PRS Portability Problem: Solutions and further work

Clive Hoggart

1st September 2024

Icahn School of Medicine, Mount Sinai, NYC

Overview of lecture

- The PRS portability problem
- Trans-ancestry PRS methods
  - PRS-CSx – Ruan et al, Nature Genetics, Vol 54, May 2022, 573–580.
  - BridgePRS – Hoggart et al, Nature Genetics, Vol 56, Jan 2024, 180–186, our method.
- Comparison of trans-ancestry PRS methods in simulated and real data
- Admixed individuals
  - What is admixture
  - PRS for admixed individuals
- Application of PRS in heterogeneous populations

The PRS portability problem

- We introduced the PRS trans-ancestry portability problem yesterday

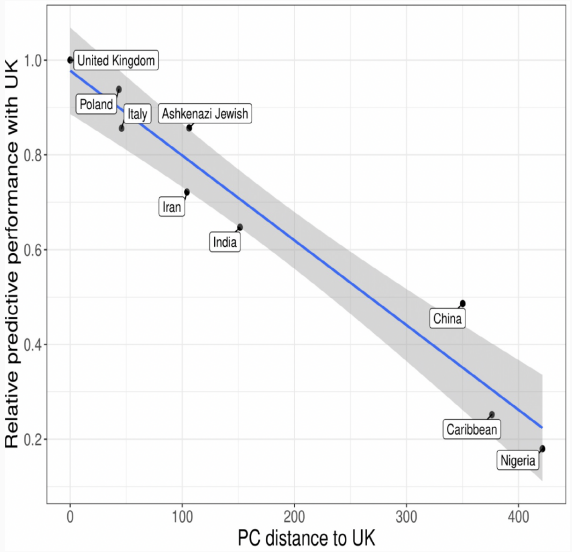
  e.g. a PRS trained using European data performs bady applied to non-Europeans

- Causes:
  - **differences in LD**: good tag for causal variants vary between population
  - **differences in MAF**: variation in power to detect causal variants
  - **differences in environmental contribution to phenotype** results in:

    proportion of phenotypic variance explained by genetics, i.e. h^2

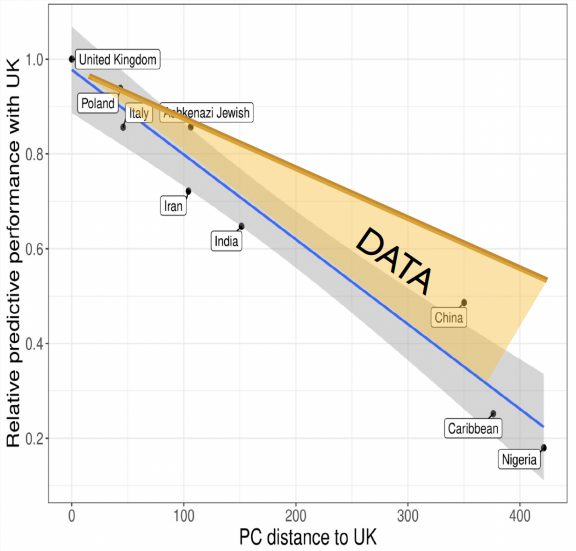
    differences in effect size in pressence of G$\times$E
- PRS portability problem occurs despite sharing of causal genetic effects globally

PRS portability decreases linearly in PC distance from training population



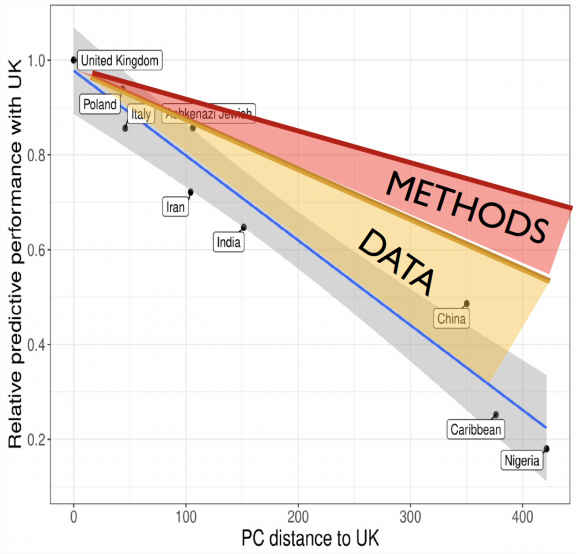
- Prive et al AJHG 2022

PRS portability problem



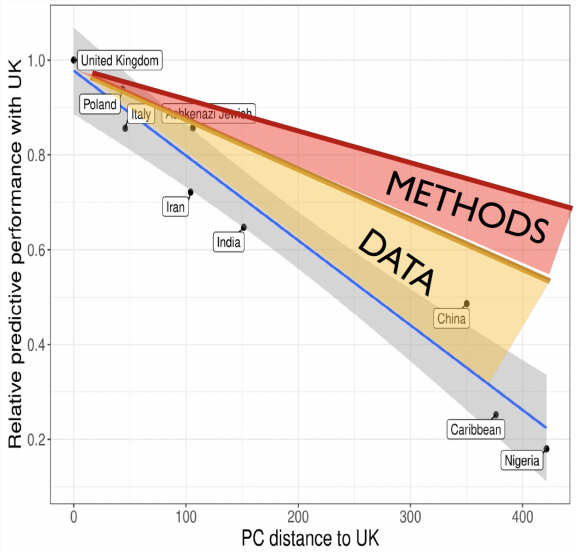
- more diverse population data will help solve this

PRS portability problem



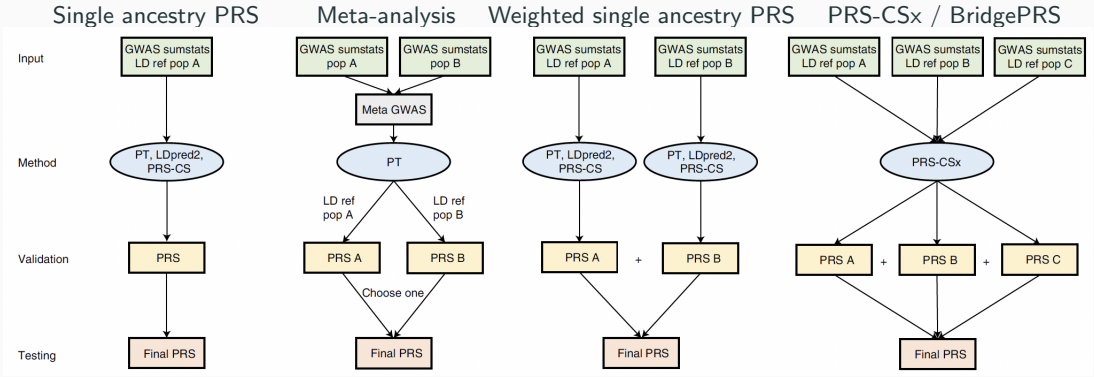
- more diverse population data will help solve this
- statistical genetics methods reduce gap further

PRS portability problem



- more diverse population data will help solve this
- statistical genetics methods reduce gap further
- ultimately, the best methods will leverage all available data to **optimise prediction in everyone**

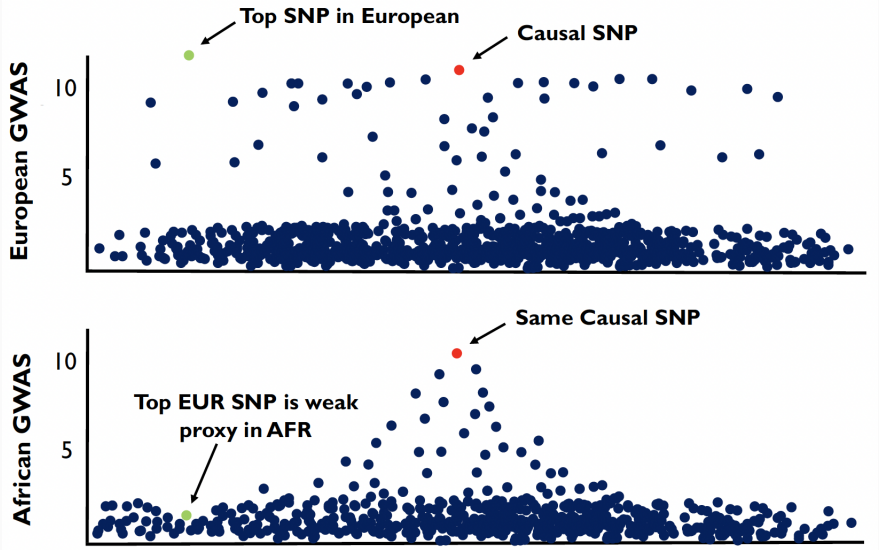
Broad overview of trans-ancestry PRS methods



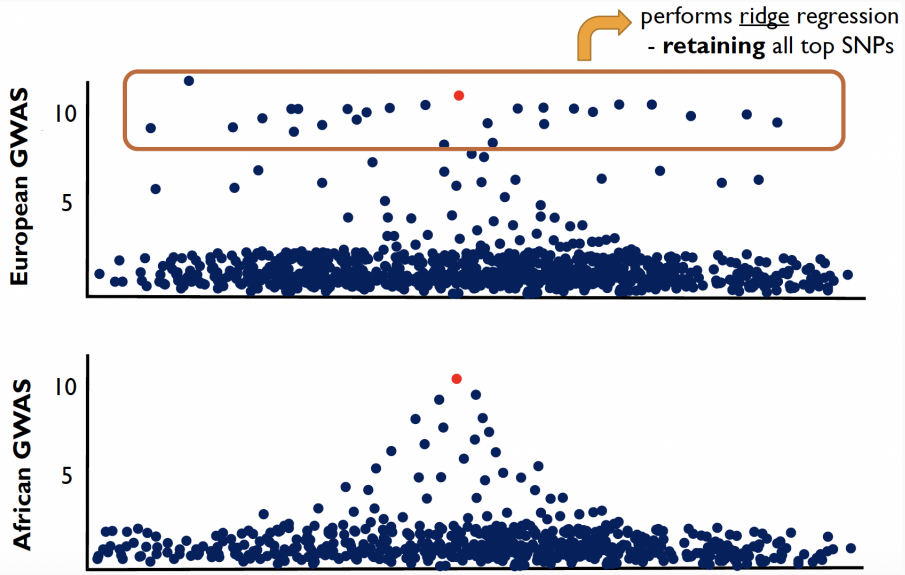
- Taken from PRS-CSx paper, Ruan et al, Nature Genetics, VOL 54, May 2022, 573–580
- BridgePRS only applicable to two populations

The BridgePRS solution

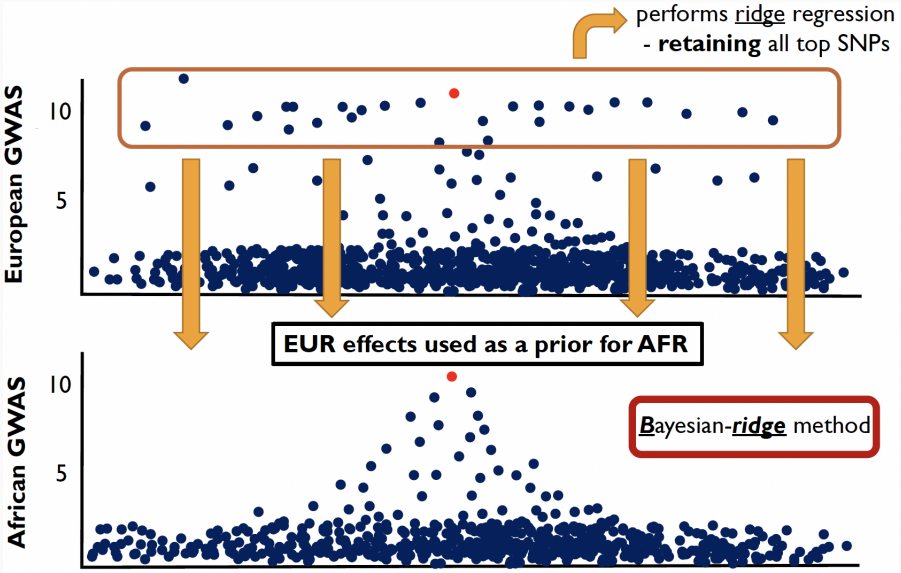
The BridgePRS solution – accounting for LD differences between population



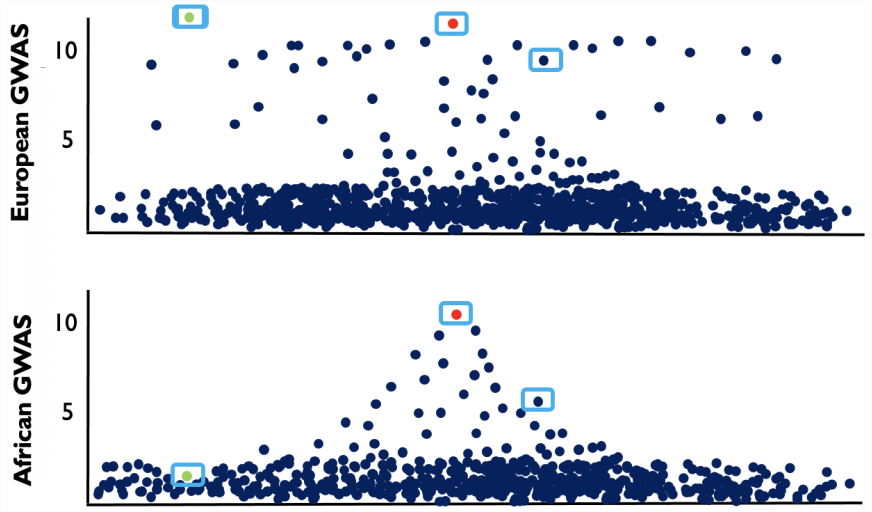
BridgePRS – Stage 1



BridgePRS – Stage 2



PRS-CSx tries to ‘pick a winner’



BridgePRS and PRS-CSx modelling strategies

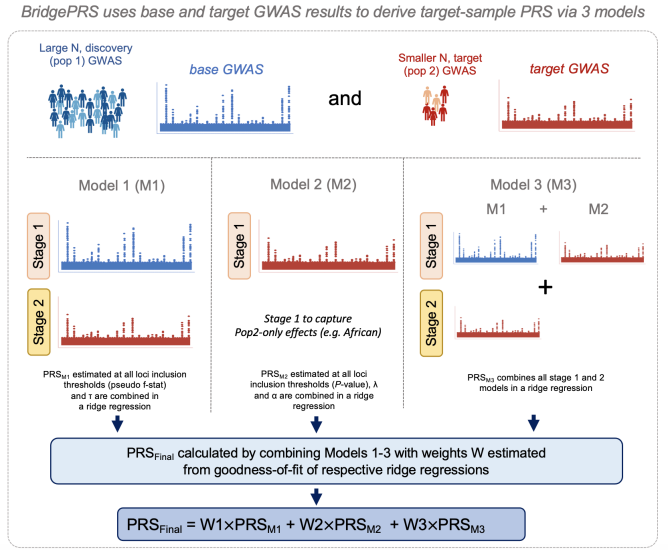
- **PRS-CSx strategy:** Fine mapping
  - refine causal variant to a single or minimal set of candidate SNPs
  - causal variant is the best to use, if known
  - may be sub-optimal when causal variant is not included in data
  - shares information between populations on inclusion of SNPs in the PRS
  - does **not** share information between populations on SNP effect size
- **BridgePRS strategy:** Aggregate information across putative loci
  - estimate optimal SNP weights to tag causal variants
  - estimating variant effect sizes (not location) is key when prediction is the goal
  - not so reliant on inclusion of causal variant
  - **does** share information between populations on SNP effect size
- Both methods combine GWAS summary statistics across ancestrally diverse populations

BridgePRS modelling – Capture target population unique effects

- Two-stage modelling (using base population as a prior) will only capture genetic effect variants present in the base population (at sufficiently high allele frequency)

- Therefore BridgePRS also applies stage 1 modelling to the target population to capture effects unique to the target population

BridgePRS model overview



BridgePRS – Interpretting the models

- **Stage 2 PRS** – using base population as prior for target population
  - reflects the belief that the target population GWAS is only informative in conjugtion with the base population GWAS.
- **Stage 1 PRS** – only using the target population
  - reflects the belief that the base population GWAS gives no additional information, only the target population GWAS is informative
- **Stage 1+2 PRS** – combining both stage 1 and stage 2 models
  - reflects the belief both the base and target population GWAS contribute independent information
- **Weighted PRS** – calculated by combining the above models with weights estimated by each of the models' goodness-of-fit to test data
- **Typically advise to use the Weighted PRS**

Benchmarking via Simulation

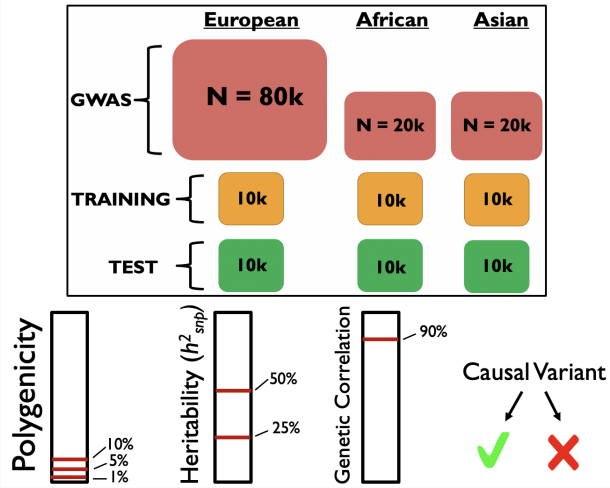
BridgePRS vs PRS-CSx

Also compare with adapted single ancestry PRS methods:

PRSice-meta – C+T applied to meta-analysis

PRS-CS-mult – optimised weighted mean of single ancestry PRS-CS fits

Benchmarking via Simulation – overview

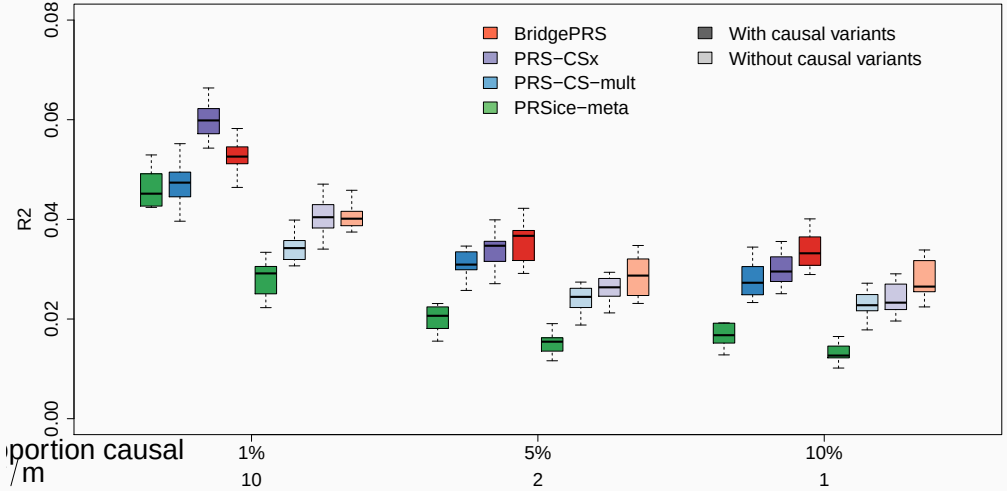


- Doubling heritability $\equiv$ to doubling sample size

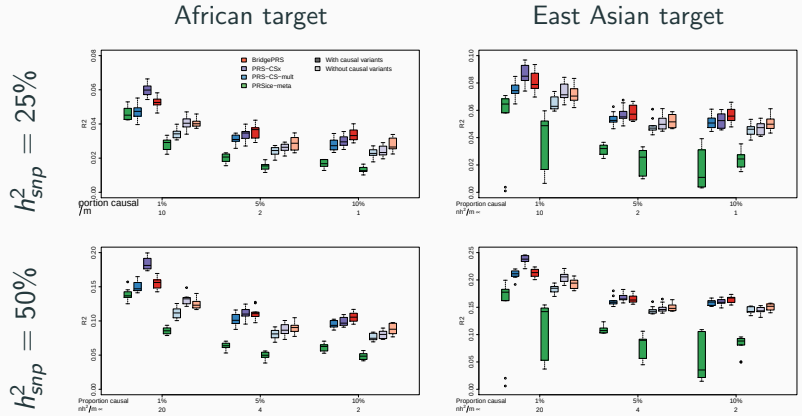
$\Rightarrow$ 50% heritability and 80K Eur GWAS $\equiv$ 25% heritability and 160K Eur GWAS

- 12 simulation scenarios – 3 architectures $\times$ 2 heritabilities $\times$ with/without causal variant
- 10 replicates of each

Simulations results – African target, $h^2_{snp} = 25\%$



Simulations results

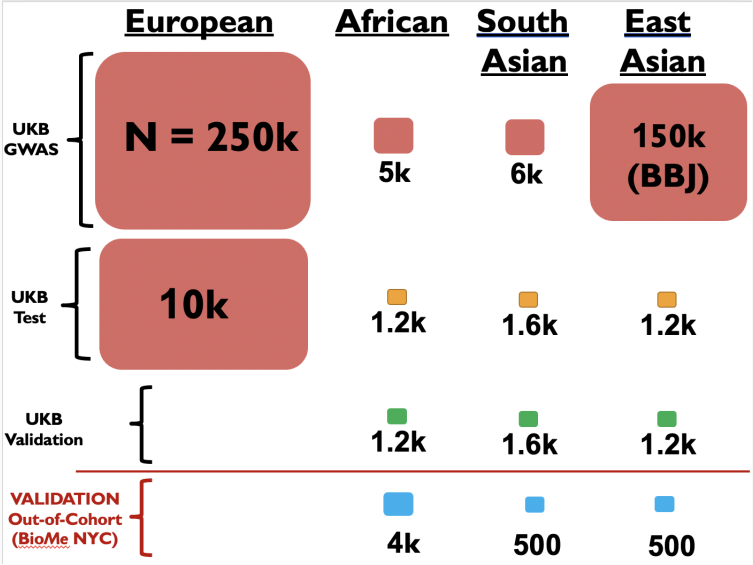


BridgePRS $>$ PRS-CSx: higher polygenicity, lower h^2, causal SNP missing, African target

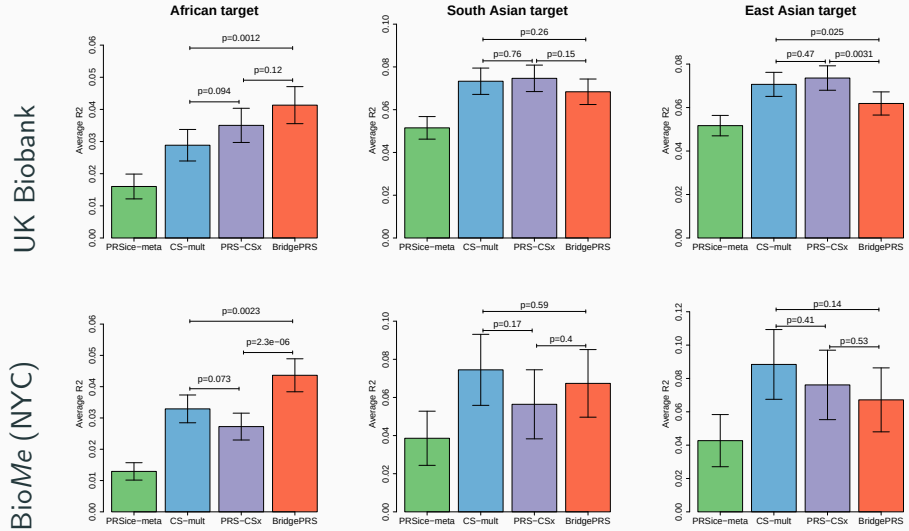
BridgePRS $>$ PRS-CSx when there is greater uncertainty

Benchmarking via Real Data

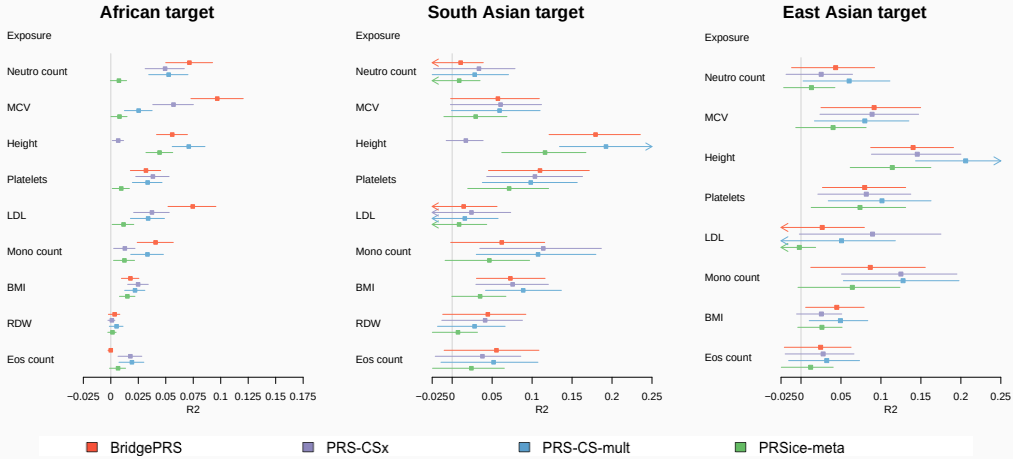
Benchmarking via Real Data – overview



Real data results



Real data results – BioMe NYC



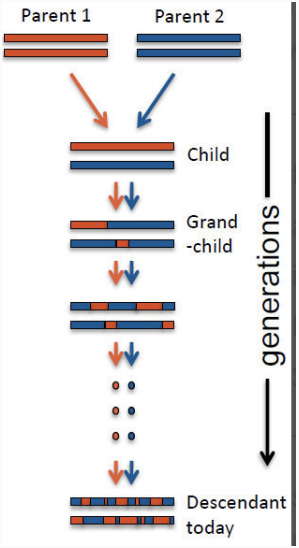
Wide variety of performance depending on trait analysed

Admixture and PRS for admixed populations

Limitations of BridgePRS and PRS-CSx

- BridgePRS and PRS-CSx are designed to optimise PRS within single homogeneous populations
- However, many individuals are **admixed**
- Admixed individuals have genetic ancestry from two or more populations
  - *admixture* is typically measured between **continental** population, e.g. Africans and European
- Examples of two-way African and European admixed populations include African-Americans and African-Carribeans
- Hispanics have three-way admixture of American, European and African ancestry

What is admixture?

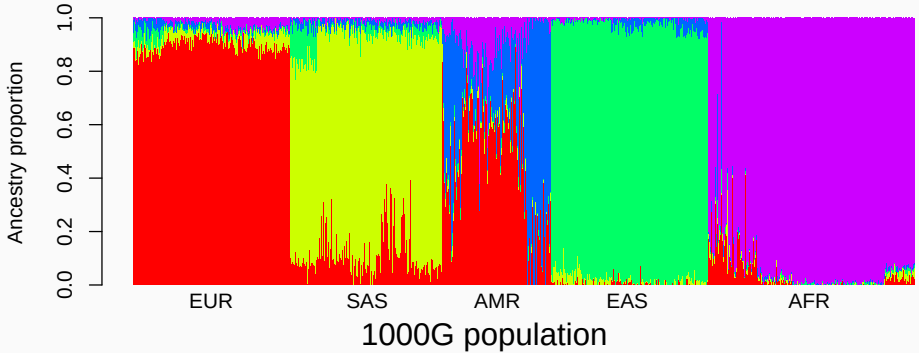


- Red and blue represant ancestral populations, e.g. African and European
- "Parents" are unadmixed individuals from the ancestral populations
- "Child" has one chromosome from each ancestral population
- Subsequent generations recombinations occur between chromosomes from the different ancestral populations
- This results in a "mosaic" of segments from the founding ancestral populations
- Because of recombinations in each generation, segments become smaller in successive generations

Describing admixture

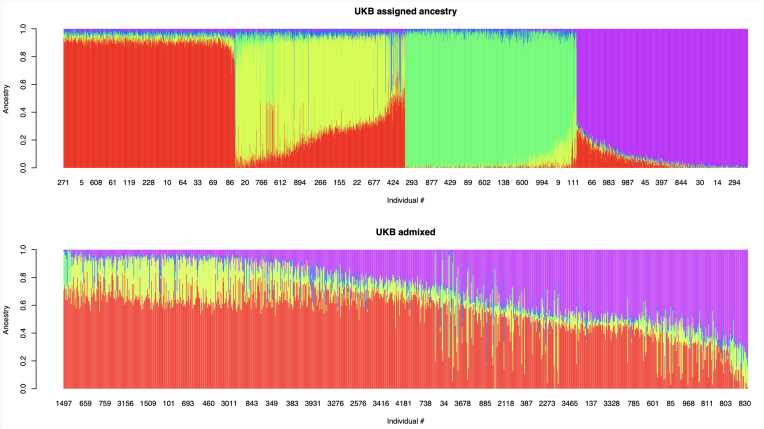
- Previous slide depicted admixed as a single event of two ancestral populations mating
- Typically admixture occurs by unadmixed, ancestral population, individuals contributing to the gene pool over many generations
- This results in different proportions of admixture in individuals in the admixed population
- Therefore, individuals' admixture is described by both:
  - Individuals' proportion of admixture from the ancestral population
  - Ancestry of chromosomal segments, local ancestry
- Software:
  - `ADMIXTURE` estimates individual admixture proportions – `https://dalexander.github.io/admixture/`
  - `RFMix` estimates local ancestry using a Random Forest algorithm – `https://github.com/slowkoni/rfmix/blob/master/MANUAL.md`

Admixture in 1000 Genomes



- `ADMIXTURE` analysis of 1000G – unsupervised learning
- American populations exhibit most admixture, particularly Mexican-American, Columbians and Puerto Ricans
- Least admixed Americans, most American ancestry (blue), are Peruvian
- Admixed AFR population are African-Americans

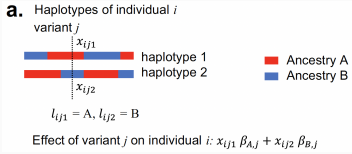
Admixture in UK Biobank



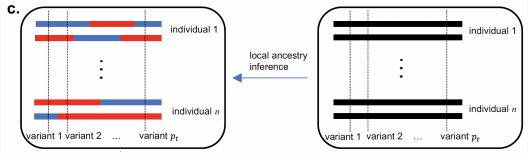
- Top row: samples we assigned to ancestral populations in BridgePRS paper
  - Many of these individuals are admixed
- Bottom row: Many admixed individuals, predominately two-way AFR/EUR
  - These individuals not included in BridgePRS paper

PRS for admixture samples – GAUDI

- GAUDI, Sun et al. Nature Comm, 2024, developed specifically to estimate PRS in admixed individuals
- GAUDI estimates ancestry specific locus effects



- PRS sums ancestry specific locus effects according to inferred local ancestry

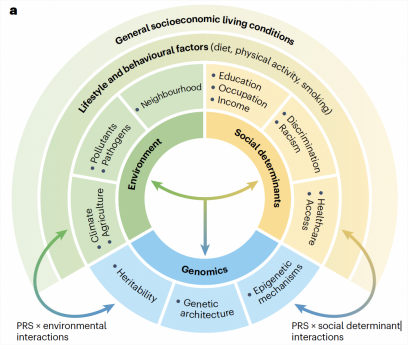


- Local ancestry estimated using RFMix

PRS for heterogeneous cohorts

Applying PRS to heterogeneous cohorts

- UK Biobank, BioMe (recruited in New York City) and many real world cohorts where health care providers may want to apply PRS have individuals from diverse genetic backgrounds:
  - unadmixed Europeans, Africans, East Asians, South Asians **AND** admxied individuals
- Heterogeneous populations have varying environmnetal exposures and social determinants which can affect health outcomes and PRS performance



Kachuri et al (2024) Nature Reviews Genetics

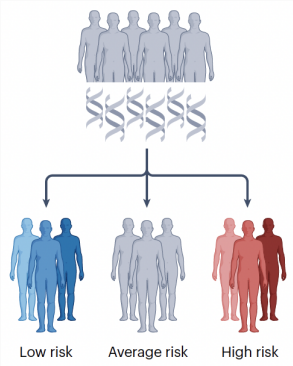
Effect on genetic prediction

- Disease risk can vary between ethnic groups because of both
  - genetic differences, i.e. higher allele frequencies of risk variants
  - differences in environmnetal exposures and social determinants of health
- PRS should capture real causal genetic effects which vary in allele frequency between populations
  - but should be robust to confounding effect of differences in environmnetal exposures and social determinants of health between populations
- These two effects difficult/impossible to dissantangle studying populations with different genetic and environmnetal exposures separately
- Here admixed populations with similar environmnetal exposures and social determinants of health but varying proportions of admixture can dissentangle these two factors

Incorporating PRS in the clinical

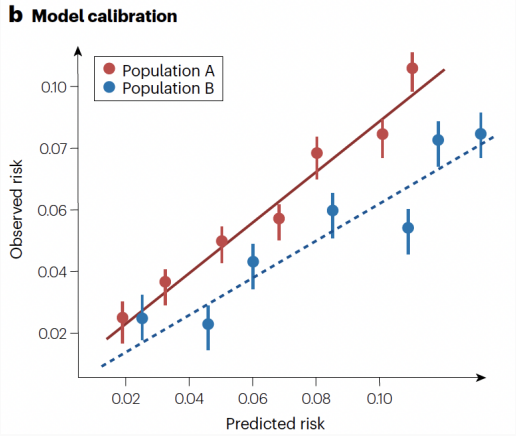
- A risk model used in the clinic should include PRS and other known risk factors
  - it should also be interpreatable for clinicians and patients, e.g. not based on a neural network, random forest, etc.
- Predictions should be interpretable on the observed scale
  - e.g. probability of disease in given time period
- Well calibrated: predicted probability of disease = observed probablity of disease

Risk stratification



Kachuri et al (2024) Nature Reviews Genetics

Model calibration



Kachuri et al (2024) Nature Reviews Genetics

- Model is well calibrated in population A
- Risks are systematically overestimated in population B

Consideration of distribution of risk between populations

Even if risk is well callibrated in populations problems in cross population implementation could persist

Conclusions

- BridgePRS power relative PRS-CSx increases as uncertainty increases, eg:
  - causal varant not included in data
  - sample size decreases
  - heritability decreases
  - polygenicity increases
  - more distant populations, eg European base and African target
- PRS-CSx power relative to BridgePRS increases as power to fine-map increases
- Recommend using both methods
- Present implemenation of PRS-CSx is restricted to HapMap variants
  - code relies on pre-computed LD, decreases likelihood of including causal variant
- BridgePRS computes LD on the fly within loci using all available genotype data

Conclusions and future work

- Extend BridgePRS to Bayesian hierarchical model
  - utilise multiple ancestral populations
- Extend to model PRS for admixed individuals
- **BridgePRS** website `https://www.bridgeprs.net`
- Application of PRS to heterogeneous population in an equitable manner is an ongoing reseach topic

  equitable – without bias and similar accuracy across populations

Thanks

- **Paul O'Reilly**
- **Sam Choi**
- **Judit García González**
- **Tade Souaiaia**

- Fits continuous shrinkage prior to SNP effects
  - normal-exponential-gamma (NEG) aka Strawderman–Berger distribution
  - generalisation of double expontential dist. (DE) $\equiv$ LASSO penalty

$$\text{DE}(\beta \mid \xi) = \int_0^\infty \text{N}(\beta \mid 0, \psi)\text{Ga}(\psi \mid 1, \xi^2/2)\ d\psi = \frac{\xi}{2}\exp\{-\xi|\beta|\}$$

$$\text{NEG}(\beta \mid \lambda, \gamma) = \int_0^\infty \int_0^\infty \text{N}(\beta \mid 0, \psi)\text{Ga}(\psi \mid 1, \phi)\text{Ga}(\phi \mid \lambda, \gamma^2)\ d\psi d\phi$$

  - large mass at zero inducing strong shrinkage of small (noisy) effects
  - strong shrinkage – variable selection – **analogous to fine-mapping**
  - heavy tails results in minimal shrinkage of large effects

BridgePRS – Modelling at locus level

- Modelling applied to loci defined by clumping and thresholding (in plink)
- Unlike C+T retain all SNPs within clumps for ridge regression
- Jointly model SNP effects within loci and assume loci are independent
- PRS is sum of effects across all M contributing loci

$$PRS = \sum_{j=1}^{M} \boldsymbol{X}_j \boldsymbol{\beta}_j$$

$\boldsymbol{X}_j$ and $\boldsymbol{\beta}_j$ are genotypes and vector of effects at locus j

PRS-CSx: a focus on fine-mapping

- Fits continuous shrinkage prior to SNP effects
- A hierarchical model is used to share information between populations k

$$\text{PRS-CSx prior}(\beta_{jk}) = \int_0^\infty \int_0^\infty \text{N}\left(\beta_{jk} \mid 0, \psi_j \frac{\sigma_K^2}{N_k}\right) \text{Ga}(\psi_j \mid 1, \phi_j)\text{Ga}(\phi_j \mid \lambda, \gamma^2)\ d\psi_j d\phi_j$$

- Information shared on shrinkage/inclusion of each SNP j via parameter ψ_j
  - infomation not shared on effect size
- Large mass at zero inducing strong shrinkage (to 0) of small effects
  - **analogous to fine-mapping**
- **MCMC** accounts for uncertainty in location of causal variants
- Individual level data used for paramter optimisation

BridgePRS – Aggregate informaton across putative loci

BridgePRS – is broken into two stages

- BridgePRS – stage 1
  - Applies **ridge** regression – equivalent to zero centred multivariate Gaussian priors
  - Applied to powerful single population GWAS, eg Europeans
  - Spreads signal from causal variants across loci accounting for their unknown location
  - Tackles "winner's curse" by shrinking effect estimates
- BridgePRS – stage 2
  - Output from stage 1 (also a multivariate Gaussian) used as prior for 2nd population
  - Updates PRS given *target* population data
- Modelling applied to independent loci defined by clumping and thresholding
- Unlike C+T retain **all SNPs** within loci for ridge regression

BridgePRS – details of modelling

- Stage 1: Zero centred Gaussian priors applied to pop 1 GWAS sumstats

$$\beta_{pop1} \sim \mathrm{N}(0, \mathrm{diag}(\lambda^{-1}(\theta_k(1-\theta_k))^{-\alpha}))$$

λ – shrinkage to zero
α – relative contribution of rarer and more common variants, θ_k=MAF

- Stage 2: Prior for pop 2 PRS defined by Gaussian posterior from Stage 1

$$\beta_{pop2} \sim \mathrm{N}(\tilde{\beta}_{pop1}, (\tau\lambda_{pop1})^{-1})$$

τ – degree of shrinkage to pop 1 PRS