Coalescent Processes with Skewed Offspring Distributions and Nonequilibrium Demography

Sebastian Matuszewski,[*,†,‡,1] Marcel E. Hildebrandt,[*,‡,1] Guillaume Achaz,[§,] and Jeffrey D. Jensen[*,†,††,2]**

[*]School of Life Sciences, and [†]School of Basic Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland, [‡]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland, [§]Institut de Systématique, Evolution, Biodiversité, ISYEB, UMR 7205 CNRS MNHN UPMC EPHE, Paris, France, [**]Centre Interdisciplinaire de Recherche en Biologie, CIRB, UMR 7241 CNRS Collège de France INSERM, Paris, France, and [††]Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Tempe, Arizona 85287

ORCID IDs: 0000-0002-4393-9283 (S.M.); 0000-0003-4514-5935 (G.A.); 0000-0002-4786-8064 (J.D.J.)

ABSTRACT Nonequilibrium demography impacts coalescent genealogies leaving detectable, well-studied signatures of variation. However, similar genomic footprints are also expected under models of large reproductive skew, posing a serious problem when trying to make inference. Furthermore, current approaches consider only one of the two processes at a time, neglecting any genomic signal that could arise from their simultaneous effects, preventing the possibility of jointly inferring parameters relating to both offspring distribution and population history. Here, we develop an extended Moran model with exponential population growth, and demonstrate that the underlying ancestral process converges to a time-inhomogeneous psi-coalescent. However, by applying a nonlinear change of time scale—analogous to the Kingman coalescent—we find that the ancestral process can be rescaled to its time-homogeneous analog, allowing the process to be simulated quickly and efficiently. Furthermore, we derive analytical expressions for the expected site-frequency spectrum under the time-inhomogeneous psi-coalescent, and develop an approximate-likelihood framework for the joint estimation of the coalescent and growth parameters. By means of extensive simulation, we demonstrate that both can be estimated accurately from whole-genome data. In addition, not accounting for demography can lead to serious biases in the inferred coalescent model, with broad implications for genomic studies ranging from ecology to conservation biology. Finally, we use our method to analyze sequence data from Japanese sardine populations, and find evidence of high variation in individual reproductive success, but few signs of a recent demographic expansion.

KEYWORDS coalescent theory; multiple mergers; population growth; maximum likelihood; site-frequency spectrum

THE origins of the coalescent in the early 1970s mark a milestone for evolutionary theory (Kingman 2000). More than 45 years after Kingman formally proved the existence of the "*n*-coalescent" (Kingman 1982a,b,c), the so-called Kingman-*n*-coalescent has gradually become the key theoretical tool to study the complex interplay of mutation, genetic drift, gene flow, and selection. Closely linked to its underlying forward-in-time population model, *e.g.*, the Wright-Fisher (WF; Fisher 1930; Wright 1931) and the Moran model (Moran 1958, 1962), the Kingman coalescent has been used to derive expected levels of neutral variation, including the number of segregating sites, the average number of pairwise differences, and the expectation of the allele frequencies in a population sample (*i.e.*, the site-frequency spectrum; SFS). In fact, these predictions apply not only to the WF and Moran model, but extend to a large class of Cannings exchangeable population models (Cannings 1974) that all converge to the Kingman coalescent in the ancestral limit (Möhle and Sagitov 2001). Furthermore, the Kingman coalescent forms the basis for many population genetic statistics—such as Tajima's D (Tajima 1989), Fay and Wu's H (Fay and Wu 2000), or, more generally, any SFS-based test statistic (Achaz 2009; Ferretti *et al.* 2010)—and subsequent inferences (Irwin *et al.* 2016) to

doi: https://doi.org/10.1534/genetics.117.300499

Manuscript received May 12, 2017; accepted for publication October 30, 2017; published Early Online November 10, 2017.

Supplemental material is available online at www.genetics.org/lookup/suppl/doi:10.1534/genetics.117.300499/-/DC1.

[1] These authors contributed equally to this work.

[2] Corresponding author: Center for Evolution and Medicine, School of Life Sciences, Arizona State University, Jensen Lab, Tempe, AZ 85287. E-mail: jeffrey.d.jensen@asu.edu

detect deviations from the assumption of a neutrally evolving, constant-sized, panmictic population (Wakeley 2009).

While the Kingman coalescent has been shown to be robust to violations of its assumptions (Möhle 1998, 1999), such as constant population size, random mating, and nonoverlapping generations, and has been extended to accommodate selection, migration, and population structure (Neuhauser and Krone 1997; Nordborg 1997; Wilkinson-Herbots 1998), it breaks down in the presence of skewed offspring distributions (Eldon and Wakeley 2006), strong positive selection (Neher and Hallatschek 2013; Schweinsberg 2017), recurrent selective sweeps (Durrett and Schweinsberg 2004, 2005), and large sample sizes (Wakeley and Takahashi 2003; Bhaskar *et al.* 2014). In particular, all of these effects can cause more than two lineages to coalesce at a time, resulting in so-called multiple mergers. Hence, the underlying coalescent topology (*i.e.*, the gene genealogy) is no longer represented by a bifurcating tree as in the "standard" Kingman case, but can take more complex tree shapes that can also feature several simultaneous mergers. Taking these points into account, a more general class of models, so-called multiple-merger coalescent (MMC) models, have been developed (*e.g.*, Bolthausen and Sznitman 1998; Pitman 1999; Sagitov 1999; Schweinsberg 2000; Möhle and Sagitov 2001; reviewed in Tellier and Lemaire 2014), aiming to generalize the Kingman coalescent model (Wakeley 2013). As for the latter, these MMC models can often be derived from Moran models, generalized to allow multiple offspring per individual (Eldon and Wakeley 2006; Huillet and Möhle 2013; see also review of Irwin *et al.* 2016).

Starting from such an extended Moran model, Eldon and Wakeley (2006) proved that the underlying ancestral process converges to a psi-coalescent (sometimes also called Dirac coalescent; Eldon *et al.* 2015), and that population genetic parameters inferred from genetic data from Pacific oysters (*Crassostrea gigas*) under this model differ vastly from those inferred assuming the Kingman coalescent. Their study—being the first to link MMC models to actual biological questions, molecular data and population genetic inferences—highlighted that high variation in individual reproductive success drastically affect both genealogical history and subsequent analyses; this has been observed in many marine organisms such Atlantic cod (*Gadus morhua*) and Japanese sardines (*Sardinops melanostictus*), but should also occur more generally in any species with type III survivorship curves that undergo so-called sweepstake-reproductive events (Hedgecock 1994; Hedgecock and Pudovkin 2011). Fundamentally, the problem is that an excess of low-frequency alleles (*i.e.*, singletons), a ubiquitous characteristic of many marine species (Niwa *et al.* 2016), could be explained by either models of recent population growth or skewed offspring distributions when analyzed under the Kingman coalescent, assuming neutrality, which can result in serious mis-inference (*e.g.*, a vast overestimation of population growth).

In developing a SFS-based maximum likelihood framework, Eldon *et al.* (2015) demonstrated that multiple merger coalescents and population growth can be distinguished from their genomic footprints in the higher-frequency classes of the SFS with high statistical power (see also Spence *et al.* 2016). However, there is currently neither a modeling framework that considers the genomic signal arising from the joint action of both reproductive skew and population growth, nor is there any *a priori* reason to believe that the two could not act simultaneously.

Here, we develop an extension of the standard Moran model that accounts for both reproductive skewness and exponential population growth, and prove that its underlying ancestral process converges to a time-inhomogeneous psi-coalescent. By (nonlinearly) rescaling branch lengths this process can—analogous to the Kingman coalescent (Griffiths and Tavaré 1998)—be transformed into its time-homogeneous analog, allowing efficient large-scale simulations. Furthermore, we derive analytical formulae for the expected site-frequency spectrum under the time-inhomogeneous psi-coalescent and develop an approximate-likelihood framework for the joint estimation of the coalescent and growth parameters. We then perform extensive validation of our inference framework on simulated data, and show that both the coalescent parameter and the growth rate can be estimated accurately from whole-genome data. In addition, we demonstrate that, when demography is not accounted for, the inferred coalescent model can be seriously biased, with broad implications for genomic studies ranging from ecology to conservation biology (*e.g.*, due to its effects on effective population size or diversity estimates). Finally, using our joint estimation method, we reanalyze mtDNA from Japanese sardine (*Sardinops melanostictus*) populations, and find evidence for considerable reproductive skew, but only limited support for a recent demographic expansion.

Methods

Here, we will first present an extended, discrete-time Moran model (Moran 1958, 1962; Eldon and Wakeley 2006) with exponential population growth that will serve as the forward-in-time population genetic model underlying the ancestral limit process. We will then give a brief overview of coalescent models, with special focus on the psi-coalescent (Eldon and Wakeley 2006), before revisiting SFS-based maximum likelihood methods to infer coalescent parameters and population growth rates.

An extended Moran model with exponential growth

We consider the idealized, discrete-time model with variable population size shown generally in Figure 1. Furthermore, let $N_n \in \mathbb{N}$ be the deterministic and time-dependent population size $n \in \mathbb{N}$ time steps in the past, where, by definition, $N = N_0$ denotes the present population size. In particular, defining $\boldsymbol{\nu}(n)$ as the exchangeable vector of family sizes—with

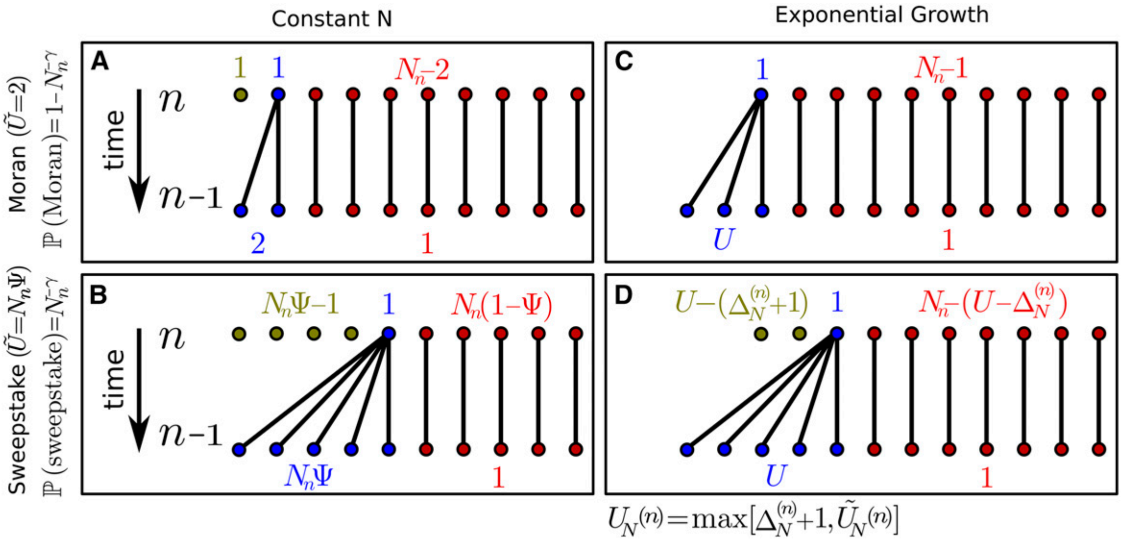
Constant N
Exponential Growth
Moran ($\tilde{U}$=2)
$\mathbb{P}$ (Moran)=1−$N_n^{-\gamma}$
Sweepstake ($\tilde{U}$=$N_n\Psi$)
$\mathbb{P}$ (sweepstake)=$N_n^{-\gamma}$
A
1
1
N_n−2
time
n
n−1
2
1
C
1
N_n−1
U
1
B
$N_n\Psi$−1
1
N_n(1−Ψ)
time
n
n−1
$N_n\Psi$
1
D
U−($\Delta_N^{(n)}$+1)
1
N_n−(U−$\Delta_N^{(n)}$)
U
1
$U_N(n)$=max[$\Delta_N^{(n)}$+1, $\tilde{U}_N(n)$]

Figure 1 Illustration of the extend Moran model with exponential growth. Shown are the four different scenarios of population transition within a single discrete time step. (A) The population size remains constant and a single individual produces exactly two offspring ("Moran-type" reproductive event). (B) The population size remains constant and a single individual produces ψN_n offspring ("sweepstake" reproductive event). (C) The population size increases by $\Delta_N^{(n)}$ individuals and a single individual produces exactly max$[\Delta_N^{(n)} + 1, 2]$ offspring. (D) The population size increases by $\Delta_N^{(n)}$ individuals and a single individual produces exactly max$\left[\Delta_N^{(n)} + 1, \psi N_n\right]$ offspring. Note that n denotes the number of steps in the past, such that $n = 0$ denotes the present. An overview of the notation used in this model is given in Table 1.

components $\nu_i(n)$ indicating the number of descendants of the ith individual—the (variable) population size can be expressed as

$$N_{n-1} = \sum_{i=1}^{N_n} \nu_i(n) \quad \text{with} \quad (\nu_1(n), \nu_2(n), \ldots, \nu_N(n)) \in \mathbb{N}^{N_n}. \tag{1}$$

Furthermore, we assume that the reproductive mechanism follows that of an extended Moran model (Eldon and Wakeley 2006; Huillet and Möhle 2013). In particular, as in the original Moran model, at any given point in time $n \in \mathbb{N}$, only a single individual reproduces and leaves $U_N(n)$ offspring (including itself). Formally, the number of offspring can be written as a sequence of random variables $(U_N(n))_{n\in\mathbb{N}}$ [where each $U_N(n)$ is supported on $\{0, 1, \ldots, N_{n-1}\}$], such that $\boldsymbol{\nu}(n)$ – up to reordering – is given by

$$\nu_i(n) := \begin{cases} 0 & \text{if } i < U_N(n) \\ U_N(n) & \text{if } i = U_N(n) \\ 1 & \text{otherwise.} \end{cases} \tag{2}$$

However, since population size varies over time, the sequence $(U_N(n))_{n\in\mathbb{N}}$ is generally not identically distributed. On a technical note though, we require that the (U_n) are independently distributed, which ensures that the corresponding backward process satisfies the Markov property.

An illustration of our model, and the four different scenarios for forming the next generation (*i.e.*, within a single discrete time step), is shown in Figure 1. Generally, we differentiate between two possible reproductive events: a classic "Moran-type" reproductive event (Figure 1, A and C), and a "sweepstake" reproductive event (Figure 1, B and D) occurring with probabilities $1 - N_n^{-\gamma}$ and $N_n^{-\gamma}$, respectively. If the population size remains constant between consecutive generations (Figure 1, A and B), we reobtain the extended Moran model introduced by Eldon and Wakeley (2006), in which a single randomly chosen individual either leaves exactly two offspring and replaces one randomly chosen individual (Moran-type), or replaces a fixed proportion $\psi \in (0, 1]$ of the population (of size N_n). Note that, throughout, without loss of generality, we assume that $N_n\psi$ is integer-valued. In both reproductive scenarios, the remaining individuals persist. However, if the population size increases between consecutive generations (Figure 1, C and D), the reproductive mechanism needs to be adjusted accordingly. Let

$$\Delta_N^{(n)} := N_{n-1} - N_n \tag{3}$$

denote the increment in population size between two consecutive time points. Then, the number of offspring at time n is given by

$$U_N(n) := \max\left[\Delta_N^{(n)} + 1, \tilde{U}_N(n)\right] \tag{4}$$

where $\tilde{U}_N(n)$ denotes number of offspring for the constant-size population. Thus, independent of the type of reproductive event, *i.e.*, Moran-type or sweepstake, and, in the spirit of the original Moran model, additional individuals are always assigned to be offspring of the single reproducing individual of the previous generation.

Following Eldon and Wakeley (2006), the distribution of the number of offspring $\mathbb{P}(U_N(n) = u)$ can be written as

Table 1 Summary of notation and definitions

Notation	Definition
U_N	Number of offspring of a reproductive event in an extended Moran model with population size N
$\boldsymbol{\nu}$	Vector of family sizes
Λ	Probability measure on $[0,1]$
$\lambda_{i,x}$	Coalescent rate for x out of i active lineages
$G_{i,x}$	Probability of an $x-$merger among i active lineages
$c_N^{(n)}$	Coalescence probability
$(\mathcal{A}_{n,k}^{\psi,\rho})_{n\in\mathbb{N}} \subset \mathcal{P}_k$	Ancestral process of the extended Moran model sweepstake parameter ψ ($\psi=0$ implying Kingman's coalescent), and exponential population growth at rate ρ for a sample of size k defined on $\mathcal{P}_k$, *i.e.*, the collection of partitions of the set $[k]=\{1,\dots,k\}$.
$(\Pi_{t,k}^{\psi,\rho})_{t\geq 0} \subset \mathcal{P}_k$	$\psi-$coalescent ($\psi=0$ implying Kingman's coalescent) with exponential growth at rate ρ and sample of size k defined on $\mathcal{P}_k$, *i.e.*, the collection of partitions of the set $[k]=\{1,\dots,k\}$.
$\mathcal{G}(\cdot)$	Time-change function
$T_{\text{MRCA}}^{(k)}$	Time until the MRCA for a sample of size k
T_i	Sum of the length of all branches with i descendants
T_{tot}	Total branch length of the coalescent tree
$\boldsymbol{\eta}^{(k)}=(\eta_1^{(k)},\dots,\eta_{k-1}^{(k)})$	SFS for a sample of size k
$\boldsymbol{\varphi}^{(k)}=(\varphi_1^{(k)},\dots,\varphi_{k-1}^{(k)})$	Normalized expected SFS for a sample of size k
$\mathfrak{c}_k=(\mathfrak{c}_{2,2},\dots,\mathfrak{c}_{k,k})$	Expected time to the first coalescence for a sample of size $i\in\{2,\dots,k\}$

$$\mathbb{P}(U_N(n)=u)=\begin{cases} N_n^{-\gamma} & \text{if } u=\max\left[\Delta_N^{(n)}+1, N_n\psi\right] \\ 1-N_n^{-\gamma} & \text{if } u=\max\left[\Delta_N^{(n)}+1, 2\right] \\ 0 & \text{otherwise,} \end{cases} \tag{5}$$

for some $\gamma>0$ that—for a given fixed population size—determines the probability of a sweepstake reproductive event. Here, we will consider only the case where $1<\gamma<2$, such that sweepstake events happen frequently enough that the ancestral process will be characterized by multiple mergers, and that all coalescent events are due to sweepstake reproductive events, but not so frequently that the population is devoid of genetic variation (Eldon and Wakeley 2006). Note that, while the numbers of offspring and replaced individuals are no longer (necessarily) equal when the population size increases, the general reproductive mechanism remains unaltered.

Throughout the paper, and following Griffiths and Tavaré (1994), we will assume that the population is growing exponentially over time at rate ϱ, and, in particular, that the population size, n steps in the past, is given by

$$N_n := \lfloor N(1-\varrho)^n \rfloor, \tag{6a}$$

with

$$\varrho=\rho\frac{\psi^2}{N^\gamma} \tag{6b}$$

if the ancestral process is dominated by sweepstake events (*i.e.*, if $1<\gamma<2$), or

$$\varrho=\rho\frac{1}{N^2} \tag{6c}$$

if Moran-type reproductive events dominate (*i.e.*, if $\gamma>2$), and the growth rate ρ is measured in units of the corresponding coalescent time. A discussion and details about the derivation of the coalescent-time scaling are given below in the *Derivation of the ancestral limit process* section.

Multiple merger coalescents: the Psi-coalescent

The most general class of coalescent processes that allows for multiple lineages to coalesce per coalescent event (but not for multiple coalescent events at the same time) is the so-called Λ-coalescent. These processes are partition-valued exchangeable stochastic processes defined by a finite measure Λ on the $[0,1]$ interval (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999). In particular, the rate at which x out of i active lineages merge is given by

$$\lambda_{i,x}=\binom{i}{x}\int_0^1 y^{x-2}(1-y)^{i-x}\Lambda(\mathrm{d}y). \tag{7}$$

Special instances of the Λ-coalescent are Kingman's coalescent (Kingman 1982a,b) with

$$\Lambda(\mathrm{d}y)=\delta_0(\mathrm{d}y) \tag{8}$$

and the psi-coalescent (Eldon and Wakeley 2006) with

$$\Lambda(\mathrm{d}y)=\delta_\psi(\mathrm{d}y), \tag{9}$$

where the measure Λ is entirely concentrated at 0 and ψ, respectively.

Under a (constant-size) extended Moran model, as proposed by Eldon and Wakeley (2006) (corresponding to Figure 1, A and B), the scaled coalescence rates of the ancestral process become

$$\lambda_{i,x} = \begin{cases} \binom{i}{x} \psi^{x-2}(1-\psi)^{i-x} & \text{if } 0<\gamma<2 \\ \binom{i}{x}\left(\frac{2}{2+\psi^2}1_{x=2} + \frac{\psi^x}{2+\psi^2}(1-\psi)^{i-x}\right) & \text{if } \gamma = 2 \\ \binom{i}{x}\mathbb{1}_{x=2} & \text{otherwise,} \end{cases} \tag{10}$$

where $\mathbb{1}_{x=2}$ denotes the indicator function, which is 1 if $x=2$ and 0 otherwise. Accordingly, the corresponding rate matrix of the ancestral process $\boldsymbol{Q}_\psi \in \mathbb{R}^{k\times k}$ with sample size k is given by

$$\boldsymbol{Q}_\psi(i,j) = \begin{cases} \lambda_{i,j} & \text{if } i>j \\ -\frac{1}{\psi^2}\left(1-(1-\psi)^i - i\psi(1-\psi)^{i-1}\right) & \text{if } i=j \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where $j=i-x-1$. Note that the diagonal entries of $\boldsymbol{Q}_\psi(i,j)$ (*i.e.*, when $i=j$) is given by the (negative) sum over all coalescent rates, *i.e.*, $-\sum_{m=2}^{i}\lambda_{i,m}$, which evaluates to the closed-form representation given in the second line of Equation 11.

In particular, in the boundary case $\psi=0$, we recover the rate matrix under the Kingman coalescent as

$$\boldsymbol{Q}_0(i,j) = \begin{cases} \binom{i}{2} & \text{if } j=i-1 \\ -\binom{i}{2} & \text{if } i=j \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

Note that, in the infinite population size limit, γ defines the time scale of the ancestral process. In particular, if $0<\gamma<2$, all coalescence events are due to sweepstake reproductive events, whereas sweepstake events do not happen frequently enough if $\gamma>2$, such that all (2-) mergers are due to Moran-type reproductive events. Moreover, in the latter case, the ancestral process of the Moran model can be described accurately by the Kingman coalescent (when scaled appropriately). Note that, for the special case $\gamma=2$, both reproductive events happen on the same time scale (Eldon and Wakeley 2006).

SFS-based maximum likelihood inference

In order to infer the coalescent model and its associated coalescent parameter, and to (separately) estimate the demographic history of the population, Eldon *et al.* (2015) recently derived an (approximate) maximum likelihood framework based on the SFS [see also Birkner and Blath (2008) and Koskela *et al.* (2015) for alternative inference approaches based on a full likelihood framework and approximate conditional sampling distributions, respectively].

In the following, we will give a concise overview of their approach, which forms the basis for the joint inference of coalescent parameters and population growth rates.

First, let k denote the number of sampled (haploid) individuals (*i.e.*, the number of leaves in the coalescent tree). Furthermore, let $\boldsymbol{\eta}^{(k)} = (\eta_1^{(k)},\ldots,\eta_{k-1}^{(k)})$ denote the number of segregating sites with derived allele count of $i=1,\ldots,k-1$ of all sampled individuals (*i.e.*, the SFS), and let $s=\sum_{i=1}^{k-1}\eta_i$ be the total number of segregating sites. Provided that $s>0$, we define the normalized expected SFS $\boldsymbol{\varphi}^{(k)} = (\varphi_1^{(k)},\ldots,\varphi_{k-1}^{(k)})$ as

$$\varphi_i^{(k)} = \frac{\mathbb{E}\left[\eta_i^{(k)}\right]}{\sum_{i=1}^{k-1}\mathbb{E}\left[\eta_i^{(k)}\right]}, \tag{13}$$

which, given a coalescent model $(\Pi_{t,k}^{\psi,\rho})_{t\geq 0}$, and, assuming the infinite-sites model (Watterson 1975), can be interpreted as the probability that a mutation appears i times in a sample of size k (Eldon *et al.* 2015). Furthermore, note that $\varphi_i^{(k)}$ is a function of $(\Pi_{t,k}^{\psi,\rho})_{t\geq 0}$ (*i.e.*, of the coalescent process and the demographic population history), but, unlike $\mathbb{E}[\boldsymbol{\eta}^{(k)}]$, is not a function of the mutation rate, and should provide a good first-order approximation of the expected SFS as long as the sample size and the mutation rate are not too small (Eldon *et al.* 2015).

Then, the likelihood function $\mathcal{L}((\Pi_{t,k}^{\psi,\rho})_{t\geq 0}, \tilde{\boldsymbol{\eta}}^{(k)}, s)$ for the observed frequency spectrum $\tilde{\boldsymbol{\eta}}^{(k)}$ and given coalescent model $(\Pi_{t,k}^{\psi,\rho})_{t\geq 0}$ is given by

$$\begin{aligned} \mathcal{L}\left(\left(\Pi_{t,k}^{\psi,\rho}\right)_{t\geq 0}, \tilde{\boldsymbol{\eta}}^{(k)}, s\right) &= \mathbb{P}^{\left(\Pi_{t,k}^{\psi,\rho}\right)_{t\geq 0}, s}\left(\eta_i^{(k)} = \tilde{\eta}_i^{(k)}, i\in[k-1]\right) \\ &= \mathbb{E}^{\Pi}\left[\frac{s!}{\tilde{\eta}_1^{(k)}!\ldots\tilde{\eta}_{k-1}^{(k)}!}\prod_{i=1}^{k-1}\left(\frac{T_i^{(k)}}{T_{\mathrm{tot}^{(k)}}}\right)^{\tilde{\eta}_i^{(k)}}\right] \\ &\approx \frac{s!}{\tilde{\eta}_1^{(k)}!\ldots\tilde{\eta}_{k-1}^{(k)}!}\prod_{i=1}^{k-1}\left(\varphi_i^{(k)}\right)^{\tilde{\eta}_i^{(k)}} \\ &\propto \prod_{i=1}^{k-1}\exp\left[-s\varphi_i^{(k)}\right]\frac{\left(s\varphi_i^{(k)}\right)^{\tilde{\eta}_i^{(k)}}}{\tilde{\eta}_i^{(k)}!} \end{aligned} \tag{14}$$

(Eldon *et al.* 2015). Note that, in the third line, we approximated $\mathbb{E}\left[\left(T_i^{(k)}/T_{\mathrm{tot}}^{(k)}\right)\right] \approx \mathbb{E}\left[T_i^{(k)}\right]\big/\mathbb{E}\left[T_{\mathrm{tot}}^{(k)}\right] = \varphi_i^{(k)}$. In fact, Bhaskar *et al.* (2015) recently used a Poisson random field approximation to derive an analogous, structurally identical likelihood function for estimating demographic parameters under the Kingman coalescent. Notably though, their approximation assumes that the underlying coalescent tree is independent at each site, under which condition Equation 14 is exact.

As an alternative to the likelihood approach, we followed Eldon *et al.* (2015) and also implemented a minimal-distance statistic approach where

$$\hat{\psi}, \hat{\rho} = \arg\min_{\psi,\rho} d_p\left(\tilde{\boldsymbol{\eta}}^{(k)}, \mathbb{E}\left[\boldsymbol{\eta}^{(k)}\right]\right), \tag{15}$$

where d_p is some metric on $\mathbb{R}^{p-1}$ calculated between the observed and the expected SFS under the generating coalescent process.

Note, though, that both the likelihood and the distance-based approach require expressions for the normalized expected SFS $\boldsymbol{\varphi}^{(k)}$. Instead of performing Monte Carlo simulations to obtain these quantities, we adapted an approach recently proposed by Spence *et al.* (2016), who derived analytical formulas for the expected SFS under a given (general) coalescent model $(\Pi^{\psi,\rho}_{t,k})_{t\geq 0}$, and an intensity measure $\xi(t) : \mathrm{R}_{\geq 0} \rightarrow \mathrm{R}_{>0}$. In particular, the authors showed that

$$\mathbb{E}\left[\boldsymbol{\eta}^{(k)}\right] = \frac{\theta}{2}\mathbf{BCL}\mathfrak{c}_k, \tag{16}$$

where $\mathbf{B} \in \mathbb{R}^{k-1\times k-1}$ and $\mathbf{C} \in \mathbb{R}^{k-1\times k-1}$ are both $\Lambda -$ independent (and thus easy to calculate) matrices, $\mathbf{L} \in \mathbb{R}^{k-1\times k-1}$ is a $\Lambda -$ dependent lower triangular matrix that depends on the rate matrix $\mathbf{Q}$ and its spectral decomposition, θ is the population-scaled mutation rate, and $\mathfrak{c}_k = (\mathfrak{c}_{2,2}, \ldots, \mathfrak{c}_{k,k})$ denotes the expected time to the first coalescence for a sample of size $i \in \{2, \ldots, k\}$. Importantly, the time-inhomogeneity of the underlying coalescent process only enters through the first coalescence times $\mathfrak{c}_k$. For example, the first coalescence times for the Kingman coalescent with an exponentially growing population are given by

$$\mathfrak{c}_{i,i} = -\frac{1}{\rho}\exp\left[\frac{\binom{i}{2}}{\rho}\right]\mathrm{Ei}\left(-\frac{\binom{i}{2}}{\rho}\right), \tag{17}$$

where $\mathrm{Ei}(x) := -\int_{-x}^{\infty}(\exp[-t]/t)\mathrm{d}t$ denotes the exponential integral (Polanski *et al.* 2003; Polanski and Kimmel 2003; Bhaskar *et al.* 2015). Finally, plugging Equation 16 into Equation 13 leads to

$$\varphi_i^{(k)} = \frac{(\mathbf{BCL}\mathfrak{c}_k)_i}{\sum_{i=1}^{k-1}(\mathbf{BCL}\mathfrak{c}_k)_i}, \tag{18}$$

highlighting that θ cancels, and that the likelihood function (Equation 14) is independent of the mutation rate.

To obtain the coalescent parameter ψ and population growth rate ρ that maximize the likelihood function (Equation 14) or, respectively, minimize the distance function (Equation 15), we used a grid search procedure over an equally spaced two-dimensional grid with $\psi_{\text{grid}} = \{0, 0.01, \ldots, 1\}$ and $\rho_{\text{grid}} = \{0, 1, \ldots, 1024\}$, and evaluated the value of the likelihood, respectively, distance function, at each grid point.

Data availability

The empirical raw data used have been downloaded from GenBank (accession numbers LC031518–LC031623; data from Niwa *et al.* (2016)). The empirical SFS can be downloaded from Supplemental Material, File S5. The simulation program and the inference program were written in C++ and can be downloaded from GitHub under https://github.com/Matu2083/MultipleMergers-PopulationGrowth.

Results and Discussion

The aim of this work was to derive the ancestral process for an exponentially expanding population that undergoes sweepstake reproductive events. We first derive the time-inhomogeneous Markovian ancestral process that underlies the extended Moran model, and show that, analogous to the Kingman coalescent, it can be described by a time-homogeneous Markov chain on a nonlinear time scale. In particular, we derive the coalescent rates and the time-change function, and prove convergence to a $\Lambda -$ coalescent with Dirac measure at ψ. Detailed derivations of the results, which in the main text have been abbreviated to keep formulas concise, can be found in File S1. On the basis of these results, we derive a maximum likelihood inference framework for the joint inference of the coalescent parameter and the population growth rate, and assess its accuracy and performance through large-scale simulations. Furthermore, we quantify the bias of coalescent and population growth parameter estimates when mistakenly neglecting population demography or reproductive skew. Finally, we apply our approach to mtDNA from Japanese sardine (*S. melanostictus*) populations. where patterns of sequence variation were shown to be more consistent with sole influence from sweepstake reproductive events, again highlighting the potential mis-inference of growth if reproductive skew is not properly accounted for (Grant *et al.* 2016; Niwa *et al.* 2016).

Derivation of the ancestral limit process

Unlike in the case of a constant-size population, the sequence of the number of offspring $\left(U_N(n)\right)_{n\in\mathrm{N}}$ changes along with the (time-dependent) population size. Thus, the ancestral process is characterized by an inhomogeneous Markov chain with transition probabilities

$$G_{i,x}^{(n)} = \binom{i}{x}\sum_{u=2}^{N_n}\frac{\mathbb{P}^N\left(U_N(n) = u\right)(u)_x(N_n - u)_{i-x}}{(N_n)_i} \tag{19a}$$

where $(z)_j$ is the descending factorial, $z(z-1)\ldots(z-j+1)$ with $(z)_0 = 1$, and $\mathbb{P}^N$ denotes the rescaled distribution of U_N given by

$$\mathbb{P}^N(U_N = u) := \frac{\mathbb{P}\left(U_N = u\right)u\left(u-1\right)/\left(N(N-1)\right)}{c_N^{(n)}}. \tag{19b}$$

Note that $\mathbb{P}^N$ is scaled by the time-dependent coalescence probability $c_N^{(n)}$, which scales the unit of time in the limit process such that it is equal to $G_{2,2}$ steps in the discrete-time model, and thus serves as the "natural" time scale for the corresponding ancestral process—defined as

$$c_N^{(n)} \coloneqq G_{2,2}^{(n)} = \frac{\mathbb{E}\left[U_N(n)^2 - U_N(n)\right]}{N_n(N_n-1)} \tag{20}$$

for all $n \in \mathbb{N}$.

Plugging Equation 5 into Equations 19a and 20, and using Equation 4 then yields

$$G_{i,x}^{(n)} = \binom{i}{x} \frac{1}{N_n(N_n-1)}\Big((1-N_n^{-\gamma})\Big(\max\Big[2, \Delta_N^{(n)}+1\Big]\Big)_x \times \Big(N_n - \max\Big[2, \Delta_N^{(n)}+1\Big]\Big)_{i-x} + N_n^{-\gamma}\Big(\max\Big[\psi N_n, \Delta_N^{(n)}+1\Big]\Big)_x \times \Big(N_n - \max\Big[\psi N_n, \Delta_N^{(n)}+1\Big]\Big)_{i-x}\Big) \tag{21}$$

and

$$c_N^{(n)} = \frac{(1-N_n^{-\gamma})\Big(\max\Big[2, \Delta_N^{(n)}+1\Big]\Big)_2 + N_n^{-\gamma}\Big(\max\Big[\psi N_n, \Delta_N^{(n)}+1\Big]\Big)_2}{N_n(N_n-1)}, \tag{22}$$

respectively. Note that Equation 22 is the weighted sum of the number of offspring for the two different reproductive events. Furthermore, taking the limit $N \to \infty$ in Equation 3

$$\begin{aligned} \lim_{N\to\infty} \Delta_N^{(n)} &= \lim_{N\to\infty} (N_n - N_{n-1}) \\ &\leq \lim_{N\to\infty} N\rho \frac{\psi^2}{N^\gamma} \\ &= 0 \end{aligned} \tag{23}$$

shows that $\Delta_N^{(n)}$ is bounded for all $n \in \mathbb{N}$ under the exponential growth model, and thus allows dropping of the maxima condition in Equations 21 and 22. Furthermore, for sufficiently large N, Equation 22 becomes

$$c_N^{(n)} = \frac{(1-N_n^{-\gamma})2 + \psi N_n^{1-\gamma}(\psi N_n - 1)}{N_n(N_n-1)} \sim \frac{\psi^2}{N_n^\gamma}. \tag{24}$$

To prove that the time-scaled ancestral process of the underlying extended Moran model converges to a continuous-time Markov chain as the initial population size approaches infinity, we apply *Theorem 2.2* in Möhle (2002), which requires the following definitions: First, consider a step function $F_N : [0, \infty) \to [0, \infty)$ given by

$$F_N(s) \coloneqq \sum_{n=1}^{\lfloor s \rfloor} c_N^{(n)}. \tag{25}$$

Furthermore, let $\mathcal{G}_N^{-1}$ denote a modification of the right-continuous inverse of F_N

$$\mathcal{G}_N^{-1}(t) \coloneqq \inf\{s > 0 | F_N(s) > t\} - 1 \tag{26}$$

which will constitute the time-change function in the following. Since by assumption $\lim_{s\to\infty} F_N(s) = \infty$ it follows that $\mathcal{G}_N^{-1}(t)$ is finite for all $t \in [0, \infty)$. Finally, *Theorem 2.2* (Möhle 2002) requires that for all $t \in [0, \infty)$

$$\lim_{N\to\infty} \inf_{1 \leq n \leq \mathcal{G}_N^{-1}(t)} N_n = \infty \tag{27}$$

and

$$\lim_{N\to\infty} \sup_{1 \leq n \leq \mathcal{G}_N^{-1}(t)} c_N^{(n)} = 0 \tag{28}$$

holds, *i.e.*, that—on the new time scale—the population size remains large while the coalescent probabilities become small.

Then, let $(\mathcal{A}_{n,k}^{\psi,\rho})_{n\in\mathbb{N}}$ denote the ancestral process of the extended Moran model with exponential growth (see *Model and Methods*), and let φ and ξ denote two partitions of $[k]$ with $\xi \subset \phi$ of size a and $b = b_1 + \ldots + b_a$ (where $b_1 \geq b_2 \geq \ldots \geq b_a \geq 1$), respectively. The transition probability of $(\mathcal{A}_{n,k}^{\psi,\rho})_{n\in\mathbb{N}}$ at time $n \in \mathbb{N}$ is given by

$$\begin{aligned} &\Phi_a^{(N)}(n; b_1, b_2, \ldots, b_a) \\ &\coloneqq \frac{1}{(N_{n-1})_b} \sum_{\text{all distinct}^{i_1,\ldots,1_a=1}}^{N_n} \mathrm{E}\Big(\big(\nu_{i_1}(n)\big)_{b_1} \cdots \big(\nu_{i_a}(n)\big)_{b_a}\Big). \end{aligned} \tag{29}$$

Thus, for the extended Moran model with exponential growth *Theorem 2.2* in Möhle (2002) states:

Theorem 1. (Theorem 2.2; Möhle 2002). Assume that Equations 27 and 28 hold, and for all $t \in \mathbb{R}_{>0}$ the limit

$$\pi_a((0,t]; b_1, \ldots, b_a) \coloneqq \lim_{N\to\infty} \sum_{n=1}^{\mathcal{G}_N^{-1}(t)} \Phi_a^{(N)}(n; b_1, b_2, \ldots, b_a) \tag{30}$$

exists. Then, for each sample size, $k \in \mathbb{N}$, the ancestral process $(\mathcal{A}_{\mathcal{G}_N^{-1}(t),k}^{\psi,\rho})_{t\geq 0}$ converges as N tends to infinity to a time-continuous, and, in general, a time-inhomogeneous Markov chain $(\Pi_{t,k}^{\psi,\rho})_{t\geq 0}$.

Note, though, that in its general form Theorem 1 was derived for any generic Cannings model as well as any kind of population size change (Möhle 2002).

We will now derive our first main result, and show that the ancestral limiting process $\lim_{N\to\infty}(\mathcal{A}_{n,k}^{\psi,\rho})_{t\geq 0}$ converges to a $\Lambda - k -$ coalescent on a nonlinear time scale. First, we derive the time-change function $\mathcal{G}_N^{-1}(t)$ for the ancestral process by considering the step function (Equation 25)

$$\begin{aligned} F_N(s) &= \sum_{n=1}^{s} c_N^{(n)} \sim \sum_{n=1}^{s} \frac{\psi^2}{N_n^\gamma} \\ &= \frac{\psi^2}{N^\gamma}\Big(1 - \rho\frac{\psi^2}{N^\gamma}\Big)^{-\gamma} \frac{\Big(\Big(1-\rho\frac{\psi^2}{N^\gamma}\Big)^{-\gamma s} - 1\Big)}{\Big(1-\rho\frac{\psi^2}{N^\gamma}\Big)^{-\gamma} - 1}. \end{aligned} \tag{31}$$

Solving for s then gives

$$\begin{aligned} \mathcal{G}_N^{-1}(t) &= \inf\{s>0 : F_N(\lfloor s>t \rfloor)\} \\ &\sim \left\lfloor \frac{\log[1+\rho\gamma t]}{\rho\gamma}\,\frac{N^\gamma}{\psi^2} \right\rfloor, \end{aligned} \tag{32}$$

where we have used $\log[1-\rho\psi^2/N^\gamma] \sim -\rho\psi^2/N^\gamma$ for sufficiently large N. In particular, we have

$$\mathcal{G}^{-1}(t) \coloneqq \lim_{N\to\infty} \mathcal{G}_N^{-1}(t) c_N^{(0)} = \frac{\log[1+\rho\gamma t]}{\rho\gamma}. \tag{33}$$

Furthermore, Equations 27 and 28 hold, since

$$\begin{aligned} &\lim_{N\to\infty} \inf_{1\le n\le \mathcal{G}_N^{-1}(t)} N\left(1-\rho\frac{\psi^2}{N^\gamma}\right)^n \\ &= \lim_{N\to\infty} N \inf_{1\le n\le \mathcal{G}_N^{-1}(t)} \exp\left(-\frac{\rho\psi^2}{N^\gamma}n\right) \\ &= \lim_{N\to\infty} N(1+\rho\gamma t)^{-\frac{1}{\gamma}} \\ &= \infty \quad \forall t \in (0,\infty), \end{aligned} \tag{34}$$

and, by the same reasoning

$$\lim_{N\to\infty} \sup_{1\le n\le \mathcal{G}_N^{-1}(t)} \frac{\psi^2}{N\left(1-\rho\frac{\psi^2}{N^\gamma}\right)^n} = 0, \quad \forall t \in (0,\infty). \tag{35}$$

Finally, to show that Equation 30 holds, we first note that

$$\Phi_a^{(N)}(n; b_1, \ldots, b_a) = 0, \tag{36}$$

and for $a \ge 2$, and there are two indices $1 \le i < j \le a$ with $b_i, b_j \ge 2$

$$\pi_a((0,t]; b_1, \ldots, b_a) = 0, \tag{37}$$

since the extended Moran model does not allow for more than one reproductive event at a time.

Thus, $\lim_{N\to\infty} (\mathcal{A}^{\psi,\rho}_{\mathcal{G}_N^{-1}(t),k})_{t\ge 0}$ is well defined and does not feature any simultaneous coalescent events, implying that the limiting process must be a (possibly time-inhomogeneous) $\Lambda - k -$ coalescent. Further, for $a = 1$,

$$\begin{aligned} \lim_{N\to\infty} \sum_{n=1}^{\mathcal{G}_N^{-1}(t)} \Phi_1^{(N)}(n; b) &= \lim_{N\to\infty} \sum_{n=1}^{\mathcal{G}_N^{-1}(t)} G_{b,b} \\ &= \psi^{b-2} t. \end{aligned} \tag{38}$$

Hence, Theorem 1 implies that, for each sample size $k \in \mathbb{N}$ the limit of the time-scaled ancestral process $(\mathcal{A}^{\psi,\rho}_{\mathcal{G}_N^{-1}(t),k})_{t\ge 0}$ exists, and, from Equation 38 it follows that $\lim_{N\to\infty} (\mathcal{A}^{\psi,\rho}_{\mathcal{G}_N^{-1}(t),k})_{t\ge 0}$ is a time-homogeneous $\Lambda - k -$ coalescent. Further,

$$t\psi^{b-2} = t\int_0^1 x^{b-2}\Lambda(dx) \tag{39}$$

holds for all $b \in \mathbb{N}$, if, and only if, Λ is the Dirac measure at ψ. Thus, in the large population-size limit, the ancestral process, $\lim_{N\to\infty} (\mathcal{A}^{\psi,\rho}_{\mathcal{G}_N^{-1}(t),k})_{t\ge 0}$, converges to a psi-$k$-coalescent – $(\Pi^{\psi,\rho}_{t,k})_{t\in\mathbb{R}_{\ge 0}} \coloneqq \lim_{N\to\infty} (\mathcal{A}^{\psi,\rho}_{\lfloor t/c_N^{(0)} \rfloor,k})_{t\ge 0}$, which is equal (in distribution) to a regular psi-k-coalescent $(\Pi^{\psi,0}_{\mathcal{G}(t),k})$ with time (nonlinearly) rescaled by

$$\mathcal{G}(t) = \frac{\exp(\rho\gamma t) - 1}{\rho\gamma}. \tag{40}$$

Put differently, analogous to the results obtained for the Kingman coalescent (Griffiths and Tavaré 1994, 1998; Kaj and Krone 2003), the time-inhomogeneous ancestral limiting process of the extended Moran model with exponential growth can be transformed into a time-homogeneous psi-coalescent with coalescent rates given by Equation 10, with branches rescaled by Equation 33, allowing it to be simulated easily and efficiently. Intuitively, the transformation sums over the coalescence intensities of the time-inhomogeneous process, and weighs them by the time they were effective, such that, on the new time-scale coalescent intensities are constant across time, and the (rescaled) process is time-homogeneous (see also Kaj and Krone 2003). Thus, changing the time-scale by Equation 40 compensates for the shrinking population sizes (going backward in time) and the effect of increasing (total) coalescent rates.

To highlight the duality between the two processes, *i.e.*, the (forward in time) extended Moran model and the corresponding coalescent, key properties (*e.g.*, the summed length of all branches with i descendants T_i and the total tree length T_{tot}) are compared in the File S2. Finally, note that Equation 33 is—except for the additional factor γ that is proportional to the coalescent time scale—structurally identical to the time-change function in the Kingman case (see Equation 2.7 in Griffiths and Tavaré 1998). However, since $\mathcal{G}^{-1}(t)$ depends on the product $\rho\gamma$, it is impossible to obtain a direct estimate of ρ (or γ) without additional information, and thus—analogous to the case of the population scaled mutation rate θ—only the compound parameter can be estimated. To keep notation simple, though, we will refer to ρ (instead of the compound parameter $\rho\gamma$) when referring to growth rate estimates.

Joint inference of coalescent parameters and population growth rates

In this paragraph we modify the likelihood function

$$\mathcal{L}\left(\left(\Pi^{\psi,\rho}_{t,k}\right)_{t\ge 0}, \tilde{\boldsymbol{\eta}}^{(k)}, s\right) \propto \sum_{i=1}^{k-1} \exp\left[-s\varphi_i^{(k)}\right] \frac{\left(s\varphi_i^{(k)}\right)^{\tilde{\eta}_i^{(k)}}}{\tilde{\eta}_i^{(k)}!} \tag{41}$$

derived in the *Methods* section to jointly infer the coalescent parameter ψ and the population growth rate ρ. Note that, while the general form of the likelihood function (Equation 14) is independent of the generating coalescent process, changes in ψ and ρ affect the normalized expected SFS, as given by

$$\varphi_i^{(k)} = \frac{(\mathbf{BCL}\mathfrak{c}_k)_i}{\sum_{i=1}^{k-1}(\mathbf{BCL}\mathfrak{c}_k)_i}. \tag{42}$$

Recall that **B** and **C** depend neither on ψ nor ρ, and that **L** does depend on ψ but not on ρ, and that the time-inhomogeneity of the underlying coalescent process enters only through the first coalescence times $\mathfrak{c}_k$, which are given by

$$\begin{aligned}\mathfrak{c}_{i,i} &= \int_0^\infty \mathbb{P}(\text{time of first coalescence for } i \text{ individuals} > t)dt \\ &= \int_0^\infty \exp\left((\mathbf{Q})_{i,i}\int_0^t (1/\xi(s))ds\right)dt,\end{aligned} \tag{43}$$

where $\xi(s)$ denotes the intensity measure (Polanski and Kimmel 2003; Bhaskar *et al.* 2015; Spence *et al.* 2016). For the psi-coalescent with exponential growth, $\xi(t) = e^{-\rho\gamma t}$, such that Equation 43 becomes

$$\begin{aligned}\mathfrak{c}_{i,i} &= \int_0^\infty \exp\left((\mathbf{Q})_{i,i}\frac{\exp(\rho\gamma t)-1}{\rho\gamma}\right)dt \\ &= -\frac{\exp\left((\mathbf{Q})_{i,i}\,\rho\gamma\right)}{\rho\gamma}\mathrm{Ei}\left(-(\mathbf{Q})_{i,i}\,\rho\gamma\right),\end{aligned} \tag{44}$$

where $\mathrm{Ei}(x) := -\int_{-x}^{\infty}(\exp[-t]/t)\mathrm{d}t$ denotes the exponential integral. Thus, when growth rates are measured on their corresponding coalescent scale, *i.e.*, $\rho\gamma$ under the psi-coalescent *vs.* ρ under the Kingman coalescent, Equation 44 is a generalization of the Kingman coalescent result (Equation 17) derived by Polanski and Kimmel (2003). Finally, combining Equation 44 with Equation 42 allows for the exact computation of the normalized expected SFS $\boldsymbol{\varphi}^{(k)}$, avoiding the simulation error that would be introduced by Monte Carlo simulations.

Figure 2 shows the normalized expected SFS obtained from Equation 42, where higher frequency classes have been aggregated (*i.e.*, lumped) for different values of ψ and ρ. In line with previous findings, both multiple mergers and population growth lead to an excess in singletons (Durrett and Schweinsberg 2005; Eldon *et al.* 2015; Niwa *et al.* 2016). Furthermore, this excess increases as sample size increases under the psi-coalescent (Figure S1 in File S3), while it decreases for the Kingman coalescent independent of the presence or absence of exponential growth. These qualitative differences stem from the different footprints reproductive skew and exponential growth leave on a genealogy. While the latter is a simple rescaling of branch lengths, leaving the topology unchanged, multiple-merger coalescents by definition affect the topology of the genealogical tree (Eldon *et al.* 2015). In particular, when ψ is large, adding samples will disproportionally increase the number of external branches T_1, such that the genealogy will become more star-like, rendering disproportionately more singletons.

Though the excess in singletons characterizes either process, their higher frequency classes will typically differ (Eldon *et al.* 2015). When both processes—reproductive skew and exponential growth—act simultaneously though, their joint effects on the SFS (nontrivially) combine. As expected, increasing growth under the psi-coalescent further exacerbates the excess in singletons. More generally, exponential growth leads to a systematic left shift in the SFS toward lower frequency classes that is independent of ψ. Increasing ψ, on the other hand, changes the SFS—and in particular the higher frequency classes—nonmonotonically even if there is no population growth (Figure S2 in File S3). Interestingly, for $\rho = 0$, the last entry of the normalized expected SFS $\mathbb{E}[\eta_{k-1}]$ initially increases with ψ, and takes an intermediate maximum, decreases monotonically until $\psi \approx 0.85$, peaks again, and then quickly reduces to 0 as ψ approaches 1. This effect prevails as sample size increases (Figure S2 in File S3), even though the intermediate maximum shifts slightly toward lower ψ. However, this intermediate maximum is effectively washed out by increasing ρ, such that the second peak becomes the maximum. Furthermore, the shape of the peak becomes more pronounced as the sample size increases. Thus, reproductive skew and exponential growth leave complex and distinct genomic footprints on the SFS. While, in theory, population growth and reproductive skew should be identifiable, in practice this strongly depends on sample size (Spence *et al.* 2016). In the next section, we will assess the accuracy of our joint estimation framework, and perform extensive validation (Equation 14) on large-scale simulated data.

Simulated coalescent and demographic models

To test our inference framework, we followed two different simulation approaches, each corresponding to two biological limiting cases. In both, data were simulated for the Cartesian product set over $\psi = \{0, 0.15, 0.3, 0.45, 0.6, 0.75, 0.9\}$, $\rho = \{0, 1, 10, 100\}$, $k = \{20, 50, 100, 200\}$, and $s = \{100, 1\,000, 10\,000\}$ per locus over $10,000$ replicates each. In order to make results comparable across different coalescent models, and, thus, across different values of ψ and ρ, we calculated the population-scaled mutation rate θ based on Watterson's estimator (Watterson 1975),

$$\theta = \frac{2s}{\mathbb{E}\left[T_{\text{tot}}^{\psi,\rho}\right]} \tag{45}$$

for a fixed number of segregating sites s over the expected total tree length under the generating coalescent model (given by the denominator in Equation 42). Note that T_{tot} decreases with both increasing ψ and ρ. Thus, keeping s constant implies that θ effectively increases with ψ and ρ. We will discuss the latter point in more detail in light of the results below. Data were simulated for the following two underlying genetic architectures:

Case 1 (Independent-sites simulations): Under the Poisson random field assumption, the underlying coalescent tree at each site is independent (Sawyer and Hartl 1992; Bhaskar *et al.* 2015). Thus, by averaging over independent

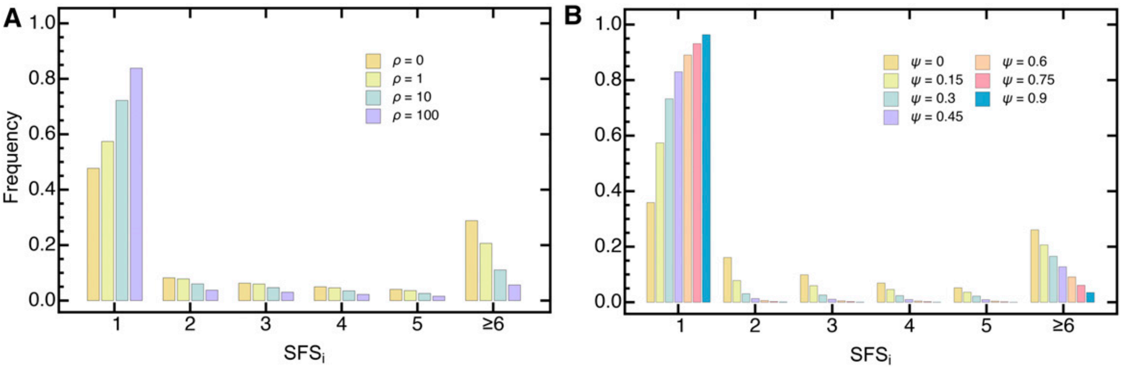

Figure 2 The normalized expected (lumped) SFS for the psi-coalescent for an exponentially growing population (Equation 18) with sample size $k = 20$ (A) for different values of ρ and fixed $\psi = 0.15$, and (B) for different values of ψ and fixed $\rho = 1$. The sixth entry in the SFS contains the aggregate of the higher frequency classes.

realizations of the (shared) underlying coalescent process, the SFS can be obtained by randomly drawing from a multinomial distribution such that $\boldsymbol{\eta} \sim \text{Multinomial}(s, \boldsymbol{\varphi})$.

Case 2 (Whole-genome simulations): In this scenario, we consider a genome of $\ell = 100$ independent loci, where sites within each locus share the same genealogy (*i.e.*, coalescent tree). Thus, for each locus, we draw a random genealogy according to Equations 10 and 33, superimpose $s \sim \text{Poisson}(\theta/2)$ random mutations onto the ancestral tree by multinomial sampling, and aggregate the individual locus SFS into a single genome-wide SFS.

Finally, data sets where $s = \eta_1$ (*i.e.*, where all segregating sites were singletons) were discarded, and simulated again since these do not allow the underlying coalescent parameter and demographic history to be identified. Note that both types of simulations are merely for checking the robustness and accuracy of the inference framework (and might not always necessarily be biologically realistic). A discussion of the independence assumption between loci and its biological implications is given below.

Accuracy of the joint estimation framework

Next, we evaluated the accuracy of the joint estimation framework by means of the mean absolute deviation (MAD) $MAD = 1/n\sum_{i=1}^{n}|x_i - \hat{x}_i|$, the mean deviation (MD) $MD = 1/n\sum_{i=1}^{n}x_i - \hat{x}_i$, the mean squared error (MSE) $MSE = 1/n\sum_{i=1}^{n}(x_i - \hat{x}_i)^2$, and the median deviation (MDD), where x and $\hat{x}$ denote the true and the estimated parameter, respectively. If not stated otherwise, results in the main text are shown for the default parameters $k = 100$ and θ (Equation 45), with $s = 10,000$. More results are given in File S3 and File S4. Recall that, for notational simplicity, we will refer to ρ (instead of the compound parameter $\rho\gamma$) when referring to growth rate estimates.

Inference under the independent-sites assumption: First, for a consistency check, we applied our grid-search algorithm to estimate ψ and ρ from an idealized SFS (*i.e.*, where the SFS accurately reflects the expected branch length under the generating coalescent and demographic model $\boldsymbol{\varphi}$, except for distortions due to rounding). An exemplary likelihood surface (Equation 14) for such an idealized SFS is depicted in Figure 3, which shows that the likelihood surface—up to the resolution of the grid point—is smooth, and generally unimodal, and that the true parameters can be estimated accurately. Furthermore, Figure 3 shows that there is generally a negative correlation between ψ and ρ, and that the likelihood surface tends to be steeper and more concentrated along the ψ direction, which suggests that growth rate estimates might show a larger variance, and could, in general, be more difficult to estimate. The steepness of the likelihood surface along the ψ axis tends to increase with ψ, and sample size k, suggesting that the accuracy for estimating ψ should increase as well, while it should become more difficult to estimate ρ accurately.

An exemplary distribution of the jointly inferred maximum likelihood estimates $(\hat{\psi}, \hat{\rho})$ assuming independent sites is shown in Figure 4. The shape of this distribution resembles that of the likelihood surface (Figure 3), indicating that there is some variance—in particular along the ρ-axis—in the maximum likelihood estimates. However, the median and the mean of the distribution match the true underlying coalescent and growth rate parameters (*i.e.*, ψ and ρ) very well, implying that, if sites are independent, $\hat{\psi}$ and $\hat{\rho}$ are unbiased estimators.

Generally, as expected from the shape of the likelihood surface, ψ is estimated with high accuracy and precision, even for large sample sizes ($k = 200$) with only a few segregating sites ($s = 100$) and (nearly) independent of ρ (Figure 5A, Figure S3A, Figure S4A in File S3, and Table S1 in File S4). Growth rate estimates $\hat{\rho}$, however, show a larger variance, and, for some parameters—namely large k and small s—might be slightly upwardly biased when both the coalescent parameter and the growth rate are large (Figure 5B, Figure S3B, Figure S4B in File S3, and Table S2 in File S4). Though, as the number of segregating sites increases, this bias vanishes and the variance decreases (Figure S5 in File S3), highlighting that the joint estimation procedure gives asymptotically unbiased estimators.

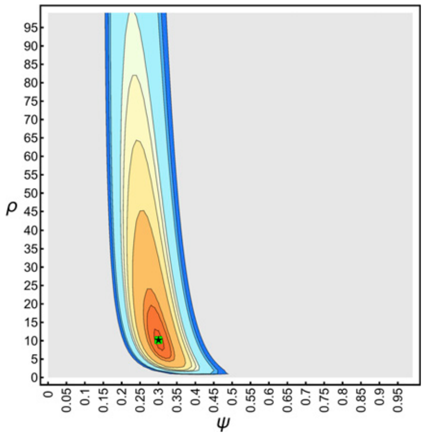

Figure 3 Likelihood surface (Equation 14) of the idealized SFS with $k = 100$, $\psi = 0.3$, $\rho = 10$, and $s = 10,000$. Contours show the 0.95, 0.9675, 0.975, 0.99, 0.99225, 0.9945, 0.99675, 0.999, 0.99945, and 0.9999 quantiles. Likelihoods below the 0.95 quantile are uniformly colored in gray. The green square shows the true ψ and ρ. The black star shows the maximum likelihood estimates $\widehat{\psi}$ and $\hat{\rho}$.

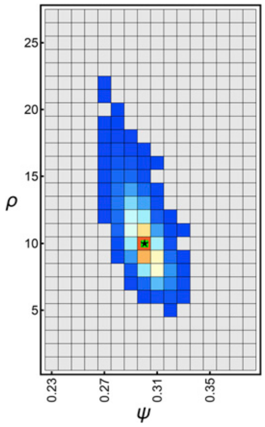

Figure 4 Heatplot of the frequency of the maximum likelihood estimates for 10,000 data sets, assuming independent sites with $k = 100$, $\psi = 0.3$, $\rho = 10$, and θ (Equation 45) with $s = 10,000$. Counts increase from blue to red with gray squares showing zero counts. The green square shows the true ψ and ρ. The black star shows the median (and mean) of the maximum likelihood estimates $\widehat{\psi}$ and $\hat{\rho}$.

For a given s, increasing sample size k increases the signal-to-noise ratio, and, thus, the error in both $\widehat{\psi}$ and $\hat{\rho}$ (Table S1, Table S2, Table S3, and Table S4 in File S4) which is most noticeable in growth rate estimates, in particular when ρ is large (Figure S6 in File S3). This increase in estimation error can (partially) be compensated by increasing the number of segregating sites s (Figure S7 in File S3 and Table S5 in File S4). Specifically, if the true underlying ψ is large (*i.e.*, if the offspring distribution is heavily skewed), an increasing number of segregating sites is needed to accurately infer ρ. However, the total tree length $\overline{T}_{\text{tot}}$—and thus the number of segregating sites s—is expected to decrease sharply with ψ (Eldon and Wakeley 2006), implying that trees tend to become shorter under heavily skewed offspring distributions. This effect could (again, partially) be overcome by increasing sample size since $\overline{T}_{\text{tot}}$—unlike the Kingman coalescent—scales linearly with k as ψ approaches 1 (Eldon and Wakeley 2006). However, population growth will reduce $\overline{T}_{\text{tot}}$ and the number of segregating sites even further.

Calculating θ based on a fixed and constant (expected) number of segregating sites for the assessment of the accuracy of the estimation method evades this problem to some extent. However, by making this assumption, we effectively increase θ in our simulations as ψ and ρ increases. Our results suggest, though, that even more segregating sites than considered in this study (*i.e.*, an even larger θ) would be necessary to infer population growth accurately. Thus, unless (effective) population sizes and/or genome-wide mutation rates are large, it might be very difficult to infer population growth if the offspring distribution is heavily skewed (*i.e.*, if ψ is large). On the other hand, the few studies that have estimated ψ generally found it to be small (Eldon and Wakeley 2006; Birkner *et al.* 2013; Árnason and Halldórsdóttir 2015), leaving it unresolved whether this problem is of any practical importance when studying natural populations.

Inference from genome-wide data: We next tested the accuracy of our joint estimation framework when applied to genome-wide data obtained from $\ell = 100$ independent loci. An exemplary distribution of the jointly inferred maximum likelihood estimates $(\widehat{\psi}, \hat{\rho})$ is depicted in Figure 6, and Figure 7 shows the overall performance of the joint estimation method when applied to genome-wide data. While the whole-genome simulations are designed such that each site in a given locus shares the same underlying genealogy, and, thus, violate the assumption of (statistical) independence between sites, we find that coalescent and growth rate parameters (*i.e.*, ψ and ρ) can be estimated robustly and accurately. In concordance with the independent-sites simulations, the variance in $\widehat{\psi}$ is typically small, whereas $\hat{\rho}$ spreads considerably, and increasingly so if ψ is large. The mean and median of the coalescent parameter and growth rate estimates are again centered around the true value, implying that $\widehat{\psi}$ and $\hat{\rho}$ are unbiased estimators (see also Table S6, Table S7, Table S8, and Table S9 in File S4).

Next, we assessed how the precision of the coalescent and growth rate parameter estimates depends on the number of (independent) loci (*i.e.*, the number of independent coalescent realizations), while keeping the number of segregating sites constant. We find that coalescent and growth rate estimates obtained from a single locus display a huge variance, in particular, when the true underlying growth rate is large (Figure S8 in File S3), warranting caution when interpolating population trends from a single statistical realization as is common practice in studies fitting a multiple-merger coalescent models. Expectedly, the precision of the coalescent and growth rate parameter estimates increases (Figure S9 in File S3 and Table S10 in File S4) when considering estimates obtained from $\ell = 1000$ independent loci (*i.e.*, from independent coalescent realizations), suggesting that sequencing efforts should be put on covering the genome in its entirety

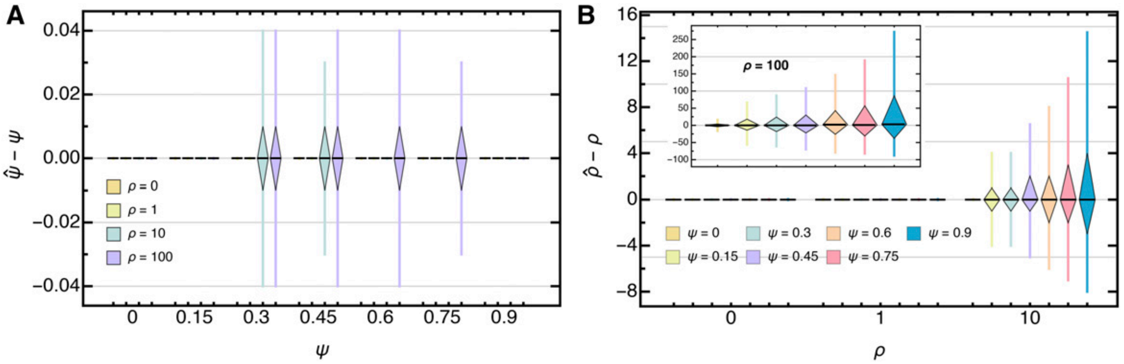

Figure 5 Boxplot of the deviation of the maximum likelihood estimate from the true (A) ψ and (B) ρ for $10,000$ data , assuming independent sites with $k = 100$ and θ (Equation 45) with $s = 10,000$. Boxes represent the interquartile range (*i.e.*, the 50% C.I.) and whiskers extend to the highest/lowest data point within the box ±1.5 times the interquartile range.

rather than on increasing coverage of individual genomic regions.

Distance-based inference and the effect of lumping: As an alternative to the likelihood-based method, Eldon *et al.* (2015) proposed an ABC approach based on a minimum-distance statistic (Equation 15). In this section, we assess the accuracy of $\widehat{\psi}_d$ and $\hat{\rho}_d$ when estimated from d_1 and d_2 distances (*i.e.*, the l_1 and l_2 distance). A surface plot of the l_1 and the l_2 distance is shown in Figure S10 in File S3. We find that, for the l_1 and the l_2 distance, results are comparable to those of the likelihood-based estimates, but generally display a larger variance (Figure S11, Figure S12, and Figure S13 in File S3). Likelihood-based estimates, $\widehat{\psi}_{ML}$, tend to be more accurate across the entire parameter space, though differences between the two are marginal.

Over the majority of the parameter space, the same holds true for $\hat{\rho}_{ML}$. Particularly for small-to-intermediate ψ, the likelihood-based approach outperforms both distance-based approaches considerably (Table S11 and Table S12 in File S4). Interestingly though, for large ψ and ρ (*i.e.*, in the part of the parameter space, where estimating ρ is generally difficult) the l_1 distance approach gives more accurate estimates. When increasing the number of segregating sites, though, the likelihood approach becomes more accurate again, suggesting that the l_1 distance-based approach only outperforms the likelihood-based approach when there is insufficient data (data not shown). These general findings are also upheld when considering genome-wide data (Figure S14 and Figure S15 in File S3). Despite the slightly reduced power as compared to the maximum likelihood approach, our results indicate that, given the asymptotic properties, both the l_1 and the l_2 distance should perform reasonably well when used in a rejection-based ABC analysis.

Finally, we investigated the effect of lumping (*i.e.*, aggregating the higher-frequency classes of the SFS into a single entry after a given threshold i) on the performance of our estimator. In contrast to Eldon *et al.* (2015), who found that lumping can improve the power to distinguish between multiple-merger coalescent models and models of population growth, we find that estimates based on the lumped SFS (using $i = 5$ and $i = 15$) show considerably more error (Table S13 and Table S14 in File S4). While ψ can again be reasonably well estimated, $\hat{\rho}$—in particular when ψ and/or ρ are large—is orders of magnitude more inaccurate when higher frequency classes are lumped. The reason is that, when trying to differentiate between different coalescent or growth models, lumping can reduce the noise associated with the individual higher frequency classes, and, thus, increases the power, provided that the different candidate models show different mean behaviors in the lumped classes (Eldon *et al.* 2015). While this seems to hold true when considering "pure" coalescent or growth models, the joint footprints of skewed offspring distributions and (exponential) population growth are more subtle. In particular, since growth induces a systematic left shift in the SFS toward lower frequency classes, most of the information to distinguish between a psi-coalescent, with or without growth, is lost when aggregated.

Mis-inference of coalescent parameters when neglecting demography

As argued above, both reproductive skew and population growth result in an excess of singletons (*i.e.*, low-frequency mutations) in the SFS. However, topological differences between the two generating processes in the right tail of the SFS allows distinguishing between the two. In particular, fitting an exponential growth model and not accounting for reproductive skewness results in a vastly (and often unrealistically) overestimated growth rate (Eldon *et al.* 2015).

Here, we investigate how coalescent parameter estimates (*i.e.*, $\widehat{\psi}$) are affected when not accounting for (exponential) population growth (*i.e.*, assuming $\rho = 0$) when both processes act simultaneously. As expected, we find that $\widehat{\psi}$ is consistently overestimated (Figure 8) and that the estimation error—independent of ψ—increases with larger (unaccounted for) growth rates. This is because, unless the underlying genealogy is star-shaped (*e.g.*, when $\psi = 1$), growth will always left-shift the SFS, and, hence, increase the singleton class. Thus, when assuming $\rho = 0$, increasing ψ compensates for the "missing" singletons.

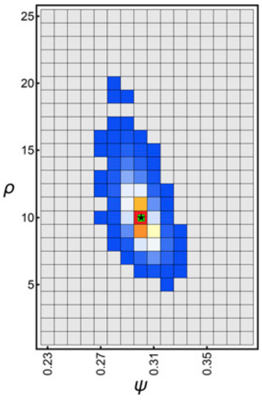

Figure 6 Heatplot of the frequency of the maximum likelihood estimates for 10, 000 whole-genome data sets assuming with $\ell = 100, k = 100$, $\psi = 0.3$, $\rho = 10$, $\gamma = 1.5$ and θ (Equation 45) with $s = 1000$. Counts increase from blue to red with gray squares showing zero counts. The green square shows the true ψ and ρ. The black star shows the median (and mean) of the maximum likelihood estimates $\widehat{\psi}$ and $\hat{\rho}$

Interestingly though, the estimation error changes nonmonotonically with ψ, and, for large ρ, can be as great as twice the value of the true underlying coalescent parameter. Furthermore, for low-to-intermediate ψ, even small growth rates can result in a relative error of up to 23%. Overall, not accounting for demography can lead to serious biases in ψ with broad ecological implications when trying to understand the variation in reproductive success.

Application to sardine data

Finally, we applied our joint inference framework to a derived SFS for the control region of mtDNA in Japanese sardine (*S. melanostictus*; File S5). Niwa *et al.* (2016) recently analyzed this data to test whether the observed excess in singletons was more likely caused by a recent population expansion or by sweepstake reproductive events, and found that the latter is the more likely explanation. However, there is of course no *a priori* reason to believe that both reproductive skew and population growth could not have acted simultaneously.

When estimated jointly, the maximum likelihood estimate is $(\widehat{\psi}, \hat{\rho}) = (0.46, 0)$, which implies considerable reproductive skew, but no (exponential) population growth (Figure 9; see Figure S16 in File S3 for the corresponding l_1 and l_2 distance estimates). While our analysis confirms their results at first glance, there are two points that warrant caution with this interpretation. First, as indicated by the contour lines in the plot, there is some probability that the Japanese sardine population underwent a recent population expansion, though, if it did, it only grew at a very low rate. Second, our inference is based on a single nonrecombining locus (*i.e.*, mtDNA), implying that there is correlation between sites. Our approximation, though, is exact only if there is independence between sites. While violations of the independence assumption seem to be robust on the genome-wide scale (see above; Figure 7), per-locus estimates can vary drastically, and might not be representative for the true underlying coalescent process (Figure S8 in File S3).

Concluding remarks

This study marks the first multiple-merger coalescent with time-varying population sizes derived from a discrete time random mating model, and provides the first in-depth analyses of the joint inference of coalescent and demographic parameters. Since the Kingman coalescent represents a special case of the general class of multiple-merger coalescents (Donnelly and Kurtz 1999; Pitman 1999; Sagitov 1999; Schweinsberg 2000; Spence *et al.* 2016), it is interesting and encouraging to see that our analytical results—*i.e.*, the time-change function (Equation 33) and the first expected coalescence times (Equation 44)—are generalizations of results derived for the Kingman coalescent (Griffiths and Tavaré 1998; Polanski and Kimmel 2003). In fact, when growth rates are measured within the corresponding coalescent framework (*e.g.*, as $\rho\gamma$ for the psi-coalescent or ρ for the Kingman coalescent), these formulas should extend to other, more general multiple-merger coalescents. This also holds true for the challenges arising when calculating the normalized expected SFS (Equation 13), which is central to estimating coalescent parameters and growth rates: Because of catastrophic cancellation errors—due mainly to summing terms involving large binomial coefficients and numerical

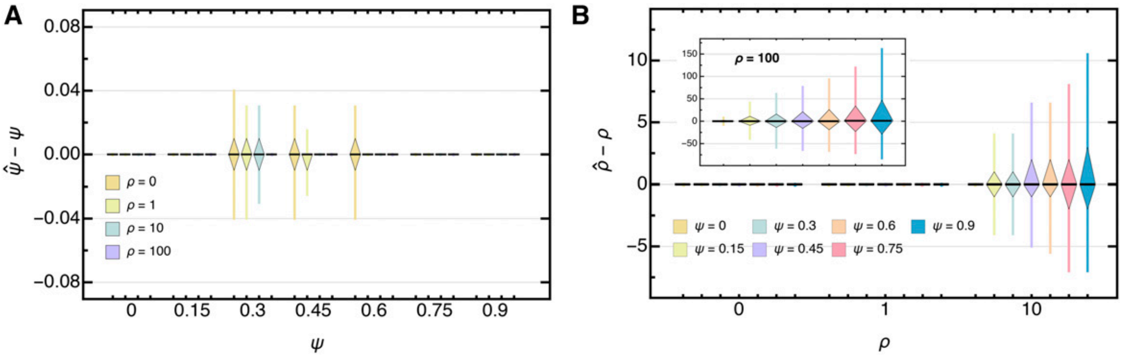

Figure 7 Boxplot of the deviation of the maximum likelihood estimate from the true (A) ψ and (B) ρ for 10, 000 whole-genome data sets with $\ell = 100, k = 100$, $\gamma = 1.5$, and θ (Equation 45) with $s = 1000$. Boxes represent the interquartile range (*i.e.*, the 50% C.I.), and whiskers extend to the highest/lowest data point within the box ±1.5 times the interquartile range.

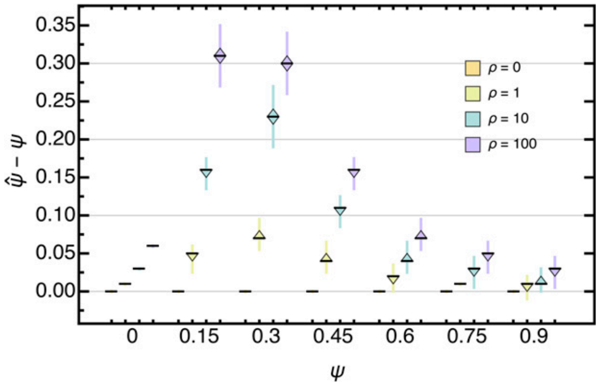

Figure 8 Boxplot of the deviation of the maximum likelihood estimate from the true ψ for 10, 000 data sets, assuming independent sites with $k = 100$ and θ (Equation 45) with$s = 10,000$ when not accounting for population growth. Boxes represent the interquartile range (*i.e.*, the 50% C.I.), and whiskers extend to the highest/lowest data point within the box ±1.5 times the interquartile range.

representations of the exponential integral Ei(x) with alternating signs—computations have to be carried out using multi-precision libraries (Spence *et al.* 2016).

While both ψ and ρ can generally be estimated precisely, accurate estimation of the latter requires sufficient information (*i.e.*, a large number of segregating sites), especially when offspring distributions are heavily skewed (*i.e.*, if ψ is large). However, since strong recurrent sweepstake reproductive events—analogous to recurrent selective sweeps—constantly erase genetic variation (*i.e.*, reduce the number of segregating sites), there might be little power to accurately infer ρ in natural populations in these cases. In accordance with previous findings derived for the Kingman coalescent (Terhorst and Song 2015), increasing sample size does not improve the accuracy of demographic inference (*i.e.*, estimating ρ) for a fixed (expected) number of segregating sites s. However, unlike in the Kingman coalescent, where s increases logarithmically with sample size, genetic variation in ψ increases linearly for large ψ, which could offset—or at least hamper—this effect.

More importantly, these results have proven to be robust to violations of the assumptions underlying the approximate likelihood framework (Equation 14), namely, that the expectation of a ratio can be approximated by the ratio of two expectations (*i.e.*, $\varphi_i^{(k)}$), allowing ψ and ρ to be estimated accurately on a genome-wide scale. Interestingly, the performance of the estimators seemed to improve when considering more independent loci (while keeping the number of segregating sites constant; see also Figure S8, Figure S9 in File S3, and Table S10 in File S4). Note, though, that we have used a very simplistic genetic architecture, in particular one where sites within each locus are maximally dependent, and there is no correlation among genealogies across different loci (*i.e.*, where loci are independent). While these assumptions might be met for some loci and sites, they generally mark the endpoint of a continuum of correlations, and might not always be biologically realistic. Importantly, these linkage disequilibria (*i.e.*, the extent of statistical independence between sites) depend not only on the rate of recombination, but also on the specifics of the reproduction parameters (*i.e.*, ψ)—and can potentially be elevated, despite frequent recombination, or largely absent, despite infrequent recombination in the MMC setting (Eldon and Wakeley 2008; Birkner *et al.* 2012), potentially biasing results. For instance, when trying to estimate the duration and the rate of exponential growth under the Kingman coalescent, Bhaskar *et al.* (2015) found that linkage equilibria cause the approximate likelihood approach (Equation 14) to become increasingly inaccurate, and, thus, bias estimates. Likewise, Schrider *et al.* (2016) recently found that linked positive selection can severely bias demographic estimates. While their analyses assumed a Kingman framework, positive selection and recurrent selective sweeps typically result in multiple merger events (Durrett and Schweinsberg 2004, 2005; Neher and Hallatschek 2013; Schweinsberg 2017). Thus, if neutral regions are tightly linked to a selected site they will—at least partially—share the genealogical relationship with the selected region, and potentially skew inference. Similarly, large reproductive skew (*i.e.*, large ψ) will induce correlations between coalescent trees across loci (*i.e.*, linkage), which will reduce the number of "effective independent loci," suggesting an increased variance in both coalescent parameter and growth rate estimates. However, due to the lack of explicit coalescent simulators that allow for multiple-mergers, nonconstant population sizes, and varying recombination rates, the effects of linkage on the joint estimation of coalescent and demographic parameters cannot directly be assessed, and remain open for future research.

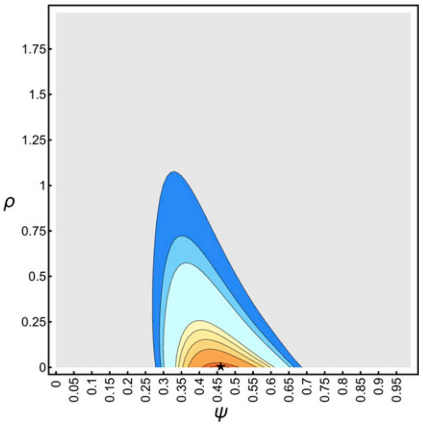

Figure 9 Likelihood surface (Equation 14) of the unfolded SFS (given the ML rooted tree) of the sardine mtDNA sequences with $k = 106$ and $s = 78$. Contours show the 0.95, 0.9675, 0.975, 0.99, 0.99225, 0.9945, 0.99675, 0.999, 0.99945, and 0.9999 quantiles. Likelihoods below the 0.95 quantile are uniformly colored in gray. The black star shows the maximum likelihood estimates $\widehat{\psi} = 0.46$ and $\hat{\rho} = 0$.

Despite the fact that our model here considers organisms with skewed offspring distributions under neutrality owing to the specifics of their reproductive biology, increasing ψ is tantamount to increasing the strength of positive selection under a non-neutral model, which is thus relevant to a very broad class of organisms indeed. It is important to note that while both processes—selection and sweepstake reproductive events—have a similar effect on the SFS (*i.e.*, an excess of low-frequency alleles and a slight increase in high-frequency alleles), there are of course vast qualitative differences in the underlying processes and their causes. First, in the presence of selection, offspring no longer choose their parents at random, such that selected alleles need to be tracked along the genealogy (*e.g.*, see the ancestral selection graph under the Kingman coalescent (Krone and Neuhauser 1997), or under the Λ-coalescent (Etheridge *et al.* 2010)). Second, similar to the effects of demography, sweepstake reproductive events should have a genome-wide impact, whereas traces of selection should remain local, unless selection is very strong, such that only a single individual gives rise to the entire next generation. Thus, it should in principle be possible to discriminate between the two processes, though, also analogous to demography, it will be important to investigate the conditions under which positively selected loci will be expected to reside in the tails of genomic distributions under such models (see Thornton and Jensen 2007).

Overall, our analyses emphasize the importance of accounting for demography and illuminates the serious biases that can arise in the inferred coalescent model if ignored. Such bias can have broad implications on inferred patterns of genetic variation (Eldon and Wakeley 2006; Tellier and Lemaire 2014; Niwa *et al.* 2016), including misguiding conservation efforts (Montano 2016), and obscuring the extent of reproductive skew.

Finally, most of the current analytical and computational tools have been derived and developed under the Kingman coalescent. In order to achieve the overall aim of generalizing the Kingman coalescent model (Wakeley 2013), these tools, though often computationally challenging, need to be extended. Great efforts have recently been undertaken toward developing a statistical inference framework, allowing for model selection (Birkner and Blath 2008; Eldon 2011; Birkner *et al.* 2011, 2012, 2013; Steinrücken *et al.* 2013; Eldon *et al.* 2015; Spence *et al.* 2016). By setting up a discrete-time random mating model, and deriving the ancestral process, along with providing the analytical tools necessary to enable the joint inference of offspring distribution and demography, this study makes an important contribution toward this goal.

Acknowledgments

We thank Kristen Irwin, Adamandia Kapopoulou, Martin Möhle, Sylvain Mousset, and Meike Wittmann for helpful discussion and comments on earlier versions of this manuscript, Hiro-Sato Niwa for providing the sardine data, and two anonymous reviewers for their insightful comments and constructive criticism, which greatly helped to improve this manuscript. This project was funded by grants from the Swiss National Science Foundation (FNS) and a European Research Council (ERC) Starting Grant to J.D.J.

Literature Cited

Achaz, G., 2009 Frequency spectrum neutrality tests: one for all and all for one. Genetics 183: 249–258.

Árnason, E., and K. Halldórsdóttir, 2015 Nucleotide variation and balancing selection at the *Ckma* gene in Atlantic cod: analysis with multiple merger coalescent models. PeerJ 3: e786.

Bhaskar, A., A. G. Clark, and Y. S. Song, 2014 Distortion of genealogical properties when the sample is very large. Proc. Natl. Acad. Sci. USA 111: 2385–2390.

Bhaskar, A., Y. R. Wang, and Y. S. Song, 2015 Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Res. 25: 268–279.

Birkner, M., and J. Blath, 2008 Computing likelihoods for coalescents with multiple collisions in the infinitely many sites model. J. Math. Biol. 57: 435–465.

Birkner, M., J. Blath, and M. Steinrücken, 2011 Importance sampling for lambda-coalescents in the infinitely many sites model. Theor. Popul. Biol. 79: 155–173.

Birkner, M., J. Blath, and B. Eldon, 2012 An ancestral recombination graph for diploid populations with skewed offspring distribution. Genetics 193: 255–290.

Birkner, M., J. Blath, and B. Eldon, 2013 Statistical properties of the site-frequency spectrum associated with λ-coalescents. Genetics 195: 1037–1053.

Bolthausen, E., and A.-S. Sznitman, 1998 On Ruelle's probability cascades and an abstract cavity method. Commun. Math. Phys. 197: 247–276.

Cannings, C., 1974 The latent roots of certain Markov chains arising in genetics: a new approach, i. haploid models. Adv. Appl. Probab. 6: 260–290.

Donnelly, P., and T. G. Kurtz, 1999 Particle representations for measure-valued population models. Ann. Probab. 27: 166–205.

Durrett, R., and J. Schweinsberg, 2004 Approximating selective sweeps. Theor. Popul. Biol. 66: 129–138.

Durrett, R., and J. Schweinsberg, 2005 A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stoch. Proc. Appl. 115: 1628–1657.

Eldon, B., 2011 Estimation of parameters in large offspring number models and ratios of coalescence times. Theor. Popul. Biol. 80: 16–28.

Eldon, B., and J. Wakeley, 2006 Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics 172: 2621–2633.

Eldon, B., and J. Wakeley, 2008 Linkage disequilibrium under skewed offspring distribution among individuals in a population. Genetics 178: 1517–1532.

Eldon, B., M. Birkner, J. Blath, and F. Freund, 2015 Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics 199: 841–856.

Etheridge, A. M., R. C. Griffiths, and J. E. Taylor, 2010 A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. Theor. Popul. Biol. 78: 77–92.

Fay, J. C., and C.-I. Wu, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405–1413.

Ferretti, L., M. Pérez-Enciso, and S. E. Ramos-Onsins, 2010 Optimal neutrality tests based on the frequency spectrum. Genetics 186: 353–365.

Fisher, R., 1930 *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford.

Grant, W. S., E. Árnason, and B. Eldon, 2016 New DNA coalescent models and old population genetics software. ICES J. Mar. Sci. 73: 2178–2180.

Griffiths, R., and S. Tavaré, 1998 The age of a mutation in a general coalescent tree. Commun. Stat. Stoch. Models 14: 273–295.

Griffiths, R. C., and S. Tavaré, 1994 Sampling theory for neutral alleles in a varying environment. Philos. Trans. R. Soc. Lond. B Biol. Sci. 344: 403–410.

Hedgecock, D., 1994 Does variance in reproductive success limit effective population sizes of marine organisms? pp. 1222–1344 in *Genetics and Evolution of Aquatic Organisms*, edited by A. Beaumont. Chapman and Hall, London.

Hedgecock, D., and A. I. Pudovkin, 2011 Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and commentary. Bull. Mar. Sci. 87: 971–1002.

Huillet, T., and M. Möhle, 2013 On the extended Moran model and its relation to coalescents with multiple collisions. Theor. Popul. Biol. 87: 5–14.

Irwin, K. K., S. Laurent, S. Matuszewski, S. Vuilleumier, L. Ormond *et al.*, 2016 On the importance of skewed offspring distributions and background selection in virus population genetics. Heredity 117: 393–399.

Kaj, I., and S. M. Krone, 2003 The coalescent process in a population with stochastically varying size. J. Appl. Probab. 40: 33–48.

Kingman, J. F. C., 1982a The coalescent. Stoch. Proc. Appl. 13: 235–248.

Kingman, J. F. C., 1982b On the genealogy of large populations. J. Appl. Probab. 19: 27–43.

Kingman, J. F. C., 1982c Exchangeability and the evolution of large populations, pp. 97–112 in *Exchangeability in Probability and Statistics*, edited by G. Koch, and F. Spizzichino. North-Holland, Amsterdam.

Kingman, J. F. C., 2000 Origins of the coalescent: 1974–1982. Genetics 156: 1461–1463.

Koskela, J., P. Jenkins, and D. Spanò, 2015 Computational inference beyond Kingman's coalescent. J. Appl. Probab. 52: 519–537.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. Popul. Biol. 51: 210–237.

Möhle, M., 1998 Robustness results for the coalescent. J. Appl. Probab. 35: 438–447.

Möhle, M., 1999 Weak convergence to the coalescent in neutral population models. J. Appl. Probab. 36: 446–460.

Möhle, M., 2002 The coalescent in population models with time-inhomogeneous environment. Stoch. Proc. Appl. 97: 199–227.

Möhle, M., and S. Sagitov, 2001 A classification of coalescent processes for haploid exchangeable population models. Ann. Probab. 29: 1547–1562.

Montano, V., 2016 Coalescent inferences in conservation genetics: should the exception become the rule? Biol. Lett. 12: 20160211.

Moran, P. A. P., 1958 Random processes in genetics. Proc. Camb. Philos. Soc. 54: 60.

Moran, P. A. P., 1962 *The Statistical Processes of Evolutionary Theory*. Clarendon Press, Oxford.

Neher, R. A., and O. Hallatschek, 2013 Genealogies of rapidly adapting populations. Proc. Natl. Acad. Sci. USA 110: 437–442.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Niwa, H.-S., K. Nashida, and T. Yanagimoto, 2016 Reproductive skew in Japanese sardine inferred from DNA sequences. ICES J. Mar. Sci. 73: 2181–2189.

Nordborg, M., 1997 Structured coalescent processes on different time scales. Genetics 146: 1501–1514.

Pitman, J., 1999 Coalescents with multiple collisions. Ann. Probab. 27: 1870–1902.

Polanski, A., and M. Kimmel, 2003 New explicit expressions for relative frequencies of single-nucleotide polymorphisms with application to statistical inference on population growth. Genetics 165: 427–436.

Polanski, A., A. Bobrowski, and M. Kimmel, 2003 A note on distributions of times to coalescence, under time-dependent population size. Theor. Popul. Biol. 63: 33–40.

Sagitov, S., 1999 The general coalescent with asynchronous mergers of ancestral lines. J. Appl. Probab. 36: 1116–1125.

Sawyer, S. A., and D. L. Hartl, 1992 Population genetics of polymorphism and divergence. Genetics 132: 1161–1176.

Schrider, D. R., A. G. Shanku, and A. D. Kern, 2016 Effects of linked selective sweeps on demographic inference and model selection. Genetics 204: 1207–1223.

Schweinsberg, J., 2000 Coalescents with simultaneous multiple collisions. Electron. J. Probab. 5: 1–50.

Schweinsberg, J., 2017 Rigorous results for a population model with selection II: genealogy of the population. *Electron. J. Probab.* 22: 54.

Spence, J. P., J. A. Kamm, and Y. S. Song, 2016 The site frequency spectrum for general coalescents. Genetics 202: 1549–1561.

Steinrücken, M., M. Birkner, and J. Blath, 2013 Analysis of DNA sequence variation within marine species using beta-coalescents. Theor. Popul. Biol. 87: 15–24.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.

Tellier, A., and C. Lemaire, 2014 Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Mol. Ecol. 23: 2637–2652.

Terhorst, J., and Y. S. Song, 2015 Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proc. Natl. Acad. Sci. USA 112: 7677–7682.

Thornton, K. R., and J. D. Jensen, 2007 Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737–750.

Wakeley, J., 2009 *Coalescent Theory: An Introduction*. Roberts & Company Publishers, Greenwood Village, CO.

Wakeley, J., 2013 Coalescent theory has many new branches. Theor. Popul. Biol. 87: 1–4.

Wakeley, J., and T. Takahashi, 2003 Gene genealogies when the sample size exceeds the effective size of the population. Mol. Biol. Evol. 20: 208–213.

Watterson, G., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7: 256–276.

Wilkinson-Herbots, H. M., 1998 Genealogy and subpopulation differentiation under various models of population structure. J. Math. Biol. 37: 535–585.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Communicating editor: R. Nielsen