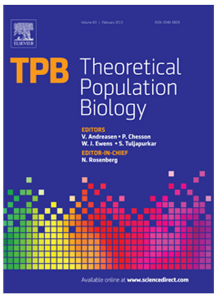

Single and simultaneous binary mergers in Wright-Fisher genealogies

Andrew Melfi, Divakar Viswanath

PII: S0040-5809(17)30173-9
DOI: https://doi.org/10.1016/j.tpb.2018.04.001
Reference: YTPBI 2639

To appear in: *Theoretical Population Biology*

Received date : 13 October 2017

Please cite this article as: Melfi A., Viswanath D., Single and simultaneous binary mergers in Wright-Fisher genealogies. *Theoretical Population Biology* (2018), https://doi.org/10.1016/j.tpb.2018.04.001

Single and Simultaneous Binary Mergers in Wright-Fisher Genealogies

Andrew Melfi and Divakar Viswanath

March 28, 2018

Department of Mathematics, University of Michigan (melfi/divakar@umich.edu).

Abstract

The Kingman coalescent is a commonly used model in genetics, which is often justified with reference to the Wright-Fisher (WF) model. Current proofs of convergence of WF and other models to the Kingman coalescent assume a constant sample size. However, sample sizes have become quite large in human genetics. Therefore, we develop a convergence theory that allows the sample size to increase with population size. If the haploid population size is N and the sample size is $N^{1/3-\epsilon}$, $\epsilon > 0$, we prove that Wright-Fisher genealogies involve at most a single binary merger in each generation with probability converging to 1 in the limit of large N. Single binary merger or no merger in each generation of the genealogy implies that the Kingman partition distribution is obtained exactly. If the sample size is $N^{1/2-\epsilon}$, Wright-Fisher genealogies may involve simultaneous binary mergers in a single generation but do not involve triple mergers in the large N limit. The asymptotic theory is verified using numerical calculations. Variable population sizes are handled algorithmically. It is found that even distant bottlenecks can increase the probability of triple mergers as well as simultaneous binary mergers in WF genealogies.

1 Introduction

The Kingman coalescent (Kingman, 1982a,b) is a mathematical model of the genealogy of n haploid samples. If k lineages are present in some earlier generation, those lineages induce a partition of the n current samples into k. For convenience, we will refer to lineages present in earlier generations as ancestral samples.[1]

One of Kingman's motivations in deriving the coalescent (Kingman, 1982a,b, 2000) was to gain an understanding of the structure of Ewens' sampling formula (Ewens, 1972, Durrett, 2008). The coalescent gives an almost instantaneous derivation of Ewen's sampling formula, and Ewens' sampling formula is exact under the coalescent approximation. The coalescent is perfectly memoryless in the following sense: at every coalescence exactly two ancestral samples are picked at random (without regard to the number or inter-relationship of their descendants)

---

[1] The "ancestral sample" nomenclature is more intuitive for our purposes. However, in the context of the coalescent, the same concept is referred to as "lineage" or "ancestral lineage" (Griffiths, 2006, Griffiths and Tavaré, 1998, Tavaré, 1984).