Aiding the discovery of susceptibility variants for complex phenotypes:

Development of a simulation framework for the genetic-architecture-informed selection of appropriate algorithms

Serena Nicoll[1], Timothy Mitchell[3], Athena Hadjixenofontos[2,3]

[1]REU Computing for Structure Student, [2]Center for Computational Science, [3]University of Miami

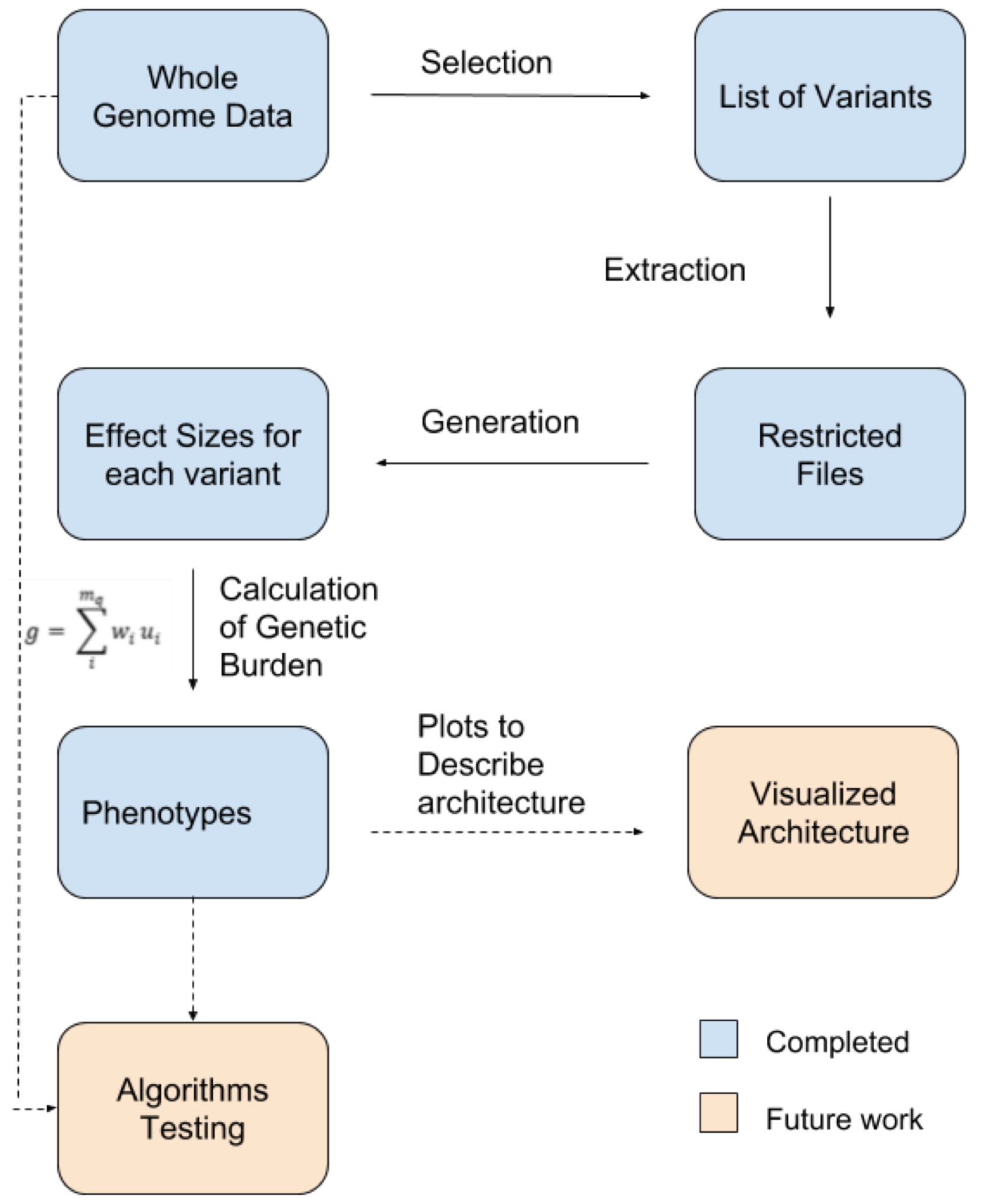
Introduction

Complex phenotypes, such as Autism Spectrum Disorders and multiple sclerosis, are the result of the polygenic action of hundreds to thousands of genetic variants and environmental factors. The presence of each susceptibility variant contributes additively to the risk of manifesting the phenotype. Identifying complete sets of variants that contribute to susceptibility for a complex phenotype has proven difficult. Researchers point to different lines of evidence to justify the reasons why many of the effects remain unaccounted for, including inflated heritability, underpowered sample sizes, and a majority of small effects. Recently the omnigenic model was put forward, suggesting a continuum of contribution to susceptibility by all variants present in an individual. This also predicts an inner circle of relevant variants that are most closely associated with the phenotype.

We propose that the models and algorithms used so far in genetic epidemiology (primarily mixed-effects regression models, and random forests) are not well suited to the fundamental characteristics of the genetic architectures of complex phenotypes. Genetic architecture is defined as the number of variants underlying susceptibility, the distribution of their allele frequencies, and the distribution of their effect sizes. We seek to characterize the relationship between features of a collection of algorithms and parameters of genetic architecture, given a particular phenotype. In the present study, we began to develop a simulation framework which will form the basis for this characterization as a systematic endeavor.

Methods

- We have built a series of scripts to select variants according to a user-defined genetic architecture.
- Starting with whole genome sequencing data from the 1000 Genomes Project avoids unrealistic features of simulated genomes.
- Scripts can bias variant selection towards user-defined distributions of allele frequency.
- The scripts also calculate predicted phenotypes for each individual, based on their genetic load for the simulated susceptibility variants.



Results

Simulated effect sizes as a function of allele frequency

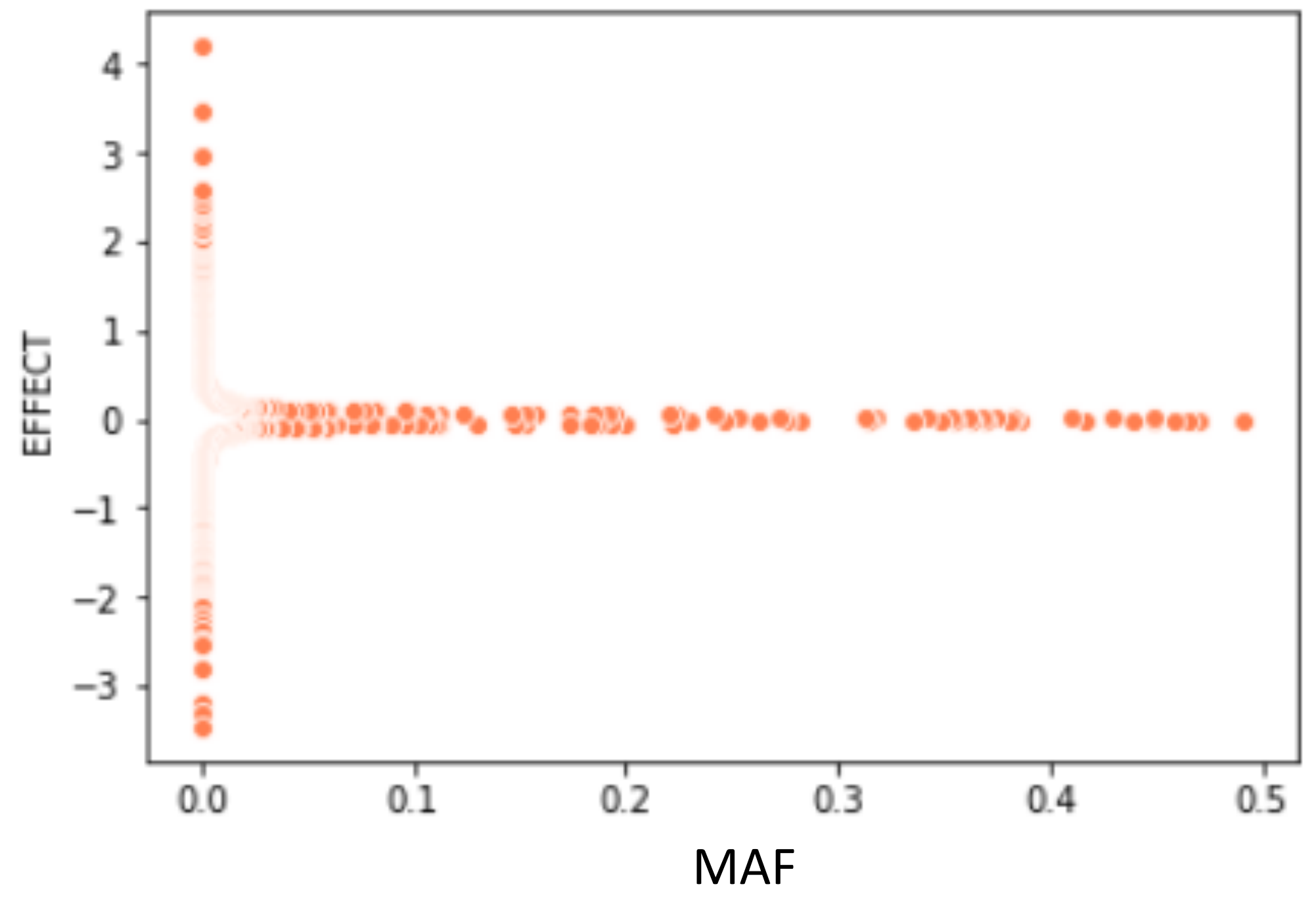


Figure 1: The effect sizes are generated from a standard normal distribution, with effects expressed as Cohen's d to account for positive risk effects and negative, protective effects.

Allele Frequency Distribution of selected risk variants

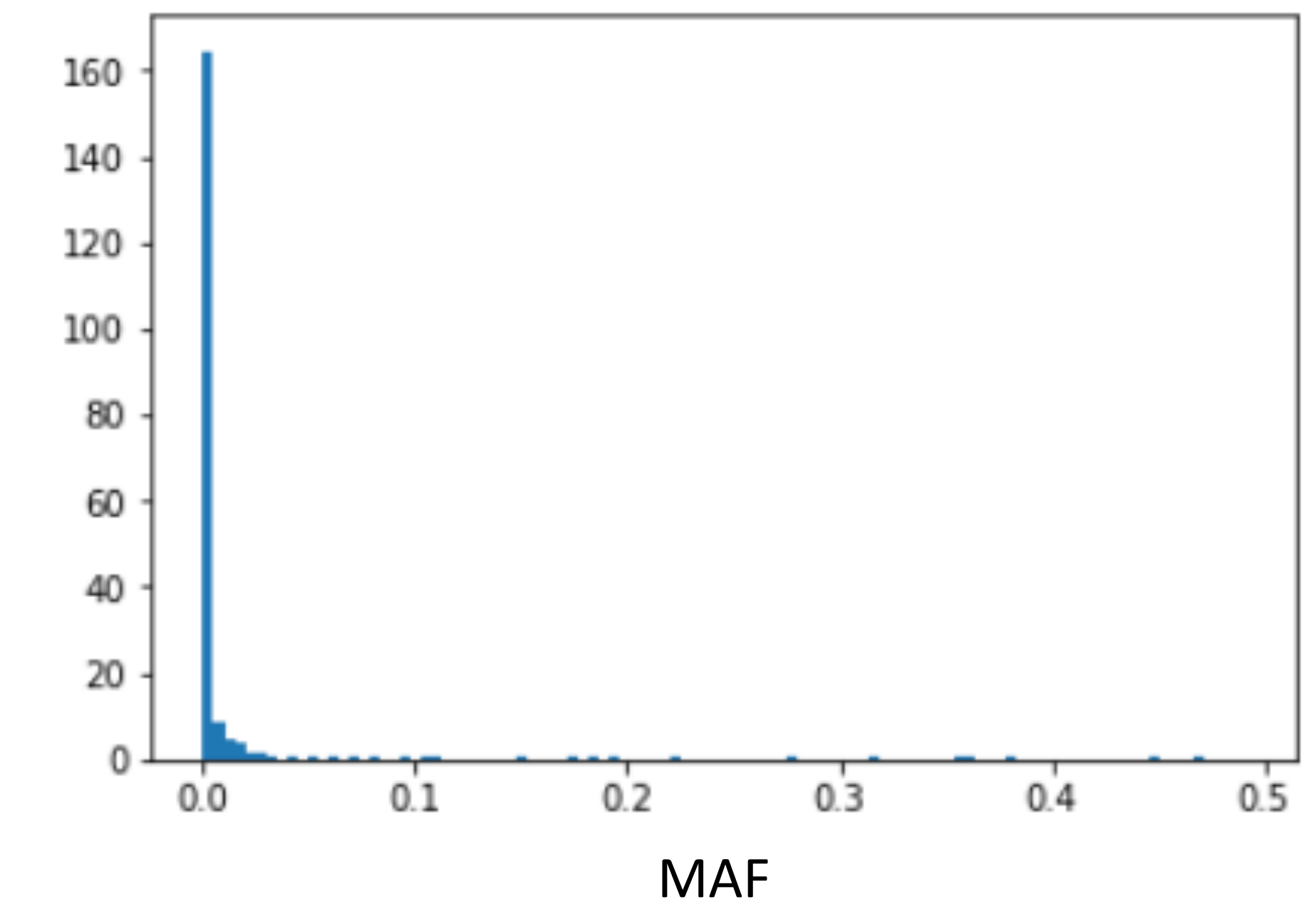
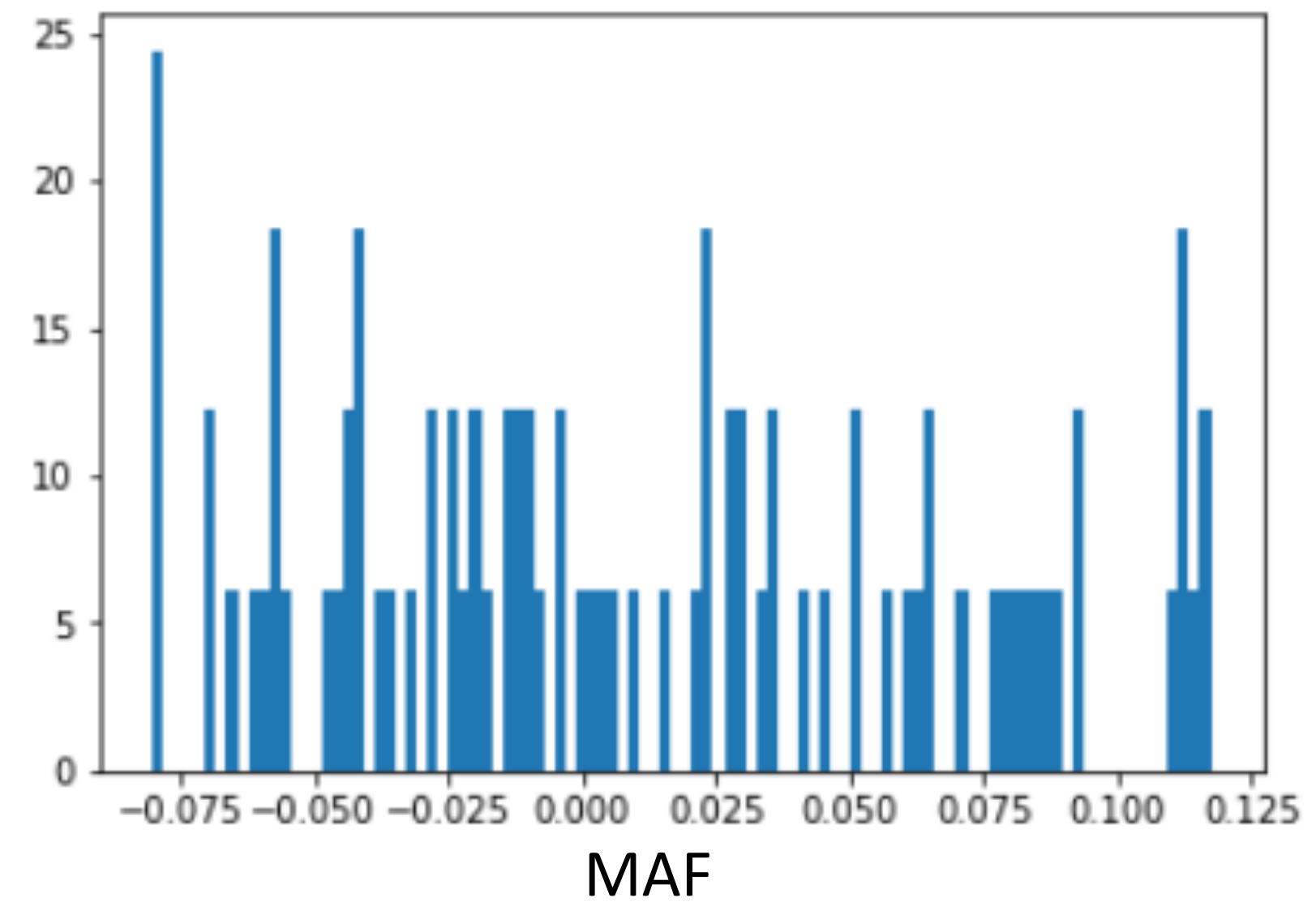


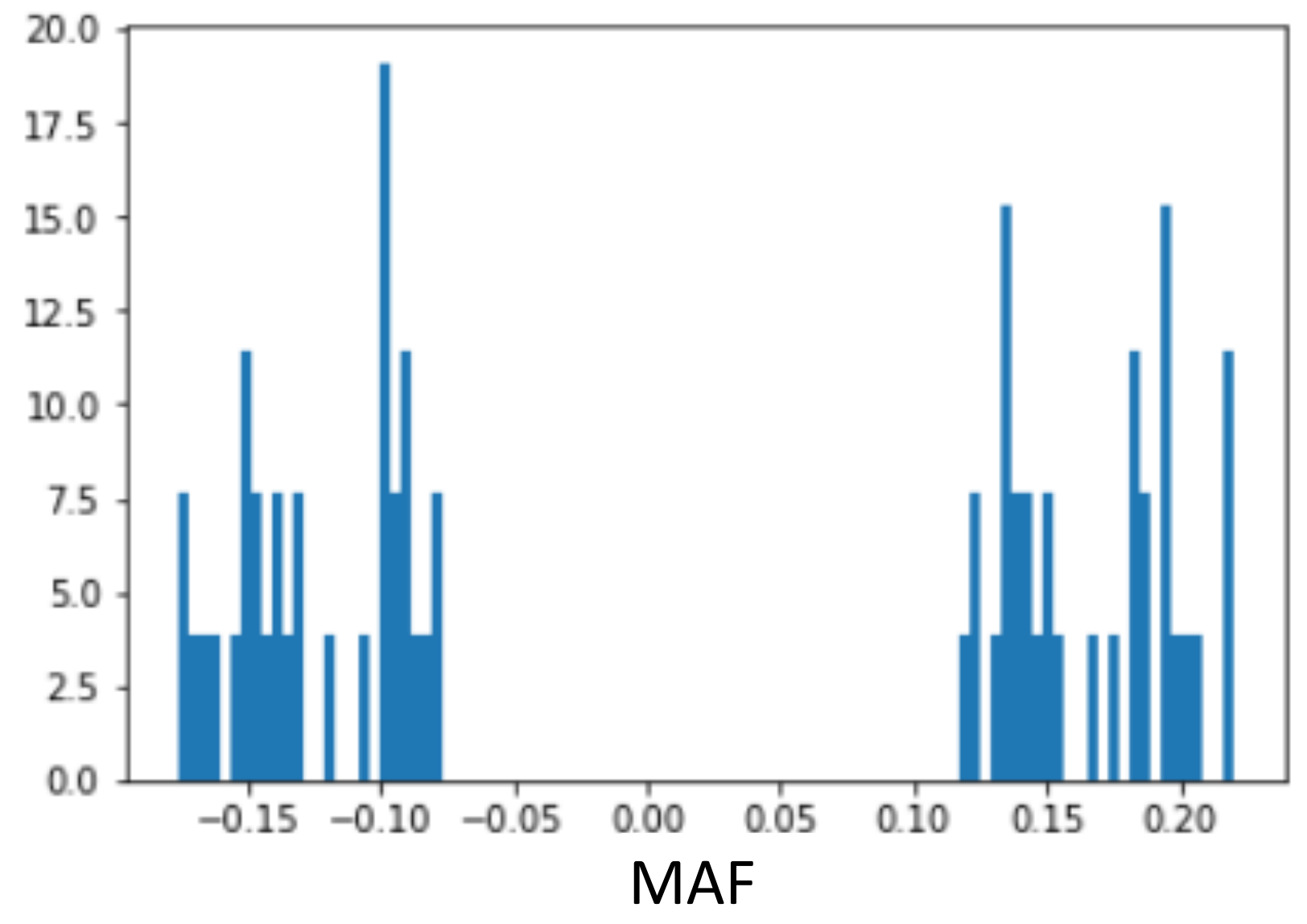
Figure 2: Allele frequencies for the 1000 variants match what is expected for a random sample. The majority of variation in the human genome is rare, recent variation that has not had time to be impacted by genetic drift and other population genetic forces.

As an example, we present the genetic architecture of a phenotype with 1000 risk variants, randomly distributed allele frequencies, and effect sizes that are inversely correlated with allele frequency, such that common variants have smaller effects, as expected based on population genetics models.

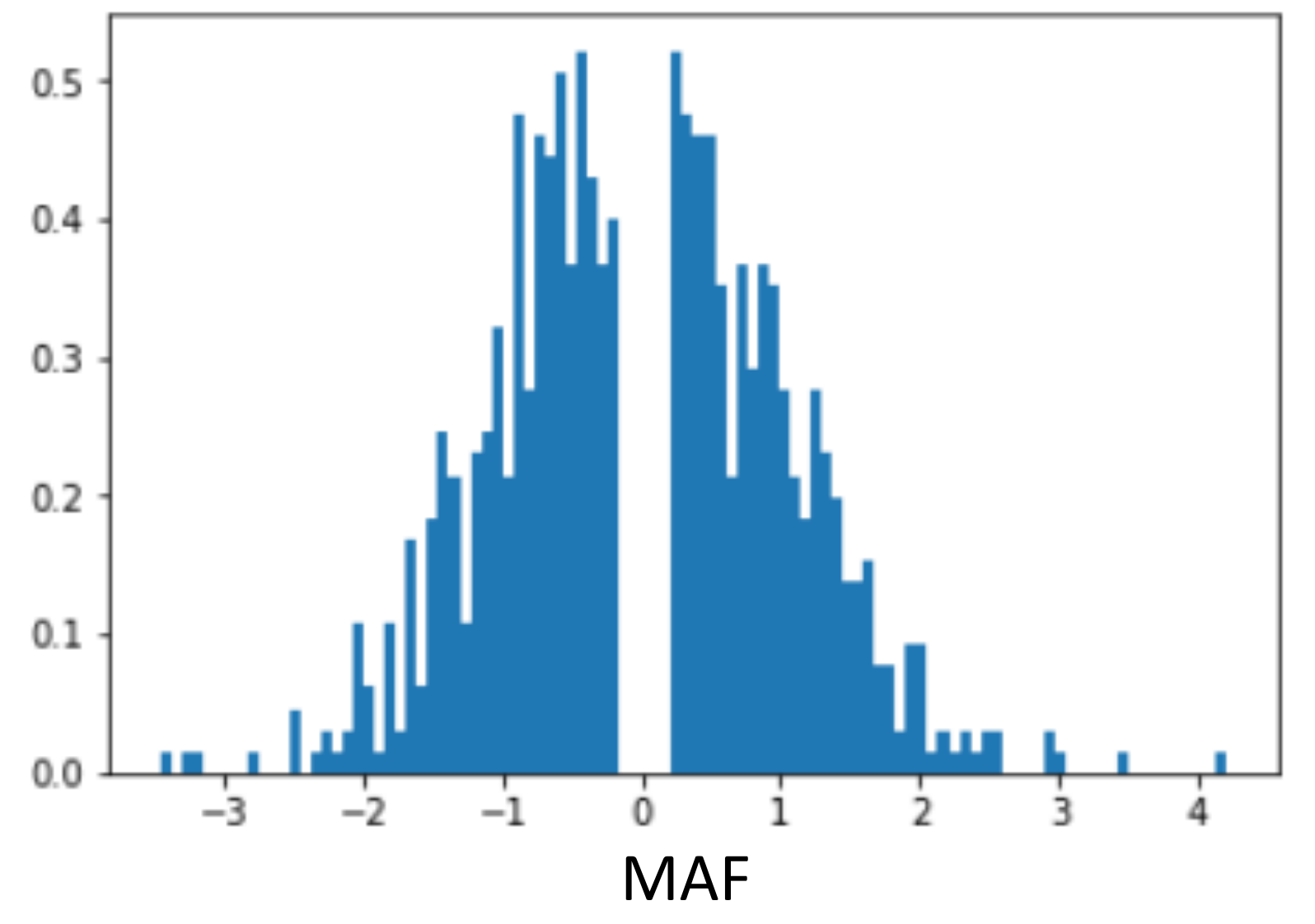
Effect Sizes for MAF < 0.01



Effect Sizes for 0.01 < MAF < 0.05



Effect Sizes for MAF > 0.05



Methods (Cont'd)

- Example genetic architecture scenarios include:
  - Scenario I: N randomly sampled variants
  - Scenario 2: N random variants + Q common variants
  - Scenario 3: N random variants + Q rare variants

Discussion

This motivating idea behind this project is the need to use algorithmic tools in an informed way based on the underlying biology and epidemiology. To this end, we launched on the creation of a tool to help evaluate algorithm performance given genetic architectures defined by specific parameters. Simulations are essential because they provide a way to evaluate the suitability of algorithms. Our simulation framework is based on real data, avoiding concerns due to unrepresentative linkage disequilibrium, allele frequencies, and other features of simulated genomes. In current implementation, it uses simplified models of effect sizes. The key next step in evaluating the performance of the developed simulation framework in assessing the suitability of WINNOW, a machine learning algorithm that uses multiplicative updates as a first algorithm.

Future Work

Future work includes:

- Implement additional options for effect size distributions to model risk contributions of selected variants (2).
- Including a bias towards specific gene ontologies suspected to be more relevant for a particular phenotype.
- Scale up number of simulations to account for variability in individual runs and create ways to visualize overall results.
- Evaluate consistency of the genetic architecture between populations.
- Apply rigorous analysis on algorithms based on whole genome and phenotypic data to determine optimal pairings of algorithms and architectures.

References

[1]: Yang, J., A. Bakshi, Z. Zhu, G. Hemani, A. A. Vinkhuyzen, S. H. Lee, M. R. Robinson, et al. 2015. "Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index." Nature genetics 47 (10): 1114-1120. doi:10.1038/ng.3390. http://dx.doi.org/10.1038/ng.3390.

[2]: Ju-Hyun,Mitchell H.,R.,J.Carroll,C.,,J.Chanock,F.,. 2011. "Distribution of allele frequencies and effect sizes and their interrelationships for common genetic susceptibility variants" Proceedings of the National Academy of Sciences(44)-18031;:.1073/pnas.1114759108

[3]: D. G. Haegert. 2004. "Analysis of the threshold liability model provides new understanding of causation in autoimmune diseases." *Med Hypotheses.*2004;63:257–61.

Acknowledgements

I would like to offer special thanks to:

- Athena Hadjixenofontos, PhD for her dedicated mentorship and support
- The Center for Computational Science, for their generous assistance
- The University of Miami Department for Computer Science, for hosting this REU
- Dr. Burt Rosenberg and all mentors of the CFS REU for their wisdom and teachings
- This material is based upon work supported by the National Science Foundation under Grant No. CNS-1659144