Annales de l'Institut Henri Poincaré - Probabilités et Statistiques
2019, Vol. 55, No. 2, 1087–1115
https://doi.org/10.1214/18-AIHP910

ANNALES
DE L'INSTITUT
HENRI
POINCARÉ
PROBABILITÉS
ET STATISTIQUES
www.imstat.org/aihp

Bayesian nonparametric analysis of Kingman's coalescent

Stefano Favaro[a,*], Shui Feng[b] and Paul A. Jenkins[c]

[a]*Department of Economics and Statistics, University of Torino, Torino 10134, Italy. E-mail:* stefano.favaro@unito.it
[b]*Department of Mathematics and Statistics, McMaster University, Hamilton L8S4K1, Canada. E-mail:* shuifeng@mcmaster.ca
[c]*Department of Statistics & Department of Computer Science, University of Warwick, Coventry CV4 7AL, United Kingdom. E-mail:* P.Jenkins@warwick.ac.uk

Received 26 June 2017; revised 7 April 2018; accepted 17 April 2018

Dedicated to the memory of Paul Joyce

Abstract. Kingman's coalescent is one of the most popular models in population genetics. It describes the genealogy of a population whose genetic composition evolves in time according to the Wright–Fisher model, or suitable approximations of it belonging to the broad class of Fleming–Viot processes. Ancestral inference under Kingman's coalescent has had much attention in the literature, both in practical data analysis, and from a theoretical and methodological point of view. Given a sample of individuals taken from the population at time $t > 0$, most contributions have aimed at making frequentist or Bayesian parametric inference on quantities related to the genealogy of the sample. In this paper we propose a Bayesian nonparametric predictive approach to ancestral inference. That is, under the prior assumption that the composition of the population evolves in time according to a neutral Fleming–Viot process, and given the information contained in an initial sample of m individuals taken from the population at time $t > 0$, we estimate quantities related to the genealogy of an additional unobservable sample of size $m' \geq 1$. As a by-product of our analysis we introduce a class of Bayesian nonparametric estimators (predictors) which can be thought of as Good–Turing type estimators for ancestral inference. The proposed approach is illustrated through an application to genetic data.

Résumé. La coalescence de Kingman est l'un des modèles les plus populaires en génétique des populations. Il décrit la généalogie d'une population dont la composition génétique évolue dans le temps selon le modèle de Wright–Fisher, ou des approximations appropriées de celle-ci appartenant à la grande classe des processus de Fleming–Viot. L'inférence ancestrale sous la coalescence de Kingman a reçu beaucoup d'attention dans la littérature, à la fois dans l'analyse des données, et d'un point de vue théorique et méthodologique. Étant donné un échantillon d'individus échantillonnés dans la population au temps $t > 0$, la plupart des contributions existantes visaient l'inférence paramétrique, fréquentiste ou bayésienne, sur des quantités liées à la généalogie de l'échantillon. Dans cet article, nous proposons une approche prédictive bayésienne non paramétrique de l'inférence ancestrale. C'est-à-dire, sous l'hypothèse préalable que la composition de la population évolue dans le temps selon un processus de Fleming–Viot neutre, et compte tenu de l'information contenue dans un échantillon initial de m individus dans la population au temps $t > 0$, nous estimons des quantités liées à la généalogie d'un échantillon additionnel non observable de taille $m' \geq 1$. En corollaire de notre analyse, nous introduisons une classe d'estimateurs bayésiens non paramétriques (prédicteurs) qui peuvent être considérés comme des estimateurs de type Good–Turing pour l'inférence ancestrale. L'approche proposée est illustrée par une application sur données génétiques.

MSC: Primary 62C10; secondary 62M05

Keywords: Ancestral inference; Bayesian nonparametrics; Dirichlet process; Kingman's coalescent; Lineages distributions; Predictive probability

[*] Also affiliated to Collegio Carlo Alberto (Torino, Italy) and IMATI-CNR "Enrico Magenes" (Milan, Italy).

1. Introduction

The Wright–Fisher (WF) model is a popular discrete-time model for the evolution of gene frequencies in a population. Consider a population of individuals, i.e. chromosomes, and assume that each individual has an associated genetic type, with $\mathcal{X}$ being the set of possible types. In the classical WF model the population has constant (large) size N and it evolves in discrete non-overlapping generations according to the following random processes: (i) each individual in the next generation chooses, uniformly at random, an individual in the current generation and copies it, with the choice made by different individuals being independent; (ii) the type of each progeny of an individual of type $i \in \mathcal{X}$ is i with probability $1-\delta$, and j with probability $\delta p_{i,j}$, that is mutations occur with probability $\delta \in (0,1)$ per individual per generation, according to a Markov chain with zero-diagonal transition matrix $P=(p_{i,j})_{i\geq 1, j\geq 1}$. Two additional common assumptions are that P has a unique stationary distribution and that the evolution of the population is neutral, namely all variants in the population are equally fit and are thus equally likely to be transmitted. The assumption of neutrality allows for a crucial simplification of the above WF evolution. Indeed under this assumption the random process describing the demography of the population becomes independent of the random process describing the genetic types carried by the individuals. Although rather simple, the neutral WF model captures many important features of the evolution of human and other populations, thus providing a statistical model which is at the basis of most existing inference methods in population genetics. We refer to the monographs by Ewens [9] and Tavaré [34] for a comprehensive and stimulating account on the WF model.

In the WF model one can describe the genetic composition of the population at any point in time by giving a list of the genetic types currently present, and the corresponding proportion of the population currently of each type. Note that such a description corresponds to giving a probability measure on the set $\mathcal{X}$ of possible population types. In such a framework one obtains a discrete time probability-measure-valued Markov process, namely a discrete time Markov process whose state space corresponds to the space of the probability measures on $\mathcal{X}$. As the population size N becomes large a suitable rescaling of the Markov process converges to a diffusion limit: time is measured in units of N generations, and the mutation rates are rescaled as N^{-1}. The limiting process, called the Fleming–Viot (FV) process, is formulated as a diffusion process whose state space is the space of probability measures on an arbitrary compact metric space $\mathcal{X}$. See Ethier and Kurtz [7] and references therein for a rigorous treatment with a view towards population genetics. Intuitively, the FV process can thus be thought of as an approximation to a large population evolving in time according to the WF model. For instance, the classical WF diffusion on the set [0, 1] is a special case of the FV process which arises when there are only two possible genetic types and one tracks the population frequency of one of the types.

The coalescent arises by looking backward in time at the evolution described by the WF model. See, e.g., the seminal works by Griffiths [13], Kingman [22], Kingman [23] and Tavaré [33], and the monographs by Ewens [9] and Tavaré [34]. Consider a population evolving in time as a WF model with scaled mutation rate $\alpha=\theta/(2N)$, for $\theta>0$, and with parent-independent transition matrix P. In the large N population limit the genealogical history of a sample of m individuals from the population may be described by a rooted random binary tree, where branches represent genealogical lineages, i.e., lines of descent. The tree initially has m lineages for a period of time T_m, after which a lineage is lost for the occurrence of one of the following events: (i) a mutation according to the transition matrix P; (ii) a coalescence, namely a pair of lineages, chosen uniformly at random and independently of all other events, join. Recursively, the times T_k, for $k=m,m-1,\dots,2$ for which the tree has k lineages are independent exponential random variables with parameter $2^{-1}k(k-1+\theta)$, after which a lineage is lost by mutation with probability $k\theta/(k(k-1+\theta))$ or by coalescence with probability $k(k-1)/(k(k-1+\theta))$. When $\theta=0$, the resulting random tree is referred to as the m-coalescent, or simply the coalescent, and was first described in the seminal work of Kingman [22]. When $\theta>0$ the process is a coalescent with mutation, with antecedents including Griffiths [13]. As shown in Donnelly and Kurtz [5], in the large N population limit the m-coalescent describes the genealogy of a sample of m individuals from a population evolving as the FV process. There exists even a natural limit, as the sample size $m\to+\infty$, of the m-coalescent. This can be thought of as the limit of the genealogy of the whole population, or alternatively as the genealogy of the infinite population described by the FV process.

This paper considers the problem of making ancestral inference, i.e. inference on the genealogy of a genetic population, from a Bayesian nonparametric predictive perspective. The statistical setting we deal with can be described as follows. We consider a population with an (ideally) infinite number of types, and we assume that the population's composition evolves in time according to a neutral FV process whose unique stationary distribution is the law of the

Dirichlet process by Ferguson [10]. From a Bayesian perspective, the law of the FV process, or its dual law determined with respect to the Kingman's coalescent process, plays the role of a nonparametric prior for the evolution of the population. Given the observed data, which are assumed to be a random sample sample of m individuals from the population at time $t > 0$, we characterize the posterior distribution of some statistics of the enlarged $(m+m')$-coalescent induced by an additional unobservable sample of size $m' \geq 1$. Corresponding Bayesian nonparametric estimators, with respect to a squared loss function, are then given in terms of posterior expectations. Of special interest is the posterior distribution of the number of non-mutant lineages surviving from time 0 to time t, that is the number of non-mutant ancestors in generation t in a sample at time 0. This, in turn, leads to the posterior distribution of the time of the most recent common ancestor in the $(m+m')$-coalescent. As a by-product of our posterior characterizations we introduce a class of Bayesian nonparametric estimators of the probability of discovery of non-mutant lineages. This is a novel class of estimators which can be thought of as ancestral counterparts of the celebrated Good–Turing type estimators developed in Good [11] and Good and Toulmin [12].

Ancestral inference has had much attention in the statistical literature, both in practical data analysis, and from a theoretical and methodological point of view. See, e.g., Griffiths and Tavaré [16], Griffiths and Tavaré [17], Stephens and Donnelly [32], Stephens [31] and Griffiths and Tavaré [18]. Given an (observable) random sample sample of m individuals from the population at time $t > 0$, most contributions in the literature have aimed at making frequentist or Bayesian inference on quantities related to the genealogy of the sample, e.g., the number of non-mutant lineages, the age of the alleles in the sample, the time of the most recent common ancestor, the age of particular mutations in the ancestry, etc. This is typically done in a parametric setting by using suitable summary statistics of the data, or by combining the full data with suitable approximations of the likelihood function obtained via importance sampling or Markov chain Monte Carlo techniques. In this paper, instead, we introduce a Bayesian nonparametric predictive approach that makes use of the observed sample of m individuals to infer quantities related to the genealogy of an additional unobservable sample. For instance, how many non-mutant lineages would I expect a time t ago if I enlarged my initial observable sample by m' unobservable samples? How many of these non-mutant lineages have small frequencies? In the context of ancestral inference, these questions are of great interest because they relate directly to the speed of evolution via the rate of turnover of alleles. See Stephens and Donnelly [32] and references therein for a comprehensive discussion. Our approach answers these and other questions under the (prior) assumption that the genealogy of the population follows the Kingman coalescent. To the best of our knowledge, this is the first predictive approach to ancestral inference in this setting.

The paper is structured as follows. In Section 2 we recall some preliminaries on the neutral FV process and Kingman's coalescent, we introduce new results on ancestral distributions, and we characterize the posterior distribution of the number of non-mutant lineages at time t back in the enlarged $(m+m')$-coalescent. A suitable refinement of this posterior distribution and a class of Good–Turing type estimators for ancestral inference are also introduced. In Section 3 we show how to implement our results, and we present a numerical illustration based on genetic data. Section 4 contains a discussion of the proposed methodology, and outlines future research directions. Proofs are deferred to the Appendix.

2. Ancestral posterior distributions

All the random elements introduced in this section are meant to be assigned on a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, unless otherwise stated. Let $\mathcal{X}$ be a compact metric space. For any $\theta > 0$ and any non-atomic probability measure ν_0 on $\mathcal{X}$, let $\Pi(\theta\nu_0)$ be the distribution of a Dirichlet process on $\mathcal{X}$ with base measure $\theta\nu_0$. We refer to Ferguson [10] for a definition and distributional results on the Dirichlet process. In our context θ will correspond to the mutation rate, ν_0 to the stationary distribution of the mutation process, and $\Pi(\theta\nu_0)$ to the stationary distribution of the population type frequencies in the diffusion limit. For any $n \geq 0$ let $d_n(t) = \mathbb{P}[D(t) = n]$ where $\{D(t) : t \geq 0\}$ is a pure death process, $D(0) = +\infty$ almost surely, with rate $\lambda_n = 2^{-1}n(n-1+\theta)$. It is known from Griffiths [13] that

$$d_n(t) = (-1)^n \sum_{i \geq n} \rho_i(t) \frac{\binom{i}{n}(n+\theta)_{(i-1)}}{i!}, \tag{2.1}$$

where

$$\rho_i(t) = (-1)^i (2i-1+\theta) \mathrm{e}^{-\lambda_i t}$$

for each $t > 0$. Here and elsewhere, for any nonnegative x we use $x_{(0)} = x_{[0]} = 1$ and, for any $n \geq 1$, $x_{(n)} = x(x+1)\cdots(x+n-1)$ and $x_{[n]} = x(x-1)\cdots(x-n+1)$, i.e. rising and falling factorial numbers. If $\alpha = \theta/(2N)$, with α being the mutation rate of the WF model, then $D(t)$ is the number of non-mutant lineages surviving from time 0 to time $t > 0$ in the large N population limit of the WF model when the sample size $m \to \infty$, and λ_n is the total backwards-in-time rate of loss of lineages when there are currently n lineages. The pure death process $\{D(t) : t \geq 0\}$ is typically referred to as the ancestral (genealogical) process. See, e.g., Griffiths [13] and Tavaré [33] for a detailed account on the ancestral process.

Let $\mathcal{P}_{\mathcal{X}}$ be the space of probability measures on $\mathcal{X}$ equipped with the topology of weak convergence. The neutral FV process is a diffusion process on $\mathcal{P}_{\mathcal{X}}$, namely a probability-measure-valued diffusion. Here we focus on the neutral FV process $\{\mu(t) : t \geq 0\}$ whose unique stationary distribution is $\Pi(\theta\nu_0)$. Among various definitions of this FV process, the most intuitive is in terms of its transition probability functions. In particular Ethier and Griffiths [6] shows that $\{\mu(t) : t \geq 0\}$ has transition function $P(t, \mu, \mathrm{d}\nu)$ given for any $t > 0$ and $\mu \in \mathcal{P}_{\mathcal{X}}$ by

$$P(t,\mu,\mathrm{d}\nu) = \sum_{n\geq 0} d_n(t) \int_{\mathcal{X}^n} \mu(\mathrm{d}Z_1)\cdots\mu(\mathrm{d}Z_n)\Pi\left(\theta\nu_0 + \sum_{i=1}^{n} \delta_{Z_i}\right)(\mathrm{d}\nu). \tag{2.2}$$

For each $t > 0$ and $\mu \in \mathcal{P}_{\mathcal{X}}$, the transition probability function (2.2) is a (compound) mixture of distributions of Dirichlet processes. More precisely, recalling the conjugacy property of the Dirichlet process with respect to multinomial sampling (see, e.g., Ferguson [10]), Equation (2.2) reads as the posterior law of a Dirichlet process with base measure $\theta\nu_0$, where: (i) the conditioning sample is randomized with respect to the n-fold product measure μ^n; (ii) the sample size n is randomized with respect to the marginal distribution of the ancestral process, i.e. $d_n(t)$ in (2.1).

Consider a population whose composition evolves in time according to the transition probability function (2.2). Given a sample $\mathbf{Y}_m(t) = (Y_1(t), \ldots, Y_m(t))$ from such a population at time $t > 0$, the m-coalescent describes the genealogy of such a sample. We denote by $C_{\mathbf{Y}_m(t)}$ the m-coalescent of the sample $\mathbf{Y}_m(t)$, and by $D_m(t)$ the number of non-mutant lineages surviving from time 0 to time t in $C_{\mathbf{Y}_m(t)}$. The distribution of $D_m(t)$ was first introduced by Griffiths [13], and further investigated in Griffiths [14] and Tavaré [33]. In particular, Griffiths [13] showed that

$$\mathbb{P}\big[D_m(t) = x\big] = (-1)^x \sum_{i=x}^{m} \rho_i(t) \frac{\binom{m}{i}\binom{i}{x}(x+\theta)_{(i-1)}}{(\theta+m)_{(i)}} \tag{2.3}$$

for any $x = 0, \ldots, m$ and each time $t > 0$. If T_r denotes the time until there are $r \geq 1$ non-mutant lineages left in the sample, then the following identity is immediate:

$$\mathbb{P}[T_r \leq t] = \mathbb{P}\big[D_m(t) \leq r\big]. \tag{2.4}$$

Note that (2.4) with $r = 1$ gives the distribution of the time of the most recent common ancestor in the sample, which is of special interest in genetic applications. For any $m \geq 1$ the stochastic process $\{D_m(t) : t \geq 0\}$ may be thought as the sampling version of the ancestral process $\{D(t) : t \geq 0\}$. Indeed it can be easily verified that (2.3) with $m = +\infty$ coincides with the probability (2.1). There are two different, but equivalent, ways to describe the evolution of $\{D_m(t) : t \geq 0\}$ with respect to the transition probability functions (2.2). Let $\mathbf{Y}_m$ be a sample from the population at time 0, i.e. a sample from a non-atomic probability measure. The first way follows Kingman's coalescent and looks backward in time: based on $\mathbf{Y}_m$, for any $t > 0$ define $D_m^*(t)$ to be the total number of equivalent classes in the m-coalescent starting with $\{1\}, \ldots, \{m\}$, that is $D_m^*(t)$ is the total number of non-mutant ancestors of $\mathbf{Y}_m$ at time $-t$ in the past. An alternative way is forward looking in time and follow the lines of descent: based on $\mathbf{Y}_m$, for any $t > 0$ define $D_m^{**}(t)$ to be the number of individuals that have non-mutant descendants at time t. It is known from the works of Griffiths [13] and Tavaré [33] that $D_m^*(t)$ and $D_m^{**}(t)$ have the same distribution, which coincides with (2.3).

2.1. *New results on ancestral distributions*

We start by introducing a useful distributional identity for $D_m(t)$. For any $n \geq 0$ let $(Z_1^*, \ldots, Z_n^*)$ be independent random variables identically distributed according to a non-atomic probability measure and, for any $m \geq 1$, let $\mathbf{X}_m = (X_1, \ldots, X_m)$ be a random sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\leq i\leq n} \delta_{Z_i^*}$. The random variables $(Z_1^*, \dots, Z_n^*)$ will be used to denote the genetic types of the ancestors. In order to keep track of the different ancestors, we add the non-atomic requirement for the law so that different ancestors will be represented by different types. Due to the almost sure discreteness of the Dirichlet process, the composition of the sample $\mathbf{X}_m$ can be described as follows. We denote by $\{X_1^*, \dots, X_{K_m}^*\}$ the labels identifying the K_m distinct types in $\mathbf{X}_m$ which do not coincide with any of the atoms Z_i^*'s. Moreover, we set

(i) $\mathbf{M}_m = (M_{1,m}, \dots, M_{n,m})$ where $M_{j,m} = \sum_{1 \le i \le m} \mathbb{1}_{\{Z_j^*\}}(X_i)$ denotes the number of X_i's that coincide with the atom Z_j^*, for any $j = 1, \dots, n$;

(ii) $\mathbf{N}_m = (N_{1,m}, \dots, N_{K_m,m})$ where $N_{j,m} = \sum_{1 \le i \le m} \mathbb{1}_{\{X_j^*\}}(X_i)$ denotes the number of X_i's that coincide with the label X_j^*, for any $j = 1, \dots, K_m$;

(iii) $V_m = \sum_{1 \le i \le K_m} N_{i,m}$ denotes the number of X_i's which do not coincide with any of the labels $\{Z_1^*, \dots, Z_n^*\}$.

Observe that the statistic $(\mathbf{N}_m, \mathbf{M}_m, K_m, V_m)$ includes all the information of $\mathbf{X}_m$, i.e., $(\mathbf{N}_m, \mathbf{M}_m, K_m, V_m)$ is sufficient for $\mathbf{X}_m$. See Appendix A.1 for a detailed description of the distribution of $(\mathbf{N}_m, \mathbf{M}_m, K_m, V_m)$. Now, consider the random variable

$$R_{n,m} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m} > 0\}}. \tag{2.5}$$

Precisely, $R_{n,m}$ denotes the number of distinct types in the sample $\mathbf{X}_m$ that coincide with the atoms Z_i^*'s. In the next theorem we derive the distribution of $R_{n,m}$, and we introduce a distributional identity between $D_m(t)$ and a suitable randomization of $R_{n,m}$ with respect to $\{D(t) : t \ge 0\}$. See Appendix A.1 for the proof.

Theorem 2.1. *For any $m \ge 1$ let $\mathbf{X}_m$ be a sample from a Dirichlet process with atomic base measure $\theta \nu_0 + \sum_{1 \le i \le n} \delta_{Z_i^*}$, for $n \ge 0$. Then, for $x = 0, \dots, \min(n, m)$*

$$\mathbb{P}[R_{n,m} = x] = x! \frac{\binom{n}{x}\binom{m}{x}(\theta + x)_{(m-x)}}{(\theta + n)_{(m)}}. \tag{2.6}$$

Furthermore,

$$D_m(t) \overset{\mathrm{d}}{=} R_{D(t),m} \tag{2.7}$$

for each $t > 0$, where $\{D(t) : t \ge 0\}$ is the death process with marginal distribution (2.1).

The distributional identity (2.7) introduces a Bayesian nonparametric interpretation on the sampling ancestral process $\{D_m(t) : t \ge 0\}$, in the sense that it establishes an interplay between $D_m(t)$ and the sampling from a Dirichlet process prior with atomic base measure $\theta \nu_0 + \sum_{1 \le i \le n} \delta_{Z_i^*}$. Intuitively, the identity (2.7) can be explained by taking the view of the forward looking description of the evolution of $\{D_m(t) : t \ge 0\}$. Starting at time 0 with an infinite number of individuals sampled (at random) from a non-atomic probability measure, the number of non-mutant lineages that survive at time $t > 0$ is described by the ancestral process $\{D(t) : t \ge 0\}$. Now, consider a random sample $\mathbf{Y}_m$ of m individuals at time 0, that is a random sample from a non-atomic probability measure which allows us to distinguish individuals. Then the sampling ancestral process $\{D_m(t) : t \ge 0\}$ describes the number of non-mutant lineages surviving at time t, for any $t > 0$. The transition probability function (2.2) then says that conditionally on $D(t) = n$, the genetic types of the descendants (including mutants) of the m individuals at time t correspond to a sample from a Dirichlet process prior with atomic base measure $\theta \nu_0 + \sum_{1 \le i \le n} \delta_{Z_i^*}$. That the population size remains constant comes from the Wright–Fisher approximation. Given $D(t) = n$, the genetic types of the $D_m(t)$ ancestors must have types that belong to $\{Z_1^*, \dots, Z_n^*\}$, the distinct types of the $D(t) = n$ ancestors. The distribution of $D_m(t)$ is independent of the exact distribution of $Z_1^*, \dots, Z_n^*$ as long as they are distinct. Thus the conditional distribution of $D_m(t)$ given $D(t) = n$ is simply the distribution of $R_{n,m}$.

We now present a novel refinement of the ancestral distribution (2.3). Such a refinement takes into account the lines of descent frequencies at time $t > 0$ of lines beginning at individual roots at time 0 and surviving to time t; these frequencies do not include new mutants. Among the $D_m(t)$ non-mutant lineages surviving from time 0 to time

t, we denote by $D_{l,m}(t)$ the number of non-mutant lineages at time t in the past having frequency l at time 0, for any $l = 1, \dots, m$. We are interested in the distribution on $D_{l,m}(t)$. Let $\mathbf{X}_m$ be the usual random sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n} \delta_{Z_i^*}$, and define

$$R_{l,n,m} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m}=l\}}.$$

Precisely, $R_{l,n,m}$ denotes the number of distinct types in $\mathbf{X}_m$ that coincide with the atoms Z_i^*'s and have frequency l. From the discussion above it is clear that, given $D(t) = n$, $D_{l,m}(t)$ has the same distribution as $R_{l,n,m}$. Thus we obtain that

$$D_{l,m}(t) \overset{\mathrm{d}}{=} R_{l,D(t),m}, \tag{2.8}$$

and the marginal distribution of $\{D(t) : t \ge 0\}$ is given by(2.1). Note that the random variable $D_{l,m}(t)$ represents a natural refinement of $D_m(t)$ in the sense that

$$D_m(t) = \sum_{l=1}^{m} D_{l,m}(t).$$

We stress the fact that $\sum_{l=1}^{m} l D_{l,m}(t)$ may be different from m, since frequency counts do not include new mutants. Although a large amount of literature has been devoted to the study of distributional properties of $D_m(t)$, to the best of our knowledge $D_{l,m}(t)$ has never been investigated, and not even introduced, before. In the next theorem we derive the distribution of $R_{l,n,m}$. See Appendix A.1 for the proof.

Theorem 2.2. *For any $m \ge 1$ let $\mathbf{X}_m$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n} \delta_{Z_i^*}$, for $n \ge 0$. Then, for $x = 0, \dots, \min(n, \lfloor m/l \rfloor)$*

$$\mathbb{P}[R_{l,n,m} = x] = \frac{m!}{(\theta+n)_{(m)}} \sum_{i=x}^{\min(n,\lfloor m/l\rfloor)} (-1)^{i-x} \frac{\binom{i}{x}\binom{n}{i}(\theta+n-i)_{(m-il)}}{(m-il)!}, \tag{2.9}$$

where $\min(n, \lfloor m/l \rfloor)$ denotes the minimum between n and the integer part of m/l.

According to the distributional identity (2.8), the distribution of $D_{l,m}(t)$ follows by combining the ancestral process (2.1) with the distribution (2.9), for any $l = 1, \dots, m$. As a representative example of the distribution of $D_{l,m}(t)$, we consider the case $l = 1$. The distribution of $D_{1,m}(t)$ is of special interest because it corresponds to the sampling ancestral distribution of "rare" non-mutant lineages with frequency 1, i.e. non-mutant lineages composed by a unique individual. If we apply the distribution (2.1) to randomize n in (2.9) with $l = 1$, then

$$\begin{aligned}\mathbb{P}\bigl[D_{1,m}(t) = x\bigr] &\\ = \sum_{j=x}^{m} (-1)^{j-x} \binom{j}{x}\binom{m}{j} &\\ \times \sum_{i=j}^{m} \rho_i(t)(-1)^i \frac{\binom{m-j}{i-j}(\theta+i-j)_{(m-i)}(1-i-j)_{(i-j)}}{(\theta+i+j-1)_{(m-i+1)}(1-\theta+m-i)_{(i-j)}} &\end{aligned} \tag{2.10}$$

for any $x = 0, \dots, m$ and each $t > 0$. The study of finite and asymptotic properties of $D_{l,m}(t)$ is out of the scope of the present paper, and it is deferred to future work. In the rest of this section we introduce a Bayesian nonparametric predictive approach to ancestral inference under the prior assumption that the composition of the population evolves in time according to (2.2). In particular, we consider a sample $\mathbf{Y}_{m+m'}(t) = (Y_1(t), \dots, Y_m(t), Y_{m+1}(t), \dots, Y_{m+m'}(t))$ from such a population at time $t > 0$ and we make use of (2.7) and (2.3) to determine the conditional, or posterior, distribution of $D_{m+m'}(t)$ given $C_{\mathbf{Y}_m(t)}$. A natural refinement of this conditional distribution is also obtained by means of the identity (2.8).

2.2. *Ancestral conditional distributions*

Let $\mathbf{X}_m$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$ and, for any $m'\ge 0$, let $\mathbf{X}_{m'} = (X_{m+1},\dots,X_{m+m'})$ be an additional sample. More precisely $\mathbf{X}_{m'}$ may be viewed as a sample from the conditional distribution of the Dirichlet process with base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$, given the initial sample $\mathbf{X}_m$. We refer to Appendix A.2 for a detailed description of the composition of $\mathbf{X}_{m'}$. We denote by $M_{j,m'} = \sum_{1\le i\le m'} \mathbb{1}_{\{Z_j^*\}}(X_{m+i})$ the number of X_{m+i}'s that coincide with the atom Z_j^*, and we introduce the random variable

$$R_{n,m+m'} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m}+M_{i,m'}>0\}}, \tag{2.11}$$

which denotes the number of distinct types in the enlarged sample $\mathbf{X}_{m+m'} = \{\mathbf{X}_m, \mathbf{X}_{m'}\}$ that coincide with the atoms Z_i^*'s. Observe that if we set $m'=0$ then $R_{n,m+m'}$ reduces to $R_{n,m}$ in (2.5). We also introduce the following random variable

$$\tilde{R}_{l,n,m'} = \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m'}>0\}}\mathbb{1}_{\{M_{i,m}=l\}}, \tag{2.12}$$

which is the number of distinct types in the additional sample $\mathbf{X}_{m'}$ that coincide with the atoms Z_i^*'s that have frequency l in the samples $\mathbf{X}_m$. In the next theorem we derive the conditional distribution of $R_{n,m+m'}$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m)$, and the conditional distribution of $\tilde{R}_{l,n,m'}$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m)$. Interestingly, it turns out that such conditional distributions depend on $(\mathbf{N}_m, \mathbf{M}_m, K_m)$ solely through the statistics $R_{n,m}$ and $R_{l,n,m}$, respectively. See Appendix A.2 for the proof.

Theorem 2.3. *For any $m\ge 1$ and $m'\ge 0$ let $\mathbf{X}_{m+m'}$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$, for $n\ge 0$. Then one has*

(i) *for $x = y,\dots,\min(n, y+m')$*

$$\begin{aligned}
&\mathbb{P}[R_{n,m+m'} = x \mid \mathbf{N}_m = \mathbf{n}_m, \mathbf{M}_m = \mathbf{m}_m, K_m = k_m]\\
&\quad = \mathbb{P}[R_{n,m+m'} = x \mid R_{n,m} = y]\\
&\quad = (x-y)!\frac{\binom{n-y}{x-y}\binom{m'}{x-y}(\theta+m+x)_{(m'-x+y)}}{(\theta+n+m)_{(m')}};
\end{aligned} \tag{2.13}$$

(ii) *for $x = 0,\dots,\min(y, m')$*

$$\begin{aligned}
&\mathbb{P}[\tilde{R}_{l,n,m'} = x \mid \mathbf{N}_m = \mathbf{n}_m, \mathbf{M}_m = \mathbf{m}_m, K_m = k_m]\\
&\quad = \mathbb{P}[\tilde{R}_{l,n,m'} = x \mid R_{l,n,m} = y]\\
&\quad = \frac{\binom{y}{x}}{(\theta+n+m)_{(m')}}\\
&\qquad \times \sum_{i=y-x}^{y} (-1)^{i-(y-x)}\binom{x}{y-i}\big(\theta+n+m-i(1+l)\big)_{(m')}.
\end{aligned} \tag{2.14}$$

Therefore, $R_{n,m}$ and $R_{l,n,m}$ are sufficient to predict $R_{n,m+m'}$ and $\tilde{R}_{l,n,m'}$, respectively.

The predictive sufficiency of $R_{n,m}$ in (2.13) plays a fundamental role for deriving the conditional counterpart of the sampling ancestral distribution (2.3). In particular, consider a population whose composition evolves in time according to (2.2), and let $\mathbf{Y}_m(t)$ be a sample of m individuals from the population at time t. Furthermore, for any $m'>1$ let

$\mathbf{Y}_{m'}(t) = (Y_{m+1}(t), \dots, Y_{m+m'}(t))$ be an additional unobservable sample. The identity (2.7) and the sufficiency of $R_{n,m}$ imply that the conditional distribution of $D_{m+m'}(t)$ given $D_m(t)$ can be obtained by randomizing the parameter n in (2.13) with respect to the distribution

$$\mathbb{P}\big[D(t) = n \mid D_m(t) = y\big] = \frac{\mathbb{P}[R_{n,m} = y]\mathbb{P}[D(t) = n]}{\mathbb{P}[D_m(t) = y]}, \tag{2.15}$$

where $\mathbb{P}[D_m(t) = y] = \sum_{n \geq 0} \mathbb{P}[R_{n,m} = y]\mathbb{P}[D(t) = n]$, and the distributions of $R_{n,m}$, $D_m(t)$ and $D(t)$ are in (2.6), (2.3) and (2.1), respectively. Then, we can write

$$\begin{aligned}
&\mathbb{P}\big[D_{m+m'}(t) = x \mid C_{\mathbf{Y}_m(t)}\big] \\
&\quad = \mathbb{P}\big[D_{m+m'}(t) = x \mid D_m(t) = y\big] \\
&\quad = \frac{\binom{m}{y}\binom{m'}{x-y}(\theta + y)_{(x-y)}(m + m' + \theta)_{(y)}\mathbb{P}[D_{m+m'}(t) = x]}{\binom{m'+m}{x}(\theta + m)_{(x)}\mathbb{P}[D_m(t) = y]}
\end{aligned} \tag{2.16}$$

for any $x = 0, \dots, m + m'$ and each $t > 0$. Note that the probability (2.16) with $m = 0$ reduces to the unconditional ancestral distribution (2.7). Moments of (2.16) are obtained by randomizing n, with respect to (2.15), in the corresponding moments of (2.13) given in (A.25). Equation (2.16) introduces a novel sampling ancestral distribution under the Kingman coalescent. Observe that, due to the identity (2.4), the distribution (2.16) leads to the conditional distribution of the time T_r until there are r non-mutant lineages left in the $(m + m')$-coalescent.

We now consider a refinement of (2.16) which takes into account the frequency counts of non-mutant lineages. Specifically, we determine the conditional distribution of the number $\tilde{D}_{l,m'}(t)$ of non-mutant lineages surviving from time 0 to time t in $\mathbf{Y}_{m'}(t)$ whose frequency in the lineages ancestral to the initial sample $\mathbf{Y}_m$ is l. As a representative example we focus on $l = 1$. Due to (2.8) and the sufficiency of $R_{l,n,m}$ to predict $\tilde{R}_{l,n,m'}$, the conditional distribution of $\tilde{D}_{1,m'}(t)$ given $D_{1,m}(t)$ is obtained by randomizing the parameter n in (2.14) with respect to the distribution

$$\mathbb{P}\big[D(t) = n \mid D_{1,m}(t) = y\big] = \frac{\mathbb{P}[R_{1,n,m} = y]\mathbb{P}[D(t) = n]}{\mathbb{P}[D_{1,m}(t) = y]}, \tag{2.17}$$

because $\mathbb{P}[D_{1,m}(t) = y] = \sum_{n \geq 0} \mathbb{P}[R_{1,n,m} = y]\mathbb{P}[D(t) = n]$, and the distributions of $R_{1,n,m}$, $D_{1,m}(t)$ and $D(t)$ are in (2.9), (2.10) and (2.1), respectively. Then, we have

$$\begin{aligned}
&\mathbb{P}\big[\tilde{D}_{1,m'}(t) = x \mid C_{\mathbf{Y}_m(t)}\big] \\
&\quad = \mathbb{P}\big[\tilde{D}_{1,m'}(t) = x \mid D_{1,m}(t) = y\big] \\
&\quad = (-1)^x \frac{\binom{y}{x}}{\mathbb{P}[D_{1,m}(t) = y]} \sum_{k=0}^{y} (-1)^{y-k} \binom{x}{y-k} \\
&\qquad \times \sum_{j=y}^{m} (-1)^{j-y} \binom{j}{y} \binom{m}{j} \\
&\qquad \times \sum_{i=j}^{m} \frac{\rho_i(t)}{(i-j)!} \sum_{n=j}^{i} (-1)^n \frac{\binom{i-j}{i-n}(\theta + n - j)_{(m-j)}(\theta + n + m - 2k)_{(m')}}{(\theta + n + i - 1)_{(m+m'-i+1)}},
\end{aligned} \tag{2.18}$$

for any $x = 0, \dots, m'$ and each $t > 0$. Moments of (2.18) are obtained by randomizing the parameter n, with respect to (2.17), in the corresponding moments of (2.14) given in (A.28). We stress that the sufficiency of $R_{l,n,m}$ to predict $\tilde{R}_{l,n,m'}$ plays a fundamental role for determining the conditional distribution of $\tilde{D}_{l,m'}(t)$.

If we interpret the FV transition probability function (2.2) as a prior distribution on the evolution in time of the composition of the population, then the conditional distributions (2.16) and (2.18) take on a natural Bayesian nonparametric meaning. Specifically, they correspond to the posterior distributions of $D_{m+m'}(t)$ and $\tilde{D}_{1,m}(t)$, respectively, given the initial sample $\mathbf{Y}_m(t)$ whose ancestry $C_{\mathbf{Y}_m(t)}$ features $D_m(t)$ non-mutant lineages of which $D_{1,m}(t)$ are of frequency 1. Given the information on $D_m(t)$ and $D_{1,m}(t)$ from the initial observed sample, the expected values of (2.16) and (2.18) provide us with Bayesian nonparametric estimators, under a squared loss function, of $D_{m+m'}(t)$ and $\tilde{D}_{1,m}(t)$. It is worth pointing out that $D_m(t)$ and $D_{1,m}(t)$, and in general the m-coalescent $\{C_{\mathbf{Y}_m(t)} : t \geq 0\}$, are latent quantities, in the sense that they are not directly observable from the data. However, one can easily infer $D_m(t)$ and $D_{1,m}(t)$, as well as the mutation parameter θ, from the observed data and then combine their estimates with the posterior distributions (2.16) and (2.18). This approach for making predictive ancestral inference will be detailed in Section 3. We conclude this section with a proposition that introduces an interesting special case of the posterior distributions (2.16) and (2.18). See Appendix A.2 for the proof. Let

$$\tilde{D}_{m'}(t) = D_{m+m'}(t) - D_m(t)$$

be the number of new non mutant lineages, that is $\tilde{D}_{m'}(t)$ is the number non-mutant lineages at t back in the additional sample of size m' which do not coincide with any of the non-mutant lineages at time t back in the initial sample of size m.

Proposition 2.1. *Consider a population whose composition evolves in time according to the FV transition probability function* (2.2). *Then for each* $t > 0$ *one has*

$$\begin{aligned}
&\mathbb{P}\big[\tilde{D}_1(t) = 1 \mid C_{\mathbf{Y}_m(t)}\big] \\
&\quad = \mathbb{P}\big[\tilde{D}_1(t) = 1 \mid D_m(t) = y\big] \\
&\quad = \frac{(y+1)(\theta+y)\mathbb{P}[D_{m+1}(t) = y+1]}{(m+1)(\theta+m)\mathbb{P}[D_m(t) = y]}
\end{aligned} \tag{2.19}$$

and

$$\begin{aligned}
&\mathbb{P}\big[\tilde{D}_{1,1}(t) = 1 \mid C_{\mathbf{Y}_m(t)}\big] \\
&\quad = \mathbb{P}\big[\tilde{D}_{1,1}(t) = 1 \mid D_{1,m}(t) = y\big] \\
&\quad = \frac{y}{\mathbb{P}[D_{1,m}(t) = y]} \\
&\qquad \times \sum_{j=y}^{m} (-1)^{j-y} \binom{j}{y} \binom{m}{j} \\
&\qquad \times \sum_{i=j}^{m} \frac{\rho_i(t)}{(i-j)!} \sum_{n=j}^{i} (-1)^n \frac{\binom{i-j}{i-n} (\theta+n-j)_{(m-j)}}{(\theta+n+i-1)_{(m+1-i+1)}}.
\end{aligned} \tag{2.20}$$

Proposition 2.1 introduces two Bayesian nonparametric estimators for the probability of discovering non-mutant lineages surviving from time 0 to time $t > 0$. This proposition makes explicit the link between our results and the work of Good [11], where the celebrated Good–Turing estimator has been introduced. Given a sample of size m from a population of individuals belonging to an (ideally) infinite number of species with unknown proportions, the Good–Turing estimator provides with an estimate of the probability of discovering at the $(m+1)$th draw a species observed with frequency l in the initial sample. Of course $l = 0$ corresponds to the case of the probability of discovering a new species at the $(m+1)$th draw. Within our framework for ancestral inference under the FV prior assumption (2.2), the probabilities (2.19) and (2.20) may be considered as natural Bayesian nonparametric counterparts of the celebrated Good–Turing estimators. Precisely: (2.19) is the Bayesian nonparametric estimator of the probability of discovery in one additional sample a new non-mutant lineage surviving from time 0 to time $t > 0$; (2.20) is the Bayesian nonparametric estimator of the probability of discovery in one additional sample a non-mutant lineages surviving from time 0 to time $t > 0$ and whose frequency is 1 in the initial sample.

3. Illustration

In this section we show how to use the results of the previous section by applying them to real genetic dataset. Consider a population whose composition evolves in time according to the FV transition probability function (2.2), and suppose we observe a sample of m individuals $\mathbf{Y}_m$ taken from a Dirichlet process with base measure $\theta\nu_0$. Recall that, under this assumption on the evolution of the population, the law of the Dirichlet process with base measure $\theta\nu_0$ is the unique stationary distribution of the neutral FV process. The sample then consists of a collection of $K_m = k \le m$ distinct genetic types with corresponding frequencies $(N_1, \ldots, N_k) = (n_1, \ldots, n_k)$. In particular if $p^{(m)}(n_1, \ldots, n_k)$ denotes the probability of a sample $\mathbf{Y}_m$, which features k genetic types with frequencies $(n_1, \ldots, n_k)$, then

$$p^{(m)}(n_1, \ldots, n_k) = \frac{\theta^k}{(\theta)_m}\prod_{i=1}^{k}(n_i - 1)!; \tag{3.1}$$

see Ewens [8] for details. With a slight abuse of notation, we denote by $X \mid Y$ a random variable whose distribution coincides with the conditional distribution of X given Y. As we pointed out at the end of Section 2, in order to apply the posterior distributions (2.16) and (2.18) we have to estimate the unobservable quantities $(\theta, D_m(t))$ and $(\theta, D_{1,m}(t))$, respectively. Using a fully Bayesian approach, estimates of $(\theta, D_m(t))$ and $(\theta, D_{1,m}(t))$ are obtained as the expected values of the posterior distributions of $(\theta, D_m(t))$ and $(\theta, D_{1,m}(t))$ given $\mathbf{Y}_m$, with respect to some prior choice for θ. For simplicity we focus on the posterior distributions of $D_m(t) \mid \mathbf{Y}_m$ and $D_{1,m}(t) \mid \mathbf{Y}_m$, and we resort to an empirical Bayes approach for estimating θ. Specifically, we use the maximum likelihood estimate for θ originally proposed by Ewens [8], which is obtained from the likelihood (3.1) by numerically finding the root of a certain polynomial in θ. From the point of view of ancestral inference, the distributions of $D_m(t) \mid \mathbf{Y}_m$ and $D_{l,m}(t) \mid \mathbf{Y}_m$ correspond respectively to the questions: How many non-mutant genetic ancestors to the sample existed a time t ago? And how many non-mutant genetic ancestors existed whose type appeared with frequency l among those ancestors?

First we consider the posterior distribution of $D_m(t) \mid \mathbf{Y}_m$. Under the Kingman coalescent model in which mutation is parent-independent, the distribution of this random variable is straightforward: indeed it is well known that the distribution of $D_m(t) \mid \mathbf{Y}_m$ coincides with the distribution of $D_m(t)$, for any $t > 0$. This holds because the coalescent process for a sample of size m can be decomposed into its ancestral process $\{D_m(t) : t \ge 0\}$, and a skeleton chain taking values in marked partitions of the set $\{1, \ldots, m\}$. See Watterson [35] and references therein for details. These two processes are independent, and the sample $\mathbf{Y}_m$ is informative about only the skeleton chain. Thus, the distribution of $D_m(t) \mid \mathbf{Y}_m$ is given by (2.3). In particular if we denote by $\hat{\theta}$ the maximum likelihood estimate of θ, then an estimate of $D_m(t)$ is given by the following expression,

$$\hat{D}_m(t) = \sum_{i=1}^{m} \rho_i(t)(-1)^i i! \frac{\binom{m}{i}}{(\hat{\theta} + m)_{(i)}},$$

which is the expected value of the distribution (2.3) with θ replaced by its estimate $\hat{\theta}$. Thus, we can plug in the estimate $(\hat{\theta}, \hat{D}_m(t))$ to the posterior distribution (2.16) and then predict the number of non-mutant lineages surviving from time 0 to time t in the enlarged sample of size $m + m'$, given the initial observed sample of size m. Observe that under parent-independent mutation the information of the initial sample $\mathbf{Y}_m$ affects the prediction only through the estimate $\hat{\theta}$.

The posterior distribution of $D_{l,m}(t) \mid \mathbf{Y}_m$ is not trivial. Differently from the distribution of $D_m(t) \mid \mathbf{Y}_m$, which is independent of $\mathbf{Y}_m$, the sample $\mathbf{Y}_m$ is informative for $D_{l,m}(t)$. In order to derive the distribution of $D_{l,m}(t) \mid \mathbf{Y}_m$, one strategy would be to study a posterior analogue of the marked-partition-valued process introduced in the work by Watterson [35], and then project it onto its block sizes. The resulting formulas are, however, unwieldy. Our preferred approach is via Monte Carlo simulation, since the posterior transition rates of $D_{l,m}(t) \mid \mathbf{Y}_m$ evolving backwards in time are easy to describe. In particular, if we set

$$\mathbf{D}_{\cdot,m} = \big\{(D_{1,m}, D_{2,m}, \ldots, D_{m,m})(t) : t \ge 0\big\}$$

then the transition rate matrix for $\mathbf{D}_{\cdot,m}$ when currently $\sum_{1\le l\le m} l D_{l,m}(t) = x$ is

$$q_{\mathbf{D}_{\cdot,m},\mathbf{D}'_{\cdot,m}} = \frac{x(x+\theta-1)}{2} \times \begin{cases} \frac{lD_{l,m}(t)}{x} & \text{if } \mathbf{D}'_{\cdot,m}(t) = (D_{1,m},\ldots,D_{l-1,m}+1, D_{l,m}-1,\ldots,D_{m,m})(t), l=2,\ldots,m, \\ \frac{D_{1,m}(t)}{x} & \text{if } \mathbf{D}'_{\cdot,m}(t) = (D_{1,m}-1,\ldots,D_{m,m})(t), \\ -1 & \text{if } \mathbf{D}'_{\cdot,m}(t) = \mathbf{D}_{\cdot,m}(t), \\ 0 & \text{otherwise,} \end{cases}$$

with initial condition

$$\big(D_{l,m}(0) \mid \mathbf{Y}_m\big) = \left|\left\{ y \in \mathcal{X} : \sum_{i=1}^{m} \mathbb{1}_{\{y\}}(Y_i) = l \right\}\right|.$$

See Hoppe [21] for details on $q_{\mathbf{D}_{\cdot,m},\mathbf{D}'_{\cdot,m}}$. In words, lineages are lost at rate $x(x+\theta-1)/2$. At such an event, the lineage selected to be lost is chosen uniformly at random. If that lineage contributed to $D_{1,m}(t)$ then we recognise this loss as having been caused by mutation, otherwise its loss was due to coalescence. The stochastic process $\mathbf{D}_{\cdot,m}$ can be regarded as a time-evolving counterpart to the allelic random partition introduced by Ewens [8], whose stationary sampling distribution is the Ewens-sampling formula, and is clearly straightforward to simulate.

We now present a numerical illustration of our approach. We reconsider the electrophoretic dataset in Table 1 of Singh et al. [30], who sampled $m = 146$ family lines of the fruit fly *Drosophila pseudoobscura* at the xanthine dehydrogenase locus. This organism is well studied in evolutionary biology, and it is especially used to address questions on the nature of speciation. It is thought to have diverged from its sister species *Drosophila persimilis* about 589,000 years ago. See Hey and Nielsen [19] and references therein for details. It is therefore important to quantify relative levels of genetic diversity either shared between the two species or private to one of them. One might ask how much of the genetic diversity observed by Singh et al. [30] existed intact at the time t_{div} the two species diverged. In other words, what is the distribution of $(\mathbf{D}_{\cdot,m}(t_{\text{div}}) \mid \mathbf{Y}_m)$? The process commences back in time from an initial allelic partition inferred from the sample $\mathbf{Y}_m$:

$$\mathbf{D}_{\cdot,146}(0) = (10, 3, 7, 0, 2, 2, 0, 1, 0, 0, 1, 0, \ldots, 0, 1, 0, \ldots, 0),$$

where the most common allele (the rightmost 1) has multiplicity equal to 68. The reader is referred to the work of Singh et al. [30] for details on these data. Using a different dataset, Hey and Nielsen [19] estimated $t_{\text{div}} = 0.34$ in units of $2N_e$ generations, where N_e is the diploid effective population size (these units are appropriate when appealing to the coalescent timescale). The use of the maximum likelihood estimate $\hat{\theta} = 9.5$, and the application of the simulation process described above, result in a Monte Carlo sample of $\mathbf{D}_{\cdot,m}(t_{\text{div}}) \mid \mathbf{Y}_m$ which is summarized in Figure 1. In particular, from Figure 1, posterior means are $\hat{D}_m(t_{\text{div}}) = 2.31$, with narrowest 95% credible interval $[0, 4]$; and $\hat{D}_{1,m}(t_{\text{div}}) = 1.58$, with a narrowest 95% credible interval $[0, 3]$. In other words, with high probability almost all genetic variability, as summarised by the total number of lineages $D_m(t)$ and the total number of singleton lineages $D_{1,m}(t)$, is lost as far back as t_{div}.

As discussed above, Equation (2.16) and Equation (2.18) provide us with a quick predictive distribution for the following question: if we take an additional sample of size m', how much additional genetic variability in the historical population that existed at the divergence time is uncovered? This question is informative because it provides a window into levels of diversity in an unobservable historical population with respect to alleles existing in the modern day. This in turn governs the levels of divergence that we might expect between the two modern species. Equation (2.16) provides a distribution on the total number of non-mutant lineages ancestral to the enlarged sample given $D_m(t)$ lineages ancestral to the original sample, while Equation (2.18) provides a distribution on the number of singleton (frequency 1) lineages ancestral to the original sample that are also discovered in the additional sample. If we plug the (rounded) posterior means $\hat{D}_m(t_{\text{div}}) = 2$ and $\hat{D}_{1,m}(t_{\text{div}}) = 2$ to (2.16) and (2.18) respectively, along with the

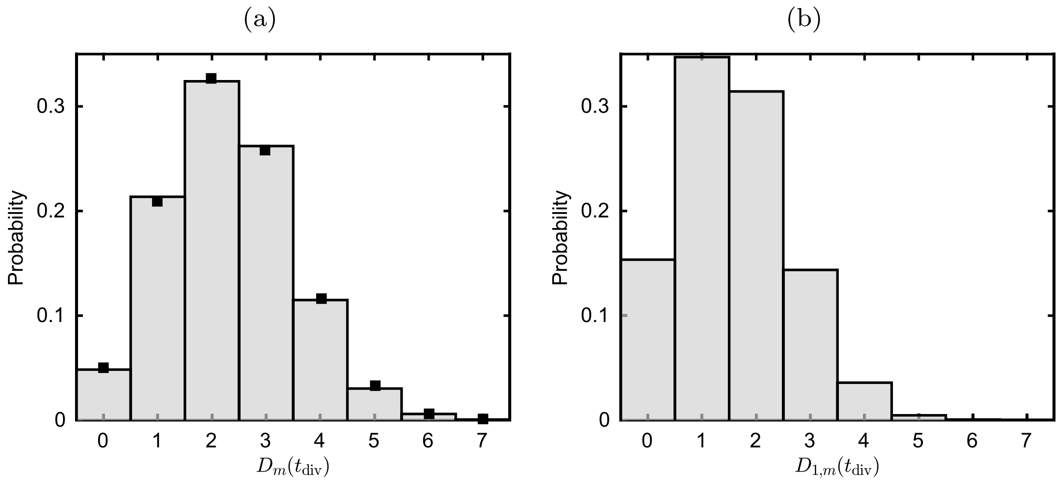

Fig. 1. An approximation of $\mathbf{D}_{\cdot,m}(t_{\text{div}}) \mid \mathbf{Y}_m$ using the data of Singh et al. [30] and 10^4 Monte Carlo replicates, summarized by (a) $D_m(t_{\text{div}}) \mid \mathbf{Y}_m$ and (b) $D_{1,m}(t_{\text{div}}) \mid \mathbf{Y}_m$. Also shown are the predictions from equation (2.3) (black squares).

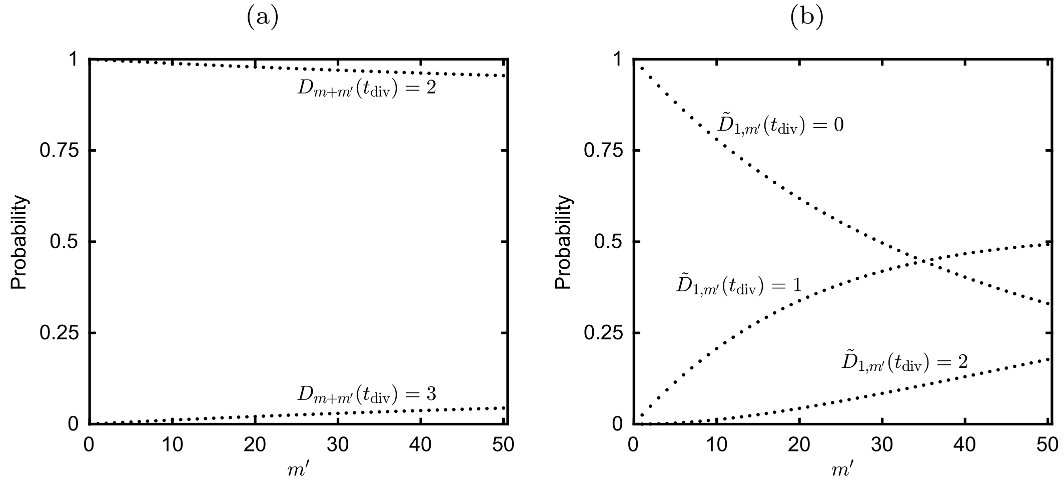

Fig. 2. (a) The probability $\mathbb{P}[D_{m+m'}(t_{\text{div}}) = x \mid D_m(t_{\text{div}}) = 2]$ that there are x non-mutant lineages ancestral to an enlarged sample of size $m + m'$, given that there were two lineages ancestral to the original sample. (b) The probability $\mathbb{P}[\tilde{D}_{1,m'}(t_{\text{div}}) = x \mid D_{1,m}(t_{\text{div}}) = 2]$ that x of the two singleton lineages ancestral to an original sample of size m are also ancestral to some members of the additional sample of size m'.

maximum likelihood estimate $\hat{\theta} = 9.48$, then we obtain the predictive distributions shown in Figure 2. It is clear that, if we regard increasing the initial sample size by m' as a method of "ancestral lineage discovery", then this method is rather inefficient. With high probability, the total number of ancestral lineages is still two, and at most increases to three, even if the sample size is increased by 50. Figure 2(b) shows that at least some of this inefficiency is due to the fact that the two singleton alleles ancestral to the original sample are also ancestral to members of the additional sample; it is moderately easy for these alleles to be rediscovered in the additional sample, at least for sufficiently large m'. Note that these observations are not surprising since the additional lineages coalesce rapidly with each other as we go back in time. Because of shared ancestry, taking additional samples in a coalescent framework is far less informative than the random sampling typically possible in other statistical models.

It is worth remarking that the idea of estimating the allelic configuration of an unobservable historical population, as described above, has broader utility. Very sophisticated models have been formulated in population genetics, encompassing a variety of phenomena we have ignored in this paper: nucleotide-level mutation, changes in historical population size, population substructure, and so on. It turns out that inference under these models can be phrased in terms of the predictive, or conditional sampling, distributions associated with an additional sample of size m', which in turn depend on the genetic types of lineages ancestral to the original sample. See Stephens and Donnelly [32] for a detailed account. Under these more sophisticated models, such predictive distributions are usually intractable, but Stephens and Donnelly [32] showed that if they can be approximated then exact Monte Carlo based inference is still possible by applying an importance sampling correction to this approximation. Thus, many population genetic inference problems can be reduced to the following: design a decent approximation to the predictive distribution associated with an enlarged sample. See, e.g., De Iorio and Griffiths [4], Griffiths et al. [15], Hobolth et al. [20], Li and Stephens [24], Paul and Song [27], Paul et al. [28], Sheehan et al. [29] and Stephens and Donnelly [32]. Now the tractability of the model studied in this paper becomes crucial: it can be used as a guide for more complex models. Indeed, this is the strategy taken by Stephens and Donnelly [32] and Hobolth et al. [20].

4. Discussion

We introduced a Bayesian nonparametric predictive approach to ancestral inference. This approach relies on the FV transition probability function (2.2) as a nonparametric prior assumption for the evolution in time of the composition of a genetic population. That is, backward in time the Kingman coalescent is assumed to be the prior model for the genealogy of the population. Under this prior assumption, and given a sample of m individuals from the population at time $t > 0$, we showed how to derive the posterior distributions of some quantities related to the genealogy of an additional unobservable sample of size $m' \geq 1$. Our posterior analysis built upon the distributional identity for $D_m(t)$ introduced in Theorem 2.2, which provides a Bayesian nonparametric interpretation of the sampling ancestral process. In particular, we determined the posterior distribution of the number $D_{m+m'}(t)$ of non-mutant lineages surviving from time 0 to time t in the enlarged sample of size $m + m'$ which, in turn, leads to the posterior distribution of the time of the most recent common ancestor. This result has then been extended to the number $\tilde{D}_{l,m'}(t)$ of non-mutant lineages having frequency l, for any $l = 1, \dots, m$, surviving from time 0 to time t. Our results allowed us to introduce a novel class of Bayesian nonparametric estimators which can be thought as Good–Turing estimators in the context of ancestral inference.

This paper paves the way for future work towards predictive ancestral inference under the Kingman coalescent. A first important problem consists in investigating the asymptotic behavior of the statistics introduced in this paper. The asymptotic behaviour of $D_m(t)$ for small time t was first investigated by Griffiths [14]. If $m \to +\infty$ and $t \to 0$ such that mt is constant, then $D_m(t)$ appropriately scaled converges in distribution to a Gaussian random variable. Furthermore, if t goes to 0 faster with m^2t being bounded above, then $m - D_m(t)$ will be approximated in distribution by a Poisson random variable. Besides extending these asymptotic results to $D_{l,m}(t)$, it would be interesting to characterize the large m' asymptotic behavior of the posterior distributions of $D_{m+m'}(t)$ and $\tilde{D}_{l,m'}(t)$. Such a characterization would be useful to obtain large m' approximations of these posterior distributions. Work on this is ongoing. Another problem consists in investigating the posterior distribution of other statistics of the additional sample. Apart from statistics related to the age of alleles, it seems natural to complete our analysis by determining a posterior counterpart of the distribution of $D_{l,m}(t)$. This requires one to study the conditional distribution of $R_{l,n,m+m'} = \sum_{1 \leq i \leq n} \mathbb{1}_{\{M_{i,m}+M_{i,m'}=l\}}$ given $\mathbf{X}_m$. While this conditional distribution can be derived by means of techniques similar to those developed in this paper, we expect that $R_{l,n,m+m'}$ will be a function of $\mathbf{X}_m$ through $R_{n,m}$ and $(R_{1,n,m}, \dots, R_{l,n,m})$. Such unwieldy sufficient statistics could make difficult the randomization of $R_{l,n,m+m'} \mid R_{n,m}, (R_{1,n,m}, \dots, R_{l,n,m})$ over the parameter n.

Kingman's coalescent is a special case of a broader class of so-called Λ-coalescent models, which have been generalized further to the Ξ-coalescents, whose genealogies allow for simultaneous coalescence events each involving possibly more than two lineages. We refer to the monograph by Berestycki [1] for a comprehensive and stimulating account of these generalizations of the Kingman coalescent. In particular Ξ-coalescents arise in genetics as models of diploid populations with high fecundity and highly skewed offspring distributions; that is, in some generations the offspring of a single individual can replace a substantial fraction of the whole population. See Möhle and Sagitov [26] for a biological interpretation of the Ξ-coalescent. Another direction for future work is in Bayesian nonparametric predictive ancestral inference for these broader classes of coalescent models. This seems to be a much more challenging task, as far fewer results are available. For example, Möhle [25] showed that there is no simple analogue of the Ewens-sampling formula except in a few special cases. Furthermore, although there exists a construction of the Ξ-Fleming–Viot process to describe the forwards-in-time evolution of the population, there seems to be no tractable expressions for its transition function as in (2.2), nor even a known stationary distribution in general. See Birkner et al. [2] for details. It may be therefore difficult to give a natural Bayesian nonparametric interpretation of the underlying genealogical process.

Appendix

A.1. *Proofs of Section* 2.1

Let $\mathbf{X}_m$ be a sample from a Dirichlet process with atomic base measure $\theta\nu_0 + \sum_{1\le i\le n}\delta_{Z_i^*}$. Recall from Section 2.1 that we $\{X_1^*,\dots,X_{K_m}^*\}$ the labels identifying the K_m distinct types in $\mathbf{X}_m$ which do not coincide with any of the atoms Z_i^*'s. Moreover, we defined the following quantities: (i) $\mathbf{M}_m=(M_{1,m},\dots,M_{n,m})$ where $M_{j,m}=\sum_{1\le i\le m}\mathbb{1}_{\{Z_j^*\}}(X_i)$ denotes the number of X_i's that coincide with the atom Z_j^*, for any $j=1,\dots,n$; (ii) $\mathbf{N}_m=(N_{1,m},\dots,N_{K_m,m})$ where $N_{j,m}=\sum_{1\le i\le m}\mathbb{1}_{\{X_j^*\}}(X_i)$ denotes the number of X_i's that coincide with the label X_j^*, for any $j=1,\dots,K_m$; (iii) V_m denotes the number of X_i's which do not coincide with any of the labels $\{Z_1^*,\dots,Z_n^*\}$, i.e., $V_m=\sum_{1\le i\le K_m}N_{i,m}$. Note that the sample $\mathbf{X}_m$ may be viewed as a a random sample from a posterior Dirichlet process given the sample $(Z_1^*,\dots,Z^*)$ featuring n distinct types. Since the distribution of a sample of size n from a Dirichlet process (Ewens [8]) featuring J_n distinct types with corresponding frequency $\mathbf{Q}_n=(Q_{1,n},\dots,Q_{J_n,n})$ is

$$\mathbb{P}[J_n=j_n,\mathbf{Q}_n=\mathbf{q}_n]=\frac{1}{j_n!}\binom{n}{q_{1,n},\dots,q_{j_n,n}}\frac{\theta^{j_n}}{(\theta)_{(n)}}\prod_{i=1}^{j_n}(q_{i,n}-1)!,$$

then the distribution of $\mathbf{X}_m$ from a Dirichlet process with atomic base measure $\theta\nu_0+\sum_{1\le i\le n}\delta_{Z_i^*}$ is

$$\begin{aligned}
&\mathbb{P}[\mathbf{N}_m=\mathbf{n}_m,\mathbf{M}_n=\mathbf{m}_m,K_m=k_m,V_m=v_m]\\
&\quad=\frac{\frac{\theta^{n+k_m}}{(\theta)_{(n+m)}}}{\frac{\theta^n}{(\theta)_{(n)}}\prod_{i=1}^n(1-1)!}\\
&\qquad\times\binom{m}{v_m}\binom{m-v_m}{m_{1,m},\dots,m_{n,m}}\prod_{i=1}^n(1+m_{i,m}-1)!\\
&\qquad\times\frac{1}{k_m!}\binom{v_m}{n_{1,m},\dots,n_{k_m,m}}\prod_{i=1}^{k_m}(n_{i,m}-1)!\\
&\quad=\frac{\theta^{k_m}}{(\theta+n)_{(m)}}\binom{m}{v_m}\binom{m-v_m}{m_{1,m},\dots,m_{n,m}}\prod_{i=1}^n m_{i,m}! \qquad\text{(A.1)}\\
&\qquad\times\frac{1}{k_m!}\binom{v_m}{n_{1,m},\dots,n_{k_m,m}}\prod_{i=1}^{k_m}(n_{i,m}-1)!. \qquad\text{(A.2)}
\end{aligned}$$

In addition to the above preliminaries on the distribution of the random sample $\mathbf{X}_m$, for any $n\ge 1$, $x\in\{1,\dots,n\}$ and $1\le\tau_1<\dots<\tau_x\le n$, let $\mathbf{M}_{(\tau_1,\dots,\tau_x),m}=(M_{\tau_1,m},\dots,M_{\tau_x,m})$ be a collection of x components of $\mathbf{M}_m$. We denote by $S(n,k)$ the Stirling number of the second kind, and by $\mathcal{C}_{n,x}$ the set of x-combinations without repetition of $\{1,\dots,n\}$, i.e., $\mathcal{C}_{n,x}=\{(c_1,\dots,c_x): c_k\in\{1,\dots,n\}, c_k\neq c_l, \text{if } k\neq l\}$ for any $x\ge 1$, and $\mathcal{C}_{n,0}=\varnothing$. See Charalambides [3] for details.

Proof of Theorem 2.1. We start by determining the distribution of the random variable $R_{n,m}$, and then we show the distributional identity for $D_m(t)$. In order to compute the distribution of $R_{n,m}$, we start by computing the corresponding

rth descending factorial moments. By the Vandermonde formula, we can write

$$\mathbb{E}\big[(R_{n,m})_{[r]}\big]=\sum_{s=0}^{r}\binom{r}{s}(-1)^s(n-s)_{[r-s]}\mathbb{E}\big[\big(R^*_{n,m}\big)_{[s]}\big], \tag{A.3}$$

where $R^*_{n,m}=\sum_{1\le i\le n}\mathbb{1}_{\{M_{i,m}=0\}}$. A repeated application of the Binomial theorem leads to write the rth moment of the random variable $R^*_{n,m}$ as follows

$$\begin{aligned}
&\mathbb{E}\big[\big(R^*_{n,m}\big)^r \mid V_m=v_m, K_m=k_m\big]\\
&\quad=\sum_{x=1}^{n}\sum_{i_1=1}^{r-1}\sum_{i_2=1}^{i_1-1}\cdots\sum_{i_{x-1}=1}^{i_{x-2}-1}\binom{r}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{x-2}}{i_{x-1}}\\
&\quad\quad\times\sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}}\mathbb{E}\left[\prod_{t=1}^{x}(\mathbb{1}_{\{M_{c_t,m}=0\}})^{i_{x-t}-i_{x-t+1}}\;\middle|\; V_m=v_m, K_m=k_m\right]\\
&\quad=\sum_{x=1}^{r}S(r,x)x!\sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}}\mathbb{P}\big[\mathbf{M}_{\mathbf{c}_x,m}=(\underbrace{0,\dots,0}_{x})\mid V_m=v_m, K_m=k_m\big],
\end{aligned} \tag{A.4}$$

where

$$\mathbb{P}\big[\mathbf{M}_{\mathbf{c}_x,m}=(\underbrace{0,\dots,0}_{x})\mid V_m=v_m, K_m=k_m\big]=\frac{(n-x)_{(m-v_m)}}{(n)_{(m-v_m)}}, \tag{A.5}$$

and

$$\mathbb{P}[V_m=v_m]=\frac{\binom{m}{v_m}}{(\theta+n)_m}(n)_{(m-v_m)}(\theta)_{v_m}. \tag{A.6}$$

Therefore, by combining Equation (A.4) with Equation (A.5) and Equation (A.6) one has

$$\mathbb{E}\big[\big(R^*_{n,m}\big)_{[r]}\big]=r!\binom{n}{r}\frac{(\theta+n-r)_{(m)}}{(\theta+n)_{(m)}},$$

and from (A.3)

$$\begin{aligned}
\mathbb{E}\big[(R_{n,m})_{[r]}\big]&=\sum_{s=0}^{r}\binom{r}{s}(-1)^s(n-s)_{[r-s]}s!\binom{n}{s}\frac{(\theta+n-s)_{(m)}}{(\theta+n)_{(m)}}\\
&=\frac{r!}{(\theta+n)_{(m)}}\sum_{s=0}^{r}\binom{n-s}{r-s}(-1)^s\binom{n}{s}(\theta+n-s)_{(m)}.
\end{aligned} \tag{A.7}$$

The distribution of the random variable $R_{n,m}$ follows from the factorial moments in Equation (A.7). In particular, for any $x=0,\dots,\min(n,\lfloor m\rfloor)$, we can write the following

$$\begin{aligned}
&\mathbb{P}[R_{n,m}=x]\\
&\quad=\sum_{y\ge 0}\frac{(-1)^y}{x!y!}\mathbb{E}\big[(R_{n,m})_{[x+y]}\big]\\
&\quad=\frac{1}{(\theta+n)_{(m)}}\sum_{y\ge x}\frac{1}{x!}(-1)^{y-x}(y)_{[x]}\sum_{s=0}^{y}\binom{n-s}{y-s}(-1)^s\binom{n}{s}(\theta+n-s)_{(m)}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{(\theta+n)_{(m)}} \sum_{s=0}^{n} (-1)^{s} \binom{n}{s} (\theta+n-s)_{(m)} \sum_{y=s}^{n} (-1)^{y-x} \binom{y}{x} \binom{n-s}{y-s} \\
&= \frac{(-1)^{-x}}{(\theta+n)_{(m)}} \sum_{s=0}^{n} (-1)^{n-s} \binom{n}{s} \binom{s}{n-x} (\theta+n-s)_{(m)} \\
&= \frac{(-1)^{n}}{(\theta+n)_{(m)}} \sum_{s=n}^{n+x} (-1)^{s} \binom{n}{s-x} \binom{s-x}{n-x} (\theta+n-s+x)_{(m)} \\
&= \frac{\binom{n}{n-x}}{(\theta+n)_{(m)}} \sum_{s=0}^{x} (-1)^{s} \binom{x}{s} (\theta-s+x)_{(m)} \\
&= x! \binom{n}{x} \binom{m}{x} \frac{(\theta+x)_{(m-x)}}{(\theta+n)_{(m)}},
\end{aligned}
$$

where the last equality arises by an application of the Vandermonde identity. This proves (2.6). As regards the distributional identity (2.7), let us randomize the distribution of $R_{n,m}$ on n with respect to the distribution (2.1). We can write

$$
\begin{aligned}
&\sum_{n \geq x} d_n(t) x! \binom{n}{x} \binom{m}{x} \frac{(\theta+x)_{(m-x)}}{(\theta+n)_{(m)}} \\
&\quad = \sum_{n \geq x} \sum_{i \geq n} \rho_i(t) (-1)^{-n} \binom{i}{n} (\theta+n)_{(i-1)} \frac{x!}{i!} \binom{n}{x} \binom{m}{x} \frac{(\theta+x)_{(m-x)}}{(\theta+n)_{(m)}} \\
&\quad = \sum_{i \geq x} \sum_{n=x}^{i} \rho_i(t) (-1)^{-n} \binom{i}{n} (\theta+n)_{(i-1)} \frac{x!}{i!} \binom{n}{x} \binom{m}{x} \frac{(\theta+x)_{(m-x)}}{(\theta+n)_{(m)}} \\
&\quad = \sum_{i \geq x} \rho_i(t) \frac{1}{i!} \binom{m}{x} x! (\theta+x)_{(m-x)} \sum_{n=x}^{i} \binom{i}{n} (-1)^{n} (\theta+n)_{(i-1)} \frac{\binom{n}{x}}{(\theta+n)_{(m)}}.
\end{aligned}
\tag{A.8}
$$

Let us focus on the second factor appearing in the last expression, namely the term $\sum_{n=x}^{i} \binom{i}{n} (-1)^n (\theta + n)_{(i-1)} \binom{n}{x} / (\theta+n)_{(m)}$. In particular we can rewrite it as

$$
\begin{aligned}
&\frac{i!(-1)^{x}}{(i-x)!x!} \sum_{n=0}^{i-x} (-1)^{n} \binom{i-x}{n} \frac{(\theta+n+x)_{(i-1)}}{(\theta+n+x)_{(m)}} \\
&\quad = \frac{i!(-1)^{x}}{(i-x)!x!(m-i)!} \sum_{n=0}^{i-x} (-1)^{n} \binom{i-x}{n} \int_0^1 y^{\theta+n+x+i-1-1} (1-y)^{m-i+1-1} \mathrm{d}y \\
&\quad = \frac{i!(-1)^{x}}{(i-x)!x!(m-i)!} \int_0^1 y^{\theta+x+i-1-1} (1-y)^{m-x+1-1} \mathrm{d}y \\
&\quad = \frac{i!(-1)^{x}}{(i-x)!x!(m-i)!} \frac{\Gamma(m-x+1)\Gamma(\theta+x+i-1)}{\Gamma(\theta+i+m)}.
\end{aligned}
\tag{A.9}
$$

Finally, by combining the expression (A.8) with the expression (A.9) one obtains what follows

$$
\begin{aligned}
&\sum_{n \geq x} d_n(t) x! \binom{n}{x} \binom{m}{x} \frac{(\theta+x)_{(m-x)}}{(\theta+n)_{(m)}} \\
&\quad = \sum_{i \geq x} \rho_i(t) \frac{1}{i!} \binom{m}{x} x! (\theta+x)_{(m-x)}
\end{aligned}
$$

$$
\begin{aligned}
&\times \frac{i!(-1)^x}{(i-x)!x!(m-i)!} \frac{\Gamma(m-x+1)\Gamma(\theta+x+i-1)\Gamma(\theta+x)\Gamma(\theta+m)}{\Gamma(\theta+i+m)\Gamma(\theta+x)\Gamma(\theta+m)} \\
&= \sum_{i\geq x} \rho_i(t)\frac{1}{i!}\binom{m}{x} x! \frac{i!(-1)^x}{(i-x)!x!(m-i)!}\frac{\Gamma(m-x+1)(\theta+x)_{(i-1)}}{(\theta+m)_{(i)}} \\
&= \sum_{i=x}^{m} \rho_i(t)\binom{m}{i}\frac{i!(-1)^x}{(i-x)!x!}\frac{(\theta+x)_{(i-1)}}{(\theta+m)_{(i)}},
\end{aligned}
$$

which, after some rearrangement of terms, coincides with the sampling ancestral distribution (2.3). This proves the distributional identity (2.7), and the proof is completed. □

Proof of Theorem 2.2. The proof is along line similar to the first part of the proof of Theorem 2.1. We start by determining the rth factorial moments of the random variable $R_{l,n,m}$. In particular, by a repeated application of the Binomial theorem

$$
\begin{aligned}
&\mathbb{E}\big[(R_{l,n,m})^r \mid V_m = v_m, K_m = k_m\big] \\
&= \sum_{x=1}^{n}\sum_{i_1=1}^{r-1}\sum_{i_2=1}^{i_1-1}\cdots\sum_{i_{x-1}=1}^{i_{x-2}-1}\binom{r}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{x-2}}{i_{x-1}} \\
&\quad\times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \mathbb{E}\left[\prod_{t=1}^{x}(\mathbb{1}_{\{M_{c_t,m}=l\}})^{i_{x-t}-i_{x-t+1}} \,\middle|\, V_m = v_m, K_m = k_m\right] \\
&= \sum_{x=1}^{r} S(r,x)x! \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \mathbb{P}\big[\mathbf{M}_{\mathbf{c}_x,m} = (\underbrace{l,\ldots,l}_{x}) \mid V_m = v_m, K_m = k_m\big],
\end{aligned}
\tag{A.10}
$$

where

$$
\mathbb{P}\big[\mathbf{M}_{\mathbf{c}_x,m} = (\underbrace{l,\ldots,l}_{x}) \mid V_m = v_m, K_m = k_m\big] = (xl)!\frac{\binom{m-v_m}{xl}(n-x)_{(m-v_m-xl)}}{(n)_{(m-v_m)}},
\tag{A.11}
$$

and the distribution of the random variable V_m is given in (A.6). Therefore, by a combination of Equation (A.10) with Equation (A.11) and Equation (A.6) one obtains

$$
\mathbb{E}\big[(R_{l,n,m})_{[r]}\big] = m!r!\frac{\binom{n}{r}}{(m-rl)!}\frac{(\theta+n-r)_{(m-rl)}}{(\theta+n)_{(m)}}.
\tag{A.12}
$$

Finally, the distribution of the random variable $R_{l,n,m}$ then follows from the factorial moments (A.12). In particular, for any $x = 0,\ldots,\min(n,\lfloor m/l\rfloor)$, we can write

$$
\begin{aligned}
\mathbb{P}[R_{l,n,m} = x] &= \sum_{y\geq 0}\frac{(-1)^y}{x!y!}\mathbb{E}\big[(R_{l,n,m})_{[x+y]}\big] \\
&= \sum_{y\geq 0}(-1)^y\frac{1}{x!y!}(x+y)!\binom{n}{x+y} \\
&\quad\times \frac{m!}{(m-(x+y)l)!}\frac{(\theta+n-x-y)_{(m-(x+y)l)}}{(\theta+n)_{(m)}}.
\end{aligned}
\tag{A.13}
$$

The expression (A.13) coincides, after some simplification and rearrangement of terms, to (2.9). In particular, the sum over the index y ranges between x and $\min(n,\lfloor m/l\rfloor)$, where $\lfloor m/l\rfloor$ denotes the integer part of m/l. The proof is completed. □

A.2. *Proofs of Section* 2.2

Let $\mathbf{X}_m$ be a random sample from a Dirichlet process with base measure $\theta\nu_0 + \sum_{1\leq i\leq n}\delta_{Z_i^*}$. Recall that $\mathbf{X}_n$ may be viewed as a a random sample from a posterior Dirichlet process given the sample $(Z_1^*,\dots,Z^*)$ featuring n distinct types, and that the distribution of $\mathbf{X}_n$ is given in Equation (A.1). We start by describing the composition of an additional sample $\mathbf{X}_{m'}$, for $m'\geq 0$. Let $\{X^*_{K_m+1},\dots,X^*_{K_m+K_{m'}}\}$ be the labels identifying the $K_{m'}$ distinct types in the sample $\mathbf{X}_{m'}$ which do not coincide with any of $\{Z_1^*,\dots,Z_n^*,X_1^*,\dots,X^*_{K_m}\}$. Moreover, let

(i) $M_{j,m'} = \sum_{1\leq i\leq m'}\mathbb{1}_{\{Z_j^*\}}(X_{m+i})$ be the number of X_{m+i}'s that coincide with Z_j^*, for any $j=1,\dots,n$,
(ii) $N_{j,m'} = \sum_{1\leq i\leq m'}\mathbb{1}_{\{X_j^*\}}(X_{m+i})$ be the number of X_{m+i}'s that coincide with X_j^*, for any $j=1,\dots,K_m+K_{m'}$.

Additionally, let $\mathbf{N}_{m'}=(N_{1,m'},\dots,N_{K_m,m'},N_{K_m+1,m'},\dots,N_{K_m+K_{m'},m'})$ and $\mathbf{M}_{m'}=(M_{1,m'},\dots,M_{n,m'})$. Also, we denote by $V_{m'}$ the number of X_{m+i}'s which do not coincide with any of the labels $\{Z_1^*,\dots,Z_n^*,X_1^*,\dots,X^*_{K_m}\}$, i.e., we can write

$$V_{m'} = \sum_{i=1}^{K_{m'}} N_{K_m+i,m'}.$$

In a similar way, we denote by $W_{m'}$ the number of X_{m+i}'s which do not coincide with any of the labels $\{Z_1^*,\dots,Z_n^*,X^*_{K_m+1},\dots,X^*_{K_m+K_{m'}}\}$, i.e., we can write $W_{m'}$ as

$$W_{m'} = \sum_{i=1}^{K_m} N_{i,m'}.$$

We can write the conditional probability of $(\mathbf{N}_{m'},\mathbf{M}_{m'},V_{m'},W_{m'},K_{m'})$ given $(\mathbf{N}_m,\mathbf{M}_m,K_m)$, where $\mathbf{N}_m$, $\mathbf{M}_m$, and K_m have been defined in Section 2. This may be viewed as the natural conditional (posterior) counterpart of (A.1). In particular, by a direct application of Equation (A.1), we can write the following probability

$$\begin{aligned}
&\mathbb{P}[\mathbf{N}_{m'}=\mathbf{n}_{m'},\mathbf{M}_{m'}=\mathbf{m}_{m'},V_{m'}=v_{m'},W_{m'}=w_{m'},K_{m'}=k_{m'}\mid \mathbf{N}_m=\mathbf{n}_m,\mathbf{M}_m=\mathbf{m}_m,K_m=k_m]\\
&\quad=\frac{\frac{\theta^{n+k_m+k_{m'}}}{(\theta)_{(n+m+m')}}}{\frac{\theta^{n+k_m}}{(\theta)_{(n+m)}}\prod_{i=1}^n(1+m_{i,m}-1)!\prod_{i=1}^{k_m}(n_{i,m}-1)!}\\
&\qquad\times\binom{m'}{v_{m'},w_{m'},m'-v_{m'}-w_{m'}}\\
&\qquad\times\binom{m'-v_{m'}-w_{m'}}{m_{1,m'},\dots,m_{n,m'}}\prod_{i=1}^n(1+m_{i,m}+m_{i,m'}-1)!\\
&\qquad\times\binom{w_{m'}}{n_{1,m'},\dots,n_{k_n,m'}}\prod_{i=1}^{k_m}(n_{i,m}+n_{i,m'}-1)!\\
&\qquad\times\frac{1}{k_{m'}!}\binom{v_{m'}}{n_{k_n+1,m'},\dots,n_{k_n+k_{m'},m'}}\prod_{i=1}^{k_{m'}}(n_{k_m+i,m'}-1)!\\
&\quad=\frac{\frac{\theta^{k_{m'}}}{(\theta+n+m)_{(m')}}}{\prod_{i=1}^n m_{i,m}!\prod_{i=1}^{k_m}(n_{i,m}-1)!}\\
&\qquad\times\binom{m'}{v_{m'},w_{m'},m'-v_{m'}-w_{m'}}
\end{aligned}$$

$$\times \binom{m' - v_{m'} - w_{m'}}{m_{1,m'}, \dots, m_{n,m'}} \prod_{i=1}^{n} (m_{i,m} + m_{i,m'})!$$

$$\times \binom{w_{m'}}{n_{1,m'}, \dots, n_{k_n,m'}} \prod_{i=1}^{k_m} (n_{i,m} + n_{i,m'} - 1)!$$

$$\times \frac{1}{k_{m'}!} \binom{v_{m'}}{n_{k_n+1,m'}, \dots, n_{k_n+k_{m'},m'}} \prod_{i=1}^{k_{m'}} (n_{k_m+i,m'} - 1)!. \tag{A.14}$$

To simplify the notation we define $A_m(\mathbf{n}_m, \mathbf{m}_m, k_m) = \{\mathbf{N}_m = \mathbf{n}_m, \mathbf{M}_m = \mathbf{m}_m, K_m = k_m\}$ and $B_{m'}(v_{m'}, w_{m'}, k_{m'}) = \{V_{m'} = v_{m'}, W_{m'} = w_{m'}, K_{m'} = k_{m'}\}$. Furthermore, with a slight abuse of notation, we denote by $X \mid Y$ a random variable whose distribution coincides with the conditional distribution of X given Y.

Lemma A.1. *For any $n \geq 1$, $x \in \{1, \dots, n\}$ and $1 \leq \tau_1 < \cdots < \tau_x \leq n$, let $\mathbf{M}_{(\tau_1,\dots,\tau_x),m'} = (M_{\tau_1,m'}, \dots, M_{\tau_x,m'})$ be a collection of x components of $\mathbf{M}_{m'}$. Then*

$$\begin{aligned}
&\mathbb{P}\big[\mathbf{M}_{(\tau_1,\dots,\tau_x),m'} = \mathbf{m}_{(\tau_1,\dots,\tau_x),m'} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&\quad = \binom{m' - v_{m'} - w_{m'}}{m_{\tau_1,m'}, \dots, m_{\tau_x,m'}, m' - v_{m'} - w_{m'} - \sum_{i=1}^{x} m_{\tau_i,m'}} \\
&\qquad \times \frac{(n + m - \sum_{i=1}^{k_m} n_{i,m} - \sum_{i=1}^{x}(1 + m_{\tau_i,m}))_{(m' - v_{m'} - w_{m'} - \sum_{i=1}^{x} m_{\tau_i,m'})}}{(n + m - \sum_{i=1}^{k_m} n_{i,m})_{(m' - v_{m'} - w_{m'})}} \\
&\qquad \times \prod_{i=1}^{x} (1 + m_{\tau_i,m})_{(m_{\tau_i,m'})}.
\end{aligned} \tag{A.15}$$

Proof. We start by determining the conditional distribution of the random variable $(V_{m'}, W_{m'}, K_{m'})$ given the sample $\mathbf{X}_m$. This is obtained by suitably marginalizing the distribution (A.14) over $(\mathbf{N}_{m'}, \mathbf{M}_{m'})$. In particular, with this regards, if

$$\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'},n} = \left\{ (m_{i,m'})_{1 \leq i \leq n} : m_{i,m'} \geq 0 \wedge \sum_{i=1}^{n} m_{i,m'} = m' - v_{m'} - w_{m'} \right\},$$

$$\mathcal{S}^{(0)}_{w_{m'},k_m} = \left\{ (n_{i,m'})_{1 \leq i \leq k_m} : n_{i,m'} \geq 0 \wedge \sum_{i=1}^{k_m} n_{i,m'} = w_{m'} \right\},$$

and

$$\mathcal{S}_{v_{m'},k_{m'}} = \left\{ (n_{k_m+i,m'})_{1 \leq i \leq k_{m'}} : n_{k_m+i,m'} \geq 1 \wedge \sum_{i=1}^{k_{m'}} n_{k_n+i,m'} = v_{m'} \right\},$$

then

$$\begin{aligned}
&\mathbb{P}\big[V_{m'} = v_{m'}, W_{m'} = w_{m'}, K_{m'} = k_{m'} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&\quad = \frac{\frac{\theta^{n+k_m+k_{m'}}}{(\theta)_{(n+m+m')}}}{\frac{\theta^{n+k_m}}{(\theta)_{(n+m)}} \prod_{i=1}^{n} m_{i,m}! \prod_{i=1}^{k_m} (n_{i,m} - 1)!} \\
&\qquad \times \binom{m'}{v_{m'}, w_{m'}, m' - v_{m'} - w_{m'}}
\end{aligned}$$

$$\times \sum_{\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'},n}} \binom{m'-v_{m'}-w_{m'}}{m_{1,m'},\dots,m_{n,m'}} \prod_{i=1}^{n}(m_{i,m}+m_{i,m'})!$$

$$\times \sum_{\mathcal{S}^{(0)}_{w_{m'},k_m}} \binom{w_{m'}}{n_{1,m'},\dots,n_{k_m,m'}} \prod_{i=1}^{k_m}(n_{i,m}+n_{i,m'}-1)!$$

$$\times \frac{1}{k_{m'}!} \sum_{\mathcal{S}_{v_{m'},k_{m'}}} \binom{v_{m'}}{n_{k_m+1,m'},\dots,n_{k_m+k_{m'},m'}} \prod_{i=1}^{k_{m'}}(n_{k_m+i,m'}-1)!. \tag{A.16}$$

Now, we apply Vandermonde formula and Theorem 2.5 in Charalambides [3] in order to solve the above summations. In particular, we have the following identities

(i)

$$\sum_{\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'},n}} \binom{m'-v_{m'}-w_{m'}}{m_{1,m'},\dots,m_{n,m'}} \prod_{i=1}^{n}(m_{i,m}+m_{i,m'})!$$

$$= \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \prod_{i=1}^{n} m_{i,m}!, \tag{A.17}$$

(ii)

$$\sum_{\mathcal{S}^{(0)}_{w_{m'},k_m}} \binom{w_{m'}}{n_{1,m'},\dots,n_{k_m,m'}} \prod_{i=1}^{k_m}(n_{i,m}+n_{i,m'}-1)!$$

$$= \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} \prod_{i=1}^{k_m}(n_{i,m}-1)!, \tag{A.18}$$

(iii)

$$\frac{1}{k_{m'}!} \sum_{\mathcal{S}_{v_{m'},k_{m'}}} \binom{v_{m'}}{n_{k_m+1,m'},\dots,n_{k_m+k_{m'},m'}} \prod_{i=1}^{k_{m'}}(n_{k_m+i,m'}-1)!$$

$$= \left|s(v_{m'},k_{m'})\right|, \tag{A.19}$$

where $|s(n,k)|$ denotes the signless Stirling number of the first type (see Charalambides [3]). By combining (A.16) with identities (A.17), (A.18) and (A.19) we obtain

$$\mathbb{P}\left[V_{m'}=v_{m'}, W_{m'}=w_{m'}, K_{m'}=k_{m'} \mid A_m(\mathbf{n}_m,\mathbf{m}_m,k_m)\right]$$

$$= \frac{\frac{\theta^{n+k_m+k_{m'}}}{(\theta)_{(n+m+m')}}}{\frac{\theta^{n+k_m}}{(\theta)_{(n+m)}} \prod_{i=1}^{n} m_{i,m}! \prod_{i=1}^{k_m}(n_{i,m}-1)!}$$

$$\times \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}}$$

$$
\begin{aligned}
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \prod_{i=1}^{n} m_{i,m}! \\
&\times \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} \prod_{i=1}^{k_m} (n_{i,m}-1)! \left|s(v_{m'},k_{m'})\right| \\
&= \frac{\theta^{k_{m'}}}{(\theta+n+m)_{(m')}} \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}} \\
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} \left|s(v_{m'},k_{m'})\right|. \qquad \text{(A.20)}
\end{aligned}
$$

Accordingly, from the probability (A.20) we can write the following marginal probability

$$
\begin{aligned}
&\mathbb{P}\left[V_{m'}=v_{m'}, W_{m'}=w_{m'} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\right] \\
&= \sum_{k_{m'}=0}^{v_{m'}} \frac{\theta^{k_{m'}}}{(\theta+n+m)_{(m')}} \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}} \\
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} \left|s(v_{m'},k_{m'})\right| \\
&= \frac{1}{(\theta+n+m)_{(m')}} \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}} \\
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} (\theta)_{(v_{m'})}. \qquad \text{(A.21)}
\end{aligned}
$$

By combining (A.14) with (A.20) one obtains the conditional distribution of the random variable $(\mathbf{N}_{m'}, \mathbf{M}_{m'})$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m, V_{m'}, W_{m'}, K_{m'})$. In particular,

$$
\begin{aligned}
&\mathbb{P}\left[\mathbf{N}_{m'}=\mathbf{n}_{m'}, \mathbf{M}_{m'}=\mathbf{m}_{m'} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\right] \\
&= \Bigg[\frac{\theta^{k_{m'}}}{(\theta+n+m)_{(m')}} \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}} \\
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}\right)_{(m'-v_{m'}-w_{m'})} \left(\sum_{i=1}^{k_m} n_{i,m}\right)_{(w_{m'})} \left|s(v_{m'},k_{m'})\right|\Bigg]^{-1} \\
&\times \frac{\frac{\theta^{n+k_m+k_{m'}}}{(\theta)_{(n+m+m')}}}{\frac{\theta^{n+k_m}}{(\theta)_{(n+m)}} \prod_{i=1}^{n} m_{i,m}! \prod_{i=1}^{k_m} (n_{i,m}-1)!} \\
&\times \binom{m'}{v_{m'}, w_{m'}, m'-v_{m'}-w_{m'}} \\
&\times \binom{m'-v_{m'}-w_{m'}}{m_{1,m'}, \dots, m_{n,m'}} \prod_{i=1}^{n} (m_{i,m}+m_{i,m'})!
\end{aligned}
$$

$$
\begin{aligned}
&\times \binom{w_{m'}}{n_{1,m'}, \dots, n_{k_n,m'}} \prod_{i=1}^{k_m} (n_{i,m} + n_{i,m'} - 1)! \\
&\times \frac{1}{k_{m'}!} \binom{v_{m'}}{n_{k_n+1,m'}, \dots, n_{k_n+k_{m'},m'}} \prod_{i=1}^{k_{m'}} (n_{k_m+i} - 1)! \\
&= \frac{[(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})} (\sum_{i=1}^{k_m} n_{i,m})_{w_{m'}} |s(v_{m'}, k_{m'})|]^{-1}}{\prod_{i=1}^{n} m_{i,m}! \prod_{i=1}^{k_m} (n_{i,m} - 1)!} \\
&\times \binom{m' - v_{m'} - w_{m'}}{m_{1,m'}, \dots, m_{n,m'}} \prod_{i=1}^{n} (m_{i,m} + m_{i,m'})! \\
&\times \binom{w_{m'}}{n_{1,m'}, \dots, n_{k_n,m'}} \prod_{i=1}^{k_m} (n_{i,m} + n_{i,m'} - 1)! \\
&\times \frac{1}{k_{m'}!} \binom{v_{m'}}{n_{k_n+1,m'}, \dots, n_{k_n+k_{m'},m'}} \prod_{i=1}^{k_{m'}} (n_{k_m+i} - 1)!. \qquad \text{(A.22)}
\end{aligned}
$$

The distribution (A.22) leads to the conditional distribution (A.15). For any $x \in \{1, \dots, n\}$, let $1 \le \tau_1 < \cdots < \tau_x \le n$ let $\mathcal{J}_{n,x} = \{0, \dots, n\}/\{\tau_1, \dots, \tau_x\}$. Also, let

$$
\begin{aligned}
&\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'},n-x} \\
&\quad = \Biggl\{ (m_{i,m'})_{i \in \mathcal{J}_{n,x}} : m_{i,m'} \ge 0 \wedge \sum_{i \in \mathcal{J}_{n,x}} m_{i,m'} = m' - v_{m'} - w_{m'} - \sum_{i=1}^{x} m_{\tau_i,m'} \Biggr\}.
\end{aligned}
$$

Marginalizing (A.22) over $\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'},n-x}$, $\mathcal{S}^{(0)}_{w_{m'},k_m}$ and $\mathcal{S}_{v_{m'},k_{m'}}$ one obtains, by simple algebraic manipulations, the marginal conditional distribution

$$
\begin{aligned}
&\mathbb{P}\bigl[\mathbf{M}_{(\tau_1,\dots,\tau_x),m'} = \mathbf{m}_{(\tau_1,\dots,\tau_x),m'} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\bigr] \\
&\quad = \frac{[(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})} (\sum_{i=1}^{k_m} n_{i,m})_{(w_{m'})} |s(v_{m'}, k_{m'})|]^{-1}}{\prod_{i=1}^{n} m_{i,m}! \prod_{i=1}^{k_m} (n_{i,m} - 1)!} \\
&\qquad \times \sum_{\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'},n-x}} \binom{m' - v_{m'} - w_{m'}}{m_{1,m'}, \dots, m_{n,m'}} \prod_{i=1}^{n} (m_{i,m} + m_{i,m'})! \\
&\qquad \times \sum_{\mathcal{S}^{(0)}_{w_{m'},k_m}} \binom{w_{m'}}{n_{1,m'}, \dots, n_{k_n,m'}} \prod_{i=1}^{k_m} (n_{i,m} + n_{i,m'} - 1)! \\
&\qquad \times \frac{1}{k_{m'}!} \sum_{\mathcal{S}_{v_{m'},k_{m'}}} \binom{v_{m'}}{n_{k_n+1,m'}, \dots, n_{k_n+k_{m'},m'}} \prod_{i=1}^{k_{m'}} (n_{k_n+i,m'} - 1)! \\
&\quad = \frac{[(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}]^{-1}}{\prod_{i=1}^{n} m_{i,m}!}
\end{aligned}
$$

$$
\begin{aligned}
&\times \sum_{\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'},n-x}} \binom{m'-v_{m'}-w_{m'}}{m_{1,m'},\dots,m_{n,m'}} \prod_{i=1}^{n}(m_{i,m}+m_{i,m'})! \\
&= \frac{[(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}]^{-1}}{\prod_{i=1}^{n} m_{i,m}!} \\
&\times \frac{(m'-v_{m'}-w_{m'})!\prod_{i=1}^{x}(m_{\tau_i,m}+m_{\tau_i,m'})!}{(m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'})!\prod_{i=1}^{x} m_{\tau_i,m'}!} \\
&\times \sum_{\mathcal{S}^{(0)}_{m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'},n-x}} \frac{(m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'})!}{\prod_{i\in\mathcal{J}_{n,x}} m_{i,m'}!} \\
&\times \prod_{i\in\mathcal{J}_{n,x}} (1+m_{i,m}+m_{i,m'}-1)! \\
&= \frac{[(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}]^{-1}}{\prod_{i=1}^{n} m_{i,m}!} \\
&\times \frac{(m'-v_{m'}-w_{m'})!\prod_{i=1}^{x}(m_{\tau_i,m}+m_{\tau_i,m'})!}{(m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'})!\prod_{i=1}^{x} m_{\tau_i,m'}!} \\
&\times \left(n+m-\sum_{i=1}^{k_m} n_{i,m}-\sum_{i=1}^{x}(1+m_{\tau_i,m})\right)_{(m'-v_{m'}-w_{m'}-\sum_{i=1}^{x} m_{\tau_i,m'})} \\
&\times \prod_{i\in\mathcal{J}_{n,x}} m_{i,m}!, \qquad \text{(A.23)}
\end{aligned}
$$

follows by a direct application of Theorem 2.5 in Charalambides [3]. The expression (A.23) coincides with the conditional distribution of $\mathbf{M}_{m(\tau_1,\dots,\tau_x),m'}$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m, V_{m'}, W_{m'}, K_{m'})$ displayed in (A.15), and the proof is completed. □

Proof of Theorem 2.3(i). We compute the rth descending factorial moment of the random variable $R_{n,m+m'}$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m)$. In particular, by a direct application of the Vandermonde formula, we can write the following identity

$$
\begin{aligned}
&\mathbb{E}\big[(R_{n,m+m'})_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&= \sum_{s=0}^{r} \binom{r}{s} (-1)^s (n-s)_{[r-s]} \mathbb{E}\big[\big(R^*_{n,m+m'}\big)_{[s]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big],
\end{aligned}
$$

where $R^*_{n,m+m'} = \sum_{1\le i\le n} \mathbb{1}_{\{M_{i,m}+M_{i,m'}=0\}}$. A repeated application of the Binomial theorem leads to write the rth moment of the random variable $R^*_{n,m+m'}$ as follows

$$
\begin{aligned}
&\mathbb{E}\big[\big(R^*_{n,m+m'}\big)^r \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&= \sum_{x=1}^{n} \sum_{i_1=1}^{r-1} \sum_{i_2=1}^{i_1-1} \cdots \sum_{i_{x-1}=1}^{i_{x-2}-1} \binom{r}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{x-2}}{i_{x-1}} \\
&\times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \mathbb{E}\left[\prod_{t=1}^{x} (\mathbb{1}_{\{M_{c_t,m}+M_{c_t,m'}=0\}})^{i_{x-t}-i_{x-t+1}} \,\middle|\, A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\right]
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_{x=1}^{r} S(r,x)x! \\
&\quad \times \sum_{\mathbf{c}^{(x)} \in \mathcal{C}_{n,x}} \mathbb{E}\left[\prod_{t=1}^{x} \mathbb{1}_{\{M_{c_t,m}+M_{c_t,m'}=0\}} \,\middle|\, A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'}) \right] \\
&= \sum_{x=1}^{r} S(r,x)x! \\
&\quad \times \sum_{\mathbf{c}^{(x)} \in \mathcal{C}_{n,x}} \mathbb{P}\big[\mathbf{M}_{\mathbf{c}^{(x)},m} + \mathbf{M}_{\mathbf{c}^{(x)},m'} = (\underbrace{0,\ldots,0}_{x}) \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big].
\end{aligned}
$$

Then we can use (A.15) to obtain an expression for the conditional probability of $\mathbf{M}_{\mathbf{c}^{(x)},m'}$ given $A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)$ and $B_{m'}(v_{m'}, w_{m'}, k_{m'})$. In particular, from (A.15) we have

$$
\begin{aligned}
&\mathbb{E}\big[\big(R^*_{n,m+m'}\big)_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&= r! \sum_{\mathbf{c}^{(r)} \in \mathcal{C}_{n,r}} \prod_{i=1}^{r} \mathbb{1}_{\{m_{c_i,m}=0\}} \\
&\quad \times \frac{(n+m-\sum_{i=1}^{k_m} n_{i,m} - r)_{(m'-v_{m'}-w_{m'})}}{(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}}.
\end{aligned}
\tag{A.24}
$$

Finally, we marginalize the last expression with respect to the distribution of the random variable $(V_{m'}, W_{m'}) \mid (\mathbf{N}_m, \mathbf{M}_m, K_m)$. Them by combining (A.24) with (A.21), and using the fact that (A.24) does not depend on $K_{m'}$, we can write

$$
\begin{aligned}
&\mathbb{E}\big[\big(R^*_{n,m+m'}\big)_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&= r! \sum_{\mathbf{c}^{(r)} \in \mathcal{C}_{n,r}} \frac{(\theta+n+m-r)_{(m')}}{(\theta+n+m)_{(m')}} \prod_{i=1}^{r} \mathbb{1}_{\{m_{c_i,m}=0\}},
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\big[(R_{n,m+m'})_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&= \sum_{s=0}^{r} \binom{r}{s} (-1)^s (n-s)_{[r-s]} s! \sum_{\mathbf{c}^{(s)} \in \mathcal{C}_{n,s}} \frac{(\theta+n+m-s)_{(m')}}{(\theta+n+m)_{(m')}} \prod_{i=1}^{s} \mathbb{1}_{\{m_{c_i,m}=0\}} \\
&= \sum_{s=0}^{r} \binom{r}{s} (-1)^s (n-s)_{[r-s]} s! \frac{(\theta+n+m-s)_{(m')}}{(\theta+n+m)_{(m')}} \sum_{\mathbf{c}^{(s)} \in \mathcal{C}_{n,s}} \prod_{i=1}^{s} \mathbb{1}_{\{m_{c_i,m}=0\}} \\
&= \sum_{s=0}^{r} \binom{r}{s} (-1)^s (n-s)_{[r-s]} \frac{(\theta+n+m-s)_{(m')}}{(\theta+n+m)_{(m')}} (n-r_m)_{[s]} \\
&= \frac{r!}{(\theta+n+m)_{(m')}} \sum_{s=0}^{r} \binom{n-s}{r-s} (-1)^s \binom{n-r_m}{s} (\theta+n+m-s)_{(m')},
\end{aligned}
\tag{A.25}
$$

where $r_m = \sum_{1\le i\le n} \mathbb{1}_{\{m_{i,m}>0\}}$. The distribution of $R_{n,m+m'} \mid (\mathbf{N}_m, \mathbf{M}_m, K_m)$ follows from (A.25). In particular, for any $x = r_m, \dots, \min(n, m' + r_m)$, we can write

$$\begin{aligned}
&\mathbb{P}\big[R_{n,m+m'} = x \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&\quad = \mathbb{P}[R_{n,m+m'} = x \mid R_{n,m} = r_m] \\
&\quad = \sum_{l\ge 0} \frac{(-1)^l}{x!l!} \mathbb{E}\big[(R_{n,m+m'})_{[x+l]} \mid R_{n,m} = r_m\big] \\
&\quad = \frac{1}{(\theta+n+m)_{(m')}} \sum_{l\ge x} \frac{1}{x!} (-1)^{l-x} (l)_{[x]} \\
&\qquad \times \sum_{s=0}^{l} \binom{n-s}{l-s} (-1)^s \binom{n-r_m}{s} (\theta+n+m-s)_{(m')} \\
&\quad = \frac{1}{(\theta+n+m)_{(m')}} \sum_{s=0}^{n} (-1)^s \binom{n-r_m}{s} (\theta+n+m-s)_{(m')} \\
&\qquad \times \sum_{l=s}^{n} (-1)^{l-x} \binom{l}{x} \binom{n-s}{l-s} \\
&\quad = \frac{(-1)^{-x}}{(\theta+n+m)_{(m')}} \sum_{s=0}^{n} (-1)^{n-s} \binom{n-r_m}{s} \binom{s}{n-x} (\theta+n+m-s)_{(m')} \\
&\quad = \frac{(-1)^{n}}{(\theta+n+m)_{(m')}} \sum_{s=n}^{n+x} (-1)^{s} \binom{n-r_m}{s-x} \binom{s-x}{n-x} (\theta+n+m-s+x)_{(m')} \\
&\quad = \frac{\binom{n-r_m}{n-x}}{(\theta+n+m)_{(m')}} \sum_{s=0}^{x-r_m} (-1)^{s} \binom{x-r_m}{s} (\theta+m-s+x)_{(m')}.
\end{aligned} \tag{A.26}$$

The expression (A.26) coincides, after applying the Vandermonde identity, to the conditional distribution of $R_{n,m+m'}$ given $R_{n,m}$ in (2.13), and the proof is completed. □

Proof of Theorem 2.3(ii). Similarly to Part (i), we compute the rth descending factorial moment of the random variable $R_{n,m+m'}$ given $(\mathbf{N}_m, \mathbf{M}_m, K_m)$. As a first step, we observe that we can rewrite Equation (2.12) in the following way

$$\begin{aligned}
\tilde{R}_{l,n,m'} &= \sum_{i=1}^{n} (1 - \mathbb{1}_{\{M_{i,m'}=0\}}) \mathbb{1}_{\{M_{i,m}=l\}} \\
&= R_{l,n,m} - \sum_{i=1}^{n} \mathbb{1}_{\{M_{i,m'}=0\}} \mathbb{1}_{\{M_{i,m}=l\}},
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{E}\big[(\tilde{R}_{l,n,m'})_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&\quad = \sum_{s=0}^{r} \binom{r}{s} (-1)^s (r_{l,m} - s)_{[r-s]} \mathbb{E}\big[\big(\tilde{R}^*_{l,m,m'}\big)_{[s]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big],
\end{aligned}$$

where $r_{l,m} = \sum_{1\le i\le n} \mathbb{1}_{\{m_{i,m}=l\}}$ and $\tilde{R}^{*}_{l,m,m'} = \sum_{1\le i\le n} \mathbb{1}_{\{M_{i,m'}=0\}}\mathbb{1}_{\{M_{i,m}=l\}}$. By a repeated application of the Binomial theorem we can write the following expression

$$
\begin{aligned}
&\mathbb{E}\big[\big(\tilde{R}^{*}_{l,m,m'}\big)^{r} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&\quad = \sum_{x=1}^{n}\sum_{i_1=1}^{r-1}\sum_{i_2=1}^{i_1-1}\cdots\sum_{i_{x-1}=1}^{i_{x-2}-1}\binom{r}{i_1}\binom{i_1}{i_2}\cdots\binom{i_{x-2}}{i_{x-1}} \\
&\qquad \times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \mathbb{E}\left[\prod_{t=1}^{x}\big(\mathbb{1}_{\{M_{c_t,m'}=0\}}\mathbb{1}_{\{M_{c_t,m}=l\}}\big)^{i_{x-t}-i_{x-t+1}} \,\Big|\, A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\right] \\
&\quad = \sum_{x=1}^{r} S(r,x)x! \\
&\qquad \times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \mathbb{E}\left[\prod_{t=1}^{x}\mathbb{1}_{\{M_{c_t,m'}=0\}}\mathbb{1}_{\{M_{c_t,m}=l\}} \,\Big|\, A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\right] \\
&\quad = \sum_{x=1}^{r} S(r,x)x! \\
&\qquad \times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \prod_{t=1}^{x}\mathbb{1}_{\{m_{c_t,m}=l\}}\mathbb{E}\left[\prod_{t=1}^{x}\mathbb{1}_{\{M_{c_t,m'}=0\}} \,\Big|\, A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\right] \\
&\quad = \sum_{x=1}^{r} S(r,x)x! \\
&\qquad \times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \prod_{t=1}^{x}\mathbb{1}_{\{m_{c_t,m}=l\}}\mathbb{P}\big[\mathbf{M}_{\mathbf{c}^{(x)},m'} = (\underbrace{0,\dots,0}_{x}) \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&\quad = \sum_{x=1}^{r} S(r,x)x! \\
&\qquad \times \sum_{\mathbf{c}^{(x)}\in\mathcal{C}_{n,x}} \prod_{t=1}^{x}\mathbb{1}_{\{m_{c_t,m}=l\}}\frac{(n+m-\sum_{i=1}^{k_m} n_{i,m} - \sum_{i=1}^{x}(1+m_{c_i,m}))_{(m'-v_{m'}-w_{m'})}}{(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}},
\end{aligned}
$$

i.e.,

$$
\begin{aligned}
&\mathbb{E}\big[\big(\tilde{R}^{*}_{l,m,m'}\big)_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m), B_{m'}(v_{m'}, w_{m'}, k_{m'})\big] \\
&\quad = r! \sum_{\mathbf{c}^{(r)}\in\mathcal{C}_{n,r}} \prod_{t=1}^{r}\mathbb{1}_{\{m_{c_t,m}=l\}}\frac{(n+m-\sum_{i=1}^{k_m} n_{i,m} - \sum_{i=1}^{r}(1+m_{c_i,m}))_{(m'-v_{m'}-w_{m'})}}{(n+m-\sum_{i=1}^{k_m} n_{i,m})_{(m'-v_{m'}-w_{m'})}}. \qquad \text{(A.27)}
\end{aligned}
$$

Finally, we marginalize the last expression with respect to the distribution of the random variable $(V_{m'}, W_{m'}) \mid (\mathbf{N}_m, \mathbf{M}_m, K_m)$. By combining (A.27) with (A.21) one has

$$
\begin{aligned}
&\mathbb{E}\big[\big(\tilde{R}^{*}_{l,m,m'}\big)_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&\quad = r! \sum_{\mathbf{c}^{(r)}\in\mathcal{C}_{n,r}} \prod_{t=1}^{r}\mathbb{1}_{\{m_{c_t,m}=l\}}\frac{(\theta+n+m-\sum_{i=1}^{r}(1+m_{c_i,m}))_{(m')}}{(\theta+n+m)_{(m')}} = r!\binom{r_{l,m}}{r}\frac{(\theta+n+m-r(1+l))_{(m')}}{(\theta+n+m)_{(m')}}
\end{aligned}
$$

and

$$\begin{aligned}
&\mathbb{E}\big[(\tilde{R}_{l,m,m'})_{[r]} \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&\quad = \sum_{s=0}^{r} \binom{r}{s} (-1)^s (r_{l,m} - s)_{[r-s]} s! \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}}.
\end{aligned} \tag{A.28}$$

Accordingly, the distribution of $R_{n,m+m'} \mid (\mathbf{N}_m, \mathbf{M}_m, K_m)$ follows from (A.28). In particular, for any $x = 0, \dots, \min(r_{l,m}, m')$, we can write the following expression

$$\begin{aligned}
&\mathbb{P}\big[\tilde{R}_{l,m,m'} = x \mid A_m(\mathbf{n}_m, \mathbf{m}_m, k_m)\big] \\
&\quad = \mathbb{P}[\tilde{R}_{l,m,m'} = x \mid R_{l,n,m} = r_{l,m}] \\
&\quad = \sum_{y \geq 0} \frac{(-1)^y}{x! y!} \mathbb{E}\big[(\tilde{R}_{l,m,m'})_{[x+l]} \mid R_{l,n,m} = r_{l,m}\big] \\
&\quad = \sum_{y \geq 0} (-1)^y \frac{1}{x! y!} \sum_{s=0}^{x+y} \binom{x+y}{s} (-1)^s (r_{l,m} - s)_{[x+y-s]} \\
&\qquad \times s! \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} \\
&\quad = \sum_{y \geq x} (-1)^{y-x} \frac{1}{x!(y-x)!} \sum_{s=0}^{y} \binom{y}{s} (-1)^s (r_{l,m} - s)_{[y-s]} \\
&\qquad \times s! \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} \\
&\quad = \sum_{y=0}^{r_{l,m}} (-1)^{y-x} \binom{y}{x} \sum_{i=0}^{y} \binom{r_{l,m} - s}{y - s} (-1)^s \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} \\
&\quad = \sum_{s=0}^{r_{l,m}} (-1)^{s-x} \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} \sum_{y=s}^{r_{l,m}} (-1)^y \binom{y}{x} \binom{r_{l,m} - s}{y - s} \\
&\quad = \sum_{s=0}^{r_{l,m}} (-1)^{-x} \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} \\
&\qquad \times \sum_{y=0}^{r_{l,m}-s} (-1)^y \binom{y+s}{x} \binom{r_{l,m} - s}{y} \\
&\quad = \sum_{s=0}^{r_{l,m}} (-1)^{-x} \binom{r_{l,m}}{s} \frac{(\theta + n + m - s(1+l))_{(m')}}{(\theta + n + m)_{(m')}} (-1)^{r_{l,m}-s} \binom{s}{x - r_{l,m} + s}.
\end{aligned}$$

□

Proof of Proposition 2.1. Recalling the definitions of the random variables $R_{n,m}$ and $R_{n,m+m'}$, let $\tilde{R}_{n,m'} = R_{n,m+m'} - R_{n,m}$, that is the number of distinct types in the additional sample $\mathbf{X}_{m'}$ that coincide with the atoms Z_i^* that are not in the initial sample $\mathbf{X}_m$. In other terms $\tilde{R}_{n,m'}$ denotes the number of new types induced by $\mathbf{X}_{m'}$ that coincide with the atoms Z_i^*. From (2.13), we can write

$$\mathbb{P}[\tilde{R}_{n,1} = 1 \mid R_{n,m} = y] = \mathbb{E}[\tilde{R}_{n,1} \mid R_{n,m} = y] = \frac{n - y}{\theta + n + m}. \tag{A.29}$$

See also the factorial moment formula (A.25) with $r = 1$ and $m' = 1$. Also, from (2.14),

$$\begin{aligned}\mathbb{P}[\tilde{R}_{l,n,1} = 1 \mid R_{l,n,m} = y] &= \mathbb{E}[\tilde{R}_{l,n,1} \mid R_{l,n,m} = y] \\ &= y\left(1 - \frac{\theta + n + m - (1 + l)}{\theta + n + m}\right).\end{aligned} \tag{A.30}$$

See also the factorial moment formula in Equation (A.28) with $r = 1$ and $m' = 1$. The proof is completed by simply randomizing the parameter n appearing in (A.29) and in (A.30) with respect to the distribution (2.15) and (2.17), respectively. □

Acknowledgements

We are grateful to two anonymous referees, whose comments and suggestions have helped us to improve the paper substantially. S. Favaro and S. Feng acknowledge the hospitality of the Banff International Research Station, Canada, where this project has been completed during the Research in Team Programme "Random partitions and Bayesian nonparametrics". S. Favaro is supported by the European Research Council through StG N-BNP 306406. Shui Feng is supported by the Natural Sciences and Engineering Research Council of Canada. P. Jenkins is supported in part by Engineering and Physical Sciences Research Council Grant EP/L018497/1.

References

[1] N. Berestycki. *Recent Progress in Coalescent Theory*. *Ensaios Matemáticos*. SBM, Rio de Janeiro, 2009. MR2574323

[2] M. Birkner, J. Blath, M. Möhle, M. Steinrücken and J. Tams. A modified lookdown construction for the Xi–Fleming–Viot process with mutation and populations with recurrent bottlenecks. *ALEA* **6** (2009) 25–61. MR2485878

[3] C. A. Charalambides. *Combinatorial Methods in Discrete Distributions*. Wiley, Hoboken, 2005. MR2131068

[4] M. De Iorio and R. C. Griffiths. Importance sampling on coalescent histories I. *Adv. in Appl. Probab.* **36** (2004) 417–433. MR2058143

[5] P. Donnelly and T. G. Kurtz. A countable representation of the Fleming–Viot measure-valued diffusion. *Ann. Probab.* **24** (1996) 698–742. MR1404525

[6] S. N. Ethier and R. C. Griffiths. The transition function of a Fleming–Viot process. *Ann. Probab.* **21** (1993) 1571–1590. MR1235429

[7] S. N. Ethier and T. G. Kurtz. Fleming–Viot processes in population genetics. *SIAM J. Control Optim.* **31** (1993) 345–386. MR1205982

[8] W. J. Ewens. The sampling theory of selectively neutral alleles. *Theor. Popul. Biol.* **3** (1972) 87–112. MR0325177

[9] W. J. Ewens. *Mathematical Population Genetics*. Springer, Berlin, 2004. MR0554616

[10] T. S. Ferguson. A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** (1973) 209–230. MR0350949

[11] I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika* **40** (1953) 237–264. MR0061330

[12] I. J. Good and G. H. Toulmin. The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** (1956) 45–63. MR0077039

[13] R. C. Griffiths. Lines of descent in the diffusion approximation of neutral Wright–Fisher models. *Theor. Popul. Biol.* **17** (1980) 37–50. MR0568666

[14] R. C. Griffiths. Asymptotic line of descent distributions. *J. Math. Biol.* **21** (1984) 67–75. MR0770713

[15] R. C. Griffiths, P. A. Jenkins and Y. S. Song. Importance sampling and the two-locus model with subdivided population structure. *Adv. in Appl. Probab.* **40** (2008) 473–500. MR2433706

[16] R. C. Griffiths and S. Tavaré. Ancestral inference in population genetics. *Statist. Sci.* **9** (1994) 307–319. MR1325431

[17] R. C. Griffiths and S. Tavaré. Simulating probability distributions in the coalescent. *Theor. Popul. Biol.* **46** (1994) 131–159.

[18] R. C. Griffiths and S. Tavaré. The genealogy of a neutral mutation. In *Highly Structured Stochastic Systems*, P. J. Green, N. L. Hjort and S. Richardson (Eds). Oxford University Press, Oxford, 2003. MR2082417

[19] J. Hey and R. Nielsen. Multilocus methods for estimating population sizes, migration rates and divergence time, with applications to the divergence of *Drosophila pseudoobscura* and *D. persimilis*. *Genetics* **167** (2004) 747–760.

[20] A. Hobolth, M. Uyenoyama and C. Wiuf. Importance sampling for the infinite sites model. *Stat. Appl. Genet. Mol. Biol.* **7** (2008) 32. MR2457045

[21] F. M. Hoppe. The sampling theory of neutral alleles and an urn model in population genetics. *J. Math. Biol.* **25** (1987) 123–159. MR0896430

[22] J. F. C. Kingman. The coalescent. *Stochastic Process. Appl.* **13** (1982) 235–248. MR0671034

[23] J. F. C. Kingman. On the genealogy of large populations. *J. Appl. Probab.* **19** (1982) 27–43. MR0633178

[24] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165** (2003) 2213–2233.

[25] M. Möhle. On sampling distributions for coalescent processes with simultaneous multiple collisions. *Bernoulli* **12** (2006) 35–53. MR2202319

[26] M. Möhle and S. Sagitov. Coalescent patterns in diploid exchangeable population models. *J. Math. Biol.* **47** (2003) 337–352. MR2024501

[27] J. S. Paul and Y. S. Song. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. *Genetics* **186** (2010) 321–338.

[28] J. S. Paul, M. Steinrücken and Y. S. Song. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. *Genetics* **187** (2011) 1115–1128.

[29] S. Sheehan, K. Harris and Y. S. Song. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. *Genetics* **194** (2013) 647–662.

[30] R. S. Singh, R. C. Lewontin and A. A. Felton. Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in *Drosophila pseudoobscura*. *Genetics* **84** (1976) 609–629.

[31] M. Stephens. Inference under the coalescent. In *Handbook of Statistical Genetics*, D. J. Balding, M. Bishop and C. Cannings (Eds). Wiley, New York, 2001.

[32] M. Stephens and P. Donnelly. Inference in molecular population genetics. *J. Roy. Statist. Soc. Ser. B* **62** (2000) 605–655. MR1796282

[33] S. Tavaré. Line-of-descent and genealogical processes, and their applications in population genetics models. *Theor. Popul. Biol.* **26** (1984) 119–164. MR0770050

[34] S. Tavaré. Ancestral inference in population genetics. In *Ecole d'Eté de Probabilités de Saint-Flour XXXI*. *Lecture Notes in Mathematics*. Springer, New York, 2004.

[35] G. A. Watterson. Lines of descent and the coalescent. *Theor. Popul. Biol.* **26** (1984) 77–92. MR0760232