

Polygenic Risk Scores for SSNS, SRNS

Alejandro Ochoa

StatGen, Biostatistics & Bioinformatics — Duke University

2024-03-20 — NS U01 working group

Overview

- ▶ Added CT: clump and threshold, and a “stacking” variant
 - ▶ Data setup is very different, took me a while to adapt
 - ▶ Not better than previous best, but much more interpretable:
 - ▶ Only 8-31 SNPs from HLA lead to these good predictions!
- ▶ Updated CureGN (removed two patients that withdrew consent, one of which was SRNS)

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

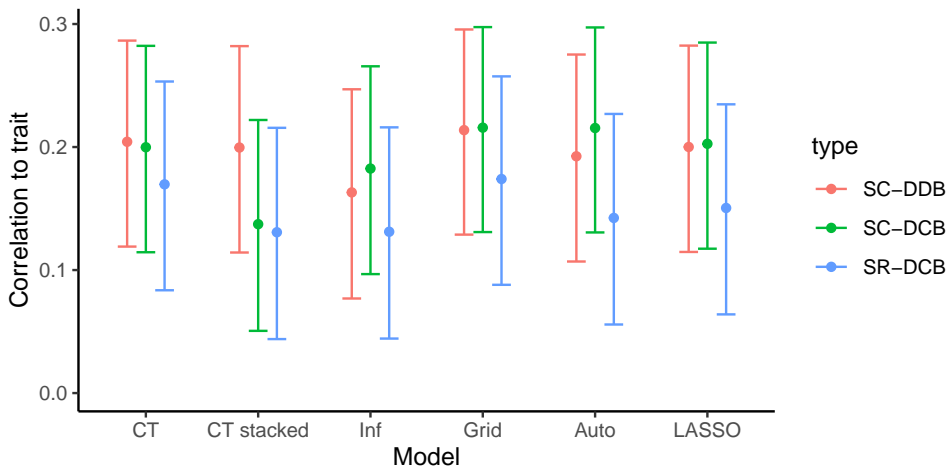
Name	Base	Train	Test
SC-DDB	D SC (532/3553)	D SR (193/193)	B SR (365/149)
SC-DCB	D SC (725/3553)	C SR (250/170)	B SR (365/149)
SR-DCB	D SR (725/193)	C SR (250/170)	B SR (365/149)

SR=SSNS-SRNS; SC=SSNS-Ctrl

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

- ▶ SSNS: MCD and $\text{age} \leq 21$
- ▶ SRNS: FSGS and $\text{age} \leq 21$

Test results: SSNS-Ctrl base with CureGN best, SSNS-SRNS base worst



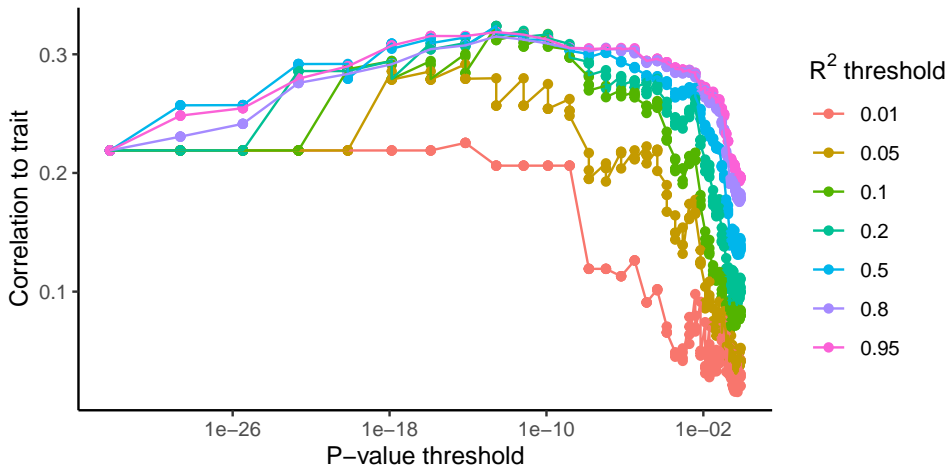
Clump and Threshold notes

- ▶ Best models for each training setup were all small
 - ▶ SC-DDB: 31
 - ▶ SC-DCB: 8
 - ▶ SR-DCB: 15
- ▶ All were chr6 SNPs except first model had one chr10 SNP

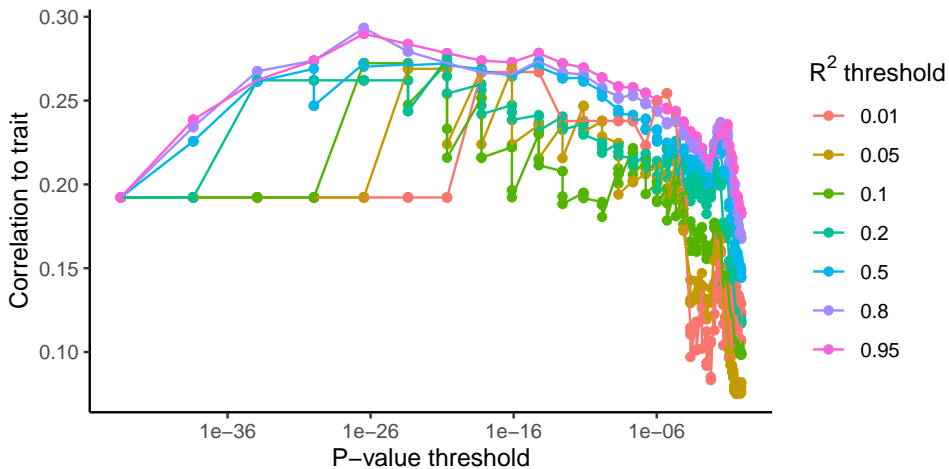
Next steps

- ▶ Rerun with SAIGE data
- ▶ Use HLA haplotypes!
- ▶ Vary SNP set filters
 - ▶ Due to LD runtime, only using array SNPs right now
 - ▶ Could enrich for more significant ones or higher severity variants
- ▶ Try Barry et al., 2023 base data

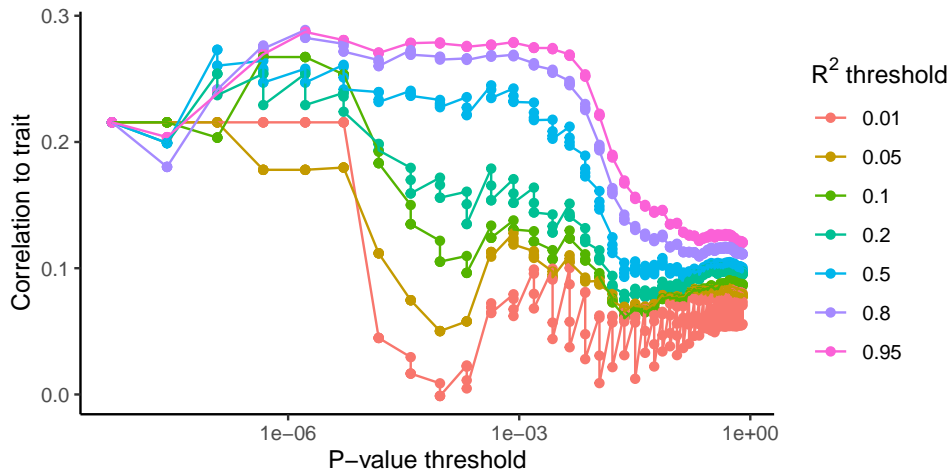
Train results: SC-DDB Idpred2-ct



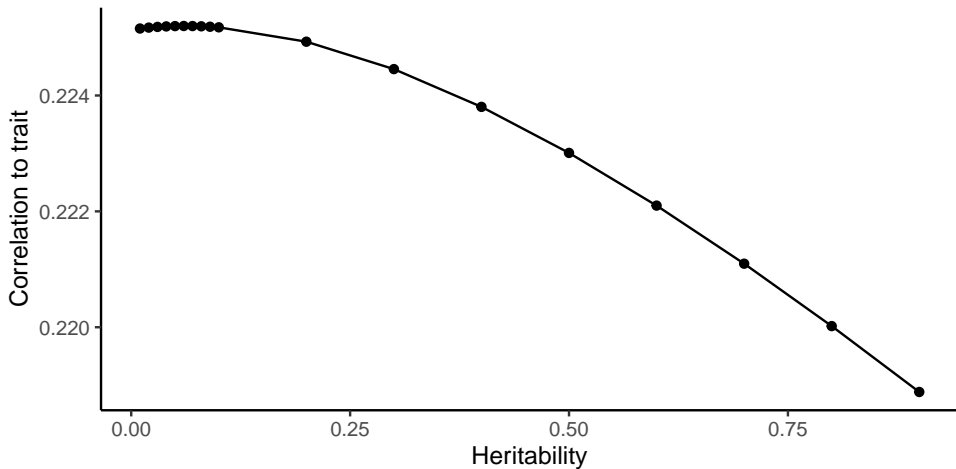
Train results: SC-DCB ldpred2-ct



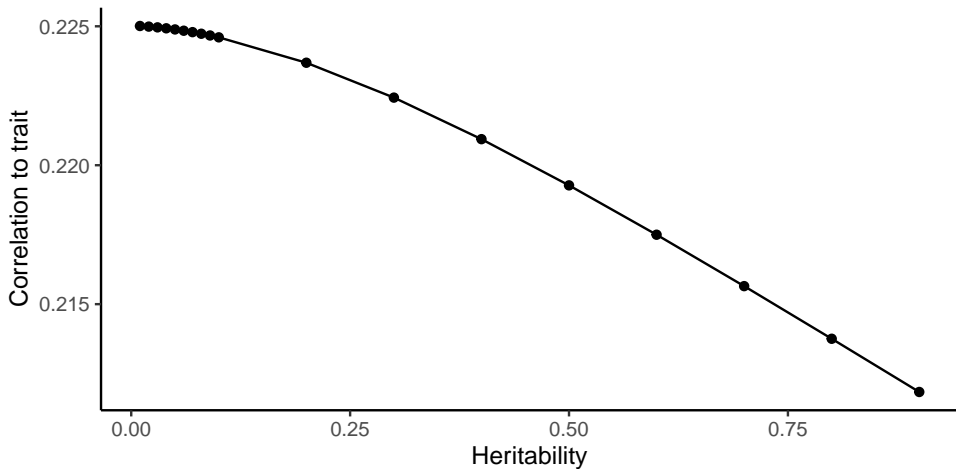
Train results: SR-DCB ldpred2-ct



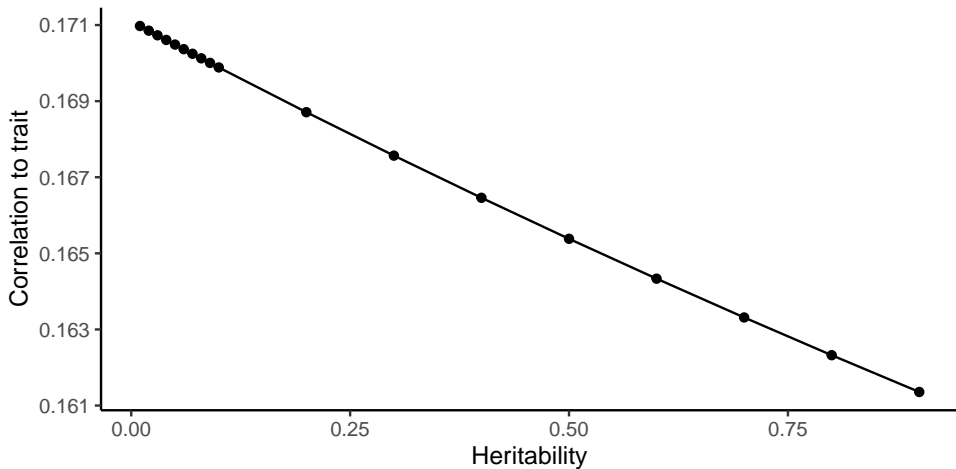
Train results: SC-DDB Idpred2-inf



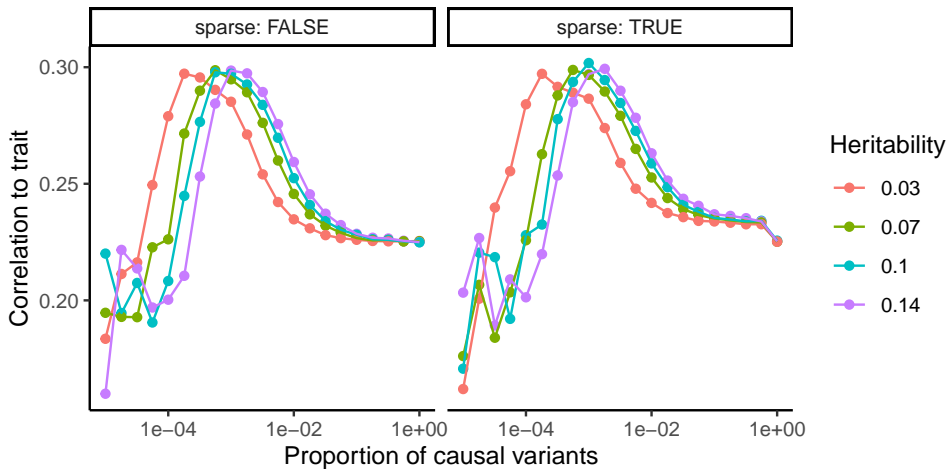
Train results: SC-DCB ldpred2-inf



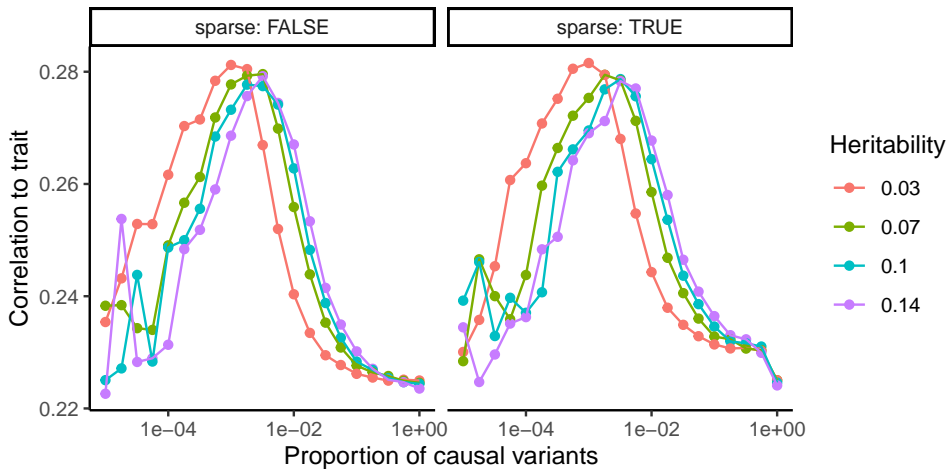
Train results: SR-DCB ldpred2-inf



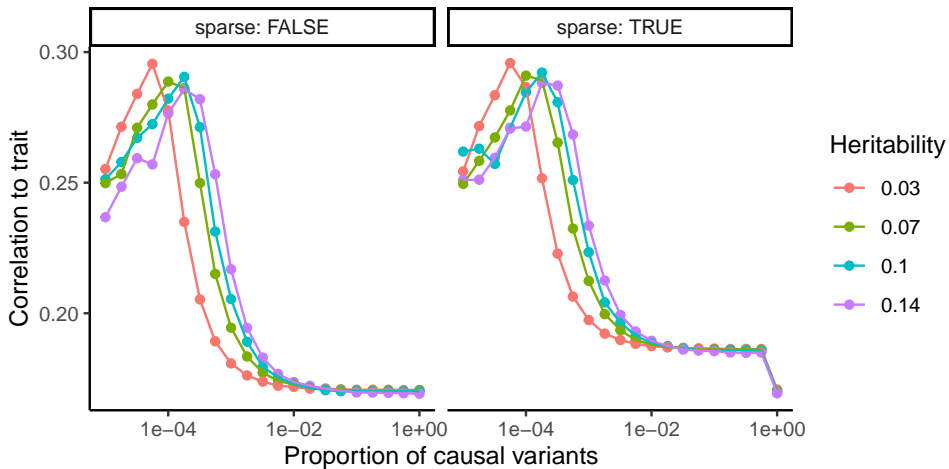
Train results: SC-DDB Idpred2-grid-h0.1



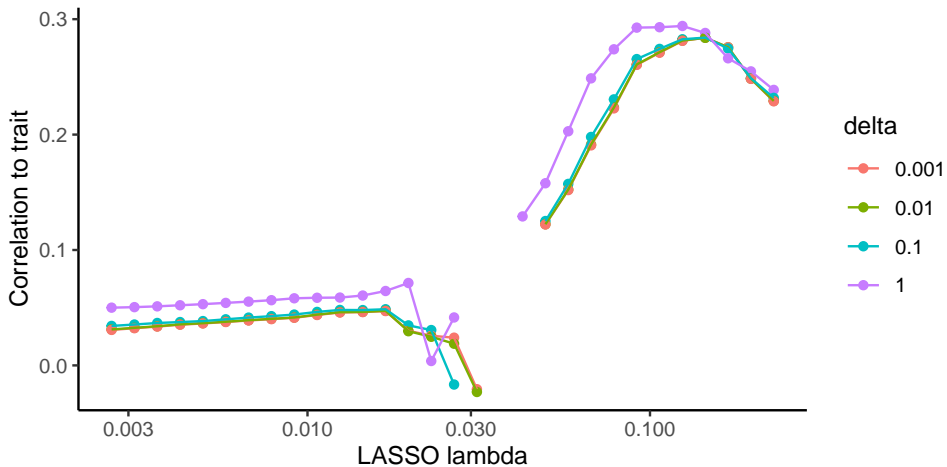
Train results: SC-DCB ldpred2-grid-h0.1



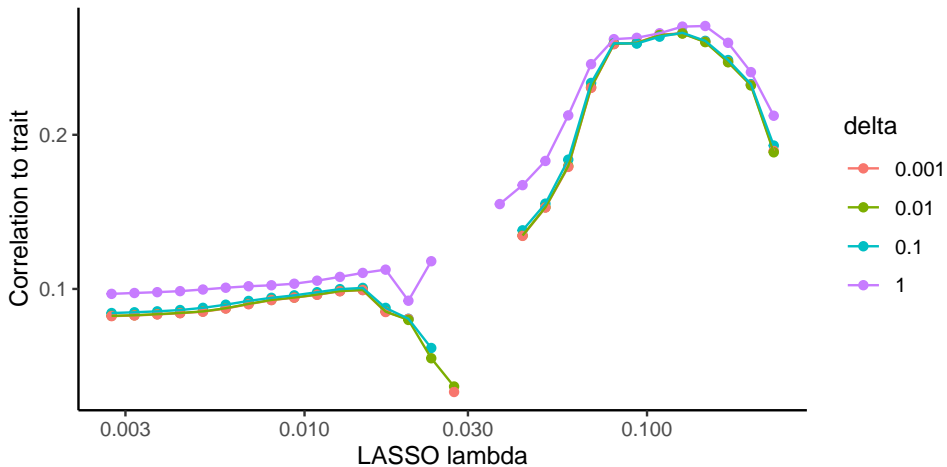
Train results: SR-DCB ldpred2-grid-h0.1



Train results: SC-DDB Idpred2-lassosum



Train results: SC-DCB ldpred2-lassosum



Train results: SR-DCB ldpred2-lassosum

