Journal of Statistical Physics
https://doi.org/10.1007/s10955-018-2128-4

Exact Steady-State Distributions of Multispecies Birth–Death–Immigration Processes: Effects of Mutations and Carrying Capacity on Diversity

Renaud Dessalles[1,2] · **Maria D'Orsogna**[1,3] · **Tom Chou**[1,4]

Received: 14 May 2018 / Accepted: 30 July 2018

Abstract
Stochastic models that incorporate birth, death and immigration (also called birth–death and innovation models) are ubiquitous and applicable to many problems such as quantifying species sizes in ecological populations, describing gene family sizes, modeling lymphocyte evolution in the body. Many of these applications involve the immigration of new species into the system. We consider the full high-dimensional stochastic process associated with multispecies birth–death–immigration and present a number of exact and asymptotic results at steady state. We further include random mutations or interactions through a carrying capacity and find the statistics of the total number of individuals, the total number of species, the species size distribution, and various diversity indices. Our results include a rigorous analysis of the behavior of these systems in the fast immigration limit which shows that of the different diversity indices, the species richness is best able to distinguish different types of birth–death–immigration models. We also find that detailed balance is preserved in the simple noninteracting birth–death–immigration model and the birth–death–immigration model with carrying capacity implemented through death. Surprisingly, when carrying capacity is implemented through the birth rate, detailed balance is violated.

Keywords Birth–death–immigration processes · Multispecies · Steady-state probability distributions · Diversity · Mutations

1 Introduction

In recent years, stochastic Birth–death–immigration (BDI) models have emerged as effective descriptions of the evolution of multi-species populations. BDI models assume that each indi-

✉ Tom Chou
tomchou@ucla.edu

[1] Department of Biomathematics, University of California, Los Angeles, CA 90095-1766, USA

[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[3] Department of Mathematics, California State University, Northridge, CA 91330, USA

[4] Department of Mathematics, University of California, Los Angeles, CA 90095-1555, USA

Published online: 07 August 2018

vidual belongs to a given "species" and undergoes a classical birth–death process; offspring populate the same species as their parent, while new species are introduced via immigration and/or mutation. The body of work on BDI models in the mathematical, ecological and biological literature is rich, and many results have been independently discovered in the context of different disciplines. Arguably, the first BDI model can be traced to [22] who described the evolution of different alleles in a genetic population. Later, similar tools were applied to ecology in the context of the "Neutral Theory of Biodiversity" [18,23,26,34], where BDI models were used to study the abundance distribution of island populations that undergo continuous immigration from the mainland. BDI models have also been used under the name birth, death and innovation models by Karev et al. [21] to describe gene domain family size in genomes. Here, each domain is part of a family, and can be duplicated or deleted; new domains of new families can be added to the genome via horizontal gene transfer. Desponds et al. [10] and Lythe et al. [25] have instead employed BDI formalisms to study lymphocyte populations in an organism. T-cells expressing the same surface receptor are assumed to belong to the same "clone" (the species). Each T-cell can divide, generating receptor-identical daughter cells, and die through apoptosis. In this context, immigration is represented by the export of new naïve T-cells from the thymus. Due to the large number of theoretically possible T-cell clonotypes that can be generated, with estimates ranging from 10^{15}–10^{20} [27], one can assume that each new export almost surely generates a new clone rather than contribute to an existing clonotype. Another application of BDI models arises in the study of microbiota populations in the gut of metazoa [31]. Finally, counting "clones" is also used in stochastic models of nucleation, where a high- or infinite-dimensional distribution function can be used to describe states comprised of certain numbers of clusters (the "clones") of specific size [8,12,13]. In the rest of this paper, we will use both "individuals" (or "particles") and "species" to describe the two types of quantities (individuals of a given species and the number of species of a given size) in all of the above-mentioned examples.

Note that we will focus exclusively on "neutral" BDI representations in the sense of the Neutral Theory of Biodiversity [4,18], that is all individuals within a population are subject to the same birth and death rates so that there is no fitness difference in the population. Our first model is the simple BDI (sBDI) model where each individual evolves independently of all others and where the only possible processes are birth, death and immigration. The second model (BDIM) further includes mutations, whereby the dynamics of each individual is still uncoupled from that of others, but where new species can arise via mutations. The last model (BDICC) includes a carrying capacity on the death rate to represent the sharing of limited resources. In this case, the dynamics of each individual is coupled to that of others, and the overall mathematical analysis is more complex. Thus, for simplicity, when including a carrying capacity term, we exclude mutations. The three major BDI processes we will analyze are depicted in Fig. 1.

Since measures of diversity in a population are also of significant interest in ecology [7,9,29], we will also analyze species diversity through three commonly used indexes [28]: the species richness (the total number of species in the system), Shannon's entropy, and Simpson's diversity index, and we will contrast and compare these quantities among the different models.

The goal of this paper is to provide an accessible, yet rigorous, theoretical analysis of each of the three types of BDI models outlined above. In particular, we determine the conditions for the existence of an equilibrium distribution and derive analytical expressions for the steady state distributions of the total number of individuals, the numbers if clones of each size, the total number of species in the system, as well as the expected species diversities predicted by model. Some results presented here can be recovered from previous work.

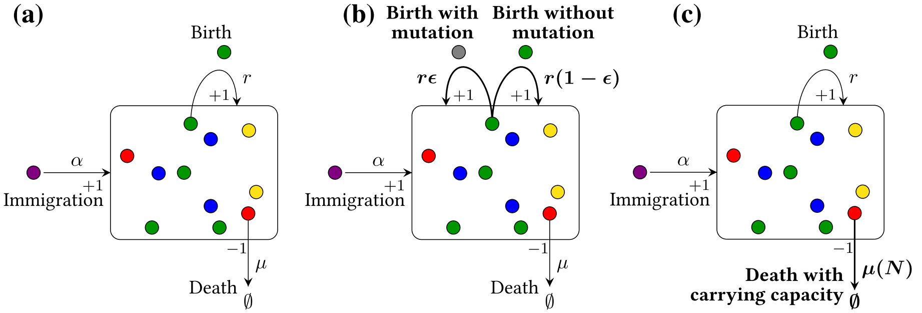

Fig. 1 Schematic of various birth–death–immigration processes. Three distinct variants are considered, including **a** simple birth–death–immigration (sBDI) without mutation, where r, μ are constants, **b** birth–death–immigration with mutation (BDIM) where r, μ are constants and mutation rate $\epsilon > 0$, and **c** birth–death–immigration without mutation but with carrying capacity (BDICC) where r is constant and $\mu = \mu(N)$ is an increasing function of total population N. We will also analyze a variant of the BDICC model, the BDICC-bis model, where μ is constant but the growth $r = r(N)$ is assumed to be a decreasing function of the total population N

More precisely, time-dependent versions of our sBDI model can be found in Karlin and McGregor [22], Travaré [33], Lambert [23] and one particular version the BDIM model (with somatic mutations) is described in [23]. In each case, it is possible to recover the steady state distributions of the total population and the total number of species by evaluating the infinite-time limits of their results.

Our work provides a number of additional results in the steady state limit: (i) the theoretical analysis of an interacting BDI model with carrying capacity is new, to the best of our knowledge; (ii) we provide "full probability distributions" that completely describe the properties of each model and that can be used to derive general quantities of interest, (in particular, the moments of the species counts); (iii) we analytically quantify diversity indices predicted by each model; (iv) we provide systematic quantitative comparisons between the models and (v) we derive simpler limiting forms of our results in the large immigration rate limit. A summary of all our results can be found in Table 1 in the general case and in Table 2 for the fast immigration limit. The interested reader will find more details of the methods and the proofs of the derivation of these expressions in the Mathematical Appendix.

2 Basic Definitions

In this section we outline some general assumptions and introduce the mathematical notation to describe our three BDI models. First, we assume new individuals immigrate to the system following a Poisson point process of rate α, i.e. the time interval between successive immigration events is given by a random variable exponentially distributed with rate α. Each arriving individual will define a new species not yet present in the system. The random variable representing the total number of individuals in the system will be denoted by N and the total number of species by C. We consider both "particle-count" and "species-count" representations (n_i and c_k respectively) of the system: in the particle count representation, n_i (with $1 \leq i \leq C$) represents the number of individuals in the ith species; in the species-count representation, c_k (with $k \geq 1$) represents the number of species having exactly k individuals. In principle, there can be species with infinite population and hence both n_i and the index

Table 1 Table summarizing our analytical results

	Simple birth–death–immigration model (sBDI)	Birth–death–immigration model with mutation (BDIM)	Birth–death–immigration model with carrying capacity (BDICC)
Defn.	$u \equiv \alpha/r,\, v \equiv r/\mu$ $Z = (1-v)^{-u}$ $f_k(x) = \log\left[\frac{x+k}{k}\right]/(x+k)$	$u = \alpha/r,\, v = r/\mu$ $p \equiv \frac{r}{\mu}(1-\epsilon)$	$u \equiv \alpha/r,\, v(x) \equiv r/\mu(x)$ $Z_{\alpha,r} = \sum_{n=0}^{\infty} \frac{1}{n!} \prod_{k=0}^{n-1} \frac{\alpha+rk}{\mu(k+1)},$ $S_n(x) = \sum_{k=0}^{x-1} (\alpha+rk)^{-n}$ $f_k(x) = \log\left[\frac{x+k}{k}\right]/(x+k)$
Cond.	$v < 1$	$v < 1$	$\lim_{x\to\infty} v(x) < 1$ (sufficient cond.*)
N	$N \sim \mathrm{NegBinom}(u, v)$	$N \sim \mathrm{NegBinom}(u, v)$	$P(N) = \frac{1}{Z_{\alpha,r}} \frac{1}{N!} \prod_{k=0}^{N-1} (u+k)\, v(k+1)$
$\vec{c}$	$P(\vec{c}) = \frac{1}{Z} \frac{u^C v^N}{\prod_{i=1}^{\infty} i^{c_i} c_i!}$	N.A.	$P(\vec{c}) = \frac{u^C}{Z_{\alpha,r}} \prod_{n=1}^{N} \frac{v(n)}{\prod_{i=1}^{\infty} i^{c_i} c_i!}$
c_k	$c_k \sim \mathrm{Poisson}\left(\frac{uv^k}{k}\right)$	$\mathbb{E}[c_k] = \frac{uv}{1-v}(1-p)\frac{p^{k-1}}{k}$ $\mathrm{var}[c_k] = \mathbb{E}[c_k] + \epsilon \frac{uv^2 p^{2(k-1)}}{(1-v)^2 k^2}$	$\mathbb{E}[c_k] = \frac{u}{k}\mathbb{E}\left[\prod_{m=1}^{k} v(N+m)\right]$
C	$C \sim \mathrm{Poisson}\left(u \log\left[\frac{1}{1-v}\right]\right)$	$\mathbb{E}[C] = \frac{uv}{1-v}\frac{1-p}{p}\log\frac{1}{1-p}$ $\mathrm{var}[C] = \mathbb{E}[C]\left[\frac{\mathbb{E}[C]}{u} + 1 + \log(1-p)\right]$	$\mathbb{E}[C] = \alpha\mathbb{E}[S_1(N)]$ $\mathrm{var}[C] = \mathbb{E}[C](1-\mathbb{E}[C])$ $+\alpha^2\mathbb{E}\left[(S_1(N))^2 - S_2(N)\right]$
$\vec{n}$	$n_i \sim \mathrm{LogSeries}(v)$	$n_i \sim \mathrm{LogSeries}(p)$	N.A.
H	$\mathbb{E}[H] = u\sum_{k=1}^{\infty} v^k \mathbb{E}[f_k(N)]$	$\mathbb{E}[H] \simeq (1-p)\sum_{k=1}^{\infty} p^{k-1}\log\left[\frac{uv}{k(1-v)}\right]$	$\mathbb{E}[H] = u\sum_{k=1}^{\infty}\mathbb{E}\left[f_k(N)\prod_{m=1}^{k} v(N+m)\right]$
S	$\mathbb{E}[S] = 1 - u\sum_{k=1}^{\infty} k v^k \mathbb{E}\left[\left(\frac{1}{N+k}\right)^2\right]$	$\mathbb{E}[S] \simeq 1 - \frac{1}{uv}\frac{1-v}{1-p}$	$\mathbb{E}[S] = 1 - u\sum_{k=1}^{\infty} k\mathbb{E}\left[\frac{\prod_{m=1}^{k} v(N+m)}{(N+k)^2}\right]$

Poisson: Poisson distribution; NegBinom: negative binomial distribution, LogSeries: logarithmic distribution. In each case, the quantities C and N implicitly depend on the vector $\vec{c}$ through Eq. (1). *Indicates a sufficient condition, for a necessary and sufficient condition, see [2, Chapter 1]. The functions $f_k(x)$, $S_1(N) = \sum_{k=0}^{N-1}(\alpha+rk)^{-1}$ and $S_2(N) = \sum_{k=0}^{N-1}(\alpha+rk)^{-2}$ are defined in entries of the first row

Table 2 Table summarizing model results in the fast immigration limit defined by $\alpha = \widetilde{\alpha}\Omega$, $\Omega \to \infty$

	Simple birth–death–immigration model (sBDI)	Birth–death–immigration model with mutation (BDIM)	Birth–death–immigration model with carrying capacity (BDICC)
Cond.	$\Omega \to \infty$	$\Omega \to \infty$	$\Omega \to \infty$ $\mu(x) = \widetilde{\mu}(x/\Omega)$ n^* : positive soln of $x\widetilde{\mu}(x) = \widetilde{\alpha} + r\,x$
N/Ω	$\frac{\widetilde{\alpha}}{\mu - r}$	$\frac{\widetilde{\alpha}}{\mu - r}$	n^*
c_k/Ω	$\frac{\widetilde{\alpha}}{r}\ \frac{(r/\mu)^k}{k}$	$\frac{\widetilde{\alpha}}{\mu - r}\ \frac{\mu - r(1-\epsilon)}{r(1-\epsilon)}\ \frac{\left(\frac{r}{\mu}(1-\epsilon)\right)^k}{k}$	$\frac{\widetilde{\alpha}}{k}\ \frac{r^{k-1}}{\widetilde{\mu}(n^*)^k}$
C/Ω	$\frac{\widetilde{\alpha}}{r} \log\left[\frac{1}{1-r/\mu}\right]$	$\frac{\widetilde{\alpha}}{\mu - r}\ \frac{\mu - r(1-\epsilon)}{r(1-\epsilon)} \log\left[\frac{1}{1-\frac{r}{\mu}(1-\epsilon)}\right]$	$\frac{\widetilde{\alpha}}{r} \log\left[\frac{1}{1-r/\widetilde{\mu}(n^*)}\right]$
n_i	$n_i \sim \mathrm{LogSeries}\,(r/\mu)$	$n_i \sim \mathrm{LogSeries}\left(\frac{r}{\mu}\,(1-\epsilon)\right)$	$n_i \sim \mathrm{LogSeries}\left(r/\widetilde{\mu}(n^*)\right)$
$H/\log\Omega$	1	1	1
S	1	1	1

$H/\log\Omega$ and S are expanded to the first nontrivial term

k are unbounded. The sequence of all numbers $(n_i)_{i\le C}$, the infinite vector $\vec{c} = (c_k)_{k\ge 1}$, as well as N and C, are related by the following expressions:

$$c_k = \sum_{i=1}^{C} \boldsymbol{I}\,(n_i, k) \quad \text{for } k \ge 1, \qquad C = \sum_{k\ge 1} c_k, \qquad N = \sum_{i=1}^{C} n_i = \sum_{k\ge 1} k\,c_k, \tag{1}$$

where $\boldsymbol{I}$ is the indicator function such that $\boldsymbol{I}\,(x, y) = 1$ if $x = y$ and 0 otherwise. Effectively, the first relation in Eq. (1) will count the number of species that carry k individuals. The second relation describes the total number of clones C that are present in the system, while N is the total number of particles in the system. For many applications C and N are large. For example, in humans, the richness of naive T-cells $C \sim 10^6$–10^{10} [24,25].

The particle-count and species-count representations are related since a given sequence $(n_i)_{i\le C}$ corresponds to a unique vector $\vec{c}$ (determined via the first relation in Eq. (1)). However, given a vector $\vec{c}$ one can determine the sequence $(n_i)_{i\le C}$ only up to permutations of the species identities. More information is intrinsically contained in $(n_i)_{i\le C}$ than in $\vec{c}$.

As mentioned, we will also be interested in the statistics of the population diversity, as described by, e.g., Shannon's entropy H and Simpson's diversity index S. These quantities can be defined using either the particle-count or the species-count representations:

$$H = -\sum_{i=1}^{C} \frac{n_i}{N} \log\left[\frac{n_i}{N}\right] = -\sum_{k\ge 1} c_k \frac{k}{N} \log\left[\frac{k}{N}\right] \quad \text{and}$$
$$S = 1 - \sum_{i=1}^{C} \left(\frac{n_i}{N}\right)^2 = 1 - \sum_{k\ge 1} c_k \left(\frac{k}{N}\right)^2. \tag{2}$$

While many variants of Simpson's diversity exist, we have chosen the "Gini-Simpson" index [20] with replacement, also known as the probability of interspecific encounter [19], Gini-Martin, or Blau indices [15], so that more diverse populations have a higher value of S. Our choice also allows for analytic derivations not available for other diversity indices.

We shall analytically derive, whenever possible, probability distributions over all the quantities defined above. Our results will be limited to distributions in steady state. Henceforth, we will define probabilities of a quantity X having a value x as $P(X = x)$, but we will interchangeably also use the imprecise notation $P(X)$ when no ambiguity exists.

After determining results in steady state for each of the three neutral BDI models (sBDI, BDIM, and BDICC) we will also analyze the asymptotically large immigration limit. This regime may relevant to applications where the per-individual immigration rate is higher than their birth and death rates, such as in the case of lymphocyte production and maintenance. In particular, we will assume the immigration rate α that defines the Poisson point process described above to be proportional to a scaling factor Ω (i.e., the immigration rate $\alpha \equiv \widetilde{\alpha}\Omega$ with $\widetilde{\alpha}$ being a proportionality constant) and then study the $\Omega \to \infty$ limit. We will show that as Ω increases, the above quantities also diverge, however, their scaled values

$$N/\Omega, \quad C/\Omega \quad \text{and} \quad (c_k/\Omega)_{k\geq 1}$$

will be shown to converge in distribution. For example, the convergence in distribution of N/Ω to a given constant limit ℓ will be denoted $N/\Omega \xrightarrow[\Omega\to\infty]{\mathcal{D}} \ell$ and, when ℓ is a constant, the convergence can be characterized by

$$\text{for any } \delta > 0, \quad \lim_{\Omega\to\infty} P(|N/\Omega - \ell| < \delta) = 1 \tag{3}$$

(for a more general definition of the convergence in distribution where ℓ is an arbitrary random variable, see [5, Chapter 5]). This type of convergence implies that

$$\mathbb{E}[N/\Omega] \xrightarrow[\Omega\to\infty]{} \ell \quad \text{and} \quad \text{var}[N/\Omega] \xrightarrow[\Omega\to\infty]{} 0.$$

3 Simple Birth–Death–Immigration Model (sBDI)

We start with the neutral and independent simple birth–death–immigration model (sBDI) where individuals are assumed to be identical, subject to the same birth, death and immigration rates (neutral), and where the dynamics of each individual is independent of that of others (independent). Mutations are not included. One of the most immediate applications of this sBDI model is within the study of island biodiversity [18,34] where individuals follow classical birth and death processes, and new species are introduced to the island via immigration. The ensuing species abundances are then determined. The main ingredients of the sBDI model are depicted in Fig. 1a and include (i) immigration, in which an individual of a new species is added to the system at rate α, (ii) birth, in which each individual gives birth to an offspring of the same species at rate r, and (iii) death, where each individual dies and is removed from the system at rate μ.

3.1 Derivation of Steady State Statistics

We now determine the steady-state probability distribution $P(N)$ of the total number of individuals N in the simple BDI model and the full probability distribution $P(\vec{c}) \equiv P(c_1, \ldots, c_k, \ldots)$ at steady-state. This quantity will lead us to the derivation of the marginal steady-state probability distributions $P(c_k)$ and $P(n_i)$ in the individuals and species count representations, respectively. From $P(\vec{c})$ we will also be able to obtain the probability distribution $P(C)$ of the total number of species C at steady-state. Finally, Shannon's entropy and Simpson's diversity index will be calculated.

The total number of particles N is a random variable that follows a birth and death process with non-constant rates $\alpha + rN$ and μN, respectively. The properties of this birth and death process are known (see for instance Méléard [2]); in particular for a finite steady state to exist the condition $r < \mu$ must hold. This constraint implies that death dominates reproduction so that the number of individuals N does not diverge at long times. At steady state, and for $r < \mu$, detail balance leads to the following condition

$$\mu N P(N) = (\alpha + (N-1)r)P(N-1). \tag{4}$$

This equation can be solved iteratively to yield

$$P(N) = \begin{cases} \left(1 - \dfrac{r}{\mu}\right)^{\alpha/r}, & N = 0 \\ P(0)\left(\dfrac{r}{\mu}\right)^{N} \dfrac{1}{N!} \displaystyle\prod_{k=0}^{N-1} \left(\dfrac{\alpha}{r} + k\right), & N \geq 1 \end{cases} \tag{5}$$

which we recognize as a negative binomial distribution with parameters α/r and r/μ, and mean and variance

$$\mathbb{E}[N] = \frac{\alpha/\mu}{1 - r/\mu}, \quad \text{var}[N] = \frac{\alpha/\mu}{(1 - r/\mu)^2}. \tag{6}$$

Equation (5) does not resolve how the subpopulations are distributed within the different species. To determine this distribution we must derive the distribution $P(\vec{c})$ over the species-count vector $\vec{c} = (c_1, \ldots, c_k, \ldots)$ by explicitly writing down all possible BDI events and their relative rates:

$$\begin{array}{llll} & (c_1, c_2, \ldots) \overset{\alpha}{\to} (c_1 + 1, c_2, \ldots) & & \text{Immigration} \\ \text{for } k \geq 1 & (c_1, \ldots, c_k, c_{k+1}, \ldots) \xrightarrow{rkc_k} (c_1, \ldots, c_k - 1, c_{k+1} + 1, \ldots) & & \text{Birth} \\ \text{for } k \geq 2 & (c_1, \ldots, c_{k-1}, c_k, \ldots) \xrightarrow{\mu k c_k} (c_1, \ldots, c_{k-1} + 1, c_k - 1, \ldots) & \rbrace & \text{Death} \\ & (c_1, c_2, \ldots) \xrightarrow{\mu c_1} (c_1 - 1, c_2, \ldots) & & \end{array}$$

Since each clone is populated by k individuals the total clone population is kc_k, within which each cell can duplicate or die with rate r or μ. The overall birth and death rates for all clones of size k are thus given by rkc_k and μkc_k, respectively. We can thus write for every $k \geq 2$,

$$(k-1)\, c_{k-1} \mu P(c_1, \ldots, c_{k-1}, c_k, \ldots) = k\,(c_k - 1)\, r P(c_1, \ldots, c_{k-1} + 1, c_k - 1, \ldots), \tag{7}$$

and for $k = 1$,

$$\mu c_1 P(c_1, c_2, \ldots) = \alpha P(c_1 - 1, c_2, \ldots). \tag{8}$$

As shown in Appendix C.1 for the more general case of the BDICC model, by recursion of Eq. (8) and using Eq. (7), we find

$$P(c_1, \ldots, c_k, \ldots) = P(0, 0, \ldots)\left(\frac{\alpha}{r}\right)^{C} \left(\frac{r}{\mu}\right)^{N} \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!}, \tag{9}$$

with $C = \sum_{k\geq 1} c_k$ and $N = \sum_{k\geq 1} kc_k$ as defined in Eq. (1). The prefactor $P(0, 0, \ldots)$ is simply the normalization constant and can be computed as

$$P(0,0,\ldots)^{-1} = \sum_{\vec{c}} \left(\frac{\alpha}{r}\right)^C \left(\frac{r}{\mu}\right)^N \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!} = \exp\left(\frac{\alpha}{r}\sum_{i=1}^{\infty}\frac{1}{i}\left(\frac{r}{\mu}\right)^i\right) = \left(1-\frac{r}{\mu}\right)^{-\alpha/r}, \tag{10}$$

so that finally

$$P(\vec{c}) = P(c_1,\ldots,c_k,\ldots) = \left(1-\frac{r}{\mu}\right)^{\alpha/r}\left(\frac{\alpha}{r}\right)^C\left(\frac{r}{\mu}\right)^N \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!}. \tag{11}$$

Note that $P(0, 0, \ldots)$ as expressed in Eq. (10) corresponds to the $N = 0$ case in Eq. (5) since the state with no individuals present in the population can only be represented by the configuration $\vec{c} = (0, 0, \ldots)$.

We can now use Eq. (9) to determine the distribution for the total number of species C. To do this, we consider its moment generating function $M_C(\xi)$ defined as the average $\mathbb{E}\left[\exp(\xi C)\right]$

$$M_C(\xi) \equiv \mathbb{E}\left[\exp(\xi C)\right] = \sum_{c_1,\ldots,c_k,\ldots} e^{\xi C}\left(1-\frac{r}{\mu}\right)^{\alpha/r}\left(\frac{\alpha}{r}\right)^C\left(\frac{r}{\mu}\right)^N \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!},$$

with $C = \sum_{k\geq 1} c_k$ and $N = \sum_{k\geq 1} kc_k$. We find

$$M_C(\xi) = \left(1-\frac{r}{\mu}\right)^{(1-e^{\xi})\alpha/r} \sum_{c_1,\ldots,c_k,\ldots}\left(1-\frac{r}{\mu}\right)^{\alpha e^{\xi}/r}\left(\frac{\alpha e^{\xi}}{r}\right)^C\left(\frac{r}{\mu}\right)^N \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!}. \tag{12}$$

Upon comparing Eq. (11) with the terms in the last summation in Eq. (12) we can easily see that the terms within the sum define the probability $P(c_1, \ldots, c_k, \ldots)$ of another simple independent BDI model with immigration rate $\alpha \to \alpha e^{\xi}$. Thus, from normalization, the sum in Eq. (12) is equal to one. By writing $M_C(\xi)$ in the form

$$M_C(\xi) = \exp\left[\left(e^{\xi}-1\right)\frac{\alpha}{r}\log\left(\frac{1}{1-r/\mu}\right)\right],$$

which is a moment generating function of a Poisson random variable with parameter $(\alpha/r)\log\left[1/\left(1-r/\mu\right)\right]$ (see [17, Chapter 4]) we find

$$P(C) = \left(1-\frac{r}{\mu}\right)^{\frac{\alpha}{r}} \frac{\left(\frac{\alpha}{r}\log\left[1/\left(1-\frac{r}{\mu}\right)\right]\right)^C}{C!}. \tag{13}$$

Using this distribution, we find

$$\mathbb{E}[C] = \mathrm{var}[C] = \left(\frac{\alpha}{r}\right)\log\left[\frac{1}{1-r/\mu}\right]. \tag{14}$$

We now find the marginal probability $P(c_k)$ for the number of species c_k with k individuals regardless of all others. By using Eq. (11), separating out the c_k terms, we obtain

$$\begin{aligned} P(c_k) &= \sum_{(c_i)_{i \neq k}} P(c_1, c_2, \ldots, c_{k-1}, c_k, c_{k+1}) \\ &= \left(1 - \frac{r}{\mu}\right)^{\alpha/r} \sum_{(c_i)_{i \neq k}} \prod_{j=1}^{\infty} \frac{1}{c_j!} \left(\frac{1}{j} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{j}\right)^{c_j} \\ &= \frac{1}{c_k!} \left(\frac{1}{k} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{k}\right)^{c_k} \left(1 - \frac{r}{\mu}\right)^{\alpha/r} \prod_{j \neq k}^{\infty} \exp\left(\frac{1}{j} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{j}\right) \\ &= \frac{1}{c_k!} \left(\frac{1}{k} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{k}\right)^{c_k} \exp\left(-\frac{1}{k} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{k}\right), \end{aligned} \tag{15}$$

which is a Poisson distribution with parameter equal to the mean and variance

$$\mathbb{E}\left[c_k\right] = \operatorname{var}\left[c_k\right] = \frac{\alpha}{r} \frac{1}{k} \left(\frac{r}{\mu}\right)^{k}. \tag{16}$$

Next, we determine the marginal distribution of the number n_i of individuals belonging to species i. By taking the mean of the first relation in Eq. (1), we find

$$\mathbb{E}\left[c_k\right] = \mathbb{E}\left[\sum_{i=1}^{C} \boldsymbol{I}\left(n_i, k\right)\right] = \mathbb{E}\left[\sum_{i=1}^{C} \mathbb{E}\left[\boldsymbol{I}\left(n_i, k\right) | C\right]\right].$$

Since species are assumed to be non-interacting, the random variables $(n_i)_{i \leq C}$ are independent and identically distributed (iid) and are also independent of C. Thus, for every $1 \leq i \leq C$ we can write

$$\mathbb{E}\left[\boldsymbol{I}\left(n_i, k\right) | C\right] = \mathbb{E}\left[\boldsymbol{I}\left(n_1, k\right)\right] = P(n_1 = k)$$

for which $P(n_1 = k)$ is still undetermined. The above relation yields

$$\mathbb{E}\left[c_k\right] = \mathbb{E}\left[C\right] P(n_1 = k), \tag{17}$$

in which $\mathbb{E}\left[c_k\right]$ and $\mathbb{E}\left[C\right]$ are determined by Eqs. (16) and (14). Since all n_i values are identically distributed and $P(n_i = k) = P(n_1 = k)$, we can finally write $P(n_i)$ for any species i:

$$P(n_i) = \frac{1}{n_i} \left(\frac{r}{\mu}\right)^{n_i} \frac{-1}{\log\left[1 - r/\mu\right]}. \tag{18}$$

Thus, every n_i follows Fisher's logarithmic series distribution [14] with parameter r/μ. Note that although the distribution $P(N)$ for the total population N depends on the immigration rate, the distribution in Eq. (18) is *independent* of α. This is because each immigration event necessarily introduces a new species but does not influence the dynamics of a species already present. Once introduced, the evolution of any species depends solely on its birth and death rates r and μ.

Finally, we can use Eq. (11) to determine the expected Shannon's entropy and Simpson's diversity index, as defined in Eq. (2). Using a similar procedure to the one used to determine $P(c_k)$, we isolate the c_k term in the definition of $P(\vec{c})$ and find the same form in terms of $c_k - 1$. Note that we can write the mean of $c_k f(N)$, for any function $f(N)$, as

$$\mathbb{E}\left[c_k f(N)\right] = \frac{1}{k} \frac{\alpha}{r} \left(\frac{r}{\mu}\right)^{k} \mathbb{E}\left[f(N + k)\right].$$

By considering the functions $f(x) = \log(x/k)/x$ and $f(x) = (k/x)^2$ we find the respective expressions for Shannon's Entropy and Simpson's diversity index

$$\mathbb{E}[H] = \frac{\alpha}{r}\sum_{k=1}^{\infty}\left(\frac{r}{\mu}\right)^k \mathbb{E}\left[\frac{\log\left(\frac{N+k}{k}\right)}{N+k}\right] \quad \text{and} \quad \mathbb{E}[S] = 1 - \frac{\alpha}{r}\sum_{k=1}^{\infty} k\left(\frac{r}{\mu}\right)^k \mathbb{E}\left[\left(\frac{1}{N+k}\right)^2\right]. \tag{19}$$

Since the distribution of N is known and given by Eq. (5), we can use Eq. (19) to numerically compute $\mathbb{E}[H]$ and $\mathbb{E}[S]$.

For completeness, we also derive results for the sBDI process with a finite number of clones Q that each carry a finite immigration rate into the system. In Appendix A.1, we use the detailed balance conditions to derive explicit steady state probability distributions over the particle count vector n_i, the species count vector c_k, and the number of clones in the sample C.

3.2 Fast Immigration Limit

We now consider the large immigration limit of the sBDI model. While at steady-state the distribution of the number of individuals in each species $P(n_i)$, given by Eq. (18), is independent of α, the distributions $P(N)$, $P(C)$ and $P(c_k)$ do depend on the total immigration rate α as indicated in Eqs. (5), (13) and (15), respectively. Since immigration always introduces a new species, the *per clone* immigration rate is zero. To study the large immigration regime in which each clone has a finite immigration rate, we assume $\alpha \equiv \widetilde{\alpha}\,\Omega$ scales as the parameter $\Omega \to \infty$, which can be thought of as the total number of different clones that can immigrate into the system per unit time. Increasing α will introduce new individuals and new species to the system, so one can intuitively conclude that the total population N and number of species C, as well as the number of species with k individuals c_k, will also increase. We also show that, as Ω increases, the scaled values N/Ω, C/Ω and c_k/Ω converge in distribution to a constant, as described in Eq. (3), with average values given by

$$\frac{N}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{\mu - r}, \quad \frac{C}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\log\left[\frac{1}{1 - r/\mu}\right] \quad \text{and} \quad \frac{c_k}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\frac{(r/\mu)^k}{k},$$

and vanishing variances. A rigorous proof is given in Appendix A.2. Moreover, we can also write the convergence in distribution of the scaled Shannon's entropy $H/\log\Omega$ and Simpson's diversity index S,

$$\frac{H}{\log\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1 \quad \text{and} \quad S \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1,$$

as also derived in Appendix A.2. We can now use the scaling results above to infer that

$$\frac{\mathbb{E}[H]}{\log\Omega} = 1 - \mathcal{O}(\Omega^{-1}), \quad \mathbb{E}[S] = 1 - \mathcal{O}(\Omega^{-1}).$$

3.3 Interpretation of Results

All distributions computed above depend on two nondimensional quantities: $u \equiv \alpha/r$ and $v^{-1} \equiv \mu/r$ (with $0 \leq v < 1$). Also note that the ratios $\mathbb{E}[C]/\mathbb{E}[N]$ and $\text{var}[C]/\text{var}[N]$ are

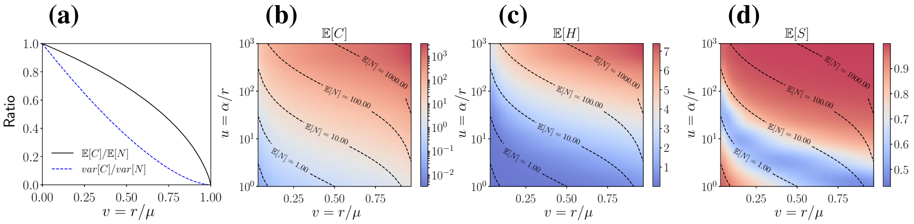

Fig. 2 **a** The ratios $\mathbb{E}[C]/\mathbb{E}[N]$ and $\mathrm{var}[C]/\mathrm{var}[N]$ as a function of $v = r/\mu$. This ratio is independent of the immigration rate α and can be explicitly determined by our theoretical results. **b**–**d** Analytical results for the total richness $\mathbb{E}[C]$, Shannon's entropy $\mathbb{E}[H]$ and Simpson's index $\mathbb{E}[S]$, respectively, as functions of $u = \alpha/r$ and $v = r/\mu$. Both $\mathbb{E}[C]$ and $\mathbb{E}[H]$ increase with α/r, r/μ, and $\mathbb{E}[N]$. Simpson's index $\mathbb{E}[S]$ has a similar behavior only if $\mathbb{E}[N] = uv/(1-v)$ is greater than ~ 10

"immigration-invariant" in the sense that they depend only on $v = r/\mu$ and are independent of the immigration rate α and/or u. These ratios are plotted in Fig. 2a. For systems where the immigration rate is difficult to estimate, such quantities can be useful probes of the system. For example, $\mathbb{E}[C]/\mathbb{E}[N]$ represents the relative abundance of species with respect to the total number of individuals in the system. By construction, this ratio cannot exceed unity. The limiting case of $\mathbb{E}[C]/\mathbb{E}[N] \to 1$ corresponds to a completely heterogeneous system where every individual belongs to a different species while $\mathbb{E}[C]/\mathbb{E}[N] \to 0$ corresponds to homogeneous systems where the entire population is dominated by very few species.

In Fig. 2b–d we show the respective heat-maps of $\mathbb{E}[C]$, $\mathbb{E}[H]$ and $\mathbb{E}[S]$ as a function of $u = \alpha/r$ and $v = \alpha/r$, along with level-sets of $\mathbb{E}[N]$. These figures have been generated using Eqs. (13) and (19) and show that both $\mathbb{E}[C]$ and $\mathbb{E}[H]$ increase with u, v, and $\mathbb{E}[N]$. In our plots, $\mathbb{E}[S]$ is not strictly monotonic despite the expectation that $\mathbb{E}[S]$, as a measure of diversity, would follow the same trend as $\mathbb{E}[C]$ and $\mathbb{E}[H]$. However, this qualitative discrepancy occurs only in the small $\mathbb{E}[N]$ regime where there is a high probability that there are no individuals in the system and diversity loses its meaning. In the extreme limit of $N \to 0$, we find $C \to 0$ and $H \to 0$, but $S \to 1$, giving rise to the nonmonotonic pattern for $\mathbb{E}[S]$.

4 Birth–Death–Immigration Model with Mutation (BDIM)

In this section we consider a birth–death–immigration Model with Mutation (BDIM). Mutation events are particularly relevant in ecology as they lead to speciation within populations [23], and in studies of gene domain family evolution [21]. In the BDIM process, we still assume individuals and species are non-interacting and that birth, death, immigration, and mutation rates do not depend on the state of the system. We allow an individual of a given species to mutate and give rise to a new, yet unrepresented, species. Mutations are assumed to be neutral in that an individual arising from mutation maintains the same birth and death rates as the rest of the population.

We start by allowing mutations only in offspring arising immediately after their birth, as illustrated in Fig. 1b. For each birth event there is a probability ϵ (with $0 \leq \epsilon \leq 1$) that the offspring is mutated, representing a completely new species. This mechanism is applicable to e.g., bacterial populations where DNA replication can induce a gene mutation that will be carried by the newborn cell. The subsequent theoretical analysis will be carried out within the framework of a single mutation at birth as described here. However our mathematical treatment is not limited to this specific case and, in Sect. 4.4, we will apply the same tools to study other relevant scenarios such as "somatic" mutations that can occur any time during the lifetime of an individual, and "double" (or "symmetric") mutations where both parent and offspring mutate upon birth.

4.1 Derivation of Steady State Statistics

In Sect. 3, we were able to use reversibility and detailed balance to determine $P(\vec{c})$, the probability for a given species-count configuration $\vec{c} = (c_1, \ldots, c_k, \ldots)$ to occur. The introduction of mutation, however, makes the system irreversible and analytically evaluating $P(\vec{c})$ becomes prohibitively complex. We can nonetheless exploit some general features of the BDIM model, such as neutrality and independence, to extract results such as the mean and the variance of C and c_k. The evaluation of other quantities such as the mean of the diversity indices $\mathbb{E}[H]$ and $\mathbb{E}[S]$ will require numerical simulations. Our theoretical analysis relies on two important features of the model:

- Because mutations do not affect overall birth or death rates but only the species to which newborns belong, the distribution for $P(N)$ remains identical to the one derived in Eq. (5) for the simple BDI model. Hence, in the BDIM model, the overall growth rate due to immigration and birth is still $\alpha + Nr$ and the overall death rate is still μN. The resulting $P(N)$ is independent of mutation events and Eq. (5) still holds.
- The marginal distribution $P(n_i)$ of the number of particles n_i of species i still follows a logarithmic series distribution as in Eq. (18), but with the replacement $r \to r(1-\epsilon)$. Intuitively, this can be understood by noting that under mutation a new individual is introduced into the n_i population with rate $r(1-\epsilon)$ instead of r, since the "remainder" $r\epsilon$ is the rate at which a new individual in a new species arises. The dynamics of the n_i individuals thus remains unchanged, provided the birth rate is modified to $r(1-\epsilon)$ to account for the diminished births within the given species. In Appendix B.1 we provide a more rigorous justification of this argument. Also, in Fig. 6 of Appendix B.1 we plot the probability distribution for the number of individuals in a given species as determined from simulations of the BDIM model, compare our findings to the expected logarithmic distribution, and show good agreement between the two. Thus, both theoretically and numerically, we verify that $P(n_i)$ follows a logarithmic series distribution with parameter $r(1-\epsilon)/\mu$ for all values of $0 \le \epsilon \le 1$:

$$P(n_i) = \frac{1}{n_i}\left(\frac{r(1-\epsilon)}{\mu}\right)^{n_i} \frac{-1}{\log\left[1 - r(1-\epsilon)/\mu\right]}. \tag{20}$$

Once the $P(n_i)$ and $P(N)$ distributions are known for the BDIM model we can use Eq. (1) and the fact that the $(n_i)_{i\le C}$ are iid and independent of C to express the mean of the third relation of Eq. (1) as

$$\mathbb{E}[N] = \mathbb{E}\left[\sum_{i=1}^{C} \mathbb{E}[n_i | C]\right] = \mathbb{E}[C]\,\mathbb{E}[n_1],$$

so that

$$\mathbb{E}[C] = \frac{\alpha/\mu}{1 - r/\mu} \log\left(\frac{1}{1 - r(1-\epsilon)/\mu}\right) \frac{1 - r(1-\epsilon)/\mu}{r(1-\epsilon)/\mu}. \tag{21}$$

Using the moment generating function of N, we can similarly determine the variance of C as shown in detail in Appendix B.2

$$\text{var}\,[C] = \mathbb{E}\,[C]\left[\frac{\mathbb{E}\,[C]}{(\alpha/r)} + \log\left(1 - \frac{r(1-\epsilon)}{\mu}\right) + 1\right]. \tag{22}$$

Finally, we take the mean of the first relation in Eq. (1). Since all the $(n_i)_{i\leq C}$ are iid and independent of C we can write

$$\mathbb{E}\,[c_k] = \mathbb{E}\,[C]\,P(n_1 = k) = \frac{\alpha/\mu}{1 - r/\mu}\frac{1}{k}\left(1 - \frac{r(1-\epsilon)}{\mu}\right)\left(\frac{r(1-\epsilon)}{\mu}\right)^{k-1}. \tag{23}$$

For the variance of c_k, we also use the definition in the first relation in Eq. (1) to find

$$c_k c_\ell = \boldsymbol{I}\,(k,\ell)\sum_{i=1}^{C}\boldsymbol{I}\,(n_i,k) + \sum_{i=1}^{C}\sum_{j\neq i}\boldsymbol{I}\,(n_i,k)\,\boldsymbol{I}\,\left(n_j,k\right).$$

Upon using Eq. (20) to take the mean of this expression, and recalling that n_i is independent of $n_{j\neq i}$ and C, we find

$$\begin{aligned}\mathbb{E}\,[c_k c_\ell] &= \boldsymbol{I}\,(k,\ell)\,\mathbb{E}\,[c_k] + \mathbb{E}\,[C(C-1)]\,P(n_i = k)P(n_i = \ell),\\ \text{var}\,[c_k] &= \mathbb{E}\,[c_k] + \frac{\text{var}\,[C] - \mathbb{E}\,[C]}{\mathbb{E}\,[C]^2}\,\mathbb{E}\,[c_k]^2.\end{aligned} \tag{24}$$

These expressions are also valid for the sBDI model, but since C and c_k are Poisson-distributed in that case, var $[C] = \mathbb{E}\,[C]$ and var $[c_k] = \mathbb{E}\,[c_k]$ (see below).

We can use Eqs. (20), (23), (21), and Appendix B.2 to further develop the second moments. For example,

$$\mathbb{E}\left[c_k c_{\ell\neq k}\right] = \left(\frac{r(1-\epsilon)}{\mu}\right)^{k\ell}\frac{1}{k\ell}\frac{\left[\frac{\alpha}{\mu}\left(1 - \frac{r(1-\epsilon)}{\mu}\right) + \epsilon\frac{r}{\mu}\right]\left[\frac{\alpha}{\mu}\left(1 - \frac{r(1-\epsilon)}{\mu}\right)\right]}{\left(1 - \frac{r}{\mu}\right)^2\left(\frac{r(1-\epsilon)}{\mu}\right)^2},$$

which reduces to the simple BDI result $\mathbb{E}\,[c_k]^2$ when $\epsilon = 0$. For $k = \ell$, we have

$$\text{var}\,[c_k] = \mathbb{E}\,[c_k] + \epsilon\frac{(\alpha/\mu)(r/\mu)}{(1 - r/\mu)^2 k^2}\left(\frac{r(1-\epsilon)}{\mu}\right)^{2(k-1)}.$$

4.2 Fast Immigration Limit

We now study the large immigration limit of the BDIM model. As done in Sect. 3.2 we set $\alpha = \widetilde{\alpha}\,\Omega$ and consider the $\Omega \to \infty$ limit. Since the dynamics of N/Ω remain unchanged in the BDIM model compared to the dynamics in the simple BDI model, we recover convergence in distribution for N/Ω towards the constant $\widetilde{\alpha}/(\mu - r)$ as $\Omega \to \infty$. Following the same procedures illustrated in Appendix A.2 for the simple BDI model, and using the moment generating functions of C and c_k we can also prove the convergence in distribution of C/Ω and c_k/Ω towards the following

$$\begin{aligned}\frac{C}{\Omega} &\xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{(\widetilde{\alpha}/\mu)}{1 - r/\mu}\,\frac{1 - r(1-\epsilon)/\mu}{r(1-\epsilon)/\mu}\log\left(\frac{1}{1 - r(1-\epsilon)/\mu}\right),\\ \frac{c_k}{\Omega} &\xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{(\widetilde{\alpha}/\mu)}{1 - r/\mu}\,\frac{1 - r(1-\epsilon)/\mu}{k}\left(\frac{r(1-\epsilon)}{\mu}\right)^{k-1}.\end{aligned}$$

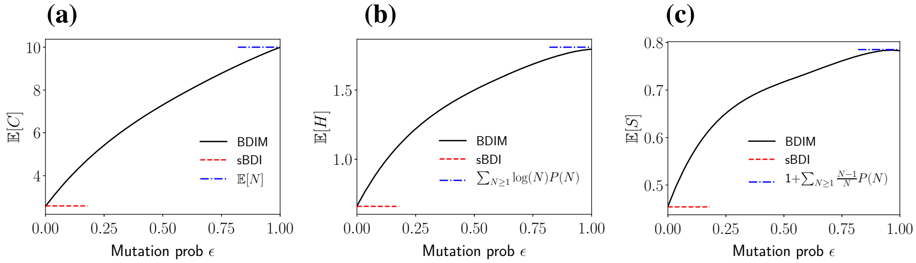

Fig. 3 Comparison of various diversity indices between the sBDI model and the BDIM model at varying values of mutation probability ϵ. For both models, we set $\alpha = 1$, $\mu = 1$, and $r = 0.9$. **a** Total richness C as determined using Eqs. (21) and (22). **b** Expected Shannon's entropy $\mathbb{E}[H]$, determined with simulations of the model with mutation at division. **c** Expected Simpson's diversity index $\mathbb{E}[S]$, determined with simulations of the model with mutation at division. In each case, mutation increases diversity (black curves) relative to that of the sBDI model (blue dot-dashed lines). The maximum diversity is obtained in the limit where mutation always occurs (i.e. $\epsilon = 1$) and all individuals belong to different species such that $C = N$ (red dashed lines) (Color figure online)

Finally, the convergence in distribution of the scaled Shannon's entropy $H/\log\Omega$ and Simpson's diversity index S are

$$\frac{H}{\log\Omega}\xrightarrow[\Omega\to\infty]{\mathcal{D}}1 \quad\text{and}\quad S\xrightarrow[\Omega\to\infty]{\mathcal{D}}1.$$

4.3 Interpretation of Results and Comparison with sBDI Model

In this subsection we compare features of the sBDI and BDIM models. Note that only three parameters are necessary to characterize all results obtained in both models: $u \equiv \alpha/r \geq 0$, $v \equiv r/\mu$ $(0 \leq v \leq 1)$, and $0 \leq \epsilon \leq 1$. In Fig. 3a–c we plot the total richness C, Shannon's entropy H, and Simpson's index S as defined in Eq. (2). In these figures, $\mathbb{E}[C]$ was determined using Eqs. (21) and (22), while $\mathbb{E}[H]$ and $\mathbb{E}[S]$ were found from simulations.

The expected diversity indices for the simple BDI model are shown by the dashed red horizontal lines. As can be seen from Fig. 3, all measures of diversity in the BDIM model increase with ϵ and reach their maximum at $\epsilon = 1$ (shown by the blue dot-dashed lines) when all individuals give rise to mutant offspring. Upon setting $\epsilon = 1$, and assuming a nonzero population N, Eq. (2) yields $C = N$, $H = \log N$ and $S = (N-1)/N$, mirroring the fact that each species only has one individual. By noting that $H = 0$ and $S = 1$ for $N = 0$, we find for $\epsilon = 1$

$$\mathbb{E}[C] = \mathbb{E}[N], \quad \mathbb{E}[H] = \sum_{N\geq1}\log N\, P(N) \quad\text{and}\quad \mathbb{E}[S] = 1 + \sum_{N\geq1}\frac{N-1}{N}P(N),$$

with $P(N)$ being the probability distribution of N given in Eq. (5).

For general ϵ in the BDIM model, since we cannot analytically determine the probability for the species-count vector $\vec{c}$, we cannot derive explicit formulae for $\mathbb{E}[H]$ and $\mathbb{E}[S]$ as we did for the simple BDI model in Sect. 3. However, we can estimate both $\mathbb{E}[H]$ and $\mathbb{E}[S]$ by approximating c_k with $\mathbb{E}[c_k]$ in Eq. (2) to find

$$\mathbb{E}\left[H\right] \simeq -\left(1-\frac{r(1-\epsilon)}{\mu}\right)\sum_{k\geq 1}\left(\frac{r(1-\epsilon)}{\mu}\right)^{k-1}\log\left[\frac{\mu}{\alpha}k\left(1-\frac{r}{\mu}\right)\right] \quad \text{and}$$

$$\mathbb{E}\left[S\right] \simeq 1-\frac{\mu}{\alpha}\frac{1-r/\mu}{1-r(1-\epsilon)/\mu}. \tag{25}$$

We compare these approximations with results obtained from numerical simulations in Fig. 7 in Appendix B.1. As can be seen our analytical estimates become more accurate as the average number of individuals $\mathbb{E}\left[N\right] = \alpha/(\mu - r)$ increases, that is, for $\alpha \gg \mu$ and $\mu \gtrsim r$. For $\mathbb{E}\left[N\right] \geq 5$, both estimates for Shannon's entropy and Simpson's diversity index fall within 10% of their simulated values.

4.4 Alternative Mutation Mechanisms

The BDIM model, as described above, assumes that mutations occur with probability ϵ during each birth event. We can very easily adapt the mathematical reasoning used in Sect. 4.1 to characterize other types of mutation processes. Note that if mutation events add more species to the system, but do not change the overall birth and death rates of the population, the total-population distribution $P(N)$ will remain unchanged from the expression found in Eq. (5) for the simple BDI model. This will be the case for the two alternative mutation mechanisms described below.

Somatic mutation: Each individual may spontaneously mutate at constant rate $\eta > 0$ over its lifetime, giving rise to an individual of a new species. Such a birth-independent mutation might be a reasonable model for e.g., DNA damage or epigenetic changes in a cell. In this scenario, for a given n_i population, new individuals are added to the same i species at rate rn_i and removed at rate $(\mu + \eta)n_i$ since mutation events will effectively transfer an individual from a given species to a new one. Hence, the distribution for $P(n_i)$ should remain a logarithmic series distribution as in Eq. (18) but with parameter $r/(\mu + \eta)$. All theoretical results found in Sect. 4.1 remain the same in this case provided we replace $\epsilon \rightarrow \eta/(\mu + \eta)$.

Double mutation: Both parent and offspring may spontaneously mutate at birth, as for example in symmetric stem cell differentiation. More generally, we can assume that one of the two individuals mutates to a new species with probability ϵ_1 and that both mutate into two new species with probability ϵ_2. In this case, for a given n_i population, new individuals are added at rate $r(1-\epsilon_1-\epsilon_2)n_i$ to species i and removed at rate $(\mu+r\epsilon_2)n_i$. The number of individuals in species i should thus still be logarithmically distributed as in Eq. (18), but with parameter $r(1-\epsilon_1-\epsilon_2)/(\mu + r\epsilon_2)$. All theoretical results found in Sect. 4.1 remain the same, provided we replace $\epsilon \rightarrow (r\epsilon_2 + \mu(\epsilon_1 + \epsilon_2))/(r\epsilon_2 + \mu)$.

5 Birth–Death–Immigration Model with Carrying Capacity (BDICC)

In the third and final model analyzed in this paper, we include an important interaction within the total population—a carrying capacity that is typically used to represent resource limitations. The more individuals present in a system, the more they need to share resources, potentially affecting survival or reproduction rates. The carrying-capacity concept is ubiquitous in ecology such as for species on an island with finite resources that limit the total population. Other applications may include lymphatic growth which is known to be induced by several molecules, in particular cytokines [32] that may become insufficient to sustain further proliferation of T-cells if the population becomes too large.

We first consider a carrying capacity on the death rate of each individual and derive analytical results; more general cases will be addressed via numerical simulations. As shown in Fig. 1c, the only difference between our BDI model with carrying capacity (BDICC) and the sBDI model is that the death rate now depends on the total number of individuals in the system N. We assume that $\mu(N)$ is an increasing function with N as dwindling resources led by population increases will also increase the death rate. It is important to remark that populations described by the BDICC model do not evolve independently. Since the dynamics of each individual now depends on that of all others, there is a global, but "neutral" interaction. In contrast to the two previous models, the number of individuals in each species $(n_i)_{i\leq C}$, can no longer be considered an independent random variable so that

$$\mathbb{E}\left[c_k\right] \neq \mathbb{E}\left[C\right] P(n_1 = k).$$

The equality of the quantities on the left and right hand sides above was used in the previous analysis to determine Eqs. (17) and (23) and is no longer applicable to the BDICC model.

5.1 Derivation of Steady State Statistics

We first consider the dynamics of the total number of individuals N and study how $P(N)$ is modified in the BDICC model. In this case, the overall population still undergoes a birth and death process with rates $\alpha + rN$ and $\mu(N)N$, respectively. The properties of birth and death process with non homogeneous rates are known [2]. In particular, in the case of an increasing function $\mu(N) > 0$, the conditions for the existence of a steady state is

$$\lim_{N\to\infty} \mu(N) > r.$$

More general conditions for the existence of a steady-state configuration have been detailed in the case of a non-increasing death rate $\mu(N)$ [2]. If a steady-state exists, then $P(N)$ can be found using detailed balance, similar to what was done in Sect. 3

$$P(N) = \begin{cases} \dfrac{1}{Z_{\alpha,r}}, & N = 0, \\ \dfrac{1}{Z_{\alpha,r}} \dfrac{1}{N!} \displaystyle\prod_{k=0}^{N-1} \dfrac{\alpha + rk}{\mu(k+1)}, & N \geq 1 \end{cases} \tag{26}$$

with $Z_{\alpha,r}$ a normalization constant given by

$$Z_{\alpha,r} = 1 + \sum_{n=1}^{\infty} \frac{1}{n!} \prod_{k=0}^{n-1} \frac{\alpha + rk}{\mu(k+1)}.$$

To determine $P(\vec{c})$, $P(C)$ and $P(c_k)$, we rely on reversibility of the system and detailed balance. Interestingly, while a non-constant death rate $\mu(N)$ preserves detailed balance, a non-constant growth function $r(N)$ does not strictly obey detailed balance. We will come back to this point further in the discussion, in Sect. (5.3.2). For now, we consider $\mu(N)$ and constant r and write all possible transitions of the system as was done in Sect. 3.1

$$\begin{array}{lll}
 & (c_1, c_2, \ldots) \xrightarrow{\alpha} (c_1+1, c_2, \ldots) & \text{Immigration} \\
\text{for } k \geq 1 & (c_1, \ldots, c_k, c_{k+1}, \ldots) \xrightarrow{rkc_k} (c_1, \ldots, c_k - 1, c_{k+1}+1, \ldots) & \text{Birth} \\
\text{for } k \geq 2 & (c_1, \ldots, c_{k-1}, c_k, \ldots) \xrightarrow{\mu(N)kc_k} (c_1, \ldots, c_{k-1}+1, c_k - 1, \ldots) & \multirow{2}{*}{$\Big\}$ Death} \\
 & (c_1, c_2, \ldots) \xrightarrow{\mu(N)c_1} (c_1 - 1, c_2, \ldots) &
\end{array}$$

which differ from the ones written in Sect. 3.1 by virtue of $\mu \to \mu(N)$ with $N = \sum_{k\geq 1} kc_k$. By assuming detailed balance, we write

$$\mu(N)kc_k P(c_1, \ldots, c_{k-1}, c_k, \ldots) = (k-1)\,(c_{k-1}+1)\, r P(c_1, \ldots, c_{k-1}+1, c_k - 1, \ldots), \tag{27}$$

for $k \geq 2$, while for $k = 1$ the following holds

$$\mu(N)c_1 P(c_1, c_2, \ldots) = \alpha P(c_1 - 1, c_2, \ldots). \tag{28}$$

We follow the same procedure as in Sect. 3.1 and iterate Eq. (7) using Eq. (8). After imposing normalization, we obtain

$$P(\vec{c}) = P(c_1, \ldots, c_k, \ldots) = \frac{1}{Z_{\alpha,r}} \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=1}^N \mu(n)} \frac{1}{\prod_{i=1}^\infty i^{c_i} c_i!}, \tag{29}$$

where $C = \sum_{k\geq 1} c_k$ as defined in Eq. (1) and where $Z_{\alpha,r}$ is the normalization constant that can be obtained by evaluating $P(N = 0)$ in Eq. (26) so that

$$Z_{\alpha,r} = \sum_{c_1,\ldots,c_k,\ldots}^{\infty} \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=1}^N \mu(n)} \frac{1}{\prod_{i=1}^\infty i^{c_i} c_i!} = 1 + \sum_{n=1}^\infty \frac{1}{n!} \prod_{k=0}^{n-1} \frac{\alpha + rk}{\mu(k+1)}. \tag{30}$$

More details can be found in Appendix C.1. We can now use the expression for $P(\vec{c})$ in Eq. (29) to evaluate the moment generating function of C and related moments

$$M_C(\xi) \equiv \mathbb{E}\left[\exp(\xi C)\right] = \frac{1}{Z_{\alpha,r}} \sum_{c_1,\ldots,c_k,\ldots} \left(\frac{\alpha}{r} e^\xi\right)^C \frac{r^N}{\prod_{n=1}^N \mu(n)} \frac{1}{\prod_{i=1}^\infty i^{c_i} c_i!}.$$

Since the argument of the sum in the above expression is the same as in Eq. (30) provided $\alpha \to \alpha e^\xi$ we can write

$$M_C(\xi) = \frac{Z_{\alpha e^\xi, r}}{Z_{\alpha,r}},$$

for any $\xi < 0$. We can now differentiate $M_C(\xi)$ with respect to ξ and take the limit $\xi \to 0$ to find the following expressions for the mean and the variance of C

$$\mathbb{E}[C] = \alpha\, \mathbb{E}\left[\sum_{k=0}^{N-1} (\alpha + rk)^{-1}\right], \tag{31}$$

$$\operatorname{var}[C] = \mathbb{E}[C]\,(1 - \mathbb{E}[C]) + \alpha^2\, \mathbb{E}\left[\left(\sum_{k=0}^{N-1} \frac{1}{\alpha + rk}\right)^2 - \sum_{k=0}^{N-1} \frac{1}{(\alpha + rk)^2}\right]. \tag{32}$$

We can use the above expressions and $P(N)$ as determined in Eq. (26) to evaluate the mean and variance of C. Note that setting a uniform $\mu(N) = \mu$ in Eqs. (29) and (30) reduces the results to those of the sBDI model (Sect. 3.1). We can now evaluate $\mathbb{E}[c_k]$ using Eq. (29):

$$\mathbb{E}[c_k] = \frac{1}{Z_{\alpha,r}} \sum_{c_1,\ldots,c_k,\ldots} c_k \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=1}^{N} \mu(n)} \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!},$$

which can be rearranged to yield

$$\begin{aligned}\mathbb{E}[c_k] &= \frac{1}{Z_{\alpha,r}} \frac{\alpha}{kr} \sum_{c_1,\ldots,c_k,\ldots} \left(\frac{\alpha}{r}\right)^C \frac{r^{N+k}}{\prod_{n=1}^{N+k} \mu(n)} \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!} \\ &= \frac{\alpha r^{k-1}}{k} \sum_{c_1,\ldots,c_k,\ldots} \frac{P(\vec{c})}{\prod_{m=1}^{k} \mu(N+m)} \\ &= \frac{\alpha r^{k-1}}{k} \mathbb{E}\left[\prod_{m=1}^{k} \frac{1}{\mu(N+m)}\right]. \end{aligned} \tag{33}$$

A uniform $\mu(n) = \mu$ returns $\mathbb{E}[c_k] = (\alpha/\mu)(r/\mu)^{k-1}/k$, as previously determined in Sect. 3.1. We can also verify that for any function $f(N)$,

$$\mathbb{E}[c_k f(N)] = \frac{\alpha r^{k-1}}{k} \mathbb{E}\left[\frac{f(N+k)}{\prod_{m=1}^{k} \mu(N+m)}\right].$$

For $f(x) = \log(x/k)/x$ and $f(x) = (k/N)^2$ the expressions for Shannon's Entropy and Simpson's diversity index become

$$\mathbb{E}[H] = \alpha \sum_{k=1}^{\infty} r^{k-1} \mathbb{E}\left[\frac{\log\left[\frac{N+k}{k}\right]}{(N+k)\prod_{m=1}^{k} \mu(N+m)}\right], \tag{34}$$

$$\mathbb{E}[S] = 1 - \alpha \sum_{k=1}^{\infty} k r^{k-1} \mathbb{E}\left[\frac{1}{(N+k)^2 \prod_{m=1}^{k} \mu(N+m)}\right]. \tag{35}$$

Once again, setting $\mu(N) = \mu$ a constant allows us to recover the results in Eq. (19) for the sBDI model.

5.2 Fast Immigration Limit

To analyze the large immigration limit, $\alpha = \widetilde{\alpha}\,\Omega$, $\Omega \to \infty$, we need to assume a specific form for the death rate. For a given Ω, we take the death rate as a function of N/Ω:

$$\mu(N) = \widetilde{\mu}(N/\Omega).$$

The reason behind this scaling is that we want to keep $\mu(N)$ at the same order of magnitude as Ω increases. As in the previous models, we will show that $\mathbb{E}[N]$ diverges as Ω increases, but the random variable N/Ω will be shown to converge in distribution to a constant. As a consequence, the death rate $\widetilde{\mu}(N/\Omega)$ will also converge in distribution to a constant.

Given $\widetilde{\mu}(x)$ is continuous and strictly increasing, and that $\lim_{x\to\infty} r/\widetilde{\mu}(x) < 1$, one can show that there exists a unique, positive solution n^* to the fixed-point equation

$$n^* \widetilde{\mu}(n^*) = \widetilde{\alpha} + r n^*. \tag{36}$$

In Appendix C.2, we show that for every $\delta > 0$,

$$P(|N/\Omega - n^*| > \delta) \xrightarrow{\Omega\to\infty} 0,$$

thus proving that

$$\frac{N}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} n^* \tag{37}$$

in which n^* is defined by Eq. (36). The proof of this convergence is analogous to the one in [11, Proposition 4]). Intuitively, n^* can be identified with the steady-state solution to the deterministic approximation of the dynamics of $n(t) \equiv N(t)/\Omega$ given by

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = \widetilde{\alpha} + rn(t) - \widetilde{\mu}(n(t))n(t).$$

Using the convergence of Eq. (37), we can show convergence in distribution of C/Ω and c_k/Ω as follows

$$\frac{C}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\log\left[\frac{1}{1 - r/\widetilde{\mu}(n^*)}\right] \quad \text{and} \quad \frac{c_k}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\frac{1}{k}\left(\frac{r}{\widetilde{\mu}(n^*)}\right)^k.$$

The complete proofs of these convergences are given in Appendix C.3, but one can also verify them by inspecting Eqs. (31) and (33) respectively to determine the convergence of $\mathbb{E}\left[C/\Omega\right]$ and $\mathbb{E}\left[c_k/\Omega\right]$ using convergence of N/Ω to n^*.

Even though the dynamics of all n_i are coupled through the death rate $\mu(N) = \mu\left(\sum n_i\right)$, all n_i remain identically distributed: $P(n_i = k) = P(n_j = k)$ for all $i, j \leq C$ and $k \geq 1$. This "neutrality" allows us to determine the convergence of n_i in the $\Omega \to \infty$ limit as detailed in Appendix C.4:

$$P(n_i = k) \xrightarrow[\Omega\to\infty]{} \frac{1}{k}\left(\frac{r}{\widetilde{\mu}(n^*)}\right)^k \frac{-1}{\log[1 - r/\widetilde{\mu}(n^*)]},$$

which shows that for $\Omega \to \infty$, n_i converges to a logarithmic-series distribution with parameter $r/\mu(n^*)$.

Finally, we can use the convergence in distribution of both N/Ω and c_k/Ω, to determine the convergence in distribution for the rescaled Shannon's Entropy $H/\log\Omega$ and Simpson's diversity index S:

$$\frac{H}{\log\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1 \quad \text{and} \quad S \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1.$$

These convergence results are identical for all three models and their proofs are similar to the ones for the sBDI model as described in Appendix A.2.

5.3 Interpretation and Analysis of Results

5.3.1 Comparison with the sBDI Model

To properly compare the sBDI and BDICC models, we fix their immigration rates α and birth rates r to be the same. For the BDICC model, we use a linear death rate function $\mu(N) = \mu_1 N$ and tune both μ_1 and the constant death rate μ in the sBDI model to yield the same average total number of individuals $\mathbb{E}\left[N\right]$.

In Fig. 4 we plot the distributions $P(N)$ and $P(C)$ as well as the average $\mathbb{E}\left[c_k\right]$ in a low immigration regime ($\alpha = 0.2$ and $r = 0.99$) for both the sBDI and the BDICC models. We

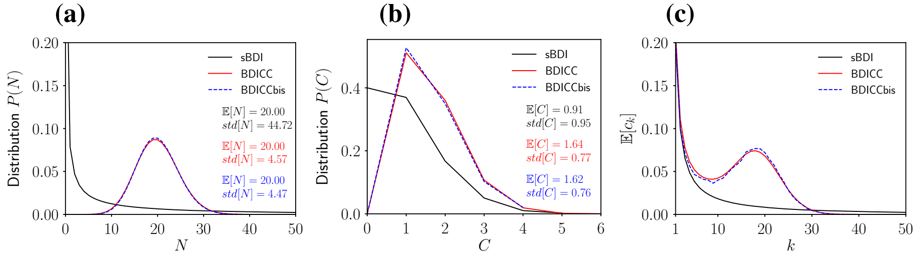

Fig. 4 Comparison of the sBDI model with the two carrying capacity models (BDICC with carrying capacity on the death rate and BDICC-bis with carrying capacity on birth the rate) for slow immigration (parameters are chosen such as $\mathbb{E}[N] = 20$ in both cases: $\alpha = 0.2$, $r = .99$, $\mu = 1$ for the sBDI; $\alpha = 0.2$, $r = .99$, $\mu(N) = 0.0475N$ for the BDICC). **a** Theoretical distributions $P(N)$ for the three models. **b** Distributions of the richness $P(C)$ obtained from Monte-Carlo simulations. **c** The theoretical expected species-count vector $\left(\mathbb{E}\left[c_k\right]\right)_{k\geq 1}$ calculated from Eqs. (33) and (26). Contrary to the sBDI model, the BDICC model is dominated by only one species ($C \simeq 1$) with around 20 individuals (the peak of $\mathbb{E}\left[c_k\right]$ arises at $k \simeq 20$). This attribute is completely missed in a mean-field approximation to $\mathbb{E}\left[c_k\right]$ [16]. Negligible differences between the BDICC and BDICC-bis models are observed (Color figure online)

adjusted μ for the sBDI model and μ_1 for the BDICC model so that $\mathbb{E}[N] = 20$ in both cases. Clear differences emerge. First, since $\mu(N)$ is proportional to the existing population N in the BDICC model, very rarely will the population reach vanishingly small levels: as $N \to 0$ so will $\mu(N) \to 0$ allowing birth and immigration to replenish N. This is in contrast to the sBDI model where μ is a constant independent of N.

Another feature of a low immigration rate is that it allows one species to "invade the niche" of the BDICC model before the arrival of another species. The result is that only one species ($C \simeq 1$) represents the whole population and $\mathbb{E}[c_k]$ has a peak around $k \approx \mathbb{E}[N] = 20$. This exclusion effect does not arise in the sBDI model since the presence of species already in the system does not influence the dynamics of the newly arriving ones. These exclusionary interactions are also the origin of the peak observed in Fig. 4c. Note that this difference is not only due to the sBDI model's high probability of extinction ($N = 0$): we checked that the distributions of the sBDI model, conditioned on $N > 0$, also fail to display the exclusionary effect where one clone dominates. Direct mean-field approximations, $\mathbb{E}[c_k c_\ell] \approx \mathbb{E}[c_k]\,\mathbb{E}[c_\ell]$, lead to monotonic decreasing $\mathbb{E}[c_k]$ [16], completely missing the peak around the carrying capacity ($k \approx 20$). Global carrying capacity interactions can also have a significant influence on Shannon's entropy and Simpson's diversity index.

The qualitative differences between the two models diminish as the immigration rate α increases. This confirms our theoretical analysis through which we showed that the sBDI and the BDICC models follow similar trends as α increases. If we fix μ of the sBDI model and μ_1 in the BDICC model such that $\lim_{\Omega\to\infty}\mathbb{E}[N/\Omega]$ remains the same for both models, we find that N/Ω, C/Ω and c_k/Ω converge to the same constants in the two models and that n_i converges to the same the log-series distribution as well.

5.3.2 Carrying Capacity on Birth (BDICC-bis Model)

Our BDICC model included an interaction only through the death rate $\mu(N)$. This choice, as opposed to, say, $r(N)$ was made because the detailed balanced assumption can be shown to hold between all pairs of states, rendering our analytic results for the probability distribution $P(\vec{c})$ exact.

Alternatively, one can impose an interaction through a population-dependent birth term. It is well-known that even if the mean populations are equal, models using $\mu(N)$ yield different higher order statistics from those using $r(N)$ [1]. The interacting model with μ constant, but a growth rate $r(N)$ is dubbed the BDICC-bis model. For the BDICC-bis model, the equilibrium distribution of N can still be determined as

$$P(N) = \begin{cases} \dfrac{1}{Z_{\alpha,\mu}}, & N = 0, \\ \dfrac{1}{Z_{\alpha,\mu}} \dfrac{1}{N!} \displaystyle\prod_{k=0}^{N-1} \dfrac{\alpha + r(k)k}{\mu}, & N \geq 1, \end{cases}$$

with $Z_{\alpha,\mu}$ a normalizing constant. However, as shown in Sect. C.5 of the Appendix the BDICC-bis model with population-dependent growth is no longer reversible when enumerated by the species counts c_k and we cannot use detailed balance properties to exactly determine the probability distribution $P(\vec{c})$. Consequently, neither means nor variances of c_k and C can be determined. We thus perform numerical simulations by setting $r(N) = r_1/N$, while keeping α, μ uniform.

We compare results of the BDICC-bis model to those of the sBDI model (α, r, μ uniform) and the previous BDICC model ($\alpha, r, \mu(N) = \mu_1 N$). As in Sect. 5.3 we consider a low immigration rate $\alpha = 0.2$, set $\mu = 1$, and adjust the parameter r_1 so that $\mathbb{E}[N]$ is the same across the three models. Results for the BDICC-bis model are plotted as the blue dashed curves in Fig. 4. Observed trends for the $P(N)$ and $P(C)$ distributions within the BDICC and the BDICC-bis models are similar, as well as for $\mathbb{E}[c_k]$. Shannon's entropy and Simpson's diversity index also remain similar, $\mathbb{E}[H] = 0.25$ and $\mathbb{E}[S] = 0.15$ for the BDICC-bis model, and $\mathbb{E}[H] = 0.26$ and $\mathbb{E}[S] = 0.16$ for the BDICC model.

5.3.3 Quasi-steady State and Reflecting Boundary Conditions

When $\alpha = 0$ in the BDICC model, the $N = 0$ state is a perfect sink. In the absence of immigration, a system cannot escape from the "absorbing" $N = 0$ state. However, in the deterministic limit, the $N = 0$ state is unstable while the finite-population state with N^* individuals is stable (for $\mu(N) = \mu_1 N$, $N^* = r/\mu_1$). Even though the true steady-state of the stochastic problem is $N = 0$, it may take an exponentially long time for a population initially at $N \sim N^*$ to become extinct. Therefore, given a system initiated with a large population $N \sim N^*$, we expect that a quasi-steady state is established before extinction.

To find distributions associated with the long-lived quasi-steady state of the BDICC model, we modify the absorbing boundary condition at $N = 0$ to a reflecting boundary condition by simply preventing the last individual from dying by setting $\mu(N = 1) = 0$. We can now compute the steady state distribution of N using detailed balance to find

$$P(N) = P(1) \frac{1}{N!} \prod_{k=1}^{N-1} \frac{\alpha + rk}{\mu(k+1)},$$

with $P(1)$ being the probability of having one individual. Contrary to the BDICC model with an absorbing boundary condition, we can no longer recurse the detailed balance equations down to $N = 0$, since the last individual cannot die (in other words $P(0) = 0$). By denoting

$$Z'_{\alpha,r} = \sum_{n=1}^{\infty} \frac{1}{n!} \prod_{k=1}^{n-1} \frac{\alpha + rk}{\mu(k+1)},$$

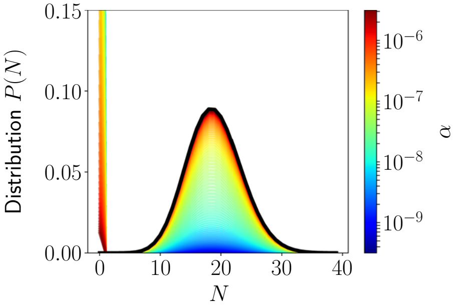

Fig. 5 Comparison of $P(N)$ for the BDICC model with absorbing and reflecting boundary conditions in the small α limit. For both submodels, $r = 0.995$ and $\mu(N) = 0.0498N$, leading to a carrying capacity of $N^* \approx 20$. The thick black curve corresponds to the quasi-steady state for $\alpha = 0$ computed by using a reflecting boundary condition approximation (Eq. (38)). The colored curves correspond to the steady-state distribution of the absorbing model using different values of α. When $\alpha = 0$, the standard absorbing BDICC model leads to an equilibrium "vacuum" or "extinct" state (dark blue), while the BDICC model approximated with reflecting boundary condition leads to a the quasi-steady state distribution $P(N)$ centered about N^* (Color figure online)

we find

$$P(N) = \frac{1}{Z'_{\alpha,r}} \frac{1}{N!} \prod_{k=1}^{N-1} \frac{\alpha + rk}{\mu(k+1)}.$$

Similarly, using the detailed balance equations, we find the distribution of $\vec{c}$,

$$P(\vec{c}) = \frac{P(1,0,\ldots)}{\alpha} \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=2}^{N} \mu(n)} \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!},$$

where $P(1,0,\ldots) = 1/Z'_{\alpha,r}$.

The importance of the quasi-steady state is most discernible in the $\alpha \to 0$ limit where initial conditions determine long-lived configurations. With absorbing boundary conditions, the equilibrium state is the trivial empty state even if it is deterministically unstable. However, by using a reflecting boundary condition on the total population, we can approximate the long-lived quasi-steady state distributions with

$$P(N) \xrightarrow[\alpha\to 0]{} \frac{1}{Z'_{0,r}} \frac{r^{N-1}}{N!} \prod_{k=2}^{N} \frac{1}{\mu(k+1)} \tag{38}$$

$$P(\vec{c}) \xrightarrow[\alpha\to 0]{} \begin{cases} \frac{1}{Z'_{0,r}} \frac{r^{N-1}}{N!} \prod_{k=2}^{N} \frac{1}{\mu(k+1)} & \text{if } \ C = \sum_k c_k = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{39}$$

In this limit, only one species survives and occupies the whole system before final extinction at exponentially long times.

Intuitively, without immigration, new species cannot be introduced in the system, and with probability 1 there will be at some point only one individual in the system. This single-species population persists for a long time before final extinction. This long time persistence is approximated by the reflecting boundary condition that prevents true extinction. Note that this limit is related to species extinction and coarsening in a multispecies Moran model with fixed population size [3]. The distributions $P(N)$ for absorbing and reflecting boundary conditions are compared in Fig. 5 for small α.

6 Summary and Conclusions

In this paper we analyzed three stochastic, neutral birth–death–immigration (BDI) models: the simple BDI (sBDI), BDI with mutations (BDIM), and BDI with carrying capacity (BDICC). Where possible, we derived analytical expressions for the steady-state distribution $P(N)$ of the total population and the steady-state distribution $P(C)$ for the total number of species in the system. In many cases, we were also able to derive expressions for the steady-state distributions of individual subpopulations $P(n_i)$ and $P(c_k)$, given in terms of cells counts n_i and species counts c_k, respectively.

All three models (sBDI, BDIM, and BDICC) analyzed show similar species abundance distribution functions. In particular, we find that the number of individuals in one species n_i follows a strict log-series distribution $P(n_i)$, or, in the case of the BDICC model, can be approximated by one. The prediction that species could follow this type of distribution dates to the early days of theoretical ecology. For example, after analyzing insect abundances in the field, Fisher et al. [14] proposed that the distribution of insect species in an area should follow a geometric or, possibly, a log-series distribution. The log-series distribution has since been widely used in theoretical ecology [4,26,34], but has also been challenged. For instance Preston [30] speculated that actual species abundances would be better described by a log-normal, or possibly a Poisson log-normal distribution [6]. Within immunology, the abundance of T-cell clones appears to follow a power-law distribution, incompatible with a log-series distributions [10]. The log-series characteristic of our BDI models can be linked to their neutrality, i.e. that replication and death rates are independent of the given species.

We also evaluated diversity metrics such as Shannon's entropy and Simpson's diversity index and provided expectations and variances of a number of quantities. Stochastic simulations were also performed and matched with our analytical results. Our analytical results are summarized in Table 1, while Table 2 lists the same results in the large immigration regime. Interestingly, we show that in the fast immigration limit, the diversity indices H and S converge to values independent of the model, but the richness C converges to values that are model-dependent. Only the richness can distinguish the different processes in the fast immigration limit, implying that in this limit it is a more useful diversity metric.

Finally, we confirmed the consistency of detailed balance for a carrying capacity model in which the global interaction is implemented through the death rate (BDICC) but demonstrated that detailed balance is violated if carrying capacity is effected through the birth rate (BDICC-bis model). Nonetheless, this asymmetry generates almost no qualitative difference in the statistical properties when comparing the two models using equal mean total populations.

Many related applications motivate us to extend our work towards non-neutral BDI models. We expect that lifting the neutrality condition will typically generate longer tails in species abundance distributions.

Acknowledgements This work was supported in part by an INRA Contrat Jeune Scientifique Award (RD) and by the National Science Foundation through grants DMS-1814364 (TC) and DMS-1814090 (MD). The authors also thank Song Xu for clarifying discussions.

Mathematical Appendices

A: Simple Birth–Death–Immigration Models (sBDI)

A.1: Finite Number of Species

So far, we have assumed immigration events introduce completely new species to the system, regardless of the existing population structure. Within the context of island biodiversity, this assumption corresponds to the mainland hosting an unlimited number of species, so that individuals who emigrate to the island are always part of a new species. Mathematically, we are assuming that each species immigrates only once.

In this Appendix, we consider an alternative model where the number of mainland species Q is finite. In this case, the probability that a newly immigrated individual belongs to species i (with $1 \leq i \leq Q$) is $1/Q$ and the number of species in the island cannot exceed Q. As a consequence, the total number of species $C \leq Q$, and the number of species with k individuals $c_k \leq Q$ for all k.

The dynamics of the total number of individuals N remains unchanged with respect to the sBDI model, as the type of species immigrating from the mainland does not affect overall birth or death rates. Therefore, the distribution for $P(N)$ remains identical to the one derived in Eq. (5) for the simple BDI model. We can now determine the distribution of $\vec{c}$ in the alternative model using the same approach taken for the sBDI model. Transitions are given by

$$
\begin{array}{lll}
 & (c_1, c_2, \ldots) \xrightarrow{\alpha(1-C/Q)} (c_1+1, c_2, \ldots) & \left.\right\}\begin{array}{l}\text{Immigration}\\ \text{(of new species)}\end{array} \\
\text{for } k \geq 1 & (c_1, \ldots, c_k, c_{k+1}, \ldots) \xrightarrow{(rk+\alpha/Q)c_k} (c_1, \ldots, c_k - 1, c_{k+1}+1, \ldots) & \left.\right\}\begin{array}{l}\text{Birth+Immigration}\\ \text{(ofexistingspecies)}\end{array} \\
\text{for } k \geq 2 & (c_1, \ldots, c_{k-1}, c_k, \ldots) \xrightarrow{\mu k c_k} (c_1, \ldots, c_{k-1}+1, c_k - 1, \ldots) & \multirow{2}{*}{$\left.\right\}$ Death.} \\
 & (c_1, c_2, \ldots) \xrightarrow{\mu c_1} (c_1 - 1, c_2, \ldots) &
\end{array}
$$

Note that the birth process rate is effectively augmented by α/Q, due to the possibility of a new individual immigrating into an existing species. Conversely, the corresponding immigration rate for new species is decreased by $\alpha C/Q$. Also note that the limit $Q \to \infty$ reduces the current model to the original sBDI. Using detailed balanced equations, similarly as in the sBDI model, we can write $P(\vec{c})$ as follows

$$
P(\vec{c}) = \left(1 - \frac{r}{\mu}\right)^{\alpha/r} \frac{Q!}{(Q-C)!} \left(\frac{r}{\mu}\right)^{N} \left(\frac{1}{\prod_{i=1}^{\infty} c_i!}\right) \prod_{\ell=1}^{\infty} \prod_{j=0}^{\ell-1} \left(\frac{j + \frac{\alpha}{Qr}}{j+1}\right)^{c_\ell}.
$$

One can verify that this distribution satisfies all the required transition equations. Yet, contrary to the sBDI model, it is more difficult to determine the distributions of C, c_k and n_i based on this formulation; in particular the factor $Q!/(Q-C)!$ prevents us from applying the same mathematical procedure used in the sBDI case.

We can however take a different route, namely invoking neutrality and the independence of the system, to deduce the distributions of C and c_k. Since each species behaves independently from all others, we can consider the number m_i of individuals in the i^{th} species (with $1 \leq i \leq Q$) independently from the rest. Note that m_i is a random variable that can be zero when there are no individuals of species i present in the system. The quantity m_i is the counterpart to n_i

introduced for the sBDI model with the caveat that n_i represents the number of individuals of a species *actually present* on the island (i.e. $P\left(n_i=0\right)=0$). In the current model n_i can be expressed as a function of m_i via

$$P\left(n_i=k\right)=P\left(m_i=k \mid m_i>0\right) \qquad \text{for } k \geq 1, \tag{40}$$

describing the distribution of the i^{th} species provided that at least one of its individuals is on the island. The random variable m_i follows a birth and death process: its birth rate is $\alpha/Q+rm_i$ and its death rate is μm_i. The α/Q rate corresponds to immigration, the rate rm_i corresponds to actual reproduction. We already determined the steady state distribution of this process in Eq. (5), yielding a negative binomial distribution with parameters $\alpha/(rQ)$ and r/μ as follows

$$P(m_i)=\left(1-\frac{r}{\mu}\right)^{\alpha/(Qr)}\left(\frac{r}{\mu}\right)^{m_i}\frac{1}{m_i!}\prod_{k=0}^{m_i-1}\left(\frac{\alpha}{Qr}+k\right).$$

The $P(n_i)$ distribution can be determined from $P(m_i)$ expressed above, using Eq. (40)

$$P\left(n_i=k\right)=\frac{P\left(m_i=k\right)}{1-P\left(m_i=0\right)}=\frac{\left(1-\frac{r}{\mu}\right)^{\alpha/(Qr)}}{1-\left(1-\frac{r}{\mu}\right)^{\alpha/(Qr)}}\left(\frac{r}{\mu}\right)^{k}\frac{1}{k!}\prod_{k'=0}^{k-1}\left(\frac{\alpha}{Qr}+k'\right)$$

$$\text{for any } k \geq 1.$$

Finally, the number of species c_k with k individuals and the total number of species C can be expressed as a function of m_i as follows

$$c_k=\sum_{i=1}^{Q} I\left(m_i=k\right) \quad \text{and} \quad C=\sum_{i=1}^{Q} I\left(m_i>0\right).$$

Since all m_i are i.i.d., the probability distributions of c_k and C are given by

$$P(c_k)=\binom{Q}{c_k}P\left(m_i=k\right)^{c_k}\left(1-P\left(m_i=k\right)\right)^{Q-c_k},$$

$$P(C)=\binom{Q}{C}\left(1-P\left(m_i=0\right)\right)^{C}P\left(m_i=k\right)^{Q-C},$$

which are binomial distributions of respective parameters Q and $P\left(m_i=k\right)$ for c_k, and Q and $1-P\left(m_i=0\right)$ for C. Note that this approach does not allow us to determine the diversity indices H and S.

A.2: Convergences in the Large Immigration Regime

In this section, we will prove the convergence of

$$N/\Omega, \quad C/\Omega, \quad \left(\frac{c_1}{\Omega},\frac{c_2}{\Omega},\ldots\right), \quad \text{and} \quad H/\log\Omega$$

in the large immigration regime defined by $\alpha=\widetilde{\alpha}\Omega$, $\Omega\to\infty$.

Proposition 1 *The scaled total number of individuals N/Ω converges in distribution to the constant $\widetilde{\alpha}/(\mu-r)$.*

Proof The definition of the convergence in distribution described in Eq. (3) is equivalent to the convergence of its moment generating function. One is left with showing that

$$\text{for any } \xi < 0, \quad \lim_{\Omega\to\infty} \mathbb{E}\left[e^{\xi N/\Omega}\right] = \frac{\widetilde{\alpha}}{\mu - r}$$

(see for instance [5, Chapter 5]). Since $N \sim \text{NegBinom}\,(\widetilde{\alpha}\Omega/r, r/\mu)$ for which the moment generating function is known, we have for any $\xi < 0$:

$$\mathbb{E}\left[e^{\xi N/\Omega}\right] = \left(\frac{1 - r/\mu}{1 - e^{\xi/\Omega} r/\mu}\right)^{\widetilde{\alpha}\Omega/r}.$$

Upon taking the logarithm of the previous expression, we find

$$\begin{aligned} \log\left[\mathbb{E}\left[e^{\xi N/\Omega}\right]\right] &= \frac{\widetilde{\alpha}\Omega}{r}\left[\log\left(1 - r/\mu\right) - \log\left(1 - e^{\xi/\Omega} r/\mu\right)\right] \\ &\times \underset{\Omega\to\infty}{\sim} -\frac{\widetilde{\alpha}\Omega}{r} \log\left[1 - \frac{\xi}{\Omega}\frac{r/\mu}{1 - r/\mu}\right] \\ &\times \underset{\Omega\to\infty}{\sim} \frac{\widetilde{\alpha}\Omega}{r}\frac{\xi}{\Omega}\frac{r/\mu}{1 - r/\mu} = \xi\frac{\widetilde{\alpha}}{\mu - r}, \end{aligned}$$

so

$$\mathbb{E}\left[e^{\xi N_\Omega/\Omega}\right] \xrightarrow[\Omega\to\infty]{} \exp\left[\xi\frac{\widetilde{\alpha}}{\mu - r}\right],$$

thus proving the proposition. □

Proposition 2 *The scaled total number of species C/Ω converges in distribution to*

$$\frac{C}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r} \log\left[\frac{1}{1 - r/\mu}\right].$$

Proof The proof is similar to Proposition 1. □

Proposition 3 *For each $k > 0$, c_k/Ω converges in distribution to*

$$\frac{c_k}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\frac{(r/\mu)^k}{k}.$$

Proof For any vector $\vec{c}$ and $k \geq 1$, we have that

$$c_k = \sum_{i=1}^{C} \boldsymbol{I}\left(n_i, k\right).$$

Consider the moment generating function of the random variable c_k. For any $\xi < 0$, we have

$$\mathbb{E}\left[e^{\xi c_k/\Omega}\right] = \mathbb{E}\left[\exp\left(\frac{\xi}{\Omega}\sum_{i=1}^{C} \boldsymbol{I}\left(n_i, k\right)\right)\right].$$

Since n_i are identical and independently distributed and independent of C, and since their distributions do not depend on the parameter Ω, it follows that

$$\begin{aligned}\mathbb{E}\left[e^{\xi c_k/\Omega}\right] &= \mathbb{E}\left[\left(\mathbb{E}\left[\exp\left(\frac{\xi}{\Omega}\boldsymbol{I}\left(n_1, k\right)\right)\right]\right)^C\right] \\ &= \mathbb{E}\left[\left(e^{\xi/\Omega}P(n_1 = k) + (1 - P(n_1 = k))\right)^C\right] \\ &= \mathbb{E}\left[\left(\left(e^{\xi/\Omega} - 1\right)P(n_1 = k) + 1\right)^C\right].\end{aligned}$$

Since the probability distribution of n_1 is known, we have

$$\mathbb{E}\left[e^{\xi c_k/\Omega}\right] = \mathbb{E}\left[\left(1 - \frac{1}{k}\left(\frac{r}{\mu}\right)^k \frac{(e^{\xi/\Omega} - 1)}{\log(1 - r/\mu)}\right)^C\right].$$

Note that for any real A,

$$\begin{aligned}C\log\left[1 - \left(e^{\xi/\Omega} - 1\right)A\right] &\underset{\Omega\to\infty}{\sim} -C\left(e^{\xi/\Omega} - 1\right)A, \\ &\underset{\Omega\to\infty}{\sim} -\frac{C}{\Omega}\xi A.\end{aligned}$$

Considering the exponential of this expression, we have

$$\mathbb{E}\left[e^{\xi c_k/\Omega}\right] = \mathbb{E}\left[\exp\left[-\frac{C}{\Omega}\frac{(r/\mu)^k}{k}\frac{\xi}{\log(1 - r/\mu)}\right]\right].$$

Finally, since we have already shown that C/Ω converges in distribution (Proposition 2 above), we find

$$\lim_{\Omega\to\infty}\mathbb{E}\left[e^{\xi c_k/\Omega}\right] = \exp\left(\xi\frac{\widetilde{\alpha}}{r}\frac{(r/\mu)^k}{k}\right).$$

□

Proposition 4 *The Shannon's Entropy H converges in distribution as*

$$\frac{H}{\log\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1.$$

Proof Using the definition of H,

$$\frac{H}{\log\Omega} = \sum_{k=1}^{\infty} k\frac{c_k}{\Omega}\frac{\Omega}{N}\frac{\log N - \log k}{\log\Omega},$$

where c_k/Ω and N/Ω converge in distribution to known constants, we find

$$\frac{H}{\log\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\mu - r}{r}\sum_{k=1}^{\infty}\left(\frac{r}{\mu}\right)^k = 1$$

□

Proposition 5 *The Simpson's diversity index S converges in distribution as*

$$S \xrightarrow[\Omega\to\infty]{\mathcal{D}} 1.$$

Proof By the definition of S (Eq. (2))

$$S = 1 - \frac{1}{\Omega} \sum_{k=1}^{\infty} \frac{c_k}{\Omega} \left(\frac{k}{N/\Omega} \right)^2 ,$$

and since c_k/Ω and N/Ω converge in distribution to known constants, we find

$$S \xrightarrow[\Omega \to \infty]{\mathcal{D}} -\frac{1}{\Omega} \sum_{k=1}^{\infty} \frac{\widetilde{\alpha}}{r} k \left(\frac{r}{\mu} \right)^k \left(\frac{\mu - r}{\widetilde{\alpha}} \right)^2 = 1 - \frac{(\mu - r)^2}{\Omega r \widetilde{\alpha}} \sum_{k=1}^{\infty} k \left(\frac{r}{\mu} \right)^k .$$

One can then recognize the power series identity

$$\frac{r/\mu}{(1 - r/\mu)^2} = \sum_{k=1}^{\infty} k \left(\frac{r}{\mu} \right)^k$$

and hence show that the second term vanishes as $\Omega \to \infty$ and deduce the result $S \xrightarrow[\Omega \to \infty]{\mathcal{D}} 1$. □

B: BDI Model with Mutation (BDIM)

B.1: Distribution of the Number of Individual in One Species

We propose an argument for a Log-series distribution of any species

$$\pi_k = P(n_i = k)$$

when all species are independent of each other. There are several ways to interpret π_k. First consider the explicit dynamics of each species. Denote by $m_q(t)$ the number of individuals of species q at time t and define a_q as the time of arrival (by convention, we order the species such as $a_0 = 0 < a_1 < a_2 < \ldots$) and d_q its "lifespan", i.e. the species will be extinct at time $a_q + d_q$ (see the example in Fig. 8a). Note that the index q indicates the order of arrival (and not the species identity index i used in the main article), and that the distribution of the times a_q is not specified and can be adapted to any rate of species creation (either by immigration or by mutation). The evolution of each species is independent of each other, and each of them defines an identically distributed birth–death process characterized by the following transitions

$$\begin{cases} m_q \to m_q + 1 & \text{at rate } m_q \, r(1 - \epsilon), \\ m_q \to m_q - 1 & \text{at rate } m_q \, \mu. \end{cases} \tag{41}$$

Due to the $r < \mu$ assumption, this process will become extinct almost surely [2, Chapter 2] and the lifespan d_q of each species is finite (Figs. 6 and 7).

In the main article, we interpreted π_k as the number of individuals in a given species at steady state, that is to say, we considered the $T \to \infty$ limit

$$\pi_k = \lim_{T \to \infty} P\left(m_{J_T}(T) = k \right)$$

where J_T is the index of a randomly sampled species among those that exist at time T; i.e., J_T is uniformly chosen among all the species q such that $a_q < T < a_q + d_q$.

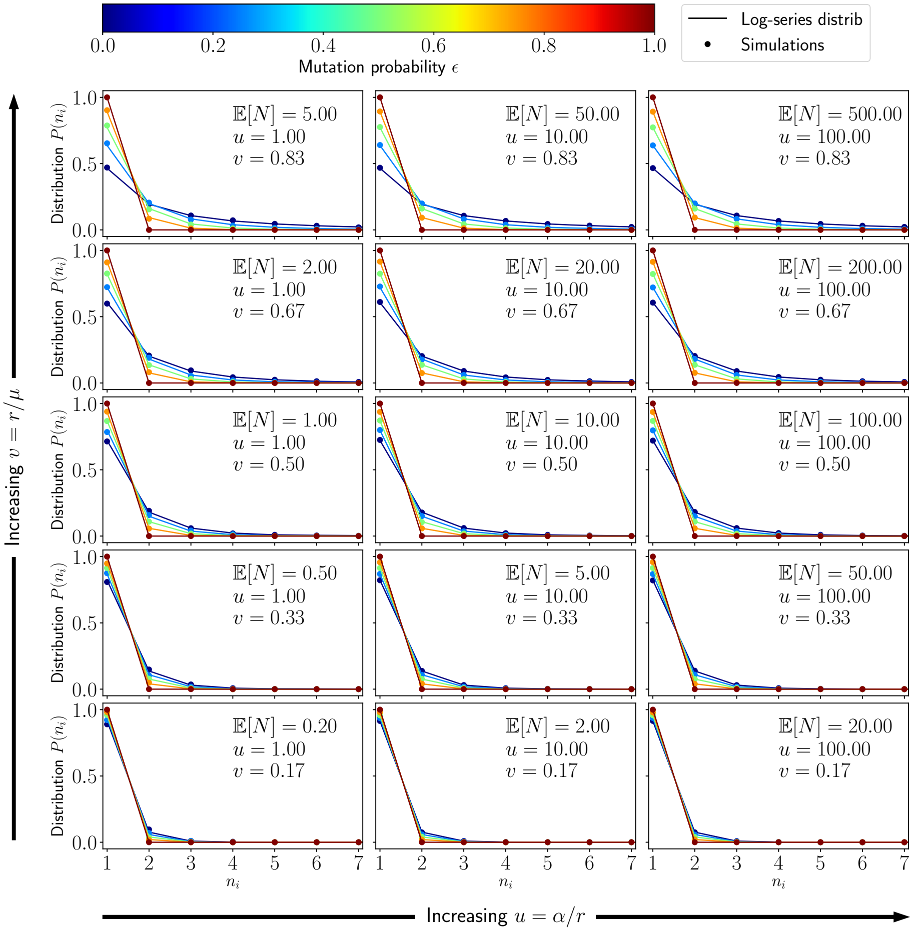

Fig. 6 Distribution of the number of individuals in one species n_i under different parameter choices. Dots represent simulations for various values of $u = \alpha/r$, $v = r/\mu$, $(r = 1)$ and ϵ; solid lines depict logarithmic distributions with parameter $r(1-\epsilon)/\mu$. As expected, the logarithmic distributions match the simulations n_i, and the distributions of n_i do not depend on u

However, there is another way to interpret π_k. Consider all species that *exist or have existed* up to time T and then randomly select one of them, species I_T. The number of individuals in species I_T at a randomly chosen time τ_{I_T} between the introduction of the species (at time a_{I_T}) and the extinction (at time $a_{I_T} + d_{I_T}$) is denoted m_{I_T}. In this picture, we can characterize π_k according to

$$\pi_k = \lim_{T\to\infty} P\left(m_{I_T}(\tau_{I_T}) = k\right). \tag{42}$$

The main difference between the two approaches is that, in the first case, we sample among the species that exist at a precise time T before taking $T \to \infty$, while in the second case, we sample among all the species that existed before time T (before taking $T \to \infty$).

For a fixed time T, the last species introduced in the system is given by

$$Q_T = \underset{q\in\mathbb{N}}{\operatorname{argmax}} \left(a_q < T\right).$$

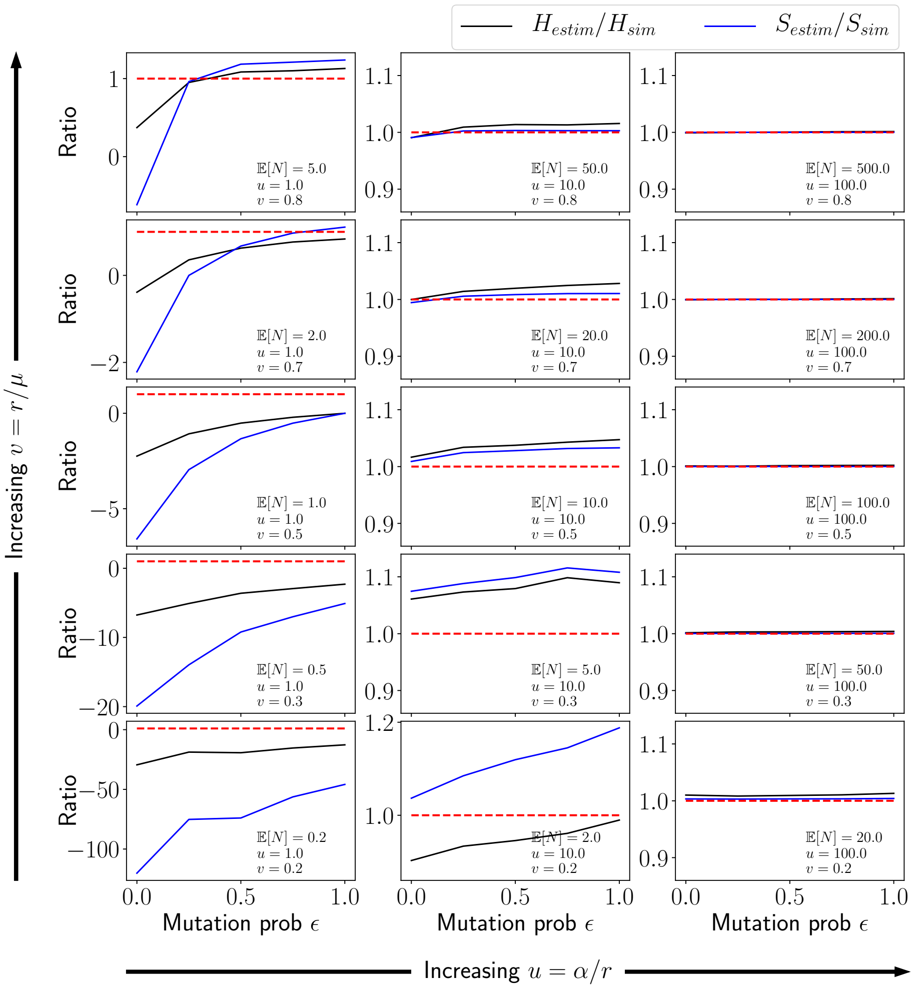

Fig. 7 Accuracy of Shannon's entropy and Simpson's index (as defined in Eq. (25)). We plot the ratio of the estimates of Shannon's entropy and Simpson's index and their respective values measured via simulation for different $u = \alpha/r$, $v = r/\mu$ (by taking $r = 1$), and different ϵ. The estimates become more accurate as $\mathbb{E}[N]$ increases: the error is below 10% for any parameters u, v, ϵ such that $\mathbb{E}[N]$ is larger than 5

All species that exist or have existed before time T are in the set $\{0, \ldots, Q_T\}$. Note that since a_q are increasing in q, $\lim_{T\to\infty} Q_T = \infty$. As per Eq. (42), we have to sample one species among the set $\{0, \ldots, Q_T\}$. One key point is that the random selection is not uniform: there is a higher chance of selecting species with longer lifespans. If I_T is the index of the randomly chosen species, we can write

$$P\left(I_T = q\right) = \boldsymbol{I}\left(q \leq Q_T\right) \frac{d_q}{\sum_{j=0}^{Q_T} d_j}.$$

The first term $\boldsymbol{I}\left(q \leq Q_T\right)$ ensures that the species q exists before time T while the second term proportionally weights the probability of sampling according to their lifespans. Con-

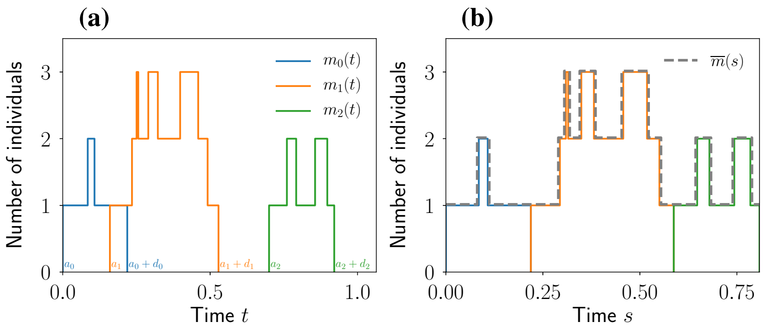

Fig. 8 **a** A representative trajectory of three immigrated species. The q-th species is introduced at time a_q and extinguishes at time $a_q + d_q$. **b** Construction of the process $\overline{m}$ (defined in Eq. (43)) by stacking and concatenating the trajectories of each species $\left(m_q\right)_{q\in\mathbb{N}}$

ditioned on species I_T having been sampled, we then randomly chose a time τ_{I_T} uniformly distributed between a_{I_T} and $a_{I_T} + d_{I_T}$.

Proposition 6 *The limiting distribution becomes*

$$\pi_k = \lim_{T\to\infty} P\left(m_{I_T}(\tau_{I_T}) = k\right) = \frac{1}{\log\left(1 - \frac{r(1-\epsilon)}{\mu}\right)} \frac{1}{k} \left(\frac{r(1-\epsilon)}{\mu}\right)^k .$$

Proof By summing over all possible species q, we can write

$$\begin{aligned} P\left(m_{I_T}(\tau_{I_T}) = k\right) &= \sum_{q\in\mathbb{N}} \mathbb{E}\left[\boldsymbol{I}\left(q \le Q_T\right) \frac{d_q}{\sum_{j=0}^{Q_T} d_j} \boldsymbol{I}\left(m_q(\tau_q), k\right)\right] \\ &= \mathbb{E}\left[\sum_{q=0}^{Q_T} \frac{d_q}{\sum_{j=0}^{Q_T} d_j} \frac{1}{d_q} \int_{a_q}^{a_q+d_q} \boldsymbol{I}\left(m_q(t), k\right) \mathrm{d}t \right] \\ &= \mathbb{E}\left[\frac{\sum_{q=0}^{Q_T} \int_{a_q}^{a_q+d_q} \boldsymbol{I}\left(m_q(t), k\right) \mathrm{d}t}{\sum_{j=0}^{Q_T} d_j} \right] \end{aligned}$$

Next, consider the process $\overline{m}(s)$ defined as

$$\overline{m}(t) = m_{\nu_t}\left(t - \overline{d}_{\nu(t)} + a_{\nu(t)}\right) \tag{43}$$

with

$$\overline{d}_k = \sum_{q=1}^{k-1} d_q \quad \text{and} \quad \nu(t) = \underset{q}{\operatorname{argmax}} \left\{ \sum_{j=0}^{q} d_j < t \right\}.$$

The process $\overline{m}$ is simply the stacking of all the processes m_q in the sense that the process $\overline{m}(t)$ for t between $\overline{d}_q$ and $\overline{d}_{q+1}$ will be equal to the process $m_q(s)$ for $s = t - \overline{d}_q + a_q$ between a_q and its extinction time $a_q + d_q$ (see the example on Fig. 8b). With this stacked process,

$$P\left(m_{I_T}(\tau_{I_T}) = k\right) = \mathbb{E}\left[\frac{\int_0^{a_{\delta_T}+d_{\delta_T}} \boldsymbol{I}\left(\overline{m}(t), k\right) \mathrm{d}t}{\overline{d}_{Q_T}} \right].$$

By ergodicity of the process $\overline{m}$, we have

$$\lim_{T\to\infty} P\left(m_{I_T}(\tau_{I_T}) = k\right) = \lim_{T\to\infty} P\left(\overline{m}(T) = k\right).$$

Finally, we have to determine the steady state of the process $\overline{m}$. Since the transitions of the process $\overline{m}$ are a simple birth–death process

$$\begin{cases} \overline{m} \to \overline{m}+1 & \text{at rate} \quad \overline{m}\, r(1-\epsilon) \\ \overline{m} \to \overline{m}-1 & \text{at rate} \quad \overline{m}\, \mu \boldsymbol{I}\, (\overline{m} > 0)\,. \end{cases}, \tag{44}$$

we have that its equilibrium distribution is a logarithmic series distribution with parameter $p \equiv r(1-\epsilon)/\mu$ (by imposing equations of detailed balance). □

B.2: Moments of C

The third relation of Eq. (1) yields the following expression for the moment generating function of N:

$$\mathbb{E}\left[e^{\xi N}\right] = \mathbb{E}\left[\prod_{i=1}^{C} e^{\xi n_i}\right],$$

for any $\xi < 0$. Since all the $(n_i)_{i\leq C}$ are identical and independently distributed and independent of C, we have

$$\mathbb{E}\left[e^{\xi N}\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\xi n_1}\right]^C\right] = \mathbb{E}\left[\left(\frac{\log\left(1-pe^{\xi}\right)}{\log\left(1-p\right)}\right)^C\right]. \tag{45}$$

Equation (20) shows that the distribution over n_1 is a log-series distribution with parameter $p = r\left(1-\epsilon\right)$. By redefining the variable ξ' such that $e^{\xi'} := \log\left(1-pe^{\xi}\right)/\log\left(1-p\right)$ and eliminating ξ for ξ', Eq. (45) becomes an expression for the moment generating function of C,

$$\mathbb{E}\left[e^{\xi' C}\right] = \left(\frac{1-r/\mu}{1-\frac{r}{\mu}\frac{1-(1-p)^{e^{\xi'}}}{p}}\right)^{\alpha/r}.$$

By differentiating this expression, we can determine the second moment of C:

$$\mathbb{E}\left[C^2\right] = \lim_{\xi'\to 0}\frac{\mathrm{d}^2}{\mathrm{d}\xi'^2}\mathbb{E}\left[e^{\xi' C}\right] = \mathbb{E}\left[C\right]\left[1+\log\left(1-p\right)+\left(1+\frac{r}{\alpha}\right)\mathbb{E}\left[C\right]\right],$$

which yields the expression for $\operatorname{var}\left[C\right]$ in Eq. (22).

C: BDI Model with Carrying Capacity (BDICC)

C.1: Steady State Distribution of $\vec{c}$

To determine $P(\vec{c})$, the probability of occurrence of the species-count state $\vec{c}$, first consider a finite $K = \operatorname{argmax}_i(c_i > 0)$. As explained in the main text, if the system is reversible, one instance of Eq. (27) is

$$\mu(N) c_K K P(\vec{c}) = (K-1)\,(c_{K-1}+1)\, r P(c_1, \ldots, c_{K-1}+1, c_K - 1, \vec{0}).$$

Recursively unwinding this relationship, we find

$$P(\vec{c}) = P(c_1, \ldots, c_{K-1}+1, c_K - 1, \vec{0}) \frac{r}{\mu(N)} \frac{K-1}{K} \frac{c_{K-1}+1}{c_K},$$

$$P(\vec{c}) = P(c_1, \ldots, c_{K-1}+c_K, 0, \vec{0}) \frac{r^{c_K}}{\mu(N) \ldots \mu(N - c_K + 1)} \left(\frac{K-1}{K}\right)^{c_K} \frac{(c_{K-1}+c_K)!}{c_K! c_{K-1}!},$$

$$P(\vec{c}) = P(C, \vec{0}) \frac{r^{N-C}}{\mu(N) \ldots \mu(N-(K-1)c_K - \ldots - c_2 + 1)} \prod_{i=1}^{K-1} \prod_{j=i+1}^{K} \left(\frac{i}{i+1}\right)^{c_j} \frac{C!}{\prod_{i=1}^{K} c_i!},$$

$$P(\vec{c}) = P(C, \vec{0}) \frac{r^{N-C}}{\prod_{n=1}^{N-C} \mu(N-n+1)} \frac{C!}{\prod_{i=1}^{K} i^{c_i} c_i!}.$$

After applying Eq. (28), we have by recursion

$$P(C, 0, \ldots) = \frac{\alpha}{\mu(C)} \frac{1}{C} P(C-1, 0, \ldots) = \frac{\alpha^C}{C! \prod_{i=1}^{C} \mu(i)} P(0, \ldots),$$

and

$$P(\vec{c}) = P(0, \ldots) \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=1}^{N} \mu(n)} \frac{1}{\prod_{i=1}^{K} i^{c_i} c_i!}.$$

Since the state $\vec{c} = \vec{0}$ uniquely corresponds to the state $N = 0$ and the above expression holds for K arbitrarily large, it follows that

$$P(\vec{c}) = \frac{1}{Z_{\alpha,r,\mu}} \left(\frac{\alpha}{r}\right)^C \frac{r^N}{\prod_{n=1}^{N} \mu(n)} \frac{1}{\prod_{i=1}^{\infty} i^{c_i} c_i!}. \tag{46}$$

One can verify that this steady-state distribution satisfies the detailed balanced conditions connecting all pairs of states:

$$\begin{cases} \mu\left(\sum_k k c_k\right) k c_k P(\vec{c}) = (k-1)\,(c_{k-1}+1)\, r P(c_1, \ldots, c_{k-1}+1, c_k - 1, \ldots) & \forall k > 1, \\ \mu\left(\sum_k k c_k\right) c_1 P(\vec{c}) = \alpha P(c_1 - 1, \ldots, c_k, \ldots). \end{cases} \tag{47}$$

C.2: Convergence of N/Ω

Theorem 7 *The random variable N/Ω converges in probability to the real n^* which is the only solution of the fixed point Eq.* (36).

To prove this Theorem, first define

$$f(x) := \frac{\widetilde{\alpha} + rx}{x \widetilde{\mu}(x)} \quad \text{and} \quad f_k := \frac{\widetilde{\alpha} + (k-1)r/\Omega}{(k/\Omega)\widetilde{\mu}(k/\Omega)} \quad \forall k \in \mathbb{N}_*.$$

The function f defines the steady-state constraint on $n = N/\Omega$ given by Eq. (36) where $x = n^*$ is the only real solution to $f(x) = 1$. With these definitions, the probability distribution over N can be expressed as

$$\forall n \in \mathbb{N}, \qquad P\,(N = n) = \frac{\exp\left(\sum_{k=1}^{n} \log f_k\right)}{\sum_{n'=0}^{\infty} \exp\left(\sum_{k=1}^{n'} \log f_k\right)}.$$

Now, consider the following lemma:

Lemma 8 *The function f is strictly decreasing and there exists a Ω^* for which $\forall \Omega \geq \Omega^*$, $(f_k)_{k\geq 1}$ is a decreasing sequence.*

Proof The decrease of the function f is a direct implication of the increase of $\widetilde{\mu}$. For, $(f_k)_{k\geq 1}$ we have

$$(k+1)\,(\widetilde{\alpha}\Omega + r(k-1)) - k\;(\widetilde{\alpha}\Omega + rk) = \widetilde{\alpha}\Omega - r,$$

which is positive for large enough Ω. Since $\widetilde{\mu}$ is increasing,

$$\frac{f_k}{f_{k+1}} = \frac{\widetilde{\mu}((k+1)\,/\,\Omega)}{\widetilde{\mu}(k/\,\Omega)}\,\frac{(k+1)\,(\widetilde{\alpha}\Omega + (k-1)r)}{k\;(\widetilde{\alpha}\Omega + rk)} > 1.$$

□

To prove Theorem 7, we have to show that $\forall \delta > 0$,

$$P\left(\left|N/\Omega - n^*\right| > \delta\right) \xrightarrow[\Omega\to\infty]{} 0$$

that is to say, we have to show that

$$P\left(N/\Omega > n^* + \delta\right) \xrightarrow[\Omega\to\infty]{} 0, \tag{48}$$

$$P\left(N/\Omega < n^* - \delta\right) \xrightarrow[\Omega\to\infty]{} 0. \tag{49}$$

The proofs of convergence for both limits above are very similar so we will focus on the proof of Eq. (48). To simplify notation, we define $a_{\Omega,\delta} \equiv \lceil \Omega\,(n^* + \delta)\rceil$, (where $\lceil\cdot\rceil$ is the ceiling function). Since the distribution of N is known, we have

$$\begin{aligned} P\left(N/\Omega > n^* + \delta\right) &= \frac{\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right)}{\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(\sum_{k=1}^{n} \log f_k\right) + \sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right)} \\ &= \left(\frac{\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(\sum_{k=1}^{n} \log f_k\right)}{\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right)} + 1\right)^{-1}. \end{aligned}$$

Thus, it is enough to show

$$\frac{\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(\sum_{k=1}^{n} \log f_k\right)}{\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right)} \xrightarrow[\Omega\to\infty]{} \infty$$

in order to prove the convergence of Eq. (48).

Proposition 9 *In the $\Omega \to \infty$ limit, the following equivalence holds*

$$\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right) \underset{\Omega\to\infty}{\sim} \exp\left(\sum_{k=1}^{a_{\Omega,\delta}-1} \log f_k\right) \frac{1}{1 - f\,(n^* + \delta)}$$

Proof We first decompose the sum according to

$$\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=1}^{n} \log f_k\right) = \exp\left(\sum_{k=1}^{a_{\Omega,\delta}-1} \log f_k\right) \sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=a_{\Omega,\delta}}^{n} \log f_k\right).$$

The second term of the decomposition can be rewritten as

$$\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=a_{\Omega,\delta}}^{n} \log f_k\right) = \sum_{n=0}^{\infty} \exp\left(\sum_{k=0}^{n} \log f_{k+a_{\Omega,\delta}}\right).$$

Since $a_{k,\Omega}/\Omega \xrightarrow[\Omega\to\infty]{} n^* + \delta$, it follows that

$$\sum_{k=0}^{n} \log f_{k+a_{\Omega,\delta}} \underset{\Omega\to\infty}{\sim} n \log f(n^* + \delta).$$

As f is a strictly decreasing function (cf. Lemma 8), and since n^* is the only point where $f(n^*) = 1$, it follows that $f(n^* + \delta) < 1$. Therefore, the sum over n converges, and we have

$$\sum_{n=a_{\Omega,\delta}}^{\infty} \exp\left(\sum_{k=a_{\Omega,\delta}}^{n} \log f_k\right) \underset{\Omega\to\infty}{\sim} \frac{1}{1 - f(n^* + \delta)}$$

□

With the previous Proposition, it is enough to prove that the ratio

$$\frac{\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(\sum_{k=1}^{n} \log f_k\right)}{\exp\left(\sum_{k=1}^{a_{\Omega,\delta}-1} \log f_k\right)} = \sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(-\sum_{k=n+1}^{a_{\Omega,\delta}-1} \log f_k\right)$$

diverges to infinity in order to prove the convergence of Eq. (48).

Proposition 10 *The sum*

$$\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(-\sum_{k=n+1}^{a_{\Omega,\delta}-1} \log f_k\right) \xrightarrow[\Omega\to\infty]{} \infty$$

diverges.

Proof Since $(f_k)_{k\geq 1}$ is decreasing for large Ω (cf. Lemma 8), we have

$$\sum_{k=n+1}^{a_{\Omega,\delta}-1} \log f_k \leq \left(a_{\Omega,\delta} - n - 1\right) \log f_{a_{\Omega,\delta}-1}$$

for sufficiently large Ω. Therefore,

$$\sum_{n=0}^{a_{\Omega,\delta}-1} \exp\left(-\sum_{k=n+1}^{a_{\Omega,\delta}-1} \log f_k\right) \geq \sum_{n'=1}^{a_{\Omega,\delta}} \left(\frac{1}{f_{a_{\Omega,\delta}-1}}\right)^{n'}.$$

Since

$$f_{a_{\Omega,\delta}-1} \xrightarrow[\Omega\to\infty]{} f(n^* + \delta),$$

for large enough Ω and since f is decreasing, we have that $f_{a_{\Omega,\delta}-1} < 1 - \eta$ for η small enough. Therefore, we conclude the divergence

$$\sum_{n'=1}^{a_{\Omega,\delta}} \left(\frac{1}{f_{a_{\Omega,\delta}-1}}\right)^{n'} \xrightarrow[\Omega\to\infty]{} \infty$$

and proof of the proposition. □

With this Proposition, we have proven the convergence of Eq. (48). The convergence of Eq. (49) can be proved using exactly the same methods by considering $b_{\Omega,\delta} = \lfloor \Omega\,(\delta + n^*) \rfloor$ instead of $a_{\Omega,\delta}$.

C.3: Convergence of C/Ω

Theorem 11 *The scaled total number of species C/Ω converges in distribution to*

$$\frac{C}{\Omega} \xrightarrow[\Omega\to\infty]{\mathcal{D}} \frac{\widetilde{\alpha}}{r}\log\left[1+\frac{r}{\widetilde{\alpha}}n^*\right] = \frac{\widetilde{\alpha}}{r}\log\left[\frac{1}{1-r/\widetilde{\mu}(n^*)}\right],$$

in which n^* is the only real solution of the fixed point Eq. (36).

Proof One has to prove that

$$\mathbb{E}\left[\exp\left[\xi C/\Omega\right]\right] = \frac{Z_{\alpha e^{\xi/\Omega},r,\mu}}{Z_{\alpha,r,\mu}} \xrightarrow[\Omega\to\infty]{} \left(\frac{1}{1-r/\widetilde{\mu}(n^*)}\right)^{\xi\widetilde{\alpha}/r} = \left(1+\frac{r}{\widetilde{\alpha}}n^*\right)^{\xi\widetilde{\alpha}/r}$$

with

$$Z_{\alpha,r,\mu} = \sum_{n'=0}^{\infty}\exp\left(\sum_{k=1}^{n'}\log\frac{\widetilde{\alpha}+r(k-1)/\Omega}{k/\Omega\,\widetilde{\mu}(k/\Omega)}\right).$$

First note that

$$\begin{aligned}
\mathbb{E}\left[\exp\left[\xi C/\Omega\right]\right] &= \frac{1}{Z_{\alpha,r,\mu}}\sum_{n=0}^{\infty}\exp\left(\sum_{k=1}^{n}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+r(k-1)/\Omega}{k/\Omega\,\widetilde{\mu}(k/\Omega)}\right)\\
&= \sum_{n=0}^{\infty}P\left(N=n\right)\exp\left(\sum_{k=1}^{n}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+r(k-1)/\Omega}{\widetilde{\alpha}+r(k-1)/\Omega}\right)\\
&= \mathbb{E}\left[\exp\left(\sum_{k=1}^{N}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+r(k-1)/\Omega}{\widetilde{\alpha}+r(k-1)/\Omega}\right)\right]
\end{aligned}$$

Since N/Ω converges in probability to n^*,

$$\mathbb{E}\left[\exp\left[\xi C/\Omega\right]\right] \underset{\Omega\to\infty}{\sim} \exp\left(\sum_{k=1}^{n^*\Omega}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+r(k-1)/\Omega}{\widetilde{\alpha}+r(k-1)/\Omega}\right).$$

Since the function $\log\left(\frac{\widetilde{\alpha}e^{\xi/\Omega}+r(x-1)/\Omega}{\widetilde{\alpha}+r(x-1)/\Omega}\right)$ is decreasing in x, we can bound the sum with its lower and upper integral bounds

$$\begin{aligned}
\int_1^{n^*\Omega+1}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+(x-1)r/\Omega}{\widetilde{\alpha}+(x-1)r/\Omega}\,\mathrm{d}x &\le \sum_{k=1}^{n^*\Omega}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+(k-1)r/\Omega}{\widetilde{\alpha}+(k-1)r/\Omega}\\
&\le \frac{\xi}{\Omega}+\int_1^{n^*\Omega}\log\frac{\widetilde{\alpha}e^{\xi/\Omega}+(x-1)r/\Omega}{\widetilde{\alpha}+(x-1)r/\Omega}\,\mathrm{d}x.
\end{aligned}$$

After rescaling $y = (x-1)/\Omega$, the bounds can be expressed as

$$\Omega \int_0^{n^*+1/\Omega} \log \frac{\widetilde{\alpha} e^{\xi/\Omega} + ry}{\widetilde{\alpha} + ry}\,\mathrm{d}y \leq \sum_{k=1}^{n^*\Omega} \log \frac{\widetilde{\alpha} e^{\xi/\Omega} + (k-1)r/\Omega}{\widetilde{\alpha} + (k-1)r/\Omega}$$
$$\leq \frac{\xi}{\Omega} + \Omega \int_0^{n^*} \log \frac{\widetilde{\alpha} e^{\xi/\Omega} + ry}{\widetilde{\alpha} + ry}\,\mathrm{d}y$$

Upon taking $\Omega \to \infty$ and expanding the above expression, we find that both bounds converge to

$$\xi \int_0^{n^*} \frac{\widetilde{\alpha}}{\widetilde{\alpha} + ry}\,\mathrm{d}y.$$

Thus, we find

$$\mathbb{E}\left[\exp\left[\xi C/\Omega\right]\right] \underset{\Omega\to\infty}{\sim} \exp\left(\xi\widetilde{\alpha} \int_0^{n^*} \frac{1}{\widetilde{\alpha} + ru}\,\mathrm{d}u\right) = \exp\left(\xi \frac{\widetilde{\alpha}}{r} \log\left[1 + \frac{r}{\widetilde{\alpha}} n^*\right]\right).$$

□

C.4: Convergence of n_i

Proposition 12 *The marginal probability over each particle count n_i converges according to*

$$P\left(n_1 = k\right) \xrightarrow[\Omega\to\infty]{} \frac{1}{k}\left(\frac{r}{\widetilde{\mu}(n^*)}\right)^k \frac{-1}{\log\left[1 - r/\widetilde{\mu}(n^*)\right]}.$$

Proof The n_i values are identically distributed, so that for any $i, j \leq C$,

$$\text{for any } k \geq 1, \quad P\left(n_i = k\right) = P\left(n_j = k\right).$$

We can then compute the expectation

$$\mathbb{E}\left[\frac{c_k}{C}\right] = \mathbb{E}\left[\sum_{i=1}^{C} \frac{\mathbb{E}\left[\boldsymbol{I}\left(n_i, k\right)|C\right]}{C}\right] = \mathbb{E}\left[\mathbb{E}\left[\boldsymbol{I}\left(n_1, k\right)|C\right]\right] = P\left(n_1 = k\right).$$

This expectation is over a product of two converging quantities:

$$\mathbb{E}\left[\frac{c_k}{C}\right] = \mathbb{E}\left[\frac{c_k}{\Omega}\frac{\Omega}{C}\right] = P(n_1 = k),$$

where c_k/Ω and C/Ω converge in distribution to constants

$$\left(\frac{c_k}{\Omega}, \frac{C}{\Omega}\right) \xrightarrow[\Omega\to\infty]{\mathcal{D}} \left(\frac{\widetilde{\alpha}}{r}\log\left[\frac{1}{1 - r/\widetilde{\mu}(n^*)}\right], \frac{\widetilde{\alpha}}{r}\frac{1}{k}\left(\frac{r^k}{\widetilde{\mu}(n^*)}\right)^k\right).$$

We now apply the mapping theorem (see [5, Chapter 5]) to $\mathbb{E}\left[g\left(\frac{c_k}{\Omega}, \frac{C}{\Omega}\right)\right]$ for any continuous function g to obtain

$$P(n_1 = k) \xrightarrow[\Omega\to\infty]{} \frac{1}{k}\left(\frac{r}{\widetilde{\mu}(n^*)}\right)^k \frac{-1}{\log[1 - r/\widetilde{\mu}(n^*)]}.$$

□

C.5: Explicit Breakdown of Detailed Balance in the BDICC-bis Model with Birth-Mediated Carrying Capacity

Here, we consider a birth–death–immigration model with carrying capacity but contrary to the BDICC model presented in Fig. 1c, the carrying capacity is on the birth rate $r(N)$, and the death rate μ is a constant. By analogy with the BDICC analysis, we find a sufficient condition for a steady state to exist

$$\lim_{N\to\infty} r(N) < \mu.$$

The distribution $P(N)$ of the total number of individuals is given by

$$P(N) = \begin{cases} \dfrac{1}{Z_{\alpha,\mu}}, & N = 0, \\ \dfrac{1}{Z_{\alpha,\mu}} \dfrac{1}{N!} \displaystyle\prod_{k=0}^{N-1} \dfrac{\alpha + r(k)k}{\mu}, & N \geq 1, \end{cases}$$

where

$$Z_{\alpha,\mu} = 1 + \sum_{N=1}^{\infty} \frac{1}{N!} \prod_{k=0}^{N-1} \frac{\alpha + r(k)k}{\mu}.$$

All possible transitions of the BDICC-bis model are given by

$$\begin{array}{lll} & (c_1, c_2, \ldots) \xrightarrow{\alpha} (c_1 + 1, c_2, \ldots) & \text{Immigration} \\ \text{for } k \geq 1 & (c_1, \ldots, c_k, c_{k+1}, \ldots) \xrightarrow{r(N)kc_k} (c_1, \ldots, c_k - 1, c_{k+1} + 1, \ldots) & \text{Birth} \\ \text{for } k \geq 2 & (c_1, \ldots, c_{k-1}, c_k, \ldots) \xrightarrow{\mu k c_k} (c_1, \ldots, c_{k-1} + 1, c_k - 1, \ldots) & \multirow{2}{*}{\Big\}\text{ Death}} \\ & (c_1, c_2, \ldots) \xrightarrow{\mu c_1} (c_1 - 1, c_2, \ldots) & \end{array}$$

If we assume detailed balance between pairs of states with maximum clone size K, we can recurse the relations

$$\mu c_k k P(c_1, \ldots, c_{k-1}, c_k, \ldots) = r(N)(k-1)\,(c_{k-1}+1)\,P(c_1, \ldots, c_{k-1}+1, c_k - 1, \ldots)$$

for $2 \leq k \leq K$ down to the states

$$\mu c_1 P(c_1, \vec{0}) = \alpha P(c_1 - 1, \vec{0})$$

to give

$$P(\vec{c}) = \frac{1}{Z_{\alpha,\mu}} \frac{\alpha^C}{\mu^N} \frac{\prod_{n=1}^{N-C} r(N-n)}{\prod_{i=1}^{\infty} i^{c_i} c_i!}. \tag{50}$$

Using these chosen pairs of states to impose detailed balance, we find a unique distribution $P(\vec{c})$. However, this form of $P(\vec{c})$ will not obey detailed balance between all pairs of states. For example, balancing the transitions

$$(c_1, c_2, \ldots) \underset{\alpha}{\overset{\mu c_1}{\rightleftharpoons}} (c_1 - 1, c_2, \ldots)$$

would also require

$$\mu c_1 P(c_1, c_2 \geq 1, \ldots) = \alpha P(c_1 - 1, c_2 \geq 1, \ldots).$$

However, using the $P(\vec{c})$ from Eq. (50), we find

$$\frac{\mu c_1 P(c_1, c_2 \geq 1, \ldots)}{\alpha P(c_1 - 1, c_2 \geq 1, \ldots)} = \frac{r(C-1)}{r(N-1)} \neq 1$$

because generally, $N \neq C$. Remarkably, the analogous exercise for the BDICC model where $\mu = \mu(N)$ does satisfy detailed balance between all pairs of states and the $P(\vec{c})$ we derived for the BDICC model, Eq. (29), is exact.

References

1. Allen, L.J.S.: An Introduction to Stochastic Processes with Applications to Biology. Taylor and Francis, Boca Raton (2010)
2. Bansaye, V., Méléard, S.: Birth and death processes. In: Stochastic Models for Structured Populations, Mathematical Biosciences Institute Lecture Series, pp. 7–17. Springer, Cham (2015)
3. Baxter, G.J., Blythe, R.A., McKane, A.J.: Exact solution of the multi-allelic diffusion model. Math. Biosci. **209**, 124–170 (2007)
4. Bell, G.: Neutral macroecology. Science **293**(5539), 2413–2418 (2001)
5. Billingsley, P.: Probability and Measure, 4th edn. Wiley, Hoboken (2012)
6. Bulmer, M.G.: On fitting the Poisson lognormal distribution to species-abundance data. Biometrics **30**(1), 101–110 (1974)
7. Chiu, C.-H., Wang, Y.-T., Walther, B.A., Chao, A.: An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. Biometrics **70**, 671–682 (2014)
8. Chou, T., D'Orsogna, M.R.: Coarsening and accelerated equilibration in mass-conserving heterogeneous nucleation. Phys. Rev. E **84**, 011608 (2011)
9. Colwell, R.K., Coddington, J.A.: Estimating terrestrial biodiversity through extrapolation. Philos. Trans. R. Soc. B **345**, 101–118 (1994)
10. Desponds, J., Mora, T., Walczak, A.M.: Fluctuating fitness shapes the clone-size distribution of immune repertoires. Proc. Natl. Acad. Sci. USA **113**, 274–279 (2016)
11. Dessalles, R., Fromion, V., Robert, P.: A stochastic analysis of autoregulation of gene expression. J. Math. Biol. **75**, 1–31 (2017)
12. D'Orsogna, M.R., Lakatos, G., Chou, T.: Stochastic self-assembly of incommensurate clusters. J. Chem. Phys. **136**, 084110 (2012)
13. D'Orsogna, M.R., Zhao, B., Berenji, B., Chou, T.: Combinatoric analysis of heterogeneous stochastic self-assembly. J. Chem. Phys. **137**, 121918 (2013)
14. Fisher, R.A., Corbet, A.S., Williams, C.B.: The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. **12**, 42–58 (1943)
15. Gibbs, J.P., Martin, W.T.: Urbanization, technology, and the division of labor: international patterns. Am. Sociol. Rev. **27**, 667–677 (1962)
16. Goyal, S., Kim, S., Chen, I.S.Y., Chou, T.: Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques. BMC Biol. **13**(1), 85 (2015)
17. Grimmett, G., Stirzaker, D.: Probability and Random Processes. Oxford University Press, Oxford (2001)
18. Hubbell, S.: The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32) (Monographs in Population Biology). Princeton University Press, Princeton (2001)
19. Hurlbert, S.H.: The nonconcept of species diversity: A critique and alternative parameters. Ecology **52**, 577–586 (1971)
20. Jost, L.: Entropy and diversity. Oikos **113**, 363–375 (2006)
21. Karev, G.P., Wolf, Y.I., Rzhetsky, A.Y., Berezovskaya, F.S., Koonin, Eugene V.: Birth and death of protein domains: a simple model of evolution explains power law behavior. BMC Evolut. Biol. **2**, 18 (2002)
22. Karlin, S., McGregor, J.: The number of mutant forms maintained in a population. In: Proceedings of the Fifth Berkeley Symposium on Mathematics, Statistics and Probability, vol. 4, pp. 415–438 (1967)
23. Lambert, A.: Species abundance distributions in neutral models with immigration or mutation and general lifetimes. J. Math. Biol. **63**, 57–72 (2011)
24. Laydon, D.J., Bangham, C.R.M., Asquith, B.: Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach. Philos. Trans. R. Soc. B **370**, 20140291 (2015)
25. Lythe, G., Callard, R.E., Hoare, R.L., Molina-París, C.: How many TCR clonotypes does a body maintain? J. Theor. Biol. **389**, 214–224 (2016)

26. MacArthur, R.H., Wilson, E.O.: The Theory of Island Biogeography. Princeton University Press, Princeton (2016)
27. Miles, J.J., Douek, D.C., Price, D.A.: Bias in the $\alpha\beta$ T-cell repertoire: implications for disease pathogenesis and vaccination. Immunol. Cell Biol. **89**, 375–387 (2011)
28. Morris, E.K., Caruso, T., Buscot, F., Fischer, M., Hancock, Christine, Maier, Tanja S, Meiners, Torsten, Müller, Caroline, Obermaier, Elisabeth, Prati, Daniel, Socher, Stephanie A, Sonnemann, Ilja, Wäschke, Nicole, Wubet, Tesfaye, Wurst, Susanne, Rillig, Matthias C: Choosing and using diversity indices: insights for ecological applications from the German Biodiversity Exploratories. Ecol. Evol. **4**(18), 3514–3524 (2014)
29. Palmer, M.W.: The estimation of species richness by extrapolation. Ecology **71**, 1195–1198 (2003)
30. Preston, F.W.: The commonness, and rarity, of species. Ecology **29**(3), 254–283 (1948)
31. Sala, C., Vitali, S., Giampieri, E., do Valle, I.F., Remondini, D., Garagnani, P., Bersanelli, M., Mosca, E., Milanesi, L., Castellani, G.: Stochastic neutral modelling of the Gut Microbiota's relative species abundance from next generation sequencing data. BMC Bioinform. **17**, S16 (2016)
32. Tan, J.T., Dudl, E., LeRoy, E., Murray, R., Sprent, Jonathan, Weinberg, Kenneth I., Surh, Charles D.: IL-7 is critical for homeostatic proliferation and survival of naïve T cells. Proc. Natl. Acad. Sci. USA **98**(15), 8732–8737 (2001)
33. Travaré, S.: The genealogy of the birth, death, and immigration process. In: Feldman, M.W. (ed.) Mathematical Evolution Theory, pp. 41–56. Princeton University Press, Princeton (1989). ISBN 0-691-08502-1
34. Volkov, I., Banavar, J.R., Hubbell, S.P., Maritan, A.: Neutral theory and relative species abundance in ecology. Nature **424**(6952), 1035–1037 (2003)