A novel, scalable, and efficient blockLASSO PGS method in All of Us and the UK Biobank

Timothy G. Raben [1] Louis Lello [1,2] Erik Widen [1,2] Stephen D.H. Hsu [1,2]

[1] Michigan State University [2] Genomic Prediction, Inc.

Polygenic scores (PGS) are becoming important tools for understanding genetic architecture, identifying potential genetic risk of disease, and now have clinical applications. PGS are traditionally built in one of two ways: (1) starting from single marker regression and adding in additional genomic information or (2) training algorithms executed directly on a subset of the genome. Among *sparse* methods, the simple least absolute shrinkage and selection operator (LASSO) regularly performs among the best methods[1]. **We present an approximate method to approach (2) that can be used in exploratory stages and methods development to very efficiently estimate the performance of polygenic scores without high computational costs.**

Block LASSO

Here we present simple, but novel results about a "block" LASSO[2] which can be run on small chunks of the genome (e.g., individual chromosomes) and then merged together. Comparing separate LASSO regressions can be challenging because the overall scale of effects can be unknown. Over the past decade there have been efforts to speed up LASSO computations by using "safe" and "strong" screening rules[3]. Screening refers to identifying features that will remain with zero weight at successive LASSO hyper-parameter steps, thus reducing the effective dimensionality of the problem at that step. Several works have shown possible improved computational efficiency. The most recent advance involves combining strong screening with an early stopping criterion to find the exact LASSO solution[4]. However, exact solutions can still require $\mathcal{O}(10^2)$ GB of memory and require hours to run. As many biobanks transition to only support third-party cloud computing this can lead to prohibitive costs for researchers who already have access to a university based high performance computing cluster.

trait		AoU		UKB*		UKB	
		block	global	block	global	block	global
asthma	AUC	$0.53_{0.02}$	$\mathbf{0.555_{0.008}}$	$0.57_{0.02}$	$0.579_{0.007}$	$0.606_{0.006}$	$0.623_{0.005}$
gout		$\mathbf{0.57_{0.03}}$	$\mathbf{0.59_{0.01}}$	$\mathbf{0.58_{0.03}}$	$\mathbf{0.61_{0.01}}$	$0.65_{0.01}$	$0.65_{0.01}$
hyperlipidemia		$0.56_{0.01}$	$0.606_{0.007}$	$0.65_{0.01}$	$0.644_{0.003}$	$0.642_{0.004}$	$0.660_{0.003}$
hypertension		$\mathbf{0.53_{0.01}}$	$\mathbf{0.57_{0.01}}$	$\mathbf{0.55_{0.02}}$	$\mathbf{0.57_{0.01}}$	$0.614_{0.004}$	$0.633_{0.003}$
psoriasis		$0.54_{0.02}$	$0.58_{0.02}$	–	–	$0.67_{0.01}$	$0.68_{0.01}$
type 1 diabetes		$\mathbf{0.63_{0.02}}$	$0.65_{0.03}$	$\mathbf{0.62_{0.03}}$	$0.66_{0.01}$	$0.66_{0.02}$	$0.67_{0.02}$
type 2 diabetes		$\mathbf{0.57_{0.01}}$	$\mathbf{0.62_{0.02}}$	$\mathbf{0.58_{0.01}}$	$\mathbf{0.60_{0.02}}$	$0.63_{0.01}$	$0.635_{0.007}$
bmi	corr.	$\mathbf{0.21_{0.01}}$	$\mathbf{0.19_{0.03}}$	$\mathbf{0.19_{0.02}}$	$0.210_{0.007}$	$0.308_{0.008}$	$0.350_{0.005}$
hdl		$\mathbf{0.30_{0.01}}$	$\mathbf{0.37_{0.03}}$	$\mathbf{0.33_{0.02}}$	$\mathbf{0.33_{0.02}}$	$0.429_{0.007}$	$0.458_{0.004}$
height		$0.45_{0.01}$	$0.49_{0.03}$	$0.49_{0.01}$	$0.529_{0.005}$	$0.595_{0.004}$	$0.630_{0.004}$
total bilirubin		$0.41_{0.01}$	$0.52_{0.03}$	$0.56_{0.02}$	$0.57_{0.006}$	$0.578_{0.005}$	$0.590_{0.004}$

Table 1. PGS metrics for results in AoU, the UKB trained with sets matching the size of those found in AoU (UKB*), and for the UKB using the maximum possible training size. All predictors are trained and tested on European populations. Blocks colored 🟩 indicate that the block and traditional results agree within uncertainty. Blocks colored 🟧 indicate that the block and traditional results disagree by less than 20%. Finally, **bold text** indicates that the results between AoU and UKB (either block or traditional) are in agreement within uncertainty.

PGS via Penalized Regression

We can model phenotypes as a linear combination of genetics, environment, and then some error: $\vec{y} = \vec{\beta}\bar{X} + \vec{\theta}\bar{K} + \vec{\epsilon}$. After modeling the covariates, we can residualize the phenotypes: $\vec{y} \to \vec{y}^*$. Penalized regression algorithms can be described by minimizing the objective function

$$\mathcal{O}(\lambda) = \frac{1}{2N}||\vec{y^*} - \bar{X}\cdot\vec{\beta}||^2_{L_2} + P(\lambda, \vec{\beta})\,,$$

where $\vec{y^*}$ is the residualized phenotype, $\bar{X}$ is the genotype matrix, $\vec{\beta}$ are the model weights, and λ is a hyper-parameter. The final term is the penalty which for LASSO is The LASSO algorithm can be described by minimizing the objective function $P(\lambda, \vec{\beta}) = |\vec{\beta}|_{L_1}$.

We can also compute the covariance between each feature and sum the contribution from all correlated features. Within each block, the encoded genotype can be written ${}^bX_{i,j}$ where b labels the block, i labels the sample, and j labels the feature (SNV). We then include the weights for each feature, ${}^b\beta_j$, and each block, α_b: $\alpha_b\,{}^b\beta_j\,{}^bX_{i,j} \equiv {}^bH_{i,j}$. Finally we compute the covariance matrix $K_{j,k} = COV(\vec{H}_j, \vec{H}_k)$,where $\vec{H}_j$ is the column of sample values for the jth feature. Note that when computing the covariance between features (SNVs) on different blocks the covariance is assumed 0 by definition. By comparing the block LASSO to the traditional LASSO we see that the block method recovers the same important regions with similar weights.

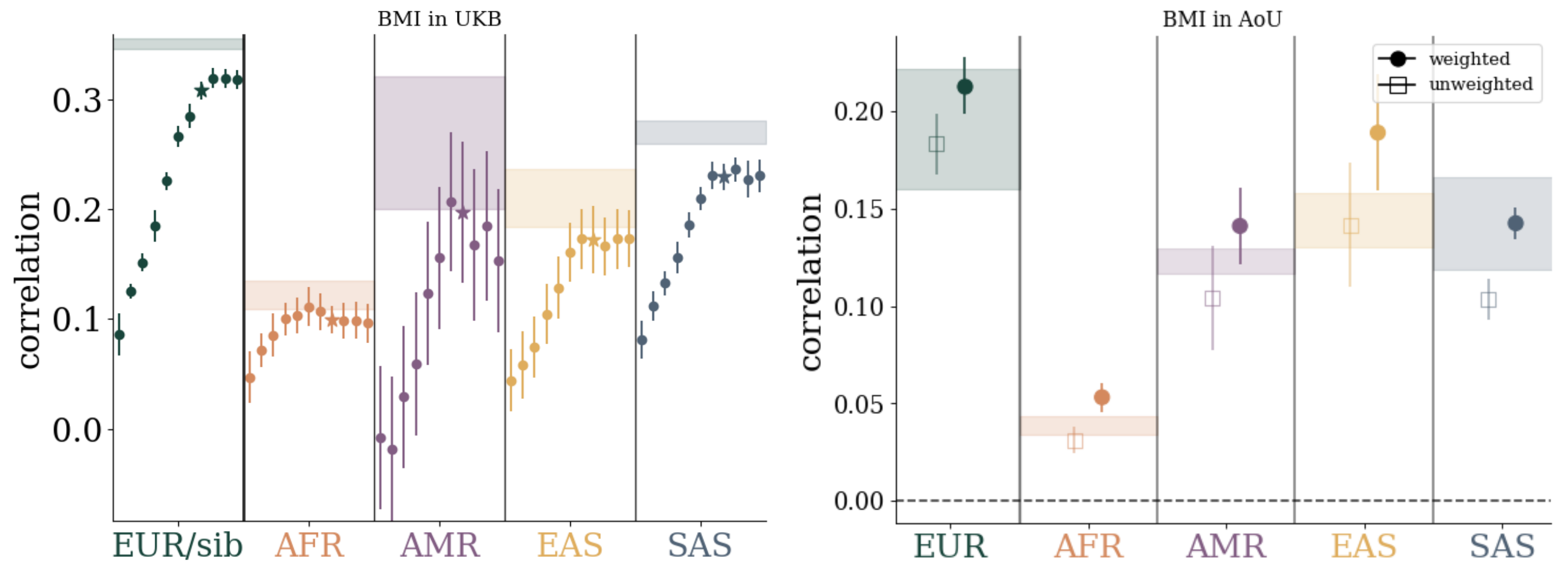

Figure 1. Left: performance as a function of training SNV size in UKB and applied to different ancestry groups. Within each ancestry group, dots correspond to training with different numbers of SNVs per chromosome. Right: performance in AoU before and after the re-weighting step of the blockLASSO and compared to the predicted[2] global result (shaded bands).

Computational Resources

The main advantage of the block LASSO approach is massively reduced computational requirements. Global memory requirements scale linearly (i.e., with the number of features). blockLASSO can use orders of magnitude less features per block leading to large time, memory, and third-party cost savings. Over a wide range of sparsities, i.e., predictors ranging from a few hundred to a tens of thousands of features, the block approach shows significant gains.

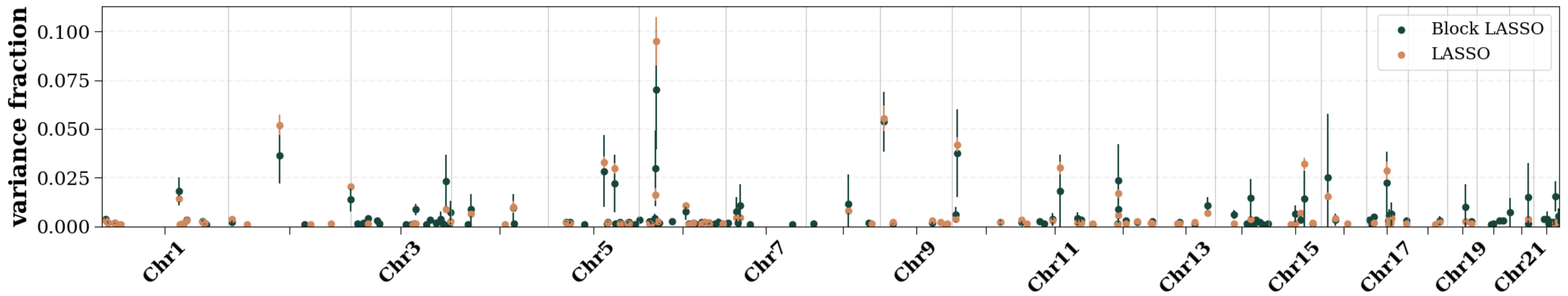

Figure 2. Fraction of predicted variance accounted for of an asthma predictor as a function of approximate location across the autosome. Error bars represent averaging over cross-validation. The block approach finds similarly important regions as the global approach (e.g., the HLA/MHC region above), but includes more spurious signals accounting for small amounts of variance.

Conclusions

- ML algorithms exploiting block correlational structure (e.g., chromosome by chromosome), generate best-in-class PGS while requiring only modest compute resources. Examples include LDpred2, PRS-cs, and now *penalized regression*. Many of these computations can be run on a laptop.
- A chromosome-by-chromosome (block) LASSO can produce PGS comparable to traditional LASSO based approaches in less than 10 minutes and only using 8GB of RAM.
- Block approaches are: more "parallelizable", faster run times, reduced computational demand, cheaper, more environmentally friendly.
- This approach relies on the fact that correlations *between* SNPs on different chromosomes are generally small.
- Different machine learning methods, novel re-weighting approaches, and incorporating additional information (e.g., functional info, other -omics, etc.) can still improve block approaches.

References, Acknowledgments, & Github

We gratefully acknowledge both All of Us and UK Biobank participants for their contributions, without whom this research would not have been possible. We also thank the National Institutes of Health's All of Us Research Program for making available the participant data examined in this study. This work is carried out using data sets via UK Biobank main application 15326 and in the All of Us Research Workbench workspace "Scalable and efficient polygenic scores in diverse populations".

Additional thanks to the Michigan State University Institute for Cyber-Enabled Research and collaborators at UC San Francisco and Academia Sinica.

[1.] **T.G. Raben**, et al. doi.org/10.1038/s41598-023-37580-5
[2.] **T.G. Raben**, et al. doi.org/10.1101/2024.06.25.24309482
[3.] R. Tibshirani, et al. doi.org/10.1111/j.1467-9868.2011.01004.x
[4.] F. Privé. et al. doi.org/10.1534/genetics.119.302019
[5.] L. Lello, **T.G. Raben**, et al. doi.org/10.1038/s41598-020-69927-7

Hsu group at Michigan State University ASHG 2024 Denver CO rabentim@msu.edu