Polygenic Risk Scoring

Peter Sørensen

Center for Quantitative Genetics and Genomics (QGG)

Who Am I?

Senior Scientist - Center for Quantitative Genetics and Genomics (QGG) - Aarhus University (2003-present)

- Statistical models and methods for better understanding of genetic architecture and prediction of complex traits and diseases
- Implemented some of these methods into an R software package (qgg) that can be used for risk prediction

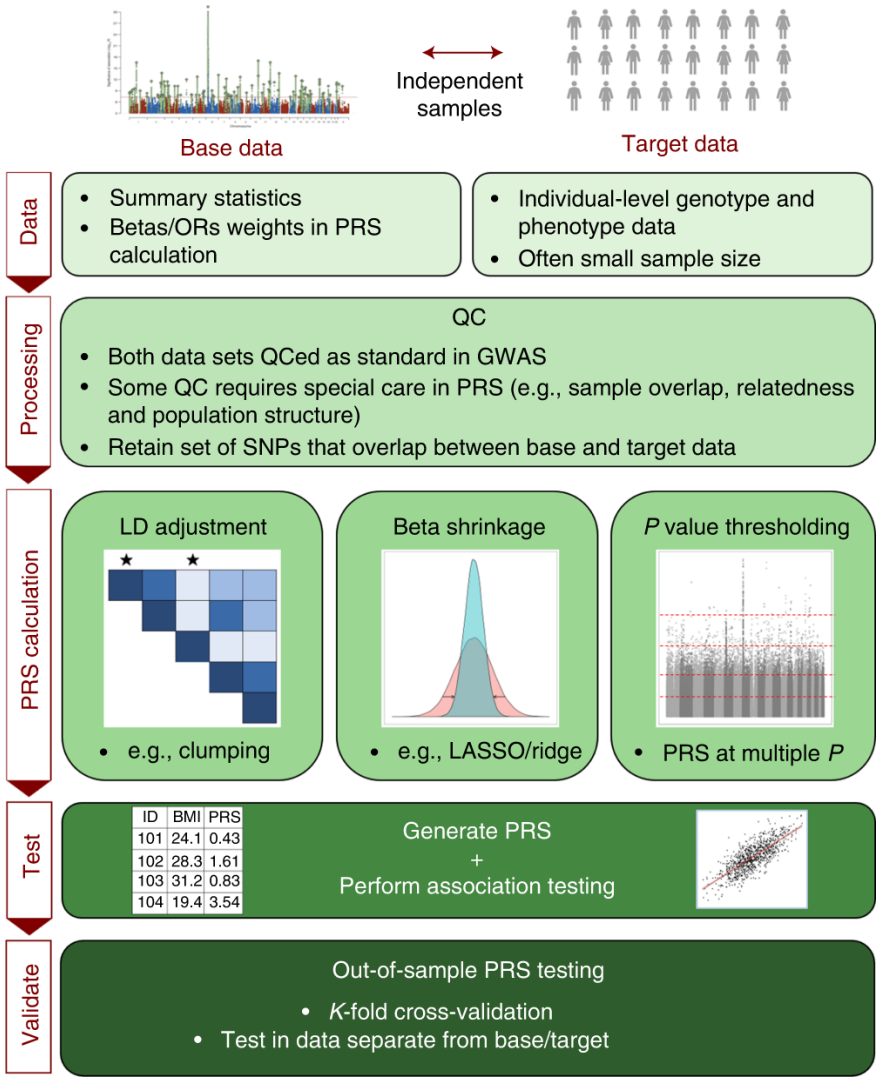
Senior Scientist - Genomics Plc - Oxford (2016-2020)

- Genomics Plc is a company focused on using genomics for drug discovery and precision health in human health care systems
- Developed statistical methods and contributed to implementing a computational platform for genetic risk prediction used in Precision Health

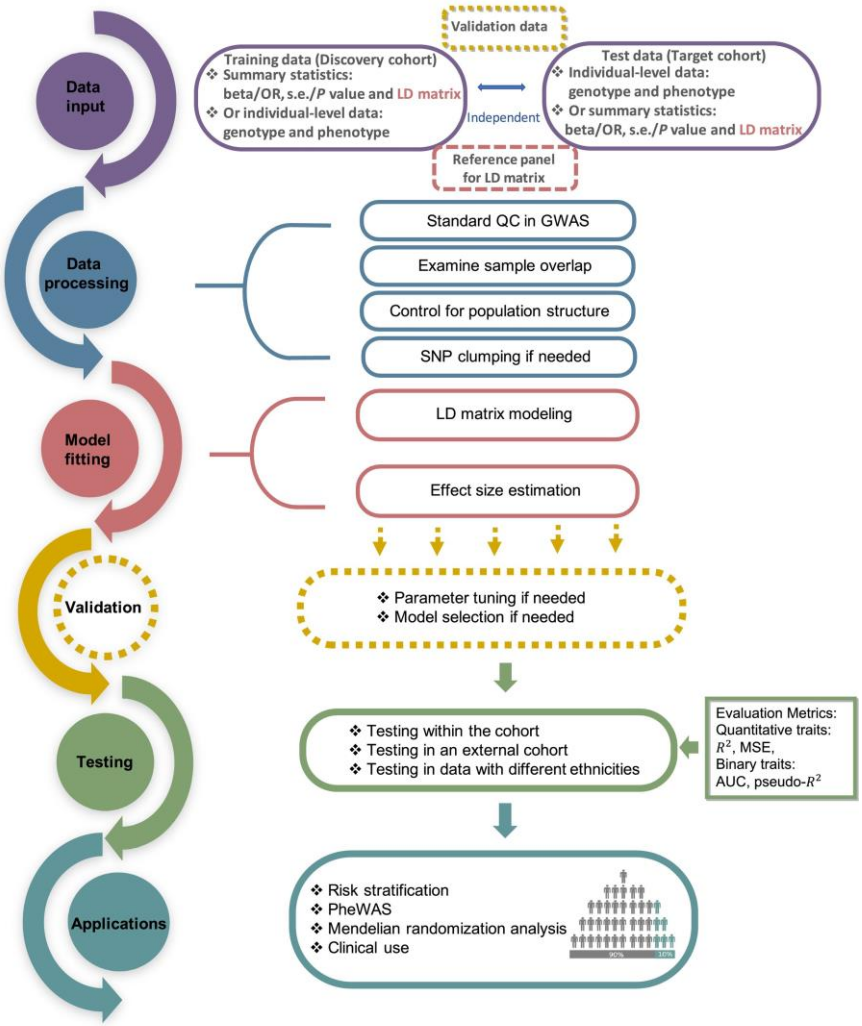
Outline

1. Introduction to Polygenic Risk Scoring
2. Data used for Polygenic Risk Scores
3. Methods for Computing Polygenic Risk Scores
4. Methods for Evaluating Polygenic Risk Scores
5. Clinical Utility of Polygenic Risk Scores
6. Brief Introduction to using R-package qgg for PRS

Introduction Polygenic Risk Scoring



(Choi et al 2020)



(Ma & Zhou 2021)

Introduction Polygenic Risk Scoring

Many complex traits and diseases in humans are heritable:

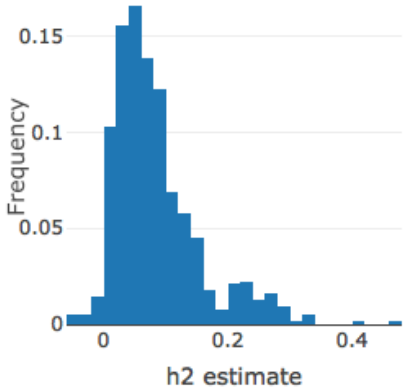
- Degree of heritability determines the value of using genetics for risk prediction

- For many traits and diseases there will be thousands of genetic variants that each contribute with a small effect on disease risk

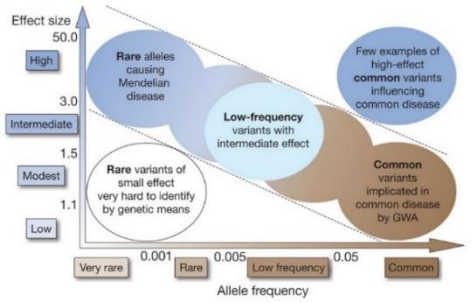
- Rare variant with large effects will only explain small proportion of h^2 (low predictive potential)

- Common variants with small effects can explain larger proportion of h^2 (high predictive potential)

Need large data sets to accurately estimate small to moderate effects => improve prediction accuracy



(http://ldsc.broadinstitute.org/)



(Manolio et al. 2009)

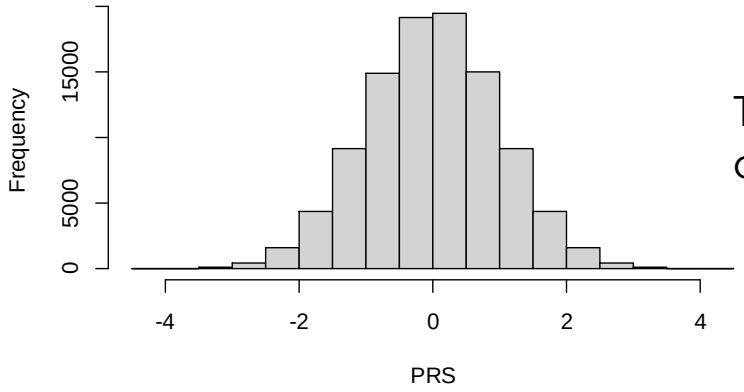
Introduction Polygenic Risk Scoring

Polygenic risk scoring combines information from large numbers of markers across the genome (hundreds to millions) to give a single numerical score for individual's risk for developing a specific disease on the basis of the DNA variants they have inherited.

For a particular disease or trait a polygenic risk score (PRS) is calculated as:

$$\mathrm{PRS} = \sum_{i=1}^{m} X_i \, b_i$$

where X_i is the genotype vector, and b_i the weight of the i'th single genetic marker.



The PRS tends to follow a normal distribution

Introduction Polygenic Risk Scoring

Terminology: Polygenic risk scores, polygenic scores, genomic risk score, genetic scores, genetic predispostion, genetic value, genomic breeding value is (more or less) the same thing.

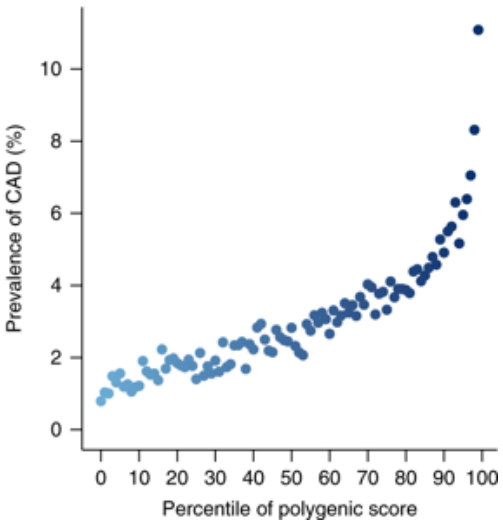
Genomic prediction used for many years in animal and plant breeding (e.g. Meuwissen et al. 2001)

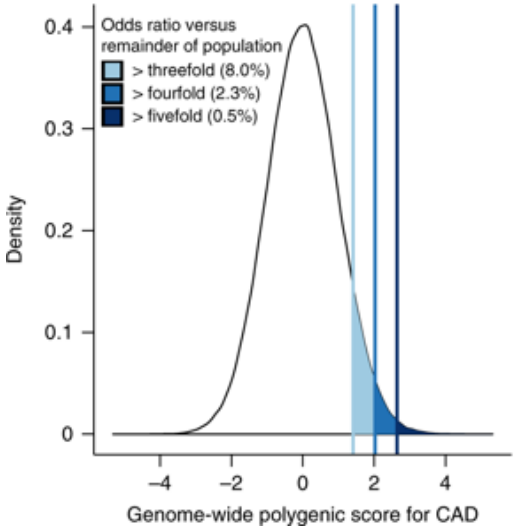
Genomic prediction (i.e. polygenic risk scoring) in humans:

- Larger GWAS sample size = more precision for effect estimates
- Development of methods that combine genome-wide sets of variants
- Large Biobanks for validation and testing of genetic risk scores
- Ability to identify clinically meaningfull increases in disease risk predictions

Introduction Polygenic Risk Scoring

Polygenic risk scores can be a powerful approach to identify individuals with higher (or lower) risk of particular diseases:





$$PRS = \sum_{i=1}^{m} X_i \, b_i$$

(Khera et al. 2018)

- Individuals ranked according to their polygenic risk score for coronary artery disease (CAD)
- Individuals in the extreme upper tail of the PRS distribution have increased risk for CAD

Introduction Polygenic Risk Scoring

Which data do we need to perform polygenic risk scoring?

Which methods should we use to compute the polygenic risk scores?

How do we evaluate the predictive ability of the polygenic risk scores?

Outline

Introduction to Polygenic Risk Scoring

Data used for Polygenic Risk Scores
- Training/Validation
- Inivididual level or summary statistic data

Methods for Computing Polygenic Risk Scores
- Standard approach based on LD pruning and thresholding (P+T)
- Bayesion approach using shrinkage estimation (e.g. Ldpred, BayesR)
- Multiple trait approaches

Methods for Evaluating Polygenic Risk Scores
- Population or individual level measures
- Quantitative and binary traits
- Expected accuracies

Clinical Utility of Polygenic Risk Scores

Data for Polygenic Risk Scoring

Which data do we need to perform polygenic risk scoring?

Training/Discovery/Base population

- used for obtaining marker weights (b_i)
- individual level phenotype and genotype data from which we directly can obtain marker effects
- Or genome-wide marker association summary statistics (e.g. beta's, standard errors of beta's, z-scores, p-values) from which we can derive marker weights

$$PRS = \sum_{i=1}^{m} X_i \, b_i \qquad\qquad y_{obs} <-> PRS$$

Validation/Testing/Target population

- used for evaluating the polygenic risk scores (or other risk predictors)
- individidual level phenotype and genotype data

Data for Polygenic Risk Scoring

Training/Discovery/Base population

- How was the disease phenotype defined?
- What are the number of observations, cases/controls?
- Which co-factors was used in the GWAS analyses?
- Which ancestry and environment characteristics?
- What are the characteristics of the study population (males, females, both)?
- Can I get access to the data?

Validation/Testing/Target population

- Same considerations as above
- Overlap in markers used in discovery/target population?
- Same ancestry and environment characteristics will increase accuracy
- Independence of individuals in discovery and target population

Important to always perform extensive quality control of your training and validation data

(Choi et al. 2020)

Data for Polygenic Risk Scoring

Quality control is a critical step for working with summary statistics (in particular external).

Processing and quality control of GWAS summary statistics includes:

- map marker ids (rsids/cpra (chr, pos, ref, alt)) to LD reference panel data
- check effect allele (flip EA, EAF, Effect)
- check effect allele frequency
- thresholds for MAF and HWE
- exclude INDELS, CG/AT and MHC region
- remove duplicated marker ids
- check which build version
- check for concordance between marker effect and LD data

R functions (qcStat and adjLDStat) for processing of summary statistics is available in our qgg package.

Data for Polygenic Risk Scoring

Identifying summary statistics that does not fit to the LD matrix:

It is assumed that the z statistics ($z = \frac{\hat{b}}{\sigma_{\hat{b}}}$) follows a multivariate Gaussian distribution with mean zero and variance covariance matrix that is determined by the LD matrix. Impute z statistics (z_{pred}) based on LD reference and observed z statistics (z_{obs}) in a region (e.g. Pasaniuc et al. 2015, Chen et al. 2021):

$$\mathrm{z_{i|-i}} = \mathrm{D_{i,-i}}\mathbf{D^{-1}_{-i,-i}}\mathrm{z_{-i}}$$

To test if z_{obs} is an outlier the following is used:

$$\frac{\left(\mathrm{z_i} - \mathrm{z_{i|-i}}\right)^2}{1 - \mathrm{D_{i,-i}}\mathbf{D^{-1}_{-i,-i}}\mathrm{D'_{i,-i}}} \sim X_1^2$$

Ideally the values should be on the diagonal and values that are too far away from the diagonal are considered to be outliers.

Outline

Introduction to Polygenic Risk Scoring

Data used for Polygenic Risk Scores

- Training/Validation
- Inivididual level or summary statistic data

Methods for Computing Polygenic Risk Scores

- Standard approach based on LD pruning and thresholding (P+T)
- Bayesion approach using shrinkage estimation (e.g. Ldpred, BayesR)
- Multiple trait approaches

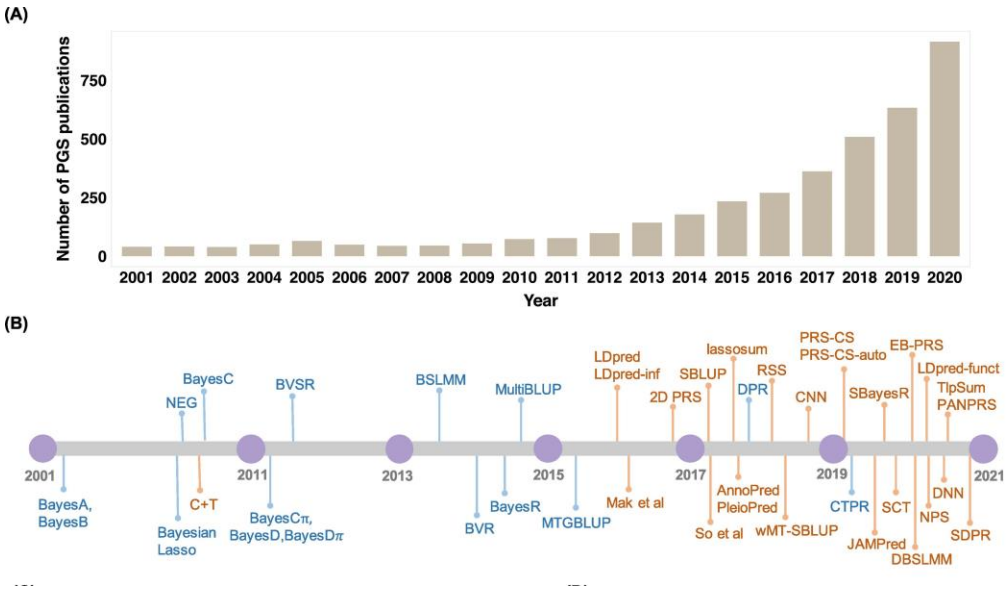
Methods for Evaluating Polygenic Risk Scores

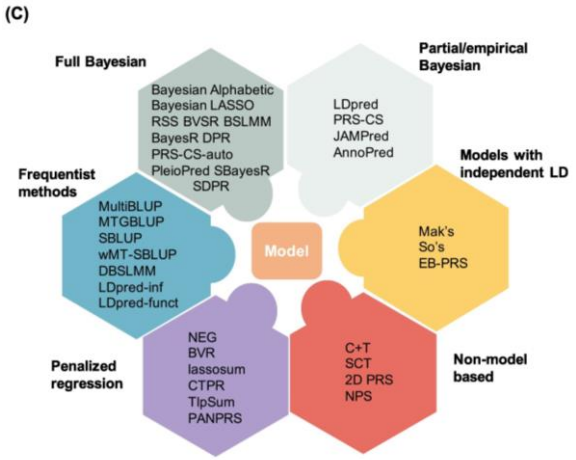
- Population or individual level measures
- Quantitative and binary traits
- Expected accuracies

Clinical Utility of Polygenic Risk Scores

Methods for Computing Polygenic Risk Scores

Which methods should we use to compute the marker weights used in the polygenic risk scores?





(Ma & Zhou 2021)

Methods for Computing Polygenic Risk Scores

For a particular disease or trait a polygenic risk score (PRS) is constructed as:

$$\mathrm{PRS} = \sum_{i=1}^{m} X_i \, b_i$$

where X_i is the genotype, and b_i the weight of the i'th single genetic marker

- weights could be -1,0,1
- or beta's (or log(OR)) from a standard genome-wide association analysis
- or adjusted beta's (or log(OR)) from multiple regression models

Methods for Computing Polygenic Risk Scores

Which markers should be included in the polygenic risk score?

Including only genome-wide significant SNPs in a prediction model usually leads to poor prediction:

- polygenic nature of many complex traits means that many true predictors do not reach genome-wide significance significance
- each individually marker conveys little information, but collectively they can be important.

On the other hand, including many predictors in a model risks over-fitting:

- parameter estimates achieve close matching of fitted values to the observed data, which appears good but …
- much of this apparent success amounts to “fitting statistical noise”: parameters are tuned to irreproducible features of the data, leading to poor fit to new data

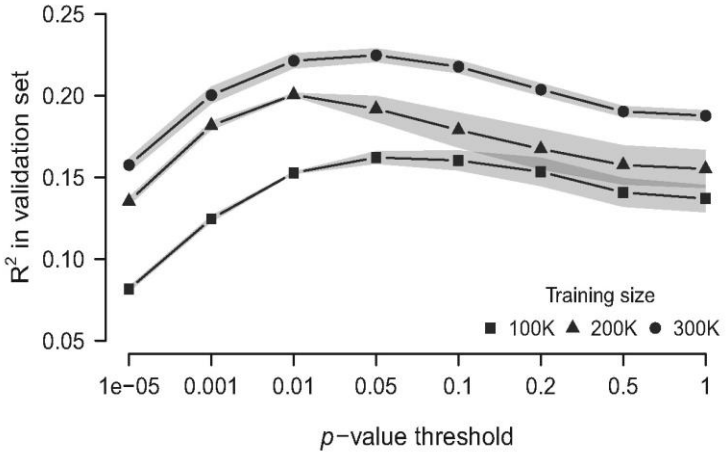
Methods for Computing Polygenic Risk Scores

Standard approach based on LD pruning and thresholding (P+T):

- Test each SNP one-at-a-time in the training sample and record those that are significant at level α and their estimated effect sizes.

- Account for LD (i.e. correlation) between markers using r2 threshold based on a LD reference panel (e.g. 1000G).

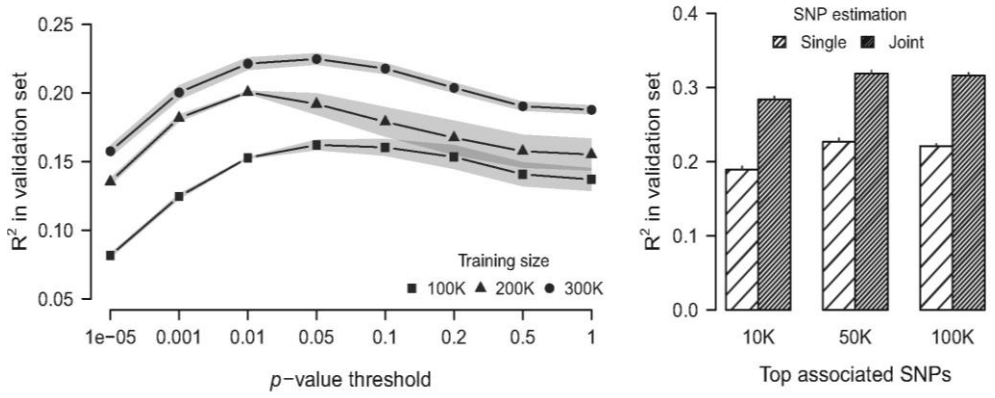
- It is common to repeat for different α and r2 in order to try to maximise predictive success.



(Rohde et al. 2020)

Methods for Computing Polygenic Risk Scores

A better solution to the over-fitting problem is offered by penalised (or shrinkage) regression in which a penalty in the residual sum of squares or log-likelihood "shrinks" parameter estimates towards zero:



It can also be motivated in Bayesian terms: the penalty function should reflect available knowledge about the true distribution of effect sizes of marker alleles, i.e. your prior distribution.

(de los Campos et al. 2013, Rohde et al. 2020)

Methods for Computing Polygenic Risk Scores

Bayesian linear regression (BLR) models:

- unified mapping of genetic variants, estimation of genetic parameters (e.g. heritability) and prediction of disease risk
- handles different genetic architectures (few large, many small effects)
- scale to large data (e.g. sparse LD)

(e.g. Lloyd-Jones et al. 2019, Vilhjálmsson et al. 2015)

Methods for Computing Polygenic Risk Scores

Estimation of marker effects based on a multiple linear regression model in which the phenotype is related to the set of genetic markers:

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

y: is an n×1 vector of trait phenotypes (centered)

X: is an n×m matrix of genotypes (suitable coded)
b: is an m×1 vector of marker effects

Z: is an n×p design matrix for covariates (e.g. age, location, treatment)
c: is an p×1 vector of covariate effects

e: is an n×1 vector of residuals assumed to be independent and normally distributed

Methods for Computing Polygenic Risk Scores

Estimation of marker effects based on a multiple linear regression model in which the phenotype is related to the set of genetic markers:

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

- Most often only additive genetics effects are modelled, thus ignoring dominance and epistasis.
- Independence of the residuals implies that all kinship effects are assumed to be accounted for through the markers.

Methods for Computing Polygenic Risk Scores

Estimation of marker effects is based on a multiple linear regression model fitted to the phenotypes in the training data:

$$\mathbf{y}_T = \mathbf{X}_T\mathbf{b}_T + \mathbf{Z}_T\mathbf{c}_T + \mathbf{e}_T \qquad \text{(Training data)}$$

Computation of polygenic risk scores are based on the marker effects estimated in the training data and the genotypes of the individuals in the validation data:

$$\hat{\mathbf{y}} = \boldsymbol{X}_V\hat{\mathbf{b}}_T + \boldsymbol{Z}_V\hat{\boldsymbol{c}}_T \qquad \text{(Validation data)}$$

$$\widehat{\boldsymbol{PRS}} = \sum_{i=1}^{m} \mathbf{X}_i\,\hat{b}_i = \mathbf{X}_V\hat{\mathbf{b}}_V$$

In practice covariates can be very important in prediction, but from now on we ignore them and focus on prediction from genomic data

Methods for Computing Polygenic Risk Scores

Relationship between single and multiple marker models:

Single marker effects: $\hat{b}_i = (\mathbf{x}_i'\mathbf{x}_i)^{-1}\mathbf{x}_i'\mathbf{y}$ (marginal)

Multi marker effects: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ (joint)

Multi marker effects with regularization (needed if m>n):

$$\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{X} + \mathbf{I}\frac{\sigma_e^2}{\sigma_b^2}\right)^{-1}\mathbf{X}'\mathbf{y}$$ (joint with shrinkage)

- joint estimation of marker effects
- account for linkage disequilibrium (LD)
- uniform shrinkage of marker effects if σ_b^2 is the same for all markers
- differential shrinkage of marker effects if σ_b^2 is not the same for all markers

Methods for Computing Polygenic Risk Scores

Multiple linear regression models are based on individual level data (i.e. **y** and **X**) or summary statistic data.

$\mathbf{X'X}$ is derived from an LD matrix **B** (population matched reference) and summary statistics:

$$\mathbf{X'X = D^{0.5}BD^{0.5}}$$

where $\mathrm{D_i} = \frac{1}{\widehat{\sigma}^2_{\mathrm{b_i}} + \widehat{\mathrm{b}}^2_{\mathrm{i}}/\mathrm{n_i}}$ if the genotypes have been centered to mean 0 or $\mathrm{D_i} = \mathrm{n_i}$ if the genotypes have been centered to mean 0 and scaled to unit variance

$\mathbf{X'y}$ is derived from univariate marker effects:

(Lloyd-Jones et al. 2019
Vilhjálmsson et al. 2015)

$$\mathbf{b_{uni} = D^{-1}X'y \Rightarrow X'y = Db_{uni}}$$

Methods for Computing Polygenic Risk Scores

In the Bayesian multiple regression model the posterior density of the model parameters $(\mathbf{b}, \sigma^2_\beta, \sigma^2_e)$ depend on the likelihood of the data given the parameters and a prior probability for the model parameters:

$$p(\mathbf{b}, \sigma^2_\beta, \sigma^2_e | \mathbf{y}) \propto p(\mathbf{y}|\sigma^2_e, \sigma^2_\beta, \mathbf{b}, \dots)p(\mathbf{b}|\sigma^2_\beta \dots)p(\sigma^2_\beta| \dots)p(\sigma^2_e| \dots)$$

The prior density of marker effects, $p(\mathbf{b}|\sigma^2_\beta \dots)$, defines whether the model will induce variable selection and shrinkage or shrinkage only.

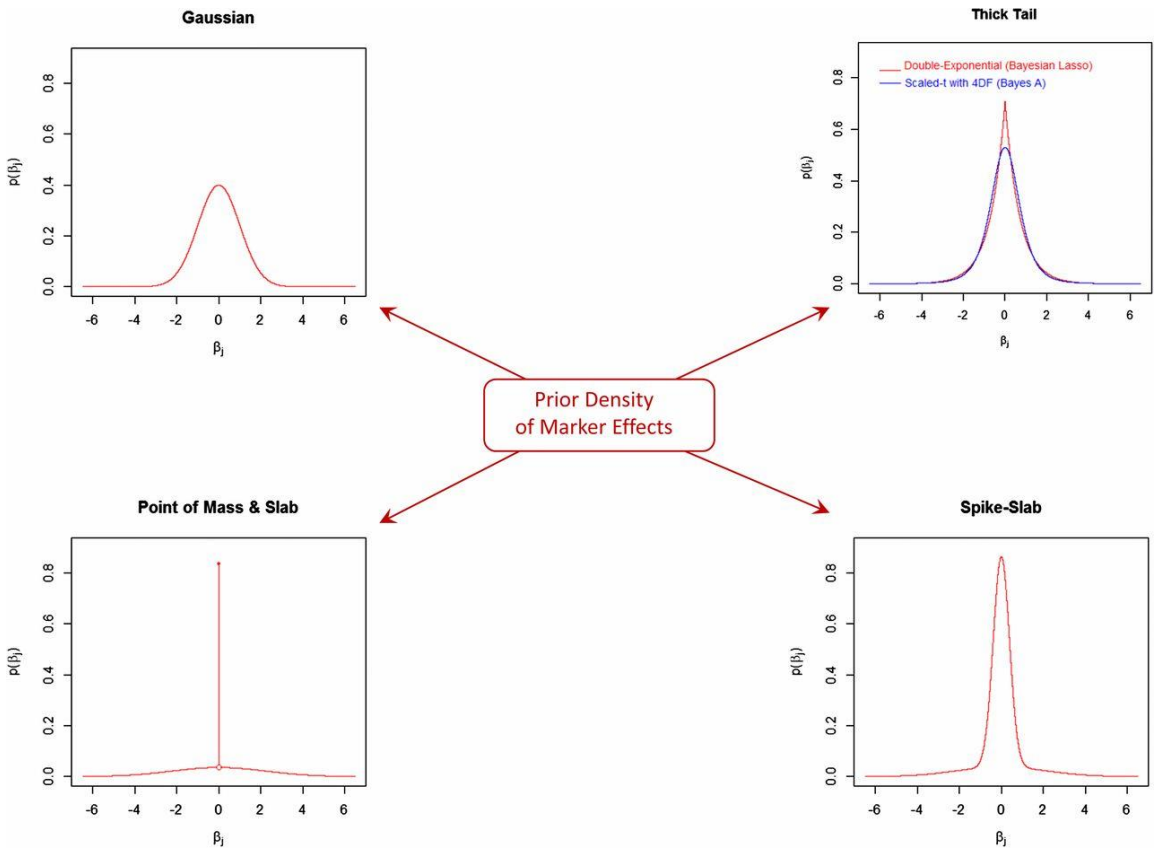
Also, the choice of prior will define the extent and type of shrinkage induced.

Ideally the choice of prior for the marker effect should reflect the genetic architecture of the trait, and will vary (perhaps a lot) across traits.

(de los Campos et al. 2013)

Methods for Computing Polygenic Risk Scores

Commonly used prior densities of marker effects (all with zero mean and unit variance):



The densities are organized in a way that, starting from the Gaussian in the top left corner, as one moves clockwise, the amount of mass at zero increases and tails become thicker and flatter.

(de los Campos et al. 2013)

Methods for Computing Polygenic Risk Scores

Commonly used prior densities of marker effects :

BLUP: Assigning a Gaussian prior to β implies that the posterior means are the BLUP estimates (same as Ridge Regression).

Bayesian Lasso: Assigning a double-exponential or Laplace prior is the density used in the Bayesian LASSO

Bayes A: similar to ridge regression but t-distribution prior (rather than Gaussian) for the βj ; variance comes from an inverse-χ 2 instead of being fixed. Estimation via Gibbs sampling.

Bayes Cπ: uses a "rounded spike" (low-variance Gaussian) at origin many small effects can contribute to polygenic component, reduces the dimensionality of the model (makes Gibbs sampling feasible).

Bayes R: Hierarchical Bayesian mixture model with 4 Gaussian components, with variances scaled by 0, 0.0001 , 0.001 , and 0.01 .

(de los Campos 2013, Lloyd-Jones et al. 2019, Vilhjálmsson et al. 2015)

Methods for Computing Polygenic Risk Scores

In the Bayesian multiple regression model the marker effects, $\mathbf{b}$, are a priori assumed to be sampled from a mixture with a point mass at zero and univariate normal distributions conditional on common marker effect variance σ^2_β, and variance scaling factors, $\boldsymbol{\gamma}$:

$$b_j|\sigma^2_\beta, \boldsymbol{\gamma} = \begin{cases} 0 \text{ with probability } \pi_1 \\ \sim N\left(0, \gamma_2\sigma^2_\beta\right) \text{ with probability } \pi_2 \\ \ldots \\ \sim N\left(0, \gamma_C\sigma^2_\beta\right) \text{ with probability } \pi_C \end{cases}$$

where $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_C)$ is a vector of probabilities and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_C)$ is a vector of variance scaling factors for each of C marker variance classes (e.g. SBayesR or LDPred).

(Lloyd-Jones et al. 2019, Vilhjálmsson et al. 2015)

Methods for Computing Polygenic Risk Scores

- Estimation of the joint marker effects depend on additional model parameters such as a probability of being causal ($\boldsymbol{\pi}$), an overall marker variance (σ^2_β), and residual variance (σ^2_e).

- An iterative algorithm for estimating joint marker effects. The model parameters can be pre-specified or can be estimated as part of the iterative estimation procedure.

- Estimation of model parameters can be done by sampling from fully conditional posterior distributions or by using a grid search over potential model parameter values.

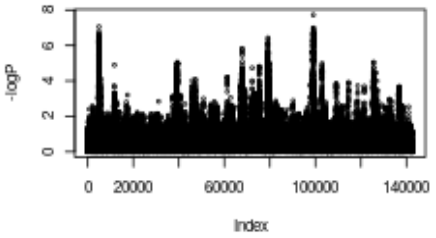
Methods for Computing Polygenic Risk Scores

Weights used in polygenic risk scores estimated using Bayesian linear regression (BLR) models:

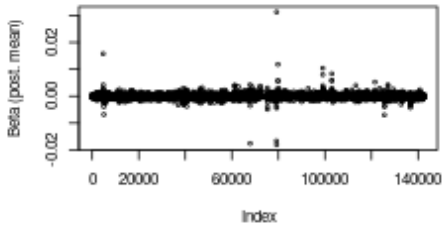
$$b_j|\sigma_\beta^2, \boldsymbol{\gamma} = \begin{cases} 0 \text{ with probability } \pi_1 \\ \sim N(0, \gamma_2\sigma_\beta^2) \text{ with probability } \pi_2 \\ \cdots \\ \sim N(0, \gamma_C\sigma_\beta^2) \text{ with probability } \pi_C \end{cases}$$

Effect size
No effect
Small effect
.....
Large effect

- BLR handles different genetic architectures (few large, many small effects)
- BLR models leads to much clearer genetic signal leading to better predictive power for genetic risk predictors (Lloyd-Jones et al. 2019, Vilhjálmsson et al. 2015)



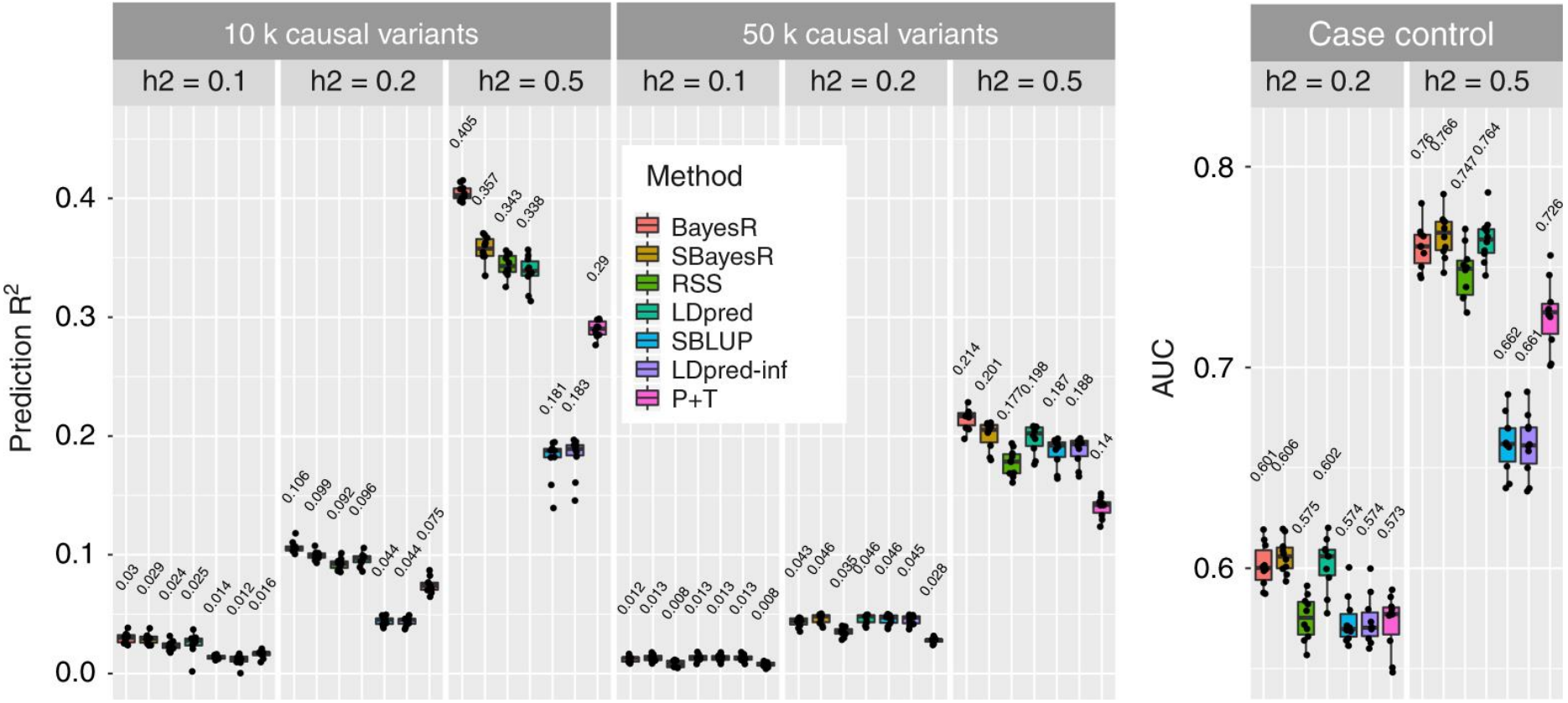
Standard GWAS signal



BLR adjusted GWAS signal

Methods for Computing Polygenic Risk Scores

Comparison of polygenic risk score methods on simulated data:

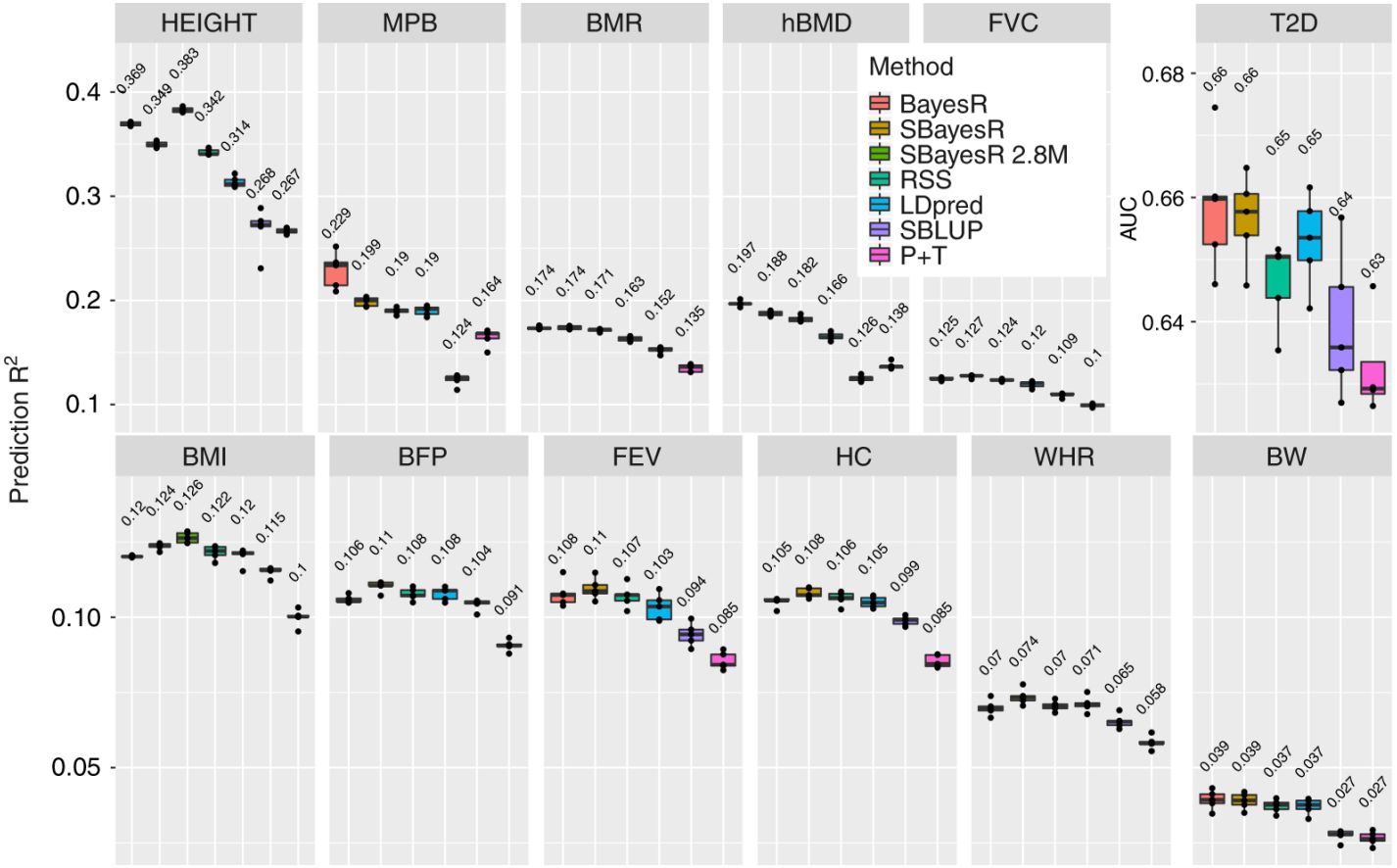


2500 variants ~ N(0, 0.01)
5000 variants ~ N(0, 0.1)
2500 variants ~ N(0, 1)

(Lloyd-Jones et al. 2019)

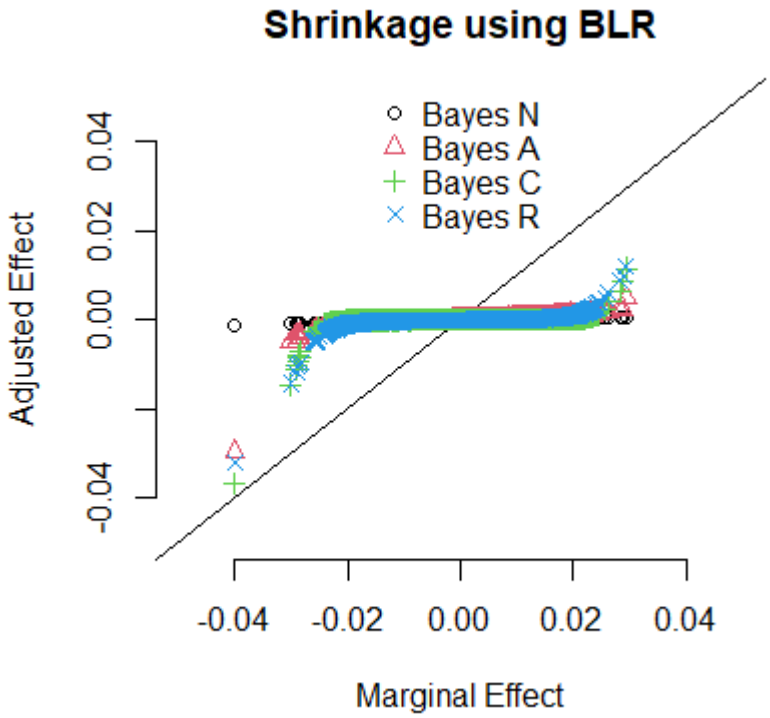
Methods for Computing Polygenic Risk Scores

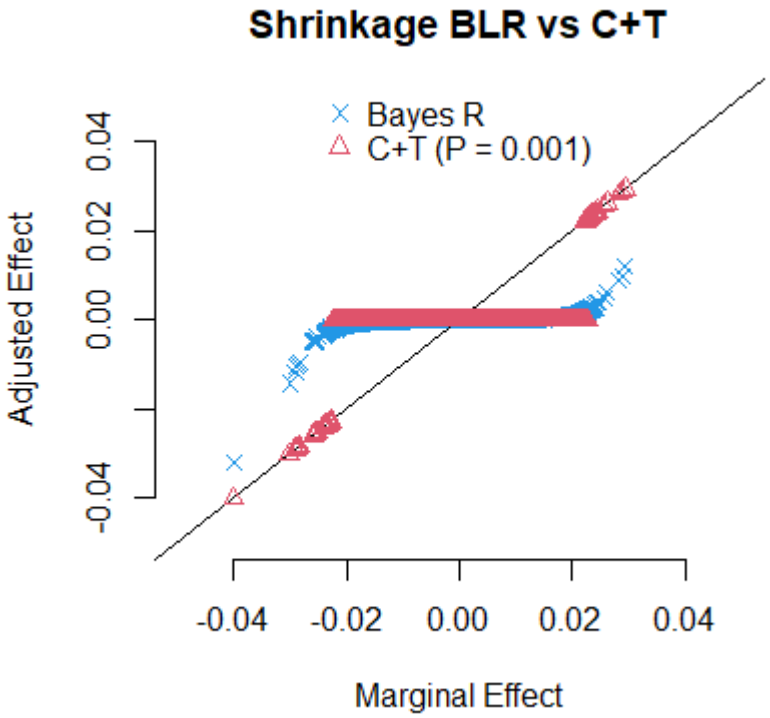
Comparison of polygenic risk score methods on real data:



(Lloyd-Jones et al. 2019)

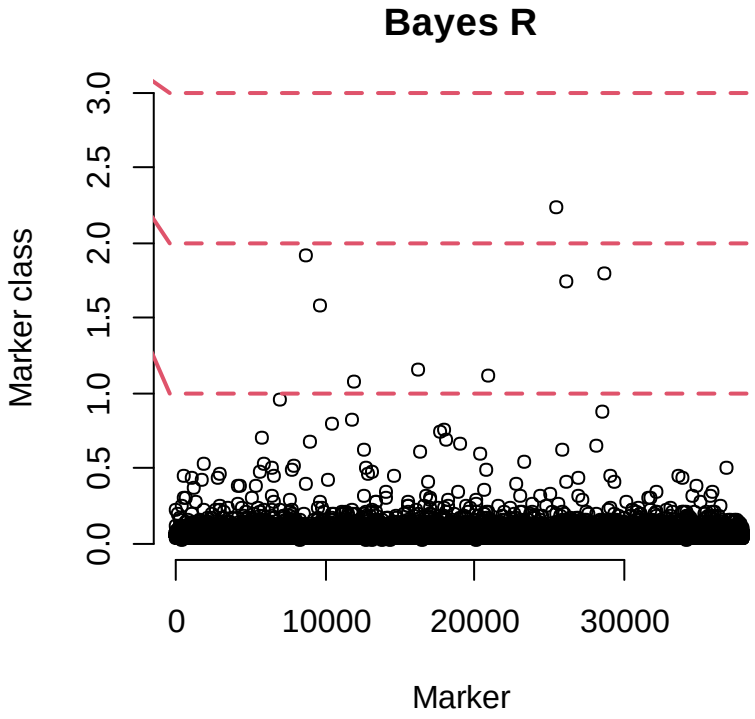
BLR Models – Estimation of marker effects

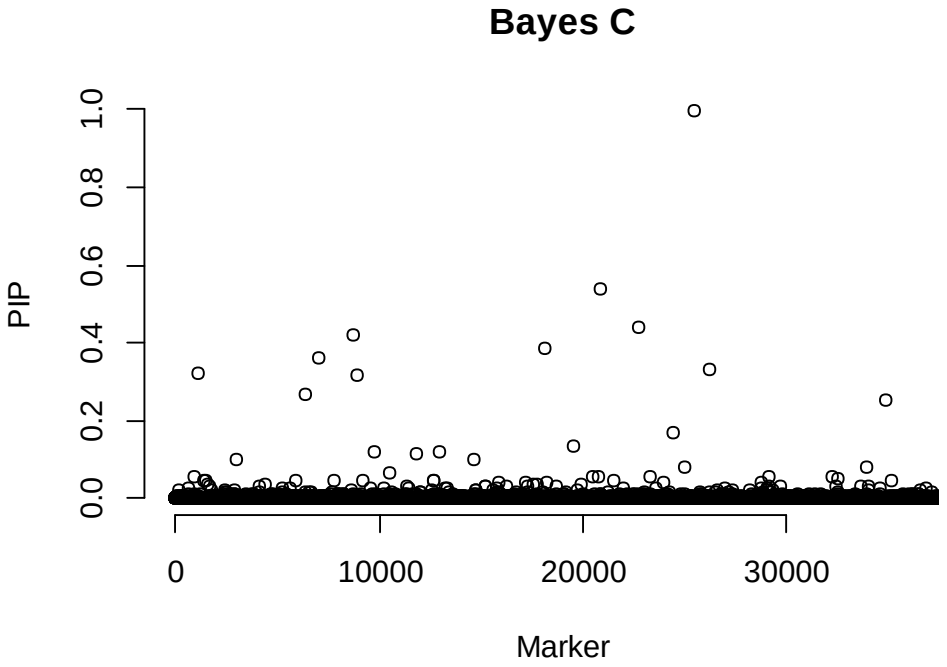




- Marginal effects are shrunk towards zero
- Bayes R/C less shrinkage compared to Bayes A/N
- Clumping and Thresholding (C+T) is also ”a shrinkage method”

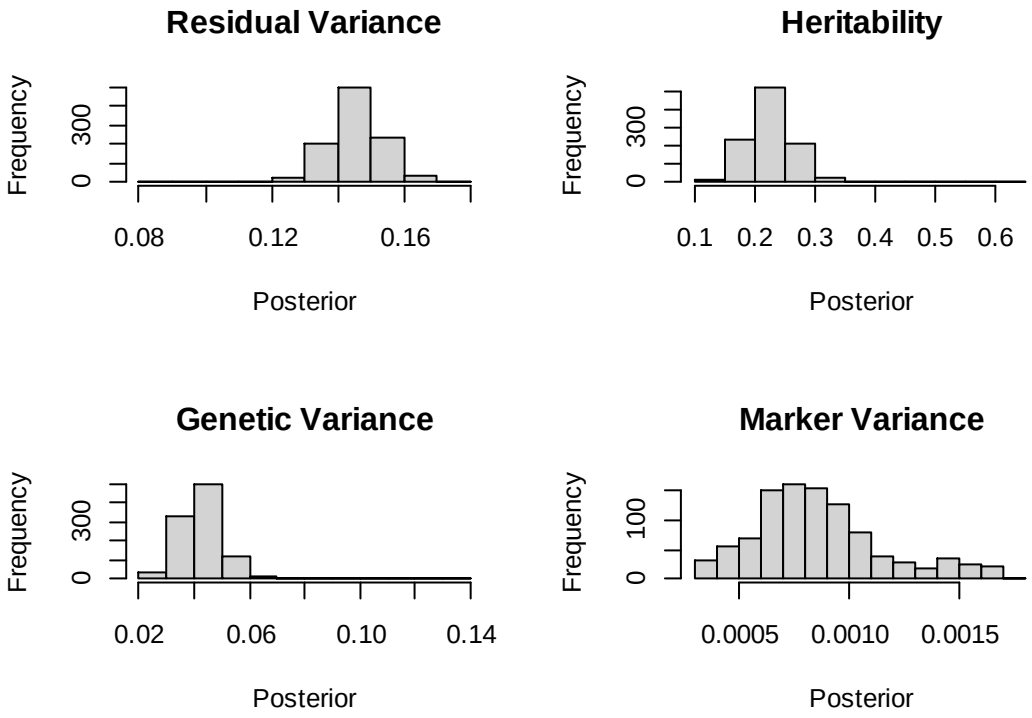
BLR Models - Finemapping





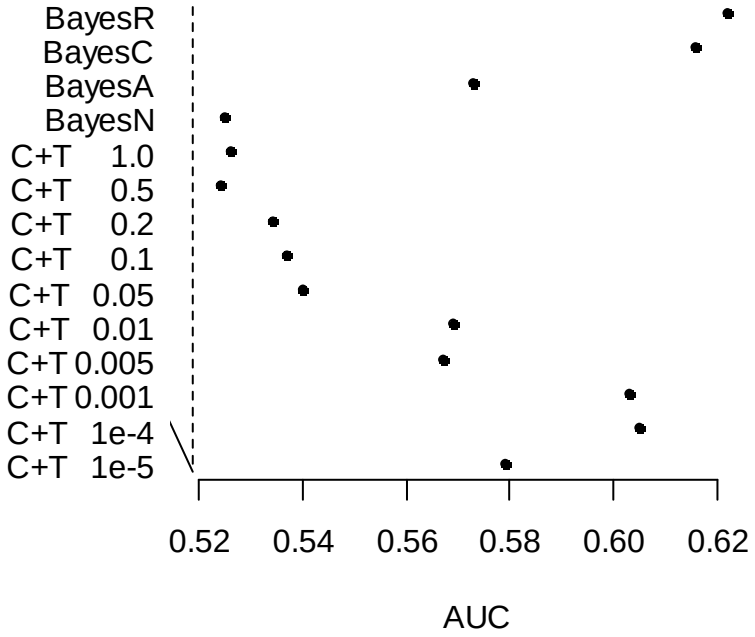
- BayesR posterior mean for variance class
- BayesC posterior inclusion probability (PIP)
- Bayes R/C can be used for finemapping, but mapping power and precision needs to be determined and compared to existing methods

BLR Models – Estimation of Genetic Parameters



- estimation of variance components and genetic parameters (e.g. h2)
- posterior mean and variance of parameters

BLR Models – Genomic Prediction



- BLR models can be used for prediction (AUC=Area Under the Curve)
- Bayes R/C most accurate, but C+T is pretty close!
- ”Best model” depends on the true genetic architecture and how well we can estimate the parameters in the model

Methods for Computing Polygenic Risk Scores

Bayesian linear regression (BLR) models:

- unified mapping of genetic variants, estimation of genetic parameters (e.g. heritability) and prediction of disease risk
- handles different genetic architectures (few large, many small effects)
- scale to large data (e.g. sparse LD)

(e.g. Lloyd-Jones et al. 2019, Vilhjálmsson et al. 2015)

Multi-trait and multi-component Bayesian linear regression (BLR) models:

- handle multiple traits => use correlated trait information
- extend to multiple marker groups => use functional marker information
- extend to joint analysis of individual level and summary data

(Sørensen et al. 2015, Cheng et al. 2018)

Methods for Computing Polygenic Risk Scores

Multi marker effects with regularization using correlated trait information:

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X} + \mathbf{I} \otimes \mathbf{B}^{-1}\mathbf{E})^{-1}\mathbf{X}'\mathbf{y}$$ (multiple trait BLR)

Genetic (co)variances

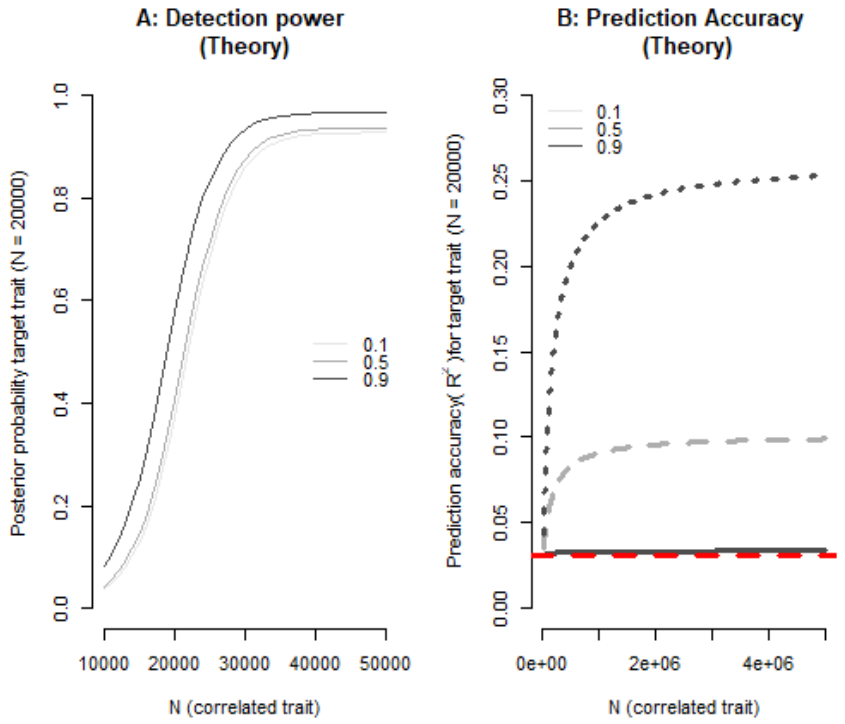
Residual (co)variance

$$\mathbf{B} = \begin{bmatrix} \sigma^2_{b_1} & \boldsymbol{\sigma^2_{b_{12}}} \\ \boldsymbol{\sigma^2_{b_{21}}} & \sigma^2_{b_2} \end{bmatrix} \qquad \mathbf{E} = \begin{bmatrix} \sigma^2_{e_1} & \sigma^2_{e_{12}} \\ \sigma^2_{e_{21}} & \sigma^2_{e_2} \end{bmatrix}$$

If genetic covariance ($\boldsymbol{\sigma^2_{b_{21}}}$) is different from zero information can be borrowed across traits!

Methods for Computing Polygenic Risk Scores

The value (increase in prediction accuracy or detection power) of using correlated trait information depend on:



- number of observations
- heritability of the traits
- genetic correlation between traits
- disease prevalence

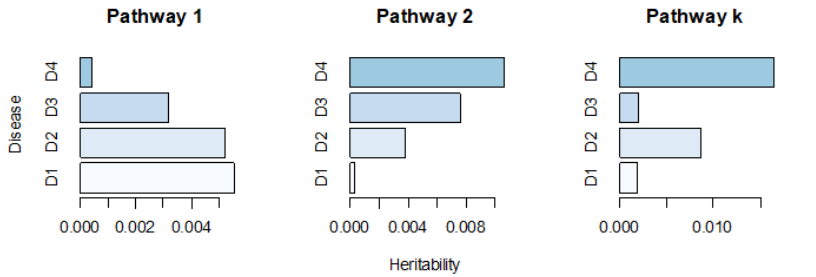
In particular, low heritability traits with small number of observations will benefit most from a multiple trait analysis with a high heritability traits with large number of observations.

Rheenen et al 2019

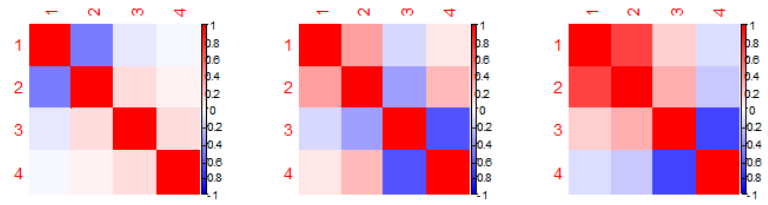
Panel A: MT-BLR, marker effect b=0.025 for two continous traits, r_g=0.1,0.3,0.9
Panel B: MT-BLUP, h^2=0.3 for two continous traits, r_g=0.1,0.3,0.9

Methods for Computing Polygenic Risk Scores

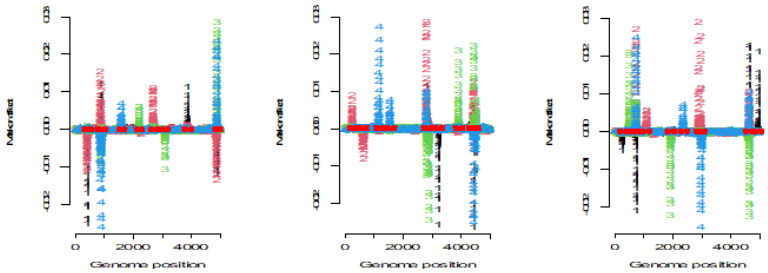
MT-BLR models is used for simultaneously estimation of genetic parameters (heritability and correlation) and gene mapping:



Heritability for pathways may differ across traits (e.g. mapping of multiple genetic variants each with small effects)



Genetic correlation between traits may differ across pathways (”How joint effects of multiple markers correlates across traits”)



Mapping of genetic variants with moderate to large effects (increased detection power)

Methods for Computing Polygenic Risk Scores

A range of biological relevant multiple trait BLR models can be specified including:

- correlated traits and/or diseases in the same population
- correlated traits and/or diseases across populations
- correlated traits and/or diseases across environments

- use information on functional marker groups to detect functional related genetic variants each contributing small effects across traits, environments or populations

- use summary statistics or individual level genotype and phenotype data (i.e. combine data from GWAS consortia and Biobanks)

The multi-trait and multi-component trait BLR models are implemented using fast and memory efficient algorithms in C++ and made publicly available in the R software package qgg (https://cran.r-project.org/web/packages/qgg).

Outline

Introduction to Polygenic Risk Scoring

Data used for Polygenic Risk Scores
- Training/Validation
- Inivididual level or summary statistic data

Methods for Computing Polygenic Risk Scores
- Standard approach based on LD pruning and thresholding (P+T)
- Bayesion approach using shrinkage estimation (e.g. Ldpred, BayesR)
- Multiple trait approaches

Methods for Evaluating Polygenic Risk Scores
- Population or individual level measures
- Quantitative and binary traits
- Expected accuracies

Clinical Utility of Polygenic Risk Scores

Methods for Evaluating Polygenic Risk Scores

After fitting a prediction model in a training sample, we can measure success using a validation sample for which the phenotype is available (but these individuals must not form part of the training population).

Suppose that in a validation sample of size k we have the predicted values $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_k]$, and the observed values $y = [y_1, \ldots, y_k]$.

- The closer the $\hat{y}_i$ to the y_i the better, but there are many ways to measure closeness.
- The different metrics are typically highly correlated but they are not equivalent.
- Different measures for predictive accuracy for continuous traits and binary traits

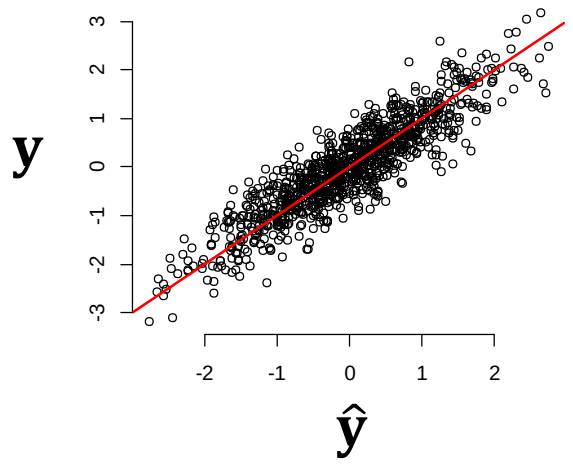
Methods for Evaluating Polygenic Risk Scores

Some measures of predictive accuracy for a continuous outcome includes the correlation $\mathrm{cor}(\hat{\mathbf{y}}, \mathbf{y})$ or else the squared correlation (which is related to variance explained in a regression):

$$R^2 = \mathrm{cor}(\hat{\mathbf{y}}, \mathbf{y})^2$$

The slope and intercept from a regression:

$$\mathbf{y} = \mathrm{intercept} + \mathrm{slope} \cdot \hat{\mathbf{y}}$$



The mean absolute error or the (root) mean square error:

$$\mathrm{MSE} = \frac{1}{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{k}} (\hat{\mathrm{y}}_{\mathrm{i}} - \mathrm{y}_{\mathrm{i}})^2$$

Methods for Evaluating Polygenic Risk Scores

Some measures of predictive accuracy for a binary outcome includes the the area under the receiver operating characteristic curve (AUC):

$$\mathrm{AUC} = \frac{1}{\mathrm{n_{control}}}\left(\bar{\mathrm{r}}_{\mathrm{case}} - \frac{\mathrm{n_{case}}}{2} - \frac{1}{2}\right)$$

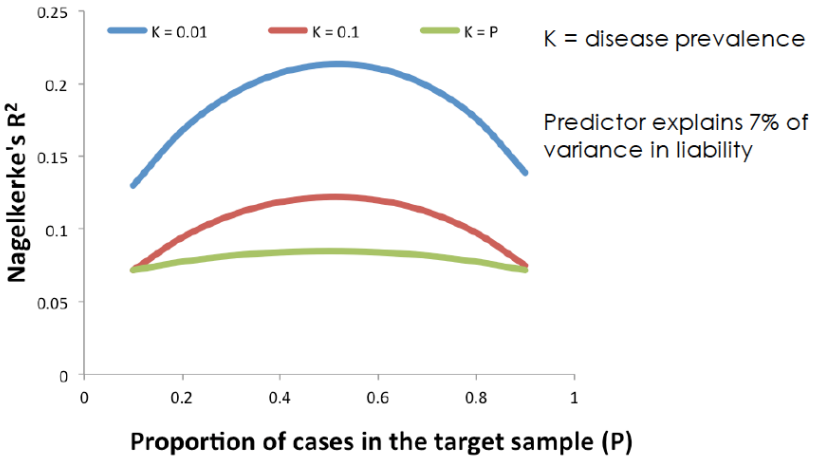
(Wray et al 2010)

- $\bar{\mathrm{r}}_{\mathrm{case}}$ is the average rank of cases
- $\mathrm{n_{case}}$ and $\mathrm{n_{control}}$ are the number of case and controls
- takes a value from 0.5 to 1 where 1 is optimal
- can be interpreted as the probability that a randomly selected case will have a higher polygenic risk score than a randomly selected control

Methods for Evaluating Polygenic Risk Scores

Nagelkerke's R^2_{NAG} for logistic regression for case-control disease status:

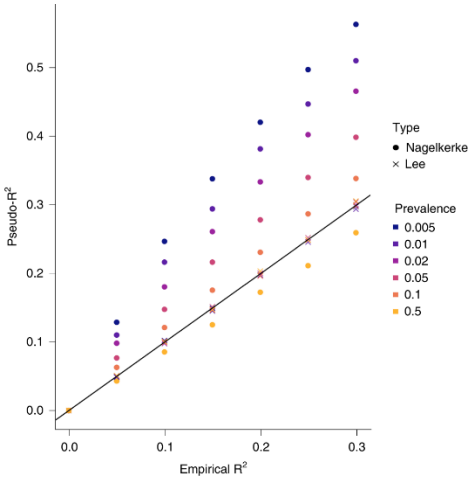
$$R^2_{NAG} = \frac{1-e^{-LR/n}}{1-e^{-(-2L_0)/n}}$$



- LR is the likelihood ratio comparing two nested logistic regression models
- L_0 is the log-likelihood of a model neglecting the GRS
- n is the number of observations

But R^2_{NAG} has an unfortunate property of depending on disease prevalence and proportion of cases

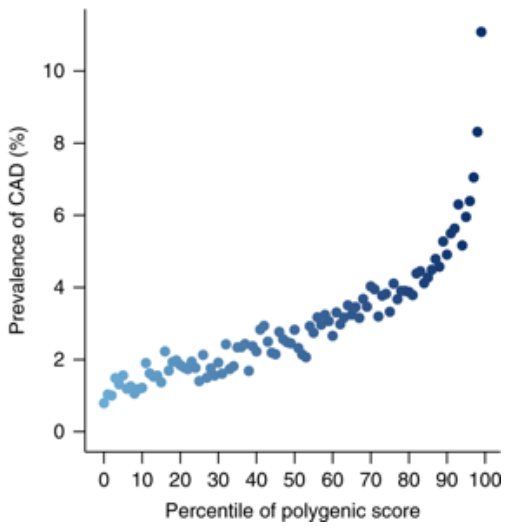
Better alternative is to use R^2 on a liability scale: Lee & Wray 2013, Wray et al. 2013, Choi et al. 2020)

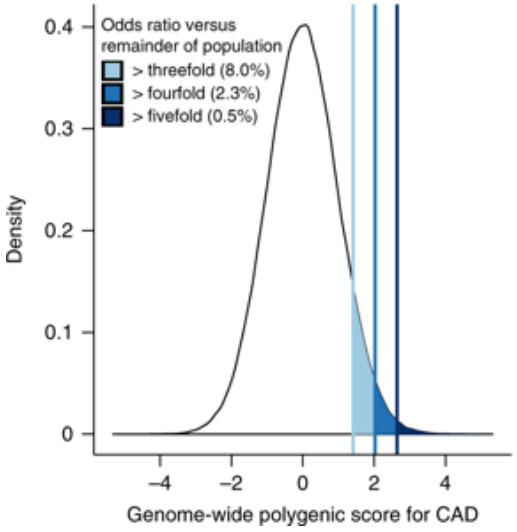


Methods for Evaluating Polygenic Risk Scores

The proportion of the population that has a k-fold increased odds (k = 2, 3, ...), compared to the population disease risk.

Odds ratio of disease for an individual in the top PRS decile (or other quantiles) compared to individuals in a different part of the PRS distribution.





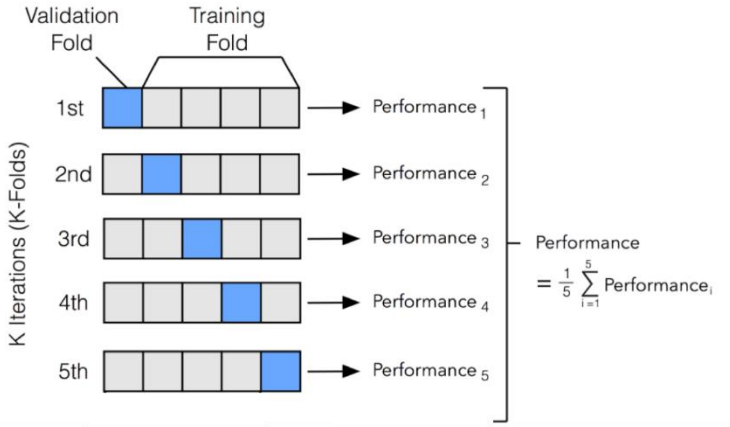
Methods for Evaluating Polygenic Risk Scores

Cross validation (CV) is a statistical procedure where prediction accuracy is estimated by holding back a fraction of the training population:

- held-back individuals may be resampled at random each time, or sampled systematically so that each individual is a member of the test sample a fixed number of times (k-fold CV)

- predictive accuracy then tends to be understated because the full training sample is not used to fit the model

- individuals in validation and training populations must be independent otherwise prediction accuracy will be inflated

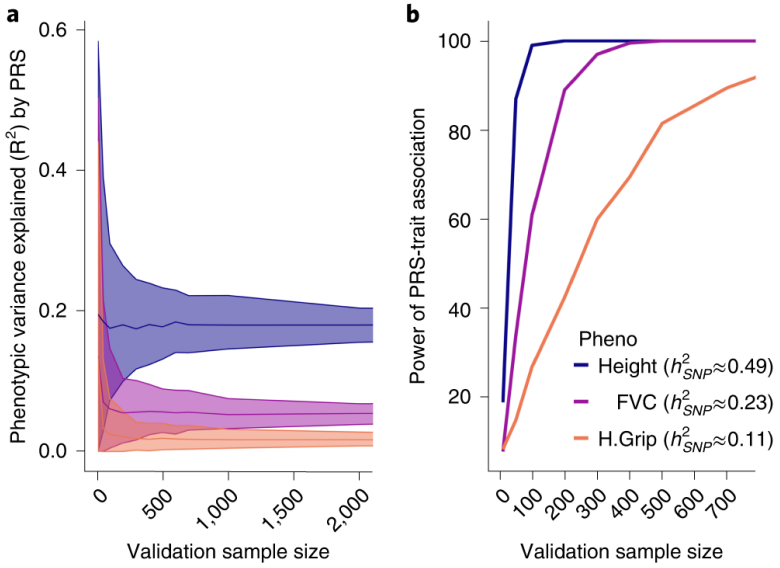


Methods for Evaluating Polygenic Risk Scores

Optimal design of cross validation procedure (i.e. how should we optimally split n_T and n_V):

n_T increase => increase $E[R_V^2]$
n_V increase => no influence $E[R_V^2]$

n_T increase => increase power$[R_V^2]$
n_V increase => increase power$[R_V^2]$



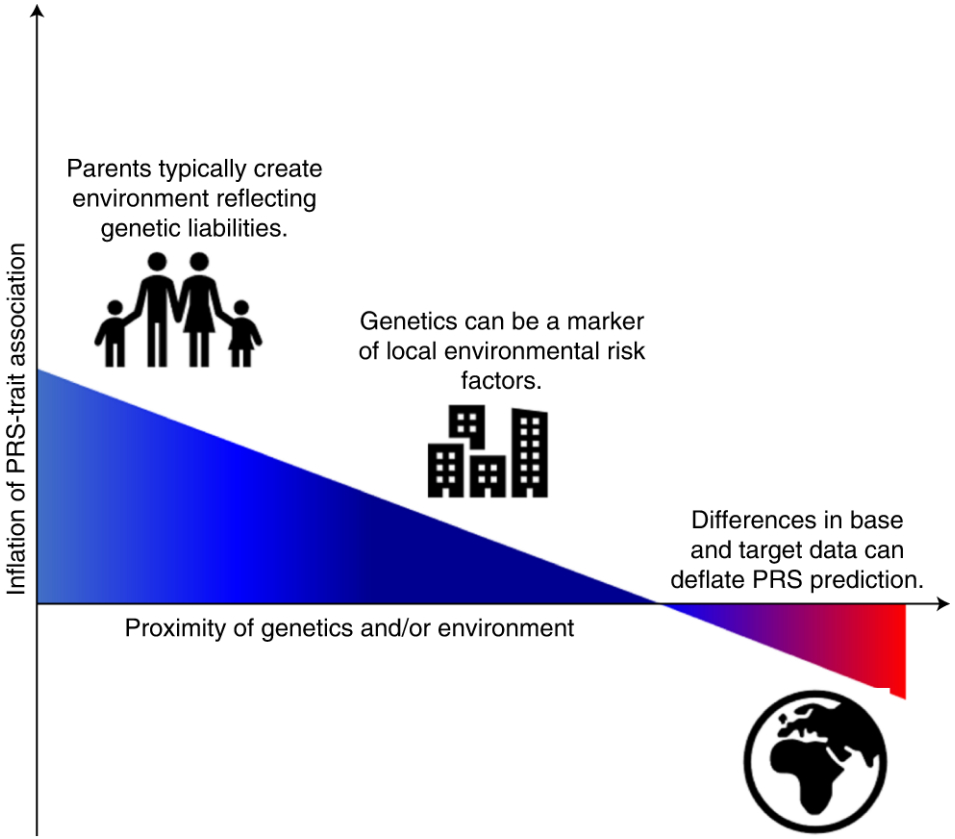
To maximize power, split discovery & target equally
To maximize prediction accuracy we should maximize n_T

Similar for AUC

(Dudbridge 2013, Choi et al. 2020)

Methods for Evaluating Polygenic Risk Scores

Potential sources of biases in the evaluation of polygenic risk scores:

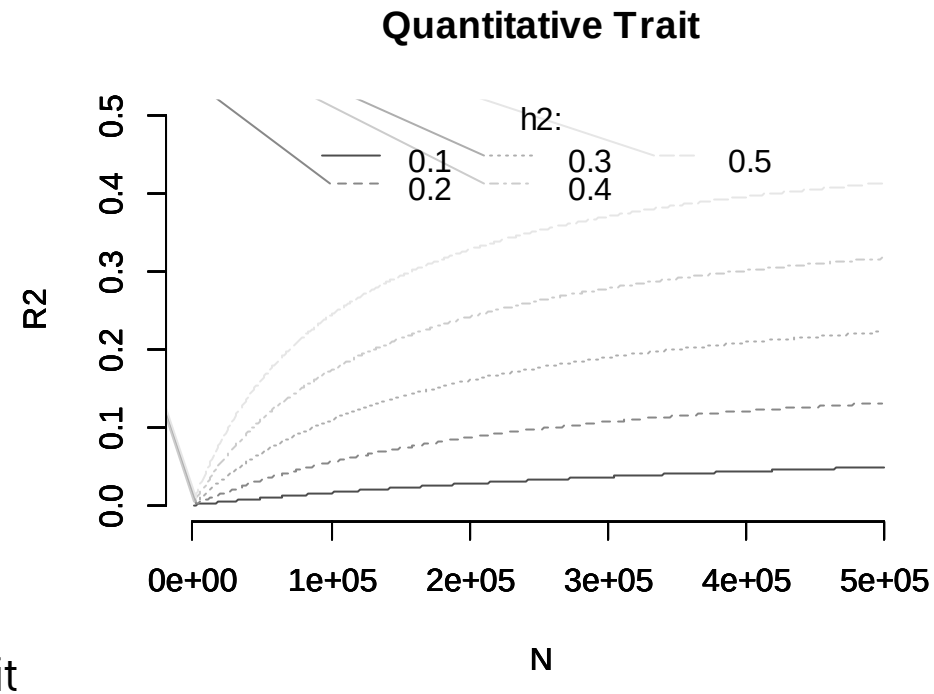


(Choi et al 2020)

Methods for Evaluating Polygenic Risk Scores

Expected R^2 for the phenotypic variation explained by SNPs for a quantitative trait:

$$R^2 = h^2 \frac{bh^2}{bh^2 + \frac{M_e}{N}}$$



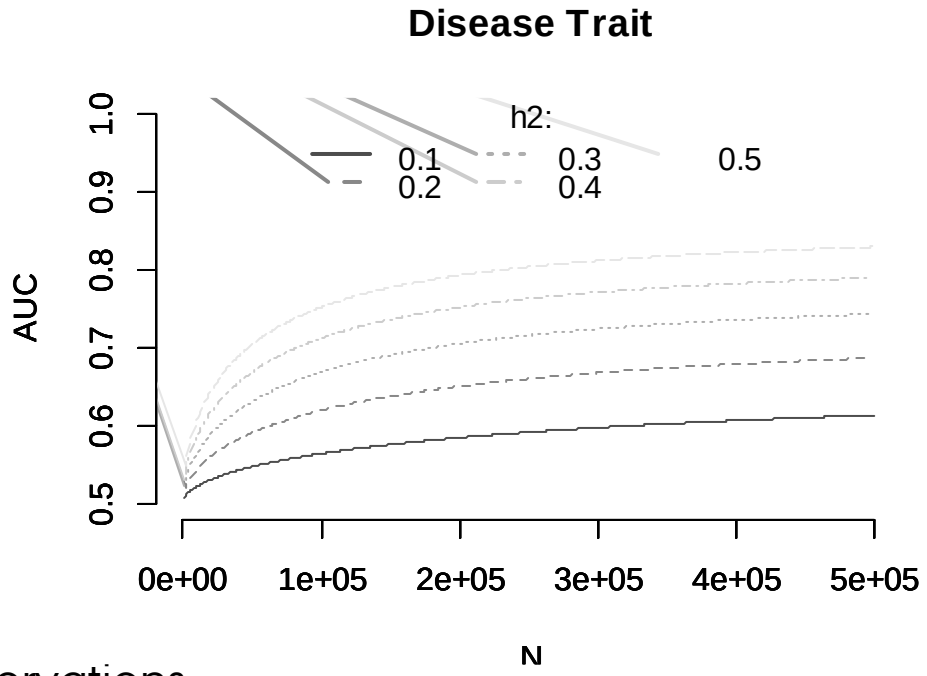
h^2 is the heritability of the trait
N is the number of phenotypic observations
M is the number of markers used in the analysis
M_e is the effective number of chromosome segments
$b = \frac{M}{M+M_e}$ is the proportion of genetic variance captured by markers

(Lee et al. 2017a,
Lee et al. 2017b,
van Rheenen et al. 2019)

Methods for Evaluating Polygenic Risk Scores

Expected AUC explained by SNPs for a disease trait:

$$\mathrm{AUC} = f(\mathrm{h}^2, N, M, M_e, K, P)$$



h^2 is the heritability of the trait
N is the number of phenotypic observations
M is the number of markers used in the analysis (1M)
M_e is the effective number of chromosome segments (50K)
K is prevalence of target trait (0.1)
P is case-control proportion of target trait (0.5)

(Lee et al. 2017a,
Lee et al. 2017b,
van Rheenen et al. 2019)

Outline

Introduction to Polygenic Risk Scoring

Data used for Polygenic Risk Scores
- Training/Validation
- Inivididual level or summary statistic data

Methods for Computing Polygenic Risk Scores
- Standard approach based on LD pruning and thresholding (P+T)
- Bayesion approach using shrinkage estimation (e.g. Ldpred, BayesR)
- Multiple trait approaches

Methods for Evaluating Polygenic Risk Scores
- Population or individual level measures
- Quantitative and binary traits
- Expected accuracies

Clinical Utility of Polygenic Risk Scores

Clinical Utility of Polygenic Risk Scores

Precision health account for differences in people's genes, environments and lifestyles and formulates treatment and prevention strategies based on patients' unique backgrounds and conditions:

- Aims to predict disease risk or response to medical treatment based on an individual's DNA profile (or other types of biomarkers) and other risk factors
- Potential to improve decision-making in health care systems which could improve patient health and lower health care costs
- It is not a new concept, but growing amounts of genetic and health care data and development of sophisticated analytical tools are bringing it closer to the clinics
- Because we inherit our unique pattern of DNA variation at birth, genetics (e.g. polygenic risk scoring) has a special role to play in disease risk prediction and in patient stratification for medical treatment.

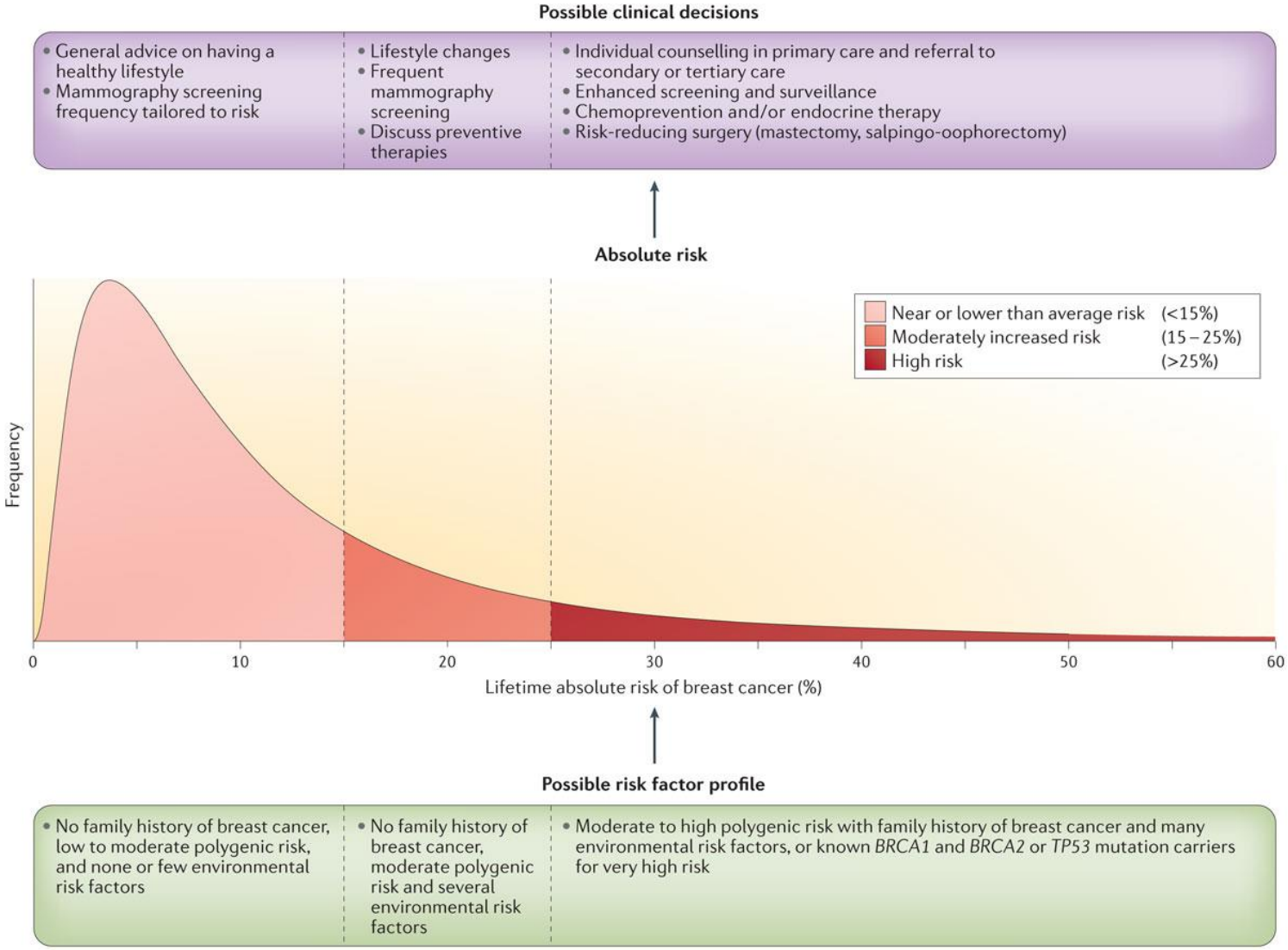
Clinical Utility of Polygenic Risk Scores

Lifeline of the potential relevance of polygenic risk scores showing points through disease trajectory where polygenic risk scores have the potential to impact clinical care:

- As PRS remains constant over the life course it could be used to guide disease prevention earlier in life before standard risk factors have an appreciable impact
- Given the plummeting costs of genetic tests current disease risk prediction tools could be enhanced with the addition of polygenic risks

(Lewis & Vassos. 2020)

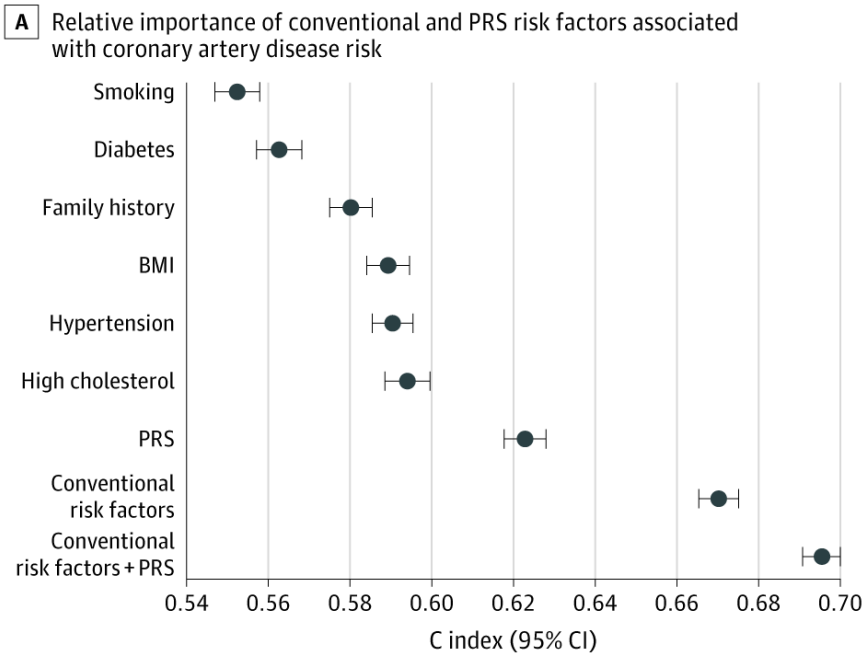
Clinical Utility of Polygenic risk Scores

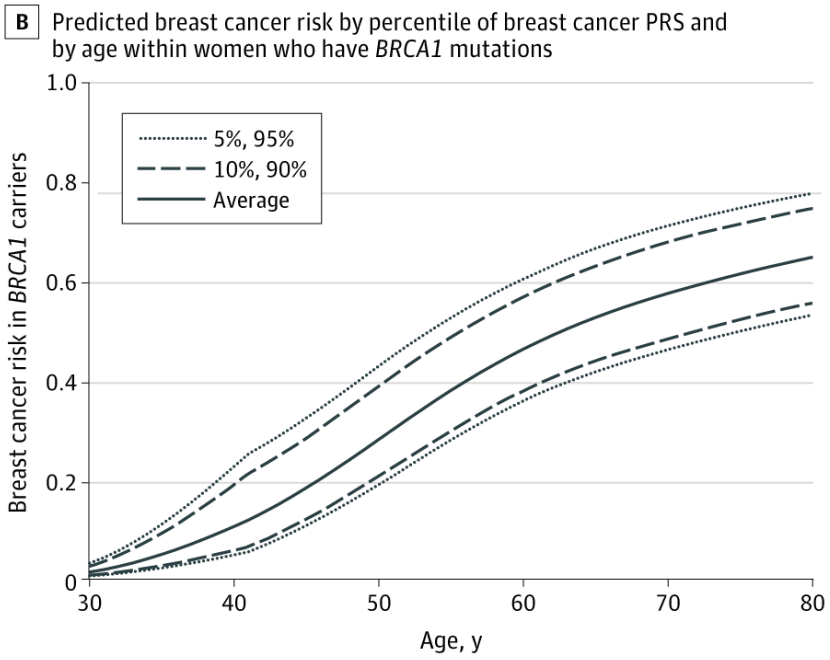


(Chatterjee et al. 2016)

Clinical Utility of Polygenic Risk Scores

Relative importance of conventional and polygenic risk scores:

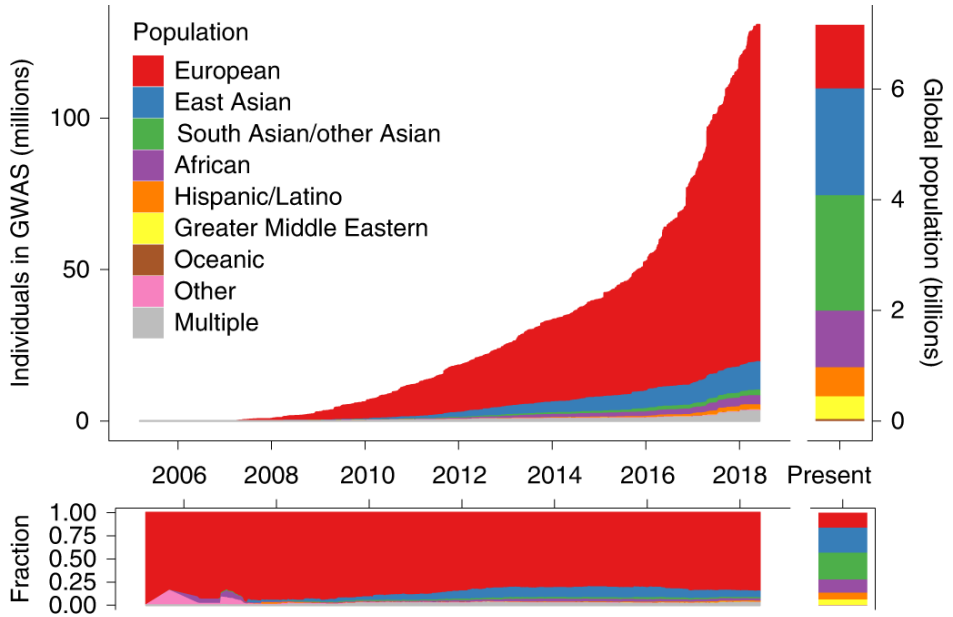




(Wray et al 2021)

Clinical Utility of Polygenic Risk Scores

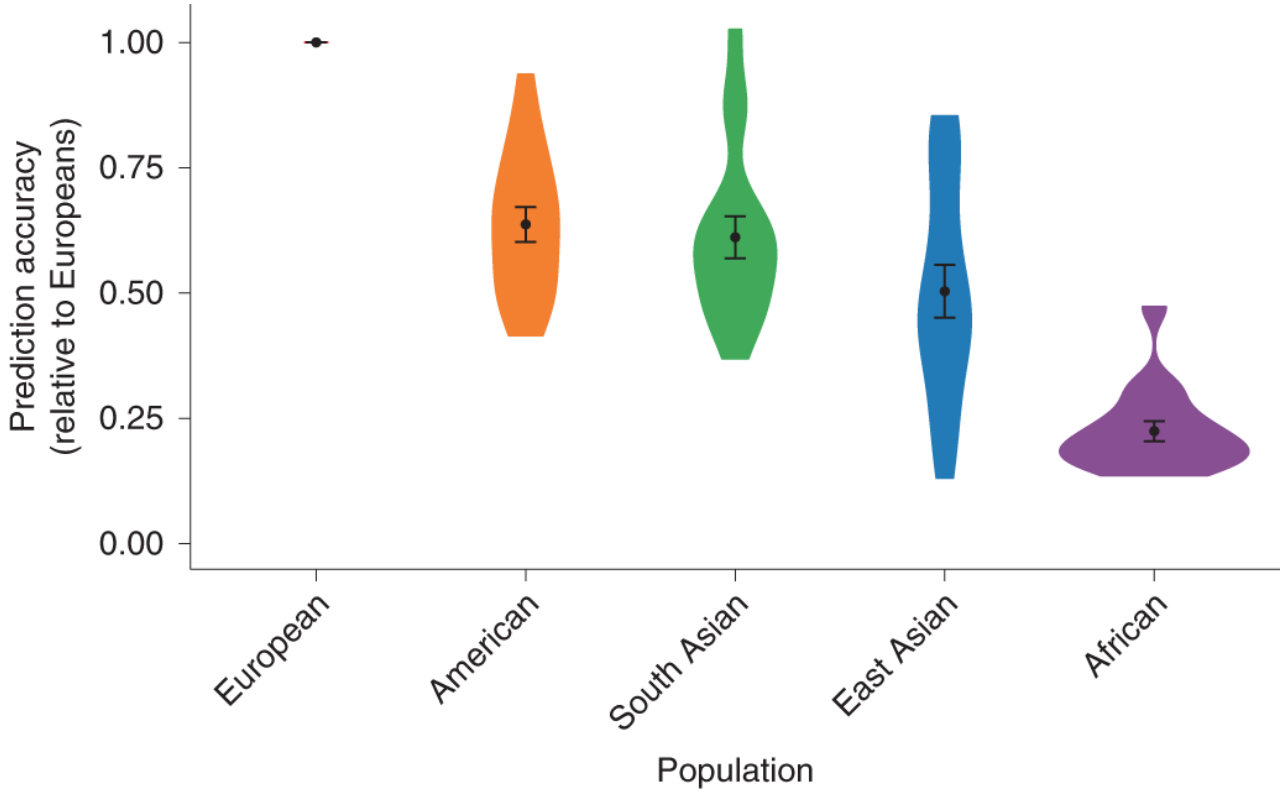
Ancestry of GWAS participants over time, as compared with the global population:



(Martin et al 2019)

Clinical Utility of Polygenic Risk Scores

Prediction accuracy relative to European-ancestry individuals across 17 quantitative traits and 5 continental populations in the UKBB:



(Martin et al 2019)

Summary

Introduction to Polygenic Risk Scoring

Data used for Polygenic Risk Scores

- Training/Validation
- Inivididual level or summary statistic data

Methods for Computing Polygenic Risk Scores

- Standard approach based on LD pruning and thresholding (P+T)
- Bayesion approach using shrinkage estimation (e.g. Ldpred, BayesR)
- Multiple trait approaches

Methods for Evaluating Polygenic Risk Scores

- Population or individual level measures
- Quantitative and binary traits
- Expected accuracies

Clinical Utility of Polygenic Risk Scores

References

Dudbridge 2013:	10.1371/journal.pgen.1003348
Daetwyler et al. 2008:	10.1371/journal.pone.0003395
Lee et al. 2017:	10.1371/journal.pone.0189775
Lee & Wray 2013:	10.1371/journal.pone.0071494
de los Campos et al. 2013:	10.1534/genetics.112.143313
Cheng et al. 2018:	10.1534/genetics.118.300650
Wray et al. 2010:	10.1371/journal.pgen.1000864
Martin et al. 2020:	10.1038/s41588-019-0379-x
Riveros-Mckay et al. 2021:	10.1161/CIRCGEN.120.003304
Wray et al 2013:	10.1038/nrg3457
van Rheenen et al 2019:	10.1038/s41576-019-0137-z
Lloyd et al. 2019:	10.1038/s41467-019-12653-0
Vilhjalmson 2015:	10.1016/j.ajhg.2015.09.001
Sørensen et al. 2015	10.1093/genetics/201.2.NP
Rohde et al. 2020	10.1093/bioinformatics/btz955
Rohde et a.l 2021	10.3389/fmed.2021.711208
Maier et al. 2018:	10.1038/s41467-017-02769-6
Ma & Zhou 2021:	10.1038/s41467-017-02769-6
Choi et al. 2020:	10.1038/s41596-020-0353-1
Wray et al. 2021:	10.1001/jamapsychiatry.2020.3049

R package qgg
(psoerensen.github.io/qgg/)

qgg provides an infrastructure for efficient processing of large-scale genetic and phenotypic data including core functions for:

- BLUP/REML/BLR methods
- fitting linear (mixed) models
- estimating genetic parameters (heritability and correlation)
- genomic prediction (polygenic risk scoring)
- single marker association analysis
- gene set enrichment analysis

qgg handles large-scale data by taking advantage of:

- fast and memory efficient algorithms implemented using C++
- multi-core processing using openMP
- multithreaded matrix operations implemented in BLAS libraries (e.g. OpenBLAS, ATLAS or MKL)
- batch processing of genotype data stored in binary files (e.g. PLINK bedfiles)

Analyses workflow in R

```
# Prepare genotype information, quality control
Glist <- gprep(bed/bim/famfiles, task="prepare")    # summarise genotype information
rsids <- gfilter(Glist,excludeMAF=0.01)             # filter markers based on MAF, HWE,...

# Compute sparse LD matrices and ldscores
Glist <- gprep(Glist, rsids, ids, ldfiles, task="sparseld") # sparseLD using using markers in rsids

# Compute summary statistics
stat <- glma(y=y[train], X=X[train,], Glist=Glist)  # fit single marker regression model
stat <- qcStat(stat=stat, Glist=Glist)              # quality control of summary statistics

# BLR model analysis based on summary statistics    # estimate marker effects
fit <- gbayes( stat=stat, Glist=Glist, method="bayesR")  # and genetic parameters

# Genomic prediction
grs <- gscore(Glist=Glist, stat=fit$stat)           # compute genomic scores
acc(yobs=y[valid], ypred=grs[valid,], typeoftrait="binary")  # assess accuracy (e.g. AUC)

# Gene Set Enrichment analysis
gsea(stat=fit$stat, sets=sets)                      # marker set association statistics
```

?