The coalescent with replication-independent mutations

Stephen M. Krone and Beth M. Tuschhoff

Department of Mathematics and Statistical Science, University of Idaho,
Moscow, Idaho, United States

ABSTRACT

We develop the mathematical structure of the neutral coalescent with both replication-dependent and replication-independent mutations. This allows us to explain and quantify empirical results that explore differences in genetic diversity in bacterial cultures with different growth rates. We also derive an unbiased and consistent estimator for the replication-independent mutation rate that is based on a comparison of total single nucleotide polymorphism counts for two independent well-mixed cultures with different growth rates. In addition to explaining differences in genetic diversity between well-mixed cultures with different (but constant) growth rates, our coalescent also quantifies the effects of fluctuating growth rates—a situation that can be common in natural populations.

INTRODUCTION

There are two types of mutations: *replication-dependent* mutations occur during reproduction events, and *replication-independent* mutations arise independently of reproduction. This is seen most clearly in empirical work on bacteria in growth-limited conditions (*Ryan, 1955*; *Kivisaar, 2010*; *France, Ridenhour & Forney, 2018*). Replication-independent mutations arise in a variety of ways, including error-prone repair of DNA that has become damaged through UV radiation, oxidation, or alkylation (*France, Ridenhour & Forney, 2018*; *Kivisaar, 2010*). Such mutations are sometimes attributed to a stress response, though that need not be the case. Microbial populations can have growth rates that vary widely—with doubling times ranging from a fraction of an hour in lab cultures to days, weeks, or longer in environmental settings (*Morita, 1988*; *France, Ridenhour & Forney, 2018*). This suggests that the role of replication-independent mutations in the generation of microbial diversity should also vary widely. In this paper, we derive the mathematical structure of the coalescent describing the genealogy of a sample for a version of the neutral Wright–Fisher model that includes both replication-dependent and replication-independent mutations. We ignore selective effects in the generation and fixation of mutants, focusing on the accumulation of neutral genetic diversity. This diversity, of course, can provide standing genetic variation that contains the seeds of adaptive evolution once selective pressure is applied, say through the introduction of an antibiotic (*France, Ridenhour & Forney, 2018*).

Our interest in this topic was inspired by empirical work of *France (2018)*, *France, Ridenhour & Forney (2018)* and *France et al. (2019)* that demonstrated the effects of reduced growth rates on the presence of mutations in bacterial populations that were grown in chemostats and biofilms. In particular, they compared two chemostat cultures with the same population density, one with a 'slow' growth rate and the other with a 'fast' growth rate. They found enhanced bacterial diversity in the slowly growing culture compared to faster growing culture when the populations were compared after the same number of generations. When comparisons were made at the same amount of 'real time,' the faster-growing culture displayed the larger number of mutations. Coalescent theory, which is based on sampling existing genetic diversity in the present based on the shared ancestry of individuals in the sample, is a generation-based construct. As such, a coalescent that accounts for replication-independent mutations should predict more genetic diversity in slowly growing populations.

Blath et al. (2015) studied a related situation involving spore-forming cells in which a given microbial population is divided into subpopulations, with one containing actively growing cells and the other dormant cells. They model this with a neutral seed bank structure and derive a coalescent describing the ancestry of a sample. Our approach contrasts with the one in *Blath et al. (2015)* in that we deal with populations that (1) experience periods of reduced growth rather than complete dormancy and (2) are not spatially or otherwise structured in a way that keeps individuals in the reproducing or non-reproducing subpopulations over an extended period. For example, our model corresponds to bacteria growing in a chemostat in which nutrient levels can be changed to alter growth rates; the models of *Blath et al. (2015)* are more appropriate for bacteria growing in a biofilm, where parts of the population do not have access to nutrients and do not grow, and others are actively growing. Although *Blath et al. (2015)* have replication-independent mutations in their model, their analyses focus on cases where these mutations are ignored. We find that replication-independent mutations are important in explaining certain empirical results, and so we emphasize their role in our analyses.

In theoretical population genetics, forward-time mathematical models typically include mutation as a component of generational events (reproduction). The standard approach in diffusion theory and coalescent theory is to assume a Wright–Fisher model (*Wakeley, 2009*) with mutation probability per individual per reproduction of the form $u = \theta/(2N)$, where N is population size (assumed constant in the simplest cases) and θ is the scaled (replication-dependent) mutation rate. With this scaling and a speed up of time (t units of continuous time corresponding to $\lfloor Nt \rfloor$ generations), one obtains a Wright–Fisher diffusion in forward time and Kingman's coalescent in reverse time. This theme has been expanded upon to include the effects of more complex dynamics by incorporating additional structure in the Wright–Fisher dynamics. Two examples of such additional structure are pertinent to the present study. In one example, *Kaj & Krone (2003)* and *Sjödin et al. (2005)* investigated the structure of the coalescent in the presence of stochastically fluctuating population sizes. The appropriately scaled ancestral process converges to a coalescent with constant mutation rate $\theta/2$ along each lineage and pairwise

coalescence rates that change with time. An equivalent rescaling (in terms of predicting the same mutation frequencies in a sample) of this coalescent has constant pairwise coalescence rates but mutation rates that vary with time. A second example is the seed bank coalescent introduced by *Kaj, Krone & Lascoux (2001)*. In this case, not all offspring come from parents in the previous generation, with some "seeds" being sequestered for a time before they result in progeny. This sequestration of part of the population results in a slowing of coalescence rates and a consequent increase in diversity. *Blath et al. (2015)* expanded on this by allowing for a more general "strong seed bank" structure, and they were the first to propose its use to model microbial diversity.

MODEL DEVELOPMENT AND COALESCENT STRUCTURE

We derive the mathematical structure of the coalescent when both replication-independent and replication-dependent mutations are present. We will see how the two types of mutations operate on different time scales—a result that can be put to use in estimating rates of replication-independent mutations and explaining recent empirical data on microbial diversity. Reproduction (or replication) events involve replication of the genome of the reproducing "parent," and this process is subject to replication errors, leading to replication-dependent mutations. Replication-independent mutations arise over time independently of replication. In order to account for both sources of mutation in our model, we must be able to partially decouple time steps and replication events.

Reproduction: Our forward-time model is an extension of the discrete-time haploid Wright–Fisher model with fixed population size N in which, instead of replacing all N individuals with their offspring each time step, only a fraction of them (randomly chosen) is replaced—resulting in a decoupling of time steps and "generations," as will be discussed later. This replacement fraction is assumed to depend on growth conditions (something easily controlled in a lab, for example, by regulating nutrient levels and dilution rate for bacteria in a chemostat). Since we are interested in a retrospective approach, we denote by $R_N(k)$ the number of individuals in the population who are replaced k time steps in the past; thus $0 \le R_N(k) \le N$. The N individuals in the population alive k time steps in the past thus arose from the previous time step through $R_N(k)$ replacements and $N - R_N(k)$ individuals surviving to the next time step. Thus, the replicating part of the population undergoes the usual Wright–Fisher dynamics with all $R_N(k)$ parents dying and being replaced with $R_N(k)$ offspring. The special case $R_N(k) \equiv N$ yields the usual Wright–Fisher model. Two specific individuals in time step $k - 1$ will have come from a reproduction event from the same parent in time step k if they both were the products of reproducing parents and, given this, they had the same parent. Thus

P(a given pair in time step $k - 1$ coalesces to a common parent in time step k) =

$$\left(\frac{R_N(k)}{N}\right)^2 \cdot \frac{1}{R_N(k)} = \frac{R_N(k)}{N^2}.$$

Note that the intermediate step in this equation does not work in the special case $R_N(k) = 0$, but the final result holds trivially since there can be no coalescing in a time step

where there was no replication. The special case of Wright–Fisher dynamics, $R_N(k) = N$, yields the usual $1/N$ pairwise coalescence probability.

To construct the coalescent process, assume that the scaled 'replacement' process converges to a *replication rate* process $\psi(t) \in [0, 1]$:

$$\frac{R_N(\lfloor Nt \rfloor)}{N} \rightarrow \psi(t) \tag{1}$$

as $N \rightarrow \infty$. Roughly speaking, this means that the average number replaced per time step is of order N; *e.g.*, $R_N(k) = c_k N$, where $c_k \in [0, 1]$. The convergence in Eq. (1) is weak convergence of stochastic processes and includes as special cases of $\psi(t)$ continuous-time Markov chains, diffusion processes, and deterministic functions of time. See (*Kaj & Krone, 2003*; *Sjödin et al., 2005*) for examples and the technical details of convergence in a model with stochastically fluctuating population sizes, which has some features similar to those in the model here.

With the above convergence, the probability that two specific lineages do not coalesce in $\lfloor Nt \rfloor$ time steps, given the sequence of $R_N(k)$ values, is

$$
\begin{aligned}
P_2(\text{no coalescence in} \lfloor Nt \rfloor \text{time steps} \mid R_N(\cdot)) &= \prod_{k=1}^{\lfloor Nt \rfloor} \left(1 - \frac{R_N(k)}{N^2} \right) \\
&\sim \exp\left(-\frac{1}{N} \sum_{k=1}^{\lfloor Nt \rfloor} \frac{R_N(k)}{N} \right) \\
&\rightarrow \exp\left(-\int_0^t \psi(s)\,ds \right),
\end{aligned}
$$

as $N \rightarrow \infty$. Thus, on the coalescent time scale, $\psi(t)$ is the pairwise coalescence rate at time t in the past. This agrees with the standard pairwise coalescence rate of 1 when we have full replacement each generation.

A simple example of the framework indicated by Eq. (1) has $R_N(k)$ taking values in $\{0, N\}$, so that we get full replacement in some time steps and no replacement in others. If the transition probabilities for the process $(R_N(k))$ are of the form $p(0, N) = \lambda_1/N$ and $p(N, 0) = \lambda_0/N$, then the replication rate process $(\psi(t))$ takes values in $\{0, 1\}$ and has transition rates $q(0, 1) = \lambda_1$ and $q(1, 0) = \lambda_0$. In this situation, coalescence events and replication-dependent mutations would arise according to a standard coalescent while $\psi(t) = 1$ and they would pause while $\psi(t) = 0$. Replication-independent mutations would arise independently of the value of $\psi(t)$.

We remark briefly that, in the above model description, $R_N(k) \equiv 1$ is formally equivalent to the Moran model (*Wakeley, 2009*), though this model assumes that all parents are equally likely to serve as the lone parent in a given time step, and hence the coalescence of lineages going back in time requires a slightly different interpretation. If we had chosen $R_N(k)$ to be of order 1, instead of N, we would need a different scaling of time ($\lfloor N^2 t \rfloor$ instead of $\lfloor Nt \rfloor$) in (1). We will not pursue the Moran scaling in this paper.

Mutation: All mutations in our model are assumed to follow the "infinite sites" assumption, with each mutation being distinguishable from all others (*Wakeley, 2009*). With the usual *replication-dependent mutation* probability per individual *per reproduction* scaled as $u = \theta/(2N)$, we obtain a scaled replication-dependent mutation rate $\frac{\theta}{2} \cdot \psi(t)$ along each lineage in the limiting coalescent process. This follows by a calculation similar to the one in the above displayed equation but, at time step k in the past, the probability that a given lineage experiences a replication-dependent mutation is

$$\frac{R_N(k)}{N} \cdot \frac{\theta}{2N},$$

the probability that a specific individual replicates, and the offspring acquires a new replication-dependent mutation.

To account for the different patterns of mutations seen in empirical data for growth-limiting cultures, we need another source of mutations that is not tied to reproduction events. We thus further assume that, in the Wright–Fisher construction, there is an additional probability $w = v/(2N)$ of a *replication-independent mutation* for each individual in the population *per time step*, regardless of their status as a new offspring or a surviving individual from the previous time step. An even simpler calculation of the above form shows that, in the limiting coalescent, this will correspond to a constant rate $v/2$ of replication-independent mutations along each lineage.

Remark: Note that the same scaling of coalescence and replication-dependent mutation obtains. If replication-independent mutations were not present ($v = 0$), patterns of replication-dependent mutations in such a tree (or in a sample of SNPs from such a population with varying replication rates) would be indistinguishable from those in the standard coalescent. This is not the case, for example, in the coalescent with stochastically varying population size (*Kaj & Krone, 2003*; *Sjödin et al., 2005*) or in the seed bank coalescent (*Kaj, Krone & Lascoux, 2001*). In the present model, it is the additional replication-independent mutations that lead to differences from the standard coalescent.

Combining the effects of reproduction and the two types of mutation, in the limit as $N \to \infty$, we obtain a coalescent process that can be viewed in two ways:

I. The coalescent process with replication-dependent and replication-independent mutations, defined by the rates

- pairwise coalescence rate $\psi(t)$,
- replication-dependent mutation rate per lineage $\frac{\theta}{2} \cdot \psi(t)$,
- replication-independent mutation rate per lineage $\frac{v}{2}$.

These rates correspond to a scaling of coalescence time in units of "time steps" in the revised Wright–Fisher model; t units of coalescent time corresponds to $\lfloor Nt \rfloor$ time steps in the WF model.

If we re-scale coalescence rates and mutation rates by the same amount (even if that scaling depends on t), the distribution of branch lengths will change but we obtain the

same distribution of mutations along the branches. Thus, as far as polymorphism data are concerned, it is equivalent to use.

II. The re-scaled coalescent with replication-dependent and replication-independent mutations, defined by the rates

- pairwise coalescence rate 1,
- replication-dependent mutation rate per lineage $\frac{\theta}{2}$,
- replication-independent mutation rate per lineage $\frac{v}{2\psi(t)}$.

In this scaling, of course, we need $\psi(t) > 0$ for all t. These re-scaled rates correspond to a scaling of coalescence time in units of "generations" (the average time it takes to replace N individuals) in the revised Wright–Fisher model; t units of coalescent time corresponds to $\lfloor Nt \rfloor$ generations in the WF model. (Note that time steps and generations are no longer the same in our revised Wright–Fisher model; not all individuals are involved in the reproduction-replacement dynamics.)

In other words, we can obtain the re-scaled coalescent process with replication-dependent and replication-independent mutations by generating the standard Kingman coalescent and then adding replication-independent mutations at rate $v/(2\psi(t))$ per lineage at time t in the past. This version is easier to work with since the replication rate appears only in the "new" replication-independent mutation rate. Note that it also implies that the number of replication-independent mutations increases as v increases or $\psi(t)$ decreases; only the ratio matters.

APPLICATIONS

Two distinct settings are relevant for application of this coalescent:

1. *Constant replication rate*: the replication rate $\psi(t)$ is constant, but we compare polymorphism data for several different values of this constant;
2. *Fluctuating replication rate*: the replication rate $\psi(t)$ is changing with time quickly enough that we see changes of replication rate in the coalescent.

Comparing populations with different constant replication rates

In all that follows, we use abbreviations RD and RI for replication-dependent and replication-independent mutations. We begin by considering how RI mutations and replication rate combine to affect the number of segregating sites in a sample from a single population with constant replication rate. A constant replication rate, $\psi(t) = c \in (0, 1]$, will yield a "generation" (defined as the time it takes for replacement of N individuals in a population of size N) every $1/c$ time steps on average. We will exploit this difference between time steps and generations to provide a method for determining the rate of replication-independent mutations, even though we cannot directly distinguish these from replication-dependent mutations in practice.

Recall first some basic results from coalescent theory (*Wakeley, 2009*). Writing T_i for the time (in the coalescent) until the next coalescence when there are i lineages, and $L_n = \sum_{i=2}^{n} iT_i$ the total branch length in the tree for a sample of size n, we have $E(L_n) = 2a_n$

and $\text{Var}(L_n) = 4b_n$, where $a_n = \sum_{i=1}^{n-1} 1/i$ and $b_n = \sum_{i=1}^{n-1} 1/i^2$. Assume the infinite sites model of mutation and let $S_n^{(RD)}$ denote the number of single nucleotide polymorphisms (SNP) in a sample of size n that were generated by RD mutations. We have $E(S_n^{(RD)}) = \theta a_n$ and $\text{Var}(S_n^{(RD)}) = \theta a_n + \theta^2 b_n$, where $\theta/2$ is the per-lineage RD mutation rate.

The quantities T_i and L_n in the coalescent correspond to NT_i and NL_n generations when the population size is fixed at N and all N individuals are replaced by offspring each generation. Now, if the replication rate is fixed at $\psi(t) = c \in (0, 1)$, the rates of coalescence and RD mutation both slow by a factor of c compared to a standard coalescent with replication rate 1, so $S_n^{(RD)}$ has the same distribution, independent of $c > 0$. Since we are interested in the pattern of SNPs in a sample, only the relative rates matter and we can use the re-scaled coalescent process (II) when considering the added effects of RI mutations. In particular, the mean and variance of $S_n^{(RD)}$ do not depend on c. If, however, we let $S_n^{(RI,c)}$ denote the number of RI mutations in a sample of size n, this quantity depends on actual time instead of generations, and so is a function of the replication rate c. In particular, the total branch length L_n, in generations, corresponds to L_n/c in real time units. So the effective rate of RI mutations on the coalescent scale is $v/2c$. Hence $E(S_n^{(RI,c)}) = E[E(S_n^{(RI,c)}|L_n)] = \frac{v}{2c} \cdot 2a_n = \frac{v}{c} \cdot a_n$ and, similarly, $\text{Var}(S_n^{(RI,c)}) = E[\text{Var}(S_n^{(RI,c)}|L_n)] + \text{Var}[E(S_n^{(RI,c)}|L_n)] = \frac{v}{c} \cdot a_n + \frac{v^2}{c^2} \cdot b_n$.

Combining the two types of mutations, the *total number of SNPs* in a sample of size n is

$$S_n^{(c)} = S_n^{(RD)} + S_n^{(RI,c)}.$$

When the replication rate is fixed at c,

$$E(S_n^{(c)}) = a_n \left(\theta + \frac{v}{c} \right) \tag{2}$$

and

$$\text{Var}(S_n^{(c)}) = \text{Var}(S_n^{(RD)}) + \text{Var}(S_n^{(RI,c)}) + 2\text{Cov}\left(S_n^{(RD)}, S_n^{(RI,c)} \right)$$

$$= a_n \left(\theta + \frac{v}{c} \right) + b_n \left(\theta^2 + \frac{v^2}{c^2} \right) + \frac{2\theta v}{c} b_n$$

$$= a_n \left(\theta + \frac{v}{c} \right) + b_n \left(\theta + \frac{v}{c} \right)^2. \tag{3}$$

Of course, these results follow even more directly by recognizing that the rescaled version of our coalescent (II), in the special case where $\psi(t) = c$, is equivalent to the standard coalescent with (total) mutation rate $(\theta + \frac{v}{c})/2$. Similarly, it follows (*Wakeley, 2009*) that the average number of pairwise differences, $\pi_n^{(c)}$, in a sample of size n satisfies

$$E(\pi_n^{(c)}) = \theta + \frac{v}{c} \tag{4}$$

and

$$\text{Var}(\pi_n^{(c)}) = \frac{n+1}{3(n-1)} \left(\theta + \frac{v}{c} \right) + \frac{2(n^2 + n + 3)}{9n(n-1)} \left(\theta + \frac{v}{c} \right)^2 \tag{5}$$
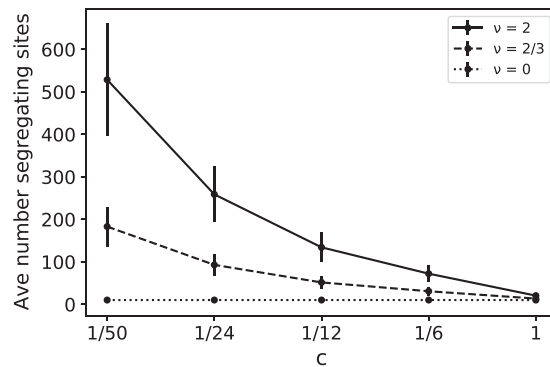
Figure 1 Average number of segregating sites, $S_{100}^{(c)}$, for a sample of size $n = 100$ as a function of replication rate c for three different values of RI mutation rate v. The RD mutation rate is $\theta = 2$ and each data point corresponds to 100,000 runs. For $v = 0$, the constant value of the sample means for $S_{100}^{(c)}$ was approximately 10.35 with sample standard deviation $\sigma \approx 4.1$. The predicted expected value for the case $v = 0$ is $E(S_{100}^{(c)}) = a_{100}\theta \approx 5.177 \times 2 \approx 10.35$ RD mutations for each value of c, clearly within one standard deviation of the mean. The other simulated means in the graph (corresponding to $v = 2/3$ and 2) were also found to be within one standard deviation of the theoretical mean $E(S_{100}^{(c)}) = a_{100}(\theta + v/c)$. Full-size ⬚ DOI: 10.7717/peerj.12926/fig-1

and the ith element in the site frequency spectrum has mean

$$E(\xi_i) = (\theta + \frac{v}{c})/i. \tag{6}$$

Note that the effects of v and c are intertwined, always appearing together in the ratio v/c. For example, doubling the RI mutation rate has the same effect as reducing the replication rate by 1/2. In Fig. 1 we plot simulated values of $S_n^{(c)}$ as a function of c for several different values of the RI mutation rate v. These simulations provide a coalescent-based rendering of the types of data found in *France (2018)*. In Fig. 1, for a given $v > 0$, the increase in $S_n^{(c)}$ as c decreases is due to the fact that slower replication provides more opportunity for RI mutations to accumulate.

Estimating the RI mutation rate. Consider now the case where we are able to compare SNP data from two independent cultures of the same species, with constant replication rates: one slow with rate c_1 and one fast with rate c_2 ($0 < c_1 < c_2$). Since one expects the same number of replication-dependent mutations in the two cultures, the average difference between total numbers of mutations, $S_n^{(c_1)} - S_n^{(c_2)}$, will be accounted for by the rate of replication-independent mutations. Indeed,

$$E\left(S_n^{(c_1)} - S_n^{(c_2)}\right) = a_n v\left(\frac{1}{c_1} - \frac{1}{c_2}\right), \tag{7}$$

which leads to an unbiased *estimator of the RI mutation rate*:

$$\hat{v} = \frac{S_n^{(c_1)} - S_n^{(c_2)}}{a_n\left(\frac{1}{c_1} - \frac{1}{c_2}\right)} . \tag{8}$$
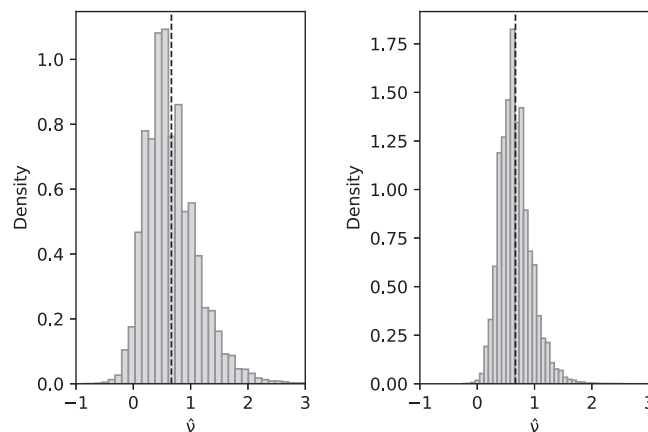
Figure 2 **Simulated values of $\hat{v}$ with sample size $n = 10$ (left) and $n = 100$ (right) when actual value is $v = 2/3$ and parameters are $\theta = 2$, $c_1 = 1/12$ (for slow-growth population), $c_2 = 1$ (for fast-growth population).** Left: sample mean 0.6648, sample variance 0.2060; Right: sample mean 0.6677, sample variance 0.0744. These are a close match to the theoretical values $E(\hat{v}) = 2/3$ and $\text{Var}(\hat{v}) = 0.2073$ for $n = 10$ and 0.0742 for $n = 100$. Full-size ◨ DOI: 10.7717/peerj.12926/fig-2

Thus $E(\hat{v}) = v$ and, since $S_n^{(c_1)}$ and $S_n^{(c_2)}$ are independent,

$$
\text{Var}(\hat{v}) = \frac{\text{Var}\left(S_n^{(c_1)} - S_n^{(c_2)}\right)}{a_n^2\left(\frac{1}{c_1} - \frac{1}{c_2}\right)^2} = \frac{1}{a_n^2\left(\frac{1}{c_1} - \frac{1}{c_2}\right)^2}\left[a_n\left(2\theta + v\left(\frac{1}{c_1} + \frac{1}{c_2}\right)\right)\right.
$$
$$
\left. + b_n\left(2\theta^2 + v^2\left(\frac{1}{c_1^2} + \frac{1}{c_2^2}\right)\right) + 2\theta v b_n\left(\frac{1}{c_1} + \frac{1}{c_2}\right)\right]. \tag{9}
$$

Asymptotically, since $a_n \sim \ln(n)$ and $b_n \sim \pi^2/6$, $\text{Var}(\hat{v}) \to 0$ as $n \to \infty$, and hence $\hat{v}$ is also a consistent estimator of v. Compare with Fig. 2.
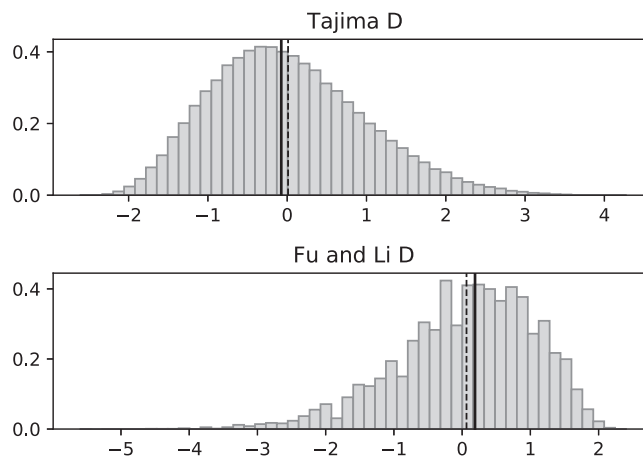
Fluctuating replication rates

In the case of fast fluctuations in replication rate, we can compute the *coalescent effective population size* as in *Kaj & Krone (2003)* and *Sjödin et al. (2005)*. Intuitively, if the growth rate changes fast enough, it will reach equilibrium between coalescence and mutation events. If $\psi(t)$ has unique stationary distribution π, then the resulting ancestral dynamics will be equivalent to that of a population with constant growth rate $c = \int_{[0,1]} x\pi(dx)$ and the coalescent effective size will be $N_e = N/c$. The process thus behaves as though there was a constant growth rate and an overall effective mutation rate $c \cdot \theta/2 + v/2$ per lineage.

In the case of growth rates that fluctuate on the same time scale as coalescence events, there will be no coalescent effective size and we expect our coalescent to generate polymorphism data that deviate from what one expects in a standard, constant population size, neutral coalescent. In Figs. 3–6, we compare the effects of fluctuating replication rates as measured by Tajima's D (D_T) and Fu and Li's D (D_{FL}) statistics. We use the same versions of D_T and D_{FL} that are given in *Blath et al. (2015)*. In particular, SNP data that
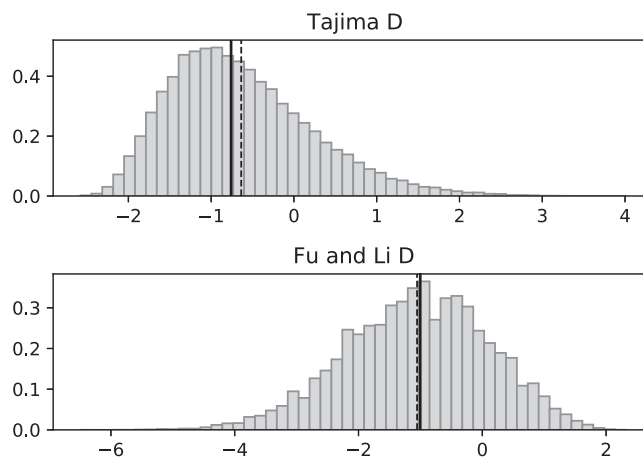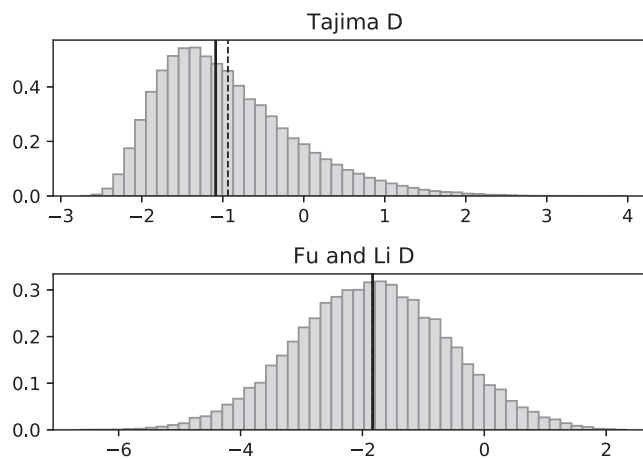
Figure 3 Fluctuating replication rates with RI mutation rate $v = 2/3$ and slow initial replication. Simulated values of Tajima and Fu and Li D statistics for fluctuating replication rates $\psi(t)$ switching between $c_1 = 1/12$ and $c_2 = 1$ at rate 2 (starting with slower replication rate c_1). Simulations used $n = 100$, $\theta = 2$, $v = 2/3$, and 100,000 replicates. The means (dashed vertical lines) are $D_T = -0.6369$ and $D_{FL} = -1.0539$. The medians (solid vertical lines) are $D_T = -0.7593$ and $D_{FL} = -1.0027$.

Full-size 🖼 DOI: 10.7717/peerj.12926/fig-3

Figure 4 Fluctuating replication rates with RI mutation rate $v = 2/3$ and fast initial replication. Simulated values of Tajima and Fu and Li D statistics for fluctuating replication rates $\psi(t)$ switching between $c_1 = 1/12$ and $c_2 = 1$ at rate 2, as in Fig. 3, but this time starting with the faster replication rate c_2. Simulations used $n = 100$, $\theta = 2$, $v = 2/3$, and 100,000 replicates. The means (dashed vertical lines) are $D_T = 0.0111$ and $D_{FL} = 0.0640$. The medians (solid vertical lines) are $D_T = -0.0746$ and $D_{FL} = 0.1889$.

Full-size 🖼 DOI: 10.7717/peerj.12926/fig-4

come from a model consistent with the standard Kingman coalescent (in particular, with constant replication rates) should have a mean near 0.

Simulated SNP data were produced from simulations of the coalescent with RI and RD mutations. Deviations from the SNP patterns that would be found in a standard constant-replication-rate coalescent are evident in negative mean and median values of D_T and D_{FL} when the coalescent begins (at the time of the sample) with the slower rate (Fig. 3). By contrast, starting at the faster replication rate masks many of the effects of RI mutations since the tree experiences more coalescing of lineages before the replication

Figure 5 Fluctuating replication rates with RI mutation rate $v = 2$ and slow initial replication. Increasing the RI mutation rate enhances the effects of fluctuating replication rates. Simulations used $n = 100$, $\theta = 2$, $v = 2$, ψ switches between 1/12 and 1 at rate 2 (starting with 1/12), and 100,000 replicates. The means (dashed vertical lines) are $D_T = -0.9360$ and $D_{FL} = -1.8354$. The medians (solid vertical lines) are $D_T = -1.0873$ and $D_{FL} = -1.8331$. Full-size ⬜ DOI: 10.7717/peerj.12926/fig-5
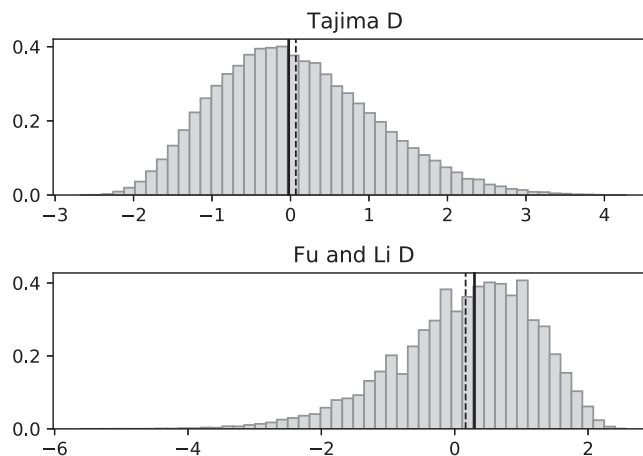
Figure 6 Fluctuating replication rates with RI mutation rate $v = 2$ and fast initial replication. Simulations used $n = 100$, $\theta = 2$, $v = 2$, ψ switches between 1/12 and 1 at rate 2 (starting with 1), and 100,000 replicates. The means (dashed vertical lines) are $D_T = 0.0680$ and $D_{FL} = 0.1636$. The medians (solid vertical lines) are $D_T = -0.0233$ and $D_{FL} = 0.2965$. Full-size ⬜ DOI: 10.7717/peerj.12926/fig-6

rate fluctuations begin, as seen in mean D_T and D_{FL} values near 0 (Fig. 4). Thus, sampling from a population that is in a slow-growth phase at the time of the sample is more likely to exhibit differences from standard SNP patterns. This effect is enhanced when the slow-growth phase is even slower. For example, under the conditions of Fig. 3, if we change c_1 to 1/50, the means become $D_T = -0.9521$ (instead of -0.6369) and $D_{FL} = -2.2670$ (instead of -1.0539). Because of its emphasis of mutations along external branches, D_{FL} is more sensitive to these patterns than D_T.

Figures 5 and 6 repeat the fluctuating size experiments of Figs. 3 and 4, but have a higher RI mutation rate ($v = 2$ instead of $v = 2/3$). As expected, the deviations from constant replication rate expectations are enhanced when the RI mutation rate is larger.

DISCUSSION

Diversity is purged by directional selection and drift—both requiring reproduction. When growth rates are very low, these forces play a minor role relative to replication-independent mutations, allowing diversity to accumulate despite a paucity in the replication-dependent mutations that most models focus on.

It is through the partial decoupling of growth and generation (with a fraction of the population being replaced by offspring in each time step) that we are able to use discrete time steps as a proxy for real time in the implementation of the replication-independent mutation process. In the Wright–Fisher model with population size N, one time step is the same as one generation; all N individuals are replaced every generation and, retrospectively, it takes an average of N generations for a pair of lineages to coalesce. As a simple comparison, consider our revised Wright–Fisher model with $R_N(k) = N/3$ individuals replaced each time step (*i.e.*, the case $c = 1/3$). It takes three time steps to replace N individuals in this model, and it takes an average of $3N$ time steps for two lineages to coalesce. This matches the Wright–Fisher model if we now consider one "generation" to be three time steps. The RI mutations accumulate on the basis of time steps, not generations, and hence we expect more of these mutations in the case where the replication rate is slower.

In the limiting coalescent process, one can think of the accumulation of replication-independent mutations as providing a real-time molecular clock (*Kumar & Subramanian, 2002*). In the absence of replication-independent mutations ($v = 0$), there is no notion of real time in the coalescent, just generational time. This is seen from the fact that coalescence and replication-dependent mutation rates both scale identically with $\psi(t)$. In other words, with $v = 0$, there is no way to discern differences in growth rates over time or across spatial locations (or different experimental conditions) from the pattern of mutations in a sample.

Our coalescent provides insight into the empirical work of *France (2018)*, *France, Ridenhour & Forney (2018)* and *France et al. (2019)* that tracked the accumulation of mutations in several slow-growth settings in biofilms and chemostats. In particular, they observed the accumulation of mutations in *E. coli* forward in time for slow-growth (24 h doubling time) and fast-growth (2 h doubling time) chemostats and found that the fast-growth cultures had more mutations when compared over the same amount of (real) time. However, when they were compared over the same number of generations, the slow-growth cultures had significantly more mutations due to the fact that replication-independent mutations had more time to accumulate over their longer growth period (*i.e.*, generation time). The growth rates in our simulations were chosen to match the relative rates in these experiments. It is worth remarking that population sizes were held fixed in our model and in the experiments of *France (2018)*. This means that when comparing fast and slow growth populations, the fast growth population has faster turnover (*i.e.*, replacement) of individuals; growth and death rates are both faster. If we were to only increase the growth rate, but not the death rate, we would be comparing populations with different population sizes. *Kaj & Krone (2003)* and *Sjödin et al. (2005)*
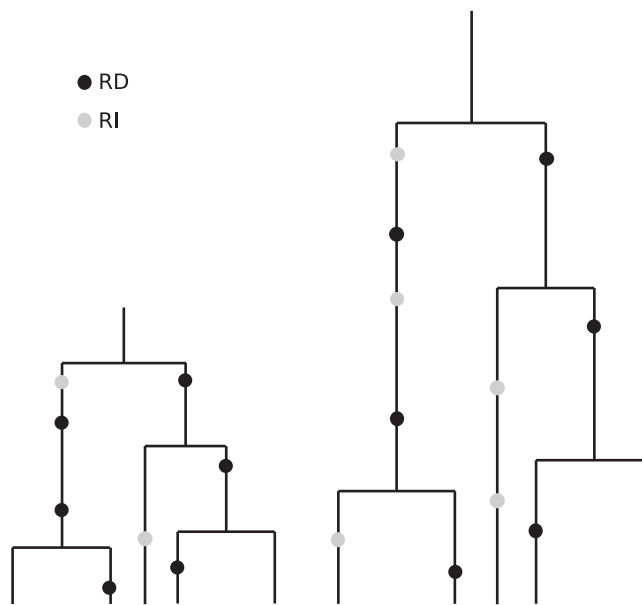
Figure 7 Illustration of the effects of "real" time and generation time. This drawing indicates two coalescent trees having the same sample size and the same topology, with the one on the right corresponding to a slower replication rate. Although the number of generations is the same (and we have put the black RD mutations in the same relative positions for visual clarity), the number of time steps–real time–per generation is greater in the tree on the right. Since the grey RI mutations occur at a rate that is based on real time, we see more RI mutations in the tree on the right.

Full-size 🖼 DOI: 10.7717/peerj.12926/fig-7

showed how an increase in population size increases diversity in a sample due to a slowing of coalescence rate.

The coalescent, which is designed to account for the statistical patterns of SNPs in samples from a population, is generation-based and hence conforms more closely to comparisons that are made between cultures having the same number of generations. Specifically, coalescence events (that determine the branch lengths in the genealogy of a sample) occur according to generational time. These branches then accommodate replication-dependent mutation events (also generation based) and replication-independent mutations (occurring in real time). See Fig. 7 for an artist's rendering of two coalescent trees that illustrate how slower replication rate leads to more RI mutations. The patterns of diversity generated by the coalescent with replication-independent mutations are in broad agreement with the empirical results from *France (2018)* and provide insight into the underlying mechanisms.

Our coalescent results for the case of constant, but reduced, growth rate show that the expected number of RI mutations in a sample will be proportional to v/c. So even a very small RI mutation rate, v, will significantly influence the diversity of a sample provided that the replication rate, c, is small enough. In environmental samples, for example, c can be extremely small. Using our unbiased estimator of v, it will now be possible to quantify in the lab actual RI mutation rates for various bacterial species.

Blath et al. (2015) studied a related seed bank model and coalescent that focuses on dormancy instead of slow growth. Their simulations were limited to the case of no

replication-independent mutations ($v = 0$), though their model included the possibility of $v > 0$. Unlike our model, which produces negative values of Tajima's D and Fu and Li's D when replication-independent mutations occur at positive rates, theirs yielded positive values of Tajima's D and Fu and Li's D even in the absence of replication-independent mutations. In their model, the dormant cells are sequestered (as would occur with spatial structure) and cells migrate slowly between this dormant group and the actively growing group; it is this spatial segregation that produces positive D statistics. As such, their model appears to be especially appropriate for the interesting case of biofilm growth. Our model is appropriate for well-mixed cultures. The fact that the two models make different predictions about SNP data opens the possibility of using SNP data to infer the historical level of spatial structure or mixing in an environmental microbial population.

There is no selection in our model or in the Blath seed bank model. *France (2018)*, *France, Ridenhour & Forney (2018)* and *France et al. (2019)* have demonstrated increased mutations in slow growth chemostat cultures, which are presumably neutral, and in biofilms, where differences in growth rate are depth dependent. Some of these mutations turned out to provide resistance to antibiotics—a selective pressure that was not present when the mutations were generated. This suggests that even a neutral coalescent with replication-independent mutations can inform the evolutionary dynamics of bacterial populations that eventually undergo selective pressure—such as through the use of antibiotics.

ACKNOWLEDGEMENTS

ADDITIONAL INFORMATION AND DECLARATIONS

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.12926#supplemental-information.

REFERENCES

Blath J, González Casanova A, Eldon B, Kurt N, Wilke-Berenguer M. 2015. Genetic variability under the seedbank coalescent. *Genetics* **200(3)**:921–934 DOI 10.1534/genetics.115.176818.

France MT. 2018. Evolutionary mechanisms that facilitate the maintenance of microbial diversity. Ph.D. dissertation, University of Idaho.

France MT, Cornea A, Hanna K-D, Forney LJ. 2019. Spatial structure facilitates the accumulation and persistence of antibiotic-resistant mutants in biofilms. *Evolutionary Applications* **12(5)**:498–507 DOI 10.1111/eva.12728.

France MT, Ridenhour BJ, Forney LJ. 2018. Effects of spatial structure and reduced growth rates on evolution in bacterial populations. In: Rampelotto P, ed. *Molecular Mechanisms of Microbial Evolution, Grand Challenges in Biology and Biotechnology.* Cham: Springer, 175–197.

Kaj I, Krone SM. 2003. The coalescent process in a population with stochastically varying size. *Journal of Applied Probability* **40(1)**:33–48 DOI 10.1017/S0021900200022257.

Kaj I, Krone SM, Lascoux M. 2001. Coalescent theory for seed bank models. *Journal of Applied Probability* **38(2)**:285–300 DOI 10.1239/jap/996986745.

Kivisaar M. 2010. Mechanisms of stationary-phase mutagenesis in bacteria: mutational processes in pseudomonads. *FEMS Microbiology Letters* **312(1)**:1–14 DOI 10.1111/j.1574-6968.2010.02027.x.

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proceedings of The National Academy of Sciences* **99(2)**:803–808 DOI 10.1073/pnas.022629899.

Morita RY. 1988. Bioavailability of energy and its relationship to growth and starvation survival in nature. *Canadian Journal of Microbiology* **34(4)**:436–441 DOI 10.1139/m88-076.

Ryan FJ. 1955. Spontaneous mutation in non-dividing bacteria. *Genetics* **40(5)**:726–738 DOI 10.1093/genetics/40.5.726.

Sjödin P, Kaj I, Krone SM, Lascoux M, Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics* **169(2)**:1061–1070 DOI 10.1534/genetics.104.026799.

Wakeley J. 2009. *Coalescent theory: an introduction.* Greenwood Village: Roberts and Company Publishers.