Interpreting the pervasive observation of U-shaped Site Frequency Spectra

Fabian Freund[a,*], Elise Kerdoncuff[b,k,*], Sebastian Matuszewski[c], Marguerite Lapierre[b,d], Marcel Hildebrandt[e], Jeffrey D. Jensen[f], Luca Ferretti[g], Amaury Lambert[b,h], Timothy B. Sackton[i] and Guillaume Achaz[b,j]

April 12, 2022

[a]: Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Germany

[b]: SMILE group, Center for Interdisciplinary Research in Biology (CIRB), Collège de France, CNRS UMR 7241, INSERM U 1050, Paris, France

[c]: Accenture, Vienna, Austria

[d]: Lycée Louis le Grand, Paris, France

[e]: Siemens AG, Munich, Germany

[f]: Center for Evolution & Medicine, School of Life Sciences, Arizona State University, USA

[g]: Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, Nuffield Department of Medicine, University of Oxford, Oxford, UK

[h]: Institut de Biologie de l'ENS (IBENS), École Normale Supérieure, UMR CNRS 8197, INSERM U 1024, Paris, Franc

[i]: Informatics Group, Harvard University, Cambridge, MA, 02138, USA

[j]: Éco-anthropologie, Muséum National d'Histoire Naturelle, CNRS UMR 7206, Université Paris-Cité, Paris, France

[k]: Department of Genetics, University of California, Berkeley, CA 94620 USA

[*]: these authors contributed equally to this work

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.

Corresponding author: guillaume.achaz@mnhn.fr
Running title: U-shaped Site Frequency Spectra

1 Introduction

The Kingman coalescent, [Kin82], 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 [Wak09]. Under the standard neutral theory [Kim68, Kim83], 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 [Hud83], population structure [WH98], and natural selection [KDH88]. 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 [Fu95]. The θ parameter of the SNM is defined as $\theta = 2pN\mu$, where p is the ploidy (typically 1 or 2), N the 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 [JPS+19]. 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 [ME20], population structure [LBL+16], misorientation of ancestral and derived alleles [BD03], biased gene conversion [PATE18], recent positive selection at many targets across the genome [BWSH01], background selection [CGD18], [JCJ20], temporally-fluctuating selection [HSDB08], and various reproductive strategies [TL14].

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) [Sag99, Pit99, DK99, MS01, Sch00]. Briefly, MMCs may arise when the number of offspring per individual has very high variance across the population. Such effects of concentrations of ancestrality (resulting in polytomies in the trees) have been reported in various species across all kingdoms of life [Mon16], and MMC-like genealogies have been observed for species ranging from bacteria (e.g. for *Mycobacterium tuberculosis* [MASSJ20, MGF20]) to viruses (e.g. for influenza [SHJ19]) to animals (e.g. for the nematode *Pristionchus pacificus* [RNW+14], multiple fish species, e.g. [ÁH14, NNY16, VPC+21]) and even to cancer cells [KVS+17].

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 macroscopic fraction of the population by chance, thus resulting in MMC genealogies (for a review, see [Eld20]). 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. [BD13, NH13, DWF13, BBS13, Sch17]. Importantly, other biological factors can also lead to

MMC-like genealogies, including large rapid demographic deviations [BBM$^+$09], seed banks [CCSWB22], extinction-recolonisation in metapopulations [TV09] and range expansions [BHK21]. 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 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 $n > 10$) 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 [Sch03] and Psi-MMC [EW06], both tuned by a single parameter that interpolates between pure 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 [Nie00] 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 [GT94, MHAJ18, Fre20]). 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 $\nu_t = N_t/N_0$ 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

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

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 [Pit99]: 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 $\nu(t) = \exp(-gt)$ for growth rate $g \geq 0$ (see Appendix A.1 for interpretation of g in the initial reproduction model). As underlying reproduction models, we use modified Moran models [HM13, EW06, MHAJ18]. 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 [EW06, MHAJ18]), 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 [HM13, Fre20], 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 Appendix A.1 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.

For statistical inference, we treat the observed SFS of s mutations as s independent multinomial draws from the expected SFS (see [Nie00] and [EBBF15, Eq. 11] [MHAJ18, 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 [Lap17, Section 4.2] or [BD03, 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 Appendix A.3). 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 [JC+69] to predict for the number $s_{\neq}$ of non-polarizable tri-allelic variants. This leads to a slight variant of [MHAJ18, Eq. 14]. For any coalescent model with a specific set of coalescent, exponential growth and misorientation parameters,

the pseudolikelihood is:

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

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 Appendix A.3 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 [MHAJ18], using the approach from [SKS16]. 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 0.05, $e \in [0, 0.15]$ 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 0.05 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:

$$BF_1 = \max(\log \max_{\alpha,g,e} PsL, \log \max_{\Psi,g,e} PsL) - \log \max_{\alpha=2,g,e} PsL, \quad BF_2 = \log \max_{\alpha,g,e} PsL - \log \max_{\Psi,g,e} PsL \tag{3}$$

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.
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 $(\frac{E[T_1]}{E[T_{tot}]}, \ldots, \frac{E[T_{n-1}]}{E[T_{tot}]})$. 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 Appendix A.4 for details.

2.3 Data

We collected 45 genome-wide SFS that are described in Tables 2 and A.5. The collected SFS come from public data sets or private communications. 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. Supplementary files 1 and 2 provide the shapes of the empirically-observed SFS.

3 Results

We have first demonstrated the power of the methodology using extensive simulations, and then applied 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, sampling over 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 Appendix A.5.

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 A.3). 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.

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 (Figure 1, Figures A.2–A.4). The growth rate, in contrast, is only estimated well for situations where the simulated growth rate was low (Figures A.9, A.12, A.15, A.18).

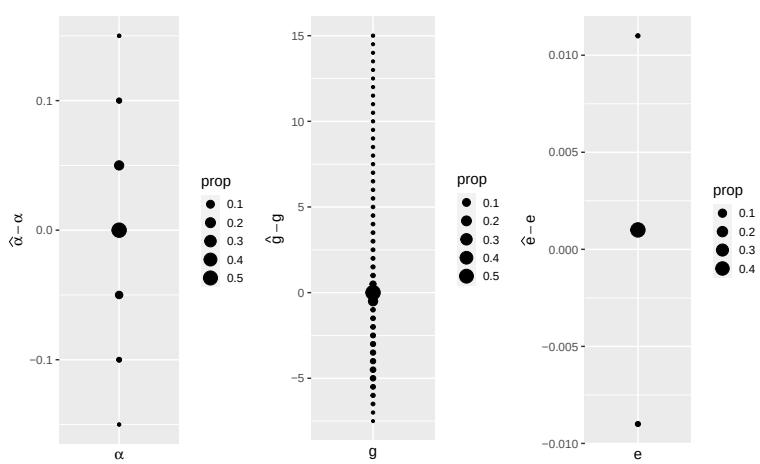


Figure 1: Error for estimating parameters for Beta coalescents with exponential growth and allele misorientation across the parameter grid for (α, g, e). Sample size $n = 100$, 50 independent loci with 100 mutations on average. 500 simulations were performed per parameter triplet.

True model	Within the grid?	Fraction model inferred as 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	

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 hows whether the parameters used for simulation were included in the inference grid. Fractions are rounded to two digits. 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 A.2. For details on simulations and inference parameters see Appendix A.5.

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 (73%) 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 (51%), followed by the Kingman coalescent (27%) 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). Combined with visual inspections, (supplementary files 1,2), we designed 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 71% of data sets: 32/45 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.

Table 2: Data sets description: Taxa, Species, number of individuals (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	7,167,395	Beta	1.95	1	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/Ψ	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/Ψ	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	480	KM	∅	0	0.11	B
	Ciona intestinalis B	20	1,883	Beta	1.65	0	0.02	B
	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/Ψ	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/Ψ	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

Model \ Grade	A	B	C	D	Total
Kingman	2	2	3	5	12
Beta	8	11	4		23
Psi	2	4			6
MMC		3	1		4
Total	12	20	8	5	45

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: $V \in [0.1 : \infty[$.

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, 68% (31/45) 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 (Table A.4, 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, Fig A.7).

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 A.4), although estimates of g tend to converge in empirical datasets where the MMC parameter estimate approaches Kingman (Fig A.5a). 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. [MHAJ18].
In contrast, the allele misorientation parameters e are almost identical between the Kingman model and the MMC (Fig A.5b), 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.6a, Spearman correlation: $\rho = -0.73$). The estimated growth and misorientation parameters are highly positively correlated (Spearman correlations $\rho = 0.74$ and $\rho = 0.96$). 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.

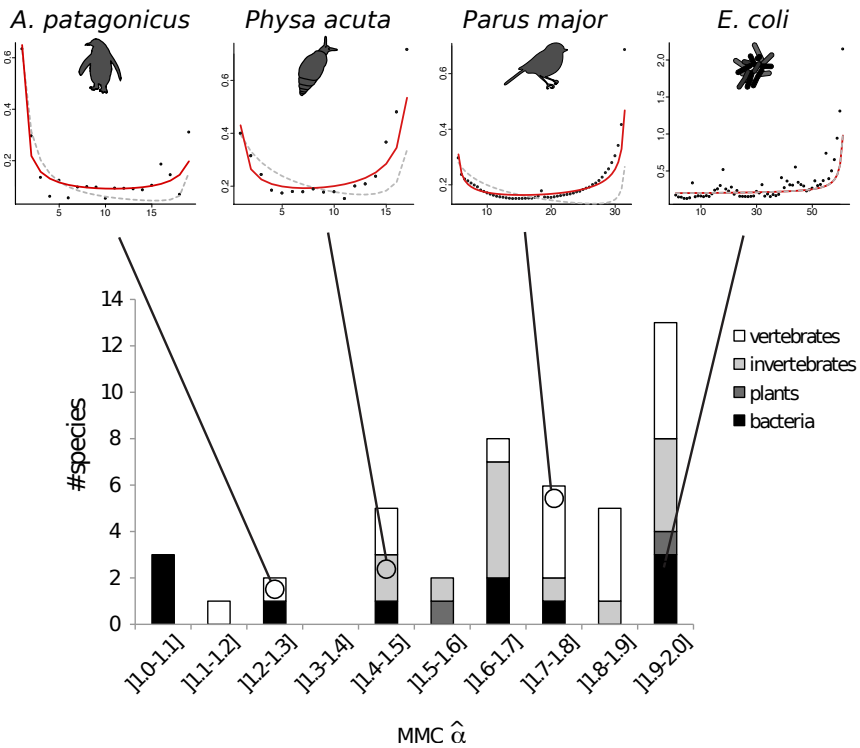


Figure 2: Distribution of α in function of the order of the species. The four top panels represent transformed ϕ-SFS ($\phi_i = i\xi_i$ as in [Ach09, LLA17]) for four species from different taxa: two vertebrates *Aptenodytes patagonicus* (left) and *Parus major* (center right) an invertebrate *Physa acuta* (center left), and a bacteria *Escherichia coli* (right). 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.

3.3 Code and data availability

All simulation and analysis codeis available upon request and additionally will be made available as a public GitHub repository upon publication.

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 good 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 modeling, 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 [EW06, Sch03]. 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 [BD13, NH13, DWF13, BBS13, Sch17]. The Beta-coalescent has also been associated with range expansions [BHK21]. In addition, Psi n-coalescents have been successfully used as proxy models for detecting regions experiencing positive selection [HJ20].

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 [HM13], which is a generally good candidate for sweepstakes reproduction. Other alternative models exist that more closely mimic recurrent selective sweeps [DS05] or appear as variants of Psi- and Beta- coalescents, but for diploid reproduction [BCEH16, BLS18, KB19].

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 [FSJ21]. 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 [MGF20]. In this regard, our inference approach is based on computing $E(T_i)$ from Eq. (2) via the method from [SKS16], 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 [BBE13] and Appendix A.8), 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 order of the number of generations compressed to form one unit of coalescent time (which depends on the probability that a coalescence event happens in a generation) is not affected by the demographic changes. 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 population-level

waiting times c_k (for the next coalescence of k lineages) with a simple analytical form:

$$\mathrm{E}[\xi_f] = \theta \sum_{k=2}^{\infty} k(k-1)c_k \cdot (1-f)^{k-2}, \tag{4}$$

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'). More details are provided in Appendix A.11. 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 [BBM$^+$09, CPSJ22].

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 two particularly notable factors, population structure (e.g. gene flow or admixture) and biased gene conversion.

First, to explore population structure, we performed a PCA analysis of all datasets, followed by a k-means clustering (results in Table A.6). 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.7, the unbiased SFS are added in supplementary files 1,2). 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 ≥ 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 32 species that display a good statistical fit (with grades A and B), 28 point to MMC models whereas only 4 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 [MASSJ20] and [MGF20] (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 (Figure A.21, Table A.7). 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 (28 point to MMC and only 4 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.

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 [Wak13].

Ackowledgments

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. 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.

References

[ABAB+16] Carlos Alonso-Blanco, Jorge Andrade, Claude Becker, Felix Bemm, Joy Bergelson, Karsten M. Borgwardt, Jun Cao, Eunyoung Chae, Todd M. Dezwaan, Wei Ding, Joseph R. Ecker, Moises

Exposito-Alonso, Ashley Farlow, Joffrey Fitz, Xiangchao Gan, Dominik G. Grimm, Angela M. Hancock, Stefan R. Henz, Svante Holm, Matthew Horton, Mike Jarsulic, Randall A. Kerstetter, Arthur Korte, Pamela Korte, Christa Lanz, Cheng-Ruei Lee, Dazhe Meng, Todd P. Michael, Richard Mott, Ni Wayan Muliyati, Thomas Nägele, Matthias Nagler, Viktoria Nizhynska, Magnus Nordborg, Polina Yu. Novikova, F. Xavier Picó, Alexander Platzer, Fernando A. Rabanal, Alex Rodriguez, Beth A. Rowan, Patrice A. Salomé, Karl J. Schmid, Robert J. Schmitz, Ümit Seren, Felice Gianluca Sperone, Mitchell Sudkamp, Hannes Svardal, Matt M. Tanzer, Donald Todd, Samuel L. Volchenboum, Congmao Wang, George Wang, Xi Wang, Wolfram Weckwerth, Detlef Weigel, and Xuefeng Zhou. 1,135 genomes reveal the global pattern of polymorphism in arabidopsis thaliana. *Cell*, 166(2):481–491, 2016.

[Ach09] Guillaume Achaz. Frequency spectrum neutrality tests: one for all and all for one. *Genetics*, 183(1):249–58, Sep 2009.

[ÁH14] Einar Árnason and Katrín Halldórsdóttir. Nucleotide variation and balancing selection at the ckma gene in atlantic cod: Analysis with multiple merger coalescent models. *PeerJ PrePrints*, 2, 2014.

[BBE13] Matthias Birkner, Jochen Blath, and Bjarki Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

[BBM+09] Matthias Birkner, Jochen Blath, Martin Möhle, Matthias Steinrücken, and Johanna Tams. A modified lookdown construction for the Ξ-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *Alea*, 6:25–61, 2009.

[BBS13] Julien Berestycki, Nathanaël Berestycki, and Jason Schweinsberg. The genealogy of branching brownian motion with absorption. *The Annals of Probability*, 41(2):527–618, 2013.

[BCEH16] Jochen Blath, Mathias Christensen Cronjäger, Bjarki Eldon, and Matthias Hammer. The site-frequency spectrum associated with Ξ-coalescents. *Theoretical Population Biology*, 110:36–50, 2016.

[BD03] Emmanuelle Baudry and Frantz Depaulis. Effect of misoriented sites on neutrality tests with outgroup. *Genetics*, 165(3):1619–1622, 2003.

[BD13] Éric Brunet and Bernard Derrida. Genealogies in simple models of evolution. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(01):P01006, 2013.

[BHK21] Gabriel Birzu, Oskar Hallatschek, and Kirill S Korolev. Genealogical structure changes as range expansions transition from pushed to pulled. *Proceedings of the National Academy of Sciences*, 118(34), 2021.

[BLS18] Matthias Birkner, Huili Liu, and Anja Sturm. Coalescent results for diploid exchangeable population models. *Electronic Journal of Probability*, 23, 2018.

[BMHR+17] Jean-Tristan Brandenburg, Tristan Mary-Huard, Guillem Rigaill, Sarah J. Hearne, Hélène Corti, Johann Joets, Clémentine Vitte, Alain Charcosset, Stéphane D. Nicolas, and Maud I. Tenaillon. Independent introductions and admixtures have contributed to adaptation of european maize and its american counterparts. *PLOS Genetics*, 13(3):e1006666, Mar 2017.

[BNK+15] Reto Burri, Alexander Nater, Takeshi Kawakami, Carina F Mugal, Pall I Olason, Linnea Smeds, Alexander Suh, Ludovic Dutoit, Stanislav Bureš, Laszlo Z Garamszegi, et al. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of ficedula flycatchers. *Genome research*, 25(11):1656–1665, 2015.

[BWSH01] Carlos D Bustamante, John Wakeley, Stanley Sawyer, and Daniel L Hartl. Directional selection and the site-frequency spectrum. *Genetics*, 159(4):1779–1788, 2001.

[Can74] C. Cannings. The latent roots of certain markov chains arising in genetics: A new approach, i. haploid models. *Advances in Applied Probability*, 6(2):260–290, 1974.

[CCSWB22] Fernando Cordero, Adrián González Casanova, Jason Schweinsberg, and Maite Wilke-Berenguer. Λ-coalescents arising in a population with dormancy. *Electronic Journal of Probability*, 27:1–34, 2022.

[CGD18] Ivana Cvijović, Benjamin H Good, and Michael M Desai. The effect of strong purifying selection on genetic diversity. *Genetics*, 209(4):1235–1278, 2018.

[Con15] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

[CPSJ22] Adrián González Casanova, Verónica Miró Pina, and Arno Siri-Jégousse. The symmetric coalescent and Wright–Fisher models with bottlenecks. *The Annals of Applied Probability*, 32(1):235–268, 2022.

[Cra16] Harald Cramér. *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton university press, 2016.

[DAA+11] Petr Danecek, Adam Auton, Goncalo Abecasis, Cornelis A Albers, Eric Banks, Mark A DePristo, Robert E Handsaker, Gerton Lunter, Gabor T Marth, Stephen T Sherry, et al. The variant call format and vcftools. *Bioinformatics*, 27(15):2156–2158, 2011.

[DK99] Peter Donnelly and Thomas G Kurtz. Particle representations for measure-valued population models. *The Annals of Probability*, 27(1):166–205, 1999.

[DS05] Rick Durrett and Jason Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628 – 1657, 2005.

[DWF13] Michael M. Desai, Aleksandra M. Walczak, and Daniel S. Fisher. Genetic diversity and the structure of genealogies in rapidly adapting populations. *Genetics*, 193(2):565–585, 2013.

[EBBF15] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the site-frequency spectrum distinguish exponential population growth from multiple-merger coalescents? *Genetics*, 199(3):841–856, 2015.

[Eld20] Bjarki Eldon. Evolutionary genomics of high fecundity. *Annual Review of Genetics*, 54, 2020.

[ESB+12] Hans Ellegren, Linnea Smeds, Reto Burri, Pall I Olason, Niclas Backström, Takeshi Kawakami, Axel Künstner, Hannu Mäkinen, Krystyna Nadachowska-Brzyska, Anna Qvarnström, et al. The genomic landscape of species divergence in ficedula flycatchers. *Nature*, 491(7426):756–760, 2012.

[ETT+17] Tore O Elgvin, Cassandra N Trier, Ole K Tørresen, Ingerid J Hagen, Sigbjørn Lien, Alexander J Nederbragt, Mark Ravinet, Henrik Jensen, and Glenn-Peter Sætre. The genomic mosaicism of hybrid speciation. *Science advances*, 3(6):e1602996, 2017.

[EW06] Bjarki Eldon and John Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, 2006.

[FKR+18] Luca Ferretti, Alexander Klassmann, Emanuele Raineri, Sebastian E Ramos-Onsins, Thomas Wiehe, and Guillaume Achaz. The neutral frequency spectrum of linked sites. *Theoretical population biology*, 123:70–79, 2018.

[FLW+17] Luca Ferretti, Alice Ledda, Thomas Wiehe, Guillaume Achaz, and Sebastian E Ramos-Onsins. Decomposing the site frequency spectrum: the impact of tree topology on neutrality tests. *Genetics*, 207(1):229–240, 2017.

[Fre20] Fabian Freund. Cannings models, population size changes and multiple-merger coalescents. *Journal of mathematical biology*, 80(5):1497–1521, 2020.

[FSJ21] Fabian Freund and Arno Siri-Jégousse. The impact of genetic diversity statistics on model selection between coalescents. *Computational Statistics & Data Analysis*, 156:107055, 2021.

[Fu95] Yun-Xin Fu. Statistical properties of segregating sites. *Theoretical population biology*, 48(2):172–197, 1995.

[GJQP+19] Hugh G Gauch Jr, Sheng Qian, Hans-Peter Piepho, Linda Zhou, and Rui Chen. Consequences of PCA graphs, SNP codings, and PCA variants for elucidating population structure. *PloS one*, 14(6):e0218306, 2019.

[GT94] Robert C Griffiths and Simon Tavare. Sampling theory for neutral alleles in a varying environment. *Philosophical transactions: biological sciences*, pages 403–410, 1994.

[HJ20] Rebecca B. Harris and Jeffrey D. Jensen. Considering genomic scans for selection as coalescent model choice. *Genome biology and evolution*, 12(6):871–877, 2020.

[HM13] Thierry Huillet and Martin Möhle. On the extended moran model and its relation to coalescents with multiple collisions. *Theoretical population biology*, 87:5–14, 2013.

[HSDB08] Emilia Huerta-Sanchez, Rick Durrett, and Carlos D Bustamante. Population genetics of polymorphism and divergence under fluctuating selection. *Genetics*, 178(1):325–337, 2008.

[Hud83] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[IM02] Alex Iksanov and Martin Möhle. On the number of jumps of random walks with a barrier. *Advances in Applied Probability*, 40(01):206–228, 2002.

[JA11] Thibaut Jombart and Ismaïl Ahmed. adegenet 1.3-1: new tools for the analysis of genome-wide snp data. *Bioinformatics*, 27(21):3070–3071, 2011.

[JC+69] Thomas H Jukes, Charles R Cantor, et al. Evolution of protein molecules. *Mammalian protein metabolism*, 3:21–132, 1969.

[JCJ20] Parul Johri, Brian Charlesworth, and Jeffrey D Jensen. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *Genetics*, 215(1):173–192, 2020.

[JPS+19] Jeffrey D. Jensen, Bret A. Payseur, Wolfgang Stephan, Charles F. Aquadro, Michael Lynch, Deborah Charlesworth, and Brian Charlesworth. The importance of the neutral theory in 1968 and 50 years on: A response to kern and hahn 2018. *Evolution*, 73(1):111–114, 2019.

[KB19] Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical biosciences*, 311:1–12, 2019.

[KDH88] Norman L Kaplan, Thomas Darden, and Richard R Hudson. The coalescent process in models with selection. *Genetics*, 120(3):819, 1988.

[KHM+16] Marty Kardos, Arild Husby, S Eryn McFarlane, Anna Qvarnström, and Hans Ellegren. Whole-genome resequencing of extreme phenotypes in collared flycatchers highlights the difficulty of detecting quantitative trait loci in natural populations. *Molecular Ecology Resources*, 16(3):727–741, 2016.

[Kim68] Motoo Kimura. Evolutionary rate at the molecular level. *Nature*, 217(5129):624–626, 1968.

[Kim83] Motoo Kimura. *The Neutral Theory of Molecular Evolution*. Cambridge University Press, 1983.

[Kin82] J.F.C. Kingman. The coalescent. *Stochastic Processes and their Applications*, 13(3):235–248, Sep 1982.

[Kos18] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *Statistical applications in genetics and molecular biology*, 17(3), 2018.

[KVS+17] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society Open Science*, 4(9), 2017.

[Lap17] Marguerite Lapierre. *Extensions du modèle standard neutre pertinentes pour l'analyse de la diversité génétique*. PhD thesis, Université Pierre et Marie Curie-Paris VI, 2017.

[LBL+16] Marguerite Lapierre, Camille Blin, Amaury Lambert, Guillaume Achaz, and Eduardo PC Rocha. The impact of selection, gene conversion, and biased sampling on the assessment of microbial demography. *Molecular biology and evolution*, 33(7):1711–1725, 2016.

[LCC+15] Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-Detig, Kristian A Stevens, Charles H Langley, and John E Pool. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. *Genetics*, 199(4):1229–1241, 01 2015.

[LGS+16] Veronika N Laine, Toni I Gossmann, Kyle M Schachtschneider, Colin J Garroway, Ole Madsen, Koen JF Verhoeven, Victor De Jager, Hendrik-Jan Megens, Wesley C Warren, Patrick Minx, et al. Evolutionary signals of selection on cognition from the great tit genome and methylome. *Nature communications*, 7(1):1–9, 2016.

[LLA17] Marguerite Lapierre, Amaury Lambert, and Guillaume Achaz. Accuracy of demographic inferences from the site frequency spectrum: The case of the Yoruba population. *Genetics*, 206(1):439–449, 05 2017.

[LLC+14] Shengbin Li, BO Li, Cheng Cheng, Zijun Xiong, Qingbo Liu, Jianghua Lai, Hannah V Carey, Qiong Zhang, Haibo Zheng, Shuguang Wei, et al. Genomic signatures of near-extinction and rebirth of the crested ibis and other endangered bird species. *Genome biology*, 15(12):1–17, 2014.

[LLL+17] Max Lundberg, Miriam Liedvogel, Keith Larson, Hanna Sigeman, Mats Grahn, Anthony Wright, Susanne Åkesson, and Staffan Bensch. Genetic differences between willow warbler migratory phenotypes are few and cluster in large haplotype blocks. *Evolution Letters*, 1(3):155–168, 2017.

[MASSJ20] Ana Y Morales-Arce, Susanna J Sabin, Anne C Stone, and Jeffrey D Jensen. The population genomics of within-host Mycobacterium tuberculosis. *Heredity*, pages 1–9, 2020.

[ME20] Nina Marchi and Laurent Excoffier. Gene flow as a simple cause for an excess of high-frequency-derived alleles. *Evolutionary applications*, 13(9):2254–2263, 2020.

[MGF20] Fabrizio Menardo, Sébastien Gagneux, and Fabian Freund. Multiple Merger Genealogies in Outbreaks of Mycobacterium tuberculosis. *Molecular Biology and Evolution*, 38(1):290–306, 07 2020.

[MHAJ18] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 208(1):323–338, 2018.

[MKB+18] Jakob C Mueller, Heiner Kuhl, Stefan Boerno, Jose L Tella, Martina Carrete, and Bart Kempenaers. Evolution of genomic variation in the burrowing owl in response to recent colonization of urban areas. *Proceedings of the Royal Society B: Biological Sciences*, 285(1878):20180206, 2018.

[Mon16] Valeria Montano. Coalescent inferences in conservation genetics: should the exception become the rule? *Biology letters*, 12(6):20160211, 2016.

[MS01] Martin Möhle and Serik Sagitov. A classification of coalescent processes for haploid exchangeable population models. *The Annals of Probability*, 29(4):1547–1562, 2001.

[NH13] Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proc. Natl. Acad. Sci. USA*, 110(2):437–442, 2013.

[Nie00] R Nielsen. Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics*, 154(2):931–942, 02 2000.

[NNY16] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

[Par10] Emmanuel Paradis. pegas: an r package for population genetics with an integrated–modular approach. *Bioinformatics*, 26(3):419–420, 2010.

[PATE18] Fanny Pouyet, Simon Aeschbacher, Alexandre Thiéry, and Laurent Excoffier. Background selection and biased gene conversion affect more than 95% of the human genome and bias demographic inferences. *Elife*, 7:e36317, 2018.

[Pit99] Jim Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, 1999.

[PMSK+13] Javier Prado-Martinez, Peter H. Sudmant, Jeffrey M. Kidd, Heng Li, Joanna L. Kelley, Belen Lorente-Galdos, Krishna R. Veeramah, August E. Woerner, Timothy D. O'Connor, Gabriel Santpere, Alexander Cagan, Christoph Theunert, Ferran Casals, Hafid Laayouni, Kasper Munch, Asger Hobolth, Anders E. Halager, Maika Malig, Jessica Hernandez-Rodriguez, Irene Hernando-Herraez, Kay Prüfer, Marc Pybus, Laurel Johnstone, Michael Lachmann, Can Alkan, Dorina Twigg, Natalia Petit, Carl Baker, Fereydoun Hormozdiari, Marcos Fernandez-Callejo, Marc Dabad, Michael L. Wilson, Laurie Stevison, Cristina Camprubí, Tiago Carvalho, Aurora Ruiz-Herrera, Laura Vives, Marta Mele, Teresa Abello, Ivanela Kondova, Ronald E. Bontrop, Anne Pusey, Felix Lankester, John A. Kiyang, Richard A. Bergl, Elizabeth Lonsdorf, Simon Myers, Mario Ventura, Pascal Gagneux, David Comas, Hans Siegismund, Julie Blanc, Lidia Agueda-Calpena, Marta Gut, Lucinda Fulton, Sarah A. Tishkoff, James C. Mullikin, Richard K. Wilson, Ivo G. Gut, Mary Katherine Gonder, Oliver A. Ryder, Beatrice H. Hahn, Arcadi Navarro, Joshua M. Akey, Jaume Bertranpetit, David Reich, Thomas Mailund, Mikkel H. Schierup, Christina Hvilsom, Aida M. Andrés, Jeffrey D. Wall, Carlos D. Bustamante, Michael F. Hammer, Evan E. Eichler, and Tomas Marques-Bonet. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

[PVB+14] Jelmer W Poelstra, Nagarjun Vijay, Christen M Bossu, Henrik Lantz, Bettina Ryll, Inge Müller, Vittorio Baglione, Per Unneberg, Martin Wikelski, Manfred G Grabherr, et al. The genomic landscape underlying phenotypic integrity in the face of gene flow in crows. *Science*, 344(6190):1410–1414, 2014.

[QTH+15] Yanhua Qu, Shilin Tian, Naijian Han, Hongwei Zhao, Bin Gao, Jun Fu, Yalin Cheng, Gang Song, Per GP Ericson, Yong E Zhang, et al. Genetic responses to seasonal variation in altitudinal stress: whole-genome resequencing of great tit in eastern himalayas. *Scientific Reports*, 5(1):1–10, 2015.

[RdSB+18] Olaya Rendueles, Jorge A. Moura de Sousa, Aude Bernheim, Marie Touchon, and Eduardo P. C. Rocha. Genetic exchanges are more frequent in bacteria encoding capsules. *PLOS Genetics*, 14(12):1–25, 12 2018.

[RET+18] Mark Ravinet, Tore Oldeide Elgvin, Cassandra Trier, Mansour Aliabadian, Andrey Gavrilov, and Glenn-Peter Sætre. Signatures of human-commensalism in the house sparrow genome. *Proceedings of the Royal Society B*, 285(1884):20181246, 2018.

[RGB+14] J. Romiguier, P. Gayral, M. Ballenghien, A. Bernard, V. Cahais, A. Chenuil, Y. Chiari, R. Dernat, L. Duret, N. Faivre, and et al. Comparative population genomics in animals uncovers the determinants of genetic diversity. *Nature*, 515(7526):261–263, Aug 2014.

[RNW+14] Christian Rödelsperger, Richard A Neher, Andreas M Weller, Gabi Eberhardt, Hanh Witte, Werner E Mayer, Christoph Dieterich, and Ralf J Sommer. Characterization of genetic diversity in the nematode pristionchus pacificus from population-scale resequencing data. *Genetics*, 196(4):1153–1165, 2014.

[RTE+18] Anna Runemark, Cassandra N Trier, Fabrice Eroukhmanoff, Jo S Hermansen, Michael Matschiner, Mark Ravinet, Tore O Elgvin, and Glenn-Peter Sætre. Variation and constraints in hybrid genome formation. *Nature Ecology & Evolution*, 2(3):549–556, 2018.

[RZL+18] Aurélien Richaud, Gaotian Zhang, Daehan Lee, Junho Lee, and Marie-Anne Félix. The local coexistence pattern of selfing genotypes in caenorhabditis elegans natural metapopulations. *Genetics*, 208(2):807–821, 2018.

[Sag99] Serik Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, 1999.

[Sch00] Jason Schweinsberg. Coalescents with simultaneous multiple collisions. *Electronic Journal of Probability*, 5:1–50, 2000.

[Sch03] Jason Schweinsberg. Coalescent processes obtained from supercritical Galton–Watson processes. *Stochastic Proc. Appl.*, 106(1):107–139, 2003.

[Sch17] Jason Schweinsberg. Rigorous results for a population model with selection ii: genealogy of the population. *Electronic Journal of Probability*, 22, 2017.

[SHJ19] Andrew M. Sackman, Rebecca B. Harris, and Jeffrey D. Jensen. Inferring demography and selection in organisms characterized by skewed offspring distributions. *Genetics*, 211(3):1019–1028, 2019.

[SKS16] Jeffrey P. Spence, John A. Kamm, and Yun S. Song. The site frequency spectrum for general coalescents. *Genetics*, 202(4):1549–1561, 2016.

[SLS+15] Sonal Singhal, Ellen M Leffler, Keerthi Sannareddy, Isaac Turner, Oliver Venn, Daniel M Hooper, Alva I Strand, Qiye Li, Brian Raney, Christopher N Balakrishnan, et al. Stable recombination hotspots in birds. *Science*, 350(6263):928–932, 2015.

[SMQE16] Linnéa Smeds, Carina F Mugal, Anna Qvarnström, and Hans Ellegren. High-resolution mapping of crossover and non-crossover recombination events by whole-genome re-sequencing of an avian pedigree. *PLoS genetics*, 12(5):e1006044, 2016.

[SW08] Ori Sargsyan and John Wakeley. A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms. *Theoretical Population Biology*, 74(1):104–114, Aug 2008.

[TL14] Aurelien Tellier and Christophe Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular ecology*, 23(11):2637–2652, 2014.

[TV09] Jesse E Taylor and Amandine Véber. Coalescent processes in subdivided populations subject to recurrent mass extinctions. *Electron. J. Probab*, 14:242–288, 2009.

[VPC+21] David L. J. Vendrami, Lloyd S. Peck, Melody S. Clark, Bjarki Eldon, Michael Meredith, and Joseph I. Hoffman. Sweepstake reproductive success and collective dispersal produce chaotic genetic patchiness in a broadcast spawner. *Science Advances*, 7(37):eabj4713, 2021.

[Wak09] John Wakeley. *Coalescent Theory: An Introduction*. Greenwood Village: Roberts & Company Publishers, 2009.

[Wak13] John Wakeley. Coalescent theory has many new branches, 2013.

[WH98] Hilde M. Wilkinson-Herbots. Genealogy and subpopulation differentiation under various models of population structure. *Journal of Mathematical Biology*, 37(6):535–585, Dec 1998.

[WZH+18] Yan Wu, Yaolei Zhang, Zhuocheng Hou, Guangyi Fan, Jinsong Pi, Shuai Sun, Jiang Chen, Huaqiao Liu, Xiao Du, Jie Shen, et al. Population genomic data reveal genes related to important traits of quail. *GigaScience*, 7(5):giy049, 2018.

Appendix A Appendix

A.1 Reproduction models linked to MMC and time scalings

The coalescent approximations from the main text are the coalescent limits for population size $N \to \infty$ (with changed time-scale) of genealogical trees given some reproduction model. We focus on Cannings models [Can74] of reproduction, which are discrete-generation models, usually with fixed population size, and exchangeable offspring numbers between individuals. This is a standard model choice, see e.g. [Sag99], and the modified Moran models present in the Methods are Cannings models. Different reproduction models can lead to the same coalescent limit, e.g. the Wright-Fisher and Moran model both lead to Kingman's coalescent. If the coalescent limit is identical for two constant population size reproduction models (and the number of generations to form one coalescent time unit is of order N^η), we can describe the limit as in Eq. (1) for both models. Thus, adding population size changes can still lead to a difference in coalescent limit via changing the power η of the population size ratio ν. For instance, $\eta = 2$ for the standard Moran model but $\eta = 1$ for the Wright-Fisher model (Λ the point mass in 0 in both cases). In the case of exponential growth (on the coalescent time scale), we see that the factor influenced by η in Eq. (1) equals $\nu(t)^{-\eta} = \exp(\eta g t)$. This means that we can still interpret parameters assuming one reproduction model (model 1) leading to the coalescent (with scale parameter $\eta = x_1$) under the assumption of an alternative reproduction model (model 2 with scale parameter $\eta = x_2$) by simply re-scaling the exponential growth parameter g from model 1 as $g' = g\frac{\eta_1}{\eta_2}$. For instance, a growth rate of $2g$ in the Wright-Fisher model corresponds to a growth rate of g in the Moran model.

For our two MMC models, this also means that we could analyze the models based on alternative reproduction models. For instance, we set $\gamma = 1.5$ for the discrete reproduction model leading to the Psi-coalescent, but we could also choose any other $1 < \gamma < 2$. For the Beta-coalescent, there is indeed a very appealing alternative reproduction model due to Schweinsberg [Sch03]. This alternative model assumes, for $1 < \alpha < 2$, that each individual at each generation independently produces a number of offspring following a power law distribution with tail parameter α (infinite variance), and that the next generation (the individuals surviving long enough to reproduce) is sampled from these offspring. In this model, one unit of coalescent time corresponds to an order of $N^{\alpha-1}$ generations. As discussed above, if $\alpha > 1$, we can interpret any growth rate g when seeing the Beta-coalescent as the genealogy model based on the modified Moran model with growth rate $g' = \left(1 + \frac{1}{\alpha-1}\right) g$ under Schweinsberg's model.

A.2 Properties of the reproduction model underlying the Beta-coalescent

The modified Moran model with distribution given on p.4 leading to the Beta($2-\alpha$,α)-coalescent was introduced in [HM13] and the properties of U have been additionally analyzed in [IM02]. U, or more precisely U_N since it depends on N, is distributed as the number of lineages merged at the first merger in a Beta($2-\alpha$,α)-coalescent starting with sample size N. Since when increasing the sample size, the first merger can only include more lineages, $U_N \leq U_M$ holds for $M \geq N$ (we can assume that, with increasing sample sizes, coalescent events just add branches to the tree from smaller sample sizes, see [Pit99]). This then also holds for the expected values, so $E(U_N) \leq E(U_M) \leq E(U_\infty) = \frac{\alpha}{\alpha-1}$, where U_∞ is the limit of U for $N \to \infty$. See [HM13, p.9] for the existence of the limit, whose properties including its mean are described on the cited page combined with [IM02, p.226], including its infinite variance. See Table A.1 for some properties of U_N for different N and α, computed from the definition of U_N and the listed properties of its limit.

N	α	$E(U_N)$	$\sqrt{Var(U_N)}$	$P(U \leq x_{min}) \geq 0.99$
5000	1.1	6.54	46.87	62
10000	1.1	6.83	64.24	62
25000	1.1	7.20	97.26	62
∞	1.1	11.0	-	62
5000	1.5	2.96	9.39	15
10000	1.5	2.97	11.27	15
25000	1.5	2.98	14.29	15
∞	1.5	3.00	-	15
5000	1.9	2.11	1.39	4
10000	1.9	2.11	1.50	4
25000	1.9	2.11	1.64	4
∞	1.9	2.11	-	4

Table A.1: Properties of U_N for the modified Moran model underlying the Beta-coalescents. x_{min}: Minimal integer x such that $P(U \leq x) \geq 0.99$.

A.3 Mathematical derivation of the pseudolikelihood function Eq. (2)

We follow the derivation from [EBBF15, Eq. 11]. We want to compute the likelihood of seeing the observed SFS $s_1, \ldots, s_{n-1}$ under a given coalescent (here a Beta-n-coalescent or a Psi-n-coalescent with exponential growth, but the derivation works for any coalescent model). Let $s = \sum_{i=1}^{n-1} s_i$ be the number of observed segregating sites. We assume the fixed-s approach, e.g. we assume that the distribution of the SFS is given by placing s mutations at random on the genealogical tree. Under the fixed-s assumption, the probability of observing the SFS is given by the multinomial distribution

$$P(SFS = (s_1, \ldots, s_{n-1})) = \mathbb{E}\left[\frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{T_i}{T_{tot}}\right)^{s_i}\right], \tag{5}$$

since a segregating site has mutant allele frequency i if it lands on a branch that supports i leaves (T_i is the sum of lengths of branches supporting i leaves, $T_{tot} = \sum_{i=1}^{n-1} T_i$ is the total length of the genealogy). Under further assumptions of independence of the different fractions $\frac{T_i}{T_{tot}}$ of the total branch length and approximating $E\left(\frac{T_i}{T_{tot}}\right) \approx \frac{E(T_i)}{E(T_{tot})}$, we have a further approximation

$$P(SFS = (s_1, \ldots, s_{n-1})) = \frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{E(T_i)}{E(T_{tot})}\right)^{s_i}, \tag{6}$$

Next, we consider the addition of a misorientation probability, e, describing the switch of ancestral and derived states. Eq. (6) constitutes a multinomial distribution, which can be interpreted as throwing s balls into compartments 1,...,$n-1$, where compartment i is hit with probability $\frac{E(T_i)}{E(T_{tot})}$. Misorienting the allele in this interpretation means that a ball that originally lands in compartment i is placed in compartment $n-i$ instead. If this happens with probability e, a ball consequently lands in compartment i with probability $(1-e)\frac{E(T_i)}{E(T_{tot})} + e\frac{E(T_{n-i})}{E(T_{tot})}$. So the probability to observe a specific SFS when ancestral and derived types can be confused is

$$P(SFS{=}(s_i)_{i=1}^{n-1}) = \frac{s!}{s_1! \cdots s_{n-1}!} \prod_{i=1}^{n-1} \left(\frac{(1-e)E(T_i) + eE(T_{n-i})}{E(T_{tot})}\right)^{s_i}, \tag{7}$$

where $(s_i)_{i=1}^{n-1} = (s_1, \ldots, s_{n-1})$ is the observed SFS. This is again a multinomial distribution.
Simulations showed that inferring parameters via a pseudolikelihood approach based on Eq. 7 tends to overestimate e to fit the U-shape. To counteract this, we couple this equation with an alternative estimation of e by using polymorphic sites discarded in the process of polarizing the SFS due to having a third allele in the outgroup. As described in [Lap17, Section 4.2] or [BD03, p. 1620], these sites carry information about e. Let $S_0 = S + S_{\neq}$ be the total number of biallelic SNPs in the sample, where S is the (random) number of sites where the outgroup does not show a third allele not observed in the sample (left and central trees in figure A.1) and $S_{\neq}$ the number of sites where it does (right trees in Figure A.1). Observe that s is the observed outcome of S, the total sum of the observed SFS.

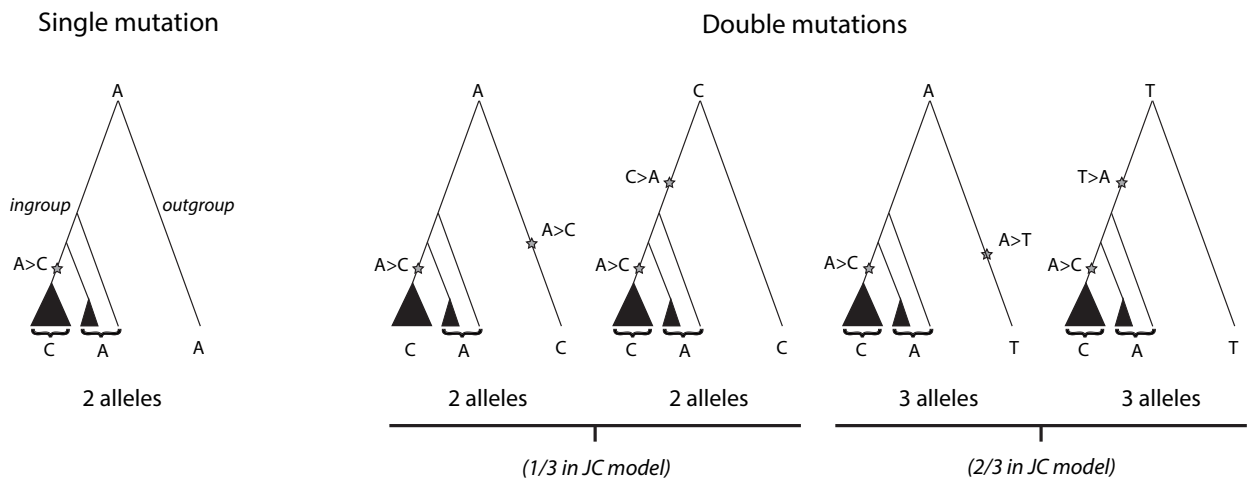


Figure A.1: Sketch of trees with mutations to illustrate how tri-allelic sites relates to the probability of misorientation. Left and central trees have one of their variant equal to the outgroup (counted in s) whereas the left tree have 3 differents alleles (counted in $s_{\neq}$). Under a Jukes and Cantor setting, the expected number of misoriented variants (central trees) equals half the number of tri-allelic sites.

Consider a polymorphic site in the sample (ingroup). If one of the allele is the same than the outgroup, there was either a single mutation (most likely for closely related outgroup) or two mutations, the second one masking the effect of the first (central trees of figure A.1). In the Jukes-Cantor model where all mutations are equally likely, if the probability of having a specific mutation $X > Y$ masking the effect of the first is p, then the probabilties of observing 2 alelles is p and the probability of observing 3 alleles is $2p$. We emphasise that while we assume an infinite sites model within the sample, we allow reverse mutations here due to the considerably longer branch lengths.

Following this, we can compute the probability $P(S_{\neq} = s_{\neq} | S_0 = s + s_{\neq})$ that we observe exactly $s_{\neq}$ sites which are biallelic within the sample but have a third allele for the outgroup. This is simply binomial sampling from S_0 biallelic sites with success probability $2p$. We can also express this probability in terms of the misorientation probability e. Let $v \in SFS$ be the event that a bialllelic site (variant) can be polarized via outgroup (which means it has one of the two alleles of the sample also for the outgroup) and $mis(v)$ the event that the ancestral state of v is misidentified. The probability that a site of the SFS can be polarized (displaying one allele similar to the outgroup) is $1 - p$. We have

$$e = P(mis(v) | v \in SFS) = \frac{P(mis(v), v \in SFS)}{P(v \in SFS)} = \frac{p}{1 - 2p},$$

and thus equivalently $p = \frac{e}{1+2e}$. This leads to

$$P(S_{\neq} = s_{\neq} | S_0 = s + s_{\neq}) = \binom{s + s_{\neq}}{s_{\neq}} \left(\frac{2e}{1+2e}\right)^{s_{\neq}} \left(\frac{1}{1+2e}\right)^{s}. \tag{8}$$

We now assume a composite likelihood, multiplying Eqs. (7) and (8). Conditional on observing $s + s_{\neq}$ segregating sites from which s can be polarized via outgroup and form the SFS, the pseudo-likelihood of observing a specific SFS is given by Eq. (2).

Remark A.1. *Eq.* (8) *shows that S given S_0 is binomially distributed with S_0 draws with success probability $1 - 2p = (1 + 2e)^{-1}$. Let $X(S_0)$ be a r.v. with this distribution. We will use this to simulate S_0 based on S and the mis-classification probability e: The maximum likelihood estimate for the number of trials S_0 of the binomial r.v. $X(S_0) \sim Bin(S_0, (1 + 2e)^{-1})$, in the sense of maximising $P(X(S_0) = s)$, is $\hat{S}_0 = \lfloor (1 + 2e)s \rfloor$, since*

$$\frac{P(X(S_0 + 1) = s)}{P(X(S_0) = s)} = \frac{S_0 + 1}{S_0 + 1 - s} \frac{2e}{1 + 2e} \geq 1 \Leftrightarrow S_0 \leq s(1 + 2e) - 1.$$

We will use this estimate to simulate a reasonable $s_{\neq}$. If we simulate a SFS with s mutations, and we flip each mutation in it with frequency e from class i to $n - i$, we then simulate $s_{\neq}$ as a binomial draw from $\hat{S}_0 = \lfloor (1 + 2e)s \rfloor$ Bernoulli r.v.'s with success probability $\frac{2e}{1+2e}$. This is denoted as the $\hat{S}_0$ approach.

A.4 Cramér's V as a goodness-of-fit measure

Our assumptions leading to Equations (6), (7) can be interpreted that each variant observed for the SFS is sampled from a multinomial distribution from the 'true' allele frequency spectrum. In the following, we denote the multinomial approximation of the SFS entry frequencies, the 'true' spectrum, by $(p_1, \ldots, p_{n-1})$. Since assuming sampling from a multinomial distribution is also the statistical model behind the χ^2 goodness-of-fit test, we chose the effect size measure Cramér's V [Cra16, ch. 21] of this test, defined as

$$V = \sqrt{\sum_{i=1}^{n-1} \frac{(o_i - p_i)^2}{p_i(n-2)}},$$

to quantify the lack of goodness of fit (o_i is the observed frequency of mutations with frequency i/n among all mutations). This measure can be interpreted as a dimensionless version of the χ^2 test statistic, since the mutation counts do not enter, just the mutation frequencies and the additional factor $n - 2$ corrects for unequal sample sizes.

A.5 Assessing estimation errors

A.5.1 Simulation and inference setup

As a rough approximation of a genome, we simulated 100 independent loci (ignoring the fine structure of weakly physically linked loci and long range LD, see Appendix A.8). This means that the genealogical trees of the loci are independent and follow the same tree distribution, e.g. realisations of a Beta coalescent with exponential growth with rate g and coalescent parameter α. The mutations on each tree are independent of all trees (and mutations on other trees) and given by a Poisson process with rate $\frac{\theta}{2}$. We assumed three different sample sizes $n = 20$, $n = 25$ and $n = 100$. For each locus, we set the mutation rate so that on average 50 mutations appear, *i.e.* we set $\theta = 100/E[T_{tot}]$ (generalized Watterson estimate), where T_{tot} is the sum of all branch lengths of the locus' genealogy. Mutations are interpreted under the infinite-sites model, resulting in simulated SNP sequences (ancestral vs. derived type). For each SNP, we then flip ancestral and

derived allele with probability e. We simulate 500 SNP sequences as described above for each combination of coalescent parameter α or Ψ, growth rate g and misorientation probability e from the following two sets (the first set has α, Ψ and g on the inference grid, the second uses off-grid values).

- Set 1: equidistant $\alpha \in \{1, 1.05, 1.1, \ldots, 2\}$ and $\Psi \in \{0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$
  Set 2: $\alpha \in \{1.025, 1.325\}$, $\Psi \in \{0.025, 0.075\}$ (additionally $\Psi = 0.005$ for $n = 25$)
- Set1: $g \in \{0, 0.5, 1, 10\}$, Set 2: $g \in \{0.25, 2.25, 11.25\}$
- Set 1: $e \in \{0, 0.01, 0.05, 0.1\}$ (essentially on grid), Set 2: $e \in \{0, 0.015, 0.045, 0.095\}$

To infer via Eq. (2), we also need the total number of segregating sites $s + s_{\neq}$, adding the number of segregating positions not included in the SFS due to not being able to polarize them. For this, we use the $\hat{S}_0$ approach described in Remark A.1.

A.5.2 Parameter and model selection accuracy

First, for $n = 20$ and $n = 100$, we estimate parameters using Eq. 2 using the same coalescent model (Beta or Psi) on equidistant grids with $\alpha \in \{1, 1.05, 1.1, \ldots, 2\}$ or $\Psi \in \{0, 0.5, 0.1, \ldots, 1\}$, $g \in \{0, 0.05, \ldots, 25\}$, $e \in \{0.001, 0.011, \ldots, 0.201\}$. For this, we only used the on-grid values (Set 1). Results are shown in Figures 1, A.2 – A.4,A.8 – A.19.
Second, we assess the error of our model selection approach based on approximated Bayes factors for $n \in \{20, 25, 100\}$. For this, we fixed different values of Ψ and α from Set 1 and Set 2 including $\alpha = 2$.

We then picked 2,000 simulations at random from all parameter combinations with this fixed coalescent parameter (as described above) and performed model selection via Bayes factors as described in the method described in the main document (section 2.2). The maximum was taken on the same equidistant grids as for the parameter estimation. Expanded results are provided in Tables A.2 and A.3.

A.5.3 Parameter estimation accuracy - results

For inferring parameters under the Beta-coalescent or the Psi-coalescent, Figures 1, A.2–A.4 show the error distribution of all three parameters for $n \in \{20, 100\}$ across all simulation parameter choices. While g cannot be estimated precisely in some cases, e, Ψ and, to a lesser degree, α, can generally be estimated rather well, especially if sample size $n = 100$.

These errors distribute over the different parameter settings as shown in Figures A.8 – A.19. Most notably, large errors when estimating growth rates only happen if the growth rate is also large. For Psi-coalescents, we see that choosing Ψ between grid points are still mostly captured by the adjacent Ψ grid points.

sample size	true model	grid	Fraction model inferred as			
			Kingman	Beta	Psi	MMC
$n = 20$	$\alpha = 2$	yes	0.77	0.22	0.00	0.00
	$\alpha = 1.9$	yes	0.37	0.58	0.02	0.02
	$\alpha = 1.8$	yes	0.06	0.79	0.09	0.08
	$\alpha = 1.625$	no	0	0.82	0.09	0.08
	$\alpha = 1.025$	no		0.99	0.01	0
	$\alpha = 1$	yes		0.99	0	0
	$\Psi = 0.025$	no	0.14	0.59	0.15	0.12
	$\Psi = 0.05$	yes	0.01	0.17	0.70	0.11
	$\Psi = 0.075$	no		0.12	0.82	0.06
	$\Psi = 0.1$	yes		0.04	0.94	0.02
$n = 25$	$\alpha = 2$	yes	0.79	0.21		
	$\alpha = 1.9$	yes	0.34	0.66	0	0
	$\alpha = 1.8$	yes	0.02	0.91	0.04	0.03
	$\alpha = 1.625$	no		0.9	0.06	0.05
	$\alpha = 1.025$	no		1		0
	$\alpha = 1$	yes		1		0
	$\Psi = 0.005$	no	0.55	0.45		
	$\Psi = 0.025$	no	0.05	0.72	0.14	0.09
	$\Psi = 0.05$	yes	0	0.12	0.82	0.06
	$\Psi = 0.075$	no		0.06	0.91	0.03
	$\Psi = 0.1$	yes		0.02	0.98	0
$n = 100$	$\alpha = 2$	yes	0.87	0.13		
	$\alpha = 1.9$	yes	0.12	0.88		
	$\alpha = 1.8$	yes		1		
	$\alpha = 1.625$	no		1		
	$\alpha = 1.025$	no		1		
	$\alpha = 1$	yes		1		
	$\Psi = 0.025$	no		0.92	0.06	0.02
	$\Psi = 0.05$	yes			1	
	$\Psi = 0.075$	no			1	
	$\Psi = 0.1$	yes			1	

Table A.2: Model selection via two-step Bayes factor criterion. Based on 2,000 simulations for each true model assuming 100 loci with 50 observed mutations. 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 column grid shows whether the parameters used for simulation were included in the inference grid. For details on both simulations and inference parameters see Appendix A.5. Fractions are rounded to two digits.

A.6 Visual correlations between the estimated parameters

We provide a graphical view of the correlation between the parameters inferred between the Kingman coalescent and the Beta-coalescent (Figure A.5) or between both MMC models (Figure A.6).

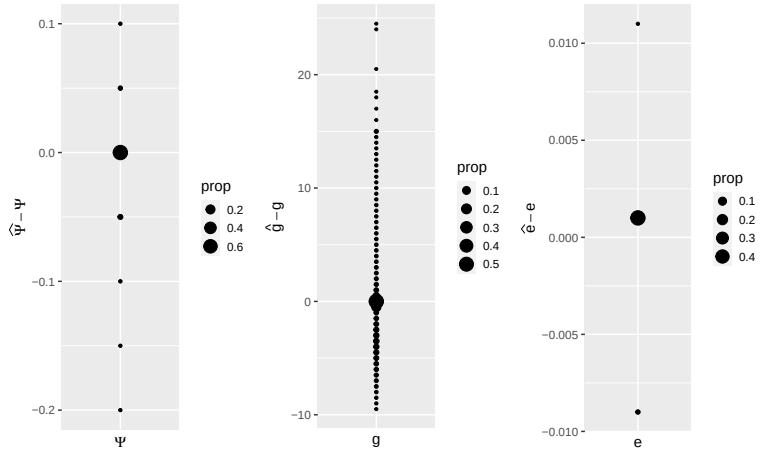


Figure A.2: Error for estimating parameters for Psi-coalescents ($n = 100$, with growth and misclassification) across all simulation scenarios

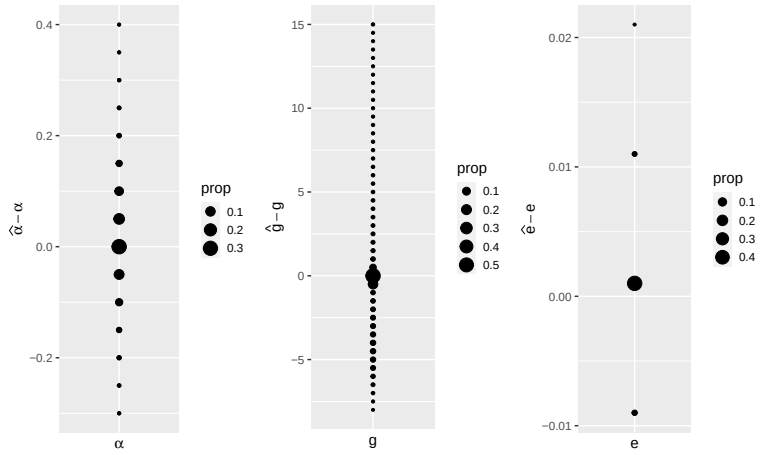


Figure A.3: Error for estimating parameters for Beta-coalescents ($n = 20$, with growth and misclassification) across all simulation scenarios

sample size	$n = 20$	$n = 100$
$\alpha = 1.75$	0.01	
$\alpha = 1.8$	0.04	
$\alpha = 1.85$	0.22	0.04
$\alpha = 1.9$	0.37	0.53
$\alpha = 1.95$	0.34	0.43
$\Psi = 0.05$	0.02	

Table A.3: Fractions of estimated parameters of model-misidentified coalescent simulations with $\alpha = 2$. If the two-step Bayes factor model inference recorded "MMC", the Beta parameter is reported.

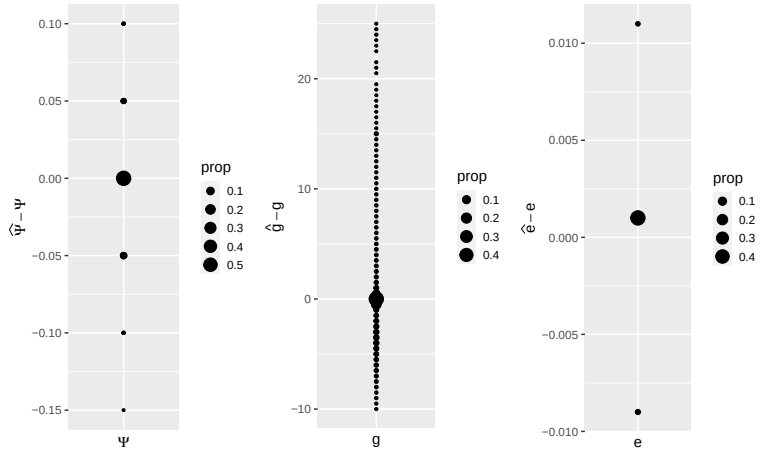
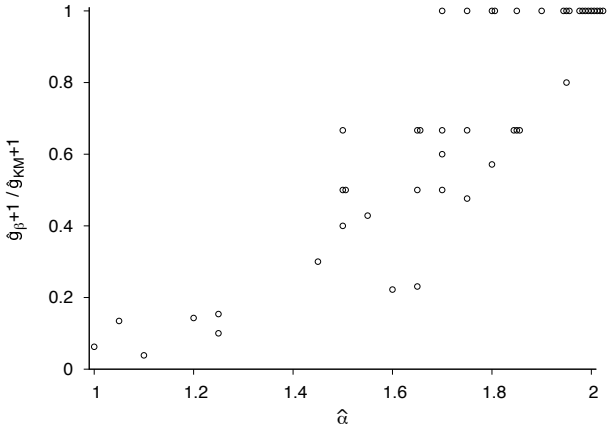
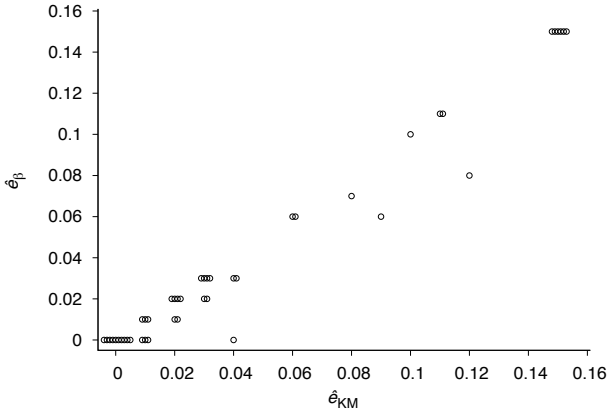


Figure A.4: Error for estimating parameters for Psi-coalescents ($n = 20$, with growth and misclassification) across all simulation scenarios



(a) Kingman model and Beta-coalescent model $\hat{g}+1$ ratio in function of $\hat{\alpha}$.



(b) Kingman model $\hat{e}$ in function of the corresponding Beta-coalescent allele misorientation parameter.

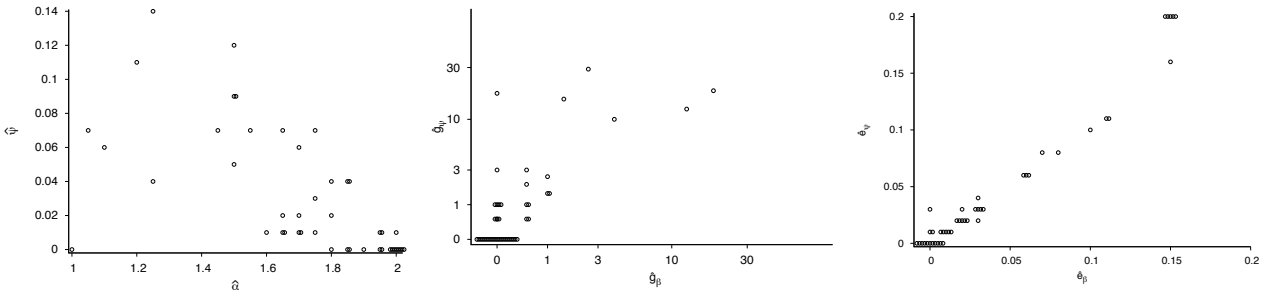
Figure A.5: Comparison of parameters between Kingman model and Beta-coalescent model.

A.7 Correction for GC-bias

We use the approach from [PATE18] and consider the subset of SNPs corresponding to $A \leftrightarrow T$ and $G \leftrightarrow C$ substitutions, which are not affected by biased gene conversion. We overlaid these neutralized SFS to the observed SFS and the predictions of the fitted models in supplementary files 1,2.

A.8 Non-independence of unlinked loci under multiple merger genealogies

Here, we address the issue that physically unlinked loci in multiple merger genealogies still have dependent genetic diversity. For Λ-coalescents, which all our coalescent models are, the issue can be easily understood within the approximate multi-unlinked-locus model from the appendix of [Kos18]. In this model, multiple mergers result from large families appearing in a short amount of evolutionary time (see also a more thorough explanation in [MGF20]), so these families affect not only one, but all loci. Due to the model definition of MMC, each ancestral lineage can join one of such events with the same probability. Thus, if this probability



(a) α/Ψ (b) g (c) e

Figure A.6: Comparison of Beta-coalescent (x-axis) and Psi-coalescent (y-axis) parameters inferred from each SFS.

is high, there will be a merger of similar size at each or nearly each locus in the genome, introducing a dependency between loci. The strength of this dependency should be correlated to the probability with which an ancestral lineage merges, following Remark 1 in [Kos18]. This probability x is generated by a Poisson process whose rate is proportional to $x^{-2}\Lambda(dx)$, where Λ is the associated measure of the coalescent (a Beta distribution or the point mass in Ψ for our model classes). This probability is rather small for Beta coalescents, but high for high Ψ values, see also Figure A.20.

A.9 Population structure scans

We performed two simple checks for population structure: PCA and `find.clusters` from the R package `adegenet` [JA11]. For PCA, we coded alleles as 0 and 1, imputed missing data as the mean allele at the site, to then perform a double-centered PCA: PCA, as implemented in `adegenet`, was performed on the SNP matrix after subtracting row means and column means (and adding the overall mean), see [GJQP+19, p.20]. The approach behind `find.clusters` is to first perform a standard PCA and then group individuals by running the k-means clustering algorithm on the principal component coordinates for different numbers of clusters. Based on the goodness-of-fit criterion BIC, we chose the 'optimal' k as the smallest value of k that is visibly a local minimum (essentially the elbow criterion). For large data sets of more than 1 million SNPs, we performed the analysis with a reduced data set by filtering down the number of SNPs by only retaining each xth SNP where $x = \frac{\#\ SNPs}{1000000}$, rounded to the lower integer. Results are shown in Supplementary file 3 and Table A.6. For diploids, PCA and `find.clusters` results were not qualitatively affected by performing them on either haplotypes or diploid genotypes. For *D. melanogaster*, we performed the population structure scan separately on Chromosome 2L, 2R, 3L, 3R (with filtering down as described above). For the human data, we omitted the X and Y chromosomes.

A.10 Nucleotide diversity across the genome

We recorded the (sample) mean and standard deviation of the per-site nucleotide diversity in non-overlapping windows of 15,000 sites along the genome (resp. the sequenced part of it). For computation, we used the R package `pegas` [Par10] and `vcftools` [DAA+11] (for haploid data presented as vcf files, we used J. Dutheils fork of `vcftools` `https://github.com/jydu/vcftools`). For the human data, we omitted the X and Y chromosomes and for *D. melanogaster*, we used chromosomes 2L, 2R, 3L, 3R.
Results are shown in Table A.7 and Figure A.21.

A.11 Effect of non-extreme demography on the SFS

The expected SFS $(E(S_1), \ldots, E(S_{n-1}))$ for a given genealogy tree (conditional on waiting times and topology) is a linear function of the waiting times C_k for the next coalescence event of the sample genealogy if k ancestral lineages are present, with coefficients $kP_{n,k}(i)$ dependent on the topology, where $P_{n,k}(i)$ is the probability that a random branch at level k has i descendants in the sample ([Fu95, FLW+17, SW08]). For a sample from a population with panmictic, neutral dynamics and finite variance in offspring number, corresponding to a Kingman coalescent where time is rescaled by a deterministic strictly monotonic function, all tree topologies are equally probable and independent of waiting times. The expected coefficients are given by $\mathrm{E}[kP_{n,k}(i)] = k\binom{n-i-1}{k-2}/\binom{n-1}{k-1}$. It should be noted that the assumption of monotonic time change ensures that the genealogy stays bifurcating: extreme changes in population size violate this and may lead to multiple merger genealogies.

The SFS for a large population described by the time-rescaled Kingman coalescent can be obtained as the large sample limit $n \to \infty$ of the above spectrum [FKR+18]. For large n, the probability that a random lineage at level k takes a fraction f of the descendants is $\mathrm{E}[P_{n,k}(fn)] \to (k-1)(1-f)^{k-2}df$. Hence the continuous expected SFS is given by equation (4), which depends on the expected population-level waiting times $c_k = \mathrm{E}[C_k] > 0$ for $k = 2 \ldots \infty$. The positivity of all coefficients in this expansion implies that for a finite expected TMRCA $\sum_{k=2}^{\infty} T_k < \infty$, the expected SFS for populations with non-extreme demography is an absolutely monotonic function of $1-f$ in $[0,1)$, and therefore a completely monotonic function of the frequency f in $(0,1]$. This is the case for all non-extreme demographies with bounded past population size, since all of them have finite expected TMRCA.

A.12 Psi-coalescent graphs

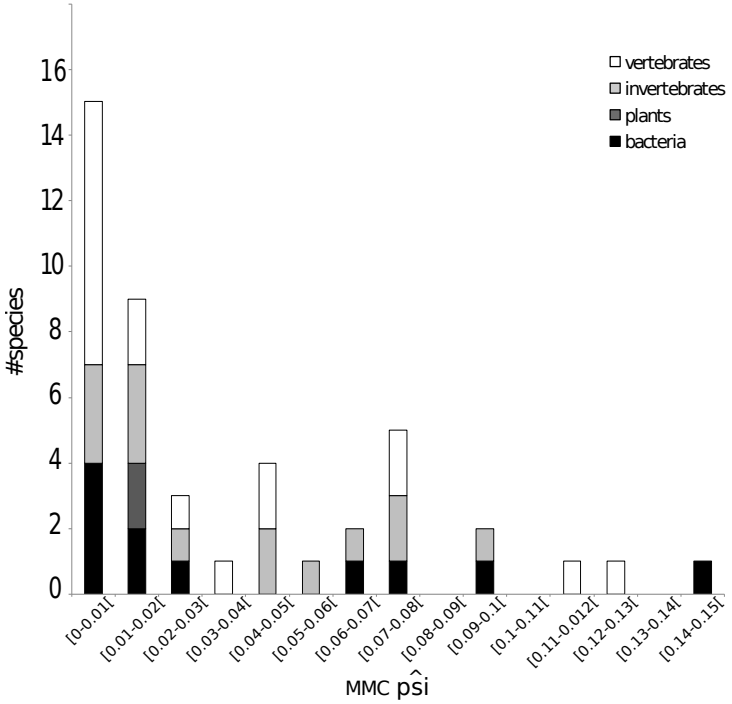


Figure A.7: Distribution of psi parameter in function of the order of the species (white: vertebrates, light grey: invertebrates, dark grey: plants, black: bacteria).

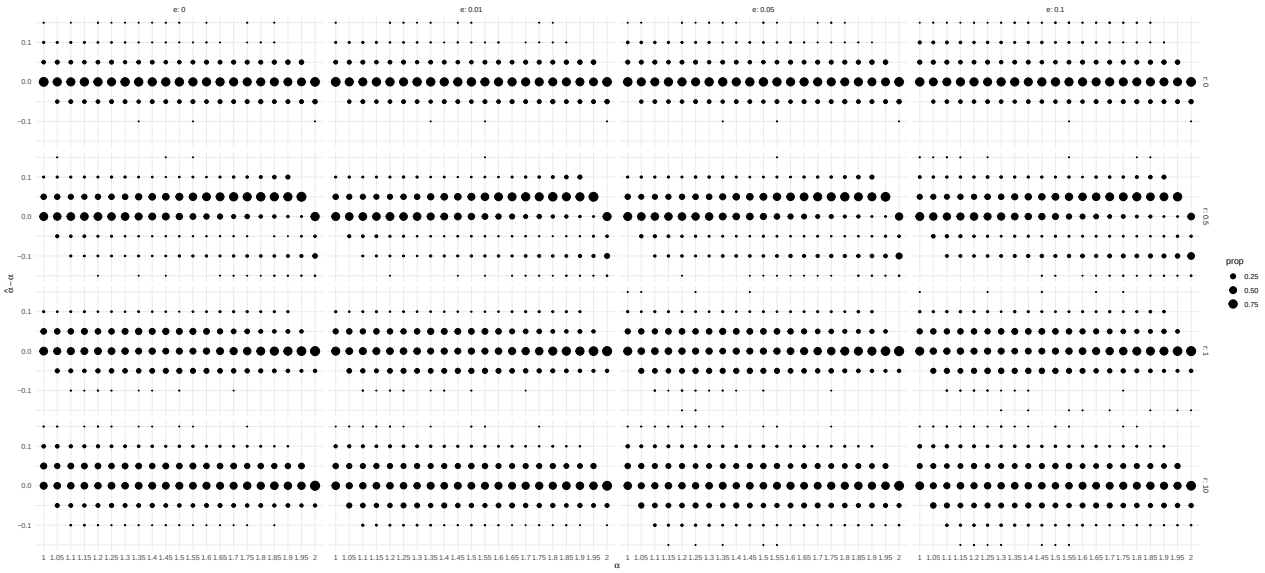


Figure A.8: Error for estimating coalescent parameter α for Beta-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

Table A.4: Parameters Estimations

Species	g_{KM}	e_{KM}	V_{KM}	α	g_{Beta}	e_{Beta}	V_{Beta}	Ψ	g_{Ψ}	e_{Ψ}	V_{Ψ}	Model
Acinetobacter baumannii	0	0.1	0.064	1.8	0	0.1	0.059	0.02	0	0.1	0.061	Beta
Aptenodytes patagonicus	24	0.01	0.071	1.25	1.5	0	0.047	0.04	15.5	0.01	0.063	Beta
Arabidopsis thaliana	3.5	0.08	0.019	1.6	0	0.07	0.010	0.01	1	0.08	0.017	Beta
Armadillidium vulgare	0	0.03	0.089	1.7	0	0.03	0.069	0.06	0	0.02	0.083	Beta
Artemia franciscana	0.5	0.03	0.067	1.65	0	0.03	0.055	0.02	0.5	0.03	0.065	Beta
Athene cunicularia	2.5	0.03	0.040	1.8	1	0.03	0.037	0	2.5	0.03	0.040	Beta
Bacillus subtilis	5.5	0.15	0.085	1.25	0	0.15	0.079	0.14	0	0.2	0.062	Ψ
Caenorhabditis brenneri	1.5	0.09	0.094	1.5	0	0.06	0.086	0.09	0	0.06	0.105	Beta
Caenorhabditis elegans	0	0.06	0.142	2	0	0.06	0.1422	0	0	0.06	0.1422	KM
Chlamydia trachomatis	0	0.11	0.105	2	0	0.11	0.105	0	0	0.11	0.105	KM
Ciona intestinalis A	0	0.11	0.053	1.95	0	0.11	0.052	0	0	0.11	0.053	KM
Ciona intestinalis B	0.5	0.03	0.085	1.65	0	0.02	0.061	0.07	0	0.02	0.068	Beta
Clostridium difficile	15	0.15	0.214	1	0	0.15	0.221	0	17.5	0.2	0.214	KM
Corvus cornix	1.5	0	0.023	1.95	1	0	0.020	0	1.5	0	0.023	Beta
Coturnix japonica	4	0.02	0.048	1.45	0.5	0.01	0.020	0.07	1.5	0.01	0.044	Beta
Culex pipiens	2.5	0.02	0.069	1.55	0.5	0.01	0.057	0.07	1	0.01	0.063	Beta
Drosophila melanogaster	5.5	0.02	0.019	1.65	0.5	0.02	0.005	0.01	3	0.02	0.017	Beta
Egretta garzetta	0	0.02	0.055	1.75	0	0.02	0.037	0.07	0	0.02	0.039	Beta
Emys orbicularis	0.5	0	0.068	1.85	0	0	0.060	0.04	0	0	0.059	KM
Escherichia coli	0	0.06	0.054	2	0	0.06	0.054	0	0	0.06	0.054	KM
Ficedula albicollis	0.5	0.01	0.029	2	0.5	0.01	0.029	0.01	0.5	0.01	0.028	Ψ
Gorilla gorilla gorilla	0	0	0.042	1.9	0	0	0.040	0	0	0	0.042	Beta
Halictus scabiosae	0	0.01	0.069	1.85	0	0.01	0.064	0.04	0	0.01	0.062	MMC
Helicobacter pilori	1	0.15	0.052	1.65	0	0.15	0.060	0.01	1	0.2	0.050	Ψ
Homo sapiens	0.5	0.01	0.010	1.85	0	0	0.011	0	0.5	0.01	0.010	Beta
Klebsiella pneumoniae	18.5	0.15	0.122	2	18.5	0.15	0.126	0	18.5	0.16	0.122	KM
Lepus granatensis	0.5	0.04	0.102	1.5	0	0.03	0.069	0.12	0	0.03	0.066	MMC
Melitaea cinxia	1.5	0.04	0.061	1.7	0.5	0.03	0.059	0.01	2	0.04	0.061	Beta
Messor barbarus	0.5	0	0.069	2	0.5	0	0.069	0	0.5	0	0.069	KM
Mycobacterium tubercolosis	25	0.01	0.118	1.05	2.5	0	0.090	0.07	29	0	0.126	Beta
Nipponia nippon	0	0.03	0.160	2	0	0.03	0.160	0	0	0.03	0.160	KM
Ostrea edulis	0	0.02	0.052	1.8	0	0.02	0.044	0.04	0	0.02	0.042	MMC
Pan paniscus	2	0	0.068	1.85	1	0	0.056	0	2	0	0.068	Beta
Pan troglodytes ellioti	0.5	0	0.052	1.7	0	0	0.028	0.02	0.5	0	0.045	Beta
Parus major	0.5	0.01	0.031	1.75	0	0.01	0.010	0.03	0	0.01	0.022	Beta
Parus caeruleus	6	0.04	0.062	1.2	0	0	0.037	0.11	1.5	0.03	0.031	MMC
Passer domesticus	0	0	0.022	2	0	0	0.022	0	0	0	0.022	KM
Phylloscopus trochilus	12.5	0	0.022	2	12.5	0	0.022	0	12.5	0	0.022	KM
Physa acuta	1	0.03	0.068	1.5	0	0.02	0.035	0.05	0.5	0.03	0.055	Beta
Pseudomonas aeruginosa	25	0.15	0.073	1.1	0	0.15	0.063	0.06	3	0.2	0.050	Ψ
Sepia officinalis	0	0.02	0.091	1.95	0	0.02	0.090	0.01	0	0.02	0.090	KM
Staphylococcus aureus	1	0.15	0.054	1.7	0	0.15	0.059	0.01	1	0.2	0.055	Ψ
Streptococcus pneumoniae	1	0.12	0.103	1.5	0	0.08	0.099	0.09	0	0.08	0.102	Beta
Taeniopygia guttata	9.5	0	0.034	1.75	4	0	0.019	0.01	10	0	0.030	Beta
Zea mays	0	0	0.033	1.95	0	0	0.031	0.01	0	0	0.030	Ψ

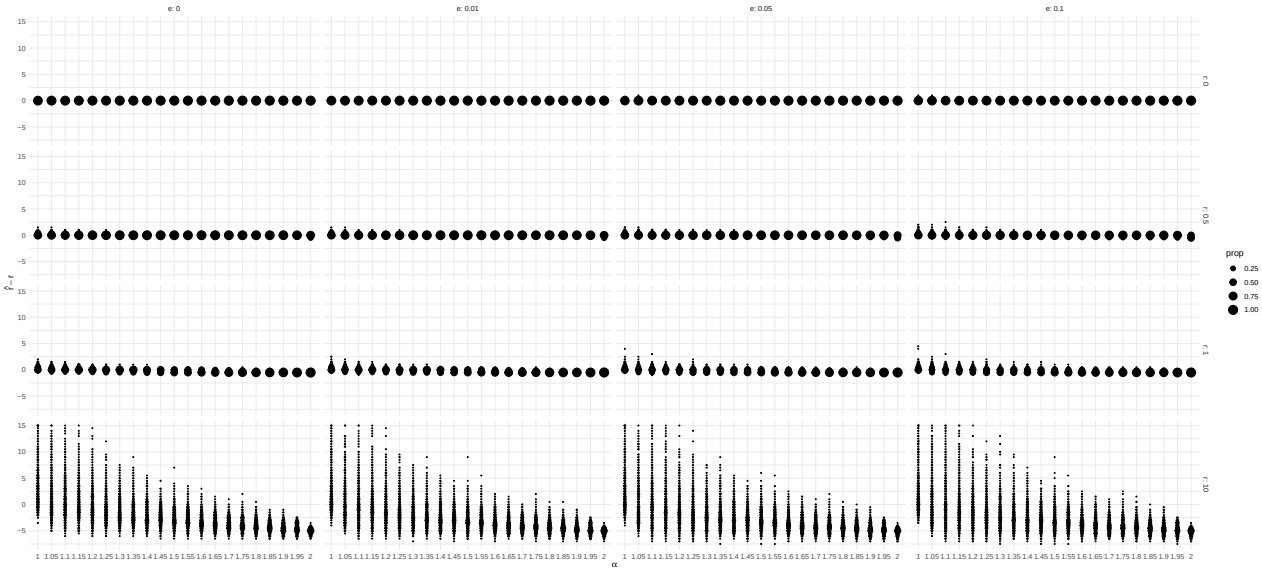


Figure A.9: Error for estimating growth rate g for Beta-coalescents with growth and misclassification ($n = 100$)

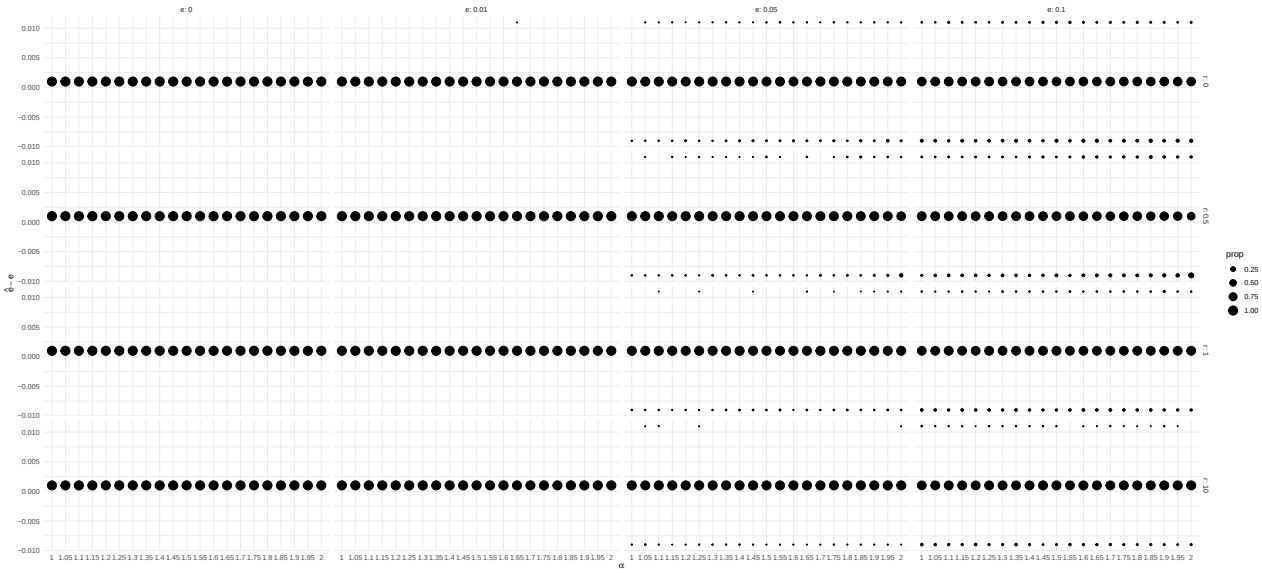


Figure A.10: Error for estimating misorientation rate e for Beta-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

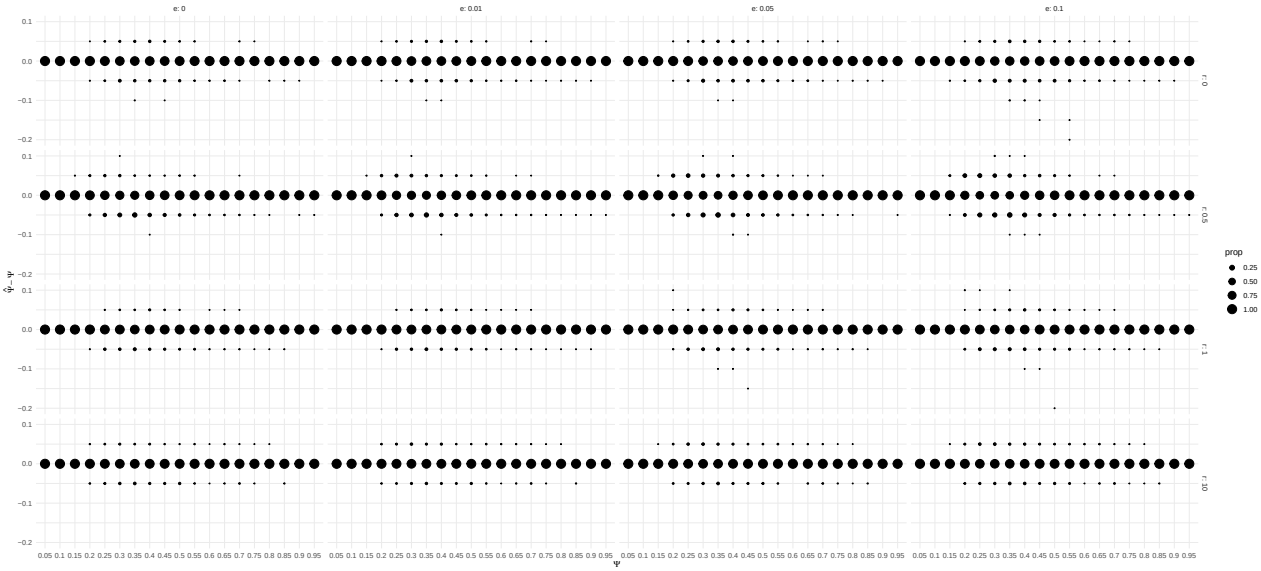


Figure A.11: Error for estimating coalescent parameter Ψ for Psi-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

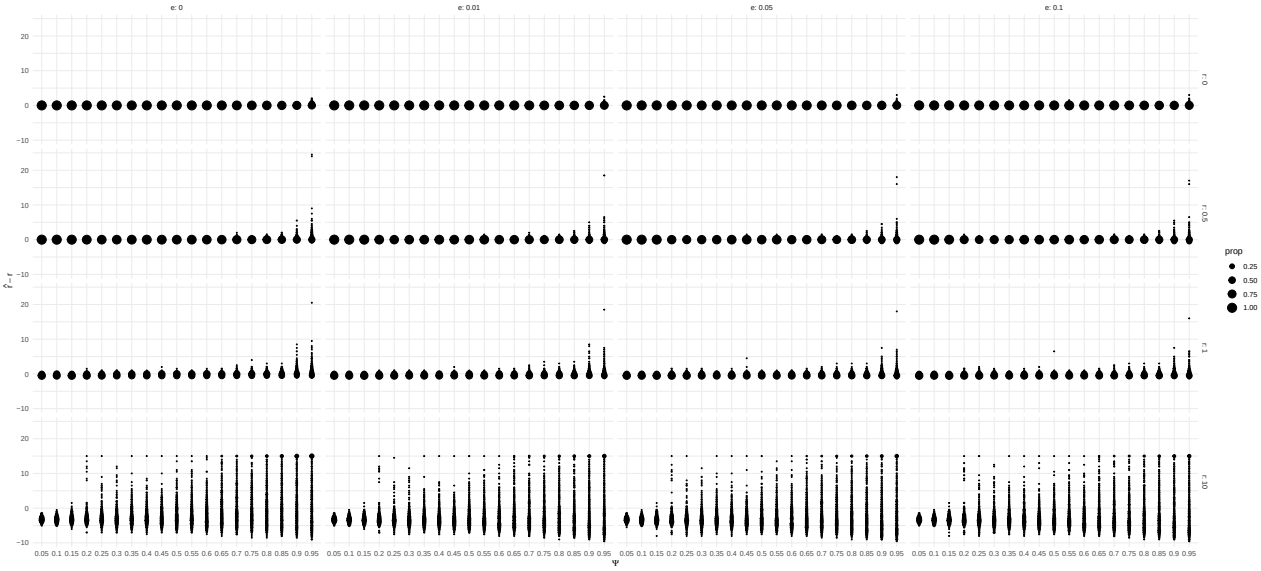


Figure A.12: Error for estimating growth rate g for Psi-coalescents with growth and misclassification ($n = 100$).

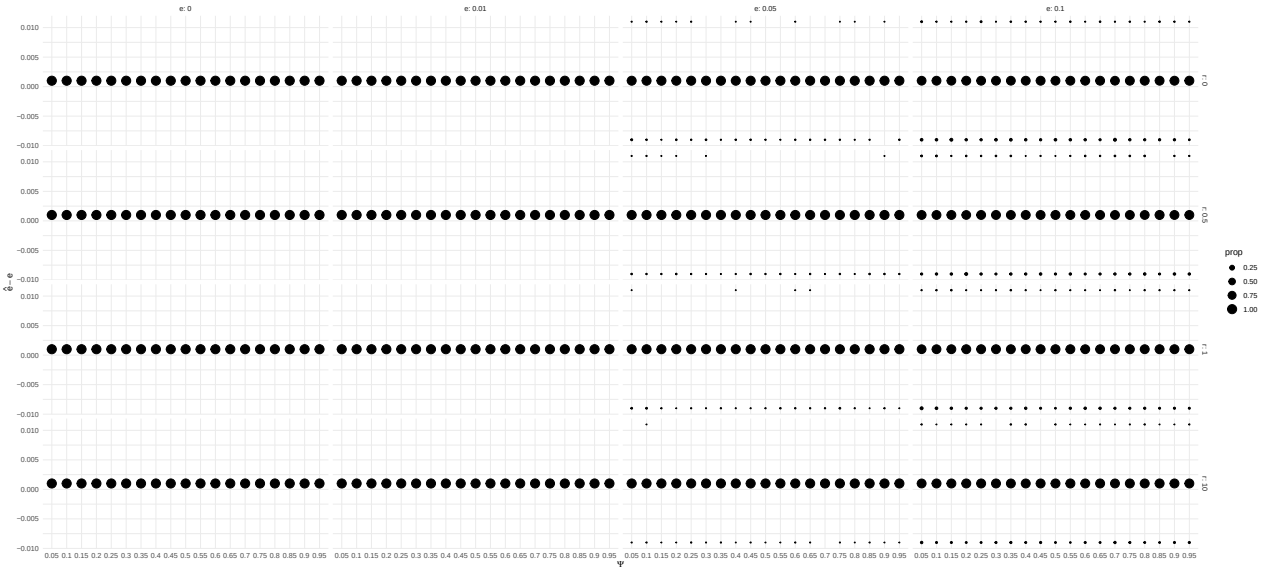


Figure A.13: Error for estimating misorientation rate e for Psi-coalescents with growth and misclassification ($n = 100$). Growth rate is denoted by g.

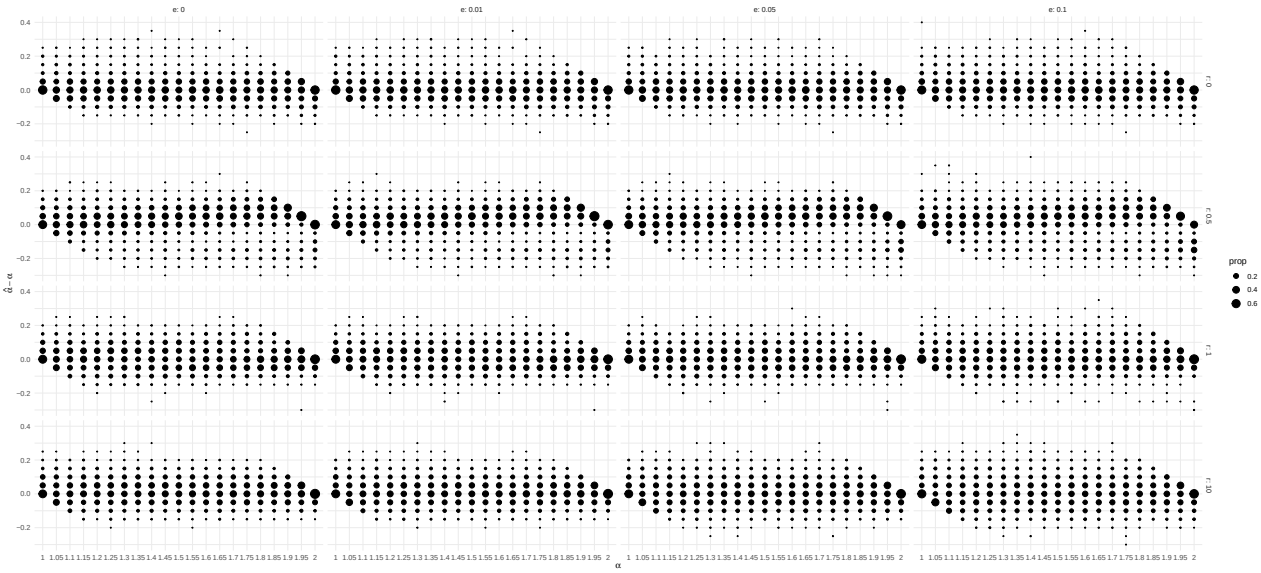


Figure A.14: Error for estimating coalescent parameter α for Beta coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

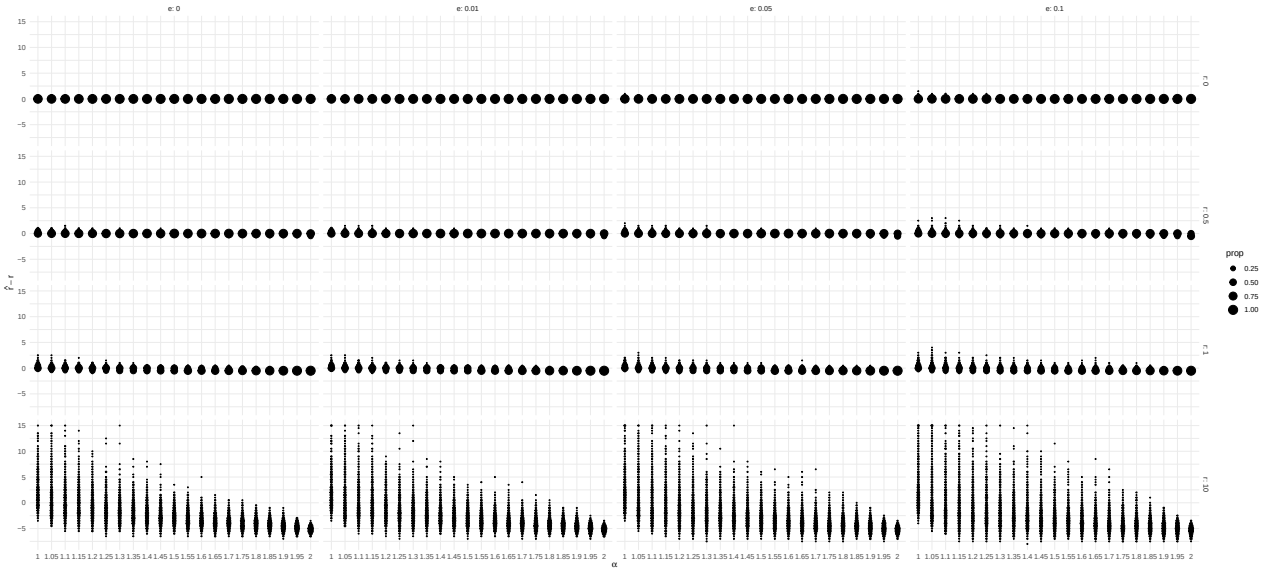


Figure A.15: Error for estimating growth rate g for Beta-coalescents with growth and misclassification ($n = 20$)

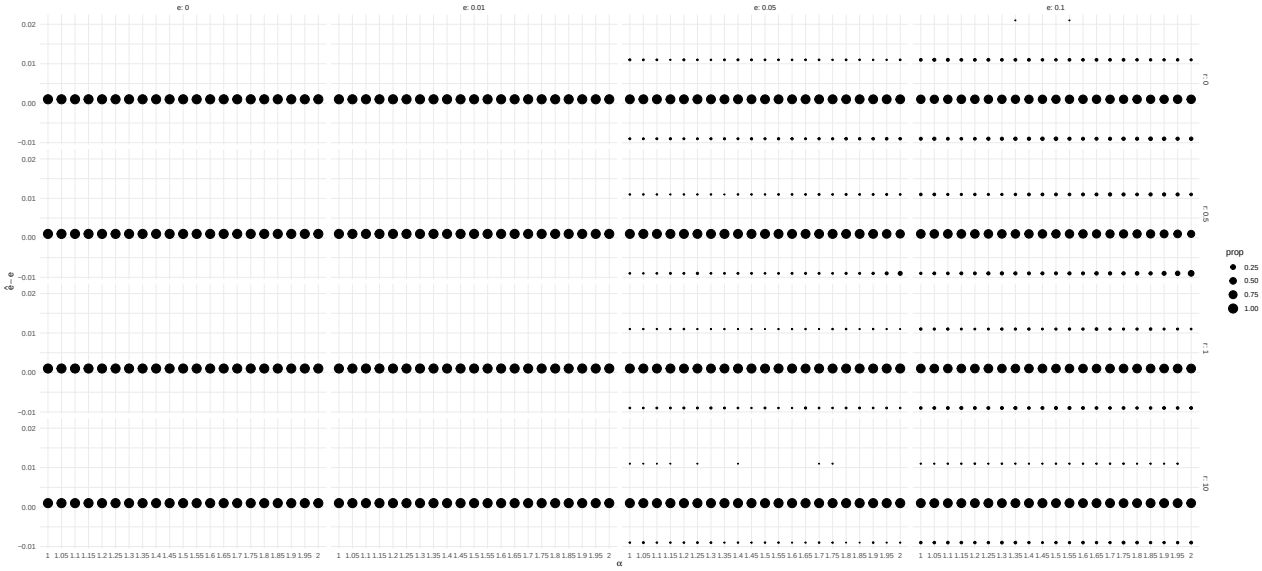


Figure A.16: Error for estimating misorientation rate e for Beta-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

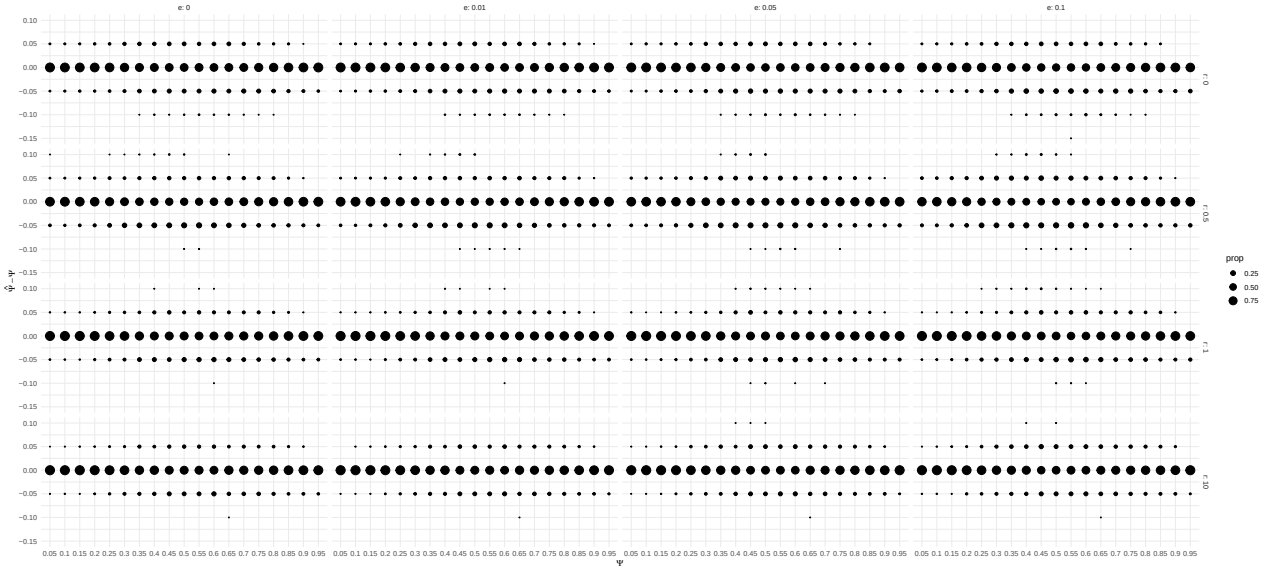


Figure A.17: Error for estimating coalescent parameter Ψ for Psi-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.

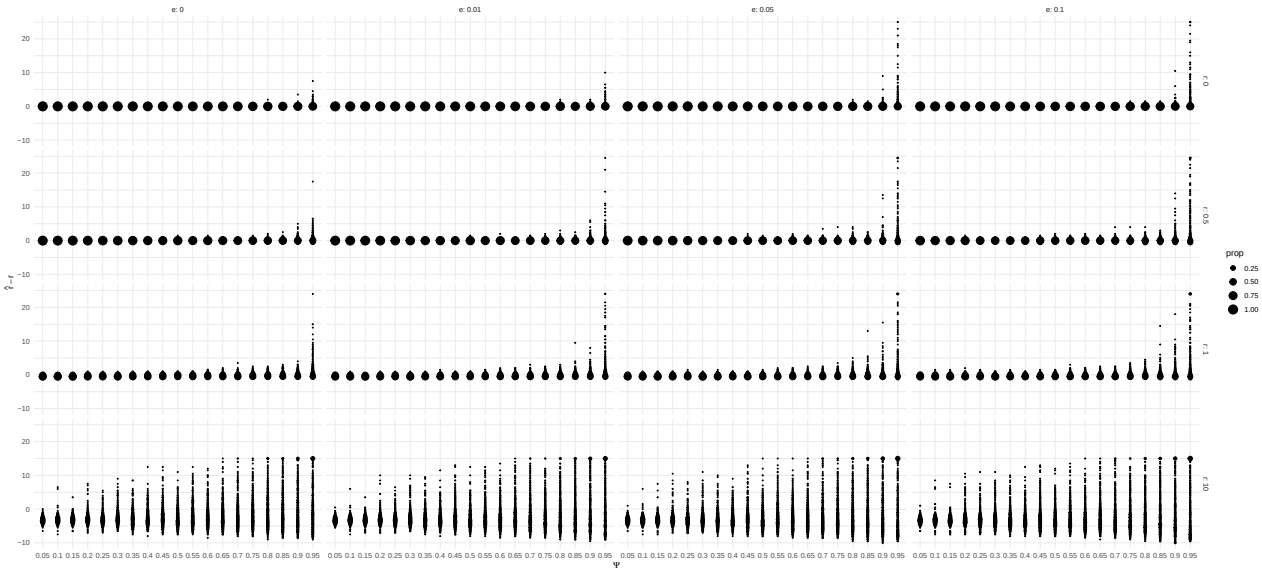


Figure A.18: Error for estimating growth rate g for Psi-coalescents with growth and misclassification ($n = 20$).

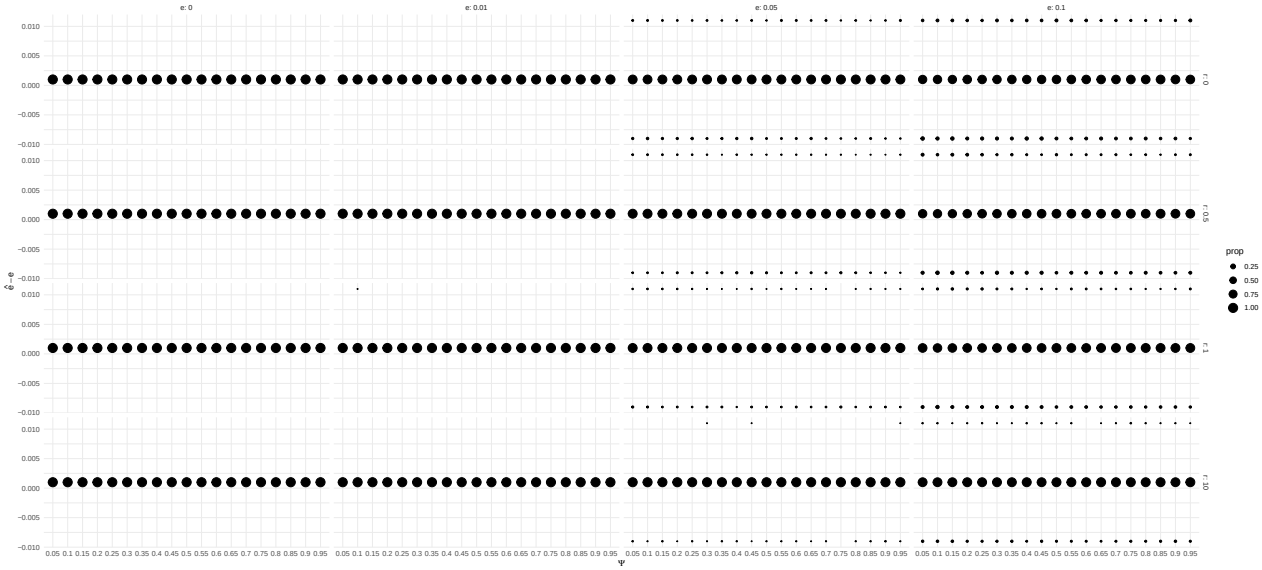
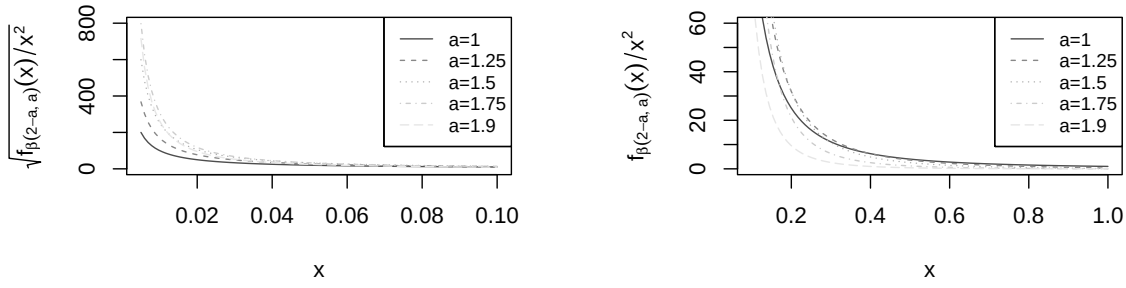


Figure A.19: Error for estimating misorientation rate e for Psi-coalescents with growth and misclassification ($n = 20$). Growth rate is denoted by g.



(a) (Improper) distribution of mergP x close to 0

(b) (Improper) distribution of bigger mergP x

Figure A.20: Distribution of merger rates of Beta-coalescents: Each lineage merges with merger probability x (abbeviated as mergP), where x is chosen with rate $x^{-2} * \Lambda(dx)$, where Λ is a Beta distribution with parameters $2 - a$ and a. Mergers only are realized if at least two lineages merge. The figures depict the corresponding (improper) density $x^{-2} * f_\beta(2 - a, a)$, where f_β is the density of the Beta distribution used. The detailed (Poisson) construction can be found in [Pit99].

Table A.5: Data set information

Species	Outgroup	n	Polarized SNP	$\neq$ outgroup	Diallelic outgroup	Source
Acinetobacter baumannii	*A. nosocomialis*	79	78175	6006		[RdSB+18]*
Aptenodytes patagonicus	*A. forsteri*	20	1278	12	32	[RGB+14]+
Arabidopsis thaliana	*A. lyrata*	345	10322757	1023148	398365	[ABAB+16]
Armadillidium vulgare	*A. nasatum*	20	23323	745		[RGB+14]+
Artemia franciscana	*A. sinica*	20	5548	247		[RGB+14]+
Athene cunicularia	*Strix occidentalis*	40	11268203	383702	68196	[MKB+18]⋆
Bacillus subtilis	*B. atrophaeus*	38	105523	29934		[RdSB+18]*
Caenorhabditis brenneri	*Caenorhabditis sp. 10*	20	1339	106		[RGB+14]+
Caenorhabditis elegans (Orsay population)	*C. elegans* ECA396 ECA723 ECA744	573	165	5	22	[RZL+18]
Chlamydia trachomatis	*C. muridarum*	59	9924	1694		[RdSB+18]*
Ciona intestinalis A	*C. intestinalis B*	20	480	94	1641	[RGB+14]+
Ciona intestinalis B	*C. intestinalis A*	20	1883	59	490	[RGB+14]+
Clostridium difficile	*Anaerococcus prevotii*	11	192	49		[RdSB+18]*
Corvus cornix	*C. monedula*	38	7167395	25949	664205	[PVB+14]⋆
Coturnix japonica	*Gallus varius*	20	5061864	87069	220450	[WZH+18]⋆
Culex pipiens	*C. torrentium*	20	5442	106		[RGB+14]+
Drosophila melanogaster	*D. simulans*	196	4662706	151138		[LCC+15]
Egretta garzetta	*Pelecanus crispus*	10	9318499	361539	10242	[LLC+14]⋆
Emys orbicularis	*Trachemys scripta*	20	515	14		[RGB+14]+
Escherichia coli	*E. fergusonii*	62	84222	6903		RefSeq used in [LBL+16]
Ficedula albicollis	*F. hypoleuca*	24	14697230	269430	229260	[ESB+12, KHM+16, SMQE16, BNK+15]⋆
Gorilla gorilla	ancestral allele call from [PMSK+13]	54	9878547	42	569321	[PMSK+13]
Halictus scabiosae	*H simplex*	22	712	10		[RGB+14]+
Helicobacter pilori	*H. felis*	70	27498	8235		[RdSB+18]*
Homo sapiens (Yoruba population)	ancestral allele call from [Con15]	216	19441528	105146		[Con15]
Klebsiella pneumoniae	*K. varicola*	156	203601	375		[RdSB+18]*
Lepus granatensis	*L americanus*	20	769	31		[RGB+14]+
Melitaea cinxia	*M didyma*	18	1695	101		[RGB+14]+
Messor barbarus	*M structor*	20	9651	50		[RGB+14]+
Mycobacterium tubercolosis	MYCN001 - MYCN005	33	7142	13	78	RefSeq
Nipponia nippon	*Pelecanus crispus*	16	1140694	44153	2034	[LLC+14]⋆
Ostrea edulis	*O. chilensis*	20	939	28		[RGB+14]+
Pan paniscus	ancestral allele call from [PMSK+13]	26	6293657	63	284527	[PMSK+13]
Pan troglodytes ellioti	ancestral allele call from [PMSK+13]	20	10009190	44	459884	[PMSK+13]
Parus caeruleus	*P. major*	20	866	51	19	[RGB+14]+
Parus major	*Cyanistes caeruleus*	54	14174305	143760	126876	[QTH+15, LGS+16]⋆
Passer domesticus	*P. montanus*	16	18501992	90623	633399	[RTE+18, ETT+17, RET+18]⋆
Phylloscopus trochilus	*P. tristis*	24	33401127	8605	6092936	[LLL+17]⋆
Physa acuta	*P. gyrina*	18	4286	176		[RGB+14]+
Pseudomonas aeruginosa	*P. knackmussii*	86	90258	17208		[RdSB+18]*
Sepia officinalis	*Sepiella japonica*	18	1740	52		[RGB+14]+
Staphylococcus aureus	*S. epidermis*	152	30052	8694		[RdSB+18]*
Streptococcus pneumoniae	*S. mitis*	32	49917	2468		[RdSB+18]*
Taeniopygia guttata	*Poephila acuticauda*	38	53263038	118506	4346767	[SLS+15]⋆
Zea mays	*Tripsacum dactyloides*	66	520310	214398		[BMHR+17]

*Core genomes were computed and aligned as is [RdSB+18], with the difference that for every species we added to the set of a species genome one further genome from the closest species (to be able to orient the changes).

⋆ `https://github.com/harvardinformatics/shortRead_mapping_variantCalling`

+SFS were extracted from transcriptomes as described in [RGB+14].

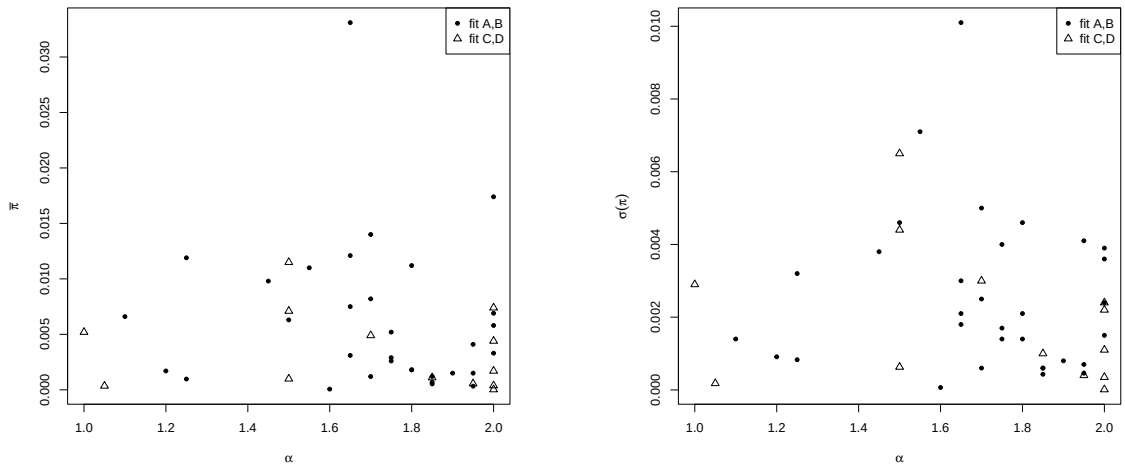


Figure A.21: Comparison of estimated α parameter $\hat{\alpha}$ (x-axis) and mean $\bar{\pi}$ (standard deviation $\sigma(\pi)$) of windowed nucleotide diversity π (y-axis). See Sect. A.10) for details.

Table A.6: Population structure inference and genetic diversity π (nucleotide diversity per site in 15 kb windows) for the fitted data sets. Model fitted, grade of fit, and biological order repeated from Table 2. Number of clusters: k inferred by BIC criterion for subsequent k-means clustering. PCA eye-test: Visual inspectation of PCA plot. Clear structure: "yes" if PCA eye-test=yes and DAPC cluster=1, "no" if PCA eye-test=no and DAPC cluster>1, otherwise "?".

Species (clade)	best model, grade	PCA eye-test	DAPC clusters	clear structure	mean π (s.d. π) $\times 10^{-3}$
Aptenodytes patagonicus (V)	Beta, B	unclear	5	?	1.0 (0.8)
Athene cunicularia (V)	Beta, B	yes	1	?	1.8 (1.4)
Corvus cornix (V)	Beta, A	no	1	no	1.5 (0.7)
Coturnix japonica (V)	Beta, A	no	1	no	9.8 (3.8)
Egretta garzetta (V)	Beta, B	no	4	?	2.6 (1.7)
Emys orbicularis (V)	KM, C	unclear	2	?	1.1 (1.0)
Ficedula albicollis (V)	Ψ, A	no	1	no	3.3 (1.5)
Gorilla gorilla (V)	Beta, B	no	1	no	1.5 (0.8)
Homo sapiens (Yoruba population) (V)	Beta, A	no	1	no	1.2 (0.6)
Lepus granatensis (V)	MMC/Ψ, C	yes	2	yes	1.0 (0.6)
Nipponia nippon (V)	KM, D	unclear	7	?	0.4 (0.3)
Pan paniscus (V)	Beta, B	no	1	no	0.7 (0.4)
Pan troglodytes ellioti (V)	Beta, A	unclear	1	?	1.2 (0.6)
Parus caeruleus (V)	MMC/Ψ, B	unclear	1	?	1.7 (0.9)
Parus maior (V)	Beta, A	no	1	no	2.9 (1.4)
Passer domesticus (V)	KM, A	no	1	no	5.8 (2.4)
Phylloscopus trochilus (V)	KM, A	no	1	no	6.9 (3.6)
Taeniopygia guttata (V)	Beta, A	no	1	no	5.2 (4.0)
Armadillidium vulgare (I)	Beta, C	no	1	no	4.9 (3.0)
Artemia franciscana (I)	Beta, B	no	1	no	3.1 (1.8)
Caenorhabditis brenneri (I)	Beta, C	no	1	no	7.1 (4.4)
Caenorhabditis elegans (I)	KM, D	no	10	?	0.02 (0.01)
Ciona intestinalis A (I)	KM, B	unclear	3	?	4.1 (4.1)
Ciona intestinalis B (I)	Beta, B	unclear	9	?	12.1 (10.1)
Culex pipiens (I)	Beta, B	unclear	2	?	11 (7.1)
Drosophila melanogaster (I)	Beta, A	yes	2	yes	7.5 (3)
Halictus scabiosae (I)	MMC/Ψ, B	unclear	2	?	0.5 (0.6)
Melitaea cinxia (I)	Beta, B	unclear	3	?	13.5 (5.3)
Messor barbarus (I)	KM, C	unclear	2	?	1.7 (1.1)
Ostrea edulis (I)	MMC/Ψ, B	yes	2	yes	1.8 (2.1)
Physa acuta (I)	Beta, B	no	1	no	6.3 (4.6)
Sepia officinalis (I)	KM, C	no	2	?	0.6 (0.4)
Zea mays (P)	Ψ, A	unclear	3	?	0.3 (0.5)
Arabidopsis thaliana (P)	Beta, A	yes	1	?	0.06 0.07
Acinetobacter baumannii (B)	Beta, B	yes	14	yes	11.2 (4.6)
Bacillus subtilis (B)	Ψ, B	unclear	5	?	11.9 (3.2)
Chlamydia trachomatis (B)	KM, D	yes	8	yes	4.4 (2.2)
Clostridium difficile (B)	KM, D	unclear	5	?	5.2 (2.9)
Escherichia coli (B)	KM, B	yes	17	yes	17.4 (3.9)
Helicobacter pilori (B)	Ψ, B	yes	2	yes	33.1 (2.1)
Klebsiella pneumoniae (B)	KM, D	unclear	9	?	7.4 (2.4)
Mycobacterium tubercolosis (B)	Beta, C	yes	12	yes	0.4 (0.2)
Pseudomonas aeruginosa (B)	Ψ, B	yes	9	yes	6.6 (1.4)
Staphylococcus aureus (B)	Ψ, B	yes	15	yes	8.2 (2.5)
Streptococcus pneumoniae (B)	Beta, C	yes	2	yes	11.5 (6.5)

x	$\rho(\hat{\alpha}, x)$ all data sets	$\rho(\alpha, x)$ data sets with fits A,B
$\bar{\pi}$	-.11	-.11
$\sigma(\pi)$	-.12	-.06

Table A.7: Correlation coefficient ρ of estimated α parameter $\hat{\alpha}$ and mean $\bar{\pi}$ (standard deviation $\sigma(\pi)$) of windowed nucleotide diversity π (Sect. A.10)