Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in
Genetics
& Development

Inference of population history using coalescent HMMs: review and outlook

Jeffrey P Spence[1], Matthias Steinrücken[2], Jonathan Terhorst[3] and Yun S Song[4,5]

Studying how diverse human populations are related is of historical and anthropological interest, in addition to providing a realistic null model for testing for signatures of natural selection or disease associations. Furthermore, understanding the demographic histories of other species is playing an increasingly important role in conservation genetics. A number of statistical methods have been developed to infer population demographic histories using whole-genome sequence data, with recent advances focusing on allowing for more flexible modeling choices, scaling to larger data sets, and increasing statistical power. Here we review coalescent hidden Markov models, a powerful class of population genetic inference methods that can utilize linkage disequilibrium information effectively. We highlight recent advances, give advice for practitioners, point out potential pitfalls, and present possible future research directions.

Addresses
[1] Computational Biology Graduate Group, University of California, Berkeley, United States
[2] Department of Ecology and Evolution, University of Chicago, United States
[3] Department of Statistics, University of Michigan, United States
[4] Computer Science Division and Department of Statistics, University of California, Berkeley, United States
[5] Chan Zuckerberg Biohub, San Francisco, United States

Corresponding author: Song, Yun S (yss@berkeley.edu)

Current Opinion in Genetics & Development 2018, **53**:70–76

This review comes from a themed issue on **Genetics of human origins**

Edited by **Lluis Quintana-Murci** and **Brenna Henn**

https://doi.org/10.1016/j.gde.2018.07.002https://doi.org/10.1016/j.gde.2018.07.002

Introduction

Using genetic data to understand the history of a population has been a long-standing goal of population genetics [1], and the emergence of massive data sets with individuals from many populations [2,3,4••], often including ancient samples [5], have enabled the inference of increasingly realistic models of the genetic history of human populations [6–8]. The progress in other species is no less impressive, with demographic models inferred for dogs [9], horses, [10], pigs [11], and many others.

These demographic models are frequently of interest in their own right for historical or anthropological reasons, and failing to account for demographic history when performing tests of neutrality [12], disease associations, [13], or recombination rate inference [14,15] can lead to spurious results. Demographic models also play an important role in conservation genetics, informing breeding strategies for maintaining genetic diversity in endangered populations [16].

Yet, inferring complex demographic models — often including multiple populations with continuous migration, admixture events, and changes in effective population size — is challenging both statistically and computationally, and numerous methods have been developed to address this problem. Even under neutral evolution, computing the likelihood of observing a set of genotypes given a demographic model is computationally and analytically intractable. Hence, demographic inference methods must make simplifying approximations and generally fall into three classes: those based on allele frequencies; those based on identity-by-descent (IBD) or identity-by-state (IBS); and coalescent hidden Markov models (coalescent-HMMs).

Allele frequency-based methods use the multipopulation sample frequency spectrum (SFS) to infer either parametric [17–19,20•,21•] or non-parametric [22] models. For computational purposes, these methods assume that all loci are independent, an assumption violated by physically-linked loci, and thus ignore the rich information contained in such linkage (although [23] relaxes this to allow pairwise dependencies). Yet, these methods are very fast, with recent methods scaling to data sets with hundreds of individuals from tens of populations [21•], making them ideal for quickly exploring many potential models (e.g. testing models with different number of admixture events). Nevertheless, there are concerns about statistical identifiability ([24], but see [25]), power [26,27•], and stability [28].

IBD-based and IBS-based methods use patterns of pairwise haplotype sharing to infer demographic models, matching the distribution of observed IBD or IBS tract lengths to the distribution expected under the inferred demographic model. While IBD-based methods, such as [29–31], can be powerful — especially for learning about the recent past — they rely on having access to unobserved IBD tracts. Many methods have been developed for inferring IBD tracts [32,33], but these rely either explicitly or implicitly on the unknown demographic history of the samples, resulting in a chicken/egg problem. The effect of these assumptions on IBD-based methods has not been thoroughly explored, although see [34]. To sidestep this issue, [35] works directly with IBS tracts, a promising direction for further methodological development.

The focus of this review is the final class of methods: coalescent-HMMs. Below, we provide a historical overview of coalescent-HMMs; explore recent advances; discuss caveats, pitfalls, and best practices for applying coalescent-HMMs to data; and conclude with open problems and promising future research directions.

A brief history of coalescent-HMMs

Coalescent-HMMs can trace back to the seminal work of Wiuf and Hein [36]. The coalescent — a stochastic model of the genealogy of a sample of homologous chromosomes — was first developed for a single non-recombining locus [37] and then extended to incorporate recombination [38]. The coalescent had been thought of as a process through time, but Wiuf and Hein [36] formulated it as a process along the genome. This sequential coalescent is very complex and non-Markovian (the genealogy at a locus depends on the genealogies at all previous loci), but simple, yet highly accurate, Markovian approximations were subsequently proposed (the *sequentially Markovian coalescent*; SMC) [39–42].

Under the SMC, observed sequence data are modelled in a hidden Markov model (HMM) [43] framework by treating the genealogy of the sampled individuals at a given locus as an unobserved, latent variable. Because the demographic model impacts the distribution of genealogies (e.g. without migration, samples from different populations cannot have a common ancestor more recent than the divergence of those populations) and the observed sequence data are directly dependent on the underlying genealogy, coalescent-HMM methods can be extremely powerful. Furthermore, the HMM framework integrates over all possible genealogies when inferring demographic models — even if there is substantial uncertainty about the genealogy of a given sample, the set of genealogies likely to have given rise to that sample is still informative about its demographic history.

In principle, the HMM framework enables efficient inference of demographic parameters, but there are a number of complications. First, except for rare special cases (e.g. Kalman Filters [44] and iHMMs [45]), HMM algorithms require a finite state space for the latent variables; this is problematic in the coalescent-HMM case since the branch lengths of the genealogy at a given locus are continuous and can take an uncountably infinite number of values. All coalescent-HMMs avoid this issue by discretizing time. Having a finite state space is not sufficient for efficient inference, however, as the number of tree topologies grows super-exponentially in the sample size, making the full coalescent-HMM impractical for all but the smallest sample sizes. The menagerie of coalescent-HMM methods then arises by making different approximations to this idealized coalescent-HMM: instead of tracking the entire genealogy of the sample as a latent variable, these methods only track some features or subset of the genealogy.

Briefly, CoalHMM [46,47], developed to study different species, tracks only the topology of the genealogy and in which branch of the species tree the lineages coalesce. CoalHMM cannot scale to more than a few species. PSMC [48] can only be applied to a pair of haplotypes, but tracks their genealogy exactly, up to the discretization of time. MSMC [49] can use more than two haplotypes, but only tracks the time to the first coalescence event and the individuals involved in it. The first version of diCal [50], inspired by the copying model of [51] and subsequent work on conditional sampling distributions (CSDs) [52,53], considers a particular haplotype and tracks when and with which other haplotype it first coalesces. PSMC makes the fewest simplifying assumptions, but as it can only be applied to two haplotypes it is less powerful than MSMC or diCal, especially in the recent past.

Furthermore, these methods differ in the types of demographic models they can infer. PSMC, MSMC, and diCal v1 all infer piece-wise constant population size histories for a single panmictic population. CoalHMM and MSMC are capable of making inferences about multiple populations: CoalHMM fits simple parametric models, and MSMC performs non-parametric inference, reporting 'cross-coalescence rate' curves (CCRs). While CCRs have been interpreted in terms of divergence times [4••,49], an exploration of what models give rise to a particular CCR has not been performed: if the goal of a study is to fit a particular demographic model (e.g. a two population isolation migration model), CCR curves can be a useful diagnostic, but are difficult to interpret and cannot replace parametric model fitting. All of the coalescent-HMMs discussed here are summarized visually in Figure 1.

Recent advances

In response to the aforementioned shortcomings, there has been much progress in coalescent-HMM methodology. In particular, diCal version 2 allows for the parametric inference of more complex demographic models involving multiple populations, and SMC++ and ASMC push the boundaries of scalability for coalescent-HMMs.