Polygenic Risk Scores for SSNS, SRNS

Alejandro Ochoa

StatGen, Biostatistics & Bioinformatics — Duke University

2024-04-03 — NS U01 working group

Overview

- Rerun with SAIGE "base" data
  - results were surprisingly similar to GMMAT
- Added CureGN as a testing dataset
  - Was already a training dataset, but will only use to test when training is something else
  - Two versions: pediatric only, and all MCD/FSGS (performs slightly worse)
- Evaluate separately per ancestry
  - Was hoping to show that accuracy is similar across ancestries, but small sample sizes left it inconclusive

How PRS works

Score is generally a linear model:

$$\text{PRS}_j = \sum_i \beta_i x_{ij}.$$

- i: variant index
- j: individual index
- β_i: coefficient of variant i
- x_{ij}: genotype (0,1,2) at variant i, individual j

Challenge is about picking β_i:

- Not all variants are in all datasets
- If starting from GWAS, need to decorrelate (LD or clumping), shrink (p-value threshold or fancier models)

Basics of PRS construction and evaluation

- PRS construction and validation requires 3 disjoint datasets:
  - Base set: Used to fit "GWAS summary statistics": variant coefficients (betas), standard errors, p-values
  - Training set: Used to fit PRS parameters: p-value threshold, or heritability and sparsity
    - Modifies betas, usually by shrinking them to zero and reducing correlation due to LD
  - Testing set: Data where nothing was trained, reveals true performance (correlation to trait)

Testing setups

Name	Base	Train	Test
SC-DDB	D SC (532/3553)	D SR (193/193)	B SR (365/149)
SC-DCB	D SC (725/3553)	C SR (250/170)	B SR (365/149)
SR-DCB	D SR (725/193)	C SR (250/170)	B SR (365/149)
SC-DDC	D SC (532/3553)	D SR (193/193)	C SR (250/170)
SC-DDC2	D SC (532/3553)	D SR (193/193)	C2 MF (415/476)

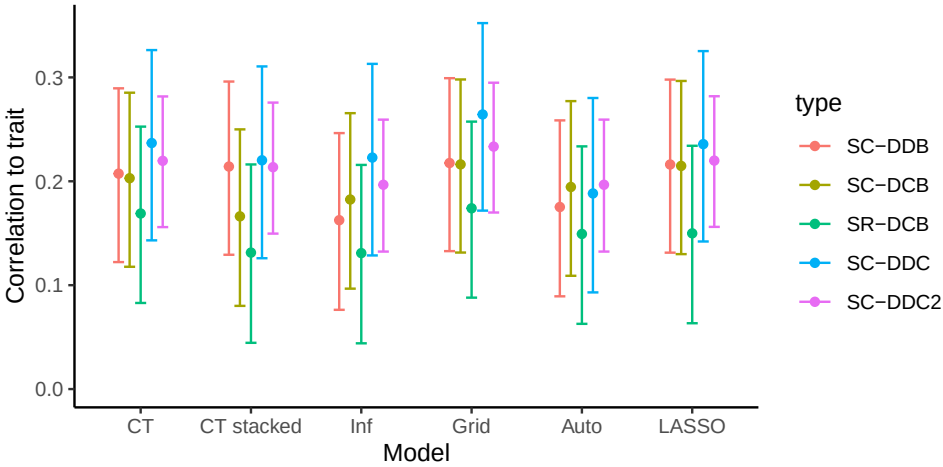
SR=SSNS-SRNS; SC=SSNS-Ctrl; MF=MCD-FSGS

D=Discovery; B=Bristol; C=CureGN, based on these rules:

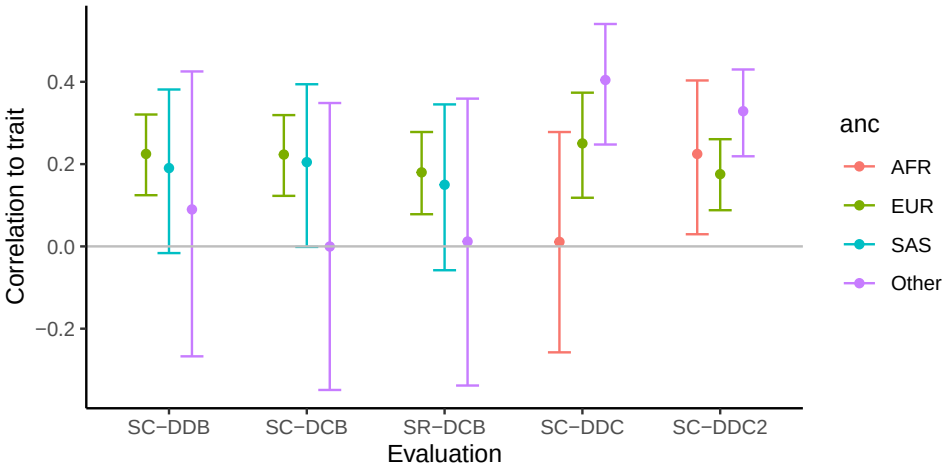
- SSNS: MCD and age ≤ 21
- SRNS: FSGS and age ≤ 21

C2=CureGN does not apply age filters!

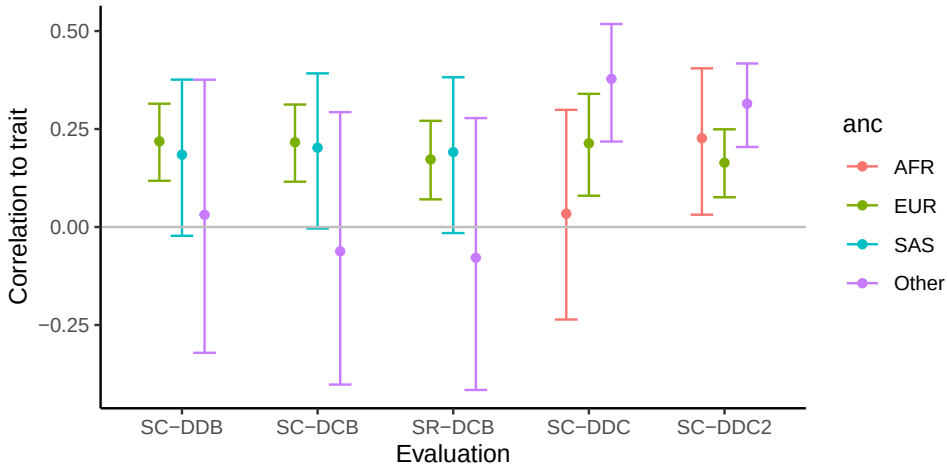
Test results: SSNS-Ctrl base with CureGN best, SSNS-SRNS base worst; CureGN age filters help



Test results by ancestry: Grid



Test results by ancestry: CT



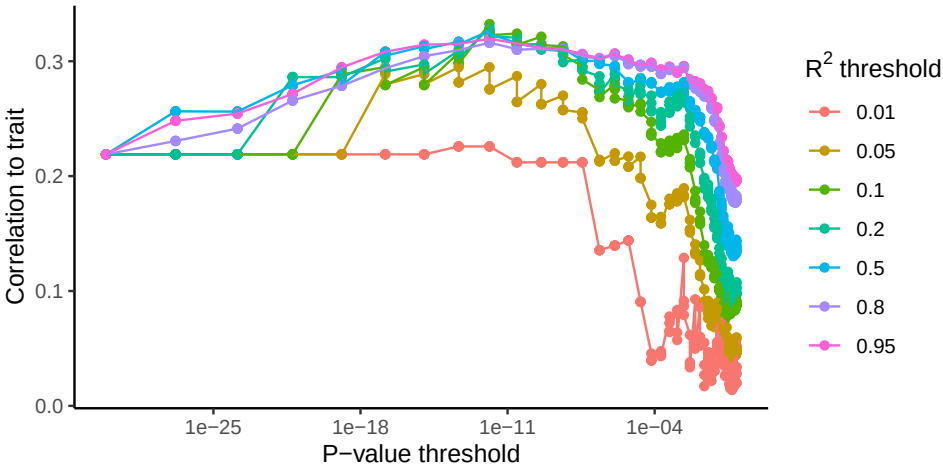
Clump and Threshold notes

- Best models for each training setup were all small
  - SC-DDB: 8
  - SC-DCB: 6
  - SR-DCB: 12
- All were chr6 SNPs!

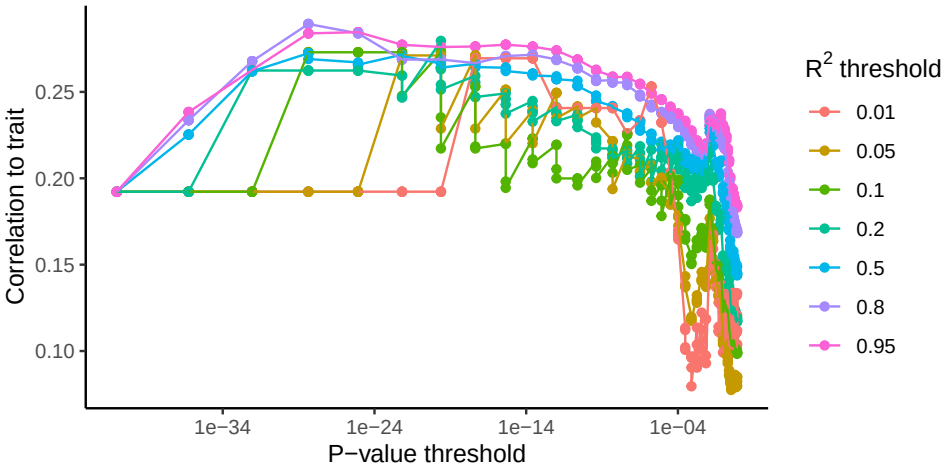
Next steps

- Use HLA haplotypes!

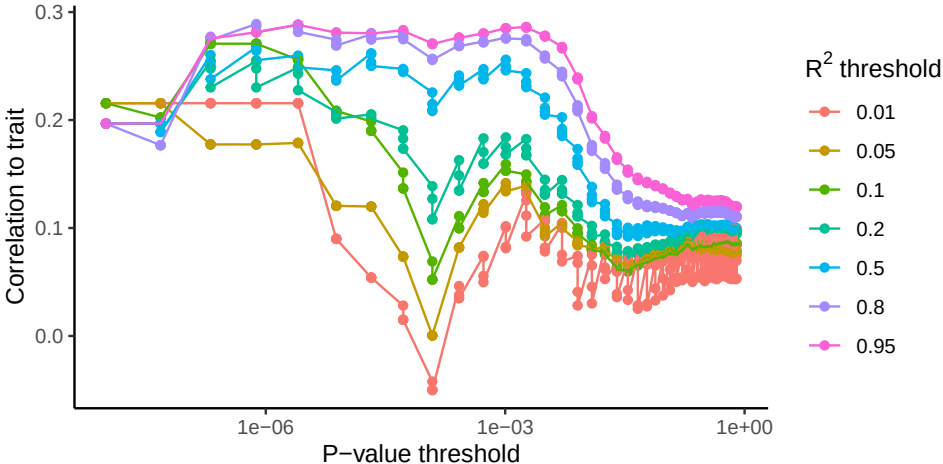
Train results: SC-DDB ldpred2-ct



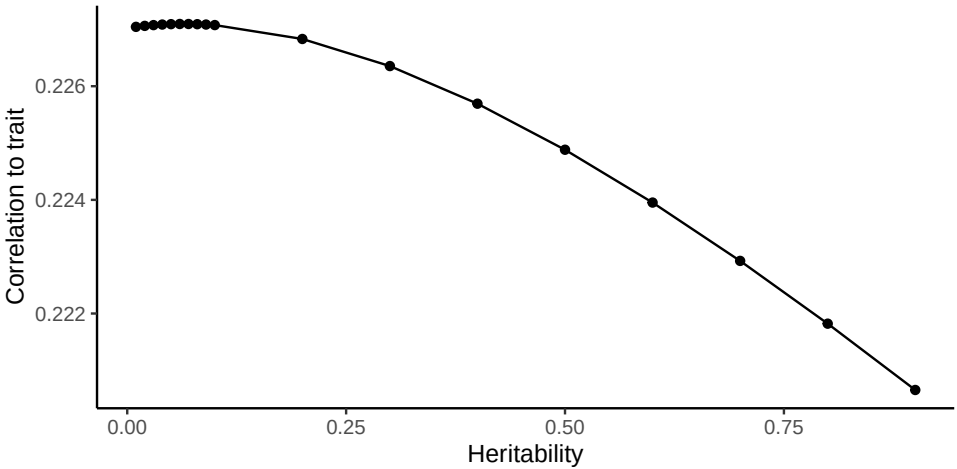
Train results: SC-DCB ldpred2-ct



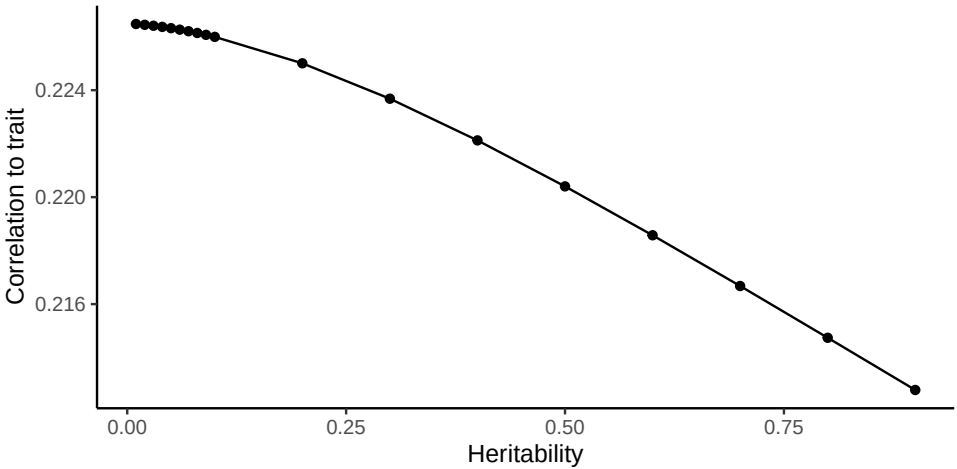
Train results: SR-DCB ldpred2-ct



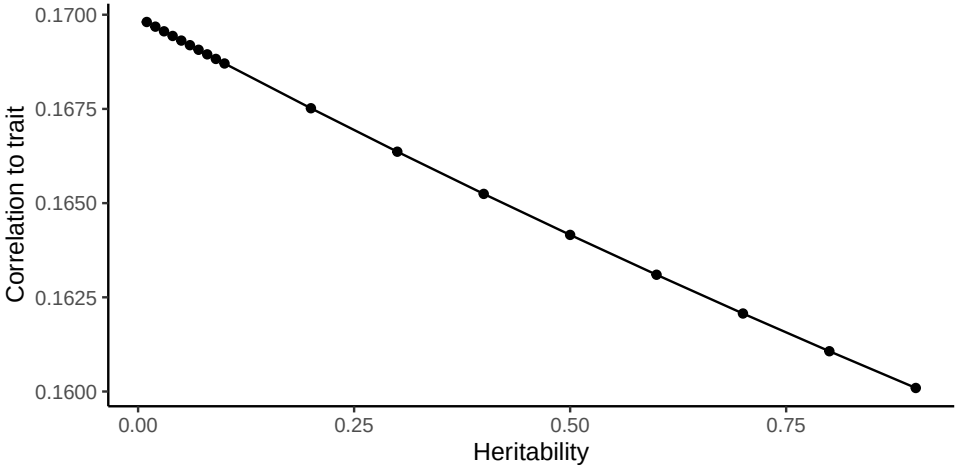
Train results: SC-DDB ldpred2-inf



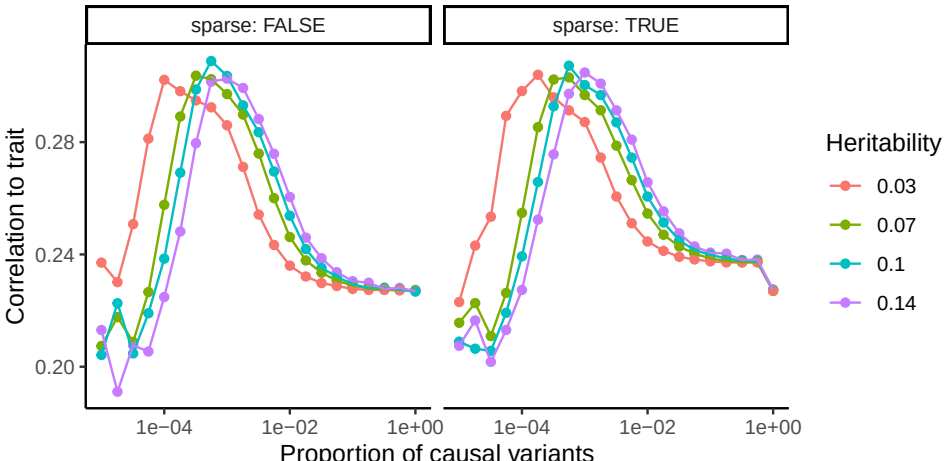
Train results: SC-DCB ldpred2-inf



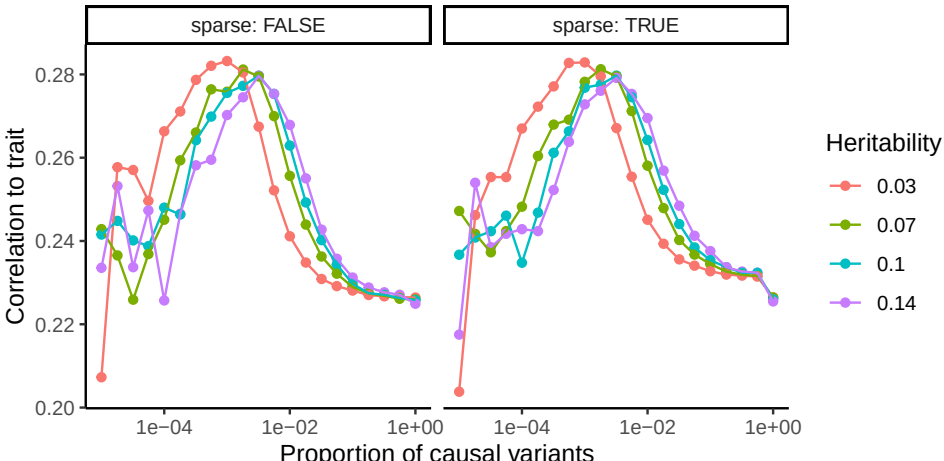
Train results: SR-DCB ldpred2-inf



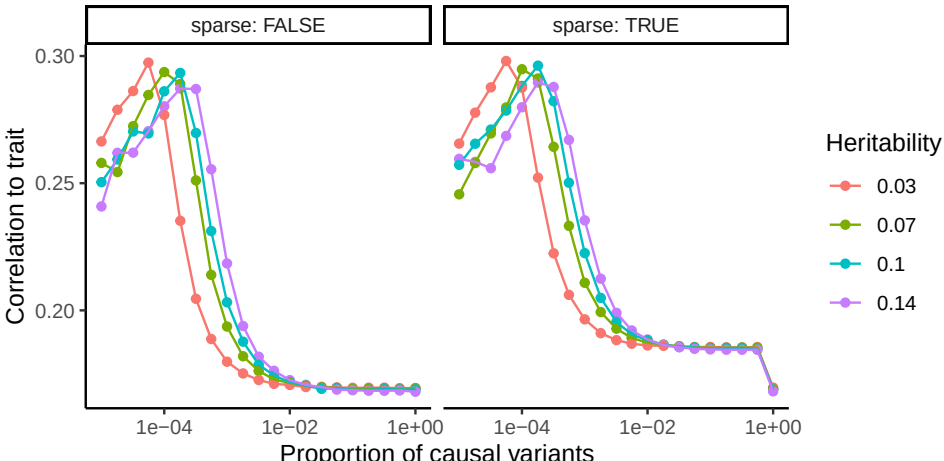
Train results: SC-DDB ldpred2-grid-h0.1



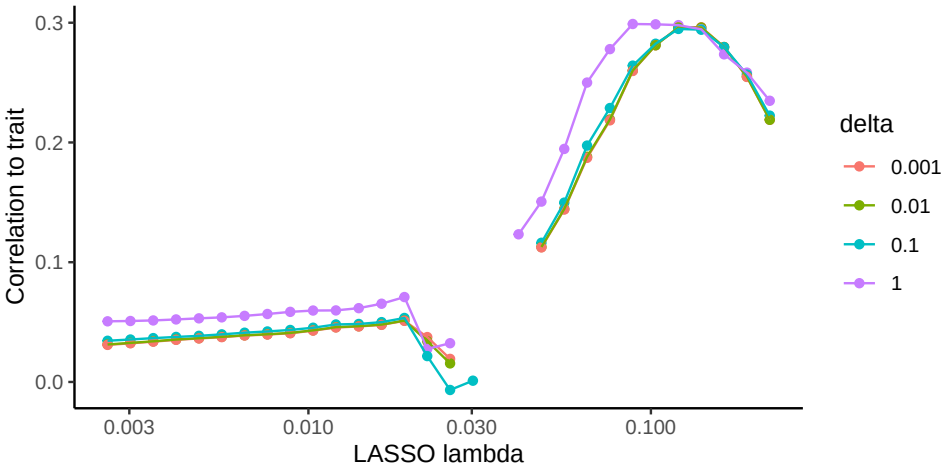
Train results: SC-DCB ldpred2-grid-h0.1



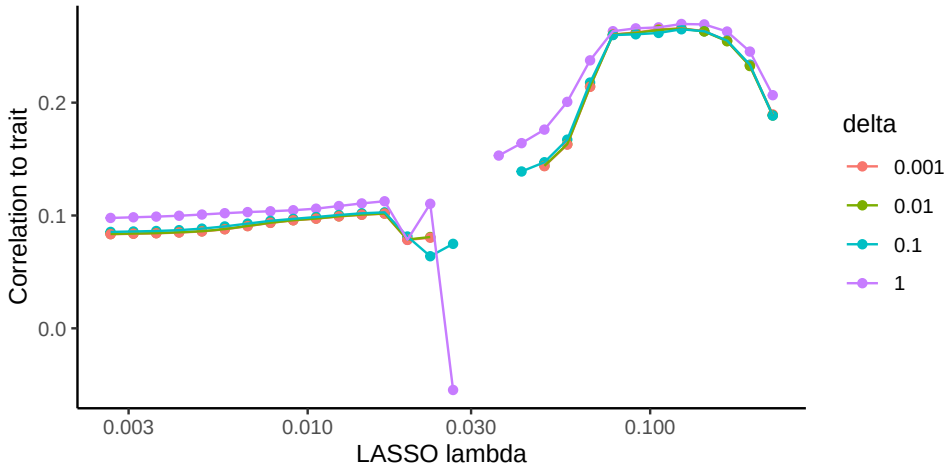
Train results: SR-DCB ldpred2-grid-h0.1



Train results: SC-DDB ldpred2-lassosum



Train results: SC-DCB ldpred2-lassosum



Train results: SR-DCB ldpred2-lassosum