2017 SISG Brisbane Module 10: Statistical & Quantitative Genetics of Disease

Lecture 7

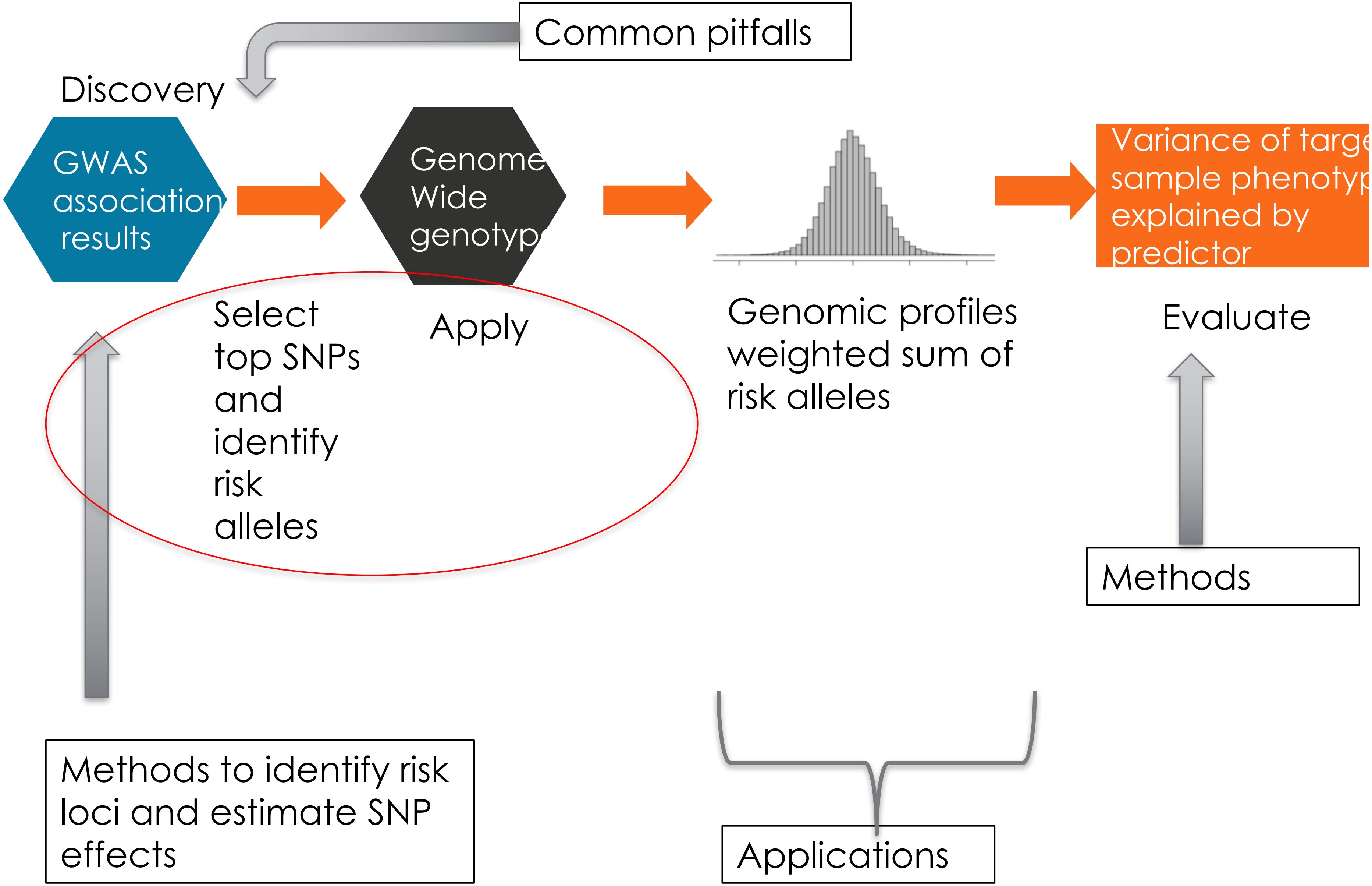
Risk profile scores

Naomi Wray

Aims of Lecture 7

1. Statistics to evaluate risk profile scores
   a. Nagelkerke's R^2
   b. AUC
   c. Decile Odds Ratio
   d. Variance explained on liability scale
   e. Risk stratification
2. Examples of Use of Risk Profile Scores

SNP profiling schematic



Common pitfalls

Discovery

GWAS association results

Genome Wide genotyp

Variance of targe sample phenotyp explained by predictor

Select top SNPs and identify risk alleles

Apply

Genomic profiles weighted sum of risk alleles

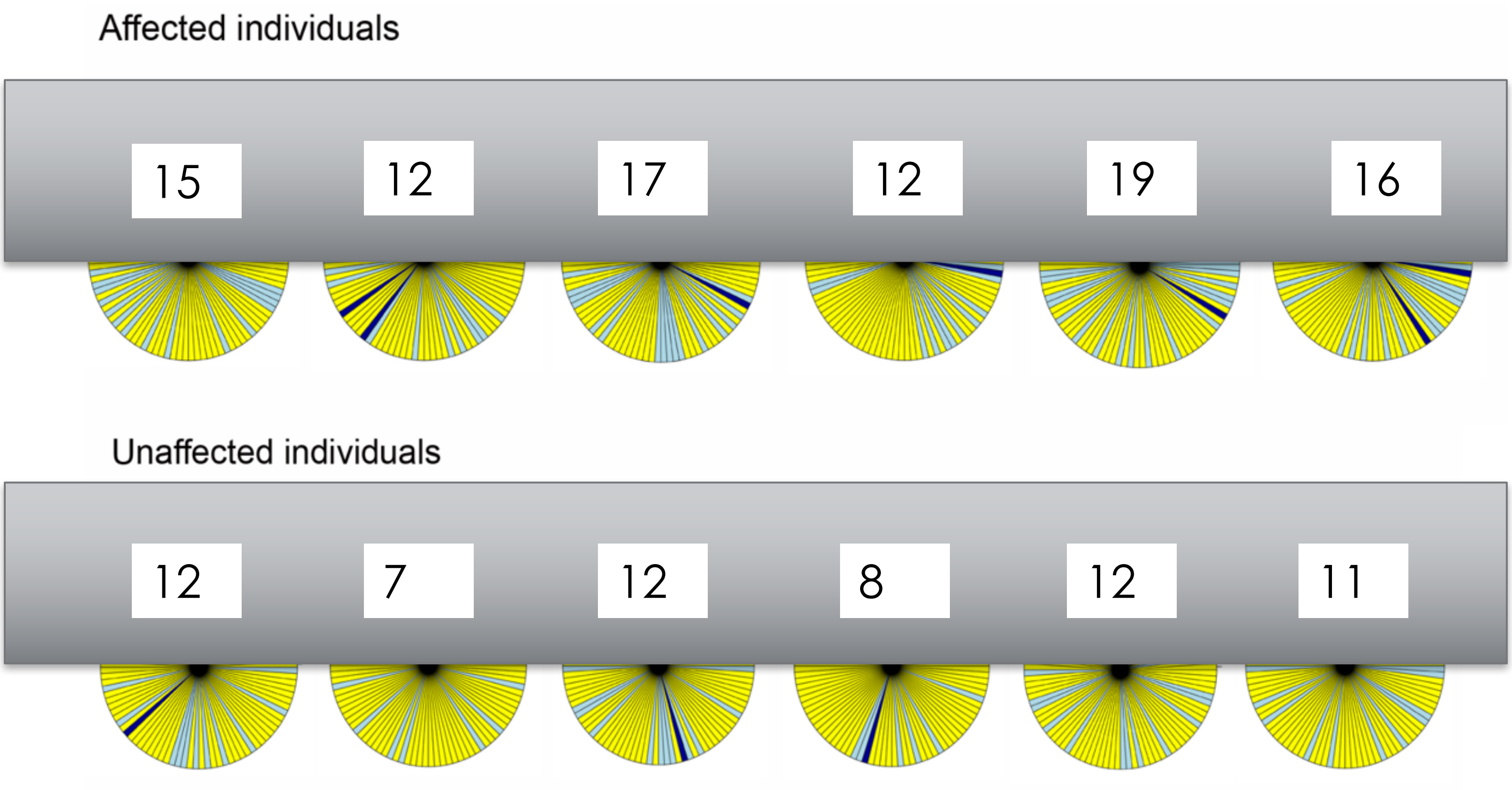
Evaluate

Methods

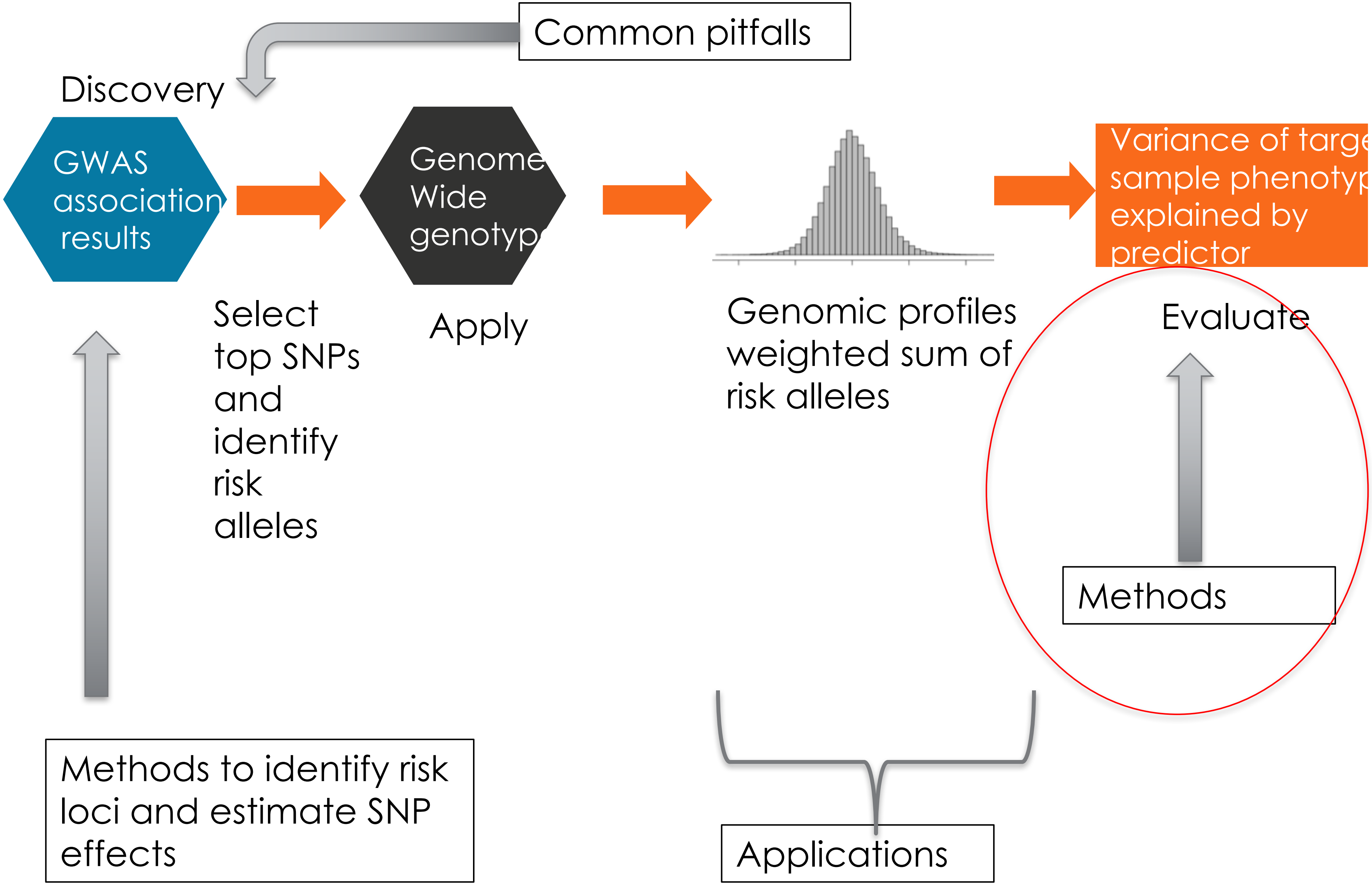
Methods to identify risk loci and estimate SNP effects

Applications

Visualising variation between individuals for common complex genetic diseases



SNP profiling schematic



Common pitfalls

Discovery

GWAS association results

Genome Wide genotyp

Variance of targe sample phenotyp explained by predictor

Select top SNPs and identify risk alleles

Apply

Genomic profiles weighted sum of risk alleles

Evaluate

Methods

Methods to identify risk loci and estimate SNP effects

Applications

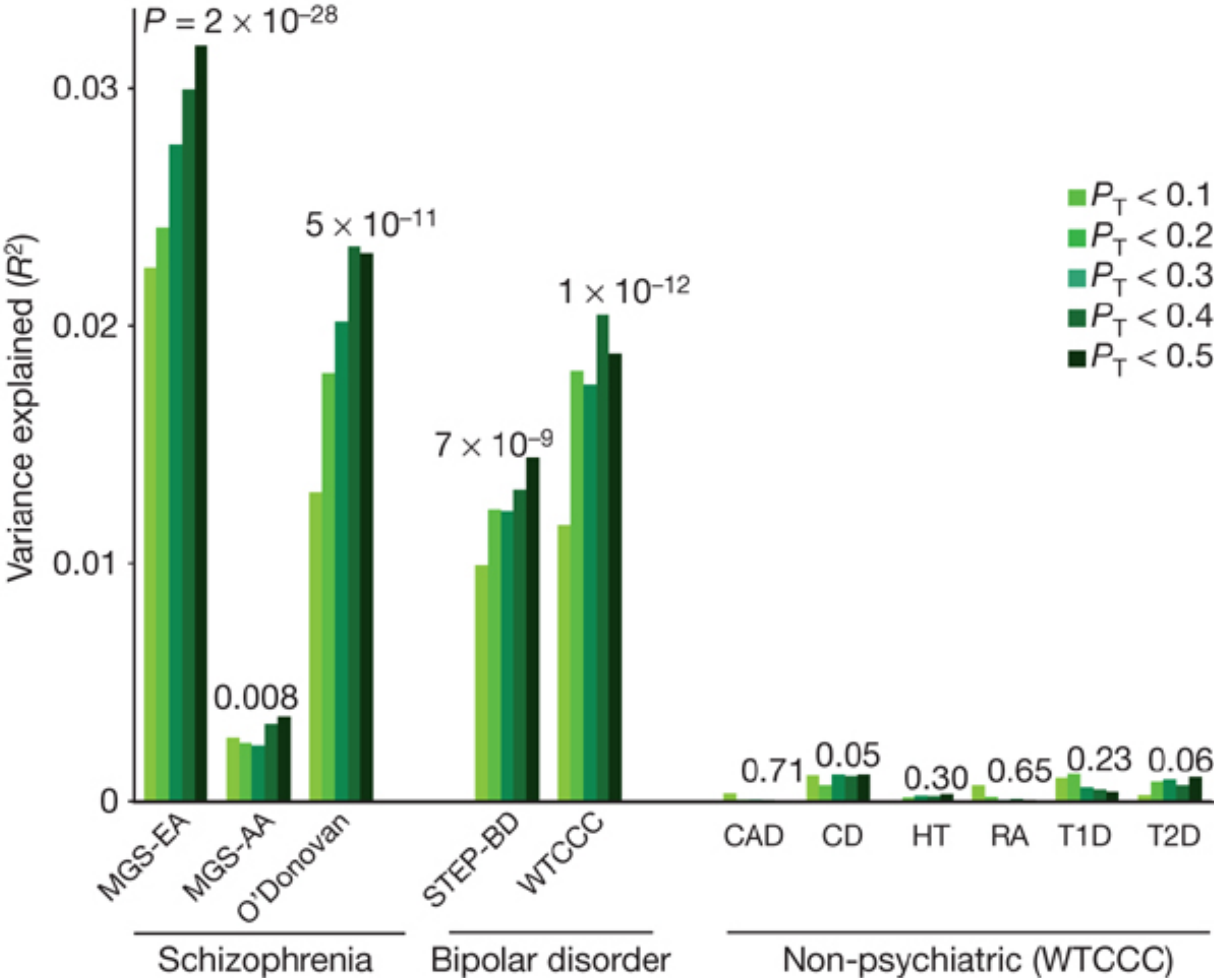
Evaluate efficacy of score predictor

Regression analysis:

- y= phenotype, x = profile score.
- Compare variance explained from the full model (with x) compared to a reduced model (covariates only).
- Check the sign of the regression coefficient to determine if the relationship between y and x is in the expected direction.

- BINARY TRAIT

First Application of Risk Profile Scoring



Purcell / ISC et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder *Nature* 2009

Statistics to evaluate polygenic risk scoring 1.

1. Nagelkerke's R^2
   - Pseudo-R^2 statistic for logistic regression

http://www.ats.ucla.edu/stat/mult_pkg/faq/general/Psuedo_RSquareds.htm

Cox & Snell R^2

$$= 1 - \exp\left(\frac{2}{N}\right)(LogLikelihood\ (Reduced\ model) - LogLikelihood(Full\ model))$$

Full model: y ~ covariates + score Logistic, y= case/control = 1/0

Reduced model: y ~ covariates

N: sample size

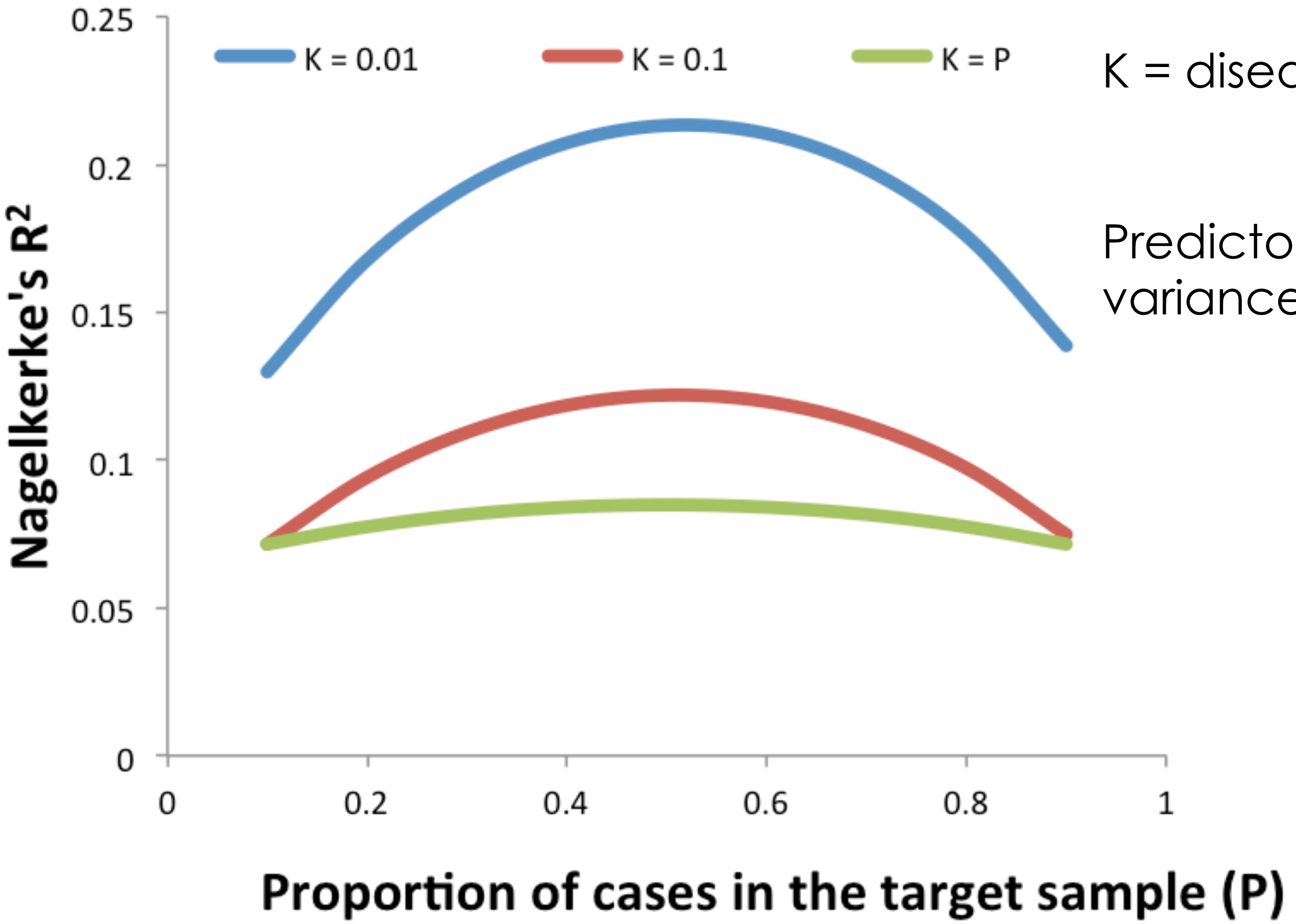
This definition gives R^2 for a quantitative trait.

For a binary trait in logistic regression, C&S R^2 has maximum

$$= 1 - \exp\left(\frac{2}{N}\right)(LogLikelihood\ (Reduced\ model))$$

Nagelkerke's R^2 divides Cox & Snell R^2 by its maximum to give an R^2 with usual properties of between 0 and 1.

Problem with Nagelkerke's R^2



K = disease prevalence

Predictor explains 7% of variance in liability

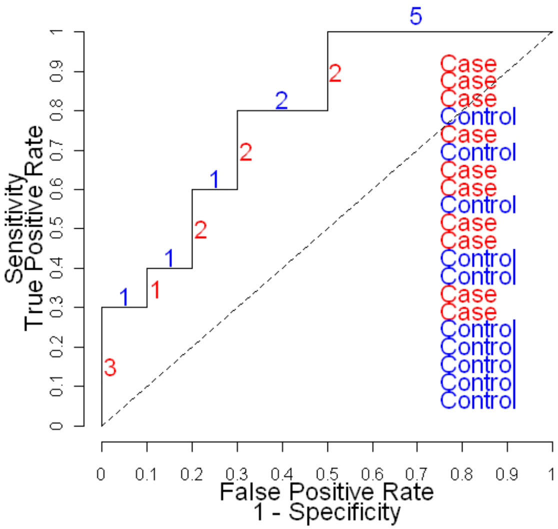
Statistics to evaluate polygenic risk scoring 2.

2. Area Under Receiver Operator Characteristic Curve

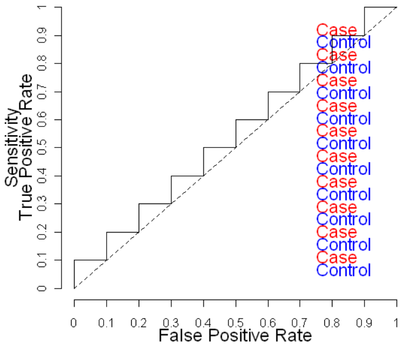
- Well established measure of validity of tests for classifier diseased vs non-diseased individuals
- Nice property – independent to proportion of cases and controls in sample
- Range 0.5 to 1
- 0.5 the score has no predictive value
- Probability that a randomly selected case has a score higher than a randomly selected control

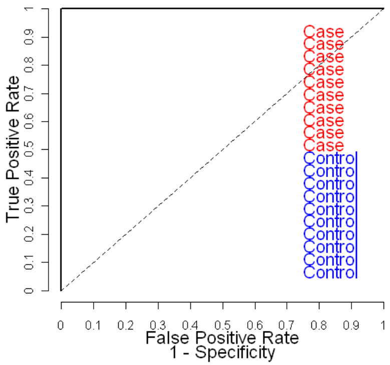
Visualising AUC

- Rank individuals on score from highest ranked to lowest
- Start at origin on graph
- Work through list of ranked individuals
- Move one unit along y-axis if next individual is a case
- Move one unit along x-axis if next individual is a control



AUC = Probability that a randomly selected case has a higher test score than a randomly selected control





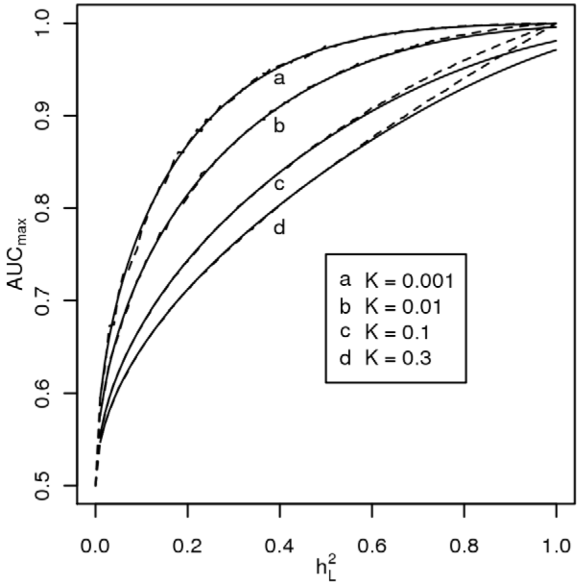
Problem with AUC

Well recognised as a measure of clinical validity

A measure of how well genomic profile predicts yes/no phenotype

But hides the fact that is should be judged as a measure of analytic validity

A measure of how well genomic profile predicts genotype



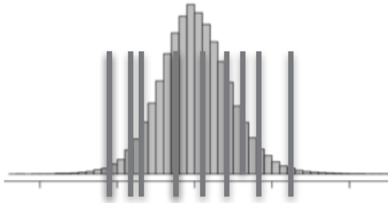
The maximum AUC achievable depends on the heritability of the disease

Many useful properties
Problem is genetic interpretation

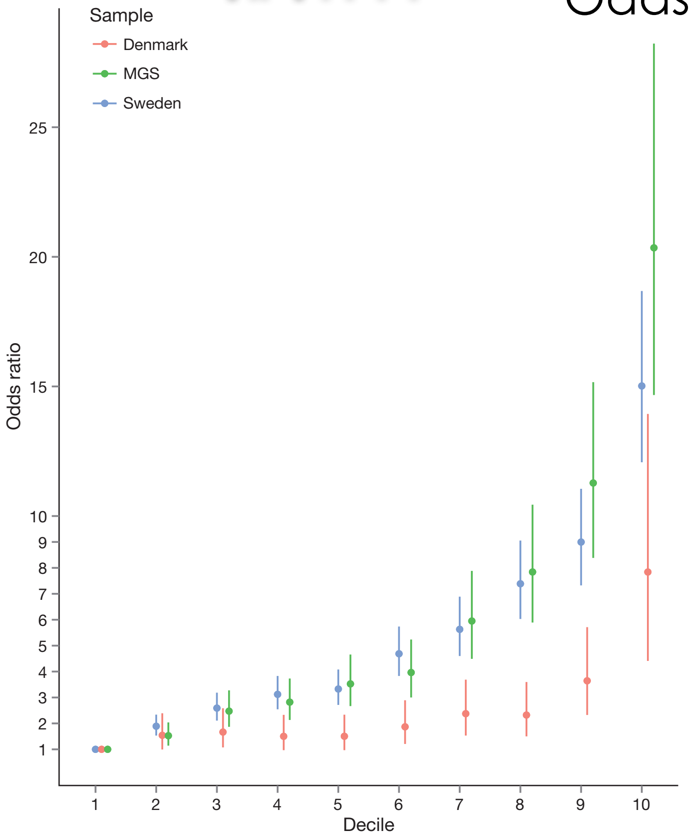
Wray et al (2010) The genetic interpretation of area under the receiver operator characteristic curve in genomic profiling. PLoS Genetics

Statistics to evaluate polygenic risk scoring 3.

3. Odds Ratio



Cut distribution into deciles
Each decile will include both cases and controls
Odds of being a case in each decile
Odds ratio for each decile compared to the 1st decile

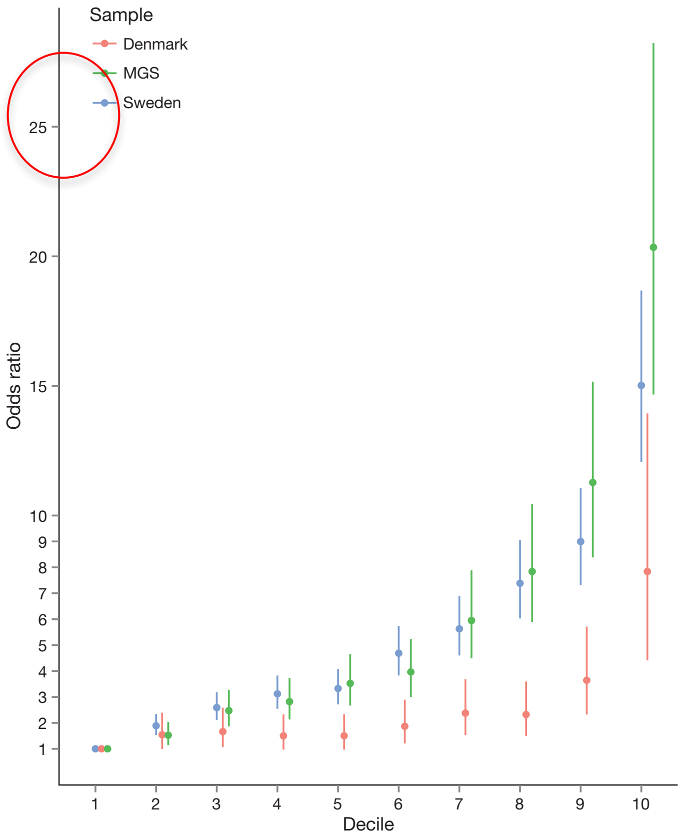


- Good visualisation
- Shows that there could be utility in using high vs low profile risk scores
- But remember case-control samples are 50% cases
- Would look less impressive if a population sample

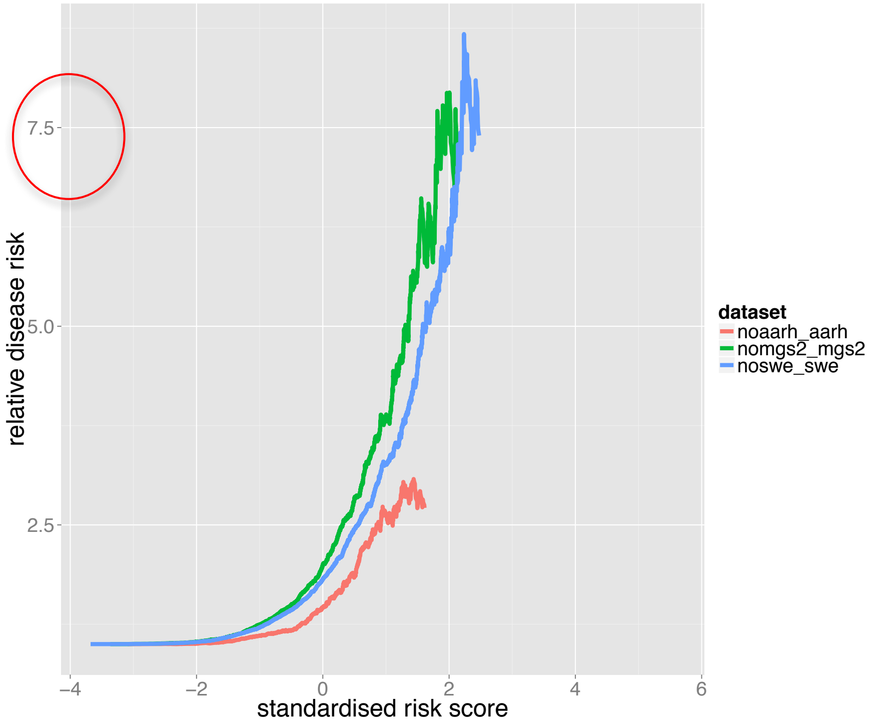
PGC-SCZ 2014 108 loci Nature

Statistics to evaluate polygenic risk scoring 3.

In case control samples



Same data scaled to population risk



Statistics to evaluate polygenic risk scoring 4.

3. R^2 on liability scale

Linear model

Full model: y ~covariates + score   y = case/control = 1/0

Reduced model: y~ covariates

Calculate R^2 attributable to score

If target sample is a population sample i.e. prevalence of cases in sample = prevalence of cases in controls

Then R^2 is a measure of the proportion of variance in case-control status attributable to the genomic risk profile score

= heritability attributable to genomic profile score $h^2_{GRPS-01}$ on the disease scale

Convert to liability scale

$$h^2_{GRPS} = \frac{h^2_{GRPS-01} K(1-K)}{z^2}$$

Lee et al (2012) A better coefficient of determination for genetic profile analysis. Genetic Epidemiology

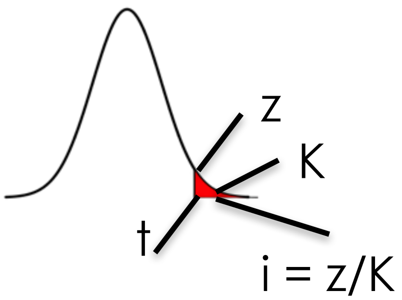
Relationship between heritabilities on disease and liability scales

Consider a linear regression of genetic values on the disease scale (A_{01}) on genetic values on the liability scale (A_L):

$$A_{01} = \mu + bA_L \qquad b = \frac{\mathrm{cov}(A_{01}, A_L)}{\mathrm{var}(A_L)}$$

$$\mathrm{Var}(A_{01}) = b^2\mathrm{Var}(A_L) = \frac{\mathrm{cov}(A_{01}, A_L)^2}{\mathrm{var}(A_L)}$$ by differential calculus normal distribution theory....

$$h^2_{01} = \frac{z^2 h^2_L}{K(1-K)} = \frac{i^2 K h^2_L}{(1-K)}$$



$$h^2_L = \frac{(1-K)h^2_{01}}{i^2 K}$$

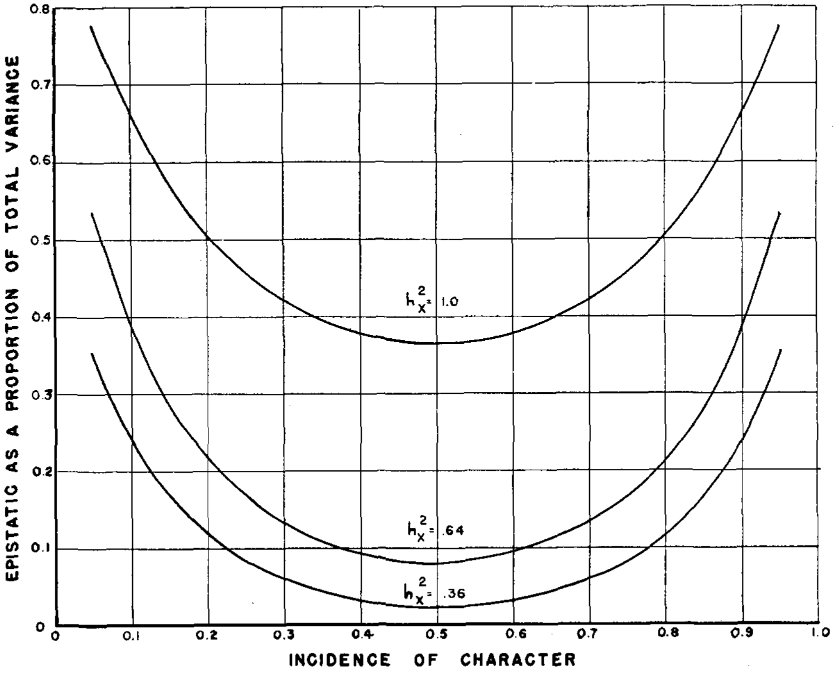
NB Estimates of narrow heritability on observed scale from family data often contaminated by non-additive heritability

Robertson (1950) Appendix of Dempster & Lerner (1950) Heritability of threshold characters. Genetics 35

Relationship between heritability on the disease and liability scales

$$h^2_{01} = \frac{z^2 h^2}{K(1-K)} = \frac{i^2 K h^2}{(1-K)}$$

On the disease scale



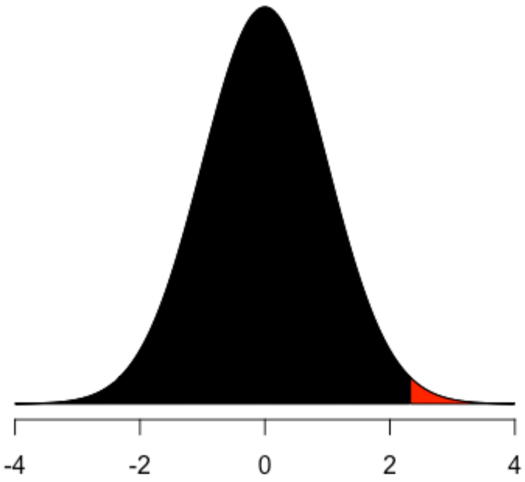
Lines are heritability of liability

= Prevalence

Dempster & Lerner (1950) Appendix by Alan Robertson. Heritability of threshold characters. Genetics 35

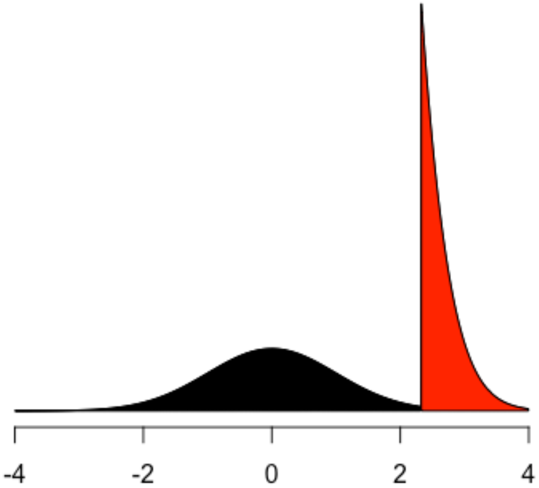
Ascertainment in case-control studies

$\widehat{h^2_{o_{cc}}}$ · Estimate of proportion of variance explained by SNP between cases and controls



Unaffected (1-K)

Affected (K)



Control (1-P)

Case (P)

$$h_l^2 = h_o^2 \frac{K(1-K)}{z^2}$$

Robertson (1950)
Appendix of Dempster and Lerner (1950)
See Lecture 1

$$h_l^2 = \widehat{h^2_{o_{cc}}} \frac{K(1-K)}{z^2} \frac{K(1-K)}{P(1-P)}$$

Lee et al (2011)AJHG
Zhou & Stephens (2013) Polygenic Modeling with Bayesian Sparse Linear Mixed Models PLoSG Text S3
Golan et al (2014) Measuring missing heritability: Inferring the contribution of common variants PNAS

Statistics to evaluate polygenic risk scoring 4 cont.

3. R^2 on liability scale cont.

If target sample is a case-control sample

i.e. prevalence of cases in sample >> prevalence of cases in controls

Then R^2 is a measure of the proportion of variance in case-control status attributable to the genomic risk profile score

= heritability attributable to genomic profile score on the case-control scale

$$h^2_{GRS-CC}$$

Convert to the liability scale

$$h^2_{GRS} = \frac{h^2_{GRS-CC}C}{1 + h^2_{GRS-CC}C}$$

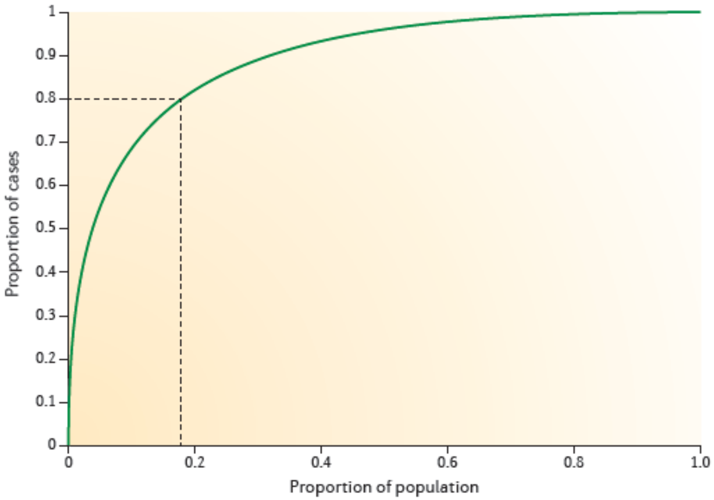
Where C is:

$$C = \frac{K(1-K)}{z^2}\frac{K(1-K)}{P(1-P)}$$

h^2_{GRS} is on the same scale as heritability estimated from family studies and GREML SNP-chip heritability

Lee et al (2012) A better coefficient of determination for genetic profile analysis. Genetic Epidemiology

Statistics to evaluate polygenic risk scoring
5.Stratification & health economics



Population risk of 1%

80% of cases in
top 18% of genetic risk

For every 1,000 people treated with intervention could "save" 10
Treat only 18% = 180 and "save" 8 (4%)

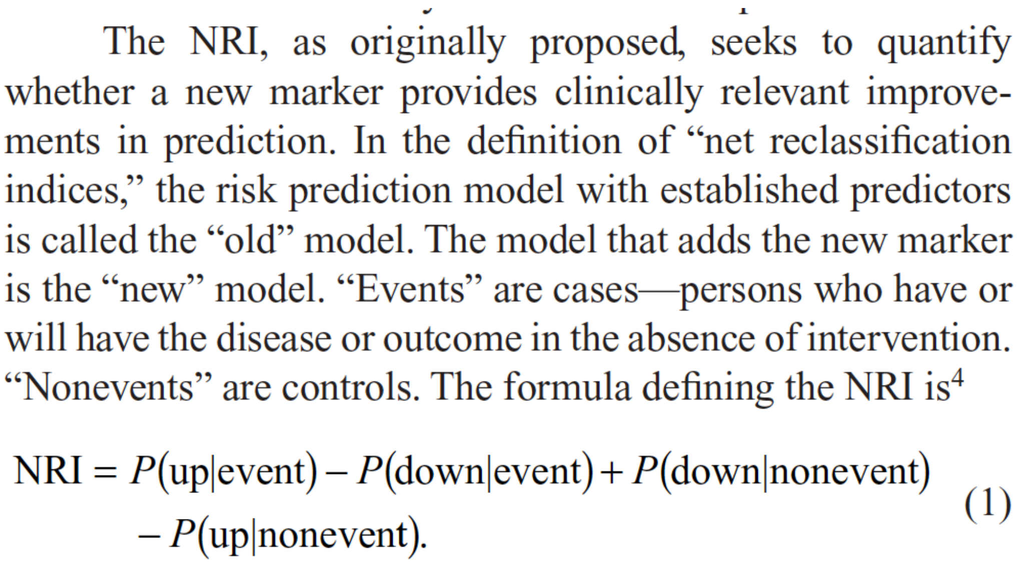
Number of people treated to save 1 reduced from 100 to 22.5

Polychronakos & Li NRG (2011) Understanding Type I Diabetes through genetics. Nat Rev Genetics

Improvement between predictors

Difference in AUC

Net reclassification index



The NRI, as originally proposed, seeks to quantify whether a new marker provides clinically relevant improvements in prediction. In the definition of “net reclassification indices,” the risk prediction model with established predictors is called the “old” model. The model that adds the new marker is the “new” model. “Events” are cases—persons who have or will have the disease or outcome in the absence of intervention. “Nonevents” are controls. The formula defining the NRI is[4]

$$\mathrm{NRI} = P(\mathrm{up}|\mathrm{event}) - P(\mathrm{down}|\mathrm{event}) + P(\mathrm{down}|\mathrm{nonevent}) - P(\mathrm{up}|\mathrm{nonevent}). \tag{1}$$

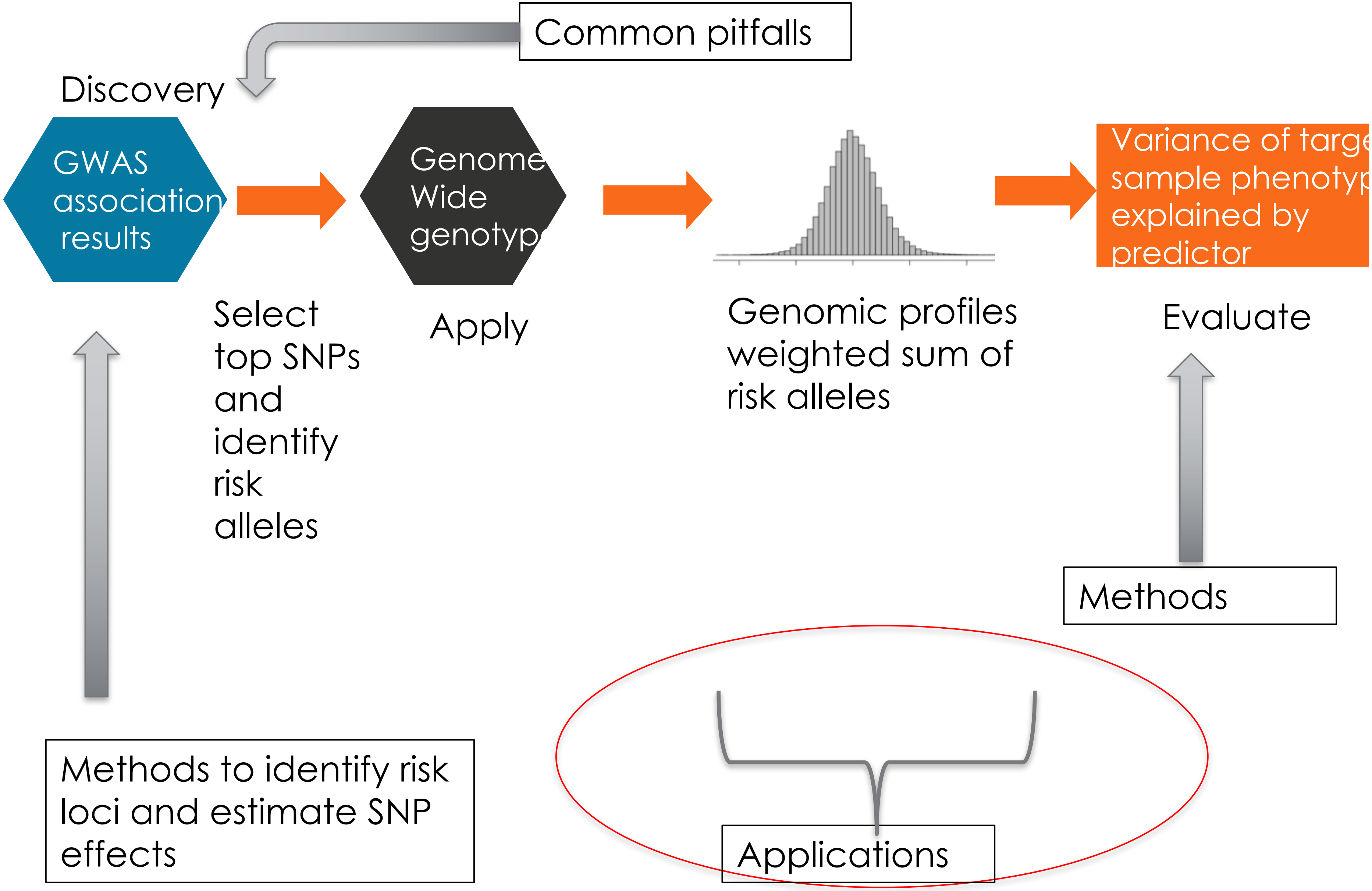
$$\mathrm{NRI}_{\mathrm{e}} = P(\mathrm{up}|\mathrm{event}) - P(\mathrm{down}|\mathrm{event})$$

$$\mathrm{NRI}_{\mathrm{ne}} = P(\mathrm{down}|\mathrm{nonevent}) - P(\mathrm{up}|\mathrm{nonevent})$$

Topic of debate
Needs more research

Kerr et al (2014) NRI for evaluating risk prediction indices.

SNP profiling schematic



Common pitfalls

Discovery

GWAS association results

Genome Wide genotyp

Variance of targe sample phenotyp explained by predictor

Select top SNPs and identify risk alleles

Apply

Genomic profiles weighted sum of risk alleles

Evaluate

Methods

Methods to identify risk loci and estimate SNP effects

Applications

Applications of polygenic Risk Profile Scoring

Discovery & Target samples could be:

A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations

B. Different disorders - demonstrates genetic overlap between disorders

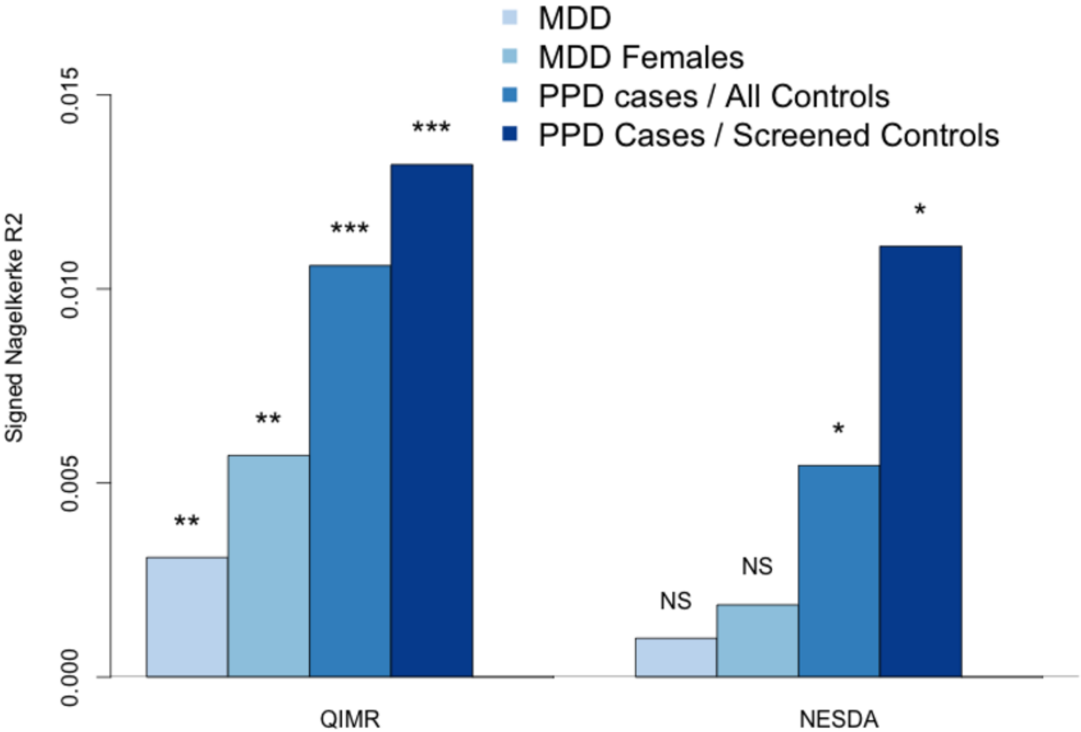
C. Target samples are disorder subtypes

- investigates genetic genetic heterogeneity
- think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

Example Disorder Sub-types. Discovery: PGC-BPD Target: Postnatal depression in MDD

Postnatal depression – a more homogeneous subtype of depression?

Female only
Same bio-social stressor



Enda Byrne
Tania Carillo-Roa
Samantha Meltzer-Brody
Nick Martin
Brenda Penninx

NB. Null result in the ALSPAC community sample measured for PND but not MDD

Byrne et al (2014) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Archives of Women's Health. In press

Applications of polygenic Risk Profile Scoring

Discovery & Target samples could be:

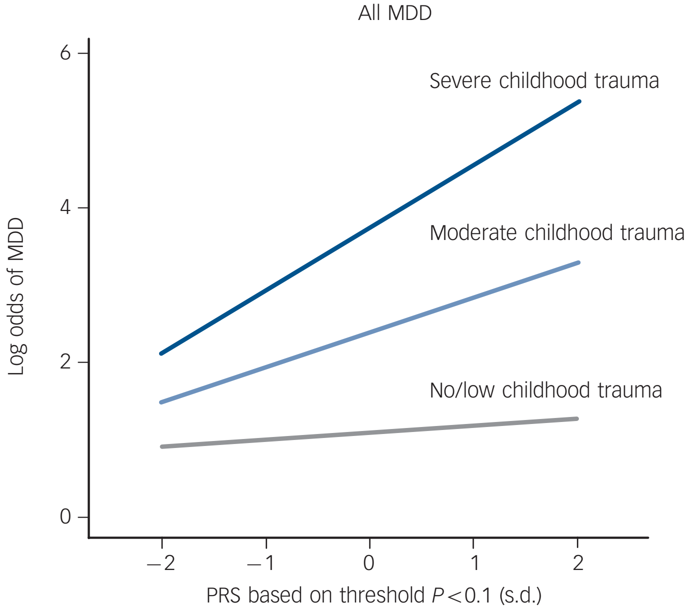
A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations

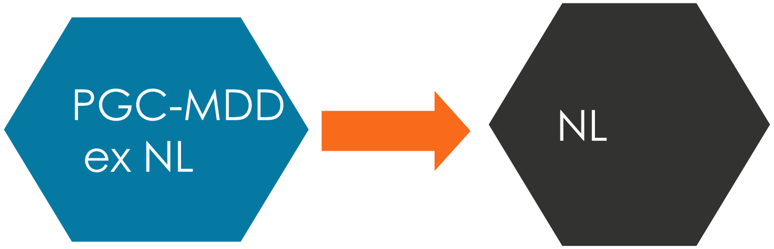
B. Different disorders - demonstrates genetic overlap between disorders

C. Target samples are disorder subtypes
   - investigates genetic genetic heterogeneity
   - think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

D. Target samples have the same disease as the discovery sample and have environmental risk factors recorded
   - investigate GxE
   - think carefully about how the environmental risk factor is represented in the Discovery sample

Application of Polygenic Risk Profiling Scores to investigate GxE, e.g., depression and childhood trauma




PGC-MDD ex NL

NL

Peyrot et al (2014) Effect of polygenic risk scores on depression in childhood trauma Biol Psychiatry

Applications of polygenic Risk Profile Scoring

Discovery & Target samples could be:

A. Same Disorder - demonstrates polygenicity even in absence of genome-wide significant SNP associations

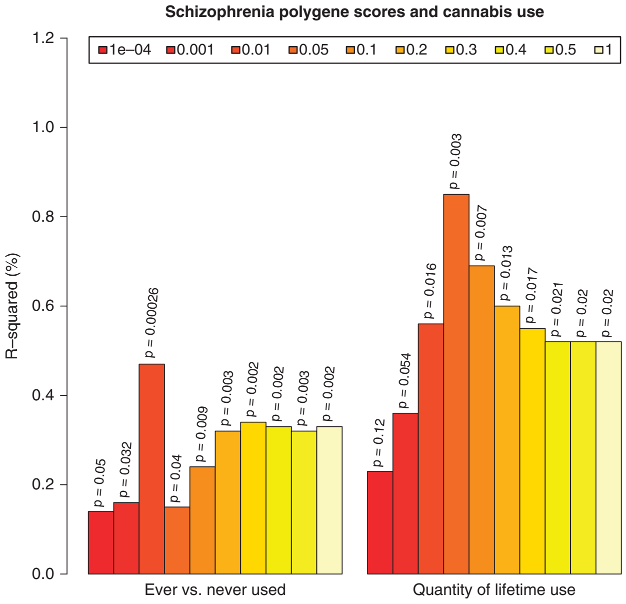
B. Different disorders - demonstrates genetic overlap between disorders

A. Target samples are disorder subtypes
  - investigates genetic genetic heterogeneity
  - think carefully about how the heterogeneity is represented in the Discovery sample if Target and Discovery are the same disease

D. Target samples have the same disease as the discovery sample and have environmental risk factors recorded
  - investigate GxE
  - think carefully about how the environmental risk factor is represented in the Discovery sample

E. Target samples are recorded for an environmental risk factor
  - insight into GxE

Example: E in target sample
Discovery: schizophrenia
Target: Cannabis use



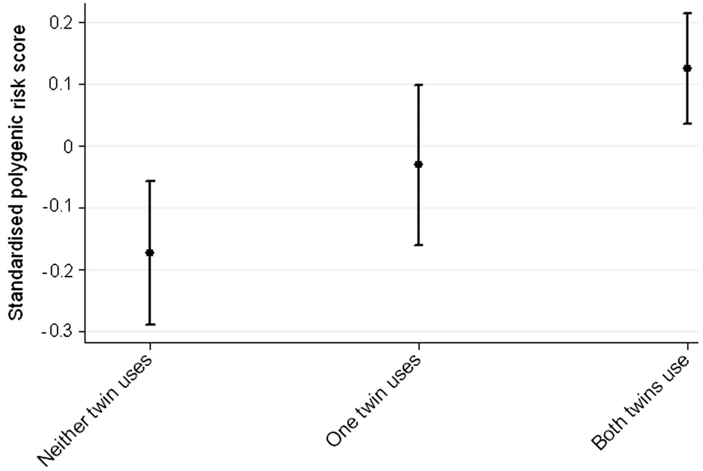
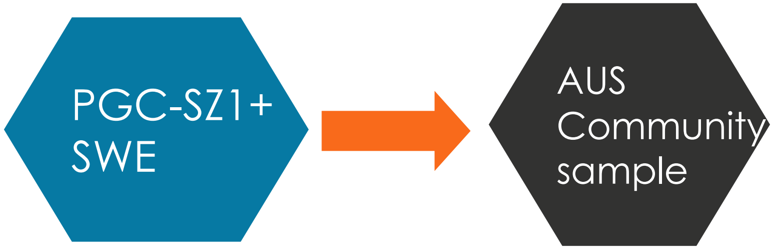


Figure 2. Mean standardized polygenic risk scores for pairs of twins when neither ($n = 272$), one ($n = 273$) or both twins ($n = 445$) had reported use of cannabis. An ordinal regression reported a significant association ($P = 0.001$).



PGC-SZ1+ SWE

AUS Community sample

Power et al (2014) Effect of polygenic risk scores on depression in childhood trauma Mol Psychiatry

Factors affecting accuracy of risk prediction

Genetic architecture of the trait – unknown

- Number, frequency, effect size
- How well marker effects are correlated with causal variants (LD)

Sample size of discovery sample – maximise

- How well marker effects are estimated

Sample size of target sample – be sufficiently large (once achieved not so much gained by increasing further)

- Precision of estimation of R^2

Number of SNPs in GWAS panel

P-value thresholds to select SNPs predictor/ Method to estimate SNP effects

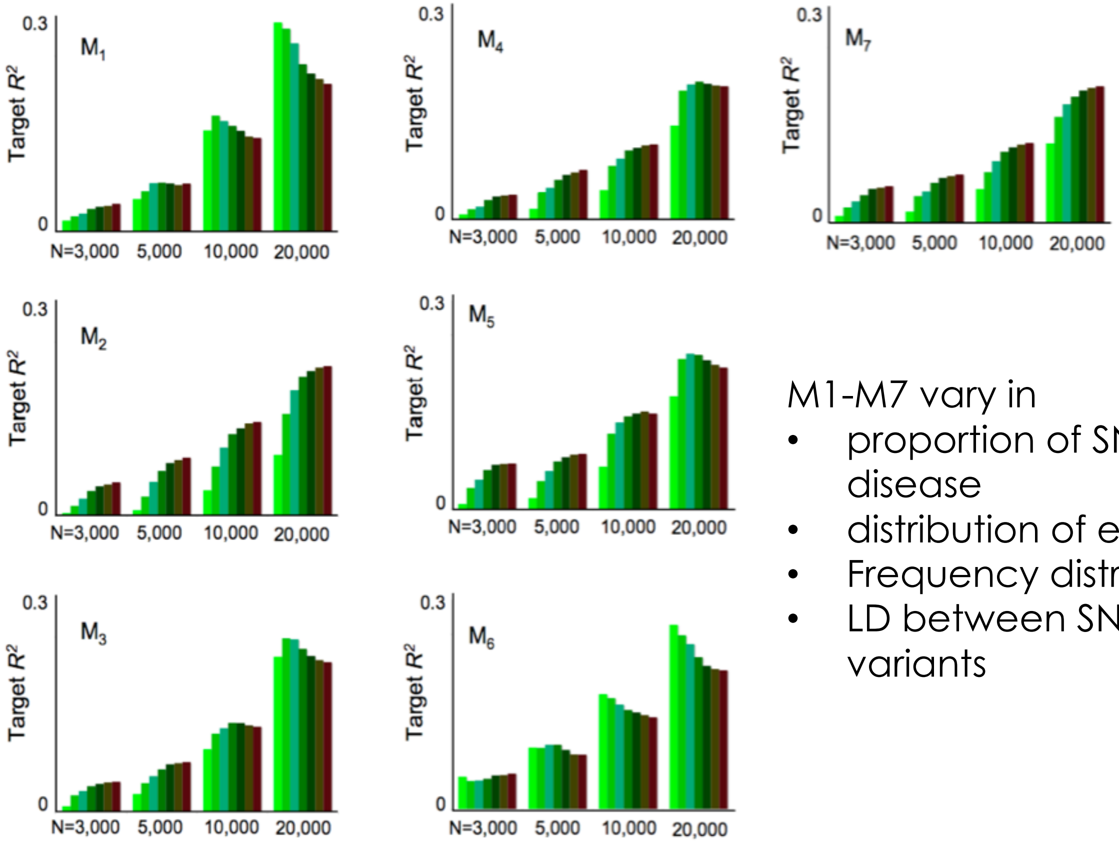
Disease lifetime risk and case/control sampling fractions

Dudbridge (2013) Power and predictive accuracy of polygenic risk scores. PLoS Genetics

Wray et al (2014) Polygenic methods and their application to psychiatric traits. Journal of Child Psychology & Psychiatry (in press)

Simulation study demonstrating the impact of sample size and genetic architecture on profile scoring

Figure S8: *Impact of increasing sample size on score analysis.*



M1-M7 vary in

- proportion of SNPs associated in disease
- distribution of effect sizes
- Frequency distribution
- LD between SNPs and causal variants

Purcell et al (2009) Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. Nature

Sampling error in the polygenic risk score

- The weights in the risk score must be estimated from finite discovery sample data: we have the *estimated* risk score

$$\hat{S} = \sum_i \hat{\beta}_i x_i$$

$$\operatorname{var}(\hat{S}) = \operatorname{var} \sum_i \hat{\beta}_i x_i = \sum_i \operatorname{var}(\hat{\beta}_i)$$

- The more SNPs in the score:
  - The more variation we could explain ☺
  - The greater its sampling error ☹
- A trade-off

Prediction – errors in estimating single SNP effect

$$y_i = bx_i + e_i$$

$$\hat{y} = \hat{b}x_i$$

$$\hat{R}^2_{y,\hat{y}} = \mathrm{cov}(y,\hat{y})^2 / \{\mathrm{var}(y)\,\mathrm{var}(\hat{y})\}$$

$$E[\mathrm{cov}(y,\hat{y})] = E[\mathrm{cov}(xb, x\hat{b})] = \mathrm{var}(x_i)E(\hat{b})b$$

$$= \mathrm{var}(x)b^2$$

$$E[\mathrm{var}(\hat{y})] = E[\mathrm{var}(x\hat{b})] = \mathrm{var}(x)E[\hat{b}^2]$$

$$= \mathrm{var}(x)[b^2 + \mathrm{var}(\hat{b})] \approx \mathrm{var}(x)b^2 + \mathrm{var}(x)\,\mathrm{var}(y) / [N\,\mathrm{var}(x)]$$

$$= \mathrm{var}(x)b^2 + \mathrm{var}(y) / N$$

$$E(\hat{R}^2_{y,\hat{y}}) \approx R^2_{SNP} / [1 + 1 / \{NR^2_{SNP}\}]$$

Prediction – errors in estimating SNP effects

with *m* causal variants together explain h^2 proportion of variance

$$E(\hat{R}^2_{y,\hat{y}}) \approx h^2 / [1 + m / \{Nh^2\}]$$

even if we knew all m causal variants but needed to estimate their effect sizes then the variance explained by the predictor is less than the variance explained by the causal variants in the population.

A perfect predictor of genetic component can be a lousy predictor of a phenotype

The regression R^2 has a maximum that depends on heritability (or in this context variance explained by all SNPs, SNP-heritability)

Daetwyler et al. 2008, PLoS Genetics; Visscher, Yang, Goddard 2010, Twin Research Human Genetics 2010

Disease

Several things to take account of compared to quantitative traits:

- Binary trait (disease prevalence K)
- Over-ascertainment of cases (proportion of cases P)

$$r^2_{u,\hat{u}} = \frac{h^2 z^2}{h^2 z^2 + (K(1-K))^2/(\tau P(1-P))}$$

h^2 proportion of variance of case control status attributable to predictor

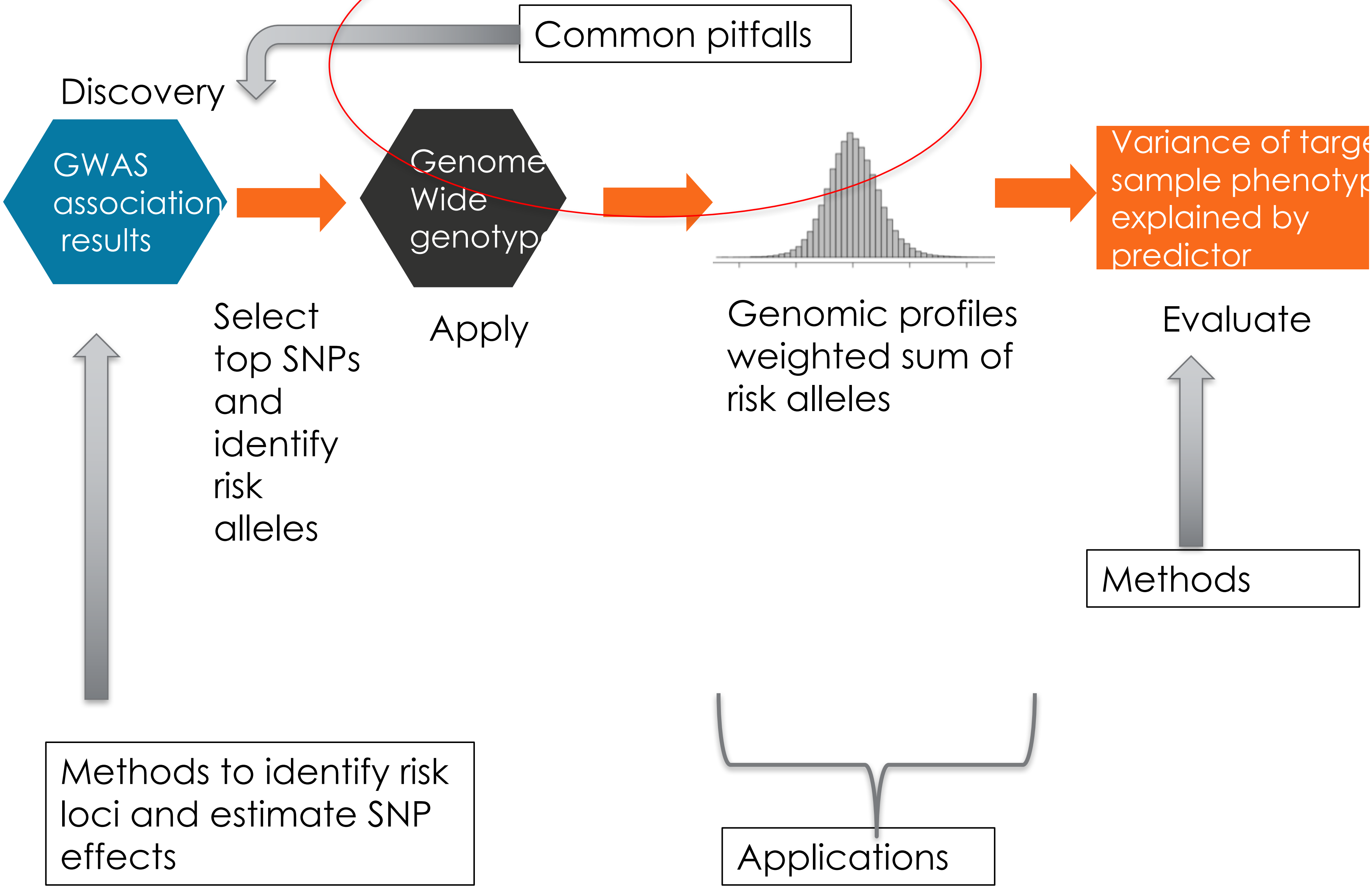
z height of normal curve at K

τ = M/N

M is the number of markers, N is the discovery sample size

Lee & Wray (2013) Novel genetic analysis for case-control GAS: quantification of power and genomic prediction accuracy. PLoS One

SNP profiling schematic



Common pitfalls

Discovery

GWAS association results

Select top SNPs and identify risk alleles

Genome Wide genotyp

Apply

Genomic profiles weighted sum of risk alleles

Variance of targe sample phenotyp explained by predictor

Evaluate

Methods

Methods to identify risk loci and estimate SNP effects

Applications

Pitfall 1: No target (=validation) sample

- Report R^2 or AUC from discovery sample only
- Small n large p problem
- Even under null can get high R^2 within discovery sample when p >> n

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

Variance explained by a predictor under the null hypothesis

y = Σb*x + e

m markers, sample size *N*

All b = 0, ie null hypothesis

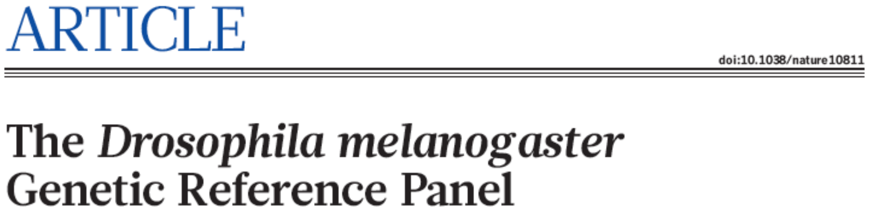
Multiple linear regression of y on m markers

E(R^2) = m/N {strictly $m/(N-1)$}

→ Variation “explained” by chance

Wishart, 1931

Selection bias



ARTICLE

doi:10.1038/nature10811

The *Drosophila melanogaster* Genetic Reference Panel

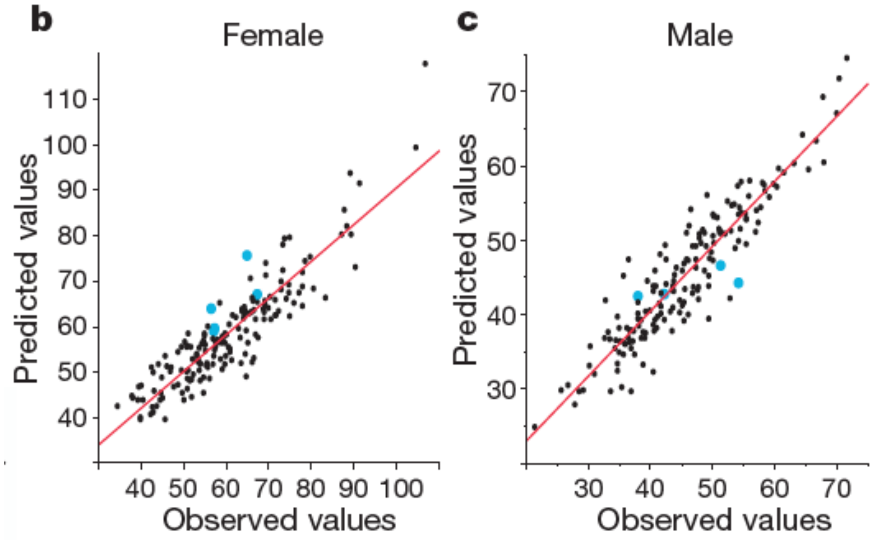
Select m 'best' markers out of M in total

'Prediction' in same sample

~15 best markers selected from 2.5 million markers

$E(R^2) >> m/N$

→ Lots of variation explained by chance



Pitfall 2: Overlapping Discovery & Target Sample

- Overlapping discovery & target samples
- Greater similarity between discovery & target samples than discovery & true validation samples
  - E.g. cross-validation samples
  - Not a pitfall, as such, but to be aware

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

$\hat{b}$ estimated in discovery sample and applied to target sample

$$\operatorname{cov}(\hat{y}_i, y_i) = \operatorname{cov}\{\sum_{j=1}^{m}(x_{ij}\hat{b}_j), \sum_{j=1}^{m} x_{ij}b_j + e_i\}$$

$$= \sum_{j=1}^{m} \operatorname{var}(x_{ij})\hat{b}_j b_j + \sum_{j=1}^{m} x_{ij} \operatorname{cov}(\hat{b}_j, e_i)$$

If b estimated from the same data in which prediction is made, then the second term is non-zero

Pitfall 3: Less obvious non-independence

- Cross-validation but select associated SNPs from total sample

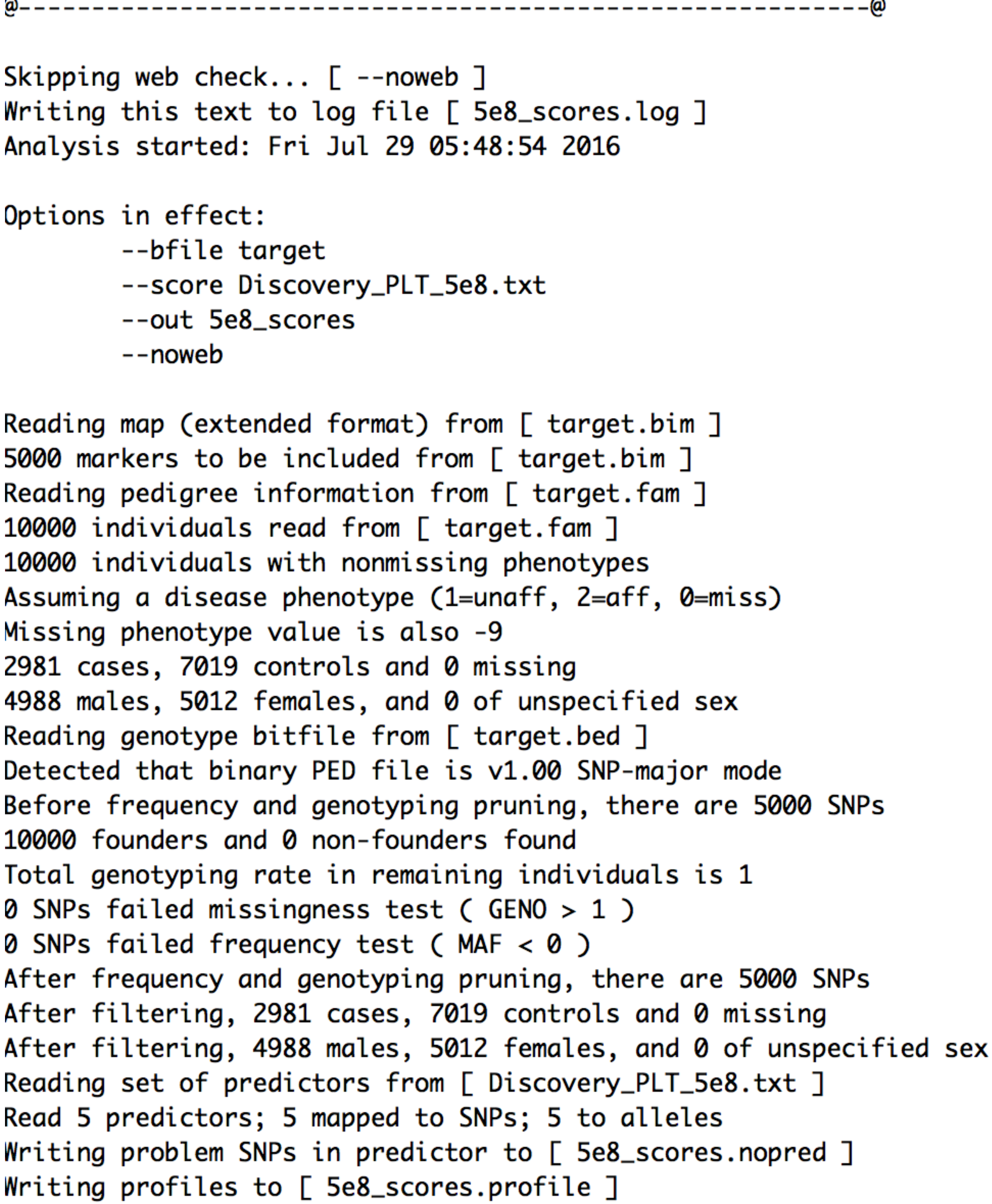
- Select SNPs in discovery sample, for those SNPs re-estimate effects in the target sample

Wray, Yang, Hayes, Price, Goddard, Visscher (2013) Pitfalls of predicting complex traits from SNPs. Nat Rev Genetics

Practical

- Have Folder Practical 7
  - Practical7_ProfileScoring.R
  - Plink binary file for executing plink
  - target.bim, target.bam, target.bed
  - = PLINK genotype files –binary cant open (simulated)
  - See http://pngu.mgh.harvard.edu/~purcell/plink/binary.shtml
  - Discovery_PLT_x.txt x= pvalue cut-offs from Discovery GWAS

- Open R script
- Set working directory
- Run PLINK from within R to generate scores per person in the target sample based on weights from Discovery sample



```
@----------------------------------------------------------@

Skipping web check... [ --noweb ]
Writing this text to log file [ 5e8_scores.log ]
Analysis started: Fri Jul 29 05:48:54 2016

Options in effect:
        --bfile target
        --score Discovery_PLT_5e8.txt
        --out 5e8_scores
        --noweb

Reading map (extended format) from [ target.bim ]
5000 markers to be included from [ target.bim ]
Reading pedigree information from [ target.fam ]
10000 individuals read from [ target.fam ]
10000 individuals with nonmissing phenotypes
Assuming a disease phenotype (1=unaff, 2=aff, 0=miss)
Missing phenotype value is also -9
2981 cases, 7019 controls and 0 missing
4988 males, 5012 females, and 0 of unspecified sex
Reading genotype bitfile from [ target.bed ]
Detected that binary PED file is v1.00 SNP-major mode
Before frequency and genotyping pruning, there are 5000 SNPs
10000 founders and 0 non-founders found
Total genotyping rate in remaining individuals is 1
0 SNPs failed missingness test ( GENO > 1 )
0 SNPs failed frequency test ( MAF < 0 )
After frequency and genotyping pruning, there are 5000 SNPs
After filtering, 2981 cases, 7019 controls and 0 missing
After filtering, 4988 males, 5012 females, and 0 of unspecified sex
Reading set of predictors from [ Discovery_PLT_5e8.txt ]
Read 5 predictors; 5 mapped to SNPs; 5 to alleles
Writing problem SNPs in predictor to [ 5e8_scores.nopred ]
Writing profiles to [ 5e8_scores.profile ]
```