simplePHENOTYPES: SIMulation of Pleiotropic, Linked and Epistatic PHENOTYPES

Samuel B. Fernandes[1], Alexander E. Lipka[1,*]

1 Department of Crop Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

* **To whom correspondence should be addressed** *alipka@illinois.edu*

Abstract

Motivation: Advances in genotyping and phenotyping techniques have enabled the acquisition of a great amount of data. Consequently, there is an interest in multivariate statistical analyses that identify genomic regions likely to contain causal mutations affecting multiple phenotypes (i.e., pleiotropy). As the demand for multivariate analyses increases, it is imperative that optimal tools are available to validate and compare different implementations of these analyses. To facilitate the testing and validation of these multivariate approaches, we implemented simplePHENOTYPES, an R package that simulates pleiotropy, partial pleiotropy, and spurious pleiotropy in a wide range of genetic mechanisms, including additive, dominance and epistatic models.
Results: We compared simplePHENOTYPES to other simulation packages and we showed that, under comparable circumstances, all packages produce equivalent results. Using standard GWAS software (GEMMA), we confirmed that the genetic signal is correctly simulated under different genetic architectures. Finally, we illustrate simplePHENOTYPE's ability to simulate thousands of phenotypes in less than one minute.
Availability: simplePHENOTYPES is freely available from CRAN. Developmental version, as well as vignettes, are available on GitHub (https://github.com/samuelbfernandes/simplePHENOTYPES).

1 Introduction

The wealth of data available from high-throughput phenotyping platforms used in modern agronomical experiments is enabling unprecedented study into the genomic underpinnings of genotype-to-phenotype relationships [2, 12, 10]. In particular, these data are making it possible to gain insight into the simultaneous contributions of genomic loci to multiple phenotypes, a phenomenon known as pleiotropy [11]. To ensure that accurate inferences are being made from these data, the most appropriate statistical approaches must be used. One avenue towards assessing the performance of such approaches is to simulate correlated phenotypic data in which the pleiotropic and non-pleiotropic quantitative trait nucleotides (QTNs) underlying the genomic sources of phenotypic variability are known.

Efforts in developing software for simulating correlated phenotypes resulted in several useful packages [6, 8, 9]. However, there is not a single package to bring together all the parameters involved in the genetic architecture of complex traits. For instance, to the best of our knowledge, no package is currently set for simulating spurious pleiotropy as defined by Solovieff et al. [11]. In this work, we present the R/CRAN package simplePHENOTYPES. This package will use real marker data to simulate additive, dominance, and additive x additive epistatic QTN controlling multiple traits via pleiotropy, partial pleiotropy, and spurious pleiotropy. We maximize its accessibility to the research community by ensuring its compatibility with popular data formats and state-of-the-art GWAS and GS software. Developmental versions, as well as vignettes, may be found at https://github.com/samuelbfernandes/simplePHENOTYPES.

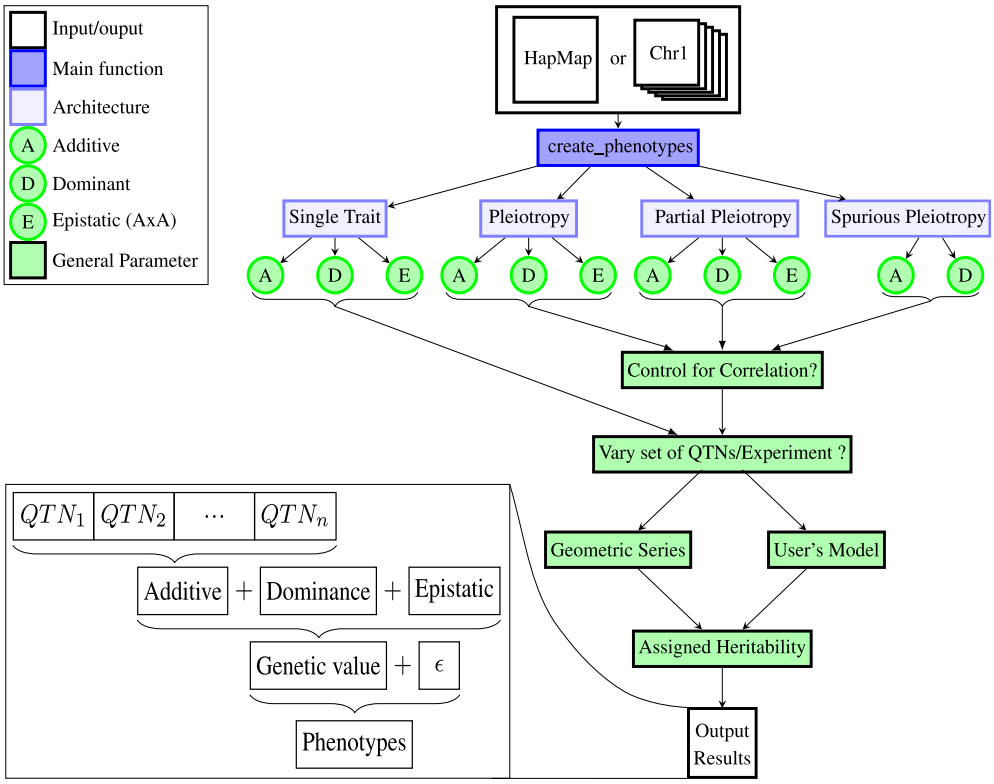

Figure 1: simplePHENOTYPES workflow.

2 Methods and Implementation

simplePHENOTYPES is capable of simulating traits with a wide range of genetic architectures, as depicted in Figure 1. After reading biallelic marker data into R, the user specifies the desired genetic architecture using the *create_phenotypes()* function. There is substantial flexibility with respect to specifying the genetic architecture of these traits, specifically with the number of additive and non-additive (i.e. dominance and additive x additive epistasis) QTNs controlling each trait and whether each QTN makes pleiotropic or non-pleiotropic contributions. Additionally, it is possible to assign specific Pearson correlations between genetic values of different traits. simplePHENOTYPES also allows the user to specify if the markers randomly selected to be QTNs are to differ across each replicate trait. Next, the user will either manually input vectors of the desired effect sizes of each QTN or allow the effect sizes to follow a geometric series [5]. Finally, user-specified trait heritabilities will be used to simulate normally distributed random error terms that represent non-genetic sources of trait variability. Thus, for a given individual, the resulting simulated trait value is the sum of the simulated genetic effects and this random error term.

Figure 2 describes the three different types of pleiotropy that can be simulated. For illustration, let us consider a pool of 20 QTNs and two traits ($Trait\ j$ and $Trait\ j'$). In the pleiotropy scenario, all simulated traits will be controlled by the same 20 QTNs. In the partially pleiotropy scenario, four QTNs control both traits, seven QTNs only control $Trait\ j$ and nine QTNs only control $Trait\ j'$. The final scenario, spurious pleiotropy, simulates two correlated traits where each is controlled by 10 different QTNs. For this genetic architecture, simplePHENOTYPES randomly selects ten markers; each of these markers is in a user-specified amount of maximum linkage disequilibrium with one QTN controlling $Trait\ j$, as well as another QTN controlling $Trait\ j'$.

We took several measures to ensure the quality, reproducibility, and accessibility of the results produced by simplePHENOTYPES. Upon the completion of simulating traits, simplePHENOTYPES will create a log file that will compare the estimated sample heritabilities and (when appropriate) trait correlations with those

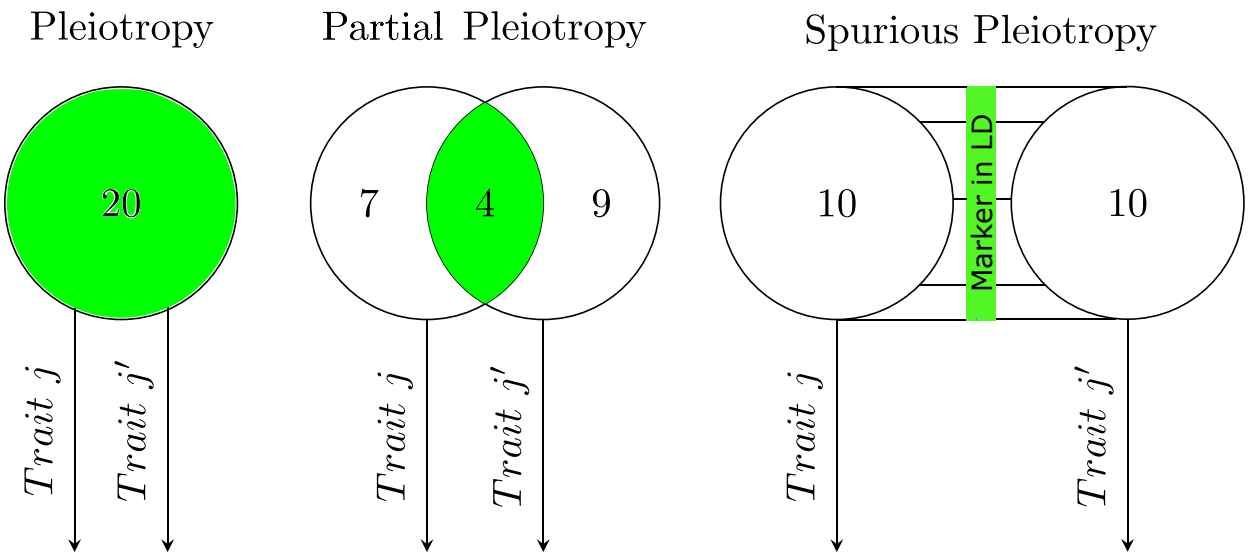

Figure 2: Multi-trait genetic architectures.

specified by the user, as well as confirm details on the specified genetic architecture. The seed numbers used to select QTNs and for simulating the random non-genetic error terms for each simulated trait are saved as text files. These, along with genomic information on the markers selected to be QTNs, should facilitate the regeneration of the simulated traits whenever the need arises. Finally, the simulated traits are saved in user-specified formats that are ready for downstream evaluation in external quantitative genetics software packages such as GEMMA [13] and TASSEL [1], as well as saved as an R object that can be read into R packages including GAPIT [7] and rrBLUP [4].

3 Performance assessment

We compared traits simulated from simplePHENOTYPES to SimPhe [6] and PhenotypeSimulator [8]. Despite the different methods used by the three packages, under some specific settings, namely the additive model and the dominance model for SimPhe, they produced equivalent results. When phenotypes with the same QTNs and effect sizes where simulated under these settings with a heritability of 1, the simulated phenotypes were identical (Supplementary File). Additionally, we simulated phenotypes with large-effect QTNs and ran multi-trait GWAS on GEMMA [13]. As expected, the simulated QTNs were identified as peak associations (Supplementary File, Figures 1, 2 and 3). Finally, when we simulated phenotypes from a data set of 36,392 markers genotyped on 280 maize lines, a subset of the 282 Maize Diversity Panel [3], on a single core with 70 GB of RAM, the median time across all settings was 11.9255 seconds to simulate 1000 replicate phenotypes (Supplementary File, Figure 4).

4 Conclusion

simplePHENOTYPES makes it possible to simulate multiple traits controlled by loci with varying degrees of pleiotropy. Its ability to interface with both commonly-used marker data formats and downstream quantitative genetics software and packages should facilitate the rigorous assessment of both existing and emerging statistical GWAS and GS approaches.

Acknowledgements

We acknowledge Candice Hirsch, Rafael Della Coletta and Brian Rice for their independent testing and suggestions that improved the usability of simplePHENOTYPES.

Funding: National Science Foundation, Plant Genome Research Project (Grant Number 1733606)

References

[1] P. J. Bradbury, Z. Zhang, D. E. Kroon, T. M. Casstevens, Y. Ramdoss, and E. S. Buckler. TASSEL: Software for association mapping of complex traits in diverse samples. *Bioinformatics*, 23(19):2633–2635, 2007. ISSN 13674803. doi: 10.1093/bioinformatics/btm308.

[2] L. Cabrera-Bosquet, J. Crossa, J. von Zitzewitz, M. D. Serret, and J. Luis Araus. High-throughput Phenotyping and Genomic Selection: The Frontiers of Crop Breeding Converge. *Journal of Integrative Plant Biology*, 54(5):312–320, 2012. ISSN 16729072. doi: 10.1111/j.1744-7909.2012.01116.x.

[3] J. P. Cook, M. D. McMullen, J. B. Holland, F. Tian, P. Bradbury, J. Ross-Ibarra, E. S. Buckler, and S. A. Flint-Garcia. Genetic architecture of maize kernel composition in the nested association mapping and inbred association panels. *Plant Physiology*, 158(2):824–834, 2012. ISSN 00320889. doi: 10.1104/pp.111.185033.

[4] J. B. Endelman. Ridge Regression and Other Kernels for Genomic Selection with R Package rrBLUP. *The Plant Genome Journal*, 4(3):250, 2011. ISSN 1940-3372. doi: 10.3835/plantgenome2011.08.0024.

[5] R. A. Fisher. The Genetical Theory of Natural Selection, 1930.

[6] B. Jiang and B. Pütz. SimPhe, 2018.

[7] A. E. Lipka, F. Tian, Q. Wang, J. Peiffer, M. Li, P. J. Bradbury, M. A. Gore, E. S. Buckler, and Z. Zhang. GAPIT: Genome association and prediction integrated tool. *Bioinformatics*, 28(18):2397–2399, 2012. ISSN 13674803. doi: 10.1093/bioinformatics/bts444.

[8] H. V. Meyer and E. Birney. Phenotype Simulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics*, 34(17):2951–2956, 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty197.

[9] H. F. Porter and P. F. O'Reilly. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Scientific Reports*, 7(October 2016):1–12, 2017. ISSN 20452322. doi: 10.1038/srep38837. URL `http://dx.doi.org/10.1038/srep38837`.

[10] D. Singh, X. Wang, U. Kumar, L. Gao, M. Noor, M. Imtiaz, R. P. Singh, and J. Poland. High-throughput phenotyping enabled genetic dissection of crop lodging in wheat. *Frontiers in Plant Science*, 10(April): 1–11, 2019. ISSN 1664462X. doi: 10.3389/fpls.2019.00394.

[11] N. Solovieff, C. Cotsapas, P. H. Lee, S. M. Purcell, and J. W. Smoller. Pleiotropy in complex traits: Challenges and strategies. *Nature Reviews Genetics*, 14(7):483–495, 2013. ISSN 14710056. doi: 10.1038/nrg3461.

[12] W. Yang, Z. Guo, C. Huang, L. Duan, G. Chen, N. Jiang, W. Fang, H. Feng, W. Xie, X. Lian, G. Wang, Q. Luo, Q. Zhang, Q. Liu, and L. Xiong. Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nature Communications*, 5:1–9, 2014. ISSN 20411723. doi: 10.1038/ncomms6087.

[13] X. Zhou and M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics*, 44(7):821–824, 2012. ISSN 10614036. doi: 10.1038/ng.2310.