Federal State Autonomous Educational Institution
for the Higher Education
National Research University "Higher School of Economics"
Faculty of Mathematics

Dudkovskaya Anastasiya Vadimovna

Coalescent Model and its Approximations for Population Genetics

Bachelor's thesis

Field of study: 01.03.01 "— Mathematics,
Degree programme: bachelor's educational programme "Mathematics"

Reviewer:
PhD, Associate Professor
Evgeniy Andreevich Burovskiy

Scientific supervisor:
PhD, Associate Professor
Vladimir Lvovich Shchur

Moscow 2022

Abstract

Coalescent theory is a probabilistic model for genealogical trees. It models genealogies as a Markovian process with backwards time (ancestral lineages go from descendants to their ancestors). Structured coalescence is a generalisation of a basic model to multiple populations with migration. Though efficient for simulations, the inference is computationally challenging under structured coalescent. MASCO is an approximation of coalescent which allows to calculate tree likelihoods efficiently for complicated population scenarios. In our work we develop an approach for computing Ancestral Recombination Graphs' likelihood using Importance Sampling and MASCO.

Contents

1 Introduction

Phylogenetics is a science of evolutionary history within or among species or other groups of organisms like animals, viruses, bacteria. It has been widely done to study migration patterns and events, spread of viruses or another pathogens, relatives between species.

Many models were developed to learn about populations structure, migrations, common ancestors.The basic Wright-Fisher model was formulated in 1930-1931 years [1], [2]. It describes an evolution of a genetic locus with two alleles A and a that have the same fitness in a diploid population of constant size N with nonoverlapping generations that undergoes random mating.

But there are a lot of species and organisms, which generations are overlapping, so there was developed new model - Moran model [11].

In these models we measure time in generations. In some situations it is more convenient to look at a process as a continuous-time process. This is how the Kingman theory of coalescence appears [5]. Also Kingman Coalescent model is the limit case of Moran and Wright-Fisher models - with large population size.

Also *recombination events* happen during evolutionary process, so different parts of genomes may have different phylogenetics trees. We can look at these ancestral recombination graphs as at different possible scenarios of coalescence process of some samples. In 1983 Hudson [13] developed a model that simulate coalescence with recombination. But it is computationally hard to use this model for real big data, so there were received approximation models - "The sequentially markov coalescent" [9] and "SMC'"[10].

To simulate Ancestral recombination graphs on the base of real DNA sequences it can be used `ARGweaver` [4] for simple situation with one population with constant size or `ARGweaver-D` [16] for demographic scenarios.
For example ARGs are using to detect the natural selection [17], [20] or to estimate coalescence time [18].

At present, there is a huge amount of genomic data from which one we can obtain information about ancestral populations and their structure. There are already exists databases with whole genome of some organism. For example for human being it is a 1000 genomes project[12].

However, interaction with a large amount of data make such developments impossible due to computing power. Therefore, new phylogenetic and phylogeographic methods are arised.

The methods that we will consider work mainly with specific populations. Some of them are based on the theory of structured coalescence [3]. Such methods assume that migration and coalescence tree formations are dependent. Some of them look at the complete migration history using Monte Carlo Markov chains [14] [19], however, complex scenarios are difficult to implement on the big data in this way. Others keep track of the

probabilities of finding each line in different states [15].

The article "The Structured Coalescent and Its Approximations" [3] shows one of the methods developed on the basis of structured coalescence theory.

The object of this research is to analyze how MASCO method of calculating the likelihood of structured coalescent trees works with ancestral recombination graphs (**ARGs**).

2 Preliminaries

2.1 Coalescent theory

Let us define a *structured coalescent tree.*

Definition 2.1. *A diploid organism is defined as an organism that has two copies of the same gene in the system. For example humans have 23 pairs of homologous chromosomes, with 2 copies of each gene in a pair of chromosomes.*

Definition 2.2. *A locus is a fixed position on a chromosome.*

We consider a diploid population of size N and look at one locus. We observe only n objects of this population, so we have n samples (alleles). We explore ancestors of these samples. Then we suppose that in the previous generation all samples have a parent, maybe some of them have a common ancestor. We ignore events when more than one pair of samples or three samples have a common ancestor.

So when we look at many previous generations there are two variant of development at every step: only one pair has a common parent or all samples have different parents.

Lemma 1. *Suppose we have a population of size $2N$ and k samples, $N \gg k$. Then the probability that a pair had a common parent in the previous generation is $\approx \frac{k(k-1)}{2}\frac{1}{2N}$.*

Two samples have a common ancestor in previous generation with probability $\frac{k(k-1)}{2}\frac{1}{2N}$ and other possible situations have probability with the order $1/N^2$ so we can ignore it.

Proposition 2.1. *The time measured in units of $2N$ generations during which k lines coexist is approximately exponentially distributed with the mean $\frac{2}{k(k-1)}$.*

Proof. Let us take a look at the probability that during n generations there are k coexisting lineages. The probability is approximately equal to

$$\approx \Big(1 - \frac{k(k-1)}{4N}\Big)^n \approx exp\Big(-\frac{k(k-1)}{2}\frac{n}{2N}\Big).$$

When we start measure time in $t = \frac{n}{2N}$ with enough big N then we can see that t has approximately exponential distribution with the mean $\frac{2}{k(k-1)}$.

We look backward in time to the point where one pair of two lineages coalesce and call this point *the most recent common ancestor* ***(MRCA)***.

Definition 2.3. *This process of lineages coalescent is called "Kingman Coalescent" [5].*

2.2 Structured coalescence

The structured coalescent differs from Kingman by adding *migration*. We suppose that every sample lives in some deme and can migrate between demes. Also coalescent event can be only between lineages in the same demes. So now in every previous generation we add a new type of event: one lineage can migrate from one deme to another. We can set the migration rates by ourselves.

Suppose we have n sample $\{1, 2, ..., n\}$ and m demes $\{1,..., m\}$. Then there are n^m configurations $(L_1 = l_1,\ldots,L_i = l_i,\ldots,L_n = l_n)$; $l_i \in \{1,\ldots,m\}$. This model can be perceived as a markovian chain, where configurations are the state space and probabilities of transitions depend on the migration and coalescent rates.

The migration event that moves the sample L_i from the deme a to b changes configuration as follows:

$$(L_1 = l_1,...,L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ...,L_n = l_n) \rightarrow$$

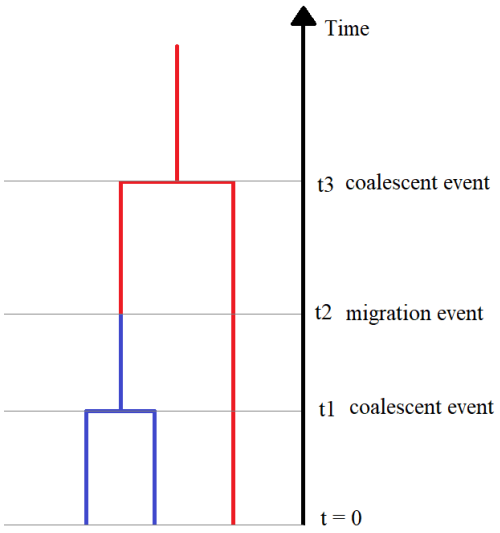
$$\rightarrow (L_1 = l_1,..., L_{i-1} = l_{i-1}, L_i = b, L_{i+1} = l_{i+1},..., L_n = l_n).$$

The coalescent event between L_i and L_j, where $i < j$ and they are in the same deme a changes configuration as follows:

$$(L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ..., L_{j-1} = l_{j-1}, L_j = a, L_{j+1} = l_{j+1}, ..., L_n = l_n) \rightarrow$$

$$\rightarrow (L_1 = l_1, ..., L_{i-1} = l_{i-1}, L_i = a, L_{i+1} = l_{i+1}, ..., L_{j-1} = l_{j-1}, L_j = l_{j+1}, ..., L_{n-1} = l_n).$$

Example 1. *Structured Coalescent Tree*



So here we can see three events:

1. a coalescent event between two first lineages at the time t_1;
2. a migration event of the first lineage from the blue state to the red state at the time t_2;
3. a coalescent event between the first two lineages at the time t_3.

There are more details of coalescent theory in the book [6]

2.3 Coalescence with recombination and Ancestral recombination graph

Recombination is a process that forms new combinations alleles or genes, new orders of genes as a result of inversions, duplications and deletions of genes. While germ cells grow up, alleles received from parents are shuffled and each gamete gets only half of parental alleles. Then alleles of each parents are randomly combined in a zygote. The combination of these two random processes - "shuffling" of alleles in generative cells and meeting gametes - provides phenotypic diversity, those differences between organisms that play a decisive role in their struggle for existence.

Definition 2.4. *Recombination is a process by which pieces of DNA are broken and recombined to produce new combinations of alleles. This recombination process creates genetic diversity at the level of genes that reflects differences in the DNA sequences of different organisms.*

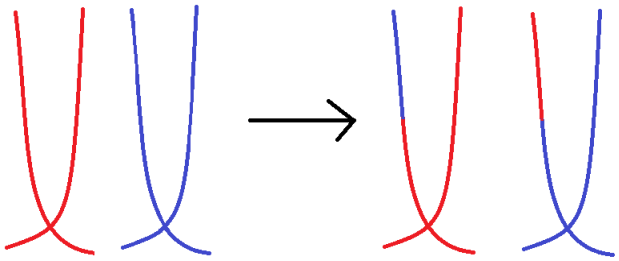


Figure 1: Example of recombination

It is quite hard to get analitical results from so sophisticated process, but we can simply model it. In 1983 Hudson[13] do this.

Let us consider a segment of DNA and rescale it to be the unit interval $[0, 1]$. We suppose that more than one recombination event could not occur. We look forward in time and "choose a parent" for our samples. We want to construct a chromosome or another segment. Suppose that with probability r a recombination occurs. So to generate a segment in the next generation we copy the hole content from one randomly chosen sample with probability $1 - r$. If the recombination occurs, then we pick a point uniformly along the segment and two individuals at random from the population. We copy the genetic material to the left of that point from the first individual and copy the material to the right from the second. The most important realisation of this model *Ancestral recombination graph.*

Definition 2.5. *An ancestral recombination graph (**ARG**) is the collection of coalescence and recombination events that describes the history of sampled chromosomes until the most recent common ancestor of each non-recombining block is reached.*

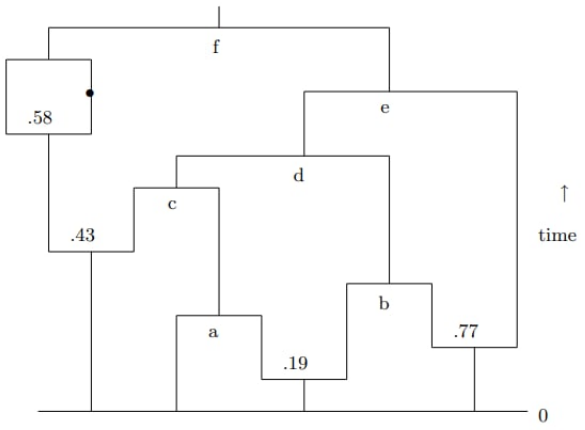


Figure 2: Example of ARG, from Durrett[6]

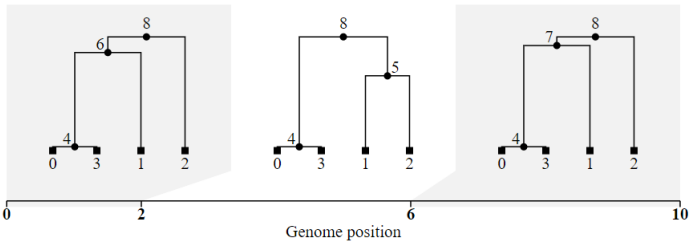


Figure 3: Example of ARG, simulated by software `msprime` [8]

2.4 MASCO

In the article [3] the authors presented the method, that calculates a likelihood of a structured tree and its approximation. They observe the coalescent and migration process like a Markovian chain and wrote Kolmogorov's equations [7]. They obtained the following equations for the Exact Structured Coalescent (ESCO) process.

$$\frac{d}{dt}P_t(L_1 = l_1,..., L_i = l_i,..., L_n = l_n, T) =$$
$$= \sum_{i=1}^{n}\sum_{a=1}^{m}(\mu_{al_i}P_t(L_1 = l_1,..., L_i = a,..., L_n = l_n, T) - \mu_{lia}P_t(L_1 = l_1,..., L_i = l_i,..., L_n = l_n, T)) -$$
$$- \sum_{a=1}^{m}\lambda_a C_{k_a}^2 P_t(L_1 = l_1,..., L_i = l_i,..., L_n = l_n, T); \quad (2.1)$$

$$P_t^{\text{after coal}}(L_1 = l_1, \ldots, L_i = a, \ldots, L_{n-1} = l_n, T) =$$
$$= \lambda_a P_t(L_1, \ldots, L_i = a, \ldots, L_j = a, \ldots, L_n = l_n, T) \quad (2.2)$$

These ones equations for the Marginal Structured Coalescent process (MASCO, ap-

proximation of Exact Structured Coalescent process)

$$\frac{d}{dt}P_t(L_i = l_i,T) = \sum_{a=1}^{m}(\mu_{al_i}P_t(L_i = a,T) - \mu_{l_ia}P_t(L_i = l_i,T))-$$
$$-P_t(L_i = l_i,T)(\lambda_{l_i}\sum_{k=1,k\neq i}^{n} P_t(L_k = l_i|T)+\sum_{a=1}^{m}(\frac{\lambda_a}{2}\sum_{j=1,j\neq i}^{n}\sum_{k=1,k\neq i,j}^{n} P_t(L_j = a|T)P_t(L_k = a|T)); \tag{2.3}$$

$$P_t^{\text{after coal}}(L_k = a, T) = P_t(L_k = a|T)P_t(L_s = a|T)\lambda_a P_t(T).$$

The second equations are received from the consideration

$$P_t(L_i = l_i, T) = \sum_{K/i} P_t(K, T),$$

where we take the sum over all configuration K such that the lineage i is in the fixed state l_i;

and the assumption, that lineages i, j, and k and their states l_i, l_j, and l_k are uncorrelated, that is:

$$P_t(L_k = l_k, L_i = l_i, L_j = l_j|T) = P_t(L_i = l_i|T)P_t(L_j = l_j|T)P_t(L_k = l_i|T).$$

Also we can write an approximation for ESCO equations using the probability of pairs of lineages.

$$\frac{d}{dt}P_t(L_i = l_i, L_j = l_j, T) =$$
$$= \sum_{a=1}^{m}[\mu_{al_i}P_t(L_i = a, L_j = l_j, T)+\mu_{al_j}P_t(L_i = l_i, L_j = a, T)-(\mu_{l_ia}+\mu_{l_ja})P_t(L_i = l_i, L_j = l_j, T)]-$$
$$-\frac{1}{2}P_t(L_i = l_i, L_j = l_j, T)[\sum_{a=1}^{m}\lambda_a\sum_{s=1,s\neq i,j}^{n}\sum_{k=1,k\neq s,i,j}^{n} P_t(L_s = a, L_k = a|T)+$$
$$+\sum_{k=1,k\neq i,j}^{n}(2\lambda_{l_i}P_t(L_k = l_i|T) + 2\lambda_{l_j}P_t(L_k = l_j|T)) + 2\lambda_i\cdot\delta_{l_il_j}]; \tag{2.4}$$

$$P_t^{\text{after}}(L_i = a, L_k = l_k, T) = \lambda_a P_t(L_i = a, L_j = a, T)P_t(L_k = l_k|T); \tag{2.5}$$

$$P_t^{\text{after}}(L_k = l_k, L_s = l_s, T) = P_t(L_k = l_k, L_s = l_s|T)\sum_{a=1}^{m}\lambda_a P_t(L_i = a, L_j = a, T) \tag{2.6}$$

Proposition 2.2. *Without migration solutions of approximation MASCO equations become solutions of ESCO equations.*

Proof. Let us look at the assumption:

$$P_t(L_k = l_k, L_i = l_i, L_j = l_j|T) = P_t(L_i = l_i|T)P_t(L_j = l_j|T)P_t(L_k = l_i|T).$$

If there is no migration then every of probabilities can be only equal to 1 or 0. The direct check gets us what we want.

2.5 Likelihood of tree

When we have solutions of ESCO equations we can find a likelihood of a tree: $P(T|S, M, \Lambda) = \sum_{a=1}^{m} P_{t_{mrca}}(L_1 = a,T)$, where t_{mrca} - is a time to MRCA of all n samples, and Λ and M - collections of coalescent and migration rates.

There are n^m 'interval contributions', so find the solution is impossible for large n, m. In the article Muller et al [3] develop a method, MASCO, that can find the likelihood of a tree approximately and this method has only $n \cdot m$ equations.

2.6 ARGweaver

We want to recalculate Likelihood of trees from ancestral recombination graph, so we need to sample ARGs. We do it using the software `ARGweaver`[4].

`ARGweaver` is a method that simulates ARG in the cases of one population of constant population size. It uses a Markov chain Monte Carlo (MCMC) algorithm to sample ARGs. Observing some data like DNA sequences and model parametrs `ARGweaver` samples ARGs with frequencies proportional to their probabilities. Distribution of ARGs depends on the proposed data. As in all MCMC algorithms, `ARGweaver` starts with some initial ARG and then, respectevly to the previous graph, receives a new one.

Often first ARGs have low probability, but from some moment distribution of obtained graphs is similar to real distribution, so the probability of these graphs will be much higher.

Also we should notice, that consecutive ARGs are not independent, so we choose only every 10th sample.

The main idea of `ARGweaver` is to remove a single haploid genome from the ARG, and then to "thread" this genome back through the ARG, by sampling both its coalescence points with the remaining sequences and the associated recombination points.

2.7 Importance Sampling

To recalculate trees' Likelihood we use Bayesian inference. Bayesian inference is a method of statistical inference in which Bayes' theorem is used to update the probability for a hypothesis as more data or another information becomes available. The data in our case is genealogies, sequences, parameters like migration rates, recombination rate and mutation rate.

Bayesian inference derives the posterior probability as result of two parts : a prior probability and a likelihood function. Bayesian inference computes the posterior probability according to Bayes' theorem:

$$P(\Theta|D) = \frac{P(D|\Theta)P(\Theta)}{P(D)};$$

where D is available data and Θ is parameters which we want to estimate;

$P(\Theta)$ is the probability of parameters, it is independent of observed data, it is called the *prior probability*;

$P(\Theta|D)$ is the probability of parameters Θ after D is observed, it is called the *posterior probability*;

$P(D|\Theta)$ is the probability of observing data D given parameters Θ, it is called the *likelihood*;

$P(D)$ is *model evidence.*

We can see that $P(D)$ is the same despite all parameters, so if we look at the posterior probability the value of $P(D)$ does not matter.

Now we want to maximize the posterior function.

Importance sampling is a method that is used to estimate some properties of one distribution while we get samples generated by another distribution.

Proposition 2.3. *Suppose we have independent genealogies $\{G_i\}$ sampled from a distribution Q. We want to estimate some parameters Θ of a distribution P_Θ. Then*

$$L_D(\Theta) = \int_{all\ genealogies\ G} Pr_\Theta(D|G) \cdot P_\Theta(G) dG \approx \frac{1}{k} \sum_{i=1}^{k} \frac{Pr_\Theta(D|G_i) \cdot P_\Theta(G_i)}{Q(G_i)},$$

where $Pr_\Theta(D|G)$ is the likelihood of data under parameters and $P_\Theta(G)$ is the density function for coalescent trees.

Proof.

$$L_D(\Theta) = \int_{\text{all genealogies G}} Pr_\Theta(D|G) \cdot P_\Theta(G) dG =$$

$$= \int_{\text{all genealogies G}} \frac{Pr_\Theta(D|G) \cdot P_\Theta(G)}{Q(G)} Q(G) dG = \mathbb{E}_Q \frac{Pr_\Theta(D|G) \cdot P_\Theta(G)}{Q(G)} \approx$$

$$\approx \frac{1}{k} \sum_{i=1}^{k} \frac{Pr_\Theta(D|G_i) \cdot P_\Theta(G_i)}{Q(G_i)},$$

where the genealogies $\{G_i\}$ are sampled from a distribution $Q(G)$.

The method depends on the choice of Q. It is the most accurate if $Q(G) \approx \text{const} \cdot Pr_\Theta(D|G) \cdot P_\Theta(G)$.

3 Main results

The main part of this work is to analyze how we can use MASCO in estimation of some parameters, like migration rates, population size using ARGs, additionally using Importance sampling.

The first part of work is to simulate such situation (and the sequences of length $3 \cdot 10^7$) that we have 2 populations which population sizes are 10000. We suppose the symmetric migration between populations with the migration rate 9.6e-1. We observe 4 samples from each population.

To do this we use a python software `msprime`. The code is on the `GitHub` repository.

The second part is to simulate a collection of ARGs of these 6 sequences. In this step we use the `ARGweaver` [4]. We get 1000 ARGs and get from them every 50th graph. In the ARG we do not take every tree, because of the correlation between neighbours trees, so choose every 100th tree.

The last part is recalculating the ARGs' Likelihood using Structured Coalescent theory and Importance sampling. To calculate ARG's Likelihood in MASCO model we use a code that solve the MASCO equations numerically using Runge-Kutta method. Then we use Importance Sampling to calculate the Likelihood of ARGs, where $Q(G_i)$ is the likelihood of tree in Kingman Coalescent model, $P_\Theta(G_i)$ is the likelihood in MASCO model and $Pr_\Theta(D|G_i)$ is the probability of observed sequences.

4 Supplementary materials

The code is available here: https://github.com/mintuit/Diplom

References

[1] Sewall Wright, EVOLUTION IN MENDELIAN POPULATIONS, Genetics, Volume 16, Issue 2, 1 March 1931, Pages 97–159, https://doi.org/10.1093/genetics/16.2.97

[2] Ronald A. Fisher Sir, R. A. Fisher, The Genetical Theory of Natural Selection, OUP Oxford, 21 окт. 1999 г.

[3] N. F. Müller, D. A. Rasmussen, T. Stadler, *The Structured Coalescent and Its Approximations*. Mol Biol Evol. 2017 Nov; 34(11): 2970–2981.

[4] Melissa Hubisz, Adam Siepel, *Inference of Ancestral Recombination Graphs Using ARGweaver*. In: Dutheil J.Y. (eds) Statistical Population Genomics. Methods in Molecular Biology, vol 2090. Humana, New York, NY.

[5] J.F.C.Kingman, *The coalescent*. Stochastic Processes and their Applications, Volume 13, Issue 3, 1982, Pages 235-248, ISSN 0304-4149,

[6] Rick Durrett, *Probability Models for DNA Sequence Evolution*. Second Edition, Copyright 2008.

[7] Wentzel E. S., *Operations Research: Problems, Principles, Methodology*. 2nd ed., Sr.— M.: Nauka. Ch. ed. Phys.-Math. lit., 1988

[8] J. Kelleher, A. M. Etheridge and G. McVean, *Efficient Coalescent Simulation and Genealogical Analysis for Large Sample Sizes*; PLOS Comput Biol 12(5): e1004842. doi: 10.1371/journal.pcbi.1004842

[9] McVean GA, Cardin NJ. Approximating the coalescent with recombination. Philos Trans R Soc Lond B Biol Sci. 2005;360(1459):1387-1393. doi:10.1098/rstb.2005.1673

[10] Marjoram, P., Wall, J.D. Fast "coalescent" simulation. BMC Genet 7, 16 (2006). https://doi.org/10.1186/1471-2156-7-16

[11] Moran, P. (1958). Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society, 54(1), 60-71. doi:10.1017/S0305004100033193

[12] Susan Fairley, Ernesto Lowy-Gallego, Emily Perry, Paul Flicek, The International Genome Sample Resource (IGSR) collection of open human genomic variation resources, Nucleic Acids Research, Volume 48, Issue D1, 08 January 2020, Pages D941–D947, https://doi.org/10.1093/nar/gkz836

[13] Hudson, R. (1987). Estimating the recombination parameter of a finite population model without selection. Genetical Research, 50(3), 245-250. doi:10.1017/S0016672300023776

[14] Lemey P, Rambaut A, Drummond AJ, Suchard M. a.. 2009. Bayesian phylogeography finds its roots. PLoS Comput Biol. 59:e1000520.

[15] Volz EM. 2012. Complex population dynamics and the coalescent under neutrality. Genetics 1901:187–201.

[16] ARGweaver-D citation: Melissa J. Hubisz, Amy L. Williams, Adam Siepel. Mapping gene flow between ancient hominins through demography-aware inference of the ancestral recombination graph. Biorxiv 2019

[17] Article Source: An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data Stern AJ, Wilton PR, Nielsen R (2019) An approximate full-likelihood method for inferring selection and allele frequency trajectories from DNA sequence data. PLOS Genetics 15(9): e1008384. https://doi.org/10.1371/journal.pgen.1008384

[18] Débora Y. C. Brandt, Xinzhu Wei, Yun Deng, Andrew H Vaughn, Rasmus Nielsen, Evaluation of methods for estimating coalescence times using ancestral recombination graphs, Genetics, Volume 221, Issue 1, May 2022, iyac044, https://doi.org/10.1093/genetics/iyac044

[19] Speidel, L., Forest, M., Shi, S. et al. A method for genome-wide genealogy estimation for thousands of samples. Nat Genet 51, 1321–1329 (2019). https://doi.org/10.1038/s41588-019-0484-x

[20] Aaron J. Stern, Leo Speidel, Noah A. Zaitlen, Rasmus Nielsen, Disentangling selection on genetically correlated polygenic traits via whole-genome genealogies, The American Journal of Human Genetics, Volume 108, Issue 2, 2021, Pages 219-239, ISSN 0002-9297, https://doi.org/10.1016/j.ajhg.2020.12.005.