IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **51** (2018) 425602 (27pp) https://doi.org/10.1088/1751-8121/aadcb4

Immigration-induced phase transition in a regulated multispecies birth-death process

Song Xu[1] **and Tom Chou**[1,2]

[1] Department of Biomathematics, UCLA, Los Angeles, CA 90095-1766, United States of America
[2] Department of Mathematics, UCLA, Los Angeles, CA 90095-1555, United States of America

E-mail: tomchou@ucla.edu

Received 20 June 2018, revised 21 August 2018
Accepted for publication 24 August 2018
Published 14 September 2018

Abstract

Power-law-distributed species counts or clone counts arise in many biological settings such as multispecies cell populations, population genetics, and ecology. This empirical observation that the number of species c_k represented by k individuals scales as negative powers of k is also supported by a series of theoretical birth–death–immigration (BDI) models that consistently predict many low-population species, a few intermediate-population species, and very high-population species. However, we show how a simple global population-dependent regulation in a neutral BDI model destroys the power law distributions. Simulation of the regulated BDI model shows a high probability of observing a high-population species that dominates the total population. Further analysis reveals that the origin of this breakdown is associated with the failure of a mean-field approximation for the expected species abundance distribution. We find an accurate estimate for the expected distribution $\langle c_k \rangle$ by mapping the problem to a lower-dimensional Moran process, allowing us to also straightforwardly calculate the covariances $\langle c_k c_\ell \rangle$. Finally, we exploit the concepts associated with energy landscapes to explain the failure of the mean-field assumption by identifying a phase transition in the quasi-steady-state species counts triggered by a decreasing immigration rate.

Keywords: species abundance, Moran model, stochastic birth–death process, immigration

(Some figures may appear in colour only in the online journal)

1. Introduction

High-dimensional stochastic models are important across many fields of science and often arise in biological contexts such as T cell receptor (TCR) diversity in immunology [1], species

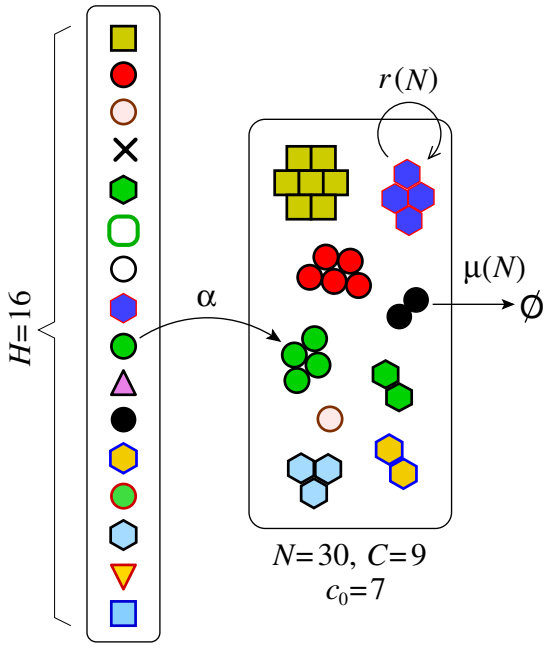

Figure 1. A simple H-species birth–death–immigration process in which an external fixed 'source' always contains H individuals; each of a different species. This source may represent uniquely tagged stem cells; a 'mainland' from which species emigrate, or the thymus that outputs naive T cell clone; each expressing a different TCR. Each cell in the source buds off a daughter cell with rate α a daughter cell into the system but remains intact. All individuals in the system can proliferate with rate $r(N)$ and die with rate $\mu(N)$, where N is the total population in the system. A specific configuration with $H = 16$ and $N = 30$ is depicted. Here, $C = 9$ represents the *number of different species* that exist in the system. c_0 represents the number of species in the source that are not represented in the system.

abundance and diversity in ecology [2], and populations in cellular barcoding experiments [3]. T cells in jawed vertebrates can be classified into multiple subpopulations, each corresponding to different TCR subtypes produced in the thymus. Here the number of T cells n_i expressing the ith receptor represents the ith dimension. In this setting the large number of different TCRs ($1 \leqslant i \leqslant \Omega$, $\Omega \sim 10^6$–10^8) present in an organism allows its adaptive immune system to recognize and respond to a wide range of antigens that it might encounter. Multispecies ecological communities are another example of high-dimensional systems. If the habitat of interest is an island, then n_i quantifies the number of animals of species i on the island. The gut is also a habitat for many coexisting species of bacteria that make up the microbiome [4]. Finally, DNA-tagging and sequencing technology has allowed *in vivo* tracking of multiple hematopoietic clones, each of which was derived from a unique hematopoietic stem cell that carries a unique DNA tag [1, 5–8], resulting in clonal-tracking data of very high dimensions [3, 9].

The simplest single-compartment mathematical structure that is common to all the multispecies systems mentioned above is the birth–death–immigration (BDI) processes shown in figure 1. The source of immigration into the system is a fixed 'source' population of H different individuals, each of a different species. In the T cell setting, the possible number of different receptors that can be produced by the thymus is $H > 10^{15}$ [10] while in typical

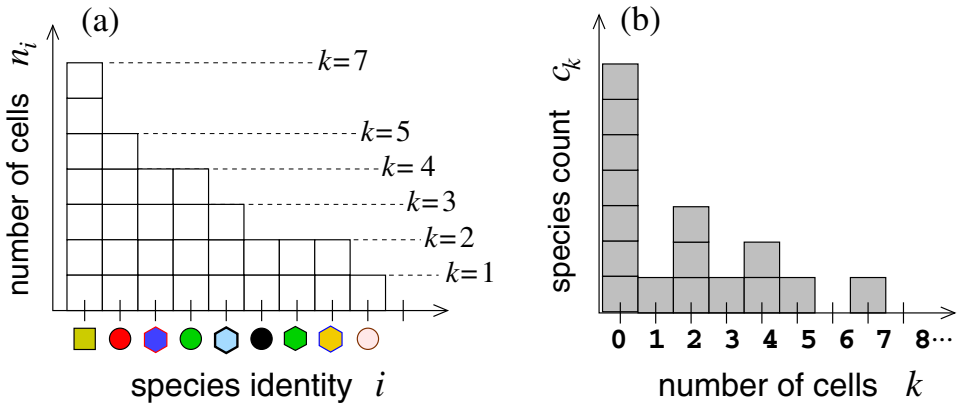

Figure 2. Definition of species counts corresponding to the configuration in figure 1. (a) In the cell-count representation, n_i is the number of cells of species i detected in a sample. (b) c_k is the number of different species that are represented by exactly k cells in a sample. A given set $\{n_i\}$ uniquely determines the corresponding $c_k \equiv \sum_{i=1}^{\infty} \mathbb{1}(n_i, k)$. However, one cannot recover n_i from c_k since species identity information is lost when transforming from n_i to c_k.

barcoding experiments $H \approx 10^3$–10^6 different tags can be implanted [5]. After immigration into the system, the individuals can proliferate with rate $r(N)$ and die with rate $\mu(N)$, both possibly functions of the total population N. In the configuration shown in figure 1, the maximum number of different species is $H = 16$ and the number of individuals of each species is $n_1 = 7, n_2 = 5, n_3 = n_4 = 4, n_5 = 3, n_6 = n_7 = n_8 = 2, n_9 = 1$. We have labeled the species i according to decreasing population. In this work, the terms 'clones' and 'species' are interchangeable and 'clones' will only be used to refer to different tags in barcoding experiments and different TCR types.

Such high-dimensional stochastic systems are generally difficult to study because of the 'curse of dimensionality'. The evolution of the full probability distribution $P(\mathbf{n}) \equiv P(\{n_1, n_2, ..., n_H\})$ is unintuitive and computationally intractable [11]. It also contains more information than necessary if we consider only neutral species and their identities are not relevant. Describing the system in terms of moments such as $\langle n_i \rangle$ and $\langle n_i n_j \rangle$ reduces the model complexity and allows one to track specific species [3], but does not directly capture the species size distribution resulting from the relevant stochastic processes. Another approach is to use single-quantity metrics such as species richness, Simpson's diversity, Shannon's diversity, or the Gini index to describe and compare various ecological communities. Such diversity measures can be overly simplistic and lead to different conclusions depending on the diversity index used. Thus, a description of intermediate complexity is desired.

In ecology, a commonly used measure is the species abundance distribution (SAD) that counts the number of different species encountered in a community [2]. In the language of clonal dynamics, it is the count of the number of species or 'clones' that are each represented by k individuals as depicted in figure 2 [3]:

$$c_k = \sum_{i=1}^{H} \mathbb{1}(n_i, k). \tag{1}$$

Here $\mathbb{1}(x, y)$ is the identity function which takes on the value 1 when $x = y$ or 0 otherwise. The species-count c_k represents a one-dimensional vector of numbers indexed by $k = 0, 1, ...$ and

gives a more comprehensive picture of how the clone/species are distributed compared to that of a single index. By construction, the species count c_k also obeys the constraints

$$c_0 + \sum_{k=1}^{\infty} c_k = H, \qquad \sum_{k=0}^{\infty} k c_k = N. \tag{2}$$

Species counts are useful in describing numbers of rare or abundant species especially when their identities are not important. Examples of such systems include genetically barcoded, virally tagged, or TCR-decorated [1, 3, 12–14] cellular clones, microbial populations [15, 16], and ecological species [2, 17, 18].

A universally observed feature in empirical studies across all these fields is a 'hollow curve distribution' for c_k, where few highly populous species and many low-population species arise [2]. Theoretical studies have attempted to explain these observations by proposing various physical models, including neutral models with constant immigration, birth, and death rates [11, 19–21]; time-dependent birth and death [22]; cell-wise and species-wise heterogeneities [14]; and intra-species carrying capacities [23].

Multi-clonal/multi-species models with fluctuating total population size commonly arise in the evolutionary biology and physics communities [24–28]. Most of the attention has been on computing expected values of population counts rather than species abundance distributions. For example, Parsons *et al* [24] used a neutral and quasi-neutral birth–death model with carrying capacity and studied the mean fixation time of any species.

In ecology and immunology, the literature on the distribution $\langle c_k \rangle$ is rich. For example, Volkov *et al* [23] considered an *intra-species* carrying capacity that balances the birth and death rates of each species. However, they did not consider a global carrying capacity that regulates populations across all species. In this case, no interactions arise among the species, and the mean-field results are accurate. Other theories considered competition for resources (such as T cell proliferation competing for stimuli from self-peptides [14, 29, 30]), but the resources were modeled as evolving variables of the system, and explicit solutions could be found only in very simple cases. None of these previous studies have treated global interactions that correlate populations across all clones/species, an important ingredient in studies of interacting populations.

Stochastic simulations of a neutral BDI model that includes a simple global carrying capacity exhibit distributions of $\langle c_k \rangle$ that differ qualitatively from those of the above-mentioned studies. Under a high immigration rate, the classical power law distribution of $\langle c_k \rangle$ (with an exponential cutoff) remains an accurate representation of our simulated results. However, as the immigration rate is decreased, a single large-population species emerges. Such single-species dominance is not captured by classical mean-field theories. Through further analysis that more accurately includes the interactions between species populations, we find that a low immigration rate induces a phase transition to bistability in species populations. The mean-field approximation, which was explicitly or implicitly assumed in previous studies, breaks down near the new local stable state where the ensemble average of c_k follows a different distribution.

In this paper, we introduce the simple idea of transforming the problem of calculating the qth moment of c_k to the problem of solving a $(q+1)$-dimensional process described by the population vector $\{n_1, , n_2, ..., n_q, N'\}$. This process can be further approximated by a q-dimensional Moran model that imposes a fixed total population. We use this approach to accurately calculate the 1st and 2nd moments of c_k under general functional forms of the carrying capacity. We then exploit ideas from energy landscapes to identify the key parameters controlling a phase transition of the general multi-species BDI process that explains the failure of previously used mean-field assumptions.

2. Classical formulation and mean-field assumption

Here, we develop the stochastic dynamics of the BDI process depicted in figure 1. In the language of clonally tracked stem cell differentiation, the probability of a stem cell carrying any specific tag asymmetrically differentiating to produce a progenitor cell within infinitesimal time $\mathrm{d}t$ is $\alpha \mathrm{d}t$. We will assume that there is a fixed number H of stem cells or 'source' individuals. The probability for any progenitor cell to divide into two new identical progenitor cells (birth) within $\mathrm{d}t$ is $r\mathrm{d}t$, and the probability of its dying in $\mathrm{d}t$ is $\mu \mathrm{d}t$.

We will further assume the particle dynamics are coupled in a species-independent way, leading to identical (but not independent) statistics of the populations of each species. The canonical implementation of such a 'global' neutral interaction is through a birth rate $r(N)$ and/or death rate $\mu(N)$ which depend only on the total population $N \equiv \sum_{i=1}^{H} n_i = \sum_{\ell} \ell c_\ell$. Thus, the total population N can be 'decoupled' and completely described by its own master equation,

$$\begin{aligned}\frac{\partial P(N,t)}{\partial t} &= \alpha H[P(N-1) - P(N)] \\ &\quad + r(N-1)(N-1)P(N-1) - r(N)NP(N) \\ &\quad + \mu(N+1)(N+1)P(N+1) - \mu(N)NP(N),\end{aligned} \tag{3}$$

from which moments of N can be computed. The higher-dimensional master equation obeyed by the full multispecies distribution $P(\{n_j\}; t)$ is explicitly given in appendix A.

Let us denote the ensemble (not time) average of a quantity by $\langle\cdot\rangle$. Thus, $\langle n_i(t)\rangle \equiv \sum_{\{n_j\}} n_i P(\{n_j\}; t)$ represents the expected population of the ith species. By using equations (3) and (A.1), we can show that the expected subpopulation $\langle n_i(t)\rangle$ and total population $\langle N(t)\rangle$ for the BDI process obeys

$$\begin{aligned}\frac{\mathrm{d}\langle n_i\rangle}{\mathrm{d}t} &= \alpha + \langle (r(N) - \mu(N)) n_i\rangle, \\ \frac{\mathrm{d}\langle N\rangle}{\mathrm{d}t} &= \alpha H + \langle (r(N) - \mu(N)) N\rangle.\end{aligned} \tag{4}$$

In appendix B, we also explicitly derive the equation for $\langle c_k(t)\rangle$,

$$\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} = \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + \langle r(N)\,[(k-1)c_{k-1} - kc_k]\rangle + \langle \mu(N)\,[(k+1)c_{k+1} - kc_k]\rangle, \tag{5}$$

from the master equation for $P(c_0, c_1, c_2, \ldots; t)$. This evolution equation indicates that immigration (at rate α) of an individual from a species with population $n_i = k$ increases its size by 1, thereby decreasing c_k by 1 but increasing the number of species with population $k+1$, c_{k+1}, by 1. Cellular birth and death have similar effects, but their corresponding rates are proportional to the species population k (the number of individuals/cells in the species). In the rest of this paper we will be interested in evaluating the steady-state values of $\langle c_k\rangle$.

2.1. Constant rates

In the simplest scenario of constant birth and death rates, one can write [3]

$$\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} = \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + r[(k-1)\langle c_{k-1}\rangle - k\langle c_k\rangle] + \mu[(k+1)\langle c_{k+1}\rangle - k\langle c_k\rangle]. \tag{6}$$

If $r < \mu$, a stable steady state can be found:

$$\langle c^*_{k\geqslant 1}\rangle = \frac{\alpha H}{rk!}\frac{(\frac{r}{\mu})^k(1-\frac{r}{\mu})^{\alpha/r}}{\frac{\alpha}{r}+k}\prod_{\ell=1}^{k}\left(\frac{\alpha}{r}+\ell\right), \quad \langle c^*_0\rangle = H - \sum_{k=1}^{\infty}\langle c^*_k\rangle = H\left(1-\frac{r}{\mu}\right)^{\alpha/r}. \tag{7}$$

In the $\alpha/r \to 0^+$ limit, the species counts monotonically decay as

$$\langle c^*_{k\geqslant 1}\rangle \approx H\left(\frac{\alpha}{r}\right)\left(1-\frac{r}{\mu}\right)^{\alpha/r}\left(\frac{r}{\mu}\right)^k\frac{1}{k}. \tag{8}$$

2.2. Carrying capacity and mean-field approximation

Now, assume that $r(N)$ decreases with N and/or $\mu(N)$ increases with N, and that $\lim_{N\to\infty} r(N)/\mu(N) < 1$. These conditions on $r(N)$ and $\mu(N)$ guarantee that n_i and c_k are bounded even if $r(N) > \mu(N)$ for some finite N. Terms of the form $\langle r(N)c_k\rangle$ in equation (5) cannot be approximated by factoring because $r(N)$ depends on c_k through the stochastic variable $N \equiv \sum_\ell \ell c_\ell$ defined in equation (2). Nonetheless, to make headway, a mean-field method is often invoked to simplify equations (4) and (5). Upon fully factorizing interaction terms such as $\langle r(N)c_k\rangle \approx r(\langle N\rangle)\langle c_k\rangle$ and $\langle r(N)N\rangle \approx r(\langle N\rangle)\langle N\rangle$, we can approximate equations (4) and (5) as

$$\frac{\mathrm{d}\langle N\rangle}{\mathrm{d}t} \approx \alpha H + \langle (r(\langle N\rangle) - \mu(\langle N\rangle))\langle N\rangle \equiv f(\langle N\rangle), \tag{9}$$

$$\begin{aligned}\frac{\mathrm{d}\langle c_k\rangle}{\mathrm{d}t} \approx{}& \alpha(\langle c_{k-1}\rangle - \langle c_k\rangle) + r(\langle N\rangle)\left[(k-1)\langle c_{k-1}\rangle - k\langle c_k\rangle\right] \\ &+ \mu(\langle N\rangle)\left[(k+1)\langle c_{k+1}\rangle - k\langle c_k\rangle\right].\end{aligned} \tag{10}$$

By first solving equation (9), we can input $\langle N(t)\rangle$ into equation (10) and explicitly solve for $\langle c_k(t)\rangle$. The steady-state solution to $\langle N\rangle$, $\langle N^*\rangle$, is defined in equation (9) by $f(\langle N^*\rangle) = 0$ and the requirement that $\langle N^*\rangle > 0$ requires $[\mathrm{d}f(\langle N\rangle)/\mathrm{d}\langle N\rangle]_{\langle N^*\rangle} \equiv f'(\langle N^*\rangle) \equiv r'(\langle N^*\rangle) - \mu'(\langle N^*\rangle) < 0$. The steady-state values of $\langle c_k\rangle$ can be reached only after the steady state of $\langle N\rangle$ is reached and $r(\langle N\rangle)$ and $\mu(\langle N\rangle)$ approach constant values.

We show in appendix C that this deterministic description breaks down after an exponentially long time when the immigration rate α is sufficiently small. The reason is that for $\alpha = 0$, $N = 0$ becomes an absorbing boundary in the full stochastic model. Thus, when $\alpha = 0$, the $\langle N^*\rangle$ we find from $f(\langle N^*\rangle) = 0$ is actually a quasi-steady state (QSS) even though equation (9) indicates a stable deterministic equilibrium $\langle N^*\rangle$ for physically reasonable functions $r(\langle N\rangle)$ and $\mu(\langle N\rangle)$.

Focusing on evaluating the QSS value of $\langle c_k\rangle$, $\langle c^*_k\rangle$, before the final extinction that occurs over exponentially long times, we denote $r(\langle N^*\rangle) \equiv r^*$ and $\mu(\langle N^*\rangle) \equiv \mu^*$ as the rates of birth and death at QSS. The QSS solution $\langle c^*_k\rangle$ can be written in the same form as equation (7),

$$\langle c^*_{k\geqslant 1}\rangle = \frac{\alpha H}{r^*k!}\frac{(\frac{r^*}{\mu^*})^k(1-\frac{r^*}{\mu^*})^{\alpha/r^*}}{\frac{\alpha}{r^*}+k}\prod_{\ell=1}^{k}\left(\frac{\alpha}{r^*}+\ell\right), \quad \langle c^*_0\rangle = H - \sum_{k=1}^{\infty}\langle c^*_k\rangle = H\left(1-\frac{r^*}{\mu^*}\right)^{\alpha/r^*}. \tag{11}$$

Here, $\langle c^*_k\rangle$ corresponds to the mean QSS species-count under the mean-field approximation which we expect to be different from the exact solution. In the $\alpha/\mu^*, \alpha/r^* \to 0^+$ limit, the expected species count $\langle c^*_{k\geqslant 1}\rangle$, as in equation (8), is monotonic in k:

$$\langle c^*_{k\geqslant 1}\rangle \approx H\left(\frac{\alpha}{r^*}\right)\left(1-\frac{r^*}{\mu^*}\right)^{\alpha/r^*}\left(\frac{r^*}{\mu^*}\right)^k\frac{1}{k}. \tag{12}$$

However, under regulation, $r^*/\mu^* \approx 1 - \mathcal{O}(\alpha/\mu^*)$, resulting in a long-tail k-dependence of $\langle c_k^* \rangle$. Although the amplitude of $\langle c_{k\geqslant 1}^* \rangle$ is proportional to α/r^*, it is constructed to obey the mean total population constraint $\langle N^* \rangle = \sum_{k=1}^{\infty} k \langle c_k^* \rangle$ which is reflected in the long-tail property of the mean-field approximation to $\langle c_k^* \rangle$.

2.3. Failure of the mean-field approximation to $\langle c_k^* \rangle$ in the slow immigration regime

To concretely investigate the errors incurred under a mean-field assumption, we first focus explicitly on a logistic growth law for the total population defined by

$$r(N) = p\left(1 - \frac{N}{K}\right), \qquad \mu(N) = \mu, \tag{13}$$

where p is the maximal birth rate and K is the carrying capacity parameter. The mean-field solution for the total population is

$$\begin{aligned} \langle N^* \rangle &= \frac{K}{2}\left(1 - \frac{\mu}{p}\right)\left[1 + \sqrt{1 + \frac{4\alpha H p}{(p-\mu)^2 K}}\right] \\ &= K\left(1 - \frac{\mu}{p}\right) + \frac{\alpha H p}{p - \mu} + \mathcal{O}(1/K). \end{aligned} \tag{14}$$

In many examples, such as progenitor cells, $K \gg 1$ and $\langle N^* \rangle \sim K$ is large except when μ approaches or exceeds p.

We are now in a position to use $\langle N^* \rangle$ to determine r^* and evaluate the mean-field approximation for $\langle c_k^* \rangle$ (equation (11)). In figure 3 we compare numerically evaluated mean-field solutions of $\langle c_k^* \rangle$ with Monte-Carlo simulations of the underlying BDI process for various values of α.

For small α, such as 10^{-8} used to generate figure 3(a), equation (11) fails to capture the peak arising in $\langle c_k^* \rangle$ at $k \approx \langle N^* \rangle$. In the singular limit $\alpha \to 0$, the mean-field solution $\langle c_{k\geqslant 1}^* \rangle \to 0$ and $\langle c_0^* \rangle \to H$ but nonetheless, by construction, satisfies $\sum_{k=1}^{\infty} k \langle c_k^* \rangle \to \langle N^* \rangle$. However, in the simulated $\langle c_k^* \rangle$, the small peak at large size $k \approx \langle N^* \rangle$ signals that a single species has come to dominate the total population. The number of species not in the system is thus $\langle c_0^* \rangle \approx H - 1$. One species, typically the first to have immigrated, has taken over the system squeezing out all others that try to immigrate when the immigration rate α is small. This peak in $\langle c_k^* \rangle$ near $k \approx \langle N^* \rangle$ is completely missed by the mean-field approximation. The mean-field approximation also inaccurately captures the rapid decay in $\langle c_k^* \rangle$ for $k > \langle N^* \rangle$ due to exhaustion of the population in the single size-k species.

When α is still relatively small as in figure 3(b), the simulated $\langle c_k^* \rangle$ is dominated by many low-population species, with a slow decay with size k followed by a faster decay at large k, again due to mass depletion. At this modest immigration rate, high-population species do not have the opportunity to establish and more intermediate-sized species arise at the expense of very high-population species but the simulated result $\langle c_k^* \rangle$, remains monotonic. Nevertheless, the mean-field approximation of equation (11) still fails to capture the fast decay of $\langle c_k^* \rangle$ for large k.

For even larger α, the preferred total population increases. Since the total number of species remains capped at H, the mean number of individuals/cells per species increases. The distribution $P(n_1)$ peaks at higher values of n_1 thereby forming a peak in $\langle c_k^* \rangle$ at size $k \ll \langle N^* \rangle$. The larger-$\alpha$ cases shown in figures 3(c) and (d) are accurately described by the mean-field approximation of $\langle c_k^* \rangle$ for all values of k.

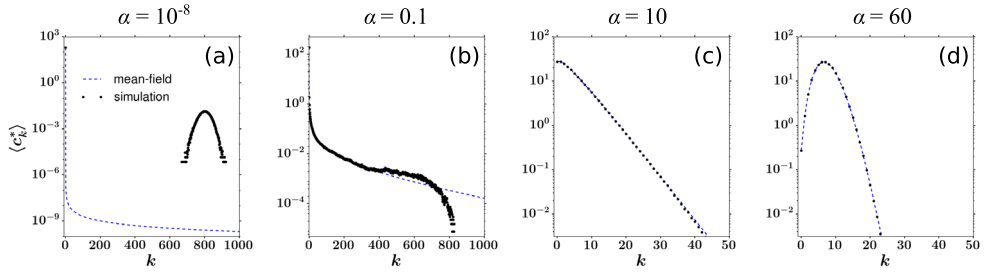

Figure 3. Comparison of steady-state species-count distributions from simulations (black dots) to those computed from the mean-field approximation (dashed blue curves) using the logistic growth model of equation (13) and (a) $\alpha = 10^{-8}$, (b) $\alpha = 0.1$, (c) $\alpha = 10$, and (d) $\alpha = 60$. Other parameters used are $\mu = 10$, $p = 20$, $K = 1600$, $H = 200$. The resulting N^* are 800, 840, 4800, and 24 800, respectively. The mean-field approximation $\langle c_k^* \rangle$ breaks down for small α completely missing the peak at $k \approx \langle N^* \rangle$ in (a). Also, note the log scale and the absence of simulations that capture the rare configurations in (a) where $k \neq \langle N^* \rangle$.

3. Proposed model for $\langle \mathbf{c}^{\boldsymbol{q}} \rangle$

The challenge in solving equation (5) lies in the nonseparable terms $\langle r(N)c_k \rangle$. Even in the simple case of logistic growth where $r(N)$ is linear, the $\langle r(N)c_k \rangle$ terms include second-moments $\langle c_k c_\ell \rangle$, which usually cannot be approximated by $\langle c_k \rangle \langle c_\ell \rangle$. If one attempts to solve equation (5) for the time-dependent or steady-state solution $\langle c_k^* \rangle$, one encounters the so-called 'moment closure' problem, where the solution of the 1st moment $\langle c_k \rangle$ depends on 2nd moments $\langle c_k c_\ell \rangle$, which in turn depends on 3rd moments, and so on [31]. There is usually no closed-form solution or easy approximation to such problems. In the rest of this section, we develop an alternative approach.

3.1. Transformation of the problem

A complete description would be an H-dimensional model for the distribution $P(\{n_1, n_2, \ldots, n_H\}; t)$. However, by using the definition of c_k in equation (1), assuming the initial populations of all species are identical $n_1(0) = n_2(0) = \ldots = n_H(0)$ and assuming indistinguishability among species, one can easily show that (appendix D)

$$\langle c_k(t) \rangle = HP(n_1 = k; t), \tag{15}$$

where $P(n_1; t) = \sum_{n_2=0}^{\infty} \cdots \sum_{n_H=0}^{\infty} P(\{n_1, n_2, \ldots, n_H\}; t)$ is the single-dimensional marginal distribution. Here, the singling out of species 1 is arbitrary. The assumption of identical initial species populations is not needed in the long time QSS limit as long as different initial distributions $c_k(0)$ converge to a unique $\langle c_k^* \rangle$. Thus, at QSS, $\langle c_k^* \rangle = HP(n_1 = k; t \to \infty)$ always holds. Intuitively, the expected fraction of all species that have size k is the probability that any one species is of size k.

We can write the master equation for the BDI process of a single species with population n_1 as

$$\begin{aligned} \frac{\partial P(n_1; t)}{\partial t} &= \alpha[P(n_1 - 1) - P(n_1)] + r(N)\left[(n_1 - 1)P(n_1 - 1) - n_1 P(n_1)\right] \\ &\quad + \mu(N)\left[(n_1 + 1)P(n_1 + 1) - n_1 P(n_1)\right] \end{aligned} \tag{16}$$

where $N(t)$ represents one trajectory of the random process $N = n_1 + n_2 + \ldots + n_H$, which we might approximate using the deterministic solution to equation (4). Equation (16) has the exact same form as the right-hand side of equation (10) for $\langle c_k \rangle$. However, in the presence of other species or clones, it is immediately clear that equation (16) is not a complete description for n_1 since the variable N depends on the population of all species. Species 'independence' breaks down through the $r(N)$ and $\mu(N)$ terms. All species compete with each other for the limited sources in the environment through their shared and regulated birth and death rates.

Equation (15) remains exact (appendix D) since the population dynamics are neutral and all species start with the same initial size. One still needs to solve any *individual* species' marginal probability distribution $P(n_1)$ given that all species, including itself, can affect it. Formally, this corresponds to first solving the full distribution $P(\{n_1,\ n_2, ...,\ n_H\})$ before summing over all other populations $\{n_2, ...,\ n_H\}$. Since we are not concerned about the detailed configurations of $\{n_2, ...,\ n_H\}$, but rather their combined effects on n_1, we can lump species 2 through H into an effective 'bath' species whose size is $N' = n_2 + ... + n_H$. This effective species has a birth rate $N'r(n_1 + N')$, a death rate $N'\mu(n_1 + N')$, and an immigration rate $(H-1)\alpha$. Equation (16) is now coupled to the master equation

$$\begin{aligned}\frac{\partial P(N';t)}{\partial t} = {} & \alpha(H-1)[P(N'-1) - P(N')] + r(N)\left[(N'-1)P(N'-1) - N'P(N')\right] \\ & + \mu(N)\left[(N'+1)P(N'+1) - N'P(N')\right].\end{aligned} \tag{17}$$

One usually combines equations (16) and (17) together into a 2D master equation

$$\begin{aligned}\frac{\partial P(n_1,N';t)}{\partial t} = {} & \alpha[P(n_1-1,N') - P(n_1,N')] \\ & + \alpha(H-1)[P(n_1,N'-1) - P(n_1,N')] \\ & + r(N-1)\left[(n_1-1)P(n_1-1,N') + (N'-1)P(n_1,N'-1)\right] \\ & + \mu(N+1)\left[(n_1+1)P(n_1+1,N') + (N'+1)P(n_1,N'+1)\right] \\ & - \left[r(N) + \mu(N)\right]\left[n_1P(n_1,N') + N'P(n_1,N')\right].\end{aligned} \tag{18}$$

The 2D problem can be approximated by a 1D problem when $n_1 \ll N'$ and $r(N) \approx r(N')$. The birth rate is approximately regulated by the 'bath' population N' which leads to a decoupling from n_1. Similarly, when $n_1 \gg N'$, $r(N) \approx r(n_1)$ and the birth rate is approximately independent of N'. In either limit, the problem is approximately one-dimensional and can be modeled using a 1D master equation for $P(n_1)$ (equation (16)) or $P(N')$ (equation (3)) correspondingly. However, when n_1 and N' are comparable in size, one needs to evaluate the full 2D distribution $P(n_1,N')$ and marginalise over N' to obtain $P(n_1) = \sum_{N'=0}^{\infty} P(n_1,N')$ and

$$\langle c_k(t) \rangle = HP(n_1 = k;t) = H\sum_{N'=0}^{\infty} P(n_1 = k, N';t). \tag{19}$$

This approach can be extended to higher dimensions to determine higher moments of $c_k(t)$, which are important for characterizing the variability of species size distributions. Covariances $\mathrm{cov}(c_k, c_\ell) \equiv \langle c_k c_\ell \rangle - \langle c_k \rangle \langle c_\ell \rangle$, in particular, will reveal the differences between the solutions to the mean-field model (equation (11)) and the exact model (equation (5)). In appendix D, we derive relationships between higher moments of c_k and the cell count distributions $P(n_1, n_2, ...)$. Specifically, for the second moments,

$$\langle c_k(t)c_\ell(t) \rangle = H(H-1)P(n_1 = k, n_2 = \ell;t) + \mathbb{1}(k,\ell)HP(n_1 = k;t). \tag{20}$$

3.2. Approximating $P(\{n_1, n_2, ..., n_q, N'\})$ by a q-dimensional Moran model

We now try to find a solution to $P(n_1, N')$. Since the 2D master equation does not usually have analytic solutions, we will show how to approximate $P(n_1, N')$ by a 1D two-species Moran model [24–28] with n_1 individuals of species 1 and N' individuals of species 2 (which, for this case, is the sum of the populations of species 2 through H in the original multispecies model). The 1D Moran model imposes $n_1 + N' \equiv N$, the total population size, to be a fixed value.

We first fix the value of N to be the quasi-steady-state value of the original unconstrained BDI process $N \to N^* := \langle N^* \rangle$, at which the condition $\alpha H + r(N^*)N^* = \mu(N^*)N^*$ is satisfied. For example, under a logistic birth law (equation (13)), the mean-field approximation equation (14) yields an accurate value of N^*. At this value of N^*, the growth and death rates take on specific values defined by $r^* := r(N^*)$, $\mu(N^*) := \mu^*$. In fact, to absolutely fix N^*, the stochastic dynamics are driven by completely coupled birth and death events. During each event, one individual is randomly chosen to die and immediately replaced by a new one. This tethering of birth and death ensures that the total population N^* is fixed. The total rate of a tethered birth-death event is $\frac{1}{2}(\alpha H + r^*N^* + \mu^*N^*) = \mu^*N^*$, where the factor 1/2 factors in the fact that two birth-death events occur simultaneously during one tethered event, so on average the arrival rate of events has to be halved. Thus, μ^*N^* is the intrinsic rate of evolution in the Moran model. The master equation for the probability distribution $P_{\rm M}(n_1; t|N^*)$ of the fixed-N^* two-species Moran model can be expressed as

$$\begin{aligned}\frac{\partial P_{\rm M}(n_1;t|N^*)}{\partial t} =& \omega_{12}(n_1-1|N^*)P_{\rm M}(n_1-1|N^*) + \omega_{21}(n_1+1|N^*)P_{\rm M}(n_1+1|N^*) \\ &- \left[\omega_{12}(n_1|N^*) + \omega_{21}(n_1|N^*)\right] P_{\rm M}(n_1|N^*),\end{aligned} \tag{21}$$

where the functions $\omega_{ji}(n|N^*)$ denote the rate that a species-i individual is replaced by a species-j individual in a Moran process of fixed total population N^*

$$\begin{aligned}\omega_{12}(n|N^*) &= n\left(1-\frac{n}{N^*}\right)r^* + \left(1-\frac{n}{N^*}\right)\alpha \\ &= \mu^*N^*\left[(1-m^*)\frac{n}{N^*}\left(1-\frac{n}{N^*}\right) + m^*Q_1\left(1-\frac{n}{N^*}\right)\right], \\ \omega_{21}(n|N^*) &= n\left(1-\frac{n}{N^*}\right)r^* + (H-1)\left(\frac{n}{N^*}\right)\alpha \\ &= \mu^*N^*\left[(1-m^*)\frac{n}{N^*}\left(1-\frac{n}{N^*}\right) + m^*(1-Q_1)\left(\frac{n}{N^*}\right)\right],\end{aligned} \tag{22}$$

where we have further defined

$$m^* \equiv \frac{\alpha H}{\mu^*N^*}, \quad Q_1 = \frac{1}{H}. \tag{23}$$

Here, m^* represents the relative total immigration rate and Q_1 is the fixed fraction of species 1 amongst those in the immigration source.

In these dynamics, it is clear that the probabilities of choosing an individual for removal/death from species 1 and species 2 (the bath species) are n_1/N^* and $1 - n_1/N^*$, respectively. The newly created (from birth) individual has probability n_1/N^* to be of species 1 and $1 - n_1/N^*$ to be of species 2, calculated from the state of the model prior to death. Thus, after one event, the population of species 1 may increase by 1 (if a species-2 individual is chosen to die, and a species-1 individual is chosen to be born) or decrease by 1 (if a species-1 individual is chosen to die, and a species-2 individual is chosen to be born). The total rate of population change of any one species includes the per-cell immigration rate α, which is equal to the

per-species immigration rate since the cells initiating immigration are unique (see figure 1). The total immigration into the 'bath' species (species 2) is thus $(H-1)\alpha$.

To solve equation (21) in steady state, we use equation (22) and invoke the detailed balance condition $\omega_{12}(n_1-1|N^*)P^*_{\rm M}(n_1-1|N^*)=\omega_{21}(n_1|N^*)P^*_{\rm M}(n_1|N^*)$ to obtain

$$P^*_{\rm M}(n_1|N^*)=P^*_{\rm M}(0|N^*)\frac{\omega_{12}(0|N^*)}{\omega_{21}(n_1|N^*)}\prod_{\ell=1}^{n_1-1}\frac{\omega_{12}(\ell|N^*)}{\omega_{21}(\ell|N^*)},\quad P^*_{\rm M}(0|N^*)=\left[\sum_{n_1=0}^{N}\prod_{\ell=1}^{n_1}\frac{\omega_{12}(\ell-1|N^*)}{\omega_{21}(\ell|N^*)}\right]^{-1}. \tag{24}$$

For general q-dimensional ($q\geqslant 2$) Moran models that involve ($q+1\geqslant 3$) subpopulations, closed-form solutions are difficult to obtain. However, we can approximate these models using a diffusion approximation that treats the species fractions $x_i=n_i/N^*$ ($1\leqslant i\leqslant q$) as continuous variables. After Taylor-expanding q-dimensional discrete master equations and assuming $m^*\equiv\frac{\alpha H}{\mu^* N^*}\ll 1$, a simple q-dimensional Fokker–Planck equation can be derived [32, 33]

$$\frac{\partial P_{\rm M}(\mathbf{x}|N^*)}{\mu^*\partial t}+\sum_{i=1}^{q}\frac{\partial\left[A_i(\mathbf{x})P_{\rm M}(\mathbf{x}|N^*)\right]}{\partial x_i}=\frac{1}{N^*}\sum_{i=1}^{q}\sum_{j=1}^{q}\frac{\partial^2\left[B_{ij}(\mathbf{x})P_{\rm M}(\mathbf{x}|N^*)\right]}{\partial x_i\partial x_j} \tag{25}$$

where

$$A_i(\mathbf{x})=\sum_{j=1}^{q}m^*(Q_i-x_i),\ B_{ii}(\mathbf{x})=x_i(1-x_i),\ B_{ij}(\mathbf{x})=-x_ix_j\ (i\neq j). \tag{26}$$

For example, when $q=2$ (three species), we have $Q_1=Q_2=\frac{1}{H}$, $Q_3=\frac{H-2}{H}$. We explicitly show the derivations for the 1D and 2D Fokker–Planck equations in appendix E. The exact steady-state solution of the general q-dimensional diffusion model is known and follows the Dirichlet distribution [34]

$$P^*_{\rm M}(\mathbf{n}|N^*)=\Gamma(N^*m^*)\prod_{i=1}^{q+1}\frac{(n_i/N^*)^{N^*m^*Q_i-1}}{\Gamma(N^*m^*Q_i)}. \tag{27}$$

3.3. Relaxing the fixed-population constraint of the Moran model

While the Moran model can be used to approximate $P^*(n_1,N')$, it includes an additional hard constraint $n_1+N'=N^*$ that is not imposed in the original BDI model. In fact, n_1 itself can fluctuate above N^*. To relax this fixed-population constraint and find an improved approximation to the reduced QSS distribution $P^*(\{n_1,n_2,...,n_q\})$, we simply allow the system size of the Moran process to vary and weight each QSS Moran process by the steady-state probability distribution

$$P^*(N)=\frac{\prod_{j=1}^{N}\frac{r(j-1)+\alpha H}{\mu(j)}}{\sum_{m=0}^{\infty}\prod_{\ell=1}^{m}\frac{r(\ell-1)+\alpha H}{\mu(\ell)}}, \tag{28}$$

which is readily obtained from solving equation (3), the master equation for the total population of the BDI process. We thus use a whole family of Moran models, each at a different value of N, weighted by $P^*(N)$ to approximate the QSS probability

$$P^*(\{n_1,n_2,...,n_q\})=\sum_{N=1}^{\infty}P^*_{\rm M}\left(\{n_1,n_2,...n_{q+1}\}|N\right)P^*(N). \tag{29}$$

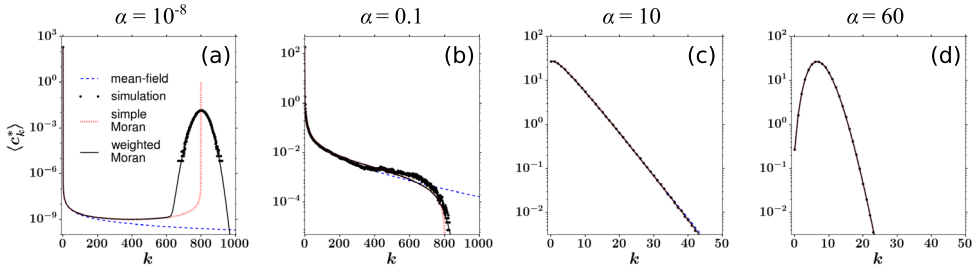

Figure 4. Simulated (black dots), mean-field (blue dashed), simple Moran (red hashes), and weighted Moran (black solid) approximations of $\langle c_k^* \rangle$ using logistic growth laws and the parameters $\mu = 10$, $p = 20$, $K = 1600$, $H = 200$. Immigration rates used were (a) $\alpha = 10^{-8}$, (b) $\alpha = 0.1$, (c) $\alpha = 10$, and (d) $\alpha = 60$, as in figure 3. In (a) we show the prediction from a single Moran model of fixed size N^*. The weighted QSS Moran model approach (solid black curves) yields a very accurate approximation to the simulated values of $\langle c_k^* \rangle$ for all values of α, including small α as shown in (a) and (b).

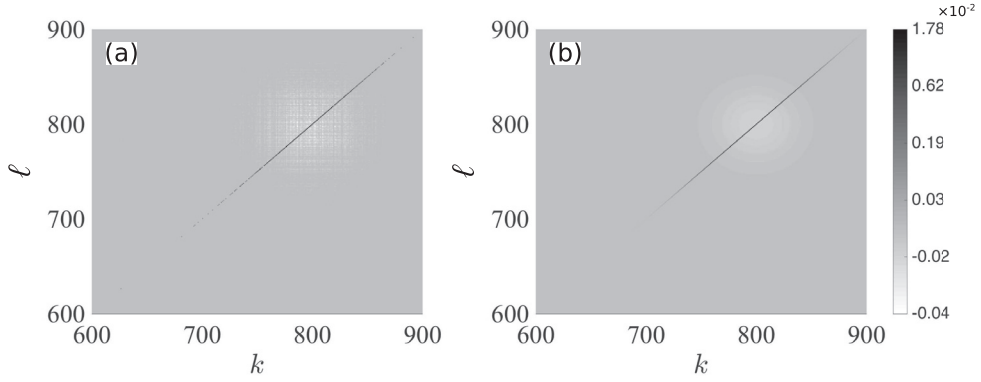

Figure 5. $\mathrm{cov}(c_k, c_\ell)$ from simulations (a) and from our calculations (b). Parameters are $\alpha = 10^{-8}$, $\mu = 10$, $p = 20$, $K = 1600$, $H = 200$. Only the interesting ranges of k and ℓ close to $N^* \approx 800$ are shown. The pattern shows that large species counts are positively self-correlated (black line) but are negatively correlated with neighboring counts (white dots). The grey background shows no correlation between species counts of significantly different population levels. Greyscale values are shown on an exponential scale.

Different values of the system size will yield different values of the rates $\omega_{ji}(n|N)$ according to equation (22). In 1D, according to equation (22), the ratio $\frac{\omega_{12}(\ell|N)}{\omega_{21}(\ell|N)}$ varies with N according to

$$\frac{\omega_{12}(\ell|N)}{\omega_{21}(\ell|N)} = \frac{(1-m^*)\frac{\ell}{N}\left(1-\frac{\ell}{N}\right) + m^* Q_1 \left(1-\frac{\ell}{N}\right)}{(1-m^*)\frac{\ell}{N}\left(1-\frac{\ell}{N}\right) + m^*(1-Q_1)\left(\frac{\ell}{N}\right)}, \tag{30}$$

where we have kept the intrinsic rates r^* and μ^* and the relative immigration rate m^* *fixed*. The only terms in equation (30) that vary with N are the relative populations ℓ/N and $1 - \ell/N$ reflecting only the changes associated with changes in system size. By keeping the r^*, μ^*, and m^* fixed, we preserve the relative tethered rates of birth, death, and immigration that define the original BDI process.

4. Results

4.1. $\langle c_k \rangle$ and $\langle c_k c_\ell \rangle$ under logistic growth

In figure 4, we plot results from Monte-Carlo simulations, mean-field solutions to equation (11), numerical solutions of the simple Moran model equation (24), and the weighted Moran model defined by equations (24), (29) and (30). As shown by figure 4(a), the simple Moran model has a sharp peak at N^* arising from the fixed-population constraint. The improved weighted solution yields accurate expected QSS species count distributions $\langle c_k^* \rangle$ for all values of α, capturing the the peak for extremely small α as well as the fast decay at large k.

To calculate the covariance between c_k^* and c_ℓ^* at QSS, we use the 2D ($q = 2$) 'continuum' solution given in equation (27) in the weighting in equation (29) in order to numerically compute equation (20). The covariances $\mathrm{cov}(c_k^*, c_\ell^*) \equiv \langle c_k^* c_\ell^* \rangle - \langle c_k^* \rangle \langle c_\ell^* \rangle$ with $\alpha = 10^{-8}$, both from Monte-Carlo simulations and from our weighted Moran model approximation, are plotted in figure 5. The results provide insight on how the true dynamics for $\langle c_k^* \rangle$ in equation (5) differs from that of the mean-field description in equation (11).

The large values (black line) along the diagonal $k = \ell$ correspond to the peak in $\langle c_k^* \rangle$ (equation (20) is dominated by the $HP(n_1 = k;t)$ term). White regions in the off-diagonal areas imply negative correlation between species counts of large neighboring sizes. In other words, whenever we observe a species with 800 individuals in a simulation at any fixed time t (at QSS), we will probably not observe another species with 801 cells at the same time. Grey areas that are farther away (such as $k = \ell = 600$) represent transient states of the system and have near-zero covariances.

4.2. Other forms of global interactions

Since global interactions across all species mediate the breakdown of the mean-field approximation, we now investigate different forms of regulation imposed through the functions $r(N)$ and $\mu(N)$. To explore how the 'stiffness' of different total population constraints affects the expected QSS species-count vector $\langle c_k^* \rangle$, we consider a simple Hill-type birth function with Hill coefficient 1:

$$r(N) = \frac{p_2 K_2}{K_2 + N}, \qquad \mu(N) = \mu_2. \tag{31}$$

This form imposes a 'softer' constraint on the total population N than the logistic birth function. In order to compare the results with those of the logistic model in section 4.1, we use the same values of α and H and use $\mu_2 = \mu$, $p_2 = p$ and $K_2 = K - N^*$ where N^* is the QSS population size obtained from the logistic model.

Another way to implement regulation is by keeping the birth rate constant but allowing the death rate to be population-dependent: [24]

$$r(N) = r_3, \qquad \mu(N) = \mu_3 \left(1 + \frac{N}{K_3}\right). \tag{32}$$

Again, we are interested in expected species counts near the same N^* as in section 4.1, and we set $K_3 = K$, $r_3 = r^*$, $\mu_3 = \mu^*/(1 + \frac{N^*}{K_3})$, where r^* and μ^* are the QSS values of the birth and death rates used in the logistic model.

Note that both the alternative regulation models, the Hill-type model and the population-dependent death model, generate the same steady-state rates r^* and μ^* at the same QSS total population size N^* as in the logistic model. Thus, we can compare the expected species-counts

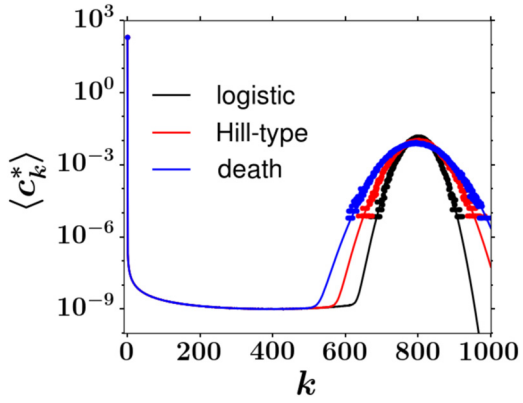

Figure 6. Comparison of $\langle c_k^* \rangle$ across the three regulation models, logistic (narrowest, black), Hill-type (intermediate, red), and population-dependent death (widest, blue). Simulations and results from the weighted Moran model approximations are shown. Here, the immigration rate is $\alpha = 10^{-8}$ where the mean-field approximation is invalid. The other parameters are $\mu_2 = 10$, $p_2 = 20$, $H = 200, K = 1600, r_3 = r^*$, and $\mu_3 = \mu^*$.

from all three models on the same footing. Since the mean-field solution given in equation (11) depends only on r^* and μ^*, all three models yield identical mean-field solutions $\langle c_k^* \rangle$. Therefore, for not-too-small values of α, for which mean-field solutions are accurate, all three models yield the same $\langle c_k^* \rangle$.

However, for small α, where the mean-field approximation breaks down, we expect that the peak in $\langle c_k^* \rangle$ near $k = N^*$ will be quantitatively different among the three models. In figure 6, we set $\alpha = 10^{-8}$ and plot the expected species count (from simulations and our weighted Moran model approximation) associated with each of the three models. Note that the peaks in $\langle c_k^* \rangle$ differ in their widths. In all examples, the underlying Moran models are identical and the differences originate in the different total-population distributions $P^*(N)$ across the three regulation models, as illustrated by the different 'widths' of the peak near N^*. According to simulations and numerical solutions of our weighted Moran model, the peak widths corresponding to each regulatory model are ranked according to population-dependent death $>$ Hill-type $>$ logistic growth.

A wider peak in $\langle c_k^* \rangle$ can be associated with a 'softer' total population constraint. As long as $f(N)$ on the right-hand side of equation (4) is a differentiable near N^*, we can define the regulatory 'stiffness' by

$$|f_*'| = -f_*' \equiv \left.\frac{\mathrm{d}f(N)}{\mathrm{d}N}\right|_{N^*} \in [0, +\infty). \tag{33}$$

Note that $f_*' < 0$ as long as N^* is a locally stable point.

The larger $|f_*'|$ is, the more likely the next event will be 'compensatory' (e.g. a new birth increases the chance for the next event to be death). This stiffness can be also thought of as the curvature of a quadratic energy profile centered about N^*. The stiffnesses of our three examples (using $\alpha = 10^{-8}$) are $|f_*'| = |p - \mu - \frac{2p}{K}N^*| = 10$ for logistic birth, $|f_*'| = |p_2 \frac{(N^*)^2 + 2K_2N^*}{(N^* + K_2)^2} - \mu_2| = 5$ for Hill-type regulation, and $|f_*'| = |r_3 - \mu_3 - \frac{2\mu_3}{K_3}N^*| = 3.3$ for population-dependent death. These stiffness values are consistent with the progression of peak widths shown in figure 6.

We may extend our definition of the stiffness to cases where $f(N)$ is not differentiable. For example, the Moran model has an infinitely 'stiff' constraint ($f'_* = -\infty$) which 'forces' an immediate death after a new birth. Nevertheless, as we have shown, even though a regulated BDI model may have much less sensitivity than the Moran model, the latter still provides insights on how such regulatory effects can induce an expected species count that exhibits a peak at N^*.

4.3. Energy landscape and phase transition in the species-count distribution

In this section, we provide an interpretation of the failure of the mean-field equation (equation (11)) as a 'phase-transition' in the statistics of populations. The full high-dimensional BDI model in QSS can be described by two processes: (1) evolution of individual species fractions via a Moran model that is equivalent across different regulation models, and (2) fluctuations of the total population size according to a QSS distribution that depends on the regulation model. Since the failure of the mean-field approach arises essentially from the emergence of a species that represents a large fraction of the whole population, we focus on the contribution of the Moran process.

A phase-transition can be conveniently visualized using a potential energy landscape ϕ, as is widely used in population genetics and developmental biology [35–41]. Its recent development in the physics community has extended its application to quantitative and systems biology [42–44]. Defined as a measure of 'generalized energy', its gradient indicates the direction of evolution of the system and its minima (potential wells) denote local stable states.

To simplify the math, we consider the $N \gg 1$ limit and use the continuum limit of the Moran model to find a continuum energy landscape $\phi(\{x_1, x_2, ...\})$ such that $P^*_{\mathrm{M}}(\mathbf{x}) \propto \mathrm{e}^{-\phi}$ satisfies equation (25) in steady-state. The shape of ϕ across $\{x_1, x_2, ...\}$ characterizes the global stability of the model. Starting from the 1D version of equation (25) we have

$$A(x) = m^* \left(\frac{1}{H} - x\right), \quad B(x) = x(1-x), \tag{34}$$

which allow us to define the 1D energy function [41]

$$\begin{aligned}\phi(x) &\equiv -N^* \int^x \frac{A(y)}{B(y)} + \ln B(x) \\ &= \left(1 - \frac{\alpha}{\mu^*}\right) \ln(x) + \left(1 - \frac{\alpha(H-1)}{\mu^*}\right) \ln(1-x) \\ &\equiv \frac{1}{H-1}\left[(H-1) - \frac{\alpha}{\alpha_{\mathrm{c}}}\right] \ln x + \left(1 - \frac{\alpha}{\alpha_{\mathrm{c}}}\right) \ln(1-x).\end{aligned} \tag{35}$$

Here, the parameter

$$\alpha_{\mathrm{c}} \equiv \frac{\mu^*}{H-1}, \tag{36}$$

is a critical immigration rate that controls a 'phase transition'. Equation (36) is unambiguous when μ is constant. If the regulation arises from a population-dependent rate $\mu(N)$ as in equation (32), the critical immigration rate α_{c} can be approximated by self-consistently solving $\alpha_{\mathrm{c}} = \mu(N^*(\alpha_{\mathrm{c}}))/(H-1)$.

When $P^*_{\mathrm{M}}(x)$ is normalisable, the energy function satisfies $\phi(x) \propto -\ln P^*_{\mathrm{M}}$. Since $\ln(0^+) \to -\infty$ and $\ln(1) = 0$, the shape of $\phi(x)$ is determined by the signs of the coefficients $H - 1 - \frac{\alpha}{\alpha_{\mathrm{c}}}$ and $1 - \frac{\alpha}{\alpha_{\mathrm{c}}}$. Assuming $\alpha_{\mathrm{c}} > 0$ (see section 4.4 for the special case $\alpha_{\mathrm{c}} = 0$), different regimes of the model can be delineated

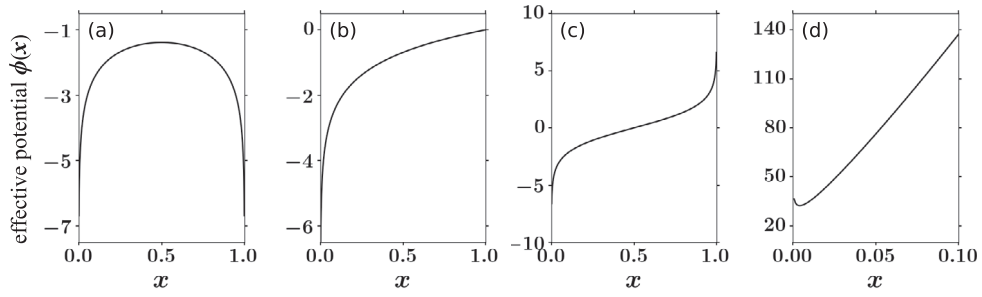

Figure 7. Energy landscapes $\phi(x)$ as a function of $x = n/N^*$ for $\mu = 10$, $p = 20$, $K = 1600$, $H = 200$ and different values of α. (a) $\alpha = 10^{-8}$ corresponding to figures 4(a) and (b) $\alpha = \alpha_c = 10/199$, (c) $\alpha = 0.1$, and (d) $\alpha = 60$. The minimum at $x > 0$ corresponds to the peak in $\langle c_k^* \rangle$ arising at $k > 1$.

- When $\alpha < \alpha_c$, we have $\alpha < (H-1)\alpha_c$ for $H \geqslant 2$. Two infinite minima in $\phi(x)$ emerge; one at $x = 0$ and one at $x = 1$. Associated with each minima is a basin of attraction as shown in figure 7(a). Even if all species start with a small fraction $x_i \ll 1$, one of them can eventually come to dominate by crossing to the attractive peak at $x = 1$ causing a failure of the mean-field description. However, this transition is different from the usual stochastically-driven 'escape' in statistical physics (see discussion). When α is extremely small, the 'extinction' state $x = 0$ is approximately absorbing for each species and the mean-field approximation fails severely.
- When $\alpha = \alpha_c$, $\alpha < (H-1)\alpha_c$ for $H > 2$. The potential $\phi(x) = \frac{H-2}{H-1} \ln x$ is monotonic and exhibits a global diverging minimum at $x = 0$ and a global maximum at 1. The whole interval $[0, 1]$ is a basin of attraction for $x = 0$ as shown by figure 7(b). The energy away from $x = 0$ is very flat and the severity of the failure of the mean-field approach is sensitive to α when it is near α_c.
- When $(H-1)\alpha_c \geqslant \alpha > \alpha_c$, the potential $\phi(x)$ has a diverging minimum at $x = 0$ and a diverging maximum at $x = 1$ as shown in figure 7(c). The mean-field approach is accurate in this regime.
- When $\alpha > (H-1)\alpha_c$, there is a single finite minimum in $\phi(x)$ appearing at $x_{\min} = \frac{\alpha - \alpha_c(H-1)}{\alpha H - 2\alpha_c(H-1)}$, which is close to $x = 0$ when $H \gg 1$. The potential has diverging maxima at both $x = 0$ and $x = 1$ so the basin of attraction for $x_{\min}$ is the whole $[0, 1]$ interval as shown by figure 7(d). The mean-field approach is accurate in this case.

Physically, a small immigration rate $\alpha < \alpha_c$ does not allow a dominant species to be replaced by new ones. When $\alpha > \alpha_c$, the immigration frequency αH is larger that the rate of coarsening of the species counts thereby filling the system with new species and preventing any one species to dominate. Here, $\langle c_k^* \rangle$ is monotonically decreasing. At even larger $\alpha > (H-1)\alpha_c$, immigration of each species is frequent enough that $\langle c_k^* \rangle$ becomes very broad and again develops an interior peak at $k \approx x_{\min} N^*$. The collapsing of $\langle c_k^* \rangle$ into a single species is reminiscent of the collapse of cluster size distributions in self-assembly under finite resources [45].

4.4. Resolving the effects of α and H

The energy landscape formulation provides a general way to visualize whether there is phase transition in the dynamics of an individual species which can be applied to study how various parameters affect the model. Equipped with the energy landscape, we can now examine the species populations as the intrinsic immigration rate α and the total number of species H are varied, keeping the total immigration rate, αH, fixed. Varying α and H in this way will not change the dynamics of the total population N but will influence the dynamics of populations of individual species. This is readily shown by the different shapes of $\phi(x)$ as α and H change. For example, let $\mu^* = 0.33$. With $H = 2,\ \alpha = 1$, the landscape exhibits a 'most probable' species size maintained by high per-species immigration rate. However, if $H = 200$ and $\alpha = 0.01$, each species has a low immigration rate. The associated landscape $\phi(x) = 0.97 \ln x - 5 \ln(1 - x)$ exhibits a unique potential well at $x = 0$ as all species are driven small. We can also consider the limit $H \to \infty$ while keeping αH fixed. This limit approximates naive T cell generation by the thymus. While total thymic output αH is finite, there are theoretically $H > 10^{15}$ different species (TCR sequences) that can be generated although only about 10^6–10^8 different species survive [29]. In any case, this large value of H means that nearly every immigration is from a new, unrepresented species and $k = 0$ is an absorbing boundary for all existing species. Species labels keep changing, but the distribution of $\langle c_k \rangle$ reaches a QSS. The energy landscape becomes (taking $\alpha \to 0$ in equation (35)) $\phi(x) \to \ln x + \left(1 - \frac{\alpha H}{\mu^*}\right) \ln(1 - x)$. There is always a potential well at $x = 0$ while the dynamics near 1 depend on the sign of $\frac{\alpha H}{\mu^*} - 1$.

Recall that if $\alpha < \alpha_c$, the mean-field approximation to $\langle c_k \rangle$ fails. From equation (36), the critical value α_c increases as H decreases, rendering the mean-field approximation invalid for a larger range of immigration rates. When $\alpha_c \to 0$ (e.g. realized when $\mu^* \to 0$), equation (35) no longer has a valid form because birth would need to be negative ($N^* > K$) in order to balance immigration. Nonetheless, we can multiply the landscape function by a constant μ^* without affecting its ability to qualitatively characterize and classify the dynamics of the system. We then take the limit $\mu^* \to 0$ and get $\phi \propto -\alpha \ln(x) - \alpha(H - 1) \ln(1 - x)$, which always has a unique minimum between $(0, 1)$, corresponding to figure 7(d).

5. Discussion and summary

In our analysis, the non-mean-field behavior of the expected clone abundances is mediated by global regulation mechanisms that act uniformly across all clones. Such population-dependent interactions break independence between clones, are difficult to model, and consequently have been rarely discussed in the context of species diversity [23]. Empirical studies have focused on the small-to-intermediate range of k, where the distribution $\langle c_k^* \rangle$ is well approximated by the mean-field model (equation (11)) as seen in figures 3 and 4. In another study, Parsons *et al* [24] considered a neutral and quasi-neutral birth-death model with carrying capacity but focussed on the mean fixation time of any species rather than species counts. However, they find that fluctuations in the total population do not affect fixation times of neutral species, which is consistent with our finding that a fixed-population Moran process can be used to accurately construct the fractions $\frac{n_1}{N}$ of any species in our BDI model. To our knowledge,

the failure of predicting a large-size clone by equation (11) has not been explicitly discussed in detail.

In many contexts such as stem or progenitor cells in a bone marrow niche, or multiple species competing for common resources, the observation of one or a few large clones or high population species is often naturally attributed to selection (differences in growth or death rates). This largest 'outlier' clone can contain most of the population and be biologically more important than all other smaller clones/species in the organism/community. Our results show that a simpler mechanism may arise from slow immigration into a neutral birth-death process with regulation, providing an initial 'null hypothesis' for selection. Otherwise, one may incorrectly argue that the existence of such a singular outlier clone suggests a species selection effect.

To quantitatively understand the regulated multispecies BDI process, we showed that an often-used mean-field approximation captures the expected steady-state species counts at low populations, but completely misses a possible peak in the species abundance near the population supported by a general regulated birth–death–immigration process. This peak arises only when the immigration rate decreases below a threshold value.

To develop a theory that approximates the clone abundance distribution accurately in all parameter regimes, we then mapped the qth moment of the species abundance distribution c_k to a $(q+1)$-dimensional cell-count BDI model, which was then approximated by weighting over q-dimensional Moran models of different system size. The expected distribution and covariances of species counts were accurately calculated in parameter regimes in which the mean-field approximations break down. By exploiting the concept of energy landscapes, we analytically describe a phase transition which explains the failure of the mean-field approach in the original model. Our analysis shows that global (inter-species) carrying capacity, when combined with a random sampling mechanism, generates a genetic-drift-like effect [46] in a Moran model that ultimately destroys the universal power-law distribution of c_k.

In equation (5), dynamics of any $\langle c_k^* \rangle$ are controlled by $r(N)$, where $N \equiv \sum_\ell \ell c_\ell$ involves contributions from all clone populations $\ell = 0, 1, 2, ..., k, ...$ Recall that in many classical scenarios, the relative strengths of these effects on the kth component decay with distance $|k-\ell|$. For example, in the constant-rate BDI model, only $c_{k\pm1}$ and c_k affect the dynamics of c_k. Here, however, the contribution from c_ℓ is proportional to the index ℓ itself instead of on $|k-\ell|$. This is a type of 'long-range' interaction or long-distance coupling arising in theoretically challenging contexts in different areas [47–50]. Thus, the structure of $\langle c_k^* \rangle$ according to equation (5) can no longer be approximated by a simple monotonic form as is shown in figure 5 where correlations between large k and its neighboring states $k \pm 1$ are negative.

To effectively find higher moments of species counts in QSS, higher-dimensional Moran models can be used to construct the related QSS cell-count distributions. Diffusion approximations to high-dimensional Moran models provide convenient analytic-form steady-state distributions [34, 51, 52]. However, the boundary values of $P_{\rm M}^*(\mathbf{x})$ in the diffusion approximation may not accurately approximate those from the discrete Moran model, especially in higher dimensions. For example, when $N^*m^* \ll 1$ in equation (27), $P_{\rm M}(x_i = 0) = +\infty$ but $P_{\rm M}(x_i = \frac{1}{N}) \approx 0$. Near the boundary, $P_{\rm M}(0 < x < \frac{1}{N})$ generally changes in a highly nonlinear fashion. Only with extremely large N do the probability distributions of the discrete and continuous Moran models match well [53, 54]. Nevertheless, our result in figure 5 is accurate because the region of interest is far away from the boundaries. Moreover, the second moment

$\langle (c_k^*)^2 \rangle$ in equation (20) turns out to be dominated by the first-moment term $H\langle c_k^* \rangle$, which was calculated based on the exact discrete solution in equation (24).

It is worth noting that the *initial* establishment of the large clone i, denoted by the transition $x_i \approx 0 \to x_i \approx 1$, is different from traditional scenarios where clone i randomly crosses the energy barrier near $x = 0.5$ and 'escapes' to the other attractive basin. Here, the potential energy profile corresponds to QSS in which there are many different clones j starting with small fractions $x_j \approx 0$. One of these small clones eventually replaces the dominating clone i ($x_j \approx 1$). The waiting time for such replacement events was obtained by [41] as $T_r \sim \mathcal{O}(\frac{N^*}{\alpha(H-1)})$, a much longer time than the waiting time $T_2 \sim \mathcal{O}(\frac{N^*}{\mu^*})$ (see appendix C) for the establishment of the first dominant clone in our BDI regulated model under $\alpha(H-1) \ll \mu^*$.

Future improvements to our analysis include more accurately determining steady-state solutions of the higher-dimensional Moran models, especially near the boundaries and extending our approaches to time-dependent approximations. To better distinguish our neutral mechanism from true selection, a careful analysis of heterogeneous populations should be explored to determine how random dominance from neutral regulation might be balanced by selection in the form of heterogeneous growth, death, and immigration parameters.

Acknowledgments

This work was supported in part by grants from the NSF (DMS-1516675 and DMS-1814364) and the Army Research Office (W911NF-18-1-0345).

Mathematical appendices

Appendix A. Cell-count Master equation for $P(n_1, ..., n_i, ..., n_H; t)$

The high-dimensional master equation obeyed by the full multispecies distribution reads

$$\begin{aligned}\frac{\partial P(\mathbf{n};t)}{\partial t} &= \alpha \sum_{i=1}^{H} [P(n_1, ..., n_{i-1}, n_i - 1, n_{i+1}, ..., n_H) - P(\mathbf{n})] \\ &+ \sum_{i=1}^{H} [r(N-1)(n_i - 1)P(n_1, ..., n_i - 1, ..., n_H) - r(N)n_i P(\mathbf{n})] \\ &+ \sum_{i=1}^{H} [\mu(N+1)(n_i + 1)P(n_1, ..., n_i + 1, ..., n_H) - \mu(N)n_i P(\mathbf{n})] ,\end{aligned} \tag{A.1}$$

where $N \equiv \sum_{i=1}^{H} n_i$.

Appendix B. Dynamical equations for $\langle c_k(t) \rangle$

Define $P(\mathbf{c};t)$ as the probability of observing the configuration $\mathbf{c} = \{c_0, c_1, c_2, ...\}$ at a specific time t. Under constant immigration and population-regulated birth and death rates, the evolution of the full probability distribution satisfies the master equation

$$\frac{\partial P(\mathbf{c};t)}{\partial t} = -\sum_{k=0}^{\infty}[\alpha + (\mu(N) + r(N))k]c_k P(\mathbf{c}) + \sum_{k=0}^{\infty}(c_{k+1}+1)(k+1)\mu(N+1)P(\{...,c_k-1,c_{k+1}+1,...\}) + \sum_{k=0}^{\infty}(c_k+1)(\alpha + kr(N-1))P(\{...,c_k+1,c_{k+1}-1,...\}). \tag{B.1}$$

Without loss of generality, let us assume constant μ, α but regulated $r = r(N) = r(\sum_{k=1}^{\infty} kc_k)$. The expected clone count is

$$\langle c_\ell(t)\rangle = \sum_{c_\ell=0}^{H} c_\ell P(c_\ell;t) = \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_\ell P(c_0,c_1,...,c_{\ell-1},c_\ell,c_{\ell+1},...;t). \tag{B.2}$$

Substituting equation (B.1) into (B.2), we obtain

$$\begin{aligned}\frac{\mathrm{d}\langle c_\ell(t)\rangle}{\mathrm{d}t} &= \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_\ell \frac{\partial P(c_0,c_1,...,c_{k-1},c_k,c_{k+1},...;t)}{\partial t} \\ &= \sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_\ell\Bigg\{\sum_{k=0}^{\infty} -[\alpha+(\mu+r(N))k]c_k P(c_0,c_1,...,c_{k-1},c_k,c_{k+1},...) \\ &\quad + \sum_{k=0}^{\infty}(c_{k+1}+1)[(k+1)\mu]P(c_0,c_1,...,c_{k-1},c_k-1,c_{k+1}+1,...) \\ &\quad + \sum_{k=0}^{\infty}(c_k+1)[\alpha+kr(N-1)]P(c_0,c_1,...,c_{k-1},c_k+1,c_{k+1}-1,...)\Bigg\}.\end{aligned} \tag{B.3}$$

By collecting only terms in equation (B.3) that involve $r(N)$, we obtain two summations

$$\begin{aligned}S_1 + S_2 \equiv &-\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_\ell \sum_{k=0}^{\infty} r(N)kc_k P(c_0,c_1,...,c_{k-1},c_k,c_{k+1},...) \\ &+\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_\ell \sum_{k=0}^{\infty} r(N-1)k(c_k+1)P(c_0,c_1,...,c_{k-1},c_k+1,c_{k+1}-1,...).\end{aligned}$$

Consider the contribution of the kth terms in both summations:

- When $k < \ell - 1$ or $k \geqslant \ell + 1$, the kth term of S_1 becomes

$$-\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}...\sum_{c_k=0}^{H}... c_2 r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_k,...) \tag{B.4}$$

and kth term of S_2 becomes

$$\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H}...\,c_\ell r(N-1)(k-1)(c_{k-1}+1)P(c_0,c_1,...,c_{k-1}+1,c_k-1,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H-1}...\,c_\ell r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_{k-1},c_k,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_{k-1}=0}^{H}\sum_{c_k=0}^{H}...\,c_\ell r(N)(k-1)c_{k-1}P(c_0,c_1,c_2,...,c_{k-1},c_k,...). \tag{B.5}$$

The last equality holds since $P(c_k = H) = 0$ due to the constraint that if $c_k = H$, then all other $c_{m\neq k} = 0$.

- When $k = \ell - 1$, the kth term of S_1 is

$$-\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}...\,c_\ell r(N)(\ell-1)c_{\ell-1}P(c_0,c_1,c_2,...,c_k,...) \tag{B.6}$$

and the kth term of S_2 is

$$\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}...\,c_\ell r(N-1)(\ell-1)(c_{\ell-1}+1)P(c_0,c_{\ell-1}+1,c_\ell-1,c_3,...,c_k,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H-1}...\,(c_\ell+1)r(N)c_{\ell-1}P(c_0,c_1,c_2,c_3,...,c_k,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}...\,(c_\ell+1)r(N)c_{\ell-1}P(c_0,c_1,c_2,c_3,...,c_k,...). \tag{B.7}$$

The two terms sum to

$$\sum_{c_0=0}^{H}\sum_{c_1=0}^{H}\sum_{c_2=0}^{H}...\sum_{c_k=0}^{H}...\,r(N)c_{\ell-1}P(c_0,c_1,c_2,...,c_k,...)=\langle r(N)c_{\ell-1}\rangle. \tag{B.8}$$

- When $k = \ell$, the kth term of S_1 is

$$-\sum_{c_0=0}^{H}...\sum_{c_{\ell-1}=0}^{H}\sum_{c_\ell=0}^{H}...\,c_\ell r(N)\ell c_\ell P(c_0,c_1,c_2,...,c_k,...) \tag{B.9}$$

while the kth term of S_2 is

$$\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H}...\,c_\ell r(N-1)\ell(c_\ell+1)P(c_0,c_1,c_\ell+1,c_{\ell+1}-1,...,c_k,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H-1}...\,(c_\ell-1)r(N)\ell c_\ell P(c_0,c_1,c_2,...,c_k,...)$$
$$=\sum_{c_0=0}^{H}...\sum_{c_\ell=0}^{H}\sum_{c_{\ell+1}=0}^{H}...\,(c_\ell-1)r(N)\ell c_\ell P(c_0,c_1,c_2,...,c_k,...). \tag{B.10}$$

These two terms sum to

$$\sum_{c_0=0}^{H} ... \sum_{c_\ell=0}^{H} \sum_{c_{\ell+1}=0}^{H} ...r(N)\ell(-c_\ell)P(c_0, c_1, c_2, ..., c_k, ...) = -\ell\langle r(N)c_\ell\rangle. \tag{B.11}$$

Summarizing, terms that involve $r(N)$ in equation (B.3) are simplified as $(\ell-1)\langle r(N)c_{\ell-1}\rangle - \ell\langle r(N)c_\ell\rangle$. Terms involving α and μ can be similarly obtained if they are regulated by N. Together, equation (B.3) becomes

$$\frac{\mathrm{d}\langle c_\ell\rangle}{\mathrm{d}t} = \alpha(\langle c_{\ell-1}\rangle - \langle c_\ell\rangle) + \langle r(N)\left[(\ell-1)c_{\ell-1} - \ell c_\ell\right]\rangle + \langle \mu(N)\left[(\ell+1)c_{\ell+1} - \ell c_\ell\right]\rangle. \tag{B.12}$$

Appendix C. Multi-timescale dynamics of $N(t)$ and $c_k(t)$

For simplicity, we first discuss the model with no immigration ($\alpha = 0$) and a large carrying capacity K. In this limit, $N^* = (1-\frac{\mu}{p})K \sim \mathcal{O}(K^{-1})$. The deterministic equation (4) gives quite a good approximation for the typical dynamics for N in its first phase of evolution as $N(t)$ quickly approaches its QSS value $\langle N^*\rangle$. To estimate this timescale, one can integrate $\frac{\mathrm{d}N}{\mathrm{d}t}$ in equation (13) under $\alpha = 0$ to find $\langle N(t)\rangle = \frac{\langle N^*\rangle N_0}{N_0 + \mathrm{e}^{-(p-\mu)t}(\langle N^*\rangle - N_0)}$. Thus N approaches $\langle N^*\rangle$ in a characteristic timescale $\mathcal{O}(\frac{1}{p-\mu})$.

As N approaches $\langle N^*\rangle$, $r(N)$ also approaches $r(\langle N^*\rangle) \equiv r^*$, allowing $\langle c_k\rangle$ to approach its QSS value $\langle c_k^*\rangle$ which has a high peak at $k = 0$ and a small peak at $k \approx \langle N^*\rangle$. Although this peak is small, the number of individuals in this clone, $kc_{k\approx\langle N^*\rangle}$ can comprise nearly the entire population. This configuration is associated with a single large-size clone that persists after the disappearance of all other $H-1$ clones. If we define the number of living clones (or 'species richness')

$$R \equiv \sum_{k=1}^{\infty} c_k, \tag{C.1}$$

this 'coarsening' or 'fixation' process [46] decreases R from its initial value H to 1 in finite time. We define the waiting time for such a fixation to take place as T_1. Since fixation is most relevant to changes of fractions of clones, we study the problem in a Moran model. In a standard textbook such as [46], the mean time for the ith clone to fix (conditioned on its fixation) is $T_{\mathrm{fix}}(i) \approx -\frac{N^*}{\mu}\frac{N(0)-n_i(0)}{n_i(0)}\ln\left[1-\frac{n_i(0)}{N(0)}\right]$. The expected time until any arbitrary clone's fixation is then calculated by averaging each clone's fixation times over its its probability of fixation ($P_{\mathrm{fix}}(i) = x_i(0)$) as $T_{\mathrm{c}} = \sum_i T_{\mathrm{fix}}(i)P_{\mathrm{fix}}(i) \approx \frac{N^*}{\mu}$.

The last clone comprises the total population that is stabilized around N^* by the regulatory effect of $f(N)$. The total population fluctuates around N^* for an exponentially long time. The variance ('width') of such fluctuation near N^* can be calculated by invoking the full stochastic model equation (3) which leads to the solution $P^*(N)$ in equation (28). The fact that $P^*(0) \neq 0$ (although it is typically exponentially small) allows for a finite probability that N may incur a large deviation to the absorbing boundary $N = 0$, resulting in extinction of the total population [54]. The expected time to extinction of the total population is $T_{\mathrm{ext}} = \sum_{m=1}^{n} a_m$ where $a_m = \frac{1}{\mu m} + \sum_{j=1}^{\infty}\frac{1}{\mu(m+j)}\prod_{i=1}^{j}\frac{r_{m+i-1}}{\mu}$ [53]. The asymptotic approximation $T_{\mathrm{ext}} \sim \mathcal{O}(\mathrm{e}^{\langle N^*\rangle})$ indicates a very long timescale for extinction, well after QSS limit of $\langle c_k^*\rangle$ is approached.

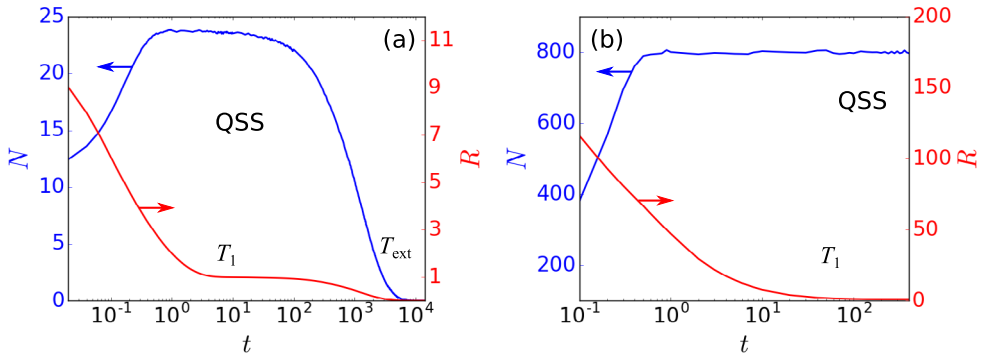

Figure C1. Simulations of the multi-timescale dynamics of a small (a) and a large (b) system. Common parameters are $\mu = 10$, $p = 20$. Different parameters are $K = 50$, $H = 11$, $\alpha = 0$ for (a) and $K = 1600$, $H = 200$, $\alpha = 10^{-8}$ for (b).

In figure C1, we plot simulations of the dynamics of both N and R under two different sets of parameters. Values of α are set to be extremely small or 0. Figure C1(a) shows that N reaches $N^* \approx 25$ within 1 unit of time and remains stable over approximately 10^2 before extinction. For $\alpha = 0$, we can identify a QSS within the time period $T_1 < t < T_{\text{ext}}$. Figure C1(b) incorporates a small immigration $\alpha = 10^{-8}$ so that $N = 0$ is technically no longer an absorbing boundary. Nonetheless, for extremely small $\alpha H T_{\text{ext}} \ll 1$, the dynamics are similar to the $\alpha = 0$ case (figure 7(a)) since the inter-immigration times $1/(\alpha H)$ are longer than the extinction time of the whole population.

Appendix D. Moments

The first moment of $\mathbf{c}$ is readily obtained by invoking its definition in equation (1) as

$$\langle c_k \rangle = \sum_{\mathbf{n}} [\mathbb{1}(n_1, k) + \mathbb{1}(n_2, k) + ... + \mathbb{1}(n_H, k)] P(\mathbf{n}) = H \sum_{\mathbf{n}} \mathbb{1}(n_1, k) P(\mathbf{n}) = HP(k).$$

The second moment, when $k \neq \ell$, is obtained as

$$\begin{aligned}
\langle c_k c_l \rangle &= \sum_{\mathbf{n}} [\mathbb{1}(n_1, k) + ... + \mathbb{1}(n_H, k)][\mathbb{1}(n_1, \ell) + ... + \mathbb{1}(n_H, \ell)] P(\mathbf{n}) \\
&= \sum_{\mathbf{n}} \sum_{i=1}^{H} \mathbb{1}(n_i, k)[\mathbb{1}(n_1, \ell) + ... + \mathbb{1}(n_H, \ell)] P(\mathbf{n}) \\
&= H \sum_{\mathbf{n}} \sum_{j \neq 1} \mathbb{1}(n_1, k)\mathbb{1}(n_j, \ell) P(\mathbf{n}) = H(H-1) \sum_{\mathbf{n}} \mathbb{1}(n_1, k)\mathbb{1}(n_2, \ell) P(\mathbf{n}) \\
&= H(H-1) P(k, \ell).
\end{aligned}$$

When $k = \ell$, we have

$$\begin{aligned}
\langle c_k c_k \rangle &= \sum_{\mathbf{n}} [\mathbb{1}(n_1, k) + ... + \mathbb{1}(n_H, k)][\mathbb{1}(n_1, k) + ... + \mathbb{1}(n_H, k)] P(\mathbf{n}) \\
&= H(H-1) P(k, k) + H \sum_{\mathbf{n}} \mathbb{1}(n_1, k)\mathbb{1}(n_1, k) P(\mathbf{n}) \\
&= H(H-1) P(k, k) + HP(k).
\end{aligned}$$

The third moment, when $k \neq \ell \neq m$, is obtained as

$$\begin{aligned}\langle c_k c_\ell c_m\rangle &= \sum_{\mathbf{n}}[\mathbb{1}(n_1,k)+...+\mathbb{1}(n_H,k)][\mathbb{1}(n_1,\ell)+...+\mathbb{1}(n_H,\ell)][\mathbb{1}(n_1,m)+...+\mathbb{1}(n_H,m)]P(\mathbf{n})\\ &= \sum_{\mathbf{n}}\sum_{i=1}^{H}\mathbb{1}(n_i,k)[\mathbb{1}(n_1,\ell)+...+\mathbb{1}(n_H,\ell)][\mathbb{1}(n_1,m)+...+\mathbb{1}(n_H,m)]P(\mathbf{n})\\ &= H\sum_{\mathbf{n}}\mathbb{1}(n_1,k)\sum_{i\neq 1}\mathbb{1}(n_i,\ell)\sum_{j\neq 1,i}\mathbb{1}(n_j,m)P(\mathbf{n})\\ &= H(H-1)(H-2)P(k,\ell,m).\end{aligned}$$

When $k=\ell\neq m$, we have

$$\begin{aligned}\langle c_k^2 c_m\rangle &= H(H-1)(H-2)P(k,k,m)+\sum_{\mathbf{n}}[\mathbb{1}(n_1,k)+...+\mathbb{1}(n_H,k)][\mathbb{1}(n_1,m)+...+\mathbb{1}(n_H,m)]\\ &= H(H-1)(H-2)P(k,k,m)+H(H-1)P(k,m).\end{aligned}$$

And finally when $k=\ell=m$, we obtain

$$\begin{aligned}\langle c_k^3\rangle &= H(H-1)(H-2)P(k,k,k)+H(H-1)P(k,k)+\sum_{\mathbf{n}}[\mathbb{1}(n_1,k)+...+\mathbb{1}(n_H,k)]\\ &= H(H-1)(H-2)P(k,k,k)+H(H-1)P(k,k)+HP(k).\end{aligned}$$

Appendix E. Diffusion approximation by the Taylor expansion

For notational simplicity, we replace x_1 with continuous variables x and neglect the subscript 'M' in the Moran model probability P_{M} in the rest of this subsection. Letting $\varepsilon=\frac{1}{N^*}\to 0$ ($N^*\to\infty$) in equation (21), we expand the transition rates to second order in ε:

$$\omega_{12}(x-\varepsilon)P(x-\varepsilon)\approx(\omega_{12}P)-\varepsilon(\omega_{12}P)'+\frac{\varepsilon^2}{2}(\omega_{12}P)'', \tag{E.1}$$

$$\omega_{21}(x+\varepsilon)P(x+\varepsilon)\approx(\omega_{21}P)+\varepsilon(\omega_{21}P)'+\frac{\varepsilon^2}{2}(\omega_{21}P)''. \tag{E.2}$$

Substituting them into equation (21), considering $\omega_{12}(x)=\alpha(1-x)+r^*N^*x(1-x)$, $\omega_{21}(x)=\alpha(H-1)x+r^*N^*x(1-x)$ in equation (22), and canceling out terms, we obtain (when $\alpha H\ll r^*N^*$)

$$\begin{aligned}\text{RHS} &\approx -\varepsilon[(\omega_{12}-\omega_{21})P]'+\frac{\varepsilon^2}{2}[(\omega_{12}+\omega_{21})P]''\\ &= -\frac{\alpha H}{N^*}\frac{\partial}{\partial x}\left(\frac{1}{H}-x\right)P+\frac{1}{2(N^*)^2}\frac{\partial^2}{\partial x^2}[\alpha(1-x)+\alpha(H-1)x+2r^*N^*x(1-x)]P\\ &\approx -\frac{\alpha H}{N^*}\frac{\partial}{\partial x}\left(\frac{1}{H}-x\right)P+\frac{r^*N^*}{(N^*)^2}\frac{\partial^2}{\partial x^2}x(1-x)P\\ &\approx \mu^*N^*\left[-\frac{1}{N^*}\frac{\partial}{\partial x}m^*\left(\frac{1}{H}-x\right)P_{\mathrm{M}}(x)+\frac{1}{(N^*)^2}\frac{\partial^2}{\partial x^2}x(1-x)P_{\mathrm{M}}(x)\right]\end{aligned} \tag{E.3}$$

where $m^*=\frac{\alpha H}{\mu^*N^*}$ is the fraction of birth that comes from immigration.

For the 2D Moran model, we have

$$\begin{aligned}\frac{\partial P(x_1,x_2)}{\partial t} &= (\omega_{21}P)(x_1+\varepsilon,x_2-\varepsilon)+(\omega_{31})P(x_1+\varepsilon,x_2)+(\omega_{12}P)(x_1-\varepsilon,x_2+\varepsilon)\\ &\quad+(\omega_{32}P)(x_1,x_2+\varepsilon)+(\omega_{13}P)(x_1-\varepsilon,x_2)+(\omega_{23}P)(x_1,x_2-\varepsilon)\\ &\quad-[(\omega_{21}+\omega_{31}+\omega_{21}+\omega_{32}+\omega_{13}+\omega_{23})P](x_1,x_2)\end{aligned} \tag{E.4}$$

where

$$\omega_{21}=\alpha x_1+r^*N^*x_2x_1,\qquad \omega_{31}=\alpha(H-2)x_1+r^*N^*x_3x_1, \tag{E.5}$$

$$\omega_{12}=\alpha x_2+r^*N^*x_1x_2,\qquad \omega_{32}=\alpha(H-2)x_2+r^*N^*x_3x_2, \tag{E.6}$$

$$\omega_{13}=\alpha x_3+r^*N^*x_1x_3,\qquad \omega_{23}=\alpha x_2+r^*N^*x_3x_2. \tag{E.7}$$

Invoking the 2D Taylor expansion on equation (E.4), we obtain terms like

$$\begin{aligned}(\omega_{21}P)(x_1+\varepsilon,x_2-\varepsilon)\approx(\omega_{21}P)&+\varepsilon\left[\frac{\partial(\omega_{21}P)}{\partial x_1}-\frac{\partial(\omega_{21}P)}{\partial x_2}\right]\\&+\frac{\varepsilon^2}{2}\left[\frac{\partial^2(\omega_{21}P)}{\partial x_1^2}-2\frac{\partial(\omega_{21}P)}{\partial x_1}\frac{\partial(\omega_{21}P)}{\partial x_2}+\frac{\partial^2(\omega_{21}P)}{\partial x_2^2}\right].\end{aligned}$$

The right-hand side of equation (E.4) is thus approximated by

$$\begin{aligned}\text{RHS}\approx{}&\varepsilon\left[\frac{\partial(\omega_{21}P)}{\partial x_1}-\frac{\partial(\omega_{21}P)}{\partial x_2}\right]+\frac{\varepsilon^2}{2}\left[\frac{\partial^2(\omega_{21}P)}{\partial x_1^2}-2\frac{\partial^2(\omega_{21}P)}{\partial x_1\partial x_2}+\frac{\partial^2(\omega_{21}P)}{\partial x_2^2}\right]\\&+\left[\varepsilon\frac{\partial(\omega_{31}P)}{\partial x_1}+\frac{\varepsilon^2}{2}\frac{\partial^2(\omega_{31}P)}{\partial x_1^2}\right]+\left[\varepsilon\frac{\partial(\omega_{32}P)}{\partial x_2}+\frac{\varepsilon^2}{2}\frac{\partial^2(\omega_{32}P)}{\partial x_2^2}\right]\\&+\varepsilon\left[-\frac{\partial(\omega_{12}P)}{\partial x_1}+\frac{\partial(\omega_{12}P)}{\partial x_2}\right]+\frac{\varepsilon^2}{2}\left[\frac{\partial^2(\omega_{12}P)}{\partial x_1^2}-2\frac{\partial^2(\omega_{12}P)}{\partial x_1\partial x_2}+\frac{\partial^2(\omega_{12}P)}{\partial x_2^2}\right]\\&+\left[-\varepsilon\frac{\partial(\omega_{13}P)}{\partial x_1}+\frac{\varepsilon^2}{2}\frac{\partial^2(\omega_{13}P)}{\partial x_1^2}\right]+\left[-\varepsilon\frac{\partial(\omega_{23}P)}{\partial x_2}+\frac{\varepsilon^2}{2}\frac{\partial^2(\omega_{23}P)}{\partial x_2^2}\right]\\={}&\varepsilon\left[\frac{\partial}{\partial x_1}(\omega_{21}+\omega_{31}-\omega_{12}-\omega_{13})P+\frac{\partial}{\partial x_2}(\omega_{12}+\omega_{32}-\omega_{21}-\omega_{23})P\right]\\&+\frac{\varepsilon^2}{2}\left[\frac{\partial^2}{\partial x_1^2}(\omega_{21}+\omega_{31}+\omega_{12}+\omega_{13})P+\frac{\partial^2}{\partial x_2^2}(\omega_{12}+\omega_{32}+\omega_{21}+\omega_{23})P-2\frac{\partial^2}{\partial x_1\partial x_2}(\omega_{12}+\omega_{21})P\right]\\={}&\mu^*N^*\left[-\frac{1}{N^*}\sum_{i=1}^{2}\frac{\partial A_i(\mathbf{x})P(\mathbf{x})}{\partial x_i}+\frac{1}{(N^*)^2}\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\partial^2 B_{ij}(\mathbf{x})P(\mathbf{x})}{\partial x_i\partial x_j}\right]\end{aligned} \tag{E.8}$$

where

$$A_i(\mathbf{x})=\sum_{j=1}^{2}m^*(Q_i-x_i),\ B_{ii}(\mathbf{x})=x_i(1-x_i),\ B_{ij}(\mathbf{x})=-x_ix_j\ (i\neq j). \tag{E.9}$$

The last step of equation (E.8) involves calculations based on equations (E.5)–(E.7) and the assumption $m^*\ll 1$. For example,

$$\begin{aligned}&\omega_{12}-\omega_{21}+\omega_{13}-\omega_{31}=\alpha(1-x_1)-\alpha Hx_1=\alpha H\left(\frac{1}{H}-x_1\right)\equiv\mu^*N^*\cdot m^*(Q_1-x_1)\\&\omega_{21}+\omega_{31}=\alpha(H-1)x_1+r^*N^*x_1(x_2+x_3)\approx\mu^*N^*\cdot x_1(1-x_1).\\&\omega_{12}+\omega_{13}=\alpha(1-x_1)+r^*N^*x_1(1-x_1)\approx r^*N^*x_1(1-x_1).\end{aligned} \tag{E.10}$$

ORCID iDs

Tom Chou https://orcid.org/0000-0003-0785-6349

References

[1] Zarnitsyna V I, Evavold B D, Schoettle L N, Blattman J N and Antia R 2013 Estimating the diversity, completeness, and cross-reactivity of the T cell repertoire *Frontiers Immunol.* **4** 485
[2] McGill B J *et al* 2007 Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework *Ecol. Lett.* **10** 995–1015
[3] Goyal S, Kim S, Chen I S Y and Chou T 2015 Mechanisms of blood homeostasis: lineage tracking and a neutral model of cell populations in rhesus macaques *BMC Biol.* **13** 85
[4] Venturelli O S, Carr A V, Fisher G, Hsu R H, Lau R, Bowen B P, Hromada S, Northen T and Arkin A P 2018 Deciphering microbial interactions in synthetic human gut microbiome communities *Mol. Syst. Biol.* **14** e8157
[5] Kim S *et al* 2014 Dynamics of HSPC repopulation in nonhuman primates revealed by a decade-long clonal-tracking study *Cell Stem Cell* **14** 473–85
[6] Sun J, Ramos A, Chapman B, Johnnidis J B, Le L, Ho Y-J, Klein A, Hofmann O and Camargo F D 2014 Clonal dynamics of native haematopoiesis *Nature* **514** 322–7
[7] Biasco L *et al* 2016 *In vivo* tracking of human hematopoiesis reveals patterns of clonal dynamics during early and steady-state reconstitution phases *Cell Stem Cell* **19** 107–19
[8] Koelle S J, Espinoza D A, Wu C, Xu J, Lu R, Li B, Donahue R E and Dunbar C E 2017 Quantitative stability of hematopoietic stem and progenitor cell clonal output in rhesus macaques receiving transplants *Blood* **129** 1448–57
[9] Xu S, Kim S, Chen I S Y and Chou T 2018 Modeling large fluctuations of thousands of clones during hematopoiesis: the role of stem cell self-renewal and bursty progenitor dynamics in rhesus macaque (in preparation) (https://doi.org/10.1101/343160)
[10] Laydon D J, Bangham C R M and Asquith B 2015 Estimating T-cell repertoire diversity: limitations of classical estimators and a new approach *Phil. Trans. R. Soc.* B **370** 20140291
[11] Dessalles R, D'Orsogna M R and Chou T 2018 Exact steady-state distributions of multispecies birth death immigration processes: effects of mutations and carrying capacity on diversity *J. Stat. Phys.* accepted (https://doi.org/10.1007/s10955-018-2128-4)
[12] Jin Q, Han H, Hu X, Li X, Zhu C, Ho S Y W, Ward R D and Zhang A 2013 Quantifying species diversity with a DNA barcoding-based method: Tibetan moth species (Noctuidae) on the Qinghai–Tibetan Plateau *PloS One* **8** e64428
[13] Qi Q, Liu Y, Cheng Y, Glanville J, Zhang D, Lee J-Y, Olshen R A, Weyand C M, Boyd S D and Goronzy J J 2014 Diversity and clonal selection in the human T-cell repertoire *Proc. Natl Acad. Sci.* **111** 13139–44
[14] Desponds J, Mora T and Walczak A M 2016 Fluctuating fitness shapes the clone-size distribution of immune repertoires *Proc. Natl Acad. Sci.* **113** 274–9
[15] Hill T C J, Walsh K A, Harris J A and Moffett B F 2003 Using ecological diversity measures with bacterial communities *FEMS Microbiol. Ecol.* **43** 1–11
[16] Hong S-H, Bunge J, Jeon S-O and Epstein S S 2006 Predicting microbial species richness *Proc. Natl Acad. Sci. USA* **103** 117–22
[17] Hubbell S P 2001 *The Unified Neutral Theory of Biodiversity and Biogeography (MPB-32)* (Princeton, NJ: Princeton University Press)
[18] Guisan A and Thuiller W 2005 Predicting species distribution: offering more than simple habitat models *Ecol. Lett.* **8** 993–1009
[19] Motomura I 1932 A statistical treatment of ecological communities *Zool. Mag.* **44** 379–83
[20] Fisher R A, Corbet A S and Williams C B 1943 The relation between the number of species and the number of individuals in a random sample of an animal population *J. Animal Ecol.* **12** 42–58
[21] Kendall D G 1948 On some modes of population growth leading to RA Fisher's logarithmic series distribution *Biometrika* **35** 6–15
[22] Kendall D G 1948 On the generalized 'birth-and-death' process *Ann. Math. Stat.* **19** 1–15
[23] Volkov I, Banavar J R, He F, Hubbell S P and Maritan A 2005 Density dependence explains tree species abundance and diversity in tropical forests *Nature* **438** 658

[24] Parsons T L, Quince C and Plotkin J B 2008 Absorption and fixation times for neutral and quasi-neutral populations with density dependence *Theor. Popul. Biol.* **74** 302–10
[25] Chotibut T and Nelson D R 2015 Evolutionary dynamics with fluctuating population sizes and strong mutualism *Phys. Rev.* E **92** 022718
[26] Constable G W A, Rogers T, McKane A J and Tarnita C E 2016 Demographic noise can reverse the direction of deterministic selection *Proc. Natl Acad. Sci.* **113** E4745–54
[27] Chotibut T and Nelson D R 2017 Population genetics with fluctuating population sizes *J. Stat. Phys.* **167** 777–91
[28] Constable G W A and McKane A J 2017 Mapping of the stochastic Lotka–Volterra model to models of population genetics and game theory *Phys. Rev.* E **96** 022416
[29] Lythe G, Callard R E, Hoare R L and Molina-París C 2016 How many TCR clonotypes does a body maintain? *J. Theor. Biol.* **389** 214–24
[30] Eftimie R, Gillard J J and Cantrell D A 2016 Mathematical models for immunology: current state of the art and future research directions *Bull. Math. Biol.* **78** 2091–134
[31] Gillespie C S 2009 Moment-closure approximations for mass-action models *IET Syst. Biol.* **3** 52–8
[32] Kimura M 1964 Diffusion models in population genetics *J. Appl. Probab.* **1** 177–232
[33] Blythe R A and McKane A J 2007 Stochastic models of evolution in genetics, ecology and linguistics *J. Stat. Mech.* P07018
[34] Baxter G J, Blythe R A and McKane A J 2007 Exact solution of the multi-allelic diffusion model *Math. Biosci.* **209** 124–70
[35] Wright S 1932 The roles of mutation, inbreeding, crossbreeding and selection in evolution *Proc. 6th Int. Congress of Genetics* **1** 356–66
[36] Waddington C H 1957 *The Strategy of the Genes: a Discussion of Some Aspect of Theoretical Biology* (London: Allen & Unwin)
[37] Sherrington D 1997 Landscape paradigms in physics and biology: introduction and overview *Physica* D **107** 117–21
[38] Arnold S J, Pfrender M E and Jones A G 2001 The adaptive landscape as a conceptual bridge between micro-and macroevolution *Microevolution Rate, Pattern, Process* (Berlin: Springer) pp 9–32
[39] Ao P 2009 Global view of bionetwork dynamics: adaptive landscape *J. Genet. Genomics* **36** 63–73
[40] Orr H A 2009 Fitness and its role in evolutionary genetics *Nat. Rev. Genet.* **10** 531–9
[41] Xu S, Jiao S, Jiang P and Ao P 2014 Two-time-scale population evolution on a singular landscape *Phys. Rev.* E **89** 012724
[42] Ao P 2004 Potential in stochastic differential equations: novel construction *J. Phys. A: Math. Gen.* **37** L25–30
[43] Qian H 2006 Open-system nonequilibrium steady state: statistical thermodynamics, fluctuations, and chemical oscillations *J. Phys. Chem.* B **110** 15063–74
[44] Wang J, Xu L and Wang E 2008 Potential landscape and flux framework of nonequilibrium networks: robustness, dissipation, and coherence of biochemical oscillations *Proc. Natl Acad. Sci.* **105** 12271–6
[45] D'Orsogna M R, Lakatos G and Chou T 2012 Stochastic self-assembly of incommensurate clusters *J. Chem. Phys.* **136** 084110
[46] Ewens W J 2012 *Mathematical population genetics 1: Theoretical introduction* vol 27 (Berlin: Springer)
[47] Morchio G and Strocchi F 1987 Mathematical structures for long-range dynamics and symmetry breaking *J. Math. Phys.* **28** 622–35
[48] Takayama H 2012 *Cooperative Dynamics in Complex Physical Systems: Proc. of the 2nd Yukawa Int. Symp. (Kyoto, Japan, 24–27 August 1988)* vol 43 (Berlin: Springer)
[49] Sanyal A, Lajoie B R, Jain G and Dekker J 2012 The long-range interaction landscape of gene promoters *Nature* **489** 109
[50] Chen J L, Voigt F F, Javadzadeh M, Krueppel R and Helmchen F 2016 Long-range population dynamics of anatomically defined neocortical networks *Elife* **5** e14679
[51] Ewens W J 1963 Numerical results and diffusion approximations in a genetic process *Biometrika* **50** 241–9
[52] Aalto E 1989 The Moran model and validity of the diffusion approximation in population genetics *J. Theor. Biol.* **140** 317–26
[53] Doering C R, Sargsyan K V and Sander L M 2005 Extinction times for birth-death processes: exact results, continuum asymptotics, and the failure of the Fokker–Planck approximation *Multiscale Modeling Simul.* **3** 283–99
[54] Kessler D A and Shnerb N M 2007 Extinction rates for fluctuation-induced metastabilities: a real-space WKB approach *J. Stat. Phys.* **127** 861–86