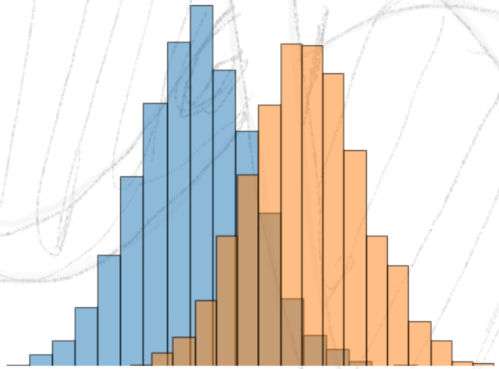
Simple regressions but big (computational) polygenic score gains.

Timothy G Raben (postdoc)



MICHIGAN STATE
UNIVERSITY

Southern California Symposium on Polygenic (Risk) Scores
no conflicts of interest

(My) Background

2010 → 2018 theoretical physicist: PhD, postdoc, etc. Lots of computational work, theoritical work, uncertainty analyses, etc.

2018 → present bioinformatics/statistical genetics. Wide range of interests: methods development, high performance computing, heritability, family studies, trans-ancestry PGS, and much more.

Frequent collaborators:

- Stephen Hsu (MSU & GP,Inc.)
- Louis Lello (GP,Inc. & MSU)
- Erik Widen (GP,Inc. & MSU)
- Academia Sinica (Taiwan)
- UC San Francisco

Office mates:

Polygenic scores

Weights applied to SNP/Vs, CNVs, genes, etc. Two common approaches:

Use GWAS weights/beta values

- "improve" weights with LD information
- "good" gwas can require millions of samples
- "easy" to combine different sources (e.g., GWAS and LD from different biobanks)
- computationally "simple" and easily parallelized
- not easy to extend to multi/trans-ancestries

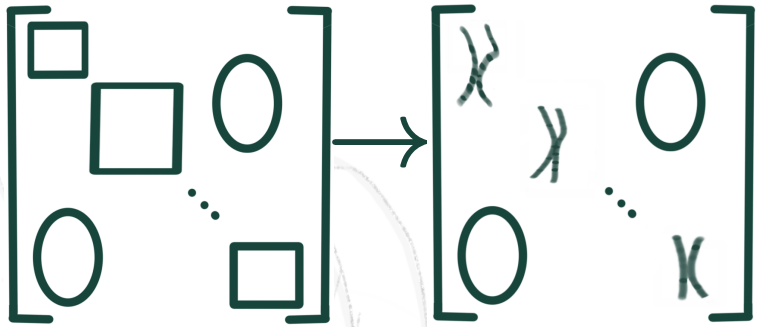
Apply machine learning directly to genotype matrices

- Train directly on correlation structure
- Rigorous compressed sensing theorems for signal recovery
- "good" PGS with < 500k samples
- computationally intensive.
- not easy to extend to multi/trans-ancestries

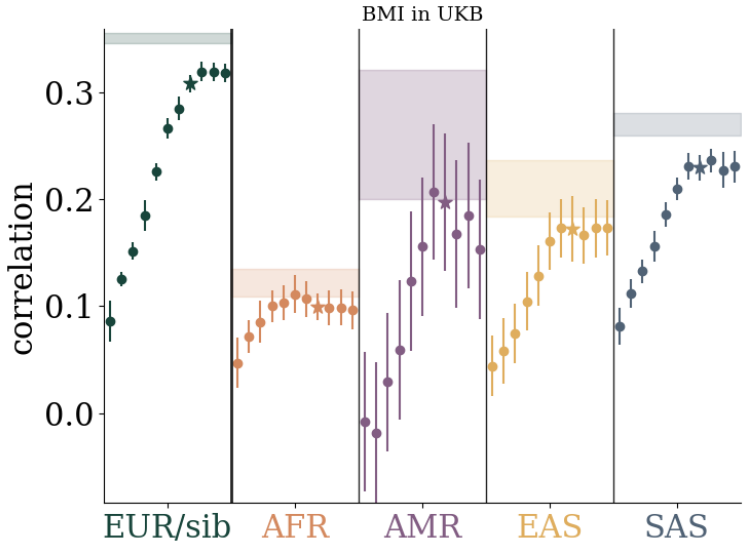
blockLASSO

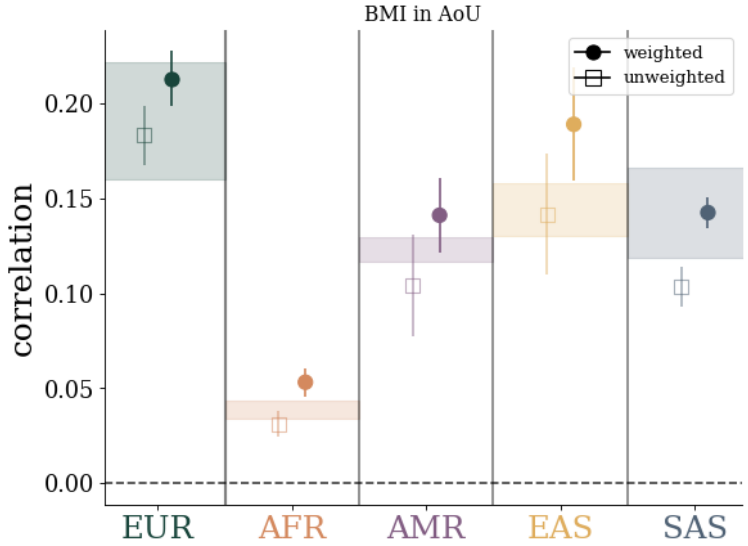
How to reduce the computational needs: use sparsity!

(1) can look at *sparse* algorithms (e.g., LASSO) (2) can enforce "screening" rules/approximations which pre-select a subset of features **(3) (NEW) utilize the approximate block diagonal structure of SNV correlation structure**



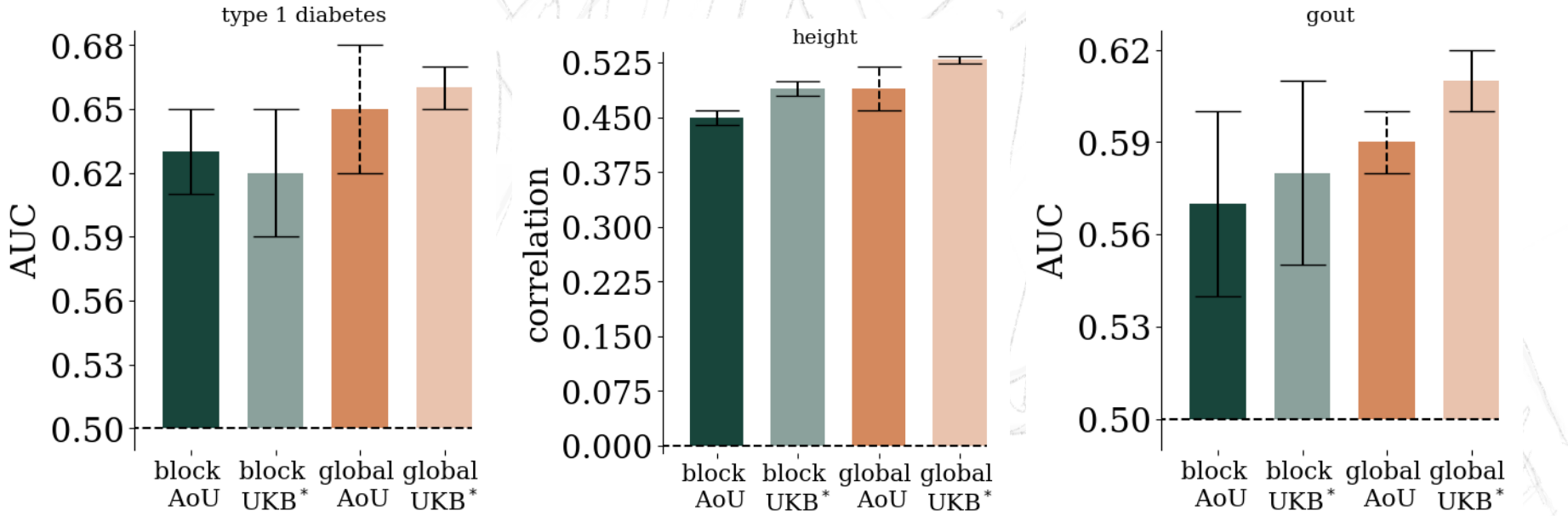
blockLASSO: run LASSO on individual chromosomes then use simple linear regression to piece back together.





blockLASSO

(*genetics only*: PGS compared to phenotypes residualized for covariates.)



Work is currently under peer review. Get a first look at the preprint:
`https://www.medrxiv.org/content/10.1101/2024.06.25.24309482v1`

Conclusions

- blockLASSO has been validated in two biobanks and 11+ phenotypes.
- Variance explained by features is simliar between LASSO and blockLASSO.
- A standard LASSO run can cost $\sim$ \$50 per via standard cloud computing rates (e.g., UKB and AoU) and take 12-24 hours.
- A blockLASSO can be run for $\sim$ \$1 and finishes within minutes
- further improvements can be made by incorporating screening rules, functional information, ancestry specific information, and utilizing warm starts from other predictors.

Interested in collaborating or learning more?
Contact me: rabentim@msu.edu or traben13@gmail.com