Автоматизация процесса вывода совместной демографической истории нескольких популяций из аллель-частотного спектра

Носкова Екатерина Эдуардовна

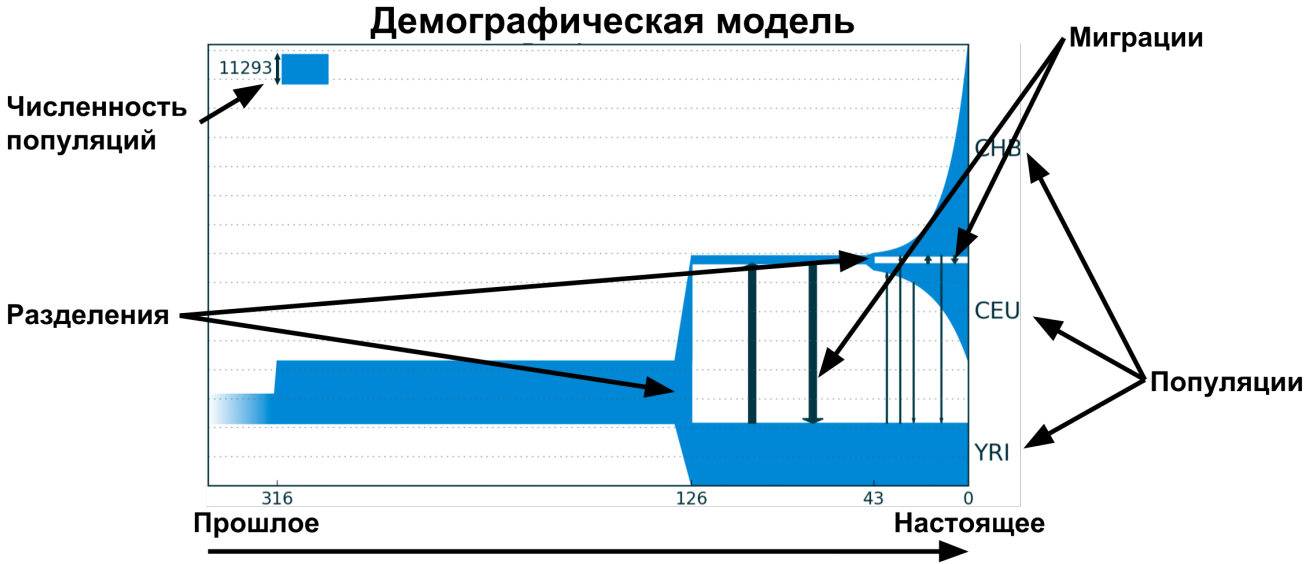
Научные руководители: к. т. н. Ульянцев Владимир Игоревич (ИТМО)
Добрынин Павел Владимирович (СПбГУ)

СПбАУ РАН

15 июня 2018 г.

Введение

- Историю развития видов и популяций можно попытаться проследить по их генетическим данным.
- Эту историю рассказывает **демографическая модель развития популяций**.



Аллель-частотный спектр

Аллель

— вариант гена или локуса генома.

Аллель-частотный спектр N популяций

— это совместное распределение частот аллелей, отличных от референса, у N популяций.

— это N-мерная гистограмма, где оси соответствуют популяциям и каждый элемент содержит число локусов, на которых аллель, отличная от референса, встретилась определенное число раз.

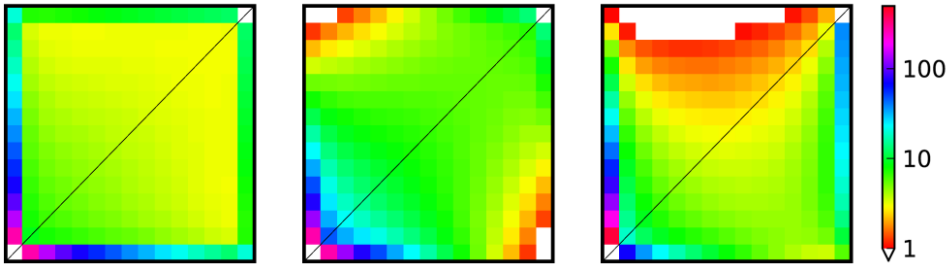


Рис. 1: Примеры аллель-частотных спектров двух популяций

Существующие подходы

Существуют два метода для симуляции ожидаемого аллель-частотного спектра M из заданной демографической модели:

- Численное решение уравнения диффузии — ∂a∂i [1].
- Аппроксимация моментов случайного процесса — *moments* [2].

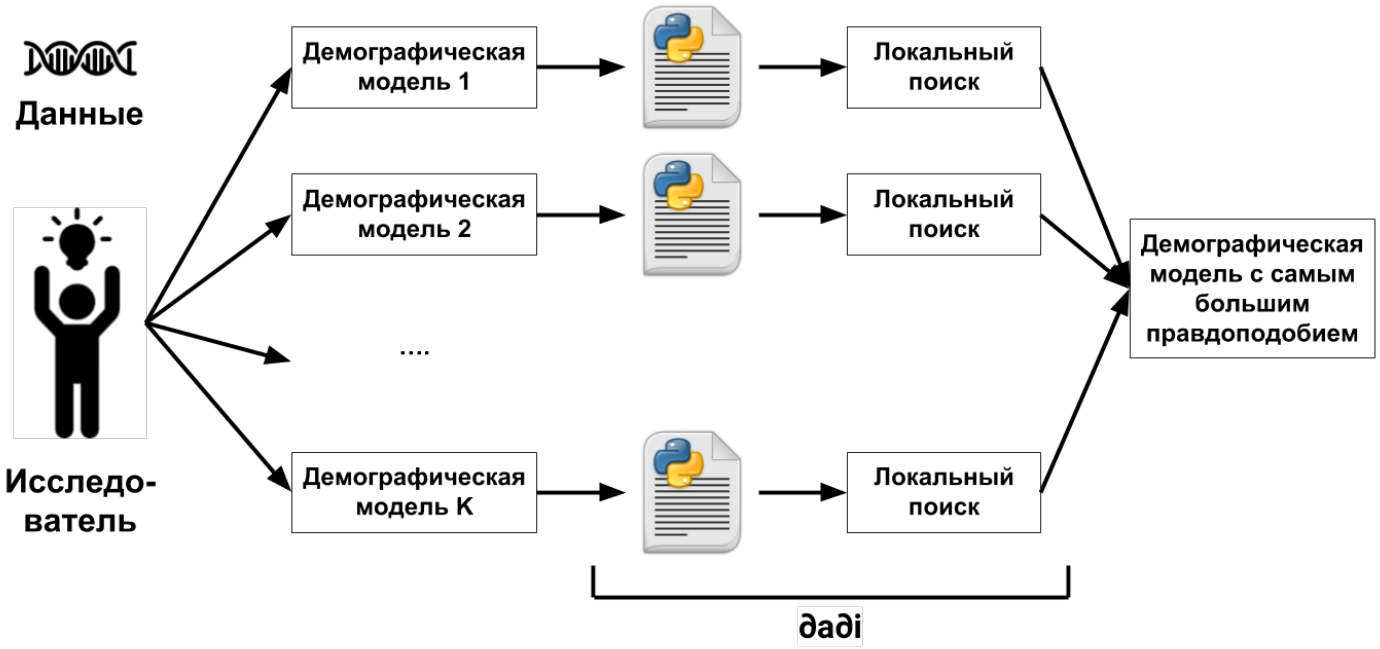
Предположение: каждый элемент аллель-частотного спектра — это независимая Пуассоновская случайная величина.

Тогда можем посчитать **правдоподобие** — вероятность получить наблюдаемый спектр S при условии, что ожидаемый спектр — M:

$$\mathcal{L}(M|S) = \prod_{i=1,\cdots P} \prod_{d_i=1,\cdots n_i} \frac{e^{-M[d_1,\cdots,d_P]} M[d_1,\cdots,d_P]^{S[d_1,\cdots,d_P]}}{S[d_1,\cdots,d_P]!}$$

[1] Gutenkunst et al., 2009
[2] Jouganous et al., 2017

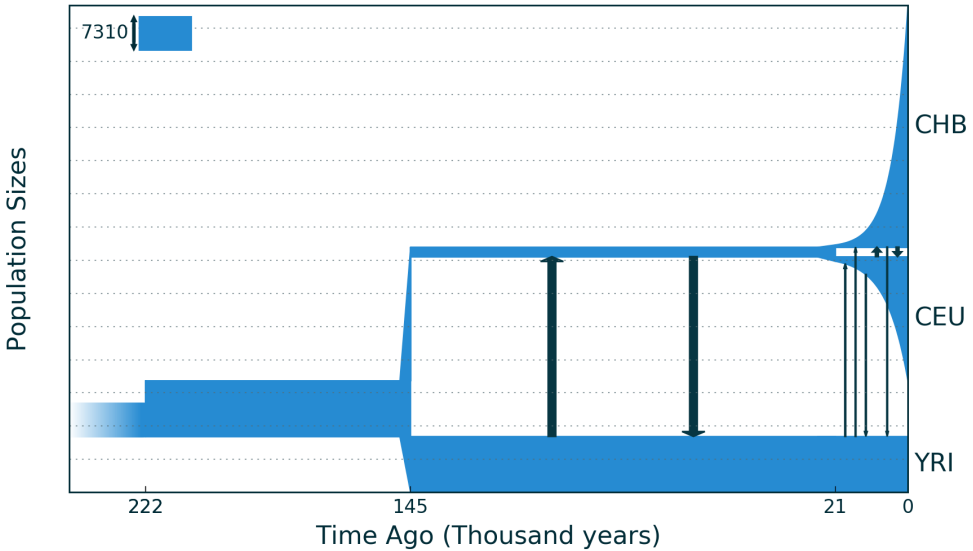
Существующие подходы: общая схема



Существующая демографическая модель выхода людей из Африки — Gutenkunst et al. (2009)

Дано 3 популяции людей:

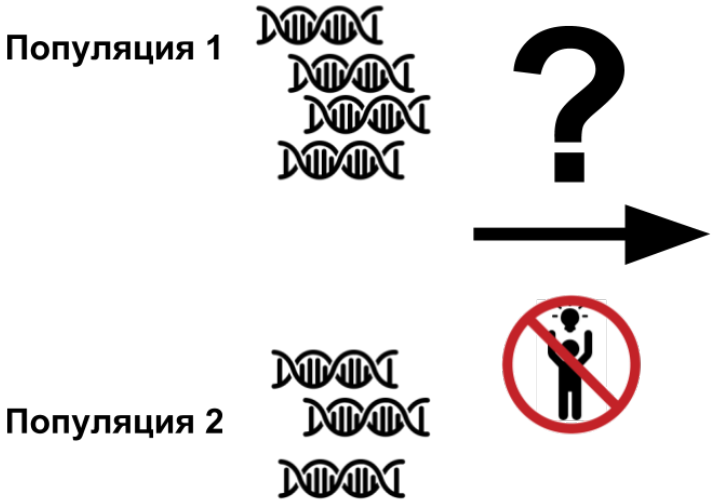
- YRI — люди народа Йоруба из Нигерии,
- CEU — жители штата Юта с предками из западной Европы,
- CHB — люди народа Хань из Пекина.

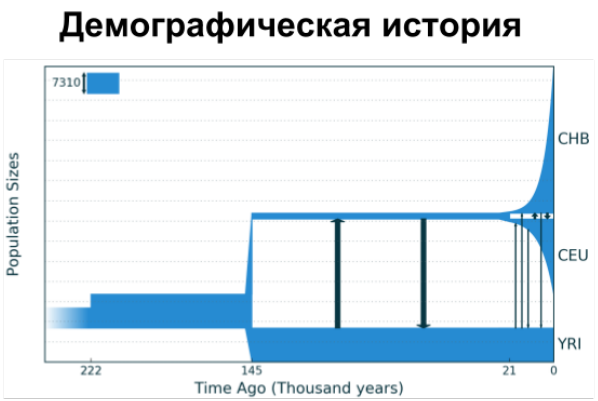


Логарифм правдоподобия: -6316.89

Цель диссертации

Цель: Автоматизация вывода демографической истории нескольких популяций из аллель-частотного спектра.



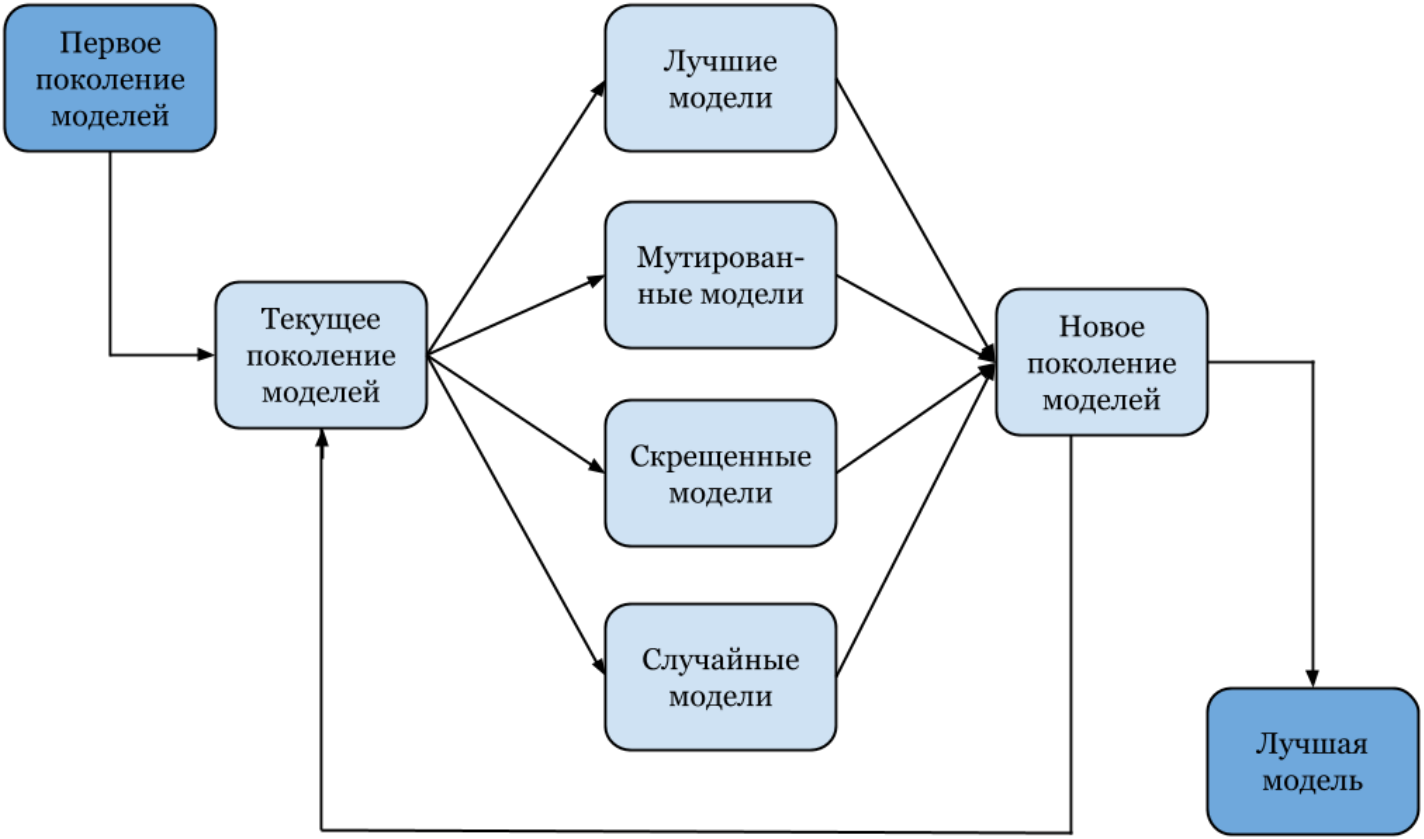


Задачи диссертации

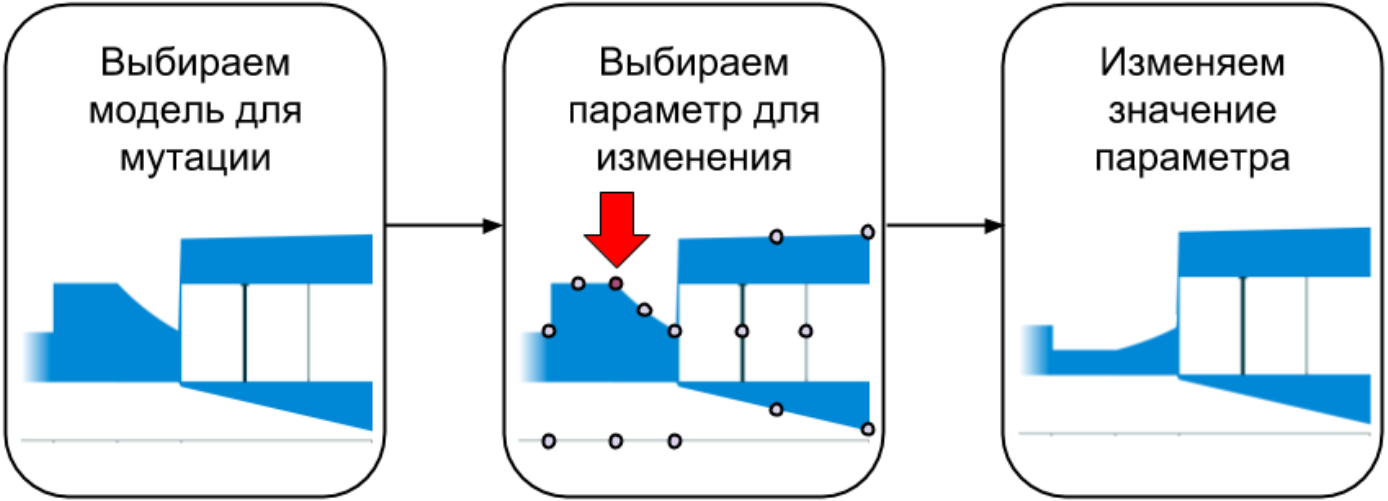
Задачи:

- Разработать метод автоматического глобального поиска демографичеякой модели по аллель-частотному спектру, основанный на генетическом алгоритме.
- Реализовать разработанный метод в прототипе программного средства.
- Провести экспериментальные исследования с использованием реальных геномов (люди *Homo sapiens*, бабочки *E. gillettii*).

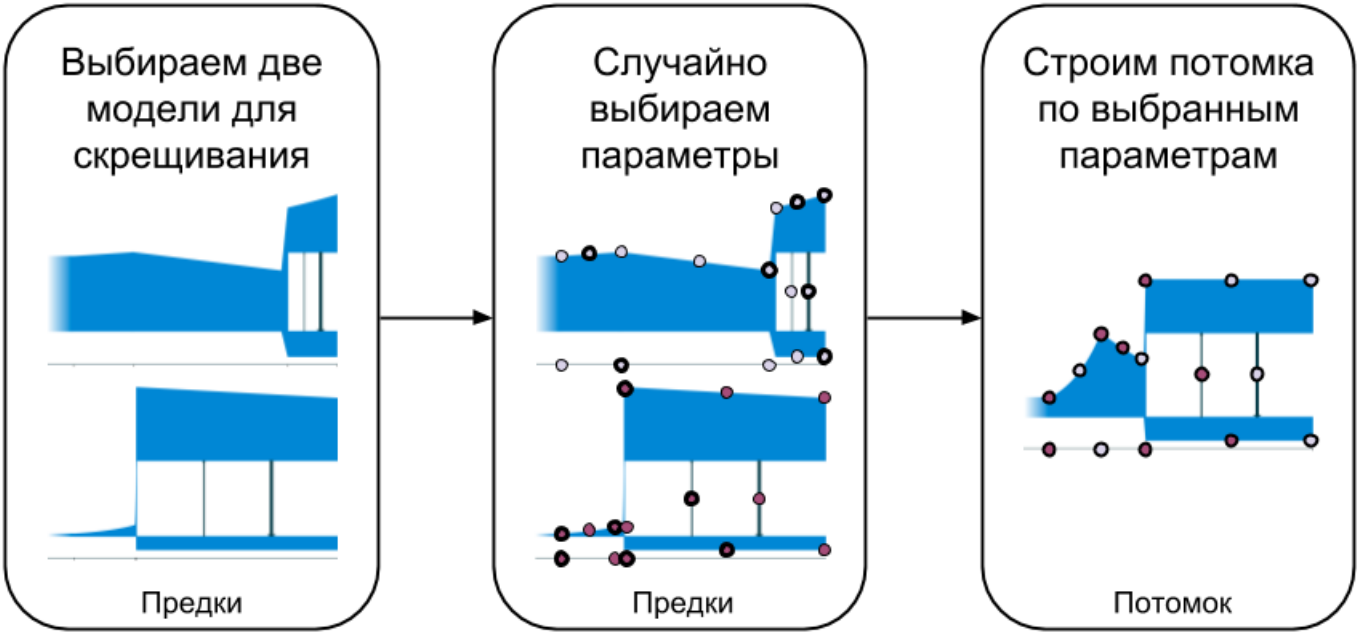
Генетический алгоритм



Мутация демографической модели

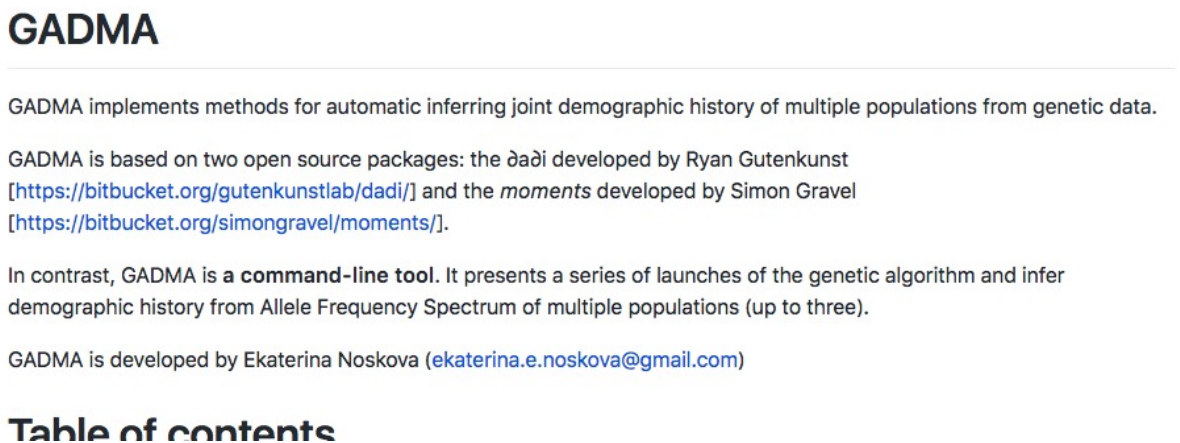


Скрещивание демографических моделей



Реализация

- GADMA (Genetic Algorithm for Demographic Model Analysis) — программное обеспечение для поиска совместной демографической модели популяций из аллель-частотного спектра.
- Язык разработки: Python.



GADMA

GADMA implements methods for automatic inferring joint demographic history of multiple populations from genetic data.

GADMA is based on two open source packages: the ∂a∂i developed by Ryan Gutenkunst [https://bitbucket.org/gutenkunstlab/dadi/] and the *moments* developed by Simon Gravel [https://bitbucket.org/simongravel/moments/].

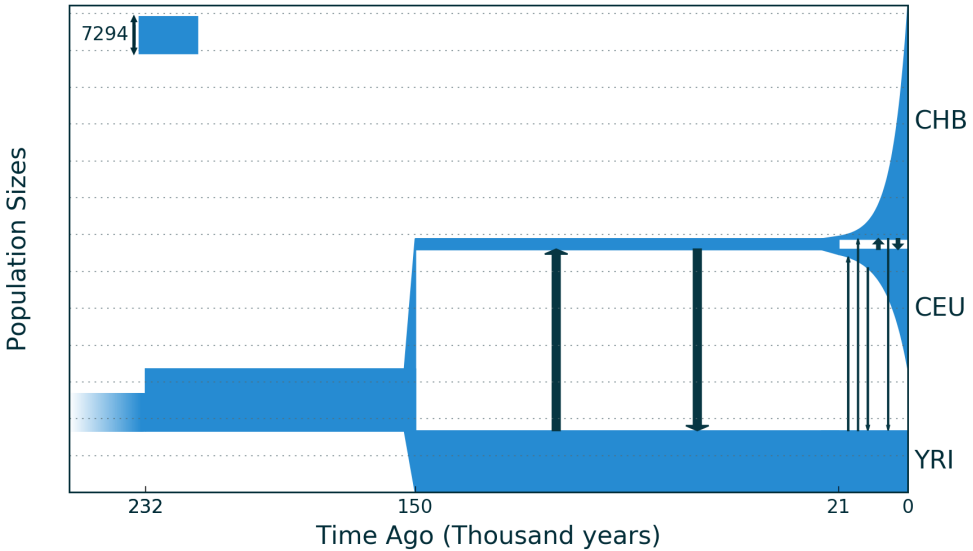
In contrast, GADMA is **a command-line tool**. It presents a series of launches of the genetic algorithm and infer demographic history from Allele Frequency Spectrum of multiple populations (up to three).

GADMA is developed by Ekaterina Noskova (ekaterina.e.noskova@gmail.com)

Table of contents

Демографическая модель выхода людей из Африки, полученная нашим методом (1)

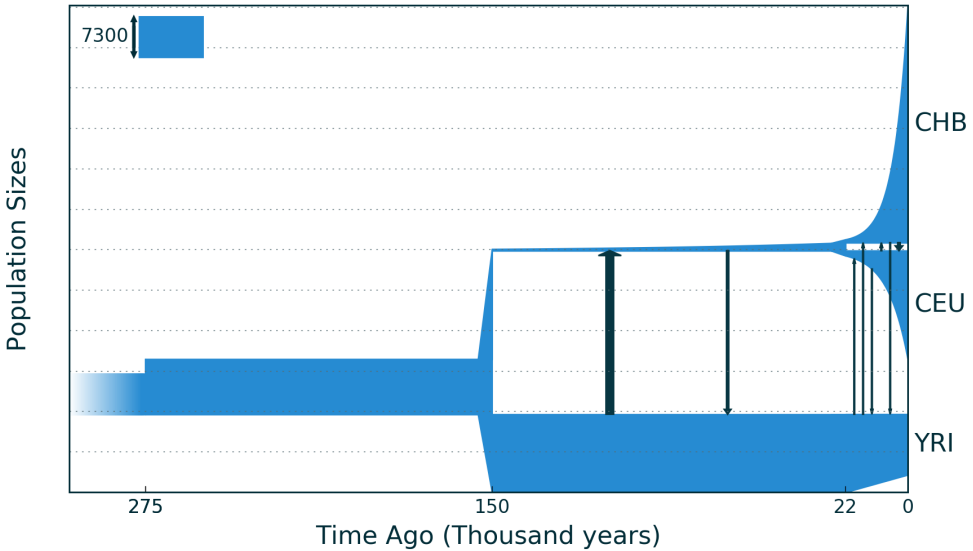
- Подбирались те же параметры (14 штук), что и в Gutenkunst et al.
- Экспертные данные: время выхода из Африки не более 150 тыс. лет назад.



Логарифм правдоподобия: -6315.86

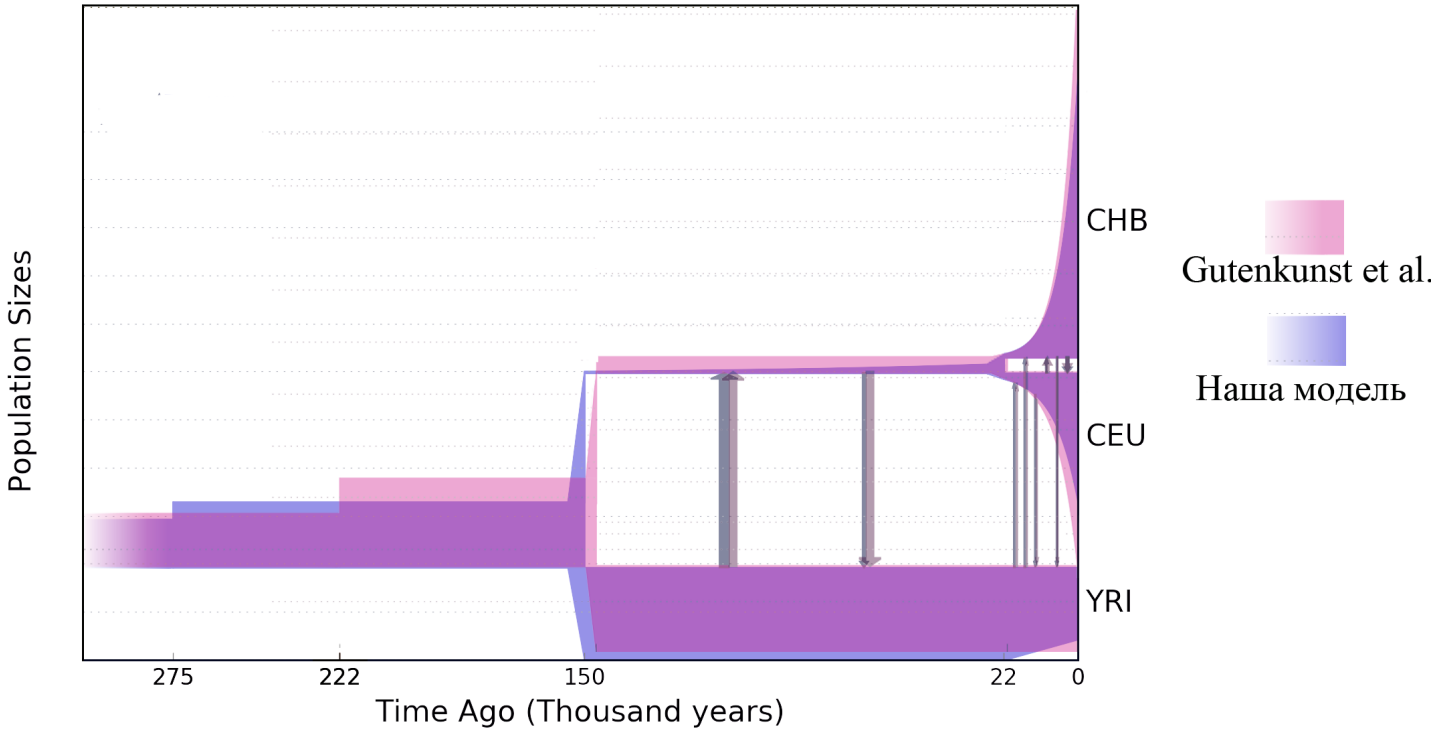
Демографическая модель выхода людей из Африки, полученная нашим методом (2)

- Подбирались все доступные параметры: 23 штуки.
- Экспертные данные: время выхода из Африки не более 150 тыс. лет назад.



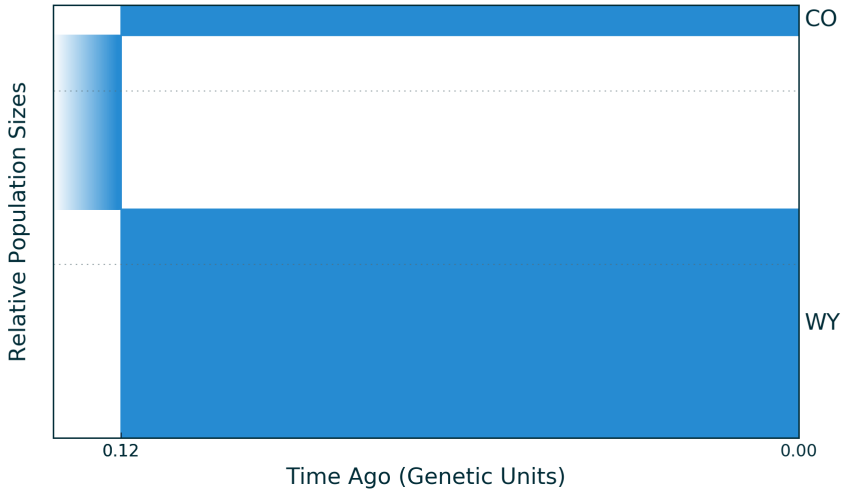
Логарифм правдоподобия: -6288.37

Сравнение существующей модели и полученной



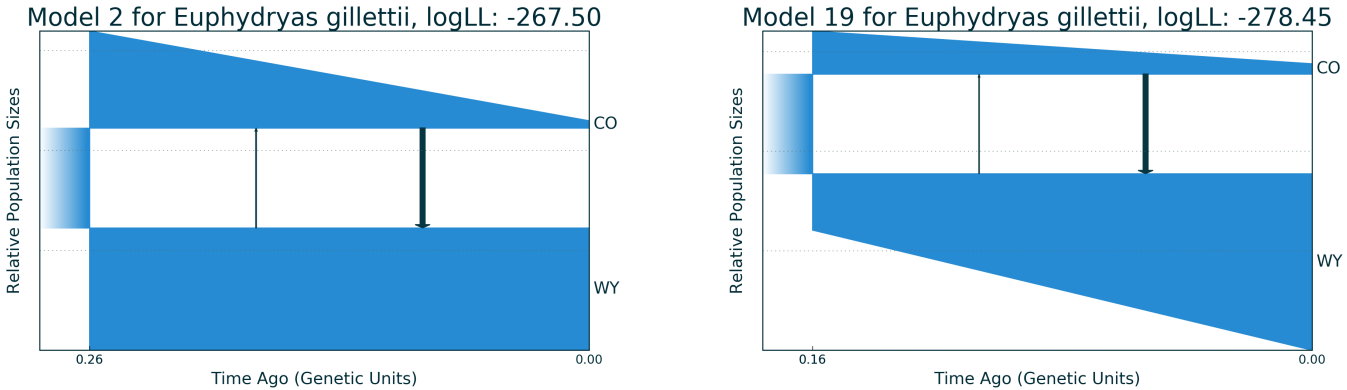
Существующая демографическая модель для бабочек *E. gillettii* McCoy et al. (2013).

- WY — популяция бабочек *Euphydryas gillettii* в штате Вайоминг,
- CO — популяция бабочек *Euphydryas gillettii* в штате Колорадо.



Логарифм правдоподобия: -284.17

Альтернативные демографические модели для бабочек *E. gillettii*, полученные нашим методом



LogLL — логарифм правдоподобия

Результаты

- Был разработан и реализован метод автоматического вывода демграфических моделей из аллель-частотного спектра, на основе генетического алгоритма.
- Были проведены экспериментальные исследования на реальных данных: выведены демографические модели для трех популяций современных людей и двух популяций бабочек *E. gillettii*.
- Метод позволил подобрать модели, лучшие по правдоподобию, чем те, что были подобраны ранее.
- Также метод предоставил несколько альтернативных демографических моделей с близким значением правдоподобия.

Спасибо за внимание!

Пример аллель-частотного спектра

Референс:	ATACG	
	1 популяция	2 популяция
1 особь	ATCCG	ACACG
2 особь	ACACG	ACACT
3 особь	GCACG	

Позиция	1	2	3	5
Аллель	G	C	C	T
Частота в 1 поп.	1	2	1	0
Частота во 2 поп.	0	2	0	1

$$\Longrightarrow \quad A = \begin{array}{c} \\ 2 \\ 1 \\ 0 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 \end{pmatrix} \end{array}$$

Аллель-частотный спектр N популяций

— это N-мерная гистограмма, где оси соответствуют популяциям и каждый элемент содержит число позиций, на которых аллель, отличная от референса, встретилась определенное число раз.

Модель Райта-Фишера

- Пусть у нас имеется одна популяция размера N.
- Локус A — две аллели A_1, A_2.
- Обозначим $X(t)$ — число аллелей A_1 в поколении t.
- Очевидно, $X(t) \in \{0, 1, ..., 2N\}$
- Тогда $X(t+1)$ — биномиальная случайная величина:

$$p_{ij} = P(X(t+1) = j | X(t) = i) = \binom{2N}{j} \left(\frac{i}{2N}\right)^j \left(1 - \frac{i}{2N}\right)^{2n-j}$$

Forward уравнение Колмогорова

- Пусть у нас имеется марковская цепь с исходами $\{0, 1/M, 2/M, \cdots, 1\}$ и матрицей переходов $P = \{p_{ij}\}$, обозначим $f(i, k, t) = p_{ij}^{(t)}$

Если выполнено:

$$E(\delta x) = a(x)\delta t + o(\delta t),$$

$$var(\delta x) = b(x)\delta t + 0(\delta t),$$

$$E(|\delta x|^3) = 0(\delta t).$$

то:

$$\frac{\partial f(x;t)}{\partial t} = -\frac{\partial}{\partial x}\{a(x)f(x;t)\} + \frac{1}{2}\frac{\partial^2}{\partial x^2}\{b(x)f(x;t)\}.$$

— forward уравнение Колмогорова или уравнение Фоккера-Планка.

Уравнение диффузии популяционной генетики

Пусть у нас имеется P популяций, где:

- ν_i — численность популяции i,
- γ_i — отбор,
- M_{ij} — темпы миграции.

Тогда можно записать следующее уравнение Фоккера-Планка:

$$\frac{\partial f(x;t)}{\partial t} = -\sum_{i=1,..,P} \frac{\partial}{\partial x}\left(\gamma_i x_i(1-x_i) + \sum_{j=1,...P} M_{ij}(x_i - x_j)\right) f(x;t) +$$

$$+\frac{1}{2}\sum_{i=1,..,P} \frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{\nu_i} f(x;t).$$