Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

Neutral diversity in experimental metapopulations

Guilhem Doulcier^{a,b,*}, Amaury Lambert^{c,d}

^a Macquarie University, Department of Philosophy, Sydney, Australia

^b Max Planck Institute for Evolutionary Biology, Department of Theoretical Biology, Plön, Germany

^c SMILE – Stochastic Models for the Inference of Life Evolution, Institut de Biologie de l'ENS (IBENS), École Normale Supérieure, CNRS UMR8197, INSERM

U1024, France

^d Centre Interdisciplinaire de Recherche en Biologie (CIRB), Collège de France, CNRS UMR7241, INSERM U1050, PSL Université, Paris, France

ARTICLE INFO	A B S T R A C T
<i>Keywords:</i> Neutral diversity Population genetics Experimental evolution	New automated and high-throughput methods allow the manipulation and selection of numerous bacterial populations. In this manuscript we are interested in the neutral diversity patterns that emerge from such a setup in which many bacterial populations are grown in parallel serial transfers, in some cases with population-wide extinction and splitting events. We model bacterial growth by a birth-death process and use the theory of coalescent point processes. We show that there is a dilution factor that optimises the expected amount of neutral diversity for a given number of cycles, and study the power law behaviour of the mutation frequency spectrum for different experimental regimes. We also explore how neutral variation diverges between two recently split populations by establishing a new formula for the expected number of shared and private mutations. Finally, we show the interest of such a setup to select a phenotype of interest that requires multiple mutations.

1. Introduction

Microbial communities are ubiquitous in nature and perform critical roles in biogeochemical cycles (Fuhrman, 2009), agricultural productivity (Trivedi et al., 2016) and human health (Pflughoeft and Versalovic, 2012). Microbial diversity has been recognised to have a strong impact on these dynamics (Kassen et al., 2000; Maron et al., 2018; Delgado-Baquerizo et al., 2016). As a consequence, considerable effort is being deployed in understanding and engineering their eco-evolutionary dynamics by either building "bottom-up" synthetic communities (Li et al., 2022), screening variants from environmental samples (Milshteyn et al., 2014), or finally developing "selective breeding" for microbial communities (Mueller and Sachs, 2015; Arias-Sánchez et al., 2019). In the following, we use stochastic processes to study some basic parameters of a device that could be used for the screening and selection of microbial communities. This device is abstract but inspired by recent advances in high-throughput microbial population manipulation methods.

Microbial populations offer numerous advantages in the domain of experimental evolution (i.e., the study of evolutionary dynamics happening in real time as a response to conditions imposed by the experimenter; Kawecki et al. 2012). These advantages include large population sizes, easily manipulable environments, the possibility to freeze and store whole populations indefinitely... Experimental evolution requires the set-up of many parallel bacterial cultures that can take several forms from bottles (in the order of 1 litre, $\approx 1L$) to tubes ($\approx 10^{-3}L$), microplates ($\approx 10^{-4}L$), or microfluidic compartments ($\approx 10^{-9}L$). Recently, new techniques for the high-throughput manipulation of bacterial populations have emerged. For instance, digital millifluidics (Cottinet, 2013; Dupin, 2018; Doulcier, 2019; Ardré et al., 2022) allows producing and imaging thousands of droplets of culture broth within a carrying fluid. The droplets amount to around $2 \times 10^{-6}L$ with a carrying capacity of 10^5 to 10^6 cells (Cottinet, 2013). Droplets can be imaged and quantitative measures be performed (e.g., optical density, fluorescence signal) during growth of the bacteria, allowing a high-throughput monitoring of ecological dynamics.

Being able to phenotype, screen and eventually artificially select microbial communities at scale could open new possibilities in experimental evolution as well as biological engineering. In particular, a new kind of experimental design becomes accessible, compared to more traditional experimental evolution set-ups (e.g., screening by plating or flow cytometry). We call this design "nested populations", in which both cells and microbial populations (e.g., physically bounded by droplets) have their own demography with birth and death events. The consequence of such a nested design is that populations themselves become units of selection (Lewontin, 1970) in their own right. Such

* Corresponding author at: Macquarie University, Department of Philosophy, Sydney, Australia. *E-mail addresses:* guilhem.doulcier@normalesup.org (G. Doulcier), amaury.lambert@ens.fr (A. Lambert).

https://doi.org/10.1016/j.tpb.2024.02.011

Received 12 July 2023

Available online 15 March 2024

0040-5809/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).







experiments are routinely performed in microcosms (Hammerschmidt et al., 2014). However, the ability of millifluidic devices to monitor in the order of a thousand cultures and retrieve some of them for analysis makes them particularly suitable for the artificial selection of microbial communities (Xie and Shou, 2021), such as through ecological scaffolding (Doulcier et al., 2020; Black et al., 2020).

Neutral diversity in experimentally nested populations is the focus of this manuscript. The aim is to build a quantitative understanding of simple diversity patterns within the experimental set-up. The main objective is to understand how they can be manipulated by changing the parameters that are accessible to the experimentalist or device designer. We consider the hypothesis that a high neutral diversity is favourable for the experiment, either because it increases the probability to screen individual variants of interest, or because populations with high diversity are more likely to have properties of interest themselves. Moreover, neutral diversity offers a solid foundation for more complex models of microbial communities (Ofiteru et al., 2010). First, a model of the device is presented. It relies on the assumption that cells undergo constant exponential growth. The optimal operating regime parameters of the machine (e.g., dilution factor, duration of droplet growth cycles, carrying capacity) are derived from characteristics of the biological material: birth and death rates. From a theoretical perspective, the system constitutes a dynamical meta-population in discrete space with explicit demography. It contrasts with simpler models, in which demography is simplified (Etheridge, 2008), as well as with more complex spatially structured models (Barton et al., 2002, 2013), in which space is continuous. Second, a coalescent model of the population across bottlenecks is proposed and coupled with a neutral mutation model with infinite alleles. This allows computation of the number of mutations and the distribution of allele frequencies within droplets after several droplet growth cycles. It shows that small bottlenecks are required to maximise diversity in one cycle, but larger bottlenecks are more favourable for diversification across many cycles. Then, the effect of splitting a droplet into several lineages is studied by computing the number of mutations accumulated in a single lineage or all the droplet lineages. Finally, a simple mutation accumulation model illustrates the interest of droplet-level selection for artificial selection.

2. Modelling nested population dynamics

Consider a device that allows the manipulation of droplets via serial transfers (Fig. 1). Cells are distributed among a train of *D* droplets (called populations or droplets). The birth and death of cells are modelled by a linear branching process with constant rates *b* for birth and *d* for death (Fig. 2). The net growth rate r := b - d is called the Malthusian parameter. After a duration *T*, a new train of *D* droplets is prepared by diluting them $\frac{1}{\delta}$ fold. Hence, for each dilution event, a cell has a probability δ of being sampled and thus being present in the new droplet (Fig. 3). This procedure is repeated periodically; each dilution followed by a growth phase constitutes a *cycle* of the experiment or a *droplet generation*. Tables 1–3 present a summary reference of all the symbols used in the text.

More formally, the initial population contains $Z_0 = c$ cells and is submitted to a bottleneck with sampling probability δ at the beginning of every cycle except the first, meaning that bottlenecks occur at times $T, 2T \dots nT$. The *n*th cycle corresponds to the slice of time [(n-1)T, nT). Thus, "the end of cycle *n*" corresponds to the moment nT^- , just before the *n*th dilution. To illustrate this, at the end of the third cycle, $t = 3T^-$, the population has experienced two bottlenecks at time *T* and 2*T*.

Birth *b* and death *d* rates depend on the biological material used (e.g., species, strain), as well as the culture medium, and are not easily controlled. However, the duration of the growth phase *T*, the dilution factor δ , and the number of droplets *D* can be changed by altering the experimental set-up. A model can help to predict the effect of those parameters and determine those that should be the focus of engineering efforts, such as the development of devices supporting larger droplets, an increased number of same-size droplets, a wider range of dilution factors, or droplets that are stable for longer times.

2.1. Optimal operating regime

When designing a serial transfer experiment, the operator has three main parameters that might be controlled: the size of the cultures (and, by extension, the carrying capacity of droplets, K, in number of cells), the duration of the growth phase separating two successive transfers T, and the dilution factor δ . Two problems must be avoided: if population sizes are too small and dilution too high, the resulting cultures might be empty. Conversely, if the population sizes are too large, and dilution too low, the population will spend most of its time in a stationary phase, with little effect of bottlenecks.

Any dilution event presents the risk of extinguishing the population. When performing a serial transfer experiment, this must be avoided at all costs because an empty microcosm brings about the end of the experiment (at least for the given independent lineage). In a nested populations' design, the presence of some empty microcosms can be tolerated because empty patches in the population can be filled by splitting a single parent droplet into several offspring droplets in the next generation.

In general, the stationary phase is not desirable for several reasons. First, a population that reaches saturation will go through fewer generations than if it was growing freely, reducing the potential evolutionary dynamics. Moreover, physiological changes in the stationary phase might result in undesired phenotypic effects on the population. Finally, in the case of millifluidic experiments, saturating densities are known to increase the risk of cross-contamination between droplets.

For all these reasons, there is an optimal dilution factor that keeps the population in an exponential phase while maximising the population size, at which selection experiments should be conducted .

A model of population dynamics can provide a first estimate of the optimal range of parameters for an experiment. In the following, a stochastic model of cells in exponential growth conditions (i.e., supercritical) with periodic bottlenecks is used to derive the probability of losing a single cell lineage, or a single droplet lineage due to the effect of dilution, as a function of experimentally accessible parameters. Saturation phenomena are not modelled explicitly because the birth and death rates are considered independent of population size, but the population dynamics are required to stay under a carrying capacity threshold.

2.1.1. Survival of a single lineage

A first quantity that can be derived from the linear branching process with periodic dilution that models the population dynamics is the probability that a single initial cell has no descent in the population after n cycles.

Proposition 1 (Survival Probability). Cells within droplets in serial transfers are modelled by a linear birth–death process, with constant parameters b and d, that is subject to periodic bottlenecks δ every duration T.

The probability s_n that a lineage spawned by a single cell at the beginning of the first cycle is not extinct at the end of the *n*th cycle is:

$$s_n = 1 - h(Q_1 Q_{\delta}^{n-1}, 0), \tag{1}$$

where $s \mapsto h(M, s)$ is the linear fractional function with coefficient $M \in M_2(\mathbb{R})$:

$$\forall M = \begin{bmatrix} v & w \\ x & y \end{bmatrix} \in M_2(\mathbb{R}), \quad h(M, s) = \frac{vs + w}{xs + y}.$$
(2)

and for $\varepsilon \in (0, 1]$ the matrix Q_{ε} is:

$$Q_{\varepsilon} = \begin{bmatrix} p - 1 + \varepsilon(1 - q) & (1 - \varepsilon) + \varepsilon q \\ p - 1 & 1 \end{bmatrix}$$
(3)

with,



Fig. 1. Sketch of the experimental set-up. Let a population of cells following a birth-death process with rates b, d be distributed within D droplets. After a growth phase duration T, a new cycle starts: the content of each droplet is diluted to seed a new droplet. Each cell has a probability δ to be transferred at the start of the next cycle. Here, the *serial transfer* regime is depicted: each droplet is diluted into exactly a single new droplet in the next cycle. In the full *nested populations'* design, a droplet can be split into several droplets in the cycle or removed altogether.



Fig. 2. Events in a Linear-Birth-Death Model. Individuals give birth to new individuals at a constant rate b, and die at constant rate d, independently. The process is super-critical if b > d.



Fig. 3. Dilution process. Individual cells are independently selected to be transferred to the next cycle (with probability δ) or to be discarded (with probability $1 - \delta$).

		p(b, d, T)	q(b,d,T)
Sub-critical cells	b < d	$\frac{-r}{d-be^{rT}}$	$\frac{d(1-e^{rT})}{d-be^{rT}}$
Critical cells	b = d	$\frac{1}{1+bT}$	$\frac{bT}{1+bT}$
Super-critical cells	b > d	$\frac{re^{-rT}}{b-de^{-rT}}$	$\frac{d(1-e^{-rT})}{b-de^{-rT}}$

Proposition 1 shows that the survival probability of a lineage depends on the birth rate *b* and death rate *d* of the cells but is also a function of the dilution factor δ , duration of the growth phase *T*, and the number of cycles *n*. When considering a single dilution and a pure birth process (s_2 , Fig. 4), the survival probability is equivalent to $\delta \mathbb{E}(Z_T)$ when δ is small; hence the linear increase with slope 1 in log–log scale.

The numerical computation of Q_{δ}^{n-1} might be problematic due to repeated multiplication of small numbers. However, since the final result only involves the ratio $(Q_{\delta}^{n-1})_{01} / (Q_{\delta}^{n-1})_{11}$, it is possible to



Fig. 4. Survival probability at the end of the second cycle s_2 . This corresponds to a first growth phase, a dilution event, and a second growth phase. The probability is presented as a function of the dilution factor δ for pure birth processes b > 0, d = 0, T = 1.

normalise Q_{δ} to have its smallest entry as 1. Indeed, this ratio does not depend on a multiplicative scalar on the matrix: $\forall \alpha > 0$, $h(Q_{\delta}^{n}, 0) = h(\alpha Q_{\delta}^{n}, 0)$. Taking $\alpha = \frac{1}{\delta}$ greatly improves the numerical stability of the computation.

The limit of this probability when the number of cycles increases gives a clearer understanding of the long-term behaviour of the population:

Proposition 2 (Long-Term Survival Probability). Let s_n be the survival probability after *n* cycles of the lineage spawned by a single cell.

$$\lim_{n \to +\infty} s_n = \begin{cases} 0 & \text{if } \delta e^{rT} \leq 1\\ \frac{r(\delta - e^{-rT})}{\delta b(1 - e^{-rT})} & \text{otherwise.} \end{cases}$$

Proposition 2 confirms that, in the long run, lineages become extinct with certainty ($s_{\infty} = 0$) if and only if the expected number δe^{rT} of cells descending from a single initial cell and surviving the first bottleneck is smaller than 1. It gives the survival probability otherwise (see Fig. 5).

In the following, only super-critical populations will be considered (b > d).

2.1.2. Optimal cycle duration and dilution

Saturation of cell dynamics is not desirable, as mentioned earlier. Depending on the nature of the cells (e.g., species, strain), and of the medium (e.g., pH, nutrient availability, temperature), it is possible to

n



Fig. 5. Survival probability for several cycles s_a . The probability is presented for pure birth processes b = 1, d = 0, T = 1. Above the critical threshold δ^* , the survival probability does not tend towards 0. The dotted line corresponds to the limit s_{∞} .



Fig. 6. Maximal Cycle Duration T^* as a function of the Malthusian parameter b - d and the initial occupancy $\frac{c}{v}$.

define an experimental carrying capacity K that corresponds to the number of cells that can be sustained in a droplet without saturation. The simple linear birth–death model cannot represent saturating populations because no density dependence is included in this model; thus, for the model to be coherent, the duration of the growth phase must be short enough that the population size does not reach the carrying capacity:

Proposition 3 (Maximal Cycle Duration). Let T^* be the maximal cycle duration before reaching saturation. Cells are following a super-critical birth-death process with growth rate b - d = r > 0. The carrying capacity is K and the initial number of cells is c:

$$T^* = -\frac{\ln(\frac{c}{K})}{r} \tag{4}$$

Proposition 3 shows that the optimal duration of the growth phase is linear in the inverse of the Malthusian parameter *r* of the population (Fig. 6), meaning that a population that grows (on average) twice as fast as another should be subject to cycles half as long as the other, for a given initial occupancy $\frac{c}{\kappa}$.

Additionally, the optimal duration of the growth phase is proportional to the logarithm of the initial occupancy $\frac{c}{K}$ of the droplet (with a minus sign, since this logarithm is always negative or zero as $c \le K$). As a consequence, for a given strain, multiplying the volume of the droplets by two, or dividing the inoculum size by two will increase the maximal duration of the growth phase by $\frac{\ln 2}{r}$.

To keep the same maximal duration T^* if the Malthusian parameter r is doubled, the carrying capacity of the droplet must be multiplied by the inverse of the previous initial occupancy K/c.

This result holds for a single cycle only. For a given dilution factor δ , the population is shrunk by an expected ratio δ , while for a given

cycle duration *T*, the population is expanded by an expected ratio δe^{rT} . To prevent the population from saturating for all cycles, the initial occupancy at the beginning of each cycle must be constant. This consideration allows discovery of the optimal dilution factor δ^* when the growth phase duration is fixed. By simply setting δ^* to be the solution of $\delta e^{rT} = 1$, we obtain $\delta^* = e^{-rT}$. We record this result in Proposition 4.

Proposition 4 (Optimal Dilution Factor). Let δ^* be the optimal dilution factor for which the expected number of cells is constant across generations. For cells following a birth-death process with growth rate b - d = r > 0 and a growth phase duration *T*:

$$\delta^* = e^{-rT}.$$
(5)

If $T = T^*$ (Proposition 3),

$$\delta^* = \frac{c}{K} \tag{6}$$

Proposition 4 shows that the dilution sampling probability should be equal to the initial occupancy when the duration of the cycle is maximal.

To summarise, the optimal operating regime of the experiment can be expressed from the Malthusian parameter of the population and the initial occupancy of the droplets $\frac{c}{K}$. As a result, the dilution sampling probability is $\delta^* = \frac{c}{K}$, and the duration of a cycle is $T^* = -\frac{\log(\delta^*)}{r}$. Fixing any two of (c, K, T, δ) values constrains the other two.

When the experiment is in the optimal regime, the expression of the survival of a lineage at cycle n is simpler:

Proposition 5 (Optimal Regime Survival Rate). Let s_n^* be the survival probability of a lineage after *n* cycles of duration *T* and with bottleneck $\delta^* = e^{-rT}$, where r = b - d > 0 is the Malthusian parameter of the population. Then:

$$s_n^* = \frac{r}{bn(1-\delta^*) + \delta^* r} \tag{7}$$

In the optimal regime, each initial cell has, on average, one descendant cell surviving the next bottleneck. The process counting the number of cells at time kT is thus a critical Galton–Watson process (as a function of k). Proposition 5 shows that, in the optimal regime, the survival probability of a lineage decreases as the inverse of the number of cycles, which is typical of critically branching populations.

Additionally, when taking a finite number of cycles *n*, the survival does not tend towards zero, even for vanishingly small bottlenecks δ . This derives from the fact that, in the optimal regime, a small bottleneck is compensated by a long cycle duration T^* ; thus, vanishingly small bottlenecks correspond to infinitely long cycles.

Finally, in the case of pure birth (i.e., d = 0), the survival of a lineage is independent of the birth rate *b*. It is certain if there is no bottleneck ($\delta = 1$), and tends towards $\frac{1}{n}$ for vanishingly small bottlenecks ($\delta \rightarrow 0$).

Overall, once the size of the droplets (which constrains *K*) and the initial occupancy (which constrains *c*) have been chosen by the operator, other parameters of the machine (duration of the growth phase *T*, dilution factor δ) can be deduced — and, conversely, fixing *T* and δ constrains *K* and *c*. The next section explores how one should select these parameters when the aim is to maximise genetic diversity within and between the droplets.

3. Modelling neutral diversity

Neutral diversity concerns mutations arising in the population of cells that are assumed not to change their birth or death rates. Neutral diversity gives rise to recognisable *patterns* that can be predicted from a mechanistic model of birth–death in the population. In experimental evolution, and *a fortiori* in artificial selection, it is desirable to increase



Fig. 7. Neutral mutations over the coalescent tree. The neutral mutations (coloured crosses) are distributed following a Poisson Process over the (real) coalescent tree (black). Each individual may carry several mutations distinguishing it from the most recent common ancestor.

the diversity within the population because this allows greater exploration of the phenotypic space. Indeed, mutations that are essentially neutral for cells might be of interest for the experimenter, or be intermediate states towards new phenotypes.

In the following, mutations follow a Poisson Point Process with constant rate θ over the lifespan of the cells, independently of their genealogy. As a consequence, the time between two mutations along a lineage (regardless of births and deaths) is exponentially distributed, and thus has no memory: the conditional expected time to the next mutation will be the same for all cells, irrespective of their age or the time of the last mutation in the lineage. This is a simplifying assumption that represents the spontaneous nature of mutations while ignoring the existence of mutations that can change the mutation rate (Sniegowski et al., 1997, 2000).

3.1. Coalescence times and the coalescent point process

The linear branching process (with constant birth rate *b* and death rate *d*) yields the full genealogy of the population (Fig. 8, left). However, the standing diversity at a given time in the population is affected (i) neither by the mutations occurring on lineages that do not have extant individuals (because their mutations have been lost) (ii) nor by mutations that are ancestral to the whole population (because they are shared by all individuals in the population). Thus, to characterise the standing diversity, it is sufficient to have knowledge of the *coalescent tree* of the population (Fig. 8, right), which is the genealogy of the extant individuals up to their most recent common ancestor.

Coalescent Point Processes (CPP) are stochastic processes whose realisations are real trees with the same probability as the coalescent tree of the corresponding branching process (Popovic, 2004; Lambert and Stadler, 2013). A CPP is defined by a time horizon $\tau \in \mathbb{R}_+$ and a node depth distribution f_H . The CPP is the sequence of independent and identically distributed variables $(H_i)_{i=1...N}$ following f_H and stopped at the first element N such that $H_N > \tau$ (the CPP is said to be "stopped at τ "). Usually, the node depth distribution is expressed in the form of the inverse tail distribution *F*:

$$F(t) := \frac{1}{P(H > t)} \tag{8}$$

3.2. Measuring neutral diversity

Neutral mutations do not affect the genealogy and can thus be superimposed *a posteriori* on the coalescent tree. Consider that mutations appear following a Poisson point process with constant rate θ over the coalescent tree. Thus, a mutation is a point on the coalescent tree, as illustrated in Fig. 7. Additionally, assume that reverse mutations are impossible (an assumption referred to as the "infinite sites model"), so that all individuals standing above the mutation in the coalescent tree (i.e., the descent of the mutation point) share the mutation (crosses at the top of Fig. 7). Individuals may carry zero, one, or several mutations.

The mutational richness of the population M (or total diversity) is the number of unique mutations found in the population. Its expected value is proportional to the total coalescent tree length.

The mutation frequency spectrum $(a_k)_{k \in \mathbb{N}}$ is another measure of diversity that counts how many mutations are represented by k individuals in the population.

All these measures require some knowledge of the shape of the coalescent tree of the population. The next paragraph is dedicated to establishing this for the simple case of serial transfer, while the next section is dedicated to the case of splitting droplets.

3.3. Diversity within droplets in serial transfer

Establishing the law of the CPP of a lineage within serial transfer requires identification of the law of the branch length. This law is well known for simple branching processes such as the linear birth–death process with parameters (b, d) (Lambert and Stadler (2013), Proposition 5).

The addition of repeated bottlenecks with period *T* is also possible within the theory (Lambert and Stadler (2013), Proposition 7) by thinning the original process (Fig. 9). Each bottleneck at time *iT*, i = 1, ..., n may remove independently each branch of the CPP with probability $(1-\delta)$ (in grey in Fig. 9). Removing a branch in the past (at time *iT*) may result in removing several branches in the present, and require an adjustment to branch length (green in Fig. 9). The number of branches removed and the adjustment to the branch length distribution can be computed from the law of the branch length of the CPP in the absence of the bottleneck, the sampling probability δ , and the period of the bottleneck *T*. As a result:

Proposition 6 (Coalescent Tree of a Lineage). Let \mathcal{T}_n be the random coalescent tree spawned by a single cell with extant descent at the end of the *n*th cycle. Then \mathcal{T}_n is a CPP stopped at *n*T, with inverse tail distribution F:

$$\begin{aligned} \forall t = kT + s, \quad F(t) &= \delta^k F(t) + (1 - \delta) \sum_{j=1}^k \delta^{j-1} F(jT) \\ &= 1 + \frac{b}{r} \left[e^{rs} (\delta e^{rT})^k - 1 + (1 - \delta) e^{rT} \frac{1 - (\delta e^{rT})^k}{1 - \delta e^{rT}} \right]. \end{aligned}$$

with b > 0, $d \ge 0$, $b \ne d$, $k \in \mathbb{N}$ and $0 \le s < T$. Additionally, in the case of critical dilution ($\delta^* = e^{-rT}$), we have:

 $\forall t = kT + s, \quad F(t) = 1 + \frac{b}{r} \left(e^{rs} - 1 + k(e^{rT} - 1) \right).$

The case b = d is covered in the appendix.

Proposition 6 gives the cumulative probability function for the node depth H: $\mathbb{P}(H < t) = 1 - \frac{1}{F(t)}$ (Fig. 10). Note that this function is defined piecewise for each cycle.

Because $F(t) \rightarrow \infty$ as $t \rightarrow \infty$, the cumulative distribution function of node depths tends to 1, which shows that *H* cannot take the value



Fig. 8. From a Birth-Death Process to a Coalescent Point Process. On the left-hand side is a birth-death process where a number of individuals give birth (eggs) and die (skulls) at different points in time, which flows from left to right. On the right-hand side is the corresponding continuous coalescent tree, where time flows from bottom to top. Note that at time *C*, lineage 3 coalesces with lineage 4, and that at time *B*, lineages 1 and 2 coalesce. Finally, at time *A*, lineage (1, 2) coalesces with lineage (3, 4).



Fig. 9. Bottlenecks are modelled by thinning the process. A bottleneck at some point in the past (red bottle) results in the extinction of some lineages (in grey), which modifies the genealogy (thick green lines).



Fig. 10. Cumulative probability for the branch length. Dotted lines represent bottlenecks.



Fig. 11. Node depth distribution. Sample of 10^8 realisations of the random variable *H* by the inversion of the cumulative probability function method. Deep nodes follow a power law distribution with parameter $\alpha = -1$ (orange line).

 $+\infty$, as is expected for critically (and also super-critically) branching populations.

The random variable *H* can be easily sampled from its cumulative probability function. As illustrated in Fig. 11, the distribution of deep nodes (deeper than depth 1) can be fitted by a power law with parameter -1 (criticality).

3.4. Number of mutations

To find the expected number of mutations within a droplet, the expected size of the full coalescent tree must be considered. This relies on the node depth distribution, conditioned to be lower than the duration of the experiment. It results in the following:

Proposition 7 (Number of Mutations). Let M_n be the expected number of mutations (compared to the ancestral phenotype) accumulated in a lineage at the end of the *n*th cycle, with dilution $\delta^* = e^{-(b-d)T}$, birth rate of cells *b*, death rate *d*, and mutation rate θ .

$$M_n = \theta s_n^* L_n,\tag{9}$$

with L_n the average length of the coalescent tree at the end of the *n*th cycle of an extant lineage started by one cell at t = 0:

$$L_n = F(nT) \int_0^{nT} \frac{dt}{F(t)},\tag{10}$$



Fig. 12. Expected number of mutations through experimental cycles. The expected number of mutations increases with the number of experimental cycles. The initial number of cells is $c = K\delta$. For one cycle (n = 1), smaller dilution factors give better results (the curve $\delta = 0.5$ is lower than the curve for $\delta = 0.001$). For more cycles, however, a larger dilution factor yields more mutations.



Fig. 13. Expected number of mutations for different dilution factors. The initial number of cells is $c = K\delta$. For a single cycle (n = 1), smaller dilution factors always yield more mutations. However, if more than one cycle is performed $(n \ge 2)$, there is a non-zero optimal dilution factor that maximises the expected number of mutations found in the population.



Fig. 14. Mutation-optimising bottleneck size as a function of the number of cycles. The initial number of cells is $c = K\delta$. The bottleneck size that optimises the expected number of mutations is increasing with the number of cycles performed.

with F the inverse tail distribution of the associated CPP. More specifically, when using the expression of F from Proposition 6:

$$L_n = \left(1 + \frac{b}{r}n(e^{rT} - 1)\right) \sum_{k=0}^n \frac{rT - \log\left(\frac{k(e^{rT} - 1) + \frac{r}{b}}{(k+1)(e^{rT} - 1) + \frac{r}{b}}\right)}{bk(e^{rT} - 1) - d}$$
(11)

Proposition 7 shows that the expected number of mutations M_n accumulated in a lineage is, naturally, proportional to the mutation rate θ . The expected number of mutations is also proportional to the expected coalescent tree length of a single extant lineage L_n (weighted by the proportion of lineages that actually survive s_n). This expected

coalescent length is parameterised by the birth and death rates of the cells, but also the duration of the growth phase *T*. The proportion of lineages that actually survives plays an important role here. Consider a massively parallel experiment with the goal of screening as many mutations as possible. To find the optimal protocol, one must not only seek to optimise the number of mutations that accumulate on one lineage θL_n but also discount this number by the proportion of them that are extinct s_n . Otherwise, one might waste a lot of resources conducting an experiment consisting mainly in serial transfer of droplets devoid of cells altogether. Finally, since the number of accumulated mutations is proportional to the number of initial lineages, increasing this number by producing more numerous or bigger droplets (with the same initial occupancy) proportionally increases the number of mutations.

The expected number of mutations increases indefinitely with the number of cycles (Fig. 12). However, the rate of increase is tied to the dilution bottleneck and tends to slow down when the number of cycles increases. Note that for a small number of cycles, the expected number of mutations increases with a higher dilution: one cycle with a dilution by two yields less diversity than one cycle with a dilution by one hundred. However, if ten cycles are performed, a dilution by two yields more diversity. This illustrates a trade-off between a harsh bottleneck, which allows long cycles and thus potentially many mutations but leads to the loss of most extant mutations (due to founder effects), and a softer bottleneck that allows for fewer mutations to accumulate during each cycle but compounds more because fewer mutations are lost.

Fig. 13 clarifies the link between the expected number of mutations and the dilution bottleneck. Note that smaller bottleneck sizes are compensated by longer cycles because the experiment is supposed to be performed in optimal conditions ($\delta = e^{-rT}$). If there is only one cycle (n = 1), the maximal expected number of mutations is reached when the dilution bottleneck is vanishingly small ($\delta \rightarrow 0$) and the cycle length adequately long ($T \rightarrow \infty$). However, if there is more than one cycle (n > 1), the expected number of mutations reaches a maximum value between $\delta = 0$ and $\delta = 1$. This maximum-diversity dilution bottleneck value increases with the number of cycles (Fig. 14). Thus, the dilution bottleneck should be adjusted to the expected duration of the experiment in terms of cycle number to maximise accumulation of neutral mutations.

Overall, the expected number of neutral mutations accumulated by the population increases through time and can be optimised by appropriately choosing a bottleneck size that optimises the trade-off between accumulating new mutations and not losing old ones.

Note that L_n behaves like $Tn \ln(n)$ (Lambert (2009), Theorem 2.4); thus, the expected neutral diversity after *n* cycles is equivalent to $\theta s_n Tn \ln(n)$.

Thanks to (7), and because $\delta = \delta^* = e^{-rT}$, we obtain:

$$M_n \sim_{n \to +\infty} \frac{\theta T r}{b(1 - e^{-rT})} \ln(n).$$

The mere number of mutations contains little information about the diversity within a droplet. Indeed, some of those mutations could be born by a single individual, while others might be shared by the whole population. The next section addresses this problem by exploring the mutation frequency spectrum.

3.5. Mutation frequency spectrum

A more precise assessment of the neutral diversity structure involves distinguishing between rare mutations (carried by few individuals) and frequent mutations (widespread within the population). The mutation frequency spectrum presents the proportion of mutations that are carried by a given number of individuals. The expected mutation frequency spectrum of a CPP can be deduced from the law of node depths H (Lambert (2009), Theorem 2.2). Indeed, the number of mutations carried by i individuals is proportional to the length of the coalescent tree subtending i leaves (Fig. 15). As a result:

1



Fig. 15. Finding the mutation frequency spectrum. The mutations shared by 3 individuals are the ones that arose within the orange region only. This region is delimited by $\max(H_{i+1}, H_{i+2}) < t < \min(H_i, H_{i+3})$.

Proposition 8 (Mutation Frequency Spectrum). Consider the CPP T_n , with overlaying mutations following a Poisson point process with intensity θ .

Let M_n^f be the expected number of mutations fixed in the population: that is, mutations shared by all individuals. Then:

$$M_n^f = \theta s_n \int_0^{nT} \frac{F(s)}{F(nT)} - \frac{F(nT)}{F(s)} ds$$
(12)

Let a_u be the expected frequency of mutations that are shared by u > 0 individuals in the limit of a large sample of the population.

$$a_{u} = \theta \int_{0}^{nT} \left(1 - \frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{u-1} \left(\frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{2} dx$$
(13)

As is expected for neutral diversity, Proposition 8 shows that the mutation frequency spectrum is proportional to the mutation rate θ , meaning that an increasing proportion of individuals carry mutations if the rate increases, but that this does not change the relative frequency of the size of groups carrying a given mutation.

Fig. 16 shows the mutation frequency spectra for one and for a hundred cycles, and for three different dilution factors. Note that for a single cycle, harsher bottlenecks (i.e., smaller δ and correspondingly longer cycle duration *T*) increase the tail of the distribution (i.e., there are more mutations that are shared by many individuals). This effect of δ and *T* is not as simple when considering several cycles. For *n* = 100, the distribution is more heavy-tailed when the bottlenecks are soft (δ = 0.5) than when they are harsh (δ = 0.001).

Fig. 17 shows the power law tails of mutation frequency spectra.

When the coalescent tree is a Kingman coalescent, corresponding to a long-lived population with approximately constant size, the mutation frequency spectrum has power law with exponent -1 (harmonic spectrum, Ewens' sampling formula, Ewens (1972)). In our setting, this happens when δ is close to 1 (constant population size) and *nT* is large (long-lived population). For a fixed growth rate *r*, because $\delta e^{rT} = 1$, this means that $(1 - \delta)$ is small but) $n(1 - \delta)$ is large, as in the yellow region of the heat map of Fig. 17.

When the coalescent tree is a Yule tree, corresponding to full, unbounded growth, the mutation frequency spectrum has power law with exponent -2 (Lambert, 2009; Dinh et al., 2020; Durrett, 2013). This is what happens for small δ , regardless of *n*, as in the turquoise region of Fig. 17. Indeed, when δ is small, *T* is large (for fixed *r*), so that all derived mutations occurred since the last bottleneck.



Fig. 16. Mutation frequency spectrum. Give the frequency of mutations carried by a given number of individuals in a large sample of the population. *Top:* After n = 1 cycle. *Bottom:* After n = 100 cycles.

A third case stands out in our setting when *n* is not too large and δ is sufficiently close to 1 that very few births/coalescences occur over the time interval of length *nT*, which happens when $n(1 - \delta) = O(1)$. In this case, conditioning the CPP to have coalescences smaller than *nT* (an event of vanishing probability) yields a CPP with uniform node depths, which gives rise to a power law mutation frequency spectrum with exponent -3, as in the deep blue region of Fig. 17.

The information entropy of the mutation frequency spectrum can be used to systematically explore the effect of δ on the shape of the distribution. Fig. 18 shows that if more than one cycle is performed, there is a value of δ that is expected to optimise the information entropy of the mutation frequency spectrum. This value is different from the value that optimises the number of mutations (Fig. 13). Thus, there is a trade-off between accumulating many mutations and having a diverse mutation frequency spectrum. The decision to fix δ in order to optimise one or the other depends on the goal of the experiment.

To sum up, higher dilution factor or longer duration of droplet growth cycles result in longer trees and increased diversity, even though the population may risk going extinct. Extinction of the population marks the "death" of the culture and is the eventual fate of a serial transfer experiment in the limit of many cycles. In a nested populations' design, the cultures can also "reproduce" and replace the extinct ones. This has far-reaching consequences for the genealogy of the cells, as shown in the next section.

4. Diversity in dividing droplets

Nested populations' design differs from simple serial transfer in parallel cultures by the opportunity for the cultures (droplets, tubes or other compartments) to be subject to birth and death themselves. At each cycle, some cultures may be removed from the experiment, while others can be duplicated, usually by dispatching samples of the original droplet in several new fresh medium compartments (rather than one in regular parallel serial transfer experiments).

This section focuses on the consequences of imposing a droplet-level birth-death process on the neutral diversity. To this end, consider the simple scenario (depicted in Fig. 19) of a pair of droplets that share a



Fig. 17. Mutation frequency spectrum power law tail. Left: Slope of the regression α , such that $log(a_u) = \alpha log(u) + b$, for u > 500. For a branching process without bottleneck, the expected value is -2; it is -1 for a Moran process and -3 for a Coalescent Point Process with uniform node depths. White dashed lines are isolines -1, -2, -3. Right: three illustrative mutation frequency spectra with associated regression lines. The location of the three spectra in the parameter space is represented by colour-matching dots on the left heat-map.



Fig. 18. Shannon Entropy of the mutation frequency spectrum. Defined as $S = -\sum_{i} a_{i} \log(a_{i})$.

common "droplet ancestor" several cycles in the past. The two droplets differ by the initial sampling performed in their common ancestor and also by all new mutations accumulated since they became isolated.

In the following, the cells follow a super-critical linear birth-death process with parameters b-d = r > 0. The parameters of the population structure are supposed to be optimal in the sense of Section 2.1: each cycle has a duration $T^* = -r^{-1} \ln (cK^{-1})$, and each lineage has an independent probability of being sampled at a bottleneck of $\delta^* = e^{-rT^*}$. Consider that the droplet split happens at cycle *m* and the observation occurs at cycle m + n.

4.1. Survival probability

First, let us consider the probability that a lineage spawned by a single cell *n* cycles in the past is not extinct within both droplets. The key to establish this probability is to recognise that the lineage undergoes a bottleneck with survival probability δ at each cycle, except the cycle of the droplet split where each cell has a probability 2δ to survive. Indeed, two inoculation volumes are concurrently sampled from the ancestral droplet and dispatched into two offspring (Fig. 20). Thus:

Proposition 9 (Survival probability - Split droplet). Cells within droplets in serial transfer are modelled by a linear birth–death process with constant parameters *b* and *d*, subject to periodic bottlenecks δ every duration *T* for m + n cycles, except at cycle *m* where *k* droplets are produced (instead of one). Each new droplet is submitted independently to the serial transfer procedure for the remaining *n* cycles.

The probability $s_{k,m,n}$ that a lineage spawned by a single cell at the beginning of the first cycle is not extinct at the end of the (n + m)th cycle is given by:

$$s_{k,m,n} = 1 - h(Q_1 Q_{\delta}^{m-1} Q_{k\delta} Q_{\delta}^{n-1}, 0),$$
(14)

where the matrix Q_{ε} for $\varepsilon \in [0,1]$ and the function h are defined in Proposition 1.

Proposition 9 is similar in its conclusion to Proposition 1, which treated the case of a simple serial transfer. However, the expression is considerably less easy to handle, as the iteration does not simplify into a single matrix power.

4.2. Total diversity

As seen in Proposition 7, quantifying the total neutral diversity in an infinitely-many sites model is a matter of finding the total length of the coalescent tree (or forest) of the population. Note that the full coalescent tree in Fig. 21 can be decomposed into a stump, before the splitting of droplets, and a corolla: another set of CPP (the corolla) sampled in one or the other droplet lineage. As a result:

Proposition 10 (Total Diversity - Split droplet). Let $M_{k,m,n}$ be the expected number of mutations accumulated in a lineage at cycle n + m after the splitting of the initial droplet at cycle m into $k = 1, 2 \dots \left\lfloor \frac{1}{\delta} \right\rfloor$ droplets. Then:

$$\begin{split} M_{k,m,n} = & \theta L_{k,m,n} \\ = & \theta s_m F^{\dagger}(mT) \left[\int_0^{mT} \frac{1}{F^{\dagger}(s)} ds + \int_0^{nT} \frac{F(nT)}{F(s)} ds \right] \end{split}$$

where $L_{k,m,n}$ is the expected length of the coalescent tree of the population, F is the inverse tail distribution of the CPP with periodic bottlenecks, and $F^{\dagger} := 1 - k\delta s_n + k\delta s_n F$, is the inverse tail distribution of the same CPP submitted to sampling with probability $k\delta s_n$ at the present.

4.3. Private diversity

To assess the divergence between split droplets, one can compute the expected number of *private* mutations (i.e., mutations that are only found in a single of the k split droplets). This number is the sum of all mutations that occur in the droplets after the splitting time



Fig. 19. Droplet and cell nested coalescent trees. Left: coalescent tree of droplets. Right: coalescent tree of cells. An ancestral droplet lineage (black) is diluted into two offspring droplets (red, blue). At the time when the droplets are split (cycle *m*), the cell lineages within the ancestral droplets are assigned a colour (red or blue) that indicates the daughter droplet to which they are sent. Mutations appear along the genealogy of cells (crosses). Some mutations appear before the split (yellow, green, purple) and are found in both droplets (green, purple) if they are sampled by both droplet lineages, or within a single droplet (yellow) if they are segregated by the dilution. Other mutations appear after the split (light blue, orange) and are only found in one of the droplet lineages.

1



Fig. 20. Droplet Splitting Process. When a droplet is split into *k* droplets, individual cells are independently selected to be transferred to the next cycle (with probability δ for each new droplet) or be discarded (with probability $1 - k\delta$).

mT (i.e., mutations in the corolla), plus all the mutations that occur before the splitting time but in a lineage that only segregates in a single droplet (i.e., mutations in the stump). Since all the droplets are interchangeable, this value is identical for the *k* droplets. Overall, this number is proportional to the red (or blue) part of the CPP in Fig. 19.

Proposition 11 (Private Mutations - Split Droplet). Let a single droplet be split into $k = 1 \dots \left| \frac{1}{\delta} \right|$ at cycle *m*. Let $M'_{k,m,n}$ be the expected number of

mutations that are private to any of the k droplets when observed at cycle n + m.

$$\begin{aligned} \mathcal{A}'_{k,m,n} &= \underbrace{S_{k,m,n}}_{stump} + \underbrace{C_{k,m,n}}_{corolla} \\ &= \theta s_m F^{\dagger}(mT) \left[\int_0^{mT} \frac{kds}{F^{\dagger}(s)(1+(k-1)F^{\dagger}(s))} + \int_0^{nT} \frac{F(nT)}{F(s)} ds \right], \end{aligned}$$

where *F* is the inverse tail distribution of the CPP with periodic bottlenecks, and F^{\dagger} is the inverse tail distribution of the same CPP submitted to sampling with probability $k\delta s_n$ at the present.

Fig. 22 shows the expected proportion of private mutations $M'_{k,m,n}/M_{k,m,n}$ in split droplets as a function of *m*, the number of cycles before splitting. If *m* is low, there are no shared mutations among droplets and the ratio is close to 1. If more cycles occur before the split, the proportion of private mutations decreases, tending towards 0 at a logarithmic rate.

Now, the main purpose of droplet splitting is to select and duplicate a phenotype of interest. The last section explores, in the context of artificial selection, the advantage offered by a droplet-splitting process over the simple screening of parallel cultures in serial transfer.

5. Artificial selection of droplets

A practical application of a device that would allow the manipulation of small cultures of microbial organisms would be the artificial selection of phenotypes of interest. Suppose that a given phenotype of interest is reached after the accumulation of Θ mutations and that it is possible to detect the number of mutations fixed so far, by sequencing or direct observation of the cultures.

To formalise, let $D \in \mathbb{N}^*$ be the number of droplets. Each droplet *i* is assigned a number $e_i = 1, 2, ..., \Theta$, corresponding to the number of fixed mutations. Suppose that the time for a droplet to switch from $e_i = j$ to $e_i = j + 1$ is exponentially distributed with parameter $\alpha = \frac{\rho}{ND}$, where ρ is the mutation rate (that could be deduced from Proposition 8), scaled



Fig. 21. Finding the total diversity at cycle n + m of droplets split at cycle m. All the lineages (triangles) spawned from the cells dispatched in one of the k = 2 droplets (here red and blue) have the same expected length L_n . The bottom part (or stump) of the tree (black) result from the sampling of a coalescent point process stopped at mT, with probability $\pi_{k,m,n}$, the probability that an extant lineage at cycle m will be sampled in one of the two droplets and survive until cycle n + m.



Fig. 22. Expected proportion of private mutations in split droplets. Droplets are grown for *m* cycles, then split into k = 2 new droplets and grown for *n* new cycles. $b = 1, d = 0, \delta = e^{-rT}$.

by the number of droplets D and the number of cycles N. We assume that the D droplets are in state 0 at time 0. We also assume that the only possible transition is to accumulate a new mutation: that is, no reversion is possible, as illustrated in Fig. 23.

To assess the advantage of droplet splitting, consider two scenarios, illustrated in Fig. 24:

- (1) Without droplet selection *D* droplet lineages are started in state 0 at *t* = 0 and undergo serial transfer independently of each other.
- (2) With droplet selection *D* droplet lineages are started in state 0 at *t* = 0. Once a fixed mutant is detected in a lineage, all the other droplets are killed and this lineage is split in *D* new lineages.

Let Γ (respectively, Γ^*) be the random variable encoding the first time for a lineage to get to the state $\Theta \in \mathbb{N}$ in the scenario without droplet selection (respectively with droplet selection). To compare them, consider their respective cumulative distribution functions:

Proposition 12 (*Cumulative Distribution Functions*). The cumulative distribution function of Γ^* is:

$$\mathbb{P}(\Gamma^* \le x) = 1 - e^{-x\rho} \left(\sum_{u=0}^{\Theta-1} \frac{(\rho x)^u}{u!} \right)$$
(15)

The cumulative distribution function of Γ is:

$$\mathbb{P}(\Gamma \le x) = 1 - e^{-\rho x} \left(\sum_{u=0}^{\Theta-1} \frac{(\rho x)^u}{D^u u!}\right)^D$$
(16)

The selective regime is always faster than the serial transfer regime, in the sense that Γ^* is stochastically smaller than Γ

$$\mathbb{P}(\Gamma^* \le x) \ge \mathbb{P}(\Gamma \le x). \tag{17}$$

In addition as the number of mutational steps $\Theta \to \infty$, the ratio Γ^*/Γ converges to 1/D almost surely.

Proposition 12 shows that droplet level selection – that is, the process of splitting a droplet in which an intermediate mutation was fixed – leads to reducing the time to reach the Θ th mutation. Fig. 25 shows the shape of the cumulative probability function for both regimes, illustrating this advantage. This constitutes a simple use case for a device that allows the automated high-throughput manipulation of numerous cultures, such as the digital millifluidic analysers (Baraban et al., 2011; Boitard et al., 2015; Cottinet et al., 2016).

Note that this result is obtained by assuming that the detection of mutation is cost-free and error-free. A more advanced model of this system should tackle the problem of imperfect detection.

6. Discussion

This manuscript has laid the foundation for a theoretical understanding of the evolution of neutral diversity in massively parallel microbial evolution experiments with population splits. It was heavily inspired by ongoing engineering efforts to bring experimental evolution to digital millifluidics (Cottinet, 2013; Boitard et al., 2015; Dupin, 2018; Doulcier, 2019; Ardré et al., 2022).

In experimental microbiology, one desirable feature is to maximise the number of mutations accumulated within the cultures: for instance, to screen phenotypes of interest. The result presented above showed that, in an optimal growth setting, where cells are growing with a constant birth and death rate, without density-dependence or competition, the population should be submitted to cycles whose duration is tailored to compensate the bottleneck imposed at each serial transfer. The choice of the bottleneck should be made according to the expected duration of the experiment: to optimise the expected number of mutations, small bottlenecks (i.e., killing most of the lineages) should be used when the number of cycles is small, while larger bottlenecks (i.e., lower dilution factor) should be used for long-term experiments. Additionally, the expected number of mutations increases linearly with increasing droplet volume and with increasing number of droplets, which is a matter of technological progress as automation and larger droplet sizes are under consideration (Dupin, 2018; Postek et al., 2022). The mutation rate also increases linearly the number of expected mutations and can be manipulated by choosing mutator lineages or adding mutating chemicals to the culture broth. However, the potentially deleterious effects of this method might prevent using it in practical cases.

In long-term evolution experiments (Kawecki et al., 2012; Van den Bergh et al., 2018), serial transfer is imposed by the need to replenish nutrients available to the cells. It is possible to build devices ensuring that a continuous flow of nutrient washes over the culture (for large volumes, see chemostats, or morbidostats (Toprak et al., 2013); in microfluidics, see mother machines (Potvin-Trottier et al., 2018). However, these methods are usually more prone to contamination. In contrast, periodically diluting the culture in a fresh medium is simple and robust.



Fig. 24. Propagation of mutations. Without droplet selection, all the lineages accumulate mutations independently. With droplet selection, once a mutation fixation is first detected, the droplet is split into *D* lineages.



Fig. 25. Cumulative probability distribution of the time to accumulate Θ mutations. With droplet splitting, accumulation of mutation is faster. D = 100.

The nested populations' design differs from traditional serial transfer of parallel cultures because it allows a birth–death process at the level of droplets. Serial transfer is pervasive in experimental evolution (Kawecki et al., 2012) and has received extensive theoretical treatment. This article focused solely on neutral diversity, considering that mutations have no effect on the demography of the cells. However, the effect of beneficial mutations has been studied in serial transfer settings: in particular, the probability of losing a beneficial mutation due to repeated bottlenecks (Wahl and Krakauer, 2000; Wahl et al., 2002; Wahl and Gerrish, 2001; Wahl and Zhu, 2015) and the effect of bottlenecks on the evolutionary path when multiple beneficial mutations exist (Gamblin et al., 2023). These results should be extended to the nested populations' design in the future.

The "death" of a droplet in our framework can have different origins. First is the experimentalist actively discarding some droplets with undesired characteristics. Second, it can arise from the fact that some cultures are effectively empty due to high dilutions in the previous cycle and may be replaced in the next cycle by cells from a nonempty culture. Finally, it can also be a consequence of the experimental protocol. Milli- and microfluidics compartments are usually produced in large numbers, while measurements are performed on all compartments. The retrieval of all the compartments' contents might not be practically possible or even desirable when they are too numerous. In all these cases, some droplets will not be used to generate the population of the next cycle, and some of them will have to be split if the droplet population is to keep a constant size.

The use of a non-saturating population dynamics in this manuscript is a simplification that should be carefully taken into account when transposing the results of this work to the design of experiments. Nonetheless, if the cycle duration is short enough that the population is stopped during the exponential phase, the heuristics developed in this manuscript should hold. There are, however, two phenomena that were not modelled here and that will probably muddy the neutral pattern that was described. First is the absence of mutations affecting birth and death rates. If most point mutations can be safely considered neutral, some rare mutations can affect the ability of the cells to reproduce. If the mutations are beneficial, they will increase in proportion within the population and will change the relative frequency of all neutral mutations, by favouring the ones carried by the same strand of DNA (a phenomenon known as "hitch-hiking", Fay and Wu 2000). Second, horizontal gene transfer might allow the uncoupling of the mutation transmission from the genealogy (Dutta and Pan, 2002), muddying the pattern even further. In general, complete models should combine neutral and non-neutral approaches to describe microbial communities (Ofițeru et al., 2010; Dumbrell et al., 2010).

The nested populations' design also differs from trait groups' (Wilson, 1975) or transient compartments' (Blokhuis et al., 2018) population structure because migration between compartments is prevented. As a consequence, it is possible to construct a non-ambiguous genealogy of the cultures. In practice, serial transfer design offers a natural way to implement the birth-death process by diluting some cultures into several new compartments (the droplet splitting) and discarding others.

Finally, this manuscript touches briefly on the problem of artificial selection using a nested populations' design. This was done by considering the accumulation of neutral mutations. A more complete model of artificial selection would, however, take into account interactions between individuals and potentially the selection of whole communities. Community-level selection has been the subject of both experimental (Swenson et al., 2000; Panke-Buisse et al., 2015) and theoretical inquiries (Mueller and Sachs, 2015; Arias-Sánchez et al., 2019; Xie et al., 2019; Doulcier et al., 2020).

Overall, the results presented in this manuscript should be considered as a way to build intuition about the experimental system while providing a null-model for diversity that could be compared to the actual patterns. Inevitable differences have to appear, but the point of comparison that is offered by neutral evolution will allow a better description of the observed diversity. Focusing on the part of the patterns that differ from this naive theoretical prediction will surely be fruitful — by showing that other mechanisms besides drift must be invoked. Table 1

Global Symbols reference.		
Symbol	Name	Reference
D	Number of droplets	
Т	Cycle duration	
n	Number of cycles	
Κ	Droplet carrying capacity	
с	Initial number of particles	
b, d, r	Particle birth rate, death rate and Malthusian parameter $(r := b - d)$	
5	Dilution factor (i.e., particle survival probability at a bottleneck)	
9	Particle mutation rate	
s _n	Probability that a lineage spawned by a single particle is not	Proposition 1
	extinct at the end of the <i>n</i> th cycle. (At time nT^- , just before the	
	<i>n</i> th dilution). $s_n = 1 - h_{Q^{n-1}R}(0)$	
s	Limit probability that a lineage spawned by a single particle is not	Proposition 2
	extinct after a large number of cycles. $s_{\infty} = \lim_{n \to \infty} s_n$.	
r *	Maximal cycle duration before reaching saturation. $T^* = r^{-1} \ln(cK^{-1})$	Proposition 3
*	Optimal dilution factor $\delta^* = e^{-rT}$	Proposition 4
s_n^*	Survival probability of a lineage spawned by a single cell before the	Proposition 5
	<i>n</i> th dilution in the optimal regime in which $\delta = e^{-rT}$.	
F	Inverse tail distribution of the CPP without bottlenecks	Proposition 6
7	Inverse tail distribution of the CPP with bottlenecks.	Eq. (8) and Proposition 6
	$F(t) = [\mathbb{P}(H > t)]^{-1}$	
'n	Coalescent tree of an extant lineage at the end of the <i>n</i> th cycle. It	Proposition 7
	is a Coalescent Point Process with inverse tail distribution F	
	stopped at the first branch length larger than <i>nT</i> .	
$N(\tau_n)$	number of leaves of the coalescent tree τ_n . Geometric random	Eqs. (18) and (19).
	variable with expected value $F(nT)$.	
n	Expected length of the coalescent tree τ_n . $L_n = F(nT) \int_0^{nT} F(x)^{-1} dx$	Proposition 7
M _n	Expected number of mutations in a lineage after n cycles.	Proposition 7
	$M_n = \theta s_n L_n$	
M_n^f	Expected number of fixed mutations in a lineage after n cycles.	Proposition 8
M_n^s	Expected number of segregating mutations in a lineage after n	Proposition 8
	cycles. $M_n^s = M_n - M_n^f$	
a _u	Expected frequency of mutations shared by $u > 0$ individuals after n	Proposition 8
	cycles.	

Table 2

Global Symbols reference (cont)

Symbol	Name	Reference
$S_{k,m,n}$	Survival probability at the end of the $(m + n)$ th cycle of a lineage spawned by a single cell in a	Proposition 9
	single droplet at the first cycle, that is was split	
	into k droplets at cycle m .	
$\tau_{k,m,n}$	Coalescent tree at the end of the $(m + n)$ th cycle of	Proposition 10
	a lineage spawned by a single cell in a single	
	droplet at the first cycle, that is was split into k	
	droplets at cycle m.	
$\pi_{k,n}$	Probability that a lineage extant at the end of	
	cycle m just before the droplet is split into k	
	droplets will be extant at cycle $m + n$. $\pi_{k,n} = k\delta s_n$.	
F^{\dagger}	Inverse tail distribution of the stump tree.	
	$F^{\dagger}(t) := F_{\pi_{k,n}}(t) = 1 - \pi_{k,n} + \pi_{k,n}F(t), t \in [0, mT]$	
$L_{k,m,n}$	Expected length of the coalescent tree $\tau_{k,m,n}$.	Proposition 10
$M_{k,m,n}$	Expected number of mutations accumulated at the	
	end of the $(m + n)$ th cycle of a lineage spawned by	
	a single cell in a single droplet at the first cycle,	
	that is was split into k droplets at cycle m .	
	$M = \theta s_{k,m,n} L_{k,m,n}$	
$M'_{k,m,n}$	Expected number of mutations that are only found	Proposition 11
	in a single droplet at the end of the $(m + n)$ th cycle	
	of a lineage spawned by a single cell in a single	
	droplet at the first cycle, that is was split into k	
	droplets at cycle m.	

CRediT authorship contribution statement

Guilhem Doulcier: Writing – review & editing, Writing – original draft, Visualization, Software, Formal analysis. **Amaury Lambert:** Writing – review & editing, Writing – original draft, Supervision, Formal analysis.

Table 3

Global Symbols reference (cont.)

Symbol	Name	Reference
α	Rate at which a lineage accumulate mutations $\alpha = \frac{\rho}{1-\alpha}$	Proposition 12
Θ	Number of mutations to accumulate	Proposition 12
Г	First time a lineage has accumulated Θ	Proposition 12
	mutations without droplet selection	
Γ^*	First time a lineage has accumulated Θ	Proposition 12
	mutations with droplet selection	

Data availability

No data was used for the research described in the article.

Acknowledgements

GD gratefully acknowledges the financial support of the Origines et Conditions d'Apparition de la Vie (OCAV) programme, PSL University (ANR-10-IDEX-001–02) during the elaboration of the manuscript and of the John Templeton Foundation, United States (#62220) during its revision. AL thanks the Center for Interdisciplinary Research in Biology (CIRB, Collège de France) for funding.

This work has been inspired by discussions with Paul B. Rainey and Silvia De Monte during the OCAV programme.

Reproducing the figures

The code used to generate all the figures in the manuscript uses python 3.11, matpolotlib 3.6.3, numpy 1.24.2 and scipy 1.24.2. It is available on the Zenodo repository under the DOI 10.5281/zenodo. 8087961.

Appendix

In the following sections we give the mathematical proofs for the results used in the main text.

Parameters range

Proof of Proposition 1 — **Survival Probability.** Let s_n be the survival probability at t = nT of a lineage started from a single cell at t = 0.

First cycle

Consider one cell at time t = 0. This cell follows a Linear Markov Branching Process $(Z_t)_{t \in [0,T)}$ with constant rates b, d until the dilution time T. We use two results from Athreya and Ney (1972). First, The branching process goes extinct $(Z_t = 0)$ with probability q(b, d, t) := $\mathbb{P}(Z_t = 0)$. Second, conditional on non-extinction Z_t follows a geometric distribution with parameter p(b, d, t), that is, $\mathbb{P}(Z_t = k | Z_t \neq 0) =$ $p(1 - p)^{k-1}$. The values of p and q are given by the following table:

		p(b,d,t)	q(b,d,t)
Subcritical cells	r < 0	$\frac{d-b}{d-be^{rt}}$	$\frac{d(1-e^{rt})}{d-be^{rt}}$
Critical cells	r = 0	$\frac{1}{1+bt}$	$\frac{bt}{1+bt}$
Supercritical cells	r > 0	$\frac{(b-d)e^{-rt}}{b-de^{-rt}}$	$\frac{d(1-e^{-rt})}{b-de^{-rt}}$

The probability generating function of Z_T is:

$$\begin{split} f_{Z_T}(s) &= \sum_{k\geq 0} \mathbb{P}(Z_T = k) s^k & \text{Definition of } f \\ &= \mathbb{P}(Z_T = 0) s^0 + \sum_{k\geq 1} \mathbb{P}(Z_T = k) s^k \\ &= \mathbb{P}(Z_T = 0) s^0 + \sum_{k\geq 1} \mathbb{P}(Z_T \neq 0) \\ &\times P(Z_T = k | Z_T \neq 0) s^k \\ &= q + \sum_{k\geq 1} (1-q) p (1-p)^{k-1} s^k & \text{Definition of } p, q \\ &= q + p s (1-q) \sum_{k\geq 0} ((1-p) s)^k \\ &= q + \frac{p s (1-q)}{1-(1-p) s} & \text{Geometric series, } (1-p) s < 1 \\ &= \frac{s (p-q)+q}{s (p-1)+1} \end{split}$$

As expected, $f_{Z_T}(1) = 1$ and $f_{Z_T}(0) = P(Z_T = 0) = q$.

Second cycle

Consider a cell at the end of the first cycle, just before the first dilution. Let B be the number of descendants of this cell at the end of the second cycle. There are two possibilities to consider:

- with probability (1δ) , the cell is discarded during the dilution at the beginning of the second cycle and B = 0,
- with probability δ , the cell is not discarded, and thus sees its descent grow until the end of the cycle following the same law as in the first circle, namely the law of Z_T .

Thus,

$$f_B = 1 - \delta + \delta f_{Z_T}.$$

We now introduce a notation for the rest of the proof: f_{Z_T} and f_B are linear fractional, a property that will be useful later when composing generating functions. We associate any linear-fractional function with a coefficient matrix M as follows:

We define Q_{ε} for all ε in [0,1) as:

$$Q_{\varepsilon} := \begin{bmatrix} p-1+\varepsilon(1-q) & 1-\varepsilon(1-q) \\ p-1 & 1 \end{bmatrix}.$$

Note that, f_{Z_T} is associated with the matrix Q_1 and f_B with the matrix Q_{δ} .

Subsequent cycles

Let X_n be the number of descendants of an ancestral cell at the end of the *n*th cycle. Let $(B_{n,i})_{n>0,i>0}$ be a collection of identical and independent random variables following the same law as *B*. Then the following recursion holds:

$$X_n = \sum_{i=1}^{X_{n-1}} B_{n,i}$$

Since *B* and X_{n-1} are independent, the generating function of X_n is given by the composition of the generating functions of X_{n-1} and *B*: $f_{X_n} = f_{X_{n-1}} \circ f_B$.

 $f_{X_n} = f_{X_{n-1}} \circ f_B$. By induction, the generating function of X_n is obtained by composing the generating function of x_1 with the n-1 times iteration of the generating function of B:

$$f_{X_n} = f_{X_1} \circ \underbrace{f_B \circ \dots \circ f_B}_{(n-1) \text{ times}}.$$

Now, recall that X_1 follows the same law as Z_t (associated with the matrix Q_1) and f_B is associated with the matrix Q_{δ} . From this observation and the composition rule outlined in the previous section, it results that:

$$f_{X_n} = h(Q_1 Q_{\delta}^{n-1}, \cdot)$$

It can be checked that this expression corresponds to *n* cycles, with n - 1 dilutions by a factor δ , since there is no dilution before the first cycle.

This leads to the survival probability of the lineage spawned by an ancestral cell, at the end of cycle *n*:

$$s_n = 1 - f_{X_n}(0) = 1 - h(Q_1 Q_{\delta}^{n-1}, 0)$$

Proof of Proposition 2 — Long Term Survival Probability. Let us look for the fixed points of f_R :

$$\begin{split} \mathcal{F}_B(s) &= s \Leftrightarrow (p-1+\delta(1-q))s+1-\delta(1-q) = s((p-1)s+1) \\ \Leftrightarrow (s-1)((p-1)s-\delta(1-q)-1) = 0 \\ \Leftrightarrow s &= \frac{\delta(1-q)-1}{p-1} \end{split} \qquad \text{or } s = 1 \end{split}$$

 $f_B(s)-s$ is a second order polynomial for *s* meaning that f_B has at most two fixed points. One of them is 1, in accordance with the definition of characteristic functions. The other is s^* , with:

 $s^* < 1 \Leftrightarrow p < \delta(1-q).$

Since f_B is continuous, $f_B(s) \ge s$ for all $s \in [0, \min(s^*, 1)]$ and $\min(s^*, 1)$ is a fixed point of f_B , the sequence $u_n = f_B(u_{n-1})$ with $u_0 = h(Q_1, 0) \in [0, 1]$ converges to $\min(s^*, 1)$, that is, to 1 if $p > \delta(1 - q)$ and to s^* otherwise.

- In Subcritical regime, $p < \delta(1-q) \Leftrightarrow e^{-rT} < \delta$, is always false since $e^{-rT} > 1$ and $\delta \le 1$.
- In **Critical regime**, $p < \delta(1-q) \Leftrightarrow \frac{1}{1+bT} < \delta\left(1-\frac{bT}{1+bT}\right) = \delta\frac{1}{1+bT}$, is always false since $\delta \le 1$.
- In Supercritical regime, $p < \delta(1-q) \Leftrightarrow e^{-rT} < \delta$,

Thus, $s^* > 1$ except if b > d and $\delta > e^{-rT}$. In this case,

$$s_{\infty} = 1 - h(Q_1, s^*)$$
$$= \frac{p - \delta(1 - q)}{\delta(p - 1)}$$
$$= \frac{r(\delta - e^{-rT})}{\delta b(1 - e^{-rT})}$$

Proof of Proposition 5 — **Critical survival Probability.** Let r = b-d, r > 0, we use the expression of p and q from the proof of Proposition 1 and let $\delta = e^{-rT}$. Then, the probability of extinction of the branching process is:

$$q = \frac{d(1-\delta)}{b-d\delta}$$

The parameter of the geometric distribution of the number of cells, conditional on non extinction, is:

 $p = \frac{r\delta}{b - d\delta}$

Thus, it is possible to rewrite Q_1 and Q_{δ} like this:

$$\begin{aligned} Q_1 &= \frac{1}{b - \delta d} \begin{bmatrix} \delta b - d & d(1 - \delta) \\ b(\delta - 1) & b - \delta d \end{bmatrix} := \frac{1}{b - \delta d} \overline{Q_1} \\ Q_\delta &= \frac{1}{b - \delta d} \begin{bmatrix} 2\delta b - \delta d - b & b(1 - \delta) \\ b(\delta - 1) & b - \delta d \end{bmatrix} := \frac{1}{b - \delta d} \overline{Q_\delta} \end{aligned}$$

By induction, we can show that:

$$s_{n}^{*} = 1 - h(Q_{1}Q_{\delta}^{n-1}, 0) = 1 - \frac{(Q_{1}Q_{\delta}^{n-1})_{01}}{(Q_{1}Q_{\delta}^{n-1})_{11}} = 1 - \frac{Q_{1}Q_{\delta}^{n-1}}{Q_{1}Q_{\delta}^{n-1}}$$
$$= \frac{b-d}{bn - \delta(b(n-1)+d)} = \frac{r}{bn(1-\delta) + \delta r} \quad \Box$$

Coalescent point processes

Proof of Proposition 6 — Coalescent tree. First, we recall the Proposition 7 from Lambert and Stadler (2013):

"Start with a CPP tree with inverse tail distribution *F*. Add extra mass extinctions with survival probabilities $e_1, \ldots e_k$ at times $T - s_1 > \cdots > T - s_k$ (where $s_1 > 0$ and $s_k < T$). Then, conditional on survival, the reconstructed tree of the phylogenetic tree obtained after the passage of mass extinctions is again a coalescent point process with inverse tail distribution F_{ϵ} given by

$$\begin{aligned} F_{\epsilon}(t) &= \epsilon_1 \dots \epsilon_m F(t) + \sum_{j=1}^m (1 - \epsilon_j) \epsilon_1 \dots \epsilon_{j-1} F(s_j), \\ t &\in [s_m, s_{m+1}], m \in \{0, 1, \dots, k\}, \end{aligned}$$

where $s_0 := 0$ and $s_{k+1} = T$ (empty sum is zero, empty product is 1)".

Consider the CPP of the population just before the *n*th cycle of dilution. The population experienced n - 1 bottlenecks at times $t_1 = (n-1)T, \dots, t_{n-2} = 2T, t_{n-1} = T$.

Let \tilde{F} be the scale function of the CPP without bottlenecks. Using Proposition 7 in Lambert and Stadler (2013) with $\epsilon_i = \delta$ and $s_j = jT$, it is possible to write F as:

$$F(t) = \delta^k \tilde{F}(t) + (1 - \delta) \sum_{j=1}^k \delta^{j-1} \tilde{F}(jT), \quad t \in [kT, (k+1)T], k \in \{0, 1, \dots, n\},$$

If the sampling at the last cycle is taken into account, the scale function becomes \overline{F} :

 $\overline{F}(t) = 1 - \delta + \delta F(t)$

The expression of \tilde{F} is known for the most common birth–death processes:

	Parameters	Scale function
Pure Birth	0 = d < b	$\tilde{F}(t) = e^{bt}$
Non-critical	$0 < d \neq b > 0$	$\tilde{F}(t) = \frac{b}{r}(e^{rt} - 1) + 1$
Critical	0 < d = b	$\tilde{F}(t) = bt + 1$

Non-critical case Let $r := b - d \neq 0, b > 0, d \ge 0, k \in \{0, ..., n\}, s < T$: $F(kT+s) = \delta^k \left[1 - \frac{b}{r} + e^{r(kT+s)} \frac{b}{r} \right] + (1-\delta) \sum_{i=1}^k \delta^{j-1} \left[e^{rjT} \frac{b}{r} + 1 - \frac{b}{r} \right]$ $=\delta^k\left(1-\frac{b}{r}\right)+\frac{b}{r}e^{rs}(\delta e^{rT})^k$ $+ (1-\delta) \left[\left(1 - \frac{b}{r}\right) \sum_{i=0}^{k-1} \delta^{j} + \frac{b}{r} e^{rT} \sum_{i=0}^{k-1} (\delta e^{rT})^{j} \right]$ $= \delta^k \left(1 - \frac{b}{r} \right) + \frac{b}{r} e^{rs} (\delta e^{rT})^k + \left(1 - \frac{b}{r} \right) (1 - \delta^k)$ + $\frac{b(1-\delta)}{r}e^{rT}\sum_{k=1}^{k-1}(\delta e^{rT})^{j}$ (Geometric series, $\delta \neq 1$) $= 1 + \frac{b}{r} \left[e^{rs} (\delta e^{rT})^k - 1 + (1 - \delta) e^{rT} \sum_{i=0}^{k-1} (\delta e^{rT})^i \right]$ If $\delta \neq \delta^*$ then, $F(kT+s) = 1 + \frac{b}{r} \left[e^{rs} (\delta e^{rT})^k - 1 + (1-\delta)e^{rT} \frac{1 - (\delta e^{rT})^k}{1 - \delta e^{rT}} \right]$ (Geometric series, $\delta^* e^{rT} \neq 1$) Otherwise, if $\delta = \delta^* = e^{-rT}$, $F(kT+s) = 1 + \frac{b}{r} \left(e^{rs} - 1 + k(e^{rT} - 1) \right)$ Critical case $b = d \neq 0$ $F(kT + s) = \delta^{k}(b(kT + s) + 1) + (1 - \delta)\sum_{k=1}^{n} \delta^{j-1}(bjT + 1)$ $= \delta^k (b(kT+s)+1) + (1-\delta) \sum_{j=1}^{k-1} \delta^j (b(j+1)T+1)$ $= \delta^{k}(b(kT+s)+1) + (1-\delta) \left[(1+bT) \sum_{i=0}^{k-1} \delta^{j} + bT \sum_{i=0}^{k-1} j \delta^{j} \right]$ $= \delta^{k} (b(kT + s) + 1) + (1 - \delta^{k})(1 + bT)$ + $(1 - \delta)bT \sum_{j=1}^{k-1} j\delta^j$ (Geometric series, $\delta \neq 1$) $= \delta^k b(T(k-1) + s) + 1 + bT + (1-\delta)bT \sum_{k=1}^{k-1} j\delta^k$ $= 1 + \delta^k b(T(k-1) + s) + bT \left[1 + (1-\delta) \sum_{j=0}^{k-1} j\delta^j \right]$ $=1+b\frac{T-s\delta^{k+1}-(T-s)\delta^k}{1-\delta}$ $(\delta \neq 1)$

Proof of Proposition 7 — Length of the Coalescent Tree. Let M_n be the expected number of mutations within the serial transfer experiment at the end of the *n*th cycle. Let also *T* be the duration of the growth phase and $\delta = \delta^*$ be the dilution factor. Suppose that the first cycle is seeded with a single ancestral cell. The initial lineage survives to cycle *n* with probability s_n^* (since $\delta = \delta^*$, see Proposition 5).

Each extant lineage after *n* cycles spawns an independent coalescent tree with expected length L_n . Moreover, mutations are accumulated following a Poisson point process on the tree with intensity θ . Hence, the expected number of mutations accumulated on one tree is θL_n . Thus,

$$M_n = \theta s_n^* L_n.$$

Let us now compute L_n .

Let $\{H_{ij}, (i, j) \in \mathbb{N}^2\}$ be a set of i.i.d. random variables following the same law as *H*, defined by its inverse tail distribution $F(t) = \frac{1}{\mathbb{P}(H > t)}$.

Let τ_n be a CPP with branches *H* stopped at *nT*. Let $N(\tau_n)$ be the number of leaves of the random tree τ_n . The length of the tree $\mathcal{L}(\tau_n)$ is the random variable:

$$\mathcal{L}(\tau_n) = nT + \sum_{j=1}^{N(\tau_n)-1} H_j,$$

where nT is the length of the spine and H_j are the length of the other $N(\tau_n) - 1$ branches.

Its expected value is:

$$L_n = \mathbb{E}(\mathcal{L}(\tau_n)) = nT + (\mathbb{E}(N(\tau_n)) - 1)\mathbb{E}(H|H < nT)$$

Number of leaves: $N(\tau_n)$ is a geometric random variable:

$$\mathbb{P}(N(\tau_n) = k) = P(H > nT)P(H \le nT)^{k-1} = \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)}\right)^{k-1}$$
(18)

Indeed, if *k* is the index of the first H_i such that $H_i > nT$, there are k - 1 branches, plus the spine, for a total of *k* leaves.

Thus, the expected number of leaves of τ_n , is:

$$\mathbb{E}(N(\tau_n)) = F(nT).$$
⁽¹⁹⁾

Note that

$$\mathbb{E}(N(\tau_n) - 1) = F(nT) - 1 = \frac{1 - \frac{1}{F(nT)}}{\frac{1}{F(nT)}} = \frac{\mathbb{P}(H < nT)}{\mathbb{P}(H > nT)}$$

so we get

$$L_n = nT + \frac{\mathbb{P}(H < nT)}{\mathbb{P}(H > nT)} \mathbb{E}(H | H < nT)$$

Length of branches:

We recall that for a positive r.v. X, $\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X > x) dx$. We can rescale the tail-distribution to take into account the conditioning:

$$\mathbb{P}(H > x | H < y) = \begin{cases} 0 & \text{if } x > y, \\ \frac{\mathbb{P}(H > x) - \mathbb{P}(H > y)}{\mathbb{P}(H < y)} & \text{otherwise.} \end{cases}$$
(20)

Thus,

$$\mathbb{E}(H|H < nT) = \int_0^{nT} \mathbb{P}(H > x|H < nT)dx$$

= $\int_0^{nT} \frac{\mathbb{P}(H > x) - \mathbb{P}(H > nT)}{\mathbb{P}(H < nT)}dx$
= $\frac{1}{\mathbb{P}(H < nT)} \left[\int_0^{nT} \frac{dx}{F(x)} - \frac{nT}{F(nT)} \right]$
(By definition, $\mathbb{P}(H > x) =: \frac{1}{F(x)}$),

so that

$$L_n = nT + \frac{\mathbb{P}(H < nT)}{\mathbb{P}(H > nT)} \mathbb{E}(H | H < nT)$$

= $nT + F(nT) \left[\int_0^{nT} \frac{dx}{F(x)} - \frac{nT}{F(nT)} \right]$
= $F(nT) \int_0^{nT} \frac{dx}{F(x)}$ (21)

Moreover, $\forall k \in \mathbb{N}$ and $\forall s \in \mathbb{R}, s < T$:

$$F(kT + s) = 1 + \frac{b}{c}(k(e^{rT} - 1) + e^{rs} - 1),$$

so that

$$\int_{kT}^{(k+1)T} \frac{dx}{F(x)} = \int_0^T \frac{1}{1 + \frac{b}{r}(k(e^{rT} - 1) + e^{rs} - 1)} ds$$
$$= \frac{rT - \log\left(\frac{k(e^{rT} - 1) + \frac{r}{b}}{(k+1)(e^{rT} - 1) + \frac{r}{b}}\right)}{bk(e^{rT} - 1) - d}$$

Finally, we get:

$$L_n = \left(1 + \frac{b}{r}n(e^{rT} - 1)\right)\sum_{k=0}^n \frac{rT - \log\left(\frac{k(e^{rT} - 1) + \frac{r}{b}}{(k+1)(e^{rT} - 1) + \frac{r}{b}}\right)}{bk(e^{rT} - 1) - d} \quad \Box$$
(22)

Proof of Proposition 8 — **Mutation Frequency Spectrum.** We now turn to a more detailed account of the mutation distribution.

Fixed and segregating mutations

Mutations compared to the ancestral type are either fixed (shared by all individuals) or segregating (not shared by all individuals). Thus, we have:

$$M_n^f + M_n^s = M_n$$

Fixed mutations are the ones found between the root of the coalescent tree (t = nT) and the first branching. Let *Y* be the random variable encoding this length. We have Y = nT - X with *X* the random variable encoding the height of the first branching. Thus,

$$M_n^f = \theta s_n \mathbb{E}(Y) = \theta s_n (nT - \mathbb{E}(X))$$

where θ is the mutation rate, s_n is the probability that the lineage spawned by one of the ancestral cells is still extant at the end of the *n*th cycle.

Now we compute $\mathbb{E}(X)$. Let $H_1...H_{N(r_n)-1}$ be the branch lengths of the CPP stopped at *nT*. The variables (H_i) are independent and identically distributed as H, conditioned to be smaller than *nT*. X is the maximum of those values:

$$X := \max \left\{ H_1, H_2 ... H_{N(\tau_n) - 1} \right\}.$$

Note that the first branching cannot be higher than the length of the spine. Thus, $\mathbb{P}(X > nT) = 0$. Otherwise:

$$\begin{split} \mathbb{P}(X > s) &= 1 - P(X < s) \\ &= 1 - \sum_{k=1}^{\infty} \mathbb{P}(N(\tau_n) = k) \mathbb{P}(X < s | N(\tau_n) = k) \\ & \text{(Total probability)} \\ &= 1 - \sum_{k=1}^{\infty} \mathbb{P}(N(\tau_n) = k) \mathbb{P}(H < s | H < nT)^{k-1} \\ & \text{(Xis the max of } (H_i) \text{ i.i.d.}) \\ &= 1 - \sum_{k=1}^{\infty} \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)} \right)^{k-1} \left(1 - \frac{\frac{1}{F(s)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{k-1} \\ & \text{(Eqs. (18) and (20)).} \\ &= 1 - \sum_{k=1}^{\infty} \frac{1}{F(nT)} \left(1 - \frac{1}{F(nT)} \right)^{k-1} \left(\frac{1 - \frac{1}{F(s)}}{1 - \frac{1}{F(nT)}} \right)^{k-1} \\ &= 1 - \frac{1}{F(nT)} \sum_{k=0}^{\infty} \left(1 - \frac{1}{F(s)} \right)^{k} \\ &= 1 - \frac{1}{F(nT)} \frac{1}{1 - 1 + \frac{1}{F(s)}} \\ & \text{(Geometric series)} \\ &= 1 - \frac{F(s)}{F(nT)} \end{split}$$

Again, we recall that for a positive r.v. *X*, $\mathbb{E}(X) = \int_0^{+\infty} \mathbb{P}(X > x) dx$. Thus,

$$\mathbb{E}(X) = \int_0^{nT} 1 - \frac{F(s)}{F(nT)} ds = nT - \int_0^{nT} \frac{F(s)}{F(nT)} ds.$$

Thus, the expected number of fixed mutations is:

$$M_n^f = \theta s_n [nT - \mathbb{E}(X)] = \theta s_n \int_0^{nT} \frac{F(s)}{F(nT)} ds$$

Finally, the expected number of segregating mutations within the droplet is the total number of mutations (M_n) minus the number of fixed mutations:

$$M_n^s = M_n - M_n^f$$

= $\theta s_n \int_0^{nT} \frac{F(nT)}{F(s)} - \frac{F(s)}{F(nT)} ds$

Full spectrum For a sample of v individuals in a CPP stopped at time nT, Theorem 2.2 in Lambert (2009) gives the expression of $\mathbb{E}(S_v(u))$ the expected number of mutant sites that are carried by exactly $1 \le u \le v - 1$ individuals.

$$\mathbb{E}(S_{\nu}(u)) = \theta \int_{0}^{nT} \left(1 - \frac{1}{W(x)}\right)^{u-1} \left(\frac{\nu - u - 1}{W(x)^{2}} + \frac{2}{W(x)}\right) dx,$$
(23)

with *W* the inverse tail distribution of the branch length in the stopped CPP. *W* is expressed, for x < nT, as a rescaling of *F*:

$$W(x) = \frac{1}{\mathbb{P}(H > x | H < nT)} = \frac{1 - \frac{1}{F(nT)}}{\frac{1}{F(x)} - \frac{1}{F(nT)}}.$$
(24)

The expected value of $S_v(u)$ converges towards a limit (Lambert, 2009) when the sample size increases, provided $\mathbb{E}(H|H < nT) < \infty$, which is the case here, since the CPP is stopped at nT.

$$\lim_{v \to \infty} v^{-1} \mathbb{E}(S_v(u)) = \theta \int_0^{nT} \left(1 - \frac{1}{W(x)}\right)^{u-1} \frac{1}{W(x)^2} dx$$
(25)

Thus,

$$a_{u} = \lim_{v \to \infty} v^{-1} \mathbb{E}(S_{v}(u)) = \theta \int_{0}^{nT} \left(1 - \frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{u-1} \left(\frac{\frac{1}{F(x)} - \frac{1}{F(nT)}}{1 - \frac{1}{F(nT)}} \right)^{2} dx \quad \Box$$
(26)

Split droplets

Proof of Proposition 9 — **Survival Probability (2 drops).** Let cells within droplets in serial transfer be modelled by a linear birth–death process with constant parameters *b* and *d*, subject to periodic bottlenecks δ every duration *T* for *m* + *n* cycles. However, consider that at cycle *m*, *k* droplets are produced (instead of one), and then each new droplet submitted independently to the serial transfer procedure for the remaining *n* cycles. We call this situation (*k*, *m*, *n*)-split droplets.

We recall from the proof of Proposition 1 (p. 36) that the generating function of the Linear Markov Branching Process at time *T* that is started with one individual with probability ϵ or with zero particles with probability $1 - \epsilon$ is the linear fractional function $s \mapsto h(Q_{\epsilon}, s)$ with the expression of Q_{ϵ} given by Eq. (3).

Let $s_{k,m,n}$, the probability that a lineage spawned by a single cell is not extinct at the end of the (n + m)th cycle of (k, m, n)-split droplets. In this situation, the cells survive surely before the first cycle ($\epsilon = 1$), then survive the dilution with probability δ before the cycles 2, 3, ... (m - 1), then survive with probability $k\delta$ before cycle *m*, and finally survive with probability δ again before cycles (m + 1), (m + 2), ..., (m + n).

By using the composition rule established in the proof of Proposition 1, the survival probability can be computed using the following linear fractional function:

$$s_{k,m,n} = 1 - h(Q_1 Q_\delta^{m-1} Q_{k\delta} Q_\delta^{n-1}, 0)$$

It can be checked that this expression corresponds to m + n cycles, one dilution with a factor $k\delta$ and (m + n - 2) dilutions with a factor δ .

Proof of Proposition 10 — **Total Diversity (2 drops).** Let $M_{k,m,n}$ be the expected number of unique mutations accumulated in a lineage at cycle n + m after the splitting of the initial droplet at cycle m into $k = 1, 2 \dots \left| \frac{1}{\delta} \right|$ droplets. Then:

$$M_{k,m,n} = \theta s_m L_{k,m,n},$$

with $L_{k,m,n}$ the expected length of the coalescent tree of the population (conditional on survival) whose expression we will establish.

Let $\mathcal{L}_{k,m,n}$ be the random variable encoding the length of the coalescent tree spawned by a single cell at cycle m + n, submitted to a bottleneck δ every T unit of time, where at cycle m, the population was diluted into k droplets instead of one, conditional on non extinction. The length of this tree, conditioned on non extinction, is the sum of the length of the stump tree \mathcal{L}^{\dagger} (i.e., the tree before the *m*th cycle) and the length of the corolla \mathcal{L}^{c} (i.e., all the trees spawned by the lineages that are extant after the dilution at cycle *m*).

First, the stump tree is a CPP stopped at Tm and sampled accordingly. Indeed, all the branches extant in Tm are not necessarily still extant in the present at time T(m + n).

Let us define $\pi_{k,n}$ the probability that a branch extant at time Tm will still be extant at time T(m + n). It is the product of the probability that the lineage be sampled in the split $(k\delta)$, and that it survives after the split (s_n) :

$$\pi_{k,n} = k\delta s_n \tag{27}$$

We recall the Proposition 2 from Lambert and Stadler (2013):

"The genealogy of a Bernoulli(p)-sample taken from a CPP with inverse tail distribution *F* is a CPP with typical node depth denoted H_p with inverse tail distribution F_p given by $F_p(t) = 1 - p + pF(t)$ ".

The stump tree is a CPP with inverse tail distribution F, stopped at time Tm and submitted to a Bernoulli sampling $\pi_{k,n}$. Thus, it is a CPP with inverse tail distribution:

$$F^{\dagger} := 1 - \pi_{k,n} + \pi_{k,n}F,$$

according to Proposition 2 from Lambert and Stadler (2013). Its expected length, L^{\dagger} , is (according to Eq. (21)):

$$\mathbb{E}(\mathcal{L}^{\dagger}) = L^{\dagger} = \int_{0}^{mT} \frac{F^{\dagger}(mT)}{F^{\dagger}(s)} ds$$

Moreover, the expected number of leaves of the stump tree is (according to Eq. (19)):

$$\mathbb{E}(N(\mathcal{L}^{\dagger})) = F^{\dagger}(mT).$$

Additionally, each leaf of the stump tree at time Tm gives rise to a CPP of expected length L_n (according to Eq. (21)):

$$\mathbb{E}(\mathcal{L}_n) = L_n = \int_0^{nT} \frac{F(nT)}{F(s)} ds,$$

Thus, the expected length of the corolla is:

$$\mathbb{E}(\mathcal{L}^{c}) = \mathbb{E}(N(\mathcal{L}^{\dagger}))\mathbb{E}(\mathcal{L}_{n}) = F^{\dagger}(mT)L_{n}.$$
(28)

Finally, the expected length of the full tree is:

$$\begin{split} L_{k,m,n} &= \mathbb{E}(\mathcal{L}_{k,m,n}) \\ &= \mathbb{E}(\mathcal{L}^{\dagger}) + \mathbb{E}(\mathcal{L}^{c}) \\ &= L^{\dagger} + F^{\dagger}(mT)L_{n} \\ &= \int_{0}^{mT} \frac{F^{\dagger}(mT)}{F^{\dagger}(s)} ds + F^{\dagger}(mT) \int_{0}^{nT} \frac{F(nT)}{F(s)} ds \\ &= F^{\dagger}(mT) \left[\int_{0}^{mT} \frac{1}{F^{\dagger}(s)} ds + \int_{0}^{nT} \frac{F(nT)}{F(s)} ds \right] \end{split}$$

This leads to the final expression:

$$M_{k,m,n} = \theta s_m F^{\dagger}(mT) \left[\int_0^{mT} \frac{1}{F^{\dagger}(s)} ds + \int_0^{nT} \frac{F(nT)}{F(s)} ds \right] \quad \Box$$

Proof of Proposition 11 — **Private Mutations - Split Droplet.** Let a single droplet be split into $k = 1 \dots \lfloor \frac{1}{\delta} \rfloor$ at cycle *m*. Let $M'_{k,m,n}$ be the expected number of mutations that are private to any of the *k* droplets when observed at cycle n+m. To correctly enumerate all these mutations, we must add (as illustrated in Fig. 19) the ones that happen in the stump tree (before *m*) and in the corolla (after *m*):

$$M'_{k,m,n} = \underbrace{S_{k,m,n}}_{\text{stump}} + \underbrace{C_{k,m,n}}_{\text{corolla}}$$

Corolla

All the mutations that happen in the corolla are private to a single droplet.

The expected number of mutations that happen in the corolla is proportional to the expected length of the corolla forest \mathcal{L}^c (given by Eq. (28)), multiplied by the mutation rate. Thus,

$$C_{k,m,n} = \theta s_m \mathbb{E}(\mathcal{L}^c) = \theta s_m F^{\dagger}(mT) L_n$$

Stump Tree

Mutations that happen in the stump tree can be carried by cell lineages present in one or several droplets (as explained in Fig. 19).

Let $S_{k,m,n}$ be the expected number of mutations occurring in the stump tree (i.e., before cycle *m*) that are only carried by cells that are found in a single droplet. This number is proportional to the mutation rate θ . Let us compute this value.

The stump tree is a Coalescent Point Process stopped at *mT*, with inverse tail distribution *F*, sampled with probability $\pi_{k,n} = k\delta s_n$ at time *mT*. Thus, it has an inverse tail distribution $F^{\dagger} = 1 - \pi_{k,n} + \pi_{k,n}F$. If the stump is not extinct (i.e., with probability s_m) it has a number of leaves $N(\mathcal{L}^{\dagger})$ that follows a geometric distribution with parameter $\frac{1}{F^{\dagger}(mT)}$.

Consider a mutation that occurred in the stump tree, and let C be the random variable encoding the number of leaves of the stump tree that bear this mutation. This variable follows a probability distribution given in Theorem 2.2 of Lambert (2009).

Finally, if C = j, the probability that all the *j* individuals carrying the focal mutation at the time of the split are sampled in the same droplet is $\frac{1}{k^{j-1}}$.

Hence, by the formula of total probabilities :

$$S_{k,m,n} = \theta s_m \sum_{i \ge 1} \mathbb{P}(N(\mathcal{L}^{\dagger}) = i) \sum_{j=1}^{i} \mathbb{P}(C = j) \frac{1}{k^{j-1}}$$
(29)

Let

$$W^{\dagger}(s) := [1 - 1/F^{\dagger}(mT)]/[1/F^{\dagger}(s) - 1/F^{\dagger}(mT)]$$

and for $\ell \ge 0$, let

 $G_{\ell}(x) := \frac{\ell - 1}{W^{\dagger}(x)^2} + \frac{2}{W^{\dagger}(x)}$

if $\ell > 0$, whereas $G_0(x) := 1$. Then,

$$\begin{split} S_{k,m,n} = &\theta s_m \sum_{i \ge 1} \frac{1}{F^{\dagger}(mT)} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{i-1} G_{i-j}(x) \frac{1}{k^{j-1}} \\ & \times \sum_{j=1}^{i} \int_{0}^{mT} dx \left(1 - \frac{1}{W^{\dagger}(x)} \right)^{j-1} G_{i-j}(x) \frac{1}{k^{j-1}} \\ & = \frac{\theta s_m}{F^{\dagger}(mT)} \int_{0}^{mT} dx \sum_{j \ge 1} \left(1 - \frac{1}{W^{\dagger}(x)} \right)^{j-1} \frac{1}{k^{j-1}} \\ & \times \sum_{i \ge j} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{i-1} G_{i-j}(x) \\ & = \frac{\theta s_m}{F^{\dagger}(mT)} \int_{0}^{mT} dx \sum_{j \ge 1} \left(1 - \frac{1}{W^{\dagger}(x)} \right)^{j-1} \frac{1}{k^{j-1}} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{j-1} \\ & \times \sum_{i \ge 0} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{i} G_{i}(x) \\ & = \frac{\theta s_m}{F^{\dagger}(mT)} \int_{0}^{mT} dx \sum_{j \ge 1} \left(1 - \frac{1}{F^{\dagger}(x)} \right)^{j-1} \frac{1}{k^{j-1}} \\ & \times \sum_{i \ge 0} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{i} G_{i}(x) \\ & = \frac{\theta s_m}{F^{\dagger}(mT)} \int_{0}^{mT} dx \frac{1}{1 - \frac{1}{k} + \frac{1}{kF^{\dagger}(x)}} \sum_{i \ge 0} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{i} G_{i}(x) \end{split}$$

Now we write for any $i \ge 1$

$$\left(1 - \frac{1}{F^{\dagger}(mT)}\right)^{i} G_{i}(x) =: A^{i}(B(i+1) + C)$$

where $A = 1 - 1/F^{\dagger}(mT)$, $B = 1/W^{\dagger}(x)^2$ and $C = 2/W^{\dagger}(x) - 2/W^{\dagger}(x)^2$, so that

$$\begin{split} \sum_{i \ge 0} \left(1 - \frac{1}{F^{\dagger}(mT)} \right)^{r} G_{i}(x) = &1 + \sum_{i \ge 1} A^{i}(B(i+1) + C) \\ = &1 - B + B \sum_{i \ge 0} (i+1)A^{i} + CA \sum_{i \ge 0} A^{i} \\ = &1 - B + \frac{B}{(1-A)^{2}} + \frac{CA}{1-A} \\ = &1 - B + BF^{\dagger}(mT)^{2} + C(F^{\dagger}(mT) - 1) \\ = &1 + \frac{F^{\dagger}(mT)^{2} - 1}{W^{\dagger}(x)^{2}} + \frac{2(F^{\dagger}(mT) - 1)}{W^{\dagger}(x)} \left(1 - \frac{1}{W^{\dagger}(x)} \right) \\ = &\left(1 - \frac{1}{W^{\dagger}(x)} + \frac{F^{\dagger}(mT)}{W^{\dagger}(x)} \right)^{2} \\ = &\left(\frac{F^{\dagger}(mT)}{F^{\dagger}(x)} \right)^{2}. \end{split}$$

Hence we eventually get

$$S_{k,m,n} = \theta s_m F^{\dagger}(mT) \int_0^{mT} \frac{k \ dx}{F^{\dagger}(x)(1+(k-1)F^{\dagger}(x))}$$

Conclusion

Collecting all the terms we get:

$$\begin{split} M'_{k,m,n} &= S_{k,m,n} + C_{k,m,n} \\ &= \theta S_m F^{\dagger}(mT) \left[\int_0^{mT} \frac{k dx}{F^{\dagger}(x)(1+(k-1)F^{\dagger}(x))} + \int_0^{nT} \frac{F(nT)}{F(x)} dx \right] \quad \Box \end{split}$$

Artificial selection of droplets

Proof of Proposition 12 — Cumulative Distribution Functions. Time to accumulate Θ mutations

Let $\left(e_{p_{ij}}^{(i,j)}\right)_{i,j\in\mathbb{N}^2}$ be independent exponential random variables with parameter p_{ij} . Without droplet selection, Γ is the minimum of a set of sums of i.i.d. exponential variables:

$$\Gamma = \inf_{j=1...D} \left(\sum_{i=1}^{\Theta} e_{\frac{\rho}{D}}^{(i,j)} \right)_j$$

With droplet selection, Γ^* is a sum of a minimum of i.i.d. exponential variables, (i.e. a sum of exponential variables):

$$\Gamma^* = \sum_{i=1}^{\Theta} \left(\inf_{j=1\dots D} e_{\frac{\rho}{D}}^{(i,j)} \right)_i = \sum_{i=1}^{\Theta} e_{\sum_{j=1}^{D} \frac{\rho}{D}}^{(i)} = \sum_{i=1}^{\Theta} e_{\rho}^{(i)},$$

the cumulative distribution function of Γ^* is the Erlang distribution with parameters Θ and ρ :

$$\mathbb{P}(\Gamma^* \leq x) = 1 - e^{-\rho x} \left(\sum_{u=0}^{\Theta-1} \frac{(\rho x)^u}{u!} \right)$$

Note that it does not depend on D any more: if the number of droplets is in the order of one over the scaled invasion rate, the time to accumulate Θ mutations with droplet selection does not depend on the invasion rate any more.

Without droplet selection, the cumulative distribution function of Γ is:

$$\begin{split} \mathbb{P}(\Gamma \leq x) &= 1 - \mathbb{P}(\Gamma > x) \\ &= 1 - \mathbb{P}\left(\bigcap_{j=1}^{D} \left[\left(\sum_{i=1}^{\Theta} e_{\frac{\beta}{D}}^{(i,j)}\right)_{j} > x \right] \right) \\ &= 1 - \left[\mathbb{P}\left(\left(\sum_{i=1}^{\Theta} e_{\frac{\beta}{D}}^{(i,j)}\right)_{j} > x \right) \right]^{D} \\ &= 1 - \left[1 - \mathbb{P}\left(\left(\sum_{i=1}^{\Theta} e_{\frac{\beta}{D}}^{(i,j)}\right)_{j} \leq x \right) \right]^{D} \\ &= 1 - \left[1 - 1 + e^{-\frac{\beta}{D}x} \left(\sum_{u=0}^{\Theta-1} \frac{(\rho x)^{u}}{D^{u} u!}\right) \right]^{D} \end{split}$$

The lineages are independent)

(sum of indep. exp. r.v.)

$$= 1 - \left[e^{-\frac{\rho}{D}x} \left(\sum_{u=0}^{\Theta-1} \frac{(x\rho)^u}{D^u u!}\right)\right]^D$$
$$= 1 - e^{-\rho x} \left(\sum_{u=0}^{\Theta-1} \frac{(\rho x)^u}{D^u u!}\right)^D$$

Comparison of the two regimes

Let us show that Γ stochastically dominates Γ^* . Denote by N^* the process counting the number of mutations in the selection regime. In the parallel regime, denote by $N^{(j)}$ the process counting the number of mutations in the *j*th lineage. We know that N^* is a Poisson process with intensity ρ , whereas the $N^{(j)}$ are independent Poisson processes with common intensity ρ/D , so that for all *t*

$$N_t^* \stackrel{(d)}{=} \sum_{j=1}^D N_t^{(j)} \ge \max_{j=1...D} N_t^{(j)}$$
(30)

Now note that Γ^* is the first time when N^* reaches Θ , whereas Γ is the first time when $\max_{j=1...D} N^{(j)}$ reaches Θ . Since the latter counting process is stochastically smaller than the former, we get the desired stochastic inequality.

Now by the law of large numbers, we have the a.s. convergence of N_t^*/t to ρ and for each j, of $N_t^{(j)}/t$ to ρ/D , which also yields the a.s. convergence of $\max_{j=1...D} N_t^{(j)}/t$ to ρ/D . As a consequence, as $\Theta \to \infty$, Γ^*/Θ converges a.s. to $1/\rho$ and Γ/Θ converges a.s. to D/ρ . This proves that Γ/Γ^* converges a.s. to 1/D.

References

- Ardré, M., Doulcier, G., Brenner, N., Rainey, P.B., 2022. A leader cell triggers end of lag phase in populations of Pseudomonas fluorescens. microLife 3, uqac022. http://dx.doi.org/10.1093/femsml/uqac022.
- Arias-Sánchez, F.I., Vessman, B., Mitri, S., 2019. Artificially selecting microbial communities: If we can breed dogs, why not microbiomes? PLOS Biol. 17 (8), e3000356. http://dx.doi.org/10.1371/journal.pbio.3000356.
- Athreya, K.B., Ney, P.E., 1972. Branching Processes. Springer Berlin Heidelberg, Berlin, Heidelberg, http://dx.doi.org/10.1007/978-3-642-65371-1.
- Baraban, L., Bertholle, F., Salverda, M.L.M., Bremond, N., Panizza, P., Baudry, J., Visser, J.A.G.M.d., Bibette, J., 2011. Millifluidic droplet analyser for microbiology. Lab Chip 11 (23), 4057–4062. http://dx.doi.org/10.1039/C1LC20545E.
- Barton, N.H., Depaulis, F., Etheridge, A.M., 2002. Neutral evolution in spatially continuous populations. Theor. Popul. Biol. 61 (1), 31–48. http://dx.doi.org/10. 1006/tpbi.2001.1557.
- Barton, N.H., Etheridge, A.M., Véber, A., 2013. Modelling evolution in a spatial continuum. J. Stat. Mech. Theory Exp. 2013 (01), P01002. http://dx.doi.org/10. 1088/1742-5468/2013/01/P01002.
- Black, A.J., Bourrat, P., Rainey, P.B., 2020. Ecological scaffolding and the evolution of individuality. Nat. Ecol. Evol. 4 (3), 426–436. http://dx.doi.org/10.1038/s41559-019-1086-9.
- Blokhuis, A., Lacoste, D., Nghe, P., Peliti, L., 2018. Selection dynamics in transient compartmentalization. Phys. Rev. Lett. 120 (15), 158101. http://dx.doi.org/10. 1103/PhysRevLett.120.158101.
- Boitard, L., Cottinet, D., Bremond, N., Baudry, J., Bibette, J., 2015. Growing microbes in millifluidic droplets. Eng. Life Sci. 15 (3), 318–326. http://dx.doi.org/10.1002/ elsc.201400089.

Cottinet, D., 2013. Diversité Phénotypique Et Adaptation Chez Escherichia coli Etudiéees En Millifluidique Digitale (Ph.D. thesis). Université Pierre et Marie Curie.

- Cottinet, D., Condamine, F., Bremond, N., Griffiths, A.D., Rainey, P.B., Visser, J.A.G.M.d., Baudry, J., Bibette, J., 2016. Lineage tracking for probing heritable phenotypes at single-cell resolution. PLoS One 11 (4), e0152395. http://dx.doi.org/10.1371/journal.pone.0152395.
- Delgado-Baquerizo, M., Giaramida, L., Reich, P.B., Khachane, A.N., Hamonts, K., Edwards, C., Lawton, L.A., Singh, B.K., 2016. Lack of functional redundancy in the relationship between microbial diversity and ecosystem functioning. J. Ecol. 104 (4), 936–946. http://dx.doi.org/10.1111/1365-2745.12585.
- Dinh, K.N., Jaksik, R., Kimmel, M., Lambert, A., Tavaré, S., 2020. Statistical inference for the evolutionary history of cancer genomes. Statist. Sci. 35 (1), 129–144. http://dx.doi.org/10.1214/19-STS7561.
- Doulcier, G., 2019. Dropsignal Millifluidic Droplet Trains Analysis. Zenodo, http: //dx.doi.org/10.5281/zenodo.1164108.
- Doulcier, G., Lambert, A., De Monte, S., Rainey, P.B., 2020. Eco-evolutionary dynamics of nested Darwinian populations and the emergence of community-level heredity. eLife 9, e53433. http://dx.doi.org/10.7554/eLife.53433.
- Dumbrell, A.J., Nelson, M., Helgason, T., Dytham, C., Fitter, A.H., 2010. Relative roles of niche and neutral processes in structuring a soil microbial community. ISME J. 4 (3), 337–345. http://dx.doi.org/10.1038/ismej.2009.122.

- Dupin, J.-B., 2018. Cultures Multi-Parallélisées En Millifluidique Digitale : Diversité Et Sélection Artificielle (Ph.D. thesis). Université Pierre et Marie Curie.
- Durrett, R., 2013. Population genetics of neutral mutations in exponentially growing cancer cell populations. Ann. Appl. Probab. 23 (1), 230–250. http://dx.doi.org/10. 1214/11-AAP824.
- Dutta, C., Pan, A., 2002. Horizontal gene transfer and bacterial diversity. J. Biosci. 27 (1), 27–33. http://dx.doi.org/10.1007/BF02703681.
- Etheridge, A.M., 2008. Drift, draft and structure: Some mathematical models of evolution. Banach Center Publ. 80, 121–144. http://dx.doi.org/10.4064/bc80-0-7.
- Ewens, W.J., 1972. The sampling theory of selectively neutral alleles. Theor. Popul. Biol. 3 (1), 87–112.
- Fay, J.C., Wu, C.-I., 2000. Hitchhiking under positive Darwinian selection. Genetics 155 (3), 1405–1413.
- Fuhrman, J.A., 2009. Microbial community structure and its functional implications. Nature 459 (7244), 193–199. http://dx.doi.org/10.1038/nature08058.
- Gamblin, J., Gandon, S., Blanquart, F., Lambert, A., 2023. Bottlenecks can constrain and channel evolutionary paths. Genetics 224 (2), iyad001. http://dx.doi.org/10. 1093/genetics/iyad001.
- Hammerschmidt, K., Rose, C.J., Kerr, B., Rainey, P.B., 2014. Life cycles, fitness decoupling and the evolution of multicellularity. Nature 515 (7525), 75–79. http: //dx.doi.org/10.1038/nature13884.
- Kassen, R., Buckling, A., Bell, G., Rainey, P.B., 2000. Diversity peaks at intermediate productivity in a laboratory microcosm. Nature 406 (6795), 508–512. http://dx. doi.org/10.1038/35020060.
- Kawecki, T.J., Lenski, R.E., Ebert, D., Hollis, B., Olivieri, I., Whitlock, M.C., 2012. Experimental evolution. Trends Ecol. Evol. 27 (10), 547–560.
- Lambert, A., 2009. The allelic partition for coalescent point processes. Markov Process. Related Fields 15 (3), 359–386. http://dx.doi.org/10.48550/arXiv.0804.2572.
- Lambert, A., Stadler, T., 2013. Birth–death models and coalescent point processes: The shape and probability of reconstructed phylogenies. Theor. Popul. Biol. 90, 113–128. http://dx.doi.org/10.1016/j.tpb.2013.10.002.
- Lewontin, R.C., 1970. The units of selection. Annu. Rev. Ecol. Syst. 1, 1-18.
- Li, X., Zhou, Z., Li, W., Yan, Y., Shen, X., Wang, J., Sun, X., Yuan, Q., 2022. Design of stable and self-regulated microbial consortia for chemical synthesis. Nature Commun. 13 (1), 1554. http://dx.doi.org/10.1038/s41467-022-29215-6.
- Maron, P.-A., Sarr, A., Kaisermann, A., Lévêque, J., Mathieu, O., Guigue, J., Karimi, B., Bernard, L., Dequiedt, S., Terrat, S., Chabbi, A., Ranjard, L., 2018. High microbial diversity promotes soil ecosystem functioning. Appl. Environ. Microbiol. 84 (9), e02738–17. http://dx.doi.org/10.1128/AEM.02738-17.
- Milshteyn, A., Schneider, J., Brady, S., 2014. Mining the metabiome: Identifying novel natural products from microbial communities. Chem. Biol. 21 (9), 1211–1223. http://dx.doi.org/10.1016/j.chembiol.2014.08.006.
- Mueller, U.G., Sachs, J.L., 2015. Engineering microbiomes to improve plant and animal health. TIM 23 (10), 606–617. http://dx.doi.org/10.1016/j.tim.2015.07.009.
- Ofiţeru, I.D., Lunn, M., Curtis, T.P., Wells, G.F., Criddle, C.S., Francis, C.A., Sloan, W.T., 2010. Combined niche and neutral effects in a microbial wastewater treatment community. Proc. Natl. Acad. Sci. 107 (35), 15345–15350. http://dx.doi.org/10. 1073/pnas.1000604107.
- Panke-Buisse, K., Poole, A.C., Goodrich, J.K., Ley, R.E., Kao-Kniffin, J., 2015. Selection on soil microbiomes reveals reproducible impacts on plant function. ISME J. 9 (4), 980–989. http://dx.doi.org/10.1038/ismej.2014.196.
- Pflughoeft, K.J., Versalovic, J., 2012. Human microbiome in health and disease. Ann. Rev. Pathol.: Mech. Dis. 7 (1), 99–122. http://dx.doi.org/10.1146/annurev-pathol-011811-132421.
- Popovic, L., 2004. Asymptotic genealogy of a critical branching process. Ann. Appl. Probab. 14 (4), 2120–2148. http://dx.doi.org/10.1214/105051604000000486.
- Postek, W., Pacocha, N., Garstecki, P., 2022. Microfluidics for antibiotic susceptibility testing. Lab Chip 22 (19), 3637–3662. http://dx.doi.org/10.1039/D2LC00394E.
- Potvin-Trottier, L., Luro, S., Paulsson, J., 2018. Microfluidics and single-cell microscopy to study stochastic processes in bacteria. Curr. Opin. Microbiol. 43, 186–192. http://dx.doi.org/10.1016/j.mib.2017.12.004.
- Sniegowski, P.D., Gerrish, P.J., Johnson, T., Shaver, A., 2000. The evolution of mutation rates: separating causes from consequences. BioEssays: News Rev. Mol., Cell. Dev. Biol. 22 (12), 1057–1066. http://dx.doi.org/10.1002/1521-1878(200012)22: 12<1057::AID-BIES3>3.0.CO;2-W.
- Sniegowski, P.D., Gerrish, P.J., Lenski, R.E., 1997. Evolution of high mutation rates in experimental populations of E. coli. Nature 387 (6634), 703–705. http://dx.doi. org/10.1038/42701.
- Swenson, W., Wilson, D.S., Elias, R., 2000. Artificial ecosystem selection. Proc. Natl. Acad. Sci. 97 (16), 9110–9114. http://dx.doi.org/10.1073/pnas.150237597.
- Toprak, E., Veres, A., Yildiz, S., Pedraza, J.M., Chait, R., Paulsson, J., Kishony, R., 2013. Building a Morbidostat: An automated continuous-culture device for studying bacterial drug resistance under dynamically sustained drug inhibition. Nat. Protoc. 8 (3), 555–567. http://dx.doi.org/10.1038/nprot.nprot.2013.021.
- Trivedi, P., Delgado-Baquerizo, M., Anderson, I.C., Singh, B.K., 2016. Response of soil properties and microbial communities to agriculture: Implications for primary productivity and soil health indicators. Front. Plant Sci. 7.
- Van den Bergh, B., Swings, T., Fauvart, M., Michiels, J., 2018. Experimental design, population dynamics, and diversity in microbial experimental evolution. Microbiol. Mol. Biol. Rev. 82 (3), http://dx.doi.org/10.1128/mmbr.00008-18.

- Wahl, L.M., Gerrish, P.J., 2001. The probability that beneficial mutations are lost in populations with periodic bottlenecks. Evolution 55 (12), 2606–2610. http: //dx.doi.org/10.1111/j.0014-3820.2001.tb00772.x.
- Wahl, L.M., Gerrish, P.J., Saika-Voivod, I., 2002. Evaluating the impact of population bottlenecks in experimental evolution. Genetics 162 (2), 961–971.
- Wahl, L.M., Krakauer, D.C., 2000. Models of experimental evolution: The role of genetic chance and selective necessity. Genetics 156 (3), 1437–1448.
- Wahl, L.M., Zhu, A.D., 2015. Survival probability of beneficial mutations in bacterial batch culture. Genetics 200 (1), 309–320. http://dx.doi.org/10.1534/genetics.114. 172890.
- Wilson, D.S., 1975. A theory of group selection. Proc. Natl. Acad. Sci. 72 (1), 143-146.
- Xie, L., Shou, W., 2021. Steering ecological-evolutionary dynamics to improve artificial selection of microbial communities. Nature Commun. 12 (1), 6799. http://dx.doi. org/10.1038/s41467-021-26647-4.
- Xie, L., Yuan, A.E., Shou, W., 2019. Simulations reveal challenges to artificial community selection and possible strategies for success. PLOS Biol. 17 (6), e3000295. http://dx.doi.org/10.1371/journal.pbio.3000295.