Author date: 2023-05-05

A Simple Estimator of Lineal Admixture Time

E. Castedo Ellerman (castedo@castedo.com)

Abstract

STAGE: Early Draft

DOCUMENT TYPE: Mathematical Results

OBJECTIVES

- Deduce estimator of lineal admixture time based on simple class of haploid lineage processes

Introduction

Lineal admixture time [1] [2] is a microscale measure of admixture timing. In this document, we derive an estimator of the average lineal admixture time of a population. This derivation is based on a simple class of haploid lineage processes [2].

Notation

- $\mathbb{N}_0$ represents the non-negative integers (0, 1, 2, 3, ...).
- $f : D \mapsto I$ denotes a function f maps domain D to image I.

Simple Model

We consider a simple model of population migration and reproduction as follows:

- discrete regular time steps,
- lifespans of only one time step,
- non-overlapping generations (like Wright-Fisher model),
- geographic “islands” of single interior zone and multiple peripheral isolated zones,
- constant expected flow of migrants from isolated zones to interior zone,
- each time step consists of migration followed by mating, and
- mating is random within each zone (following migration).

Haploid Lineage Process

We use a haploid lineage process [2] to mathematically model population migration and reproduction. The motivation for this mathematical model is to analytically derive an estimator of lineal admixture time.

A haploid lineage process is defined in terms of a fertilization function. We choose a fertilization function Fert whose image of time points Tim is the set of integers $\mathbb{Z}$.

Formally, geography is modeled via a function $\mathrm{Geo} : \mathrm{Dip} \mapsto \mathbb{N}_0$ which maps diploids to the geographic zone in which they were fertilized. Zero indexes the interior zone where admixture can occur. Positive integers index the isolated geographic zones where only non-admixed diploids are found.

> *ISSUE: Need to justify why admixture time with the following categorization function makes sense.*

Similarly, we define a categorization function $\mathrm{Cat} : \mathrm{Dip} \mapsto \mathbb{N}_0$ for lineal admixture time where zero indexes admixed diploids and positive integers index non-admixed diploids. Non-admixed diploids of having matching category and isolated zone location:

$$\mathrm{Geo}(d) \neq 0 \text{ implies } \mathrm{Geo}(d) = \mathrm{Cat}(d)$$

for all $d \in \mathrm{Dip}$.

> *ISSUE: Relying on zero to representing interior zone does not seem explicit enough.*

> *ISSUE: The following math formal details should move into the math definition of lineal admixture time.*

We only consider haploid lineage processes which satisfy the following requirements regarding the categorization function for lineal admixture time [3]:

1. For all child haploids $(d, s) \in \mathrm{dom}\,\mathrm{Par}$ if $\mathrm{Cat}(d) \neq 0$ then $\mathrm{Cat}(\mathrm{Par}((d, s))) = \mathrm{Cat}(d)$, and
2. if $\mathrm{Cat}(\mathrm{Par}((d, 0))) = \mathrm{Cat}(\mathrm{Par}((d, 1)))$ then $\mathrm{Cat}(d) = \mathrm{Cat}(\mathrm{Par}((d, 0)))$.

Random individuals and lineages

We model the population of interior zone individuals living at time t as

$$\mathrm{Pop}_t := \{d : (d, s) \in \mathrm{dom}\,\mathrm{Par}_{t-1} \text{ and } \mathrm{Geo}(d) = 0\}$$

since $\mathrm{dom}\,\mathrm{Par}_{t-1}$ is the set of children fertilized one time step prior when the previous generation was living.

We define Δ_t to be a random variable which is any member of Pop_t with equal probability. Formally, given outcome space Ω, for every $t \in \mathrm{Tim}$, $\omega \in \Omega$, and diploid $d \in \mathrm{Pop}_t(\omega)$,

$$\mathbb{P}\left(\{\omega' : \Delta_t(\omega') = d\} \mid \{\omega' : \mathrm{Pop}_t(\omega') = \mathrm{Pop}_t(\omega)\}\right) = |\mathrm{Pop}_t(\omega)|^{-1}$$

We define $S \in \{0, 1\}$ to be a Bernoulli random variable representing a random gamete or sex.

We define $\Lambda \in \mathrm{Loc}$ to be a random genomic location.

Formal model assumptions

Formally, the mathematical assumptions are:

- proportion α_i from i-th ancestral isolated populations,
- immigration such that ϕ of the interior population is new non-admixed immigrants
- stationary distribution of lineal admixture times per generation.

Generations are non-overlapping: for all $t \in \mathrm{Tim}$, $h \in \mathrm{Pop}_t \times \{0, 1\}$,

$$\mathrm{Par}(h) \in \mathrm{Pop}_{t-1} \, .$$

We assume that all mating occurs within a single geographic zone:

$$\mathrm{Geo}(\mathrm{Par}((d, 0))) = \mathrm{Geo}(\mathrm{Par}((d, 1))))$$

for all $d \in \mathrm{Dip}$.

$$\mathbb{P}\left(\mathrm{Geo}(\mathrm{Par}((\Delta_t, 0))) = 0\right) = 1 - \phi$$

and for all isolated zones $i > 0$,

$$\mathbb{P}\left(\mathrm{Geo}(\mathrm{Par}((\Delta_t, 0))) = i\right) = \phi\alpha_i \, .$$

Main Result

Formal notation

We define random lineal admixture time at time t as

$$M_t := \mathrm{Lat}_t(\mathrm{Lin}(\Lambda, (\Delta_t, S)))$$

Base Facts

For $i > 0$,

$$\mathbb{P}\left(\mathrm{Cat}(\Delta_t) = i\right) = \phi\alpha_i + (1 - \phi)\,\mathbb{P}\left(\mathrm{Cat}(\Delta_{t-1}) = i\right)^2$$

From the definition of lineal admixture time

$$\begin{aligned}
\mathrm{E}[M_{t+1}] = {} & (\mathrm{E}[M_t | M_t > 0] + 1)\,\mathbb{P}\{M_t > 0\}^2(1 - \phi) \\
& + 2\left(\frac{1}{2}\mathrm{E}[M_t | M_t > 0] + 1\right)\mathbb{P}(M_t > 0)\,\mathbb{P}(M_t = 0)(1 - \phi) \\
& + \left(\mathbb{P}(M_t = 0)^2 - \sum_i \mathbb{P}(\mathrm{Cat}(\Delta_t) = i)^2\right)(1 - \phi)
\end{aligned}$$

Derivation

Given the assumptions of stationarity, we can define:

Let $x_i := \mathbb{P}(\mathrm{Cat}(\Delta_t) = i)$ so that

$$x_i = \phi\alpha_i + (1 - \phi)x_i^2 \, .$$

By theorem 1, the only quadratic solution for x_i is

$$x_i = \frac{1 - \sqrt{1 - 4\phi(1-\phi)\alpha_i}}{2(1-\phi)}$$

Let $q := \mathbb{P}(M_t = 0)$, thus

$$q = \phi + (1-\phi)\sum_i x_i^2$$

We define

$$\mu := \mathrm{E}[M_t]$$

which is the expected lineal admixture time (and generation number).

Given the base facts, we make the following deduction using the newly defined variables μ, ϕ and x_i.

$$\begin{aligned}
\mu &= (\mathrm{E}[M_t | M_t > 0] + 1)(1-q)^2(1-\phi) \\
&\quad + 2\left(\frac{1}{2}\mathrm{E}[M_t | M_t > 0] + 1\right)(1-q)q(1-\phi) \\
&\quad + \left(q^2 - \sum_i x_i^2\right)(1-\phi) \\
&= \mu(1-q)(1-\phi) + (1-q)^2(1-\phi) \\
&\quad + \mu q(1-\phi) + 2(1-q)q(1-\phi) \\
&\quad + \left(q^2 - \sum_i x_i^2\right)(1-\phi) \\
&= \mu(1-\phi) + ((1-q)+q)^2(1-\phi) - (1-\phi)\sum_i x_i^2 \\
0 &= -\mu\phi + 1 - q \\
\mu &= \frac{1-q}{\phi}
\end{aligned}$$

Replacing q gets

$$\mu = \frac{1-\phi}{\phi}\left(1 - \sum_i x_i^2\right)$$

We conjecture that this formula serves as a consistent maximum likelihood estimator.

Estimation of ϕ

Let $\ddot{\alpha}_{i,j}$ denote the frequency of a diploid genotype with an i-th maternal ancestral source and j-ith paternal ancestral source.

Thus

$$\begin{aligned}
\ddot{\alpha}_{i,i} &= \phi\alpha_i + (1-\phi)\alpha_i^2 \\
&= \phi\alpha_i(1-\alpha_i) + \alpha_i^2 \\
\phi &= \frac{\ddot{\alpha}_{i,i} - \alpha_i^2}{\alpha_i(1-\alpha_i)}
\end{aligned}$$

Consider the case of only two ancestral sources. With $\beta := \ddot{\alpha}_{0,1} + \ddot{\alpha}_{1,0}$ we deduce that

$$\ddot{\alpha}_{0,0} + \ddot{\alpha}_{1,1} = 1 - \beta$$

$$\alpha_1 = 1 - \alpha_0$$

$$\alpha_0^2 + \alpha_1^2 = 1 - 2\alpha_0(1 - \alpha_0)$$

$$\begin{aligned}\phi &= \frac{\phi + \phi}{2} \\ &= \frac{\ddot{\alpha}_{0,0} + \ddot{\alpha}_{1,1} - \alpha_0^2 - \alpha_1^2}{2\alpha_0(1 - \alpha_0)} \\ &= 1 - \frac{\beta}{2\alpha_0(1 - \alpha_0)}\end{aligned}$$

This form is the same as the *inbreeding coefficient* but with ancestral source as the allele state rather than haplotype.

Theorem 1

The solution to x_i given stationarity etc... can not be

$$x_i = \frac{1 + \sqrt{1 - 4\phi(1 - \phi)\alpha_i}}{2(1 - \phi)}$$

when $\alpha_i < 1$.

PROOF

Assume the contrary. Since $\phi < 1$ and $x_i \leq 1$, we have

$$\begin{aligned}1 &\geq \frac{1 + \sqrt{1 - 4\phi(1 - \phi)\alpha_i}}{2(1 - \phi)} \\ 2(1 - \phi) &\geq 1 + \sqrt{1 - 4\phi(1 - \phi)\alpha_i} \\ 1 - 2\phi &\geq \sqrt{1 - 4\phi(1 - \phi)\alpha_i} \\ 1 - 4\phi + 4\phi^2 &\geq 1 - 4\phi(1 - \phi)\alpha_i \\ -4\phi(1 - \phi) &\geq -4\phi(1 - \phi)\alpha_i \\ 1 &\leq \alpha_i\end{aligned}$$

which can not be true given $\alpha_i < 1$.

References

1. Ellerman EC. Lineal admixture time: An interdisciplinary definition. 2023. Available: https://perm.pub/D9qSdCY6GPrxthT3ZnFouEU35ow
2. Ellerman EC. Haploid lineage process. Available: https://castedo.com/doc/153
3. Ellerman EC. Lineal admixture time: A mathematical definition. Available: https://castedo.com/doc/cP