GADMA: Genetic Algorithm for Automatic Inferring Joint Demographic History of Multiple Populations from Allele Frequency Spectrum

Ekaterina Noskova[1,2], Vladimir Ulyantsev[1], Stephen J. O'Brien[3, 4], and Pavel Dobrynin[3, 5]

[1]ITMO University, St. Petersburg, Russia

[2]JetBrains Research, St. Petersburg, Russia

[3]Theodosius Dobzhansky Center for Genome Bioinformatics, Saint Petersburg State University, St. Petersburg, Russia

[4]Oceanographic Center, Nova Southeastern University Ft Lauderdale, Ft Lauderdale, USA

[5]National Zoological Park, Smithsonian Conservation Biology Institute, Washington DC, USA

Abstract

During the last few years more and more genomic analysis of individuals of closely related species has appeared. The history of evolution and development of populations, so-called demographic history, is embedded into their genome and we can try to pull it out. Allele frequency spectrum or AFS, the distribution of allele frequency in populations, is a popular and convenient representation of the genetic information. The AFS is used to build evolution of multiple populations and there are some methods to do it, for example, well-known software $\partial a \partial i$ or recently appeared *moments*. Both of them provide an opportunity to simulate AFS from any specified by researcher demographic model and to fit the best-likelihood parameters using local optimizations. However there is no known global search algorithms for demographic model inferring with given AFS.

We introduce here a new method of automatic unsupervised inferring of the demographic model from the allele frequency spectrum. It is based on genetic algorithm — one of the most effective algorithms for global search. Our method was implemented in software GADMA (Genetic Algorithm for Demographic Analysis, https://github.com/ctlab/GADMA) and showed itself rather efficient: it automatically finds close or even better demographic model than the one that were previously received manually. Moreover it shows different local optima close to global one, so it is possible to find more biology corrected model during further research. Our approach confirm existing history of modern people and find some new demographic models for the history of butterfly *E. gillettii*.

1 Introduction

To understand the evolution of species and their populations, it is important to understand what events occurred in their past and when. The genetic structure and diversity of species are shaped by events, such as population divergence, and processes, such as migrations or changes in (effective) population size over the course of thousands or millions of years. Records of population history are imprinted in the genomes of individuals within species. And with the rise of next-generation sequencing (NGS) technologies and abundant genome data, it has become possible to explore complex and parameter-rich demographic models that include the estimation of mutation rate, changes in population size, nonrandom mating, admixture and selection.

There is no method that can guarantee to find demographic model that best fit observed data, all existing approaches try to find best model by likelihood. As demographic model can be arbitrarily complex, there are a lot of variants, so the problem is very complicated. There are two main methods for inferring demographic models from genomic data based either on the haplotype data or the allele frequency spectrum (AFS) also known as the site frequency spectrum [1].

The most popular methods for inferring population size history that are based on haplotype data use

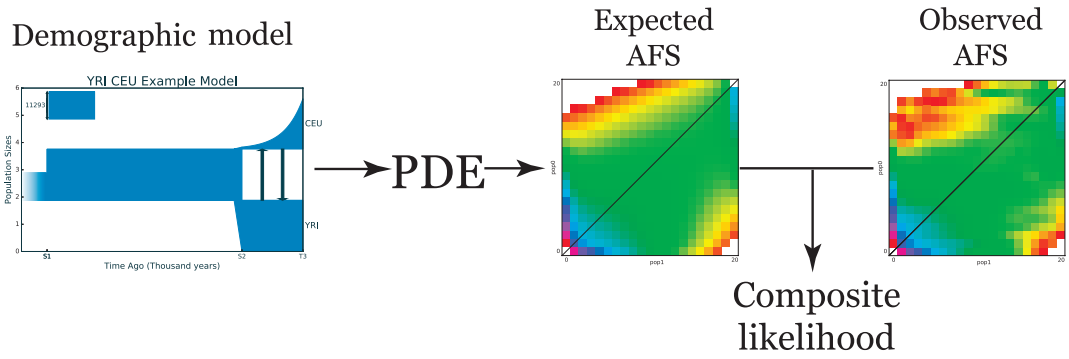


Figure 1: General scheme of ∂a∂i. It gets some demographic model as input, then partial differential equation is composed. Then it is numerically solved and expected AFS is extracted from solution of PDE. In order to get measure of similarity between demographic model and real observed AFS, ∂a∂i calculate composite likelihood between expected and observed AFS's.

Markov chain Monte Carlo (MCMC) estimation of the coalescent history of diploid genomes from sampled individuals. These methods use different models of coalescent history, and are implemented in several software programs such as PSMC (the pairwise sequentially Markovian coalescent, [2]) for single genome sequences, MSMC (the multiple sequentially Markovian coalescent, [3]) for multiple genome sequences and LAMARC (Likelihood Analysis with Metropolis Algorithm using Random Coalescence, [4]) for one or more populations. MSMC requires a large amount of RAM to run and produces cross-coalescence rates, that aren't mutation rates, which we are interested in. LAMARC is time consuming: to estimate recombination rate of 60 mtDNA of 16 kb each requires 3-4 weeks — which shows the inapplicability to large genomic data. All methods based on the haplotype data require additional data preprocessing in the form of phased data. Good phasing requires a well-sampled reference population that is close to the observed populations and is sometimes impossible. It may also be necessary to genotype families or a large number of individuals, which increases the cost of the entire project.

Since AFS requires nothing except genetic data of individuals from populations, it is more preferable than haplotype data. In essence, the AFS describes the distribution of derived allele frequencies of biallelic loci (SNVs) in a population or sample of populations based on the number of sequenced chromosomes [5]. An AFS can provide us information about how the populations developed based on observed genetic variation sampled from current individuals of those populations. Many studies have been devoted to testing and understanding the behavior of allele-frequency spectra under different demographic scenarios [6–10].

Two of the most popular methods of historical demographic inference based on AFS are the diffusion approximation (∂a∂i, [11]) and the faster continuous-time sequential Markovian coalescent approximation (fastsimcoal2, [12]). fastsimcoal2 can handle any number of populations, but it is computationally hard because it simulates multiple AFS simultaneously to estimate the most stable one. ∂a∂i simulates AFS using a numerical solution of the partial diffusion equation (PDE), which corresponds to the presented

demographic model and then provides the likelihood of the model (Figure 1). Unfortunately, PDE leads to some computational difficulties associated with analyses of complicated demographic models and large sample sizes. As a result, ∂a∂i can only handle up to three populations. More recently, a new method called *moments* has been introduced [13]. It uses ordinary differential equations to simulate AFS, which is faster and more stable than diffusion approximation in ∂a∂i, based on simulations comparing the two methods. *moments* presents a tradeoff between speed and accuracy in AFS building and can handle up to five populations. ∂a∂i and *moments* are based on two models of population genetics: Wright-Fisher generation model and infinitely-many-sites mutational model.

Whether using AFS or haplotype data, ideally, researchers seek to find the model of demographic history that best describes or "fits" their data. ∂a∂i and *moments* provide an opportunity to run multiple optimizations to help fit parameters of given demographic model, that maximizes value of composite likelihood. But optimizations based on gradient descent, for example, Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm [14–17] or its modifications, use numerical differentiation and are ineffective in practice, because of the complex structure of demographic models' search space. It also causes inefficiency of second type of methods, that offer existing solutions — local search algorithms without gradient, for example, Nelder–Mead method [18] or Powell's method [19]. As a result, all existing optimizations find local optima close to the initial values and require many runs to be performed using different initial model parameters, most of which are unknown or lack empirical data. Despite all these drawbacks, ∂a∂i and *moments* are efficient instruments, since they can simulate an allele frequency spectrum from any demographic model. In other words, the problem of finding a demographic model from the allele frequency spectrum is the inverse problem, which can be solved by solving the direct problem, that is simulation of the AFS from demographic model, with approximate numerical methods, such as diffusion or moments approximations. The lack of accurate and rapid differentiation and the complex structure of the search space lead to the usage of global optimization methods, such as the genetic algorithm.

Genetic algorithm (GA) [20] is one of the most efficient heuristic algorithm for global search. The main field of its usage is optimization of the fitness function, which is either not differentiable or it cannot be differentiated in a sufficiently effective way, for example, when the function is not representable in "closed-form expression". GA is based on principle of evolution and simulates natural selection using operations of "mutation" and "crossover", which leaves at the end the most adapted individual, the one which has the best value of fitness function. The wide variability of genetic algorithm led to its wide application including phylogenetic reconstuction [21], ancestral genome composition inference [22], and evolutionary biology in general [23].

In this paper we represent new method based on genetic algorithm to infer the best-fitting demographic model automatically from AFS data for two or three populations. Our method assumes the ability to

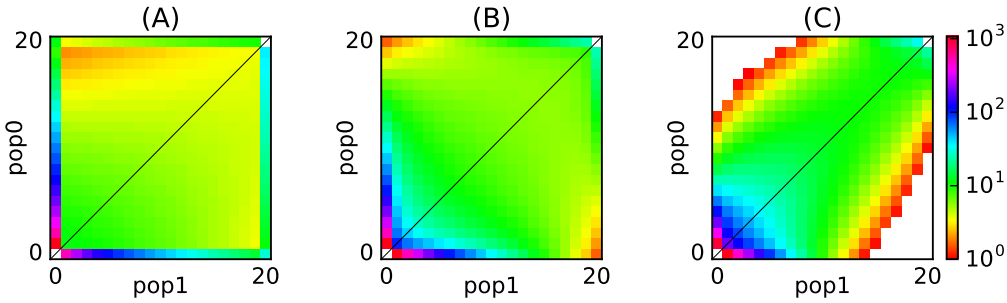


Figure 2: Examples of different AFS's. (A) Isolation of population after divergence; (B) Low symmetric migrations after populations divergence; (C) Asymmetrical stronger rates of migrations.

simulate the AFS from the demographic model, for example, using either ∂a∂i or *moments*. The GA framework overcomes limitations of local search optimizations and is more flexible in model's complexity, for example, it allows to increase number of parameters and estimate parameters such as functions of size changing of populations: sudden, linear or exponential changing. We implemented our method in the software GADMA (Genetic Algorithm for Demographic Model Analysis) written in Python and is available at Github (https://github.com/ctlab/GADMA).

2 Materials and Methods

This section provides definition of allele frequency spectrum, composite likelihood scheme, that is used in existing optimization and will be used in our method. After this introduction the problem of demographic model search from observed AFS is formulated in terms of computer science. Then there is introduced developed representation of demographic model for general approach, which is described in detail further, including genetic algorithm with its operations of "mutation" and "crossover".

2.1 Basic definitions and concepts

Assume there are P populations and for each population i there exists information about n_i chromosomes, then the AFS is the P-dimensional array A, where each entry $A[d_1, \ldots, d_P] \in \mathbb{N}$, $d_i \in [0, n_i]$, $\forall i \in [1, P]$ records the number of SNVs (relative to the common reference genome), which are exactly seen at d_1 chromosomes from population 1, d_2 chromosomes from population 2, ... and d_P chromosomes from population P. For example, if we have two populations, then the AFS is a two-dimensional matrix, where $A[i, j]$ holds the number of polymorphisms that occurred exactly in i individuals of the first population and in j individuals in the second population. Some examples of AFS's are presented on Figure 3.

Assume that we can simulate the AFS M from the proposed demographic model. Assuming no linkage between derived alleles, each element of the allele frequency spectrum $S[d_1, \ldots, d_P]$ is an independent

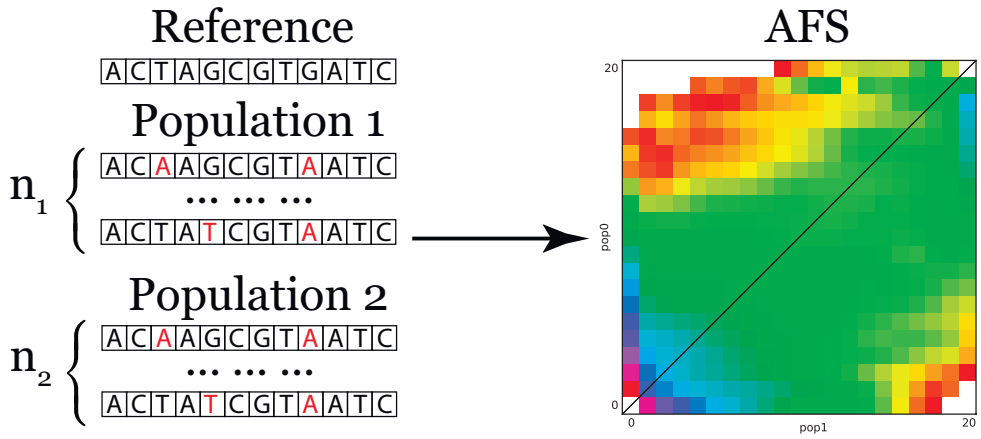


Figure 3: Process of AFS building. P populations are observed (2 on Figure), for each population i there is genetic information for n_i chromosomes. To build AFS we need to call SNV's (relative to some reference) and calculate its frequencies.

Poisson value with an mean equal to $M[d_1, \ldots, d_P]$. We then calculate the likelihood — the probability of obtaining the observed spectrum S, if the expectation spectrum is M, as the product of $(n_1+1)(n_2+1)\ldots(n_P+1)$ Poisson likelihood functions:

$$\mathcal{L}(M|S) = \prod_{d_1=0,..,n_1} \cdots \prod_{d_P=0,..,n_P} \frac{e^{-M[d_1,\ldots,d_P]} M[d_1,\ldots,d_P]^{S[d_1,\ldots,d_P]}}{S[d_1,\ldots,d_P]!} \tag{1}$$

In case of linked alleles, $\mathcal{L}(M|S)$ is the composite likelihood. Demographic models inferred by ∂a∂i and *moments* can be compared by computing the $\log(\mathcal{L}(M|S))$. Because $\mathcal{L}(M|S) \in [0,1]$, then $\log(\mathcal{L}(M|S)) \in [-\inf, 0]$ and the greater is the log-likelihood, the better the model fits the observed spectrum. In this paper, $\log(\mathcal{L}(M|S))$ was chosen as the fitness function of the genetic algorithm, discussed below.

2.2 Formulation of the problem

Consider a function $f(\Theta, A, C)$ that takes the parameters $\Theta = \{\theta^k\}_{k=1}^{N_\Theta}, \quad k \in \mathbb{R}$, the AFS $A \in \mathbb{R}^{P\times P}$, the set of constants $C = \{c^k\}_{k=1}^{N_C}$ and returns the measure of the correspondence between the parameters Θ and the AFS A.

The function $f(\Theta, A, C)$ builds a demographic model with respect to the parameters Θ that unambiguously determine it calculates the expected AFS M with respect to the constants C, and then determines the degree of similarity between M and the observed A by the composite likelihood. The constants can be various parameters of algorithms for calculating the expected spectrum, such as grid sizes for numerical solution of a differential equation in ∂a∂i, or population models parameters, including the average number of new mutated sites per individual in a generation θ_0 for the infinite-many-sites mutational model, or the time T_g for one generation in the Wright-Fisher model. The function f can have different implementation details. Here ∂a∂i and *moments* were selected for this purpose.

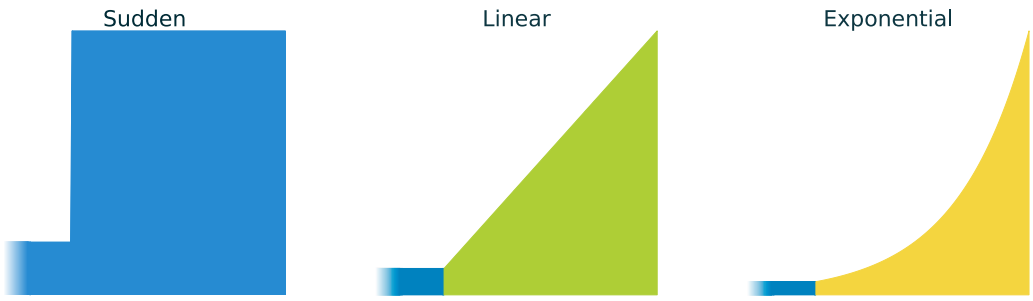


Figure 4: Three main demographic dynamics of population size change. Sudden type is usually most popular for using in demographic models, as well as the exponential type. Linear change was chosen to be some tradeoff between other two types, also it is more natural that sudden size change. Here colors are just a formality and are used for highlighting different types.

The purpose of this work is to develop an algorithm to search for the demographic model that best corresponds to the observed AFS. Formally, the problem can be formulated as follows:

Input

- $A \in \mathbb{R}^{P \times P}$ — the P-dimensional array, $P \in \{2, 3\}$.
- $C = \{c_k\}_{k=1}^{N_C}$ — the set of constants.

Output

- The set $\Theta \in \mathbb{R}^{N_\Theta}$ of values, that maximize the value of f:

$$\Theta : \quad f(\Theta, A, C) \to \max$$

There are approximate solutions of this problem with an additional input — a demographic model with a fixed number of parameters, using various local search algorithms, but in practice, as mentioned above, these algorithms have proved to be ineffective. In this paper, we present a new algorithm for the approximate solution of this more general problem automatically using one of the most effective methods of global optimization — GA.

2.3 Representation of the demographic model and its structure

Assume a division of one population into two new isolated subpopulations. Then the number of population splitting events directly depends on P, the number of considered populations.

We represent the demographic model as a sequence of "time intervals" and population splits, each of which has a fixed number of parameters. Assume fixed temporal order of the populations: from the

ancient to the most recently formed population. This temporal order is usually known or can be imputed. Then if the number of populations is ≤ 3, then each split will divide the last formed population. Thus, a split event has only one parameter — the fraction of the population, which separates to form a new population.

The next important component of the demographic model is the concept of "time interval". First, we define this as a segment of time, during which a certain dynamics of change of effective size is maintained for each population. We consider three main demographic dynamics: sudden change, linear change and exponential change of the effective population size (Figure 4). Sudden dynamic is very popular for its simplicity, exponential is often used too. We add linear type of size change, as it is tradeoff between sudden and exponential dynamic and is more natural than sudden type. Secondly, the parameters of migration rates between populations are constant during a given time interval. Thus, each time interval has the following parameters:

- time,
- effective population sizes at the end of the time interval,
- demographic dynamics of effective populations size change,
- migration between populations, if there is more than one population.

The sizes of the populations at the beginning of any time interval is equal to the number of populations at the end of the previous interval.

The first time interval is a special one: we consider that it lasts from the beginning of the existence of the species, and the demographic dynamic of change for the effective population size of the ancestral population is sudden [11]. Therefore, in this interval the only parameter is the size of the ancestral population.

Note that the number of split events is determined by the number of populations under consideration, but the number of intervals can be changed and thus change the number of parameters of the demographic model, its detail and complexity. Then we define the concept of the structure of the demographic model:

Definition 1. Structure of the demographic model — the number of time intervals as those that occur before and after a single split, in the case of two populations, in the case of three populations — the number of time intervals prior to the first split, those between the first and second split, and the ones after the second split.

For example, say we observe three populations. At the beginning there was an ancestral population (P_A) and this population started to grow in effective size. Then a split occurred that divided this ancestral population into two daughter populations (P_1 and P_2) that changed in effective size during one interval,

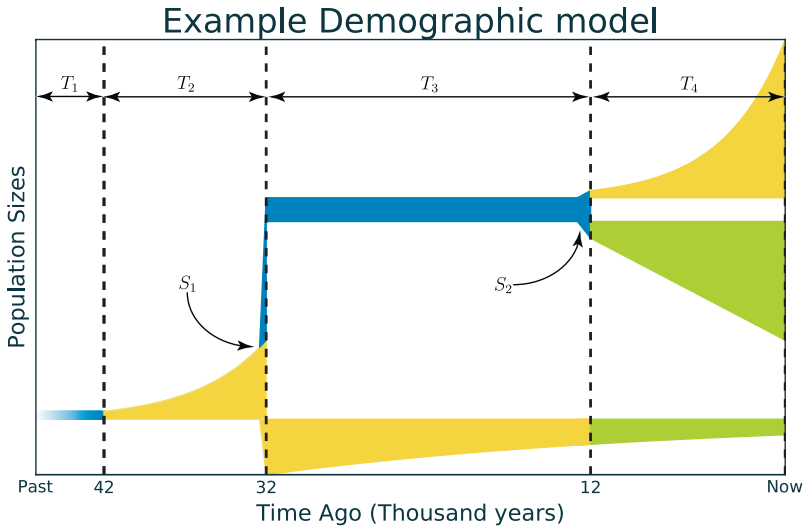


Figure 5: Example of representation of demographic model based on its plotting using *moments*. There is time on the axis of abscissa and population size on the axis of ordinates. Here there are 4 time intervals: T_1, T_2, T_3 and T_4, and two populations' splits: S_1 and S_2. The structure of that model is (2,1,1), because T_1 and T_2 are time intervals for one ancestral population, then split S_1 happens, T_3 is for two populations and T_4 is for three final populations after second split S_2. Here colors are just a formality and are used for highlighting different types.

followed by a split of the second population (into P_{2a} and P_{2b}), resulting in three descendant populations that changed within one subsequent time interval. The structure of such a model would be described as (2,1,1) (Fig. 5). The simplest model structures would be for two populations — (1,1), and for three populations — (1,1,1).

More formally, the structure of the model is a sequence of the form $S^* = \{s_i^*\}_{i=1}^{P}, \quad s_i^* \in \mathbb{N}$, where $P \in \{2, 3\}$ is the number of populations. In this case, the number of parameters $N_\Theta(S^*)$ of the demographic model with the structure S^* will be determined as follows:

$$N_\Theta(S^*) = (P-1) + \sum_{i=0}^{P} N_\theta^i(s_i^*), \quad \text{where} \quad N_\Theta^i(s_i^*) = \begin{cases} 1 + 3(s_1^* - 1), & \text{if } i = 1, \\ 7s_2^*, & \text{if } i = 2, \\ 13s_3^*, & \text{if } i = 3. \end{cases}$$

The term $(P-1)$ corresponds to the number of split parameters, and $\sum_{i=0}^{P} N_\theta^i(s_i^*)$ is the number of time intervals parameters. Thus, we can unambiguously interpret the demographic model according to the list of parameters and its structure by fixing for each time intervals a certain order of parameters.

2.4 General approach

In essence the general algorithm is a series of executions of the genetic algorithm (Figure 6). Suppose we have the initial and final structures of the demographic models, derived from considerations about the

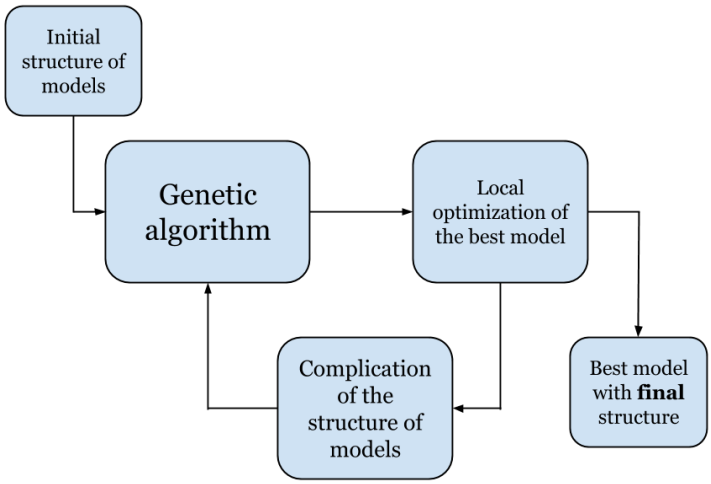


Figure 6: General algorithm

populations we are trying to model. During the genetic algorithm, the structure of the model does not change; that is, the model's parameters for the current structure are optimized. This restriction makes the procedure of crossover in the genetic algorithm simpler. As soon as the genetic algorithm stops, the parameters of the best-fitting model are adjusted in order to improve likelihood with local search algorithm. Then if the structure of the model is not quite complex, it is complicated and the genetic algorithm is run again for the new structure. Genetic algorithm and local search are executed until the best-fitting parameters of the model with final structure will be found.

Under the above scheme we begin with a simple demographic model at first and then increase its complexity (in terms of parameters) in order to estimate a more detailed model. If one try to find parameters' values for a complex demographic model at once, then the number of parameters can be too large to do it effectively.

Akaike information criterion. With the increase of the number of model parameters, we can overfit the model; a model with a large number of parameters will be better able to find parameter values corresponding to the observed data, than a model with a smaller number of parameters, but at the same time it will be less correspond to reality, for example, due to data errors. The Akaike information criterion (AIC) [24] is commonly used to compare models with different numbers of parameters:

$$AIC(M, S) = 2 \cdot k - 2 \cdot \log(\mathcal{L}(M|S)),$$

where k is number of parameters of the model and $\log(\mathcal{L}(M|S))$ — value of the log-likelihood function. The smaller AIC is, the better model fits observed data.

Obviously, it is enough to compare only the final models for each structure after local search, since

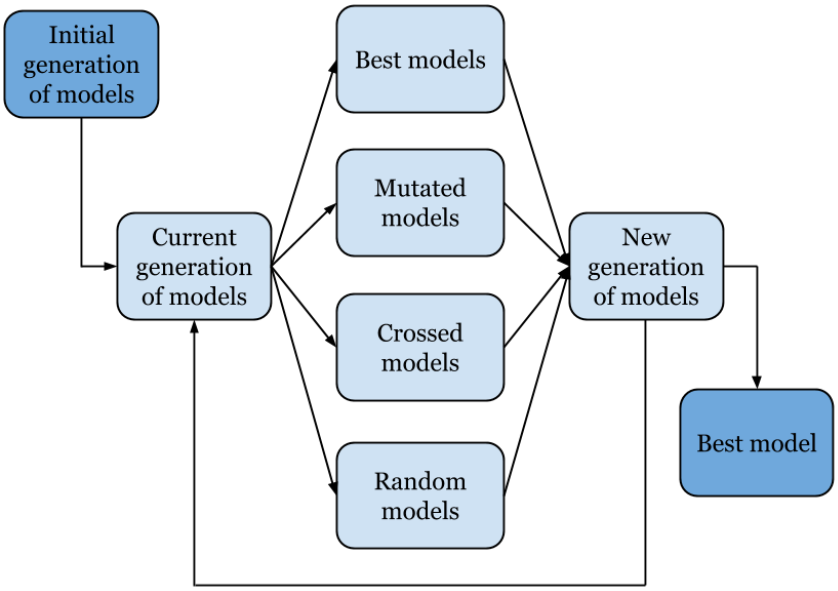


Figure 7: Scheme of the Genetic algorithm

the number of parameters between the increases in structure does not change, and therefore the value of AIC score depend only on the likelihood values. In the implementation of the method if the models with the best likelihood and best value of AIC score do not match, the user is informed about overfit.

2.5 Genetic algorithm

The genetic algorithm is one of the effective heuristic algorithms. It is based on the principles of evolution, which leads to a collision of notation with our problem. The aim of the algorithm is to find a solution to a problem that has the maximum or minimum value of the fitness function. At the beginning, a set of fixed size of random solutions, called **individuals**, is created. The set itself is called a **generation**. Each individual is given the value of its **fitness**, which is determined by the value of the fitness function. After this, new generations are iteratively built with the help of mutations, crossover and selection of the fittest individuals. All these operations can be either deterministic or random, and their order can vary in different implementations. In our case, individuals are demographic models of the same structure, and the fitness function is the log likelihood $\log(\mathcal{L}(M|S))$, which was described above.

In the first step generation of demographic models is initialized randomly, if it has not been already specified. To form a new generation of demographic models, we select the most adapted models from the previous generation, several mutated models, several crossed, and several random ones. The value of the fitness function is used to select the most adapted models. The choice of models for mutation or crossover is random, but the probability of choice is directly proportional to the value of fitness: the better fitness of the model is, the more likely it is to be selected. The genetic algorithm stops when it can no longer

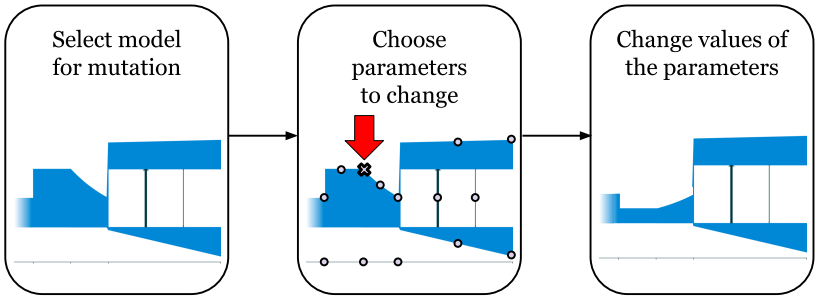


Figure 8: Mutation of demographic model.

obtain a better demographic model by the value of the fitness function for several iterations (Figure 7).

Mutation of demographic model. The mutation of the demographic model (Figure 8) is change of several parameters values. There are two constants: so-called, rate and strength of the mutation. The number of parameters to be changed is sampled from the binomial distribution with a mean equal to the mutation strength. Parameters that need to be changed are chosen with the probability that is directly proportional to their weights, which at the initial moment are equal (that is, the choice is equally probable), and then each weight can be increased when a mutation of the corresponding parameter occurred, which led to an improvement in the model. The measure of how much value of each parameter changes is determined by the sign ($+1$ or -1 equally probable) and the fraction of the change that is randomly sampled from the normal distribution with the mean equal to the mutation rate and a variance equal to the half of the mean.

Adaptive mutation rate and mutation strength. At the initial iterations of the genetic algorithm, the strong changes of a large number of parameters are much more effective than small changes of a small number, whereas when approaching the optimal solution the opposite is true. Therefore, the rate and strength of the mutation can be adaptive, that is, it changes during the operation of the algorithm. There are several ways to make an adaptive value, one of the most popular is the one-fifth algorithm [25]. First, apply it for the mutation rate: at each iteration, if we have a "successful" solution, that is, it became better after mutation, then multiply the mutation rate by the constant $C \in [1, 2]$, if not, then divide by a fourth-degree root of C, decreasing the rate. In the case of the mutation strength for the "success" of the solution, it is necessary to additionally check whether the decision has become the best for the whole time of the algorithm's work.

Thus, often getting a new best solution with a mutation that occurs at the beginning of the genetic algorithm, we will increase the number of parameters that change during the mutation operation, and the degree to which they change, and then, as we approach the optimum and decrease in frequency, they

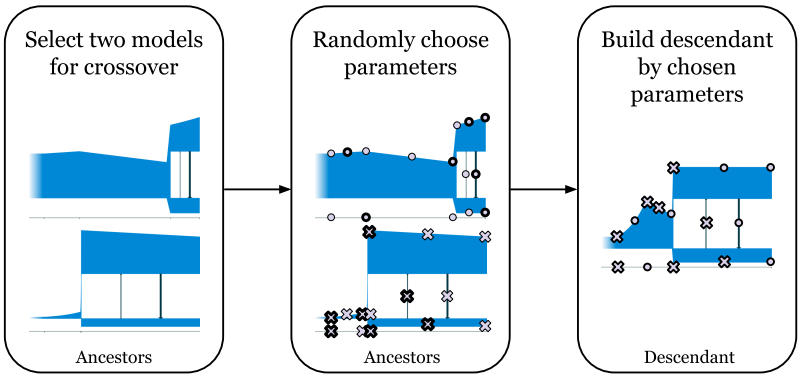


Figure 9: Crossover of two demographic models.

will also decrease and lead to a more accurate search. An increase in the number of parameters for the change leads to a more efficient crossover. At the same time, because of more weak conditions of the solution success for the adaptive mutation rate, it is changed more frequent than the mutation strength and make the mutation procedure more effective.

Crossover of two demographic models. In order to cross two demographic models, they are represented as sequences of parameters. Each parameter of their descendant is chosen randomly with equal probability from one or the second parent (Figure 9). Since the structure of models does not change during the operation of the genetic algorithm, the number of parameters for all models will be the same and they can be unambiguously interpreted by them and easily crossed by this way.

2.6 Local search algorithms.

Local optimizations are effective in cases when the initial solution is close to optimal. They are more accurate in adjusting parameters than the genetic algorithm, and can significantly improve its result.

∂a∂i and *moments* provide the following choice of algorithms for local optimization:

- Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.
- L-BFGS-B — BFGS modification, which is more efficient when the optimal parameters are close to the bounds.
- Nelder–Mead method or simplex-method.
- Powell's method.

The first two methods are gradient descent based and the last two are without using of the gradient. Powell's method was proposed by the authors of the *moments* and was noted as the most effective, so

it was adapted for usage with ∂a∂i and for our experimental studies it was chosen as a local search algorithm. In the implementation of our method, the user is provided the choice of a local optimization algorithm.

2.7 Increasing complexity of the models' structure

We need to be able to complicate the structure of the demographic models to achieve the final. To do this, a time interval is selected and it is splitted into two by division in the middle of its time. The time interval is chosen randomly on the basis that the new structure S^* should not become more than final S^F by one of its values: $s_k^* \leq s_K^F, \quad \forall k \in [1, P]$. The values of the parameters of the new intervals are calculated by the parent one: the sizes of the populations of the first more ancient time interval is equal to the sizes in the middle of the parent's time, and the parameters of the second interval are equal to the sizes at the end, the time of both intervals is equal to half of the parent's time, the migrations and the demographic dynamics of population size change remain the same. In essence the model does not change, more parameters appear, but at the same time the likelihood does not change either.

3 Results

First, we explored the efficient of our method on different data sets, that were examined before. We infer demographic models for two and three populations of modern human and for two populations of checkerspot butterfly *Euphydryas gillettii*. For human history different values of method's parameters were examined, including initial structure of models and usage either ∂a∂i or *moments*. GADMA provides demographic models close to existing ones that were received manually before. For the history of butterfly several demographic models were inferred. All of them has the likelihood close to the best one and can be valid histories.

All data, parameters for GADMA runs and results are available in the repository: https://bitbucket.org/noscode/gadma_results/src/master/.

3.1 Out of Africa model

One of the most popular demographic history is so-called "Out of Africa" demographic model of three populations [11, 13, 26, 27]:

- YRI — Yoruba individuals from Ibadan,
- CEU — Utah residents with ancestry from northern and western Europe,
- CHB — Han Chinese individuals from Beijing.

In order to show effectiveness of our method we choose allele frequency spectrum from the paper of Gutenkunst et al, 2009 [11], where ∂a∂i is introduced and demographic models for two (YRI, CEU) and three (YRI, CEU, CHB) populations are inferred from this spectrum. These models (Figure 10a and Figure 11a) were obtained as a result of a large number of local optimization launches and also have a number of restrictions on the number of parameters: the size of the YRI population does not change after the first growth of the size of the ancestral population, migrations are symmetrical and demographic dynamics of population size change are fixed by sudden, except for the last time interval for CEU, CHB, where exponential growth occurs. In total, the model for two populations has 6 parameters (Table 1), and for the three — 14 (Table 3).

The $21 \times 21 \times 21$ AFS was constructed in the paper of R. Gunenkunst et al. [11] on the basis of the Environmental Genome Project [28]. All biallelic SNVs from non-coding regions of 219 genes (total length 5.01 Mb) were considered and the effective length of the used sequence was equal to $L = 4.04 \times 10^6$.

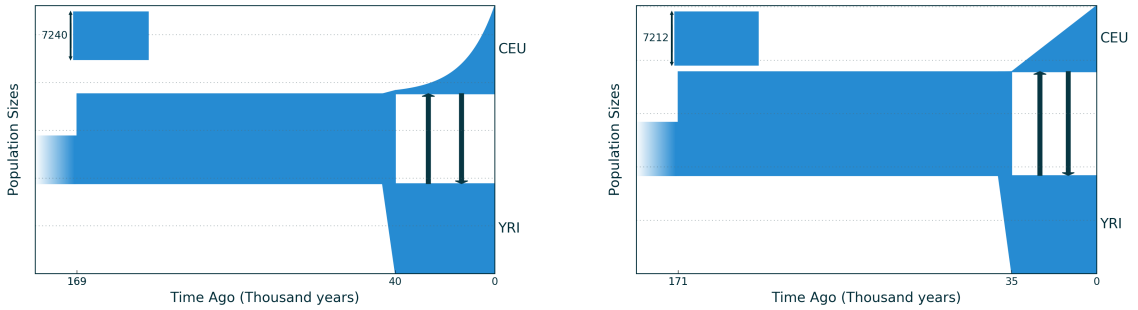
We take the same neutral mutation rate equal to $\mu = 2.35 \times 10^{-8}$ and the same time of one generation equal to $t_g = 25$ years as in the paper. Thus, the average frequency of the mutation occurrence in one individual per generation is equal to $\theta_0 = 4\mu L = 0.37976$.

The size of the generation of demographic models in genetic algorithms was chosen equal to 10, the strength and mutation rate equal to 0.2, and the proportions of the best, mutated, crossed and random models in the new generation as $0.2 : 0.3 : 0.3 : 0.2$. The strength and rate of mutation were adaptive with constants 1.05 and 1.02, respectively. The AFS was simulated using ∂a∂i with a $G = \{40, 50, 60\}$ grid size, the value of the likelihood was considered significant at the second decimal point and the genetic algorithm stopped after 100 iterations without improvement. As for local search, the Powell method was chosen. The final structures of the models corresponded to the model structures from the article and were equal to the initial ones, except for those cases where it is explicitly mentioned.

3.1.1 YRI, CEU case

The example of ∂a∂i provides an AFS, demographic model and the maximum likelihood parameters obtained for it for two populations: YRI and CEU. Three demographic models were inferred from the same AFS: with the same parameters as in the article (model 1) and two with all possible twelve parameters (models 2, 3) with different initial structures. Model 2 had a structure (1,1), which then expanded to (2,1) during the algorithm, and model 3 has initial structure equal to the final one — (2,1). Our method was launched 10 times for each of the three models (Table 1, Figure 10). All three models have likelihood better than the existing ones, the parameters of the first and second are insignificantly different, and the third has the best likelihood value. Model 3 shows a lower population size of Europeans, a higher rate of their growth (from 25 individuals to 9 thousand) and a shorter separation time than the model from

the article and models 1, 2. Migrations were chosen asymmetric, but they are practically equal to each other and coincide with values between models.



(a) Model from the example of ∂a∂i.

(b) Best likelihood model (model 3) from GADMA.

Figure 10: (a) Demographic model from ∂a∂i example. It has 6 parameters. Size of African population after its growth 169 kya is constant and migration rates between African and European populations are symmetric. (b) Demographic model from the same allele frequency spectrum using GADMA with 12 parameters.

	Model from ∂a∂i example	Model from GADMA (1)	Model from GADMA (2)	Model from GADMA (3)
Number of parameters:	6	6	12	12
Max log likelihood:	−1066.35	−1066.28	−1065.87	**−1065.15**
AIC:	2144.70	**2144.56**	2155.74	2154.30
Parameters:				
N_A	7240	7230	7200	7210
N_{AF0}	13620	13580	13400	14000
N_{EU0}	515	530	560	25
N_{EU}	13360	12400	12100	8950[2]
N_{AF}	$(=N_{AF0})$	$(=N_{AF})$	13500	13300
$m_{AF-EU}(\times 10^{-5})$	25	25	28.4	23.2
$m_{EU-AF}(\times 10^{-5})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	24.4	24.2
T_{AF} (kya)	168.5	171.5	176.5	171.1
T_{AF-EU} (kya)	40	40.8	42.3	34.8

Table 1: Maximum likelihood parameters for different demographic models of two populations. The migration rates are per generation.
[1] Linear growth

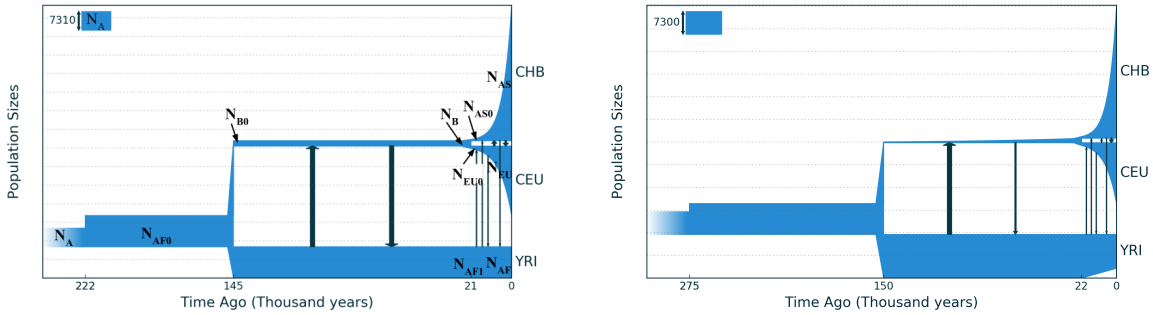
	Initial structure	Mean time of one iteration (sec)	Mean number of iterations	Mean $\log LL$	Std $\log LL$
Model 2	(1,1)	4.06	2938	**−1066.39**	**0.22**
Model 3	(2,1)	**3.77**	**1400**	−1071.16	14.35

Table 2: Comparison of runs with different initial structures: simple (10 starts) and complex (10 starts) in case of searching for demographic models of two populations YRI and CEU.
$logLL$ — log likelihood.

To demonstrate the inefficiency of the methods of local optimization for the model from the article [11] one of the proposed by ∂a∂i methods — BFGS, was launched 100 times. In each run, the initial parameters values were chosen randomly. The best value of the log-likelihood was −1629.24, which is quite far from the optimal −1066.35 offered by the article. The average time of one optimization run was about 21 minutes.

To compare the runs with different initial structures (models 2 and 3), some characteristics such as time for one iteration of genetic algorithm, number of iterations or mean and standard deviation of the log-likelihood value were calculated (Table 2). Launches with a simple initial structure show a more stable result in terms of the likelihood value, however they have a longer average time for one iteration. Also, all the models obtained for simple-structure launches have the same demographic dynamics of population size change and similar parameters as the existing model from the article, which is incorrect in case of complex-structure launches. At the same time, although the launches with a complex structure gave a model with better likelihood, it differs from the rest models in terms of parameters values, which may indicate its incorrectness.

3.2 YRI, CEU, CHB case



(a) Demographic model from Gutenkunst et al. [11] (14 parameters).

(b) Demographic model 3 from GADMA (23 parameters).

Figure 11: Demographic models for YRI, CEU, CHB populations.

Demographic models for three populations were also found from the allele frequency spectrum obtained in the work [11]. The first model (model 1) had the same parameters as in the article, and their values were inferred (Table 3). They provide a better likelihood value than the existing parameters. However, the time of people leaving the Africa, in other words, the division of the African population (YRI) and the Eurasian (the general population of CEU and CHB), was postponed up to 400 thousand years ago, which is not confirmed by any other studies. Therefore, expert data was used, which limited this time 150 thousand years ago. A demographic model (model 2) was inferred with this restriction and it also has a better likelihood value than in the article. Next, the model with all available 26 parameters was

	Model from Gutenkunst et al [11]	Model from GADMA (1)	Model from GADMA (2)	Model from GADMA (3)
Number of parameters:	13	13	14	26
Max log likelihood:	−6316.89	−6314.40	−6315.86	**−6288.36**
AIC:	12659.78	12654.80	12657.72	**12628.72**
Parameters:				
N_A	7300	6000	7300	7300
N_{AF0}	12300	11840	12200	9900
N_{B0}	2100	2050	2070	280
N_B	$(=N_{B0})$	$(=N_{B0})$	$(=N_{B0})$	1450[3]
N_{AF1}	$(=N_{AF0})$	$(=N_{AF0})$	$(=N_{AF0})$	14000
N_{AF}	$(=N_{AF0})$	$(=N_{AF0})$	$(=N_{AF0})$	11000[2]
N_{EU0}	1000	930	950	890
N_{EU}	27300	24700	23700	19600
N_{AS0}	510	500	510	560
N_{AS}	53200	50000	46200	42200
m_{AF-B} $(\times 10^{-5})$	25	26.2	25.4	68
m_{B-AF} $(\times 10^{-5})$	$(=m_{AF-B})$	$(=m_{AF-B})$	$(=m_{AF-B})$	230
m_{AF-EU} $(\times 10^{-5})$	3.0	3.1	3.1	17.8
m_{EU-AF} $(\times 10^{-5})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	$(=m_{AF-EU})$	11.3
m_{AF-AS} $(\times 10^{-5})$	1.9	1.9	2.0	3.7
m_{AS-AF} $(\times 10^{-5})$	$(=m_{AF-AS})$	$(=m_{AF-AS})$	$(=m_{AF-AS})$	8.3
m_{EU-AS} $(\times 10^{-5})$	9.6	10.2	10.2	75.4
m_{AS-EU} $(\times 10^{-5})$	$(=m_{AS-EU})$	$(=m_{AS-EU})$	$(=m_{AS-EU})$	27.4
T_{AF} (kya)	220	570	232	274.8
T_B (kya)	140	400	150	149.8
T_{EU-AS} (kya)	21.2	21.0	21.1	22.4

Table 3: Maximum likelihood parameters for YRI CEU demographic models. Migration rates are per generation.
[1]Linear growth
[2]Exponential growth

found. The trend of early exit from Africa was also preserved (the model is not shown), so the expert data was applied in this case too and the model (model 3) was obtained not only with the best likelihood, but also with the best AIC value.

As in the case of two populations, the local optimization of ∂a∂i— BFGS was tested. It was launched 100 times from randomly selected initial parameters and the best value of the log-likelihood was −6323.99, which is still far from the optimal −6326.89, although it is closer than it turned out in two populations. This example also shows the complexity of finding parameters using existing optimization algorithms.

The existing model and models 1, 2 have close likelihood values and parameters, exclude the time of separation in model 1. In model 3, which has 26 parameters and the best AIC value, some parameters are also quite similar to the values in others models. Exceptions are migration rates and the number of the Eurasian population, which exponentially grows from 200 individuals to 1500 after the split of the ancestral population. For comparison, in other models this number is a constant equal to 2,000 individuals, which seems less natural than exponential growth. Migration rates vary considerably: they are

	Mean time for one iteration (sec)	Mean number of iterations	Mean $\log LL$	Std $\log LL$
∂a∂i	136.81	5531	$\mathbf{-6293.56}$	**7.26**
moments	**18.44**	**4700**	-6305.21	12.72

Table 4: Comparison of GADMA using ∂a∂i and *moments* for inferring demographic model for YRI, CEU, CHB populations with 26 parameters. Likelihood value was calculated using ∂a∂i with $G = \{40, 50, 60\}$ grid size.
$\log LL$ — log likelihood.

higher. The model shows that the largest migrations occurred between the African and Eurasian populations, and after the division between the European and Asian populations. Moreover, the geographically farther populations are, the smaller migration rate is. This fact was expressed in other models, but not so brightly.

The algorithm was launched 10 times for each of the three models and the best solutions were observed (Table 3). There were launches (also 10 times) using *moments* to compare its effectiveness with ∂a∂i. Jourganous et al. — the authors of *moments* in their article made a similar comparison on simulated data, which show a little ambiguous results. Various characteristics of time and stability for the likelihood value based on the results of 20 GADMA runs (10 using ∂a∂i, 10 using *moments*) are presented in the table 4. To correctly comparison of the likelihood values they were recalculated for models, that were found with *moments*, using ∂a∂i with the $G = \{40, 50, 60\}$ grid size. *moments* showed its speed: faster ∂a∂i 7.5 times, however ∂a∂i is more accurate: the average and variance of the likelihood of the received models are less than for *moments*.

3.3 Butterflies *Euphydryas gillettii*

We also tested GADMA using data from McCoy et al., 2013 [29], which examined the demographic history of Gillette's checkerspot butterfly *Euphydryas gillettii*. We take the same spectra: from eight individuals from Colorado (CO) and eight individuals from native population in Wyoming (WY). Two AFS's of size 13×13 were observed: from synonymous SNVs only and from all SNVs.

McCoy et al. inferred three types of demographic models: 1) type A — models without migrations between populations, 2) type B1 — models with unidirectional migration from population CO to WY, and 3) type B2 — with bidirectional migration between CO and WY. In the paper [29] the demographic models of all types were inferred from AFS using synonymous SNVs only. Resulting model of type A had the best AIC value, so demographic model of type A was inferred from AFS using all SNV's. We used type A and type B2 to infer the demographic models from both AFS with GADMA.

Without considering migration, the models from article had the following structure: there was one population of N_A size, which at some point of time was divided into two, the size of which did not change

	A [29]	B1 [29]	B2 [29]	A	B2 (1)	B2 (2)	B2 (3)
Number of parameters:	3	4	5	5	7	7	7
Max log likelihood	−211.63	−210.83	−210.80	−211.42	**−205.94**	−210.59	−210.80
AIC	429.26	429.66	431.60	432.84	**425.88**	435.18	435.60
Parameters:							
η_{WY0}	NA	NA	NA	0.79	0.0015	0.999	$(=\eta_{WY})$
η_{WY}	0.922	0.884	0.893	0.960	0.856	0.742[1]	0.873
η_{CO}	0.104	0.119	0.121	0.064[1]	0.048[1]	0.120	0.121
m_{WY-CO}	NA	0.887	0.906	NA	2.857	0.952	0.923
m_{CO-WY}	NA	NA	0.002	NA	0.176	0.0	0.0
τ_{SPLIT}	0.066	0.080	0.081	0.075	0.241	0.079	0.080

Table 5: Demographic models for allele frequency spectrum using synonymous SNVs only.
[1] Linear change of population size

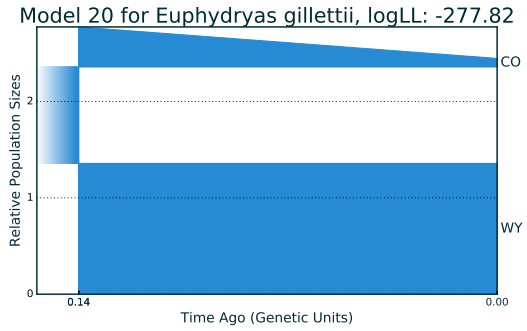
	A [29]	A	B2 (1)	B2 (2)	B2 (3)
Log likelihood	−283.53	−278.67	**−267.49**	−278.57	−282.74
Parameters:					
η_{WY0}	NA	0.584	0.028	0.566	$(=\eta_{WY})$
η_{WY}	1.320	1.358	1.225	1.773[1]	1.261
η_{CO}	0.173	0.089[1]	0.074[1]	0.105[1]	0.183
m_{WY-CO}	NA	NA	1.244	0.801	0.316
m_{CO-WY}	NA	NA	0.171	0.056	0.091
τ_{SPLIT}	0.117	0.138	0.259	0.162	0.129
Time of split in generations:					
t_{SPLIT}	45	37	**33**	40	47

Table 6: Demographic models for allele frequency spectrum using all SNVs.
[1] Linear change of population size

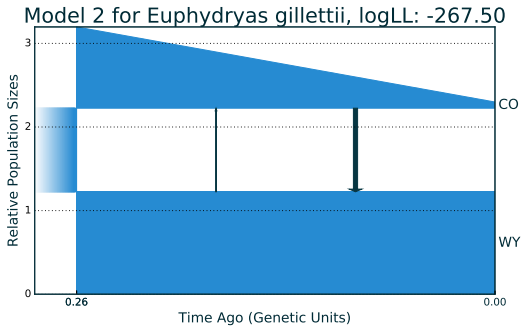
in the future (sudden change of population size). All parameters were calculated with respect to N_A and had the following notation: η_{WY}, η_{CO} — relative population sizes at nowadays; τ_{SPLIT} — time of split event; M_{WY-CO}, M_{CO-WY} — scaled migration rates. For our models, one more parameter is given in the table: $\eta_{WY0} \in [0, 1]$ — the size of the WY population right after split event or the fraction of size of ancestral population that forms population WY. The size of CO population right after split event is equal to $1-\eta_{WY0}$, because $N_A = 1$ before division. However, in the case when the demographic dynamic of population size change is sudden, the size after the separation is equal to the size at the present time.

We ran four executions of GADMA with different data inputs: 1) AFS using synonymous SNVs only without migration; 2) AFS using synonymous SNVs only with migration; 3) AFS using all SNVs without migration; and 4) AFS using all SNVs only with migration. For each execution, the analysis was repeated 50 times. *moments* was used to simulate the allele-frequency spectrum, as it showed its speed on the previous data, however, ∂a∂i with $G = \{32, 42, 52\}$ grid size was also used to calculate the likelihood value for a correct comparison with existing models.

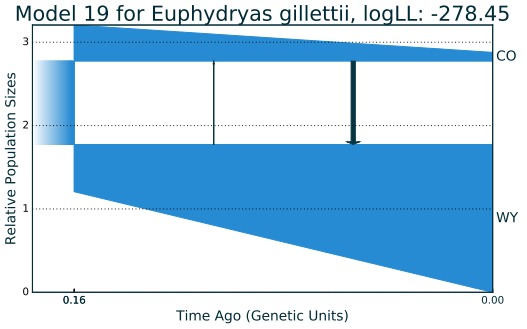
The size of the generation of demographic models in the genetic algorithm was chosen equal to 10, the strength and mutation rate equal to 0.2 with constants 1.0 and 1.02 respectively, and the proportions



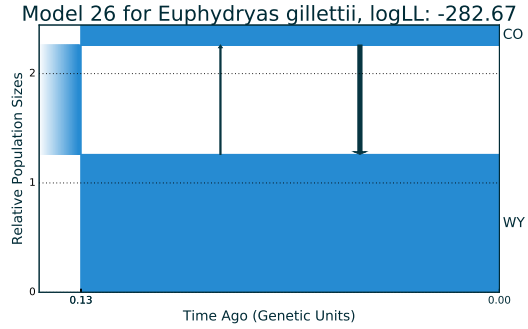
(a) Model of type A: without migrations.



(b) Model of type B2: with migrations.



(c) Model of type B2: with migrations.



(d) Model of type B2: with migrations.

Figure 12: Alternative demographic models for butterflies *Euphydryas gillettii* from Colorado (CO) and Wyoming (WY), inferred with GADMA. Models of type B2 have negligible migration rates.

of the best, mutated, crossed and random models in the new generation equal to 0.2 : 0.3 : 0.3 : 0.2. The likelihood was considered significant up to the second decimal point and the genetic algorithm stopped after 100 iterations without improvement. As for the local search, the Powell method was chosen. Since the structure of the model from the article corresponds to the simplest structure (1,1), it was chosen as the initial and final structures.

Several runs showed different local minima and the resulting alternative models (Figure 12) and their best-likelihood parameters (Table 5 and Table 6) were observed. We note that there is model (Figure 12d) selected by GADMA, that were inferred in the paper [29] with sudden growth of population size changes. For the AFS from synonymous SNVs only models of type B2 had better likelihood value than model of type A, in contrast with the resulting models from the article.

One of the results of the paper [29] was the conclusion that the model with migrations is inapplicable to the history without migrations. It was based on two factors: 1) model of type A, inferred from AFS using synonymous SNV's had best *AIC* score; 2) estimation of time of split event in models with migrations had wide 95% interval that included 0. However, during reproducing the result of the paper an error in a table with parameters was found and this conclusion was found to be incorrect. Moreover, it turned out on the contrary that the 95% interval for split time is rather good and the migration rates are so small that they can be considered as zero. So the model with migrations correctly reflects the

history without migrations. The models derived by our method also have negligible migration values and models of type B2 had better AIC value than models of type A, that confirm this fact.

The average population of butterflies from Colorado (CO) was estimated as $N_{CO} = 34$ individuals by McCoy et al. If we scale the parameters of the best resulting model so that the average size of CO is equal to this value ($x \cdot (\eta_{CO} + (1 - \eta_{WY0}))/2 = N_{CO}$), then we get 33.6 generations ($t_{SPLIT} = 2 \cdot x \cdot \tau_{SPLIT}$) after the division of population, which corresponds to the real 33 generations (1977-2010). Such estimation was made for all resulting models (Table 6), however, the best model had the best value of split time.

4 Discussion

We developed first method for automatic inferring the demographic history of several populations from the allele frequency spectrum. A method is based on a genetic algorithm and uses existing solutions either ∂a∂i or *moments* to simulate the AFS from the proposed demographic model. The existing optimizations, implemented in ∂a∂i and *moments*, turned out to be inefficient on practice. Our method was implemented in the GADMA software, which is openly accessible via the link https://github.com/ctlab/GADMA, supports up to three populations, and we verified its effectiveness on real data.

Our method proved to be efficient: it was applied to several datasets and result demographic models had better likelihood value than existing ones, that were recieved by the existing optimizations that ∂a∂i and *moments* offer. Moreover, they are consistent with the known history and other studies. It was also demonstrated the stability of the search, starting with a simpler structure than with more complex one, which reflects the profitability of using the scheme with structure increase. Additionally, we compare pipelines using *moments* or ∂a∂i and show the work time superiority of *moments*. Thus, GADMA is the first effective software that infers a demographic model from an allele frequency spectrum with nothing required from the user, except simple information, that determines how much the detailed model is required — the structure of the model.

Unfortunately, there are some limitations on how informative the AFS can be with regards to historical demographic inference. For example, it has been shown that the AFS of a single panmictic population can be matched to different demographic scenarios [10]. We should expect the same behavior in the case of many populations, which requires estimating the joint AFS. This problem can be solved by using some extra information about observed populations, for example, two-locus statistics [30]. Incorporating genetic linkage information with AFS data could improve accuracy of analysis.

In the course of experimental studies of human populations YRI, CEU and CHB, we inferred best model from the allele frequency spectrum from Gutenkunst et al. and it shows that expansion out of Africa happened 400 thousand of years ago, which does not supported by other studies. This can be caused

by noise in the spectrum, however, Gutenkunst et al. note high quality of this data set. Demographic model with the same parameters as in the paper [11] under the assumption that expansion took place not earlier than 150 thousand of years ago has close values of parameters to the parameters from article. We also inferred all possible parameters, including asymmetric migration rates and different demographic dynamics of population size changes, and obtain demographic model with best AIC score. We observe new higher asymmetric migration rates and the growth of the Eurasian population after its split from African.

Our approach finds demographic model with best-likelihood parameters, however, it is important to find best model with small number of parameters. Something similar is done in GADMA: it inform user about overfit when model with best likelihood and best AIC score do not match. But it does not sort out all possible number of parameters, so it is not guaranteed to find model with best AIC score.

Another direction of the further development of our work is the increasing the number of considered populations. Our method handle up to three populations, as ∂a∂i do the same. *moments* can simulate AFS for up to five populations. Moreover including the estimation of selection rates and the development of a user-friendly interface for various expert data, since now the time of separation only can be limited, will help to expand the capabilities of GADMA. Also it is possible to improve the proposed method, trying to use various modifications of the genetic algorithm, for example, the one, that infer deliberately different demographic models.

References

[1] Joshua G. Schraiber and Joshua M. Akey. Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics*, 16:727 EP –, 11 2015.

[2] Konrad Lohse, RJ Harrison, and Nicholas H Barton. A general method for calculating likelihoods under the coalescent process. *Genetics*, 189(3):977–987, 2011.

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

[4] Mary K Kuhner. Lamarc 2.0: maximum likelihood and bayesian estimation of population parameters. *Bioinformatics*, 22(6):768–770, 2006.

[5] Ronald Aylmer Fisher. XVII.—The Distribution of Gene Ratios for Rare Mutations. *Proceedings of the Royal Society of Edinburgh*, 50:204–219, 1931.

[6] Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *Genetics*, 166(1):351–372, 2004.

[7] Alison M Adams and Richard R Hudson. Maximum-likelihood estimation of demographic parameters using the frequency spectrum of unlinked single-nucleotide polymorphisms. *Genetics*, 168(3):1699–1712, 2004.

[8] Benjamin F Voight, Alison M Adams, Linda A Frisse, Yudong Qian, Richard R Hudson, and Anna Di Rienzo. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18508–18513, 2005.

[9] Hua Chen, Richard E Green, Svante Pääbo, and Montgomery Slatkin. The joint allele-frequency spectrum in closely related species. *Genetics*, 177(1):387–398, 2007.

[10] Simon Myers, Charles Fefferman, and Nick Patterson. Can one learn history from the allelic spectrum? *Theoretical population biology*, 73(3):342–348, 2008.

[11] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS genetics*, 5(10):e1000695, 2009.

[12] Laurent Excoffier and Matthieu Foll. Fastsimcoal: a continuous-time coalescent simulator of genomic diversity under arbitrarily complex evolutionary scenarios. *Bioinformatics*, 27(9):1332–1334, 2011.

[13] Julien Jouganous, Will Long, Aaron P. Ragsdale, and Simon Gravel. Inferring the joint demographic history of multiple populations: Beyond the diffusion approximation. *Genetics*, 206(3):1549–1567, 2017.

[14] Charles G Broyden. The convergence of a class of double-rank minimization algorithms: 2. the new algorithm. *IMA journal of applied mathematics*, 6(3):222–231, 1970.

[15] Roger Fletcher. A new approach to variable metric algorithms. *The computer journal*, 13(3):317–322, 1970.

[16] Donald Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[17] David F Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[18] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[19] Michael JD Powell. An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The computer journal*, 7(2):155–162, 1964.

[20] John Holland. Adaptation in natural and artificial systems: an introductory analysis with application to biology. *Control and artificial intelligence*, 1975.

[21] Derrick Joel Zwickl. *Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion*. PhD thesis, 2006.

[22] Nan Gao, Ning Yang, and Jijun Tang. Ancestral genome inference using a genetic algorithm approach. *PloS one*, 8(5):e62156, 2013.

[23] Steven Hamblin. On the practical usage of genetic algorithms in ecology and evolution. *Methods in Ecology and Evolution*, 4(2):184–194, 2013.

[24] Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

[25] M. Schumer and K. Steiglitz. Adaptive step size random search. *IEEE Transactions on Automatic Control*, 13(3):270–276, 1968.

[26] S. Gravel, B. M. Henn, R. N. Gutenkunst, A. R. Indap, G. T. Marth, A. G. Clark, F. Yu, R. A. Gibbs, C. D. Bustamante, D. L. Altshuler, et al. Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, 2011.

[27] K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6):e1003521, 2013.

[28] R. R. Sharp and C. J. Barrett. The environmental genome project: ethical, legal, and social implications. *Environmental Health Perspectives*, 108(4):279, 2000.

[29] Rajiv C McCoy, Nandita R Garud, Joanna L Kelley, Carol L Boggs, and Dmitri A Petrov. Genomic inference accurately predicts the timing and severity of a recent bottleneck in a nonmodel insect population. *Molecular ecology*, 23(1):136–150, 2014.

[30] Aaron P Ragsdale and Ryan N Gutenkunst. Inferring demographic history using two-locus statistics. *Genetics*, pages genetics–117, 2017.