methods that feature predefined parameter grids of a size comparable to the tree size n (Gill et al. 2012).

Using either the N or γ parametrization creates issues even when optimal design is employed. Consider the N parametrization, which has $\det[\mathcal{I}(N)]=\prod_{j=1}^{p} m_j N_j^{-2}$. We let the constant $c=\prod_{j=1}^{p} N_j^{-2}$. D-optimality is the solution to $\max_{\{m_j\}} c\prod_{j=1}^{p} m_j$ subject to $\sum_{j=1}^{p} m_j = n-1$. Our objective function is therefore $g(\{m_j\})=\prod_{j=1}^{p} m_j$ which is known to be Schur concave when all $m_j > 0$. The optimal design is uniform and given by the first equality in Eq. (6) below.

$$m_j^*|D=\frac{1}{p}(n-1), \qquad m_j^*|E=\frac{N_j^2}{\sum_{i=1}^{p} N_i^2}(n-1) \tag{6}$$

The E-optimal design solves: $\max_{\{m_j\}} \min_j m_j N_j^{-2}$. The objective function is now $g(\{m_j\})=\min(m_1 N_1^{-2},\ldots, m_p N_p^{-2})$ and is also Schur concave. The E-optimal solution satisfies $m_1^* N_1^{-2}=m_2^* N_2^{-2}=\ldots=m_p^* N_p^{-2}$ (Marshall et al. 2011), and is the second equality in Eq. (6). This optimal design assigns more coalescent events to periods with larger population size, with a square penalty. The equivalent D and E-designs for inverse population size follow by simply replacing N_j with γ_j in Eq. (6) above.

Thus, in theory, D-optimal designs that consider N or γ could result in some parameters being very poorly estimated while E-optimal ones could allocate all of the coalescent events to a single parameter, increasing the possibility of nonidentifiability. Additionally, for a given criterion, optimal N_j and γ_j designs can be contradictory. A robust design that is insensitive to both the parameter values and the choice of optimality criteria is therefore needed. This point is illustrated in the top panel of Figure 2, which presents D- and E-optimal confidence ellipsoids under the N parametrization, for the model shown in Figure 1. These ellipsoids, for some parameter vector σ, with diagonal Fisher information matrix $\mathcal{I}(\sigma)$, are given by $\sum_{j=1}^{p}\left(x_j-\sigma_j\right)^2\mathcal{I}(\sigma)_{(j,j)}=\Omega$. Here, Ω controls the confidence significance level according to a χ^2 distribution (with p degrees of freedom), and x_j is some coordinate on the jth parameter axis (Friendly et al. 2013). Under the N parametrization, D- and E-optimal designs are notably different, and sensitive to the true values of N_1 and N_2.

Robust Coalescent Design

We define a robust experimental design as being 1) insensitive to the true (unknown) parameter values and 2) minimizing both the maximum and total uncertainty over the estimated parameters. The latter condition means that a robust design is also insensitive to choice of optimality criteria. We formulate our main results as the following two-point theorem.

Theorem 1. *If the p-parameter vector σ admits a diagonal Fisher information matrix, $\mathcal{I}(\sigma)=[m_1\sigma_1^{-2},\ldots,m_p\sigma_p^{-2}]I_p$, under an isoperimetric constraint $\sum_{j=1}^{p} m_j=\kappa$, then any design that 1) works in the parametrization $[\log\sigma_1,\ldots,\log\sigma_p]$ and 2) achieves the distribution $m_1^*=\cdots=m_p^*=\frac{1}{p}\kappa$ over this $\log\sigma$ space, is provably and uniquely robust.*

Theorem 1 guarantees that inference is consistent and reliable across parameter space. We derive point 1), by maximizing how distinguishable our parameters are within their space of possible values. "Distinguishability" is a property that determines parameter identifiability and model complexity (Grunwald 2007). Assume that ψ is the true parameter vector underlying some observed tree $\mathcal{T}$, and that ψ lies in some parameter space, Ψ, of a piecewise coalescent model. Let $h(\psi)=\sigma$ define a parameter transformation, and let $\mathcal{T}$ have a total of $n-1$ coalescent events. Generally, we will be able to infer ψ from $\mathcal{T}$ with some statistical confidence. This confidence can be visualized as an ellipse around ψ. All parameter vectors that map to this ellipse are statistically indistinguishable from ψ. If we repeat this inference problem across the parameter space, we generate a lattice of ellipses (Myung et al. 2000). These ellipses define the distinguishable parameter vector subsets in Ψ and will shrink in size but increase in number as n increases estimate certainty improves with data. Thus, distinguishability is intrinsically linked to the quality of inference. More detail on these information geometric concepts is given in Myung et al. (2000) and Grunwald (2007).

We can define a volume, $\mathcal{V}:=\int_{\Psi}\det\left[\frac{1}{n-1}\mathcal{I}(\psi)\right]^{\frac{1}{2}}\mathrm{d}\psi$ to measure the total size of these ellipses over Ψ (Grunwald 2007). This volume is related to the complexity of our coalescent model and hence is unchanged by parametrization (Grunwald 2007). While $\mathcal{V}$ is invariant to parametrization choice h, different h functions control how the parameter space is discretized into distinguishable ellipses (Myung et al. 2000). For example, under $\psi=\sigma$ poor distinguishability results when any σ_j becomes large (i.e., ellipses expand as parameters take bigger values). We therefore pose the problem of finding an optimal bijective parameter transformation $h(\psi_j)=\sigma_j$, which maximizes overall parameter distinguishability in Ψ, or equivalently minimizes the sensitivity or our estimates to the unknown true values of our parameters, ψ. Geometrically, this transformation yields the smallest ellipse size that is also independent of the location of ψ in Ψ.

Applying Eq. (1), with $h':=\frac{\partial h}{\partial\psi_j}$, we get that $\mathcal{I}(\psi)_{(j,j)}=m_j h^{-2}\left(h'\right)^2$. The orthogonality of the diagonal Fisher information matrix means that ψ_j only depends on σ_j. Using the properties of determinants, we can decompose

the volume as $\mathcal{V}=\prod_{j=1}^{p}\frac{m_j}{n-1}\mathcal{V}_j$. Since $\mathcal{V}$ is constant for any parametrization, our parameters are orthogonal, and our transformation is bijective, then $\mathcal{V}_j$ is also constant. If $\sigma_j\in[\sigma_{j(1)},\sigma_{j(2)}]$, then $h(\psi_{j(1)})=\sigma_{j(1)}$ and $h(\psi_{j(2)})=\sigma_{j(2)}$. Using these endpoints and the invariance of $\mathcal{V}$, we obtain Eq. (7).

$$\mathcal{V}_j=\int_{\psi_{j(1)}}^{\psi_{j(2)}}h^{-1}h'\,\mathrm{d}\psi_j=\int_{\sigma_{j(1)}}^{\sigma_{j(2)}}\sigma_j^{-1}\,\mathrm{d}\sigma_j \tag{7}$$

This equality defines the conserved property across parametrizations of coalescent models with likelihoods given by Eq. (4). We can maximize both the insensitivity of our parametrization, h, to the unknown true parameters and our ability to distinguish between distributions across parameter space by forcing $h^{-1}h'$ to be constant irrespective of ψ_j. This is equivalent to solving a minimax problem. We choose a unit constant and evaluate Eq. (7) to obtain: $\psi_{j(2)}-\psi_{j(1)}=\log\sigma_{j(2)}-\log\sigma_{j(1)}$. Due to the bijective nature of h, this implies that our (unique) optimal parametrization is $\psi_j=\log\sigma_j$ and hence proves 1).

Point 2) follows by solving optimal design problems under the $\log\sigma$ parametrization. For consistency with Eq. (6), we set $\sigma=N$. This gives $\frac{\partial N_j}{\partial\psi_j}=e^{\psi_j}$ and results in the Fisher information matrix, $\mathcal{I}(\log N)$, in Eq. (8).

$$\mathcal{I}(\log N)=[m_1,\ldots,m_p]\mathrm{I}_p\Longrightarrow m_j^*|\mathbb{D}=\frac{1}{p}(n-1) \tag{8}$$

Let $\mathbb{D}$ be an optimal design criterion, with event distribution given by $\{m_j^*|\mathbb{D}\}$. When $\mathbb{D}\equiv$ D, we maximize $\det\left[\mathcal{I}(\log N)\right]$ to obtain the uniform coalescent distribution in Eq. (8). The D-optimal design for N, N^{-1}, and $\log N$ are therefore the same. However, we see interesting behavior under other design criteria. When $\mathbb{D}\equiv$ E, we maximize $\mathrm{mineig}\left[\mathcal{I}(\log N)\right]$ to again obtain Eq. (8). This is very different from analogous designs under N and N^{-1}. While we do not assess further optimal design criteria here, several others also yield the design of Eq. (8). Thus, under a log-parametrization optimal experimental designs converge, and parameter confidence ellipsoids are, consequently, invariant to optimality criteria. This effect is shown in the bottom panel of Figure 2 for a $p=2$ skyline model. This desirable design insensitivity emerges because $\mathcal{I}(\log N)$ is independent of N for piecewise coalescent models, and proves 2). We next apply Theorem 1 to three distinct and widely used coalescent models, to derive specific insights and recommendations.

Skyline Demographic Models

Consider a coalescent process with deterministically time-varying population size, $N(t)$, for $t\geq 0$ that features sequences sampled at different times. As with the popular "skyline" family of inference methods (Pybus et al. 2000; Strimmer and Pybus 2001; Drummond et al. 2005; Minin et al. 2008), we assume that $N(t)$ can be described by a piecewise-constant function with $p\geq 1$ values so that $N(t):=\sum_{j=1}^{p}N_j\mathbb{1}(\epsilon_{j-1}\leq t<\epsilon_j)$ with $\epsilon_0=0$ and $\epsilon_p=\infty$. N_j is the constant population size of the jth segment (or interval) which is delimited by times $[\epsilon_{j-1},\epsilon_j)$. The indicator function $\mathbb{1}(a)=1$ when a is true, and is 0 otherwise. We start by assuming that this process has generated an observable coalescent tree, $\mathcal{T}$, with $n\geq n_s+1$ tips, with $n_s\geq 1$ as the number of distinct sampling times. Each tree tip is a sample and the tuple (s_k,ϕ_k) defines a sampling protocol in which ϕ_k tips are introduced at time s_k with $1\leq k\leq n_s$ and $\sum_{k=1}^{n_s}\phi_k=n$. Since trees always start from the present then $s_1=0$ and $\phi_1\geq 2$. Figure 1 explains this notation for a $p=2$ skyline demographic model.

In keeping with the literature, we assume that sampling times are independent of $N(t)$ (Drummond et al. 2005). The choice of sampling times and the number of sequences obtained at each sampling time (i.e., the sampling protocol) is what the experimenter controls. The observed n tip tree generated under this process has $n-1$ coalescent events. We use c_i to denote the time of the ith such event with $1\leq i\leq n-1$. We define $l(t)$ as a piecewise-constant function that counts the number of lineages in $\mathcal{T}$ at t and let $\alpha(t):=\binom{l(t)}{2}$. At the kth sample time $l(t)$ increases by ϕ_k, and at every c_i it decreases by 1. The rate of producing coalescent events is defined as: $\lambda(t)=\sum_{j=1}^{p}\gamma_j\alpha(t)\mathbb{1}(\epsilon_{j-1}\leq t<\epsilon_j)$ with $\gamma_j=N_j^{-1}$ as the inverse population size in segment j. We initially work in $\gamma=[\gamma_1,\ldots,\gamma_p]$, and then transform to $N=[N_1,\ldots,N_p]$.

The log-likelihood follows from Poisson process theory as (Snyder and Miller 1991; Parag and Pybus 2018): $L(\gamma)=-\int_0^{c_{n-1}}\lambda(t)\mathrm{d}t+\sum_{i=1}^{n-1}\log\lambda(c_i)$, with $L(\gamma)=\log\mathbb{P}(\mathcal{T}|\gamma)$. By splitting the integral across the p piecewise-constant segments, we get that: $\int_0^{c_{n-1}}\lambda(t)\mathrm{d}t=\sum_{j=1}^{p}\gamma_j\int_{\epsilon_{j-1}}^{\epsilon_j}\alpha(t)\mathrm{d}t=\sum_{j=1}^{p}\gamma_j\omega_j$. Here ω_j is a constant for a given tree, and it is independent of γ. Similarly, $\sum_{i=1}^{n-1}\log\lambda(c_i)=\sum_{j=1}^{p}\sum_{i=1}^{n-1}\log(\gamma_j\alpha(c_i)\mathbb{1}(\epsilon_{j-1}\leq c_i\leq\epsilon_j))$. Expanding yields Eq. (9) with Γ_j as a constant depending on $\alpha(c_i)$ for all i such that $c_i\in[\epsilon_{j-1},\epsilon_j)$. Here m_j counts all the coalescent events within this interval.

$$L(\gamma)=\sum_{j=1}^{p}m_j\log\gamma_j-\gamma_j\omega_j+\log\Gamma_j \tag{9}$$

Equation 9 is an alternate expression of the skyline log-likelihood given in Drummond et al. (2005), except that $N(t)$ is not constrained to change only at coalescent event times. Importantly, sampling events do not contribute to the log-likelihood (Drummond et al. 2005). As a result, we can focus on defining a desired coalescent

distribution across the population size intervals, $\{m_j^*\}$. An optimal sampling protocol would then aim to achieve this benchmark distribution.

Since Eq. (9) is equivalent to Eq. (4), Theorem 1 applies. The relevant robust design is given by Eq. (8), and recommends inferring $\log N$ and sampling sequences in such a way that $\frac{n-1}{p}$ coalescent events fall in each $[\epsilon_{j-1}, \epsilon_j)$ segment. Note that the number of lineages, $l(t)$, the timing of the m_j events within $[\epsilon_{j-1}, \epsilon_j)$, and the wait between the last of these and ϵ_j are all uninformative about population size. As an illustrative example, we solve a simple skyline model design problem in the Appendix. There we apply Theorem 1 to a square wave approximation of a cyclic population size function and find practical sampling protocols that achieve robust $\{m_j^*\}$ designs.

Lastly, we comment on the impact of priors. Some inference methods, such as the Skyride (Minin et al. 2008) and Skygrid (Gill et al. 2012), use smoothing priors that ease the sharpness of the inferred piecewise-constant population profile. While these priors embed extra (implicit) information about population size, they do not alter the optimal design point, even for small n. This follows because the informativeness of a prior is unaffected by $\{m_j\}$ choices. The robust design therefore proceeds as above, independent of any contributions from the smoothing prior.

Structured Coalescent Models

Let $\mathcal{T}$ be an observed structured coalescent tree with $p \geq 1$ demes that have been sampled through time (branches are labelled according to the deme in which they exist). Our experimental variables are the placement (both in time and in deme location) of the samples, and our goal is to define robust design objectives for the inference of population size, and migration rate parameters. We set T as the number of intervals in $\mathcal{T}$, with each interval delimited by a pair of events, which can be sampling, migration or coalescent events. The ith interval has length u_i and $\sum_{i=1}^{T} u_i$ gives the time to the most recent common ancestor of $\mathcal{T}$. We use l_{ji} to count the number of lineages in deme j during interval i. Lineage counts increase on sampling or immigration events, and decrement at coalescent or emigration events. We define the migration rate from deme j into k as ζ_{jk}. We use N_j and $\gamma_j = N_j^{-1}$ for the absolute and inverse population size in deme j.

Our initial p^2 vector of parameters is $\sigma = [\gamma_1, \ldots, \gamma_p, \{\zeta_{1\bar{1}}\}, \ldots, \{\zeta_{p\bar{p}}\}] = [\gamma, \zeta]$, with $\{\zeta_{k\bar{k}}\} = [\zeta_{k1}, \zeta_{k2}, \ldots]$ as the $p-1$ subvector of all the migration rates from deme k. The log-likelihood $L(\sigma) = \log \mathbb{P}(\mathcal{T} | \gamma, \zeta)$ is then adapted from Beerli and Felsenstein (1999) and Ewing et al. (2004). We decompose $L(\sigma) = \sum_{j=1}^{p} L_j(\gamma) + L_j(\zeta)$ into coalescent and migration sums with jth deme components given in Eq. (10) and Eq. (11). Here, m_j and w_{jk}, respectively count the total number of coalescent events in subpopulation j and the sum of migrations from that deme into deme k, across all T time intervals. The factor $\alpha_{ji} := \binom{l_{ji}}{2}$ accounts for the contribution of the number of lineages to the coalescent rates. We constrain our tree to have a total of $n-1$ coalescent events so that $\sum_{j=1}^{p} m_j = n-1$.

$$L_j(\gamma) = m_j \log \gamma_j - \sum_{i=1}^{T} u_i \alpha_{ji} \gamma_j \tag{10}$$

$$L_j(\zeta) = \sum_{k=1, k\neq j}^{p} w_{jk} \log \zeta_{jk} - \sum_{i=1}^{T} u_i l_{ji} \zeta_{jk} \tag{11}$$

The log-likelihoods of both Eq. (10) and Eq. (11) are generalizations of Eq. (4) and lead to diagonal (orthogonal) Fisher information matrices like Eq. (5). This orthogonality results because migration events do not inform on population size and coalescent events tell us nothing about migrations. While migrations do change the number of lineages in a deme that can then coalesce, the lineage count component of the coalescent rate, α_{ji}, does not influence the Fisher information. Importantly, since the Fisher information is independent of the sample times and locations, we can freely modify our sampling protocols to potentially achieve optimal design objectives.

Theorem 1 implies that we should infer log population sizes and log migration rates from structured models. This ensures estimate precision is independent of the unknown population sizes and migration rates, and gives $\mathcal{I}(\psi) = [m_1, \ldots m_p, \{w_{1\bar{1}}\}, \ldots \{w_{p\bar{p}}\}] \mathrm{I}_{p^2}$ when $\psi = [\log N_1, \ldots \log N_p, \{\log \zeta_{1\bar{1}}\}, \ldots \{\log \zeta_{p\bar{p}}\}]$. The robust design under this ψ, given in Eq. (12), involves distributing coalescent and migration events uniformly among the demes. Note that the migration rate distribution, $w_{jk}^* | \mathbb{D}$, only holds if the total number of migration events are fixed, that is $\sum_{j=1}^{p} \sum_{i=1, i\neq j}^{p} w_{ji} = M$, for some constant M.

$$m_j^* | \mathbb{D} = \frac{1}{p}(n-1), \qquad w_{jk}^* | \mathbb{D} = \frac{1}{p(p-1)} M \tag{12}$$

Two points are clear from Eq. (12). First, if all the migration rates are known, so that only population sizes are to be estimated then the structured model yields exactly the same robustness results as the skyline demographic model. Second, the migration rate design is the same at both the strong and weak migration limits of the structured model (Nordborg 2001). Thus, the true (unknown) migration rates do not affect their optimal design, provided log-migration rates are inferred. If we generalize the population size function in each deme to be piecewise-constant in time, then we obtain a combination of the structured and skyline model design results. The robust design in this case maintains the log-population and log-migration recommendations, but now requires that coalescent events are divided equally

among both the demes and the piecewise-constant population segments.

Sequentially Markovian Coalescent Models

We now focus on coalescent models where recombination is applied along a genome, resulting in many correlated, hidden trees (multiple loci) (Li and Durbin 2011). This is in contrast to the skyline demographic and structured coalescent models where the coalescent trees are observable (and hence inference is more direct). Each SMC tree typically consists of a small number of lineages. Popular inference methods in this field are based on an approximation to the coalescent with recombination called the SMC (McVean and Cardin 2005). These methods typically handle SMC inference by constructing a hidden Markov model (HMM) over discretized coalescent time. If we partition time into p segments: $0=\epsilon_0<\epsilon_1<\cdots<\epsilon_p=\infty$ then, when the HMM is in state j, the coalescent time is in $[\epsilon_{j-1},\epsilon_j)$. Recombination events lead to state changes in the HMM, and the genomic sequence serves as the observed process of the HMM. Expectation–maximization type algorithms are used to iteratively infer the HMM states from the genome, as in (Li and Durbin, 2011), (Schiffels and Durbin, 2014), (Sheehan et al., 2013), (Tataru et al., 2014), and (Steinrucken et al., 2015).

A central aspect of all these techniques is the assumption that during each coalescent interval the population size is constant. If the vector $N=[N_1,\ldots,N_p]$ denotes population size, then it is common to assign N_j for the $[\epsilon_{j-1},\epsilon_j)$ interval. This not only allows an easy transformation from the inferred HMM state sequence to estimates of N (Gattepaille et al. 2016) but also controls the precision of SMC based inference. For example, if too few coalescent events fall within $[\epsilon_{j-1},\epsilon_j)$, then N_j will generally be overestimated (Sheehan et al. 2013). Thus, the choice of discretization times (and hence population size change-points) is critical to SMC (and coalescent HMM) inference performance (Tataru et al. 2014; Palacios et al. 2015; Spence et al. 2018).

Our experimental design problem involves finding an optimal criterion for choosing these discretization times. Currently, only heuristic strategies exist (e.g., choosing bins so that coalescent events are distributed evenly under a constant population size assumption, or in accordance with observed single nucleotide polymorphism spacings) (Sheehan et al. 2013; Palacios et al. 2015; Gattepaille et al. 2016). We define a vector of bins $\beta=[\beta_1,\ldots,\beta_p]$ such that $\beta_j=\epsilon_j-\epsilon_{j-1}$ and assume we have T loci (and hence coalescent trees). In keeping with Li and Durbin (2011) and Schiffels and Durbin (2014), we assume that each tree only leads to a single coalescent event. This coalescent event could correspond to different genealogical scenarios, depending on the application (e.g., to the only coalescence in a tree with two tips, or to the first event in a multilineage tree). However, we can neglect lineage counts, and hence tree topology here without loss of generality. This follows because lineage counts merely rescale time (piecewise) linearly and, more importantly, they do not contribute to the Fisher information in piecewise coalescent models (see Eq. (4)).

Let m_{ij} be the number of exponentially distributed coalescent events falling within bin β_j from the ith locus so that $\sum_{j=1}^{p} m_{ij}=1$. We further use $m_j:=\sum_{i=1}^{T} m_{ij}$ to count the total number of events from all loci falling in β_j. As before, we constrain the total number of coalescent events so that $\sum_{j=1}^{p} m_j=n-1$, and condition on the genealogies. Since each tree contributes a single coalescent event, as in Li and Durbin (2011), then $T=n-1$. Using Poisson process theory, we can write the log-likelihood of obtaining a set of coalescent event counts $\{m_{ij}\}$, within our bins $\{\beta_j\}$ for the ith locus as $L_i(\gamma,\beta)=\log\mathbb{P}(\mathcal{T}_i|\gamma,\beta)=-\int_0^\infty\lambda(t)\mathrm{d}t+\sum_{j=1}^{p} m_{ij}\log\left(\int_{\epsilon_{j-1}}^{\epsilon_j}\lambda(t)\mathrm{d}t\right)$ (Snyder and Miller 1991). Here $\lambda(t)$ is the coalescent rate at t so that $\lambda(t)=\sum_{j=1}^{p}\gamma_j 1(\epsilon_{j-1}\le t<\epsilon_j)$ and $\int_{\epsilon_{j-1}}^{\epsilon_j}\lambda(t)\mathrm{d}t=\beta_j\gamma_j$ with $\gamma_j=N_j^{-1}$. Solving this yields Eq. (13), which is analogous to Eq. (4) and has Fisher information matrix $\mathcal{I}(N)_{\mathcal{T}_i}=[m_{i1},\ldots,m_{ip}]\mathrm{I}_p$. Note that neither the waiting time until a recombination nor the time between recombination and coalescence contribute to the Fisher information (exponential memoryless property).

$$L_i(\gamma,\beta)=\sum_{j=1}^{p}-\gamma_j\beta_j+m_{ij}\log\gamma_j\beta_j \tag{13}$$

$$\mathcal{I}(N)_{\{\mathcal{T}_i\}}=\sum_{i=1}^{T}\mathcal{I}(N)_{\mathcal{T}_i|\mathcal{T}_{i-1}}=[m_1N_1^{-2},\ldots,m_pN_p^{-2}]\mathrm{I}_p \tag{14}$$

Equation 13 is an alternative form of the log-likelihood given in Weissman and Hallatschek (2017), and describes a binned coalescent process that is equivalent to the discrete one presented in Tataru et al. (2014). Interestingly, Eq. (13) is a function of the product $N_j^{-1}\beta_j$ so that we cannot identify both the bins and the population size without extra information. This explains why choosing a time discretization has been found to be as difficult as estimating population sizes (Gattepaille et al. 2016).

The total Fisher information about population size from all T trees follows from the chain rule $\mathcal{I}(N)_{\{\mathcal{T}_i\}}=\sum_{i=1}^{T}\mathcal{I}(N)_{\mathcal{T}_i|\mathcal{T}_{i-1},\ldots\mathcal{T}_1}$ (Zegers 2015). Using the Markov dependence between loci gives the first equality in Eq. (14). Here $\mathcal{T}_{i-1}$ and $\mathcal{T}_i$ are separated by a single recombination and $\mathcal{I}(N)_{\mathcal{T}_i|\mathcal{T}_{i-1}}$ is the additional Fisher information about N contained in $\mathcal{T}_i$ given $\mathcal{T}_{i-1}$ (Zamir 1998; Zegers 2015). If $\mathcal{T}_i$ contributes a new coalescent event falling within the period with population size N_j, then only the jth element of $\mathcal{I}(N)_{\mathcal{T}_i|\mathcal{T}_{i-1}}$ is nonzero, and can be computed by taking derivatives of Eq. (13).

This means that only the additional coalescent in $\mathcal{T}_i$ is important (Nordborg 2001), and $\mathcal{T}_i|\mathcal{T}_{i-1}$ is a sufficient statistic for N_j in this case (Zamir 1998). Repeating this process across all T trees and p population sizes gives the second equality in Eq. (14), for conditioned bin counts $\{m_j\}$.

If we calculate the total Fisher information with respect to β, we obtain identical expressions to Eq. (14) with the N_j simply replaced by β_j. The square dependence of these Fisher matrices means that Theorem 1 applies. We therefore find that it is optimal to infer log-bin sizes ($\psi = [\log\beta_1, \ldots, \log\beta_p]$), if population size history is known [this corresponds to the discretization results presented in (Tataru et al., 2014)], or log-population sizes ($\psi = [\log N_1, \ldots, \log N_p]$), if the bins are known. We generally assume the latter since bin end-points can often be set by the user (Palacios et al. 2015). Under either parametrization, the provably robust design recommendation is to discretize time such that the resulting bins contain equal numbers of coalescent events. Note that if we did not condition on the genealogies then the Fisher information would be different, and the robust transform would involve a square root. Robust design would, however, still involve equalising the coalescent event distribution under this new transformation.

Our results also hold for several SMC-based methods and related modifications. While the above argument assumes a different population size in each bin, some methods group bins across a common population size (Li and Durbin 2011). Although this grouping slightly changes Eq. (13), our analysis remains intact since each new tree still only contributes one coalescent event worth of information. Other methods, such as the stairway plot of Liu and Fu (2015), which combines skyline methodology with mutational site-frequency spectra, treat the T loci as independent or unlinked. In these cases, the proof is simpler as the combined log-likelihood is $L(\gamma, \beta) = \sum_{i=1}^{T} L_i(\gamma, \beta)$. This is equivalent to Eq. (4) and so Theorem 1 is valid. Our robust design principles are therefore relevant to a wide range of genomic coalescent models. This broad applicability stems from the fact that recombination events provide no information about population size (Palacios et al. 2015).

Discussion

Judicious experimental design can improve the ability of any inference method to extract useful information from observed data (Liepe et al. 2013). Despite these potential advantages, experimental design has received little attention in the coalescent inference literature (Hall et al. 2016). We therefore defined and investigated robust designs for three important, and popular coalescent models. While these models are different in composition and application, we can unite them under the key observation that longitudinal samples (through time), migration events, and recombination events all introduce additional lineages to a genealogy, in a statistically similar manner. Theorem 1, which summarises our main results, presents a clear and simple two-point robust design benchmark for the more general class of piecewise coalescent models (i.e., those with Eq. (4) type likelihoods), to which these three belong.

The first point of Theorem 1 recommends inferring the logarithm (and not the absolute value or inverse) of our parameters of interest. As this is usually effective population size, N, then $\log N$ is the uniquely robust parametrization for piecewise coalescent estimation problems. While methods using $\log N$ do exist (Minin et al. 2008; Palacios et al. 2015), the stated reasons for doing so are centered around algorithmic convenience. Here we provide sound theoretical backing for using $\log N$ in coalescent inference. The second point of Theorem 1 requires equalizing the number of coalescent events informing about each parameter. This may initially appear obvious, as apportioning data evenly among the unknowns seems wise. Indeed, the works of Sheehan et al. (2013) and Tataru et al. (2014), which focus on SMC models, state that time discretizations should aim to achieve uniform coalescent distributions. However, no proof for this statement is given. Here, we not only provide theoretical support for uniform coalescent distributions but also prove that they are only robust if the log-parameter stipulation is jointly satisfied.

Several unifying insights for piecewise coalescent models emerge as corollaries of our analysis. Because the precision with which we estimate a coalescent parameter only depends on the number of events informing about it, we can reinterpret all the designs considered here simply as different ways of allocating events to "pigeon-holes." In our three examples, these pigeon-holes respectively correspond to skyline intervals, structured coalescent demes, and SMC time-discretization bins. This perspective reveals a straightforward rule for statistical identifiability: any piecewise coalescent model with at least one empty pigeon-hole is nonidentifiable (Rothenburg 1971). This has specific ramifications. For example, it implies that we need at least one coalescent and migration event in each deme of the structured coalescent model to guarantee identifiability. This result could have interesting links to previous identifiability analyses, which considered more stringent requirements (Bhaskar and Song 2014; Kim et al. 2015).

Knowing the boundaries or change-points of our pigeon-holes (e.g., the $\{\epsilon_j\}$ for the SMC) is crucial for inference (Tataru et al. 2014). Throughout, we have assumed that these are indeed known. This is reasonable as it is often not possible to jointly infer parameters and their change-points (Sheehan et al. 2013; Tataru et al. 2014). Methods that do achieve this are usually data driven, iterative, and case specific, allowing no general design insight (Opgen-Rhein et al. 2005; Palacios et al. 2015). This

raises the question about how to derive optimal design objectives when the change-points are unknown. In the Appendix, we use Theorem 1 to derive robust change-point objectives. Intriguingly, we show that it is wise to assign change-points according to the $\frac{1}{p}$ quantiles of the normalized lineages through time plot of the observed phylogeny. This results in a maximum spacings estimator (MSE) that makes the observed tree as uniformly informative as possible, relative to the pigeon-holes (Ranneby 1984). This means that if we wish to robustly infer p log-parameters from a tree containing $n-1$ coalescent events, we should define our pigeon-holes such that they change every $r=\frac{n-1}{p}$ events. If $r=1$, we find that the classic skyline plot (Pybus et al. 2000) is the low information limit of this strategy. Our MSE design recommendation is simple, practical and guarantees robustness and identifiability.

Realization of this procedure in existing software, such as BEAST 1 and 2, would be straightforward, since pigeon-holes are already implicitly set within the implemented Bayesian skyline plot and Skygrid methods, albeit using different rules. These rules either group adjacent pigeon-holes based on an Akaike criterion to reduce noise (Strimmer and Pybus 2001; Drummond et al. 2005), or define a fixed change-point time grid for ease of use (Gill et al. 2012). Our results suggest change-points should instead be based on coalescent event counts. This guideline for grouping skyline intervals also applies to aggregating demes in structured models, or combining bins in the SMC. This MSE strategy is directly useful for epidemiological and macroevolutionary applications of skyline demographic and structured models, in which coalescent times are observable, either through a fixed time-scaled phylogeny or from a posterior set of trees that have been inferred from sampled sequences using Bayesian approaches (Drummond et al. 2005; Parag and Pybus 2017).

However, the utility of this strategy is more limited in SMC models, in which coalescent events are hidden, and depend on the unknown population size. SMC inference requires iterative co-estimation of the coalescent times and population sizes, under a preassigned time discretization (Li and Durbin 2011). This precludes direct application of the MSE strategy to optimal bin time allocation. However, Theorem 1 is still helpful. First, its two-point criterion is independent of the form of the piecewise population size function, implying that globally optimal discretizations do not exist when coalescent times are unknown.

This reveals an important constraint to SMC inference, and hints that we might achieve robustness if we could access the inferred coalescent events in each iteration and then dynamically adjust the discretization via the MSE approach. Recent methods, which decouple discretization from the demographic history, could eventually allow this flexibility (Steinrucken et al. 2015). Second, while preassigned discretizations cannot be optimal for all demographic functions, the MSE strategy validates some existing design choices. Under a constant population size null model, the MSE requires log-bin sizes and quantiles from an exponential distribution. This supports the recommendations in Li and Durbin (2011) and Schiffels and Durbin (2014). These points could be particularly useful, given that SMC models are still not well theoretically understood (Spence et al. 2018).

Another unifying insight from Theorem 1 is that any parameter entering the coalescent log-likelihood in a functionally equivalent way to γ_j in Eq. (4), should be inferred in log-space. This maximizes distinguishability in model space, and means, for example, that it is best to work with log-migration rates for structured coalescent models. Using the log of the migration matrix is uncommon and could potentially improve current structured coalescent inference algorithms. Similarly, for the SMC, this insight implies that we must decide between absolute bin sizes for inferring log-populations and absolute population sizes for estimating log-bin widths, under fixed genealogies.

Theorem 1 is also useful for finding cases where non-robust designs are inevitable. In the skyline demographic model, for example, a short interval during which population size is large would be difficult to estimate. Large N implies long coalescent times, making it unlikely that $\frac{n-1}{p}$ events can be forced to occur in such intervals (see the simulation study in the Appendix). Equally, because coalescent events tend to congregate in periods of low N, bottlenecks also disrupt robust designs, making it difficult to estimate periods of population recovery unless samples are available both before and after the bottleneck. These points provide theoretical insight into some known issues in coalescent estimation, and are corroborated by Gattepaille et al. (2016) and Palacios et al. (2015).

Similar effects occur for SMC models if the bin size is small during a period of large population size or if unknown bottlenecks are present (Sheehan et al. 2013; Palacios et al. 2015). These observations also explain why SMC inference is often underpowered in the distant past (few events fall in these periods) (Li and Durbin 2011), and why sampling more genomes can be beneficial (it allows for better coalescent coverage) (Beichman et al. 2018). The loss in performance, due to a scarcity of informative events, reflects a fundamental limit on coalescent inference, and is also an issue for related approaches such as the stairway plot (Liu and Fu 2015) and Popsicle (Gattepaille et al. 2016). For the structured coalescent model, the population size criteria are likely simpler to achieve than the migration rate ones, since controlling the distribution of $p-1$ stochastic migration event types for every deme could be challenging, and dependent on how close we are to the strong or weak migration limits (Sjodin et al. 2005; Heller et al. 2013).

While we have provided universal, robust coalescent design objectives here, we have not explored what

specific sampling or discretization protocols can be used to achieve them (e.g., whether proportional or uniform sampling is better). Existing analyses on this topic (Stack et al. 2010; Heller et al. 2013; Palacios et al. 2015; Karcher et al. 2016) tend to examine a set of reasonable but ad hoc protocols via extensive simulation. However, since no optimal design references exist, these works could only compare performance among their chosen protocols. We hope our approach provides a general robust design theorem that can be used by future studies for benchmarking and validation.

Supplementary Material

Data available from the Dryad Digital Repository: http://datadryad.org, doi: https://doi.org/10.5061/dryad.n7rm21c.

Funding

This work was supported by the European Research Council under the European Commission Seventh Framework Programme (FP7/2007-2013)/ European Research Council grant agreement (614725-PATHPHYLODYN).

Acknowledgements

The authors thank Chieh-Hsi Wu and Louis Du Plessis for their incisive comments. The authors also acknowledge three anonymous reviewers and the editors for their insightful feedback.

Appendix

Robust Coalescent Change-point Designs

Consider the class of "piecewise" coalescent models, which we define as having log-likelihoods analogous to Eq. (4) in the main text. This class includes the skyline demographic model, structured coalescent model, and the SMC. We derived a robust design theorem (Theorem 1 of the main text) for inferring the parameters (e.g., effective population size) of these models. Theorem 1 suggested that experimental designs under piecewise coalescent models could be viewed as allocations of informative events (e.g., coalescent events) to "pigeon-holes," which essentially encapsulate the different parameters that we wish to infer. These pigeon-holes, for example, are the piecewise-constant population size segments in the skyline demographic model, the demes of the structured coalescent model, and the bins in the SMC. The boundaries or change-points of these pigeon-holes effectively control the complexity of our coalescent inference problem.

The analysis behind Theorem 1 presumed that we had knowledge of the pigeon-hole change-points. This corresponds to knowing the piecewise-constant segment times of the skyline model, the number of demes in the structured coalescent, and the bin sizes in the SMC. Such assumptions are reasonable, since simultaneously inferring both change-points and parameter values is an ill-conditioned problem. For example, if we do not know anything a priori about either bin or population size, then it is impossible to derive optimal SMC time discretizations (Sheehan et al. 2013; Tataru et al. 2014). Similar identifiability problems emerge when trying to simultaneously infer the change-points of piecewise-constant segments, and their population sizes, or the number of demes, and the population sizes and migration rates within each deme. In such cases iterative and data-driven computational methods can be employed (Opgen-Rhein et al. 2005; Palacios et al. 2015). These methods will typically jointly optimize over these unknowns and produce sensible estimates, but their results will be case specific, allowing no general design insight to be derived.

While the general change-point inference problem is outside the scope of our work, we can provide some practical guidelines on how to robustly specify pigeon-hole change-points using the observed coalescent genealogy. We do this explicitly within the context of the SMC, but observe that the same results apply to all other piecewise coalescent models. It is known that if we condition on $n-1$ events from an inhomogeneous Poisson process occurring in $[0, \epsilon_p)$, with intensity $\lambda(t)$, then the event times are independently and identically distributed according to density $f(t) = \frac{\lambda(t)}{\int_0^{\epsilon_p} \lambda(u)\mathrm{d}u}$ (Snyder and Miller 1991). If we let $\lambda(t)$ be our piecewise-constant SMC rate, we find that $\int_0^{\epsilon_p} \lambda(u)\mathrm{d}u = \sum_{i=1}^{T}\sum_{j=1}^{p} \gamma_j\beta_j = \sum_{j=1}^{p}(n-1)\gamma_j\beta_j$, with $\gamma_j = N_j^{-1}$ as the inverse population size over the region $[\epsilon_{j-1}, \epsilon_j)$. The pigeon-hole size or bin width is $\beta_j = \epsilon_j - \epsilon_{j-1}$ with the ϵ_j as the change-points, and T as the number of loci. Note that, for example, in the skyline demographic model, we would have a single locus, and the β_j would correspond to scaled interval times (see ω_j in the skyline demographic log-likelihood in the main text).

We can define the cumulative distribution function (CDF) at the pigeon-hole change-points as: $F(\epsilon_j) = \int_0^{\epsilon_j} f(t)\mathrm{d}t$ and denote the consecutive spacing of this CDF as $\Delta_j = F(\epsilon_j) - F(\epsilon_{j-1})$. Empirically, this CDF corresponds to the lineage through time plot (LTT) of the observed phylogeny, normalized by its total number of coalescent events. Solving for Δ_j using the piecewise-constant coalescent rate gives the left part of Eq. (A1). If we substitute the MLE for either β_j or γ_j (depending on what is known) then we derive $\hat{\Delta}_j$. Applying the m_j^* design from Theorem 1 produces the rest of Eq. (A1). These results are precisely the same for the skyline and structured models.

$$\Delta_j = \frac{\gamma_j\beta_j}{\sum_{i=1}^{p}\gamma_i\beta_i} \Longrightarrow \hat{\Delta}_j = \frac{m_j}{n-1} \Longrightarrow \hat{\Delta}_j^* \,|\, \mathbb{D} = \frac{1}{p} \qquad \text{(A1)}$$

The robust coalescent interval spacing, $\hat{\Delta}_j^*|\mathbb{D}$, is therefore fixed by the number of pigeon-holes (and hence parameters). This has two important ramifications. First, as quantiles are defined as inverse cumulative distribution values, it means that the optimal choice of pigeon-holes is such that their boundaries are the $\frac{1}{p}$ quantiles of the normalized LTT. Robust coalescent experimental design therefore recommends assigning a new pigeon-hole after every $\frac{n-1}{p}$ coalescent events (the LTT is simply the event counting process). This quantile design clearly suggests that the largest admissible number of change-points is at $p=n-1$. This limit, for skyline demographic inference problems, corresponds to the formulation of the classic skyline plot (Pybus et al. 2000).

Second, since the spacing at the MLE is constant, robustness is achieved by the maximum spacings estimate (MSE) (Cheng and Amin 1983; Ranneby 1984). For a given set of observations, drawn from the CDF of a parameter θ, the MSE is the estimate of θ that maximizes the geometric mean of the spacing of the CDF, evaluated at each observed random sample. Our results suggest that if we view the pigeon-hole change-points as binned draws from $f(t)$ then, given a robust design, the MSE of θ results in optimal spacing. Here, θ is the effective coalescent rate with density $f(t)$. It is not difficult to prove that robust designs for the skyline demographic and structured models also imply equivalent $\frac{1}{p}$ MSEs. Under MSE designs, the observed tree, from the perspective of the pigeon-holes, will appear as uniformly informative as possible.

Simulation Study: Square Wave Populations

Here, we show how to apply Theorem 1 to a simple skyline demographic coalescent model. Let $N(t)$ define a square wave population size function with period T, with time t into the past. $N(t)$ models the harmonic mean (Pybus et al. 2000) of the fluctuating number of infected individuals across time in a seasonal epidemic. N_1 recurs on odd half-periods and N_2 on even ones ($[0, \frac{T}{2})$ is the first (odd) half-period). Given n total samples ($n-1$ coalescent events) we want to optimally infer $N(t)$. Figure 1 of the main text illustrates the experimental set-up and notation for a similar design problem. Figure A1a shows a typical $N(t)$ with its half-period numbers.

The precision with which N_1 and N_2 are estimated is an increasing function of the number of coalescent events falling within their half-periods. Let m_{1i} be the number of events in the ith recurrence of N_1 and m_{2i} be the equivalent for N_2. Theorem 1 stipulates that robust sampling schemes will distribute $\frac{1}{2}$ of all coalescent events to N_1 half-periods (Eq. (A2)). Thus, if m_1 is the observed count of coalescent events falling within N_1 half-periods, then the performance of any sampling scheme can be measured by the size of the robust deviation $d(m_1) := \left|\frac{\mathcal{I}(\log N_1)}{n-1} - \frac{1}{2}\right| = \left|\frac{m_1}{n-1} - \frac{1}{2}\right|$. Note that $d(m_1)$ increases with Fisher information skewness (higher $\mathcal{I}(\log N_1)$ means lower $\mathcal{I}(\log N_2)$), and $d(m_1^*|\mathbb{D}) = d(m_2^*|\mathbb{D}) = 0$ (robust design is symmetric).

$$\mathcal{I}(\log N_1) = m_1 = \sum_{i\geq 0} m_{1(i+1)} \Longrightarrow m_1^* = \frac{1}{2}(n-1) \qquad \text{(A2)}$$

If we define p_1 as the probability that a sampled tip is introduced in an N_1 interval then a robust sampling strategy achieves $p_1^* = \text{argmin}_{p_1} d(m_1)$. We assume p_1 is constant with time. Thus, we focus on the mapping $p_1 \to d(m_1)$ with $p_2 = 1 - p_1$. A sampling protocol involves the tuple (s_k, ϕ_k) with s_k as the time of the kth sampling event at which ϕ_k lineages are introduced. Since coalescent events are always delayed in time relative to the point in time at which samples are placed, we will always introduce our ϕ_k samples all at once, and only at the change-points so that $s_k = (k-1)\frac{T}{2}$ (the arrows in Fig. A1a). This procedure maximizes the probability that samples will coalesce within the half-period in which they are introduced.

We examine a range of deterministic sampling strategies in order to explore how p_1 controls $d(m_1)$. For a given p_1, we set the number of samples introduced in N_1 and N_2 half-periods as fractions $f_1 = \text{round}[np_1]$ and $f_2 = n - f_1$. Here round indicates the nearest integer. We allocate the f_1 and f_2 samples uniformly relative to N_1 and N_2 half-periods respectively, so that $\phi_k = a$ or 0 depending on whether samples are introduced or not. Here $p_1 = 0$ means we have placed all n samples on N_2 half-periods while $p_1 = 1$ means that they are all on N_1 ones. Intermediate p_1 values compromise between these two extremes. We illustrate these sampling strategies for $a = 1$ and $n = 10$, relative to the half-periods of $N(t)$, in Figure A1b.

Figure A1c shows the sampling protocol performance under $a = 1$ schemes at different N_1 values (scaled against T), with $N_2 = 2N_1$. We find that as N_1 becomes smaller relative to T, the optimal protocol p_1^* gets closer to $\frac{1}{2}$. This makes sense since here population changes are slow relative to the coalescent times, so that we have the greatest chance of any sample coalescing within the half-period in which it was introduced. As N_1 increases, coalescent times lengthen and we get samples coalescing outside this original half-period. This leads to a weaker, less discernible minimum with larger uncertainty [we cannot estimate fluctuations in population that are fast compared to our rate of producing coalescent events (Sjodin et al. 2005)]. The optimal strategy here is $p_1^* < \frac{1}{2}$ (if we made $N_2 = \frac{1}{2}N_1$ we would get curves skewed in the opposite direction so that $p_1^* > \frac{1}{2}$). Our robust sampling recommendation is therefore to place more samples in periods of time where larger population size is expected. This has an interesting practical implication for structured coalescent models with known, symmetric migration rates. In this case the demes are directly analogous to the N_j segments, and robust sampling would be achieved by allocating sample numbers in proportion to the deme population sizes.

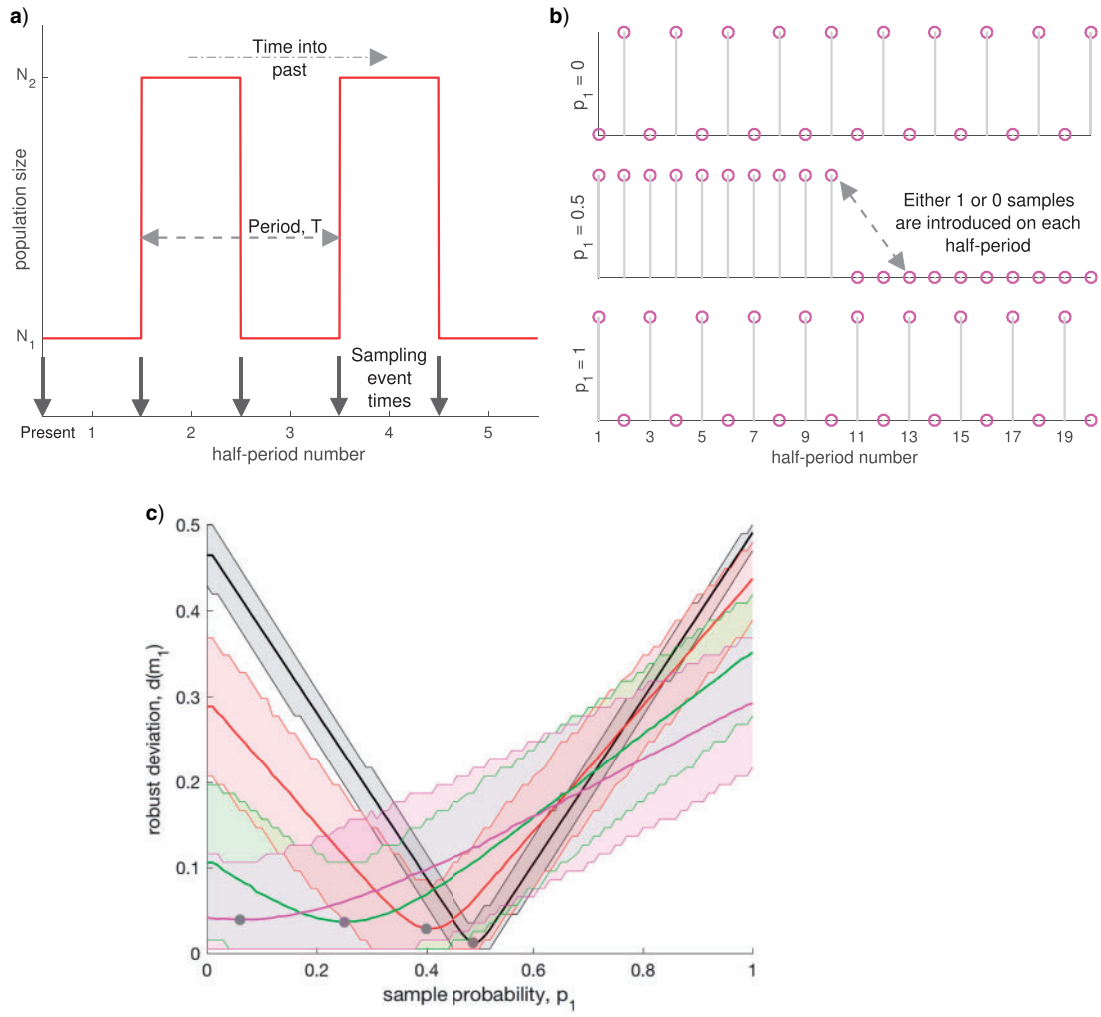

FIGURE A1. Deterministic sampling protocols for a skyline coalescent model. We apply a deterministic sampling strategy with $\phi_k = 1$ or 0 to a skyline demographic model with a population that fluctuates between N_1, and $N_2 = 2N_1$ across time. This fluctuation is described by a square wave with period T, and is shown in a) for $N_1 = \frac{T}{4}$ and $N_2 = \frac{T}{2}$. The arrows in this subplot indicate the points at which we can introduce a sample. Panel b) shows how $n = 10$ samples are allocated at these arrow points for three different p_1 protocols (p_1 controls the fraction of the n available samples that are placed in N_1 half-periods). We observe how the absolute difference, $d(m_1)$, between the Fisher information and the uniquely robust design changes with p_1 in c), for $n = 100$. The black, red, green and magenta curves (which feature increasingly wide confidence intervals) are for $N_1 = [\frac{T}{8}, \frac{T}{4}, \frac{T}{2}, T]$, respectively. Each curve gives the mean of $d(m_1)$ across 5000 repeated runs (solid line) and the 95% confidence interval around that mean. As N_1 decreases relative to T, $d(m_1)$ becomes more symmetric and maximal performance (defined as $\min d(m_1)$) improves (gets closer to 0 and has sharper confidence). The uniquely robust sampling protocol in each N_1 case, is visualized with a grey, filled circle. See the Appendix for further interpretations of these results.

REFERENCES

Atkinson A., Donev A. 1992. Optimal experimental designs. London, UK: Oxford University Press.

Banks H., Davidian M. 2009. Generalized Sensitivities and Optimal Experimental Design. Technical Report, North Carolina, USA: North Carolina State University.

Beerli P., Felsenstein J. 1999. Maximum likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics 152:763–773.

Beerli P., Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. *Proc. Natl. Acad. Sci. USA*, 98:4563–4568.

Beichman A., Huerta-Sanchez E., Lohmueller K. 2018. Using genomic data to infer historic population dynamics of nonmodel organisms. *Annu. Rev. Ecol. Evol. Syst.* 49:433–456.

Bhaskar A., Song Y. 2014. Descartes' rule of signs and the identifiability of population demographic models from genomic variation data. *Ann. Stat.* 42:2463–2493.

Box G., Cox D. 1964. An analysis of transformations. *J. R. Stat. Soc. B* 26(2):211–252.

Cheng R., Amin N. 1983. Estimating parameters in continuous univariate distributions with a shifted origin. *J. R. Stat. Soc. B* 45:394–403.

De Maio N., Wu C., O'Reilly K., Wilson D. 2015. New routes to phylogeography: a Bayesian structured coalescent approximation. *PLoS Genet.* 11:e1005421.

Drummond A., Rambaut A., Shapiro B., Pybus O. 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Mol. Biol. Evol*. 22:1185–1192.

Ewing G., Nicholls G., Rodrigo A. 2004. Using temporally spaced sequences to simultaneously estimate migration rates, mutation rate and population sizes in measurably evolving populations. *Genetics* 168:2407–2420.

Fisher R. 1956. Statistical methods and scientific induction. Edinburgh: Oliver and Boyd.

Freedman D. 1999. On the Bernstein-Von Mises theorem with infinite dimensional parameters. *Ann. Stat*. 27:1119–1140.

Friendly M., Monette G., Fox J. 2013. Elliptical insights: understanding statistical methods through elliptical geometry. *Stat. Sci*. 28:1–39.

Gattepaille L., Torsten G., Jakobsson M. 2016. Inferring past effective population size from distributions of coalescent times. *Genetics* 204:1191–1206.

Gill M., Lemey P., Faria N., Rambaut A., Shapiro B., Suchard M. 2012. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci . *Mol. Biol. Evol*. 30:713–724.

Griffiths R., Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos. Trans. R. Soc. B* 344:403–410.

Grunwald P. 2007. The minimum description length principle. Massachusetts, USA: The MIT Press.

Hall M., Woolhouse M., Rambaut A. 2016. The effects of sampling strategy on the quality of reconstruction of viral population dynamics using Bayesian skyline family coalescent methods: a simulation study. *Virus Evol*. 2(1): vew003.

Heller R., Chikhi L., Siegismund H. 2013. The confounding effect of population structure on Bayesian skyline plot inferences of demographic history. *PLoS One* 8:e62992.

Karcher M., Palacios J., Bedford T., Suchard M., Minin V. 2016. Quantifying and mitigating the effect of preferential sampling on phylodynamic inference. *PLoS Comput. Biol.* 12(3): e1004789.

Kay S. 1993. Fundamentals of statistical signal processing: estimation theory. New Jersey, USA: Prentice Hall.

Kim J., E, M., Racz M., Ross N. 2015. Can one hear the shape of a population history? *Theor. Popul. Biol.* 100:26–38.

Kingman J. 1982. On the genealogy of large populations. *J. Appl. Prob*. 19:27–43.

Le Cam L. 1986. Asymptotic methods in statistical decision theory. New York: Springer.

Lehmann E., Casella G. 1998. Theory of point estimation. 2nd ed. New York, USA: Springer.

Li H., Durbin R. 2011. Inference of human population history from individual whole-genome sequences. *Nature* 475:493–496.

Liepe J., Filippi S., Komorowski M., Stumpf M. 2013. Maximizing the information content of experiments in systems biology. *PLoS Comput. Biol.* 9:e1002888.

Liu X., Fu Y. 2015. Exploring population size changes using SNP frequency spectra. *Nat. Gen.* 47:555–562.

Marshall A., Olkin I., Arnold B. 2011. Inequalities: theory of majorization and its applications. 2nd ed. New York, USA: Springer Science + Business Media.

McVean G., Cardin N. 2005. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. B* 360:1387–1393.

Minin V., Bloomquist E., Suchard M. 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol. Biol. Evol*. 25:1459–1471.

Myung I., Balasubramanian V., Pitt M. 2000. Counting probability distributions: differential geometry and model selection. *Proc. Natl. Acad. Sci.* 97:11170–11175.

Nordborg M. 2001. Handbook of statistical genetics: coalescent theory. West Sussex, UK: John Wiley and Sons.

Notohara M. 1990. The coalescent and the genealogical process in geographically structured population. *J. Math. Biol*. 29:59–75.

Opgen-Rhein R., Fahrmeir L., Strimmer K. 2005. Inference of demographic history from genealogical trees using reversible jump Markov Chain Monte Carlo. *BMC Evol. Biol.* 5(6):1–14.

Palacios J., Wakeley J., Ramachandran S. 2015. Bayesian nonparametric inference of population size changes from sequential genealogies. *Genetics* 201:281–304.

Parag K., Pybus O. 2017. Optimal point process filtering and estimation of the coalescent process. *J. Theor. Biol.* 421:153–167.

Parag K., Pybus O. 2018. Exact Bayesian inference for phylogenetic birth-death models. *Bioinformatics* 34:3638–3645.

Pybus O., Rambaut A., Harvey P. 2000. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics* 155:1429–1437.

Ranneby B. 1984. The maximum spacing method: an estimation method related to the maximum likelihood method. *Scand. J. Stat.* 11:93–112.

Reinert G. 2009. Statistical Theory. Technical Report, Oxford, UK: University of Oxford.

Rothenburg T. 1971. Identification in parametric models. *Econometrica* 39(3):577–591.

Schiffels S., Durbin R. 2014. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* 46:919–925.

Sheehan S., Harris K., Song Y. 2013. Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach. *Genetics* 194:647–662.

Sjodin P., Kaj I., Krone S., Lascoux M., Nordborg M. 2005. On the meaning and existence of an effective population size. *Genetics* 169:1061–1070.

Snyder D., Miller M. 1991. *Random point processes in time and space*. 2nd ed. New York, USA: Springer.

Spence J., Steinrucken M., Terhorst J., Song Y. 2018. Inference of population history using coalescent HMMa: review and outlook. *Curr. Opin. Genet. Dev.* 53:70–76.

Stack J., Welch J., Ferrari M., Shapiro B., Grenfell B. 2010. Protocols for sampling viral sequences to study epidemic dynamics. *J. R. Soc. Interface* 7:1119–1127.

Steinrucken M., Kamm J., Song Y. 2015. Inference of complex population histories using whole-genome sequences from multiple populations. *BioRxiv* p. 026591.

Strimmer K., Pybus O. 2001. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Mol. Biol. Evol.* 18:2298–2305.

Tataru P., Nirody J., Song Y. 2014. diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals. *Bioinformatics* 30:3430–3431.

Vaughan T., Kuhnert D., Popinga A., Welch D., Drummond A. 2014. Efficient Bayesian inference under the structured coalescent. *Bioinformatics* 30:2272–2279.

Volz E., Kosakovsky Pond S., Ward M., Leigh Brown A., Frost S. 2009. Phylodynamics of infectious disease epidemics. *Genetics* 183:1421–1430.

Weissman D., Hallatschek O. 2017. Minimal-assumption inference from population-genomic data. *eLife* 6:e24836.

Zamir R. 1998. A proof of the Fisher information inequality via a data processing argument. *IEEE Trans. Inf. Theory* 44:1246–1250.

Zegers P. 2015. Fisher information properties. *Entropy* 17:4918–4939.