GEPSi: A Python Library to Simulate GWAS Phenotype Data

Daniel A. Reidenbach^{1,2}, Avantika Lal², Lotfi Slim³, Ohad Mosafi², Johnny Israeli²

¹ Department of Electrical Engineering and Computer Sciences, UC Berkeley

² NVIDIA Corporation, Santa Clara, CA, USA

³ NVIDIA Corporation, Reading, UK

Abstract

Motivation

Many computational methods aim to identify genetic variants associated with diseases and complex traits. Due to the absence of ground truth data, simulated genotype and phenotype data is needed to benchmark these methods. However, phenotypes are frequently simulated as an additive function of randomly selected variants, neglecting biological complexity such as non-random occurrence of causal SNPs, epistatic effects, heritability and dominance. Including such features would improve benchmarking studies and accelerate the development of methods for genetic analysis.

Results

Here, we describe GEPSi (**GWAS Epistatic Phenotype Simulator**), a user-friendly python package to simulate phenotype data based on user-supplied genotype data for a population. GEPSi incorporates diverse biological parameters such as heritability, dominance, population stratification and epistatic interactions between SNPs. We demonstrate the use of this package to compare machine learning methods for GWAS analysis.

Availability and Implementation

GEPSi is freely available under an Apache 2.0 license, and can be downloaded from <https://github.com/clara-parabricks/GEPSi>.

Supplementary information

Supplementary data are available online.

1 Introduction

Numerous computational methods have been developed to analyze Genome-Wide Association Study (GWAS) datasets, in order to identify germline genetic variants that are associated with phenotypic traits in a population. These include mixed linear models (Loh et al., 2015; Mbatchou et al., 2021), penalized logistic regression (Hoffman et al., 2013; Yang et al., 2020), decision trees (Botta et al., 2014; Schwarz et al., 2010), and neural networks (Arloth et al., 2020; van Hilten et al., 2020).

Due to the absence of ‘ground truth’ data where all the genetic variants associated with a phenotype are known with certainty, the small number of large-scale GWAS datasets available, and the protections surrounding access to such data, these methods are frequently developed using simulated genotype and phenotype data. This allows developers to benchmark their method on large and diverse datasets.

There are several existing tools to simulate genotypes based on diverse assumptions (Dimitromanolakis et al., 2019; Su et al., 2011). While phenotype simulation tools exist, most of these either randomly select causal SNPs (Meyer & Birney, 2018) or ask the user to specify them (Li & Li, 2008; Su et al., 2011). Further, it is common to assume a linear and independent additive genetic model, where the risk for an individual is a summation of risks for each causal SNP (Dimitromanolakis et al., 2019; Marigorta & Gibson, 2014; Meyer & Birney, 2018; Porter & O’Reilly, 2017).

However, causal SNPs for a trait are not randomly distributed throughout the genome, but instead are likely to directly or indirectly affect a common set of genes or pathways related to the phenotype. Moreover, various epistatic interactions between SNPs (Phillips, 2008) lead to non-linear relationships between genotype and phenotype. Although a few previous tools have included the option to specify non-additive effects for SNPs (Blumenthal et al., 2020; Tang & Liu, 2019), to our knowledge there is currently no package dedicated to complex phenotypic simulations that incorporates nonlinear phenotype models, heritability, stratification and non-random selection of causal SNPs.

Here, we describe GEPSi, a python package to simulate phenotype data for a population based on given genotype data. GEPSi takes as input the real or simulated genotypes for all the individuals in a population. It simulates the phenotype of each individual based on user-specified parameters that control causal SNP selection, heritability, dominance, stratification and epistatic interactions between SNPs.

2 Workflow and Implementation

GEPSi requires two input files - a genotype matrix file in the ‘.raw’ format that specifies the genotype for each individual, and a SNP annotation file in the ‘.bim’ format that describes each

SNP in the dataset. These files can be produced easily from other common formats using PLINK (Purcell et al., 2007).

GEPSi then simulates the phenotype for each individual causing the following steps (Figure 1A):

1. A fraction of all available SNPs are selected to be causal. Selection can be random or gene-based. In gene-based selection, causal genes are first randomly selected and then causal SNPs are randomly selected from among the SNPs in the causal genes. This gene-based selection currently cannot be used to simulate causal intergenic variants.
2. A risk is assigned to each causal SNP by sampling from a normal distribution. Causal SNPs can be dominant, recessive, or dose-dependent with the user selecting the proportion belonging to each type.
3. Epistatic interactions are assigned to a fraction of the causal SNPs. Interactions between SNPs can be masking (in which one SNP masks or negates the effect of the other) or multiplicative (where one SNP increases or decreases the effect of the other).
4. The total genetic risk for each individual is calculated as the sum of the risks for all the causal SNPs based on their genotype.
5. The genetic risk for each individual is updated based on the heritability of the phenotype, according to the following equation:

$$g' = h * g + \sqrt{(1 - h^2)} * N(0, 1)$$
 where h is the heritability and g the genetic risk score for the individual.
6. Population stratification may be incorporated by adding user-specified coefficients to the risk scores of individuals belonging to each group within the population.

User-controlled parameters include the causal SNP selection method, the number of causal SNPs of each type, the number of inter-SNP interactions of each type, the strength of inter-SNP interactions, the heritability of the phenotype, the stratification of the population into groups, and the proportion of cases and controls in the population. The runtime of GEPSi with different settings is given in Supplementary Table 1.

3 Use Case

To demonstrate GEPSi, we applied HAPGEN2 (Su et al., 2011) to simulate chromosome 1 genotypes for a population of 10,000 individuals. We filtered 38,300 protein-coding SNPs for this analysis.

We then simulated phenotype data with three settings, each time selecting 1,290 of the 38,300 SNPs to be causal and setting 50% of the population to be cases while the other 50% were designated controls. The three simulations differed as follows:

1. Causal SNPs were randomly selected. Epistatic interactions were simulated for 20% of the causal SNPs.
2. Causal SNPs were clustered within 218 causal genes. Epistatic interactions were simulated for 20% of the causal SNPs.

- Causal SNPs were clustered within 218 causal genes. Epistatic interactions were not included.

We then applied three machine learning methods - CELER Lasso (Massias et al., 2018), Xgboost, and GenNet (van Hilten et al., 2020) - to each simulated dataset, to classify individuals into cases and controls and to score the SNPs based on their association with the predicted phenotype (Supplementary Tables 2 and 3). All three methods were compared using the AUROC (Area Under the Receiver-Operator Characteristic) metric (Supplementary Tables 4 and 5).

GenNet, a deep learning method, performed best at identifying causal SNPs under all conditions (Figure 1B). While the performance of CELER and Xgboost was relatively constant across simulations, GenNet's performance improved when causal SNPs were clustered within causal genes, rather than randomly selected (Figure 1B). Further, all methods performed slightly worse when epistatic interactions were included (Figure 1B). This example illustrates how GEPSi can be used to compare GWAS methods and to understand how genetic parameters influence model performance.

Figure 1: Benchmarking GWAS methods using GEPSi. A. Schematic of the GEPSi simulation procedure. B. Performance of three methods (GenNet, CELER and Xgboost) at

identifying causal SNPs in three simulated GWAS datasets. While the genotype data was the same in all datasets, the phenotypes were simulated using GEPSi with three different settings.

4 Conclusion

GEPSi enables easy benchmarking of GWAS software on simulated phenotypes based on a range of conditions and biological assumptions. It can also be used to benchmark other methods such as those to identify epistatic interactions (Niel et al., 2015).

In future, this framework can be extended to incorporate joint simulation of multiple phenotypes and epistatic interactions involving more than two SNPs or genes. Tissue-specific functional and pathway information, such as enhancer-gene maps (Nasser et al., 2021), can also be incorporated to simulate more realistic genotype-phenotype associations for the coding as well as non-coding genome.

References

1. Arloth, J., Eraslan, G., Andlauer, T. F. M., Martins, J., Iurato, S., Kühnel, B., Waldenberger, M., Frank, J., Gold, R., Hemmer, B., Luessi, F., Nischwitz, S., Paul, F., Wiendl, H., Gieger, C., Heilmann-Heimbach, S., Kacprowski, T., Laudes, M., Meitinger, T., ... Mueller, N. S. (2020). DeepWAS: Multivariate genotype-phenotype associations by directly integrating regulatory information using deep learning. *PLoS Computational Biology*, 16(2), e1007616.
2. Blumenthal, D. B., Viola, L., List, M., Baumbach, J., Tieri, P., & Kacprowski, T. (2020). EpiGEN: an epistasis simulation pipeline. *Bioinformatics*, 36(19), 4957–4959.
3. Botta, V., Louppe, G., Geurts, P., & Wehenkel, L. (2014). Exploiting SNP correlations within random forest for genome-wide association studies. *PloS One*, 9(4), e93379.
4. Dimitromanolakis, A., Xu, J., Krol, A., & Briollais, L. (2019). sim1000G: a user-friendly genetic variant simulator in R for unrelated individuals and family-based designs. *BMC Bioinformatics*, 20(1), 26.

5. Hoffman, G. E., Logsdon, B. A., & Mezey, J. G. (2013). PUMA: A Unified Framework for Penalized Multiple Regression Analysis of GWAS Data. In *PLoS Computational Biology* (Vol. 9, Issue 6, p. e1003101). <https://doi.org/10.1371/journal.pcbi.1003101>
6. Li, C., & Li, M. (2008). GWASimulator: a rapid whole-genome simulation program. *Bioinformatics* , 24(1), 140–142.
7. Loh, P.-R., Tucker, G., Bulik-Sullivan, B. K., Vilhjálmsson, B. J., Finucane, H. K., Salem, R. M., Chasman, D. I., Ridker, P. M., Neale, B. M., Berger, B., Patterson, N., & Price, A. L. (2015). Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nature Genetics*, 47(3), 284–290.
8. Marigorta, U. M., & Gibson, G. (2014). A simulation study of gene-by-environment interactions in GWAS implies ample hidden effects. *Frontiers in Genetics*, 5, 225.
9. Massias, M., Gramfort, A., & Salmon, J. (2018). Celer: a Fast Solver for the Lasso with Dual Extrapolation. In J. Dy & A. Krause (Eds.), *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 3315–3324). PMLR.
10. Mbatchou, J., Barnard, L., Backman, J., Marcketta, A., Kosmicki, J. A., Ziyatdinov, A., Benner, C., O'Dushlaine, C., Barber, M., Boutkov, B., Habegger, L., Ferreira, M., Baras, A., Reid, J., Abecasis, G., Maxwell, E., & Marchini, J. (2021). Computationally efficient whole-genome regression for quantitative and binary traits. *Nature Genetics*.
<https://doi.org/10.1038/s41588-021-00870-7>
11. Meyer, H. V., & Birney, E. (2018). PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* , 34(17), 2951–2956.
12. Nasser, J., Bergman, D. T., Fulco, C. P., Guckelberger, P., Doughty, B. R., Patwardhan, T. A., Jones, T. R., Nguyen, T. H., Ulirsch, J. C., Lekschas, F., Mualim, K., Natri, H. M., Weeks, E. M., Munson, G., Kane, M., Kang, H. Y., Cui, A., Ray, J. P., Eisenhaure, T. M., ... Engreitz, J. M. (2021). Genome-wide enhancer maps link risk variants to disease

- genes. *Nature*, 593(7858), 238–243.
13. Niel, C., Sinoquet, C., Dina, C., & Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6.
<https://doi.org/10.3389/fgene.2015.00285>
14. Phillips, P. C. (2008). Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews. Genetics*, 9(11), 855–867.
15. Porter, H. F., & O'Reilly, P. F. (2017). Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports*, 7, 38837.
16. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, 81(3), 559–575.
17. Schwarz, D. F., König, I. R., & Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics*, 26(14), 1752–1758.
18. Su, Z., Marchini, J., & Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16), 2304–2305.
19. Tang, Y., & Liu, X. (2019). G2P: a Genome-Wide-Association-Study simulation tool for genotype simulation, phenotype simulation and power evaluation. *Bioinformatics*, 35(19), 3852–3854.
20. van Hilten, A., Kushner, S. A., Kayser, M., Arfan Ikram, M., Adams, H. H. H., Klaver, C. C. W., Niessen, W. J., & Roshchupkin, G. V. (2020). GenNet framework: interpretable neural networks for phenotype prediction. In *Cold Spring Harbor Laboratory* (p. 2020.06.19.159152). <https://doi.org/10.1101/2020.06.19.159152>
21. Yang, S., Wen, J., Eckert, S. T., Wang, Y., Liu, D. J., Wu, R., Li, R., & Zhan, X. (2020). Prioritizing genetic variants in GWAS with lasso using permutation-assisted tuning.

Bioinformatics , 36(12), 3811–3817.