PLOS GENETICS

RESEARCH ARTICLE

Interpreting the pervasive observation of U-shaped Site Frequency Spectra

Fabian Freund[1,2☯], **Elise Kerdoncuff**[3,10☯], **Sebastian Matuszewski**[4], **Marguerite Lapierre**[10], **Marcel Hildebrandt**[5], **Jeffrey D. Jensen**[6], **Luca Ferretti**[7], **Amaury Lambert**[8,10], **Timothy B. Sackton**[9], **Guillaume Achaz**[10,11]*

1 Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany, **2** Department of Genetics and Genome Biology, University of Leicester, Leicester, United Kingdom, **3** Department of Genetics, University of California, Berkeley, California, United States of America, **4** Accenture, Vienna, Austria, **5** Siemens AG, Munich, Germany, **6** Center for Evolution & Medicine, School of Life Sciences, Arizona State University, Tempe, Arizona, United States of America, **7** Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, United Kingdom, **8** Institut de Biologie de l'ENS (IBENS), École Normale Supérieure, Paris, France, **9** Informatics Group, Harvard University, Cambridge, Massachusetts, United States of America, **10** SMILE group, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, Paris, France, **11** Éco-anthropologie, Muséum National d'Histoire Naturelle, Université Paris-Cité, Paris, France

☯ These authors contributed equally to this work.
* guillaume.achaz@mnhn.fr

Citation: Freund F, Kerdoncuff E, Matuszewski S, Lapierre M, Hildebrandt M, Jensen JD, et al. (2023) Interpreting the pervasive observation of U-shaped Site Frequency Spectra. PLoS Genet 19(3): e1010677. https://doi.org/10.1371/journal.pgen.1010677

Editor: Lindi Wahl, University of Western Ontario, CANADA

Received: May 13, 2022

Accepted: February 22, 2023

Published: March 23, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pgen.1010677

Data Availability Statement: All used genomic data are publicly available. All processed material is provided in the supplementary (e.g. tables and

Abstract

The standard neutral model of molecular evolution has traditionally been used as the null model for population genomics. We gathered a collection of 45 genome-wide site frequency spectra from a diverse set of species, most of which display an excess of low and high frequency variants compared to the expectation of the standard neutral model, resulting in U-shaped spectra. We show that multiple merger coalescent models often provide a better fit to these observations than the standard Kingman coalescent. Hence, in many circumstances these under-utilized models may serve as the more appropriate reference for genomic analyses. We further discuss the underlying evolutionary processes that may result in the widespread U-shape of frequency spectra.

Author summary

This study investigates the assumed universality of the standard neutral model of molecular evolution. We demonstrate that genealogical models alternative to the widely used Kingman coalescent often provide greatly improved fits to observed genome-wide allele frequency data for taxa sampled widely from across the tree of life. As such, we argue that these more generalized multiple merger models (which contain the Kingman coalescent as a special case) may prove more fruitful and appropriate in future population genomic studies. Importantly, this modification of the standard model for interpreting genetic diversity has potentially profound implications for many population genetic inference approaches (e.g., scanning for targets of selection

plots) and in our code and data repository https://github.com/fabfreund/usfs_mmc.

Funding: We thank NSF for support through the DEB-1754397 to Timothy B. Sackton and DFG for support through FR 3633/2-1 (within Priority Program 1590: Probabilistic Structures in Evolution) to Fabian Freund. Jeffrey D. Jensen was supported by National Institutes of Health grant R35GM139383 The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: None.

across the genome and reconstructing population history), as well as for analyses in related fields.

1 Introduction

The Kingman coalescent, [1], a stochastic process describing the distribution of random, bifurcating genealogical trees in a Wright-Fisher population, has been enormously impactful in the study of natural genetic variation in populations [2]. Under the standard neutral theory [3, 4], the coalescent can be used to derive expectations of neutral diversity by tracking mutations along the branches of random genealogies, and extensions can accommodate complex processes such as recombination [5], population structure [6], and natural selection [7]. The power of this approach relies on being able to compare deviations observed in real data from expectations under the coalescent model.

One common metric used to study the consistency between the assumptions of this model and the observed data is the *Site Frequency Spectrum* (SFS)—that is, the distribution of mutational frequencies, typically computed for a sample of n haploid genomes. Under the assumptions of the Standard Neutral Model (SNM)—including constant population size and panmixia—the expected SFS, averaged across the tree space, is given by $E[\xi_i] = \theta/i$, where ξ_i is the number of sites that carry a derived variant of frequency i/n [8]. The θ parameter of the SNM is defined as $\theta = 2pN\mu$, where p is the ploidy (typically 1 or 2), N the (effective) population size, and μ the mutation rate.

Observed SFS in natural populations are often poorly fit by this expectation, owing to violations of one or more of the underlying assumptions of the SNM, including varying population sizes, population structure, direct selection, and linkage with selected sites [9]. A standard procedure in population genetics is thus to first statistically test for the SNM (treated as H_0, a null statistical model) and then, when rejected, fit a variety of alternative demographic and/or selection models.

In this article, we show that among a collection of genome-wide SFS from a diverse set of species, many show an unexpected excess of low and high frequency variants, resulting in a U-shaped SFS. Many possible factors may result in such a pattern of variation. These include recent migration from non-sampled populations [10], population structure [11], misorientation of ancestral and derived alleles [12], biased gene conversion [13], recent positive selection at many targets across the genome [14], background selection [15, 16], temporally-fluctuating selection [17], and various reproductive strategies [18].

A number of these scenarios result in an important general violation of Kingman assumptions: the presence of multiple mergers in genealogies (*i.e.*, a node with more than two descendants). Under such scenarios, these distributions are better described by a more general class of models known as the Multiple Merger Coalescent (MMC) [19–23]. Briefly, MMCs may arise when the number of offspring per individual has very high variance, even up to the order of the total population size. Such effects of concentrations of ancestrality (resulting in polytomies in the trees) have been reported in various species across all kingdoms of life [24], and MMC-like genealogies have been observed for species ranging from bacteria (e.g. for *Mycobacterium tuberculosis* [25, 26]) to viruses (e.g. for influenza [27]) to animals (e.g. for the nematode *Pristionchus pacificus* [28], multiple fish species, e.g. [29–31]) and even to cancer cells [32].

Compared to the Kingman coalescent, MMC trees have different distributions of both the branch lengths and the number of lineages that coalesce in each node. They mostly occur when the distribution of offspring numbers is highly variable: a recurring ancestor of a

substantial fraction of the population for the Ψ-coalescent or less specifically an inflated variance for the Beta-coalescent. The most extreme scenario is the presence of a single ancestor for the whole population, resulting in a star-shaped tree where a single node collapses all branches. MMC trees tend, like all star-like trees (e.g. Kingman-like trees in expanding populations), to have an excess of low frequency variants (e.g. derived singletons). Furthermore, the root MRCA node of MMC trees is more often imbalanced than it is for Kingman trees. Imbalanced trees nodes have most leaves on one side while few on the other. As a consequence, MMC trees also display an excess of ancestral rare alleles (e.g. ancestral singletons). Both effects jointly produce a U-shaped SFS (for more details refer to A.1 in S1 Appendix).

Multiple neutral and selective processes can produce MMC genealogies in natural populations. Generally, the term sweepstake reproduction has been proposed for species that have rare individuals with a high reproduction rate coupled with high early-life mortality. In these species, a single or few individuals can become ancestors of a substantial fraction of the population by chance, thus resulting in MMC genealogies (for a review, see [33]). Multiple models featuring the recurrent and rapid emergence of genotypes with high fitness also result in MMC genealogies, often modeled by the Bolthausen-Sznitman coalescent or related models, e.g. [34–38]. Importantly, other biological factors can also lead to MMC-like genealogies, including large rapid demographic deviations [39], seed banks [40], extinction-recolonisation in metapopulations [41] and range expansions [42]. Yet, the frequency of MMC genealogies in nature, and more generally whether MMC models ought to be employed as a more appropriate null for certain species, remains an open question.

In this study, we collected SFS from 45 species (Table 2) from across the tree of life (bacteria, plants, invertebrates and vertebrates), for which genome-wide polymorphism data (with sample sizes of $n \geq 10$) were available together with an outgroup to assign ancestral and derived states. We show that MMC genealogies provide a better fit than the Kingman coalescent in many cases, even when both are combined with non-constant demography and misorientation of ancestral and derived alleles. For several species, the fit is excellent. For each species, we tested two simple MMC models: Beta-MMC [43] and Psi-MMC [44], both tuned by a single parameter that interpolates between a star-shaped tree (*i.e.* a single radiation) to a Kingman-like tree. Demography is here tuned by a single parameter (a simple exponential growth), as is the frequency of misorientation errors. Using composite-likelihood maximization [45] on genome-wide data, we explore statistical power to distinguish between these contributing factors. Finally, we discuss how MMCs may be better utilized in future population genetic analysis, and what evolutionary forces may contribute to the pervasive observation of U-shaped SFS.

2 Materials and methods

2.1 Coalescent and allele misorientation models

We compared the empirically observed SFS to the theoretical SFS expected under a variety of models. The genealogical models emerge from a discrete generation reproduction model. Each is a (random) tree with n leaves which approximates the genealogy for a sample of size n in a reproduction model in which the population size N is very large ($N \to \infty$). One unit of time in the coalescent tree corresponds to many generations in the underlying reproduction model: for Kingman's coalescent one time unit corresponds to N generations of a haploid Wright-Fisher model, or order of N^2 time steps of an haploid Moran model. This correspondence affects how population size changes are reflected in the coalescent approximation (see definition below, for mathematical justification and details see [46–48]). On the genealogical tree, mutations are placed randomly via a Poisson process with rate $\theta/2$.

We compared three coalescent models: Kingman's n-coalescent, Psi-n-coalescent (also called Dirac-n-coalescent) with parameter $\Psi \in [0, 1]$ and Beta$(2-\alpha,\alpha)$-n-coalescent with $\alpha \in [1, 2]$. The parameters α or Ψ regulate the strength and frequency of multiple mergers: the smaller α or the larger the Ψ, the more frequently coalescence events are multiple mergers of increasing size. Both MMCs incorporate Kingman's n-coalescent as a special case ($\alpha = 2$ or $\Psi = 0$).

Both MMC coalescent models can be defined for demographic variation that stays of the same order, *i.e.* where the populations size ratio $v_t = N_t/N_0$ of the population size at time t in the past (in coalescent time units) is positive and finite (for large population sizes N). The coalescent merges any k of b (ancestral) lineages present at a time t with rate

$$\lambda_{n,k}(t) = v(t)^{-\eta} \int_0^1 x^{k-2}(1-x)^{n-k}\Lambda(dx), \tag{1}$$

where

- Λ could be any probability distribution on [0, 1] but is here either the Dirac distribution (point mass) in Ψ (Psi-coalescent) or the Beta$(2-\alpha,\alpha)$ distribution (Beta coalescent).
- η is a scaling factor reflecting how many time steps from the discrete reproduction model form one unit of coalescent time. More precisely, it is the power of N of the scaling factor: e.g. $\eta = 2$ for a Moran model and $\eta = 1$ for a Wright-Fisher model.

A common way of constructing the Λ-coalescent, which provides a nice interpretation of Eq (1), is the paintbox process [20]: at rate $x^{-2}\Lambda(dx)$ per time unit, paint each lineage independently with probability x and merge all painted lineages simultaneously. Note that when Λ is the Dirac mass at 0, $\lambda_{n,k}(t)$ is nonzero only when $k = 2$, recovering Kingman's coalescent.

We focused on exponentially growing populations, *i.e.* a population size ratio $v(t) = \exp(-gt)$ for growth rate $g \geq 0$ (see A.2 in S1 Appendix for interpretation of g in the initial reproduction model). As underlying reproduction models, we use modified Moran models [44, 47, 49]. At each time step, in a population of size N, a single random individual has $U + G$ offspring while $N - U$ random individuals have 1 offspring (leaving $U - 1$ individuals devoid of offspring). As a consequence, the population grows from N to $N + G$ individuals and G is chosen to fit the desired growth rate.

In a standard Moran model, $U = 2$ and $G = 0$, leaving the population size constant. However, for both MMCs, U is set to different values. In both cases, the mean of U does not grow indefinitely with N (for all parameters α and Ψ), but the resulting variance does (for $\alpha \neq 2$ and $\Psi \neq 0$).

- In the Psi-n-coalescent (essentially [44, 47]), we have $U = 2$, except when a sweepstake event occurs with a small probability of order $N^{-\gamma}$ $(1 < \gamma \leq 2)$; in this case, $U = \lfloor N\Psi \rfloor$. In the coalescent time scale, one unit of time corresponds to an order of N^{γ} time steps; this is the expected time to a sweepstake event so that η must equal γ. We chose $\gamma = \eta = 1.5$ for $\Psi > 0$, and $\gamma = \eta = 2$ for $\Psi = 0$ (standard Moran model) with $U = 2$ in every time step.
- In the Beta-n-coalescent [48, 49], U has distribution $P(U = j) = \lambda_N^{-1}\binom{N}{j}\frac{B(j-\alpha,\alpha+N-j)}{B(2-\alpha,\alpha)}$, where B is the Beta function and λ_N is the normalizing constant. Consequently, although the random variable U has a finite mean of at most $\frac{\alpha}{\alpha-1}$, it can take large values with high probability when $\alpha < 2$. See A.2 in S1 Appendix for more details. On the coalescent time scale, one unit of time corresponds to an order of N^{α} time steps, so $\eta = \alpha$. Note that $\alpha = 2$ is the classical Moran model and thus leads to Kingman's coalescent. We stress that allowing for a

exponentially growing population by setting $G > 0$ in the models above does neither change the order of the time scaling nor the resulting coalescent model, it only introduces a slight further rescaling of time in the coalescent, as reflected in the coalescence rates (Eq. 1).

For statistical inference, we treat the observed SFS of s mutations as s independent multinomial draws from the expected SFS (see [45] and [50, Eq. 11] [47, Eq. 14]). This computes an approximate composite likelihood function of the data for any combination of growth rate (g) and coalescent parameter (α or ψ). However, to include the effect of misorienting the ancestral allele with the derived allele, we introduced another parameter e. On average, a misorientation probability of e lets a fraction e of the derived allele carried by i sequences to be falsely seen as appearing in $n - i$ sequences. Additionally, as described in [51, Section 4.2] or [12, p. 1620], as misorientation stems from double-mutated sites, e also relates to the number of sites that cannot be oriented when compared with the outgroup owing to the presence of a third allele (see A.4 in S1 Appendix). We account for these two effects of e by swapping a fraction e of the variants at frequency i/n to $1 - i/n$ and we assume a Jukes-Cantor substitution model [52] to predict for the number $s_{\neq}$ of non-polarizable tri-allelic variants. This leads to a slight variant of [47, Eq. 14]. For any coalescent model with a specific set of coalescent, exponential growth and misorientation parameters, the pseudolikelihood is:

$$PsL(s_1, \ldots, s_{n-1}, s, s_{\neq}) = \frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E[T_i](1-e) + E[T_{n-i}]e}{E[T_{tot}]} \right)^{s_i} \underbrace{\binom{s + s_{\neq}}{s_{\neq}} \left(\frac{2e}{1+2e} \right)^{s_{\neq}} \left(\frac{1}{1+2e} \right)^{s}}_{\text{from non-polarizable variants}}, \tag{2}$$

where $s_1, \ldots, s_{n-1}$ is the observed SFS (so we observe s_i sites with derived allele frequency i/n), $s = \sum_i s_i$ is the total number of polarizable polymorphic sites and $s_{\neq}$ is the number of non-polarizable sites. $E[T_i]$ is the expected sum of branch lengths that support i leaves in the genealogy and $E[T_{tot}]$ is the sum of all branch lengths. For $e = 0$, we set the term estimated from non-polarizable variants to 1. See A.4 in S1 Appendix for details on the derivation.

2.2 Statistical inference

To find the best-fitting parameters, we conduct a grid-search for the highest pseudolikelihood. The expected branch lengths $E[T_i]$ in Eq (2) are computed as in [47], using the approach from [53]. We use the following grids with equidistant steps

Beta: $\alpha \in [1, 2]$ in steps of 0.05, $g \in [0, 25]$ in steps of 0.5, $e \in [0, 0.15]$ in steps of 0.01.

Psi: $\Psi \in [0, 1]$ in steps of 0.05, g, e as for Beta above, complemented with $\Psi \in [0, 0.2]$ in steps of 0.01 (further expanding $g \in [0, 30]$ by steps of 0.5 and $e \in [0, 0.2]$ by steps of 0.01) when Ψ was estimated to be close to 0.

To perform model selection between the three coalescent models, we computed the two following log Bayes factors:

$$BF_1 = \max(\log \max_{\alpha,g,e} PsL,\ \log \max_{\Psi,g,e} PsL) - \log \max_{\alpha=2,g,e} PsL, \tag{3}$$

$$BF_2 = \log \max_{\alpha,g,e} PsL - \log \max_{\Psi,g,e} PsL \tag{4}$$

from the maximum pseudolikelihoods computed for the three models. We inferred a MMC genealogy when $BF_1 > \log(10)$ and further chose a Beta coalescent or a Psi-coalescent when

(additionally) $BF_2 > \log(10)$ or $BF_2 < -\log(10)$ respectively. These arbitrary thresholds have been extensively tested using simulations (see Results), showing that they empirically point to the right model.

We appreciate that the "Bayes Factors" (BF) are computed here as "log-Likelihood Ratios" (log-LR). Interestingly, any likelihood ratio can be interpreted as a posterior probability ratio, provided that the prior on models is uniform (as it is assumed routinely in Bayesian MCMC sampling) as we do here. Thus, in our case, both denominations are equivalent.

For the best fitting parameter combinations either over the full parameter space or restricted to the Kingman coalescent with growth and allele misorientation (*i.e.*, fixing $\alpha = 2$ or $\Psi = 0$), we assessed the goodness-of-fit of the observed data. First, we graphically compare the observed SFS with the expected SFS, approximated as $\left(\frac{E[T_1]}{E[T_{tot}]}, \ldots, \frac{E[T_{n-1}]}{E[T_{tot}]}\right)$. Second, we quantified the (lack of) fit of the data by Cramér's V, a goodness-of-fit measure which accounts for different sample sizes and different numbers of polymorphic sites. See A.6 in S1 Appendix for details.

2.3 Data

We collected 45 genome-wide SFS that are described in Table 2 and Table G in S1 Appendix. The collected SFS come from public data sets. For 20 data sets, SFS were extracted from whole genome SNP data, including both coding and non-coding regions. For 16 data sets, they were extracted only from transcriptomes (equivalent to coding regions). For 9 bacterial data sets, the SFS were extracted from the core genome. The supplementary material S1 and S2 Figs provide the shapes of the empirically-observed SFS.

3 Results

We first demonstrate the power of the methodology using extensive simulations, and then apply it to 45 real SFS computed from a very large variety of taxa.

3.1 Statistical performance

Using simulations, we first assess the power of the method to retrieve the correct model and then its power to estimate the parameters. Briefly, for each simulation, we simulated 100 independent loci for each parameter combination, choosing different values for the coalescent parameter (α or Ψ), the growth rate of the demographic model (g), and the misorientation probability (e). For each locus, we then simulate SNPs under an infinite sites model, with a mutation rate such that on average 50 sites are segregating for each locus. This simulation setup is described in further detail in A.7 in S1 Appendix.

Applied to the simulated data, our method performs well. Even for small datasets ($n = 25$), the model selection approach based on Bayes factors computed from Eq (2) identifies the correct multiple merger model in most cases (Table 1), as long as multiple mergers occur with reasonable frequency. As the rate of multiple mergers becomes very low ($\alpha \approx 2$ or $\Psi \approx 0$), mis-identifications are more common (Table 1). However, even when our model prefers the beta-coalescent for data simulated with $\alpha = 2$, in 96% of such cases (with $n = 100$; 71% with $n = 20$), we estimate $\alpha \geq 1.9$, suggesting that even when model mis-idenfication occurs, parameter estimation remains reliable (Table C in S1 Appendix). Over the range of parameter combinations, larger sample sizes lead to smaller errors, as expected. This selection approach is conservative with respect to departures from the standard Kingman coalescent, as we choose a Kingman genealogy model if the Bayes factor does not distinctively point towards an MMC model.

Table 1. Model selection via two-step Bayes factor criterion. Based on 2,000 simulations for each true model assuming $n = 25$ individuals with 100 loci with 50 mutations each on average. For each simulation, the coalescent parameter is fixed and the growth parameter g and the allele misorientation rate e are randomly chosen ($g \in [0, 11.25]$, $e \in [0, 0.1]$). The second column shows whether the parameters used for simulation were included in the inference grid. Fractions are rounded to two digits. The maximum of each row is marked in bold. MMC refers to cases in which neither the Psi- nor Beta-coalescent is preferred. An expanded version with enhanced sample size is provided in Table B in S1 Appendix. For details on simulations and inference parameters see A.7 in S1 Appendix.

True model	Within	Fraction model inferred as			
	the grid?	Kingman	Beta	Psi	MMC
$\alpha = 2$	yes	**0.79**	0.21		
$\alpha = 1.9$	yes	0.34	**0.66**		
$\alpha = 1.8$	yes	0.02	**0.91**	0.04	0.03
$\alpha = 1.625$	no		**0.9**	0.06	0.05
$\alpha = 1.025$	no		**1**		
$\Psi = 0.005$	no	**0.55**	0.45		
$\Psi = 0.025$	no	0.05	**0.72**	0.14	0.09
$\Psi = 0.05$	yes		0.12	**0.82**	0.06
$\Psi = 0.075$	no		0.06	**0.91**	0.03
$\Psi = 0.1$	yes		0.02	**0.98**	

https://doi.org/10.1371/journal.pgen.1010677.t001

Parameter estimation within both the Beta- and Psi-coalescent models works well for multi-locus data for large enough samples, especially for the allele misorientation rate e and for the coalescent parameter α or ψ (Fig 1 and Figs A.4–A.6 in S1 Appendix). The growth rate, in contrast, is only estimated well for situations where the simulated growth rate was low (Figs A.11, A.14, A.17 and A.20 in S1 Appendix).

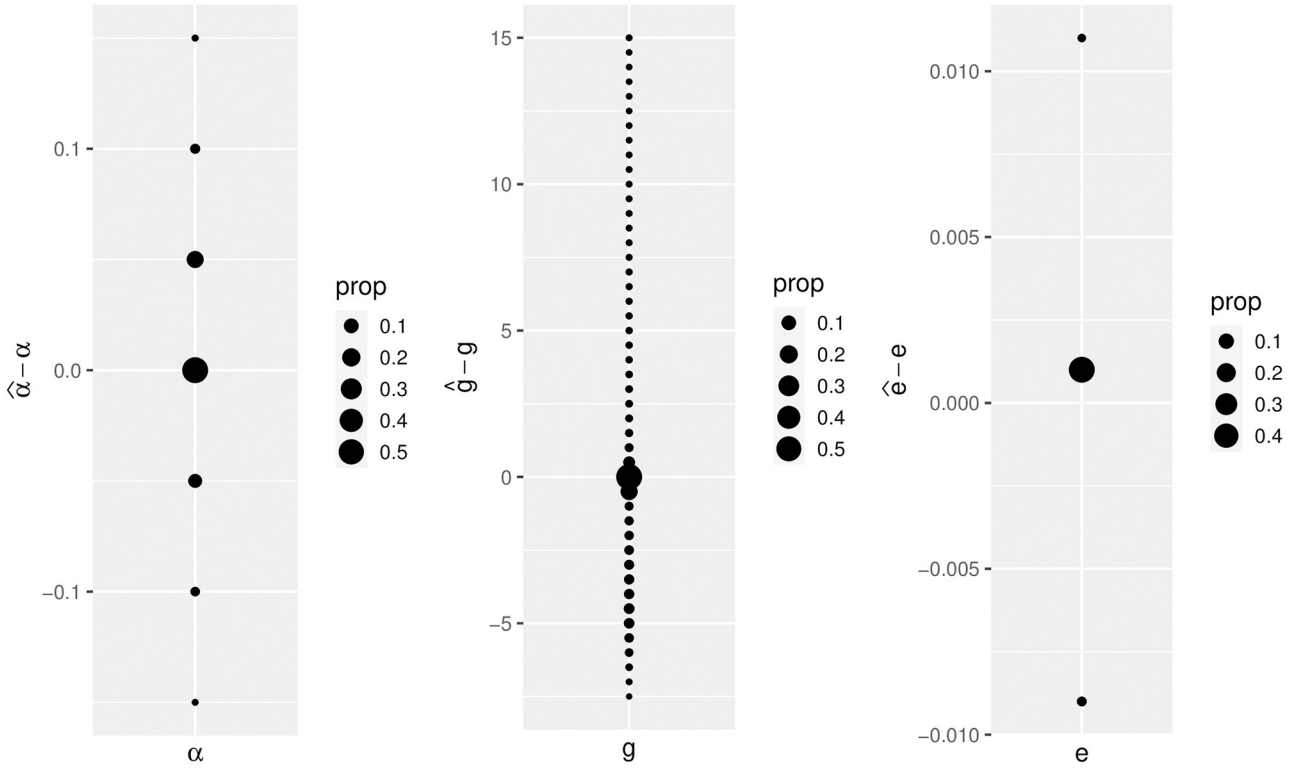

Fig 1. Error for estimating parameters for Beta coalescents with exponential growth and allele misorientation across the parameter grid for (α, g, e). The space between the points stems from the grid. Sample size $n = 100$, 50 independent loci with 100 mutations on average. 500 simulations were performed per parameter triplet.

https://doi.org/10.1371/journal.pgen.1010677.g001

3.2 Data analysis

The simulations demonstrate that the method is able to retrieve the correct model, and also correctly estimate the parameters of the MMC, provided that there is enough signal in the data. Next, we applied the method to 45 real SFS from 45 distantly related taxa. We first tested how many datasets are better fit by an MMC model than by a Kingman model, then tested the goodness of the MMC fit and estimated MMC parameters for real data.

MMC fits better than Kingman. First, we assessed the fit of each SFS to both MMC models and the Kingman coalescent, with exponential growth and misorientation. Using the Bayes Factor criterion, we selected the best fitting model for each empirical SFS in our dataset (Table 2). A large majority (76%) of the SFS produce a better fit to MMC models than to the standard Kingman coalescent model. The best model is most frequently the Beta-coalescent (53%), followed by the Kingman coalescent (24%) and the Psi-coalescent (13%). In a few cases, both MMC models produce a better fit than the coalescent, but we cannot distinguish the best fitting MMC (9%).

MMC is sometimes a good fit. While we show that MMC models produce better fits than the Kingman coalescent across many species, this could be because no model fits well. To test whether the best fit coalescent model is indeed a good model to predict the observed SFS, we calculated Cramér's V, a measure of goodness-of-fit appropriate for variable contingency tables (e.g., SFS with different sample sizes across species, see A.6 in S1 Appendix for details). Combined with visual inspections (all SFS with their fit are provided in supplementary material S1 and S2 Figs), we designed empirical grade categories from 'very accurate' fit to 'very poor' fit, as following: A: $V \in [0 : 0.033[$, B: $V \in [0.033 : 0.066[$, C: $V \in [0.066 : 0.1[$ and D: $V \in [0.1 : \infty[$. Importantly, the MMC models fit well to 67% of data sets: 30/45 SFS have grades A or B on Table 3. This demonstrates that not only is MMC a better choice than Kingman on statistical grounds but also that it appears as a good model to predict patterns of diversity for a large majority of species.

The amount of multiple mergers greatly varies among species. The MMC models we use vary in the extent of multiple mergers, from star-like to Kingman-like, scaled by a single parameter (α and Ψ respectively for the Beta- and Psi-coalescent). To determine whether the model fits suggest an appreciable level of multiple mergers, we next explore the estimated parameters for MMC models. Of the 45 empirical SFS we analyzed, 73% (33/45) have $\hat{\alpha} < 1.9$ under the Beta-coalescent, which suggests a non-trivial frequency of multiple mergers, and implies something that is not captured by the SNM is occurring in these species (see Table D in S1 Appendix for α estimates of all data sets, including those where the Kingman or Psi-coalescent are the best fit model). Nonetheless, estimates of α and Ψ are both skewed towards values that approach the Kingman coalescent (2 and 0, respectively), despite covering the full range of values across the tree of life (Fig 2 and Fig A.9 in S1 Appendix).

Assuming a Kingman coalescent leads to an overestimation of the growth rate. One potential impact of using the standard Kingman coalescent instead of better-fitting MMC models is the incorrect estimation of other parameters, including aspects of demography. To explore this issue, we compared the estimated growth rate and misorientation error assuming a Kingman model rather than an MMC model. We observe that the growth parameters are often higher when inferred under the the Kingman coalescent than in either of the MMC models (Table D in S1 Appendix), although estimates of g tend to converge in empirical datasets where MMC parameter estimate approaches Kingman (Fig A.7A in S1 Appendix). This mirrors previous results of compensating the effect of MMC when inferring under a Kingman coalescent by estimating a higher growth rate in our scenario without allele misorientation, see e.g. [47].

Table 2. Data sets description: Taxa, Species, number of haplotypes (*n*) and number of polymorphic sites (#SNP). Best fitting model (Kingman (KM), Beta, Psi-coalescent or no preference between Multiple Merger Coalescents (MMC)), its parameters (parameters describing coalescence (Coal), growth rate (*g*) and misorientation (*e*)) and goodness-of-fit grade from Cramér's *V* values.

Order	Species	*n*	#SNP	Model	Coal	g_{Model}	e_{Model}	Grade
Vertebrates	*Aptenodytes patagonicus*	20	1,278	Beta	1.25	1.5	0	B
	Athene cunicularia	40	11,268,203	Beta	1.8	1	0.03	B
	Corvus cornix	38	7,551,159	Beta	1.85	0.5	0	A
	Coturnix japonica	20	5,061,864	Beta	1.45	0.5	0.01	A
	Egretta garzetta	10	9,318,499	Beta	1.75	0	0.02	B
	Emys orbicularis	20	515	KM	∅	0.5	0	C
	Ficedula albicollis	24	14,697,230	Ψ	0.01	0.5	0.01	A
	Gorilla gorilla gorilla	54	9,878,547	Beta	1.9	0	0	B
	Homo sapiens	216	19,441,528	Beta	1.85	0	0	A
	Lepus granatensis	20	769	MMC	0.12	0	0.03	C
	Nipponia nippon	16	1,140,694	KM	∅	0	0.03	D
	Pan paniscus	26	6,293,657	Beta	1.85	1	0	B
	Pan troglodytes ellioti	20	10,009,190	Beta	1.7	0	0	A
	Parus major	54	14,174,305	Beta	1.75	0	0.01	A
	Parus caeruleus	20	866	MMC	0.04	0	0.02	B
	Passer domesticus	16	18,501,992	KM	∅	0	0	A
	Phylloscopus trochilus	24	33,401,127	KM	∅	12.5	0	A
	Taeniopygia guttata	38	53,263,038	Beta	1.75	4	0	A
Invertebrates	*Armadillidium vulgare*	20	23,323	Beta	1.7	0	0.03	C
	Artemia franciscana	20	5,548	Beta	1.65	0	0.03	B
	Caenorhabditis brenneri	20	1,339	Beta	1.5	0	0.06	C
	Caenorhabditis elegans	574	165	KM	∅	0	0.06	D
	Ciona intestinalis A	20	1491	Beta	1.9	0	0.03	C
	Ciona intestinalis B	20	2186	Beta	1.6	0	0.02	C
	Culex pipiens	20	5,442	Beta	1.55	0.5	0.01	B
	Drosophila melanogaster	196	4,662,706	Beta	1.65	0.5	0.02	A
	Halictus scabiosae	22	712	MMC	0.04	0	0.01	B
	Melitaea cinxia	18	1,695	Beta	1.7	0.5	0.03	B
	Messor barbarus	20	9,651	KM	∅	0.5	0	C
	Ostrea edulis	20	939	MMC	0.04	0	0.02	B
	Physa acuta	18	4,286	Beta	1.5	0	0.02	B
	Sepia officinalis	18	1,740	KM	∅	0	0.02	C
Plants	*Arabidopsis thaliana*	345	10,322,757	Beta	1.6	0	0.07	A
	Zea mays	66	520,310	Ψ	0.01	0	0	A
Bacteria	*Acinetobacter baumannii*	79	78,175	Beta	1.8	0	0.1	B
	Bacillus subtilis	38	105,523	Ψ	0.14	0	0.2	B
	Chlamydia trachomatis	59	9,924	KM	∅	0	0.11	D
	Clostridium difficile	11	192	KM	∅	15	0.15	D
	Escherichia coli	62	84,222	KM	∅	0	0.06	B
	Helicobacter pylori	70	27,498	Ψ	0.01	1	0.2	B
	Klebsiella pneumoniae	156	203,601	KM	∅	18.5	0.15	D
	Mycobacterium tubercolosis	33	7,142	Beta	1.05	2.5	0	C
	Pseudomonas aeruginosa	86	90,258	Ψ	0.06	3	0.2	B
	Staphylococcus aureus	152	30,052	Ψ	0.01	1	0.2	B
	Streptococcus pneumoniae	32	49,917	Beta	1.5	0	0.08	C

https://doi.org/10.1371/journal.pgen.1010677.t002

Table 3. Distribution of goodness-of-fit grades of the best-fitting models for the 45 collected SFS. Calculated from Cramér's V, A: $V \in [0 : 0.033[$, B: $V \in [0.033 : 0.066[$, C: $V \in [0.066 : 0.1[$ and D: $V \in [0.1 : \infty[$.

Model \ Grade	A	B	C	D	Total
Kingman	2	1	3	5	11
Beta	8	10	6		24
Psi	2	4			6
MMC		3	1		4
Total	12	18	10	5	45

https://doi.org/10.1371/journal.pgen.1010677.t003

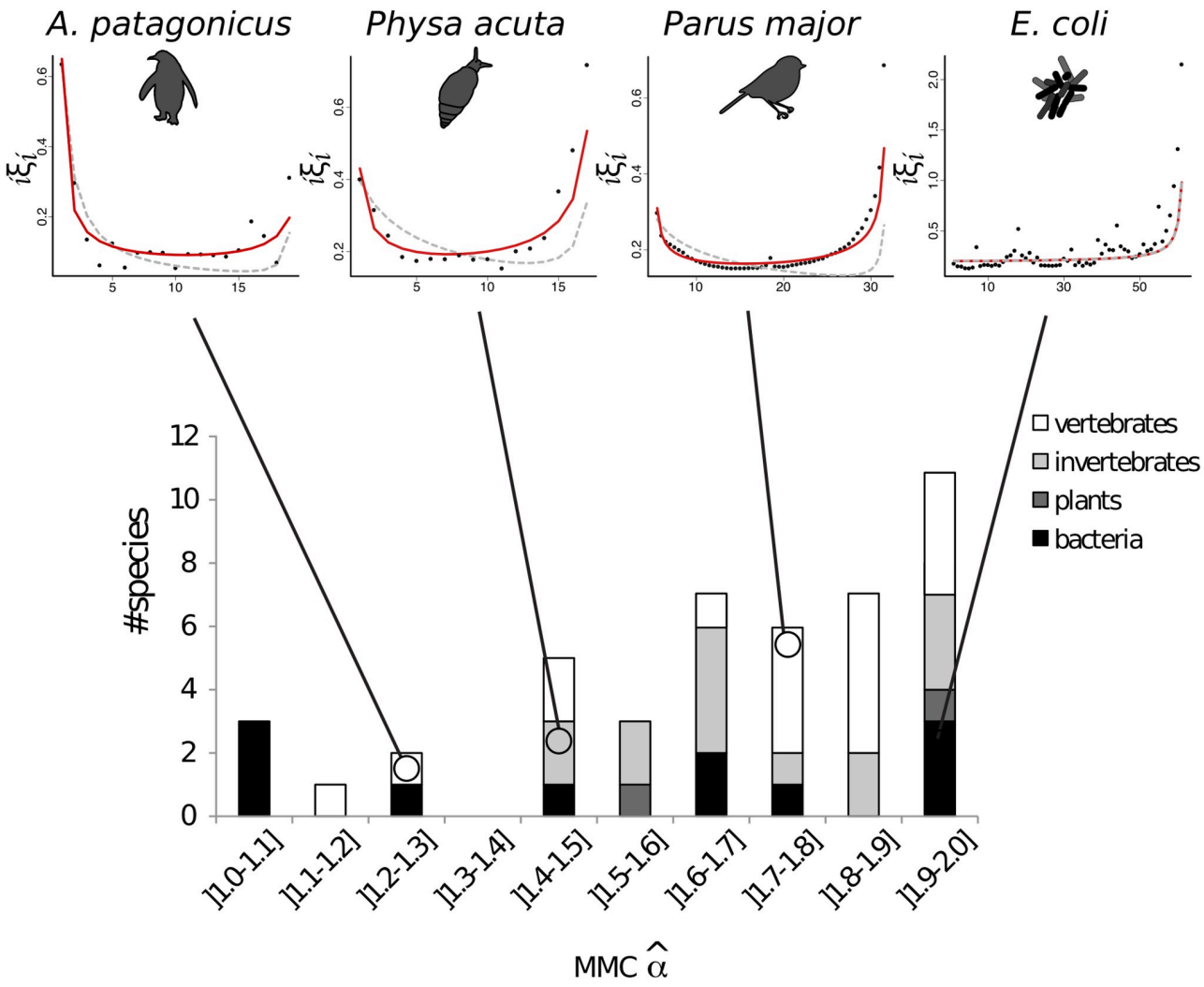

Fig 2. Estimates of α by species. The four top panels represent transformed ϕ-SFS ($\phi_i = i\xi_i$ as in [54, 55]) for four species from different taxa: two vertebrates *Aptenodytes patagonicus* (left) and *Parus major* (center right) an invertebrate *Physa acuta* (center left), and a bacteria *Escherichia coli* (right). For *E. coli*, the uptick in the spectrum comes exclusively from the allele miss-orientation, as $\hat{\alpha} = 2$. Black dots are the observed values, grey dotted lines are the best fits under the Kingman's coalescent model and red lines are the best fits under a Beta-coalescent model.

https://doi.org/10.1371/journal.pgen.1010677.g002

In contrast, the allele misorientation parameters e are almost identical between the Kingman model and the MMC (Fig A.7B in S1 Appendix), which may be a consequence of adding a second, coalescent-model-free estimation method for e to the pseudolikelihood 2. This suggests that for datasets with frequent multiple mergers, assuming a Kingman model may lead to overestimating g, but is not likely to impact estimates of e.

Both MMC models have similar parameter estimates. Finally, we compare the estimations of both MMC models to see whether using one or the other would result in qualitatively different conclusions. The parameters inferred under the two MMC models are highly correlated. The multiple merger parameters α of the Beta-coalescent and Ψ of the Psi-coalescent are negatively correlated, as expected from their definitions (Fig A.8A in S1 Appendix, Spearman correlation: $\rho = -0.72$). The estimated growth and misorientation parameters are highly positively correlated (Spearman correlations $\rho = 0.73$ and $\rho = 0.95$). The case of *Clostridium difficile* is a notable exception. The best model inferred is the Kingman, consistent with $\hat{\Psi} = 0$ inferred for the Psi-coalescent, but for the Beta-coalescent $\hat{\alpha} = 1$, the strongest MMC component, is estimated. However, this discrepancy is likely due to statistical noise: the data set is very small (192 mutations in a sample size $n = 11$) and the species has a very low recombination rate.

4 Discussion

In this study, we show that unfolded SFS for large variety of species show a characteristic U-shape, which is inconsistent with the expectations of the standard neutral model using the Kingman coalescent. One possible explanation for this observation is the prevalence of MMC and MMC-like genealogies in real populations. To explore the role of MMCs in these data, we develop a statistical framework to detect MMC models. Using simulated data, we show this approach has power to detect the correct MMC model and estimate its parameters, provided that the data are informative enough. Using real SFS collected from 45 species across the tree of life, we further show the MMC models are a better fit than the Kingman coalescent in most species, even when population growth and orientation errors are additionally modeled, although in some cases the MMC parameter suggests approximately Kingman behavior. In the following, we discuss some possible biological implications of these observations.

Chosen multiple-merger models, alternatives and limitations

We chose two commonly used haploid multiple-merger models, the Beta- and the Psi-coalescent, which were previously associated with sweepstake reproduction in the literature [43, 44]. However, these MMC models may also originate either from alternative neutral processes or from selective processes. Indeed, the Beta n-coalescent with $\alpha = 1$ is known as the Bolthausen-Sznitman n-coalescent and it (resp. a slight variant of it) emerges in a variety of models with rapid selection [34–38]. The Beta-coalescent has also been associated with range expansions [42]. In addition, Psi n-coalescents have been successfully used as proxy models for detecting regions experiencing positive selection [56].

While Beta- and Psi-coalescent models are linked to several biological properties potentially present in a considerable number of species, these are not the only MMCs used to model biological populations. For instance, in the modified Moran models presented above, one can let the Ψ be random, leading to another more general class of MMC that also belongs to the family of Λ-coalescents [49], which is a generally good candidate for sweepstakes reproduction. Other alternative models exist that more closely mimic recurrent selective sweeps [57] or appear as variants of Psi- and Beta- coalescents, but for diploid reproduction [58–60].

We have chosen to evaluate two simple classes of coalescent processes which interpolate between the two extreme tree shapes—a purely bifurcating Kingman tree ($\Psi = 0$ or $\alpha = 2$) and

a star-shaped tree ($\Psi = 1$ or $\alpha = 0$). Alternative multiple merger models could potentially be (mis)identified as Beta- or Psi-coalescents, as previously shown [61]. Our method should thus still be able to detect multiple merger signals even if caused by processes that lead to another MMC. Assessing further which MMC models are best fitting for biological populations could be informative [26]. In this regard, our inference approach is based on computing $E(T_i)$ from Eq (2) via the method from [53], so it can easily be extended to incorporate most multiple merger models (any Λ- or Ξ-coalescent) and any demographic histories, by replacing the Markov transition rate matrix of the coalescent and the population size profile ν.

To assess the quality of our inference method, we used a simplified approach where unlinked loci are assumed to be independent. This is not always true for MMC models (see [62] and A.10 in S1 Appendix), especially for Psi-coalescents caused by strong sweepstake reproduction events with Ψ well above 0. Thus, the real error rates of our techniques could be higher than anticipated by our simulation study. However, this potential increase in error rates can be offset by the presence of datasets that are larger than those assumed in our simulation study. Additionally, due to our reliance on the expected SFS entries—which are averages over the tree space—our inference method (and also our goodness-of-fit assessment) should perform worse (given identical sample sizes and mutation counts) when used on species with small genomes and low recombination rates. This tendency is clearly visible in the goodness-of-fit tests of multiple bacterial data sets.

Non-extreme demography alone cannot generate U-shaped SFS

The Kingman coalescent for a population undergoing non-extreme demographic changes corresponds simply to a monotonic time rescaling of the standard Kingman coalescent. Non-extreme changes mean that the population size changes occur at the same time scale than coalescent time. For the MMC models employed, this is for instance satisfied if the population size stays of the same order (N) throughout generations. If this is true, changes in population size correspond to changes in waiting times, but not topology, of the tree. The expected SFS for a large population and a large sample is a linear function of the expected waiting times c_k for the next coalescence of k lineages, with a simple analytical form:

$$\mathrm{E}[\xi_f] = \theta \sum_{k=2}^{\infty} k(k-1)c_k \cdot (1-f)^{k-2}, \tag{5}$$

where ξ_f is the number of variants at frequency f. Since the expected waiting times are positive $c_k > 0$, all coefficients in this expansion are positive. This means that the spectrum has a positive value, negative derivative, positive convexity (second derivative), etc., so it is a completely monotonic function ('no bumps'). A similar argument holds for finite frequencies i/n [63]. More details are provided in A.13 in S1 Appendix. As it is monotonically decreasing with i, U-shaped spectra cannot occur as a result of any non-extreme demographic dynamics alone. Note however that extreme changes in population size violate this and may lead to multiple merger genealogies [39, 64].

Alternative processes leading to *U*-shaped SFS, further confounding factors

Our model directly incorporates MMC genealogies, exponential growth combined with allele misorientation as sources of the U-shape of the SFS. However, other potential factors can also influence the SFS and produce SFS with similar shapes. We further discuss here three particularly notable factors, further sources of misorientation errors, population structure (e.g. gene flow or admixture) and biased gene conversion.

First, we tested whether other sources of misorientation errors can explain the strong support for MMC in the dataset. As sequencing errors in the in-group will likely create mostly derived singletons, they cannot explain the U-shape. Furthermore sequencing errors in the outgroup would result in the exact same patterns as natural mutations. Thus the amount of misorientation error can include both recurrent mutations and sequencing errors in the outgroup. We have also developed an extended version of the orientation errors model (see A.4.1 in S1 Appendix) taking into account different rates for transitions and transversions [65]. Even though orientation errors are then modeled by two parameters, the general picture is the same: the best supported model (Tables E and F in S1 Appendix) remains unchanged for 33 species and becomes another MMC for 6 species. However, 6 species have their statistical support swapped between an MMC and Kingman models: *A. franciscana*, *C. cornix*, *F. albicollis*, *M. cinxia*, *P. paniscus* (Beta to Kingman for the 5) and *K. pneumoniae* (Kingman to Beta), leaving 66% of the species with a better support for MMC than Kingman. We then tested whether the phylogenetic proximity of the outgroup could allow for Incomplete Lineage Sorting (ILS) that can cause ancestral polymorphisms to segregate in the sampled species (see A.5 in S1 Appendix). Results (Table H in S1 Appendix) show that ancestral polymorphisms (ILS with mutation) is not a likely contributor of orientation errors for most species as they cannot represent an appreciable amount of the polymorphic sites. However, for the 10 species for which the estimated $P(ILS) \times 0.1$ is larger than 1%, the error rate is possibly underestimated. However, 3/10 show strong support for the Kingman coalescent. Therefore, which model would have the best statistical support for these species if ILS was properly account for in a dedicated non-trivial MMC likelihood framework remains unclear.

Second, to explore population structure, we performed a PCA analysis of all datasets, followed by a *k*-means clustering (results in Table H in S1 Appendix). We acknowledge the possibility that unsampled "ghost" demes can exist and could potentially result in U-shaped SFS [10], and that some cases of metapopulation dynamics results in MMC trees [41, 42]. Assessing the presence of ghost demes from genetic data is challenging. Importantly, among the 11 species that display a clear pattern of genetic structure, only 6 have an observed U-shaped SFS that is well fitted (grades A and B) by an MMC model. Furthermore, among the 14 species with no clear structure, 10 have an observed U-shaped SFS well fitted by an MMC model. This suggests that population structure is not the main cause of the U-shape of the observed SFS. Additionally, many species with clear structure have low goodness-of-fit grades (C and D), suggesting that none of the models we compare are a good fit to these datasets. We however note that 8/11 species with a clear structure pattern are Bacteria. Indeed for the small genomes with low recombination rate (in Bacteria recombination preserves long distance linkage), the apparent structure does not necessarily equate with population structure, but may instead arise from the limited number of genealogies. At the limit, a single Kingman tree would result in a clear structure pattern due its long internal branches.

To check for the effect of biased gene conversion, we built alternative SFS only based on a subset of unbiased mutations that are immune to biased gene conversion (details in A.9 in S1 Appendix, the unbiased SFS are added in supplementary material S1 and S2 Figs). Many of these unbiased SFS were only slightly changed, and many kept their *U*-shape. However 6 species (*A. cunicularia*, *F. albicollis*, *E. garzetta*, *P. maior*, *O. edulis*, *P. troglodytes e.*, all but one vertebrates) lost their U-shape. Two have a small sample size (*E. garzetta*) or a low multiple merger component estimate (*F. albicollis*). For these species, it is nonetheless possible that the U-shape is caused by biased gene conversion.

In a very conservative approach, among the 17 data sets showing robust and strong MMC signals (category A, B in Table 2, with $\alpha \leq 1.8$ or $\Psi \geq 0.04$ and sample size $\geq$20), 6 cases may arise due to structured genetic diversity (*A. baumannii*, *D. melanogaster*, *H. pilori*, *O. edulis*,

P. aeruginosa and *S. aureus*) and 3 more lose their characteristic U-shape when biased gene conversion is accounted for (*A. cunicularia*, *P. maior*, *P. troglodytes*; *O. edulis* being in common). Thus, 8 species have strong support for MMC models with population growth. We believe that at least for these cases (and likely for more), neutral sweepstake reproduction, frequent selection, or other factors that can produce MMC-like genealogies ought to be seriously considered as underlying drivers of their genetic diversity.

Importantly and more generally, among the 30 species that display a good statistical fit (with grades A and B), 27 point to MMC models whereas only 3 point to a Kingman coalescent. Noting that MMC models encompass the Kingman coalescent as a special case, our results support the view that MMC models may often constitute better reference models.

MMC and biological properties

Although we only analyzed a small number of species sampled non-uniformly across the tree of life, we often observed signatures of multiple merger-like events. Reassuringly, our analysis supports multiple merger genealogies for *Mycobacterium tuberculosis*, which was recently proposed in [25] and [26] (the non-optimal goodness-of-fit likely stems from a small and essentially non-recombining genome). The strongest multiple merger effects estimated within the class of Beta coalescents ($\alpha \leq 1.1$) were found in two bacterial pathogens with low or intermediate recombination rates (*M. tuberculosis* and *P. aeruginosa*). There also does not seem to be a meaningful correlation between MMC effects and overall genetic diversity (Fig A.23 and Table I in S1 Appendix). We stress that links between MMC model parameters and biological properties are not always obvious. For example, while reproduction sweepstakes can lead to both Beta- and Psi-coalescents, it is not straightforward to translate the parameters α and Ψ into realistic offspring distributions. For instance the Psi-coalescent model hypothesizes that an occasional individual contributes a fraction Ψ of the next generation, though examples of such a single-individual contribution are not biologically likely. Still, the coalescent approximations do fit well to data. Importantly, different reproduction models can result in the same model on the *coalescent time scale*. The large families of the MMC models could result from the rapid accumulation of coalescences over multiple generations instead of in a single one.

Conclusion

We analyzed genomic data for 45 species across the tree of life, and showed that many exhibit a U-shaped SFS. By developing a statistical approach to distinguish the genetic signatures of different potential sources of this U-shape: allele misorientation and MMC genealogies, together with exponential population growth, our results show that while some U-shaped SFS are well-described by only allele misorientation, the majority are better described by models that include an MMC component (27 point to MMC and only 3 to Kingman coalescent, with the rest inconclusive). However, distinguishing true MMC from MMC-like processes remains challenging. For example, both biased gene conversion (evident for 6 species) and population structure (clear for 11 species, many of which had no U-shapes) could also generate U-shaped SFS, and appear to be plausible explanations for the observed data of certain species. MMC models with simple growth nonetheless represent an excellent fit for at least 8 species. More complex demographic scenario (with more parameters) can be included in the MMC framework presented here and can be statistically tested when demographic inference is being performed. However non-extreme variations of population sizes cannot explain the pervasive observation of U-shaped SFS.

This study thus invites both closer inspection for the species at hand, but also suggests that MMC genealogies may appear in a wider range of species than previously reported (e.g., a few

marine species and multiple human pathogens). For such species, their biological properties likely render MMC rather than Kingman models as the more fruitful analysis framework, highlighting the importance of further developing both theory and statistical inference procedures under these lesser-used models [66].

Supporting information

S1 Appendix. Extended methods, supplementary figures and tables.
(PDF)

S1 Fig. Observed SFS and expected SFS under best fitting models.
(PDF)

S2 Fig. Observed transformed SFS and expected transformed SFS under best fitting models.
(PDF)

S3 Fig. PCA plots w. DAPC colours.
(PDF)

S4 Fig. BIC plots of DAPC population structure analyses.
(PDF)

S5 Fig. Per-chromosome PCA+DAPC and BIC plots for D. melanogaster data.
(PDF)

Acknowledgments

We would like to thank Allison J Shultz and Brian J Arnold for help with producing VCFs for several bird species. The authors acknowledge the support by the state of Baden-Württemberg (Germany) through bwHPC.

Author Contributions

Conceptualization: Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Marguerite Lapierre, Jeffrey D. Jensen, Luca Ferretti, Amaury Lambert, Timothy B. Sackton, Guillaume Achaz.

Data curation: Fabian Freund, Elise Kerdoncuff, Timothy B. Sackton.

Formal analysis: Fabian Freund, Amaury Lambert, Guillaume Achaz.

Investigation: Jeffrey D. Jensen, Luca Ferretti, Guillaume Achaz.

Methodology: Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Marguerite Lapierre, Marcel Hildebrandt, Luca Ferretti, Amaury Lambert, Guillaume Achaz.

Project administration: Guillaume Achaz.

Resources: Timothy B. Sackton.

Software: Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Marcel Hildebrandt, Guillaume Achaz.

Supervision: Fabian Freund, Jeffrey D. Jensen, Amaury Lambert, Guillaume Achaz.

Validation: Sebastian Matuszewski, Jeffrey D. Jensen, Timothy B. Sackton, Guillaume Achaz.

Visualization: Elise Kerdoncuff, Marguerite Lapierre, Guillaume Achaz.

Writing – original draft: Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Jeffrey D. Jensen, Luca Ferretti, Amaury Lambert, Timothy B. Sackton, Guillaume Achaz.

Writing – review & editing: Fabian Freund, Jeffrey D. Jensen, Timothy B. Sackton, Guillaume Achaz.

References

1. Kingman JFC. The coalescent. Stochastic Processes and their Applications. 1982; 13(3):235–248. https://doi.org/10.1016/0304-4149(82)90011-4
2. Wakeley J. Coalescent Theory: An Introduction. Greenwood Village: Roberts & Company Publishers; 2009.
3. Kimura M. Evolutionary Rate at the Molecular Level. Nature. 1968; 217(5129):624–626. https://doi.org/10.1038/217624a0 PMID: 5637732
4. Kimura M. The Neutral Theory of Molecular Evolution. Cambridge University Press; 1983.
5. Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical population biology. 1983; 23(2):183–201. https://doi.org/10.1016/0040-5809(83)90013-8 PMID: 6612631
6. Wilkinson-Herbots HM. Genealogy and subpopulation differentiation under various models of population structure. Journal of Mathematical Biology. 1998; 37(6):535–585. https://doi.org/10.1007/s002850050140
7. Kaplan NL, Darden T, Hudson RR. The coalescent process in models with selection. Genetics. 1988; 120(3):819. PMID: 3066685
8. Fu YX. Statistical properties of segregating sites. Theoretical population biology. 1995; 48(2):172–197. https://doi.org/10.1006/tpbi.1995.1025 PMID: 7482370
9. Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the Neutral Theory in 1968 and 50 years on: A response to Kern and Hahn 2018. Evolution. 2019; 73(1):111–114. https://doi.org/10.1111/evo.13650 PMID: 30460993
10. Marchi N, Excoffier L. Gene flow as a simple cause for an excess of high-frequency-derived alleles. Evolutionary applications. 2020; 13(9):2254–2263. https://doi.org/10.1111/eva.12998 PMID: 33005222
11. Lapierre M, Blin C, Lambert A, Achaz G, Rocha EP. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. Molecular biology and evolution. 2016; 33(7):1711–1725. https://doi.org/10.1093/molbev/msw048 PMID: 26931140
12. Baudry E, Depaulis F. Effect of misoriented sites on neutrality tests with outgroup. Genetics. 2003; 165(3):1619–1622. https://doi.org/10.1093/genetics/165.3.1619 PMID: 14668409
13. Pouyet F, Aeschbacher S, Thiéry A, Excoffier L. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. Elife. 2018; 7:e36317. https://doi.org/10.7554/eLife.36317 PMID: 30125248
14. Bustamante CD, Wakeley J, Sawyer S, Hartl DL. Directional selection and the site-frequency spectrum. Genetics. 2001; 159(4):1779–1788. https://doi.org/10.1093/genetics/159.4.1779 PMID: 11779814
15. Cvijović I, Good BH, Desai MM. The effect of strong purifying selection on genetic diversity. Genetics. 2018; 209(4):1235–1278. https://doi.org/10.1534/genetics.118.301058 PMID: 29844134
16. Johri P, Charlesworth B, Jensen JD. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. Genetics. 2020; 215(1):173–192. https://doi.org/10.1534/genetics.119.303002 PMID: 32152045
17. Huerta-Sanchez E, Durrett R, Bustamante CD. Population genetics of polymorphism and divergence under fluctuating selection. Genetics. 2008; 178(1):325–337. https://doi.org/10.1534/genetics.107.073361 PMID: 17947441
18. Tellier A, Lemaire C. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. Molecular ecology. 2014; 23(11):2637–2652. https://doi.org/10.1111/mec.12755 PMID: 24750385
19. Sagitov S. The general coalescent with asynchronous mergers of ancestral lines. Journal of Applied Probability. 1999; 36(4):1116–1125. https://doi.org/10.1017/S0021900200017903
20. Pitman J. Coalescents with multiple collisions. Annals of Probability. 1999; 27(4):1870–1902. https://doi.org/10.1214/aop/1022874819
21. Donnelly P, Kurtz TG. Particle representations for measure-valued population models. The Annals of Probability. 1999; 27(1):166–205. https://doi.org/10.1214/aop/1022677258

22. Möhle M, Sagitov S. A classification of coalescent processes for haploid exchangeable population models. The Annals of Probability. 2001; 29(4):1547–1562.
23. Schweinsberg J. Coalescents with Simultaneous Multiple Collisions. Electronic Journal of Probability. 2000; 5:1–50. https://doi.org/10.1214/EJP.v5-68
24. Montano V. Coalescent inferences in conservation genetics: should the exception become the rule? Biology letters. 2016; 12(6):20160211. https://doi.org/10.1098/rsbl.2016.0211 PMID: 27330172
25. Morales-Arce AY, Sabin SJ, Stone AC, Jensen JD. The population genomics of within-host Mycobacterium tuberculosis. Heredity. 2020; p. 1–9. https://doi.org/10.1038/s41437-020-00377-7 PMID: 33060846
26. Menardo F, Gagneux S, Freund F. Multiple Merger Genealogies in Outbreaks of Mycobacterium tuberculosis. Molecular Biology and Evolution. 2020; 38(1):290–306. https://doi.org/10.1093/molbev/msaa179
27. Sackman AM, Harris RB, Jensen JD. Inferring demography and selection in organisms characterized by skewed offspring distributions. Genetics. 2019; 211(3):1019–1028. https://doi.org/10.1534/genetics.118.301684 PMID: 30651284
28. Rödelsperger C, Neher RA, Weller AM, Eberhardt G, Witte H, Mayer WE, et al. Characterization of genetic diversity in the nematode Pristionchus pacificus from population-scale resequencing data. Genetics. 2014; 196(4):1153–1165. https://doi.org/10.1534/genetics.113.159855 PMID: 24443445
29. Árnason E, Halldórsdóttir K. Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: Analysis with multiple merger coalescent models. PeerJ PrePrints. 2014; 2.
30. Niwa HS, Nashida K, Yanagimoto T. Reproductive skew in Japanese sardine inferred from DNA sequences. ICES Journal of Marine Science. 2016; 73(9):2181–2189. https://doi.org/10.1093/icesjms/fsw070
31. Vendrami DLJ, Peck LS, Clark MS, Eldon B, Meredith M, Hoffman JI. Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. Science Advances. 2021; 7(37):eabj4713. https://doi.org/10.1126/sciadv.abj4713 PMID: 34516767
32. Kato M, Vasco DA, Sugino R, Narushima D, Krasnitz A. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. Royal Society Open Science. 2017; 4(9). https://doi.org/10.1098/rsos.171060 PMID: 28989791
33. Eldon B. Evolutionary Genomics of High Fecundity. Annual Review of Genetics. 2020; 54. https://doi.org/10.1146/annurev-genet-021920-095932 PMID: 32870729
34. Brunet É, Derrida B. Genealogies in simple models of evolution. Journal of Statistical Mechanics: Theory and Experiment. 2013; 2013(01):P01006. https://doi.org/10.1088/1742-5468/2013/01/P01006
35. Neher RA, Hallatschek O. Genealogies of rapidly adapting populations. Proc Natl Acad Sci USA. 2013; 110(2):437–442. https://doi.org/10.1073/pnas.1213113110 PMID: 23269838
36. Desai MM, Walczak AM, Fisher DS. Genetic diversity and the structure of genealogies in rapidly adapting populations. Genetics. 2013; 193(2):565–585. https://doi.org/10.1534/genetics.112.147157 PMID: 23222656
37. Berestycki J, Berestycki N, Schweinsberg J. The genealogy of branching Brownian motion with absorption. The Annals of Probability. 2013; 41(2):527–618. https://doi.org/10.1214/11-AOP728
38. Schweinsberg J. Rigorous results for a population model with selection II: genealogy of the population. Electronic Journal of Probability. 2017; 22. https://doi.org/10.1214/17-EJP57
39. Birkner M, Blath J, Möhle M, Steinrücken M, Tams J. A modified lookdown construction for the Ξ-Fleming-Viot process with mutation and populations with recurrent bottlenecks. Alea. 2009; 6:25–61.
40. Cordero F, Casanova AG, Schweinsberg J, Wilke-Berenguer M. Λ-coalescents arising in a population with dormancy. Electronic Journal of Probability. 2022; 27:1–34. https://doi.org/10.1214/22-EJP739
41. Taylor JE, Véber A. Coalescent processes in subdivided populations subject to recurrent mass extinctions. Electron J Probab. 2009; 14:242–288. https://doi.org/10.1214/EJP.v14-595
42. Birzu G, Hallatschek O, Korolev KS. Genealogical structure changes as range expansions transition from pushed to pulled. Proceedings of the National Academy of Sciences. 2021; 118(34). https://doi.org/10.1073/pnas.2026746118 PMID: 34413189
43. Schweinsberg J. Coalescent processes obtained from supercritical Galton–Watson processes. Stochastic Proc Appl. 2003; 106(1):107–139. https://doi.org/10.1016/S0304-4149(03)00028-0
44. Eldon B, Wakeley J. Coalescent processes when the distribution of offspring number among individuals is highly skewed. Genetics. 2006; 172(4):2621–2633. https://doi.org/10.1534/genetics.105.052175 PMID: 16452141
45. Nielsen R. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. Genetics. 2000; 154(2):931–942. https://doi.org/10.1093/genetics/154.2.931 PMID: 10655242

46. Griffiths RC, Tavare S. Sampling theory for neutral alleles in a varying environment. Philosophical transactions: biological sciences. 1994; p. 403–410. PMID: 7800710

47. Matuszewski S, Hildebrandt ME, Achaz G, Jensen JD. Coalescent Processes with Skewed Offspring Distributions and Non-equilibrium Demography. Genetics. 2018; 208(1):323–338. https://doi.org/10.1534/genetics.117.300499 PMID: 29127263

48. Freund F. Cannings models, population size changes and multiple-merger coalescents. Journal of mathematical biology. 2020; 80(5):1497–1521. https://doi.org/10.1007/s00285-020-01470-5 PMID: 32008102

49. Huillet T, Möhle M. On the extended Moran model and its relation to coalescents with multiple collisions. Theoretical population biology. 2013; 87:5–14. https://doi.org/10.1016/j.tpb.2011.09.004 PMID: 22001353

50. Eldon B, Birkner M, Blath J, Freund F. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? Genetics. 2015; 199(3):841–856. https://doi.org/10.1534/genetics.114.173807 PMID: 25575536

51. Lapierre M. Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique; 2017. Université Pierre et Marie Curie-Paris VI.

52. Jukes TH, Cantor CR, et al. Evolution of protein molecules. Mammalian protein metabolism. 1969; 3:21–132. https://doi.org/10.1016/B978-1-4832-3211-9.50009-7

53. Spence JP, Kamm JA, Song YS. The Site Frequency Spectrum for General Coalescents. Genetics. 2016; 202(4):1549–1561. https://doi.org/10.1534/genetics.115.184101 PMID: 26883445

54. Achaz G. Frequency spectrum neutrality tests: one for all and all for one. Genetics. 2009; 183(1):249–58. https://doi.org/10.1534/genetics.109.104042 PMID: 19546320

55. Lapierre M, Lambert A, Achaz G. Accuracy of Demographic Inferences from the Site Frequency Spectrum: The Case of the Yoruba Population. Genetics. 2017; 206(1):439–449. https://doi.org/10.1534/genetics.116.192708 PMID: 28341655

56. Harris RB, Jensen JD. Considering genomic scans for selection as coalescent model choice. Genome biology and evolution. 2020; 12(6):871–877. https://doi.org/10.1093/gbe/evaa093 PMID: 32396636

57. Durrett R, Schweinsberg J. A coalescent model for the effect of advantageous mutations on the genealogy of a population. Stochastic Processes and their Applications. 2005; 115(10):1628–1657. https://doi.org/10.1016/j.spa.2005.04.009

58. Blath J, Cronjäger MC, Eldon B, Hammer M. The site-frequency spectrum associated with Ξ-coalescents. Theoretical Population Biology. 2016; 110:36–50. https://doi.org/10.1016/j.tpb.2016.04.002 PMID: 27112097

59. Birkner M, Liu H, Sturm A. Coalescent results for diploid exchangeable population models. Electronic Journal of Probability. 2018; 23. https://doi.org/10.1214/18-EJP175

60. Koskela J, Berenguer MW. Robust model selection between population growth and multiple merger coalescents. Mathematical biosciences. 2019; 311:1–12. https://doi.org/10.1016/j.mbs.2019.03.004 PMID: 30851276

61. Freund F, Siri-Jégousse A. The impact of genetic diversity statistics on model selection between coalescents. Computational Statistics & Data Analysis. 2021; 156:107055. https://doi.org/10.1016/j.csda.2020.107055

62. Birkner M, Blath J, Eldon B. An ancestral recombination graph for diploid populations with skewed offspring distribution. Genetics. 2013; 193(1):255–290. https://doi.org/10.1534/genetics.112.144329 PMID: 23150600

63. Sargsyan O, Wakeley J. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. Theoretical Population Biology. 2008; 74(1):104–114. https://doi.org/10.1016/j.tpb.2008.04.009 PMID: 18554676

64. Casanova AG, Pina VM, Siri-Jégousse A. The symmetric coalescent and Wright–Fisher models with bottlenecks. The Annals of Applied Probability. 2022; 32(1):235–268.

65. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of molecular evolution. 1980; 16(2):111–120. https://doi.org/10.1007/BF01731581 PMID: 7463489

66. Wakeley J. Coalescent theory has many new branches; 2013.