Peer Community Journal

Section: Evolutionary Biology

Research article

Published
2024-03-18

Cite as
Kevin Korfmann, Thibaut Paul Patrick Sellinger, Fabian Freund, Matteo Fumagalli and Aurélien Tellier (2024) *Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent*, Peer Community Journal, 4: e33.

Correspondence
kevin.korfmann@tum.de

Peer-review
Peer reviewed and recommended by PCI Evolutionary Biology,
https://doi.org/10.24072/pci.evolbiol.100699

Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent

Kevin Korfmann[#,1], Thibaut Paul Patrick Sellinger[#,2,1], Fabian Freund[3,4], Matteo Fumagalli[5,6], and Aurélien Tellier[1]

Volume **4** (2024), article e33

https://doi.org/10.24072/pcjournal.397

Abstract

The reproductive mechanism of a species is a key driver of genome evolution. The standard Wright-Fisher model for the reproduction of individuals in a population assumes that each individual produces a number of offspring negligible compared to the total population size. Yet many species of plants, invertebrates, prokaryotes or fish exhibit neutrally skewed offspring distribution or strong selection events yielding few individuals to produce a number of offspring of up to the same magnitude as the population size. As a result, the genealogy of a sample is characterized by multiple individuals (more than two) coalescing simultaneously to the same common ancestor. The current methods developed to detect such multiple merger events do not account for complex demographic scenarios or recombination, and require large sample sizes. We tackle these limitations by developing two novel and different approaches to infer multiple merger events from sequence data or the ancestral recombination graph (ARG): a sequentially Markovian coalescent (SMβC) and a graph neural network (GNN*coal*). We first give proof of the accuracy of our methods to estimate the multiple merger parameter and past demographic history using simulated data under the β-coalescent model. Secondly, we show that our approaches can also recover the effect of positive selective sweeps along the genome. Finally, we are able to distinguish skewed offspring distribution from selection while simultaneously inferring the past variation of population size. Our findings stress the aptitude of neural networks to leverage information from the ARG for inference but also the urgent need for more accurate ARG inference approaches.

[1] Department of Life Science Systems, Technical University of Munich, Munich, Germany, [2] Department of Environment and Biodiversity, Paris Lodron University of Salzburg, Salzburg, Austria, [3] Department of Genetics and Genome Biology, University of Leicester, Leicester, UK, [4] Institute of Plant Breeding, Seed Science and Population Genetics, University of Hohenheim, Stuttgart, Germany, [5] School of Biological and Behavioural Sciences, Queen Mary University of London, London, UK, [6] The Alan Turing Institute, London, UK, [#] Equal contribution

Peer Community Journal is a member of the Centre Mersenne for Open Scientific Publishing
http://www.centre-mersenne.org/

CENTRE MERSENNE

e-ISSN 2804-3871

Introduction

With the availability of genomes of increasing quality for many species across the tree of life, population genetics models and statistical methods have been developed to recover the past history of a population/species from whole genome sequence data from several individuals (Barroso and Dutheil, 2023; Barroso et al., 2019; Johri et al., 2022, 2020; Li and Durbin, 2011; Schiffels and Durbin, 2014; Sellinger et al., 2020; Sheehan and Song, 2016; Speidel et al., 2019; Stephan, 2019). Indeed, the inference of the past demographic history of a species, *i.e.* population expansion, contraction, or bottlenecks, extinction/colonisation, is not only interesting in its own right, but also essential to calibrate genome-wide scans to detect genes under (*e.g.* positive or balancing) selection (Johri et al., 2021; Stephan, 2019). A common feature of inference methods that make full use of whole genome sequences is the underlying assumption of a Kingman coalescent process (Kingman, 1982) to describe the genealogy distribution of a sample. The Kingman coalescent process and its properties stem from using the traditional forward-in-time Wright-Fisher (WF) model to describe the reproduction mechanism of a population. Besides non-overlapping generations, a key assumption of the neutral WF model is that an individual offspring chooses randomly (*i.e.* uniformly) its parents from the previous generation. More precisely, each chromosome chooses a parental chromosome from the previous generation. Thus, a key parameter is the distribution of the number of offspring that parents can have. In the WF model, due to the binomial sampling, the distribution of offspring number per parent is well approximated by a Poisson distribution with both mean and variance equal to one. This implies that parents will most likely have zero, one, or two offspring individuals, but it is improbable that one parent would have many offspring individuals (*i.e.* on the order of the population size, under the Wright-Fisher haploid model the probability for a parent to have 10 or more offspring is $\approx 10^{-8}$). The assumption of small variance in offspring distribution between individual parents is realistic for species with low juvenile mortality (so-called type I and II survivorship in ecology, see survivorship curves *e.g.* by Demetrius, 1978), such as mammals.

As genome sequence data become available for a wide variety of species with different biological traits and/or life cycles, the applicability of the Kingman coalescent relying on the WF model can be questioned (Árnason and Halldorsdottir, 2015; Árnason et al., 2023; Freund et al., 2023; Kato et al., 2017; Menardo et al., 2020; Morales-Arce et al., 2020; Niwa et al., 2016; Steinruecken et al., 2013; Tellier and Lemaire, 2014). Indeed, for some species, such as fish, with high fecundity and high juveniles mortality (type III survivorship, Demetrius, 1978), it is expected that the variance in reproduction between parents can be much larger than under the Poisson distribution (Tellier and Lemaire, 2014). This effect is termed as sweepstake reproduction (Árnason and Halldorsdottir, 2015; Hedgecock and Pudovkin, 2011). Neutral processes such as strong seed banking (Blath et al., 2020), high fecundity with skewed offspring distribution (Eldon and Wakeley, 2006; Hedgecock and Pudovkin, 2011), extremely strong and recurrent bottlenecks (Birkner et al., 2008; Casanova et al., 2020), and strong selective processes (*i.e.* positive selection; Árnason et al., 2023; Brunet et al., 2006, 2007; Durrett and Schweinsberg, 2005; Harris and Jensen, 2020) are theoretically shown to deviate from the classic WF model in a way that the genealogies can no longer be described by a Kingman coalescent process. Under such conditions, a new class of processes arise to describe the genealogy distribution, a class where multiple individuals can coalesce and/or multiple distinguished coalescence events can occur simultaneously (Bolthausen and Sznitman, 1998; Donnelly and Kurtz, 1999; Mohle and Sagitov, 2001; Pitman, 1999; Sagitov, 1999, 2003). Generally, this class of genealogical processes is called the Multiple Merger Coalescent (MMC). MMC models are more biologically appropriate than the Kingman coalescent to study many species of fish (Árnason and Halldorsdottir, 2015; Árnason et al., 2023; Eldon et al., 2015; Hedgecock and Pudovkin, 2011), invertebrates (insects, crustaceans, etc.), viruses (Matuszewski et al., 2017), bacteria (Menardo et al., 2020; Neher and Hallatschek, 2013), plants and their pathogens (Tellier and Lemaire, 2014). While we would like to assess which population model best describes the species genealogy, field experiments to quantify the underlying reproduction mechanism of a species can be costly and time consuming at best, or intractable at worst. Therefore, an alternative solution is to use inference methods based on genome data to identify which model best describes the genealogy of a given species/population.

In this study we use the so-called β-coalescent, a specific class of MMC models. Unlike under the WF model, under MMC models the ploidy level strongly affects the distribution of genealogies (Birkner et al., 2013). For simplicity, in this study we focus on haploid organisms. In the polyploid case, where each parent contributes multiple genomes, the SMC formulations of putative intra- and inter-individual coalescence events would need to be carefully modelled, since this effect would lead to smaller coalescence probabilities and a change of the predicted statistical power for demographic inference. It is demonstrated that if the probability of a parent to have k or more offspring is proportional to $k^{-\alpha}$, where $1 < \alpha < 2$, then the genealogy can be described by a Λ-coalescent (Schweinsberg, 2003). The latter is a general class of coalescent process describing how and how fast ancestral lineages merge (Pitman, 1999; Sagitov, 1999). When using the Beta$(2-\alpha, \alpha)$ distribution as a probability measure for the Λ-coalescent, the transition rates (*i.e.* coalescent rate) can be analytically obtained leading to the β-coalescent, a specific MMC model. If α tends to 2, then the coalescent process converges to a Kingman coalescent up to a scaling constant as specified in a more detailed way in the documentation of msprime (`https://tskit.dev/msprime/docs/stable/api.html#msprime.BetaCoalescent`). The effective population size calculations for the Beta coalescent yield $Ne = (\frac{\mu_{\text{estimated}}}{\mu_{\text{real}}})/\text{scaling constant})^{\frac{1}{(\alpha-1)}}$, where $m = 1 + \frac{1}{2^{\alpha-1}\cdot(\alpha-1)}$, scaling constant $= \frac{(m^{\alpha})}{(\alpha\cdot\beta(2-\alpha,\alpha))}$ (β being the Beta function) and $\mu_{\text{estimated}} = \frac{\theta}{\left(2\cdot\sum_{i=1}^{n_{\text{ind}}-1}\frac{1}{i}\right)\cdot L}$ (Baumdicker et al., 2022; Birkner et al., 2013; Koskela, 2018; Koskela and Berenguer, 2019; Schweinsberg, 2003) . If α tends to one, the model tends to a Bolthausen-Sznitman coalescent process (*i.e.* dominated by strong multiple merger events) (Bolthausen and Sznitman, 1998). The β-coalescent has the property that the observed polarized Site Frequency Spectrum (SFS) of a sample of single nucleotide polymorphisms (SNPs) exhibits a characteristic U-shape with an excess of rare and high frequency variants (compared to the Kingman coalescent) (Sargsyan and Wakeley, 2008). Current methods to draw inference under MMC models leverage information from the summary statistics extracted from full genome data such as Site Frequency Spectrum (SFS, or derived summary statistics; Harris and Jensen, 2020; Koskela and Berenguer, 2019; Sackman et al., 2019), minor allele frequency (Rice et al., 2018) or copy number alteration (Kato et al., 2017). It is shown that the SFS is robust to the effect of recombination (Koskela and Berenguer, 2019; Rice et al., 2018) and its shape allows to discriminate between simple demographic models (population expansion or contraction) under the Kingman coalescent and MMC models with constant population size (Eldon et al., 2015; Koskela, 2018; Koskela and Berenguer, 2019). However, methods relying on genome-wide SFS have two main disadvantages. First, in absence of strong prior knowledge, they can suffer from non-identifiability (Johri et al., 2022) as several complex neutral demographic and/or selective models under the Kingman or MMC models can generate similar SFS distributions. Second, as they summarize the collection of underlying genealogies, they require high sample sizes (>50) to produce trustworthy results (Eldon et al., 2015; Koskela, 2018; Koskela and Berenguer, 2019), relying on experimental designs which are prohibitive for the study of non-model species. To tackle these limitations, we develop two methods that integrate recombination events along the genome in order to leverage more information from full genome data, thus requiring fewer samples.

In species undergoing sexual reproduction, recombination events break the genealogy of a sample at different position of the genome (*i.e.* the genealogy of a sample varies along the genome), leading to what is called the Ancestral Recombination Graph (ARG; Birkner et al., 2013; Hudson, 1983; Lewanski et al., 2024; Wong et al., 2023). Because all the genealogical information is contained in the ARG, in this study we aim at the interpretation of the ARGs to recover model parameters in presence of multiple merger events. With the development of the sequentially Markovian coalescent theory (Marjoram and Wall, 2006; McVean and Cardin, 2005; Wiuf and Hein, 1999), it becomes tractable to integrate linkage disequilibrium over chromosomes in inferences based on the Kingman coalescent (Li and Durbin, 2011). Hence, we first develop an SMC approach based on the β-coalescent named the Sequentially Markovian β Coalescent (SMβC). The β-coalescent has the additional property that, under recombination, long range dependency can be generated between coalescent trees along the genome if multiple-merger events happen in a single generation (Birkner et al., 2013). In other words, coalescent trees which are located at different places in the genome, and expected to be unlinked from one another (Nelson et al., 2020), would show non-zero correlation in their topology and coalescent times. This is because coalescent trees from different genomic regions may all be affected by the same MMC event (merger event of multiple lineages in the past) which then leaves traces in the genome at several loci (Birkner et al., 2008). To overcome the theoretically predicted non-Markovian property of the distribution of genealogies along the genome under the β-coalescent with recombination (Birkner et al., 2013) and the increasing sparsity of genealogies and ancestral nodes with respect to α (see Supplementary Figure S18, S19 and S20), we develop a second method based on deep learning (DL) trained from efficient coalescent simulations (Baumdicker et al., 2022). In evolutionary genomics, DL approaches trained by simulations are shown to be powerful inference tools (Korfmann et al., 2023; Sheehan and Song, 2016). Previous work demonstrated that DL approach can help overcome problems mathematically insolvable or computationally intractable in the field of population genetics (Battey et al., 2020; Burger et al., 2022; Chen et al., 2018; Flagel et al., 2018; Isildak et al., 2021; Qin et al., 2021; Sheehan and Song, 2016; Wang et al., 2021; Yelmen et al., 2021). The novelty of our neural network relies on its structure (Graph Neural Network, GNN) and its training algorithm based on the ARG of a sample, or its tree sequence representation (Brandt et al., 2024; Kelleher et al., 2018; Whitehouse et al., 2024). GNNs are an emerging category of DL algorithm (Bronstein et al., 2017; Cao et al., 2020; Xu et al., 2019; Zhou et al., 2020) that benefit by using irregular domain data (*i.e.* graphs). GNNs are designed for the prediction of node features (Kipf and Welling, 2016; Yang et al., 2016), edge features (link prediction) (Schlichtkrull et al., 2017; Zhang and Chen, 2018), or additional properties of entire graphs (Lee et al., 2018; Ying et al., 2019). Therefore, GNNs represent a new tool to address the large dimensionality of ARGs, while simultaneously leveraging information from the genealogy (namely topology and age of coalescent events) as a substantial improvement over convolutions of genotype matrices, as currently done in the field (Sanchez et al., 2020).

We first quantify the bias of previous SMC methods (MSMC and MSMC2, Schiffels and Durbin, 2014; Wang et al., 2020) when performing inference of past population size variation under the β-coalescent. We then describe our two methods, SMβC and GNN*coal*, and demonstrate their statistical power as well as their respective limitations. From simulated tree-sequence (*i.e.* ARG) and sequence (*i.e.* SNPs) data, we assess the accuracy of both approaches to recover the past variation of population size and the α parameter of the Beta-distribution. This parameter indicates how frequent and strong multiple merger events occur (see Supplementary Figure S20). We demonstrate that our approaches can infer the evolutionary mechanism responsible for multiple merger events and distinguish local selection events from genome-wide effects of multiple mergers. We highlight the limits of the Markovian property of SMC to describe data generated under the β-coalescent. Finally, we show that both our approaches can model and identify the presence of selection along the genome while simultaneously accounting for non-constant population size, recombination, and skewed offspring distribution. Thus our methods represents a major and necessary leap forward in the field of population genetic inferences.

Materials and Methods

In our study we first assume the true ARG to be known. Hence, the ARG of the sample is given as input to our methods to estimate recover model parameters of interest (*e.g.* the α parameter and/or the past variation of population size). We then show the applicability of our methods by using as input simulated sequence data (*i.e.* SNPs) and/or ARG inferred using ARGweaver (Rasmussen et al., 2014) from simulated sequence data.

SMC-based method

In this study, we use different SMC-based algorithms: two previously published, MSMC and MSMC2 (Schiffels and Durbin, 2014; Wang et al., 2020), and the new SMβC. In the latter, the software backbone stems from our previous eSMC (Sellinger et al., 2020, 2021) whilst the theoretical framework originates from the MSMC algorithm (Schiffels and Durbin, 2014) (see Supplementary Text S1). All approaches can either use the ARG or sequence data as input. Providing the ARG as input for MSMC and MSMC2 is enabled by a re-implementation included in the R package eSMC2 previously published in (Sellinger et al., 2021). It is important to mention that there are no theoretical differences in the models whether sequence data or ARG is inputted (see Sellinger et al., 2021) and Supplementary Text S1 for details). The difference is that in one case the hidden states are inferred from sequence data with a forward-backward algorithm, and in the later the sequence of hidden states are directly built from reading the inputted ARG (skipping the forward-backward algorithm). The MSMC2 algorithm focuses on the coalescence time between two haploid samples along the genome. In the event of recombination, there is a break in the current genealogy and the coalescence time consequently takes a new value. A detailed description of the algorithm can be found in (Malaspinas, 2016; Wang et al., 2020). The MSMC algorithm simultaneously analyses multiple sequences (up to 10) and follows the distribution of the first coalescence event in a sample of size $n > 2$ along the sequence based on the Kingman coalescent (Kingman, 1982). A detailed description of MSMC can be found in Schiffels and Durbin, 2014.

Our new approach, SMβC, is a theoretical extension of the MSMC algorithm, simultaneously analyzing multiple haploid sequences and focusing on the first coalescence event of a sample size 3 or 4 (this parameter is named *M* throughout the manuscript). We define as *M* the number of lineages simultaneously modeled by either approach. Hence, the SMβC follows the distribution of the first coalescence event of a sample size *M* along sequences assuming a β-coalescent process. Therefore, our SMβC allows for more than two ancestral lineages to join the first coalescence event, or new lineages to join an already existing binary (or triple) coalescent event. Hence, the SMβC extends the MSMC theoretical framework by adding hidden states at which more than two lineages coalesce. Currently, the SMβC has been derived to analyze for up to 4 sequences simultaneously (due to computational load and mathematical complexity). However the SMβC can handle more than M sequences by analyzing all combination of sample size *M* before optimizing the likelihood. The emission matrix is similar to the one of MSMC. As in the MSMC software, the population size is assumed piece-wise constant in time and we discretize time in 40 bins throughout this study. A detailed description of SMβC can be found in Supplementary Text S1. To test and validate the theoretical accuracy of our approach, we first study its best case convergence (introduced in Sellinger et al., 2021) which corresponds to the model's performance when the true (exact) genealogy is given as input, *i.e.* as if the hidden states are known. Additionally, we also validate the practical accuracy of the SMβC on simulated sequence data taking the same input as the MSMC software (Schiffels and Durbin, 2014), or using the inferred ARGs by ARGweaver (Rasmussen et al., 2014). All SMC approaches used in this manuscript are found in the R package eSMC2 (`https://doi.org/10.5281/zenodo.10782372`, Sellinger, 2024).

GNN*coal* method

Inspired by results obtained from inferences based on tree sequence data (Gattepaille et al., 2016; Sellinger et al., 2021), we develop a graph neural network (GNN) taking tree sequence data as input. Our GNN is designed to infer population size along with the α parameter of the Beta distribution describing the distribution of offspring production. In practice, the ARG is reshaped into a sequence of genealogies (more precisely a sequence of undirected graphs), and then given as input to the GNN (similar to what is described above for the SMβC). In our analyses, we fixed the batch size to 500. This value represents the number of coalescence trees being processed before updating parameters of the neural network. As the batch size is fixed to 500, only simulations displaying at least 500 recombination events are considered for the training data sets. If more than 500 recombination events occur along the sequence, the ARG is truncated and the GNN will only take as input the first 500 genealogies and remove the rest. Thanks to the GNN architecture, the algorithm can account for the topology of the genealogy. Hence, the GNN leverages information from coalescence time and branch lengths but also from the topology of the ARG. This operation is known as a graph convolution. By doing so, the GNN is capable of learning from local features of the ARG and extract information from its complex structure. To learn from global genealogy patterns (which SMC-based methods cannot do), an additional pooling strategy is implemented as part of the network (Ying et al., 2019). To do so, the ARG is broken into smaller ARGs (*i.e.* subgraphs) during the forward-pass step. To illustrate the GNN strategy, we visualize the compression-like process, from the coalescent trees (1) being processed by GNN*coal* (2,3) to the inferred variable of interest (4, 5) in Figure 1.

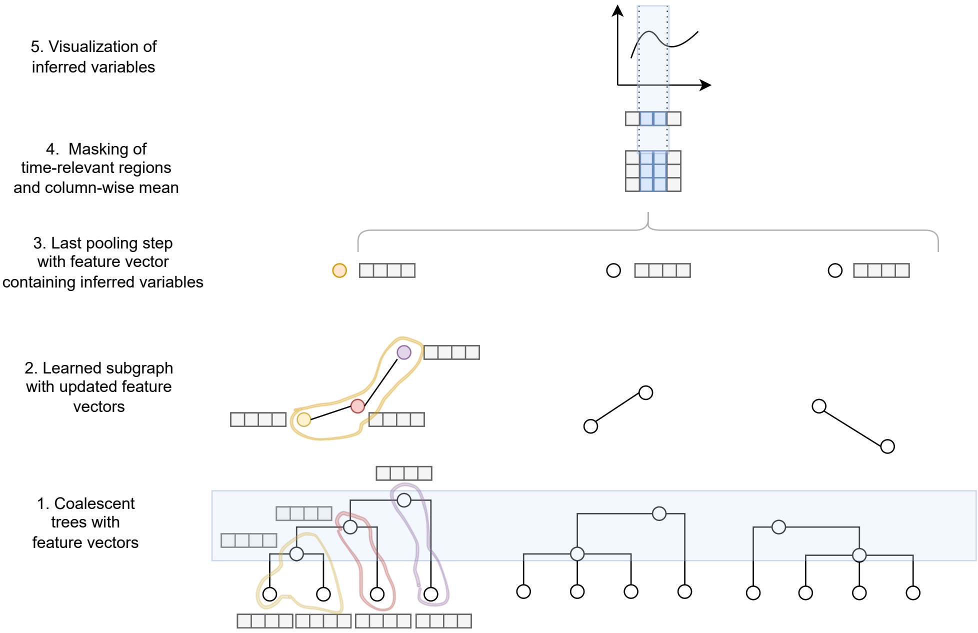
5. Visualization of inferred variables
4. Masking of time-relevant regions and column-wise mean
3. Last pooling step with feature vector containing inferred variables
2. Learned subgraph with updated feature vectors
1. Coalescent trees with feature vectors

Figure 1 – Schematic representation of GNN*coal* processing an ARG The figure represents the analogues compression of node embeddings (or feature vectors) as in Fig. 1 of (Ying et al., 2019). The pooling is hierarchical and applied to each coalescent trees until a single embedding per tree remains, which is fed into a dense neural net to obtain the inferred variable of interest (*i.e.* demographic changes). Each coalescent ancestor or leaf node is initialized by this feature vector (light grey boxes) (1). Sub-graphs are generated by a pooling network with updated feature vectors and a final compression step is performed until ideally one node per graph remains (2-3). Lastly, the column-wise mean is taken after applying a time mask (blue - based on number of coalescent events), so that single feature vector remains (4-5). Detailed description of the graph convolution, feature vector initialization, pooling methodology, coalescent time mask construction, and dataset generation can be found in Supplementary Text S2 or (Ying et al., 2019).

To infer parameters from our neural network, we need to define an objective function to be optimized. We use a masked root-mean-squared error (RMSE) loss function as objective function which is computed for each inputted ARG (*i.e.* minimizing the average square difference between predicted and true parameter value). In practice, time is discretized (as for the SMβC) and time windows are defined. The true α value and true demography at 60 predefined time points are given as input to the GNN to compute the loss function. The GNN captures the stochastic complexity arising from the underlying demographic scenario and model parameters. Furthermore, our algorithm naturally defines an appropriate time window to have sufficient observation at each time point. A more detailed description of the GNN*coal* can be found in Supplementary Text S2. The code of the model architecture is implemented in *Pytorch* (Paszke et al., 2017) using the extension *Pytorch Geometric* (Fey and Lenssen, 2019). The model is available with the simulated training dataset at `https://doi.org/10.5281/zenodo.10781643` (Korfmann, 2024b) and `https://doi.org/10.5281/zenodo.10781640` (Korfmann, 2024a).

ARGweaver and tsinfer

As the ARG is not known in practice, it needs to be inferred from sequence data. ARGweaver displays the best performance at recovering the ARG from whole genome polymorphism data at the sample sizes employed in this study (*i.e. « 50*; Brandt et al., 2022; Rasmussen et al., 2014). Briefly, ARGweaver samples the ARG of n chromosomes/scaffolds conditional on the ARG of $n-1$ chromosomes/scaffolds. To this aim, ARGweaver relies on hidden Markov models while assuming a sequentially Markov coalescent process and a discretization of time, similarly to the SMC-based methods previously described. For a more detail description of the algorithm, we refer the reader to the supplementary material of Rasmussen et al., 2014.

For distinguishing between MMC and selection we additionally applied tsinfer to estimate undated genealogical topologies in an effort to build a small training dataset for a model selection study reframed as classification task. Tsinfer has been chosen due to its computational performance and details about the algorithm can be found in the respective supplementary information of (Kelleher et al., 2019).

Simulation of data

Validation dataset for both methods. The ARG is given as input to the DL approach and the SMβC (see Sellinger et al., 2021). We use msprime (Baumdicker et al., 2022) to simulate the ARG of a sample (individuals are assumed to be haploid) under the β-coalescent based on (Birkner et al., 2013; Schweinsberg, 2003) or under the Kingman coalescent (under neutrality or selection using msprime *SweepGenicSelection* functionality with start and end frequency of $1/N_e$ and 0.99, respectively). We simulate 10 sequences of 100 Mbp under five different demographic scenarios: 1) Constant population size; 2) Bottleneck with sudden decrease of the population size by a factor 10 followed by a sudden increase of population by a factor 10; 3) Expansion with sudden increase of the population size by a factor 10, 4) Contraction with sudden decrease of the population size by a factor 10; and 5) "Saw-tooth" with successive exponential decreases and increases of population size through time, resulting in continuous population size variation (as shown in Schiffels and Durbin, 2014; Sellinger et al., 2021; Terhorst et al., 2017). We simulate data under different α values (*i.e.* parameters of the β-distribution) including values of 1.9 (almost no multiple merger events), 1.7, 1.5, and 1.3 (frequent and strong multiple merger events; Supplementary Figure S20). Mutation and recombination rate (respectively μ and r) are set to 10^{-8} per generation per bp in order to obtain the best compromise between realistic values and number of SNPs. When specified, some specific scenarios assume recombination and mutation rate set to produce sufficient data or to avoid violation of the finite site hypothesis. All python scripts used to simulate data sets are available at `https://doi.org/10.5281/zenodo.10781640` (Korfmann, 2024a). Note that the output of msprime suffers from a discontinuity in behaviour when increasing α above 1.9 and transitioning from the Beta coalescent to the Kingman coalescent ($\alpha = 2$). The coalescent process converges to a Kingman coalescent up to a scaling constant which we recover in our simulations and estimations (see description in `https://tskit.dev/msprime/docs/stable/api.html#msprime.BetaCoalescent`).

Additionally, to generate sequence data, we simulate 10 sequences of 10 Mbp under the five different demographic scenarios described above and for the same α values. For each scenario, 10 replicates are simulated. In order to obtain sufficient SNPs for inference, we simulate sequence data with mutation and recombination rate (respectively μ and r) of 10^{-8} per generation per bp when α is set to 1.9 and 1.7, 10^{-7} per generation per bp when α is set to 1.5, and 10^{-6} per generation per bp when α is set to 1.3.

Training dataset for the GNNcoal. In our study we train two GNNs, one to infer past variation of population size through time along with α, and one for model selection. The training dataset used for both GNNs is described below.

Training dataset for the GNN inferring α and demography

We generate an extensive number of ARGs to train our GNN. The ARGs are simulated under many demographic scenarios and α values. The model parameters are updated in supervised manner. The loss function is calculated for each batch with respect to how much the machine-learning estimates differ from to the true parameters used for simulation. The simulations strategy to recover past demographic history is based on the strategy described and used in (Boitard et al., 2016; Sanchez et al., 2020). The idea of this approach is to generate a representative set of demographic scenarios over which the network generalizes to consequently infer similar demographic changes after training. More details on the training strategy can be found in Supplementary Text S2.

To improve the simulated demographic history before inference, we introduce a smoothing of the demography allowing to infer continuous variation of population size through time. We do so by interpolating I time points cubically, and choosing w (set to 60) uniformly spaced new time points of the interpolation in log space. All time points more recent than ten generations in the past are discarded, since inference is too imprecise in the very recent present under our models. An example of this process can be seen in Supplementary Text S2.

Training dataset to disentangling coalescent and selection signatures

Beyond parameter inference, deep learning approaches can also be used for clustering. Hence, we train a GNN to disentangle between different scenarios and models. In total, we define eight classes, namely K (S0) (Kingman, no selection), K (WS) (Kingman, weak selection), K (MS) (Kingman, medium selection), K (SS) (Kingman, strong selection) and four different β-coalescent classes ($1.75 \leq \alpha < 2$, $1.5 \leq \alpha < 1.75$, $1.25 \leq \alpha < 1.5$, $1.01 \leq \alpha < 1.25$) without selection. The three different selection regimes are defined as: $0.01 \leq Ne \times s < 0.1$ for SS, $0.001 \leq Ne \times s < 0.01$ for MS, $0.0001 \leq Ne \times s < 0.001$ for WS and $Ne \times s = 0$ for absence of selection. Demography is kept constant and set to 10^4 and 10^6 individuals for Kingman and β-coalescent respectively and sequence length is set to 10^5 bp. The simulation is discarded if it resulted in less than 2,000 obtained trees and is rerun with twice the sequence length until the tree number required is satisfied. This procedure avoids simulating large genome segments of which only a small fraction of trees is used for the given scenario during training and inference. The selection site is introduced in the centre of the respective sequence, so that 249 trees left and 250 right of the middle tree under selection form a training sample, using 500 trees for each sample. One hundred replicates are generated for each training sample. The complete training dataset consists of 4,000 parameter sets: 2,000 for the Kingman cases and 2,000 for the β-coalescent cases (90% training dataset and 10% testing dataset). The model itself is trained for 20 epochs (number of time the data is analyzed), and the evaluation performed afterward on 1,000 randomly generated parameter sets, with one replicate per parameter set. Branches of the datasets have been normalized by population size to avoid biases in the dating. Additionally, all tree sequences have been re-inferred with tsinfer to create a separated dataset, which has been used for training and evaluation (see results below). The same architecture used for demography estimation is employed with additional linear layers to reduce the number of output dimensions from 60 to 8. The loss function is set to a Cross-Entropy-Loss for the network to be trainable for categorical labels. Otherwise all architecture and training parameters is the same as described above and detailed in Supplementary Text S2.

Results

Inference bias under the wrongly assumed Kingman coalescent

We first study the effect of assuming a Kingman coalescent when the underlying true model is a β-coalescent (*i.e.* in presence of multiple merger events) by applying MSMC and MSMC2 to our simulated data. The inference results from MSMC and MSMC2 when the population undergoes a sawtooth demographic scenario are displayed in Figure 2. For $\alpha > 1.5$ the shape of the past demography is fairly well recovered. Decreasing the parameter α of the β-coalescent (*i.e.* higher probability of multiple merger events occurring) increases the variance of inferences and flattens the demography. Yet, both methods fail to infer the correct population size, due to the scaling discrepancy between the Kingman and β-coalescent. While MSMC and MSMC2 assume an underlying Wright-Fisher model as reproduction model, whose genealogy is well approximated by a Kingman coalescent with one unit of coalescent time corresponding to *N* generations, the β-coalescent simulation are based on a different reproduction model (Schweinsberg, 2003), whose genealogy is given by a β-coalescent with a different timescale (see Introduction). Even for α close to 2, where the β-coalescent resembles the Kingman coalescent, one unit of coalescent time in the β-coalescent and one unit in a Wright-Fisher model associated Kingman coalescent still differ by a scaling factor (see Introduction and Methods for details). Hence, we perform the same analysis and correct for the scaling effect after the inference of the MMC versus a Kingman coalescent to better capture the specific effects of assuming binary mergers only. The results are displayed in Figure S1. For $\alpha > 1.5$ the demography is accurately recovered providing we know the true value of α to adjust the y-axis (population size) scale. However, for smaller α values the observed variance is extremely high and a flattened past variation of population size is observed.

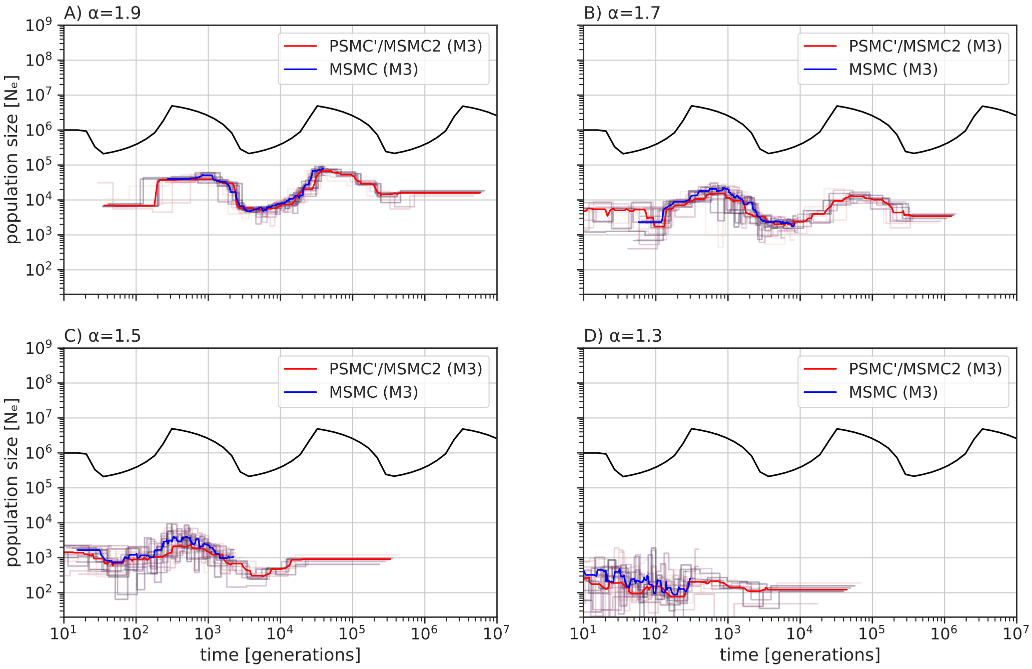

Figure 2 – Performance of MSMC and MSMC2 under a β-coalescent. Averaged estimated demographic history by MSMC (blue) and MSMC2 (red) based on 10 sequences (mean of random permutations of *M*=3) of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp over ten repetitions (while analyzing simultaneously 3 sequences, noted by M=3). Each repetition result is represented in light red (PSMC'/MSMC2) or in light blue (MSMC). Population undergoes a sawtooth demographic scenario (black) for A) $\alpha = 1.9$, B) $\alpha = 1.7$, C) $\alpha = 1.5$, and D) $\alpha = 1.3$.

The limit of the Markovian hypothesis

As SMC approaches rely on the hypothesis of Markovian change in genealogy along the genome, we study the effect of α on the linkage disequilibrium (LD) of pairs of SNPs (r^2, Miles et al., 2021; Rogers and Huff, 2009) in data simulated under the Kingman Coalescent or the β-coalescent (with $\alpha = 1.5$ and $\alpha = 1.3$) and constant population size (Figure 3). LD monotonously decreases in average with distance under the Kingman coalescent suggesting the hypothesis of Markovian change in genealogy to be a fair approximation of the genealogical process in that case (Wilton et al., 2015). Under the β-coalescent a similar shape of the distribution is observed but with a higher average amount of LD. We find a higher variance in LD for smaller α values. The increased variance results in the occurrence of high spikes of LD along the genome (*e.g.* Figure 3 B). The stochastic increase of linkage along the genome demonstrates that the Markovian hypothesis used to model genealogies along the genome is strongly violated under the β-coalescent due to the long range effect of strong multiple merger events (Birkner et al., 2013).

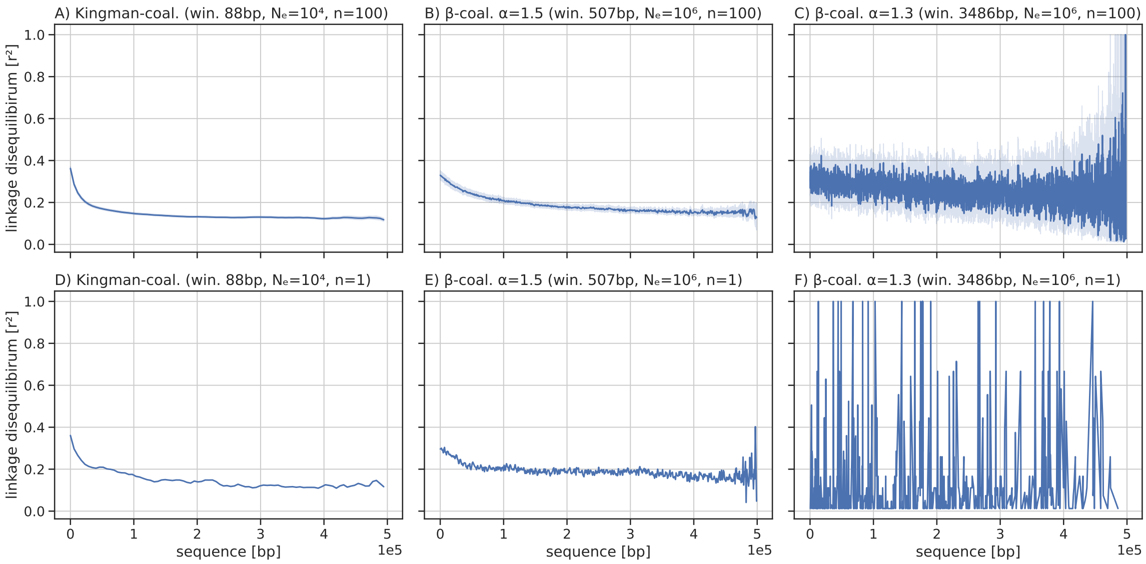

Figure 3 – Linkage disequilibrium under a Kingman and β-coalescent. Pairwise linkage disequilibrium between SNPs (r^2) under a Kingman and β-coalescent with $\alpha = 1.5$ and $\alpha = 1.3$ using 100 sequences of length 0.5 Mb for A) - C) and 1 replicate in D) - F). The population size is constant at $N = 10^4$ for the Kingman model and $N = 10^6$ for the β-coalescent, with $\mu = 1 \times 10^{-7}$ and $r = 1 \times 10^{-8}$ per generation per bp. For each LD analysis, the linkage disequilibrium is calculated by averaging it over automatically-selected window sizes, such that on average at least two mutations are in each window for A) to F), respectively.

We further investigate the effect of multiple merger events on LD. To this aim, we first assume an SMC framework (*e.g.* MSMC2 or eSMC) to predict the transition matrix (*i.e.* matrix containing the probabilities for the coalescent time to change to another value between two positions of the genome) and investigate the absolute difference between the observed transition events. Under the Kingman coalescent, the distribution of coalescent times between two positions in a sample of size two ($n = 2$) is well spread across hidden states in Figure S2 (*i.e.* absence of structured difference between observed and predicted transition events). However, under the β-coalescent (with $\alpha = 1.3$) we observe significant differences between observed and predicted transition events at times points where multiple merger events occur (Figure S3). More precisely we observed transitions at specific time points (corresponding to multiple merger events) occurring much more frequently than what is predicted by the model (dark blue lines). This plot thus shows that multiple merger events do not affect the genealogy at every time point and that multiple merger events are over represented in the distribution of transitions events due to the long range effects of MMC events (*i.e.* many positions of the genome contain the same information). This means that one multiple merger coalescent events can affect all positions in the genomes (explaining the spikes in the LD distribution). In contrast, under the Kingman coalescent with recombination, the probability for a coalescent event to affect the whole genome is negligible.

This plot thus unveils the discrepancy between the expectation from the SMC (*i.e.* approximating the distribution of genealogies along the genome by a Markov chain) and the actual effect of multiple merger events on the genealogy distribution along the genome. This discrepancy does not stem from the simulator, because it correctly generates ARG under the β-coalescent model (Baumdicker et al., 2022; Birkner et al., 2013), but from the limits of the SMC approximation to model events with long range effects on the ARG (Figure S3).

Inferring α and past demography on ARG

To test if our two approaches (GNN*coal* and SMβC) can recover the past variation of population size and the α parameter, we run both methods on simulated tree sequences under different α values and demographic scenarios. Figure 4 displays results for data simulated under a sawtooth past demography and for α ranging from 1.9, 1.7, 1.5 to 1.3. In all cases, the GNN*coal* approach exhibits low variance to infer the variation of population size and high accuracy from 1.9 to 1.5 with a noticeable drop in accuracy for 1.3 attributable to the ever increasing sparsity due to decreasing α generating stronger β-coalescent events. For high α values (>1.5), the shape of population size variation is well recovered by SMβC (4). However, for smaller values, the observed high variance demonstrates the limits of SMC inferences.

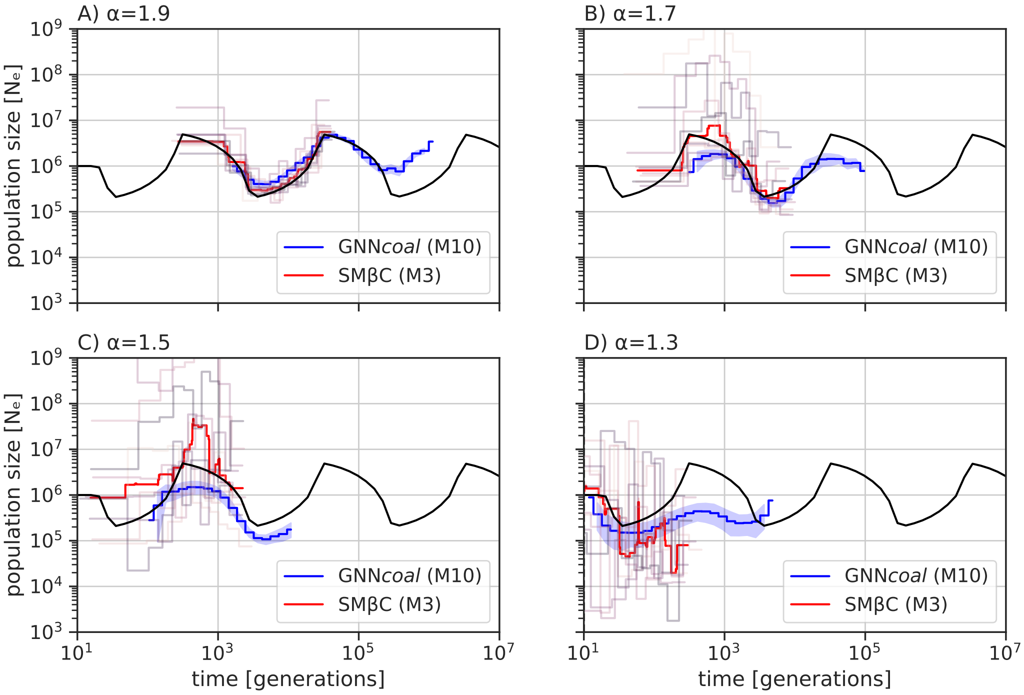

Figure 4 – Best-case convergence estimations of SMβC and GNN*coal* under a β-coalescent. Estimations of past demographic history by SMβC in red (median) and by GNN*coal* in blue (mean and 95% confidence interval, CI95; while analyzing simultaneously *M*=3 or *M*=10 sequences; individual replicates of SMβC shown as light lines) when population undergoes a sawtooth demographic scenario (black) under A) $\alpha = 1.9$, B) $\alpha = 1.7$, C) $\alpha = 1.5$ and D) $\alpha = 1.3$. SMβC runs on 10 sequences and 100 Mb, GNN*coal* runs on 10 sequences and 500 trees, and $\mu = r = 10^{-8}$ per generation per bp.

On average, both approaches seem to recover fairly well the true α value (Figure 5 and Table S1). In particular, GNN*coal* displays high accuracy and lower standard deviation. We note that the variance in the estimation of α increases with diminishing α value. Moreover, increasing the number of simultaneously analyzed sequences by SMβC does not seem to improve the inferred α value (Table S1). These conclusions are also valid for the results in Figure S4-S7 and Table S1 based on inference under four additional demographic scenarios: constant population size, bottleneck, sudden increase and sudden decrease of population size.

When α diminishes, the effective population size decreases and the number of recombination events plummets for small values of $\alpha < 1.5$. To demonstrate the theoretical convergence of SMβC to the correct values, we run SMβC on data simulated with mutation and recombination rate fifty times higher under similar scenarios as in Figure 4. This operation increases the amount of data in the form of SNPs and number of independent coalescent trees by recombination. Since branch lengths (in generations) are on average smaller in the presence of multiple merger when compared to a Kingman coalescent, we choose to increase the rates as opposed to increasing the genome lengths, which does not affect the branch lengths (but increases the number of genealogies). Results of SMβC for α values of 1.7, 1.5 and 1.3 are displayed on Table S2. Overall our results show that SMβC can recover α with higher accuracy when more data is available. To be more precise when $M = 3$ (M being the number of simultaneously haploid sequence analyzed), the overall average inferred α values improve from 1.6, 1.53 and 1.42 (Table S1) to 1.64 , 1.49 and 1.36 (for data simulated respectively under $\alpha = 1.7$,$\alpha = 1.5$ and $\alpha = 1.3$). Yet when $M = 4$ a gain in accuracy is only observed for $\alpha = 1.5$ and $\alpha = 1.3$. Indeed, the overall average inferred α values changed from 1.60, 1.54 and 1.47 (Table S1) to 1.58, 1.47 and 1.39 (for data simulated respectively under $\alpha = 1.7$, $\alpha = 1.5$ and $\alpha = 1.3$).

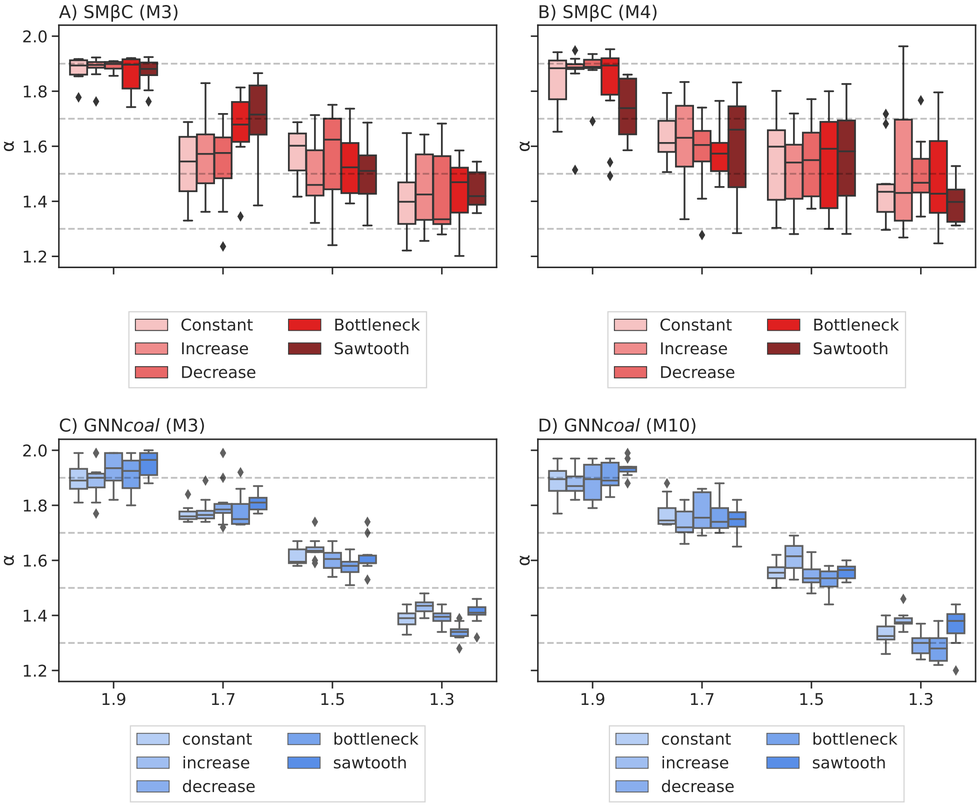

Figure 5 – Estimated α values by SMβC and GNN*coal*. Estimated values of α by SMβC and GNN*coal* over ten repetitions using 10 sequences of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp under a β-coalescent process (with different α parameter). The analysis are run on five different demographic scenarios (Constant population size, Bottleneck, Sudden increase, Sudden decrease and a Sawtooth demography) using a sample size $n = 3$ for A) and C), $n = 4$ for B), and $n = 10$ for D). Grey dashed lines indicate the true α values. For exact values and standard deviations of the respective experiment see Supplementary Table S1.

Although 10 sequences are given to SMβC in the previous analyses, the method can only analyze three or four simultaneously. On the other hand, GNN*coal* can simultaneously analyze 10 sequences, that is the whole simulated ARG. As we observe that GNN*coal* has a higher performance than SMβC, we wish to test whether the GNN*coal* better leverages information from the ARG or benefits from simultaneously analyzing a larger sample size. Thus, we run GNN*coal* on the same dataset, but downsampling the coalescent trees to a sample size three. Results for sample size ten are displayed in Figure S4 to S7 and downsampled results with sample size three (M=3) of GNN*coal*, which appear to be similar, are displayed in Figure S8, demonstrating that the GNNs can better leverage information from the ARG in presence of multiple merger events.

Additionally, we test if both approaches can recover a Kingman coalescent from the ARG when data are simulated under the Kingman coalescent, namely both approach should recover $\alpha = 2$. To do so, we simulate the same five demographic scenarios as above under a Kingman coalescent and infer the α parameter along with the past variation of population size. Estimations of α values are provided in Table 1 and are systematically higher than 1.85, suggesting mostly binary mergers. The associated inferred demographies are shown in Figures S9-S13. Both approaches correctly infer the past demographic shape up to the scaling discrepancy between the Beta and the Kingman coalescent (as previously described). Furthermore, we notice that the scaling effect only affects the y-axis for the SMβC but affect both axes for GNN*coal*.

Table 1 – Average estimated values of α by SMβC and GNN*coal* over ten repetitions under the Kingman coalescent using 10 haploid sequences of 10 Mb and $\mu = r = 10^{-8}$ per generation per bp. The standard deviation is indicated in brackets.

scenario	True α	α:SMβC,M=3	α:SMβC,M=4	α : GNN, M=3	α : GNN, M=10
Constant	2	1.97 (0.005)	1.97 (0.008)	1.99 (0.002)	1.99 (0.003)
Sawtooth	2	1.94 (0.017)	1.87 (0.019)	1.99 (0.002)	1.99 (0.003)
Bottleneck	2	1.97 (0.01)	1.97 (0.009)	1.99 (0.003)	1.99 (0.004)
Decrease	2	1.97 (0.007)	1.97 (0.008)	1.99 (0.003)	1.99 (0.004)
Increase	2	1.97 (0.007)	1.97 (0.008)	1.99 (0.004)	1.99 (0.002)

As GNN*coal* was not trained on data simulated under the Kingman coalescent (especially with such high population size), some events fall beyond the scope of the GNN due to the scaling discrepancy between the Beta and Kingman coalescence. Hence, we run GNN*coal* on data simulated under the Kingman coalescent but with smaller population size (scaled down by a factor 100) to assure that all events fall within the scope of the GNN. Values of α inferred by the GNN*coal* and the SMβC under the five demographic scenarios are available in Table S3. The associated inference of population size are plotted in Figure S9-S12. Both approaches recover high α values (*i.e.*>1.85) suggesting a genealogy with almost exclusively binary mergers. In addition, both approaches accurately recover the shape of the past variation of population size up to a scaling constant but only on the population size y-axis.

Inferring α and past demography from simulated sequence data

We first investigate results for both GNN*coal* and SMβC with the objective of evaluating the performance on ARG reconstructed from sequence data using ARGweaver (Rasmussen et al., 2014) as ARGweaver is currently being considered the best performing approach to infer ARG for sample size smaller than 20 (Brandt et al., 2022). Demographic inference results by both approaches are displayed in Figure S14, and α inference results in Table S4. GNN*coal* does not recover the shape of the demographic history from the inferred ARGs and largely overestimates α. In contrast, SMβC produces better inferences of α when giving the inferred ARG as input when compared to the GNN. SMβC recovers the shape of the past variation of population size for $\alpha > 1.3$ but displays extremely high variance for $\alpha = 1.3$. We then evaluate SMβC on simulated sequence data to compare the necessity of reconstructing the ARG for the SMC method and found that α is typically well recovered (Table 2) and that results are similar to what obtained when the true ARG is given. Furthermore, the shape of the past variation of population size is well inferred under the sawtooth demographic scenario for $\alpha > 1.3$ (Figure S15). In the other four scenarios, the shape of the demography is recovered in recent times but population sizes are underestimated in the past (Figure S16). Finally, as found above from inputted ARGs, the variance in estimates of population sizes generally increases with diminishing α.

Table 2 – Average estimated α values by SMβC on simulated sequence data over ten repetitions using 10 sequences of 10 Mb with recombination and mutation rate set to 1×10^{-8} for α 1.9 and 1.7, 1×10^{-7} for α 1.5 and 1×10^{-6} for α 1.3 per generation per bp under a Beta coalescent process. The analysis are run on five different demographic scenarios (Constant population size, Bottleneck, Sudden increase, Sudden decrease and a Sawtooth demography).

scenario	True α	α^*:SMβC,M=3
Constant	1.9	1.86 (0.16)
Bottleneck	1.9	1.89 (0.09)
Increase	1.9	1.93 (0.07)
Decrease	1.9	1.96 (0.04)
Sawtooth	1.9	1.76 (0.17)
Constant	1.7	1.82 (0.10)
Bottleneck	1.7	1.64 (0.23)
Increase	1.7	1.82 (0.10)
Decrease	1.7	1.89 (0.13)
Sawtooth	1.7	1.71 (0.27)
Constant	1.5	1.52 (0.30)
Bottleneck	1.5	1.64 (0.33)
Increase	1.5	1.57 (0.24)
Decrease	1.5	1.60 (0.18)
Sawtooth	1.5	1.66 (0.14)
Constant	1.3	1.31 (0.20)
Bottleneck	1.3	1.2 (0.17)
Increase	1.3	1.24 (0.13)
Decrease	1.3	1.57 (0.11)
Sawtooth	1.3	1.37 (0.16)

Inferring MMC and accounting for selection

As specific reproductive mechanisms and selection can lead to the occurrence of multiple merger-like events, we train our neural network on data simulated under the β-coalescent, and under the Kingman coalescent in presence or absence of selection to assess our methods capacity to distinguish between them. We then use the trained GNN*coal* to determine if multiple merger events originate from skewed offspring distribution or positive selection, or if the data follows a neutral Kingman coalescent process. The classification results are displayed in Figure 6 in the form of confusion matrices, where the percentage of times the GNN*coal* correctly assigns the true model shown on the diagonal evaluated on a test dataset of 1,000 ARGs. We tested three scenarios A) training and evaluating on known exact ARGs, B) training on exact ARGs but evaluating on inferred ARGs, and, lastly C) training and evaluating on inferred ARGs. The results indicate the necessity of integrating inference errors or instances of branch unresolvability into the training process. The network is able of distinguishing between signals of multiple merger, which translate to an estimate of α, from simple ARG-estimation uncertainties. The overall confusion between neighboring classes may be attributed to the comparably small size of training data (4,000 simulations), which enabled to build a training dataset comprised of inferred trees within few hours. To summarize our approach can accurately distinguish between Kingman and β-coalescent, but uncertainty needs to be part of the training procedure.

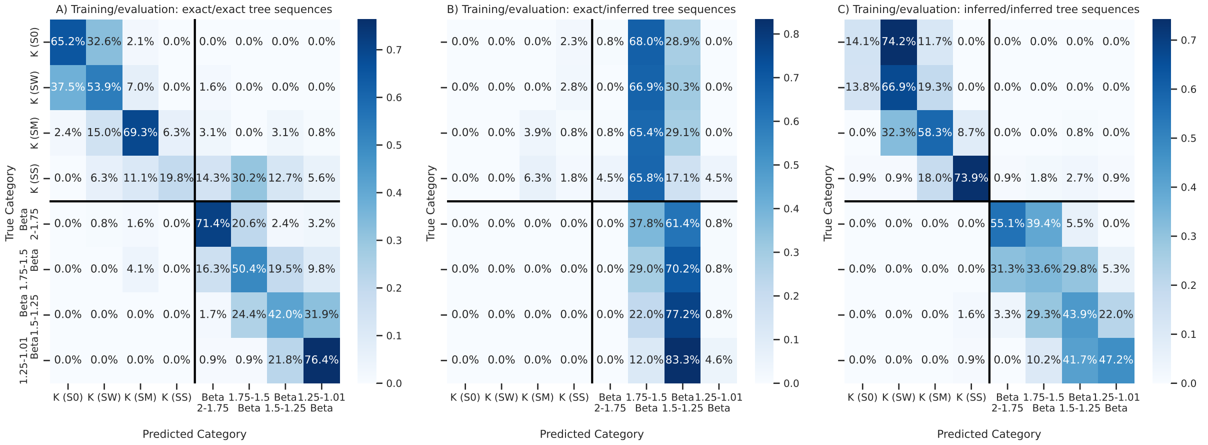

Figure 6 – Confusion matrix for Kingman and β-coalescent classification model under varying selection coefficients. Evaluation of classification accuracy for Kingman (K) and β-coalescent (B) for no selection (S0), weak selection (SW), medium selection (SM) and strong selection (SS) using a 1,000 repetition validation dataset (and small 4000 proof-of-concept repetition training set). Population size was kept constant at $N = 10^4$ individuals for the Kingman scenario and at $N = 10^6$ for the β-coalescent, using a sample size $n = 10$ and $r = 10^{-8}$ per bp per generation. Branch length are normalized by the respective population size. Classification model has been trained and evaluated either on exact or inferred tree sequences (tsinfer without dating) as indicated in the subfigure titles of A), B) and C).

Since strong selection can lead to multiple merger coalescent or rapid and successive coalescent events (as the beneficial alleles spreads very quickly in the population; Bisschop et al., 2021; Durrett and Schweinsberg, 2005; Sackman et al., 2019), we investigate if our approaches can model and recover the effect of selection. Therefore, we infer α along the genome (to model the local effect of selection on the genome) with both approaches from true genealogies simulated with strong positive selection or neutrality under a Kingman coalescent with population size being constant through time. SMβC infers α on windows of 10kbp along the genome, and GNN*coal* infers α every 20 trees along the genome. Results for GNN*coal* and SMβC are displayed in Figure 7. The SMβC approach recovers smaller α value around the locus under strong selection (while GNN*coal* displays higher variance). However under neutrality or weak selection, inferred α values remain high (>1.6) along the genome.

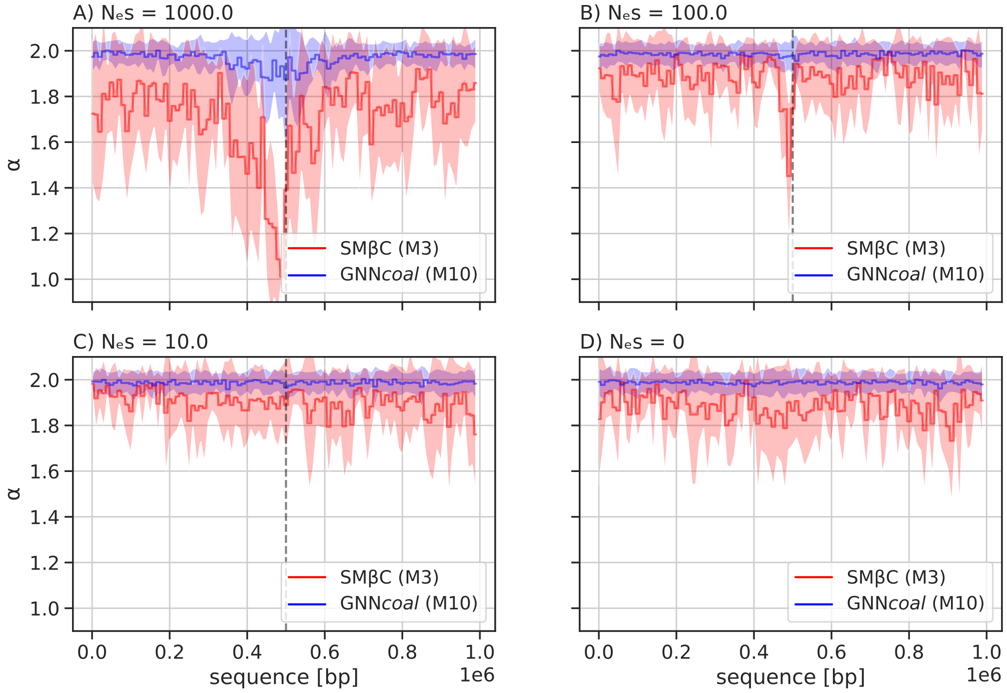

Figure 7 – Averaged estimations by GNN*coal* and SMβC under selection Estimations of α along the genome by the GNN*coal* approach and the SMβC when population undergoes as strong positive selective sweep event (at position 0.5 Mb) under different strengths of selection: A) $s = 0.01$, B)$s = 0.001$, C) $s = 0.0001$, and D) $s = 0$ meaning neutrality (mean and standard deviation for both methods). The population size is constant and set to $N = 10^5$ with $\mu = r = 10^{-8}$ per generation per bp. We hence have in A) $N_e \times s = 1000$,B) $N_e \times s = 100$, C) $N_e \times s = 10$ and D) $N_e \times s = 0$. SMβC uses 20 sequences of 1Mb (red) and GNN*coal* uses 10 sequences through down-sampling the sample nodes (blue)

Similarly, we run both approaches on genealogies simulated under the β-coalescent (assuming neutrality) and we infer the α value along the genome. Inferred α values by both approaches are plotted in Figure S17. GNN*coal* is able to recover the α value along the genome with moderate overestimation due to tree sparsity. On the contrary, SMβC systematically underestimates α values. Nevertheless, unlike in presence of positive selection at a given locus, the inferred α values are found in all cases to be fairly constant along the genome.

We finally simulate data under a Kingman coalescent (true genealogies) with a strong selective sweep or under neutrality conditioned on a sawtooth demographic scenario to test our methods' simultaneous inference capabilities. Under neutrality, our both approaches recover, as expected, high α values along the genome and can accurately recover the past variation of population size (only up to a scaling constant for GNN*coal*, since it was trained on the β-coalescent only) (Figure 8). Similarly, when the simulated data contains strong selection, a small α value is recovered at the locus under selection and the past variation of population size is accurately recovered, albeit with a small underestimation of population size in recent times for SMβC (Figure 8).

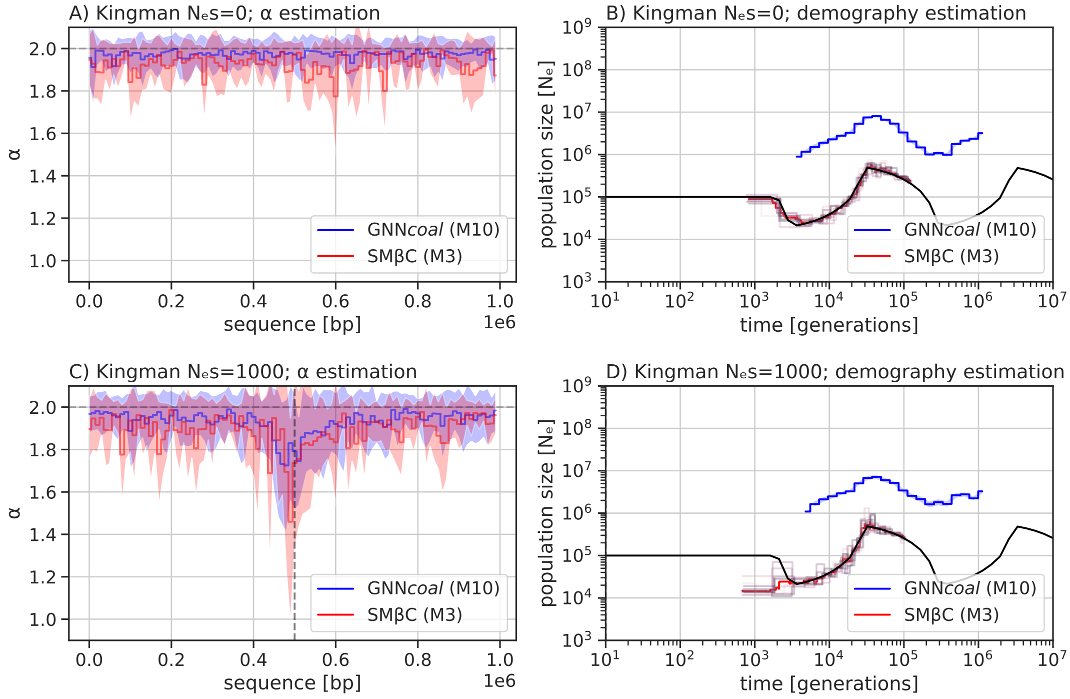

Figure 8 – Simultaneous estimations of α along the sequence under demographic change by GNN*coal* and SMβC. Simultaneous estimation of α along the genome under a partial sawtooth scenario: A) and B) in the absence of selection (mean and standard deviation for both methods), and C) and D) presence of selection with $N_e S = 1,000$ (mean and CI95 for GNN*coal* and median for SMβC). SMβC uses 20 sequences of 1Mb (red) and GNN*coal* uses 10 sequences through down-sampling the sample nodes (blue), and $\mu = r = 10^8$ per generation per bp.

Discussion

With the rise in popularity of SMC approaches for demographic inferences (Li and Durbin, 2011), most current methods leverage information from whole genome sequences by simultaneously reconstructing a portion of the ARG to infer past demographic history (Li and Durbin, 2011; Schiffels and Durbin, 2014; Terhorst et al., 2017; Upadhya and Steinrücken, 2021), migration rates (Kim et al., 2020; Wang et al., 2020), variation in recombination and mutation along the genome (Barroso and Dutheil, 2023; Barroso et al., 2019), as well as ecological life history traits such as selfing or seed banking (Sellinger et al., 2020; Struett et al., 2022). However, other previous studies proposed to uncouple both steps, namely by first reconstructing the ARG and by then inferring parameters from its distribution (Gattepaille et al., 2016; Rasmussen et al., 2014; Sellinger et al., 2021). Indeed, recent efforts have been made to improve approaches to recover the ARG (Brandt et al., 2022; Hubisz and Siepel, 2020; Kelleher et al., 2019; Mahmoudi et al., 2022; Rasmussen et al., 2014; Speidel et al., 2019), as well as its interpretation (Gattepaille et al., 2013; Sellinger et al., 2021). Our results on data simulated under the β-coalescent clearly show the strong effect of multiple merger events on the topology and branch length of the ARG. We find that the more multiple merger events occur, the more information concerning the past demography is lost. Both GNN*coal* and SMβC, whether given sequence data, the true or inferred ARG, can recover the α parameter and the variation of past population size for α values high enough (*i.e.* $\alpha \geq 1.5$). However, for lower values of α, a larger amount of data is necessary for any inference, specifically in the form of a high effective population size (correspondingly adequate mutation and recombination rates) and sufficient sequence length, which becomes nearly impossible when α tends to one. Both approaches can also recover the Kingman coalescent (*i.e.* α>1.8). We find that GNN*coal* outperforms SMβC in almost all cases when given the true ARG, and we demonstrate that GNN*coal* can be used to disentangle between β-coalescent and Kingman models with selection.

Overall, our results provide a substantial improvement in the development of inference methods for models with multiple merger events, a key step to understand the underlying reproduction mechanism of a species. While still inferring population sizes of the correct order of magnitude, SMβC is outperformed by GNN*coal* when given true ARGs as input. As ARG inference method improve, GNN models will offer a promising alternative to current SMC methods. As we directly compare our theoretical SMC to the GNN based on the same input data (coalescent trees), we are ideally placed to dissect the mechanisms underlying the power of the GNN*coal* method. We identify four main reasons for the difference in accuracy between the two methods developed. First, the SMβC approach suffers from the limit of the sequential Markovian coalescent hypothesis along the genome when dealing with strong multiple merger events (Birkner et al., 2013; Casanova et al., 2020). Second, most current SMC approaches, except XSMC (Ki and Terhorst, 2020), rely on a discretization of the coalescent times into hidden states, meaning that simultaneous mergers of three lineages may not be easily distinguished from two consecutive binary mergers occurring over a short period. Third, the SMβC relies on a complex hidden Markov model and due to computational and mathematical tractability, it cannot leverage information on a whole genealogy. In fact, as MSMC, SMβC only focuses on the first coalescent event, and therefore cannot simultaneously analyze large sample size. Furthermore, the SMβC approach leverages information from the distribution of genealogies along the genome. Whilst, in the near absence of recombination events, both approaches cannot utilize any information from the genealogy itself, GNN*coal* can overcome this limit by increasing the sample size. Fourth, the SMβC is based on a coalescent model where α is constant in time. Yet multiple merger events do not appear regularly across the genealogical timescale, but occur at few random time points. Hence, the SMC approach suffers from a strong identifiability problem between the variation of population size and the α parameter (for low α values). For instance, if during one hidden state one strong multiple merger event occurs, multiple merger events are seldom observed and SMβC may rather assume a small population size at this time point (hidden state). This may explain the high variance of inferred population sizes under the β-coalescent.

By contrast, GNN*coal* makes use of the whole ARG, and can easily scale to larger sample sizes (over 10), although it recovers α with high accuracy with sample size *M*= 3 only. Our interpretation is that GNN*coal* is able of simultaneously leveraging information from topology and the age of coalescent events (nodes) across several genealogies (here 500). GNN*coal* ultimately leverages information from observing recurrent occurrences of the same multiple merger events at different locations on the genome, while being aware of true multiple merger events from rapid successive binary mergers. We believe that our results pave the way towards the interpretability of GNN and deep learning methods applied to population genetics. For further theoretical insights into recent descriptions of multiple merger we would like to point the reader towards (Diamantidis et al., 2023).

When applying both approaches to simulated sequence data (and not to true ARGs), both approaches behave differently. GNN*coal* is not capable to accurately infer model parameters, *i.e.* past variation of population size or α. In contrast, SMβC performed better than GNN*coal* when dealing with sequence data (and not true ARG). SMβC is capable of recovering α and the shape of the demographic scenario in recent times irrespective of whether sequence data or ARG inferred by ARGweaver is given as input. This is most likely because the statistic used by SMβC (*i.e.* first coalescent event in discrete time) is coarser than the statistic used by GNN*coal* (*i.e.* the exact ARG). We therefore speculate that the theoretical framework of the SMβC, although being in theory less accurate than GNN*coal*, is more robust and suited for application to sequence data. More specifically, the issue being faced by the GNN*coal* is known as out-of-distribution inference (Hüllermeier and Waegeman, 2021), which requires the network to generalize over an untrained data distribution. This issue happens because GNN*coal* is not trained using ARG inferred by ARGweaver. Building a training data set for GNN*coal* to overcome this issue is currently impractical due to the inference speed of ARGweaver. However, future work will aim at increasing robustness of GNN inferences, for instance by adding uncertainty or multiple models during the training process. Improving the performance of GNN*coal* on sequence data requires more efficient and accurate ARG inference methods, such as to incorporate inferred (non-exact) genealogies into the training, thereby accounting for inference errors and for the evaluation of the algorithm on a broader spectrum of common population genetic research questions. The former observation is important to avoid bias from potential hypothesis violations of the chosen ARG inference approach.

Past demographic history, reproductive mechanisms, and natural selection are among the major forces driving genome evolution (Johri et al., 2022). Hence, in the second part of this manuscript we focus on integrating selection in both approaches. Currently, no method (especially if relying only on SFS information) can account for the presence of selection, linkage disequilibrium, non-constant population size and multiple merger events (Johri et al., 2022) although recent theoretical framework might render this possible in the future (Alberti et al., 2021).
As a first step to fill this gap, we demonstrate that GNN*coal* can be used for model selection to reduce the number of hypotheses to test. Determining which evolutionary forces are driving the genome evolution is key, as only under the appropriate neutral population model results of past demography and selection scans can be correctly interpreted (Johri et al., 2022, 2021). The high accuracy of GNN*coal* in model selection is promising, especially as other methods based on the SFS alone (Kato et al., 2017; Koskela and Berenguer, 2019) have limits in presence of complex demographic scenarios. GNN can possibly overcome these limits, as it is easier to scale the GNN to estimate more parameters. We follow a thread of previous work (Bisschop et al., 2021; Hejase et al., 2022; Sackman et al., 2019), by integrating and recovering selection, multiple merger and population size variation by simply allowing each fixed region in the genome to have its own α parameter. In presence of strong selection, we find lower α value around the selected loci and high α value in neutral neighbouring regions. Hence, our results point out that strong selection can indeed be modeled as a local multiple merger event (see Bisschop et al., 2021;

Durrett and Schweinsberg, 2005; Sackman et al., 2019). In presence of weak selection, no effect on the estimated α value is observed, demonstrating that weak selection can be modeled by a binary merger and has only a local effect on the branch length by shortening it. In theory, both approaches should be able to infer the global α parameter linked to the reproductive mechanism, as well as the local α parameter resulting from selection jointly with the variation of population size. However, the absence of a simulator capable of simulating data with selection and non-constant population size under a β-coalescent model prevents us from delivering such proofs. We show strong evidence that under neutrality our approaches can recover a constant (and correct) α along the genome as well as the past variation of the population size. We further predict that, while selective processes may preferentially occur in coding regions or regulatory potentially non-coding regions, local variations in α (as a consequence of sweepstake events) should be indifferent to the genomic functionality (coding or non-coding). Hence, we suggest that current sequence simulators (Baumdicker et al., 2022; Haller et al., 2019) could be extended to include the aforementioned factors and *de facto* facilitate the development of machine learning approaches.

Our study is unique in developing a state-of-the-art SMC approach and demonstrating that computational and mathematical problems can be overcome by deep learning (here GNN) approaches. The GNN*coal* approach is, in principle, not limited to the β-coalescent, and should work for other multiple merger models (*e.g.*, Dirac coalescents (Eldon and Wakeley, 2006)) with the appropriate training. Furthermore, our SMβC approach is the first step to build a full genome method with an underlying model accounting for positive selection. In the future, further implementations may be added for a more realistic approach. The α parameter should be varying along the genome (as a hidden state), as the recombination rate in the iSMC (Barroso et al., 2019). This would allow to account for the local effect of strong and weak selection (Alberti et al., 2021). The effect of the α parameter could be also changing through time to better model the non uniform occurrence of multiple merger events through time. Although it is mathematically correct to have α as a constant in time, it is erroneous in practice (Figure S2). We speculate that those additional features will allow to accurately model and infer multiple merger events, variation of population size, and selection at each position on the genome. We believe that deep learning approaches could also be improved to recover more complex scenarios, providing in depth development on the structure of the graph neural networks, for example, by accounting for more features. At last, further investigation are required to make progress in the interpretability of the GNN methods, namely which statistics and convolution of statistics are used by GNN*coal* to infer which parameters.

As our approaches are the first of their kind, we chose to restrain our study to haploid models of β and Kingman coalescent as a proof of principle. However, the GNN*coal* and SMβC approaches can be extended to higher ploidy levels. Diploid versions of the haploid reproduction models whose genealogies are given by the β-coalescent lead to slightly different MMC coalescent models which can exhibit simultaneous multiple mergers (Birkner et al., 2013, 2018). Thus, our GNN approach should be directly applicable when trained on these diploid models which are implemented in *msprime* (Baumdicker et al., 2022). However, to adjust the SMβC approach would be somewhat more cumbersome (but doable), since we would need to extend the underlying HMM to account for simultaneous multiple mergers. We emphasise that while there is growing evidence that MMC models produce better fitting genealogies for various species (Freund et al., 2023), there is ongoing discussions about which mathematical models are better suited to which species (for example see Árnason et al., 2023 for cod). We advocate that the life-cycle and various ecological factors determine whether a haploid or diploid MMC model can be chosen. On the one hand, a diploid MMC model is likely realistic if the species has a diploid life-cycle and balanced sex-ratio, so that multiple merger events do indeed happen in both sexes. On the other hand, if species are mostly haploid or clonal/asexual during their life-cycle (with periodically one short diploid phase for sexual reproduction) or exhibit strongly imbalanced sex-ratio, a haploid MMC model may be better suited. In their current form, our approaches are applicable to data from species with the latter characteristics such as many fungal and micro-parasites of plants and animals (including humans) as well as invertebrates (*e.g. Daphnia* or aphids) which undergo several clonal or parthenogenetic phases of reproduction (and one short sexual phase) per year. This represents a non-negligible set of study organisms which are of importance for medicine and agriculture (Tellier and Lemaire, 2014).

Our results on inferred ARGs stress the need for improving ARG inference (Brandt et al., 2022). Thanks to the SMC we are close to model the ARG allowing to infer demographic history, selection and specific reproductive mechanism. Moreover, the comparison of deep learning approaches with model driven *ad hoc* SMC methods may have the potential to help us solve ongoing challenges in the field. These include simultaneously inferring and accounting for recombination, variation of population size, different type of selection, population structure and the variation of the mutation and recombination rate along the genome. These issues have puzzled theoreticians and statisticians since the dawn of population genetics (Johri et al., 2022).

On a final note, as environmental changes hit us all, we suggest that decreasing the computer and power resources needed to perform DL/ GNN analyses should be attempted (Sapoval et al., 2022). Based on our study, we suggest that population genetics DL methods could be built as a two step process: 1) inferring ARGs, and 2) inferring demography and selection based on the ARGs. We speculate that general training sets based on ARGs could be build and be widely applicable for inference across many species with different life cycles and life history traits, while the inference of ARGs could be undertaken by complementary deep learning or Hidden Markov methods.

Acknowledgements

Preprint version 5 of this article has been peer-reviewed and recommended by Peer Community In Evolutionary Biology (`https://doi.org/10.24072/pci.evolbiol.100699`; Dutheil, 2024).

Fundings

This work was supported by the BMBF-funded de.NBI Cloud within the German Network for Bioinformatics Infrastructure (de.NBI) (031A532B, 031A533A, 031A533B, 031A534A, 031A535A, 031A537A, 031A537B, 031A537C, 031A537D, 031A538A). KK is supported by a grant from the Deutsche Forschungsgemeinschaft (DFG) through the TUM International Graduate School of Science and Engineering (IGSSE), GSC 81, within the project GENOMIE QADOP. TS is supported by the Austrian Science Fund (project no. TAI 151-B). AT acknowledges funding from the DFG grant TE809/1-4 (project 254587930) and TE809/7-1 (project 317616126). FF and AT acknowledge funding from the DFG Priority Program SPP1590 on "Probabilistic Structures in Evolution". MF and AT acknowledge the support from the Imperial College - TUM Partnership award.

Conflict of interest disclosure

The authors declare that they comply with the PCI rule of having no financial conflicts of interest in relation to the content of the article.

Data, script, code, and supplementary information availability

Code used to generate the simulated data for analysis, training and validation alongside (trained) deep learning models can be found at https://doi.org/10.5281/zenodo.10781643 (Korfmann, 2024b) and https://doi.org/10.5281/zenodo.10781640 (Korfmann, 2024a). Code for SMC approaches used in this manuscript are available in the R package eSMC2 (https://doi.org/10.5281/zenodo.10782372, Sellinger, 2024).
Msprime and its documentation can be found:
https://tskit.dev/msprime/docs/stable/quickstart.html
Supplementary information is available online, including Supplementary Text S1 and S2 as well as Supplementary Figures and Tables (https://doi.org/10.5281/zenodo.10792836, Korfmann et al., 2024).

References

Alberti F, Herrmann C, Baake E (2021). *Selection, recombination, and the ancestral initiation graph*. *Theoretical population biology* **142**, 46–56. https://doi.org/10.1016/j.tpb.2021.08.001.

Árnason E, Halldorsdottir K (2015). *Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models*. *peerj* **3**. https://doi.org/10.7717/peerj.786.

Árnason E, Koskela J, Halldórsdóttir K, Eldon B (2023). *Sweepstakes reproductive success via pervasive and recurrent selective sweeps*. *Elife* **12**, e80781. https://doi.org/10.7554/elife.80781.sa2.

Barroso GV, Dutheil JY (2023). *The landscape of nucleotide diversity in Drosophila melanogaster is shaped by mutation rate variation*. *Peer Community Journal* **3**, e40. https://doi.org/10.24072/pcjournal.267.

Barroso GV, Puzovic N, Dutheil JY (2019). *Inference of recombination maps from a single pair of genomes and its application to ancient samples*. *PLOS Genetics* **15**. https://doi.org/10.1371/journal.pgen.1008449.

Battey C, Ralph PL, Kern AD (2020). *Predicting geographic location from genetic variation with deep neural networks*. *eLife* **9**. Ed. by Patricia J Wittkopp, Magnus Nordborg, Christian Huber, Arbel Harpak, and Leo Speidel, e54507. https://doi.org/10.7554/eLife.54507.

Baumdicker F, Bisschop G, Goldstein D, Gower G, Ragsdale AP, Tsambos G, Zhu S, Eldon B, Ellerman EC, Galloway JG, Gladstein AL, Gorjanc G, Guo B, Jeffery B, Kretzschumar WW, Lohse K, Matschiner M, Nelson D, Pope NS, Quinto-Cortes CD, et al. (2022). *Efficient ancestry and mutation simulation with msprime 1.0*. *Genetics* **220**. https://doi.org/10.1093/genetics/iyab229.

Birkner M, Blath J, Eldon B (2013). *An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution*. *Genetics* **193**, 255–290. https://doi.org/10.1534/genetics.112.144329.

Birkner M, Blath J, Moehle M, Steinruecken M, Tams J (2008). *A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks*. https://doi.org/10.48550/arXiv.0808.0412.

Birkner M, Liu H, Sturm A (2018). *Coalescent results for diploid exchangeable population models*. *Electronic Journal of Probability* **23**. https://doi.org/10.1214/18-EJP175.

Bisschop G, Lohse K, Setter D (2021). *Sweeps in time: leveraging the joint distribution of branch lengths*. *Genetics* **219**. https://doi.org/10.1093/genetics/iyab119.

Blath J, Gonzalez Casanova A, Kurt N, Wilke-Berenguer M (2020). *The seed bank coalescent with simultaneous switching*. *Electronic Journal of Probability* **25**. https://doi.org/10.1214/19-EJP401.

Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F (2016). *Inferring Population Size History from Large Samples of Genome-Wide Molecular Data - An Approximate Bayesian Computation Approach*. *PLOS Genetics* **12**. Ed. by Mark A Beaumont, e1005877. https://doi.org/10.1371/journal.pgen.1005877.

Bolthausen E, Sznitman AS (1998). *On Ruelle's probability cascades and an abstract cavity method*. *Communications in mathematical physics* **197**, 247–276. https://doi.org/10.1007/s002200050450.

Brandt DYC, Huber CD, Chiang CWK, Ortega-Del Vecchyo D (2024). *The Promise of Inferring the Past using the Ancestral Recombination Graph (ARG)*. *Genome Biology and Evolution*. Ed. by Lucy Van Dorp, evae005. https://doi.org/10.1093/gbe/evae005.

Brandt DYC, Wei X, Deng Y, Vaughn AH, Nielsen R (2022). *Evaluation of methods for estimating coalescence times using ancestral recombination graphs*. *Genetics* **221**. https://doi.org/10.1093/genetics/iyac044.

Bronstein MM, Bruna J, LeCun Y, Szlam A, Vandergheynst P (2017). *Geometric Deep Learning: Going beyond Euclidean data*. *IEEE Signal Processing Magazine* **34**, 18–42. https://doi.org/10.1109/msp.2017.2693418.

Brunet E, Derrida B, Mueller AH, Munier S (2006). *Noisy traveling waves: Effect of selection on genealogies*. *Europhysics Letters* **76**, 1–7. https://doi.org/10.1209/epl/i2006-10224-4.

Brunet E, Derrida B, Mueller AH, Munier S (2007). *Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization*. *Physical Review E* **76**. https://doi.org/https://doi.org/10.1103/physreve.76.041104.

Burger KE, Pfaffelhuber P, Baumdicker F (2022). *Neural networks for self-adjusting mutation rate estimation when the recombination rate is unknown*. *PLOS Computational Biology* **18**, 1–17. https://doi.org/10.1371/journal.pcbi.1010407.

Cao W, Yan Z, He Z, He Z (2020). *A Comprehensive Survey on Geometric Deep Learning*. *IEEE Access* **8**, 35929–35949. https://doi.org/10.1109/ACCESS.2020.2975067.

Casanova AG, Pina VM, Siri-Jégousse A (2020). *The Symmetric Coalescent and Wright-Fisher models with bottlenecks*. *arXiv*. https://doi.org/10.48550/arXiv.1903.05642.

Chen J, Ni P, Li X, Han J, Jakovlic I, Zhang C, Zhao S (2018). *Population size may shape the accumulation of functional mutations following domestication*. *BMC Evolutionary Biology* **18**. https://doi.org/https://doi.org/10.1186/s12862-018-1120-6.

Demetrius L (1978). *Adaptive value, entropy and survivorship curves*. *Nature* **275**, 213–214. https://doi.org/10.1038/275213a0.

Diamantidis D, Fan WT, Birkner M, Wakeley J (2023). *Bursts of coalescence within population pedigrees whenever big families occur*. *bioRxiv*. https://doi.org/10.1101/2023.10.17.562743.

Donnelly P, Kurtz T (1999). *Particle representations for measure-valued population models*. *Annals of Probability* **27**, 166–205. https://doi.org/10.1214/aop/1022677258.

Durrett R, Schweinsberg J (2005). *A coalescent model for the effect of advantageous mutations on the genealogy of a population*. *Stochastic Processes and their Applications* **115**, 1628–1657. https://doi.org/10.1016/j.spa.2005.04.009.

Dutheil JY (2024). *Beyond the standard coalescent: demographic inference with complete genomes and graph neural networks under the beta coalescent*. *Peer Community in Evolutionary Biology*, 100699. https://doi.org/10.24072/pci.evolbiol.100699.

Eldon B, Wakeley J (2006). *Coalescent processes when the distribution of offspring number among individuals is highly skewed*. *Genetics* **172**, 2621–2633. https://doi.org/10.1534/genetics.105.052175.

Eldon B, Birkner M, Blath J, Freund F (2015). *Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents?* *Genetics* **199**, 841+. https://doi.org/10.1534/genetics.114.173807.

Fey M, Lenssen JE (2019). *Fast Graph Representation Learning with PyTorch Geometric*. *arXiv*. https://doi.org/10.48550/arXiv.1903.02428.

Flagel L, Brandvain Y, Schrider DR (2018). *The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference*. *Molecular Biology and Evolution* **36**, 220–238. https://doi.org/10.1093/molbev/msy224.

Freund F, Kerdoncuff E, Matuszewski S, Lapierre M, Hildebrandt M, Jensen JD, Ferretti L, Lambert A, Sackton TB, Achaz G (2023). *Interpreting the pervasive observation of U-shaped Site Frequency Spectra*. *PLOS Genetics* **19**. Ed. by Lindi Wahl, e1010677. https://doi.org/10.1371/journal.pgen.1010677.

Gattepaille LM, Jakobsson M, Blum MGB (2013). *Inferring population size changes with sequence and SNP data: lessons from human bottlenecks*. *Heredity* **110**, 409–419. https://doi.org/10.1038/hdy.2012.120.

Gattepaille L, Günther T, Jakobsson M (2016). *Inferring Past Effective Population Size from Distributions of Coalescent Times*. *Genetics* **204**, 1191–1206. https://doi.org/10.1534/genetics.115.185058.

Haller BC, Galloway J, Kelleher J, Messer PW, Ralph PL (2019). *Tree-sequence recording in SLiM opens new horizons for forward-time simulation of whole genomes*. *Molecular Ecology Resources* **19**, 552–566. https://doi.org/10.1111/1755-0998.12968.

Harris RB, Jensen JD (2020). *Considering Genomic Scans for Selection as Coalescent Model Choice*. *Genome Biology and Evolution* **12**, 871–877. https://doi.org/10.1093/gbe/evaa093.

Hedgecock D, Pudovkin AI (2011). *Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and Commentary*. *Bulletin of Marine Science* **87**, 971–1002. https://doi.org/10.5343/bms.2010.1051.

Hejase HA, Mo Z, Campagna L, Siepel A (2022). *A Deep-Learning Approach for Inference of Selective Sweeps from the Ancestral Recombination Graph*. *Molecular Biology and Evolution* **39**. https://doi.org/10.1093/molbev/msab332.

Hubisz M, Siepel A (2020). *Inference of Ancestral Recombination Graphs Using ARGweaver*. In: *Statistical Population Genomics*. Ed. by JY Dutheil. Vol. 2090. Methods in Molecular Biology. ISSN: 1064-3745, pp. 231–266. https://doi.org/10.1007/978-1-0716-0199-0_10.

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

Hüllermeier E, Waegeman W (2021). *Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods*. *Machine Learning* **110**, 457–506. https://doi.org/10.1007/s10994-021-05946-3.

Isildak U, Stella A, Fumagalli M (2021). *Distinguishing between recent balancing selection and incomplete sweep using deep neural networks*. *Molecular Ecology Resources* **21**, 2706–2718. https://doi.org/10.1111/1755-0998.13379.

Johri P, Aquadro CF, Beaumont M, Charlesworth B, Excoffier L, Eyre-Walker A, Keightley PD, Lynch M, McVean G, Payseur BA, Pfeifer SP, Stephan W, Jensen JD (2022). *Recommendations for improving statistical inference in population genomics*. *PLOS Biology* **20**, e3001669. https://doi.org/10.1371/journal.pbio.3001669.

Johri P, Charlesworth B, Jensen JD (2020). *Toward an Evolutionarily Appropriate Null Model: Jointly Inferring Demography and Purifying Selection*. *Genetics* **215**, 173–192. https://doi.org/10.1534/genetics.119.303002.

Johri P, Riall K, Becher H, Excoffier L, Charlesworth B, Jensen JD (2021). *The Impact of Purifying and Background Selection on the Inference of Population History: Problems and Prospects*. *Molecular Biology and Evolution* **38**, 2986–3003. https://doi.org/10.1093/molbev/msab050.

Kato M, Vasco DA, Sugino R, Narushima D, Krasnitz A (2017). *Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer*. *Royal Society of Open Science* **4**. https://doi.org/10.1098/rsos.171060.

Kelleher J, Thornton KR, Ashander J, Ralph PL (2018). *Efficient pedigree recording for fast population genetics simulation*. *PLOS Computational Biology* **14**, e1006581. https://doi.org/10.1371/journal.pcbi.1006581.

Kelleher J, Wong Y, Wohns AW, Fadil C, Albers PK, McVean G (2019). *Inferring whole-genome histories in large population datasets*. *Nature Genetics* **51**, 1330–1338. https://doi.org/10.1038/s41588-019-0483-y.

Ki C, Terhorst J (2020). *Exact decoding of the sequentially Markov coalescent*. *bioRxiv*. https://doi.org/10.1101/2020.09.21.307355.

Kim Y, Koehler F, Moitra A, Mossel E, Ramnarayan G (2020). *How Many Subpopulations Is Too Many? Exponential Lower Bounds for Inferring Population Histories*. *Journal of Computational Biology* **27**, 613–625. https://doi.org/10.1089/cmb.2019.0318.

Kingman J (1982). *The Coalescent*. *Stochastic Processes and their Applications* **13**. https://doi.org/10.1016/0304-4149(82)90011-4.

Kipf TN, Welling M (2016). *Semi-Supervised Classification with Graph Convolutional Networks*. *arXiv*. https://doi.org/10.48550/arXiv.1609.02907.

Korfmann K (2024a). *Kevin Korfmann/GNNcoal-analysis: GNNcoal-analysis (PCI recommendation)*. *Zenodo*. https://doi.org/10.5281/zenodo.10781640.

Korfmann K (2024b). *Kevin Korfmann/GNNcoal: GNNcoal (PCI recommendation)*. *Zenodo*. https://doi.org/10.5281/zenodo.10781643.

Korfmann K, Gaggiotti OE, Fumagalli M (2023). *Deep Learning in Population Genetics*. *Genome Biology and Evolution* **15**. Ed. by Andrea Betancourt, evad008. https://doi.org/10.1093/gbe/evad008.

Korfmann K, Sellinger T, Freund F, Fumagalli M, Tellier A (2024). *Supplementary Information for "Simultaneous Inference of Past Demography and Selection from the Ancestral Recombination Graph under the Beta Coalescent" publication. In PCJ*. *Zenodo*. https://doi.org/10.5281/zenodo.10792836.

Koskela J (2018). *Multi-locus data distinguishes between population growth and multiple merger coalescents*. *Statistical applications in genetics and molecular biology* **17**. https://doi.org/10.1515/sagmb-2017-0011.

Koskela J, Berenguer MW (2019). *Robust model selection between population growth and multiple merger coalescents*. *Mathematical Biosciences* **311**, 1–12. https://doi.org/10.1016/j.mbs.2019.03.004.

Lee JB, Rossi R, Kong X (2018). *Graph Classification using Structural Attention*. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. London United Kingdom, pp. 1666–1674. https://doi.org/10.1145/3219819.3219980.

Lewanski AL, Grundler MC, Bradburd GS (2024). *The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics*. *PLOS Genetics* **20**, e1011110. https://doi.org/10.1371/journal.pgen.1011110.

Li H, Durbin R (2011). *Inference of human population history from individual whole-genome sequences*. *Nature* **475**, 493–U84. https://doi.org/10.1038/nature10231.

Mahmoudi A, Koskela J, Kelleher J, Chan Yb, Balding D (2022). *Bayesian inference of ancestral recombination graphs*. *PLOS Computational Biology* **18**. https://doi.org/10.1371/journal.pcbi.1009960.

Malaspinas Aea (2016). *A genomic history of Aboriginal Australia*. *Nature* **538**, 207+. https://doi.org/10.1038/nature18299.

Marjoram P, Wall J (2006). *Fast "coalescent" simulation*. *BMC Genetics* **7**. https://doi.org/10.1186/1471-2156-7-16.

Matuszewski S, Hildebrandt ME, Achaz G, Jensen JD (2017). *Coalescent Processes with Skewed Offspring Distributions and Non-equilibrium Demography*. *Genetics*. https://doi.org/10.1534/genetics.117.300499.

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

Menardo F, Gagneux S, Freund F (2020). *Multiple merger genealogies in outbreaks of Mycobacterium tuberculosis*. *Molecular Biology and Evolution*. https://doi.org/10.1093/molbev/msaa179.

Miles A, pyup.io bot, R M, Ralph P, Harding N, Pisupati R, Rae S, Millar T (2021). *cggh/scikit-allel: v1.3.3*. https://doi.org/10.5281/zenodo.4759368.

Mohle M, Sagitov S (2001). *A classification of coalescent processes for haploid exchangeable population models*. *Annals of Probability* **29**, 1547–1562. https://doi.org/10.1214/aop/1015345761.

Morales-Arce AY, Harris RB, Stone AC, Jensen JD (2020). *Evaluating the contributions of purifying selection and progeny-skew in dictating within-host Mycobacterium tuberculosis evolution*. *Evolution* **74**, 992–1001. https://doi.org/10.1111/evo.13954.

Neher RA, Hallatschek O (2013). *Genealogies of rapidly adapting populations*. *Proceedings of the National Academy of Sciences* **110**, 437–442. https://doi.org/10.1073/pnas.1213113110.

Nelson D, Kelleher J, Ragsdale AP, Moreau C, McVean G, Gravel S (2020). *Accounting for long-range correlations in genome-wide simulations of large cohorts*. *PLOS Genetics* **16**. https://doi.org/10.1371/journal.pgen.1008619.

Niwa HS, Nashida K, Yanagimoto T (2016). *Reproductive skew in Japanese sardine inferred from DNA sequences*. *ICES Journal of Marine Science* **73**, 2181–2189. https://doi.org/10.1093/icesjms/fsw070.

Paszke A, Gross S, Chintala S, Chanan G, Yang E, DeVito Z, Lin Z, Desmaison A, Antiga L, Lerer A (2017). *Automatic differentiation in PyTorch*. https://openreview.net/forum?id=BJJsrmfCZ.

Pitman J (1999). *Coalescents with multiple collisions*. *Annals of Probability* **27**, 1870–1902. https://doi.org/10.1214/aop/1022677552.

Qin X, Chiang CWK, Gaggiotti OE (2021). *Deciphering signatures of natural selection via deep learning*. *bioRxiv*. https://doi.org/10.1101/2021.05.27.445973.

Rasmussen MD, Hubisz MJ, Gronau I, Siepel A (2014). *Genome-Wide Inference of Ancestral Recombination Graphs*. *PLOS Genetics* **10**. https://doi.org/10.1371/journal.pgen.1004342.

Rice DP, Novembre J, Desai MM (2018). *Distinguishing multiple-merger from Kingman coalescence using two-site frequency spectra*. *bioRxiv*. https://doi.org/10.1101/461517.

Rogers AR, Huff C (2009). *Linkage Disequilibrium Between Loci With Unknown Phase*. *Genetics* **182**, 839–844. https://doi.org/10.1534/genetics.108.093153.

Sackman AM, Harris RB, Jensen JD (2019). *Inferring Demography and Selection in Organisms Characterized by Skewed Offspring Distributions*. *Genetics* **211**, 1019–1028. https://doi.org/10.1534/genetics.118.301684.

Sagitov S (1999). *The general coalescent with asynchronous mergers of ancestral lines*. *Journal of Applied Probability* **36**, 1116–1125. https://doi.org/10.1239/jap/1032374759.

Sagitov S (2003). *Convergence to the coalescent with simultaneous multiple mergers*. *Journal of Applied Probability* **40**, 839–854. https://doi.org/10.1239/jap/1067436085.

Sanchez T, Cury J, Charpiat G, Jay F (2020). *Deep learning for population size history inference: Design, comparison and combination with approximate Bayesian computation*. *Molecular Ecology Resources*. https://doi.org/10.1111/1755-0998.13224.

Sapoval N, Aghazadeh A, Nute MG, Antunes DA, Balaji A, Baraniuk R, Barberan CJ, Dannenfelser R, Dun C, Edrisi M, Elworth RAL, Kille B, Kyrillidis A, Nakhleh L, Wolfe CR, Yan Z, Yao V, Treangen TJ (2022). *Current progress and open challenges for applying deep learning across the biosciences*. *Nature Communications* **13**, 1728. https://doi.org/10.1038/s41467-022-29268-7.

Sargsyan O, Wakeley J (2008). *A coalescent process with simultaneous multiple mergers for approximating the gene genealogies of many marine organisms*. *Theoretical population biology* **74**, 104–114. https://doi.org/10.1016/j.tpb.2008.04.009.

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

Schlichtkrull M, Kipf TN, Bloem P, Berg Rvd, Titov I, Welling M (2017). *Modeling Relational Data with Graph Convolutional Networks*. *arXiv*. https://doi.org/10.48550/arXiv.1703.06103.

Schweinsberg J (2003). *Coalescent processes obtained from supercritical Galton-Watson processes*. *Stochastic Processes and their Applications* **106**, 107–139. https://doi.org/10.1016/S0304-4149(03)00028-0.

Sellinger T (2024). *Sellinger, Thibaut/eSMC2: eSMC2 : 5.1.3*. *Zenodo*. https://doi.org/10.5281/zenodo.10782372.

Sellinger TPP, Abu Awad D, Moest M, Tellier A (2020). *Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data*. *PLOS Genetics* **16**. https://doi.org/10.1371/journal.pgen.1008698.

Sellinger TPP, Abu-Awad D, Tellier A (2021). *Limits and convergence properties of the sequentially Markovian coalescent*. *Molecular Ecology Resources* **21**, 2231–2248. https://doi.org/10.1111/1755-0998.13416.

Sheehan S, Song YS (2016). *Deep Learning for Population Genetic Inference*. *PLOS Computational Biology* **12**. https://doi.org/10.1371/journal.pcbi.1004845.

Speidel L, Forest M, Shi S, Myers SR (2019). *A method for genome-wide genealogy estimation for thousands of samples*. *Nature Genetics* **51**, 1321+. https://doi.org/10.1038/s41588-019-0484-x.

Steinruecken M, Birkner M, Blath J (2013). *Analysis of DNA sequence variation within marine species using Beta-coalescents*. *Theoretical Population Biology* **87**, 15–24. https://doi.org/10.1371/journal.pcbi.1010419.

Stephan W (2019). *Selective Sweeps*. *Genetics* **211**, 5–13. https://doi.org/10.1534/genetics.118.301319.

Struett S, Sellinger T, Glémin S, Tellier A, Laurent S (2022). *Inference of evolutionary transitions to self-fertilization using whole-genome sequences*. *bioRxiv*.

Tellier A, Lemaire C (2014). *Coalescence 2.0: a multiple branching of recent theoretical developments and their applications*. *Molecular Ecology* **23**, 2637–2652. https://doi.org/10.1111/mec.12755.

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

Upadhya G, Steinrücken M (2021). *Robust Inference of Population Size Histories from Genomic Sequencing Data*. *bioRxiv*. https://doi.org/10.1101/2021.05.22.445274.

Wang K, Mathieson I, O'Connell J, Schiffels S (2020). *Tracking human population structure through time from whole genome sequences*. *PLOS Genetics* **16**. https://doi.org/10.1371/journal.pgen.1008552.

Wang Z, Wang J, Kourakos M, Hoang N, Lee HH, Mathieson I, Mathieson S (2021). *Automatic inference of demographic parameters using generative adversarial networks*. *Molecular Ecology Resources* **21**, 2689–2705. https://doi.org/10.1111/1755-0998.13386.

Whitehouse LS, Ray D, Schrider DR (2024). *Tree sequences as a general-purpose tool for population genetic inference*. *bioRxiv*. https://doi.org/10.1101/2024.02.20.581288.

Wilton PR, Carmi S, Hobolth A (2015). *The SMC' Is a Highly Accurate Approximation to the Ancestral Recombination Graph*. *Molecular Biology and Evolution* **200**, 343–U637. https://doi.org/10.1534/genetics.114.173898.

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

Wong Y, Ignatieva A, Koskela J, Gorjanc G, Wohns AW, Kelleher J (2023). *A general and efficient representation of ancestral recombination graphs*. *bioRxiv*. https://doi.org/10.1101/2023.11.03.565466.

Xu K, Hu W, Leskovec J, Jegelka S (2019). *How Powerful are Graph Neural Networks? arXiv*. https://doi.org/10.48550/arXiv.1810.00826.

Yang Z, Cohen WW, Salakhutdinov R (2016). *Revisiting Semi-Supervised Learning with Graph Embeddings*. *arXiv*. https://doi.org/10.48550/arXiv.1603.08861.

Yelmen B, Decelle A, Ongaro L, Marnetto D, Tallec C, Montinaro F, Furtlehner C, Pagani L, Jay F (2021). *Creating artificial human genomes using generative neural networks*. *PLOS Genetics* **17**, 1–22. https://doi.org/10.1371/journal.pgen.1009303.

Ying R, You J, Morris C, Ren X, Hamilton WL, Leskovec J (2019). *Hierarchical Graph Representation Learning with Differentiable Pooling*. *arXiv*. https://doi.org/10.48550/arXiv.1806.08804.

Zhang M, Chen Y (2018). *Link Prediction Based on Graph Neural Networks*. *arXiv*. https://doi.org/10.48550/arXiv.1802.09691.

Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, Wang L, Li C, Sun M (2020). *Graph neural networks: A review of methods and applications*. *AI Open* **1**, 57–81. https://doi.org/10.1016/j.aiopen.2021.01.001.