Received: 28 May 2019 | Revised: 11 October 2019 | Accepted: 12 November 2019

DOI: 10.1002/ece3.5888

REVIEW ARTICLE

Ecology and Evolution Open Access WILEY

A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data

Niklas Mather | Samuel M. Traves | Simon Y. W. Ho

School of Life and Environmental Sciences, University of Sydney, Sydney, NSW, Australia

Correspondence
Niklas Mather, School of Life and Environmental Sciences, University of Sydney, Sydney, NSW 2006, Australia.
Email: niksmather@gmail.com

Funding information
Australian Research Council, Grant/Award Number: FT160100167

Abstract

A common goal of population genomics and molecular ecology is to reconstruct the demographic history of a species of interest. A pair of powerful tools based on the sequentially Markovian coalescent have been developed to infer past population sizes using genome sequences. These methods are most useful when sequences are available for only a limited number of genomes and when the aim is to study ancient demographic events. The results of these analyses can be difficult to interpret accurately, because doing so requires some understanding of their theoretical basis and of their sensitivity to confounding factors. In this practical review, we explain some of the key concepts underpinning the pairwise and multiple sequentially Markovian coalescent methods (PSMC and MSMC, respectively). We relate these concepts to the use and interpretation of these methods, and we explain how the choice of different parameter values by the user can affect the accuracy and precision of the inferences. Based on our survey of 100 PSMC studies and 30 MSMC studies, we describe how the two methods are used in practice. Readers of this article will become familiar with the principles, practice, and interpretation of the sequentially Markovian coalescent for inferring demographic history.

KEYWORDS
coalescent, demographic history, mutation rate, population genomics, population size

1 | INTRODUCTION

The genomes of organisms carry a record of the evolutionary and ecological forces that have shaped their populations. In principle, one can reconstruct the demographic history of a species from the genome sequences of its present-day representatives (Beichman, Huerta-Sanchez, & Lohmueller, 2018). These reconstructions can be used to answer various biological questions, such as the influence of climatic events on population size and structure (Miller et al., 2012), the timing of major events in the evolutionary history of modern humans (Fu et al., 2014; Li & Durbin, 2011), human impacts on wild animal populations (Johnson et al., 2018; Pujolar, Dalén, Hansen, & Madsen, 2017), and the effects of domestication (Frantz et al., 2016; Yu et al., 2018).

Demographic inference is rarely straightforward; there are many challenges in extracting the relevant historical signal from the population of interest. A number of methods have been developed for this purpose, such as those based on site frequency spectra and approximate Bayesian computation (reviewed by Beichman et al., 2018). Here, we focus on a pair of methods designed to analyze genome

sequences from small samples of individuals: the pairwise sequentially Markovian coalescent (PSMC; Li & Durbin, 2011) and multiple sequentially Markovian coalescent (MSMC; Schiffels & Durbin, 2014). We focus on these methods because they are widely used, they share a common theoretical framework, and a large amount of work has been done on elucidating their strengths and weaknesses. Even as new methods of demographic inference are published, interpreting the large number of studies that have already used PSMC and MSMC is difficult without a solid understanding of the assumptions and approximations made by their underlying models.

The PSMC method can be used to analyze unphased sequence data from a single diploid individual, whereas MSMC can use sequences from several individuals. The two methods are particularly useful when studying deeper population timescales and when there are very limited numbers of samples (Beichman et al., 2018; Spence, Steinrücken, Terhorst, & Song, 2018). For example, PSMC has been employed to great effect in studies of individual ancient samples, including those of an ancient horse (Orlando et al., 2013), an ancient wolf (Skoglund, Ersmark, Palkopoulou, & Dalén, 2015), and two woolly mammoths (Palkopoulou et al., 2015).

In addition to reconstructing demographic history, PSMC and MSMC have been used to infer the timing of population divergence and to estimate mutation rates from ancient genomes. However, various studies have demonstrated that inferences from the two methods can be sensitive to violations of the underlying assumptions (e.g., Mazet, Rodríguez, Grusea, Boitard, & Chikhi, 2016), and the methods do not offer an explicit framework for testing hypotheses. As a consequence, the user needs some appreciation of the underlying biological theory and statistical methods to ensure that the results are interpreted appropriately. In this review, we give a brief description of coalescent theory, the mathematical framework behind demographic inference, before discussing how it is applied in PSMC and MSMC. We describe some of the key practical issues that arise when these two methods are applied to genomic data.

1.1 | Coalescent theory

Coalescent theory provides a statistical framework that relates the size of a population to the coalescence times in the genealogy of the sampled individuals (Hudson, 1983; Kingman, 1982; Tajima, 1983). Coalescent models envision the evolutionary process running backwards in time: We start from the leaves of the genealogy, which represent the individuals that have been sampled in the data set. We then trace out their full evolutionary history by following their lineages back in time. Coalescence events occur whenever two lineages combine to become one ancestral lineage. If the population is panmictic (random mating), then all possible pairs of lineages have an equal probability of coalescing.

The rate of coalescence can tell us about population size because coalescence events are more likely to occur when the population is small. For example, if we select a few people at random from a small, isolated village, they are likely to share an ancestor in recent generations. If a few people are chosen at random from the entire human population, they are unlikely to be closely related; we would probably need to look much further into the past before we would find a coalescence event.

Coalescent theory makes this idea mathematically rigorous by providing formulae that relate the rate of coalescence to the effective population size. Thus, if we have a genealogy that shows when the coalescent events occurred, we can infer how the size of the population changed over time (Pybus, Rambaut, & Harvey, 2000).

Box 1 Hidden Markov models

A hidden Markov model is a pair of stochastic processes, X_t and Y_t, where X_t is the "hidden process" and cannot be directly observed, but Y_t can. At each point t, X_t takes on one of N possible states according to some specified probability distribution. Because X_t is a Markov process, the state it takes on depends only on the state at X_{t-1}. After X_t has moved to its new state, the value of Y_t is generated by a probability distribution that depends on the value that X_t takes on at that time. The values that Y_t can take are typically referred to as the "observation symbols" of the process.

To create a hidden Markov model, we need to define the key ingredients of the process that we described above:

1. The possible states of X_t, q_i, $i \in \{0, \ldots, N\}$
2. The possible observation symbols $v_i, i \in \{0, \ldots, M\}$
3. A probability distribution, the "transition probabilities," that describes how we move between the states of X_t: $P(X_{t+1} = q_j | X_t = q_i)$.
4. A set of probability distributions, called the "emission probabilities," that describe how the states of X_t generate the values of Y_t. Each of these will be of the form $b_j(k) = P(Y_t = v_k | X_t = q_j)$.
5. A probability distribution describing how the system looked when $t = 1$: $P(q_i | t = 1)$.

In the case of the sequentially Markovian coalescent models, t indexes locations along the genome. The hidden states are characterized by the local genealogies at each locus. For PSMC, the possible states are the possible coalescence times of the two alleles. For MSMC, it is the coalescence time of the two alleles in the sample that coalesce first. The observation symbols are features of the genetic data. For PSMC, the data are partitioned into bins of 100 bp; we observe a 1 if a heterozygous locus occurred in that bin and a 0 otherwise. For MSMC, there are a few more observation symbols to account for the extra complexity introduced by multiple genomes. The emission probabilities are determined by the mutation rate and the transition probabilities by the recombination rate.

Time periods with many coalescence events indicate a smaller population, and vice versa. A number of tools have been developed for inferring demographic history from the coalescence times in individual gene trees; these methods have been reviewed elsewhere (Ho & Shapiro, 2011).

The coalescent framework usually involves the assumption of neutral evolution. However, various forms of selection act across genomes, and these can affect specific mutations as well as any neutral variants that are linked (Maynard Smith & Haigh, 1974). Estimates of population size can be distorted by natural selection (Ewing & Jensen, 2016; Schrider, Shanku, & Kern, 2016). In particular, purifying selection is likely to be predominant and tends to remove genetic variation, which would lead to an apparent reduction in population size (Charlesworth, 1994; Charlesworth, Morgan, & Charlesworth, 1993). These potential impacts need to be borne in mind when using any methods of inference based on the coalescent, especially if the data set includes large sections of coding sequence.

1.2 | The sequentially Markovian coalescent

How can PSMC use coalescent theory to infer demographic history from a single genome, given that we need data from multiple coalescence events that each requires two alleles? The answer is that each genome contains large numbers of loci, which can split apart from each other during recombination and so trace out distinct evolutionary histories. By tracking the coalescence events between the two alleles at every locus, we can infer how many of them have occurred across the genome within a given time interval. PSMC and MSMC use this information to reconstruct the effective population size through time, provided that we make some assumptions about the mutation rate.

The original coalescent theory did not provide a framework for efficiently estimating population sizes from individual genomes, so it required modification before it could be used to infer population-size history using the approach described above. Before explaining the theoretical advances that solved this problem, we should consider what sort of model would be useful for inference. Obviously, the model should be biologically reasonable. It should also be mathematically tractable, in the sense that the computation of the likelihood functions for parameters is feasible. One class of models that fit this criterion, and hence are used widely in biology, are hidden Markov models (Box 1; Rabiner, 1989; Zucchini, MacDonald, & Langrock, 2016). In this framework, the data that we see are generated by a hidden background process. The process switches between a number of states, each of which has a specified probability of producing each of the observations. We cannot know with certainty what the background process is doing when we observe the data, but the observations provide probabilistic information about the process. Additionally, the background process must be Markovian, which means that the next state of the process depends only on the current state.

Much of the power of hidden Markov models for inference is that an algorithm—the Baum–Welch algorithm—exists that allows us to compute estimates of all of the free parameters simultaneously, provided that we specify the underlying structure of the process and how the process relates to the observations (Zucchini et al., 2016). Thus, the problem of inference under the coalescent could be solved by finding a biologically accurate description of the coalescent with recombination as a hidden Markov model. This was achieved by a change in perspective. Instead of starting from extant samples and building the full genealogy by working backwards, we work our way along the genome (Wiuf & Hein, 1999). We start from one end of the genome and then generate a "local" genealogy for each locus by incorporating new information as we move along the genome and encounter recombination events (Marjoram & Wall, 2006; McVean & Cardin, 2005; Wiuf & Hein, 1999). In the language of hidden Markov models, the local genealogy is the background process that generates the data, whereas the sequences are the observations. This process was named the sequentially Markovian coalescent (McVean & Cardin, 2005).

In PSMC, the local genealogy is completely characterized by the time to the most recent common ancestor of the two alleles, because there is only one possible tree topology for two leaves (Figure 1a). Analyzing multiple genomes is much more computationally challenging, but the MSMC simplifies this task by using only a subset of the local tree that describes the time to the most recent common ancestor of the two alleles that coalesce first at that locus (Figure 1b). The complexities associated with the analysis of multiple genomes have been addressed differently in other coalescent hidden Markov models (Dutheil, 2017).

2 | APPLICATIONS OF THE METHODS

2.1 | Reconstructing population sizes

The PSMC and MSMC methods were originally designed to infer past changes in effective population sizes, including the timing of expansions and bottlenecks. The output of each method is a plot showing how the effective size of the population to which that individual belonged has changed over time. PSMC and MSMC can combine information from much larger numbers of loci than is computationally feasible with tree-based methods such as skyline plots. For this reason, sequentially Markovian coalescent methods can probe deeper timescales, because the data are much more likely to include loci that have their most recent common ancestor far into the past. For example, application of the PSMC has been able to shed light on more than a million years of the demographic histories of modern humans and other great apes (Prado-Martinez et al., 2013). Although neither PSMC nor MSMC provides a formal framework for testing specific hypotheses of the causes of population-size changes, bootstrap replicates can be used to approximate a confidence interval around the estimate of effective population size (Li & Durbin, 2011).

In principle, MSMC has superseded PSMC, even when only a single diploid genome is available for analysis. MSMC estimates the recombination rate correctly in circumstances where PSMC fails to do so (Schiffels & Durbin, 2014). When multiple genomes are available, MSMC has better power to resolve recent changes in effective population size, because adding alleles increases the chance that there will be a coalescence event in the recent past. In analyses of human genomes, MSMC is informative for events that occurred as recently as 2,000 years ago (Schiffels & Durbin, 2014), whereas PSMC does not reliably resolve changes in effective population size that had occurred within the last 20,000 years (Li & Durbin, 2011).

There are limits to how far in the past the sequentially Markovian coalescent can make reliable estimates of demography. For the models to infer the population size in a given time period, the genome must have loci at which alleles coalesced in that period. However, coalescent theory shows that alleles with deep coalescences are relatively rare (Takahata & Nei, 1985). As we look further back in the past, it becomes increasingly unlikely that we have any data on the coalescence rate. Consequently, the inference of coalescence rates, and hence population sizes, becomes much noisier for the deep demographic past.

A critical consideration when using the PSMC and MSMC is that their output cannot always reliably be interpreted as plots of population-size changes. Recall from our description of the model that it estimates the rate of coalescence at each point in time. Coalescent theory shows that the inverse of this rate, sometimes called the inverse instantaneous coalescence rate (Mazet et al., 2016), can be used as a proxy for population size under certain assumptions. However, if the study population does not meet these assumptions, then other factors can affect the coalescence rate and must be taken into account when we attempt to interpret apparent changes in population size (Chikhi et al., 2018; Mazet, Rodríguez, & Chikhi, 2015; Mazet et al., 2016). For example, the relationship between coalescence times and population sizes can be confounded by natural selection and by nonrandom mating (Mazet et al., 2016).

One intuitive example of nonrandom mating is the *n*-island model, where *n* panmictic populations are separated except for some fixed rate of migration (Wright, 1931). Because two alleles cannot coalesce while they are in different islands, the expected coalescence time is determined by the number of islands and the migration rate between them, as well as by the population size (Mazet et al., 2016; Pannell, 2003). In particular, when the migration rate between islands is low, the coalescent effective population size—the parameter inferred by sequential methods—can be much greater than the true population size (Li & Durbin, 2011; Nei & Takahata, 1993). Thus, peaks on the demographic plot might correspond to periods of increased population structure rather than increased population size.

The assumption of panmixia can also be violated by inbreeding, which increases the rate of coalescence and hence lowers the effective population size. It should manifest in the genome as long runs of homozygous sequence (Ceballos, Joshi, Clark, Ramsay, & Wilson, 2018). One strategy for testing whether inbreeding has affected demographic inference is to identify and remove such runs of homozygosity, then repeating the analysis and checking for any changes in the inferences (Freedman et al., 2014).

In general, there is great difficulty in differentiating between the changes that are attributable to shifts in population size and those that are caused by changes to other demographic parameters (such as increased migration or a strengthening of population structure). Changes in some demographic parameters can alter the demographic curve in ways that are not well localized to the point at which the changes occurred. For example, because the migration rate is changed instantaneously in a constant-sized *n*-island population, the demographic curve rises and falls over many thousands of generations (Mazet et al., 2016). When many changes happen over a short space of time, they

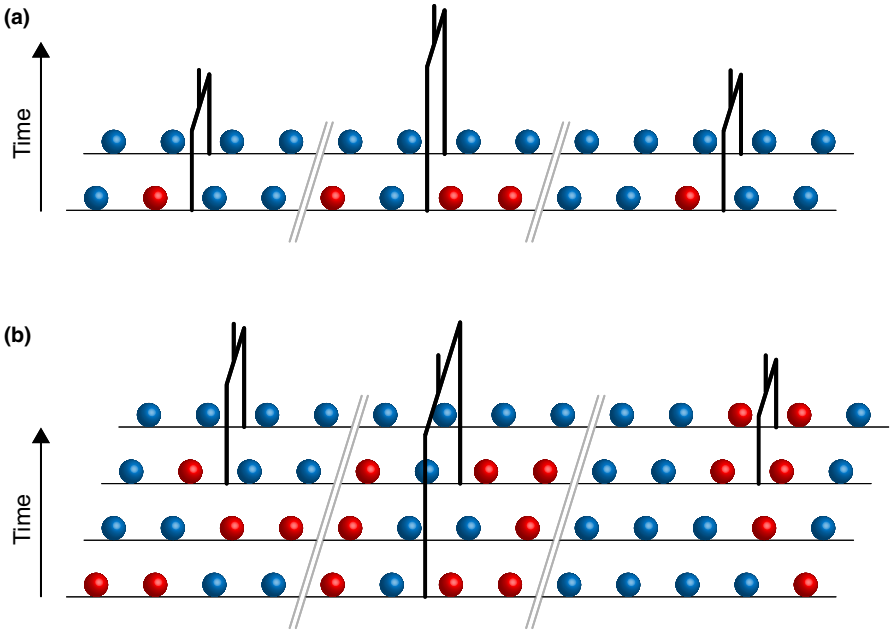

FIGURE 1 The sequentially Markovian coalescent. The colored circles represent nucleotide states belonging to the alleles at each locus. Double gray lines denote recombination breakpoints, which separate the loci along the genome. The time to the most recent common ancestor of the two alleles at each locus is reflected in the local tree. (a) In PSMC, there are only two haplotypes. Thus, the topology of the local tree is fixed, but the time to the most recent common ancestor differs among loci. (b) In MSMC, there are multiple haplotypes. MSMC ignores most of the local tree topology, focusing only on the most recent coalescence event at each locus

can interact in complex ways to produce the demographic plot, so it is difficult to attribute any feature of the plot to a specific change.

We do not know exactly what degree and type of structure will preclude the reliable interpretation of the inferences from PSMC and MSMC, but their use for highly structured populations is likely to be extremely misleading. Unfortunately, the impact of population structure cannot be assessed directly from the observed coalescence times. Even when the population size is constant but the coalescence time varies purely as a result of changes in population structure, we can always find a set of (false) population-size changes that would explain the observed coalescence times arbitrarily well (Mazet et al., 2016). Additionally, testing for population structure typically requires data from many individuals, which might not be available when methods based on the sequentially Markovian coalescent are used.

Finally, there are reasons to believe that the sequentially Markovian coalescent might perform poorly on realistic data. For example, when genetic data are produced by simulation under demographic models inferred by MSMC from human genomes, they fail to resemble the empirical data in important ways (Beichman, Phung, & Lohmueller, 2017). Other methods that use data from many individuals, such as δaδi (Gutenkunst, Hernandez, Williamson, & Bustamante, 2009) and SMC++ (Terhorst, Kamm, & Song, 2017), perform substantially better in this regard. Concerningly, this problem appears to grow worse as larger numbers of genomes are used for the MSMC, so the problem cannot necessarily be overcome by adding more data.

Since the release of MSMC, new methods have been designed that expand on the framework provided by the sequentially Markovian coalescent (Dutheil, 2017; Spence et al., 2018). These methods can outperform PSMC and MSMC in certain ways. The SMC++ method allows much larger numbers of genomes to be used for inference than is computationally possible with MSMC and does not require the genomes to be phased (Terhorst et al., 2017). SMC++ is more accurate than MSMC, especially for population sizes in the recent past and when phasing error is present. Another method, MAGIC (minimal-assumption inference from population-genomic data; Weissman & Hallatschek, 2017), is conceptually similar to the sequentially Markovian coalescent in that it infers the coalescent history of a sample from the distribution of polymorphisms in the genome. However, it does not use an explicit model of coalescence and recombination and can estimate many different parameters from the empirical data. Given the different strengths and weaknesses of the various methods for inferring demographic history, best practice should include the use of multiple methods and comparison of their inferences (Spence et al., 2018).

2.2 | Studying changes in population structure

A common task in population genetics is to infer the timing of divergence between closely related populations or species. One method that can be used for this purpose is the multispecies coalescent (e.g., Ogilvie, Bouckaert, & Drummond, 2017). However, this method requires data from multiple individuals per species and is not computationally feasible for data sets comprising large numbers of loci. Both PSMC and MSMC can be used to infer divergence times while incorporating information from whole genomes.

Although it is not an intended use of the method, PSMC can be adapted to identify the point at which gene flow ceased between a pair of populations (Cahill, Soares, Green, & Shapiro, 2016; Li & Durbin, 2011). A simple approach is to compare PSMC plots obtained from representatives of the two populations or species. The point at which their plots become identical indicates when they represent the same ancestral population (Figure 2a).

An alternative PSMC-based approach uses a synthetic diploid genome that is constructed from phased or unphased haplotypes from the two populations or species. The synthetic genome is intended to mimic that of an F_1 hybrid and so the approach is known as hPSMC (Cahill et al., 2016). Coalescence between the two alleles at each locus can only occur in the ancestral population. Because the rate of coalescence drops to zero after the populations have become reproductively isolated from each other, the effective population size will then be inferred to be infinite (Figure 2a). This technique can only provide a maximum bound on the divergence time, because it is possible that the populations diverged later than the most recent coalescence event. On the other hand, the estimated timing of divergence can be misled if the populations have not achieved complete reproductive isolation. Given the sensitivity of hPSMC to any gene flow that has occurred after the divergence between the two populations, the method is more appropriately used to date the cessation of gene flow rather than population divergence (Cahill et al., 2016).

Multiple sequentially Markovian coalescent tends to perform better across a range of population divergence scenarios (Zhou & Teo, 2016) and can be used to provide a more complete picture of how gene flow between populations has changed through time. If we know which samples came from which populations, MSMC can calculate the cross-coalescence rate between each pair of populations, as well as the coalescence rate within each population (e.g., Fan et al., 2019; Liang et al., 2019). The ratio of these terms will grow or shrink depending on the amount of gene flow between subpopulations: When migration between populations is high for some period of time, the cross-coalescence rate between populations should increase gradually during that interval as the populations come to share more alleles (Schiffels & Durbin, 2014).

2.3 | Estimating mutation rate

The PSMC framework has been used to estimate mutation rates from genome sequences (e.g., Fu et al., 2014; Palkopoulou et al., 2015; Skoglund et al., 2015). The technique requires two sequences that have distinct ages and that have been sampled from the same population. Given their shared demographic history, the two genomes should yield very similar plots of effective population size. However, the two plots will be offset along the time axis because of the difference in sampling times (Figure 2b). The mutation rate can be estimated by finding the number of mutations that need to be added to the ancient sequences in order to

superimpose the two plots, then dividing this by the difference in the ages of the two samples. One can do this numerically by simply adding mutations to the older genome until its demographic plot coincides with that of the younger genome, but the estimation can be performed more rigorously by using a maximum likelihood approach (Fu et al., 2014).

Inferring mutation rates using PSMC plots should only be done when an estimate of the mutation rate for the target species is otherwise unavailable. The use of demographic plots for this purpose is subject to a host of confounding factors, including those that affect the standard use of PSMC to infer population-size history. Other more generally accepted methods for estimating mutation rates, particularly those based on genome sequences from parent-offspring sets (e.g., Besenbacher, Hvilsom, Marques-Bonet, Mailund, & Schierup, 2019; Roach et al., 2010) or mutation-accumulation lines (e.g., Ossowski et al., 2010), are likely to be substantially more accurate.

3 | DATA REQUIREMENTS

3.1 | Genome sequences

Both PSMC and MSMC require data in the form of one or more genome sequences aligned to a reference sequence from the target species or from a closely related species. Using a reference sequence that is too dissimilar can reduce the accuracy of heterozygote calls (Günther & Nettelblad, 2019). Prior to analysis, genome data should be filtered to remove sites that have a high probability of being called inaccurately (Li, 2014). Failing to filter appropriately, especially when coverage is low, can lead to the demographic signal being obscured (Nadachowska-Brzyska, Burri, Smeds, & Ellegren, 2016).

The PSMC method requires a diploid consensus genome and does not require that haplotypes are phased, because it only needs to know the nucleotide positions that are heterozygous. The genome must be sequenced to a sufficiently high degree of coverage

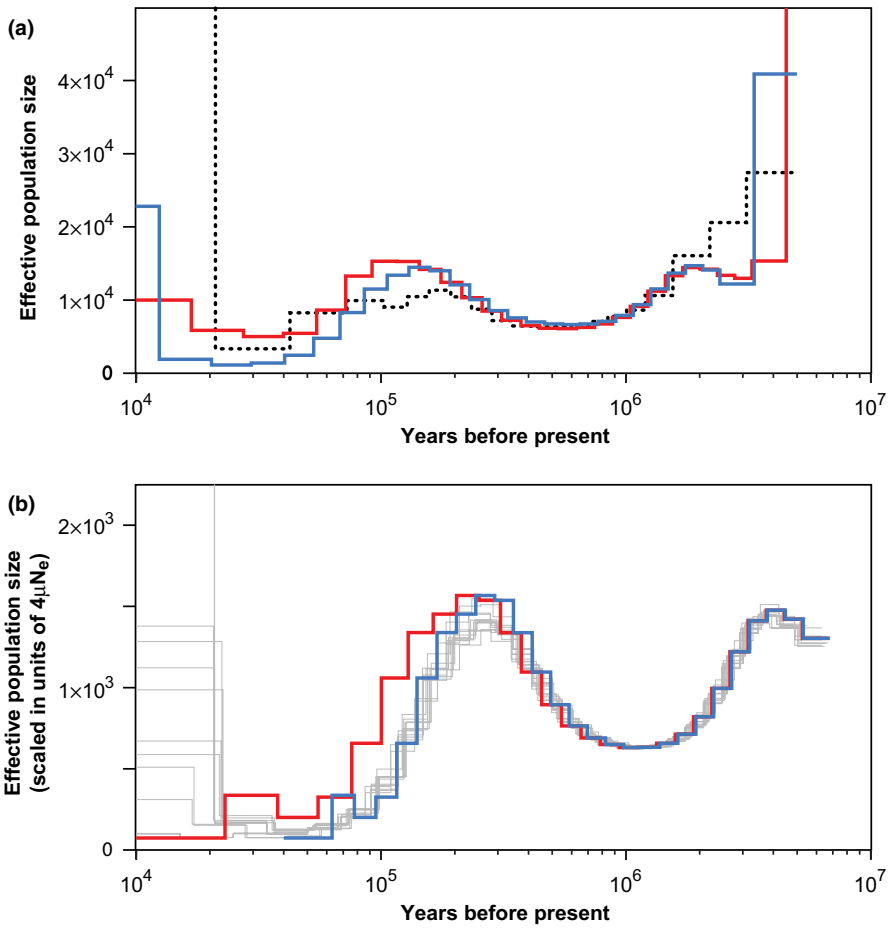

FIGURE 2 Illustrations of two different uses of PSMC and MSMC methods. (a) Dating speciation events using the PSMC (Li & Durbin, 2011). The solid lines represent PSMC plots from the genomes of two modern humans: Yoruba (red) and Chinese (blue). Looking backwards in time, the plots become identical from about 100–120 thousand years ago, indicating a shared population history. The dotted line shows a PSMC plot from a hybrid genome constructed from the X chromosomes of the Yoruba and Chinese genomes, scaled by 0.75. Looking forwards in time, the infinite population size at about 20 thousand years ago suggests cessation of gene flow between the two populations. (b) Estimating the mutation rate using PSMC (Fu et al., 2014). The two thick lines are plots for a 45,000-year-old modern human from Ust'-Ishim in western Siberia, either uncorrected (red) or shifted horizontally (blue) to align with the plots from present-day non-African modern humans (thin gray lines). The magnitude of the horizontal shift indicates the number of mutations that have occurred in 45,000 years, providing a means of estimating the mutation rate

that heterozygous sites can be called accurately; one estimate is that 18-fold coverage and <25% missing data are required for reliable inference (Nadachowska-Brzyska et al., 2016). PSMC might still be able to recover the essential features of a demographic history at lower coverage, but the shape of the graph tends to be flattened so that estimates of the effective population size might not be accurate. The effect of low coverage is worse when the amount of data is small, as when the analysis is based on only a subset of the genome.

Whole genome sequences should be used in principle, but informative PSMC plots can be obtained from individual chromosomes from the human genome (Li & Durbin, 2011). A recent simulation study also showed that PSMC can obtain accurate estimates even when using the products of short-read sequencing assembled into short scaffolds (Gower et al., 2018). If a subset of the genome is chosen, coding DNA should not constitute the majority of the data set because of the confounding effects of selection on demographic inference. Excluding sequences that are close to coding regions or that have low recombination rates can also help to mitigate the confounding impacts of selection (Schrider et al., 2016).

The MSMC method was designed to use data from multiple haplotypes but can use as few as two, in which case it reduces to a variant of PSMC (referred to as PSMC'). There is no upper bound on the number of haplotypes that can be used, but the method is feasible for at least eight haplotypes and the computational complexity increases rapidly as more are added (Schiffels & Durbin, 2014). When there are too many haplotypes to be used efficiently in MSMC, other methods can handle data sets comprising larger numbers of genomes (e.g., SMC++ and MAGIC; Terhorst et al., 2017; Weissman & Hallatschek, 2017).

In principle, MSMC requires that the sequence data are phased (i.e., haplotypes are specified), unless working with data from a parent–parent–offspring trio or from a single diploid genome. However, the importance of the phasing depends on the question that one seeks to answer: MSMC performs reasonably well when inferring the shape of a demographic curve from unphased data, but there is a substantial reduction in its resolution of the recent past and its ability to infer population divergence times (Schiffels & Durbin, 2014).

Where phasing is performed, it needs to be highly accurate because even a relatively small rate of phasing error can seriously mislead the effective population sizes estimated by MSMC, especially for the recent past (Song, Sliwerska, Emery, & Kidd, 2017; Terhorst et al., 2017). This can be problematic for MSMC, because phasing to the required level of accuracy will typically need a larger sample size or access to an external reference panel of haplotype information (Browning & Browning, 2011). When robust phasing is not possible, using unphased data might be a better option if one is only interested in the qualitative shape of the demographic curve. Alternatively, one can use a phasing-invariant method of demographic inference, such as SMC++, if the goal is to obtain precise estimates of population size or to explore demographic events from the recent past (Browning & Browning, 2011; Terhorst et al., 2017).

3.2 | Restriction-site-associated DNA data

Both PSMC and MSMC can be used with restriction site-associated DNA (RAD) data (Liu & Hansen, 2017). RAD sequencing is a reduced-representation method that gives the sequences of regions flanking the cutting sites of a chosen restriction enzyme (Miller, Dunham, Amores, Cresko, & Johnson, 2007). The smaller the fraction of the genome that this subset covers, the greater the reduction in accuracy and increase in variance. As with inferences from other reduced data sets, the demographic curve obtained from RAD data is flatter, with peaks and troughs that are less pronounced (Liu & Hansen, 2017).

Based on evidence from simulations, a rule of thumb is that PSMC can recover the broad shape of the demographic curve if $\mu p/r > 0.5$, where μ is the mutation rate, p is the fraction of the genome covered by the RAD sequencing, and r is the recombination rate (Liu & Hansen, 2017). In practice, when RAD data are used for demographic inference, the read length and sampling density should be maximized.

4 | PRACTICAL CONSIDERATIONS

4.1 | Parameter selection

In PSMC and MSMC analyses, a number of settings need to be specified by the user. The two methods assume that the history of the population is divided into discrete time intervals on which the population size is constant. The user must specify the length and number of these intervals; a poor choice of intervals can lead to over- or underfitting of the model. Repeating the analysis using different numbers of time intervals can show whether there is any impact on the inferences (Nadachowska-Brzyska et al., 2016).

Intuitively, splitting a time period into many intervals can lead to greater stochastic error in the population-size estimate for each time interval, because there will be higher variance in the number of coalescent events occurring in smaller intervals. In addition, because coalescent events are rarer when the population is large, model overfit will be most pronounced at the "peaks" of the plot. One way to reduce the impact of this problem is to check that a sufficiently large number of coalescent events fall into each interval. This threshold is somewhat subjective, but a minimum of 20 events per interval has been suggested (Li & Durbin, 2011). To choose an appropriate number of time intervals, it might be useful to begin with some number of evenly spaced intervals. After running the analysis, one can identify sections of the plot with too few coalescent events and reduce the number of intervals in those periods before repeating the analysis.

The above argument also shows a weakness of the PSMC and MSMC methods. We cannot distinguish between the noise induced by overfitting and the signals of genuine changes in population size. Thus, one should be very cautious in interpreting changes in effective population size on small timescales, especially around peaks of the demographic plot. Creating bootstrap replicates of the analysis

can provide an estimate of the variance in the estimated effective population size (Li & Durbin, 2011).

Both PSMC and MSMC also require the user to supply an initial value for the ratio of the mutation rate and recombination rate. In theory, this should not affect the outcome of the analysis because it only provides a starting value for the algorithm. Simulations have shown that even when the analysis fails to estimate this value, there are no negative impacts on the estimates of population sizes (Li & Durbin, 2011).

4.2 | Scaling the graphs

Analyses using PSMC and MSMC give plots of effective population sizes that are scaled to the per-generation mutation rate. To allow a time axis to be added to the plot, the mutation rate (per generation) and generation interval need to be specified. Generation interval can be difficult to define precisely, even for well-studied taxa such as modern humans (Scally & Durbin, 2012). Reliable estimates of mutation rates are not easily obtained, because phylogenetic estimates of long-term evolutionary rates are not necessarily applicable at the population level (Ho, Duchêne, Molak, & Shapiro, 2015); rates of spontaneous mutation have been inferred for a limited number of eukaryote species (Besenbacher et al., 2019; Smeds, Qvarnström, & Ellegren, 2016). If estimates of the mutation rate are unavailable for the target species, common practice has been to employ the phylogenetically closest estimate (see Section 5). An alternative approach is to derive an approximation of the mutation rate based on its covariation with other biological quantities, such as genome size and population size (Lynch et al., 2016). This method has been used in a number of PSMC and MSMC studies (e.g., Hall et al., 2017).

Using an incorrect value for either the mutation rate or generation interval does not affect the qualitative shape of the plot, and so will not affect analyses that do not require the precise dating of demographic events. A higher mutation rate will cause the estimated population size to be scaled down linearly and will shift the curve closer to the present, whereas a longer generation interval will scale the population size down (Nadachowska-Brzyska et al., 2016). If precise dating is required, it might be best to run the analyses under a plausible range of mutation rates, to provide upper and lower bounds for the dating of demographic events.

5 | USAGE SURVEY

To present a picture of how PSMC and MSMC are used in scientific studies, we surveyed their usage in peer-reviewed journal articles. We randomly sampled 100 of the ~200 studies that have performed PSMC analysis and 30 of the ~60 studies that have performed MSMC analysis. We identified these studies by scanning the approximately 1,400 papers that have cited the original descriptions of the two methods (Li & Durbin, 2011; Schiffels & Durbin, 2014), according to Google Scholar.

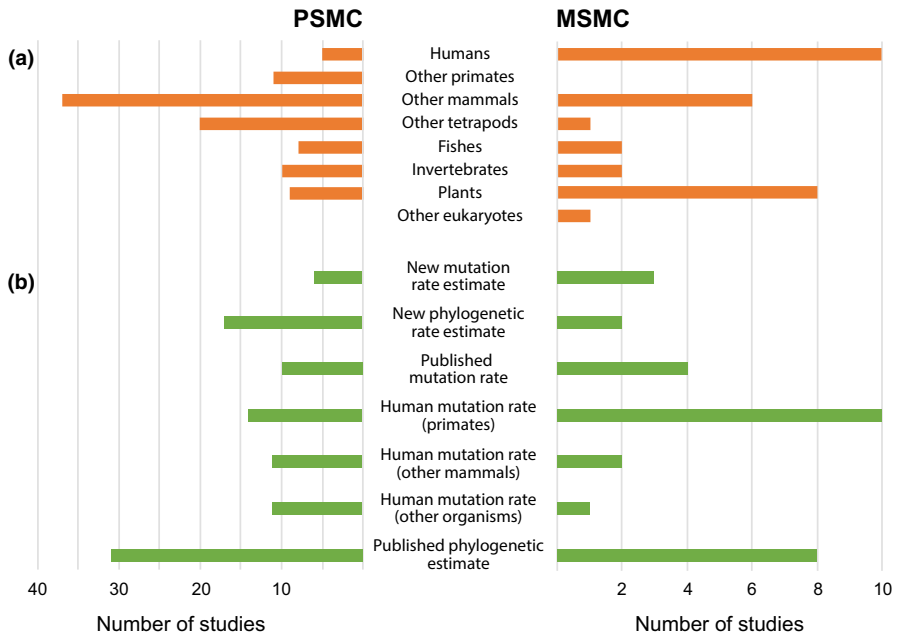

FIGURE 3 Data from random samples of 100 studies that implemented the pairwise sequentially Markovian coalescent (PSMC) method and 30 studies that implemented the multiple sequentially Markovian coalescent (MSMC) method. (a) Taxonomic affiliations of the organisms studied. (b) Source of mutation rates used to scale the demographic plots. The choice of mutation rate affects the scale of the horizontal axis and the estimate of the effective population size, but does not affect the qualitative shape of the plot. For example, a higher mutation rate will cause the estimated population size to be scaled down linearly and will shift the curve closer to the present. Details of the 130 studies surveyed are provided in Appendix S1 on Dryad (https://doi.org/10.5061/dryad.0vt4b8gv2)

Among the 100 PSMC studies that we examined in detail (Figure 3a), the majority focused on the genomes of mammals (53%) and other vertebrates (28%). Although many of these studies investigated population structure and estimated rates of gene flow between subpopulations, these demographic features were very rarely taken into account during interpretation of the PSMC plots. Similarly, the potential confounding impacts of selection were sometimes acknowledged, but were not addressed explicitly. Among the 30 MSMC studies that we examined in detail, there tended to be a greater focus on modern humans and plants (Figure 3a).

The vast majority of the PSMC and MSMC studies analyzed whole genome sequences, although some produced separate demographic plots for autosomes and sex chromosomes (e.g., Ekblom et al., 2018; Foote et al., 2016). Most studies did not provide explicit justification for the choice of discrete time intervals, which was usually the default number of 64 based on the initial application of PSMC to human genomes (Li & Durbin, 2011). Among the 30 MSMC studies, 11 analyzed data sets with two haplotypes (equivalent to the PSMC) and 11 analyzed data sets with eight haplotypes.

The PSMC and MSMC plots were usually scaled according to mutation rates that had been estimated in previous studies (Figure 3b). Many of these mutation rates were estimated on phylogenetic scales, but they were more often obtained from pedigree-based analyses. Estimates of human mutation rates were applied to PSMC and MSMC plots in 49 studies, although only 24 of these involved analyses of humans or other primates. Studies of birds, flies, lepidopterans, and plants tended to apply mutation rates estimated from *Ficedula* (Nadachowska-Brzyska et al., 2016), *Drosophila* (Haag-Liautard et al., 2007), *Heliconius* (Keightley et al., 2015), and *Arabidopsis* (Ossowski et al., 2010), respectively; together these accounted for 11% of the studies surveyed. Even in these cases, however, the scaling of the demographic plots can be severely misled if there is substantial rate heterogeneity among species. Only 7% of studies produced novel estimates of short-term mutation rates for the target species, for example using analyses of parent–parent–offspring trios (Künstner et al., 2016; Martin et al., 2018). Our usage survey highlights the challenges in identifying suitable estimates of mutation rates for rescaling the demographic plots produced by PSMC and MSMC.

6 | CONCLUSIONS

Sequentially Markovian coalescent methods provide powerful means of inferring demographic histories from genomic data. They are particularly useful when the genome sequences are restricted to only a few individuals, or when the aim is to probe timescales that are inaccessible to other methods (Beichman et al., 2018). However, there are substantial challenges in interpreting the output of PSMC and MSMC, because testing the assumptions of the underlying models usually requires more data than are available when these methods are used. Nevertheless, by being aware of the underlying theoretical framework and identifying the assumptions made in the interpretation of the results, researchers can glean valuable demographic information that might otherwise be unavailable.

ACKNOWLEDGMENTS

This work was partially supported by the Australian Research Council (grant FT160100167).

CONFLICT OF INTEREST

None declared.

AUTHOR CONTRIBUTIONS

S.M.T. and S.Y.W.H. performed the usage survey. N.M. and S.Y.W.H. wrote the paper, with input from S.M.T.

ORCID

Simon Y. W. Ho https://orcid.org/0000-0002-0361-2307

DATA AVAILABILITY STATEMENT

Data collected for Figure 3 are available provided in Appendix S1 on Dryad (https://doi.org/10.5061/dryad.0vt4b8gv2).

REFERENCES

Beichman, A. C., Huerta-Sanchez, E., & Lohmueller, K. E. (2018). Annual review of ecology. *Evolution, and Systematics*, *49*, 433–456.

Beichman, A. C., Phung, T. N., & Lohmueller, K. E. (2017). Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3: Genes, Genomes, Genetics*, *7*, 3605–3620. https://doi.org/10.1534/g3.117.300259

Besenbacher, S., Hvilsom, C., Marques-Bonet, T., Mailund, T., & Schierup, M. H. (2019). Direct estimation of mutations in great apes reconciles phylogenetic dating. *Nature Ecology and Evolution*, *3*, 286–292. https://doi.org/10.1038/s41559-018-0778-x

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: Existing methods and new developments. *Nature Reviews Genetics*, *12*, 703–714. https://doi.org/10.1038/nrg3054

Cahill, J. A., Soares, A. E., Green, R. E., & Shapiro, B. (2016). Inferring species divergence times using pairwise sequential Markovian coalescent modelling and low-coverage genomic data. *Philosophical Transactions of the Royal Society of London B*, *371*, 20150138. https://doi.org/10.1098/rstb.2015.0138

Ceballos, F. C., Joshi, P. K., Clark, D. W., Ramsay, M., & Wilson, J. F. (2018). Runs of homozygosity: Windows into population history and trait architecture. *Nature Reviews Genetics*, *19*, 220–234. https://doi.org/10.1038/nrg.2017.109

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetics Research*, *63*, 213–227. https://doi.org/10.1017/S0016672300032365

Charlesworth, B., Morgan, M. T., & Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, *134*, 1289–1303.

Chikhi, L., Rodríguez, W., Grusea, S., Santos, P., Boitard, S., & Mazet, O. (2018). The IICR (inverse instantaneous coalescence rate) as a summary of genomic diversity: Insights into demographic inference and model choice. *Heredity*, *120*, 13–24. https://doi.org/10.1038/s41437-017-0005-6

Dutheil, J. Y. (2017). Hidden Markov models in population genomics. In D. R. Westhead, & M. S. Vijayabaskar (Eds.), *Hidden Markov models: Methods and protocols* (pp. 149–164). New York, NY: Springer Science + Business Media.

Ekblom, R., Brechlin, B., Persson, J., Smeds, L., Johansson, M., Magnusson, J., … Ellegren, H. (2018). Genome sequencing and conservation genomics in the Scandinavian wolverine population. *Conservation Biology*, *32*, 1301–1312. https://doi.org/10.1111/cobi.13157

Ewing, G. B., & Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Molecular Ecology*, *25*, 135–141. https://doi.org/10.1111/mec.13390

Fan, S., Kelly, D. E., Beltrame, M. H., Hansen, M. E. B., Mallick, S., Ranciaro, A., … Tishkoff, S. A. (2019). African evolutionary history inferred from whole genome sequence data of 44 indigenous African populations. *Genome Biology*, *20*, 82.

Foote, A. D., Vijay, N., Ávila-Arcos, M. C., Baird, R. W., Durban, J. W., Fumagalli, M., … Wolf, J. B. W. (2016). Genome-culture coevolution promotes rapid divergence of killer whale ecotypes. *Nature Communications*, *7*, 11693. https://doi.org/10.1038/ncomms11693

Frantz, L. A. F., Mullin, V. E., Pionnier-Capitan, M., Lebrasseur, O., Ollivier, M., Perri, A., … Larson, G. (2016). Genomic and archaeological evidence suggests a dual origin of domestic dogs. *Science*, *352*, 1228–1231.

Freedman, A. H., Gronau, I., Schweizer, R. M., Ortega-Del Vecchyo, D., Han, E., Silva, P. M., … Novembre, J. (2014). Genome sequencing highlights the dynamic early history of dogs. *PLoS Genetics*, *10*, e1004016. https://doi.org/10.1371/journal.pgen.1004016

Fu, Q., Li, H., Moorjani, P., Jay, F., Slepchenko, S. M., Bondarev, A. A., … Pääbo, S. (2014). Genome sequence of a 45 000-year-old modern human from western Siberia. *Nature*, *514*, 445–449. https://doi.org/10.1038/nature13810

Gower, G., Tuke, J., Rohrlach, A. B., Soubrier, J., Llamas, B., Bean, N., & Cooper, A. (2018). Population size history from short genomic scaffolds: How short is too short? *bioRxiv*. https://doi.org/10.1101/382036

Günther, T., & Nettelblad, C. (2019). The presence and impact of reference bias on population genomic studies of prehistoric human populations. *PLoS Genetics*, *15*, e1008302. https://doi.org/10.1371/journal.pgen.1008302

Gutenkunst, R. N., Hernandez, R. D., Williamson, S. H., & Bustamante, C. D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genetics*, *5*, e1000695. https://doi.org/10.1371/journal.pgen.1000695

Haag-Liautard, C., Dorris, M., Maside, X., Macaskill, S., Halligan, D. L., Charlesworth, B., & Keightley, P. D. (2007). Direct estimation of per nucleotide and genomic deleterious mutation rates in *Drosophila*. *Nature*, *445*, 82–85. https://doi.org/10.1038/nature05388

Hall, M. R., Kocot, K. M., Baughman, K. W., Fernandez-Valverde, S. L., Gauthier, M. E. A., Hatleberg, W. L., … Degnan, B. M. (2017). The crown-of-thorns starfish genome as a guide for biocontrol of this coral reef pest. *Nature*, *544*, 231–234. https://doi.org/10.1038/nature22033

Ho, S. Y. W., Duchêne, S., Molak, M., & Shapiro, B. (2015). Time-dependent estimates of molecular evolutionary rates: Evidence and causes. *Molecular Ecology*, *24*, 6007–6012. https://doi.org/10.1111/mec.13450

Ho, S. Y. W., & Shapiro, B. (2011). Skyline-plot methods for estimating demographic history from nucleotide sequences. *Molecular Ecology Resources*, *11*, 423–434. https://doi.org/10.1111/j.1755-0998.2011.02988.x

Hudson, R. R. (1983). Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, *23*, 183–201. https://doi.org/10.1016/0040-5809(83)90013-8

Johnson, R. N., O'Meally, D., Chen, Z., Etherington, G. J., Ho, S. Y. W., Nash, W. J., … Belov, K. (2018). Adaptation and conservation insights from the koala genome. *Nature Genetics*, *50*, 1102–1111. https://doi.org/10.1038/s41588-018-0153-5

Keightley, P. D., Pinharanda, A., Ness, R. W., Simpson, F., Dasmahapatra, K. K., Mallet, J., … Jiggins, C. D. (2015). Estimation of the spontaneous mutation rate in *Heliconius melpomene*. *Molecular Biology and Evolution*, *32*, 239–243. https://doi.org/10.1093/molbev/msu302

Kingman, J. F. C. (1982). The coalescent. *Stochastic Processes and their Applications*, *13*, 235–248. https://doi.org/10.1016/0304-4149(82)90011-4

Künstner, A., Hoffmann, M., Fraser, B. A., Kottler, V. A., Sharma, E., Weigel, D., & Dreyer, C. (2016). The genome of the Trinidadian guppy, Poecilia reticulata, and variation in the Guanapo population. *PloS one*, *11*(12), e0169087.

Li, H. (2014). Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics*, *30*, 2843–2851. https://doi.org/10.1093/bioinformatics/btu356

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*, 493–496. https://doi.org/10.1038/nature10231

Liang, Z., Duan, S., Sheng, J., Zhu, S., Ni, X., Shao, J., … Dong, Y. (2019). Whole-genome resequencing of 472 *Vitis* accessions for grapevine diversity and demographic history analyses. *Nature Communications*, *10*, 1190. https://doi.org/10.1038/s41467-019-09135-8

Liu, S., & Hansen, M. M. (2017). PSMC (pairwise sequentially Markovian coalescent) analysis of RAD (restriction site associated DNA) sequencing data. *Molecular Ecology Resources*, *17*, 631–641. https://doi.org/10.1111/1755-0998.12606

Lynch, M., Ackerman, M. S., Gout, J.-F., Long, H., Sung, W., Thomas, W. K., & Foster, P. L. (2016). Genetic drift, selection and the evolution of the mutation rate. *Nature Reviews Genetics*, *17*, 704–714. https://doi.org/10.1038/nrg.2016.104

Marjoram, P., & Wall, J. D. (2006). Fast "coalescent" simulation. *BMC Genetics*, *7*, 16.

Martin, H. C., Batty, E. M., Hussin, J., Westall, P., Daish, T., Kolomyjec, S., … Donnelly, P. (2018). Insights into platypus population structure and history from whole-genome sequencing. *Molecular Biology and Evolution*, *35*, 1238–1252.

Maynard Smith, J., & Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetics Research*, *23*, 23–35. https://doi.org/10.1017/S0016672300014634

Mazet, O., Rodríguez, W., & Chikhi, L. (2015). Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, *104*, 46–58. https://doi.org/10.1016/j.tpb.2015.06.003

Mazet, O., Rodríguez, W., Grusea, S., Boitard, S., & Chikhi, L. (2016). On the importance of being structured: Instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, *116*, 362–371. https://doi.org/10.1038/hdy.2015.104

McVean, G. A. T., & Cardin, N. J. (2005). Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society of London B*, *360*, 1387–1393. https://doi.org/10.1098/rstb.2005.1673

Miller, M. R., Dunham, J. P., Amores, A., Cresko, W. A., & Johnson, E. A. (2007). Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. *Genome Research*, *17*, 240–248. https://doi.org/10.1101/gr.5681207

Miller, W., Schuster, S. C., Welch, A. J., Ratan, A., Bedoya-Reina, O. C., Zhao, F., … Lindqvist, C. (2012). Polar and brown bear genomes reveal ancient admixture and demographic footprints of past climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *109*, E2382–E2390. https://doi.org/10.1073/pnas.1210506109

Nadachowska-Brzyska, K., Burri, R., Smeds, L., & Ellegren, H. (2016). PSMC analysis of effective population sizes in molecular ecology and its application to black-and-white *Ficedula* flycatchers. *Molecular Ecology*, *25*, 1058–1072.

Nei, M., & Takahata, N. (1993). Effective population size, genetic diversity, and coalescence time in subdivided populations. *Journal of Molecular Evolution*, *37*, 240–244. https://doi.org/10.1007/BF00175500

Ogilvie, H. A., Bouckaert, R. R., & Drummond, A. J. (2017). StarBEAST2 brings faster species tree inference and accurate estimates of substitution rates. *Molecular Biology and Evolution*, *34*, 2101–2114. https://doi.org/10.1093/molbev/msx126

Orlando, L., Ginolhac, A., Zhang, G., Froese, D., Albrechtsen, A., Stiller, M., … Willerslev, E. (2013). Recalibrating *Equus* evolution using the genome sequence of an early Middle Pleistocene horse. *Nature*, *499*, 74–78. https://doi.org/10.1038/nature12323

Ossowski, S., Schneeberger, K., Lucas-Lledo, J. I., Warthmann, N., Clark, R. M., Shaw, R. G., … Lynch, M. (2010). The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, *327*, 92–94. https://doi.org/10.1126/science.1180677

Palkopoulou, E., Mallick, S., Skoglund, P., Enk, J., Rohland, N., Li, H., … Dalén, L. (2015). Complete genomes reveal signatures of demographic and genetic declines in the woolly mammoth. *Current Biology*, *25*, 1395–1400. https://doi.org/10.1016/j.cub.2015.04.007

Pannell, J. R. (2003). Coalescence in a metapopulation with recurrent local extinction and recolonization. *Evolution*, *57*, 949–961. https://doi.org/10.1111/j.0014-3820.2003.tb00307.x

Prado-Martinez, J., Sudmant, P. H., Kidd, J. M., Li, H., Kelley, J. L., Lorente-Galdos, B., … Marques-Bonet, T. (2013). Great ape genetic diversity and population history. *Nature*, *499*, 471–475. https://doi.org/10.1038/nature12228

Pujolar, J. M., Dalén, L., Hansen, M. M., & Madsen, J. (2017). Demographic inference from whole-genome and RAD sequencing data suggests alternating human impacts on goose populations since the last ice age. *Molecular Ecology*, *26*, 6270–6283. https://doi.org/10.1111/mec.14374

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

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*, 257–286. https://doi.org/10.1109/5.18626

Roach, J. C., Glusman, G., Smit, A. F. A., Huff, C. D., Hubley, R., Shannon, P. T., … Galas, D. J. (2010). Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, *328*, 636–639. https://doi.org/10.1126/science.1186802

Scally, A., & Durbin, R. (2012). Revising the human mutation rate: Implications for understanding human evolution. *Nature Reviews Genetics*, *13*, 745–753. https://doi.org/10.1038/nrg3295

Schiffels, S., & Durbin, R. (2014). Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, *46*, 919–925. https://doi.org/10.1038/ng.3015

Schrider, D. R., Shanku, A. G., & Kern, A. D. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, *204*, 1207–1223. https://doi.org/10.1534/genetics.116.190223

Skoglund, P., Ersmark, E., Palkopoulou, E., & Dalén, L. (2015). Ancient wolf genome reveals an early divergence of domestic dog ancestors and admixture into high-latitude breeds. *Current Biology*, *25*, 1515–1519. https://doi.org/10.1016/j.cub.2015.04.019

Smeds, L., Qvarnström, A., & Ellegren, H. (2016). Direct estimate of the rate of germline mutation in a bird. *Genome Research*, *26*, 1211–1218. https://doi.org/10.1101/gr.204669.116

Song, S., Sliwerska, E., Emery, S., & Kidd, J. M. (2017). Modeling human population separation history using physically phased genomes. *Genetics*, *205*, 385–395. https://doi.org/10.1534/genetics.116.192963

Spence, J. P., Steinrücken, M., Terhorst, J., & Song, Y. S. (2018). Inference of population history using coalescent HMMs: Review and outlook. *Current Opinion in Genetics and Development*, *53*, 70–76. https://doi.org/10.1016/j.gde.2018.07.002

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. *Genetics*, *105*, 437–460.

Takahata, N., & Nei, M. (1985). Gene genealogy and variance of interpopulational nucleotide differences. *Genetics*, *110*, 325–344.

Terhorst, J., Kamm, J. A., & Song, Y. S. (2017). Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nature Genetics*, *49*, 303–309. https://doi.org/10.1038/ng.3748

Weissman, D. B., & Hallatschek, O. (2017). Minimal-assumption Inference from population-genomic Data. *eLife*, *6*, e24836.

Wiuf, C., & Hein, J. (1999). Recombination as a point process along sequences. *Theoretical Population Biology*, *55*, 248–259. https://doi.org/10.1006/tpbi.1998.1403

Wright, S. (1931). Evolution in Mendelian populations. *Genetics*, *16*, 97–159.

Yu, Y., Fu, J., Xu, Y., Zhang, J., Ren, F., Zhao, H., … Xie, H. (2018). Genome re-sequencing reveals the evolutionary history of peach fruit edibility. *Nature Communications*, *9*, 5404. https://doi.org/10.1038/s41467-018-07744-3

Zhou, J., & Teo, Y.-Y. (2016). Estimating time to the most recent common ancestor (TMRCA): Comparison and application of eight methods. *European Journal of Human Genetics*, *24*, 1195–1201. https://doi.org/10.1038/ejhg.2015.258

Zucchini, W., MacDonald, I. L., & Langrock, R. (2016). *Hidden Markov models for time series: An introduction using R*. Boca Raton, FL: Chapman and Hall/CRC.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

How to cite this article: Mather N, Traves SM, Ho SYW. A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecol Evol*. 2020;10:579–589. https://doi.org/10.1002/ece3.5888