Joint estimation of DFE and time-varying mutation rates using paired data of allele frequency and allele age

Vivaswat Shastry[1] and Jeremy Berg[2]

[1]Committee on Genetics, Genomics & Systems Biology,
[2]Department of Human Genetics,
University of Chicago, Chicago IL
vivaswat@uchicago.edu

Introduction

- Typically, selection coefficients are estimated using allele frequency information from the **site frequency spectrum** (SFS)
- But, can we gain more information about selection & mutation parameters by **exploring the correlation structure between frequency and age** in the form of the **site frequency-age spectrum** (SFAS) i.e., incorporating more information about the trajectory of these alleles?

Methods

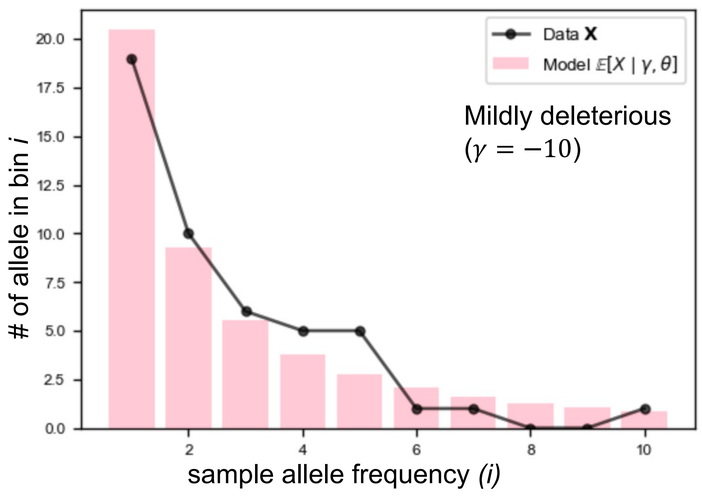
- We use **forward-in-time Wright-Fisher diffusion** to compute the expected SFS & SFAS (*moments* from Jouganous *et al*, 2017) under a specific set of parameter values
- Estimation of parameters is done via a **maximum-likelihood framework** using a Poisson likelihood on data simulated with *PReFerSim* (Ortega-Del Vecchyo *et al*, 2016)

Results

- We find a **modest shrinkage of 5-25% in variance** when estimating selection coefficients and a **huge improvement in accuracy of ~10x** when jointly estimating selection coefficients and time-varying mutation rates using paired data of allele frequency & age versus frequency alone

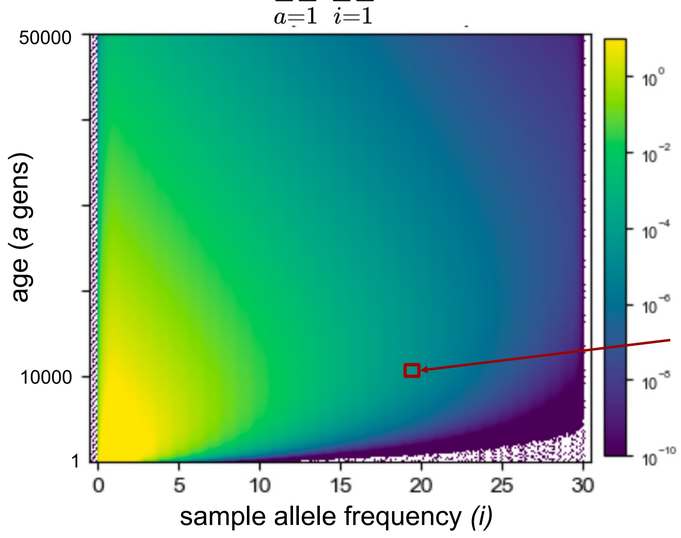
Frequency only approach using SFS

$$\mathcal{L}(\gamma;\, \mathbf{X}) = \prod_{i=1}^{2n-1} \text{Pois}(\mathbb{E}[X_i \mid \gamma, \theta])$$

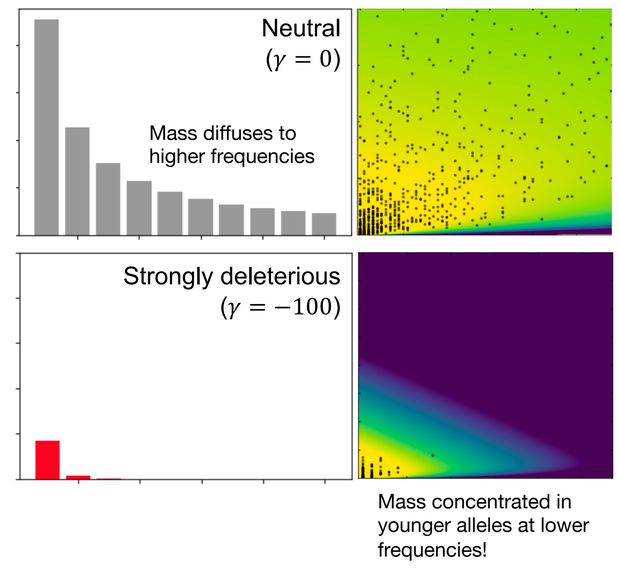


Frequency & age approach using SFAS

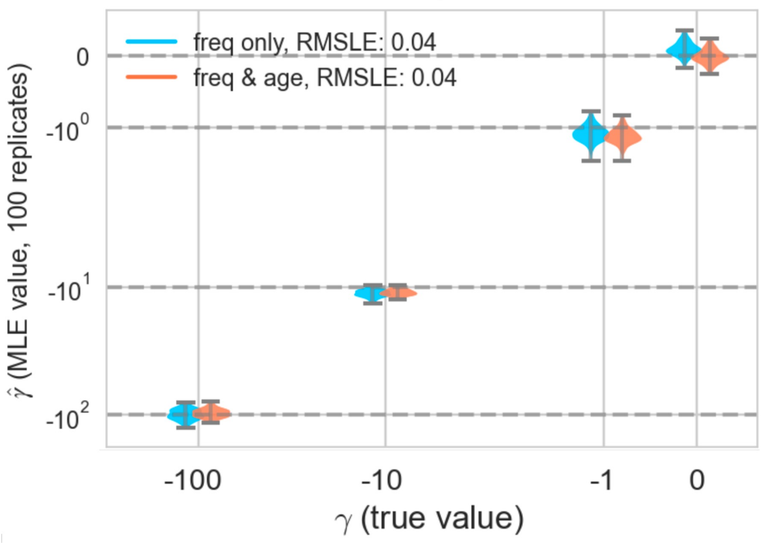
$$\mathcal{L}(\gamma;\, \mathbf{X}) = \prod_{a=1}^{A} \prod_{i=1}^{2n-1} \text{Pois}(\mathbb{E}[X_{i,a} \mid \gamma, \theta])$$



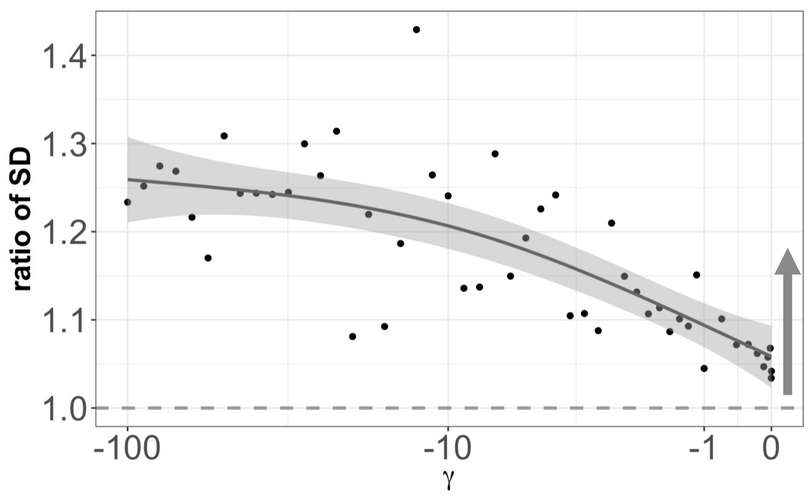
Shape of SFS & SFAS under varying strengths of selection



Estimates similarly unbiased…

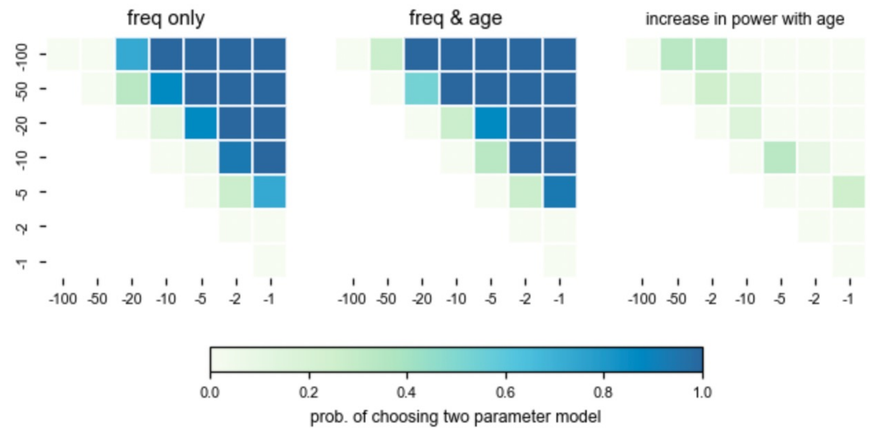


…but adding ages gives us (slightly) lower variance!

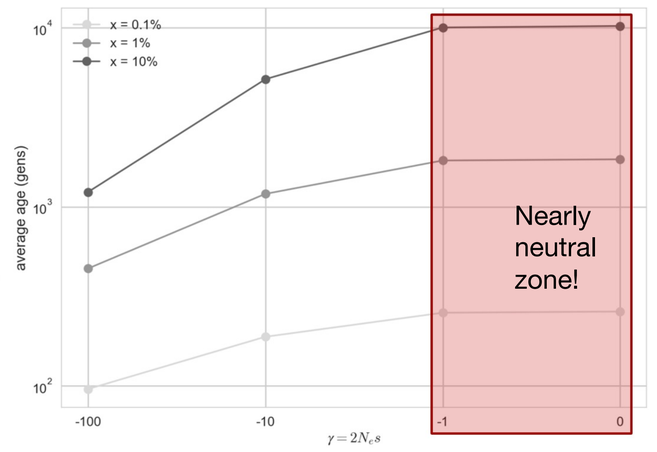


Estimates using ages are **tighter** around MLE than using frequency only

Higher discriminability in data sets containing two similar selection coefficients



Results makes sense given that average age (conditional on frequency) decreases quickly with selection greater than -1 (**selection-drift boundary**, Maruyama 1974)



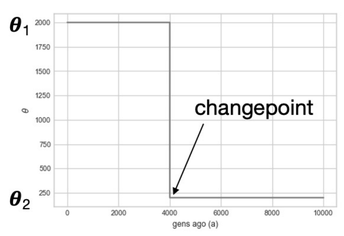
Glossary:
N_e effective population size
$\gamma = 2N_e s$ population-scaled selection coefficient
$\theta = 4N_e \mu$ population-scaled mutation rate

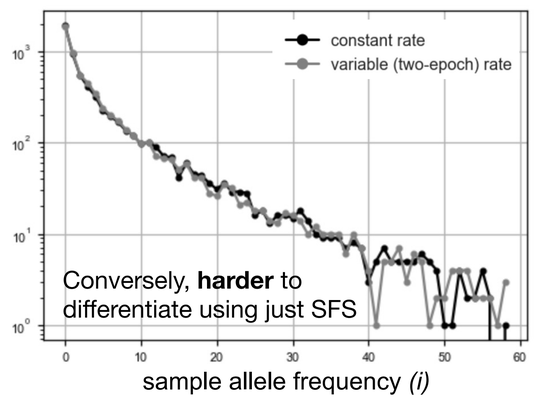
Assumptions:
1. **Allele ages estimated without error**
2. Point DFE (distribution of fitness effects)
3. Constant population size
4. Unlinked loci

Joint estimation of selection coefficient & time-varying mutation rate

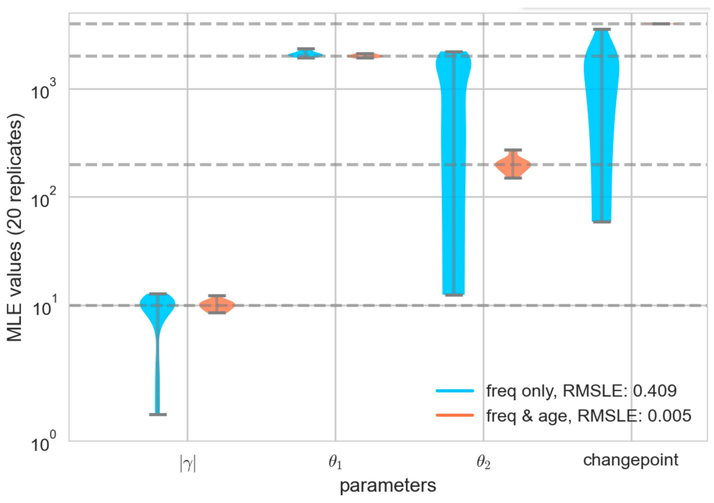
$$\mathcal{L}(\gamma, \theta;\, \mathbf{X}) = \prod_{a=1}^{A} \prod_{i=1}^{2n-1} \text{Pois}(\mathbb{E}[X_{i,a} \mid \gamma, \theta])$$

Simple two-epoch mutation rate

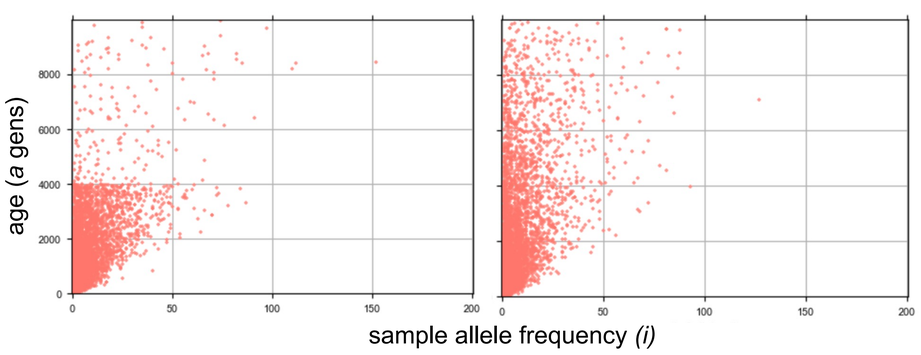




Huge improvement on adding ages



Which SFAS comes from a two-epoch history?



Future directions

1. Incorporate importance sampling scheme to account for age estimation being performed under neutral prior (using ARG-based methods)
   - Selected alleles tend to be **younger** than their neutral counterparts
2. Apply joint estimation procedure to families of transposable elements (TE) in maize (Stitzer *et al,* 2021) to estimate time-varying rate histories

References

- Maruyama, T. (1974). The age of an allele in a finite population. *Genetics Research*, *23*(2), 137-143.
- Jouganous, J., Long, W., Ragsdale, A. P., & Gravel, S. (2017). Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, *206*(3), 1549-1567.
- Vecchyo, O. D., Marsden, C. D., & Lohmueller, K. E. (2016). PReFerSim: fast simulation of demography and selection under the Poisson Random Field model. *Bioinformatics*, *32*(22), 3516-3518.
- Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *PLoS genetics*, *17*(10), e1009768.