perm.pub/dsi: Fs4CjTor07Ssbb69Qcggp5yiBA0/0.4
Additional formats and editions available online.

Edition 0.4

Author date: 2023-04-24

Archive date: 2023-10-08

Citation:
E. Castedo Ellerman (2023) "Haploid lineage process" perm.pub https://perm.pub/dsi: Fs4CjTor07Ssbb69Qcggp5yiBA0/0.4

Haploid lineage process

E. Castedo Ellerman (castedo@castedo.com)

Abstract

STAGE: WORKING DRAFT

DOCUMENT TYPE: Formal Mathematical Definition

OBJECTIVES:

- Formal mathematical definition of stochastic process used by statistical estimator of admixture timing under development.
- Precise mathematical definitions for technical discussions relating to ancestral recombination graphs.

Notation

- "$\operatorname{dom} f$" denotes the domain of function f
- "$[\![x \in S]\!]$" is Iverson bracket notation for the indicator function of S, namely

$$[\![x \in S]\!] = \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise.} \end{cases}$$

Fertilization Function

A *haploid lineage process* is formally defined in terms of a given fixed *fertilization function*, which we denote with the symbol Fert. This function maps possible occurrences of fertilization to points in time.

From a given fertilization function Fert, we define three convenient symbols:

- Tim: the range of Fert (a set of real numbers representing points in time),
- Dip: the domain of Fert (a set of possible diploid organisms), and
- Hap: the set $\mathrm{Dip} \times \{0, 1\}$ representing fertilizing gametes.

For a diploid $d \in \mathrm{Dip}$, the members $(d, 0)$ and $(d, 1)$ of Hap index the fertilizing egg and sperm gametes, respectively.

For convenience, we map Hap to fertilization times with:

$$\mathrm{Fert}_H((d,s)) := \mathrm{Fert}(d)$$

for all $d \in \mathrm{Dip}$ and $s \in \{0,1\}$.

We denote the following inverse images as:

$$\begin{aligned}\mathrm{Dip}_t &:= \{d \in \mathrm{Dip} : \mathrm{Fert}(d) = t\} \\ \mathrm{Hap}_t &:= \{d \in \mathrm{Hap} : \mathrm{Fert}_H(d) = t\} \\ \mathrm{Dip}_{<t} &:= \{d \in \mathrm{Dip} : \mathrm{Fert}(d) < t\}\ .\end{aligned}$$

A technical requirement on any given Fert in this document is that Dip_t must be countable for every $t \in \mathrm{Tim}$.

An example of a valid fertilization function is $\mathrm{Fert} : \mathbf{N}^2 \mapsto \mathbf{N}$ where $\mathrm{Fert}((t,n)) = t$.

Haploid lineage process

Given

- a fertilization function Fert,
- a set Loc of genomic locations,
- and probability space $(\Omega, \mathcal{F}, \mathbb{P})$,

a *Haploid lineage process* is a stochastic process defined by a time-indexed family of two random variables

$$\{(\mathrm{Par}_t, \mathrm{Pat}_t)\}_{t \in \mathrm{Tim}}\ .$$

Each Par_t is a random function from a subset of Hap_t to $\mathrm{Dip}_{<t}$. $\mathrm{Par}_t(h)$ represents the diploid **Par**ent that produced the haploid gamete h.

Each Pat_t is a random function from the domain of Par_t to subsets of Loc. $\mathrm{Pat}_t(h)$ represents the genome locations replicated into the gamete h from the **Pat**ernal haploid genome of the parent $\mathrm{Par}_t(h)$ (inherited from the father of the parent producing the gamete).

For convenience we define:

$$\begin{aligned}\mathrm{Par}(h) &:= \mathrm{Par}_t(h) \\ \mathrm{Pat}(h) &:= \mathrm{Pat}_t(h)\end{aligned}$$

where $t = \mathrm{Fert}_H(h)$.

Haploid lineages

A haploid lineage process induces a random function Lin which maps a genomic location in a descendant haploid genome to a haploid lineage. A haploid lineage consists of all the haploids transmitting genetic information via a genomic location to a descendant haploid. For every genomic location $\ell \in \mathrm{Loc}$ and haploid $h \in \mathrm{Hap}$, its haploid lineage is

$$\mathrm{Lin}(\ell, h) := \bigcup_i \{a_i\}$$

where a_i is defined inductively as follows:

$$a_0 := h$$

and for integers $i > 0$,

$$a_{i+1} := \begin{cases} (\mathrm{Par}(a_i), [\![\ell \in \mathrm{Pat}(a_i)]\!]) & \text{if } a_i \in \mathrm{dom}\,\mathrm{Par} \\ a_i & \text{otherwise.} \end{cases}$$

An embedded Ancestral Recombination Graph

An ancestral recombination graph [1] [2] [3] of a sampled population is embedded in any outcome of any haploid lineage process. We formally show the exact embedding using the gARG formalism [4].

We start by defining the *genetic legacy* of an ancestral haploid $h \in \mathrm{Hap}$ for sample population $S \subseteq \mathrm{Hap}$ to be

$$\mathrm{Leg}(h, S) := \{(\ell, d) \in \mathrm{Loc} \times S : h \in \mathrm{Lin}(\ell, d)\} .$$

This genetic legacy is the genetic material that survives in the sample population S originally copied from ancestral haploid h (with or without mutations).

Genetic legacy for a sample population S induces the following equivalence relationship over pairs of haploids h_1 and h_2 in Hap:

$$h_1 \simeq_S h_2 \;:=\; \mathrm{Leg}(h_1, S) = \mathrm{Leg}(h_2, S) .$$

We denote the resulting equivalence class containing $h \in \mathrm{Hap}$ as

$$[h]_S \;:=\; \{h' : \mathrm{Leg}(h', S) = \mathrm{Leg}(h, S)\} .$$

In this equivalence relationship, haploids are considered equivalent if they have the same genetic legacy for the sample population S.

A convenient choice for an embedded gARG [4] is to set the gARG nodes (vertices) to be the equivalence classes:

$$\mathrm{Nodes}(S) := \{[h]_S : h \in \mathrm{Hap}\} .$$

The (unannotated) graph edges of the gARG are chosen as child-parent node pairs $(C, P) \in \mathrm{Nodes}(S)^2$ where

$$(\mathrm{Par}(h), i) \in P \text{ for some } h \in C \text{ and some } i \in \{0, 1\} .$$

In the gARG, annotations are added for each graph edge (pair of child and parent nodes). This annotation is the set of locations through which genetic information has been copied from parent to child. In the following interpretation, the only locations of interest are those for which genetic information has been transmitted into the sample population S. With this interpretation, the annotation for edge (C, P) is

$$\{\ell \in \mathrm{Loc} : C \subseteq \mathrm{Lin}(\ell, h) \text{ and } P \subseteq \mathrm{Lin}(\ell, h) \text{ for some } h \in S\} .$$

Acknowledgements

Thanks to Daria Shipilina and Nick Barton for sharing their preprint [5] and discussing the conjecture in edition 0.1 of this document relating to their preprint.

Changes from edition 0.1

- add section about embedded ARG
- removed conjecture relating to [5]

References

1. Griffiths RC, Marjoram P. An Ancestral Recombination Graph. In: Friedman A, Miller W, Donnelly P, Tavaré S, editors. Progress in Population Genetics and Human Evolution. New York, NY: Springer New York; 1997. pp. 257–270.
2. Hein J, Schierup MH, Wiuf C. Gene genealogies, variation and evolution: A primer in coalescent theory. Oxford ; New York: Oxford University Press; 2005.
3. Wakeley J. Coalescent theory: An introduction. Greenwood Village, Colo: Roberts & Co. Publishers; 2009.
4. Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J. A general and efficient representation of ancestral recombination graphs. https://archive.softwareheritage.org/swh:1:rev:7df4f1995028cc676a6c1b231e8d7a024666b5fc; 2022.
5. Shipilina D, Stankowski S, Pal A, Chan YF, Barton N. On the origin and structure of haplotype blocks. Preprints; 2022 Feb. doi:10.22541/au.164425910.09070763/v1