Varistat, a library for statistical analysis in population genetics

Dante Dam[1], Kevin Thornton[2] ([1]Department of Computer Science, [2]Department of Ecology and Evolutionary Biology)

I. Introduction

Modern genetic datasets are rapidly growing in number and scale, faster than our ability to analyze them effectively. A decade ago, the 1000 Genomes project characterized thousands of human genomes at millions of sites[1]. In 2021, the UK Biobank released 200,000 whole human genomes; whole genome sequencing is ongoing for another 300,000 individuals. Such studies have also been repeated on other taxa, including *E. coli*[2] and *Oryza sativa* (rice)[3].

The existing software infrastructure is still developing the ability to handle datasets of this scale. For example, Python-based solution `scikit-allel` is slow and unwieldy at these scales due to memory constraints. Many such tools are limited by over-reliance on particular input formats which are overspecialized to a particular workflow. We introduce a new flexible software library for the computation of summary population genetics statistics designed with these modern data scales in mind.

II. Big ideas for our solution

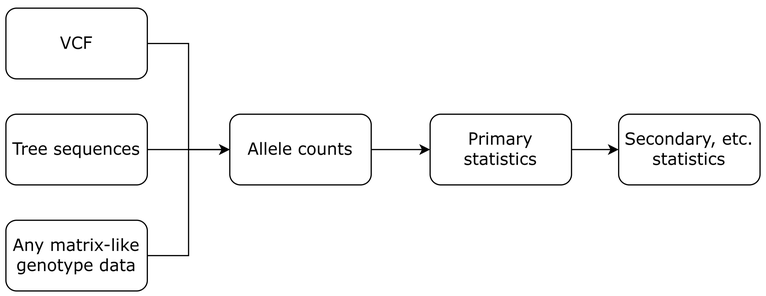

Figure 1: Our library's overall design, showing what data are required at each step.

i. Modularity

We divide the overall workflow into several independent steps; the first is input. Multiple formats are natively supported. First, the Variant Calling Format (VCF), a common representation of measured genotypic data. We also support input from tree sequences, an efficient data encoding based on ancestral relationships among sampled genomes. These often represent the result of *in silico* simulation. Any other data format which can be read as a collecton of sites and variants is also accepted. After computing allele counts from this data, statistics can be computed using said allele counts. These allele counts can also be easily accessed for any other processing task the end user wishes to implement.

ii. Rust

A unique feature of our tool is that it is written in Rust, a low-level language focusing on memory safety, correctness, and low- and zero-cost abstractions. This allows it to be used from C, C++, Python, R, and anywhere else C bindings are accepted, allowing it to follow bioinformaticians to their language of choice.

iii. Testing

To verify the correctness of our statistics, we create parallel naive implementations which adhere closely to the formal definition of a statistic. We then demonstrate that our optimized code is equivalent to the naive implementation.

III. Results

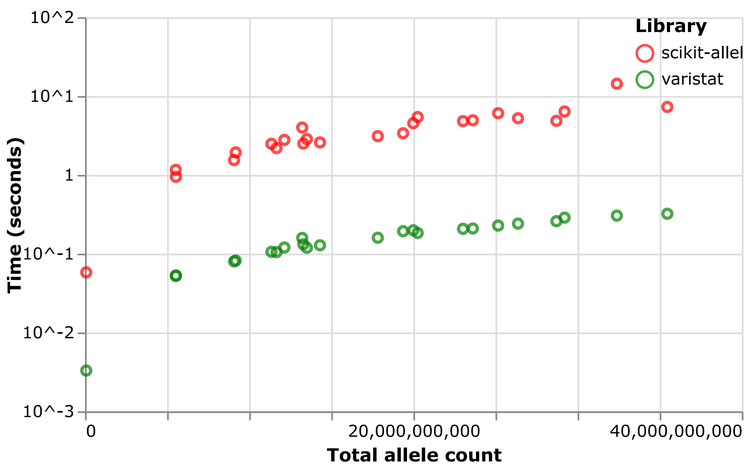

Figure 2: Our new solution computes expected heterozygosity about 10× faster (median of $n = 10$ runs) than `scikit-allel`. Note that the time axis is on a $\log_{10}$ scale. Test data is the 24 distinct human chromosomes[1].

Following these design principles, support for these statistics has been fully realized and tested:

- Expected heterozygosity/diversity, π or θ_π
- Watterson's[4] estimator θ, measuring mutation rate
- F_{ST}, measuring migration rate (by multiple definitions[5], [6], [7])
- Tajima's D[8], summarizing skew in allele frequencies

Our library is available at the GitHub link at bottom left.

IV. Next steps

Work is ongoing to add parallelism, unlocking additional performance gains over existing solutions, particularly those utilizing NumPy parallelism already. This requires support for indexed input formats and a multithreaded engine for statistic computation.

Statistics involving genome position have yet to be explored, e.g. those involving sliding genomic windows or normalization by sequence length. This library currently does not record site positions and relies on the user to select desired sites before calculating a statistic.

V. Acknowledgements

Thanks to the UROP Research Experience Fellowship for funding the production of this poster. Thanks to the School of ICS OpenLab HPC for providing resources for benchmarking of this library and its counterparts.

References

[1] A. Auton *et al.*, "A global reference for human genetic variation," *Nature*, vol. 526, no. 7571, pp. 68–74, Oct. 2015, doi: 10.1038/nature15393.

[2] D. G. Mets and M. Morin, "Creating a 7,000-strain E. coli genotype dataset with antimicrobial resistance phenotypes," *Arcadia Science*, vol. 1, Aug. 2024, [Online]. Available: https://research.arcadiascience.com/pub/dataset-ecoli-amr-genotype-phenotype

[3] J.-Y. Li, J. Wang, and R. S. Zeigler, "The 3,000 rice genomes project: new opportunities and challenges for future rice research," *Gigascience*, vol. 3, no. 1, p. 8, May 2014.

[4] G. Watterson, "On the number of segregating sites in genetical models without recombination," *Theoretical Population Biology*, vol. 7, no. 2, pp. 256–276, 1975, doi: https://doi.org/10.1016/0040-5809(75)90020-9.

[5] B. S. Weir and C. C. Cockerham, "ESTIMATING F-STATISTICS FOR THE ANALYSIS OF POPULATION STRUCTURE1," *Evolution*, vol. 38, no. 6, pp. 1358–1370, 1984, doi: 10.1111/j.1558-5646.1984.tb05657.x.

[6] M. Slatkin, "ISOLATION BY DISTANCE IN EQUILIBRIUM AND NON-EQUILIBRIUM POPULATIONS," *Evolution*, vol. 47, no. 1, pp. 264–279, 1993, doi: 10.1111/j.1558-5646.1993.tb01215.x.

[7] R. R. Hudson, D. D. Boos, and N. L. Kaplan, "A statistical test for detecting geographic subdivision.," *Molecular Biology and Evolution*, vol. 9, no. 1, pp. 138–151, 1992, doi: 10.1093/oxfordjournals.molbev.a040703.

[8] F. Tajima, "Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.," *Genetics*, vol. 123, no. 3, pp. 585–595, 1989, doi: 10.1093/genetics/123.3.585.