SimPrily:
A Python framework to simplify genome simulation with priors

Ariella L. Gladstein[1], Consuelo D. Quinto-Cortés[2], Julian L. Pistorius[3], David Christy[4], Logan Gantner[5], August E. Woerner[6], and Blake L. Joyce[3,7]

[1]Department of Ecology and Evolutionary Biology, University of Arizona, USA, [2]National Laboratory of Genomics for Biodiversity (LANGEBIO), CINVESTAV, Mexico, [3]CyVerse, University of Arizona, USA, [4] Department of Computer Science, University of Arizona, USA, [5]Graduate Interdisciplinary Program in Applied Mathematics, University of Arizona, USA, [6] Center for Human Identification, University of North Texas Health Science Center, USA, [7]BIO5 Institute, University of Arizona, USA.

Availability: Source code, HTC workflow, documentation, and examples are available at https://github.com/agladstein/SimPrily

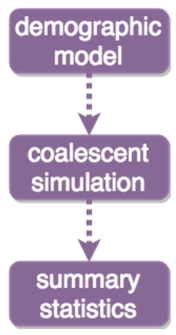
Introduction

What can you use 1000's – millions of simulations for?

- Approximate Bayesian Computation to infer demographic history
- Null demographic model to find regions under selection
- Truth datasets for testing software

Features

- Specify demographic model with priors
- Create pseudo array from simulations
- Calculate population genetics statistics
- Run 1000's of simulations with GUI in CyVerse Discovery Environment
- Run millions of simulations with Pegasus workflow on the Open Science Grid



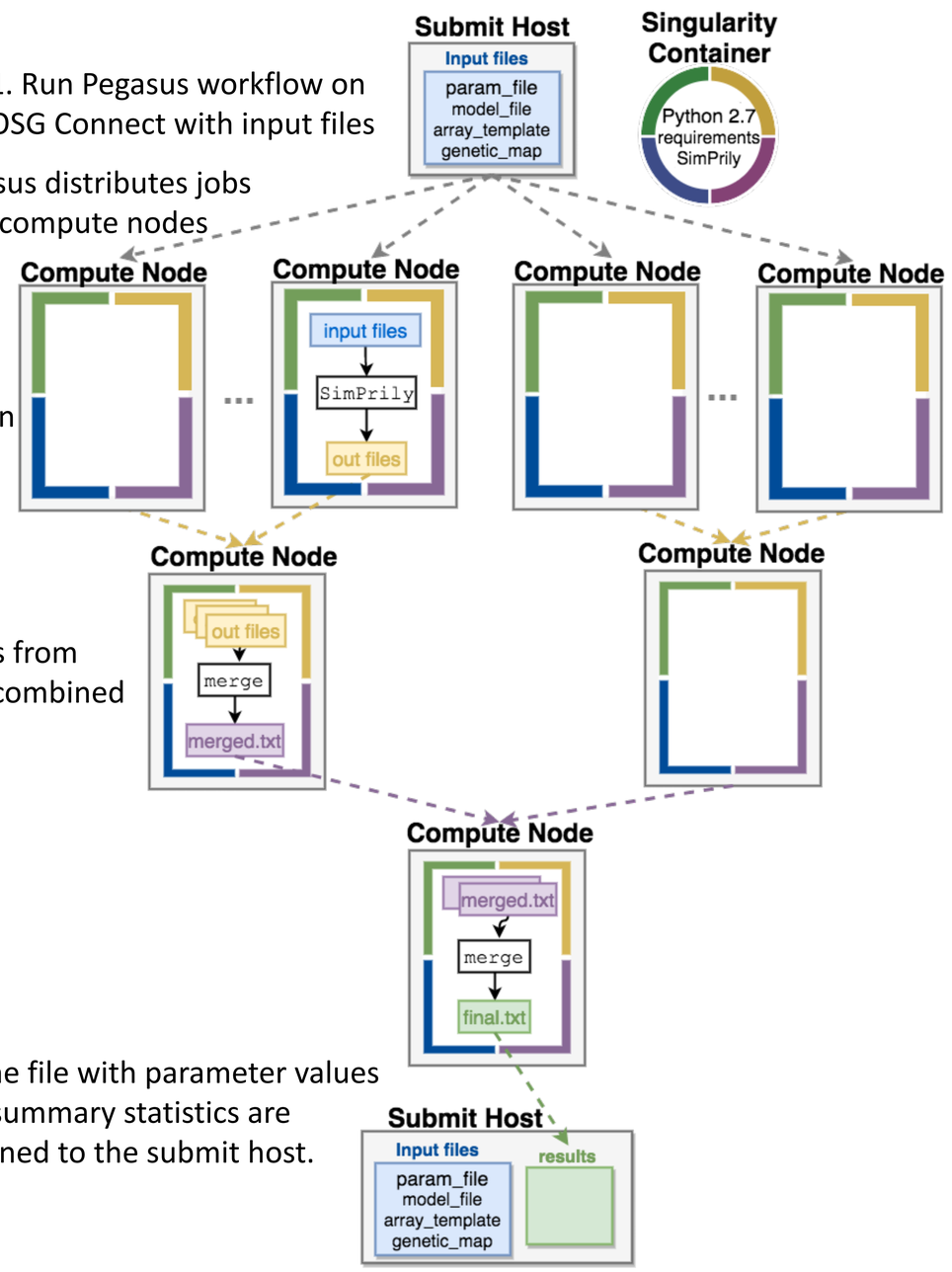
Methods

How to submit jobs to the Open Science Grid:

```
./submit param_file.txt model_file.csv array_template genetic_map number_jobs
```

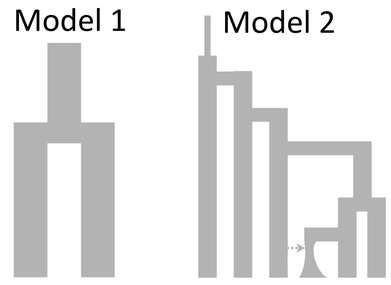
High throughput workflow

1. Run Pegasus workflow on OSG Connect with input files
2. Pegasus distributes jobs to OSG compute nodes
3. SimPrily runs in Singularity container on compute nodes
4. Results from jobs are combined
5. One file with parameter values and summary statistics are returned to the submit host.



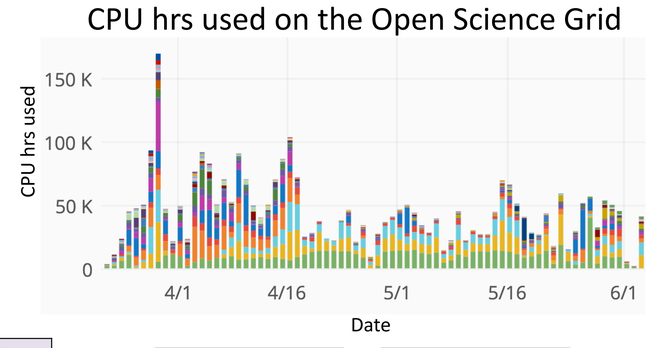
Results

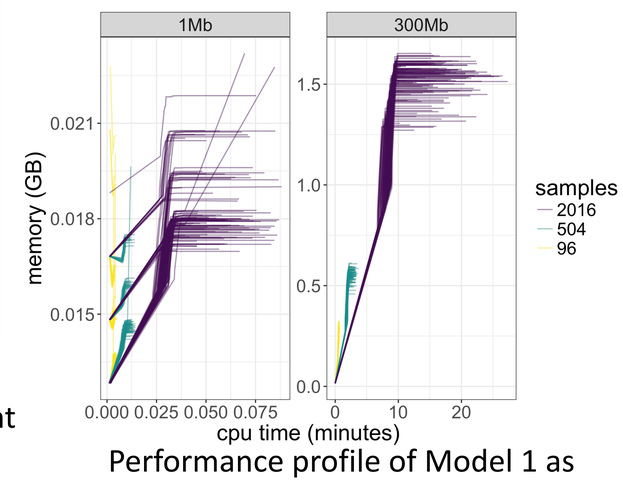
Performance



Locus Size	Sample Size	Model 1 Mem (GB)	Model 1 Time (min)	Model 2 Mem (GB)	Model 2 Time (min)
1 Mb	96	0.01	0.005	0.05	0.2
	504	0.01	0.01	0.1	0.7
	2016	0.02	0.06	0.3	6.9
300 Mb	96	0.3	0.8	3.1	12.6
	504	0.5	2.9	5.2	46.9
	2016	1.5	17.4	13.0	337.8

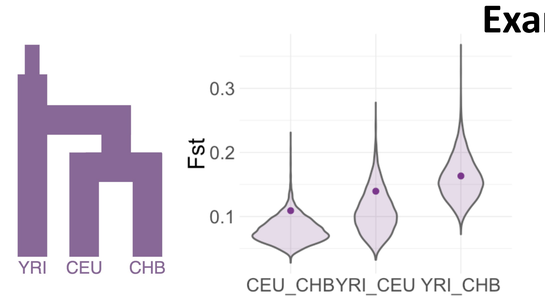
Performance of simple and more complicated model run with different sample and locus sizes.
mu = 2.5e-8, T = (0:5000), N = (1e3:1e6)



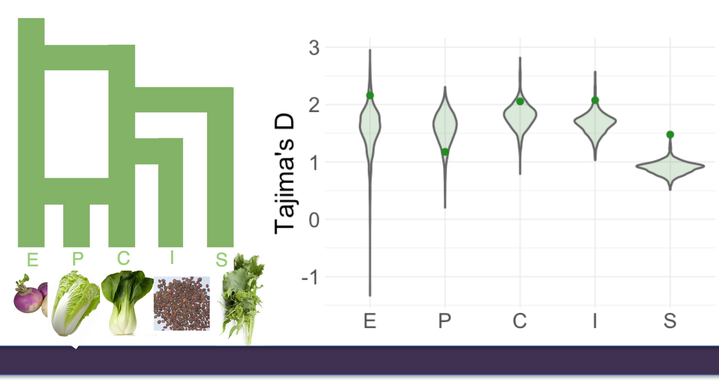


Performance profile of Model 1 as SimPrily runs. 100 simulations each.

Examples



Distribution of Fst values from 5000 simulations of chr1 of a model of Yoruba (YRI), Europeans (CEU), and Han Chinese (CHB). The dots indicate the Fst values from 1000 Genomes data.



Distribution of Tajima's D values from 1000 simulations of chrA01 of a B. *rapa* model. The dots indicate the Tajima's D values from real data.

Conclusions

Major Benefits

- No experience with HTC required - Ready to use HTC workflow
- Minimal storage required – simulations are not printed

Future Work

- Options of a variety of simulators
- Serial or parallel multilocus simulations

References

[1]Chen, G. et al. 2009. *Genome research,* 136-142.
[2]Deelman, E. et al. 2015. *Future Generation Computer Systems*, 46, 17-35.
[3]Merchant, N. et al. 2016. *Plos Biology*, 14(1), 1-9.
[4]Pordes, R. et al. 2007. *Journal of Physics: Conference Series*, 78, 012057.
[5]Qi, X. et al. 2017. *Molecular Ecology*, 26:3373-3388.
[6]Quinto-Cortés et al. 2017. Submitted
[7]The 1000 Genomes Project Consortium 2015. *Nature,* 526, 68-74.

Acknowledgments

We would like to thank Mats Rynge for his valuable help with setting up the Pegasus workflow and running it on the Open Science Grid.
This material is based upon work supported by the National Science Foundation under Award Numbers DBI-0735191 and DBI-1265383. URL: www.cyverse.org.Pegasus is funded by The National Science Foundation under OAC SI2-SSI program, grant \#1664162.This research was done using resources provided by the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

THE UNIVERSITY OF ARIZONA
Open Science Grid
CYVERSE™
Pegasus
docker