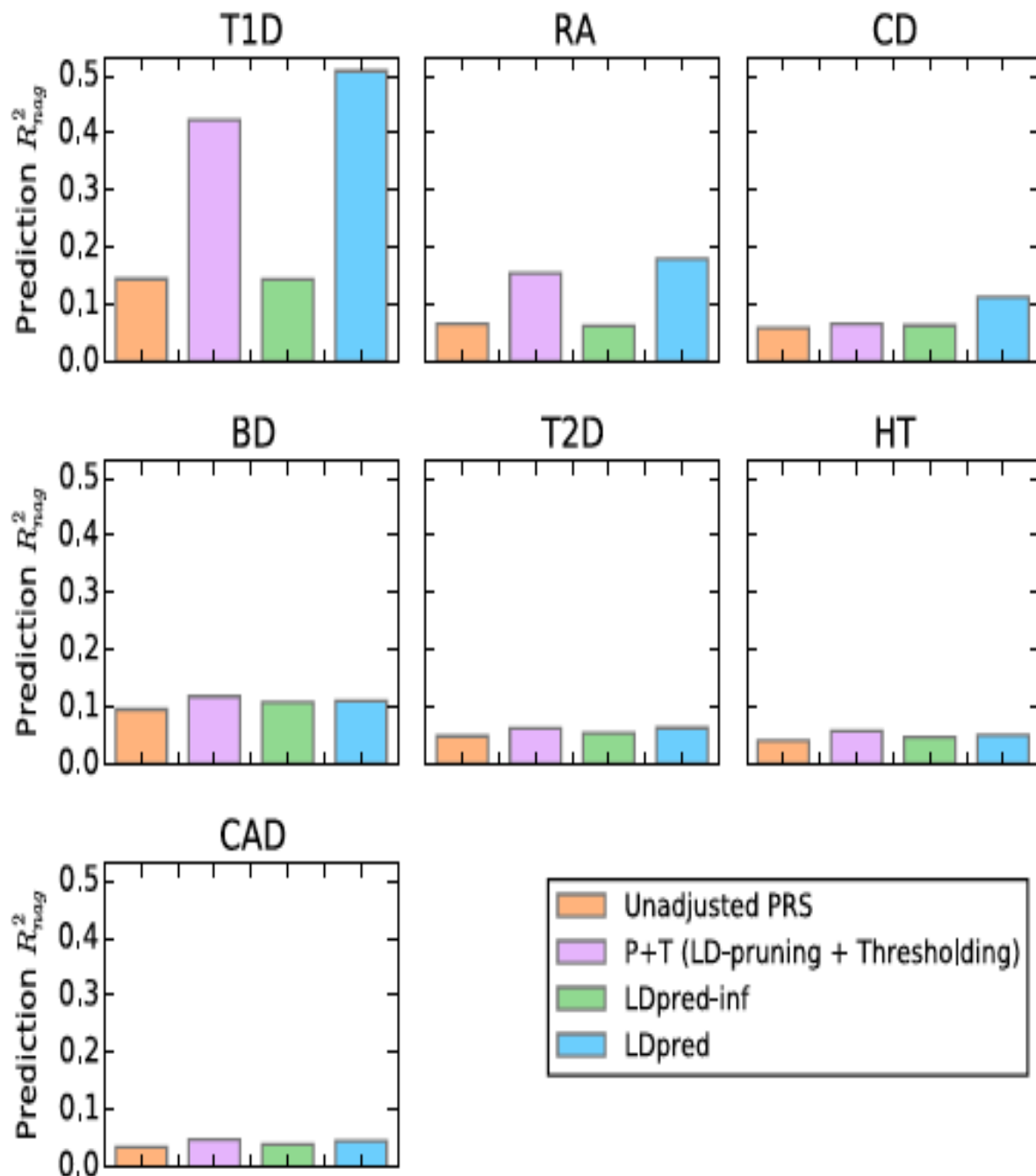
**Figure 2. Comparison of Four Prediction Methods Applied to Simulated Traits**

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  $N_{\text{eff}} = (N/M_{\text{sim}})M$ , where  $N = 10,786$  is the training sample size,  $M = 376,901$  is the total number of SNPs, and  $M_{\text{sim}}$  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

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.

Euesden et al., 2014

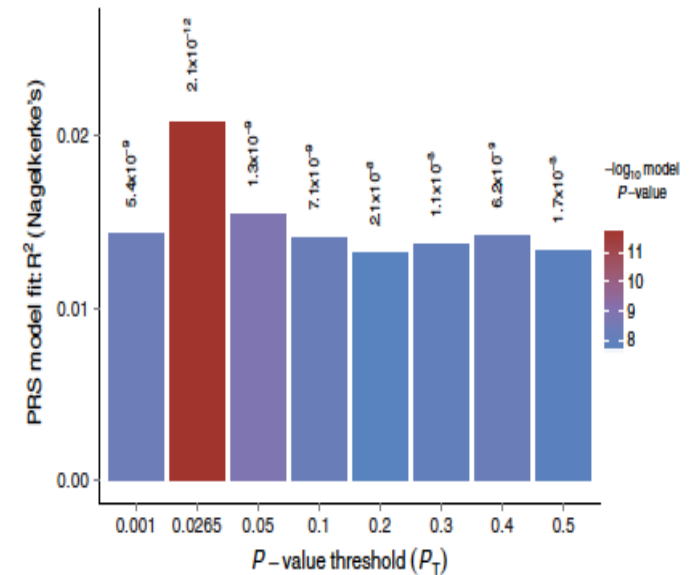


Fig. 1. Bar plot from PRSice showing results at broad  $P$ -value thresholds for Schizophrenia PRS predicting MDD status. A bar for the best-fit PRS from the high-resolution run is also included

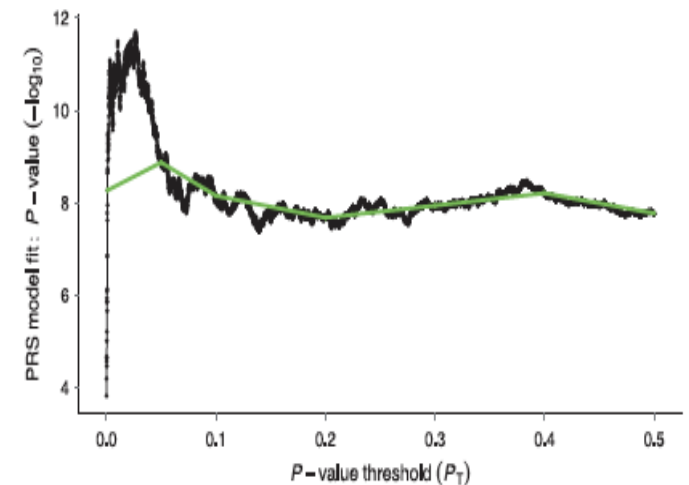


Fig. 2. High-resolution PRSice plot for SCZ predicting MDD status. The thick line connects points at the broad  $P$ -value thresholds of Fig. 1

<b>Classic / OLS</b>	<b>BLUP</b>	<b>PRSice</b>
Dosage or best guess clumping	Best guess	Dosage or best guess clumping
Multiple PRS by p-value thresholds	BLUP effects summed over all SNPs	Unique PRS
Bonferroni correction		All p-value threshold tested
		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)

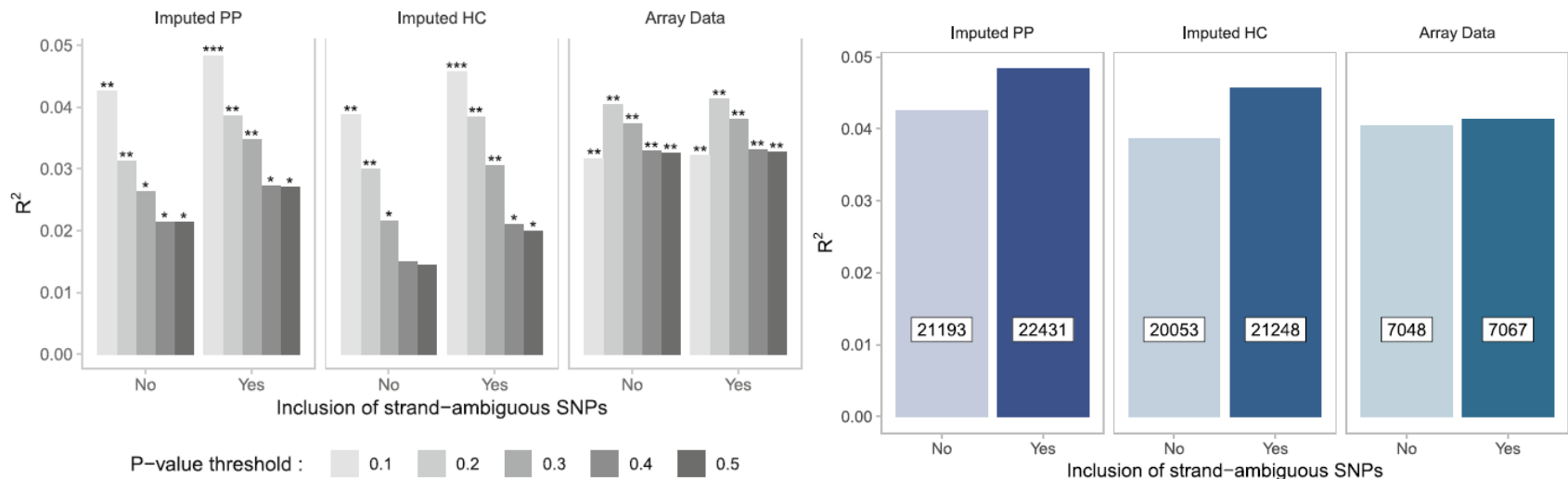
SOFTWARE

Open Access



# PRS-on-Spark (PRSoS): a novel, efficient and flexible approach for generating polygenic risk scores

Lawrence M. Chen<sup>1,2</sup>, Nelson Yao<sup>1,2</sup>, Erika Garg<sup>1,2</sup>, Yuecai Zhu<sup>1,2</sup>, Thao T. T. Nguyen<sup>1,2</sup>, Irina Pokhvisneva<sup>1,2</sup>, Shantala A. Hari Dass<sup>1,2</sup>, Eva Unternaehrer<sup>1,2</sup>, Hélène Gaudreau<sup>1,2</sup>, Marie Forest<sup>2,3</sup>, Lisa M. McEwen<sup>4</sup>, Julia L. Maclsaac<sup>4</sup>, Michael S. Kobor<sup>4</sup>, Celia M. T. Greenwood<sup>2,3,5,6,7</sup>, Patricia P. Silveira<sup>1,2,8,9</sup>, Michael J. Meaney<sup>1,2,8,9,10,11</sup> and Kieran J. O'Donnell<sup>1,2,8,9,10\*</sup>



---

A couple of “worked” examples

**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, PhD<sup>1</sup>; Baptiste Couvy-Duchesne, PhD<sup>1,2,3</sup>; John B. Whitfield, PhD<sup>1</sup>; [et al](#)

[» Author Affiliations](#)

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



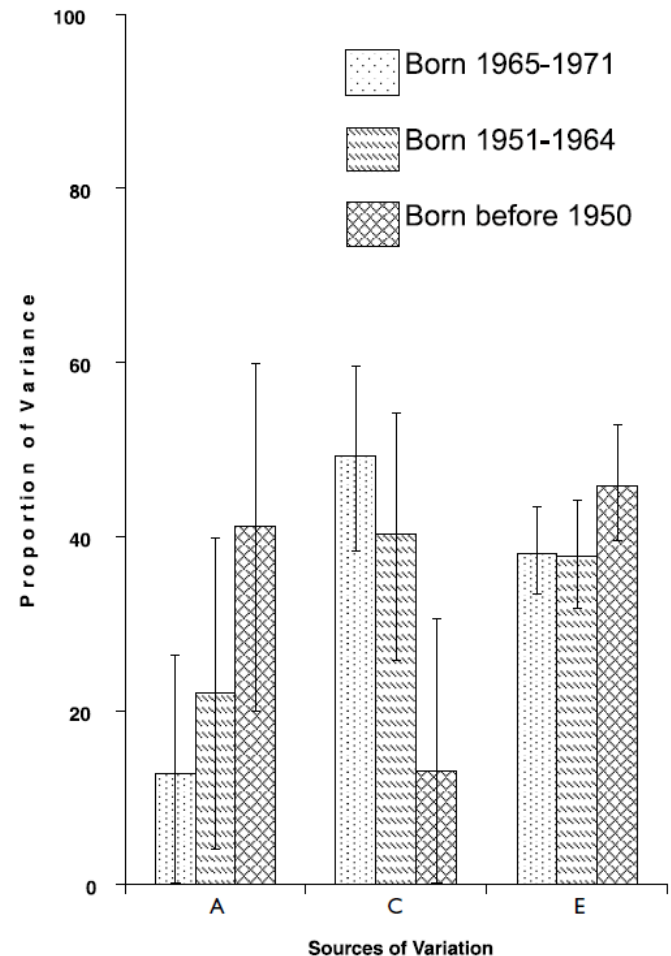
# Background

- The prevalence of schizophrenia is higher in urban areas than in rural areas → O.R. = 2.39 (1.62–3.51), (Vassos et al 2012, *Schizophrenia Bulletin*).
- Two major hypotheses have been proposed to explain this phenomenon:
  - (1) **causation hypothesis**: the stress of city life and undefined factors in the urban environment increase the risk of this disease.
  - (2) **selection hypothesis**: individuals with genetic liability for schizophrenia move into urban areas.



- Twin models have shown genetic factors have a higher impact on the country vs. city living as people grow older, while the impact of family background decreases.

Whitfield et al. 2005, *Twin Research and Human Genetics*

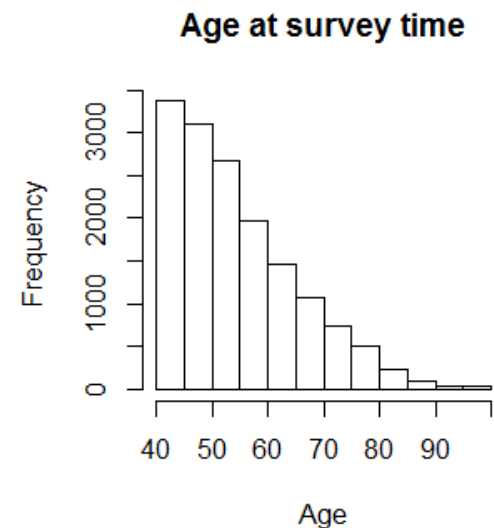


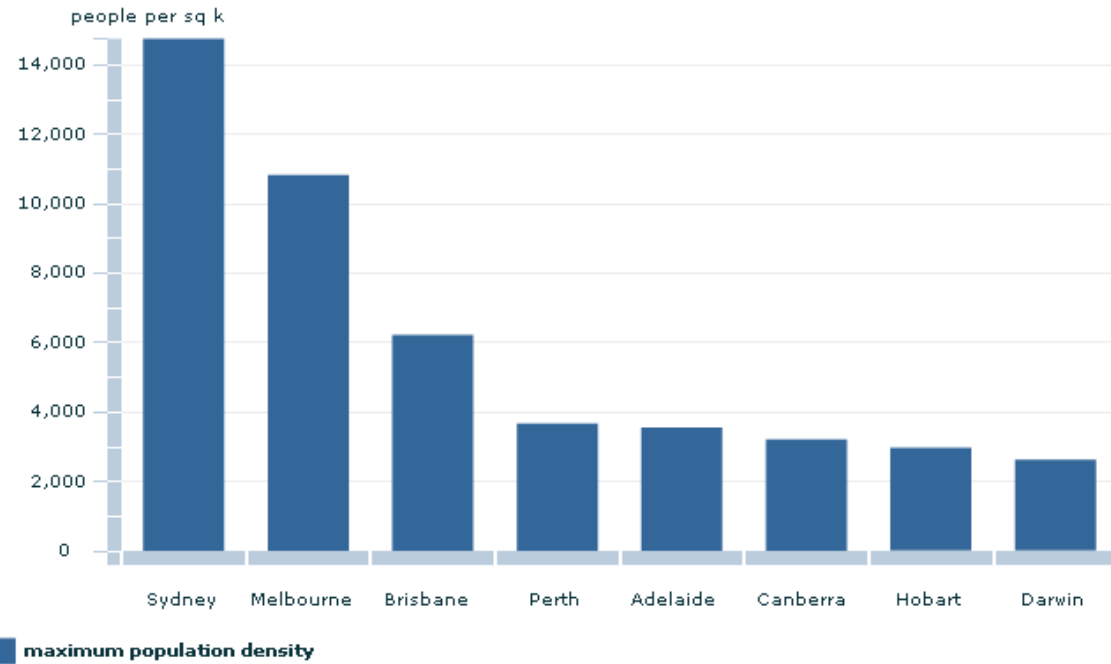
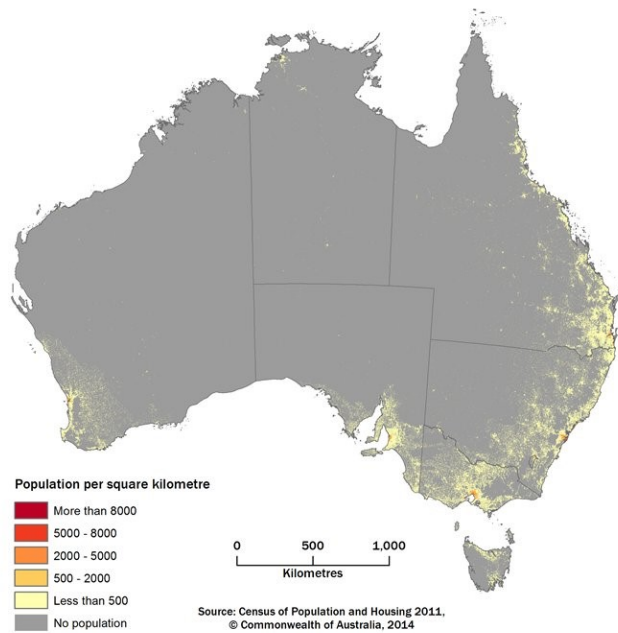
# Hypothesis

**Adults with higher genetic risk for schizophrenia are more likely to live in urbanised and populated areas than those with lower risk.**

# Methods

- 15,544 individuals in 7,015 families (65.6% females, age mean: 54.4, SD: 13.2) living in Australia.
- Participants were genotyped genome-wide and imputed to 1000G v.3.
- Reported their **postcode** as part of the protocols of several studies on health and wellbeing conducted from QIMR.





Measures of urbanicity:

Population density

Remoteness

+Socio economic status (SES)

(data from the Australian Bureau of Statistics)



**phenotype= intercept + beta0\*covariates + beta1\*g + e with  $g \sim N(0, GRM)$**

phenotype: population density or remoteness

covariates: PRS-SCZ, age, sex, (SES),

4 first genetic principal components, imputation chip

e: error

GRM: Genetic correlation matrix

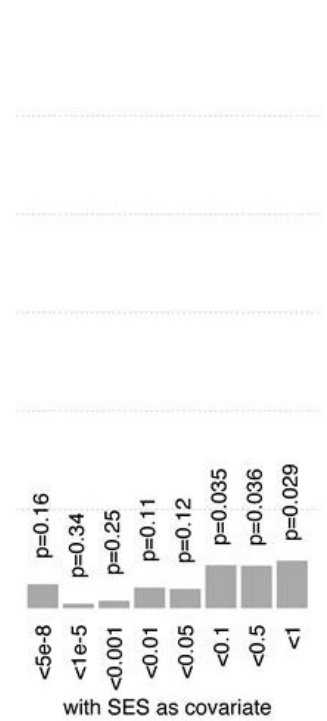
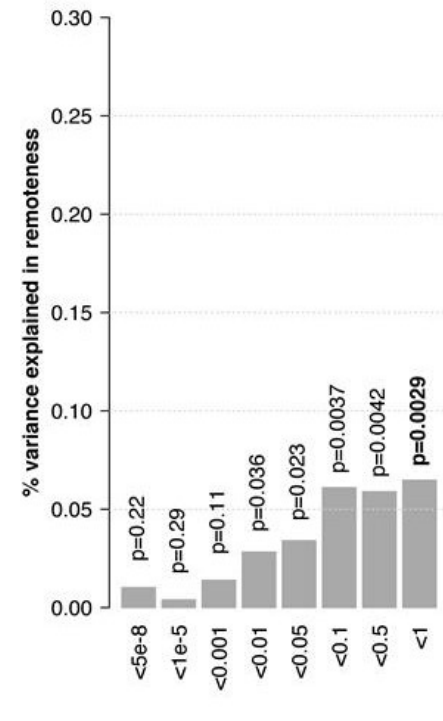
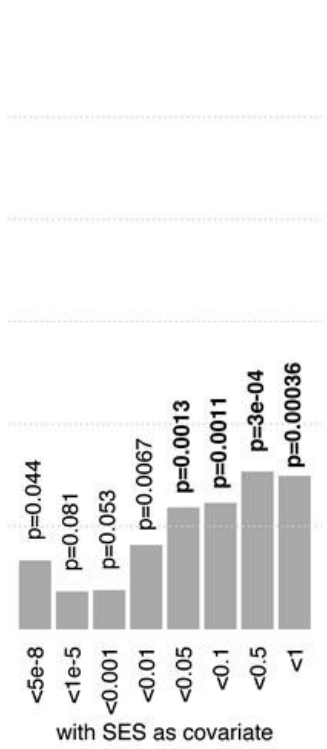
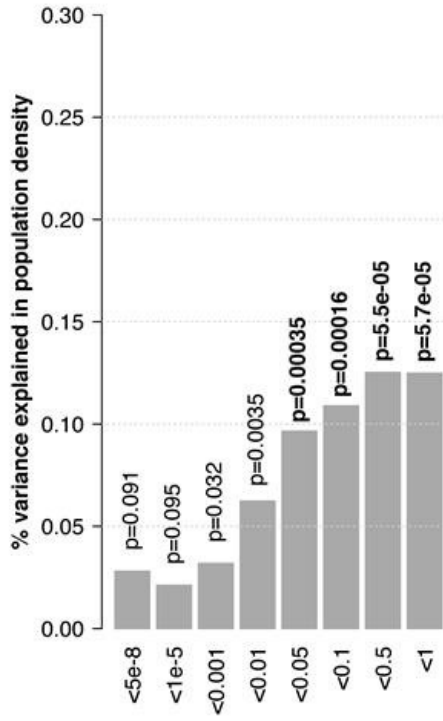
We calculated p-values using the t-statistic calculated on the basis of the Fix\_eff and SE from the GCTA output.

We then applied Bonferroni correction (Sidak method) for multiple testing yielding a significant threshold of 0.004.

**Genome-wide Complex Trait Analysis v. 1.22**

(Yang J et al 2011, Am J Hum Genet )

(a)

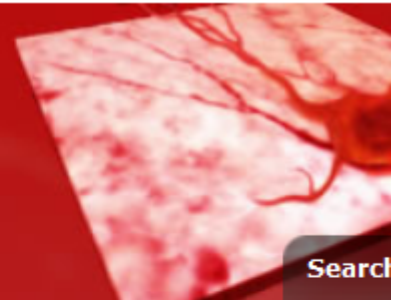


# Conclusions

- People with a **higher genetic risk for schizophrenia** may prefer to live in more **urban and populated areas**.
  - Importantly, this study does not use a case-control sample but an unselected population sample where the genetic risk for schizophrenia was estimated.
- Greater genetic predisposition to schizophrenia is at least **one mechanism** explaining why this illness is more prevalent in city environments.
- Future research should test if this effect is replicated in another countries, analyse migration effects and identify what aspects of urbanised life correlate with SCZ genetic risk.



# Molecular Psychiatry



Journal home > Advance online publication > 11 July 2017 > Full text

Journal home

Advance online publication

About AOP

Current issue

Archive

Web focus

Press releases

Online submission

## Original Article

Molecular Psychiatry advance online publication 11 July 2017; doi: 10.1038/mp.2017.130

### A direct test of the diathesis–stress model for depression

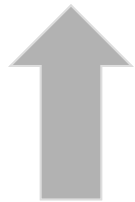
OPEN

L Colodro-Conde<sup>1,2,12</sup>, B Couvy-Duchesne<sup>1,3,12</sup>, G Zhu<sup>1</sup>, W L Coventry<sup>4</sup>, E M Byrne<sup>5</sup>, S Gordon<sup>1</sup>, M J Wright<sup>3,6</sup>, G W Montgomery<sup>5</sup>, P A F Madden<sup>7</sup>, Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium<sup>13</sup>, S Ripke<sup>8,9,10</sup>, L J Eaves<sup>11</sup>, A C Heath<sup>7</sup>, N R Wray<sup>3,5</sup>, S E Medland<sup>1</sup> and N G Martin<sup>1</sup>

# Diathesis-Stress model in depression

**Depression = Diathesis + Stress + Diathesis\*Stress**  
(Predisposition, Vulnerability) (Disruption of psychological equilibrium)

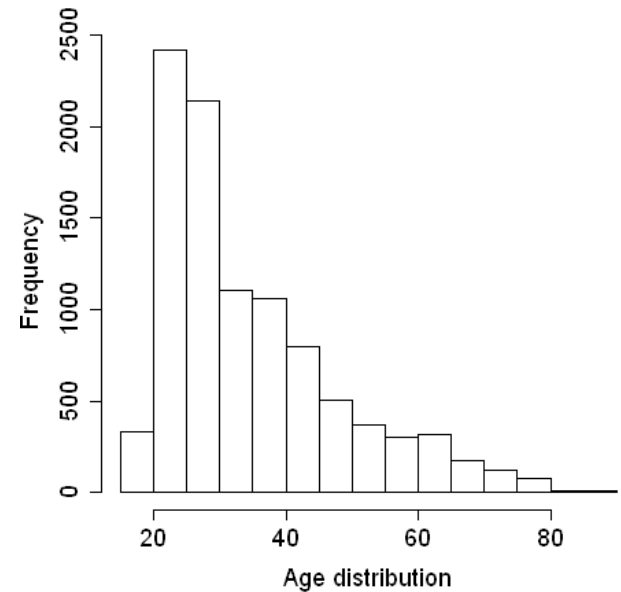
Hypothesised contribution to risk



**Depression = Polygenic risk scores (PRS) + Personal stressful life events (PSLE) Network stressful life events (NSLE) lack of social support (SS) + PRS\*PSLE PRS\*NSLE PRS\*SS**

# Sample and data

- **5,221 twins from 3,083 twin families**
- **European ancestry** (<6SD from PC1/PC2 centroid)
- **Mean age 35.7**, range 17-85, **66% females**



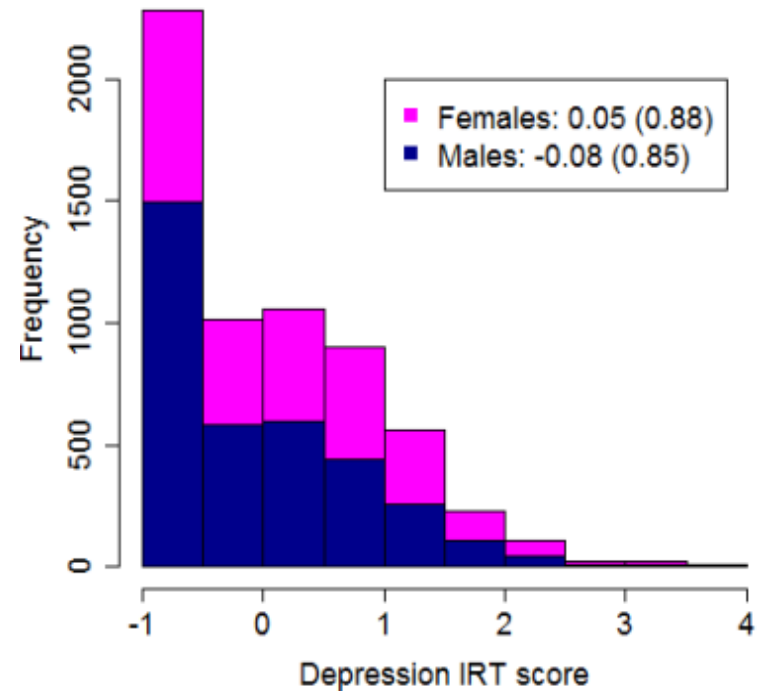
- **Depression, Personal & Network stressful life events, Perceived social support**
- **GWAS arrays, imputed to 1000G reference**

# Measure of Depression

## 12 depression items

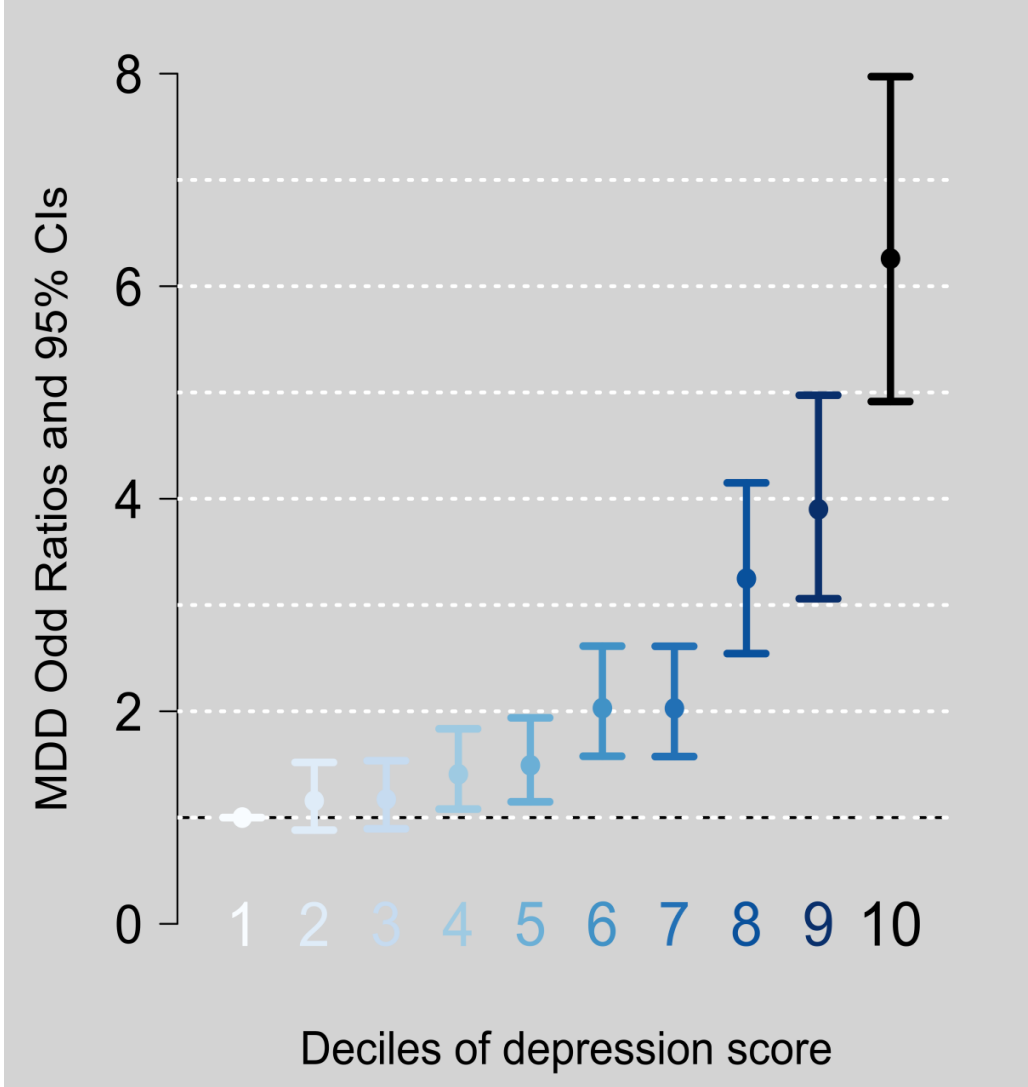
Selected from:

Delusion Symptoms Scale Inventory (DSSI)  
+  
Symptoms Check List (SCL)



**All scores were estimated using Item Response Theory (IRT)  
– improves distribution, deals with missingness**

# REALITY CHECK - Increased odds of DSM-IV MDD diagnosis per decile of depression IRT score



- Association between Depression score and lifetime DSM-IV MDD diagnoses from telephone interview studies conducted 4-12 years after DSSI/SCL)
  - p-value = 3.0e-108
- **Disease odds >6x in top decile of depression score compared to first decile**

# Measures of Stress

## Personal stressful life events (PSLE)

(adapted from List of threatening experiences (Brugha et al. 1985,))

**Serious problem** with spouse, family member, friend, neighbour, workmate  
**Event:** Divorce, separation, illness, injury, accident, burgled, robbed, lost job, financial problems, legal troubles...

## Network stressful life events (NSLE)

(adapted from List of threatening experiences)

**Illness, Injury, death or personal crisis in close network** (spouse, child, mother, father, twin, sibling, someone else close)

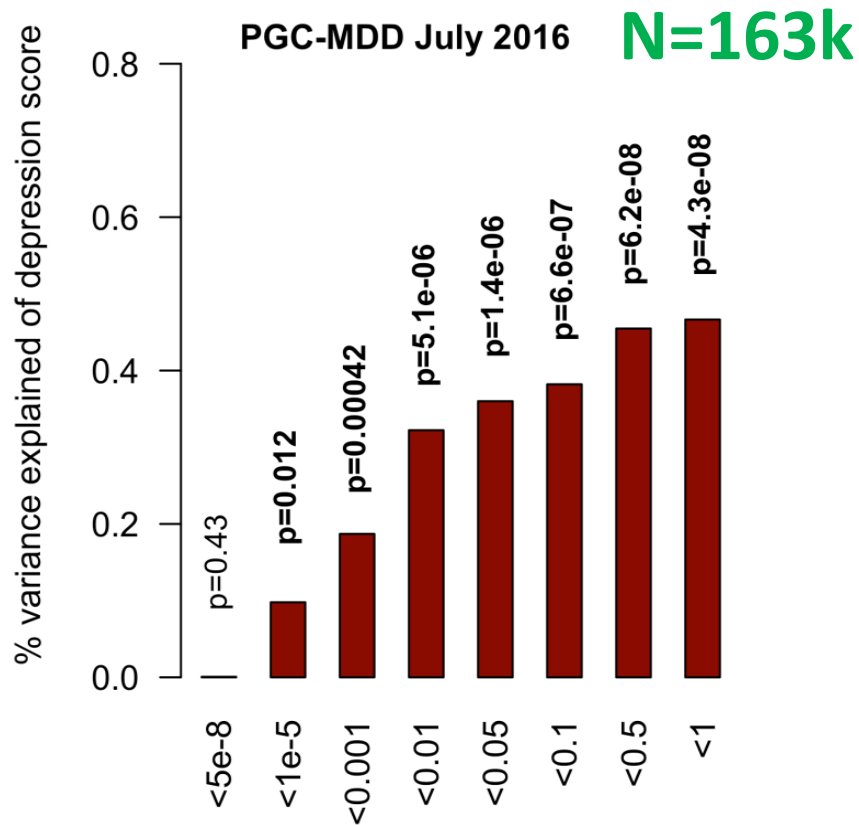
## Social Support (SS)

(Kessler Perceived SS, KPSS)

How much your **close network:** listens to your **worries**, **understands** the way you feel / think, **helps you** if needed, **shares private feelings** with you

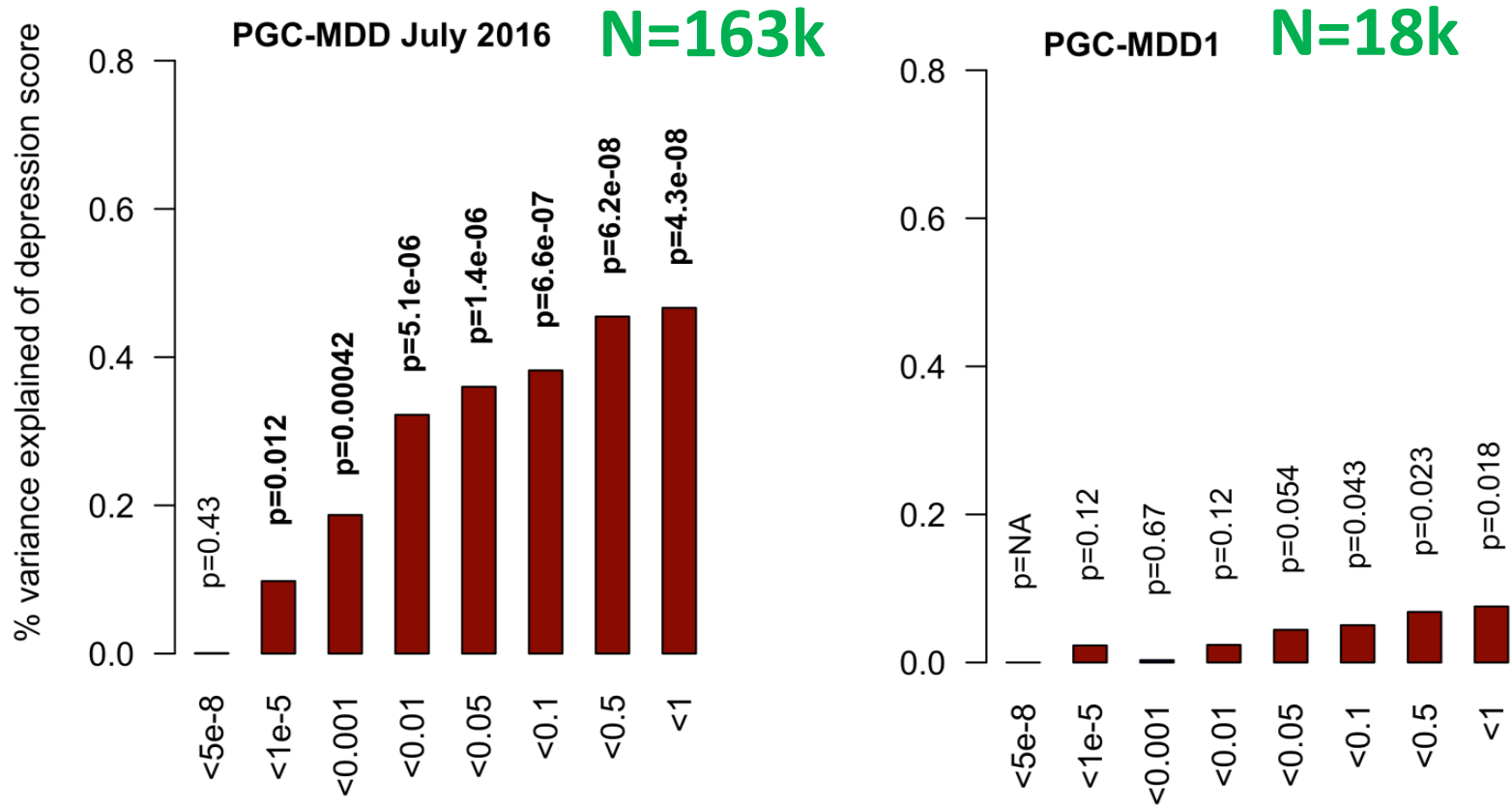
**All scores were estimated using Item Response Theory (IRT)**  
**– improves distribution, deals with missingness**

# MAIN EFFECTS - POLYGENIC RISK SCORES



(max variance explained = **0.46%**,  
p = 4.3e-08)

# Main effects - Polygenic Risk Scores

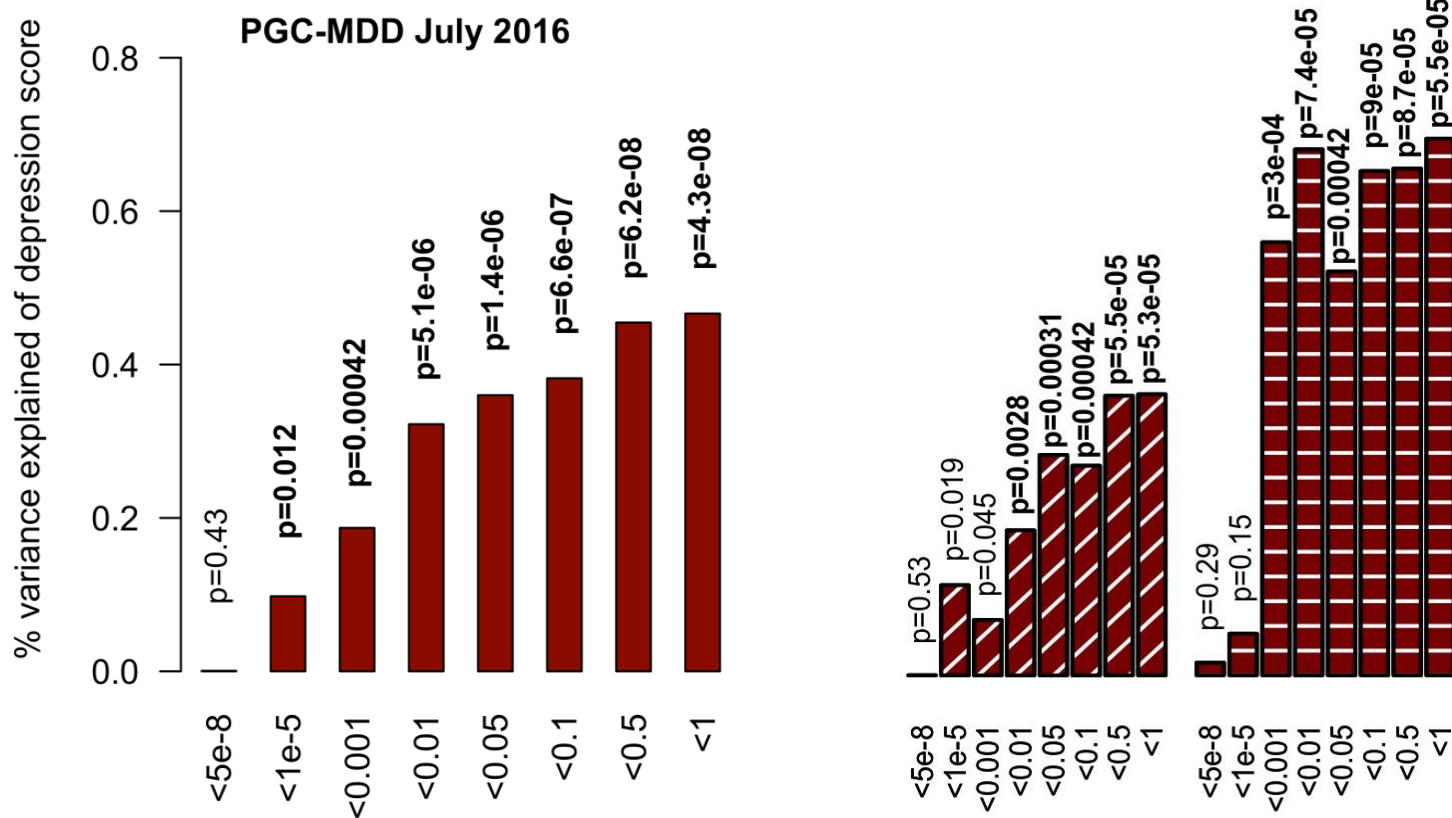


Note increased variance accounted for with larger N



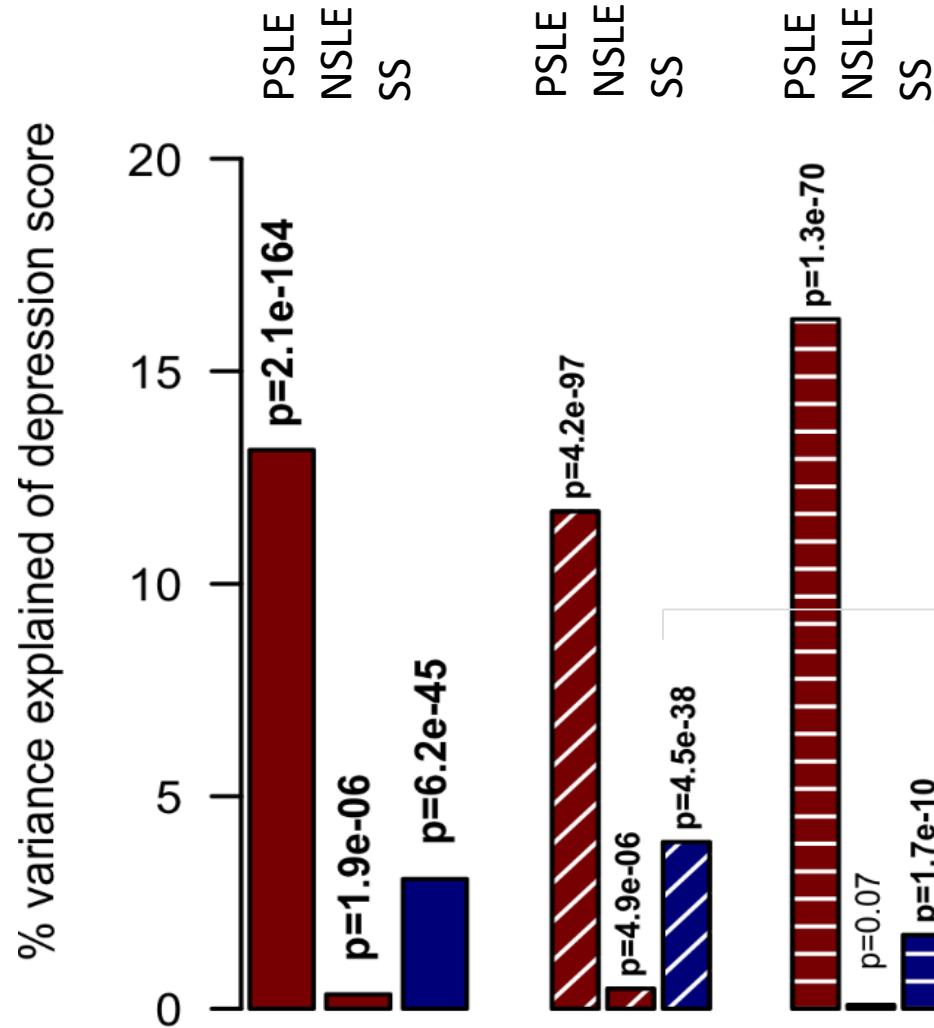
# MAIN EFFECTS - POLYGENIC RISK SCORES

PRS main effects appear larger in males



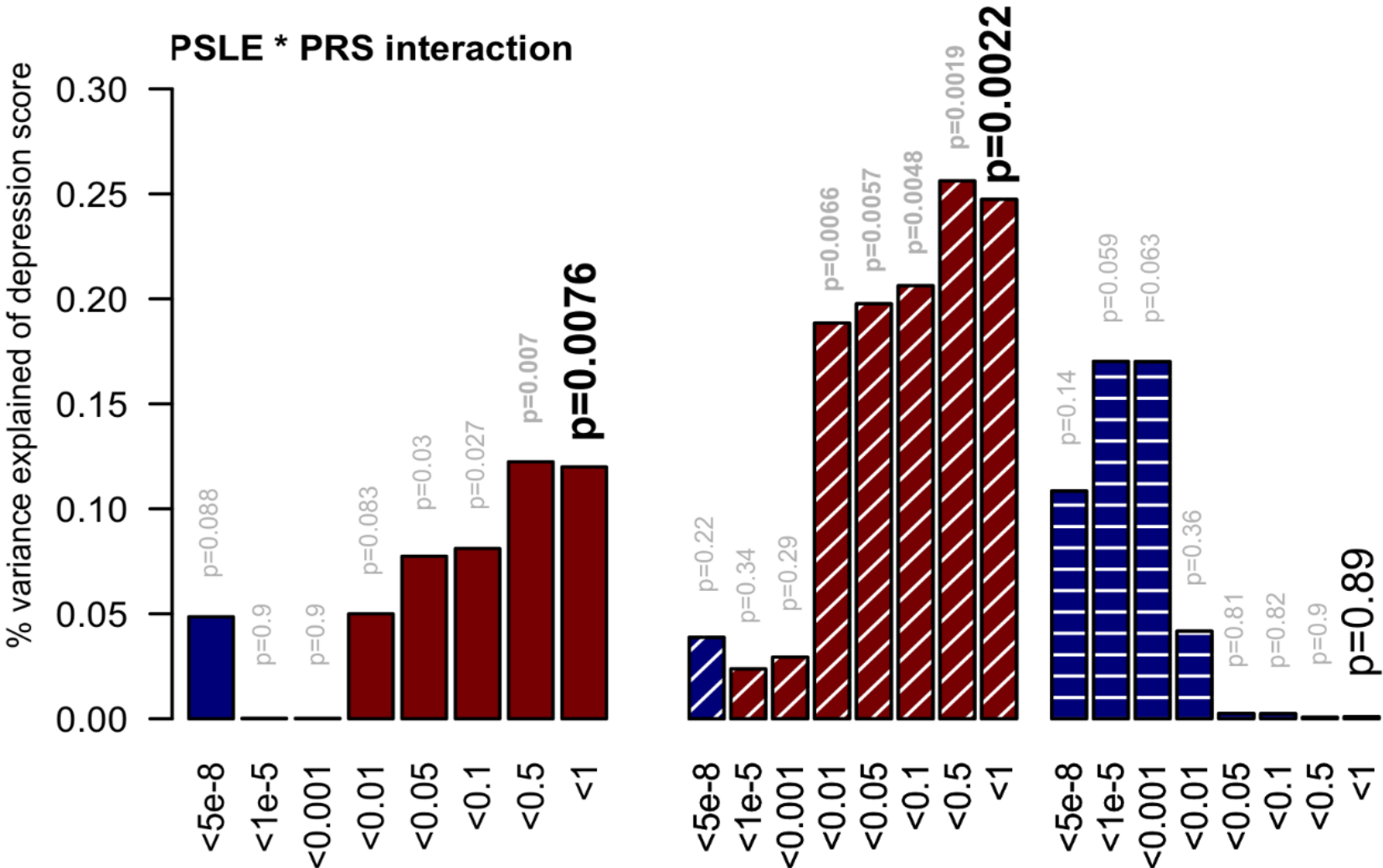
# MAIN EFFECTS - STRESSORS

Note sex differences for SS

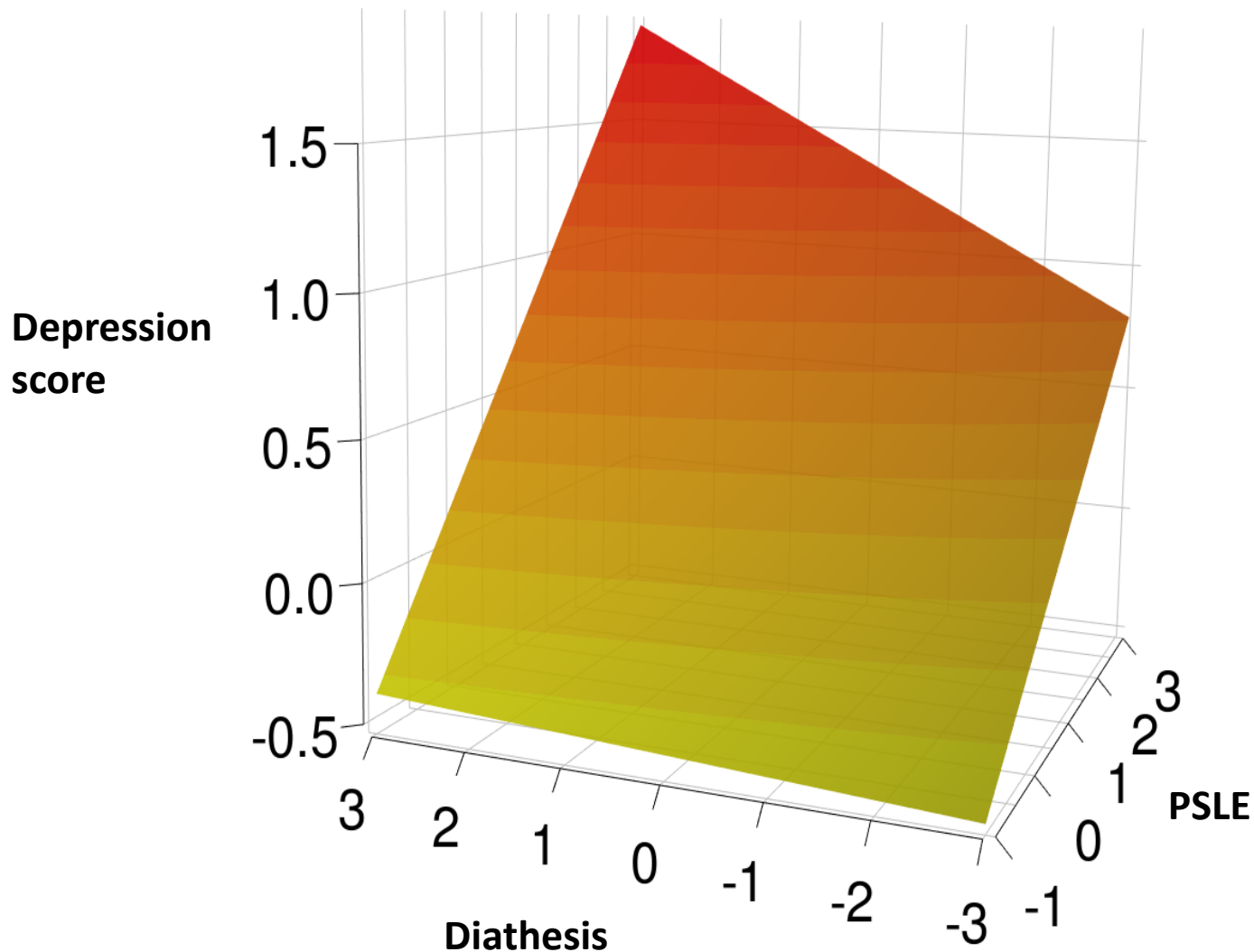


\* Blue is negative

# Test of Interaction - it's all coming from females !



*the effect of the PSLE-diathesis interaction is visible when comparing the bottom (minimal PSLE) and top (maximal PSLE) edges of the surface*



---

Practical

# Today's Data



Journal home > Archive > Letter > Abstract

<b>Journal content</b>
→ <b>Journal home</b>
→ <b>Advance online publication</b>
→ <b>Current issue</b>
→ <b>Archive</b>
→ <b>Focuses and Supplements</b>
→ <b>Press releases</b>
<b>Free Association</b>

## Letter abstract

*Nature Genetics* **39**, 1494 - 1499 (2007)  
Published online: 4 November 2007 | doi:10.1038/ng.2007.16

### A survey of genetic human cortical gene expression

Amanda J Myers<sup>1,2,10</sup>, J Raphael Gibbs<sup>1,3,10</sup>, Jennifer A Webster<sup>4,5,10</sup>, Kristen Rohrer<sup>1</sup>, Alice Zhao<sup>1</sup>, Lauren Marlowe<sup>1</sup>, Mona Kaleem<sup>1</sup>, Doris Leung<sup>1</sup>, Leslie Bryden<sup>1</sup>, Priti Nath<sup>1</sup>, Victoria L Zismann<sup>4,5</sup>, Keta Joshipura<sup>4,5</sup>, Matthew J Huentelman<sup>4,5</sup>, Diane Hu-Lince<sup>4,5</sup>, Keith D Coon<sup>4,5,6</sup>, David W Craig<sup>4,5</sup>, John V Pearson<sup>4,5</sup>, Peter Holmans<sup>7</sup>, Christopher B Heward<sup>8</sup>, Eric M Reiman<sup>4,5,9</sup>, Dietrich Stephan<sup>4,5,9</sup> & John Hardy<sup>1,3</sup>



- <http://labs.med.miami.edu/myers/LFuN/LFuN.html>
- post-mortem gene expression in 'brain' tissue
- N=364
- Real data – unfiltered!

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



The banner features a blue background with a glowing DNA double helix. 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 Broad Institute, with the text "UK Biobank GWAS Results" next to it.

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.

- for i in {1..22}
  - do
  - echo \$i
  - rm chr"\$i".pass
  - zcat chr"\$i".info.gz | awk '{ if (\$5>=.01 && \$7 >=.6) print \$1}' > chr"\$i".pass
  - done
- 
- for i in {1..22}
  - do
  - echo \$i
  - ~/bin/plink2 --vcf chr"\$i".dose.vcf.gz --extract chr"\$i".pass --make-bed --out QCchr"\$i" --threads 5
  - done
- 
- for i in {2..22}
  - do
  - echo QC"\$i".bed QC"\$i".bim QC"\$i".fam >>join.list
  - done
- 
- for i in 20160 20161 20162 2887 3466 3476
  - do
  - echo SNP P A1 A2 Beta > "\$i".4clumping
  - zless "\$i".assoc.tsv.gz | awk '{print \$1, \$6, \$9}' | sed 's:/:/g' | awk '{ if (NR>1) print \$1 ":" \$2, \$6, \$3, \$4, \$5}' >> "\$i".4clumping
  - done
- 
- ~/bin/plink1.9 --bfile QC1 --merge-list join.list --make-bed --out gwide



- for i in 20160 20161 20162 2887 3466 3476
- do
- `~/bin/plink1.9 --bfile ../imputed/gwide --clump "$i".4clumping --clump-p1 1 --clump-p2 1 --clump-r2 .2 --clump-kb 2000 --clump-verbose --clump-annotate A1 A2 Beta --out ind"$i"`
- done
  
- for i in {1..22}
- do
- echo \$i
- `~/bin/plink2 --bfile QCchr"$i" --exclude 3alleles --make-bed --out QC"$i" --threads 5`
- done
  
- for i in 20160 20161 20162 2887 3466 3476
- do
- `grep INDEX ind"$i".clumped | awk '{print $2, $7, $8, $9, $10}' | sed 's/,//g' >> ind"$i".scores`
- done
  
- `~/bin/plink1.9 --bfile ../imputed/gwide --score ind"$i".scores 1 4 5 sum no-mean-imputation include-cnt --out tobacco"$i"`
  
- for i in 20160 20161 20162 2887 3466 3476
- do
- `awk '{ if ($2 <= .01) print $0 }' ind"$i".scores > ind"$i".b`
- `awk '{ if ($2 <= .0001) print $0 }' ind"$i".scores > ind"$i".c`
- `awk '{ if ($2 <= .000001) print $0 }' ind"$i".scores > ind"$i".d`
- `cp ind"$i".scores ind"$i".a`
- done
  
- for i in 20160 20161 20162 2887 3466 3476
- do
- for j in a b c d
- do
- `~/bin/plink1.9 --bfile ../imputed/gwide --score ind"$i"."$j" 1 4 5 sum include-cnt --out "$j"."$i"`
- done
- done

# Today's data

- PRS for Ever Smoked and Pack Years
  - a no threshold
  - b  $\leq .01$
  - c  $\leq .0001$
  - d  $\leq .000001$
- Phenotypes expression of BDNF, CHRNA5 & HTR2A in Cortex

## ◆ RESIDUAL EXPRESSION DATA:

TAB DELIMITED .TXT FILE OF RESIDUAL CORRECTED PROFILES FOR EACH INDIVIDUAL FOR EACH TRANSCRIPT. CORRECTIONS WERE MADE FOR GENDER, APOE STATUS, AGE AT DEATH, CORTICAL REGION, DAY OF EXPRESSION HYBRIDIZATION, INSTITUTE SOURCE OF SAMPLE, POSTMORTEM INTERVAL, AND TRANSCRIPT DETECTION RATE USING R (SEE PUBLICATION). FILE IS IN THE FORMAT OF AN ALTERNATE PHENOTYPES FILE DESCRIBED AT [HTTP://PNGU.MGH.HARVARD.EDU/~PURCELL/PLINK/DATA.SHTML#PHENO](http://PNGU.MGH.HARVARD.EDU/~PURCELL/PLINK/DATA.SHTML#PHENO). NOTE THAT MISSING VALUES (NAN ON OTHER FILES) ARE CODED AS -9.0.

- Covariate AD status & Ancestry MDS