

Methods and Tutorials for Building Polygenic Risk Scores 1

Course # 140.721

Ziqiao Wang

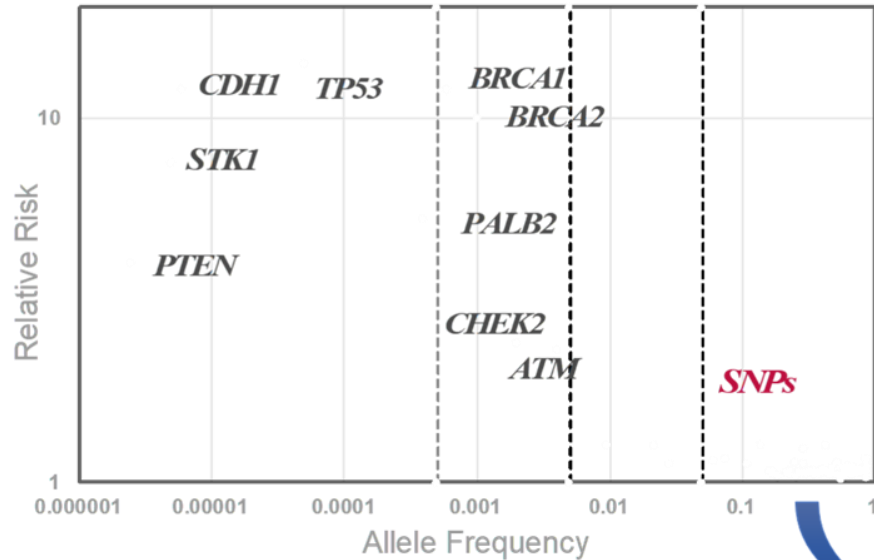
Department of Biostatistics
Johns Hopkins University

Genetic Architecture of Complex Diseases

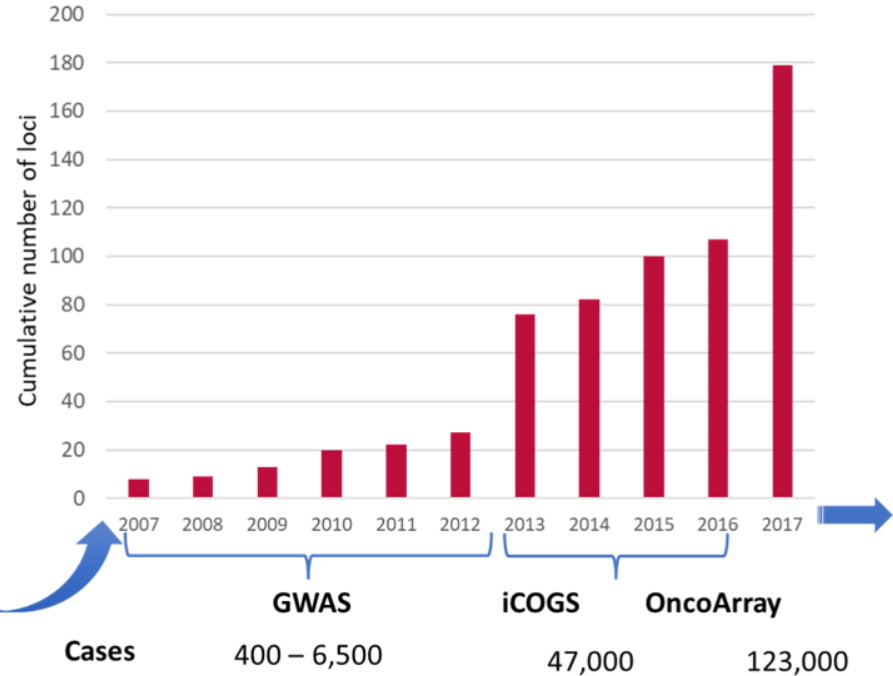
- Rare high-penetrant mutations identified through linkage studies
 - Individuals who carry the mutations are at high-risk, but the mutations explain a small fraction of cases (5-10%) in the general population
 - BRCA1/2 mutations for breast and ovarian cancer
 - P53 mutations for lynch syndrome
 - Individuals in highly affected families are recommended for testing for these mutations
- Common low-penetrant variants identified through GWAS
 - Complex traits are extremely polygenic with heritability defined by small additive effects of thousands to tens of thousands of genetics variants

Progress in defining genetic architecture of breast cancer

Breast cancer susceptibility alleles



Discoveries through 2007-17



Polygenic risk score (PRS) or polygenic score (PGS)

- Quantitative measure of the total genetic risk burden of the disease/trait over multiple susceptibility variants

Weighted average of the number of risk alleles

$$PRS = \beta_1 SNP_1 + \beta_2 SNP_2 + \dots + \beta_n SNP_n$$

Effect size

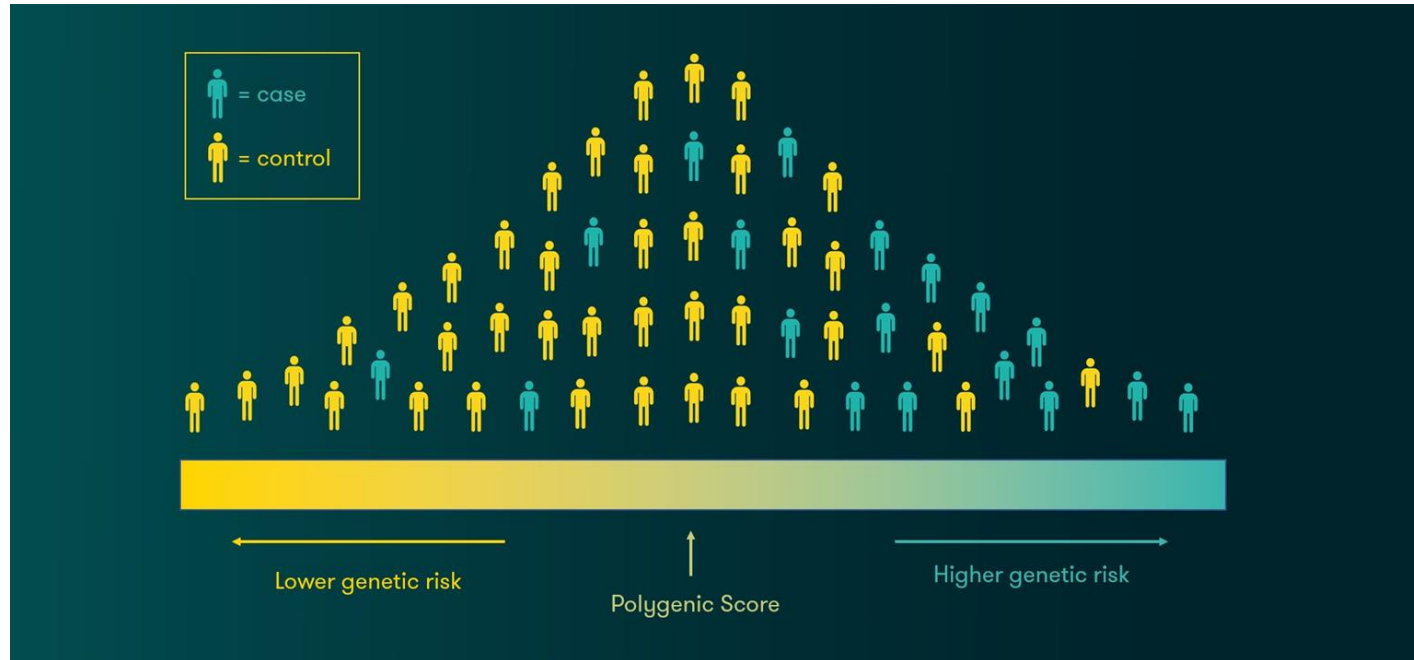
Number of risk alleles

Number of SNPs

Polygenic Score / Polygenic Risk Score (PGS/PRS) for Risk Stratification

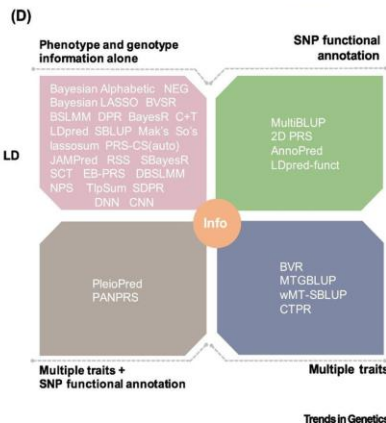
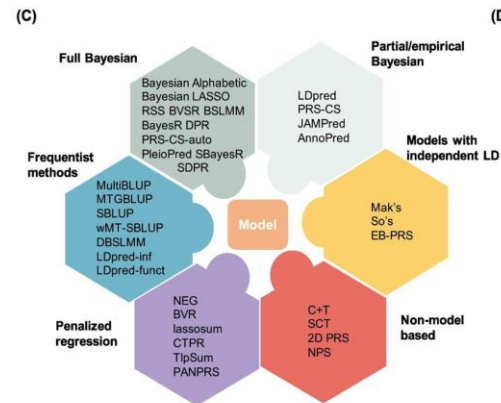
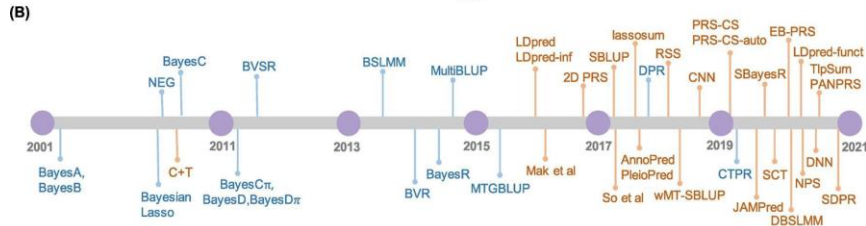
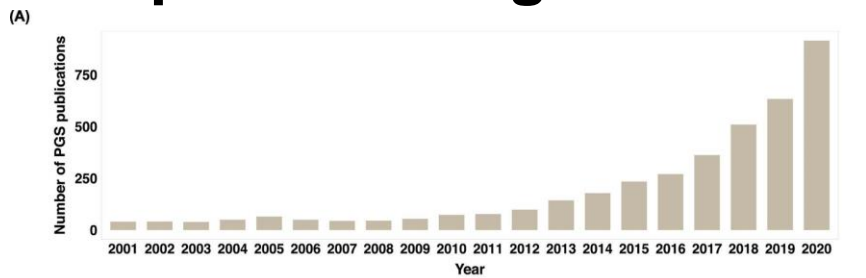
$$\text{PGS} = \sum_{j=1}^P G_j \hat{\beta}_j$$

← $\hat{\beta}_j$: Marginal effect of SNP j associated with outcome trait estimated from external GWAS



A multiple linear regression framework

$$y = X\beta + \epsilon$$



Most PGS methods can be viewed as making distinct **modeling assumptions** for the SNP effect size β in the model and rely on different algorithms to obtain the estimates $\hat{\beta}$

Ma et al., 2021 Genetic prediction of complex traits with polygenic scores: a statistical review

Outline

- Model-Free PRS algorithm
- Advanced PRS algorithm (Bayesian, penalized regression, deep learning PRS algorithms)
- Multi-ethnic PRS algorithm (CT-SLEB, PROSPER, PRSCSx, weighted by race, Bayesian methods, continuum ancestry PRS)
- Incorporating SNP functional annotations (to improve transferability of multi-ancestry PRS – presented at ASHG 2024 – extension of GAUDI for admixed population, etc)
- Interpretations and Applications of PGS, risk prediction, MR analysis

Definitions

- GWAS summary-statistics: Estimates of association coefficients, standard errors, p-value, z-statistics from one-SNP-at-a-time analysis
 - Easily available from GWASCatalog
- Linkage disequilibrium (LD): Correlation across SNPs
 - SNPs within small regions of genome will be in LD due to lower rate recombination
 - Long range LD can arise due to population structure
- Training dataset: Large GWAS studies for building PRS
- Test/validation dataset: Smaller GWAS for tuning PRS parameters or/and evaluating the performance of the final PRS

PRS with Liberal Threshold (Clumping + thresholding)

- Initially PRS were typically built with independent SNPs that reach genome-wide significance
 - Most parsimonious model
- Many loci with small effects are currently undetectable at genome-wide significance level
 - Can we take advantage of them to improve risk prediction?

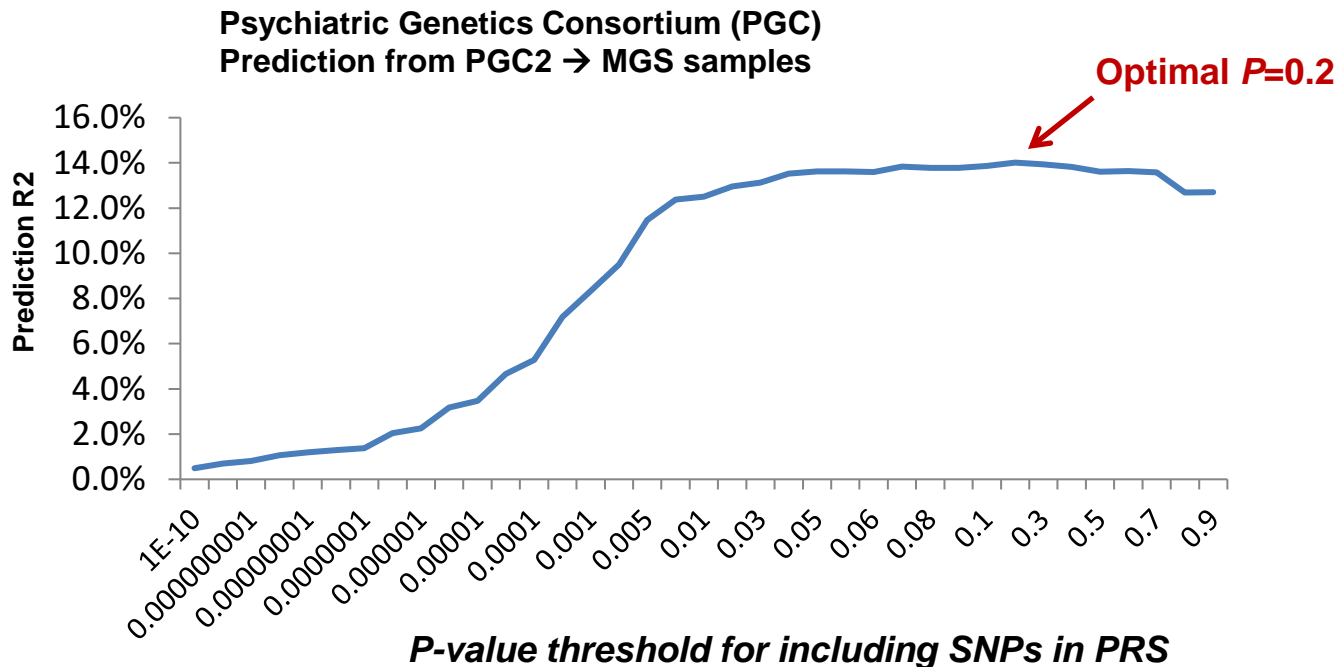
$$\begin{aligned} PRS_i(\alpha) &= \sum_k \hat{\beta}_k G_{ik} I(p_k < \alpha) \\ &= \sum_k \hat{\beta}_k G_{ik} I(|\hat{\beta}_k| > \lambda) \end{aligned}$$

- Needs stringent LD-pruning!
 - Correlated SNPs not containing independent signals add “noise”
 - association coefficients are not adjusted for LD

Clumping and thresholding (C+T) method

- Build a series of PRS using nested sets of **independent SNPs** with increasingly relaxed significance thresholds and
- Chose the optimal PRS that performs the best in an independent dataset

PRS Analysis of Schizophrenia



Purcell et al., Nature, 2009
Shi et al., Nature, 2009
Ripke et al., Nature, 2014

Towards advanced methods

- Ad-hoc removal of correlated SNPs may lead to removal of secondary signals around sentinel SNPs
- Effect-size estimation is not optimal from bias-variance tradeoff perspective
- Ideally, more principled methods which can fit “whole-genome” model will do better
 - Accounts for LD
 - Perform “shrinkage” estimation for association coefficients
 - Can’t expect magic as **sample size is the biggest rate limiting factor**

Methods related to improving PRS

- Jointly model genetic markers across the genome to make full use of the available information while accounting for local linkage disequilibrium (LD) structures
- Accommodate varying effect size distributions across complex traits and diseases, from highly polygenic genetic architectures (e.g., height and schizophrenia), to a mixture of small effect sizes and clusters of genetic loci that have moderate to larger magnitudes of effects (e.g., autoimmune diseases and Alzheimer's disease)
- Produce prediction from **GWAS summary statistics** without access to individual-level data
- Retain computational scalability

Bayesian methods

- Genome-wide multivariable model

$$g\{E(Y)\} = \sum_{m=1}^M \beta_m \times G_m$$

Link function Mean outcome Association coef SNP genotype

- Prior Distribution and hyperparameter (tuning parameter)

$$\beta \sim \pi(\theta)$$

Prior distribution Hyperparameter May incorporate functional annotation

- Bayesian Inference

$$\hat{\beta}_m(\theta) = E\{\beta_m | \text{Data}, \theta\}$$

- θ may be estimated from training data or “tuned” based on test-sample
- All SNPs may have non-zero weight

Different Choice of Priors

Method	Prior	Formula	Reference
LDpred	Spike and slab	$\beta_i \sim_{iid} \begin{cases} N\left(0, \frac{h_s^2}{Mp}\right) & \text{with probability } p \\ 0 & \text{with probability } (1-p), \end{cases}$	Vilhjálmsón, Bjarni J., et al. "Modeling linkage disequilibrium increases accuracy of polygenic risk scores." <i>The American Journal of Human Genetics</i> 97.4 (2015): 576-592.
SBayesR	Spike and slab, replace normal with mixture normal	$\beta_j \pi, \sigma_\beta^2 = \begin{cases} 0 & \text{with probability } \pi_1, \\ \sim N(0, \gamma_2 \sigma_\beta^2) & \text{with probability } \pi_2, \\ \vdots & \\ \sim N(0, \gamma_C \sigma_\beta^2) & \text{with probability } 1 - \sum_{c=1}^{C-1} \pi_c, \end{cases}$	Lloyd-Jones, Luke R., et al. "Improved polygenic prediction by Bayesian multiple regression on summary statistics." <i>Nature Communications</i> 10.1 (2019): 1-11.
PRS-CS	continuous shrinkage prior (global-local scale mixtures of normals)	$\beta_j \psi_j \sim N(0, \phi \psi_j), \quad \psi_j \sim g,$ g: an absolutely continuous density function, in contrast to a discrete mixture of densities.	Ge, Tian, et al. "Polygenic prediction via Bayesian regression and continuous shrinkage priors." <i>Nature Communications</i> 10.1 (2019): 1-10.
DPR	Dirichlet process	$\beta_i \sim \sum_{k=1}^{+\infty} \pi_k N(0, \sigma_k^2),$ $\pi_k = \nu_k \prod_{l=1}^{k-1} (1 - \nu_l), \nu_k \sim \text{Beta}(1, \lambda),$	Zeng, Ping, and Xiang Zhou. "Non-parametric genetic prediction of complex traits with latent Dirichlet process regression models." <i>Nature Communications</i> 8.1 (2017): 1-11.

Shrinkage with Different Priors

- Infinitesimal model: $pr(\beta^J) \sim N(0, \sigma^2)$ (single normal distribution)

- Independent SNPs:

$$E\{\beta^{(J)} | \hat{\beta}\} = \frac{h_g^2}{h_g^2 + M/N} \hat{\beta}, \quad \hat{\beta} = \text{vector of GWAS summary statistics}$$

- SNPs in LD

$$E\{\beta^{(J)} | \hat{\beta}\} = \left(\frac{M}{Nh_g^2} I + D\right)^{-1} \hat{\beta}, \quad D = \text{matrix of LD (correlation) coefficients}$$

- Spike and slab model (LD-pred)

- Independent SNPs

$$E\{\beta^{(J)} | \hat{\beta}\} = \frac{h_g^2}{h_g^2 + M\pi_1/N} \hat{\beta} \cdot \hat{p}_1, \quad \hat{p}_1 = \text{vector of posterior prob of association}$$

- SNP in LD

Analytic approximation not available, evaluated using MCMC Gibbs sampler

Penalized regressions

- Also builds genome-wide multivariate model accounting for LD across SNPs
- Performs shrinkage through incorporation of penalty function and associated tuning parameters
 - Closely related to prior distribution and hyperparameters
 - Certain popularly used penalty functions, such as Lasso and Elastic nets can produce sparse solutions (parsimonious PRS)
 - Typically requires validation sample for selection of tuning parameters

Penalized regressions using summary statistics: Lassosum

Individual-level data:
$$f(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) + 2\lambda ||\beta||_1^1$$
$$= \mathbf{y}^T \mathbf{y} + \beta^T \mathbf{X}^T \mathbf{X} \beta - 2\beta^T \mathbf{X}^T \mathbf{y} + 2\lambda ||\beta||_1^1$$

Summary statistics:
$$f(\beta) = \mathbf{y}^T \mathbf{y} + (1 - s)\beta^T \mathbf{X}_r^T \mathbf{X}_r \beta - 2\beta^T \mathbf{r}$$
$$+ s\beta^T \beta + 2\lambda ||\beta||_1^1,$$

LD from a reference panel

GWAS summary statistics

Becomes an elastic net problem!

Development of PRSs for breast cancer

Association between PRS and breast cancer risk in the validation set

PRS	OR per SD	95% CI	AUC
77-SNP (old)	1.49	(1.44-1.56)	60.3
313-SNP ($p < 10^{-5}$)	1.65	(1.59-1.72)	63.0
3,820-SNP (Lasso)	1.71	(1.64-1.79)	63.6

The 313-SNP PRS performs better than LD-pred derived “6 million”-SNP PRS reported by Khera et al (AUC=0.642 vs 0.627 in UKBiobank)

Moving beyond additive models: Deep learning based PRS

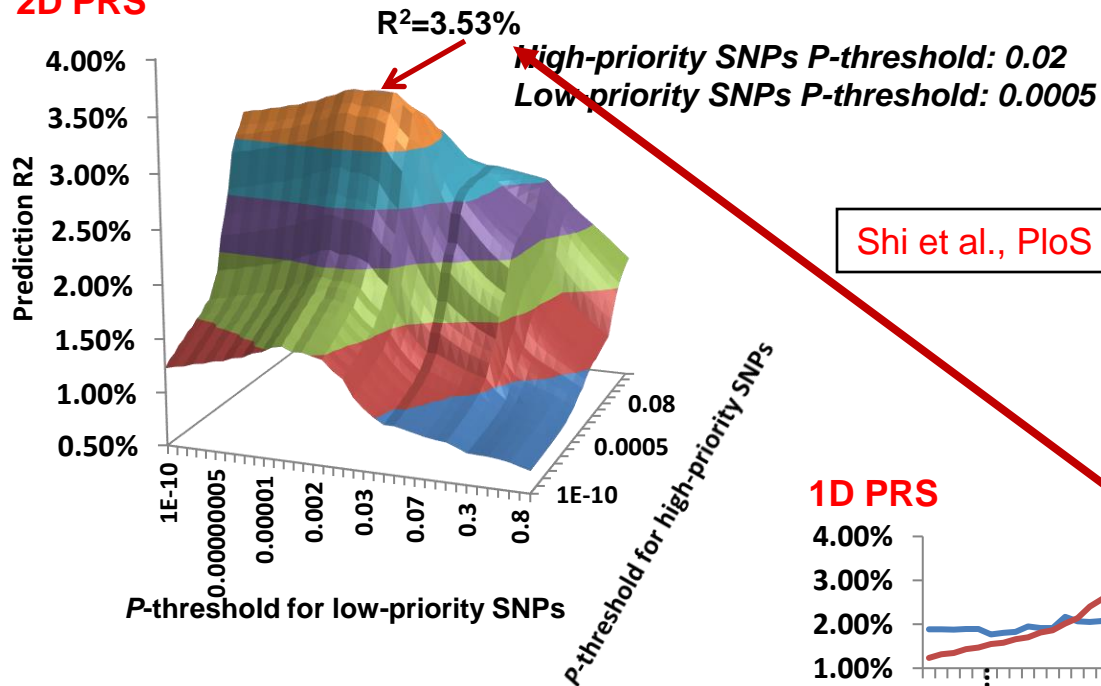
- Neural-network (NN) based deep learning has emerged as a method of intense interest to model complex, nonlinear phenomena, which may be adapted to exploit gene-gene (GxG) and gene-environment (GxE) interactions.

$$Y = G + E + G' \times E' + G' \times G' + E' \times E' + e$$

- Joint tagging effects/haplotype-effects: technical artefact, not indicate an encoding of biological processes.
 - Example: False positive epistatic effects caused by two ‘interacting’ variants imperfectly and differentially tagging causal variants which had not been genotyped
- Limited usefulness of NNs for PGS for common traits, confounded by joint tagging effects and low levels of nonlinearity

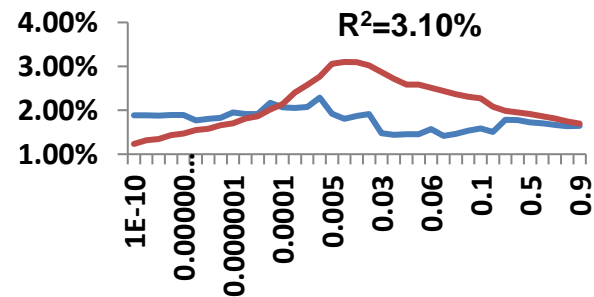
Prioritization of SNPs based on Functional and Annotation Information May Lead to Improved PRS: Type 2 Diabetes

2D PRS



Shi et al., PloS Genetics, 2017

1D PRS

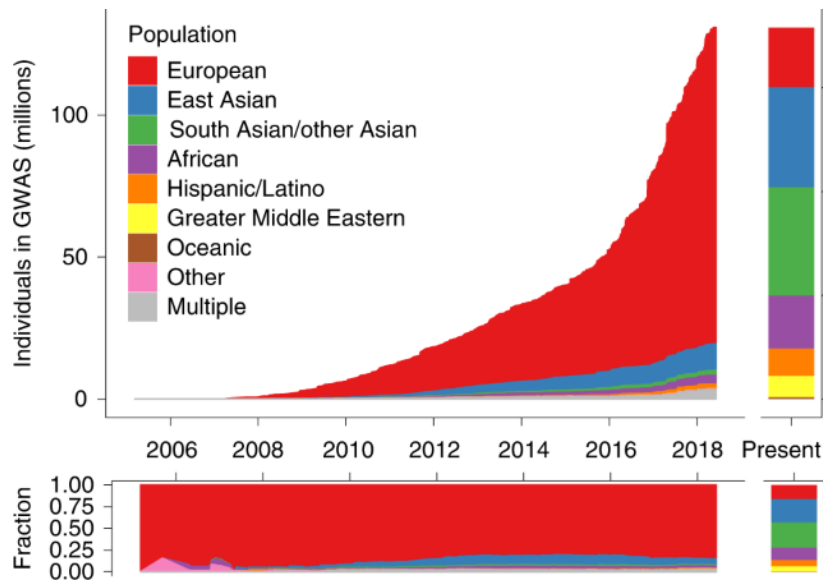


2D PRS with LASSO correction
eSNPs/meSNPs and H3K4me3 in islet cell line

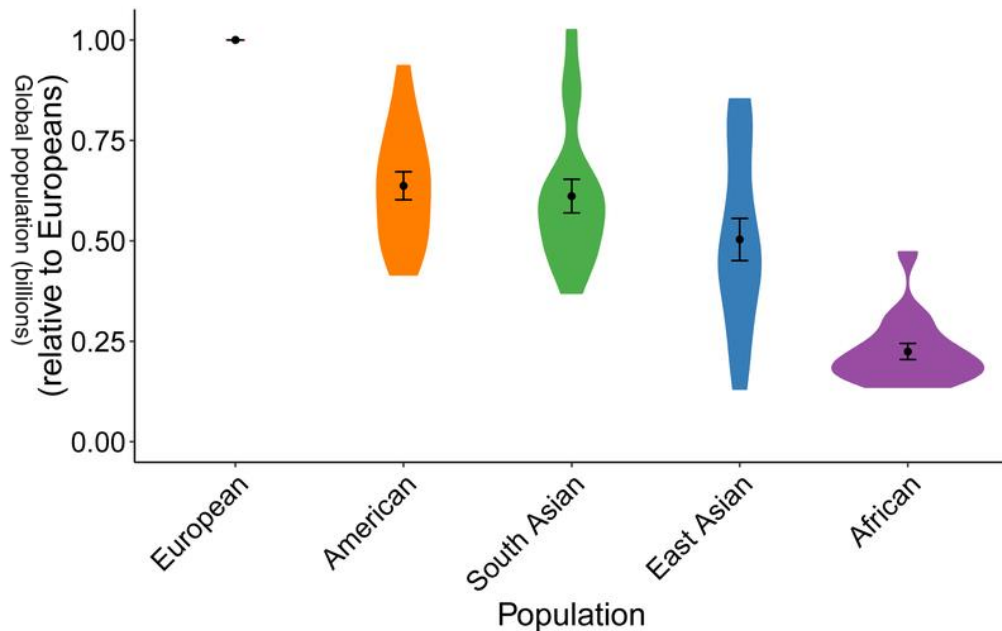
Polygenic Prediction in Diverse Population

Euro-centric bias in genetic studies

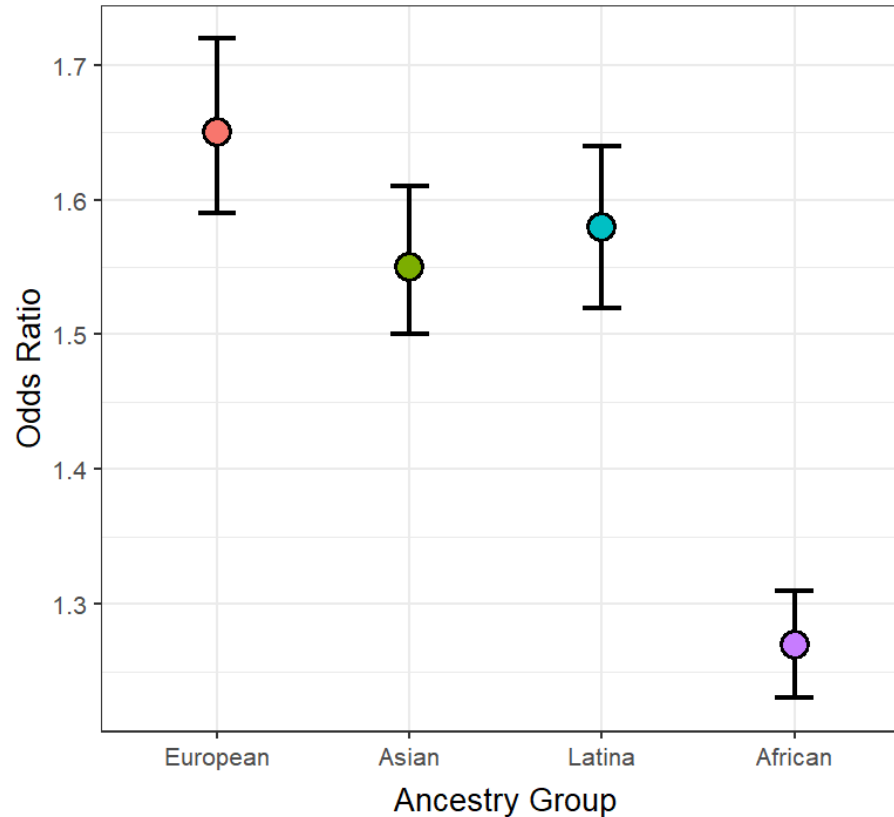
Sample size by continental groups



Predictive performance of PRS across traits



Performance of 313-SNP PRS for BrCA prediction in diverse population groups



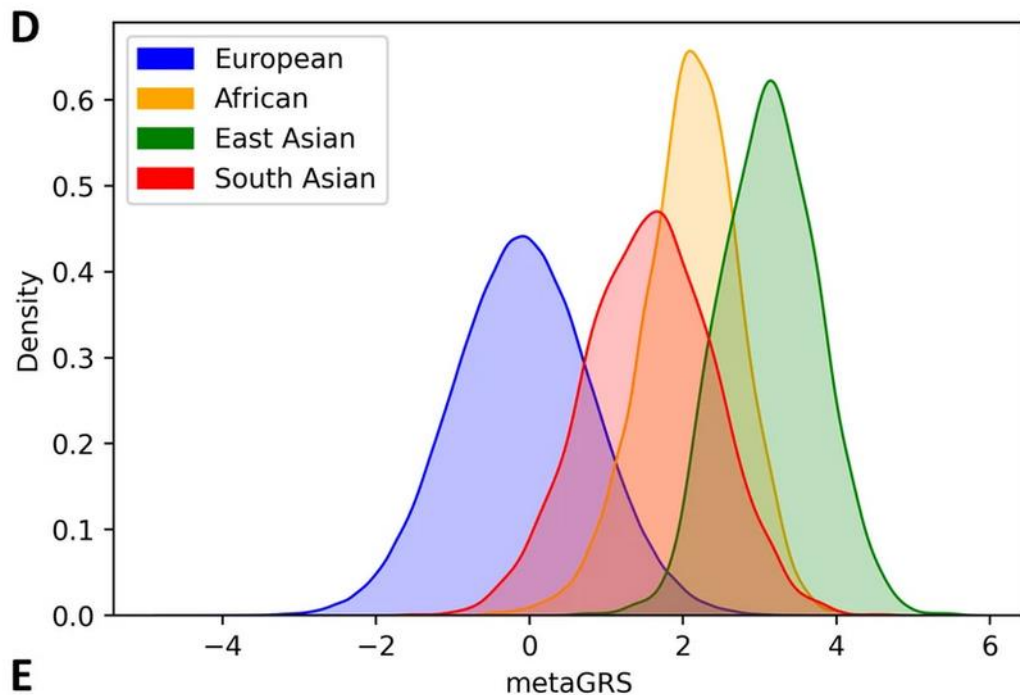
Mavaddat et al., AJHG, 2019

Ho et al., Nat Com, 2020

Sheis et al., JNCI, 2020

Du et al., JNCI

Challenges in PRS reporting and standardization



Factors affecting performance of PGS

- Sample size of training data
- Heritability
- Effect-size distribution

Total Risk due to G = (# of G's) × (Average Risk per G)

Heritability

Polygenicity

Heritability per G

- Methods characteristics ([with borrowing](#))

Multi-ethnic PRS methods

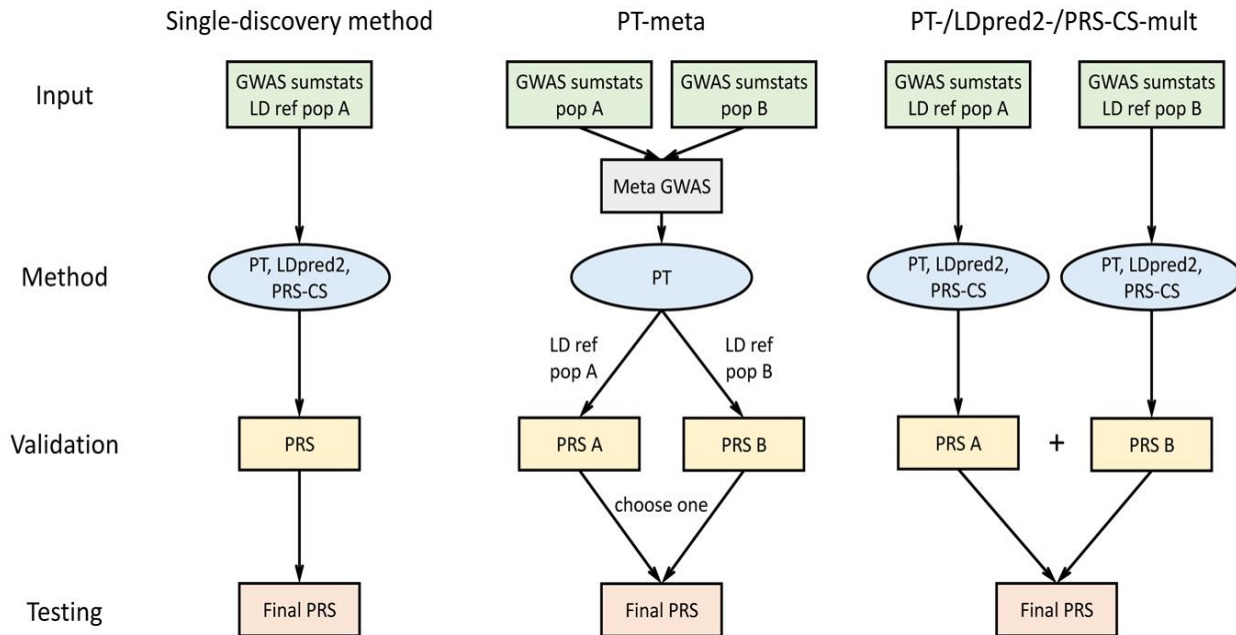
- PRS have been developed using data primarily from European ancestry populations
- Limited accuracy in individuals from non-European ancestries
 - Differences in minor allele frequencies (MAF)
 - Linkage disequilibrium (LD) structures
 - Ancestry-specific genetic effects

Multi-ethnic PRS methods

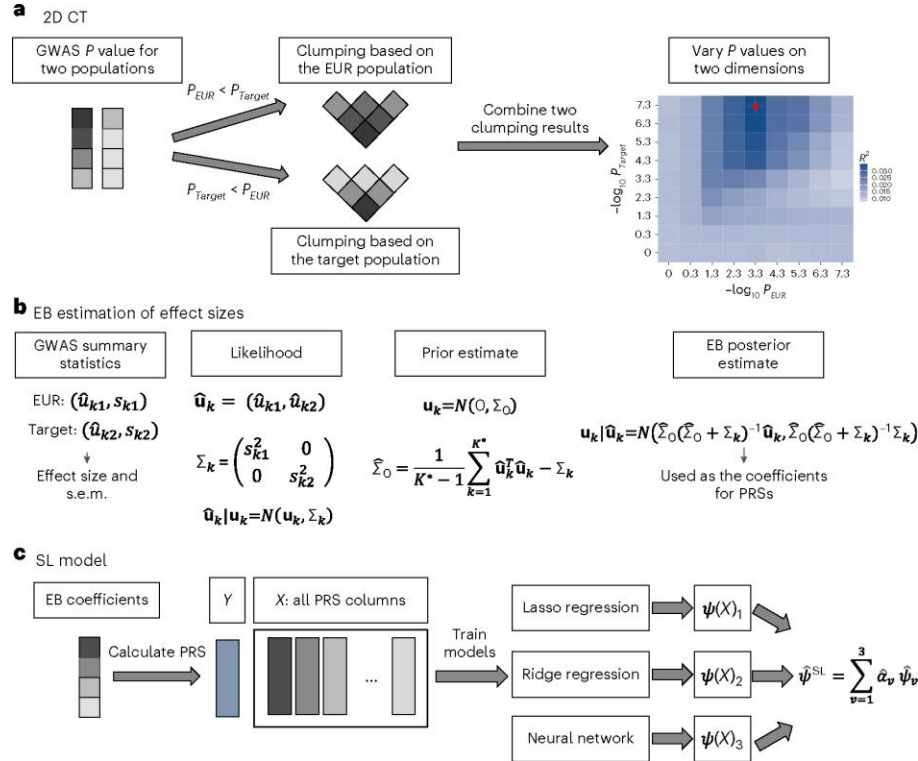
- Clumping-based methods (model-free)
 - Weighted PRS
 - CT-SLEB
- Bayesian Methods
 - PRS-CSx: jointly model ancestry effects
 - MUSSEL
- Penalized regression methods
 - PROSPER

Weighted PRS Method

- PRS for each ancestry group trained using CT or LDpred2
- $PRS_{EUR+AFR} = \alpha_1 PRS_{EUR} + \alpha_2 PRS_{AFR}$
- α_1 α_2 are mixing weights for each population
- Can be obtained using cross-validation



CT-SLEB



1. CT method for selecting SNPs to be included in a PRS for the target population;
2. Empirical Bayes method for SNP coefficient estimation;
3. Superlearning model to combine a series of PRSs generated under different SNP selection thresholds.

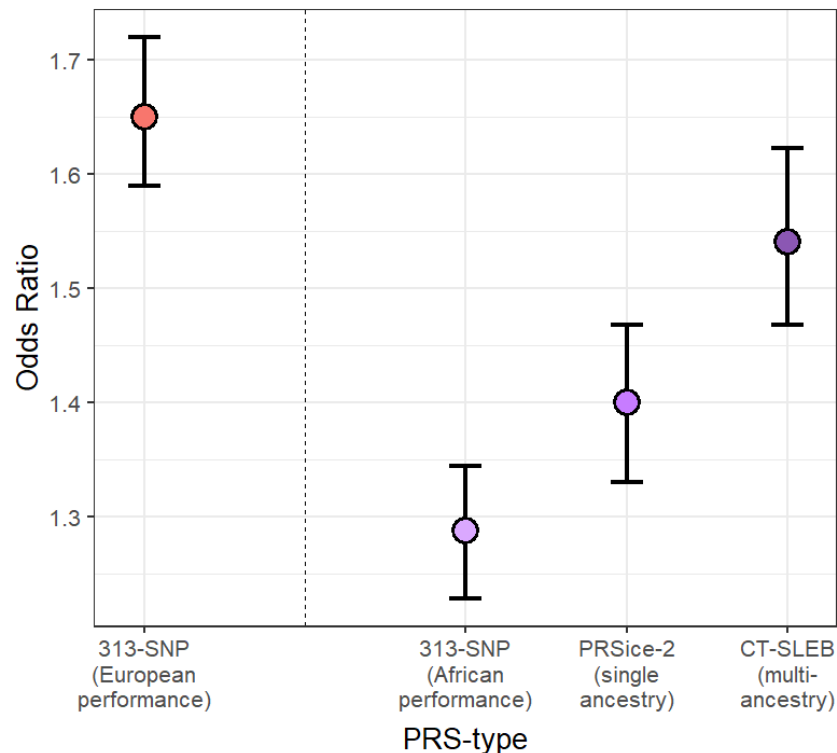
Improving BrCA PRS in African Ancestry Population

African American Breast Cancer Genomics Consortium

- 17K cases and 19K controls

Breast Cancer Association consortium

- 120K cases and 120K controls



Bayesian Method: PRS-CSx

$$\mathbf{y}_k = \mathbf{X}_k \boldsymbol{\beta}_k + \boldsymbol{\epsilon}_k, \quad \boldsymbol{\epsilon}_k \sim \text{MVN}(\mathbf{0}, \sigma_k^2 \mathbf{I}), \quad \pi(\sigma_k^2) \propto \sigma_k^{-2}, \quad k = 1, 2, \dots, K,$$

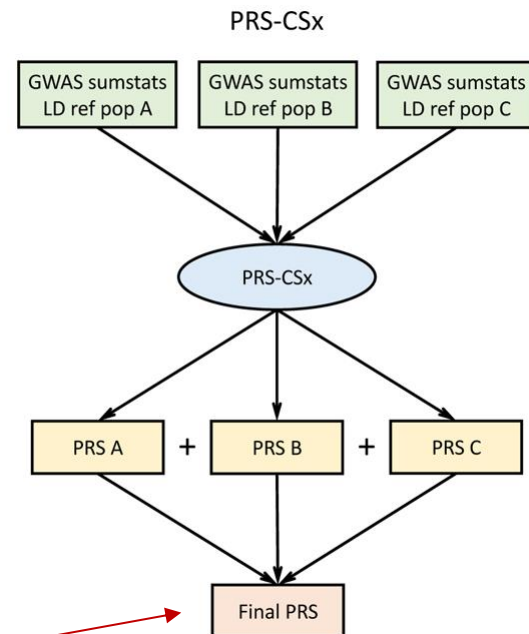
- shared continuous shrinkage prior to couple SNP effects across populations

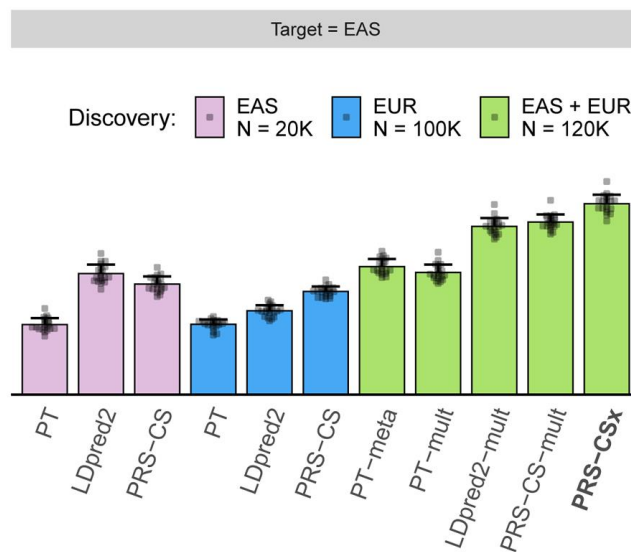
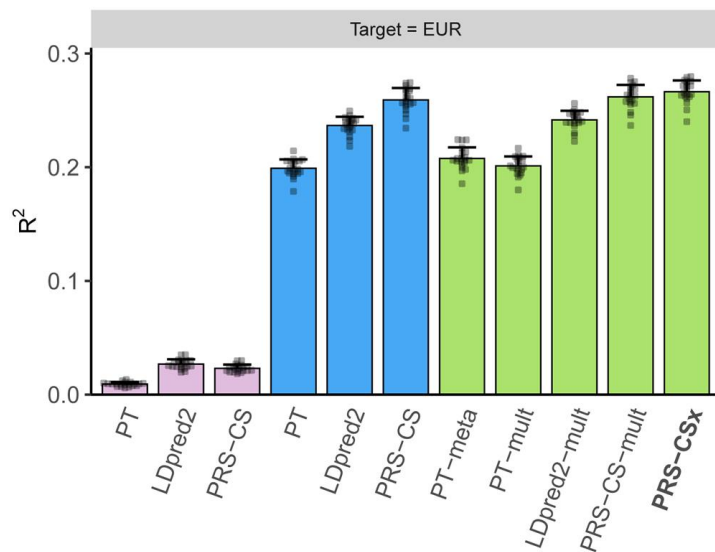
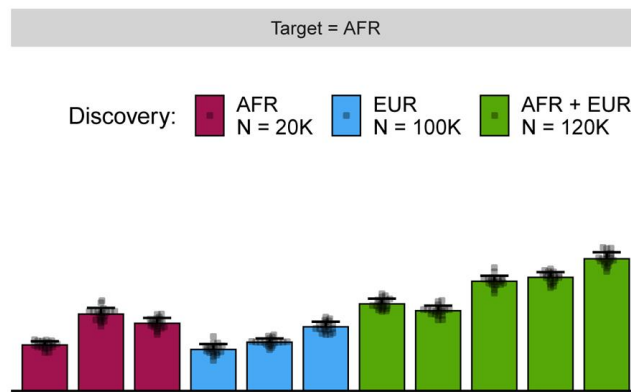
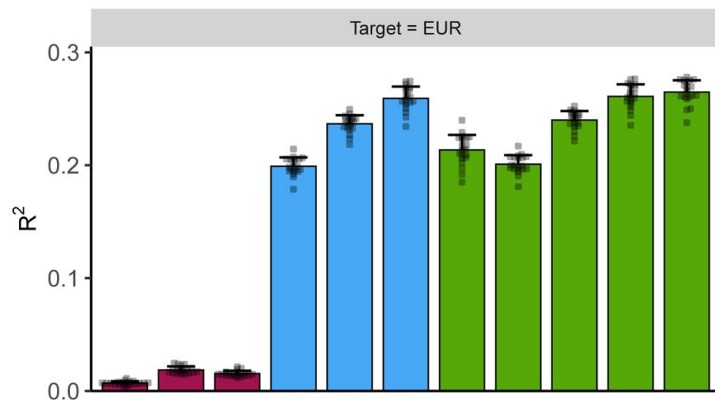
$$\beta_{jk} \sim \text{N}\left(0, \frac{\sigma_k^2}{N_k} \psi_j\right), \quad \psi_j \sim \text{Gamma}(a, \delta_j), \quad \delta_j \sim \text{Gamma}(b, \phi),$$

$$\text{E}[\boldsymbol{\beta}_k | \hat{\boldsymbol{\beta}}_k] = (\mathbf{D}_k + \boldsymbol{\Psi}^{-1})^{-1} \hat{\boldsymbol{\beta}}_k$$

$$\mathbf{D}_k = \mathbf{X}_k^T \mathbf{X}_k / N_k$$

$$\mathbf{PRS} = \hat{w}_{\hat{\phi},1} \mathbf{PRS}_{\hat{\phi},1} + \hat{w}_{\hat{\phi},2} \mathbf{PRS}_{\hat{\phi},2} + \dots + \hat{w}_{\hat{\phi},K} \mathbf{PRS}_{\hat{\phi},K}.$$





Penalized regression: PROSPER

$$\begin{aligned} \mathbf{L}(\beta_1, \dots, \beta_m) = & \sum_{1 \leq i \leq M} (\beta_i^T (\mathbf{R}_i + \delta_i \mathbf{I}) \beta_i - 2\beta_i^T \mathbf{r}_i + 2\lambda_i \|\beta_i\|_1^1) \\ & + \sum_{1 \leq i_1 < i_2 \leq M} c_{i_1 i_2} \|\beta_{i_1}^{s_{i_1 i_2}} - \beta_{i_2}^{s_{i_1 i_2}}\|_2^2 \end{aligned}$$

Idea of fused lasso, but with L2 penalty

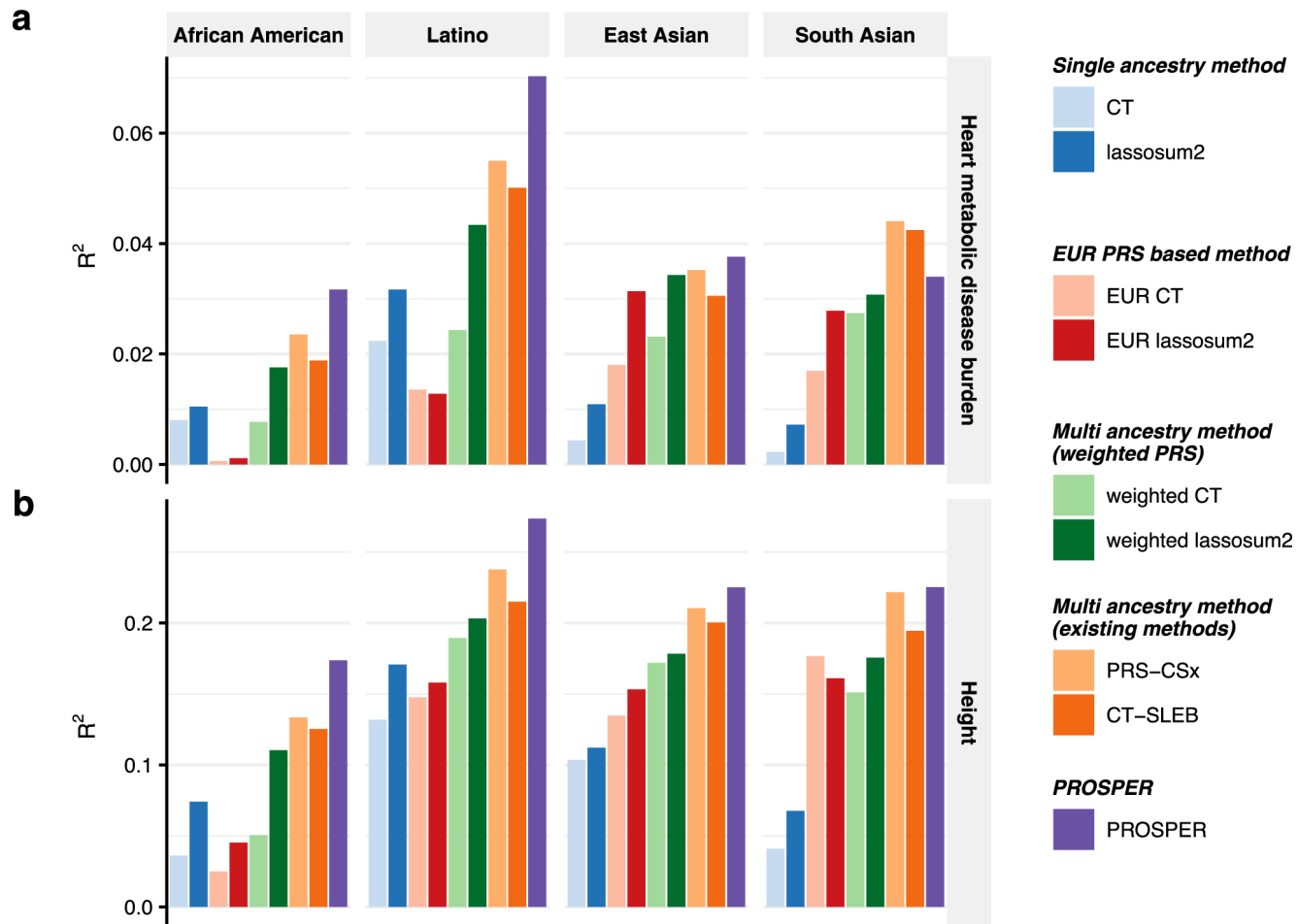
Effect-size similarity across pairs of populations



Can be further simplified assuming all pairs of populations have similar degree of homogeneity, c

A computational scalable method:

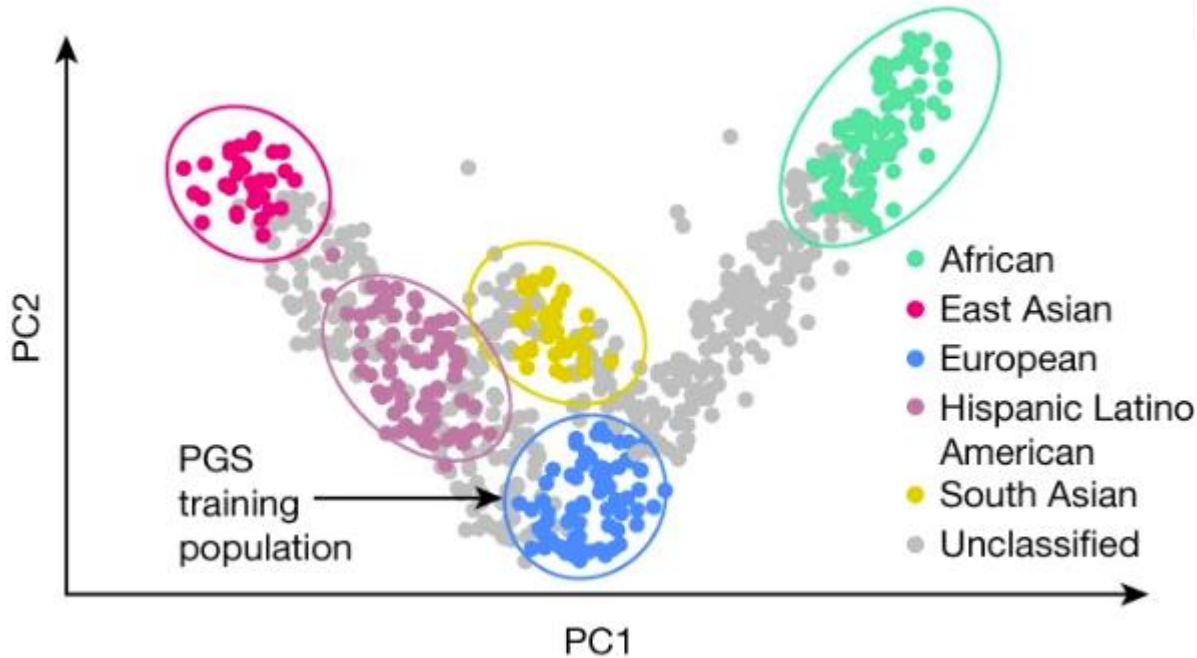
$$\sum_{1 \leq i \leq M} (\beta_i^T (\mathbf{R}_i + \delta_i^0 \mathbf{I}) \beta_i - 2\beta_i^T \mathbf{r}_i + 2\lambda \lambda_i^0 \|\beta_i\|_1^1) + \sum_{1 \leq i_1 < i_2 \leq M} c \|\beta_{i_1}^{s_{i_1 i_2}} - \beta_{i_2}^{s_{i_1 i_2}}\|_2^2$$



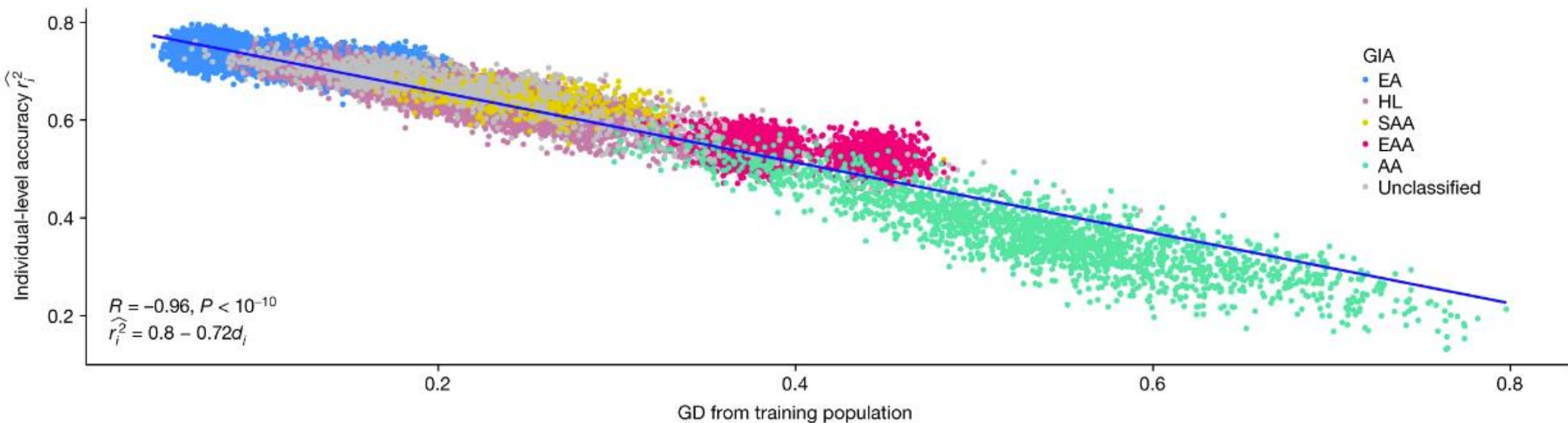
Future directions of PGS methods to improve transferability in multi-ethnic populations

- Continuous genetic ancestry, admixed populations
- Functional annotations, context-dependent effects

Genetic ancestries are **continuous and cluster-free!**



PGS accuracy decays **continuously** across genetic ancestries in UCLA-ATLAS





Continuous genetic ancestry method (for individual-level data)

New Results

 [Follow this preprint](#)

SPLENDID incorporates continuous genetic ancestry in biobank-scale data to improve polygenic risk prediction across diverse populations

 Tony Chen,  Haoyu Zhang, Rahul Mazumder, Xihong Lin

doi: <https://doi.org/10.1101/2024.10.14.618256>

G x PC interaction model:

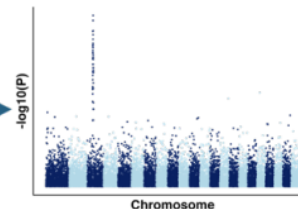
$$Y_i = \sum_{j=1}^M \beta_{j0} G_{ij} + \sum_{j \in H} \sum_{k=1}^{K_0} \beta_{jk} G_{ij} PC_{ik} + \sum_{k=1}^K \gamma_k PC_{ik} + \sum_{\ell=1}^L \alpha_{\ell} Z_{i\ell} + \epsilon_i$$

B. GxPC interaction GWAS: screen for interactions with the first K_0 ancestry PCs

$$Y \sim G_j + \underbrace{G_j \times PC_1 + \dots + G_j \times PC_{K_0}} + Z$$

H_0 : no GxPC interaction effects for variant j

p_j

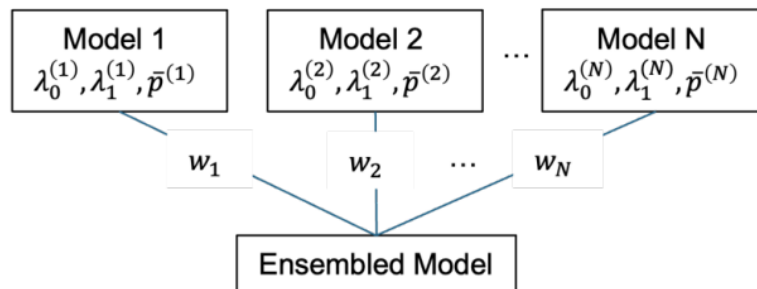


C. Group L0L1 penalized regression with tuning parameters λ_0 , λ_1 , and \bar{p}

$$\min_{\beta} \mathcal{L}(\beta | Y, G, PC, Z) + \lambda_0 \sum_{j=1}^M I(\beta_j \neq 0) + \lambda_1 \sum_{j=1}^M \|\beta_j\|$$

$$\beta_j = \begin{cases} \beta_{j0} & p_j \geq \bar{p} \\ \beta_{j0}, \beta_{j1}, \dots, \beta_{jK_0} & p_j < \bar{p} \end{cases} \quad \begin{array}{l} \text{Main genetic effects only} \\ \text{Main + GxPC interaction effects} \end{array}$$

D. Ensemble learning: combine PRS models from different tuning parameter combinations



Admixed population method (for individual-level data)

Article | [Open access](#) | Published: 03 February 2024

Improving polygenic risk prediction in admixed populations by explicitly modeling ancestral-differential effects via GAUDI

[Quan Sun](#), [Bryce T. Rowland](#), [Jiawen Chen](#), [Anna V. Mikhaylova](#), [Christy Avery](#), [Ulrike Peters](#), [Jessica Lundin](#), [Tara Matise](#), [Steve Buyske](#), [Ran Tao](#), [Rasika A. Mathias](#), [Alexander P. Reiner](#), [Paul L. Auer](#), [Nancy J. Cox](#), [Charles Kooperberg](#), [Timothy A. Thornton](#), [Laura M. Raffield](#) & [Yun Li](#) 

[Nature Communications](#) **15**, Article number: 1016 (2024) | [Cite this article](#)

4969 Accesses | 6 Citations | 8 Altmetric | [Metrics](#)

Local-ancestry specific effects (A, B)

$$y_i = \sum_{j=1}^p [\beta_{A,j}(x_{ij1} I(l_{ij1} = A) + x_{ij2} I(l_{ij2} = A)) + \beta_{B,j}(x_{ij1} I(l_{ij1} = B) + x_{ij2} I(l_{ij2} = B)) + \varepsilon_i]$$

Fused lasso on individual-level data

$$\hat{\beta}(p, \lambda, \gamma) = \operatorname{argmin}_{\beta} \frac{1}{2} \| \mathbf{Y} - \mathbf{G}\beta \|_2^2 + \lambda \| \mathbf{D}_{3p \times 2p} \beta \|_1$$

Very recent research – presented at ASHG 2024!

Platform Talks

Integrative polygenic score modeling
with tissue-specific annotation improves
polygenic scores transferability

[Schedule](#) [Notes](#)

Thu, Nov 07
11:15am - 11:30am (Mountain)
[See in my timezone](#)

Room 501
Presenter 007

Xiaohe Tian
MS
Computer Science and Artificial Intelligence Laborator...

Poster Presentations

Board 4090F: Incorporating functional
annotations to improve polygenic risk
prediction in admixed individuals

[Schedule](#) [Notes](#)

Fri, Nov 08
2:30pm - 4:30pm (Mountain)
[See in my timezone](#)

Exhibit & Poster Hall/Upper Level
Presenter 087

Brian Chen
BS/BA
University of North Carolina at Chapel Hill, Chapel Hill...

Platform Talks

GenESIS: enhancing transferability of
polygenic scores with gene-by-sex
interactions

[Schedule](#) [Notes](#)

Wed, Nov 06
11:30am - 11:45am (Mountain)
[See in my timezone](#)

Room 505
Presenter 008

Yosuke Tanigawa
PhD
Massachusetts Institute of Technology, Cambridge, M...

Large-scale integration of tissue-specific functional annotations through statistical learning improves PGS transferability

- Incorporating nonlinear and context-dependent effects, such as genetic dominance, gene-by-environment (GxE), and gene-by-sex (GxS) interaction effects, can better capture likely causal effects and improve the transferability of PGS.

Summary

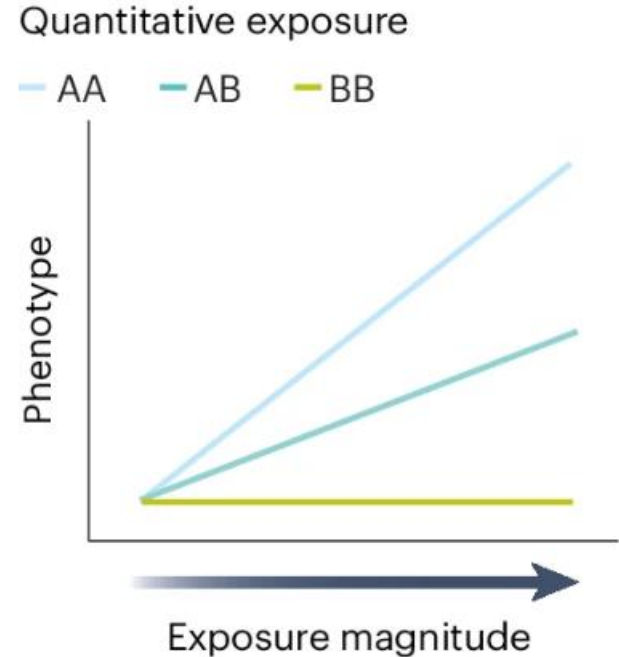
- Advanced methods for building PRS based on statistical high-dimensional modeling (machine learning) techniques is an active area of research
- Can lead to improved performance over simpler more “transparent” methods, but degree of improvement will vary of size of GWAS and underlying genetic architecture of traits
 - Can be applied to summary-statistics data available from GWAS, but tuning and validation may need individual level data
- Current PRS are biased toward European origin populations and thus may exacerbate health disparity, active research are being done to reduce this difference!

Interpretations and Applications of PGS

- Combined Utility of 25 Disease and Risk Factor Polygenic Risk Scores for Stratifying Risk of All-Cause Mortality (DOI: [10.1016/j.ajhg.2020.07.002](https://doi.org/10.1016/j.ajhg.2020.07.002))
- PGSxE interactions (Meisner et al., 2019; Stalder et al., 2017; Wang et al., 2024)
- Estimate PGS in family-based studies (Wang et al., 2024
doi: <https://doi.org/10.1101/2024.10.08.24315066>)

PRS and Environmental Variables Interplay

- The interplay between PRS and environmental variables comes in two forms
 - Interaction
 - Correlation
- Applications of PRS require better characterization of its interplay with environmental factors



Recently Proposed Methods for PRSXE Interaction

- Case-only method (Meisner et al., 2019)
 - PRS indep of E and rare disease assumption so that the disease model is assumed to follow a log-linear model with

$$(S_D|E, D = 1) \sim N(\mu + \sigma^2(\theta_S + \theta_I E), \sigma^2)$$

- Nonparametric Method (Stalder et al., 2017): an extension of the retrospective likelihood approach for arbitrary PRS distribution
- Limitations: Both methods require GXE independence assumptions to achieve high efficiency of the interaction parameter estimation, in addition, case-only method cannot estimate the main effect of environmental variables

Efficient Retrospective Likelihood Method

- Assume

$$\text{pr}(Z|\mathbf{E}, \mathbf{S}) = \text{pr}(Z|\mathbf{S})$$

- Disease model

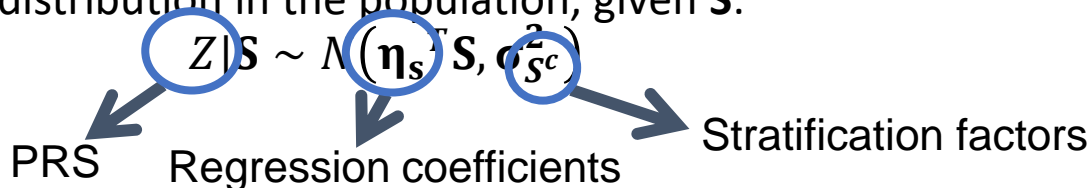
$$\text{pr}(D = 1|Z, \mathbf{E}, \mathbf{S}) = H\{\alpha + m(Z, \mathbf{E}, \mathbf{S}; \boldsymbol{\beta})\}$$

$$H(x) = \{1 + \exp(-x)\}^{-1}$$

In a special case, define $\mathbf{X} = (\mathbf{E}, \mathbf{S})$ and $\mathbf{X}^I \subseteq \mathbf{X}$:

$$m(Z, \mathbf{X}; \boldsymbol{\beta}) = \beta_Z Z + \boldsymbol{\beta}_{\mathbf{X}}^T \mathbf{X} + \boldsymbol{\beta}_{\mathbf{X}^I}^T \mathbf{X}^I Z$$

- PRS follows normal distribution in the population, given \mathbf{S} :



The Case-Control Sampling Design and Retrospective Profile Likelihood

- The retrospective likelihood is

$$L^R = \prod_{i=1}^{N_0+N_1} pr(Z_i, E_i, S_i | D_i)$$

- Using Bayes Theorem and profile-likelihood techniques,

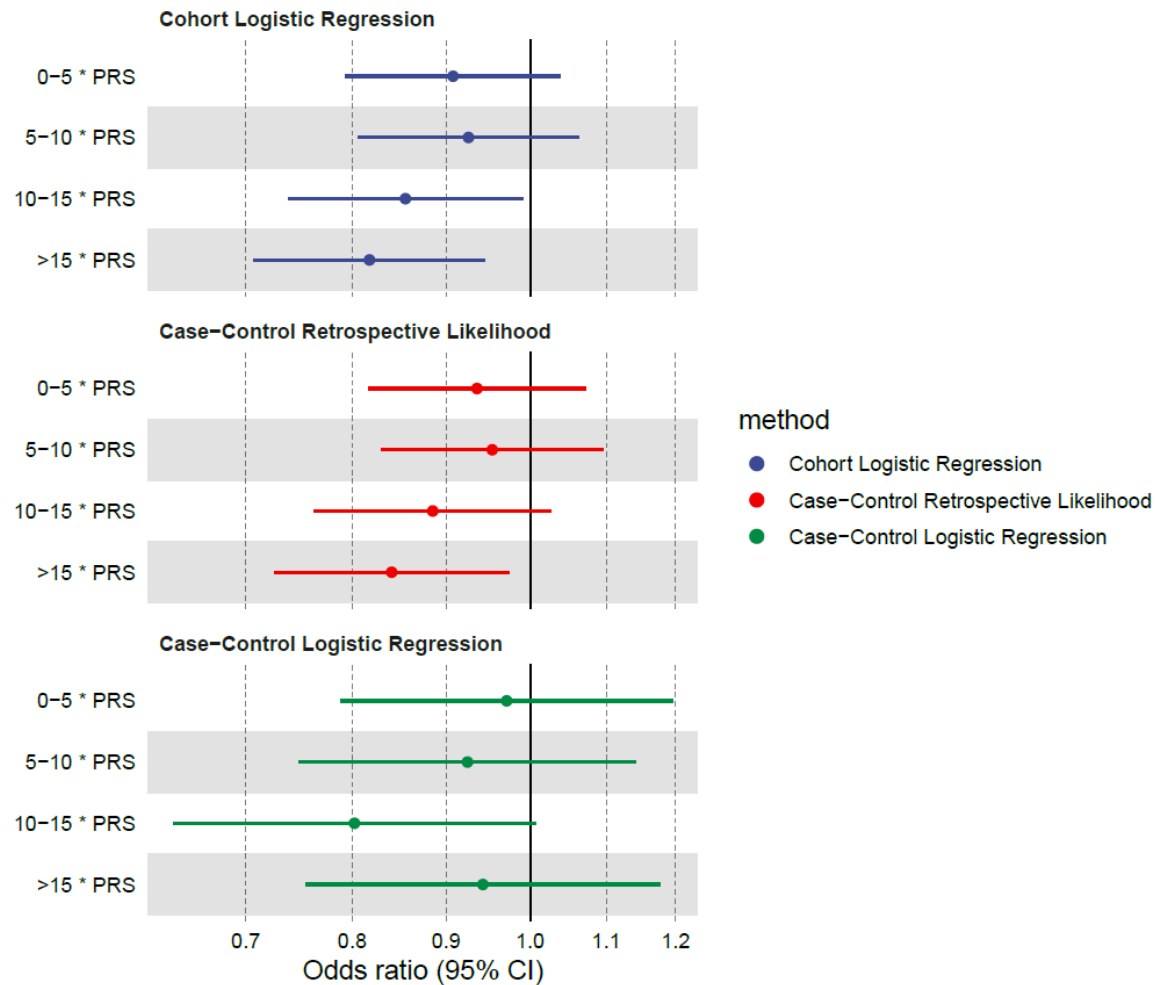
$$\begin{aligned} L_i^R &= \frac{pr(D_i | Z_i, E_i, S_i) f_{Z_i | S_i}(Z_i | S_i) pr(E_i, S_i)}{\int_{Z, E, S} pr(D_i | z', E_i, S_i) f_{Z | S}(z' | S_i) dz' dF(E, S)} \\ &= \frac{f_{Z | S}(Z_i | S_i) \exp(D_i \{\kappa + m(Z_i, E_i, S_i; \beta)\})}{1 + \exp \left\{ \kappa + \frac{1}{2} \sigma_{Sc}^2 (\beta_Z + \beta_{XZ}^T X_i')^2 + \eta_S^T S_i (\beta_Z + \beta_{XZ}^T X_i') + \beta_X^T X_i \right\}} \end{aligned}$$

where $\kappa = \alpha + \log(N_1/N_0) - \log(\pi_1/\pi_0)$, $\pi_1 = \Pr(D = 1)$.

- Parameters in $X=(E,S)$ are profiled out and therefore can be assumed to follow any arbitrary distribution

(Wang et al., AJE 2024)

Years of OC use by categories



Next Class: Tutorials for Building Polygenic Risk Scores

Evaluating PGS methods: cross-validation and cross-ethnicity performance

AUC, R², log odds ratio, log hazard ratio

How to use PGSCatalog and download PGS weights

How to use PLINK in Linux to calculate PRS

PRS-CS tutorial (download LD panel, etc)

Please check the below references related to this class:

Tutorial: a guide to performing polygenic risk score analyses (<https://www.nature.com/articles/s41596-020-0353-1>)

<https://choishingwan.github.io/PRS-Tutorial/>

