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 **Brenna M Henn** and **Lluis Quintana-Murci**

For a complete overview see the Issue and the Editorial

Available online 26th July 2018

https://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.

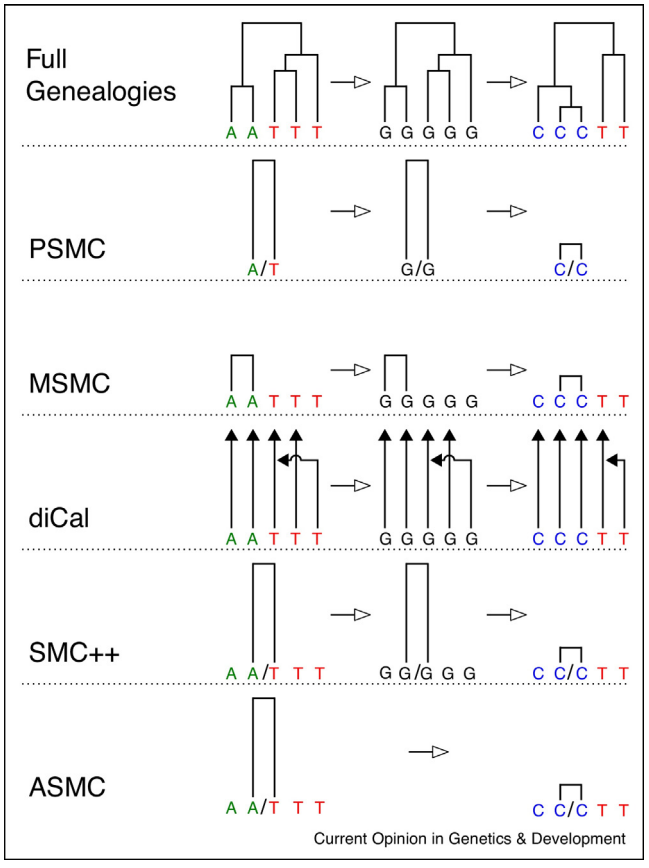
Briefly, CoalHMM [46,47], developed to study different species, tracks only the topology of the genealogy and in which branch of the species tree the lineages coalesce. CoalHMM cannot scale to more than a few species. PSMC [48] can only be applied to a pair of haplotypes, but tracks their genealogy exactly, 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.

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.

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 [7,57•].

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.

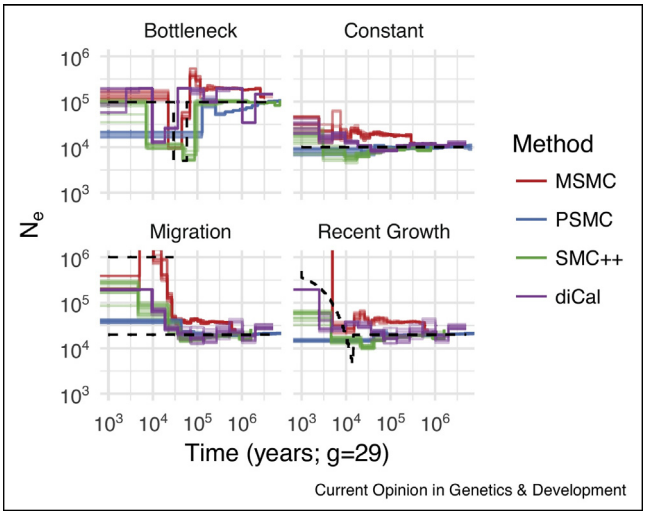
Recently, ASMC [61•] extended SMC++ to genotype array data by accounting for SNP ascertainment bias. ASMC also takes advantage of certain symmetries when computing likelihoods in the underlying HMM to achieve extremely fast runtimes — an idea first explored in [62]. Its speed allowed ASMC to be run on all pairs of haplotypes from 113,756 phased British individuals [[61•]] 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

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 four haploids were used for diCal v2, and inference was performed by taking the composite likelihood over all pairs of those four haplotypes, and running 30 EM iterations. SMC++ was run with the '–timepoints 33' and '–thinning 500' options.

reproducible and available at https://github.com/terhorst/coal_hmm_review.

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; for example, 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 — for example, 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, for example, 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.

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 [24–26,27•,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.

Conflict of interest statement

Nothing declared.

Acknowledgements

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 and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Cavalli-Sforza LL, Menozzi P, Piazza A: *The History and Geography of Human Genes. Princeton paperbacks*. Princeton University Press; 1996.

2. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M *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 Med* 2015, **12**:1-10.

3. The 1000 Genomes Project Consortium: **A global reference for human genetic variation**. *Nature* 2015, **526**:68-74.

4. •• Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, Zhao M, Chennagiri N, Nordenfelt S, Tandon A *et al.*: **The Simons Genome Diversity Project: 300 genomes from 142 diverse populations**. *Nature* 2016, **538**:201-206.
One of the most extensive and diverse set of human genomes to date. Includes samples from 142 populations located throughout the world.

5. Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, Harney E, Stewardson K, Fernandes D, Novak M *et al.*: **Genome-wide patterns of selection in 230 ancient Eurasians**. *Nature* 2015, **528**:499-503.

6. Moorjani P, Thangaraj K, Patterson N, Lipson M, Loh P-R, Govindaraj P, Berger B, Reich D, Singh L: **Genetic evidence for recent population mixture in India**. *Am J Hum Genet* 2013, **93**:422-438.

7. Raghavan M, Steinrücken M, Harris K, Schiffels S, Rasmussen S, DeGiorgio M, Albrechtsen A, Valdiosera C, Ávila-Arcos MC, Malaspinas A-S *et al.*: **Genomic evidence for the Pleistocene and recent population history of Native Americans**. *Science* 2015, **349**.

8. Malaspinas A-S, Westaway MC, Muller C, Sousa VC, Lao O, Alves I, Bergström A, Athanasiadis G, Cheng JY, Crawford JE *et al.*: **A genomic history of Aboriginal Australia**. *Nature* 2016, **538**:207-214.

9. vonHoldt BM, Pollinger JP, Lohmueller KE, Han E, Parker HG, Quignon P, Degenhardt JD, Boyko AR, Earl DA, Auton A *et al.*: **Genome-wide SNP and haplotype analyses reveal a rich history underlying dog domestication**. *Nature* 2010, **464**:898-902.

10. Warmuth V, Eriksson A, Bower MA, Barker G, Barrett E, Hanks BK, Li S, Lomitashvili D, Ochir-Goryaeva M, Sizonov GV *et al.*: **Reconstructing the origin and spread of horse domestication in the Eurasian steppe**. *Pro Natl Acad Sci U S A* 2012, **109**:8202-8206.

11. Frantz LAF, Schraiber JG, Madsen O, Megens H-J, Cagan A, Bosse M, Paudel Y, Crooijmans RPMA, Larson G, Groenen MAM: **Evidence of long-term gene flow and selection during domestication from analyses of Eurasian wild and domestic pig genomes**. *Nat Genet* 2015, **47**:1141-1148.

12. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C: **Genomic scans for selective sweeps using SNP data**. *Genome Res* 2005, **15**:1566-1575.

13. Mathieson I, McVean G: **Differential confounding of rare and common variants in spatially structured populations**. *Nat Genet* 2012, **44**:243-246.

14. Johnston HR, Cutler DJ: **Population demographic history can cause the appearance of recombination hotspots**. *Am J Hum Genet* 2012, **90**:774-783.

15. Kamm JA, Spence JP, Chan J, Song YS: **Two-locus likelihoods under variable population size and fine-scale recombination rate estimation**. *Genetics* 2016, **203**:1381-1399.

16. Mays HL Jr, Hung C-M, Shaner P-J, Denvir J, Justice M, Yang S-F, Roth TL, Oehler DA, Fan J, Rekulapally S, Primerano DA: **Genomic analysis of demographic history and ecological niche modeling in the endangered Sumatran rhinoceros *Dicerorhinus sumatrensis***. *Curr Biol* 2018, **28** 70–76.e4.

17. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD: **Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data**. *PLOS Genet* 2009, **5**:e1000695.

18. Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa VC, Foll M: **Robust demographic inference from genomic and SNP data**. *PLOS Genet* 2013, **9**:1-17.

19. Bhaskar A, Wang YXR, Song YS: **Efficient inference of population size histories and locus-specific mutation rates from large-sample genomic variation data**. *Genome Res* 2015, **25**:268-279.

20. • Jouganous J, Long W, Ragsdale AP, Gravel S: **Inferring the joint demographic history of multiple populations: beyond the diffusion approximation**. *Genetics* 2017, **206**:1549-1567.
Uses a sparse approximation to Wright-Fisher dynamics to efficiently compute the SFS for multiple populations allowing for possible selection.

21. • Kamm JA, Terhorst J, Durbin R, Song YS: **Efficiently inferring the demographic history of many populations with allele count data**. *bioRxiv* 2018 http://dx.doi.org/10.1101/287268.
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 BL, Hobolth A: **Non-parametric estimation of population size changes from the site frequency spectrum**. *Stat Appl Genet Mol Biol* 2018, **17**.

23. Ragsdale AP, Gutenkunst RN: **Inferring demographic history using two-locus statistics**. *Genetics* 2017, **206**:1037-1048.

24. Myers S, Fefferman C, Patterson N: **Can one learn history from the allelic spectrum?** *Theor Popul Biol* 2008, **73**:342-348.

25. Bhaskar A, Song YS: **Descartes' rule of signs and the identifiability of population demographic models from genomic variation data**. *Ann Stat* 2014, **42**:2469-2493.

26. Terhorst J, Song YS: **Fundamental limits on the accuracy of demographic inference based on the sample frequency spectrum**. *Proc Natl Acad Sci U S A* 2015, **112**:7677-7682.

27. • Baharian S, Gravel S: **On the decidability of population size histories from finite allele frequency spectra**. *Theor Popul Biol* 2018, **120**:42-51.
Presents classes of piece-wise constant population size histories that are qualitatively and quantitatively dissimilar but produce provably similar frequency spectra.

28. Rosen Z, Bhaskar A, Roch S, Song YS: **Geometry of the sample frequency spectrum and the perils of demographic inference**. *bioRxiv* 2017 http://dx.doi.org/10.1101/233908.

29. Palamara PF, Lencz T, Darvasi A, Pe'er I: **Length distributions of identity by descent reveal fine-scale demographic history**. *Am J Hum Genet* 2012, **91**:809-822.

30. Palamara PF, Pe'er I: **Inference of historical migration rates via haplotype sharing**. *Bioinformatics* 2013, **29**:i180-i188.

31. Browning SR, Browning BL: **Accurate non-parametric estimation of recent effective population size from segments of identity by descent**. *Am J Hum Genet* 2015, **97**:404-418.

32. Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I: **Whole population, genome-wide mapping of hidden relatedness**. *Genome Res* 2009, **19**:318-326.

33. Browning BL, Browning SR: **Detecting identity by descent and estimating genotype error rates in sequence data**. *Am J Hum Genet* 2013, **93**:840-851.

34. Tataru P, Nirody JA, Song YS: **diCal-IBD: demography-aware inference of identity-by-descent tracts in unrelated individuals**. *Bioinformatics* 2014, **30**:3430-3431.

35. Harris K, Nielsen R: **Inferring demographic history from a spectrum of shared haplotype lengths**. *PLOS Genet* 2013, **9**:1-20.

36. Wiuf C, Hein J: **Recombination as a point process along sequences**. *Theor Popul Biol* 1999, **55**:248-259.

37. Kingman JFC: **The coalescent**. *Stochast Process Appl* 1982, **13**:235-248.

38. Griffiths RC, Marjoram P: **Ancestral inference from samples of DNA sequences with recombination**. *J Comput Biol* 1996, **3**:479-502 PMID: 9018600.

39. McVean GAT, Cardin NJ: **Approximating the coalescent with recombination**. *Philos Trans R Soc Lond B Biol Sci* 2005, **360**:1387-1393.

40. Marjoram P, Wall JD: **Fast "coalescent" simulation**. *BMC Genet* 2006, **7**:16.

41. Hobolth A, Jensen JL: **Markovian approximation to the finite loci coalescent with recombination along multiple sequences**. *Theor Popul Biol* 2014, **98**:48-58.

42. Wilton PR, Carmi S, Hobolth A: **The SMC' is a highly accurate approximation to the ancestral recombination graph**. *Genetics* 2015, **200**:343-355.

43. Rabiner LR: **A tutorial on hidden Markov models and selected applications in speech recognition**. *Proc IEEE* 1989, **77**:257-286.

44. Kalman RE: **A new approach to linear filtering and prediction problems**. *Trans ASME – J Basic Eng* 1960, **82(Series D)**:35-45.

45. Beal MJ, Ghahramani Z, Rasmussen CE: **The infinite hidden Markov model**. In *Advances in Neural Information Processing Systems*, vol. 14. Edited by Dietterich TG, Becker S, Ghahramani Z . MIT Press; 2002:577-584.

46. Dutheil JY, Ganapathy G, Hobolth A, Mailund T, Uyenoyama MK, Schierup MH: **Ancestral population genomics: the coalescent hidden Markov model approach**. *Genetics* 2009, **183**:259-274.

47. Mailund T, Halager AE, Westergaard M: **Using colored petri nets to construct coalescent hidden Markov models: automatic translation from demographic specifications to efficient inference methods**. In *Application and Theory of Petri Nets*. Edited by Haddad S, Pomello L. Berlin, Heidelberg: Springer Berlin Heidelberg; 2012:32-50.

48. Li H, Durbin R: **Inference of human population history from individual whole-genome sequences**. *Nature* 2011, **475**:493-496.

49. Schiffels S, Durbin R: **Inferring human population size and separation history from multiple genome sequences**. *Nat Genet* 2014, **46**:919-925.

50. Sheehan S, Harris K, Song YS: **Estimating variable effective population sizes from multiple genomes: a sequentially Markov conditional sampling distribution approach**. *Genetics* 2013, **194**:647-662.

51. Li N, Stephens M: **Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data**. *Genetics* 2003, **165**:2213-2233.

52. Paul JS, Song YS: **A principled approach to deriving approximate conditional sampling distributions in population**

genetics models with recombination. *Genetics* 2010, **186**:321-338.

53. Paul JS, Steinrücken M, Song YS: **An accurate sequentially Markov conditional sampling distribution for the coalescent with recombination**. *Genetics* 2011, **187**:1115-1128.

54. Davison D, Pritchard JK, Coop G: **An approximate likelihood for genetic data under a model with recombination and population splitting**. *Theor Popul Biol* 2009, **75**:331-345.

55. Steinrücken M, Paul JS, Song YS: **A sequentially Markov conditional sampling distribution for structured populations with migration and recombination**. *Theor Popul Biol* 2013, **87**:51-61.

56. Steinrücken M, Kamm JA, Song YS: **Inference of complex population histories using whole-genome sequences from multiple populations**. *bioRxiv* 2015 http://dx.doi.org/10.1101/026591.

57. • Moreno-Mayar JV, Potter BA, Vinner L, Steinrücken M, Rasmussen S, Terhorst J, Kamm JA, Albrechtsen A, Malaspinas A-S, Sikora M *et al.*: **Terminal Pleistocene Alaskan genome reveals first founding population of Native Americans**. *Nature* 2018, **553**:203-207.
Studies the peopling of the Americas, making use of ancient genomes and combining frequency-based and coalescent-HMM methods for robust demographic inference.

58. Steinrücken M, Spence JP, Kamm JA, Wieczorek E, Song YS: **Model-based detection and analysis of introgressed Neanderthal ancestry in modern humans**. *Mol Ecol* 2018.

59. •• Terhorst J, Kamm JA, Song YS: **Robust and scalable inference of population history from hundreds of unphased whole genomes**. *Nat Genet* 2017, **49**:303-309.
Presents a coalescent-HMM that essentially combines PSM, with frequency-based methods for a powerful, yet scalable tool for demographic inference.

60. Paul JS, Song YS: **Blockwise HMM computation for large-scale population genomic inference**. *Bioinformatics* 2012, **28**:2008-2015.

61. • Palamara PF, Terhorst J, Song YS, Price AL: **High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability**. *Nat Genet* 2018 http://dx.doi.org/10.1101/276931. (in press).
Extends ideas from SMC++ to data from genotype array data sets and presents the largest-scale application of coalescent-HMMs to date.

62. Harris K, Sheehan S, Kamm JA, Song YS: **Decoding coalescent hidden Markov models in linear time**. In *Research in Computational Molecular Biology*. Edited by Sharan R. Cham: Springer International Publishing; 2014:100-114.

63. •• Kelleher J, Etheridge AM, McVean G: **Efficient coalescent simulation and genealogical analysis for large sample sizes**. *PLOS Comput Biol* 2016, **12**:e1004842.
Presents`msprime`: simulation software capable of simulating data under the full coalescent with recombination orders of magnitude faster than other simulators.

64. Hawks J: **Introgression makes waves in inferred histories of effective population size**. *Hum Biol* 2017, **89**:67-80.

65. Schrider DR, Shanku AG, Kern AD: **Effects of linked selective sweeps on demographic inference and model selection**. *Genetics* 2016, **204**:1207-1223.

66. Beichman AC, Phung TN, Lohmueller KE: **Comparison of single genome and allele frequency data reveals discordant demographic histories**. *G3 Genes Genomes Genet* 2017, **7**:3605-3620.

67. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D: **Principal components analysis corrects for stratification in genome-wide association studies**. *Nat Genet* 2006, **38**:904-909.

68. Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR *et al.*: **Genes mirror geography within Europe**. *Nature* 2008, **456**:98-101.

69. Pritchard JK, Stephens M, Donnelly P: **Inference of population structure using multilocus genotype data**. *Genetics* 2000, **155**:945-959.

70. Alexander DH, Novembre J, Lange K: **Fast model-based estimation of ancestry in unrelated individuals**. *Genome Res* 2009.

71. Raj A, Stephens M, Pritchard JK: **fastSTRUCTURE: variational inference of population structure in large SNP data sets**. *Genetics* 2014, **197**:573-589.

72. Cabreros I, Storey JD: **A nonparametric estimator of population structure unifying admixture models and principal components analysis**. *bioRxiv* 2017 http://dx.doi.org/10.1101/240812.

73. Dabney J, Meyer M, Pääbo S: **Ancient DNA damage**. *Cold Spring Harb Perspect Biol* 2013, **5**:a012567.

74. Miroshnikov A, Steinrücken M: **Computing the joint distribution of the total tree length across loci in populations with variable size**. *Theor Popul Biol* 2017, **118**:1-19.

75. Myers S, Bottolo L, Freeman C, McVean G, Donnelly P: **A fine-scale map of recombination rates and hotspots across the human genome**. *Science* 2005, **310**:321-324.

76. Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KTh *et al.*: **Fine-scale recombination rate differences between sexes, populations and individuals**. *Nature* 2010, **467**:1099-1103.

77. Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A *et al.*: **Rate of *de novo* mutations and the importance of father's age to disease risk**. *Nature* 2012, **488**:471-475.

78. Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA *et al.*: **Parental influence on human germline *de novo* mutations in 1,548 trios from Iceland**. *Nature* 2017, **549**:519-522.

79. Smith TCA, Arndt PF, Eyre-Walker A: **Large scale variation in the rate of germ-line *de novo* mutation, base composition, divergence and diversity in humans**. *PLOS Genet* 2018, **14**:1-30.

80. Kim J, Mossel E, Rácz MZ, Ross N: **Can one hear the shape of a population history?** *Theor Popul Biol* 2015, **100**:26-38.

81. Johndrow JE, Palacios JA: *Exact limits of inference in coalescent models*. 2017. ArXiv e-prints.