

# Robust Design for Coalescent Model Inference

KRIS V PARAG<sup>1,\*</sup> AND OLIVER G PYBUS<sup>1</sup>

<sup>1</sup> *Department of Zoology, University of Oxford, Oxford, OX1 3SY, UK*

*\*Correspondence to be sent to: The Peter Medawar Building, Department of Zoology, University of Oxford, Oxford, OX1 3SY, UK; E-mail: kris.parag@zoo.ox.ac.uk*

## ABSTRACT

The coalescent process describes how changes in the size or structure of a population influence the genealogical patterns of sequences sampled from that population. The estimation of (effective) population size changes from genealogies that are reconstructed from these sampled sequences is an important problem in many biological fields. Often, population size is characterised by a piecewise-constant function, with each piece serving as a population size parameter to be estimated. Estimation quality depends on both the statistical coalescent inference method employed, and on the experimental protocol, which controls variables such as the sampling of sequences through time and space, or the transformation of model parameters. While there is an extensive literature on coalescent inference methodology, there is comparatively little work on experimental design. The research that does exist is largely simulation-based, precluding the development of provable or general design theorems. We examine three key design problems: temporal sampling of sequences under the skyline demographic coalescent model, spatio-temporal sampling under the structured coalescent model, and time discretisation for sequentially Markovian coalescent models. In all cases we prove that (i) working in the logarithm of the parameters to be inferred (e.g. population size), and (ii)

distributing informative coalescent events uniformly among these log-parameters, is uniquely robust. ‘Robust’ means that the total and maximum uncertainty of our parameter estimates are minimised, and made insensitive to their unknown (true) values. This robust design theorem provides rigorous justification for several existing coalescent experimental design decisions, and leads to usable guidelines for future empirical or simulation-based investigations. Given its persistence among models, this theorem may form the basis of an experimental design paradigm for coalescent inference.

*Key words:* Coalescent Theory, Population Genetic Inference, Experimental Design, Skyline Models, Structured Coalescent, Sequential Markovian Coalescent.

The coalescent process (Kingman, 1982) is a popular population genetic model that describes how past changes in the size or structure of a population shape the reconstructed (observed) genealogy of genetic sequences sampled from that population. This genealogy is also referred to as a coalescent tree or phylogeny and the population size we mention here is typically known as the effective population size. The estimation of a function that describes past (effective) population size from the sequences, or from a reconstructed phylogeny, is a problem encountered in many fields including epidemiology, conservation and anthropology. Accordingly, there is an extensive and growing literature (Pybus *et al.*, 2000) (Strimmer and Pybus, 2001) (Drummond *et al.*, 2005) (Parag and Pybus, 2017) (Vaughan *et al.*, 2014) (Beerli and Felsenstein, 2001) (Volz *et al.*, 2009) (De Maio *et al.*, 2015) (Li and Durbin, 2011) (Sheehan *et al.*, 2013) (Palacios *et al.*, 2015) (Gattepaille *et al.*, 2016) 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. Variables under the control of the experimenter, such as choices of where and when sequences are sampled, or on how time is

discretised, may have a strong influence on the performance and reliability of coalescent inference methods (Stack *et al.*, 2010) (De Maio *et al.*, 2015) (Sheehan *et al.*, 2013). Good designs can result in surer inferences and sounder conclusions (Stack *et al.*, 2010), whereas bad designs, such as size-biased sampling strategies, can lead to overconfident or spurious estimates (Hall *et al.*, 2016). The best approach to coalescent inference should therefore jointly optimise experimental design and statistical methodology. Surprisingly, only a few studies have investigated optimal coalescent inference design. These works (Stack *et al.*, 2010) (Karcher *et al.*, 2016) (Palacios *et al.*, 2015) (Kim *et al.*, 2015) (Hall *et al.*, 2016), typically take a simulation-based approach, in which several designs are numerically examined and compared. While such studies can yield useful hypotheses about the components of good experimental designs, they can neither provide analytic insights nor criteria for provably optimal design. A more general and formal analysis is therefore warranted.

Additionally, there has been little consideration of what data or parameter transformations might aid experimental design. This contrasts with the development of inference theory in other fields. For example, in regression problems, research has emphasised the benefits of power transformations and regularisation procedures (Box and Cox, 1964). While some coalescent inference methods have used parameter transformations (e.g. the log transform), these are usually justified for heuristic or practical reasons, such as algorithmic stability or ease of visualisation (Palacios *et al.*, 2015) (Minin *et al.*, 2008). As a result, parameter transformations are applied inconsistently in the coalescent literature and rigorous proof of their benefits is lacking. Such proofs could help users navigate between competing methods that solve similar estimation problems, such as the Bayesian skyline plot (which infers absolute population sizes) (Drummond *et al.*, 2005) and the Skyride (which infers log-population sizes) (Minin *et al.*, 2008).

Here we take an analytical approach and formally derive optimal design criteria for coalescent inference. As we are interested in widely applicable theoretical insights, we do

not construct method-specific rules, but instead establish benchmarks which, if achieved, guarantee certain well-defined and desired properties. These benchmarks can then be used to deduce practical experimental design recommendations. We investigate three popular coalescent models, which we define as ‘piecewise’ due to the characteristic functions they infer. For each model we describe a coalescent tree as being composed of sample lineages, with time flowing from the present into the past. A coalescent event occurs when two lineages merge into a single ancestral lineage.

(1) Skyline demographic models. These infer past population size changes using piecewise-constant, time-varying functions (Griffiths and Tavaré, 1994), and usually feature genealogies with many samples from a few (usually one) loci (Pybus *et al.*, 2000) (Gattepaille *et al.*, 2016). The large sample size of these trees means that the choice of sequence sampling times is a critical design variable that can significantly influence the precision of population size inference. Skyline models are popular in infectious disease epidemiology and phylodynamics. In these applications the population describes the number of infected individuals in an epidemic, which may be sampled longitudinally during a single outbreak (e.g Ebola virus in West Africa) or over multiple epidemic seasons (e.g. human influenza viruses). Optimal sampling designs could improve disease surveillance and control strategies (Drummond *et al.*, 2005) (Stack *et al.*, 2010).

(2) Structured coalescent models. Here the population is divided into distinct but connected sub-populations (demes), which typically represent different spatial locations. Usually each deme has a constant (stable) population size. Lineages migrate between demes but only coalesce within demes. The population sizes and migration rates are the parameters of interest (Beerli and Felsenstein, 1999) (Notohara, 1990). The locations and times of sampled sequences, which are our design variables here, are known to bias inference under these models (De Maio *et al.*, 2015). The structured coalescent has been widely applied to describe and investigate the migration history of animal, plant and pathogen populations (De Maio *et al.*, 2015).

(3) Sequentially Markovian coalescent (SMC) models. These are typically applied to complete metazoan genomes, and consider many local coalescent trees (multiple correlated loci), each containing few (or two) samples (McVean and Cardin, 2005). SMC processes involve recombination, and event times are discretised to occur in finite intervals (Li and Durbin, 2011). Past population size change is often assumed to be piecewise-constant. Initially, SMC applications focussed on human demographic history, but more recently non-model organisms have been examined (Beichman *et al.*, 2018). The design variable in this context is the time discretisation, which controls the resolution with which population size changes are estimated. Poor discretisations can lead to runaway behaviour, which results in overestimation (Sheehan *et al.*, 2013).

We examine the above models using optimal design theory, which aims to optimise experimental protocols using statistical criteria that confer useful properties, such as minimum bias or maximum precision (Atkinson and Donev, 1992). As the coalescent event times contain information about population size changes, the distribution and total number of coalescent events controls the amount of information available. Within this context, we treat our sampling/discretisation choice problem as an experimental design on this coalescent event distribution. We show that it is optimal to (i) estimate the logarithm of our parameters of interest, which usually refer to effective population sizes, and (ii) sample (through time and location) or discretise time such that coalescent events are divided evenly among each log scaled parameter. If (i)-(ii) are achieved, then the resulting design is provably robust, and optimal for use with existing maximum likelihood and Bayesian coalescent inference methods.

‘Robust’ means that the maximum dimension and the total volume of the confidence ellipsoid circumscribing (asymptotic) estimate uncertainty are jointly minimised, and, further, are insensitive to the true (unknown) parameter values. Objectives (i)-(ii) hold for all piecewise coalescent models (such as those above) and therefore comprise simple, unifying rules for coalescent inference design. While some of

these criteria are already satisfied or hinted at by existing methods (e.g. some SMC approaches suggest discretising for uniform coalescent event counts (Sheehan *et al.*, 2013)), there has been no consistent or rigorous development of optimal design criteria for these models, and existing recommendations are often just sensible rules-of-thumb. We start by providing mathematical background on optimal design. We use these concepts to derive a robust design theorem for coalescent inference. This theorem is then applied to each of the three coalescent models described above, revealing new and specific insights. We close with a discussion of how our formally-derived principles relate to existing heuristics in the literature, and yield practical, experimental design recommendations.

## MATERIALS AND METHODS

Consider an arbitrary parameter vector  $\psi = [\psi_1, \dots, \psi_p]$ , which is to be estimated from a statistical model. Let  $\mathcal{T}$  represent data (a random variable sequence) generated under this statistical model (the genealogy in the case of coalescent inference) and let  $L(\psi) := \log \mathbb{P}(\mathcal{T} | \psi)$  be the log-likelihood of  $\mathcal{T}$  given  $\psi$ . The  $p \times p$  Fisher information matrix, denoted  $\mathcal{I}(\psi)$ , measures how informative  $\mathcal{T}$  is about  $\psi$  (Fisher, 1956). Since all the coalescent models used here belong to an exponential family (Lehmann and Casella, 1998) (and so satisfy necessary regularity conditions (Reinert, 2009)) then the  $(i, j)^{\text{th}}$  element of  $\mathcal{I}(\psi)$  is  $\mathcal{I}(\psi)_{(i,j)} := -\mathbb{E}_{\mathcal{T}} \left[ \frac{\partial^2 L}{\partial \psi_i \partial \psi_j} \right]$ , with the expectation taken across the data (tree branches). The Fisher information is sensitive to parametrisation choices. Eq. (1) gives the transformation between  $\psi$  and an arbitrary alternate parametrisation  $\sigma = [h(\psi_1), \dots, h(\psi_p)] = [\sigma_1, \dots, \sigma_p]$ . Here  $h$  is a differentiable function, with inverse  $f = \text{inv}[h]$  (Lehmann and Casella, 1998).

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

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 asymptotic bounds are attained so that if  $\hat{\psi}$  is the MLE then  $\text{var}(\hat{\psi}_j) = \text{inv} [\mathcal{I}(\psi)_{(j,j)}]$  is the minimum variance around the MLE of the  $j^{\text{th}}$  parameter that is achievable by any inference method (Kay, 1993). Importantly, for any given parametrisation, the Fisher information serves as a metric that we can use to compare and rank various estimation schemes (e.g. different sampling or discretisation protocols).

Since all of the statistical models that we consider have a finite number of dimensions, the Bernstein-von Mises theorem (Le Cam, 1986) (Freedman, 1999) 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 are applicable to both maximum likelihood and Bayesian approaches to coalescent inference.

We now construct our piecewise coalescent experimental design problem. If the observed data  $\mathcal{T}$  consists of  $n - 1$  coalescent events (e.g. a tree with  $n$  tips) then the set  $\{m_j\}$  for  $1 \leq j \leq p$  with  $\sum_{j=1}^p m_j = n - 1$  describes a coalescent event distribution. Here  $m_j$  counts the coalescent events that are informative of parameter  $\psi_j$ . This is illustrated for a two parameter skyline demographic model in Figure 1, for which  $\psi_j$  is the  $j^{\text{th}}$  effective population size  $N_j$ . Optimal designs are  $\{m_j\}$  sets that satisfy desirable statistical criteria. In the example of Figure 1, these sets would be achieved by appropriate sampling protocol choices. The statistical design criteria are typically functions of  $\mathcal{I}(\psi)$ , which defines our asymptotic uncertainty about  $\hat{\psi}$ .

Geometrically, this uncertainty can be represented as a confidence ellipsoid centred on  $\hat{\psi}$  (Banks and Davidian, 2009). Designing the Fisher information matrix is equivalent to controlling the shape and size of this ellipsoid. We focus on two popular criteria, known as D and E-optimality (Banks and Davidian, 2009) (Atkinson and Donev, 1992), 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. Figure 2 shows these ellipses for the  $p = 2$  skyline design problem of Figure 1.

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

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

160 Here  $\arg$ ,  $\det$  and  $\text{eig}$  are short for argument, determinant, and eigenvalues respectively.

161 D-optimal designs therefore maximise the total available information gained from the set  
 162 of parameters while E-optimal ones ensure that the worst parameter estimate is as good as  
 163 possible (Banks and Davidian, 2009) (Atkinson and Donev, 1992).

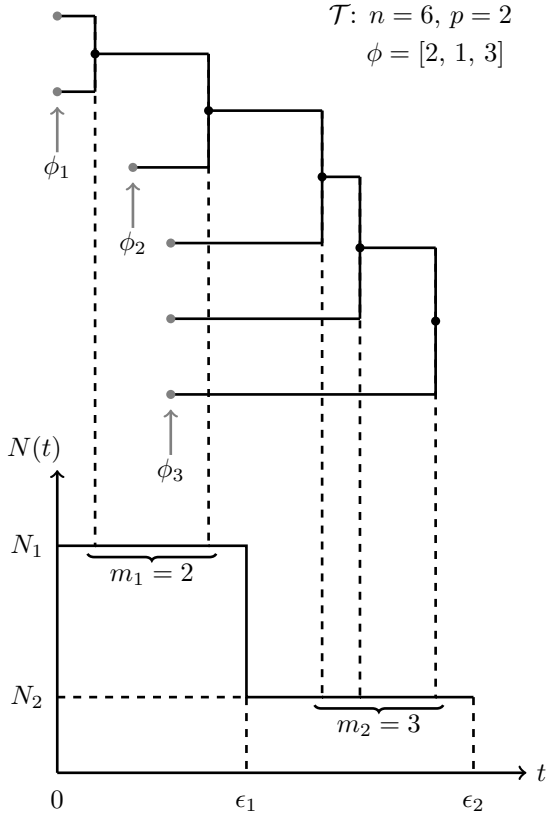


Figure 1: Piecewise coalescent design problem.



The above optimisation problems can be solved using majorization theory, which provides a way of naturally ordering vectors (Marshall *et al.*, 2011). 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  $k \in \{1, 2, \dots, p\}$ ,  $\sum_{j=1}^k \vec{a}^\downarrow_j \geq \sum_{j=1}^k \vec{b}^\downarrow_j$  and  $\sum_{j=1}^p \vec{a} = \sum_{j=1}^p \vec{b} = \kappa$ . Here  $\kappa$  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,  $\vec{a} \succ \vec{b}$  means that the elements of  $\vec{a}$  have the same mean as those of  $\vec{b}$ , but possess a higher variance.

We will also 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  $\vec{u} = [\frac{\kappa}{p}, \frac{\kappa}{p}, \dots, \frac{\kappa}{p}]$  is majorized by any arbitrary vector of sum  $\kappa$  and dimension  $p$  (Marshall *et al.*, 2011). 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 following results.

## RESULTS

### *Naive Coalescent Design*

We define a naive coalescent inference design as one that works directly in the original parametrisation of the model, which is usually effective population size or its inverse. Let  $N = [N_1, \dots, N_p]$  be the population size parameter vector to be estimated from a reconstructed genealogy,  $\mathcal{T}$ . Defining  $\gamma = [N_1^{-1}, \dots, N_p^{-1}]$ , we will find that all three of the coalescent models we consider here have log-likelihoods,  $L(\gamma) = \log \mathbb{P}(\mathcal{T} \mid \gamma)$ , of

the form of Eq. (4). We refer to these models as piecewise.

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

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  $\gamma_j$ ,  $\hat{\gamma}_j = m_j A_j^{-1}$ . 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  $\mathcal{I}(\gamma) = [m_1 \gamma_1^{-2}, \dots, m_p \gamma_p^{-2}] \mathbf{I}_p$ , with  $\mathbf{I}_p$  as a  $p \times p$  identity matrix. Using Eq. (1) we obtain the Fisher information in our original parametrisation as Eq. (5).

$$\mathcal{I}(N) = [m_1 N_1^{-2}, \dots, m_p N_p^{-2}] \mathbf{I}_p \quad (5)$$

Several points become 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 estimation confidence will rapidly deteriorate as  $N_j$  grows. Second, if our inference method directly worked in  $\gamma$ , instead of  $N$  (which is not uncommon for harmonic mean estimators (Pybus *et al.*, 2000)), then the region of parameter space in which we achieve good  $\gamma$  precision is exactly that in which we obtain poor  $N$  confidence. Third, the design variable  $m_j$  only informs on one parameter of interest,  $N_j$  or  $\gamma_j$ . Good designs must therefore 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 (Rothenburg, 1971), which can lead to issues like poor algorithmic convergence. This is particularly relevant for coalescent inference methods that feature pre-defined parameter grids of a size comparable to the tree size  $n$  (Gill *et al.*, 2012).

Using either the  $N$  or  $\gamma$  parametrisation creates issues even when optimal design is employed. 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 given by the first

equality in Eq. (6) below.

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

The E-optimal design solves:  $\max_{\{m_j\}} \min_j m_j N_j^{-2}$ . The objective function is now  $g(\{m_j\}) = \min(m_1 N_1^{-2}, \dots, m_p N_p^{-2})$  and is also Schur concave. The E-optimal solution satisfies  $m_1^* N_1^{-2} = m_2^* N_2^{-2} = \dots = m_p^* N_p^{-2}$  (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  $\gamma_j$  in Eq. (6) above.

Thus, in theory, D-optimal designs that consider  $N$  or  $\gamma$  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 non-identifiability. Additionally, for a given criterion, optimal  $N_j$  and  $\gamma_j$  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  $N$  parametrisation, for the model shown in Figure 1. These ellipsoids, for some parameter vector  $\sigma$ , 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  $\Omega$  controls the confidence significance level according to a  $\chi^2$  distribution (with  $p$  degrees of freedom) and  $x_j$  is some coordinate on the  $j^{\text{th}}$  parameter axis (Friendly *et al.*, 2013). Under the  $N$  parametrisation, 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 (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

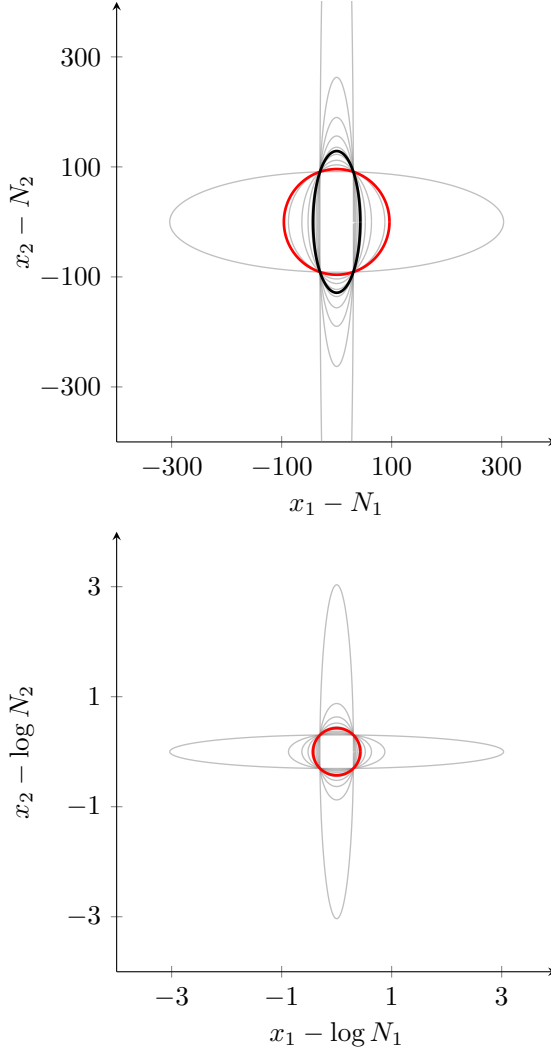


Figure 2: D and E-optimal piecewise coalescent designs.

two-point theorem.

**Theorem 1.** If the  $p$ -parameter vector  $\sigma$  admits a diagonal Fisher information matrix,  $\mathcal{I}(\sigma) = [m_1\sigma_1^{-2}, \dots, m_p\sigma_p^{-2}] \mathbf{I}_p$ , under an isoperimetric constraint  $\sum_{j=1}^p m_j = \kappa$ , then any design that (i) works in the parametrisation  $[\log \sigma_1, \dots, \log \sigma_p]$  and (ii) achieves the distribution  $m_1^* = \dots = m_p^* = \frac{1}{p}\kappa$  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 a property that determines parameter

identifiability and model complexity (Grunwald, 2007). Assume that  $\psi$  is the true parameter vector underlying some observed tree  $\mathcal{T}$ , and that  $\psi$  lies in some parameter space,  $\Psi$ , 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  $\psi$  from  $\mathcal{T}$  with some statistical confidence. This confidence can be visualised as an ellipse around  $\psi$ . All parameter vectors that map to this ellipse are statistically indistinguishable from  $\psi$ . 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  $\Psi$ , and will shrink in size but increase in number as  $n$  increases (more data improves estimation certainty). Thus, distinguishability is intrinsically linked to the quality of inference. More detail on these information geometric concepts is given in (Myung *et al.*, 2000) (Grunwald, 2007).

We can define a volume,  $\mathcal{V} := \int_{\Psi} \det \left[ \frac{1}{n-1} \mathcal{I}(\psi) \right]^{\frac{1}{2}} d\psi$  to measure the total size of these ellipses over  $\Psi$  (Grunwald, 2007). This volume is related to the complexity of our coalescent model and hence is unchanged by parametrisation (Grunwald, 2007). While  $\mathcal{V}$  is invariant to parametrisation choice  $h$ , different  $h$  functions control how the parameter space is discretised into distinguishable ellipses (Myung *et al.*, 2000). For example, under  $\psi = \sigma$  poor distinguishability results when any  $\sigma_j$  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 maximises overall parameter distinguishability in  $\Psi$ , or equivalently minimises the sensitivity of our estimates to the unknown true values of our parameters,  $\psi$ . Geometrically, this transformation yields the smallest ellipse size that is also independent of the location of  $\psi$  in  $\Psi$ .

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  $\psi_j$  only depends on  $\sigma_j$ . 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).

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

This equality defines the conserved property across parametrisations of coalescent models with likelihoods given by Eq. (4). 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  $\psi_j$ . 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 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).

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

Let  $\mathbb{D}$  be an optimal design criterion, with event distribution  $\{m_j^* | \mathbb{D}\}$ . When  $\mathbb{D} \equiv \mathbf{D}$ , 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  $\mathbb{D} \equiv \mathbf{E}$ , we maximise  $\min \text{eig}[\mathcal{I}(\log N)]$  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-parametrisation 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 (ii). 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  $N(t) := \sum_{j=1}^p N_j 1(\epsilon_{j-1} \leq t < \epsilon_j)$  with  $\epsilon_0 = 0$  and  $\epsilon_p = \infty$ .  $N_j$  is the constant population size of the  $j^{\text{th}}$  segment (or interval) which is delimited by times  $[\epsilon_{j-1}, \epsilon_j)$ . The indicator function  $1(a) = 1$  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  $(s_k, \phi_k)$  defines a sampling protocol in which  $\phi_k$  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  $i^{\text{th}}$  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  $k^{\text{th}}$  sample time  $l(t)$  increases by  $\phi_k$ , and at every  $c_i$  it decreases by 1. The rate of producing coalescent events is defined as:  $\lambda(t) = \sum_{j=1}^p \gamma_j \alpha(t) 1(\epsilon_{j-1} \leq t < \epsilon_j)$  with  $\gamma_j = N_j^{-1}$  as the inverse population in segment  $j$ . We initially work in  $\gamma = [\gamma_1, \dots, \gamma_p]$ , and then transform to  $N = [N_1, \dots, N_p]$ .

The log-likelihood follows from Poisson process theory as (Snyder and Miller, 1991) (Parag and Pybus, 2018):  $L(\gamma) = -\int_0^{c_{n-1}} \lambda(t) dt + \sum_{i=1}^{n-1} \log \lambda(c_i)$ , with  $L(\gamma) = \log \mathbb{P}(\mathcal{T} | \gamma)$ . By splitting the integral across the  $p$  piecewise-constant segments we get that:

$\int_0^{c_{n-1}} \lambda(t) dt = \sum_{j=1}^p \gamma_j \int_{\epsilon_{j-1}}^{\epsilon_j} \alpha(t) dt = \sum_{j=1}^p \gamma_j \omega_j$ . Here  $\omega_j$  is a constant for a given tree, and it is independent of  $\gamma$ . Similarly,  $\sum_{i=1}^{n-1} \log \lambda(c_i) = \sum_{j=1}^p \sum_{i=1}^{n-1} \log(\gamma_j \alpha(c_i) 1(\epsilon_{j-1} \leq c_i \leq \epsilon_j))$ . Expanding yields Eq. (9) with  $\Gamma_j$  as a constant depending on  $\alpha(c_i)$  for all  $i$  falling in the  $j^{\text{th}}$  segment. The count of all the coalescent events within  $[\epsilon_{j-1}, \epsilon_j]$  is  $m_j$ .

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

Eq. (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  $\epsilon_j$  are all non-informative 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  $i^{\text{th}}$  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 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  $i$  as  $\zeta_{ji}$ .  $N_j$  and  $\gamma_j = N_j^{-1}$  are the absolute and inverse population size in deme  $j$ .

Our initial  $p^2$  vector of parameters is  $\sigma = [\gamma_1, \dots, \gamma_p, \{\zeta_{1\bar{1}}\}, \dots, \{\zeta_{p\bar{p}}\}] = [\gamma, \zeta]$ , with  $\{\zeta_{k\bar{k}}\} = [\zeta_{k1}, \zeta_{k2}, \dots]$  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 (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  $j^{\text{th}}$  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  $\alpha_{ji} := \binom{l_{ji}}{2}$  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  $\sum_{j=1}^p m_j = n-1$ .

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

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

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

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,  $\alpha_{ji}$ , 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  $\mathcal{I}(\psi) = [m_1, \dots, m_p, \{w_{1\bar{1}}\}, \dots, \{w_{p\bar{p}}\}] \mathbf{I}_{p^2}$  when  $\psi = [\log N_1, \dots, \log N_p, \{\log \zeta_{1\bar{1}}\}, \dots, \{\log \zeta_{p\bar{p}}\}]$ . The robust design under this  $\psi$ , given in Eq. (12), involves distributing coalescent and migration events uniformly among the demes. Note that the migration rate distribution,  $w_{ji}^* | \mathbb{D}$ , 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  $M$ .

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

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 generalise 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.

### 344 *Sequentially Markovian Coalescent Models*

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

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

368 Our experimental design problem involves finding an optimal criterion for choosing  
 369 these discretisation times. Currently, only heuristic strategies exist (e.g. choosing bins so  
 370 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) (Gattepaille *et al.*, 2016) (Palacios *et al.*, 2015). We define a vector of bins  $\beta = [\beta_1, \dots, \beta_p]$  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 multi-lineage 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 coalescent events fallin within bin  $\beta_j$  from the  $i^{\text{th}}$  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  $\beta_j$ . As before, we constrain the total number of coalescent events so that  $\sum_{j=1}^p m_j = n - 1$ . 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  $i^{\text{th}}$  locus as  $L_i(\gamma, \beta) = \log \mathbb{P}(\mathcal{T}_i | \gamma, \beta) = - \int_0^\infty \lambda(t) dt + \sum_{j=1}^p m_{ij} \log \left( \int_{\epsilon_{j-1}}^{\epsilon_j} \lambda(t) dt \right)$  (Snyder and Miller, 1991). Here  $\lambda(t)$  is the coalescent rate at  $t$  so that  $\lambda(t) = \sum_{j=1}^p \gamma_j 1(\epsilon_{j-1} \leq t < \epsilon_j)$  and  $\int_{\epsilon_{j-1}}^{\epsilon_j} \lambda(t) dt = \beta_j \gamma_j$  with  $\gamma_j = N_j^{-1}$ . Solving this yields Eq. (13), which is analogous to Eq. (4) and has Fisher information matrix  $\mathcal{I}(N)_{\mathcal{T}_i} = [m_{i1}, \dots, m_{ip}] \mathbf{I}_p$ . Note that neither the waiting time until a recombination nor the time between recombination and coalescence contribute to the Fisher information

(exponential memoryless property).

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

$$\mathcal{I}(N)_{\{\mathcal{T}_i\}} = \sum_{i=1}^T \mathcal{I}(N)_{\mathcal{T}_i | \mathcal{T}_{i-1}} = [m_1 N_1^{-2}, \dots, m_p N_p^{-2}] \mathbf{I}_p \quad (14)$$

Eq. (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 discretisation 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  $\mathcal{I}(N)_{\{\mathcal{T}_i\}} = \sum_{i=1}^T \mathcal{I}(N)_{\mathcal{T}_i | \mathcal{T}_{i-1}, \dots, \mathcal{T}_1}$  (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 (expected) 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  $j^{\text{th}}$  element of  $\mathcal{I}(N)_{\mathcal{T}_i | \mathcal{T}_{i-1}}$  is non-zero, 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).

If we calculate the total Fisher information with respect to  $\beta$  we obtain identical expressions to Eq. (14) with the  $N_j$  simply replaced by  $\beta_j$ . The square dependence of these Fisher matrices means that Theorem 1 applies. We therefore find that it is optimal to infer log-bin sizes ( $\psi = [\log \beta_1, \dots, \log \beta_p]$ ), if population size history is known (this corresponds to the discretisation results presented in (Tataru *et al.*, 2014)), or log-population sizes ( $\psi = [\log N_1, \dots, \log N_p]$ ), 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

parametrisation, the provably robust design recommendation is to discretise time such that the resulting bins contain equal numbers of coalescent events.

These 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 parametrisation for piecewise coalescent estimation problems. While methods using  $\log N$  do exist (Palacios *et al.*, 2015) (Minin *et al.*, 2008), the stated reasons for doing so are centred around algorithmic convenience. Here we provide sound theoretical backing for using  $\log N$  in coalescent inference. The second point of Theorem 1 requires equalising 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 discretisations 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-discretisation bins. This perspective reveals a straightforward rule for statistical identifiability: any piecewise coalescent model with at least one empty pigeon-hole is non-identifiable (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 (Palacios *et al.*, 2015) (Opgen-Rhein *et al.*, 2005). 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 normalised 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.

Realisation of this procedure would be straightforward in existing software, such as BEAST 1 and 2, 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 pre-assigned time discretisation (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 discretisations 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 discretisation via the MSE approach. Recent methods, which decouple discretisation from the demographic history, could eventually allow this flexibility (Steinrucken *et al.*, 2015). Second, while pre-assigned discretisations 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  $\gamma_j$  in Eq. (4), should be inferred in log-space. This maximises 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.

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) (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 (Heller *et al.*, 2013) (Sjodin *et al.*, 2005).

While we have provided universal, robust coalescent design objectives here, we have not explored what specific sampling or discretisation 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) (Karcher *et al.*, 2016) (Heller *et al.*, 2013) (Palacios *et al.*, 2015) 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

Supplementary material can be found in the Dryad data repository at  
<http://datadryad.org>, <http://dx.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.

## LITERATURE CITED

- Atkinson, A. and Donev, A. (1992). *Optimal Experimental Designs*. Oxford University Press.
- Banks, H. and Davidian, M. (2009). Generalized Sensitivities and Optimal Experimental Design. Technical report, North Carolina State University.
- Beerli, P. and Felsenstein, J. (1999). Maximum Likelihood Estimation of Migration Rates and Effective Population Numbers in Two Populations using a Coalescent Approach. *Genetics*, **152**, 763–73.
- Beerli, P. and Felsenstein, J. (2001). Maximum Likelihood Estimation of a Migration Matrix and Effective Population Sizes in n Subpopulations by using a Coalescent Approach. *PNAS*, **98**(8), 4563–68.
- Beichman, A., Huerta-Sanchez, E., and Lohmueller, K. (2018). Using Genomic Data to Infer Historic Population Dynamics of Nonmodel Organisms. *Annu. Rev. Ecol. Evol. Syst.*, **49**, 433–56.
- Bhaskar, A. and Song, Y. (2014). Descartes’ Rule of Signs and the Identifiability of Population Demographic Models from Genomic Variation Data. *Ann. Stats.*, **42**(6), 2463–93.
- Box, G. and Cox, D. (1964). An Analysis of Transformations. *J. R. Statist. Soc. B*, **26**(2).
- Cheng, R. and Amin, N. (1983). Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin. *J. R. Statist. Soc. B*, **45**(3), 394–403.
- De Maio, N., Wu, C., O’Reilly, K., and Wilson, D. (2015). New Routes to Phylogeography: A Bayesian Structured Coalescent Approximation. *PLoS Genetics*, **11**(8), e1005421.

- Drummond, A., Rambaut, A., Shapiro, B., and Pybus, O. (2005). Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *Mol. Biol. Evol.*, **22**(5), 1185–92.
- Ewing, G., Nicholls, G., and Rodrigo, A. (2004). Using Temporally Spaced Sequences to Simultaneously Estimate Migration Rates, Mutation Rate and Population Sizes in Measurably Evolving Populations. *Genetics*, **168**, 2407–20.
- 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. Stats*, **27**(4), 1119–40.
- Friendly, M., Monette, G., and Fox, J. (2013). Elliptical insights: Understanding Statistical Methods through Elliptical Geometry. *Stats. Sci.*, **28**(1), 1–39.
- Gattepaille, L., Torsten, G., and Jakobsson, M. (2016). Inferring Past Effective Population Size from Distributions of Coalescent Times. *Genetics*, **204**, 1191–206g.
- Gill, M., Lemey, P., Faria, N., Rambaut, A., Shapiro, B., and Suchard, M. (2012). Improving Bayesian Population Dynamics Inference: A Coalescent-Based Model for Multiple Loci. *Mol. Biol. Evol.*, **30**(3), 713–24.
- Griffiths, R. and Tavaré, S. (1994). Sampling Theory for Neutral Alleles in a Varying Environment. *Phil. Trans. R. Soc. B*, **344**, 403–10.
- Grunwald, P. (2007). *The Minimum Description Length Principle*. The MIT Press.
- Hall, M., Woolhouse, M., and 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).
- Heller, R., Chikhi, L., and Siegmund, H. (2013). The Confounding Effect of Population Structure on Bayesian Skyline Plot Inferences of Demographic History. *PLoS ONE*, **8**(5), e62992.
- Karcher, M., Palacios, J., Bedford, T., Suchard, M., and Minin, V. (2016). Quantifying and Mitigating the Effect of Preferential Sampling on Phylodynamic Inference. *PLoS Comp. Bio.*, **12**(3).
- Kay, S. (1993). *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall.
- Kim, J., E, M., Racz, M., and Ross, N. (2015). Can one Hear the Shape of a Population History? *Theo. Pop. Bio.*, **100**, 26–38.
- Kingman, J. (1982). On the Genealogy of Large Populations. *J. App. Prob.*, **19**, 27–43.
- Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer Verlag, New York.
- Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation*. Springer-Verlag, second edition.
- Li, H. and Durbin, R. (2011). Inference of Human Population History from Individual Whole-genome Sequences. *Nature*, **475**(7357), 493–6.
- Liepe, J., Filippi, S., Komorowski, M., and Stumpf, M. (2013). Maximizing the Information Content of Experiments in Systems Biology. *PLoS Comp. Bio.*, **9**(1), e1002888.
- Liu, X. and Fu, Y. (2015). Exploring Population Size Changes using SNP Frequency Spectra. *Nat. Gen.*, **47**(5), 555–62.
- Marshall, A., Olkin, I., and Arnold, B. (2011). *Inequalities: Theory of Majorization and its Applications*. Springer Science + Business Media, second edition.
- McVean, G. and Cardin, N. (2005). Approximating the Coalescent with Recombination. *Phil. Trans. R. Soc. B*, **360**, 1387–93.
- Minin, V., Bloomquist, E., and Suchard, M. (2008). Smooth Skyride through a Rough Skyline: Bayesian Coalescent-Based Inference of Population Dynamics. *Mol. Biol. Evol.*, **25**(7), 1459–71.
- Myung, I., Balasubramanian, V., and Pitt, M. (2000). Counting Probability Distributions: Differential Geometry and Model Selection. *PNAS*, **97**(21), 11170–5.
- Nordborg, M. (2001). *Handbook of Statistical Genetics: Coalescent Theory*. 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., and Strimmer, K. (2005). Inference of Demographic History from Genealogical Trees using Reversible Jump Markov Chain Monte Carlo. *BMC Evol. Bio.*, **5**(6).
- Palacios, J., Wakeley, J., and Ramachandran, S. (2015). Bayesian Nonparametric Inference of Population Size Changes from Sequential Genealogies. *Genetics*, **201**, 281–304.

- 613 Parag, K. and Pybus, O. (2017). Optimal Point Process Filtering and Estimation of the Coalescent Process. *J. Theo. Biol.*, **421**,  
614 153–67.
- 615 Parag, K. and Pybus, O. (2018). Exact Bayesian Inference for Phylogenetic Birth-Death Models. *Bioinformatics*, **34**, 3638–45.
- 616 Pybus, O., Rambaut, A., and Harvey, P. (2000). An Integrated Framework for the Inference of Viral Population History from  
617 Reconstructed Genealogies. *Genetics*, **155**, 1429–37.
- 618 Ranneby, B. (1984). The Maximum Spacing Method: An Estimation Method Related to the Maximum Likelihood Method. *Scand. J.*  
619 *Stats*, **11**, 93–112.
- 620 Reinert, G. (2009). Statistical Theory. Technical report, University of Oxford.
- 621 Rothenburg, T. (1971). Identification in Parametric Models. *Econometrica*, **39**(3).
- 622 Schiffels, S. and Durbin, R. (2014). Inferring Human Population Size and Separation History from Multiple Genome Sequences.  
623 *Nature Genetics*, **46**(8), 919–25.
- 624 Sheehan, S., Harris, K., and Song, Y. (2013). Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially  
625 Markov Conditional Sampling Distribution Approach. *Genetics*, **194**, 647–62.
- 626 Sjodin, P., Kaj, I., Krone, S., Lascoux, M., and Nordborg, M. (2005). On the Meaning and Existence of an Effective Population Size.  
627 *Genetics*, **169**, 1061–70.
- 628 Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer-Verlag, second edition.
- 629 Spence, J., Steinrucken, M., Terhorst, J., and Song, Y. (2018). Inference of Population History using Coalescent HMMa: Review and  
630 Outlook. *Curr. Op. Gen. Dev.*, **53**, 70–6.
- 631 Stack, J., Welch, J., Ferrari, M., Shapiro, B., and Grenfell, B. (2010). Protocols for Sampling Viral Sequences to Study Epidemic  
632 Dynamics. *J. R. Soc. Interface*, **7**, 1119–27.
- 633 Steinrucken, M., Kamm, J., and Song, Y. (2015). Inference of Complex Population Histories using Whole-Genome Sequences from  
634 Multiple Populations. *BioRxiv*, page 026591.
- 635 Strimmer, K. and Pybus, O. (2001). Exploring the Demographic History of DNA Sequences using the Generalized Skyline Plot. *Mol.*  
636 *Biol. Evol.*, **18**(12), 2298–305.
- 637 Tataru, P., Nirody, J., and Song, Y. (2014). diCal-IBD: Demography-Aware Inference of Identity-by-Descent Tracts in Unrelated  
638 Individuals. *Bioinformatics*, **30**(23), 3430–1.
- 639 Vaughan, T., Kuhnert, D., Poppinga, A., Welch, D., and Drummond, A. (2014). Efficient Bayesian Inference under the Structured  
640 Coalescent. *Bioinformatics*, **30**(16), 2272–9.
- 641 Volz, E., Kosakovsky Pond, S., Ward, M., Leigh Brown, A., and Frost, S. (2009). Phylodynamics of infectious disease epidemics.  
642 *Genetics*, **183**, 1421–30.
- 643 Weissman, D. and Hallatschek, O. (2017). Minimal-assumption Inference from Population-genomic Data. *eLife*, **6**, e24836.
- 644 Zamir, R. (1998). A Proof of the Fisher Information Inequality via a Data Processing Argument. *IEEE Transactions on*  
645 *Information Theory*, **44**(3), 1246–50.
- 646 Zegers, P. (2015). Fisher Information Properties. *Entropy*, **17**, 4918–39.

## FIGURE CAPTIONS

- Figure 1: Piecewise coalescent design problem. We present 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  $\phi_k$  counting the samples  
introduced at the  $k^{\text{th}}$  sample time. The  $j^{\text{th}}$  population size parameter,  $N_j$ , is only

informed by the number of coalescent events,  $m_j$ , occurring within its duration  $[\epsilon_{j-1}, \epsilon_j]$  (with  $\epsilon_0 = 0$ ). We manipulate  $\phi$  to achieve  $m_1$  and  $m_2$  counts that guarantee desirable properties for estimates of  $N_1$  and  $N_2$ .

- Figure 2: D and E-optimal piecewise coalescent designs. 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 bivariate normal distribution that has covariance matrix equal to the inverse of the Fisher information. Each thin, light grey ellipse indicates a different  $\{m_1, m_2\}$  distribution. D and E-optimal designs are in thick red and black respectively. The top panel shows the design space in absolute population size,  $N_j$ , with  $m_1^* | D = 50$  and  $m_1^* | E = 20$ . The bottom panel is in log population size,  $\log N_j$ , and leads to a symmetrical, robust design that has coincident D and E-optimal ellipses with  $m_1^* | \mathbb{D} = 50$ .
- 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 panel a) for  $N_1 = \frac{T}{4}$  and  $N_2 = \frac{T}{2}$ . The arrows in this sub-plot 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 panel c), for  $n = 100$ . 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 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 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. See the Appendix for further interpretations of these results.

## 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 discretisations (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 (Palacios *et al.*, 2015)

(Opgen-Rhein *et al.*, 2005). These methods will typically jointly optimise 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  $f(t) = \frac{\lambda(t)}{\int_0^{\epsilon_p} \lambda(u) du}$  (Snyder and Miller, 1991). If we let  $\lambda(t)$  be our piecewise-constant SMC rate we find that  $\int_0^{\epsilon_p} \lambda(u) du = \sum_{i=1}^T \sum_{j=1}^p \gamma_j \beta_j = \sum_{j=1}^p (n - 1) \gamma_j \beta_j$ , 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  $\epsilon_j$  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  $\beta_j$  would correspond to scaled interval times (see  $\omega_j$  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:  $F(\epsilon_j) = \int_0^{\epsilon_j} f(t) dt$  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  $\Delta_j$  using the piecewise-constant coalescent rate gives the left part of Eq. (A1). This expression is precisely the same for the skyline and structured models. If we substitute the MLE for either  $\beta_j$  or  $\gamma_j$  (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).

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

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 normalised 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  $\theta$ , the MSE is the estimate of  $\theta$  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 pigeon-hole change-points as binned draws from  $f(t)$  then, given a robust design, the MSE of  $\theta$  results in optimal spacing. Here  $\theta$  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. Panel a) of Figure A1 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  $i^{\text{th}}$  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  $d(m_1^* | \mathbb{D}) = 0$ .

$$\mathcal{I}(\log N_1) = m_1 = \sum_{i \geq 0} m_{1(i+1)} \implies m_1^* | \mathbb{D} = m_2^* | \mathbb{D} = \frac{1}{2}(n-1) \quad (\text{A2})$$

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  $p_1^* = \arg \min_{p_1} d(m_1)$ . We assume  $p_1$  is constant with time. Thus, we focus on the mapping  $p_1 \rightarrow d(m_1)$  with  $p_2 = 1 - p_1$ . A sampling protocol involves the tuple  $(s_k, \phi_k)$  with  $s_k$  as the time of the  $k^{\text{th}}$  sampling event at which  $\phi_k$  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  $\phi_k$  samples all at once, and only at the change-points so that  $s_k = (k-1)\frac{T}{2}$  (the arrows in panel a) of Fig. A1). This procedure maximises 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 panel b) of Figure A1.

Panel c) of Figure A1 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, symmetrical 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.

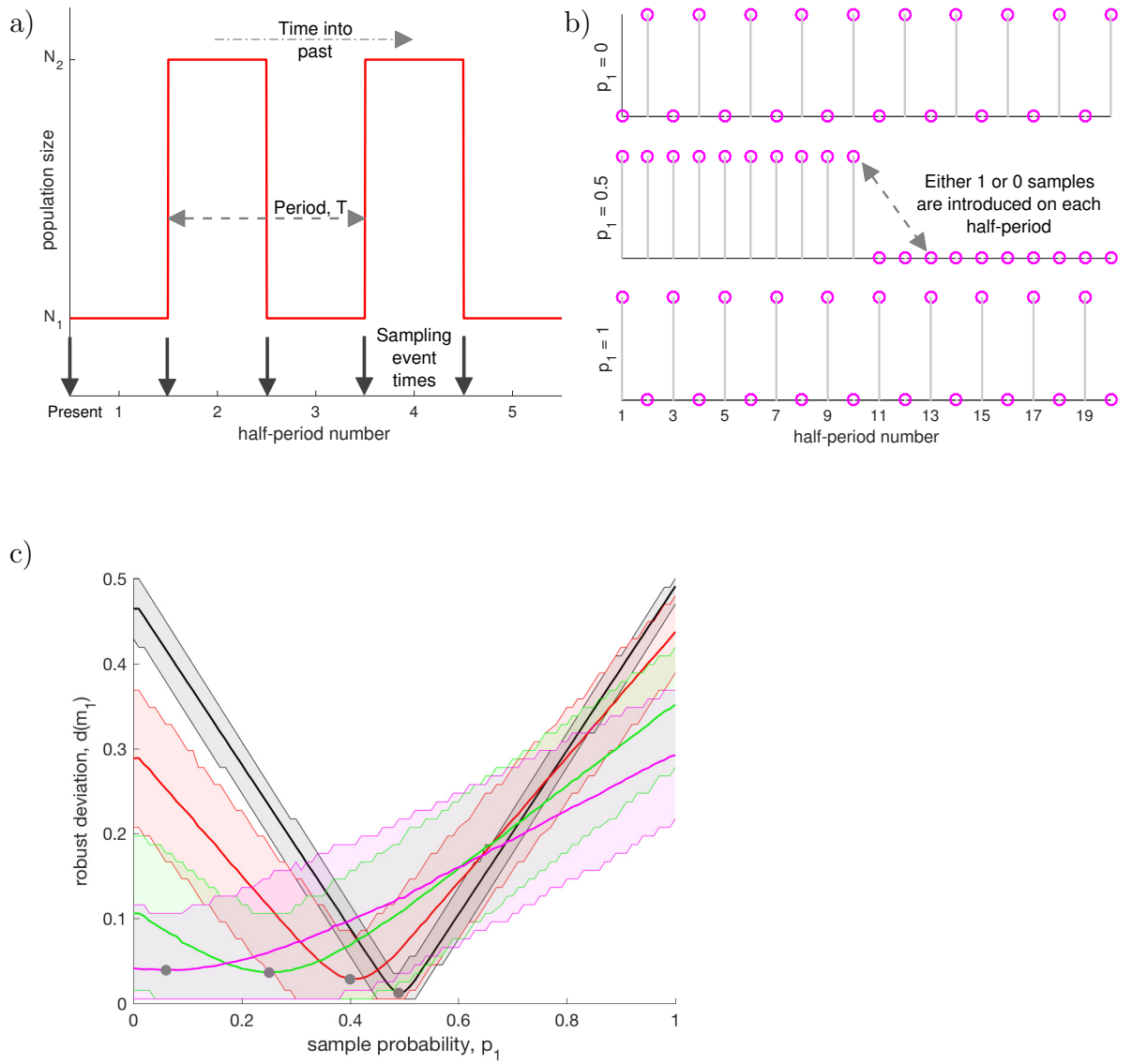


Figure A1: Deterministic sampling protocols for a skyline coalescent model.