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 25, 2025

Abstract

Placeholder

Introduction

The genetic composition of a sample of individuals is shaped by their genome biology and evolutionary history. Variation resulting from this history can be fully represented by the ancestral relationships among samples at each locus in the genome and how those gene-genealogical relationships change along a chromosome due to recombination (that is, information stored in the Ancestral Recombination Graph (Nielsen et al., 2025)). However, the ARG can be large and unwieldy, and methods for reconstructing history directly from the ARG, while showing promise (e.g., YC Brandt et al., 2022; Fan et al., 2023; Brandt et al., 2024), are in their infancy and so far limited in application and scalability. Instead, evolutionary inference using informative summaries of genetic variation remains 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 some information stored in the ARG – in this case, loci are treated independently so that haplotypic information is lost. Even so, the relative densities of allele frequencies and the overall scale of the SFS are both sensitive to demographic and non-demographic processes.

This chapter focuses on multi-population demographic inference from the SFS using `moments` (Jouganous et al., 2017). Past population processes, including changes in population size, splits, gene flow and population structure, impact the SFS in predictable ways. A population size expansion (or contraction) leads to a relative excess (or deficit) of low-frequency variants, for example, and commonly used statistics measuring population divergence, such as F_{ST}, are themselves summaries of the SFS. Thus, the development of methods to learn demographic history from the SFS came soon after the publication of whole-genome sequencing from multiple individuals (Marth et al., 2004; Williamson et al., 2005). More recent methodological advances allow for increased sample sizes and numbers of populations, providing a rich ecosystem of software for SFS-based demographic inference (Gutenkunst et al., 2009; Excoffier and Foll, 2011; Gravel et al., 2011; Jouganous et al., 2017; Ragsdale et al., 2018; Kamm et al., 2020; Dilber and Terhorst, 2024). We return to this in the final section: Considerations and Caveats.

Other evolutionary mechanisms are known to affect allele frequency dynamics and the SFS. This presents a challenge for demographic inference, as the observed SFS may be distorted by non-demographic processes. However, this also presents an opportunity to learn about different evolutionary processes, if demographic history can be controlled for. The SFS has been used to infer the distribution of fitness effects of new mutations, primarily for nonsynonymous variation in coding regions (e.g., Eyre-Walker et al., 2006; Boyko et al., 2008; Kim et al., 2017); to scan for historical selective sweeps (e.g., Kim and Stephan, 2002; Nielsen et al.,

2005); and to understand how selection on quantitative traits impacts the genetic architecture underlying those traits (eg., Patel et al., 2024; Ragsdale, 2024). The SFS has also been used to infer relative mutation rates and how they have changed over time (DeWitt et al., 2021). Each of these analyses typically requires partitioning the data in some way – by functional annotation, mutation context, local recombination rate, etc – so that comparisons can be made across SFS from different classes of mutations.

Interfacing `moments` with `demes`

Specifying demographic models requires defining populations (or "demes") and their relationships via splits and gene flow. `demes` provides a standardized and accessible format for defining demographic models (Gower et al., 2022), and has been adopted by widely used software for population genetics simulation and inference (including `msprime` (Baumdicker et al., 2022), `momi` (Dilber and Terhorst, 2024), `fwdpy11` (Thornton, 2019), `GADMA` (Noskova et al., 2023)). `demes` encourages interoperability and reuse of code across software, and ease of model implementation and reduction of errors.

Demographic models are specified in YAML format (see Gower et al. (2022) and the associated documentation). The two-population split-with-migration model depicted in Figure 1A is specified as below:

```yaml
description: split-with-migration model, loosely based on example 2
generation_time: 29
time_units: years
demes:
- name: ancestral
  epochs:
  - {start_size: 15000, end_time: 300000}
  - {start_size: 25000, end_time: 75000}
- name: popA
  ancestors: [ancestral]
  epochs:
  - {start_size: 30000}
- name: popB
  ancestors: [ancestral]
  epochs:
  - {start_size: 1200, end_size: 14000}
migrations:
- demes: [popA, popB]
  rate: 5e-5
```

In `moments`, a `demes`-specified demographic model can be directly used to compute either the SFS or multi-population LD statistics (Ragsdale and Gravel, 2019, 2020). For the SFS, we simply need to load the demographic model using `demes` and specify the number of haploid samples to draw and from which populations.

```python
import demes, moments
g = demes.load('model.yaml')
samples = {'popA': 60, 'popB': 60}
fs = moments.Demes.SFS(g, samples=samples)
```

Here, we have not specified a mutation rate. The default behavior sets $u = 1$, so that $\theta = 4N_e u = 4N_e$, where N_e is the ancestral population's initial size. With a known mutation rate, the SFS will be properly scaled by passing the mutation rate as a keyword argument: `moments.Demes.SFS(g, samples=samples, u=u)`. This is equivalent to scaling the output SFS above by multiplying by `u`, as `u*fs`.

Examples: inference in `moments` using `demes`

Below, we show two examples of multi-population demographic inference using the joint SFS. The first example simulates data with known demographic parameters, which we try to reinfer using the both the original and misspecified simpler models. In the second example, we infer a relatively simple human-Neanderthal history using data from two human populations and a high-coverage ancient sample from the Neanderthal lineage. For both of these examples, we briefly describe data processing and inference setup, highlighting the high-level components of the analyses. For detailed information for each example, including managing data, specifying models, performing inference, computing confidence intervals and visualizing fits, we refer readers to the accompanying GitHub repository (`https://github.com/StatisticalPopulationGenomics-2ndEd/moments`), which is maintained with up-to-date versioning.

- to run inference, need to specify parameters to be fit
- the parameters/options file points to values in the demographic model to be fit, along with bounds on those parameters and any constraints among parameters
- this options file is also a YAML – see snippets of examples below, and full details are given in the above URL

Inferring parameters in a simulated split-with-migration model

Using the split-with-migration model defined above, we simulated data for 30 diploid individuals from both populations using `msprime` (Baumdicker et al., 2022). These simulations consided of 500 regions, each of length 1 Mb, with per-base recombination and mutation rates of 10^{-8}. The joint SFS was computed using `tskit`'s `ts.allele_frequency_spectrum()` function separately on each replicate region and summing across regions to find the total SFS across 500 Mb of data (Figure 1B–D).

Inferring human-Neanderthal demographic parameters

Considerations and caveats

1. Include 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. Background selection [Ewing, Johri]
5. Gene dense vs gene sparse genomic architectures among species.

References

Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E Castedo Ellerman, Jared G Galloway, et al. Efficient ancestry and mutation simulation with msprime 1.0. *Genetics*, 220(3):iyab229, 2022.

Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS genetics*, 4(5):e1000083, 2008.

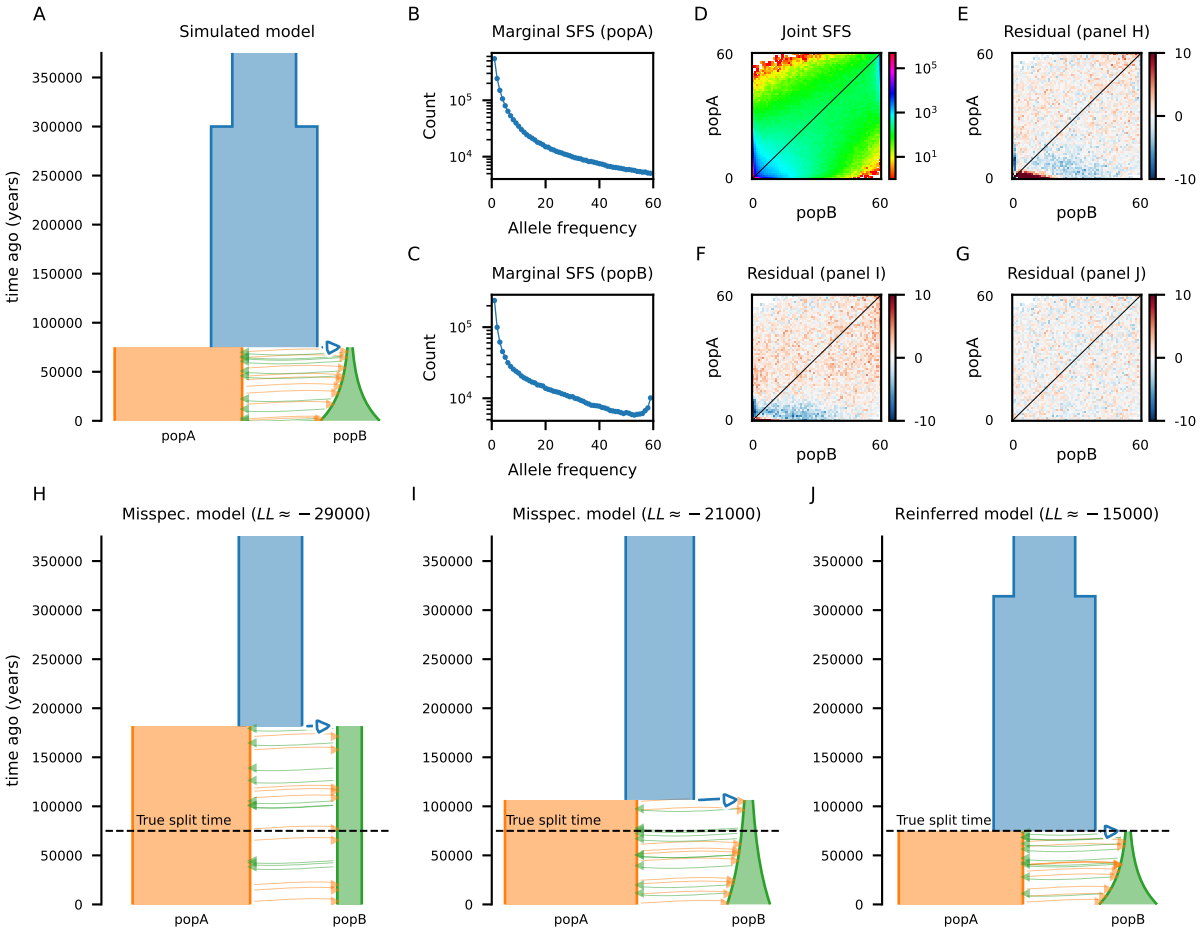

Figure 1: Caption placeholder

Figure 2: Caption placeholder

Débora YC Brandt, Christian D Huber, Charleston WK Chiang, and Diego Ortega-Del Vecchyo. The promise of inferring the past using the ancestral recombination graph. *Genome biology and evolution*, 16 (2):evae005, 2024.

William S DeWitt, Kameron Decker Harris, Aaron P Ragsdale, and Kelley Harris. Nonparametric coalescent inference of mutation spectrum history and demography. *Proceedings of the National Academy of Sciences*, 118(21):e2013798118, 2021.

Enes Dilber and Jonathan Terhorst. Faster inference of complex demographic models from large allele frequency spectra. *bioRxiv*, 2024.

Laurent Excoffier and Matthieu Foll. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, 2011.

Adam Eyre-Walker, Megan Woolfit, and Ted Phelps. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2):891–900, 2006.

Caoqi Fan, Jordan L Cahoon, Bryan L Dinh, Diego Ortega-Del Vecchyo, Christian Huber, Michael D Edge, Nicholas Mancuso, and Charleston WK Chiang. A likelihood-based framework for demographic inference from genealogical trees. *bioRxiv*, 2023.

Graham Gower, Aaron P Ragsdale, Gertjan Bisschop, Ryan N Gutenkunst, Matthew Hartfield, Ekaterina Noskova, Stephan Schiffels, Travis J Struck, Jerome Kelleher, and Kevin R Thornton. Demes: a standard format for demographic models. *Genetics*, 222(3):iyac131, 2022.

Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, 1000 Genomes Project, Carlos D Bustamante, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29): 11983–11988, 2011.

Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, 5 (10):e1000695, 2009.

Julien Jouganous, Will Long, Aaron P Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: beyond the diffusion approximation. *Genetics*, 206(3):1549–1567, 2017.

Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115 (531):1472–1487, 2020.

Bernard Y Kim, Christian D Huber, and Kirk E Lohmueller. Inference of the distribution of selection coefficients for new nonsynonymous mutations using large samples. *Genetics*, 206(1):345–361, 2017.

Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *Genetics*, 160(2):765–777, 2002.

Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using snp data. *Genome research*, 15(11):1566–1575, 2005.

Rasmus Nielsen, Andrew H Vaughn, and Yun Deng. Inference and applications of ancestral recombination graphs. *Nature Reviews Genetics*, 26(1):47–58, 2025.

Ekaterina Noskova, Nikita Abramov, Stanislav Iliutkin, Anton Sidorin, Pavel Dobrynin, and Vladimir I Ulyantsev. Gadma2: more efficient and flexible demographic inference from genetic data. *GigaScience*, 12:giad059, 2023.

Roshni A Patel, Clemens L Weiß, Huisheng Zhu, Hakhamanesh Mostafavi, Yuval B Simons, Jeffrey P Spence, and Jonathan K Pritchard. Conditional frequency spectra as a tool for studying selection on complex traits in biobanks. *bioRxiv*, 2024.

Aaron P Ragsdale. Archaic introgression and the distribution of shared variation under stabilizing selection. *bioRxiv*, pages 2024–08, 2024.

Aaron P Ragsdale and Simon Gravel. Models of archaic admixture and recent history from two-locus statistics. *PLoS genetics*, 15(6):e1008204, 2019.

Aaron P Ragsdale and Simon Gravel. Unbiased estimation of linkage disequilibrium from unphased data. *Molecular Biology and Evolution*, 37(3):923–932, 2020.

Aaron P Ragsdale, Claudia Moreau, and Simon Gravel. Genomic inference using diffusion models and the allele frequency spectrum. *Current Opinion in Genetics & Development*, 53:140–147, 2018.

Kevin R Thornton. Polygenic adaptation to an environmental shift: temporal dynamics of variation under gaussian stabilizing selection and additive effects on a single trait. *Genetics*, 213(4):1513–1530, 2019.

Scott H Williamson, Ryan Hernandez, Adi Fledel-Alon, Lan Zhu, Rasmus Nielsen, and Carlos D Bustamante. Simultaneous inference of selection and population growth from patterns of variation in the human genome. *Proceedings of the National Academy of Sciences*, 102(22):7882–7887, 2005.

Débora YC Brandt, Xinzhu Wei, Yun Deng, Andrew H Vaughn, and Rasmus Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221(1):iyac044, 2022.