Efficient variance components analysis across millions of genomes

Ali Pazokitoroudi[1], Yue Wu[1], Kathryn S. Burch[2], Kangcheng Hou[3], Bogdan Pasaniuc[4,5,6], Sriram Sankararaman[1,5,6,*]

[1]Department of Computer Science, UCLA, Los Angeles, California [2]Bioinformatics Interdepartmental Program, UCLA, Los Angeles, California [3]Department of Pathology and Laboratory Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California [4]College of Computer Science and Technology, Zhejiang University, Hangzhou, Zhejiang, China [5]Department of Human Genetics, David Geffen School of Medicine, UCLA, Los Angeles, California [6]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, California

* Contact: `sriram@cs.ucla.edu`

Abstract

While variance components analysis has emerged as a powerful tool in complex trait genetics, existing methods for fitting variance components do not scale well to large-scale datasets of genetic variation. Here we present a method for variance components analysis that is accurate and efficient: capable of estimating one hundred variance components on a million individuals genotyped at a million SNPs in a few hours. We illustrate the utility of our method in estimating and partitioning variation in a trait explained by genotyped SNPs (SNP-heritability). Analyzing 22 traits with genotypes from 300,000 individuals across about 8 million common and low frequency SNPs, we observe that per-allele effect sizes increases with decreasing minor allele frequency (MAF) and linkage disequilibrium (LD) consistent with the action of negative selection. Partitioning heritability across 28 functional annotations, we observe enrichment of heritability in FANTOM5 enhancers in asthma, eczema, thyroid and autoimmune disorders.

Multi-component Linear Mixed Models (LMMs)

$$y = \sum_{k=1}^{K} X_k\beta_k + \epsilon$$
$$\epsilon \sim \mathcal{N}(0, \sigma_e^2 I_N)$$
$$\beta_k \sim \mathcal{N}(0, \frac{\sigma_k^2}{M_k} I_{M_k}), k \in \{1, \ldots, K\} \quad (1)$$

- Each of the M SNPs is assigned to one of K non-overlapping categories.
- β_k: M_k-vector of SNP effect sizes for the k-th category.
- σ_e^2: the residual variance and σ_k^2: the variance component of the k-th category.

The total SNP heritability :

$$h_{SNP}^2 = \frac{\sum_{k=1}^K \sigma_k^2}{(\sum_{k=1}^K \sigma_k^2) + \sigma_e^2} \quad (2)$$

The SNP heritability of category k :

$$h_k^2 = \frac{\sigma_k^2}{(\sum_{k=1}^K \sigma_k^2) + \sigma_e^2}, k \in \{1, \ldots, K\} \quad (3)$$

Method of Moments

$$cov(y) = E[yy^T] - E[y]E[y^T] = \sum_k \sigma_k^2 K_k + \sigma_e^2 I_N \quad (4)$$

- $K_k = \frac{X_k X_k^T}{M_k}$: the genetic relatedness matrix (GRM) computed from all SNPs of k-th category.

Method of Moments objective function

Using yy^T as our estimate of the empirical covariance, solve the following problem to find the variance components.

$$(\hat{\sigma}_1^2, \ldots, \hat{\sigma}_K^2, \hat{\sigma}_e^2) = argmin_{(\sigma_1^2, \ldots, \sigma_K^2, \sigma_e^2)} ||yy^T - \sum_k \sigma_k^2 K_k + \sigma_e^2 I||_F^2 \quad (5)$$

The method of moments estimate satisfies the normal equations:

$$\begin{bmatrix} T & b \\ b^T & N \end{bmatrix} \begin{bmatrix} \hat{\sigma}_1^2 \\ \vdots \\ \hat{\sigma}_K^2 \\ \hat{\sigma}_e^2 \end{bmatrix} = \begin{bmatrix} c \\ y^T y \end{bmatrix} \quad (6)$$

- T: $K \times K$ matrix with entries $T_{k,l} = tr(K_k K_l), k, l \in \{1, \ldots, K\}$
- b: K-vector with entries $b_k = tr(K_k) = N$
- c: K-vector with entries $c_k = y^T K_k y$.
- Computational complexity: GRM K_k: $\mathcal{O}(N^2 M_k)$. Computing $T_{k,l}, c_k, k, l \in \{1, \ldots, K\}$: $\mathcal{O}(N^2)$. Solving the normal equation: $\mathcal{O}(K^3)$

Scalable Method-of-Moments estimator (RHE-mc)

The key bottleneck in solving the normal equation: the computation of $K_k, T_{k,l}, c_k$
Estimate $T_{k,l}, k, l \in \{1, \ldots, K\}$ using an unbiased estimator:

$$T_{k,l} = tr(K_k K_l) \approx \widehat{T_{k,l}} = \frac{1}{B}\frac{1}{M_k M_l}\sum_b z_b^T X_k X_k^T X_l X_l^T z_b \quad (7)$$

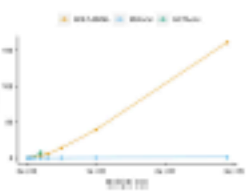
- $z_1, \ldots, z_B$: B independent random vectors with zero mean and covariance I_N.
- No explicit GRM computation.
- Estimating the values $T_{k,l}$ is $\mathcal{O}(NMB)$.
- Can improve the time complexity to $\mathcal{O}(\frac{NMB}{max(log_3 N, log_3 M)})$.

Overall time complexity: $\mathcal{O}(\frac{NMB}{max(log_3(N), log_3(M))} + K^3)$.

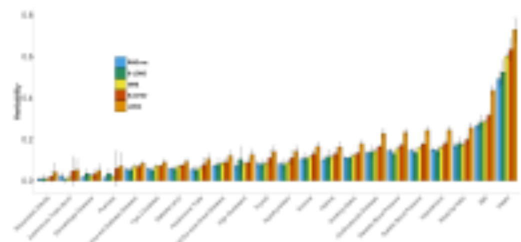
Results

RHE-mc yields relatively unbiased estimates of total SNP heritability

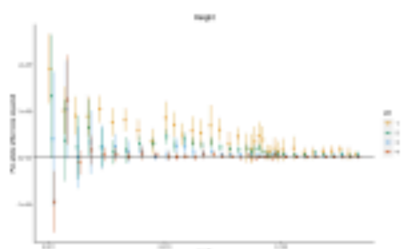
RHE-mc is scalable



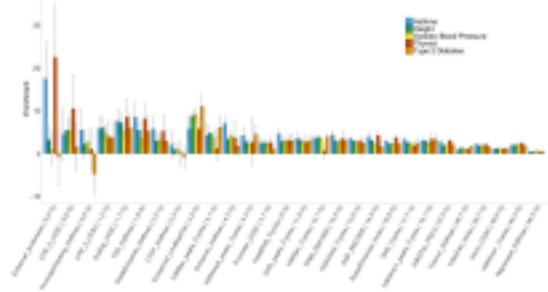
Estimates of genome-wide SNP heritability in the UK Biobank



Per-allele squared effect size of height as a function of MAF and LD (see [1] for other traits)



RHE-mc estimates of heritability enrichment across functional annotations in the UK Biobank



References

[1] Ali Pazokitoroudi, Yue Wu, Kathryn S. Burch, Kangcheng Hou, Bogdan Pasaniuc, and Sriram Sankararaman. Efficient variance components analysis across millions of genomes. *Nature Communications*, 11(4020), 2020.

Acknowledgements

This research was conducted using the UK Biobank Resource under applications 33127 and 33297. We thank the participants of UK Biobank for making this work possible. We thank Rob Brown for feedback on this manuscript. This work was funded by NIH grants R01HG009120 (B.P. and K.S.B.), R35GM125055 (S.S.), an Alfred P. Sloan Research Fellowship (S.S.), and a NSF grant III-1705121 (Y.W. and S.S.).