Joint nonparametric coalescent inference of mutation spectrum history and demography

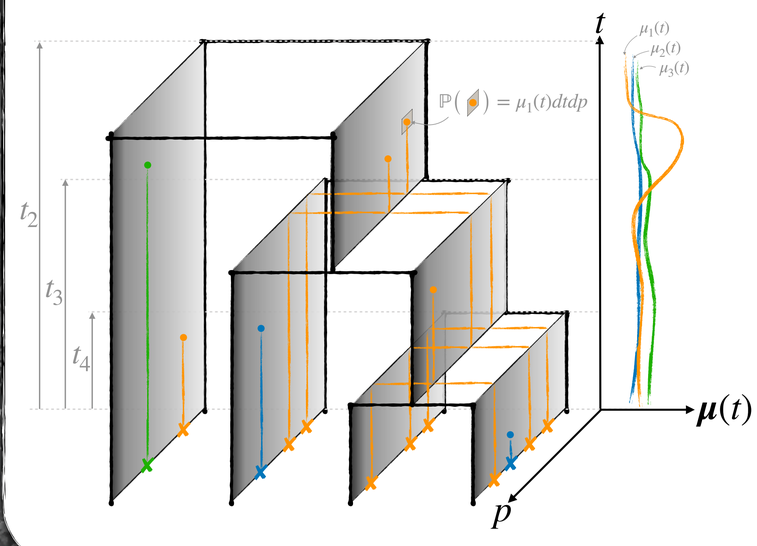
William S. DeWitt[1,4], Kameron Decker Harris[2,3], Kelley Harris[1,4]

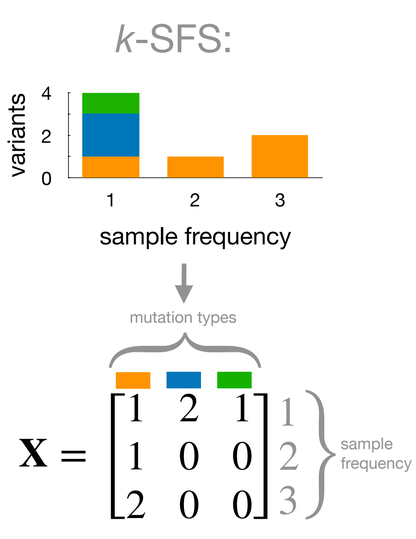
[1]Department of Genome Sciences, [2]Paul G. Allen School of Computer Science & Engineering, and
[3]Department of Biology, University of Washington, Seattle, WA
[4]Computational Biology Program, Fred Hutchinson Cancer Research Center, Seattle, WA

MUSHI: [MU]tation [S]pectrum [H]istory [I]nference

Models in evolutionary genetics typically assume that mutation rate is constant over time and between populations and closely related species. However, recent work casts doubt on this assumption in human and ape populations, and reveals that mutation is a complex and dynamic process. Whether arising from variation in replication fidelity, life history, or environmental exposures, mutation rate evolution can be accompanied by changes to the *mutation spectrum*: the mutation rate in different local nucleotide contexts. We extend theoretical tools based on Kingman's coalescent to accommodate a richly parameterized mutation process, varying in time and in spectrum. We infer human mutation spectrum histories from patterns of modern genomic diversity, allowing us to reconstruct trajectories of mutation spectrum divergence between populations, and track a transient mutation spectrum perturbation through multiple populations. We develop a tool that can perform fast, nonparametric joint inference of demographic and mutation spectrum histories from patterns of genomic diversity encoded in nucleotide context-specific sample frequency spectra, making it possible to jointly infer the contributions of two evolutionary processes that shape genetic diversity.

The k-SFS encodes mutation spectrum history





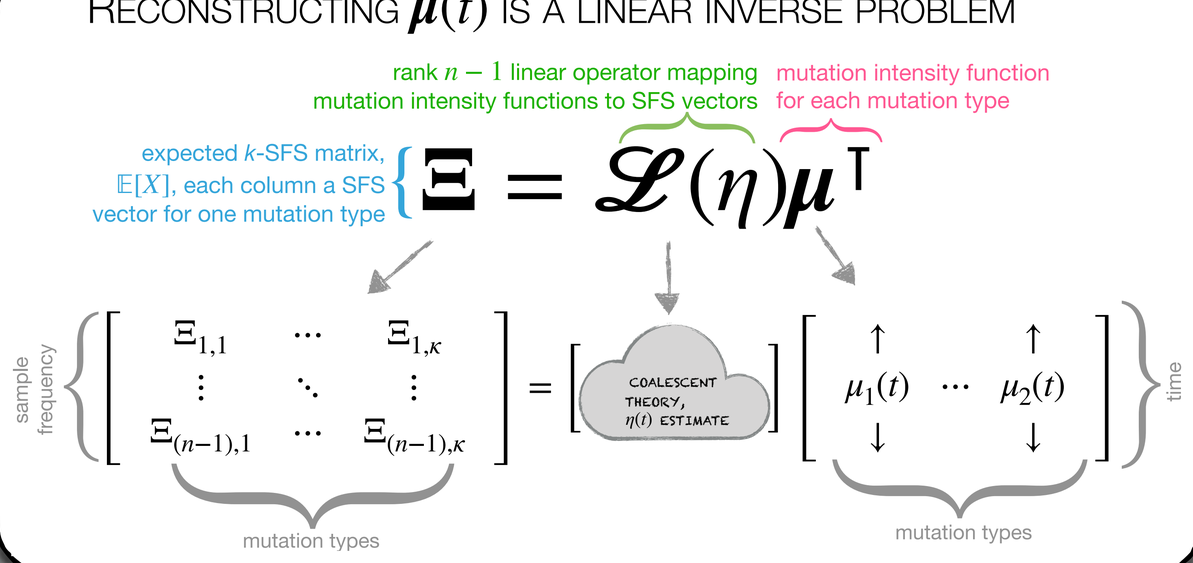
Reconstructing $\boldsymbol{\mu}(t)$ is a linear inverse problem

rank $n-1$ linear operator mapping mutation intensity functions to SFS vectors

mutation intensity function for each mutation type

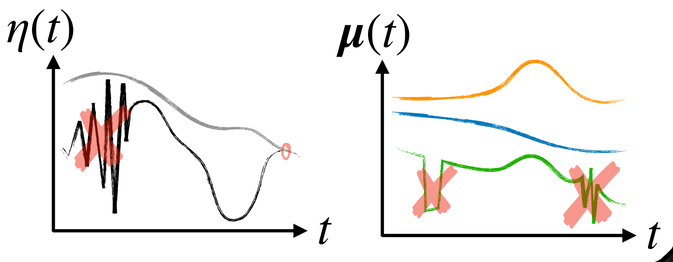
expected k-SFS matrix, $\mathbb{E}[X]$, each column a SFS vector for one mutation type

$$\boldsymbol{\Xi} = \mathscr{L}(\eta)\boldsymbol{\mu}^{\mathsf{T}}$$



Penalized maximum likelihood estimation

Penalize complexity in demographic history $\eta(t)$ and mutation spectrum history $\boldsymbol{\mu}(t)$ to regularize ill-posed inverse problems

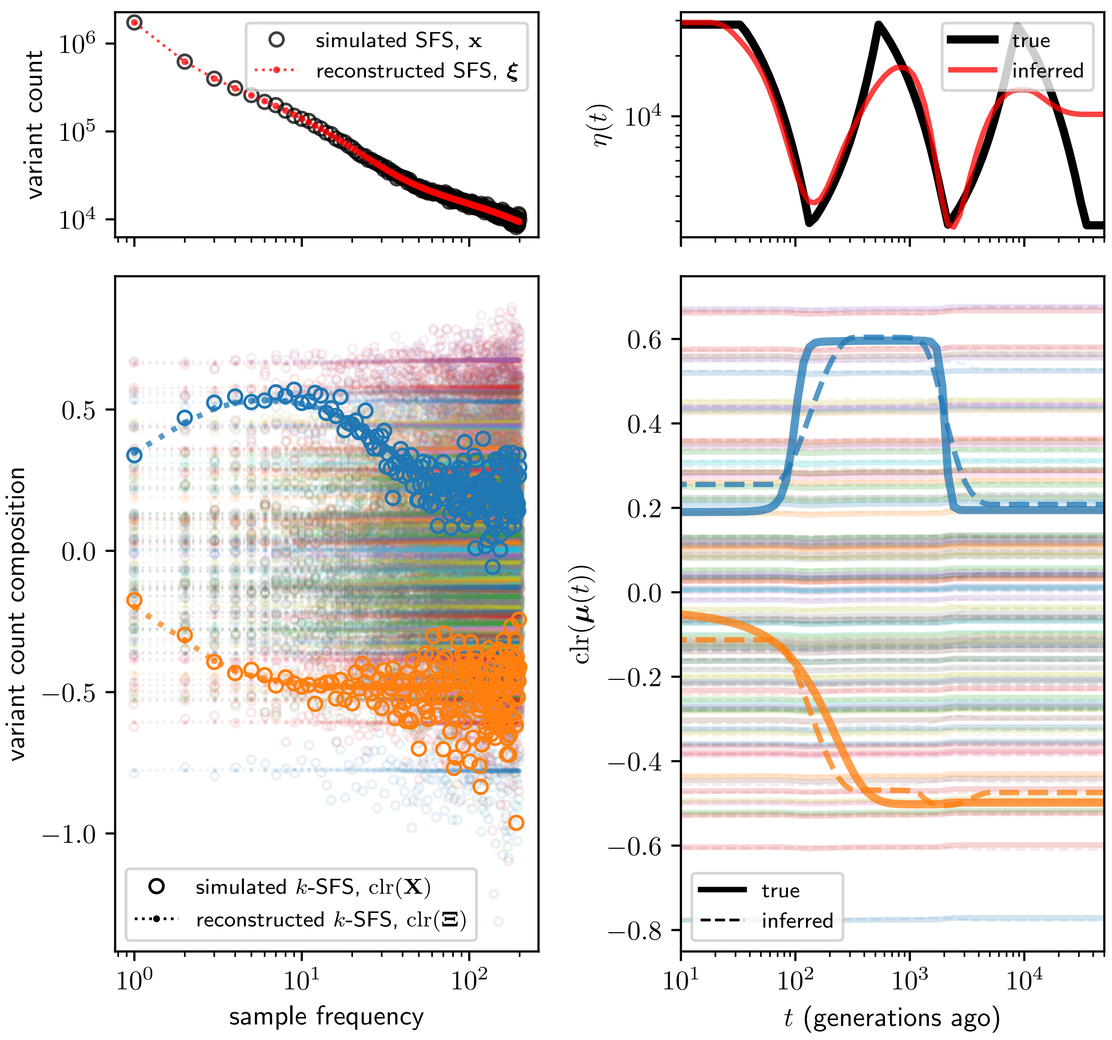


Open-source python 3 package(s)

- Source code: github.com/harrispopgen/mushi
- Installation with pip/conda
- Inference of demographic history and mutation spectrum history from population scale data takes a few seconds with modest hardware
- No model specification required
- For processing mutation spectra from VCFs check out: github.com/harrispopgen/mutyper

Reconstructing simulated histories

- Sawtooth demography simulated for human chromosome 1 (h/t `stdpopsim`)
- 96 mutation types: one pulse, one monotonic ramp, the rest flat

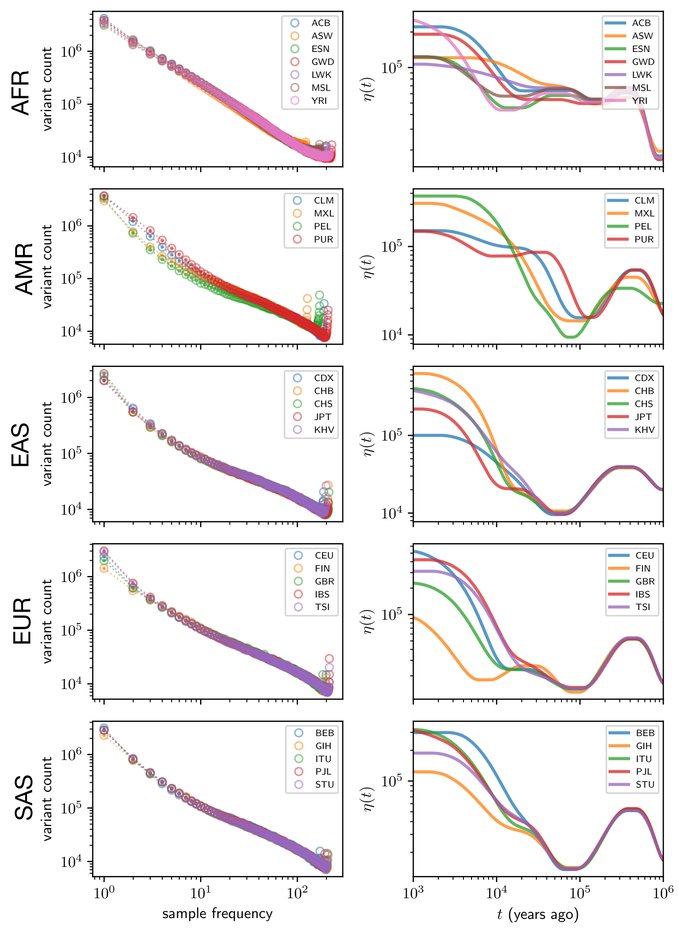


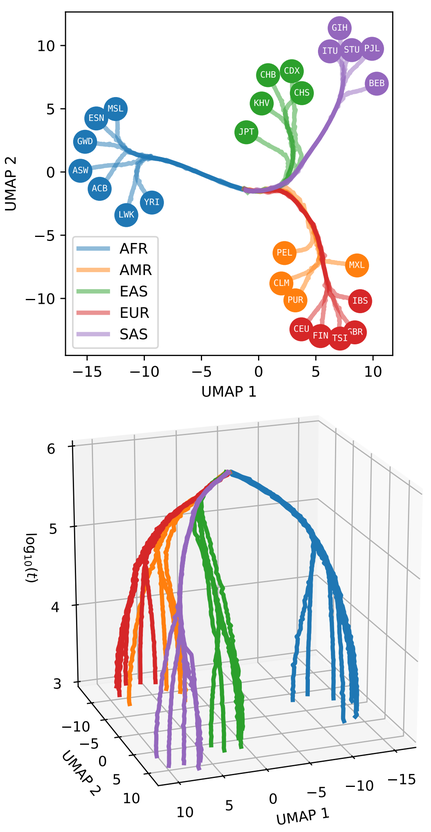
Reconstructing histories from 1000 Genomes Project

We estimate effective haploid population size histories $\eta(t) \equiv 2N_e(t)$ from the site frequency spectrum (SFS) in a few seconds without specifying a model, recovering many known features of human demography.

We estimate mutation spectrum histories $\boldsymbol{\mu}(t)$ from the k-SFS.

Naive manifold embedding of mutation spectra through time reveal tree-like population divergences in the mutational process.





Timing of `TCC>TTC` pulse in Europeans appeared younger in a previous method that used a misspecified demography. Joint estimation of $\eta(t)$ and $\boldsymbol{\mu}(t)$ resolves this.