Robust Design for Coalescent Model Inference

Kris V Parag and Oliver G Pybus

***Abstract*—The coalescent process models how unobserved changes in the size of a population influence the genealogical patterns of sequences sampled from that population. The estimation of these hidden population size changes from reconstructed sequence phylogenies, is an important problem in many biological fields. Often, population size is described by a piecewise-constant function, with each piece serving as a parameter to be estimated. Estimate quality depends on both the statistical inference method used, and on the experimental protocol, which controls variables such as the sampling or parametrisation, employed. While there is a burgeoning literature focussed on inference method development, there is surprisingly little work on experimental design. Moreover, these works are largely simulation based, and therefore cannot provide provable or general designs. As a result, many existing protocols are heuristic or method specific. We examine three key design problems: temporal sampling for the skyline demographic coalescent model; spatial sampling for the structured coalescent and time discretisation for sequentially Markovian coalescent models. In all cases we find that (i) working in the logarithm of the parameters to be inferred (e.g. population size), and (ii) distributing informative (e.g. coalescent) events uniformly among these log-parameters, is provably and uniquely robust. 'Robust' means that both the total and maximum uncertainty on our estimates are minimised and independent of their unknown true values. These results provide the first rigorous support for some known heuristics in the literature. Given its persistence among models, this two-point design may be a fundamental coalescent paradigm.**

The coalescent process [1] is a popular population genetics model that describes how past (unobserved) changes, in the size or structure of a population, shape the reconstructed (observed) genealogy of a sample of sequences, from that population. This genealogy is also known as the coalescent tree or phylogeny. The estimation of a function of the past population size from the sequences, or reconstructed phylogeny, is important in many fields including epidemiology, conservation and anthropology. Accordingly, there is an extensive and growing literature [2] [3] [4] [5] [6] [7] [8] [9] [10] [11] [12] [13], focussed on developing new statistical methods for solving coalescent inference problems.

However, the power and accuracy of the resulting coalescent estimates is not solely a function of the statistical method employed. Design variables under the control of the experimenter, such as choices of where and when sequences are sampled, or on how time is discretised, can have a strong influence on the performance and reliability of coalescent inference methods [14] [9] [11]. Good designs can result in sharper inferences and sounder conclusions [14], whereas bad designs, such as size-biased sampling strategies, can often lead to overconfident or spurious estimates [15] [16]. The best approach to coalescent inference will therefore jointly optimise experimental design and statistical methodology.

Surprisingly, only a few studies have investigated optimal coalescent inference design. These works [14] [17] [12] [18] [15], typically take a constructive, simulation based approach, in which several alternate designs are numerically examined and compared. While such studies can yield useful hypotheses about the components of good designs, they can neither provide analytic insights nor provably optimal directives. A more general and methodical analysis is therefore needed.

Additionally, there has been little consideration of what data or parameter transformations might aid experimental design. This contrasts the development of inference theory in other fields. For example, in regression or analysis of variance problems, research has emphasised the benefits of power transformations and regularisation procedures [19]. While some coalescent inference methods have used parameter transformations (e.g. the log transform), these are usually justified by heuristic or method specific reasons, such as algorithmic stability or ease of visualisation [12] [20]. As a result, any transformations present in the coalescent literature are applied inconsistently, and rigorous proof of their benefits is lacking.

Here we take a fully analytical approach and formally derive optimal design directives for coalescent inference. As we are interested in widely applicable theoretical insights, we do not construct specific protocols, but instead define objectives which, if achieved, guarantee joint inference and experiment optimality in a well-defined sense. We examine three popular coalescent models. For each model we describe a coalescent tree as being composed of lineages, with time flowing from the present into the past. A coalescent event is said to occur when two lineages merge into an ancestral lineage.

(1) Skyline demographic models. These approaches infer past population size changes using piecewise-constant time-varying functions [21], and are widely used in epidemiology, where the population is the infected class in an epidemic. The time-varying functions can describe seasonal and growth dynamics. These models are at the core of the popular 'skyline' family of inference methods [2]. The choice of sequence sampling times, which is our design variable here, can strongly impact on the quality of inference for a given epidemic. Robust inferences could improve epidemic control strategies [4] [14].

(2) Structured models. These processes describe spatial changes. The population is divided into a number of distinct but connected sub-populations (demes). Usually each deme has a constant (stable) population size. Lineages may migrate between demes but can only coalesce within demes. The parameters of interest include both the population sizes and migration rates [22] [23]. The design variable is the space-time sequence sample distribution, which is known to affect the bias with which migration rates can be inferred [9]. This model has been applied to describe the migration history of animal, plant and pathogen populations [9].

(3) Sequentially Markovian coalescent (SMC) models. These are typically applied to complete metazoan genomes, and consider many independent coalescent trees (multiple unlinked loci), each containing few (or two) samples. SMC processes involve recombination, and event times are discretised to occur in finite intervals. Past population size change is often assumed to be piecewise-constant and most applications centre on human demographic history [10] [12]. The design variable is the time discretisation, which controls the resolution with which populations are estimated. Poor discretisations can lead to overestimation or runaway behaviour [11].

We examine these three types of models using Fisher information and optimal design theory. Since the time between coalescent events contains information about population size change, the total number of observed coalescent events controls the amount of information available. We show that, under this constraint, it is optimal to (i) work in the logarithm of the parameters to be estimated, which usually relate to effective population size, and (ii) sample or discretise such that the coalescent events are divided evenly among each log scaled parameter. If (i)-(ii) are both achieved, then the resulting experimental design is provably robust and optimal for use with existing inference methods. 'Robust' means that the design minimises the maximum dimension and the total volume of the confidence ellipsoid that circumscribes asymptotic estimate uncertainty. Interestingly, these two objectives hold across all the coalescent models we investigated and therefore present simple, unifying principles for coalescent inference.

In the Preliminaries we provide mathematical background on optimal experimental design. We use these concepts to derive our main robust design theorem for piecewise coalescent inference, in

K.V. Parag and O. G. Pybus are in the Department of Zoology, University of Oxford, Oxford, OX1 3SY, UK, e-mail: kris.parag@zoo.ox.ac.uk

Results. This is then applied to each of the three previously described coalescent models, yielding new and specific insights. We close with a Discussion of how our formally derived design principles relate to existing heuristics in the coalescent inference literature.

PRELIMINARIES

Consider an arbitrary parameter vector $\psi = [\psi_1, \ldots, \psi_p]$, which is to be estimated from a statistical model. Let $\mathcal{T}$ represent data (a random variable sequence) generated under this statistical model (the tree in the case of coalescent inference) and let $L(\psi) := \log \mathbb{P}(\mathcal{T} \mid \psi)$ be the log-likelihood of $\mathcal{T}$ given ψ. The $p \times p$ Fisher information matrix, denoted $\mathcal{I}(\psi)$, is the appropriate measure for describing how informative $\mathcal{T}$ is about ψ [24]. Since all the coalescent models used here belong to an exponential family [25] (and so satisfy certain necessary regularity conditions [26]) then the (i^{th}, j^{th}) element of $\mathcal{I}(\psi)$ is defined as $\mathcal{I}(\psi)_{(i,\,j)} := -\mathbb{E}_{\mathcal{T}}\left[\frac{\partial^2 L}{\partial\psi_i \partial\psi_j}\right]$, with the expectation taken across the data (tree branches).

Thus, the Fisher information is synonymous with the curvature of the likelihood surface in our work. It is also sensitive to parametrisation choice. Eq. (1) provides the transformation between ψ and an arbitrary alternate p-parameter vector $\sigma = [h(\psi_1), \ldots, h(\psi_p)] = [\sigma_1, \ldots, \sigma_p]$. Here h is a continuously differentiable function, with inverse $f = \text{inv}[h]$ [25].

$$\mathcal{I}(\sigma)_{(i,\,j)} = \left(\frac{\partial \psi_i}{\partial \sigma_j}\right)^2 \mathcal{I}(f(\sigma))_{(i,\,j)} \tag{1}$$

The Fisher information lower bounds the best unbiased estimate precision attainable, and quantifies the confidence bounds on maximum likelihood estimates (MLEs). For exponential families, these bounds are attained so that if $\hat{\psi}$ is the MLE then $\text{var}(\hat{\psi}_j) = \text{inv}\left[\mathcal{I}(\psi)_{(j,\,j)}\right]$ is the minimum variance around the j^{th} MLE achievable by any coalescent model inference method [27]. Importantly, for any given parametrisation, the Fisher information serves as a metric with which we can compare various estimation schemes (e.g. different sampling or discretisation protocols). Thus different estimators can be ordered in performance by their Fisher information values. This ordering also incorporates the quality of the data upon which we base our analyses. Due to these ordering attributes, we propose Fisher information as our design metric.

Although the most popular coalescent estimators use Bayesian inference, it seems that we have taken a MLE or frequentist approach. However, since all our statistical models are finite dimensional, the Bernstein-von Mises theorem [28] [29] is valid. This states that, asymptotically, any Bayesian estimate will have a posterior distribution that matches that of the MLE, with equivalent confidence intervals, for any 'sensibly defined' prior. Such a prior has some positive probability mass in an interval around the true parameter value. As a result, Bayesian credible intervals also depend on the Fisher information and our designs remain applicable.

Optimal design theory aims to optimise experimental protocols, based on statistical criteria that confer useful properties such as minimum bias or maximum precision [30]. The theory centres on the notion that some measurements are potentially more informative than others. Within this context, we treat our sampling/discretisation protocol problem as an experimental design on the distribution of coalescent events. If our observed data $\mathcal{T}$ consists of a total of $n-1$ coalescent events (i.e. a tree with n tips) then we can express our coalescent event distribution as $\{m_j\}$ for $1 \leq j \leq p$ with $\sum_{j=1}^{p} m_j = n-1$. Here m_j is the count of coalescent events that are informative of parameter ψ_j. This is illustrated for a two parameter skyline demographic model in Fig. 1.

Optimality criteria are typically functions of $\mathcal{I}(\psi)$, which defines our asymptotic uncertainty about $\hat{\psi}$. Geometrically, this uncertainty maps to a confidence ellipsoid centred on $\hat{\psi}$ [31]. Designing the Fisher information matrix is equivalent the controlling the shape and size of this ellipsoid. We focus on two popular criteria, known as D and E-optimality [31] [30], the definitions of which are given in Eq. (2) and Eq. (3), with $\{m_j^*\}$ as the resulting optimal design. As we have p design variables (the m_j), our confidence ellipsoid is p-dimensional. D-optimal designs minimise the volume of this confidence ellipsoid while E-optimal ones minimise its maximum diameter. Fig. 2 shows these ellipses for a skyline demographic design problem.

$$\{m_j^* \mid D\} = \arg\max_{\{m_j\}} \det\left[\mathcal{I}(\psi)\right] \tag{2}$$

$$\{m_j^* \mid E\} = \arg\max_{\{m_j\}} \min \text{eig}\left[\mathcal{I}(\psi)\right] \tag{3}$$

Here arg, det and eig are short for argument, determinant, and eigenvalues respectively. D-optimal designs therefore maximise the total available information gained from the set of parameters while E-optimal ones ensure that the worst estimate is as good as possible [31] [30].

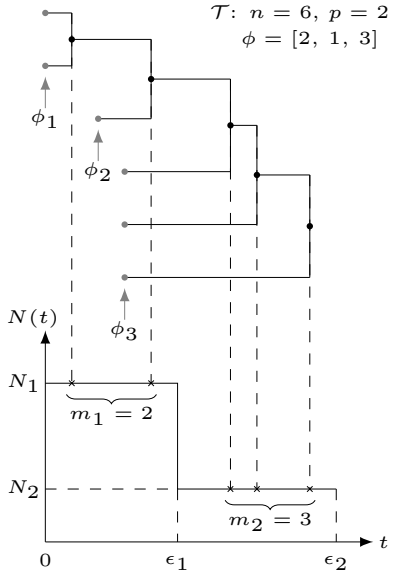

Fig. 1. **Problem set-up for a two-parameter coalescent model.** We consider a a $p = 2$ design problem for a skyline demographic coalescent model with population size parameters, N_1 and N_2. An $n = 6$ tip coalescent phylogeny, $\mathcal{T}$, is shown with the ϕ_k counting the samples introduced at the k^{th} sample time. The j^{th} population parameter is only informed by the number of coalescent events, m_j, occurring within its period $[\epsilon_{j-1},\, \epsilon_j]$, with $\epsilon_0 = 0$ as the present. Time flows into the past. We use experimental design theory to choose an m_1 and m_2 subject to $m_1 + m_2 = n - 1$, that leads to desirable properties on the resulting population size estimates. For a fixed n, we can manipulate ϕ to achieve our optimal m_j design.

The above optimisation problems can be solved using majorization theory, which provides a way of naturally ordering vectors [32]. For some p-dimensional vectors $\vec{a}$ and $\vec{b}$, sorted in descending order to form $\vec{a}^{\downarrow}$ and $\vec{b}^{\downarrow}$, $\vec{a}$ is said to majorize or dominate $\vec{b}$ if for all $k \in \{1, 2, \ldots p\}$, $\sum_{j=1}^{k} \vec{a}^{\downarrow} \geq \sum_{j=1}^{k} \vec{b}^{\downarrow}$ and $\sum_{j=1}^{p} \vec{a} = \sum_{j=1}^{p} \vec{b} = \kappa$. Here κ is a constant and this definition is written as $\vec{a} \succ \vec{b}$ for short. The total sum equality on the elements of the vectors is called an isoperimetric constraint. Conceptually, the majorization of a vector preserves its mean but increases its variance.

We will make use of Schur concave functions. A function g that takes a p-dimensional input and produces a scalar output is called Schur concave if $\vec{a} \succ \vec{b} \implies g(\vec{a}) \leq g(\vec{b})$. Importantly, it is known

that the p-element uniform vector $\vec{u} = [\frac{\kappa}{p}, \frac{\kappa}{p}, \ldots \frac{\kappa}{p}]$ is majorized by any arbitrary vector of sum κ and dimension p [32]. This means that every $\vec{a} \succ \vec{u}$. As a result, $\vec{u} = \arg\max_{\vec{a}} g(\vec{a})$ for any Schur concave function g. Thus if we can find a Schur concave function, and an isoperimetric constraint holds, then a uniform vector will maximise that function. This type of argument will underpin many of the optimisations in subsequent sections.

Results

Naive Coalescent Design

Let $N = [N_1, \ldots, N_p]$ be the parameter vector (usually effective population size values) to be estimated from a reconstructed phylogeny, $\mathcal{T}$. Defining $\gamma = [N_1^{-1}, \ldots, N_p^{-1}]$, we will find that the piecewise-constant coalescent models examined in this work admit log-likelihoods, $L(\gamma) = \log \mathbb{P}(\mathcal{T} \,|\, \gamma)$, of the form of Eq. (4).

$$L(\gamma) = \sum_{j=1}^{p} m_j \log \gamma_j - A_j \gamma_j + B_j \tag{4}$$

Here A_j and B_j are constants, for a given $\mathcal{T}$, and $\gamma_j = N_j^{-1}$. Taking partial derivatives we get $\frac{\partial L}{\partial \gamma_j} = m_j \gamma_j^{-1} - A_j$ and observe that the MLE of γ_j, $\hat{\gamma}_j = m_j A_j^{-1}$. The second derivatives follow as: $\frac{\partial^2 L}{\partial \gamma_j^2} = -m_j \gamma_j^{-2}$, $\frac{\partial^2 L}{\partial \gamma_j \partial \gamma_{i \neq j}} = 0$. This leads to a diagonal Fisher information matrix $\mathcal{I}(\gamma) = [m_1 \gamma_1^{-2}, \ldots, m_p \gamma_p^{-2}]\, \mathrm{I}_p$, with I_p as a $p \times p$ identity matrix. Using Eq. (1) we obtain the Fisher information in our original parametrisation as Eq. (5).

$$\mathcal{I}(N) = [m_1 N_1^{-2}, \ldots, m_p N_p^{-2}]\, \mathrm{I}_p \tag{5}$$

Several key points become immediately obvious. First, the achievable precision around $\hat{N}_j = \hat{\gamma}_j^{-1}$ depends on the square of its unknown true value. This is a highly undesirable property, since it means our estimate confidence is not only largely out of our control, but also will rapidly deteriorate as N_j grows. Second, if our inference method directly estimated γ instead of N (which is not uncommon for harmonic mean estimators [2]), then the region in which we achieve good γ precision is exactly that in which we obtain poor N confidence.

Third, the design variables we do control, $\{m_j\}$, only inform on one variable of interest. This means that good designs must achieve $m_j \geq 1$ for all j. Failure to attain this will result in a singular Fisher information matrix and hence parameter non-identifiability [33], which can lead to issues like poor algorithmic convergence. This is particularly relevant for coalescent inference methods that feature pre-defined parameter grids of size comparable to the tree size n [34]. Thus, naive implementations of coalescent inference methods and ad-hoc design protocols can easily result in potentially serious computational and methodological issues.

Using either the N or γ parametrisation further creates issues even when it comes to optimal design. Consider the N parametrisation 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 is 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 $m_1^* N_1^{-2} = m_2^* N_2^{-2} = \ldots = m_p^* N_p^{-2}$ [32], and is the second equality in Eq. (6). This optimal design assigns more coalescent events to larger populations 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.

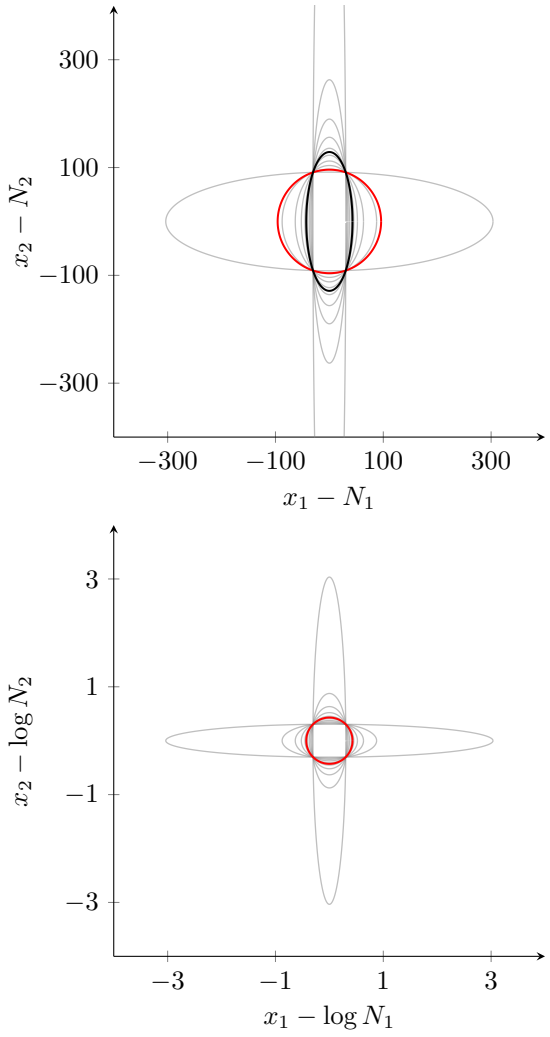

Fig. 2. **D and E-optimal designs for a two-parameter model.** We provide asymptotic 99% confidence ellipses for a $p = 2$ skyline demographic design problem (see Fig. 1) with $n - 1 = 100 = m_1 + m_2$, $N_1 = 100$ and $N_2 = 2N_1$. The ellipses depict the confidence region of the two dimensional asymptotic normal distribution that has covariance matrix equal to the inverse of the Fisher information. Each light grey ellipse indicates a different $[m_1, m_2]$ distribution. D and E-optimal designs are in red and dark grey respectively. Panel (a) shows the design space in absolute population size, N_j with $m_1^* \,|\, D = 50$ and $m_1^* \,|\, E = 20$. Panel (b) uses the log population size, $\log N_j$. The log parametrisation results in a symmetrical, robust design space that has coincident D and E-optimal ellipses with $m_1^* \,|\, \mathbb{D} = 50$.

Thus, naive D-optimal designs could result in some parameters being poorly estimated while E-optimal ones could allocate all the coalescent events to a single parameter, and hence increase the possibility of non-identifiability. 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 optimality criteria is needed.

These points are illustrated in the top panel of Fig. 2, which presents D and E-optimal confidence ellipsoids under N, for the example model in Fig. 1. These ellipsoids, for some parameter vector σ, with diagonal Fisher information matrix $\mathcal{I}(\sigma)$, are given by $\sum_{j=1}^{p} (x_j - \sigma_j)^2 \mathcal{I}(\sigma)_{(j,j)} = \Omega$. Here Ω controls the significance

level according to a p degree of freedom χ^2 distribution and x_j is some coordinate on the j^{th} parameter axis [35]. Here, the 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 design as being (i) insensitive to the true (unknown) parameter values and (ii) minimising 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. In subsequent sections we will apply this robust design to the three aforementioned coalescent models.

Theorem 1. If the p-parameter vector σ admits a diagonal Fisher information matrix, $\mathcal{I}(\sigma) = [m_1\sigma_1^{-2}, \ldots, m_p\sigma_p^{-2}]\mathrm{I}_p$, under an isoperimetric constraint $\sum_{j=1}^{p} m_j = \kappa$, then any design that (i) works in the parametrisation $[\log\sigma_1, \ldots, \log\sigma_p]$ and (ii) 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 (i), by maximising how distinguishable our parameters are within their space of possible values. 'Distinguishability' is an important property that determines parameter identifiability and model complexity [36]. Let ψ be some parametrisation with space Ψ such that $h(\psi) = \sigma$. Two vectors in Ψ, $\psi_{(1)}$ and $\psi_{(2)}$, are distinguishable, if, given $\mathcal{T}$, we can discriminate between them with some confidence. Distinguishability is therefore intrinsically linked to the quality of inference. More detail on these information geometric concepts can be found in [37] [36].

The number of distinguishable distributions in Ψ is known to be given by the volume, $\mathcal{V} = \int_{\Psi} \det\left[\frac{1}{n-1}\mathcal{I}(\psi)\right]^{\frac{1}{2}} \mathrm{d}\psi$ [36]. The $n-1$ comes from the number of informative events in $\mathcal{T}$. While $\mathcal{V}$ is invariant to the parametrisation choice h [36], different h functions control how parameter space is discretised into distinguishable segments. For example, under $\psi = \sigma$ poor distinguishability will result when any σ_j becomes large.

We therefore pose the problem of finding an optimal bijective parameter transformation $h(\psi_j) = \sigma_j$, which maximises how distinguishable our distributions across parameter space are, or equivalently minimises the sensitivity or our estimates to the unknown true values of our parameters. Note that we can always recover MLEs in our original parameters as $\hat{\sigma}_j = h(\hat{\psi}_j)$ [25], whilst taking advantage of the better estimate confidence properties less sensitive and more evenly distinguishable parametrisations provide.

Applying Eq. (1), with $h' := \frac{\partial h}{\partial \psi_j}$, we get that $\mathcal{I}(\psi)_{(j,\,j)} = m_j h^{-2}\,(h')^2$. 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 parametrisation, our parameters are orthogonal and our transformation 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).

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

This equality defines the conserved property across parametrisations of the piecewise-constant coalescent. We can maximise both the insensitivity of our parametrisation, 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 optimal parametrisation is $\psi_j = \log\sigma_j$ and hence proves (i).

Point (ii) follows by solving optimal design problems under the $\log\sigma$ parametrisation. 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).

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

Let $\mathbb{D}$ be an optimal design criterion, with resulting distribution $m_j^* \,|\, \mathbb{D}$. When $\mathbb{D} \equiv$ D, we maximise $\det[\mathcal{I}(\log N)]$ 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 behaviour under other design criteria. When $\mathbb{D} \equiv$ E, we maximise $\min\mathrm{eig}\,[\mathcal{I}(\log N)]$ to again obtain Eq. (8). This is very different from analogous designs under N and N^{-1}. Additional T and A-optimal designs (which maximise the trace of $\mathcal{I}(\log N)$ and its inverse respectively) also yield the same result or are satisfied under it.

Thus, under a log parametrisation we see an important convergence of optimality criteria to the uniform design of Eq. (8). This results in parameter confidence ellipsoids that are invariant to optimality criteria. This is shown in the bottom panel of Fig. 2 for our example model. This desirable design insensitivity emerges from the independence of $\mathcal{I}(\log N)$ from N, for piecewise-constant coalescent models, and proves (ii). We will now apply Theorem 1 to the skyline demographic, structured and sequentially Markovian coalescent models.

Skyline Demographic Models

We consider a coalescent process with piecewise-constant time-varying population size, $N(t)$, for $t \geq 0$, that features sequences sampled at different times. The coalescent tree always starts from the present, $t = 0$, with positive time going into the past. This model is often applied to epidemiological inference problems [21] [38], and is central to the popular 'skyline' family of estimation methods used in this field [2] [3] [4] [20]. We can describe $N(t)$ with $p \geq 1$ population sizes as $N(t) := \sum_{j=1}^{p} N_j \mathbb{1}(\epsilon_{j-1} \leq t < \epsilon_j)$ with $\epsilon_0 = 0$ and $\epsilon_p = \infty$. N_j is the constant population size of the j^{th} segment which is delimited by times $[\epsilon_{j-1},\, \epsilon_j)$. The indicator function $\mathbb{1}(a) = 1$ when a is true and 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 $(s_k,\, \phi_k)$ 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 $\phi_1 \geq 2$. In keeping with the literature, we assume that sampling times are independent of $N(t)$ [4]. The choice of sampling times and numbers (i.e. the temporal sampling protocol) is what the experimenter has control over in this coalescent inference problem. Fig. 1 explains this notation for a $p = 2$ skyline demographic model.

The observed n tip tree has $n-1$ coalescent events. We use c_i to denote the time of the i^{th} 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 k^{th} sample time $l(t)$ increases by ϕ_k and at every c_i it decreases by 1. The rate of producing coalescent events can then be 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 in segment j. We initially work in $\gamma = [\gamma_1, \ldots, \gamma_p]$ as

it is the natural parametrisation of the coalescent process, and then transform to N space.

The log-likelihood $L(\gamma) = \log \mathbb{P}(\mathcal{T} \,|\, \gamma)$ follows from Poisson process theory as [39] [5]: $L(\gamma) = -\int_0^{c_{n-1}} \lambda(t)\,\mathrm{d}t + \sum_{i=1}^{n-1} \log \lambda(c_i)$. Splitting the integral across the p segments we get: $\int_0^{c_{n-1}} \lambda(t)\,\mathrm{d}t = \sum_{j=1}^{p} \gamma_j \int_{\epsilon_{j-1}}^{\epsilon_j} \alpha(t)\,\mathrm{d}t = \sum_{j=1}^{p} \gamma_j \omega_j$. Here ω_j is a constant for a given tree and it is independent of γ. Similarly, $\sum_{i=1}^{n-1} \log \lambda(c_i) = \sum_{j=1}^{p} \sum_{i=1}^{n-1} \log(\gamma_j \alpha(c_i) 1(\epsilon_{j-1} \le c_i \le \epsilon_j))$. Expanding yields Eq. (9) with Γ_j as a constant depending on $\alpha(c_i)$ for all i falling in the j^{th} segment. The count of all the coalescent events within $[\epsilon_{j-1}, \epsilon_j]$ is m_j.

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

Eq. (9) is an alternate expression of the skyline log-likelihood given in [4], except that $N(t)$ is not constrained to change only at coalescent times. Importantly, sampling events do not contribute to the log-likelihood [4]. 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 coalescent distribution.

Since Eq. (9) is equivalent to Eq. (4), Theorem 1 applies, and the relevant robust design is exactly given by Eq. (8). Note that the lineage scaling, $\alpha(t)$, the timing of the m_j events falling within $[\epsilon_{j-1}, \epsilon_j]$, and the wait between the last of these and ϵ_j are all non-informative. As an illustrative example, we solve a skyline demographic design problem in the Supporting Text. There we apply Theorem 1 to a square wave approximation of a cyclic epidemic with known period, and determine what sampling protocols map to robust designs. The example conforms to the descriptions in Fig. 1 and Fig. 2.

Lastly, we comment on the impact of priors. More recent skyline inference methods use smoothing priors that ease the sharpness of the inferred piecewise constant population profile [20] [34]. While these do embed extra implicit information about N_j, they do not alter the optimal design point, even at small n. This follows as the informativeness of these priors do not change with $\{m_j\}$, so that the robust design proceeds independently of the benefits they provide.

Structured Models

The structured coalescent models the genetic relationships between samples from interconnected sub-populations, or demes. Sampled lineages in a given deme may either coalesce with others in the same deme, or migrate to any other deme [23]. When applied to empirical data, the structured model typically assumes a stable (constant) population in each deme with constant rates of migration [9]. Here, our parameters of interest are both the migration rates and population sizes in each deme, and the variables under our control are the choice of sampling times and locations.

Let $\mathcal{T}$ be an observed structured coalescent tree with $p \ge 1$ demes that have been sampled through time (branches are labelled with locations). We set T as the number of intervals in this tree, with each interval delimited by a pair of events, which can be sampling, migration or coalescent events. The i^{th} 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 denote 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 i as ζ_{ji}. N_j and $\gamma_j = N_j^{-1}$ are the absolute and inverse population size in deme j. Our initial p^2 parameter vector 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 $p-1$ sub-vector of all the migration rates from deme k. The log-likelihood $L(\sigma) = \log \mathbb{P}(\mathcal{T} \,|\, \gamma, \zeta)$ is then adapted from [22] and [40]. We decompose $L(\sigma) = \sum_{j=1}^{p} L_j(\gamma) + L_j(\zeta)$ into coalescent and migration sums with j^{th} deme components given in Eq. (10) and Eq. (11). Here m_j and w_{jk} respectively count the total number of coalescent events in sub-population j and the sum of migrations from that deme into deme k, across all T time intervals. The factor $\alpha_{ji} := \binom{l_{ji}}{2}$. 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 generalisations of Eq. (4) and lead to diagonal (orthogonal) Fisher information matrices like Eq. (5). This orthogonality make sense, since migration events do not inform on population size and coalescent events tell 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 affect the Fisher information for piecewise-constant coalescent models. Importantly, the Fisher information is independent of the spatio-temporal sampling procedure. The sampling protocol does, however, affect the time and location of coalescent and migration events and will serve as our means of achieving optimal directives.

Applying Theorem 1, we find that we should infer $\psi = [\log N_1, \ldots \log N_p, \{\log \zeta_{1\bar{1}}\}, \ldots \{\log \zeta_{p\bar{p}}\}]$ from structured models. This removes the dependence on both the unknown population sizes and migration rates, and leads to a Fisher information of $\mathcal{I}(\psi) = [m_1, \ldots m_p, \{w_{1\bar{1}}\}, \ldots \{w_{p\bar{p}}\}]\mathrm{I}_{p^2}$. The robust design is given in Eq. (12). The migration rate design, $w_{ji}^* \,|\, \mathbb{D}$, only holds if the total number of migration events are fixed, i.e. $\sum_{j=1}^{p} \sum_{i=1,\, i\neq j}^{p} w_{ji} = M$, for some constant M.

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

Distributing informative events uniformly among demes therefore results in a robust design. The separation of $\{m_j^*\}$ and $\{w_{ji}^*\}$ is a consequence of both the independent constraints on them and the orthogonality of the Fisher information matrix. Two points become 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 precisely the same robustness results as the skyline demographic model. Second, the migration rate design is exactly the same at both the strong and weak migration limits of the structured model [41]. The migration rates therefore do not affect the optimal design, provided log-migration rates are inferred.

Sequentially Markovian Coalescent Models

The previous coalescent models involved genealogies with many samples from a few (usually one) loci [13]. The large sample size of these trees meant that choice of sampling protocol was a critical design variable. This is often the case for coalescent applications in molecular epidemiology [38]. We now shift focus to coalescent inference methods for human and animal genomes. Here a coalescent model with recombination is applied along the genome, resulting in many hidden trees (multiple loci) [10]. Each tree typically consists of a small number of lineages. The most popular inference methods in this field are based on an approximation to the coalescent with recombination called the sequentially Markovian coalescent (SMC) [42].

These methods generally handle SMC inference by constructing a hidden Markov model (HMM) over discretised coalescent time [10] [43] [11]. If we partition time into p segments: $0 = \epsilon_0 < \epsilon_1 < \ldots < \epsilon_p = \infty$ then when the HMM is in state j it means that the coalescent time is in $[\epsilon_{j-1}, \epsilon_j)$ [11]. Recombinations lead to state changes and the genomic sequence serves as the observed process of the HMM. Expectation-maximisation type algorithms are used to iteratively infer the HMM states from the genome [10] [43].

A central aspect of these techniques is the assumption that in each coalescent interval the population size is constant [12]. If we use the vector $N = [N_1, \ldots, N_p]$ to denote population size then it is common to assign N_j for the $[\epsilon_{j-1}, \epsilon_j)$ interval [10]. This not only allows an easy transformation from the inferred HMM state sequence to estimates of N [13] but also controls the precision of the SMC based inference methods. For example, if too few coalescent events fall within $[\epsilon_{j-1}, \epsilon_j)$, then N_j will generally be overestimated [11]. Thus, the choice of discretisation times (and hence population size change-points) is critical to SMC inference performance [12] [44].

Our experimental design problem involves finding an optimal guideline for choosing these discretisation times. Currently, only a number of heuristic strategies exist [11] [13] [12]. 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 [10] and [43] we assume that each tree only leads to a single coalescent event, and hence we can neglect lineage counts. Since these counts merely rescale time (piecewise) linearly, we do not lose generality.

Let m_{ij} be the number of coalescent events observed in bin β_j from the i^{th} 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 $\sum_{j=1}^{p} m_j = n - 1$. Using Poisson process theory we can write the log-likelihood of observing a set of coalescent event counts $\{m_{ij}\}$, within our bins $\{\beta_j\}$ for the i^{th} locus as $L_i(\gamma, \beta) = \log \mathbb{P}(\mathcal{T}_i \,|\, \gamma, \beta) = -\int_0^\infty \lambda(t)\,\mathrm{d}t + \sum_{j=1}^{p} m_{ij} \log\left(\int_{\epsilon_{j-1}}^{\epsilon_j} \lambda(t)\,\mathrm{d}t\right)$ [39]. Here $\lambda(t)$ is the coalescent rate at t so that $\lambda(t) = \sum_{j=1}^{p} \gamma_j \mathbb{1}(\epsilon_{j-1} \leq t < \epsilon_j)$ and $\int_{\epsilon_{j-1}}^{\epsilon_j} \lambda(t)\,\mathrm{d}t = \beta_j\gamma_j$ with $\gamma_j = N_j^{-1}$. Using the independence of the T loci gives the complete log-likelihood of Eq. (13).

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

Eq. (13) is an alternative form of the log-likelihood given in [45], and describes a binned coalescent process that is analogous to the discrete one presented in [44]. 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 discretisation is seen to be as difficult as estimating population sizes [13].

Eq. (13) is analogous to Eq. (4), and so results in Fisher information matrices with square dependence on either N_j or β_j depending on what is known. Applying Theorem 1, we find that robust designs work in $\psi = [\log\beta_1, \ldots, \log\beta_p]$, if population size history is known (in keeping with discretisation results presented in [44]), or in $\psi = [\log N_1, \ldots, \log N_p]$, if the bins are known. Under either parametrisation we recover the expressions of Eq. (8) exactly. We generally assume bin sizes are known since they can often be controlled by the user [12]. However, in both scenarios, robust design requires that bins capture an equal number of coalescent events.

Discussion

Judicious experimental design can improve the ability of any inference method to extract useful and usable information from observed data [46]. In spite of these potential advantages, experimental design has received little attention in the coalescent inference literature [15]. We therefore defined and investigated robust designs for three important and popular coalescent models. Theorem 1, which summarises our main results, presents a clear and simple two-point robust design.

The first point 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 parametrisation for piecewise coalescent estimation problems. While methods using $\log N$ do exist [12] [20], their stated reasons for doing so are heuristic, and centre around algorithmic convenience or forcing estimates to be positive. To our knowledge, we have provided the first firm theoretical backing for using $\log N$ in coalescent inference.

It is worth noting that our result is closely linked to the theory of variance stabilising transforms. These transforms are often used in regression problems to make data more homoscedastic [19]. For exponentially distributed data δ, this transform is $\log\delta$ [19], and can be shown to hold for the Kingman coalescent [1], if we set δ as the coalescent inter-event times scaled by a binomial lineage count factor [5]. Generalising this scaling to each piecewise-constant segment should yield the log-transformation as optimal. Variance stabilising transforms are usually applied to the observed data, while our work focusses on the inherent parametrisation. We stabilise the estimator variance instead of the tree variance, and homoscedasticity is achieved as a by-product of maximising parameter distinguishability.

The second point of Theorem 1 requires equalising the number of coalescent events informing on each parameter. This may initially appear obvious as apportioning data evenly among the unknowns to be inferred seems wise. In fact, [11] and [44], which focus on SMC models, explicitly state that ideal time discretisations should achieve a uniform coalescent distributions. However, this appears to simply be a sensible assumption, since no proof or reference is given. Further, no mention is made in these works about log-transforms. We not only provide theoretical support for uniform coalescent distributions but also stress that they are only robust if the log-parameter stipulation is satisfied.

Several unifying insights, for piecewise-constant coalescent models, also emerge as corollaries of our robust design analysis. In particular, because the precision with which we estimate a coalescent parameter only depends of the number of coalescent events informing on it, we can reinterpret all designs as simply means of allocating events to effective 'slots'. For the models we examined, these slots represent skyline intervals, demes and time discretisation bins, respectively. From this perspective, for example, extra demes in the structured coalescent are equivalent to additional piecewise-constant population intervals in the skyline demographic model. Knowing the slot times (change-points) is crucial for inference [44].

Throughout this work, we have assumed that these effective slot times are known. This is reasonable as it is generally not possible (without side information) to simultaneously infer parameters and bin times [11] [44]. Often such low-information problems have a circular dependence in which we need to know the coalescent time distribution in order to estimate the parameters that determine that distribution [13]. Methods that do manage to achieve such joint estimates are usually data driven, iterative and case specific, allowing no general design insights [12] [47]. This raises the question about how to derive design benchmarks in such scenarios.

In the Supporting Text, we attempt to compute such benchmarks for change-points, using Theorem 1. Interestingly, we show that it is wise to assign end-points according to the $\frac{1}{p}$ quantiles of the normalised lineages through time plot of the observed phylogeny.

This results in a maximum spacings estimator (MSE) that makes the observed tree, from the perspective of the slots, as uniformly informative as possible [48]. This means that if we wish to robustly infer p log-parameters from a tree containing $n-1$ coalescent events, we should define our parameter slots such that they change every $r = \frac{n-1}{p}$ events. This has some interesting ramifications.

Optimal skyline population profiles were examined in [3], with groupings made on the basis of time. We instead propose groupings on event counts. If $r = 1$, we recover the classical skyline [2] as the low information limit of this MSE strategy. Note that grouping skyline intervals is equivalent to grouping demes in structured models. Interestingly, this MSE design provides a precise link to some popular SMC discretisation protocols. Specifically, [10] based its discretisation on a log spacing in time, while [43] used the quantiles of an exponential distribution. Our MSE result neatly connects these by recommending the use of quantiles in log-bin space.

Another unifying insight from Theorem 1 is that any parameter entering the log-likelihood in a functionally equivalent way to γ_j in Eq. (4), should be inferred in log-space. This maximises distinguishability in model space, and means that it is best to work in log-migration rates for structured models. While working in log-populations is not uncommon (albeit for heuristic reasons), using the log of the migration matrix is essentially non-existent in the literature. This could potentially improve current structured coalescent inference algorithms. Similarly, for the SMC, this insight suggested that we must trade between absolute bin sizes for inferring log-populations and absolute population sizes for estimating log-bin widths. This symmetry could influence discretisation procedures.

Theorem 1 is also useful for finding cases where robust designs are unachievable. In the skyine demographic model, for example, short intervals containing large populations 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 regions. This hypothesis is corroborated by [13]. A similar effect occurs for SMC models if the bin size is small in a period of large population size [11]. For the structured model, it is expected that the log-population criteria is simpler to achieve than the log-migration rate one since the ability to control $p-1$ stochastic migration event types per deme could be challenging, depending on how close the process is to the strong or weak migration limits [49] [50].

While we have provided directives for robust design, their realisability is not obvious. Existing analyses on this topic [14] [17] [49] [12] tend to be simulation studies that examine a set of reasonable protocols. However, since no optimal reference exists, they can only compare performance within their chosen set. We therefore took an analytical approach, and derived robust designs that could be used by future studies for benchmarking. Since coalescent data is often noisy and uncertain, we examined several coalescent models and optimised on both the total and maximum parameter credibility to ensure that our results were generalisable.

While Theorem 1 is a good first step towards defining optimal experimental benchmarks, there is still much scope for development. Future research would focus on examining the impact of parameter dependence and prior genealogical knowledge, and testing how robust design criteria deteriorate as coalescent models assumptions become invalidated.

References

[1] J. Kingman, On the Genealogy of Large Populations, Journal of Applied Probability 19 (1982) 27–43.

[2] O. Pybus, A. Rambaut, P. Harvey, An Integrated Framework for the Inference of Viral Population History from Reconstructed Genealogies, Genetics 155 (2000) 1429–37.

[3] K. Strimmer, O. Pybus, Exploring the Demographic History of DNA Sequences using the Generalized Skyline Plot, Mol. Biol. Evol 18 (12) (2001) 2298–305.

[4] A. Drummond, A. Rambaut, B. Shapiro, O. Pybus, Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences, Mol. Biol. Evol 22 (5) (2005) 1185–92.

[5] K. Parag, O. Pybus, Optimal Point Process Filtering and Estimation of the Coalescent Process, Journal of Theoretical Biology (2017) 153–67.

[6] T. Vaughan, D. Kuhnert, A. Popinga, et al., Efficient Bayesian Inference under the Structured Coalescent, Bioinformatics 30 (16) (2014) 2272–9.

[7] P. Beerli, J. Felsenstein, Maximum Likelihood Estimation of a Migration Matrix and Effective Population Sizes in n Subpopulations by using a Coalescent Approach, PNAS 98 (8) (2001) 4563–68.

[8] E. Volz, S. Kosakovsky Pond, M. Ward, et al., Phylodynamics of infectious disease epidemics, Genetics 183 (2009) 1421–30.

[9] N. De Maio, C. Wu, K. O'Reilly, D. Wilson, New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation, PLoS Genetics 11 (8) (2015) e1005421.

[10] H. Li, R. Durbin, Inference of Human Population History from Individual Whole-genome Sequences, Nature 475 (7357) (2011) 493–6.

[11] S. Sheehan, K. Harris, Y. Song, Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach, Genetics 194 (2013) 647–62.

[12] J. Palacios, J. Wakeley, S. Ramachandran, Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies, Genetics 201 (2015) 281–304.

[13] L. Gattepaille, G. Torsten, M. Jakobsson, Inferring Past Effective Population Size from Distributions of Coalescent Times, Genetics 204 (2016) 1191–206g.

[14] J. Stack, J. Welch, M. Ferrari, et al., Protocols for Sampling Viral Sequences to Study Epidemic Dynamics, J. R. Soc. Interface 7 (2010) 1119–27.

[15] M. Hall, M. Woolhouse, A. Rambaut, 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).

[16] D. Hillis, Taxonomic Sampling, Phylogenetic Accuracy, and Investigator Bias, Syst. Biol. 47 (1) (1998) 3–8.

[17] M. Karcher, J. Palacios, T. Bedford, et al., Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference, PLoS Computational Biology 12 (3).

[18] J. Kim, M. E, M. Racz, N. Ross, Can one Hear the Shape of a Population History?, Theoretical Population Biology 100 (2015) 26–38.

[19] M. Bartlett, The Use of Transformations, Biometrics 3 (1) (1947) 39–52.

[20] V. Minin, E. Bloomquist, M. Suchard, Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics, Mol. Biol. Evol 25 (7) (2008) 1459–71.

[21] R. Griffiths, S. Tavare, Sampling Theory for Neutral Alleles in a Varying Environment, Phil Trans R Soc B 344 (1994) 403–10.

[22] P. Beerli, J. Felsenstein, Maximum Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations using a Coalescent Approach, Genetics 152 (1999) 763–73.

[23] M. Notohara, The Coalescent and the Genealogical Process in Geographically Structured Population, J Math Biol 29 (1990) 59–75.

[24] R. Fisher, Statistical Methods and Scientific Induction, Edinburgh: Oliver and Boyd, 1956.

[25] E. Lehmann, G. Casella, Theory of Point Estimation, 2nd Edition, Springer-Verlag, 1998.

[26] G. Reinert, Statistical Theory, Tech. rep., University of Oxford (2009).

[27] S. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.

[28] Asymptotic Methods in Statistical Decision Theory, Springer Verlag, New York.

[29] D. Freedman, On the Bernstein-Von Mises Theorem with Infinite Dimensional Parameters, The Annals of Statistics 27 (4) (1999) 1119–40.

[30] A. Atkinson, A. Donev, Optimal Experimental Designs, Oxford University Press, 1992.

[31] H. Banks, M. Davidian, Generalized Sensitivities and Optimal Experimental Design, Tech. rep., North Carolina State University (2009).

[32] A. Marshall, I. Olkin, B. Arnold, Inequalities: Theory of Majorization and its Applications, 2nd Edition, Springer Science + Business Media, 2011.

[33] T. Rothenburg, Identification in Parametric Models, Econometrica 39 (3).

[34] M. Gill, P. Lemey, N. Faria, et al., Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci, Mol. Biol. Evol 30 (3) (2012) 713–24.

[35] M. Friendly, G. Monette, J. Fox, Elliptical insights: Understanding statistical Methods through Elliptical Geometry, Statistical Science 28 (1) (2013) 1–39.
[36] P. Grunwald, The Minimum Description Length Principle, The MIT Press, 2007.
[37] I. Myung, V. Balasubramanian, M. Pitt, Counting Probability Distributions: Differential Geometry and Model Selection, Proceedings of the National Academy of Science 97 (21) (2000) 11170–5.
[38] A. Rodrigo, J. Felsenstein, Coalescent Approaches to HIV-1 Population, The Evolution of HIV, Johns Hopkins University Press, 1999.
[39] D. Snyder, M. Miller, Random Point Procresses in Time and Space, 2nd Edition, Springer-Verlag, 1991.
[40] G. Ewing, G. Nicholls, A. Rodrigo, Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations, Genetics 168 (2004) 2407–20.
[41] M. Nordborg, Handbook of Statistical Genetics: Coalescent Theory, John Wiley and Sons, 2001.
[42] G. McVean, N. Cardin, Approximating the Coalescent with Recombination, Phil Trans R Soc B 360 (2005) 1387–93.
[43] S. Schiffels, R. Durbin, Inferring Human Population Size and Separation History from Multiple Genome Sequences, Nature Genetics 46 (8) (2014) 919–25.
[44] P. Tataru, J. Nirody, Y. Song, diCal-IBD: Demography-Aware Inference of Identity-by-Descent Tracts in Unrelated Individuals, Bioinformatics 30 (23) (2014) 3430–1.
[45] D. Weissman, O. Hallatschek, Minimal-assumption Inference from Population-genomic Data, eLife 6 (2017) e24836.
[46] J. Liepe, S. Filippi, M. Komorowski, et al., Maximizing the Information Content of Experiments in Systems Biology, PLoS Computational Biology 9 (1) (2013) e1002888.
[47] R. Opgen-Rhein, L. Fahrmeir, K. Strimmer, Inference of Demographic History from Genealogical Trees using Reversible Jump Markov Chain Monte Carlo, BMC Evolutionary Biology 5 (6).
[48] B. Ranneby, The Maximum Spacing Method: An Estimation Method Related to the Maximum Likelihood Method, Scandinavian Journal of Statistics 11 (1984) 93–112.
[49] R. Heller, L. Chikhi, H. Siegismund, The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History, PLoS ONE 8 (5) (2013) e62992.
[50] P. Sjodin, I. Kaj, S. Krone, et al., On the Meaning and Existence of an Effective Population Size, Genetics 169 (2005) 1061–70.
[51] R. Cheng, N. Amin, Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin, J. R. Statist. Soc. B 45 (3) (1983) 394–403.

SUPPORTING TEXT

Robust Coalescent Interval Spacing

In the main text we examined how to optimise experimental design for robustness, given p orthogonal parameters that need to be inferred. The log-likelihoods used in these designs have presumed some knowledge about the change-points of these parameters. This corresponds to knowing the piecewise-constant switch times, the deme number and the bin sizes (or population history) in the skyline demographic, structured and SMC models respectively. 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 discretisations [11] [44]. Similar identifiability problems emerge in trying to simultaneously infer interval limits and population sizes or deme number, population size and migration rates. Usually the best one can do in these cases is to use some data driven iterative procedure [12] [47]. These methods will optimise these unknowns in turn and produce sensible results but will be very case specific, allowing no general design insight to be derived.

A main reason for this difficulty is that, often, such low information problems require us to know the coalescent time distribution in order to estimate the parameters of interest, which themselves influence the coalescent times. This circular dependence creates a fundamental limit [11] [13]. While this general problem is outside the scope of our work, here we explore what change-point choices our robust design recommends. Theorem 1 suggests that all of our piecewise coalescent models can be viewed as allocations of events to 'slots' (which can represent bins, switch intervals or even deme demarcations). Given this perspective, we will examine change-points explicitly for the SMC model, but observe that the same results apply to the other models as they all possess analogous log-likelihoods. Note that for this analogy we only consider the deme population sizes as parameters in the structured model. A similar type of result is, however, expected to hold for migration rate inference.

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 identically and independently distributed according to density $f(t) = \frac{\lambda(t)}{\int_0^{\epsilon_p} \lambda(u)\,\mathrm{d}u}$ [39]. 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 $[\epsilon_{j-1},\, \epsilon_j]$ and slot width $\beta_j = \epsilon_j - \epsilon_{j-1}$. Here T is the number of loci. Note that, for example, in the skyline demographic model, we would have a single loci and the β_j would correspond to scaled interval times (see ω_j in the derivation of the skyline demographic log-likelihood in the main text).

We can define the cumulative distribution function (CDF) at the slot endpoints 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, normalised by its total number of coalescent events. Solving for Δ_j using the piecewise-constant coalescent rate gives the left part of Eq. (14). This expression is precisely the same for the skyline and structured models (loci based constants cancel). 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. (14).

$$\Delta_j = \frac{\gamma_j\beta_j}{\sum_{i=1}^{p}\gamma_j\beta_j} \implies \hat{\Delta}_j = \frac{m_j}{n-1} \implies \hat{\Delta}_j^* \,|\, \mathbb{D} = \frac{1}{p} \qquad (14)$$

The robust coalescent interval spacing, $\hat{\Delta}_j^* \,|\, \mathbb{D}$, is therefore fixed by the number of slots (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 slots is such that their endpoints form $\frac{1}{p}$ quantiles of the normalised LTT. Thus, robust coalescent spacing leads to a uniform histogram of coalescent event counts, with p controlling the histogram resolution. Second, since the spacing at the MLE is constant, robustness is achieved by the maximum spacings estimate (MSE) [51] [48]. For a given set of observations, drawn from the CDF of a parameter σ, the MSE is the estimate of σ that maximises the geometric mean of the spacing of the CDF, evaluated at each observed random sample. Our results suggest that if we view the endpoints as binned draws from $f(t)$ then, given a robust design, the MSE of σ results in optimal spacing. Here σ 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. We conjecture such quantile designs may be important for low information change-point problems, and propose MSE theory as a possible direction for future research into optimal binning or change-point problems in coalescent models. Note that under this MSE design, the observed tree, from the (information) perspective of the slots, will appear as uniform as possible. Lastly, we observe that the quantile design clearly suggests that the largest admissible number of change-points occur when $p = n-1$. This limit, for skyline

demographic inference problems, corresponds to the formulation of the classical skyline [2].

Simulation Study: Square Wave Populations

We show how to apply Theorem 1 to a simple skyline demographic coalescent model. Consider a population, $N(t)$, defined by a square wave with period T. $N(t)$ models the harmonic mean [2] 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 $N(t)$. Fig 1 of the main text illustrates the experimental set-up and notation for a similar design problem.

The precision with which N_1 and N_2 are estimated is an increasing function of the number of coalescent events falling within their intervals. Let m_{1i} be the number of events in the i^{th} recurrence of N_1 and m_{2i} be the equivalent for N_2. Theorem 1 stipulates that the robust sampling schemes distribute $\frac{1}{2}$ of all coalescent events to N_1 intervals (Eq. (15)). Thus, if m_1 is the observed count of coalescent events falling within N_1 intervals, the performance of any sampling scheme can be measured by the size of the scalar $d(m_1) := \left|\frac{\mathcal{I}(\log N_1)}{n-1} - \frac{1}{2}\right| = \left|\frac{m_1}{n-1} - \frac{1}{2}\right|$. Note that $d(m_1^*) = 0$ and it grows in size as the Fisher information becomes more skewed (higher $\mathcal{I}(\log N_1)$ means lower $\mathcal{I}(\log N_2)$).

$$\mathcal{I}(\log N_1) = m_1 = \sum_{i\geq 0} m_{1(i+1)} \implies m_1^* = m_2^* = \frac{1}{2}(n-1) \tag{15}$$

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 $p_1^* = \arg\min_{p_1} d(m_1)$. 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 k^{th} sampling event at which ϕ_k lineages are introduced. We will always introduce our ϕ_k samples all at once and only at the change-points so that $s_k = (k-1)\frac{T}{2}$. This procedure maximises the probability that samples will coalesce within their (desired) half-period.

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 $f_1 = \text{round}\,[p_1(n-1)]$ and $f_2 = n - 1 - f_1$. Here round indicates the nearest integer. Note that $\max_{p_j} f_j = n - 1$ as we assume that there is always an initial sample to allow the first coalescent event. We allocate the f_1 and f_2 samples uniformly across N_1 and N_2 half periods respectively, so that $\phi_i = a$ or 0 depending on whether samples are introduced or not. Here $p_1 = 0$ means we have placed all n samples uniformly 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.

Fig. 3 shows the sampling protocol performance under these schemes with $\phi_i = 1$ or 0 at different N_1 values relative to T, and $N_2 = 2N_1$. 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 in this region the population changes are slow relative to the coalescent times so that we have the greatest chance of any sample falling within its desired half period. As N_1 increases, coalescent times lengthen and we get samples falling outside their desired half period. This leads to a weaker, less discernible minimum with larger uncertainty. Further $p_1^* < \frac{1}{2}$ in this regime. If we made $N_2 = \frac{1}{2}N_1$ we would get curves skewed in the opposite direction so that $p_1^* > \frac{1}{2}$. Robust sampling therefore favours placing more samples in regions with larger population size.

This performance is not surprising because we cannot estimate fluctuations in population size that are fast compared to our information carrying events [50]. Note that if we increase ϕ_i then even at these larger population sizes p_1 gets closer to $\frac{1}{2}$ due to the reduced average coalescent times (increased speed of information). Moreover, as biological population changes are usually assumed slower than coalescent events [3], we can safely conclude that square wave robustness is achieved by protocols which assign equal sampling proportions to each population segment.

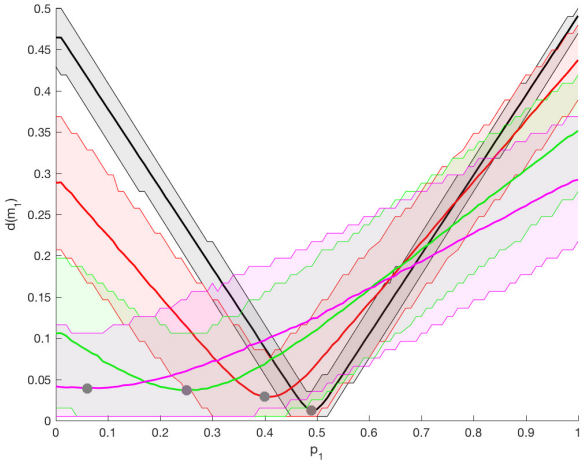

Fig. 3. **Uniform sampling protocols for a square wave population.** We apply a deterministic and uniform sampling strategy with $\phi_i = 1$ or 0 to a square wave population that fluctuates between N_1 and $N_2 = 2N_1$. We observe how the absolute difference between the Fisher information and optimally robust directive, $d(m_1)$, changes with p_1, the probability that a sampled tip lands in N_1. We set $n = 100$ and repeat this simulation 5000 times. The black, red, green and magenta curves 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 the repeated runs (solid line) and the 95% confidence interval around that mean. As N_1 decreases relative to T, $d(m_1)$ becomes more symmetrical 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 visualised with a grey, filled circle.