Received: 28 May 2024 | Accepted: 2 January 2025

DOI: 10.1111/2041-210X.14501

Methods in Ecology and Evolution

RESEARCH ARTICLE

The rarity of mutations and the inflation of bacterial effective population sizes

Rui Borges[1,2]

[1]School of Mathematics and Statistics, University of St Andrews, St Andrews, UK

[2]Institute of Population Genetics, University of Veterinary Medicine Vienna, Vienna, Austria

Correspondence
Rui Borges
Email: rb251@st-andrews.ac.uk

Present address
Rui Borges, Office 101, Mathematical Institute, St Andrews, UK

Funding information
Austrian Science Fund, Grant/Award Number: P34524 and P37050

Handling Editor: Natalie Cooper

Abstract

1. Mutations are fundamental for evolution, and their mathematical modelling in population genetics heavily relies on our perception of their frequency and the timescale over which they occur. A common assumption is that mutations are infrequent, so when a new mutation arises, the previous one has either become fixed or lost. This assumption implies that mutations occur exclusively at fixed sites and is referred to as the boundary mutation model. However, one can alternatively assume a recurrent mutation model, which additionally considers mutations contributing to shifts in allele frequency.
2. In this study, we compare these two models. By examining mutation rates and effective population sizes across the Tree of Life, we demonstrate that the boundary mutation model remains valid for most species but significantly deviates in bacteria. Our analyses further reveal that the boundary mutation model tends to overestimate the effective population size, particularly in bacteria, where estimated population sizes can be more than five times larger than those expected by the recurrent mutation model. We address these biases by proposing a Bayesian estimator for population size that accounts for recurrent mutations.
3. To illustrate how mutation models can influence the quantification of forces other than drift, we further present a case study showing that the boundary mutation model exaggerates the intensity of selective constraints acting on the three codon positions of the bacterium *Pseudomonas fluorescens*.
4. This study emphasizes the importance of considering recurrent mutations in highly diverse species for accurate population genetics inference.

KEYWORDS
bacteria and *Pseudomonas fluorescens*, boundary mutations, effective population size, mutation rate, recurrent mutations

1 | INTRODUCTION

Mutation is a key process in the study of evolution, serving as an essential source of genetic variation. Mutations are common figures in mathematical models of population genetics, and the ways to model them vary depending on how low mutation rates (denoted as μ) are assumed to be and the evolutionary timescale at which they take place. A commonly used mutation model is the infinite sites

(IS) model (Hudson, 1983; Kimura, 1969), which assumes an infinite number of sites where mutations can occur, with each new mutation necessarily occurring at a unique site. Therefore, backward or multiple mutations are not expected by this model. The IS model is quite popular and forms the basis of widely used methods for estimating the scaled population mutation rate θ. Two examples are Watterson's θ_W, based on the number of segregating or polymorphic sites in a population sample (Watterson, 1975), and π, which measures the number of nucleotide differences per site (Tajima, 1989). Despite its popularity and representing the cornerstone of important mathematical results, the IS model has a significant limitation. Desai and Plotkin (2008) demonstrated that the IS approximation is suitable for most eukaryotic populations but becomes problematic when θ exceeds 0.05, as is often the case for microbes and viruses.

At longer timescales, back mutations become more likely, and models accounting for more than one mutation per site have been proposed in the context of the diffusion limit. Wright (1931) obtained the equilibrium allele frequency distribution ψ (or site-frequency spectrum (SFS)) for a model with recurrent mutations in a Wright–Fisher haploid population expressed as:

$$\psi\left(x,\theta_{01},\theta_{10}\right) \propto x^{\theta_{01}-1}(1-x)^{\theta_{10}-1}, \quad 0<x<1, \tag{1}$$

where x is the frequency of allele 1 in the population, $\theta_{01}=2N\mu_{01}$ and $\theta_{10}=2N\mu_{10}$ are the population mutation rates between alleles 0 and 1, and N is the population size. In this model, and the ones further developed in these pages, we assume that the population follows an idealized dynamic governed by mutation and genetic drift, in which N quantifies the magnitude of genetic drift in the population and can be more specifically referred to as the *effective* population size.

An alternative approach to modelling multiple mutations is considering that μ is generally low. Vogl and Clemente (2012) proposed the boundary mutation (BM) model, which only considers mutations when the population is monomorphic. According to this model, frequency shifts between polymorphic states are driven by drift or selection but not mutation. The rationale behind this assumption is rather intuitive: The rate of drift typically outpaces μ, so by the time a new mutation occurs, the previous one would likely already be fixed or lost in the population. Using expectations from the BM model, Vogl and Clemente (2012) determined that it constitutes a good approximation to the recurrent mutation (RM) model—a model where mutations also contribute to frequency shifts—when θ is less than 0.05. This finding aligns with the one obtained by Desai and Plotkin (2008) for the IS model. The direct comparison of the SFS of the BM model with that of Equation (1) allowed for the definition of a more lenient threshold of 0.1 (Schrempf et al., 2016), a limit further supported by Vogl and Mikula (2021).

In contrast to the IS model, the BM model allows for reversible mutations; however, both models overlook recurrent mutations. The BM model thus lies between the IS and the RM models. Considering that the BM model has sound mathematical properties (Vogl & Clemente, 2012; Vogl & Mikula, 2021) and has found important applications in population genetics and phylogenetics (e.g. inferring polymorphism-aware phylogenies efficiently (Borges, Boussau, Höhna, et al., 2022), while accounting for directional and balancing selection (Borges, Boussau, Szöllösi, et al., 2022; Braichenko et al., 2024)), we here investigate the implications of assuming boundary mutations in population genetics inference by extensively comparing the BM and RM models (the IS model was thoroughly tested in Desai and Plotkin (2008)). We first leverage μ for various prokaryotic and eukaryotic species (Lynch et al., 2016; Lynch et al., 2023) to compare the SFS and estimates of N under the BM and RM models and find significant differences between the models, particularly in highly diverse species such as bacteria. Finally, we explore a case study involving the bacterium *Pseudomonas fluorescens* (Flügge, 1886), which highlights differences in inferred selective constraints when considering boundary or recurrent mutations.

2 | MATERIALS AND METHODS

2.1 | The site-frequency spectra of the boundary and recurrent mutation models

The SFS at equilibrium holds great value in population genetics and is widely used within the diffusion framework to compare population dynamics including genetic drift, mutation and directional selection (e.g. Fisher, 1930; Kimura, 1955a, 1955b; Wright, 1931). In this section, we obtain the expected SFS at equilibrium for the BM and RM models on a discrete state space. These simply represent the stationary distributions of their transition matrices, denoted as q^{BM} and q^{RM}, respectively.

Let us consider a population comprising a constant number N of individuals, each carrying one of two alleles, 0 or 1, at any given time. Let n denotes the population state where n individuals possess allele 1, and the remaining $N-n$ individuals have allele 0. When $n=0$ or $n=N$, we describe the population as being in a monomorphic state, fixed exclusively for allele 0 or 1. In all other instances, the population is in a polymorphic state. Both the BM and RM models define the transition probability $q_{u,v}$ between states u and v similarly, incorporating two population forces: genetic drift and mutation. However, the extent to which mutations contribute to frequency shifts differs between the two models. Common to both models are the transition probabilities from fixed to singleton polymorphic states:

$$q^{\mathrm{BM}}_{0,1}=q^{\mathrm{RM}}_{0,1}=\mu_{01} \quad \text{and} \quad q^{\mathrm{BM}}_{N,N-1}=q^{\mathrm{RM}}_{N,N-1}=\mu_{10}, \tag{2}$$

where μ_{01} is the mutation rate from allele 0 to 1, and vice versa for μ_{10}. Concerning the transitions between polymorphic states, the BM model considers only genetic drift, while the RM model additionally takes mutations into account. Drift is modelled using the Moran model for its tractability (Muirhead & Wakeley, 2009). In particular, we employ a model where drift and mutation are decoupled, similar to the approaches of Ellen Baake (2008) and Etheridge et al. (2010). Under the Moran model, at each time step, an individual is chosen to die, and another individual (possibly the same) is chosen to reproduce

(Moran, 1958). Let us consider the scenario where the frequency of allele 1 increases by one unit:

$$q^{\mathrm{BM}}_{n,n+1} = \frac{n}{N}\frac{N-n}{N} \quad \text{and} \quad q^{\mathrm{RM}}_{n,n+1} = \frac{n}{N}\frac{N-n}{N} + \frac{N-n}{N}\mu_{01}, \quad 0 < n < N. \tag{3}$$

The shared term in both transition probabilities reflects genetic drift and indicates the product of the probabilities of selecting an individual 1 to reproduce and an individual 0 to die. The second term in the transition probability of the RM model represents the probability of an individual 0 mutating to 1. The transition probabilities for the scenario where the allele 1 decreases by one unit remain the same in the BM model, due to the symmetry of drift rates. However, in the RM model, the mutation term models the opposite case, expressing the probability of an individual 1 mutating to 0:

$$q^{\mathrm{BM}}_{n,n-1} = \frac{N-n}{N}\frac{n}{N} \quad \text{and} \quad q^{\mathrm{RM}}_{n,n-1} = \frac{N-n}{N}\frac{n}{N} + \frac{n}{N}\mu_{10}, \quad 0 < n < N. \tag{4}$$

In both models, following the standard Moran model (Moran, 1958), frequency shifts exceeding one unit have probability 0. Acknowledging that the defined transition matrix is stochastic and represents a birth–death process, we can obtain the stationary distribution of the BM and RM models, denoted as ψ^{BM} and ψ^{RM}, by solving the detailed balance equations.

The SFS for the BM model has been previously obtained. One approach involves disregarding all the transitions with a probability of order θ^2 in a recurrent mutation model (Vogl & Clemente, 2012). Another one consists of limiting, by construction, the mutations to the fixed states (see supplementary material of Borges et al., 2019). These SFS are equivalent in the biallelic case. Let ψ_n represents the stationary frequency of the state n. The frequencies of the monomorphic and polymorphic states of the SFS of the BM model can be defined as follows:

$$\psi^{\mathrm{BM}}_n = \begin{cases} \frac{1}{k^{\mathrm{BM}}}\mu_{10} & \text{if } n = 0 \\ \frac{1}{k^{\mathrm{BM}}}\mu_{01}\mu_{10}\frac{N}{n}\frac{N}{N-n} & \text{if } 0 < n < N, \\ \frac{1}{k^{\mathrm{BM}}}\mu_{01} & \text{if } n = N \end{cases} \tag{5}$$

where k^{BM} is a normalization constant that ensures that the SFS frequencies add up to one. We have also derived the stationary frequencies of the RM model for both polymorphic and monomorphic states (see Text S1 for a detailed derivation). The SFS of the RM model is defined as follows:

$$\psi^{\mathrm{RM}}_n = \begin{cases} \frac{1}{k^{\mathrm{RM}}}\mu_{10}\Gamma(N\mu_{10}+N)\Gamma(N\mu_{01}+1) & \text{if } n = 0 \\ \frac{1}{k^{\mathrm{RM}}}N\mu_{10}\mu_{01}\binom{N}{n}\Gamma(N\mu_{10}+N-n)\Gamma(N\mu_{01}+n) & \text{if } 0 < n < N, \\ \frac{1}{k^{\mathrm{RM}}}\mu_{01}\Gamma(N\mu_{01}+N)\Gamma(N\mu_{10}+1) & \text{if } n = N \end{cases} \tag{6}$$

where $\Gamma(.)$ denotes the gamma function and k^{RM} represents the normalization constant.

Both the BM and RM models consider reversible mutations, wherein there is always the possibility that an individual carrying allele 0 may mutate to 1 and vice versa. Fixed sites are viewed as outcomes of fixation, not as sites yet to mutate, as in the IS framework. Consequently, unlike standard population genetic models, the concept of derived and ancestral alleles does not apply to the BM and RM models, and inferences can be drawn without the need to define the direction of a mutation, as becomes clear in the following section.

2.2 | The sample site-frequency spectra of the boundary and recurrent mutation models

To construct an estimator of N with boundary and recurrent mutations, we first need to obtain the sampled SFS of these models, which we respectively denote as χ^{BM} and χ^{RM}. The sampled SFS represents the distribution of allele counts in a sample of M individuals from the population.

Let us assume a population of N individuals in drift–mutation balance according to the SFS, and a hypergeometric sampling, where M individuals are sampled without replacement from a finite population of size N. At a given genomic position, allele 1 might vary between 0 and M among the sampled individuals. These allelic counts fall into three types: monomorphic count of type 0, when none of the sampled individuals have allele 1; monomorphic count of type 1, when all M individuals have allele 1; and polymorphic counts, when allele 1 is observed in 1 to $M-1$ of the sampled individuals. Assuming stationarity, the probability of sampling a monomorphic count is

$$\chi_0 = \psi_0 + \sum_{n=1}^{N-M}\frac{\binom{N-n}{M}}{\binom{N}{M}}\psi_n \quad \text{and} \quad \chi_M = \psi_N + \sum_{n=M}^{N-1}\frac{\binom{n}{M}}{\binom{N}{M}}\psi_n, \tag{7}$$

which includes the probability of sampling a monomorphic count from an originally monomorphic population state (terms ψ_0 and ψ_N), but also from a polymorphic population state, which after the sampling process resulted in a monomorphic count (the summation term). For polymorphic counts, we have:

$$\chi_m = \sum_{n=m}^{N-M+m}\frac{\binom{n}{m}\binom{N-n}{M-m}}{\binom{N}{M}}\psi_n, \tag{8}$$

which sums up over all the polymorphic population states where allele 1 is present with a frequency equal to or higher than m and vice versa for allele 0. The summations in Equations (7) and (8) can be

solved using the SFS at equilibrium of the BM and RM models. If the population is in drift–mutation balance according to the boundary mutation model (i.e. $\psi = \psi^{BM}$), allele 1 may appear with frequencies ranging from 0 to M, with probabilities proportional to the following expression:

$$\chi_m^{BM} \propto \begin{cases} \mu_{10} + \mu_{10}\mu_{01} N\left(H_{N-1} - H_{M-1}\right) & \text{if } m = 0 \\ \mu_{10}\mu_{01} \dfrac{MN}{m(M-m)} & \text{if } 0 < m < M, \\ \mu_{01} + \mu_{01}\mu_{10} N\left(H_{N-1} - H_{M-1}\right) & \text{if } m = M \end{cases} \quad (9)$$

where $H_K := \sum_{k=1}^{K} \frac{1}{k}$ represents a harmonic number. If the population is in drift–mutation balance according to the recurrent mutation model (i.e. $\psi = \psi^{RM}$), and we further assume that μ is unbiased ($\mu = \mu_{01} = \mu_{10}$), the sampled SFS of the RM model is as follows:

$$\chi_m^{RM} \propto \begin{cases} \dfrac{\Gamma(1-N-N\mu)}{\Gamma(1-N\mu-M)} \dfrac{\Gamma(1-2N\mu-M)}{\Gamma(1-N-2N\mu)} & \text{if } m = 0 \\ N\mu \binom{M}{m} \dfrac{\Gamma(N\mu+m)}{\Gamma(N\mu+1)} \dfrac{\Gamma(2N\mu+N)}{\Gamma(2N\mu+M)} \dfrac{\Gamma(N\mu-\mu+M)}{\Gamma(N\mu+N)} & \text{if } 0 < m < N. \\ \dfrac{\Gamma(1-N-N\mu)}{\Gamma(1-N\mu-M)} \dfrac{\Gamma(1-2N\mu-M)}{\Gamma(1-N-2N\mu)} & \text{if } m = N \end{cases} \quad (10)$$

2.3 | A Bayesian estimator of the effective population size with boundary and recurrent mutations

We employed the sampled SFS to develop a Bayesian estimator for N. Let us consider a scenario where we have observed S genomic positions and obtained a vector of allelic counts c, where c_m represents the number of genomic positions where m sampled individuals had allele 1 and $M - m$ had allele 0. Based on the expected sampled SFS of the BM or RM model, we can define the likelihood as

$$\log L\left(c \mid \mu_{01}, \mu_{10}, N\right) = \sum_{m=0}^{M} c_m \log\left(\chi_m\right), \quad (11)$$

where $\sum_{m=0}^{M} c_m$ equals S, the total number of sampled genomic positions. We further integrated the likelihood function into a Bayesian framework by assuming a uniform prior distribution on $\log N$ defined over the interval $[0, \infty[$. We estimated N using Markov Chain Monte Carlo (MCMC) for 100,000 generations via the Metropolis–Hastings algorithm (Hastings, 1970), using a normal proposal on $\log \mu$. The first 10% of MCMC samples were discarded as burn-in. The remaining samples were used to assess convergence and mixing by analysing two independent chains, which included visual inspection of trace plots and calculation of the Gelman–Rubin statistics (Gelman & Rubin, 1992).

To test the quality of the Bayesian estimators, we simulated data under the sampled SFS of the corresponding model and estimated N under the same model. We then assessed precision and accuracy using the MCMC samples for a range of scenarios, including varying the number of sampled individuals $M = \{5,10,20\}$ and genomic positions $S = \left\{10^5, 10^6, 10^7\right\}$.

The functions for estimating N under the BM and RM models were implemented in the R programming language (R Core Team, 2021). A tutorial on how to use these estimators is available via https://doi.org/10.5281/zenodo.14621213 (Borges, 2025). This tutorial also includes a variant of these estimators for cases where one wants to consider uncertainties, either technical or in a meta-analysis, in μ. The uncertainty associated with μ is accounted for in the inference of N by sampling μ from a normal distribution defined by user-specified mean and standard deviation during the MCMC. This approach ensures that μ is marginalized and its uncertainty fully incorporated into the posterior distribution of N.

2.4 | Gathering mutation rates and population sizes across the tree of life

To explore a biologically meaningful range of values for μ and N, we used previously published data from (Lynch et al., 2023). In particular, we employed measures of μ for 148 prokaryotic and eukaryotic species, along with the silent-site diversity θ_S and the species ploidy p, to estimate N under the expectation of the infinite sites model: $\theta_S = 2pN\mu$. Dataset S1 summarizes these data for all the species included in this study and is available via https://doi.org/10.5281/zenodo.14621213 (Borges, 2025). We note that our estimates of N and those provided by Lynch et al. (2023) are calculated differently. Lynch et al. (2023) uses the expected neutral-site heterozygosity in a diploid population under drift–mutation balance. Because we wanted an N matching the expectation of the IS model, which was theoretically shown to relate to the BM model (Vogl & Clemente, 2012), we did not use the estimates by Lynch et al. (2023). However, we note that our estimates and those of Lynch et al. (2023) are almost perfectly correlated (Figure S1).

To investigate the impact of mutation bias in the SFS of the BM and RM models, we set $\mu_{01} = f\mu_{10}$, where f is the relative difference between μ_{01} and μ_{10}, under the constraint that $\frac{1}{2}\left(\mu_{01} + \mu_{10}\right) = \mu$. In these analyses, we considered $f = \{1, 2, 5, 10\}$, values based on transition versus transversion rate biases reported by Lynch et al. (2023).

2.5 | Misspecifying the boundary mutation model

The misspecification analyses involved simulating data using the RM model and estimating N under the BM model. To achieve this, we sampled from Equation (10) and performed our inferences using the Bayesian estimator for N with the BM model. We assessed deviations between the estimates of the BM model and the true N by exploring a range of μ and N, based on Dataset S1. To consider the sampling effects on the misspecified N estimates, we varied the number of sampled individuals, $M = \{5, 10, 20\}$. The number of genomic positions was constant and equal to 1 million.

To correct the estimates of N given by the BM model, we took advantage of the bias between the N estimated in the BM model

and the true N used to simulate allele counts under the RM model. The idea is to find the RM N that would have generated the BM N. Let $f(\mu, N)$ represents the surface of N estimated by the BM model. We want to know the value of N for which $f(\mu = u, N) = \overline{N}$, where $\overline{N}$ is an estimate of the population size under the BM model, and u is the mutation rate of a particular species. The N obtained in this manner is the population size that was used to generate the allele counts under the RM model. For calculating this correction, we used the surface f calculated for a sample of 10 individuals and 1 million sites.

2.6 | The case study with Pseudomonas fluorescens

We employed Prokka (Seemann, 2014) to annotate 10 complete genomes of *Pseudomonas fluorescens*: strains L228_4038, PCL1751_2955, PICF7_2957, Ant01_21839, 2020WEIHUA_J_15229, NCTC10038_6712, W-6_15173, KF1_15196, L111_4037 and L321_4036. The strains were obtained from the Pseudomonas Genome Database (Winsor et al., 2015). Subsequently, we fed the annotated assemblies from Prokka into Roary to generate the pan-genome and the multiple sequence alignment of 2066 protein-coding core genes (Page et al., 2015). These multiple sequence alignments were later used to determine the sample SFS for each of the three codon positions, resulting in 681,518 sites per codon position. We chose to combine the counts of the AC and GT bases to mitigate compositional biases in these bases, as the Bayesian estimator created in this study ignores mutation biases. We proceeded with the inference of N using the sampled SFS, employing both the BM and RM models, as well as a variant of the RM model that considers uncertainty in μ. In these inferences, we used the *Pseudomonas fluorescens* μ reported by Long, Miller, et al. (2018): $9.31 \pm 0.66 \times 10^{-11}$. All the analyses were conducted using R (R Core Team, 2021).

3 | RESULTS

3.1 | Comparing the site-frequency spectra of the two mutation models

To identify scenarios where the SFS of the BM and RM models yield distinct predictions, we conducted a comparative analysis across a specified range for μ and N, informed by data on μ from Lynch et al. (2023) and N estimated by the IS model (Dataset S1). The dataset comprises 148 species, with μ ranging from 7.6×10^{-12} to 4.7×10^{-8} and N spanning from 1.1×10^{4} to 4.0×10^{9}. Figure 1 depicts the absolute difference between the frequencies of monomorphic and polymorphic states as expected by the SFS of both models. The absolute difference was measured as follows:

$$\varepsilon = \left|\psi_0^{RM} - \psi_0^{BM}\right| + \left|\sum_{n=1}^{N-1} \psi_n^{RM} - \sum_{n=1}^{N-1} \psi_n^{BM}\right| + \left|\psi_N^{RM} - \psi_N^{BM}\right|, \quad (12)$$

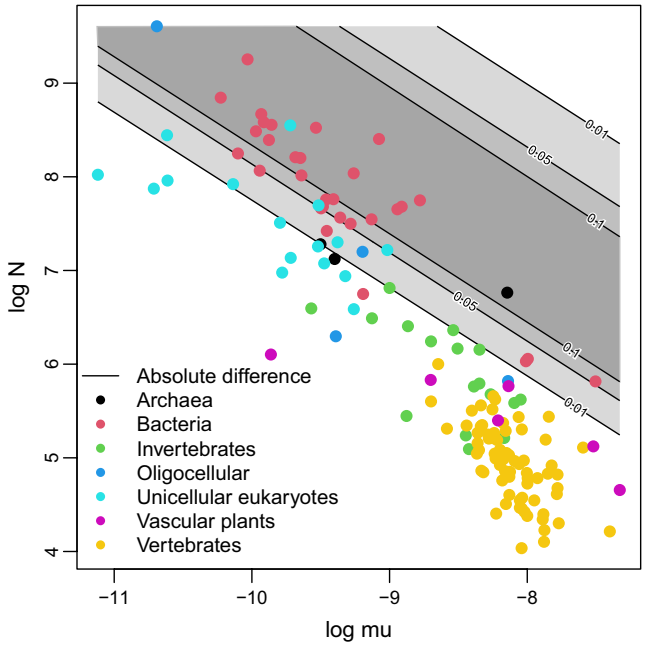

FIGURE 1 Comparing the site frequency spectra of the recurrent and boundary mutation models. The isolines depict the absolute difference between the frequencies of monomorphic and polymorphic states, as predicted by the SFS of the BM and RM models, each applied to the same μ and N. The μ and N of each species are represented by individual dots and can be found in Dataset S1.

where we chose to compare the sum of polymorphic frequencies instead of the frequency of each polymorphic state to ensure fair interspecific comparisons, given that N (and consequently the number of polymorphic states) varies significantly among the analysed species. As expected, this difference is smaller in regions where N is smaller and μ is lower. However, it is also smaller for larger N and higher μ. This is because, in the latter case, all sites are expected to be polymorphic under both models, though this does not imply that the SFS of the BM and RM models is necessarily the same. This scenario is highly unrealistic, and none of the species studied fall within this region. Instead, the majority of species predominantly fall in the region where the log (base 10 logarithm) of μ and N exhibits an inverse relationship (Lynch et al., 2016). Another noticeable pattern is that while for the majority of taxa, including vertebrates, invertebrates, vascular plants and most protists, their SFS shows almost no differences between the models, significant deviations are observed in most bacteria, as well as some archaea and protists (Figure 1). A considerable proportion of bacteria exhibit an absolute difference above 5% between the two SFS, indicating that boundary mutations inadequately account for their genetic diversity levels. Consistent with this, 53% of bacteria display a θ higher than 0.05, ranging between 0.007 and 0.42. These findings corroborate those of (Desai & Plotkin, 2008; Vogl & Clemente, 2012).

We highlight that bacteria exhibit the lowest μ but the largest N among the analysed species. Since $\theta \propto N\mu$, N is more likely to drive the high levels of genetic diversity observed in bacteria and, consequently, is likely the main factor contributing to the poor fit of the

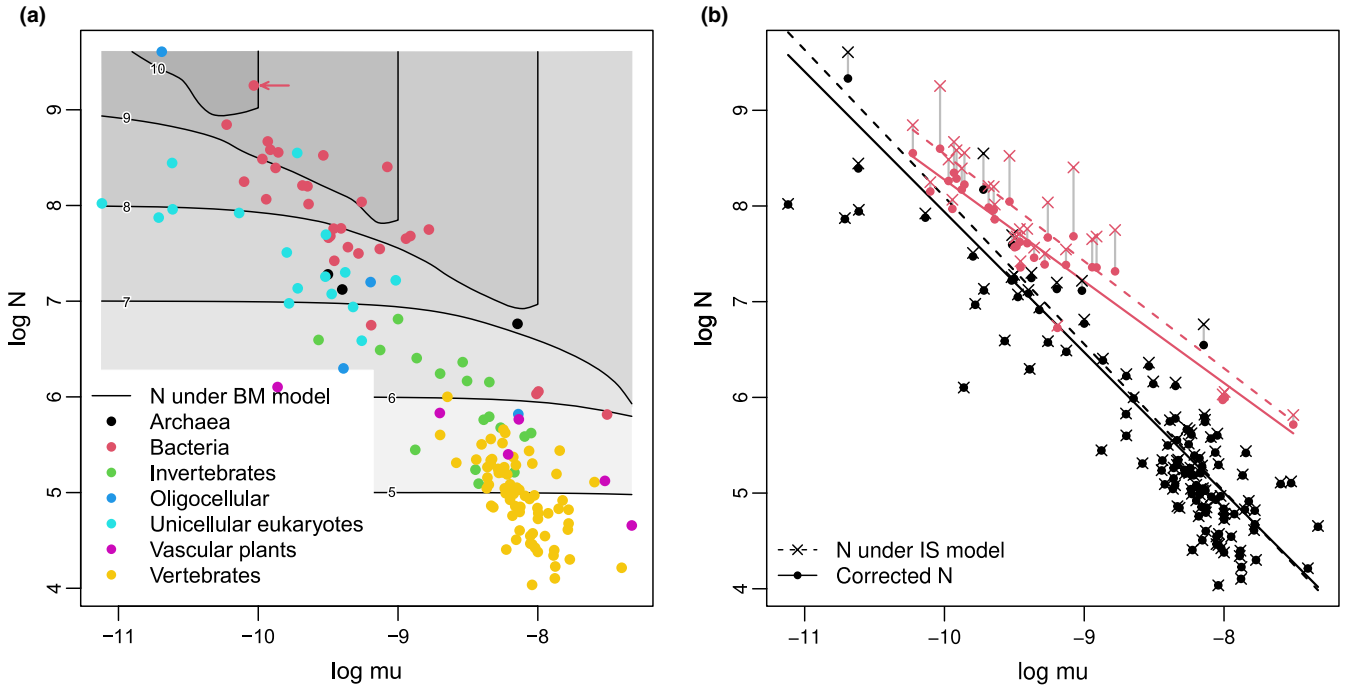

FIGURE 2 Misspecifying the boundary mutation model with recurrent mutations. (a) The isolines show the values of *N* estimated by the BM model, using allele counts simulated under the RM model. The *y*-axis represents the true *N*. The red arrow on top indicates the bacterium *Pseudomonas fluorescens*, which we explore further in a case study to highlight differences in inferred selective constraints when assuming boundary or recurrent mutations. (b) Utilizing the difference between the estimated and true *N* observed in the contour of panel (a), we corrected the estimates of *N* obtained under the IS model. Because bacteria show noticeable biases, we highlighted them in red.

BM model in general. However, we note that the coefficient of variation for μ stands at 2.89, whereas that of *N* is 1.67. This indicates that μ varies more drastically across bacteria, suggesting that it might be a key factor responsible for interspecific variation in genetic diversity.

We have also compared the SFS of the two models under mutation bias, that is, when $\mu_{01} \neq \mu_{10}$. This analysis indicates that the absolute difference between the models' SFS tends to increase with the magnitude of the bias, albeit by a small fraction (Figure S2), not altering the conclusions drawn in the previous paragraphs. To conclude, even when factoring in the likely scenario of mutation bias, the BM assumption generally holds well low-diversity taxa.

3.2 | Misspecifying the boundary mutation model with recurrent mutations

To gain a deeper understanding of the implications of wrongly assuming boundary mutations in the inference of *N*, we misspecified the BM model, by generating data using the RM model and subsequently estimating *N* assuming the BM model. *N* plays a pivotal role in population genetics, determining the level of variability within a population and the expected rate of genetic drift (Charlesworth, 2009). Therefore, this exercise helps us determine whether the assumption of boundary mutations is likely to alter the quantification of the genetic drift level in populations evolving with recurrent mutations.

The contour plot in Figure 2a illustrates the estimated *N* by the BM model across various simulated scenarios, with the true *N* represented on the *y*-axis. In the region of the plot where μ is lower, the estimated and true *N* are identical (note the horizontal isolines that match the true *N*; Figure 2a). However, in the region where μ is higher, there is a noticeable tendency for the estimated *N* by the BM model to deviate from the true *N*. For many bacterial species and some protists, the estimated *N* tends to be larger than the true *N*. We interpret this overestimation to introduce more opportunities for mutation, enabling the diversity levels in the BM model to align with those of the RM model for the same μ. It is worth noting that the observed biases are minimally influenced by the number of sampled individuals (Figure S3), suggesting that larger sample sizes would not resolve them.

Vogl and Clemente (2012) demonstrated that, in the absence of mutation bias, the BM model produces the same expectation for the number of polymorphisms as the IS model. This implies that the estimates of *N* that we obtained with the IS model using the expectation $\theta \approx N\mu$, are equal to or very close to those of the BM model. Leveraging this expectation, we utilized the surface depicted in Figure 2a to calculate the potential bias introduced by ignoring recurrent mutations in the estimates of *N* under the IS expectation. In specific, we sought to find the *N* under the RM model, which would likely have estimated the obtained *N* under the IS model, assuming the same μ. The corrected values of *N* are illustrated in Figure 2b. We observed that the estimates of *N* between the two models generally agree for the majority of species. However, bacteria once again stand out as an exception, where the values of *N* in the IS model are 1.05–5.26 times larger than would be expected by the RM model. For instance, in vertebrates, the biases are comparatively small, varying between 0.99 and 1.05.

We have also noticed that the correlation between N and μ slightly improves when using the corrected N values (Pearson's correlation ρ goes from − 0.89 to − 0.91). Nevertheless, the regression line shows no dramatic deviations from the one using estimates of N under the IS model (Figure 2b), suggesting that the correction should not fundamentally change our understanding and discussions on the relationship between μ and N (Lynch et al., 2016, 2023). Differently, when we examined the correlation among bacterial species, we observed a notable improvement (Pearson's ρ from − 0.87 to − 0.95; red dots in Figure 2b). We note that, in both cases, the correlation with the corrected N contributes to a more linear relationship between log-transformed N and μ.

We further investigated whether the bias in the estimates of N could be related to major diversification events in bacterial evolution. To address this, we conducted an analysis of variance to examine the bias across taxonomic levels (from genera to phylum) among the 30 bacterial species. The analysis revealed no significant associations (Table S1), indicating that the observed bias does not exhibit phylogenetic structure.

3.3 | Testing a Bayesian estimator of the effective population size with recurrent mutations

The estimates of N using the BM model showed a notable bias in the majority of bacterial species. To tackle this bias, we developed a Bayesian estimator for N using the sampled SFS under the RM model, χ^{RM}. It is important to note that since our approach to estimating N relies on the SFS, the estimator likely captures a historical N rather than a contemporary one. Furthermore, we assumed unbiased mutations, a simplification that allows us to derive the closed-form expression for χ^{RM} in Equation (10) and avoid the summation over N in Equations (7) and (8), which is computationally time-consuming and, for large N species such as bacteria, prone to numerical errors. It is important to note that we demonstrated that mutation biases do not significantly impact the predictions of the SFS (Figure S2).

To assess the precision and accuracy of the Bayesian estimator, we conducted simulations with χ^{RM}, varying the number of sampled individuals and genomic positions. We opted to evaluate this estimator for values of N typically observed in bacteria, where assuming RM is crucial. Figure 3 shows that the estimator delivers accurate and precise estimates of N across all simulated scenarios, even with just five sampled individuals. Notably, the precision of the estimator varies more with the number of sampled genomic positions than with the number of individuals (comparing the width of the credible intervals between lines one and two; Figure 3). The sampling conditions for achieving accurate estimates of N with the Bayesian estimator align well with current population genomics studies, which typically offer millions of genomic positions while sequencing only a limited number of individuals. A single MCMC chain of 100,000 generations was completed in 3 s across all tested scenarios (Figure S4) on a Linux machine with an Intel Core i7–7500U CPU (2.70 GHz).

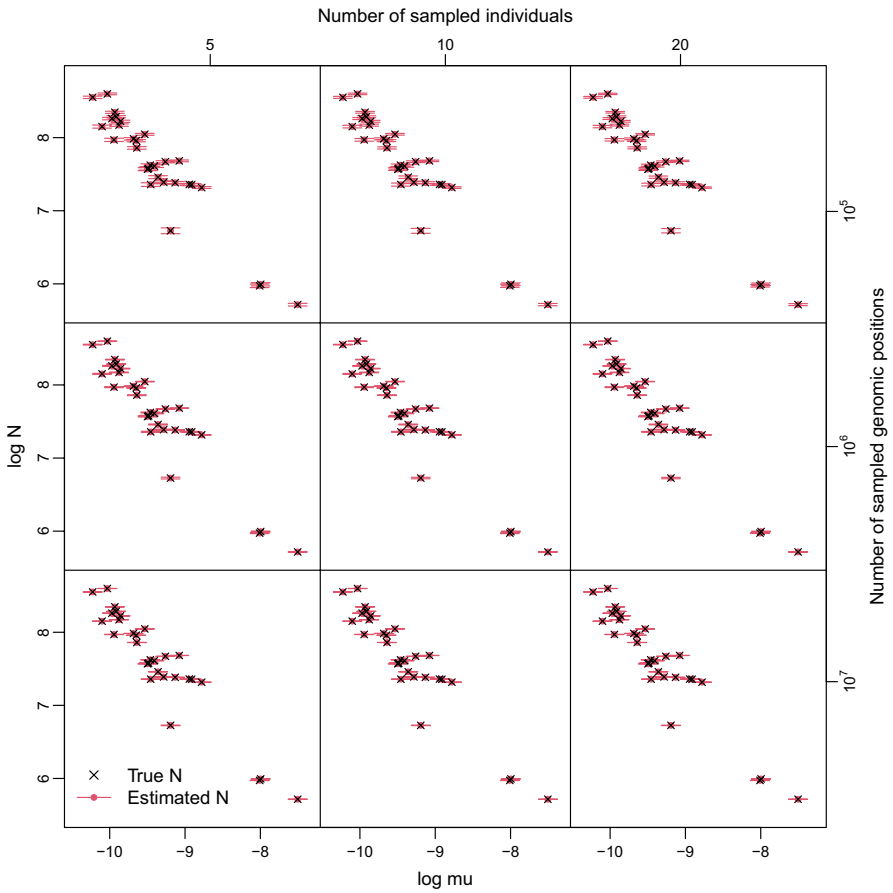

FIGURE 3 Assessing the quality of the effective population size estimator with recurrent mutations. We evaluated the precision and accuracy of the Bayesian estimator for N under the RM model across varying numbers of sampled individuals and genomic positions. The red dots represent the posterior means and the horizontal bars the 99% credible intervals. Each dot corresponds to a bacterial species.

3.4 | Measuring the effective population size of codon positions in *Pseudomonas fluorescens*

Codons exhibit complex evolutionary histories influenced by multiple factors. These include constraints imposed at the nucleotide level, such as GC-biased gene conversion (Duret & Galtier, 2009), codon level, such as codon usage biases on synonymous codons (Hershberg & Petrov, 2008), as well as structural and functional constraints at the protein level (DePristo et al., 2005). These forces interact in intricate ways, creating difficult-to-explain but often characteristic patterns of diversity across the three codon positions. Specifically, the first and second codon positions are typically more conserved, while the third position is often more variable (Bofkin & Goldman, 2006; Shapiro et al., 2005).

To illustrate the implications of assuming boundary and recurrent mutations beyond merely measuring the intensity of genetic drift, we investigated signatures of constrained evolution among the three codon positions. The idea is to use N as a proxy for the effectiveness of selection, a strategy that is frequency found in the literature (e.g., Gossmann et al., 2011). Assuming a constant μ, we expect less selectively constrained evolutionary units to exhibit higher variability and estimate a larger N. Conversely, more constrained units should estimate a lower N. We conducted this experiment in the bacterium *Pseudomonas fluorescens*, a common Gram-negative, rod-shaped microorganism, because it exhibited a particularly poor fit to the BM model (highlighted by a red arrow in Figure 2a).

The analyses revealed that the N estimated by the RM model in the third position (2.623×10^9 with a 99% credible interval of $[2.610, 2.636] \times 10^9$) is 10 times larger compared to the first position and 40 times larger than the second position (Figure 4a). This finding is consistent with prior studies by Bofkin and Goldman (2006) and Yang (1996), supporting the idea that the second codon position is highly functionally constrained due to any alteration leading to a non-synonymous change. In contrast, the third codon position is the least functionally constrained, evolving closer to the mutational process. When applying the BM model, a similar pattern emerges. However, the estimated N exhibited more pronounced variations among codon positions, with the N of the third codon position being 23 times larger than the first position and a striking 146 times larger than the second position (Figure 4a; note that the y-axis is in log scale).

It is noteworthy that the N of the third codon position stands out between the two models, being four times larger in the BM model than in the RM model (Figure 4a). This is because the third position shows a significantly higher proportion of polymorphisms compared to the first and second positions (43%, 7% and 2% respectively; Table S2), showing levels of diversity that the BM model fails to explain. The different fit of the sampled SFS of the two models becomes clear when compared to the empirical frequencies. The RM model exhibits a better fit (log-likelihoods of −1,410,991 compared to −1,617,215 in the BM model) and more accurately predicts the frequencies of observed monomorphic and polymorphic counts (Figure 4b).

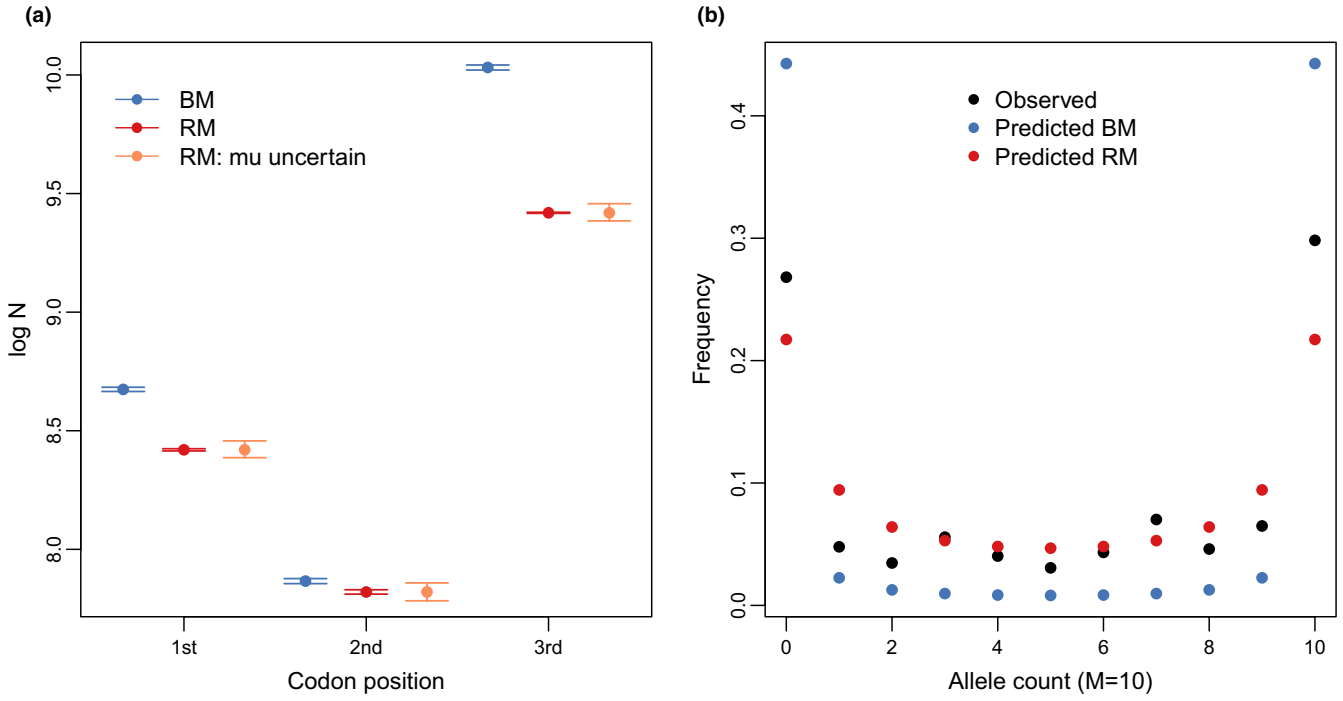

FIGURE 4 Quantifying the effective population size across codon positions in the bacterium *Pseudomonas fluorescens*. (a) Using genomic data from 10 strains and 2066 core genes of *P. fluorescens*, we estimated N across the three codon positions. We applied both the BM and RM models with a punctual estimate of μ, and additionally, an RM model that considers uncertainty in μ. For these analyses, we used the μ reported by Long, Miller, et al. (2018) as 9.31×10^{-11} with a standard error of 0.66×10^{-11}. The dots represent the posterior means and the horizontal bars the 99% credible intervals. (b) Comparison between the empirical and predicted sampled SFS on the third codon position of 10 strains of *P. fluorescens*.

The previously mentioned estimates for N were based on a single value of μ. However, according to Long, Miller, et al. (2018), the mutation rate for *P. fluorescens* is $9.31 \pm 0.66 \times 10^{-11}$. Given the adoption of a Bayesian framework, this error interval surrounding μ can be seamlessly integrated into the inferences. We observed that incorporating the uncertainty in μ results in broader credible intervals, with their width increasing by 4%–17% (Figure 4a). It is worth mentioning that considering uncertainty in μ proves valuable not only in incorporating technical errors in the measurement of μ—such as those arising from sequencing techniques (Salk et al., 2018)—but also in the broader context of meta-analysis, as it allows for the combination of μ from various independent studies. We further note that this addition does not compromise the computational performance of the estimator (Figure S4).

4 | DISCUSSION

In this study, we conducted a comparative analysis of two mutation models in population genetics: one assuming boundary mutations and the other recurrent mutations. Although these models share similarities, they differ in that the BM model only permits mutations in monomorphic or fixed states, neglecting their contribution to frequency shifts between polymorphic states. Our analyses revealed that this assumption holds true for the majority of species across the Tree of Life, with bacteria being a notable exception. The observation that the BM model fits well for most species is significant because the BM assumption leads to simple and easily extendable models. This model possesses notable analytical properties, facilitating the derivation of models that explicitly calculate equilibrium distributions, substitution rates and other informative quantities (Schrempf et al., 2016; Vogl & Clemente, 2012; Vogl & Mikula, 2021). Moreover, the BM model can seamlessly incorporate various modes of selection, such as directional and balancing selection (Borges et al., 2019; Braichenko et al., 2024; De Maio et al., 2015) and it has also found several applications in phylogenetic contexts, contributing to the development of efficient methods for phylogenetic tree inference (Borges, Boussau, Höhna, et al., 2022; Borges, Boussau, Szöllösi, et al., 2022).

In highly diverse organisms, our findings show that the BM model does not offer an accurate fit, and constraining mutations to fixed states can be problematic (Figure 1). This result aligns with other studies questioning the general validity of the IS assumption across bacteria, viruses and even in modelling intra-tumour heterogeneity during cancer evolution (e.g. Desai & Plotkin, 2008; Kuipers et al., 2017; McVean, 2002). Particularly, and perhaps more importantly, we have demonstrated a tendency for N to be overestimated when assuming the BM model, especially in populations with significant genetic diversity (Figure 2a). To compensate for the absence of recurrent mutations, the BM model inflates N so that more mutations are generated and the observed levels of diversity are explained. However, we note that this adjustment is insufficient, as demonstrated in the case study analysed here. Despite estimating a larger N, the BM model still underestimates the markedly high levels of polymorphism in the third codon position of the bacterium *P. fluorescens* (Figure 4b).

In bacteria, this overestimation of N is particularly noticeable, being more than five times larger than the true N in some species (Figure 2b). This discrepancy holds implications for bacterial studies, as they heavily rely on the quantification of N. N provides crucial information on the rate of genetic drift, efficiency of selection and evolution of μ (Lynch et al., 2016). The reported bias suggests that we might be downplaying the significance of genetic drift in bacterial evolution. As N influences the rate of evolution (Lanfear et al., 2014), we might expect that inferred evolutionary relationships and distances in bacteria could change if mutation models accounting for recurrent mutations were used. An analysis of variance showed that the observed bias in N does not exhibit phylogenetic signal across taxonomic levels (Table S1), suggesting it is unlikely to significantly alter previous phylogenetic analyses in bacteria. However, this result should be interpreted with caution, as it is based on a sample of only 30 bacterial species, which do not comprehensively represent the vast diversity within the bacterial domain. Further testing with larger datasets and proper phylogenetic methods is therefore required.

Accurately estimating N is of pivotal importance to ensure the discovery of statistically significant and robust relationships between N and other features of bacterial evolution. For instance, N plays a crucial role in pangenomic evolution (Andreani et al., 2017; McInerney et al., 2017) and may be the primary force influencing genome size and fluidity (i.e. the degree of dissimilarity of genomes evaluated at the gene level) in prokaryotes (Andreani et al., 2017; Sela et al., 2016). Additionally, N has been found to be closely tied to ecological lifestyles and life-history traits in bacteria (Bobay & Ochman, 2018; Woolfit, 2003). While we did not revisit these findings in light of the reported bias in N, we offer a Bayesian estimator for N with recurrent mutations that can be readily used in future studies. The estimator, which requires only allele frequencies, is versatile and can be applied to analyse pooled sequencing data from experimental evolution studies (e.g. Barrick & Lenski, 2013) or metagenomic recruitment data from microbial samples (e.g. Nayfach & Pollard, 2016). Moreover, the estimator is computationally efficient (Figure S4) and performed well across various sampling scenarios (Figure 3). We showed that it not only allows for the estimation of the drift rate but also enables the indirect measurement of selection efficiency, as exemplified in the case study. The findings of this case study also present a valuable opportunity to underscore the importance of incorporating realistic mutation models in population genetics. The BM model suggested much more pronounced differences in the selective constraints among the three codon positions, which the RM model refuted (Figure 4a).

We acknowledge that the proposed estimator, despite considering recurrent mutations, remains rather simplistic. For instance, it does not factor in mutation biases nor selection. This estimator, like many other strategies to estimate N (Tajima, 1989; Wang, 2005; Watterson, 1975), relies on segregating alleles at neutral sites. However, especially in bacteria, pinpointing strictly neutral sites can be challenging because selection might be acting across every nucleotide position due to their large N (Rocha, 2018; Rocha & Feil, 2010). On the one hand, unaccounted-for directional selection (e.g. as a form of linked and background selection or nucleotide usage

biases (Charlesworth et al., 1993; Long, Sung, et al., 2018; Smith & Haigh, 1974)) leads to an underestimation of *N*, as demonstrated in the case study. On the other hand, balancing selection might result in an overestimation of *N*. Given our current limited understanding of the contribution of these selection modes to genome evolution, evaluating the correctness of the current estimates of *N* remains elusive.

However, we must add that the correlation between logμ and log*N* in bacteria is stronger (from − 0.87 to − 0.95; Figure 2b) after correcting *N* for recurrent mutations. Given that selection is pervasive in bacteria and our estimator is agnostic to it, we expected unaccounted-for signatures of selection to significantly contribute to the patterns of genetic diversity in bacteria and in turn dilute the relationship between μ and *N*. By recapturing a better correlation after accounting for recurrent mutations, we have evidence that the patterns of bacterial genetic diversity are largely explained by mutation. This observation also shows that simplistic assumptions about the mutational process can significantly downplay the evolutionary significance of mutations in the evolutionary process itself.

To conclude, this study shows that accounting for recurrent mutations is crucial in highly diverse species and their omission in population genetics models can bias estimates of the rate of genetic drift and other processes whose intensity depends on this rate, such as selection. Future studies are imperative to assess the joint influence of recurrent mutations and selection, which can be achieved, for example, by jointly modelling the distribution of fitness effects.

AUTHOR CONTRIBUTIONS

Rui Borges conceived the ideas and designed methodology, analysed the data and wrote the manuscript.

ACKNOWLEDGEMENTS

We thank the two anonymous reviewers, Michael Lynch and Carolin Kosiol, for their valuable feedback on the manuscript. WolframMath was used to validate some of the mathematical results in this paper, and ChatGPT was employed to improve the clarity of the text.

FUNDING INFORMATION

This research was funded in total by the Austrian Science Fund (FWF) (10.55776/P34524 and 10.55776/P37050) to Rui Borges.

CONFLICT OF INTEREST STATEMENT

The author declares no conflicts of interest.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/2041-210X.14501.

DATA AVAILABILITY STATEMENT

The data utilized and produced in the study, along with the code and tutorials for the Bayesian estimators of the effective population size, are available via https://doi.org/10.5281/zenodo.14621213 (Borges, 2025).

ORCID

Rui Borges https://orcid.org/0000-0002-5905-3778

REFERENCES

Andreani, N. A., Hesse, E., & Vos, M. (2017). Prokaryote genome fluidity is dependent on effective population size. *The ISME Journal*, *11*(7), 1719–1721.

Barrick, J. E., & Lenski, R. E. (2013). Genome dynamics during experimental evolution. *Nature Reviews Genetics*, *14*(12), 827–839.

Bobay, L. M., & Ochman, H. (2018). Factors driving effective population size and pan-genome evolution in bacteria. *BMC Evolutionary Biology*, *18*(1), 153.

Bofkin, L., & Goldman, N. (2006). Variation in evolutionary processes at different codon positions. *Molecular Biology and Evolution*, *24*(2), 513–521.

Borges, R. (2025). mrborges23/mutation_models: Bayesian estimators for the effective population size. *Zenodo*, https://doi.org/10.5281/zenodo.14621214

Borges, R., Boussau, B., Höhna, S., Pereira, R. J., & Kosiol, C. (2022). Polymorphism-aware estimation of species trees and evolutionary forces from genomic sequences with RevBayes. *Methods in Ecology and Evolution*, *13*(11), 2339–2346.

Borges, R., Boussau, B., Szöllösi, G. J., & Kosiol, C. (2022). Nucleotide usage biases distort inferences of the species tree. *Genome Biology and Evolution*, *14*(1), evab290.

Borges, R., Szöllösi, G. J., & Kosiol, C. (2019). Quantifying GC-biased gene conversion in great ape genomes using polymorphism-aware models. *Genetics*, *212*(4), 1321–1336.

Braichenko, S., Borges, R., & Kosiol, C. (2024). Polymorphism-aware models in RevBayes: Species trees, disentangling balancing selection, and GC-biased gene conversion. *Molecular Biology and Evolution*, *41*(7), msae138.

Charlesworth, B. (2009). Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, *10*(3), 195–205.

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*(4), 1289–1303.

De Maio, N., Schrempf, D., & Kosiol, C. (2015). PoMo: An allele frequency-based approach for species tree estimation. *Systematic Biology*, *64*(6), 1018–1031.

DePristo, M. A., Weinreich, D. M., & Hartl, D. L. (2005). Missense meanderings in sequence space: A biophysical view of protein evolution. *Nature Reviews Genetics*, *6*(9), 678–687.

Desai, M. M., & Plotkin, J. B. (2008). The polymorphism frequency spectrum of finitely many sites under selection. *Genetics*, *180*(4), 2175–2191.

Duret, L., & Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. *Annual Review of Genomics and Human Genetics*, *10*(1), 285–311.

Ellen Baake, R. B. (2008). Ancestral processes with selection: Branching and Moran models. *Banach Center Publications*, *80*(1), 33–52.

Etheridge, A. M., Griffiths, R. C., & Taylor, J. E. (2010). A coalescent dual process in a Moran model with genic selection, and the lambda coalescent limit. *Theoretical Population Biology*, *78*(2), 77–92.

Fisher, R. A. (1930). *The genetical theory of natural selection*. Clarendon Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*(4), 457–472.

Gossmann, T. I., Woolfit, M., & Eyre-Walker, A. (2011). Quantifying the variation in the effective population size within a genome. *Genetics*, *189*(4), 1389–1402.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hershberg, R., & Petrov, D. A. (2008). Selection on codon bias. *Annual Review of Genetics*, *42*(1), 287–299. https://doi.org/10.1146/annurev.genet.42.110807.091442

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*(2), 183–201.

Kimura, M. (1955a). Solution of a process of random genetic drift with a continuous model. *Proceedings of the National Academy of Sciences of the United States of America*, *41*(3), 144–150.

Kimura, M. (1955b). Stochastic processes and distribution of gene frequencies under natural selection. *Cold Spring Harbor Symposia on Quantitative Biology*, *20*, 33–53.

Kimura, M. (1969). The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, *61*(4), 893–903.

Kuipers, J., Jahn, K., Raphael, B. J., & Beerenwinkel, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, *27*(11), 1885–1894.

Lanfear, R., Kokko, H., & Eyre-Walker, A. (2014). Population size and the rate of evolution. *Trends in Ecology & Evolution*, *29*(1), 33–41.

Long, H., Miller, S. F., Williams, E., & Lynch, M. (2018). Specificity of the DNA mismatch repair system (MMR) and mutagenesis bias in bacteria. *Molecular Biology and Evolution*, *35*(10), 2414–2421.

Long, H., Sung, W., Kucukyildirim, S., Williams, E., Miller, S. F., Guo, W., Patterson, C., Gregory, C., Strauss, C., Stone, C., Berne, C., Kysela, D., Shoemaker, W. R., Muscarella, M. E., Luo, H., Lennon, J. T., Brun, Y. V., & Lynch, M. (2018). Evolutionary determinants of genome-wide nucleotide composition. *Nature Ecology & Evolution*, *2*(2), 237–240.

Lynch, M., Ackerman, M. S., Gout, J. F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*(11), 704–714.

Lynch, M., Ali, F., Lin, T., Wang, Y., Ni, J., & Long, H. (2023). The divergence of mutation rates and spectra across the tree of life. *EMBO Reports*, *24*(10), e57561.

McInerney, J. O., McNally, A., & O'Connell, M. J. (2017). Why prokaryotes have pangenomes. *Nature Microbiology*, *2*(4), 17040.

McVean, G. A. T. (2002). A genealogical interpretation of linkage disequilibrium. *Genetics*, *162*(2), 987–991.

Moran, P. A. P. (1958). Random processes in genetics. *Mathematical Proceedings of the Cambridge Philosophical Society*, *54*(1), 60–71.

Muirhead, C. A., & Wakeley, J. (2009). Modeling multiallelic selection using a Moran model. *Genetics*, *182*(4), 1141–1157.

Nayfach, S., & Pollard, K. S. (2016). Toward accurate and quantitative comparative metagenomics. *Cell*, *166*(5), 1103–1116.

Page, A. J., Cummins, C. A., Hunt, M., Wong, V. K., Reuter, S., Holden, M. T. G., Fookes, M., Falush, D., Keane, J. A., & Parkhill, J. (2015). Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics*, *31*(22), 3691–3693.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Rocha, E. P. C. (2018). Neutral theory, microbial practice: Challenges in bacterial population genetics. *Molecular Biology and Evolution*, *35*(6), 1338–1347.

Rocha, E. P. C., & Feil, E. J. (2010). Mutational patterns cannot explain genome composition: Are there any neutral sites in the genomes of bacteria? *PLoS Genetics*, *6*(9), e1001104.

Salk, J. J., Schmitt, M. W., & Loeb, L. A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nature Reviews Genetics*, *19*(5), 269–285.

Schrempf, D., Minh, B. Q., De Maio, N., von Haeseler, A., & Kosiol, C. (2016). Reversible polymorphism-aware phylogenetic models and their application to tree inference. *Journal of Theoretical Biology*, *407*, 362–370.

Seemann, T. (2014). Prokka: Rapid prokaryotic genome annotation. *Bioinformatics*, *30*(14), 2068–2069.

Sela, I., Wolf, Y. I., & Koonin, E. V. (2016). Theory of prokaryotic genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(41), 11399–11407.

Shapiro, B., Rambaut, A., & Drummond, A. J. (2005). Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. *Molecular Biology and Evolution*, *23*(1), 7–9.

Smith, J. M., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, *23*(1), 23–35.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, *123*(3), 585–595.

Vogl, C., & Clemente, F. (2012). The allele-frequency spectrum in a decoupled Moran model with mutation, drift, and directional selection, assuming small mutation rates. *Theoretical Population Biology*, *81*(3), 197–209.

Vogl, C., & Mikula, L. C. (2021). A nearly-neutral biallelic Moran model with biased mutation and linear and quadratic selection. *Theoretical Population Biology*, *139*, 1–17.

Wang, J. (2005). Estimation of effective population sizes from data on genetic markers. *Philosophical Transactions of the Royal Society, B: Biological Sciences*, *360*(1459), 1395–1409.

Watterson, G. A. (1975). On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, *7*(2), 256–276.

Winsor, G. L., Griffiths, E. J., Lo, R., Dhillon, B. K., Shay, J. A., & Brinkman, F. S. L. (2015). Enhanced annotations and features for comparing thousands of Pseudomonasgenomes in the pseudomonas genome database. *Nucleic Acids Research*, *44*(D1), D646–D653.

Woolfit, M. (2003). Increased rates of sequence evolution in endosymbiotic bacteria and fungi with small effective population sizes. *Molecular Biology and Evolution*, *20*(9), 1545–1555.

Wright, S. (1931). Evolution in mendelian populations. *Genetics*, *16*(2), 97–159.

Yang, Z. (1996). Maximum-likelihood models for combined analyses of multiple sequence data. *Journal of Molecular Evolution*, *42*(5), 587–596.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Text S1: Obtaining the stationary distribution of the recurrent mutation model.

Table S1: Analysis of the variance in the bias between the estimates of N under the BM and RM models across bacterial taxonomic levels.

Table S2: The number of monomorphic and polymorphic sites for the three codon positions of *Pseudomonas florescens*.

Figure S1: Comparison of effective population size estimates.

Figure S2: Comparison of site frequency spectra between the recurrent and boundary mutation models with mutation bias.

Figure S3: Misspecifying the boundary mutation model with recurrent mutations.

Figure S4: Computational performance of the Bayesian estimators.

How to cite this article: Borges, R. (2025). The rarity of mutations and the inflation of bacterial effective population sizes. *Methods in Ecology and Evolution*, *16*, 722–732. https://doi.org/10.1111/2041-210X.14501