Demographic inference using the SFS with Moments and Demes

Nick Collier[1] and Aaron P. Ragsdale[1,*]

[1]Department of Integrative Biology, University of Wisconsin–Madison
[*]apragsdale@wisc.edu

March 18, 2025

Abstract

Placeholder

1 Introduction

The genetic composition within a sample of individuals is shaped by their genome biology and evolutionary history. Variation resulting from this history is fully described by the ancestral relationships among samples at each locus in the genome and how they change along a chromosome due to recombination (that is, information stored in the Ancestral Recombination Graph [cites]). However, the ARG can be large and unwieldy, and methods for reconstructing history directly from the ARG, while showing promise [cite], are in their infancy and so far limited in application and scalability [cite]. Instead, evolutionary inference using informative summaries provides a tractable and powerful alternative for learning parameters of population history, natural selection and genome biology.

One such summary that has seen wide use is the site frequency spectrum (SFS), which stores the counts (observed or expected) of alleles carried by a given number of genomes in a set of samples. Like any summary of the data, the SFS discards information relative to the ARG – in this case, loci are treated independently so that haplotypic information is lost. Nonetheless, a great deal can be learned from the

What moments is, what type of data it uses, general framework for inference

1. The SFS – how its shape is sensitive to population, selection and genome biological parameters

2. Demographic inference from the SFS

3. Extensions: inference of selection, mutation rates, etc

4. Extensions: Multi-allele and multi-locus models

2 Interface with Demes

1. Manipulating and plotting Demes-specified demographic models

2. Computing statistics given a Demes model

3. Inferring history on a demes model, including specifying parameters to be fit

3   Examples

Two examples, one using simulated data with known parameters, which we try to reinfer. And another with actual data (humans + neanderthal). We briefly describe these, highlighting the high-level components of the fits. For detailed information for each, including managing data, specifying models, performing inference, and visualizing results, we refer readers to the GitHub (), which is maintained with up-to-date versioning.

3.1   Inferring parameters in a simulated isolation-with-migration model

A python snippet:

```
dbg = msprime.DemographyDebugger(
  population_configurations=population_configurations,
  migration_matrix=migration_matrix,
  demographic_events=demographic_events)
dbg.print_history()
ts = msprime.simulate(
  population_configurations=population_configurations,
  migration_matrix=migration_matrix,
  demographic_events=demographic_events)
```

or

```
dbg = demography.debug()
dbg.print_history()
ts = msprime.simulate(demography=demography)
```

This was in discussion of Kelleher et al. (2016).

3.2   Inferring human-Neanderthal demographic parameters

```
def myFunc():
    pass
```

4   Considerations and caveats

1. Incluce here general ideas about the strengths and weaknesses of various approaches in population genetic inference, including when using the SFS.

2. Challenges in finding local/global optima.

3. Challenges in exploring parameter space.

4. Gene dense vs gene sparse genomic architectures among species.

References

Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.