Title

Inference of population history using coalescent HMMs: review and outlook.

Authors

Spence, Jeffrey P
Steinrücken, Matthias
Terhorst, Jonathan
et al.

Publication Date

2018-12-01

DOI

10.1016/j.gde.2018.07.002

Peer reviewed

HHS Public Access

Author manuscript

Curr Opin Genet Dev. Author manuscript; available in PMC 2019 December 01.

Published in final edited form as:

Curr Opin Genet Dev. 2018 December ; 53: 70–76. doi:10.1016/j.gde.2018.07.002.

Inference of Population History using Coalescent HMMs: Review and Outlook

Jeffrey P. Spence[a], **Matthias Steinrücken**[b], **Jonathan Terhorst**[c], and **Yun S. Song**[d,e,*]

[a]Computational Biology Graduate Group, University of California, Berkeley

[b]Department of Ecology and Evolution, University of Chicago

[c]Department of Statistics, University of Michigan

[d]Computer Science Division and Department of Statistics, University of California, Berkeley

[e]Chan Zuckerberg Biohub, San Francisco

Abstract

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.

1. 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 [2–4], 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

*To whom correspondence should be addressed: yss@berkeley.edu.

Disclosure

The authors declare no conflict of interest.

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 [17–21] 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 linkage information (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- 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.

2. 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 nonrecombining 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.

3. 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.

Building on diCal v1 [50] and advances to the CSD framework [54, 55], diCal v2 [56] was developed to perform parametric inference of essentially arbitrarily complex demographic models, including estimating divergence times, continuous and pulse migration, and population sizes with possible exponential growth. The method can scale to tens of haplotypes and has been used on models with three populations, but can handle arbitrarily many populations at increased computational cost. Like diCal v1, version 2 also considers a particular haplotype, and keeps track of when and with which other haplotype it first coalesces — these coalescence events tend to happen in the recent past making diCal well-powered to investigate recent history, such as the peopling of the Americas [7,]. diCal v2 has also been used in a hypothesis testing framework: in [57, Supplementary Information, section 18.4] a null model of a clean split between two populations was tested against a model of gene flow following that split. Furthermore, the CSD framework used by diCal v2 allows it to infer local ancestry or admixture, which was recently used to infer tracts of Neanderthal introgression in modern humans [58].

SMC++ [59] combines the scalability of SFS-based methods with the simplicity of PSMC. Like PSMC, it does not make assumptions beyond the SMC, and also does not require phased data. SMC++ tracks the coalescence time of a single “distinguished” pair of lineages, but then computes the likelihood of observing the sequence data of both the distinguished lineages and the rest of the sample. The simplicity of the hidden state allows SMC++ to scale to sample sizes in the hundreds, about an order of magnitude larger than any other coalescent-HMM presented above, giving it substantial power in both the recent and ancient past. It also achieves a substantial speedup by taking advantage of the fact that genotype data contain long stretches of non-segregating loci which may be effectively “skipped over” — an idea similar to [60]. Furthermore, instead of inferring piece-wise constant population sizes, SMC++ fits population sizes as smooth splines, reflecting a more realistic scenario of non-instantaneous population size changes. SMC++ is also capable of inferring divergence times for a pair of populations but makes the assumption that there was no migration after the populations diverged, which may not always be appropriate.

Recently, ASMC [61] extended SMC++ to genotype array data by accounting for SNP ascertainment bias. ASMC also takes advantage of certain symmetries when computing likelihoods in the underlying HMM to achieve extremely fast runtimes — an idea first

explored in [62]. Its speed allowed ASMC to be run on all pairs of haplotypes from 113,756 phased British individuals [2] although still at considerable computational cost.

To compare these methods, we performed a small simulation study shown in Figure 2. We considered four scenarios:

- A bottleneck.
- Constant size ($N_e = 10^4$).
- An isolation-with-migration model involving two populations.
- Exponential growth beginning 500 generations ago.

For each scenario, we used msprime [63] to simulate 10 replicate data sets each consisting of 30 haploids with eight 125 Mb chromosomes per haploid. The code used to simulate data and infer population sizes is fully reproducible and available at https://github.com/terhorst/coal_hmm_review.

4. Caveats, pitfalls, and best practices

Despite their power and flexibility, coalescent-HMMs are not without their pitfalls. All coalescent-HMMs contain tuning parameters that are crucial for good performance. A critical factor is the way that time is discretized. Finer discretization leads to a more accurate approximation, but the number of discretization points directly impacts the runtime, so care is needed to balance computational and accuracy considerations. Additionally, all of the methods discussed above, save SMC++, group adjacent loci and assume that they have the same genealogy. This assumption decreases the runtime substantially, but is certainly violated in practice. Depending on the method and application, it may be acceptable to perform the grouping at a kb scale, but care should be taken that such grouping does not influence the results. Furthermore, the likelihoods optimized by coalescent-HMMs — and demographic inference methods more broadly — tend to have many local optima. Thus, different initializations of the methods will likely yield different results, making it crucial to take the best of several runs, seeded with different initializations, as the final inferred model.

Users should also be careful about model choice. As an example, SMC++ infers population splits in the absence of gene flow. If there has been pervasive migration between the populations of interest, then the model inferred by SMC++ will not be reflective of reality. Additionally, even seemingly non-parametric methods, like PSMC, make implicit assumptions such as the data coming from a single panmictic, neutrally evolving population. Recent studies [64, 65] used simulated data to investigate these model violations and showed that pervasive selective sweeps or population structure bias coalescent-HMMs. Another study [66] showed that when applied to simulated data, coalescent-HMMs infer models that have an expected SFS similar to that of the data, but when applied to real data the SFS of the inferred models does not match that of the data. This suggests that real data violate the idealized models that are commonly used for simulation and inference.

We also urge caution in over-interpreting the results of any demographic inference method. For instance, all methods infer “effective population sizes”, defined as the inverse

coalescence rate for a pair of haplotypes. Under many models effective population size is correlated with census population size, but does not need to be; e.g., a structured population will have a larger effective size than a panmictic population of the same census size.

To avoid the aforementioned pitfalls, we recommend using multiple methods utilizing different aspects of the data, such as frequency-based methods *and* coalescent-HMMs. While the exact models inferred will differ between methods, one can have some confidence in aspects of the model that are robustly inferred across methods. We also recommend using the results of either a pilot run of the coalescent-HMM or the results of another method (or even PCA [67, 68], or STRUCTURE-like programs [69–72]) to inform model selection — e.g., if the data appear to come from unadmixed populations based on this initial fit, it may be appropriate to assume a clean split model instead of modeling gene flow. After fitting a model, it is crucial to measure goodness-of-fit, for example by comparing the SFS and MSMC's CCR curves for data simulated from the inferred models to those computed directly from the real data.

It is also important to understand sources of bias and noise present in data. Because most coalescent- HMMs make use of both segregating and non-segregating sites it is crucial to use "masks" indicating which regions of the genome have been reliably genotyped. Additionally, when working with ancient DNA showing an excess of transitions due to postmortem cytosine deamination [73], we have found that restricting analysis to only transversions and adjusting the mutation rate correspondingly improves inference.

Finally, as with any statistical analysis, it is important to study uncertainty in the inferred model, e.g., by bootstrapping, either parametric via simulation or non-parametric by resampling the data as in [48]. While parametric bootstrapping is more straightforward, it is only capable of estimating uncertainty in the estimation procedure, whereas non-parametric bootstrapping captures uncertainty in both modeling and estimation, but cannot reveal bias in the estimates. Note that in demographic inference, bootstrapping does not produce statistically valid confidence intervals, if the data are used to perform model selection prior to estimating statistical uncertainty. However, providing some quantification of uncertainty is still important.

5. Outlook

While there has been much recent work on improving the flexibility, and computational and statistical efficiency of coalescent-HMMs, there are still a number of open problems and interesting directions for future research.

As alluded to above, when the sample size is greater than 2, every coalescent-HMM tracks only a part of the genealogy of the whole sample. Such choices are based on intuition and are made primarily for analytic convenience to ensure computational tractability. Tree length has recently been explored as such a choice [74]. Finding optimal ways of encoding genealogical information in a small number of discrete parameters remains a challenging open problem.

While coalescent-HMMs work extremely well on simulated data, they, like most inference methods in population genetics, are less stable on real data [66]. This is likely due to rampant model misspecification: coalescent-HMMs make many unrealistic assumptions, such as assuming constant recombination [75, 76] and mutation [77–79] rates across the genome. In addition, all methods must simplify the “true” demographic model: reality is always more complicated than any model with a handful of parameters, presenting a need for more robust methods.

A major challenge, especially in studying non-model organisms, is that with the exception of PSMC and SMC++, coalescent-HMMs are currently unable to handle unphased data. Overcoming this challenge is an important task for future methods.

Lastly, despite their excellent behavior in practice, our understanding of coalescent-HMMs is based entirely on intuition and numerical experiments. In contrast to frequency-based methods, which have a rich literature on their theoretical properties [24–28], coalescent-HMMs are poorly understood from a theoretical perspective. While there has been some work on how accurately demographic history can be inferred directly from genealogies [80, 81], in the more realistic coalescent-HMM setting even the basic question of whether demographic models are statistically identifiable is unanswered

Acknowledgments

This work is supported in part by an NIH grant R01-GM094402, and a Packard Fellowship for Science and Engineering. Y.S.S. is a Chan Zuckerberg Biohub investigator.

References

[1]. Cavalli-Sforza L. Luca, Menozzi Paolo, and Piazza Alberto. The History and Geography of Human Genes. Princeton paperbacks Princeton University Press, 1996.

[2]. Sudlow Cathie, Gallacher John, Allen Naomi, Beral Valerie, Burton Paul, Danesh John, Downey Paul, Elliott Paul, Green Jane, Landray Martin, et al. UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLOS Medicine, 12(3):1–10, 2015.

[3]. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature, 526:68–74, 2015. [PubMed: 26432245]

[4]. Mallick Swapan, Li Heng, Lipson Mark, Mathieson Iain, Gymrek Melissa, Racimo Fernando, Zhao Mengyao, Chennagiri Niru, Nordenfelt Susanne, Tandon Arti, et al. The Simons Genome Diversity Project: 300 genomes from 142 diverse populations. Nature, 538(7624):201–206, 2016 ** (of outstanding interest): One of the most extensive and diverse set of human genomes to date. Includes samples from 142 populations located throughout the world. [PubMed: 27654912]

[5]. Mathieson Iain, Lazaridis Iosif, Rohland Nadin, Mallick Swapan, Patterson Nick, Roodenberg Songül Alpaslan, Harney Eadaoin, Stewardson Kristin, Fernandes Daniel, Novak Mario, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature, 528(7583):499–503, 2015. [PubMed: 26595274]

[6]. Moorjani Priya, Thangaraj Kumarasamy, Patterson Nick, Lipson Mark, Loh Po-Ru, Govin- daraj Periyasamy, Berger Bonnie, Reich David, and Singh Lalji. Genetic evidence for recent population mixture in India. American Journal of Human Genetics, 93(3):422–438, 2013. [PubMed: 23932107]

[7]. Raghavan Maanasa, Steinrücken Matthias, Harris Kelley, Schiffels Stephan, Rasmussen Simon, DeGiorgio Michael, Albrechtsen Anders, Valdiosera Cristina, Ávila-Arcos María C., Malaspinas

Anna-Sapfo, et al. Genomic evidence for the Pleistocene and recent population history of Native Americans. Science, 349(6250), 2015.

[8]. Malaspinas Anna-Sapfo, Westaway Michael C., Muller Craig, Sousa Vitor C., Lao Oscar, Alves Isabel, Bergström Anders, Athanasiadis Georgios, Cheng Jade Y., Crawford Jacob E., et al. A genomic history of Aboriginal Australia. Nature, 538:207–214, 2016. [PubMed: 27654914]

[9]. vonHoldt Bridgett M., Pollinger John P., Lohmueller Kirk E., Han Eunjung, Parker Heidi G., Quignon Pascale, Degenhardt Jeremiah D., Boyko Adam R., Earl Dent A., Auton Adam, et al. Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication. Nature, 464:898–902, 2010. [PubMed: 20237475]

[10]. Warmuth Vera, Eriksson Anders, Bower Mim Ann Graeme Barker, Barrett Elizabeth, Hanks Bryan Kent, Li Shuicheng, Lomitashvili David, Ochir-Goryaeva Maria, Sizonov Grigory V., et al. Reconstructing the origin and spread of horse domestication in the Eurasian steppe. Proceedings of the National Academy of Sciences, 109(21):8202–8206, 2012.

[11]. Frantz Laurent A. F., Schraiber Joshua G., Madsen Ole, Megens Hendrik-Jan, Cagan Alex, Bosse Mirte, Paudel Yogesh, Crooijmans Richard P. M. A., Larson Greger, and Groenen Martien A. M. Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes. Nature Genetics, 47:1141–1148, 2015. [PubMed: 26323058]

[12]. Nielsen Rasmus, Williamson Scott, Kim Yuseob, Hubisz Melissa J., Clark Andrew G., and Bustamante Carlos. Genomic scans for selective sweeps using SNP data. Genome Research, 15(11):1566–1575, 2005. [PubMed: 16251466]

[13]. Mathieson Iain and McVean Gil. Differential confounding of rare and common variants in spatially structured populations. Nature Genetics, 44:243–246, 2012. [PubMed: 22306651]

[14]. Johnston Henry R. and Cutler David J. Population demographic history can cause the appearance of recombination hotspots. The American Journal of Human Genetics, 90(5):774–783, 2012. [PubMed: 22560089]

[15]. Kamm John A., Spence Jeffrey P., Chan Jeffrey, and Song Yun S. Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. Genetics, 203(3):1381–1399, 2016. [PubMed: 27182948]

[16]. Mays Herman L., Jr., Hung Chih-Ming, Shaner Pei-Jen, Denvir James, Justice Megan, Yang Shang-Fang, Roth Terri L., Oehler David A., Fan Jun, Rekulapally Swanthana, and Primerano Donald A. Genomic analysis of demographic history and ecological niche modeling in the endangered Sumatran rhinoceros Dicerorhinus sumatrensis. Current Biology, 28(1):70–76.e4, 2018. [PubMed: 29249659]

[17]. Gutenkunst Ryan N., Hernandez Ryan D., Williamson Scott H., and Bustamante Carlos D. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLOS Genetics, 5(10):e1000695, 2009. [PubMed: 19851460]

[18]. Excoffier Laurent, Dupanloup Isabelle, Huerta-Sánchez Emilia, Sousa Vitor C., and Foll Matthieu. Robust demographic inference from genomic and SNP data. PLOS Genetics, 9(10):1–17, 2013.

[19]. Anand Bhaskar YX Wang Rachel, and Song Yun S. Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data. Genome Research, 25(2):268–279, 2015. [PubMed: 25564017]

[20]. Jouganous Julien, Long Will, Ragsdale Aaron P., and Gravel Simon. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. Genetics, 206(3):1549–1567, 2017 * (of special interst): Uses a sparse approximation to Wright-Fisher dynamics to efficiently compute the SFS for multiple populations allowing for possible selection. [PubMed: 28495960]

[21]. Kamm John A., Terhorst Jonathan, Durbin Richard, and Song Yun S. Efficiently inferring the demographic history of many populations with allele count data. bioRxiv, 2018 10.1101/287268. * (of special interest): Presents an extremely efficient method to compute the expected frequency spectrum of many populations, extending the applicability of frequency-based methods sample sizes in the hundreds for tens of populations.

[22]. Waltoft Berit Lindum and Hobolth Asger. Non-parametric estimation of population size changes from the site frequency spectrum. Stat Appl Genet Mol Biol, 17(3), 2018.

[23]. Ragsdale Aaron P. and Gutenkunst Ryan N. Inferring demographic history using two-locus statistics. Genetics, 206(2):1037–1048, 2017. [PubMed: 28413158]

[24]. Myers Simon, Fefferman Charles, and Patterson Nick. Can one learn history from the allelic spectrum? Theoretical Population Biology, 73(3):342–348, 2008. [PubMed: 18321552]

[25]. Bhaskar Anand and Song Yun S. Descartes' rule of signs and the identifiability of population demo-graphic models from genomic variation data. Annals of Statistics, 42(6):2469–2493, 2014. [PubMed: 28018011]

[26]. Terhorst Jonathan and Song Yun S. Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum. Proceedings of the National Academy of Sciences, 112(25):7677–7682, 2015.

[27]. Baharian Soheil and Gravel Simon. On the decidability of population size histories from finite allele frequency spectra. Theoretical Population Biology, 120:42–51, 2018 * (of special interest): presents classes of piece-wise constant population size histories that are qualitatively and quantitatively dissimilar but produce provably similar frequency spectra. [PubMed: 29305873]

[28]. Rosen Zvi, Bhaskar Anand, Roch Sebastien, and Song Yun S. Geometry of the sample frequency spectrum and the perils of demographic inference. bioRxiv, 2017 10.1101/233908.

[29]. Palamara Pier Francesco, Lencz Todd, Darvasi Ariel, and Pe'er Itsik. Length distributions of identity by descent reveal fine-scale demographic history. The American Journal of Human Genetics, 91(5):809–822, 2012. [PubMed: 23103233]

[30]. Palamara Pier Francesco and Pe'er Itsik. Inference of historical migration rates via haplotype sharing. Bioinformatics, 29(13):i180–i188, 2013. [PubMed: 23812983]

[31]. Browning Sharon R. and Browning Brian L. Accurate non-parametric estimation of recent effective population size from segments of identity by descent. The American Journal of Human Genetics, 97(3):404–418, 2015. [PubMed: 26299365]

[32]. Gusev Alexander, Lowe Jennifer K., Stoffel Markus, Daly Mark J., Altshuler David, Breslow Jan L., Friedman Jeffrey M., and Pe'er Itsik. Whole population, genome-wide mapping of hidden relatedness. Genome Research, 19(2):318–326, 2009. [PubMed: 18971310]

[33]. Browning Brian L. and Browning Sharon R. Detecting identity by descent and estimating genotype error rates in sequence data. The American Journal of Human Genetics, 93(5):840–851, 2013. [PubMed: 24207118]

[34]. Tataru Paula, Nirody Jasmine A., and Song Yun S. diCal-IBD: demography-aware inference of identity- by-descent tracts in unrelated individuals. Bioinformatics, 30(23):3430–3431, 2014. [PubMed: 25147361]

[35]. Harris Kelley and Nielsen Rasmus. Inferring demographic history from a spectrum of shared haplotype lengths. PLOS Genetics, 9(6):1–20, 2013.

[36]. Wiuf Carsten and Hein Jotun. Recombination as a point process along sequences. Theoretical Population Biology, 55(3):248–259, 1999. [PubMed: 10366550]

[37]. Kingman John F. C. The coalescent. Stochastic Processes and their Applications, 13(3):235–248, 1982.

[38]. Griffiths Robert C. and Marjoram Paul. Ancestral inference from samples of DNA sequences with recombination. Journal of Computational Biology, 3(4):479–502, 1996 PMID: . [PubMed: 9018600]

[39]. McVean Gilean A.T. and Cardin Niall J. Approximating the coalescent with recombination. Philosophical Transactions of the Royal Society London B: Biological Sciences, 360:1387–93, 2005.

[40]. Marjoram Paul and Wall Jeff D. Fast "coalescent" simulation. BMC Genetics, 7(1):16, 2006. [PubMed: 16539698]

[41]. Hobolth Asger and Jensen Jens Ledet. Markovian approximation to the finite loci coalescent with recombination along multiple sequences. Theoretical Population Biology, 98:48–58, 2014. [PubMed: 24486389]

[42]. Wilton Peter R., Carmi Shai, and Hobolth Asger. The SMC' is a highly accurate approximation to the ancestral recombination graph. Genetics, 200(1):343–355, 2015. [PubMed: 25786855]

[43]. Rabiner Lawrence R. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, 77(2):257–286, 1989.

[44]. Kalman Rudolph Emil. A new approach to linear filtering and prediction problems. Transactions of the ASME-Journal of Basic Engineering, 82(Series D):35–45, 1960.

[45]. Beal Matthew J., Ghahramani Zoubin, and Rasmussen Carl E. The infinite hidden Markov model In Dietterich TG, Becker S, and Ghahramani Z, editors, Advances in Neural Information Processing Systems 14, pages 577–584. MIT Press, 2002.

[46]. Dutheil Julien Y., Ganapathy Ganesh, Hobolth Asger, Mailund Thomas, Uyenoyama Marcy K., and Schierup Mikkel H. Ancestral population genomics: The coalescent hidden Markov model approach. Genetics, 183(1):259–274, 2009. [PubMed: 19581452]

[47]. Mailund Thomas, Halager Anders E., and Westergaard Michael. Using colored petri nets to construct coalescent hidden Markov models: Automatic translation from demographic specifications to efficient inference methods In Haddad Serge and Pomello Lucia, editors, Application and Theory of Petri Nets, pages 32–50, Berlin, Heidelberg, 2012 Springer Berlin Heidelberg.

[48]. Li Heng and Durbin Richard. Inference of human population history from individual whole-genome sequences. Nature, 475:493–496, 2011. [PubMed: 21753753]

[49]. Schiffels Stephan and Durbin Richard. Inferring human population size and separation history from multiple genome sequences. Nature Genetics, 46:919–925, 2014. [PubMed: 24952747]

[50]. Sheehan Sara, Harris Kelley, and Song Yun S. Estimating variable effective population sizes from multiple genomes: A sequentially Markov conditional sampling distribution approach. Genetics, 194(3):647–662, 2013. [PubMed: 23608192]

[51]. Li Na and Stephens Matthew. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics, 165(4):2213–2233, 2003. [PubMed: 14704198]

[52]. Paul Joshua S. and Song Yun S. A principled approach to deriving approximate conditional sampling distributions in population genetics models with recombination. Genetics, 186(1):321–338, 2010. [PubMed: 20592264]

[53]. Paul Joshua S., Steinriicken Matthias, and Song Yun S. An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination. Genetics, 187(4):1115–1128, 2011. [PubMed: 21270390]

[54]. Davison Dan, Pritchard Jonathan K., and Coop Graham. An approximate likelihood for genetic data under a model with recombination and population splitting. Theoretical Population Biology, 75(4):331–345, 2009. [PubMed: 19362099]

[55]. Steinracken Matthias, Paul Joshua S., and Song Yun S. A sequentially Markov conditional sampling distribution for structured populations with migration and recombination. Theoretical Population Biology, 87:51–61, 2013. [PubMed: 23010245]

[56]. Steinriicken Matthias, Kamm John A., and Song Yun S. Inference of complex population histories using whole-genome sequences from multiple populations. bioRxiv, 2015.

[57]. Moreno-Mayar J. Víctor, Potter Ben A., Vinner Lasse, Steinrücken Matthias, Rasmussen Simon, Terhorst Jonathan, Kamm John A., Albrechtsen Anders, Malaspinas Anna-Sapfo, Sikora Martin, et al. Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans. Nature, 553:203–207, 2018 * (of special interest): Studies the peopling of the Americas, making use of ancient genomes and combining frequency-based and coalescent-HMM methods for robust demographic inference. [PubMed: 29323294]

[58]. Steinrücken Matthias, Spence Jeffrey P, Kamm John A., Wieczorek Emilia, and Song Yun S. Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans. Molecular Ecology, 2018.

[59]. Terhorst Jonathan, Kamm John A., and Song Yun S. Robust and scalable inference of population history from hundreds of unphased whole genomes. Nature Genetics, 49:303–309, 2017 ** (of outstanding interest): Presents a coalescent-HMM that essentially combines PSMC with frequency-based methods for a powerful, yet scalable tool for demographic inference. [PubMed: 28024154]

[60]. Paul Joshua S. and Song Yun S. Blockwise HMM computation for large-scale population genomic inference. Bioinformatics, 28(15):2008–2015, 2012. [PubMed: 22641715]

[61]. Palamara Pier Francesco, Terhorst Jonathan, Song Yun S., and Price Alkes L. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. Nature Genetics, 2018 In press. * (of special interest): Extends ideas from SMC++ to data from genotype array data sets and presents the largest-scale application of coalescent-HMMs to date.

[62]. Harris Kelley, Sheehan Sara, Kamm John A., and Song Yun S. Decoding coalescent hidden Markov models in linear time In Sharan Roded, editor, Research in Computational Molecular Biology, pages 100–114, Cham, 2014 Springer International Publishing.

[63]. Kelleher Jerome, Etheridge Alison M, and McVean Gilean. Efficient coalescent simulation and genealogical analysis for large sample sizes. PLOS Computational Biology, 12(5):e1004842, 2016 ** (of outstanding interest): Presents msprime: simulation software capable of simulating data under the full coalescent with recombination orders of magnitude faster than other simulators. [PubMed: 27145223]

[64]. Hawks John. Introgression makes waves in inferred histories of effective population size. Human Biology, 89(1):67–80, 2017. [PubMed: 29285970]

[65]. Schrider Daniel R., Shanku Alexander G., and Kern Andrew D. Effects of linked selective sweeps on demographic inference and model selection. Genetics, 204(3):1207–1223, 2016. [PubMed: 27605051]

[66]. Beichman Annabel C., Phung Tanya N., and Lohmueller Kirk E. Comparison of single genome and allele frequency data reveals discordant demographic histories. G3: Genes, Genomes, Genetics, 7(11):3605–3620, 2017. [PubMed: 28893846]

[67]. Price Alkes L, Patterson Nick J, Plenge Robert M, Weinblatt Michael E, Shadick Nancy A, and Reich David. Principal components analysis corrects for stratification in genome-wide association studies. Nature Genetics, 38:904–909, 2006. [PubMed: 16862161]

[68]. Novembre John, Johnson Toby, Bryc Katarzyna, Kutalik Zoltán, Boyko Adam R., Auton Adam, Indap Amit, King Karen S., Bergmann Sven, Nelson Matthew R., et al. Genes mirror geography within Europe. Nature, 456:98–101, 2008. [PubMed: 18758442]

[69]. Pritchard Jonathan K., Stephens Matthew, and Donnelly Peter. Inference of population structure using multilocus genotype data. Genetics, 155(2):945–959, 2000. [PubMed: 10835412]

[70]. Alexander David H., Novembre John, and Lange Kenneth. Fast model-based estimation of ancestry in unrelated individuals. Genome Research, 2009.

[71]. Raj Anil, Stephens Matthew, and Pritchard Jonathan K. fastSTRUCTURE: Variational inference of population structure in large SNP data sets. Genetics, 197(2):573–589, 2014. [PubMed: 24700103]

[72]. Cabreros Irineo and Storey John D. A nonparametric estimator of population structure unifying admixture models and principal components analysis. bioRxiv, 2017 10.1101/240812.

[73]. Dabney Jesse, Meyer Matthias, and Pääbo Svante. Ancient DNA damage. Cold Spring Harbor Perspectives in Biology, 5(7):a012567, 2013. [PubMed: 23729639]

[74]. Miroshnikov Alexey and Steinrücken Matthias. Computing the joint distribution of the total tree length across loci in populations with variable size. Theoretical Population Biology, 118:1–19, 2017. [PubMed: 28943126]

[75]. Myers Simon, Bottolo Leonardo, Freeman Colin, Gil McVean, and Peter Donnelly. A fine-scale map of recombination rates and hotspots across the human genome. Science, 310(5746):321–324, 2005. [PubMed: 16224025]

[76]. Kong Augustine, Thorleifsson Gudmar, Gudbjartsson Daniel F., Masson Gisli, Sigurdsson Asgeir, Jonasdottir Aslaug, Walters G. Bragi, Jonasdottir Adalbjorg, Gylfason Arnaldur, Kristinsson Kari Th., et al. Fine-scale recombination rate differences between sexes, populations and individuals. Nature, 467(7319):1099–1103, 2010. [PubMed: 20981099]

[77]. Kong Augustine, Frigge Michael L., Masson Gisli, Besenbacher Soren, Sulem Patrick, Magnusson Gisli, Gudjonsson Sigurjon A., Sigurdsson Asgeir, Jonasdottir Aslaug, Jonasdottir Adalbjorg, et al. Rate of *de novo* mutations and the importance of father's age to disease risk. Nature, 488:471–475, 2012. [PubMed: 22914163]

[78]. Jónsson Hákon, Sulem Patrick, Kehr Birte, Kristmundsdottir Snaedis, Zink Florian, Hjartarson Eirikur, Hardarson Marteinn T., Hjorleifsson Kristjan E., Eggertsson Hannes P., Gudjonsson

Sigurjon Axel, et al. Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland. Nature, 549:519–522, 2017. [PubMed: 28959963]

[79]. Smith Thomas C. A., Arndt Peter F., and Eyre-Walker Adam. Large scale variation in the rate of germ-line *de novo* mutation, base composition, divergence and diversity in humans. PLOS Genetics, 14(3):1–30, 2018.

[80]. Kim Junhyong, Mossel Elchanan, Rácz Miklós Z., and Ross Nathan. Can one hear the shape of a population history? Theoretical Population Biology, 100:26–38, 2015.

[81]. Johndrow James E. and Palacios Julia A. Exact limits of inference in coalescent models. ArXiv e-prints, 2017.

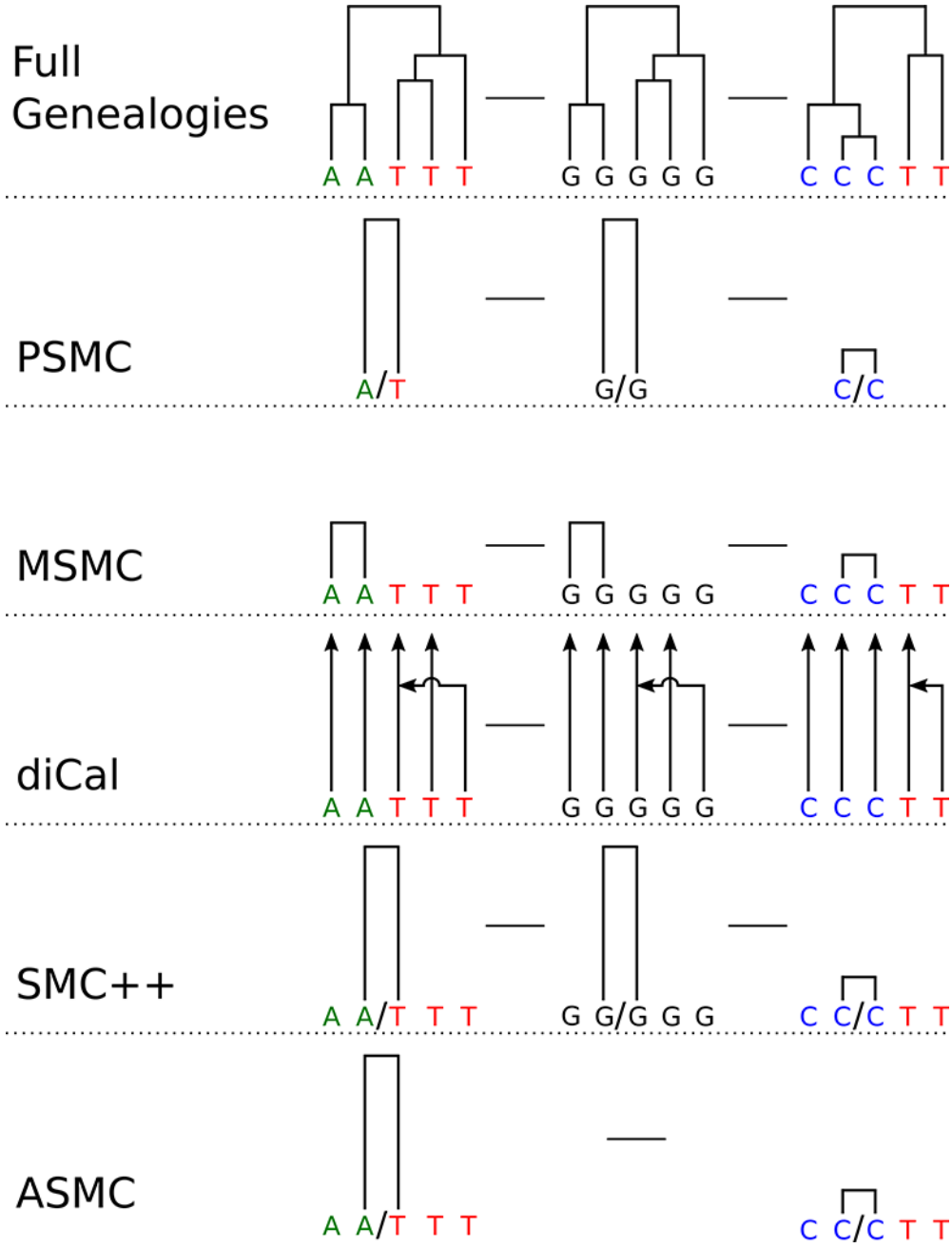

Figure 1:

The sequentially Markovian coalescent views the genealogy relating a sample of individuals as a sequence of trees along the genome. The number of possible trees relating a sample grows super exponentially with sample size, making such a model computationally intractable for inference. The commonly used coalescent-HMMs make various simplifications to this full process. PSMC, SMC++, and ASMC only track the genealogy of a “distinguished” pair of haplotypes. PSMC ignores the rest of the sample, while SMC++ and ASMC use the other samples to inform the genealogy of the distinguished pair. ASMC was designed to work on genotype array data and so skips over sites not included on the

array (middle genealogy). MSMC tracks only the most recent coalescence event in the whole sample, while diCal tracks the first coalescence event involving a particular haplotype.

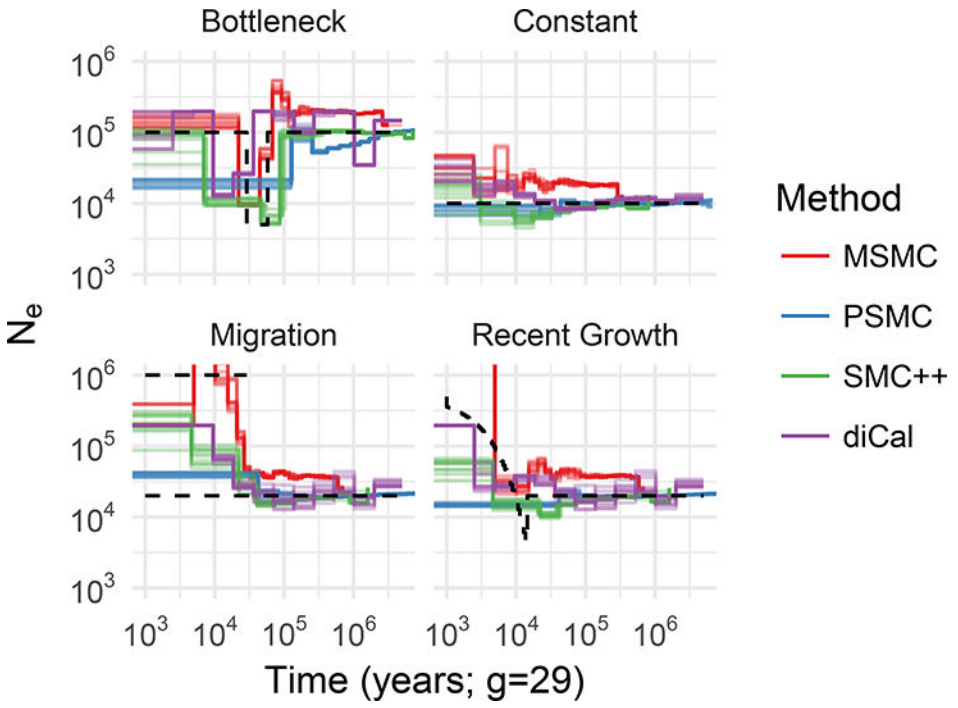

Figure 2:

Performance of various coalescent-HMMs on simulated data. The scenarios considered here are: a population experiencing a sharp bottleneck; a single panmictic population of constant size; samples from a large population that is exchanging migrants with a smaller population; and a population that has recently experienced exponential growth. Each scenario has 10 replicate data sets, with each data set containing 30 haploids with eight 125 Mb chromosomes per haploid. PSMC was run with the options '-N 25 -p 4+20*3+4' on a single pair of haploids. MSMC was run with the default hyperparameter settings with the 'fixedRecombination' flag, using only 4 of the 30 haploids. The same 4 haploids were used for diCal v2, and inference was performed by taking the composite likelihood over all pairs of those 4 haplotypes, and running 30 EM iterations. SMC++ was run with the '-timepoints 33' and '-thinning 500' options.