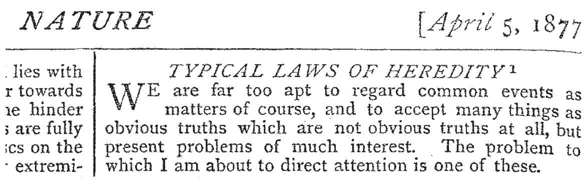
Non-Parametric Polygenic Risk Prediction

Shamil Sunyaev

Department of Biomedical Informatics
Harvard Medical School

Division of Genetics
Department of Medicine
Brigham and Women's Hospital / Harvard Medical School

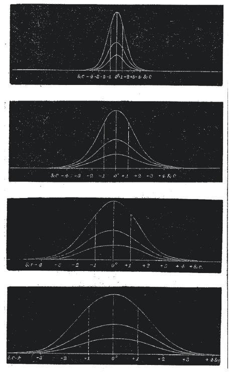


NATURE [*April* 5, 1877

. lies with
r towards
ıe hinder
; are fully
ıcs on the
· extremi-

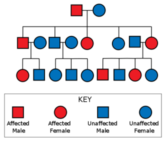
TYPICAL LAWS OF HEREDITY[1]

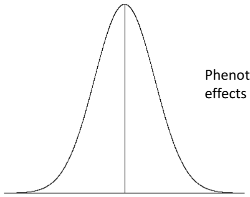
WE are far too apt to regard common events as matters of course, and to accept many things as obvious truths which are not obvious truths at all, but present problems of much interest. The problem to which I am about to direct attention is one of these.



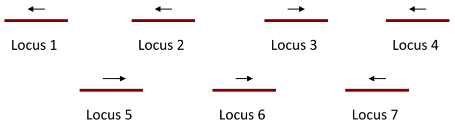
Quantitative Trait Loci (QTLs)

Inheritance at each locus is Mendelian. Loci are independent

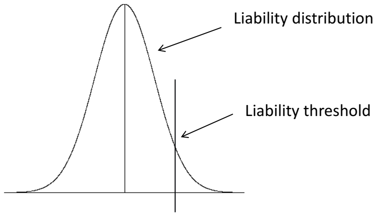


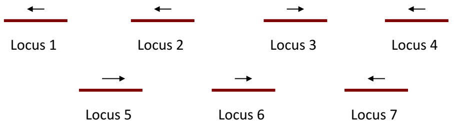


Phenotype is additive over locus effects -> normal distribution

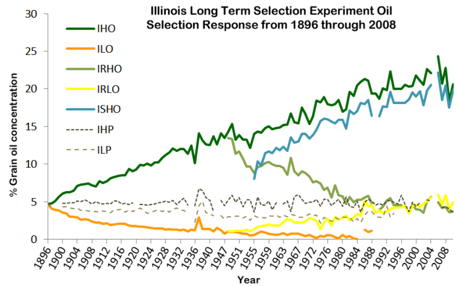


Binary traits such as diseases

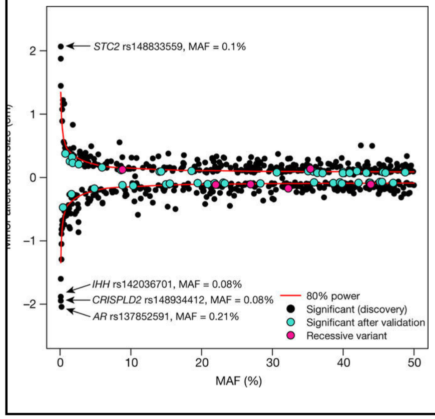


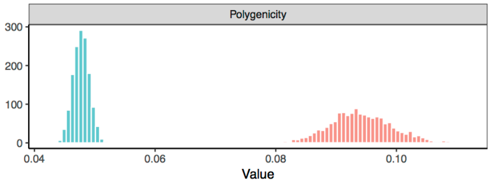


Early evidence of high polygenicity of complex traits



Evidence in favor of the highly polygenic model

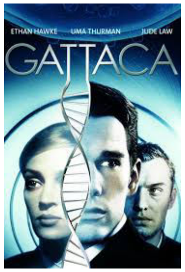




Effect sizes of individual variants are very small

- Genotype at a single locus carries very little information about phenotype.
- It does not mean that one cannot predict phenotype from genotype.
- Accuracy (***r***^{2}) of an ideal genetic predictor equals heritability.

Genetic risk prediction



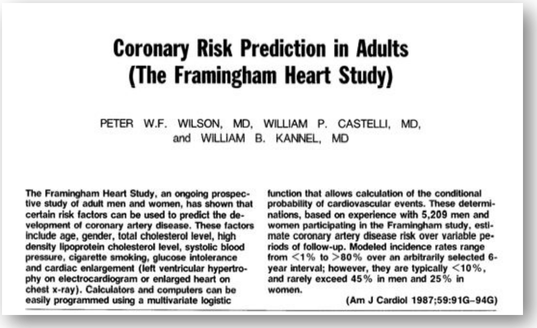
Genotype of an individual → Life-time risk of genetic disorders

(Common SNPs) (Common complex genetic disorders)

Effect sizes of individual variants are very small

- Genotype at a single locus carries very little information about phenotype.
- I does not mean that one cannot predict phenotype from genotype.
- Accuracy (***r*²**) of an ideal genetic predictor equals heritability.

Measuring risk of myocardial infarction



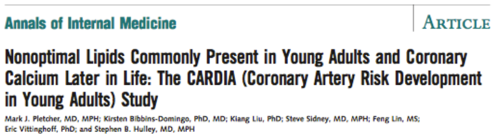
Coronary Risk Prediction in Adults (The Framingham Heart Study)

PETER W.F. WILSON, MD, WILLIAM P. CASTELLI, MD, and WILLIAM B. KANNEL, MD

The Framingham Heart Study, an ongoing prospective study of adult men and women, has shown that certain risk factors can be used to predict the development of coronary artery disease. These factors include age, gender, total cholesterol level, high density lipoprotein cholesterol level, systolic blood pressure, cigarette smoking, glucose intolerance and cardiac enlargement (left ventricular hypertrophy on electrocardiogram or enlarged heart on chest x-ray). Calculators and computers can be easily programmed using a multivariate logistic function that allows calculation of the conditional probability of cardiovascular events. These determinations, based on experience with 5,209 men and women participating in the Framingham study, estimate coronary artery disease risk over variable periods of follow-up. Modeled incidence rates range from <1% to >80% over an arbitrarily selected 6-year interval; however, they are typically <10%, and rarely exceed 45% in men and 25% in women.

(Am J Cardiol 1987;59:91G–94G)

LDL levels and risk of disease



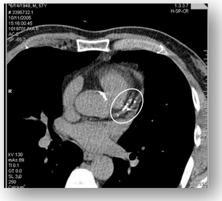
Annals of Internal Medicine | ARTICLE

Nonoptimal Lipids Commonly Present in Young Adults and Coronary Calcium Later in Life: The CARDIA (Coronary Artery Risk Development in Young Adults) Study

Mark J. Pletcher, MD, MPH; Kirsten Bibbins-Domingo, PhD, MD; Kiang Liu, PhD; Steve Sidney, MD, MPH; Feng Lin, MS; Eric Vittinghoff, PhD; and Stephen B. Hulley, MD, MPH

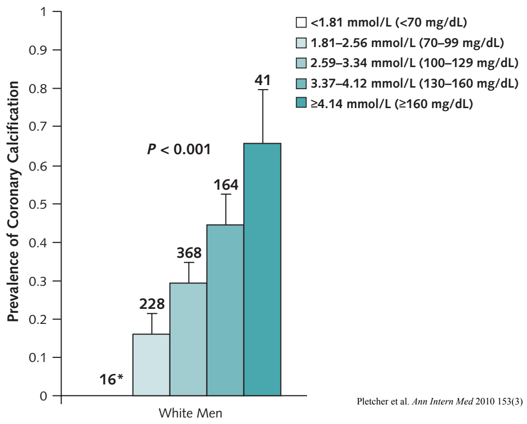
~3500 subjects < 35 years old

15-20 years



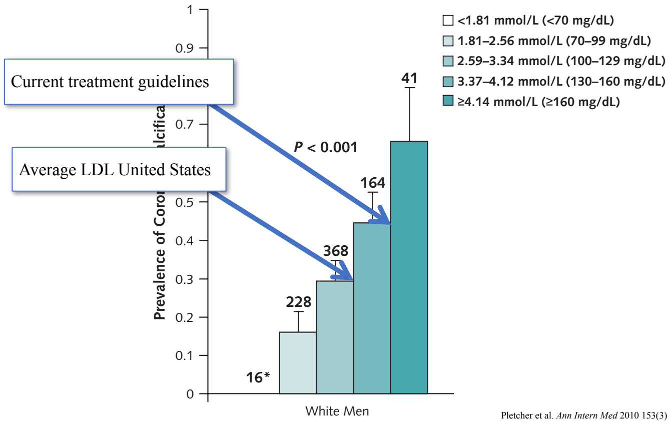
Piers *et al. BMC Cardiovascular Disorders* 2008 **8**:38

LDL levels and risk of disease



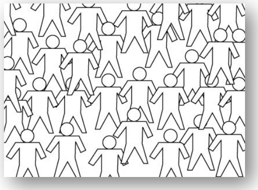
Pletcher et al. *Ann Intern Med* 2010 153(3)

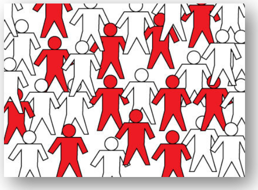
LDL levels and risk of disease



Pletcher et al. *Ann Intern Med* 2010 153(3)

Selecting populations for treatment

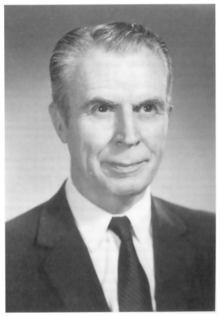




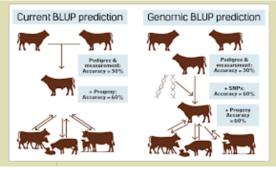
Why estimate genetic risk?

- An estimate of the long-term risk at birth
- Genetic risk can be combined together with biomarkers and clinical features
- Genetics explains about 50% of risk. One cannot predict risk any better than that but 50% is a non-trivial proportion of risk

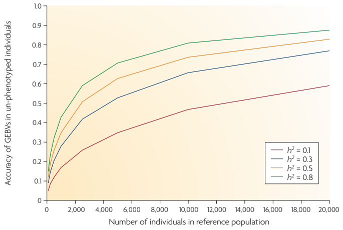
BLUP – Best Linear Unbiased Predictor



- Infinitesimal model
- Genetic effects are random
- Predict the expected genetic effect

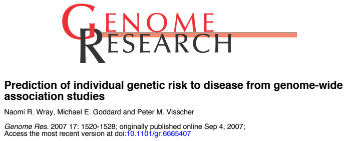

Current BLUP prediction
Pedigree & measurement: Accuracy = 30%
+ Progeny: Accuracy = 60%
Genomic BLUP prediction
Pedigree & measurement: Accuracy = 30%
+ SNPs: Accuracy = 60%
+ Progeny Accuracy = 60%

Accuracy of polygenic prediction in cattle



Poor transferability between breeds!

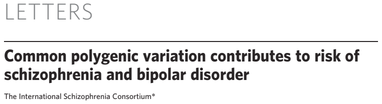
Applications in humans


GENOME RESEARCH

Prediction of individual genetic risk to disease from genome-wide association studies

Naomi R. Wray, Michael E. Goddard and Peter M. Visscher

Genome Res. 2007 17: 1520-1528; originally published online Sep 4, 2007;
Access the most recent version at doi:10.1101/gr.6665407


LETTERS

Common polygenic variation contributes to risk of schizophrenia and bipolar disorder

The International Schizophrenia Consortium*

- LD-prune
- Exclude SNPs of very small effect

Extensions of BLUP – multiple variance scales and binary phenotypes

MultiBLUP:	Speed and Balding. *Genome Research* 2014
Bayesian analysis:	MacLeod et al. *Genetics* 2014
BSLMM:	Zhou et al. *PLOS Genetics* 2013
GeRSI:	Golan and Rossett. *AJHG* 2014

Methods that work with summary statistics

- Summary statistics are easily available
- Most methods require a separate small individual level dataset to tune parameters

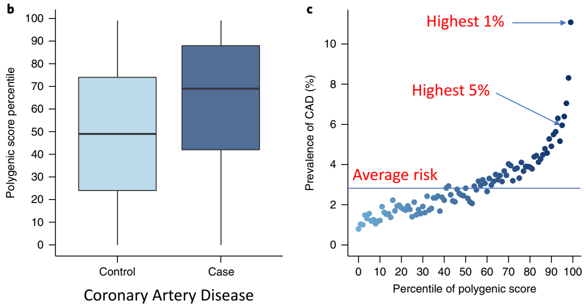
LDPred – a Bayesian method using summary statistics

$$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_g^2}{Mp}\right) \text{with probability } p \\ 0 \text{ with probability } (1-p), \end{cases}$$

Vilhjalmsson et al. 2015

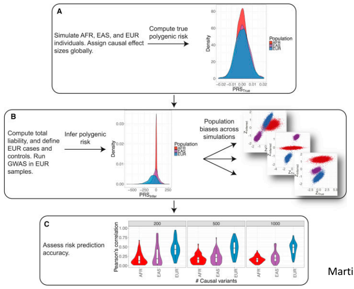
Also, check *BayesR*

Extreme tails in the distributions of genetic risk scores are highly predictive



Khera et al. 2018

With some caveats

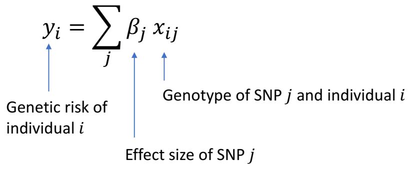

A Simulate AFR, EAS, and EUR individuals. Assign causal effect sizes globally.

B Compute total liability, and define EUR cases and controls. Run GWAS in EUR samples.

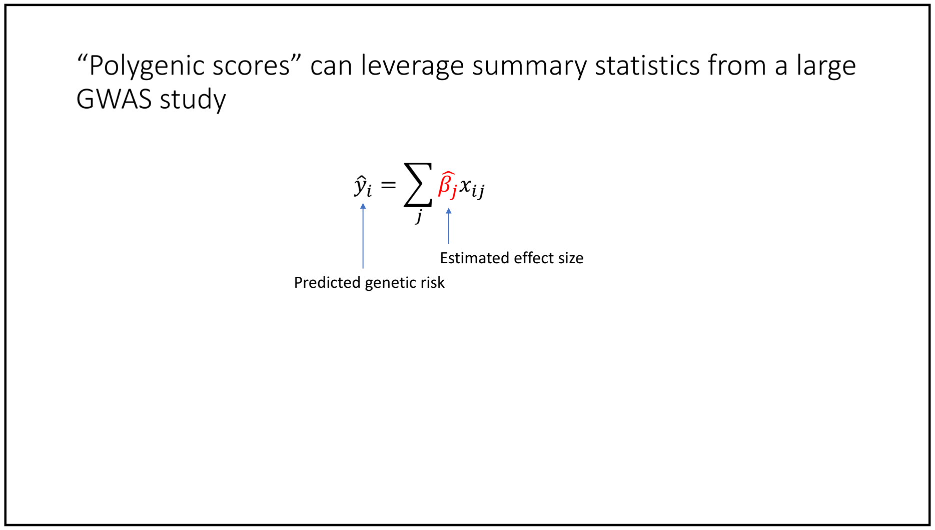
C Assess risk prediction accuracy.

Martin et al., *AJHG* 2017

Linear models for genetic risk prediction

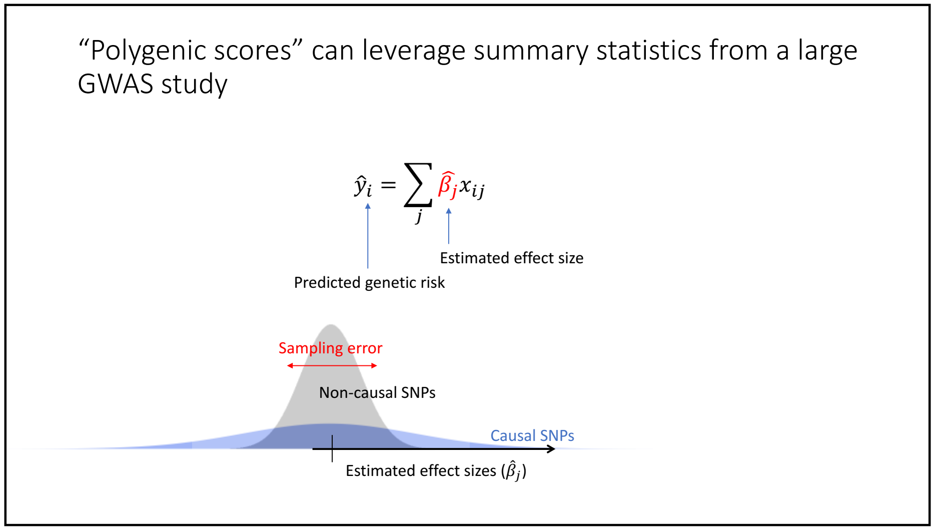

$$y_i = \sum_j \beta_j \, x_{ij}$$

Genetic risk of individual i
Effect size of SNP j
Genotype of SNP j and individual i



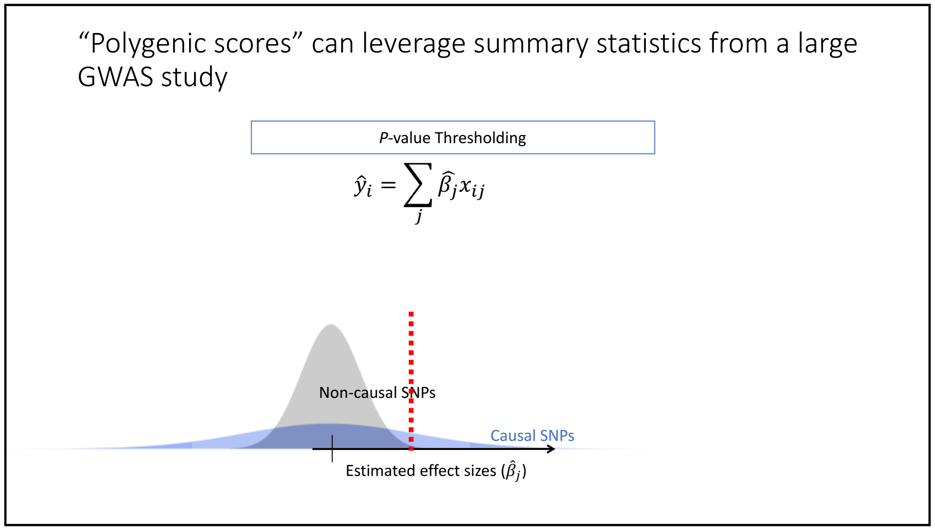
“Polygenic scores” can leverage summary statistics from a large GWAS study

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$



“Polygenic scores” can leverage summary statistics from a large GWAS study

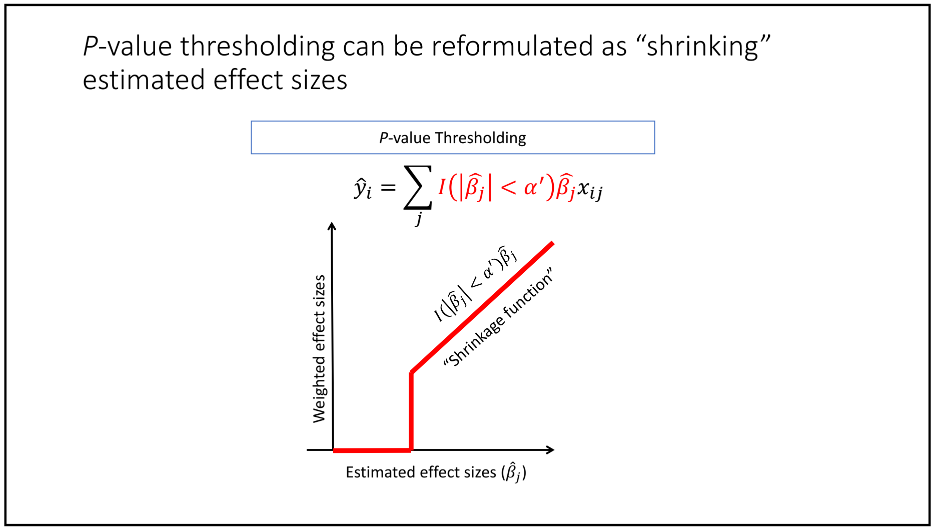
$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$



“Polygenic scores” can leverage summary statistics from a large GWAS study

P-value Thresholding

$$\hat{y}_i = \sum_j \hat{\beta}_j x_{ij}$$



P-value thresholding can be reformulated as “shrinking” estimated effect sizes

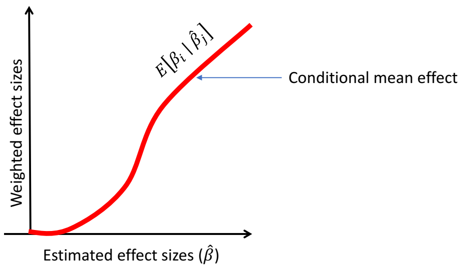
P-value Thresholding

$$\hat{y}_i = \sum_j I(|\hat{\beta}_j| < \alpha')\hat{\beta}_j x_{ij}$$

Weighted effect sizes
$I(|\hat{\beta}_j| < \alpha')\hat{\beta}_j$
“Shrinkage function”
Estimated effect sizes ($\hat{\beta}_j$)

The optimal polygenic score can be constructed with “conditional mean effects”

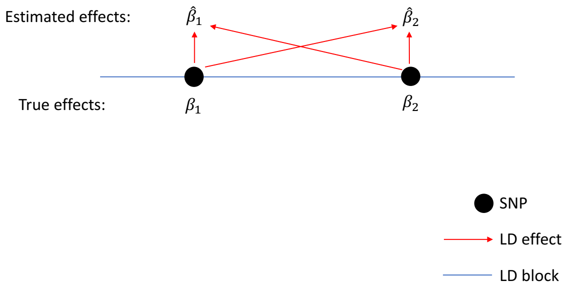
$$\hat{y}_i = \sum_j E[\beta_j \mid \hat{\beta}_j] x_{ij}$$



Goddard et al. 2009

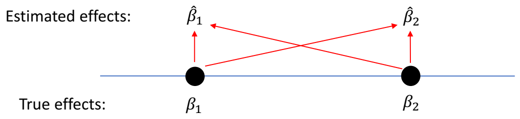
Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**

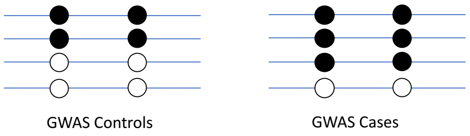


Accounting for LD in summary data is a major challenge

- Correlation between **apparent true genetic effects**



- Correlation between **sampling errors**



Our approach (“**N**on-**P**arametric **S**hrinkage” or NPS)

- No explicit specification of genetic architecture prior, thus “*non-parametric*”
- Learn conditional mean effects directly from training data
- Fully account for correlation in summary statistics

Our approach ("Non-Parametric Shrinkage" or NPS)

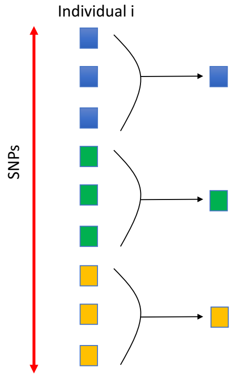
- No explicit specification of genetic architecture prior, thus "*non-parametric*"
- Learn conditional mean effects directly from training data

1. How to estimate $E\left[\beta_j \mid \hat{\beta}_j\right]$ without a Bayesian prior on $\boldsymbol{\beta}$

- Fully account for correlation in summary statistics

2. How to deal with LD

Partitioned risk scores



GWAS-significant

$G_{i,1} = \sum_j \hat{\beta}_j x_{ij} I(\alpha_1 < |\hat{\beta}_j|)$

Sub-threshold

$G_{i,2} = \sum_j \hat{\beta}_j x_{ij} I(\alpha_2 < |\hat{\beta}_j| < \alpha_1)$

Noise

$G_{i,3} = \sum_j \hat{\beta}_j x_{ij} I(|\hat{\beta}_j| < \alpha_1)$

Piecewise linear interpolation on shrinkage curve

Estimates of genetic effects in GWAS data ($\hat{\beta}_j$)

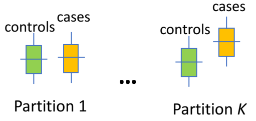
↓

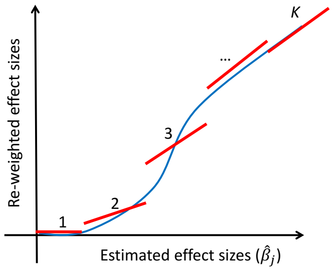
Partition SNPs into K subgroups:
$\mathcal{S}_k = \{\, j : b_{k-1} < |\hat{\beta}_j| < b_k \,\}$

↓

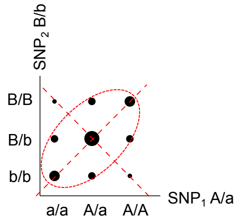
Partitioned risk scores: $G_{ik} = \sum_{j \in \mathcal{S}_k} \hat{\beta}_j x_{ij}$

↓

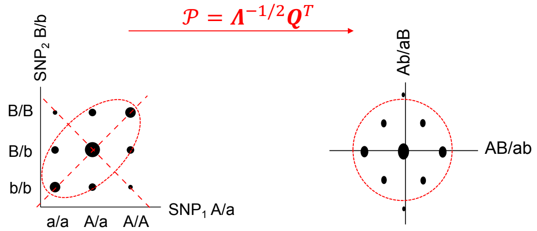




How to deal with LD?



Decorrelating linear projection $\mathcal{P}$

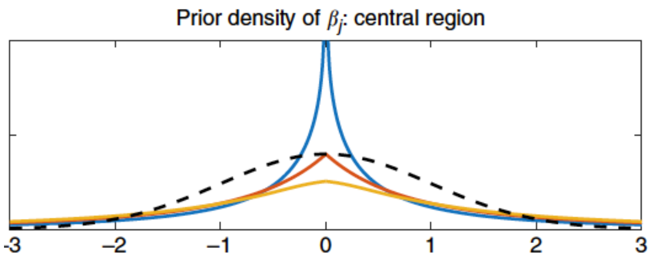


$\boldsymbol{\Sigma}$ is a local LD matrix and $\boldsymbol{\Sigma} = \boldsymbol{Q}\,\boldsymbol{\Lambda}\,\boldsymbol{Q}^T$ by eigenvalue decomposition

$$\boldsymbol{\Sigma}^{-1} = \boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1}\,\boldsymbol{Q}^T = (\boldsymbol{Q}\,\boldsymbol{\Lambda}^{-1/2})(\boldsymbol{\Lambda}^{-1/2}\boldsymbol{Q}^T)$$

Other shrinkage methods: PRS-CS

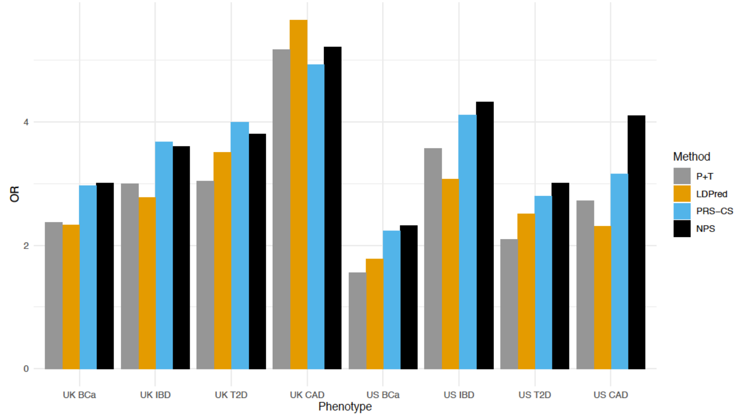
$$\beta_j \sim \mathrm{N}\left(0, \frac{\sigma^2}{N}\phi\psi_j\right), \qquad \psi_j \sim g,$$



Other shrinkage methods: PRS-CS

Lassosum – extension of *LASSO*

Accuracy of the 5% tail



Summary on the method

- NPS accounts for the correlation of sampling errors in GWAS summary statistics.
- NPS provides an extensible framework to estimate the shrinkage curve from training data.
- NPS is best-suited to take advantage of the high density of markers and imputation accuracy in latest GWAS datasets.

The preprint is available in BioRxiv:
Chun et al. “*Non-parametric polygenic risk prediction using partitioned GWAS summary statistics.”*

Software is available at: https://github.com/sgchun/nps

Is an extreme presentation with a family history Mendelian?

- It is often assumed that an extreme phenotypic presentation is due to a large effect Mendelian mutation.
- Apparently Mendelian family history is assumed to support a highly penetrant Mendelian mutation.
- Could these cases be polygenic (or, at least, not monogenic)?