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

Kevin Korfmann[1*], Thibaut Sellinger[1,2], Fabian Freund[3], Matteo Fumagalli[4], Aurélien Tellier[1]

[1] Professorship for Population Genetics,
Department of Life Science Systems,
Technical University of Munich

[2] Department of Environment and Biodiversity,
Paris Lodron University of Salzburg

[3] Institute of Plant Breeding, Seed Science and Population Genetics
University of Hohenheim

[4] Department of Biological and Behavioural Sciences,
Queen Mary University of London, London

[*] Corresponding author, kevin.korfmann@tum.de

Abstract

The reproductive mechanism of a species is a key driver of genome evolution. The standard Wright-Fisher model depict reproduction of individuals in a population and 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 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 a 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, and the added value of comparing *ad hoc* SMC and GNN methods.

Keywords— kingman coalescent, beta coalescent, selective sweep, deep learning, graph neural networks, population genetics, multiple merger coalescent, sequentially markovian coalescent, ancestral recombination graph

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 [76, 48, 71, 77, 74, 4, 3, 79, 36, 37]. 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 selection scan methods to detect genes under (positive or balancing) selection [79, 38]. A common feature of such inference methods that make full use of whole genome sequences is the underlying assumption of a Kingman coalescent process [43] to describe the genealogy distribution of a sample. The Kingman coalescent process and its properties stem from using the forward-in-time traditional 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 mean one and variance one. To give a biological intuition, it means that most parents can most likely have no, one, or two offspring, but it is improbable that one parent has many offspring (on the order of the population size, *e.g.* N/2 where N is the census size). This assumption of small variance in offspring distribution between individual parents is likely realistic for species with low juvenile (potential offspring) mortality (so-called type I and II survivorship in ecology) such as mammals.

As genome sequence data become available for many different species with as many different biological traits or life cycles, the applicability of the Kingman coalescent relying on the WF model can be questioned [78, 2, 59, 39, 56, 81, 53, 27]. Indeed, for some species with high fecundity and high juveniles mortality (type III survivorship), it is expected that the variance in reproduction between parents can be much larger than under the Poisson distribution [81]. This effect is termed as sweepstake reproduction in fish [32, 2]. Neutral processes such as strong seed banking [10], high fecundity with skewed offspring distribution [32, 22], and extremely strong and recurrent bottlenecks [8, 18] as well as strong selective processes (positive selection) [21, 15, 16, 31] 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. In such conditions, a new class of process arises to describe the genealogy distribution, a class where multiple individuals can simultaneously coalesce and/or multiple distinguished coalescence events can simultaneously occur

[68, 55, 20, 67, 61, 12]. Generally, this class of genealogical processes is called the Multiple Merger Coalescent (MMC). It is suggested that MMC models may be more biologically appropriate than the Kingman coalescent to study many species of fish [23, 2, 32], invertebrates (insects, crustaceans, etc.), viruses [51], bacteria [53, 57], plants and their pathogens [81]. 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 at best costly and time consuming and at worst intractable. Therefore, an alternative solution is to use inference methods based on genome data to distinguish which model best describes the genealogy of a given species/population.

In this study we use the so-called β-coalescent which is a specific class of MMC models. 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 [73]. The latter is a general class of coalescent process describing how and how fast ancestral lineages merge [61, 67]. When using the Beta($2-\alpha$,α) distribution as a probability measure for the Λ-coalescent, the transition rates (*i.e.* coalescent rate) can be analytically obtained leading to the thus named β-coalescent, which is a specific model of MMC. If α tends to 2, then the coalescent process tends to a Kingman coalescent (up to a scaling constant) [7, 45, 46]. If α tends to one, the model tends to a Bolthausen-Sznitman coalescent process (*i.e.* dominated by strong multiple merger events) [12]. The β-coalescent has the property that the observed polarized SFS of a sample of single nucleotide polymorphisms (SNPs) exhibits a characteristic U-shaped with an excess of rare and high frequency variants (compared to the Kingman coalescent). 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) [46, 31, 66], minor allele frequency [64] or copy number alteration [39]. It is shown that the SFS is robust to the effect of recombination [46, 64] 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 [46, 45, 23]. However, these methods relying on the genome-wide SFS present two main disadvantages: 1) they suffer from non-identifiability ([36]) as several complex neutral demographic and/or selective models under the Kingman or MMC models can generate similar SFS distributions, and 2) they summarize one genealogy underlying all SNPs and thus require high sample sizes (>50) for trustworthy results [46, 45, 23], a prohibitive requirement for the study of non-model species. To tackle these limitations, we develop two methods that integrate recombination 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) [35, 7]. Due to the fact that all the genealogical information is contained in the ARG, in this study we aim at the interpretation of the ARG to recover model parameters in presence of multiple merger events. With the development of the sequentially Markovian coalescent theory [52, 50], it becomes tractable (*i.e.* possible) to integrate linkage disequilibrium over chromosomes in inferences based on the Kingman coalescent [48]. Hence, we first develop an SMC approach based on the β-coalescent named the Sequentially Markovian β Coalescent (SMβC). However, the β-coalescent has the additional property that under recombination, long range dependency is generated between coalescent trees along the genome [7]. In other words, coalescent trees which are located at different places in the genome, and expected to be unlinked from one another [58], would show meaningful 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, a concentration of ancestry of many lineages in one family (or even individual) in the past, which then leaves traces in the genome at several loci [8]. To overcome the theoretically predicted non-Markovian property of the distribution of genealogies along the genome under the β-coalescent with recombination [7] we develop a second method based on deep learning (DL) and facilitated by the availability of an efficient coalescent simulator [6]. In evolutionary genomics, DL approaches trained by simulations are shown to be powerful inference tools [76]. Previous work demonstrated that DL approach can help overcome problems mathematically insolvable or computationally intractable in the field of population genetics [76, 5, 85, 88, 26, 19, 62]. The novelty on our neural network relies on its structure (Graph Neural Network, GNN) and that it is trained on the ARG of a sample, more precisely on the tree sequence [40]. GNNs are an emerging category of DL algorithm ([14, 86, 17, 91]), that benefit by using irregular domain data (*i.e.* graphs). GNNs are designed to the prediction of node features [44, 87], edge features (link prediction) [90, 72], or additional properties of entire graphs [89, 47]. Therefore, GNNs represent a new tool to address the large number of possibilities in multiple ARG iterations, 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 [69].

We describe our two methods, SMβC and GNN*coal*, and demonstrate their statistical power as well as their respective limitations. We first quantify the bias of previous SMC methods when performing inference of past population size variation under the β-coalescent. We then assess the accuracy of both approaches to recover the α parameter of the Beta-distribution, that is how frequent and strong multiple merger events are, and the past variation of population size from tree-sequence data. We demonstrate that our approaches can recover 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 of our approaches can model and recover the presence of selection along the genome while simultaneously accounting for non-constant population size, recombination and skewed offspring distribution.

Materials and Methods

As explained in the introduction, throughout the manuscript we assume the ARG to be perfectly known. Hence, the ARG of the sample is given as input and from this our methods recover model parameters we are interested in (*e.g.* the α parameter and/or the past variation of population size).

SMC-based method

In this study, we use different SMC-based methods: two previously published and MSMC and MSMC2 [71, 84], and the new SMβC (the software backbone stems from our previous eSMC [74, 75], see Appendix 1). All methods use the ARG (genealogies along the genome) as input throughout the manuscript. Hence MSMC and MSMC2 have been re-implemented in the R package eSMC2 to take the ARG as input. MSMC2 focuses on the coalescence events between two haploid samples along the genome, and 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 [24, 84]. MSMC simultaneously analyses multiple sequences and follows the distribution of the first coalescence event in a sample of size $n > 2$ along sequences. A detailed description of MSMC can be found in [71]. We have re-implemented MSMC and MSMC2 in the R package eSMC2 (`https://github.com/TPPSellinger/eSMC2`) so that they can take the ARG as input.

Our new method, SMβC, is based on MSMC and thus simultaneously analyses multiple sequences and focuses on the first coalescence event. The SMβC follows the distribution of the first coalescence event of a sample along sequences assuming a β-coalescent process. Therefore, the model allows for more than two individuals to join the first coalescence event, or new lineages to join an already existing binary (or triple) coalescent event. The emission matrix is similar to the one of MSMC. Hence, the SMβC extends the MSMC by simply adding hidden states where 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). As in MSMC, the population size is assumed piecewise constant in time and we discretize time in 40 bins throughout this study. A detailed description of SMβC can be found in Appendix 1. To test and validate the theoretical accuracy of our approach, we study its best case convergence throughout this manuscript. The best-case convergence, introduced in [75], corresponds to the

model's performance when the true genealogy is inputted as hidden state (*i.e.* as if the hidden states were perfectly recovered). All approaches are included in the latest version of the R package eSMC2 (`https://github.com/TPPSellinger/eSMC2`).

Tree-sequence GNN*coal* method

As coalescent trees are a subset of graphs (which forms a representation of ongoing evolutionary processes), the GNN*coal* is well suited to leverage information from data similar to what is inputted to SMC approaches. The information of a given coalescent tree which is interpreted as an undirected graph (see Appendix 2) is first the topology information of the coalescent branches and second the age of the coalescent events as encoded in the branch/edge lengths. Thus, the network input data compromised of a matrix containing the edge information, the number of edges and their lengths and a randomly initialized feature matrix, containing an embedding vector for each node, which will be continuously updated during the training in the following way: The model aggregates coalescent branch lengths by multiplying them with learnable weight parameters from the network architecture, with respect to the local neighbourhood of the nodes in the coalescent tree. Therefore, each node learns about the distance to each surrounding ancestral or sample node. This operation, known as a graph convolution, is analogous to the procedure in regular convolutional networks [44]. The aim is local feature extraction from a complex data input source into a compressed data representation. To overcome the locality of the graph convolution, an additional pooling strategy is implemented as part of the network [89]. This strategy hierarchically compresses the coalescent tree into smaller subtrees until the final representation is fed into a feed-forward network. The demography is estimated for each coalescent tree per simulation, and the batch dimension was used to integrate 500 trees of the same demography per forward pass. A masked root-mean-squared error (RMSE) loss function is computed for the entire batch with respect to the relevant time windows based on a heuristic of the number of coalescent events per window. As a consequence the gradients of these trees are being accumulated, capturing the stochastic complexity, which arises from the underlying coalescent model. A detailed description of the GNN*coal* as well as the coalescent event masking feature can be found in Appendix 2. The code of the model architecture is implemented in *Pytorch* [60], using the extension *Pytorch Geometric* [25]. The model is available along side the training dataset simulations at `https://github.com/kevinkorfmann/GNNcoal` and `https://github.com/kevinkorfmann/GNNcoal-analysis`.

Simulation of data

Validation dataset for both methods

The ARG is given as input to the DL approach and the SMβC as described in [75]. We use msprime [6] to simulate the ARG under the β-coalescent based on [73, 7] or under the Kingman coalescent (under neutrality and under selection). Throughout the manuscript, 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 [82, 71, 75]). We simulate data under different α value (*i.e.* parameter of the Beta-distribution) ranging from 1.9 (almost no multiple merger events),1.7,1.5 to 1.3 (frequent and strong multiple merger events). Throughout the manuscript the 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 amount of SNPs. In some stated and specific scenarios the recombination and mutation rate were set differently in order 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://github.com/kevinkorfmann/GNNcoal-analysis`.

Training dataset for the GNN*coal*

We generated an extensive dataset to allow the neural network to generalise over the complex branch length distributions of the ARG, including all arising variations that occur due to demographic changes in-time. These distributions are further being convoluted by variations of the multiple merger strength coefficient α, substantially decreasing the amount of retrievable information per simulation. The model parameters were updated based on a supervised manner. The loss function was calculated with respect to how much off the machine-learning based estimate was compared to the true simulator parameters after each batch. The simulation of the past demographic history is based on the strategy from [11, 69] and modified into a two step process. First $I = 18$ time points are chosen with $[t_i, t_{i+1}]$ from recent to ancient periods with $t_i = \frac{1}{\eta}((1+\eta T)^{i/(I-1)} - 1$ generations, i initialised to be between 0 and $I-1$, $T = 10^6$ and $\eta = 0.1$. The first population size N_0 is drawn for the initial time point t_0 between 10^4 and 10^7, and the next population sizes are drawn, respectively based on the formula $N_i = N_{i-1} \times 10^{\chi}$, χ likewise being sampled from a uniform distribution. Second, we introduce a smoothing of the demography, by cubically interpolating between the I time points, and choosing 60 uniformly spaced new time points of the interpolation in log space. All time points below ten generations were discarded, leaving approx. 50 time points at

which population size is estimated. The main difference being, that this demography sampling strategy leads to piece-wise constant changes of the population size compared to continuous population size changes inferred in the SMβC and used in the validation dataset. The discrepancy between the two methods was buffered by the cubic time-point smoothing described above. An example of this demography sampling process can be seen in Appendix 2. No mutations were simulated and the recombination rate was kept constant at 10^{-8} per generation per bp.

Disentangling coalescent and selection signatures

We use the classification property of GNNs to test whether it is possible to disentangle between different scenarios / models. In total, we defined 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 (]2.0-1.75],]1.75-1.5],]1.5-1-1.25],]1.25-1.01]) without selection. The three different selection regimes are defined by random uniformly sampling a selection coefficient in the following ranges: [0.1, 0.01[for SS, [0.01, 0.001[for MS, [0.001, 0.0001[for WS and [0] for absence of selection. Demography was kept constant at 10^5 individuals, sequence length was set initially to 10^5 bp. The simulation was discarded, if it resulted in less than 2,000 obtained trees and reran with twice the sequence length until the tree number required was satisfied. The selection site was introduced at the middle of the respective sequence, so that 249 trees left and 250 right of the middle tree under selection form a training sample, thus using again 500 trees for each sample. One hundred replicates were generated for each training sample. The complete training dataset consists of 1,000 parameter sets, 500 for the Kingman cases and 500 for the β-coalescent cases, with approximately 125 parameter sets per class. The model itself was trained on one epoch, and the evaluation performed afterwards on 1,000 randomly generated parameter sets, with one replicate per parameter set. The same architecture as for demography estimation was used with two added linear layers, separated by ReLU function to reduce the number of output dimensions of 60 to 8 and the loss function was exchanged for Cross-Entropy-Loss instead of RMSE. Otherwise all architecture and training parameters were the same as described above and detailed in Appendix 2.

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 1. 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 correctly infer the correct population size, due to the scaling discrepancy between the Kingman and β-coalescent. Hence, we perform the same analysis and correct for the scaling effect of the MMC versus a Kingman coalescent in order to better capture the specific effects of assuming binary mergers only. The results are displayed in Supplementary Figure 1. 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 variance observed is extremely high and a flattened past variation of population size is observed.

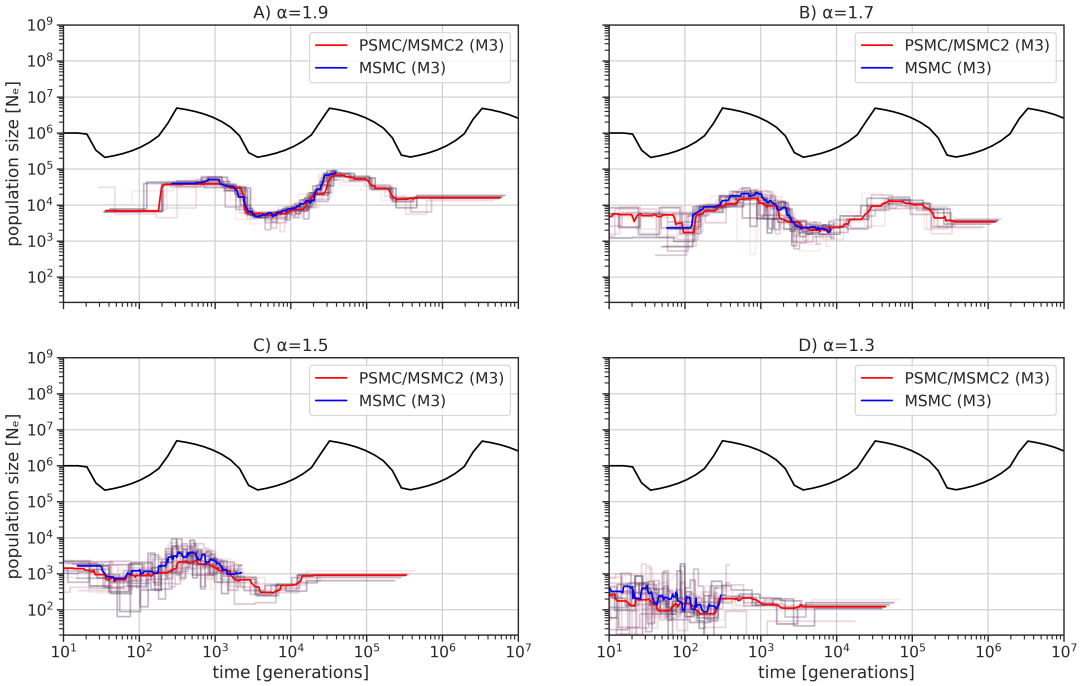

Fig. 1. Performance of MSMC and MSMC2 under a β-coalescent. Averaged estimated demographic history by MSMC (blue) and MSMC2 (red) based on 10 sequences of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp over ten repetitions (while analyzing simultaneously 3 sequences, noted by M=3). Each repetition 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 of pairs of SNPs (r^2, [65, 54]) in data simulated under the Kingman Coalescent or the β-coalescent (with $\alpha = 1.5$ and $\alpha = 1.1$) and constant population size (Figure 2). Linkage monotonously decreases with distance under the Kingman coalescent but presents a non-monotonous "U-shape" distribution under the β-coalescent. The "U-shape" linkage is more pronounced for smaller α values. This means that linkage decreases with distance until a certain point and then increases. This "U-shape" linkage disequilibrium demonstrates that the Markovian hypothesis to model genealogies along the genome is violated under the β-coalescent due to the long range effect of

strong multiple merger events [7].

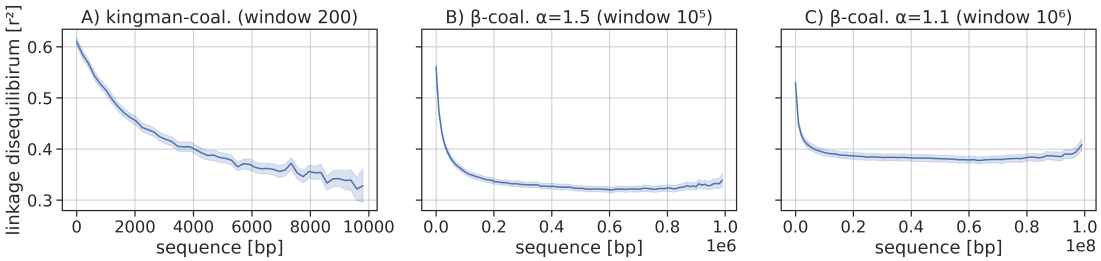

Fig. 2. Pairwise linkage disequilibrium between SNPs (r^2) under a Kingman and β-coalescent with $\alpha = 1.5$ and $\alpha = 1.1$ using 1,000 sequences of length A) 0.01 Mb, B) 1Mb, and C) 100Mb. The population size is constant at $N = 10^4$ for the Kingman model and $N = 10^6$ for the β-coalescent, with $\mu = 10^{-7}$ and $r = 5.10^{-8}$ per generation per bp.

We further investigate the effect of the observed "U-shape" linkage disequilibrium on the SMC by looking at the absolute difference between the observed coalescent time transitions and one predicted by the SMC (under a constant population size). Under the Kingman coalescent, the distribution of coalescent times between two positions in a sample of size two ($n = 2$) is well approximated by the SMC as shown in Supplementary Figure 2. However, under the β-coalescent (with $\alpha = 1.3$) we observe strong differences at times points where multiple merger events occur (Supplementary Figure 3). Indeed, multiple merger events are not evenly distributed in time, unveiling a discrepancy between the expectation from the SMC (*i.e.* approximating the distribution of genealogies along the genome by a Markov chain) and the simulated data. This discrepancy does not stem from the simulator, because it correctly generates ARG under the β-coalescent model [7, 6], but from the limits of the SMC approximation to model events with long range effects on the ARG (Supplementary Figure 3).

Inferring α and past demography

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 3 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 DL approach exhibits high accuracy and low variance to infer the variation of population size. For high α values (>1.5), the shape of population size variation is well recovered by SMβC. However, for smaller values, an extremely high variance is observed which demonstrates the limits of SMC inferences. On average, both approaches seem to recover fairly well the

true α value (Figure 3 and Supplementary Table 1), although GNN*coal* performs slightly better and displays 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 (Supplementary Table 1). These conclusions are also valid for the results in Supplementary Figures 4-7 and Supplementary Table 1 based on inference under four additional demographic scenarios: constant population size, bottleneck, sudden increase and sudden decrease of population size.

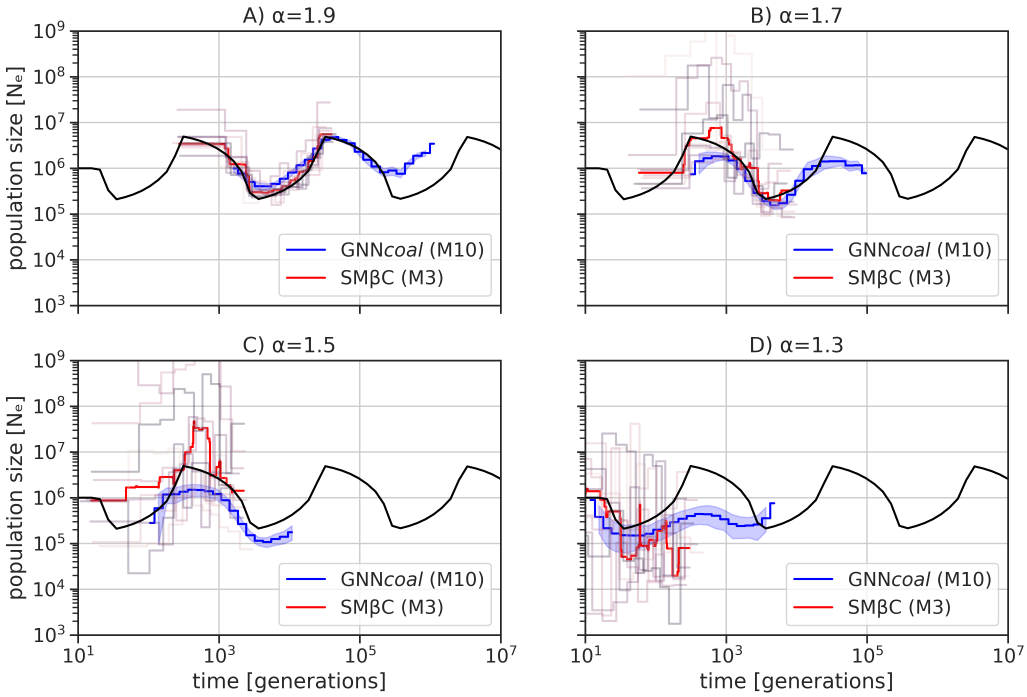

Fig. 3. Best-case convergence estimations of SMβC and GNN*coal* under a β-coalescent. Estimations of past demographic history by SMβC in red (median) and by GNN*coal* in blue (mean and 95% confidence interval (CI95); while analyzing simultaneously 3 or 10 sequences, noted by M=3 or M=10) when population 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.

Because when α diminishes, the effective population size decreases as well, the number of recombination events plummets for small α values (<1.7). To demonstrate the theoretical accuracy of the SMβC, *i.e.* the convergence to the correct

values, we run SMβC on data simulated with mutation and recombination rate fifty times higher under similar scenarios as in Figure 3, namely to increase the amount of data (SNPs and number of independent coalescent trees by recombination). Results of SMβC for α values of 1.7, 1.5 and 1.3 are displayed on Supplementary Table S2. Supplementary Table S2 shows that SMβC can recover α with higher accuracy when given more data (*i.e.* recombination events).

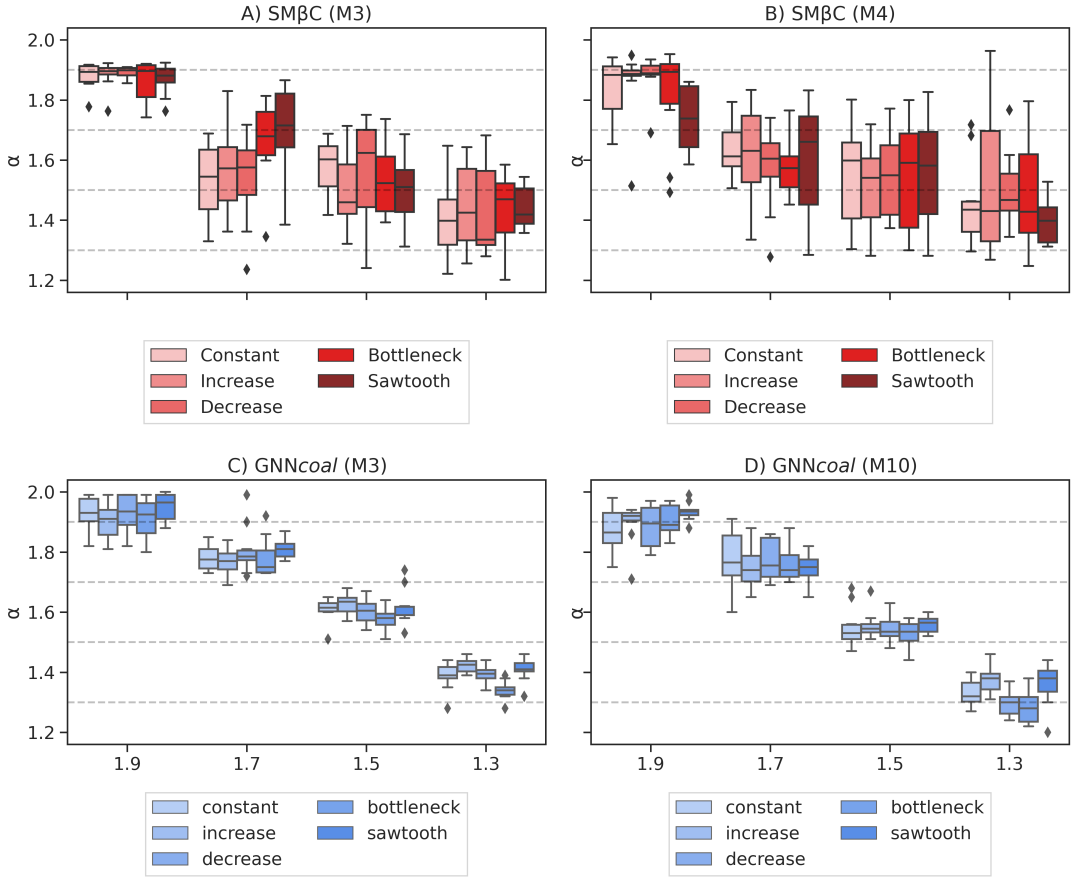

Fig. 4. Estimated values of α by SMβC and GNN*coal* over ten repetitions using 10 sequences of 100 Mb with $\mu = r = 10^{-8}$ per generation per bp under a β-coalescent process (with different α parameter). The analysis were run on five different demographic scenarios (constant population size, bottleneck, 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 inferred α values.

Although 10 sequences are given to the SMβC in the previous analyses, the

method can only simultaneously analyze three or four simultaneously. However, the GNN*coal* approach 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 whether it benefits from simultaneously analyzing bigger sample size. Thus, we run GNN*coal* on the same dataset, but downsampling the coalescent trees to a sample size three. Results are displayed in Figure 4 and Supplementary Figure 8. Results with sample size three of the GNN*coal* approach are similar to results with sample size 10, proving 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 (*i.e* an underlying Kingman coalescent process). The associated inferred demographies are show in Supplementary Figures 9-13. Both approaches correctly infer the past demographic shape up to the scaling discrepancy between the Beta and the Kingman coalescent mentioned above. Furthermore, we notice that the scaling effect only affect the y-axis for the SMβC but affect both axis for GNN*coal*.

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)

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.

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. Hence, we run the 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 Supplementary Table S3. The associated inference of population size are plotted in Supplementary Figure 8-12. 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-axis).

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. We then use the trained GNN*coal* to cluster and 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 5 in the form of a confusion matrix, that is the percentage of times the GNN*coal* correctly assigns the true model shown on the diagonal (evaluated on a test dataset of size 1000). Our approach can accurately select the model except in two cases. The model struggles to distinguish strong from average selection under the Kingman coalescent and it struggles to distinguish the β-coalescent with small amount of multiple merger event (*i.e.* $\alpha>1.75$) with the β-coalescent case with $1.75>\alpha>1.5$.

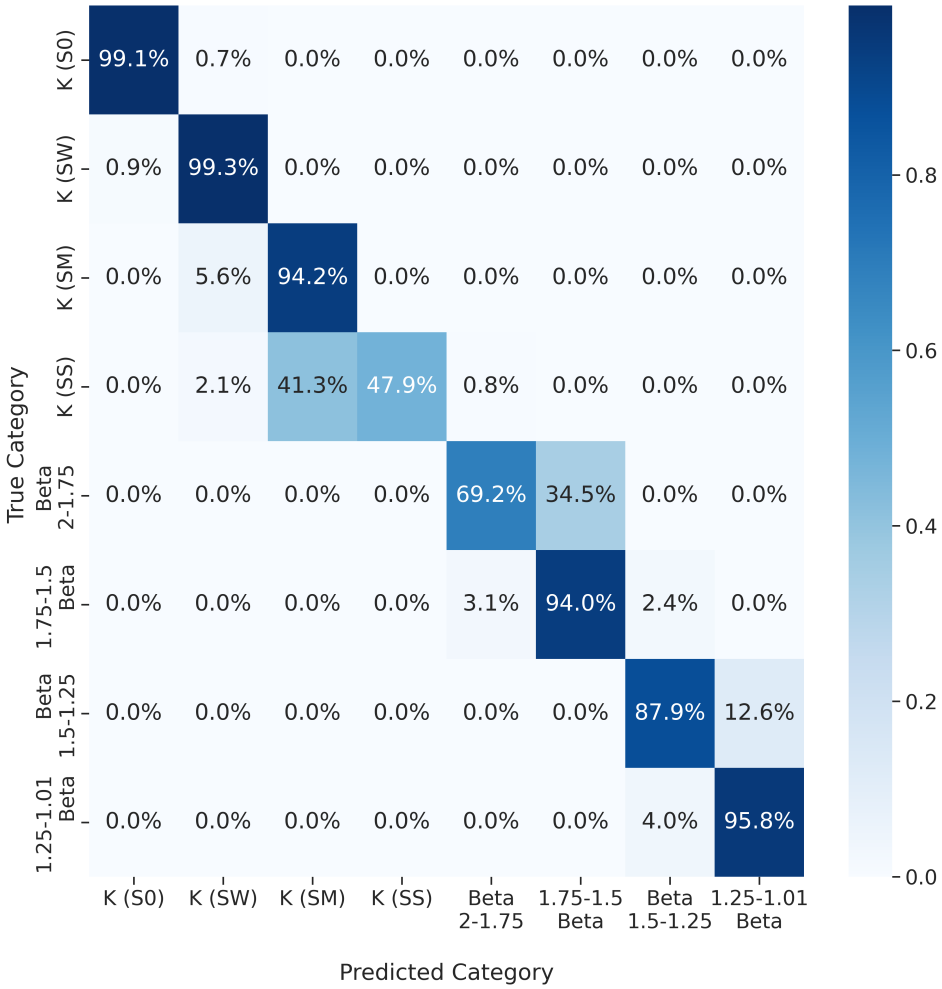

Fig. 5. Confusion matrix for Kingman and β-coalescent classification model under varying selection coefficients. Evaluation of classification accuracy for Kingman (K) and β-coalescent (B) for no selection (S0), weak selection (SW), medium selection (SM) and strong selection (SS) using a 1,000 repetition validation dataset. Each simulation is comprised of least 2,000 coalescent trees. Population size was kept constant at $N = 10^5$ individuals, using a sample size $n = 10$ and $r = 10^{-8}$ per bp per generation.

To assess the effect of selection, we infer the α along the genome with both approaches from data 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 the GNN*coal* approach and for the SMβC are displayed in Figure 6. Both approaches recover smaller α value around the locus under strong selection. However under neutrality or weak selection, inferred α values remain high (>1.6) along the genome.

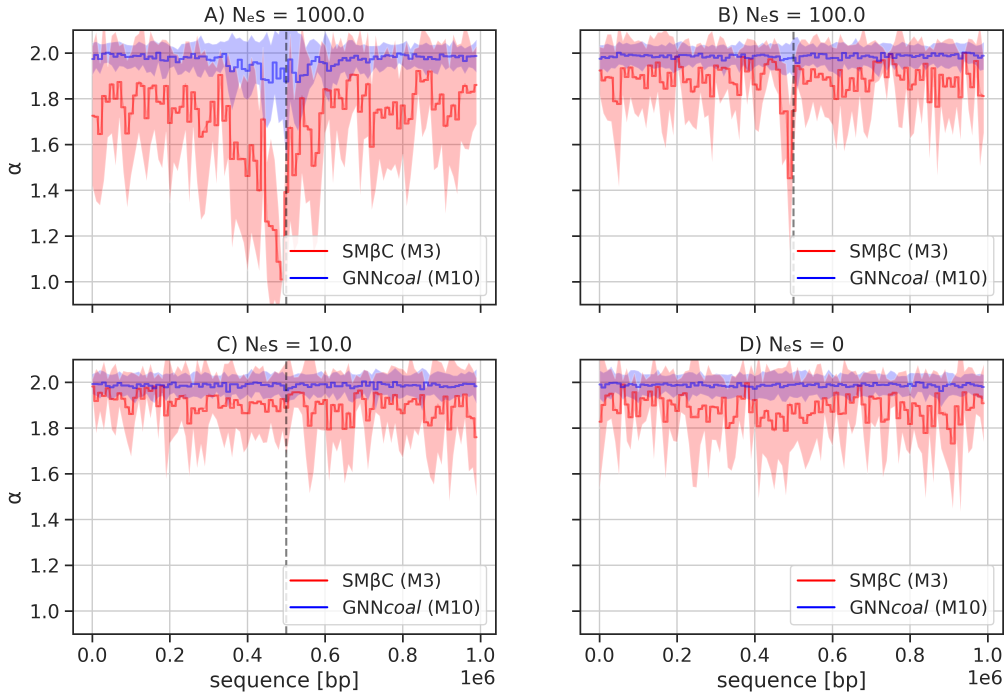

Fig. 6. Averaged estimations by GNN*coal* and SMβC under selection Estimations of α along the genome by the GNN*coal* approach and the SMβC when population undergoes as strong positive selective sweep event (at position 0.5 Mb) under different strengths of selection: A) $s = 0.01$, B)$s = 0.001$, C) $s = 0.0001$, 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 data 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 Supplementary Figure 14. GNN*coal* is able to recover the α value along the genome (also with overestimation due to tree

sparsity), and the SMβC systematically underestimate these α values. However, unlike in presence of positive selection at a given locus, the inferred α values are found in all cases to be fairly similar along the genome.

To prove our approaches can simultaneously recover population size while integrating for selection, we simultaneously recover the α value along the genome and the population size. To do so, we simulate data undergoing a strong selective sweeps or under neutrality while the variation of population size through time is a sawtooth. Under neutrality, our approach recovers high α value 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 only trained on the β-coalescent) (Figure 7). 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 7).

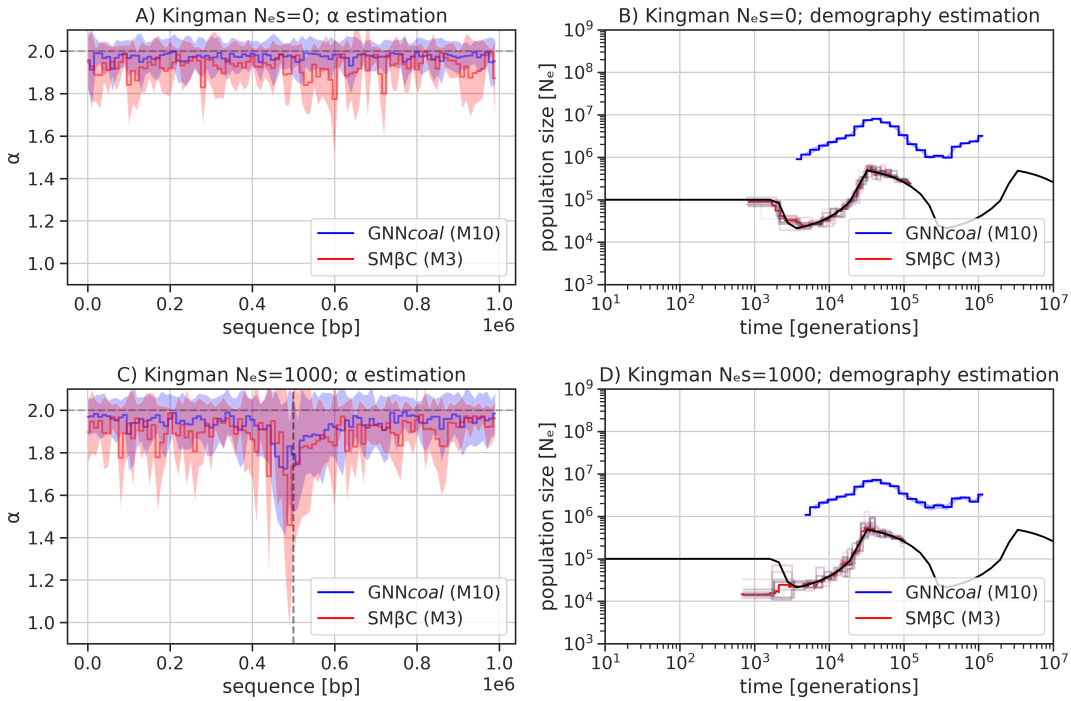

Fig. 7. Simultaneous estimations of α along the sequence under demographic change by GNN*coal* and SMβC. Simultaneous estimation of α along the genome under a partial "sawtooth" scenario: A) and B) in the absence of selection (mean and standard deviation for both methods), and C) and D) presence of selection with $N_eS = 1,000$ (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 of the SMC approaches [48], most current methods leverage information from whole genome sequences by simultaneously reconstructing a portion of the ARG in order to infer past demographic history [48, 71, 82, 83], migration rates [42, 84], variation in recombination and mutation along the genome [4], as well as ecological life history traits such as selfing or seed banking [74, 75, 80]. However in previous studies authors discussed uncoupling both steps, namely reconstructing the ARG and then inferring parameters from its distribution [75, 29, 63]. Indeed, recent efforts have been made to improve approaches recovering the ARG [77, 41, 34, 63, 49, 13], while other focus on its interpretation [28, 75]. In this study we aim at the interpretation of the ARG in presence of multiple merger events. 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 methods can recover the α parameter and the variation of past population size for α values high enough (*i.e.* $\alpha > 1.3$), however for lower values of α a larger amount of data is necessary for any inference (which becomes nearly impossible when α tends to one). Both approaches can also recover the Kingman coalescent (*i.e.* α>1.8). We find the GNN*coal* approach outperforms the SMβC approach in almost all cases, and one can use GNN*coal* to disentangle (classify) between models of β-coalescent and Kingman models with selection.

Overall, our results provide a qualitative increase in the developments of inference methods for models with multiple merger events, a key step to understand the underlying reproduction mechanism of a species. We first develop a new SMC method which takes simultaneously three or four sequences to draw inference of the coalescent tree and transitions along the genome. While fairly accurate, SMβC is outperformed by GNN*coal*. 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 both methods. First, the SMβC approach suffers from the limit of the sequential Markovian coalescent hypothesis along the genome when dealing with strong multiple merger events [7, 18]. Second, the SMC approach requires 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 of time. 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 (as MSMC, SMβC only focuses on the first coalescent event) and cannot simultaneously analyze large sample size. Furthermore, the SMβC approach leverages information from the distribution of genealogy's variation along the genome. It means that in the near absence of recombination events, both approaches cannot utilize any information from the genealogy itself. However, the GNN*coal* approach can overcome this by increasing 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). As for example, if during one hidden state one strong multiple merger event occurs, multiple merger events are seldom observed and SMβC may rather assumes 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, the GNN*coal* approach makes use of the whole ARG, and can easily scale to big sample size (>10), although it recovers α with high accuracy with sample size 3

only. Our interpretation is thus that the GNN*coal* is able of simultaneously leveraging information from topology and the age of coalescent events (nodes) across several genealogies (here 500). The GNN*coal* approach ultimately leverages information from observing recurrent occurrence of same multiple merger events at different location on the genome, while knowing true multiple merger events from rapid successive binary mergers. We believe that our results pave the way towards the interpretability of GNN and DL methods applied to population genetics.

Among the major forces driving the genome evolution there are the past demographic history, the reproductive mechanisms and natural selection [36]. Hence, in the second part of this manuscript we focus on integrating selection in both approaches. Currently, no method can account for the presence of selection, linkage, non-constant population size and multiple merger events [36] although recent theoretical framework might render this possible in the future [1]. As a first step to fill this gap, we demonstrate that the GNN*coal* approach can be used for model selection (classification), reducing the number of potential hypothesis. Determining which evolutionary forces are driving the genome evolution is key, as only under the appropriate neutral population model can results of past demography and selection scans be correctly interpreted [36, 38]. The high accuracy of the GNN*coal* approach in model selection is promising, especially as other methods based on the SFS alone [46, 39] suffer from identifiability in presence of complex demographic scenarios. We follow a thread of previous work [66, 33, 9], 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. In presence of weak selection, no effect on the estimated α value is observed, demonstrating that weak selection can be modeled by binary merger (no multiple merger) and only has a local effect on the branch length (*i.e.* by shortening it). Hence, our results point out that strong selection can indeed be seen/modeled as a local multiple merger event (see [21, 9, 66]). In theory, both approaches could be able to infer the global α parameter linked to the reproductive mechanism, 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. Hence, we suggest that current sequence simulators [6, 30] 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 new state of the art SMC approach to demonstrate that computational and mathematical problems can be overcome by

DL (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 Coalescent [22]) 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 [4]. This would allow to account for the local effect of strong and weak selection [1]. 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 (Supplementary Figure 2). 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 the DL approach 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.

Currently, the applications of our methods are constrained by the assumptions of haploidy and the insufficient accuracy of ARG estimators. Throughout the study we have assumed haploid data, but the GNN*coal* and SMβC approach can be extended to diploidy [7]. Moreover, we have assumed that the exact ARG is available and therefore used as input, but in practice this may not be realistic for many species [13]. Therefore, future efforts are to be made on testing and correcting for inaccurately inputted ARGs. We conclude that 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 DL approaches with model driven *ad hoc* SMC methods may have the potential to help us solve issues (*e.g.* 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) which have puzzled theoreticians and statisticians since the dawn of population genetics [36]. On a final note, as environmental considerations hit us all, we suggest that decreasing the computer and power resources needed to perform DL/ GNN analyses should be attempted [70]. 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 DL or Hidden Markov methods.

Acknowledgments

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). MF and AT acknowledge the support from the Imperial College - TUM Partnership award.

References

[1] Frederic Alberti, Carolin Herrmann, and Ellen Baake. Selection, recombination, and the ancestral initiation graph. *THEORETICAL POPULATION BIOLOGY*, 142:46–56, DEC 2021.

[2] Einar Arnason and Katrin Halldorsdottir. Nucleotide variation and balancing selection at the Ckma gene in Atlantic cod: analysis with multiple merger coalescent models. *PEERJ*, 3, FEB 24 2015.

[3] Gustavo V. Barroso and Julien Y. Dutheil. Mutation rate variation shapes genome-wide diversity in *Drosophila melanogaster*. preprint, Evolutionary Biology, September 2021.

[4] Gustavo V. Barroso, Natasa Puzovic, and Julien Y. Dutheil. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLOS Genetics*, 15(11), NOV 2019.

[5] CJ Battey, Peter L Ralph, and Andrew D Kern. Predicting geographic location from genetic variation with deep neural networks. *eLife*, 9:e54507, June 2020.

[6] Franz Baumdicker, Gertjan Bisschop, Daniel Goldstein, Graham Gower, Aaron P. Ragsdale, Georgia Tsambos, Sha Zhu, Bjarki Eldon, E. Castedo Ellerman, Jared G. Galloway, Ariella L. Gladstein, Gregor Gorjanc, Bing Guo, Ben Jeffery, Warren W. Kretzschumar, Konrad Lohse, Michael Matschiner, Dominic Nelson, Nathaniel S. Pope, Consuelo D. Quinto-Cortes, Murillo F. Rodrigues, Kumar Saunack, Thibaut Sellinger, Kevin Thornton, Hugo van Kemenade, Anthony W. Wohns, Yan Wong, Simon Gravel, Andrew D. Kern, Jere Koskela, Peter L. Ralph, and Jerome Kelleher. Efficient ancestry and mutation simulation with msprime 1.0. *GENETICS*, 220(3), MAR 3 2022.

[7] Matthias Birkner, Jochen Blath, and Bjarki Eldon. An Ancestral Recombination Graph for Diploid Populations with Skewed Offspring Distribution. *Genetics*, 193(1):255–290, JAN 2013.

[8] Matthias Birkner, Jochen Blath, Martin Moehle, Matthias Steinruecken, and Johanna Tams. A modified lookdown construction for the Xi-Fleming-Viot process with mutation and populations with recurrent bottlenecks. *arXiv:0808.0412*, 2008.

[9] Gertjan Bisschop, Konrad Lohse, and Derek Setter. Sweeps in time: leveraging the joint distribution of branch lengths. *GENETICS*, 219(2), OCT 2021.

[10] Jochen Blath, Adrian Gonzalez Casanova, Noemi Kurt, and Maite Wilke-Berenguer. The seed bank coalescent with simultaneous switching. *Electronic Journal of Probability*, 25, 2020.

[11] Simon Boitard, Willy Rodríguez, Flora Jay, Stefano Mona, and Frédéric Austerlitz. Inferring population size history from large samples of genome-wide molecular data - an approximate bayesian computation approach. 12(3):e1005877.

[12] Erwin Bolthausen and A-S Sznitman. On ruelle's probability cascades and an abstract cavity method. *Communications in mathematical physics*, 197(2):247–276, 1998.

[13] Debora Y. C. Brandt, Xinzhu Wei, Yun Deng, Andrew H. Vaughn, and Rasmus Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *GENETICS*, 221(1), MAY 5 2022.

[14] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, jul 2017.

[15] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Noisy traveling waves: Effect of selection on genealogies. *Europhysics Letters*, 76(1):1–7, OCT 2006.

[16] E. Brunet, B. Derrida, A. H. Mueller, and S. Munier. Effect of selection on ancestry: An exactly soluble case and its phenomenological generalization. *Physical Review E*, 76(4, 1), OCT 2007.

[17] Wenming Cao, Zhiyue Yan, Zhiquan He, and Zhihai He. A comprehensive survey on geometric deep learning. *IEEE Access*, 8:35929–35949, 2020.

[18] Adrián González Casanova, Verónica Miró Pina, and Arno Siri-Jégousse. The Symmetric Coalescent and Wright-Fisher models with bottlenecks. *arXiv:1903.05642 [math]*, September 2020. arXiv: 1903.05642.

[19] Jianhai Chen, Pan Ni, Xinyun Li, Jianlin Han, Ivan Jakovlic, Chengjun Zhang, and Shuhong Zhao. Population size may shape the accumulation of functional mutations following domestication. *BMC Evolutionary Biology*, 18, JAN 19 2018.

[20] P Donnelly and TG Kurtz. Particle representations for measure-valued population models. *Annals of Probability*, 27(1):166–205, JAN 1999.

[21] R Durrett and J Schweinsberg. A coalescent model for the effect of advantageous mutations on the genealogy of a population. *Stochastic Processes and their Applications*, 115(10):1628–1657, OCT 2005.

[22] B Eldon and J Wakeley. Coalescent processes when the distribution of offspring number among individuals is highly skewed. *Genetics*, 172(4):2621–2633, APR 2006.

[23] Bjarki Eldon, Matthias Birkner, Jochen Blath, and Fabian Freund. Can the Site-Frequency Spectrum Distinguish Exponential Population Growth from Multiple-Merger Coalescents? *Genetics*, 199(3):841+, MAR 2015.

[24] Malaspinas et al. A genomic history of Aboriginal Australia. *Nature*, 538(7624):207+, OCT 13 2016.

[25] Matthias Fey and Jan Eric Lenssen. Fast graph representation learning with PyTorch geometric.

[26] Lex Flagel, Yaniv Brandvain, and Daniel R Schrider. The Unreasonable Effectiveness of Convolutional Neural Networks in Population Genetic Inference. *Molecular Biology and Evolution*, 36(2):220–238, 12 2018.

[27] Fabian Freund, Elise Kerdoncuff, Sebastian Matuszewski, Marguerite Lapierre, Marcel Hildebrandt, Jeffrey D. Jensen, Luca Ferretti, Amaury Lambert, Timothy B. Sackton, and Guillaume Achaz. Interpreting the pervasive observation of U-shaped Site Frequency Spectra. preprint, Evolutionary Biology, April 2022.

[28] L. M. Gattepaille, M. Jakobsson, and M. G. B. Blum. Inferring population size changes with sequence and SNP data: lessons from human bottlenecks. *Heredity*, 110(5):409–419, MAY 2013.

[29] Lucie Gattepaille, Torsten Guenther, and Mattias Jakobsson. Inferring Past Effective Population Size from Distributions of Coalescent Times. *Molecular Biology and Evolution*, 204(3):1191+, NOV 2016.

[30] Benjamin C. Haller, Jared Galloway, Jerome Kelleher, Philipp W. Messer, and Peter L. Ralph. Tree-sequence recording in slim opens new horizons

for forward-time simulation of whole genomes. *MOLECULAR ECOLOGY RESOURCES*, 19(2):552–566, MAR 2019.

[31] Rebecca B. Harris and Jeffrey D. Jensen. Considering genomic scans for selection as coalescent model choice. *GENOME BIOLOGY AND EVOLUTION*, 12(6):871–877, JUN 2020.

[32] Dennis Hedgecock and Alexander I. Pudovkin. Sweepstakes reproductive success in highly fecund marine fish and shellfish: a review and Commentary. *Bulletin of Marine Science*, 87(4):971–1002, OCT 2011.

[33] Hussein A. Hejase, Ziyi Mo, Leonardo Campagna, and Adam Siepel. A deep-learning approach for inference of selective sweeps from the ancestral recombination graph. *MOLECULAR BIOLOGY AND EVOLUTION*, 39(1), JAN 7 2022.

[34] Melissa Hubisz and Adam Siepel. Inference of ancestral recombination graphs using argweaver. In JY Dutheil, editor, *STATISTICAL POPULATION GENOMICS*, volume 2090 of *Methods in Molecular Biology*, pages 231–266. 2020.

[35] RR HUDSON. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23(2):183–201, 1983.

[36] Parul Johri, Charles F. Aquadro, Mark Beaumont, Brian Charlesworth, Laurent Excoffier, Adam Eyre-Walker, Peter D. Keightley, Michael Lynch, Gil McVean, Bret A. Payseur, Susanne P. Pfeifer, Wolfgang Stephan, and Jeffrey D. Jensen. Recommendations for improving statistical inference in population genomics. *PLOS Biology*, 20(5):e3001669, May 2022.

[37] Parul Johri, Brian Charlesworth, and Jeffrey D. Jensen. Toward an evolutionarily appropriate null model: Jointly inferring demography and purifying selection. *GENETICS*, 215(1):173–192, MAY 2020.

[38] Parul Johri, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, and Jeffrey D. Jensen. The impact of purifying and background selection on the inference of population history: Problems and prospects. *MOLECULAR BIOLOGY AND EVOLUTION*, 38(7):2986–3003, JUL 2021.

[39] Mamoru Kato, Daniel A. Vasco, Ryuichi Sugino, Daichi Narushima, and Alexander Krasnitz. Sweepstake evolution revealed by population-genetic analysis of copy-number alterations in single genomes of breast cancer. *Royal Society of Open Science*, 4(9), SEP 2017.

[40] Jerome Kelleher, Kevin R. Thornton, Jaime Ashander, and Peter L. Ralph. Efficient pedigree recording for fast population genetics simulation. 14(11):e1006581.

[41] Jerome Kelleher, Yan Wong, Anthony W. Wohns, Chaimaa Fadil, Patrick K. Albers, and Gil McVean. Inferring whole-genome histories in large population datasets (vol 51, pg 1330, 2019). *Nature Genetics*, 51(11):1660, NOV 2019.

[42] Younhun Kim, Frederic Koehler, Ankur Moitra, Elchanan Mossel, and Govind Ramnarayan. How Many Subpopulations Is Too Many? Exponential Lower Bounds for Inferring Population Histories. *Journal of Computational Biology*, 27(4):613–625, APR 1 2020.

[43] JFC Kingman. The Coalescent . *Stochastic Processes and their Applications*, 13, 1982.

[44] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. 2016.

[45] Jere Koskela. Multi-locus data distinguishes between population growth and multiple merger coalescents. *STATISTICAL APPLICATIONS IN GENETICS AND MOLECULAR BIOLOGY*, 17(3), JUN 2018.

[46] Jere Koskela and Maite Wilke Berenguer. Robust model selection between population growth and multiple merger coalescents. *Mathematical Biosciences*, 311:1–12, MAY 2019.

[47] John Boaz Lee, Ryan Rossi, and Xiangnan Kong. Graph Classification using Structural Attention. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1666–1674, London United Kingdom, July 2018. ACM.

[48] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–U84, JUL 28 2011.

[49] Ali Mahmoudi, Jere Koskela, Jerome Kelleher, Yao-ban Chan, and David Balding. Bayesian inference of ancestral recombination graphs. *PLOS COMPUTATIONAL BIOLOGY*, 18(3), MAR 2022.

[50] P Marjoram and JD Wall. Fast "coalescent" simulation. *BMC Genetics*, 7, MAR 15 2006.

[51] Sebastian Matuszewski, Marcel E. Hildebrandt, Guillaume Achaz, and Jeffrey D. Jensen. Coalescent processes with skewed offspring distributions and non-equilibrium demography. *Genetics*, 2017.

[52] GAT McVean and NJ Cardin. Approximating the coalescent with recombination. *Philosophical Transactions of the Royal Society B-Biological Sciences*, 360(1459):1387–1393, JUL 29 2005.

[53] F Menardo, S Gagneux, and F Freund. Multiple merger genealogies in outbreaks of Mycobacterium tuberculosis. *Molecular Biology and Evolution*, 07 2020. msaa179.

[54] Alistair Miles, pyup io bot, Murillo R, Peter Ralph, Nick Harding, Rahul Pisupati, Summer Rae, and Tim Millar. cggh/scikit-allel: v1.3.3.

[55] M Mohle and S Sagitov. A classification of coalescent processes for haploid exchangeable population models. *Annals of Probability*, 29(4):1547–1562, OCT 2001.

[56] Ana Y. Morales-Arce, Rebecca B. Harris, Anne C. Stone, and Jeffrey D. Jensen. Evaluating the contributions of purifying selection and progeny-skew in dictating within-host Mycobacterium tuberculosis evolution. *Evolution*, 74(5):992–1001, MAY 2020.

[57] Richard A. Neher and Oskar Hallatschek. Genealogies of rapidly adapting populations. *Proceedings of the National Academy of Sciences*, 110(2):437–442, January 2013.

[58] Dominic Nelson, Jerome Kelleher, Aaron P. Ragsdale, Claudia Moreau, Gil McVean, and Simon Gravel. Accounting for long-range correlations in genome-wide simulations of large cohorts. *PLOS Genetics*, 16(5), MAY 2020.

[59] Hiro-Sato Niwa, Kazuya Nashida, and Takashi Yanagimoto. Reproductive skew in japanese sardine inferred from dna sequences. *ICES Journal of Marine Science*, 73(9):2181–2189, 2016.

[60] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. October 2017.

[61] J Pitman. Coalescents with multiple collisions. *Annals of Probability*, 27(4):1870–1902, OCT 1999.

[62] Xinghu Qin, Charleston W. K. Chiang, and Oscar E. Gaggiotti. Deciphering signatures of natural selection via deep learning. *bioRxiv*, 2021.

[63] Matthew D. Rasmussen, Melissa J. Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLOS GENETICS*, 10(5), MAY 2014.

[64] Daniel P Rice, John Novembre, and Michael M Desai. Distinguishing multiple-merger from kingman coalescence using two-site frequency spectra. *bioRxiv*, 2018.

[65] Alan R. Rogers and Chad Huff. Linkage disequilibrium between loci with unknown phase. 182(3):839–844.

[66] Andrew M. Sackman, Rebecca B. Harris, and Jeffrey D. Jensen. Inferring demography and selection in organisms characterized by skewed offspring distributions. *GENETICS*, 211(3):1019–1028, MAR 2019.

[67] S Sagitov. The general coalescent with asynchronous mergers of ancestral lines. *Journal of Applied Probability*, 36(4):1116–1125, DEC 1999.

[68] S Sagitov. Convergence to the coalescent with simultaneous multiple mergers. *Journal of Applied Probability*, 40(4):839–854, DEC 2003.

[69] Théophile Sanchez, Jean Cury, Guillaume Charpiat, and Flora Jay. Deep learning for population size history inference: Design, comparison and combination with approximate bayesian computation. 21(8):2645–2660.

[70] Nicolae Sapoval, Amirali Aghazadeh, Michael G. Nute, Dinler A. Antunes, Advait Balaji, Richard Baraniuk, C. J. Barberan, Ruth Dannenfelser, Chen Dun, Mohammadamin Edrisi, R. A. Leo Elworth, Bryce Kille, Anastasios Kyrillidis, Luay Nakhleh, Cameron R. Wolfe, Zhi Yan, Vicky Yao, and Todd J. Treangen. Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728, December 2022.

[71] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46(8):919–925, AUG 2014.

[72] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. Modeling relational data with graph convolutional networks, 2017.

[73] J Schweinsberg. Coalescent processes obtained from supercritical Galton-Watson processes. *Stochastic Processes and their Applications*, 106(1):107–139, JUL 2003.

[74] Thibaut Paul Patrick Sellinger, Diala Abu Awad, Markus Moest, and Aurelien Tellier. Inference of past demography, dormancy and self-fertilization rates from whole genome sequence data. *PLOS Genetics*, 16(4), APR 2020.

[75] Thibaut Paul Patrick Sellinger, Diala Abu-Awad, and Aurelien Tellier. Limits and convergence properties of the sequentially markovian coalescent. *MOLECULAR ECOLOGY RESOURCES*, 21(7):2231–2248, OCT 2021.

[76] Sara Sheehan and Yun S. Song. Deep Learning for Population Genetic Inference. *PLOS Computational Biology*, 12(3), MAR 2016.

[77] Leo Speidel, Marie Forest, Sinan Shi, and Simon R. Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature Genetics*, 51(9):1321+, SEP 2019.

[78] Matthias Steinruecken, Matthias Birkner, and Jochen Blath. Analysis of DNA sequence variation within marine species using Beta-coalescents. *Theoretical Population Biology*, 87:15–24, AUG 2013.

[79] Wolfgang Stephan. Selective Sweeps. *Genetics*, 211(1):5–13, January 2019.

[80] Stefan Struett, Thibaut Sellinger, Sylvain Glémin, Aurélien Tellier, and Stefan Laurent. Inference of evolutionary transitions to self-fertilization using whole-genome sequences. *bioRxiv*, 2022.

[81] Aurelien Tellier and Christophe Lemaire. Coalescence 2.0: a multiple branching of recent theoretical developments and their applications. *Molecular Ecology*, 23(11):2637–2652, JUN 2014.

[82] Jonathan Terhorst, John A. Kamm, and Yun S. Song. Robust and scalable inference of population history froth hundreds of unphased whole genomes. *Nature Genetics*, 49(2):303–309, FEB 2017.

[83] Gautam Upadhya and Matthias Steinrücken. Robust Inference of Population Size Histories from Genomic Sequencing Data. preprint, Genetics, May 2021.

[84] Ke Wang, Iain Mathieson, Jared O'Connell, and Stephan Schiffels. Tracking human population structure through time from whole genome sequences. *PLOS Genetics*, 16(3), MAR 2020.

[85] Zhanpeng Wang, Jiaping Wang, Michael Kourakos, Nhung Hoang, Hyong Hark Lee, Iain Mathieson, and Sara Mathieson. Automatic inference of demographic parameters using generative adversarial networks. *Molecular Ecology Resources*, 21(8):2689–2705, 2021.

[86] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

[87] Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. *CoRR*, abs/1603.08861, 2016.

[88] Burak Yelmen, Aurélien Decelle, Linda Ongaro, Davide Marnetto, Corentin Tallec, Francesco Montinaro, Cyril Furtlehner, Luca Pagani, and Flora Jay. Creating artificial human genomes using generative neural networks. *PLOS Genetics*, 17(2):1–22, 02 2021.

[89] Rex Ying, Jiaxuan You, Christopher Morris, Xiang Ren, William L. Hamilton, and Jure Leskovec. Hierarchical graph representation learning with differentiable pooling.

[90] Muhan Zhang and Yixin Chen. Link prediction based on graph neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[91] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. 1:57–81.