ИТМО

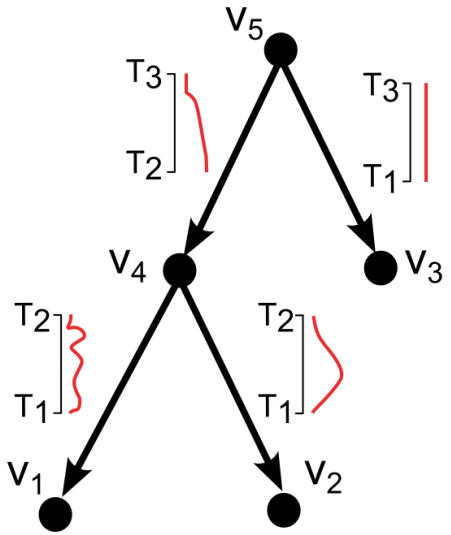
Носкова Екатерина Эдуардовна

Методы построения моделей демографических историй

Специальность 1.2.2 –
Математическое моделирование, численные методы и комплексы программ

Научный руководитель:
к.т.н. Ульянцев Владимир Игоревич

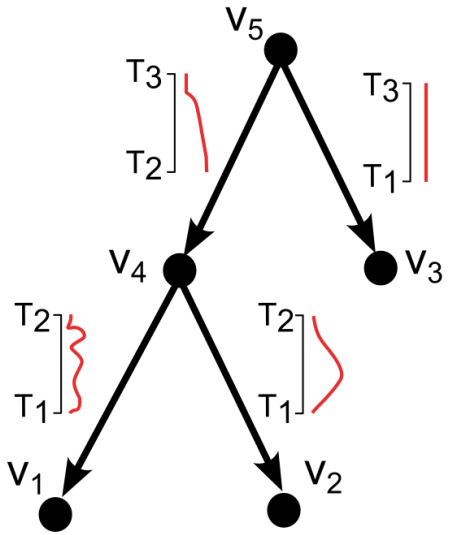
Метрическое дерево с функциями на ребрах



Метрическое дерево с функциями на ребрах

- ***Метрическое дерево*** (metric tree) — граф, являющийся деревом, где каждому ребру поставлен в соответствие интервал $[t_1, t_2]$

Метрическое дерево с функциями на ребрах

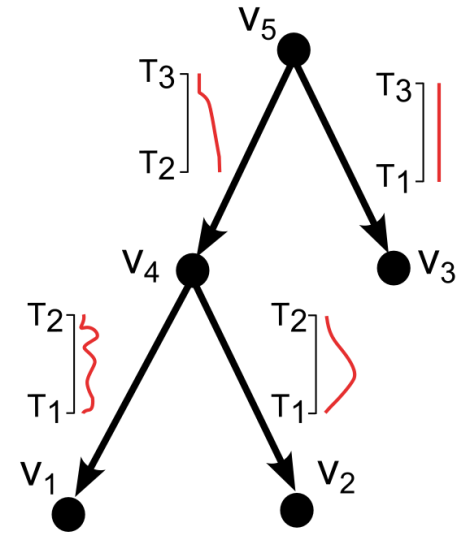


Метрическое дерево с функциями на ребрах

- ***Метрическое дерево*** (metric tree) — граф, являющийся деревом, где каждому ребру поставлен в соответствие интервал $[t_1, t_2]$

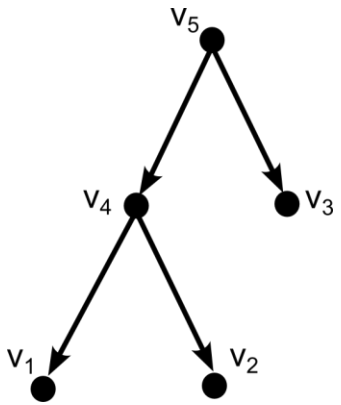
- **Метрические деревья с функциями на ребрах** широко применяются для моделирования различных объектов в физике (волноводы) и в биологии (демографические истории)

Метрическое дерево с функциями на ребрах

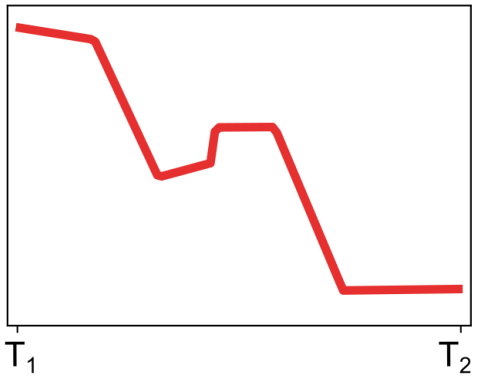


Метрическое дерево с функциями на ребрах

Модель графа



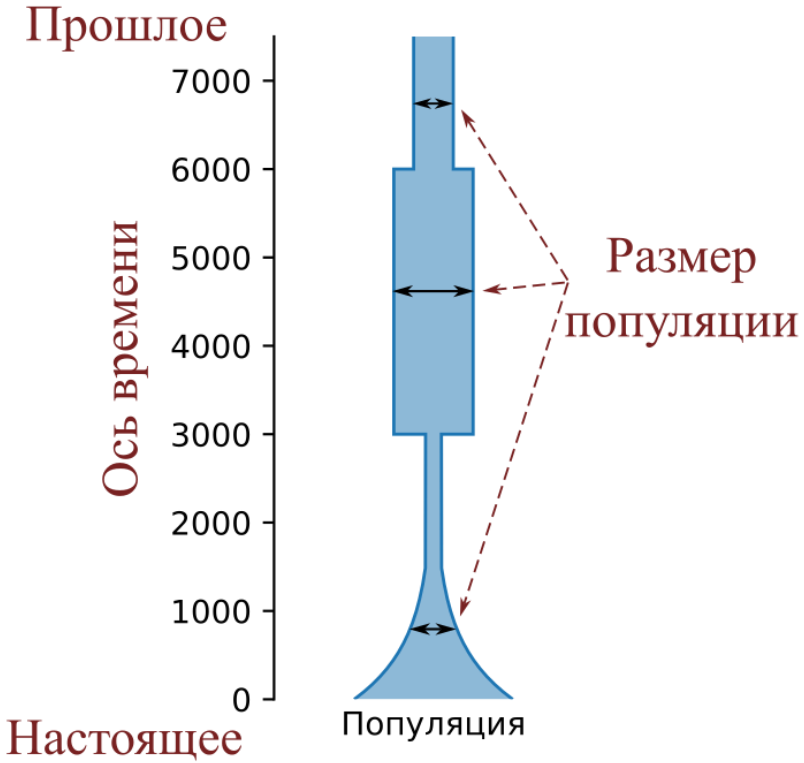
Модели функций

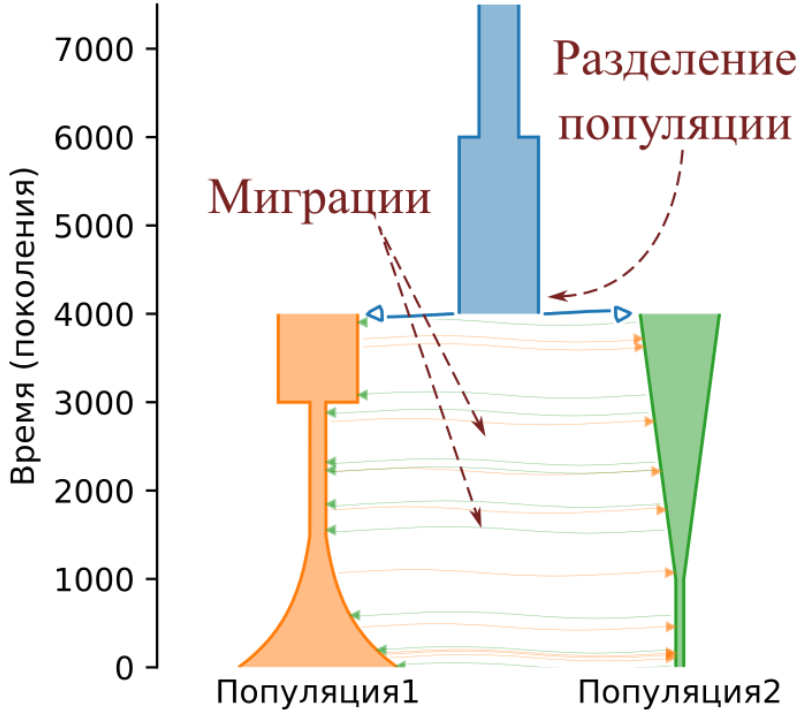


- ***Метрическое дерево*** (metric tree) — граф, являющийся деревом, где каждому ребру поставлен в соответствие интервал $[t_1, t_2]$
- **Метрические деревья с функциями на ребрах** широко применяются для моделирования различных объектов в физике (волноводы) и в биологии (демографические истории)
- При работе с такими моделями прибегают к **экспертным данным** для определения:
  - структуры графа
  - свойств функций на ребрах дерева

Демографическая история популяций

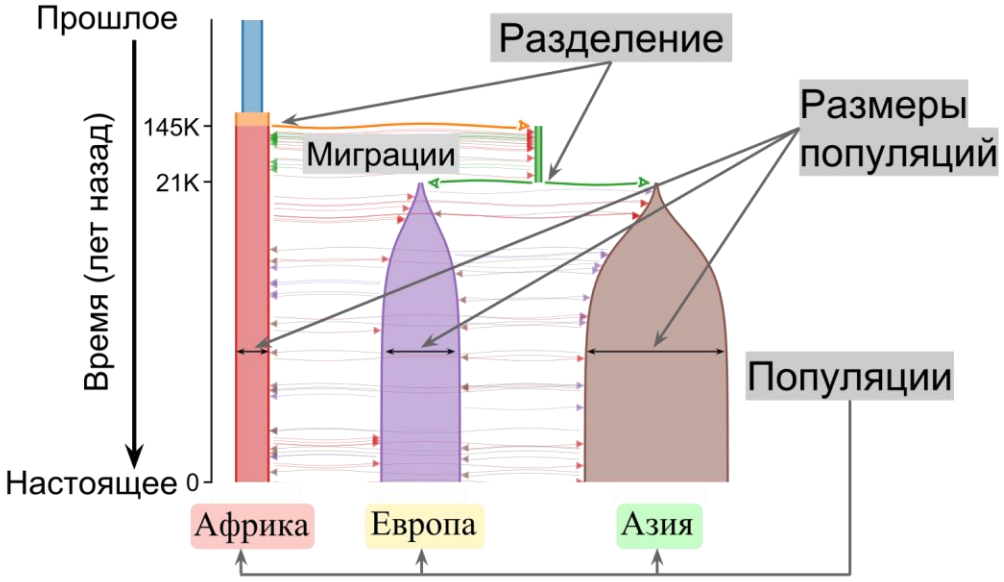
- Демографическая история популяций – это история развития этих популяций, которая включает в себя информацию о дереве разделения, времени разделений, численности популяций в прошлом и миграциях

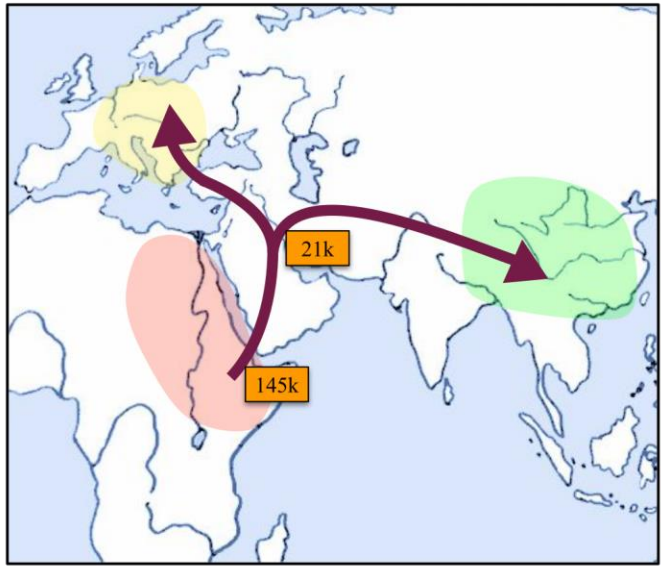




Демографическая история популяций

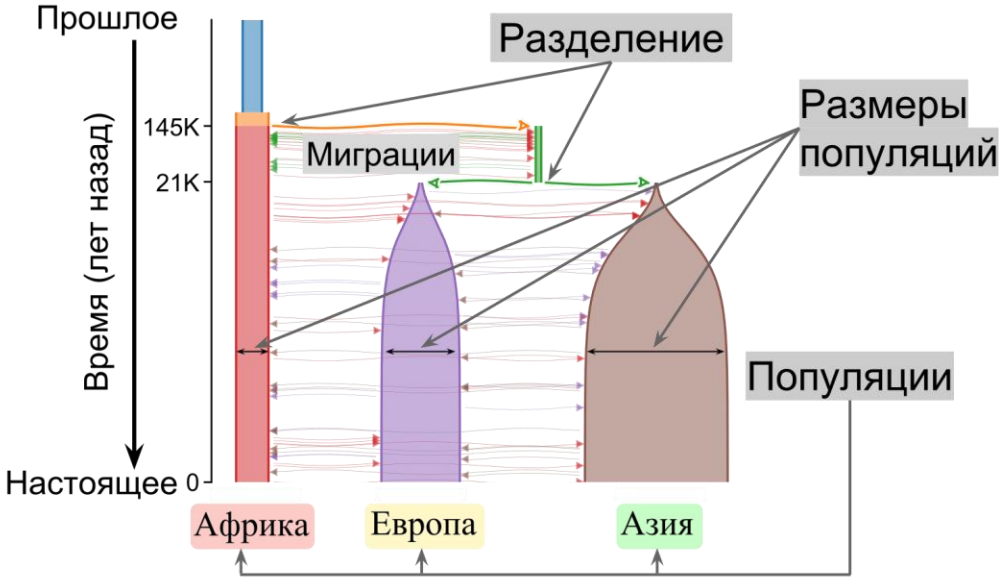
- Демографические истории используются для датирования исторических событий, не оставивших письменных свидетельств

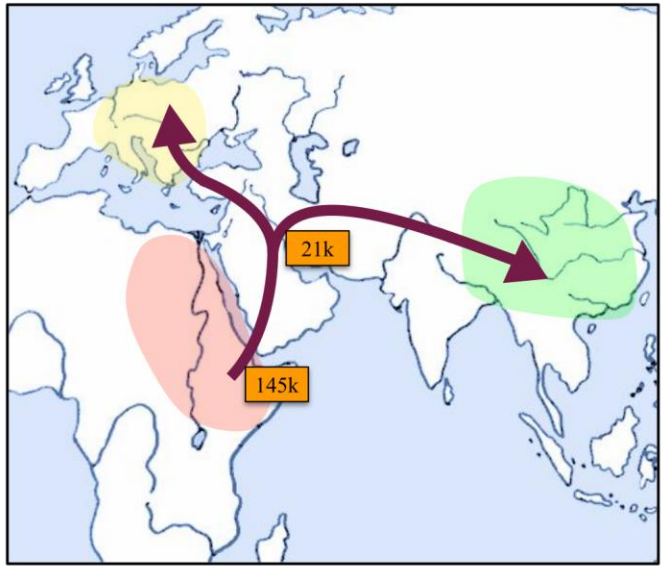


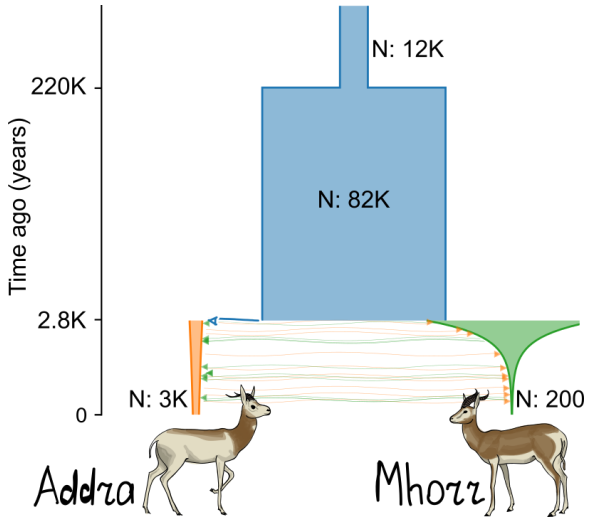


Демографическая история популяций

- Демографические истории используются для датирования исторических событий, не оставивших письменных свидетельств
- Они играют важную роль в области консервативной генетики для изучения исчезающих видов



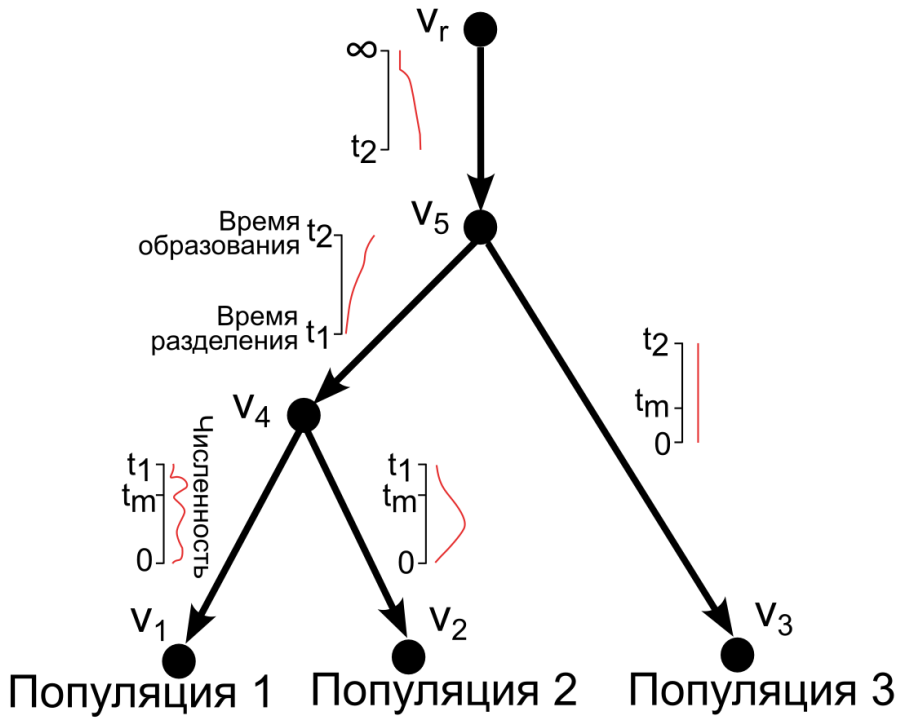




Демографическая история популяций

Определение 1. Демографическая история D для P популяций — это четверка $\langle T, G, M, C \rangle$, где:

- $T = \langle V, E \rangle$ — дерево разделения популяций
- G — отображение, которое для каждого ребра e ставит в соответствие интервал $[t_s, t_e]$ и функцию изменения численности $g(t)$
- M — набор единичных и непрерывных миграций
- C — набор дополнительных констант для популяций



Миграции

t_m, 0: $(v_4, v_1) \xrightarrow{m_{13}} (v_5, v_3)$

t_1, t_m: $(v_4, v_1) \underset{m_{21}}{\overset{m_{12}}{\rightleftarrows}} (v_4, v_2)$

t_m: $(v_4, v_2) \xrightarrow{m_{23}} (v_5, v_3)$

Константы

Коэффициенты инбридинга:

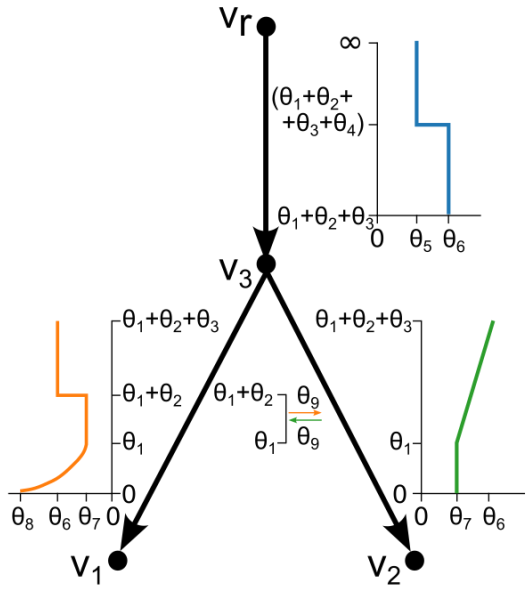
$v_1 : 0$

$v_2 : 0$

$v_3 : F$

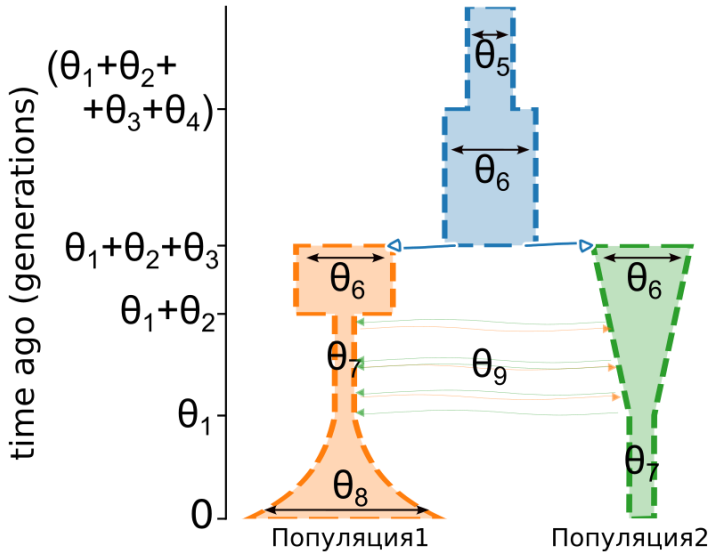
Модель демографической истории популяций

- Метрические деревья с функциями на ребрах используются для моделирования демографической истории популяций
- Функция изменения численности популяции — кусочно-заданная функция, состоящая из сегментов константной, линейной и экспоненциальной динамики

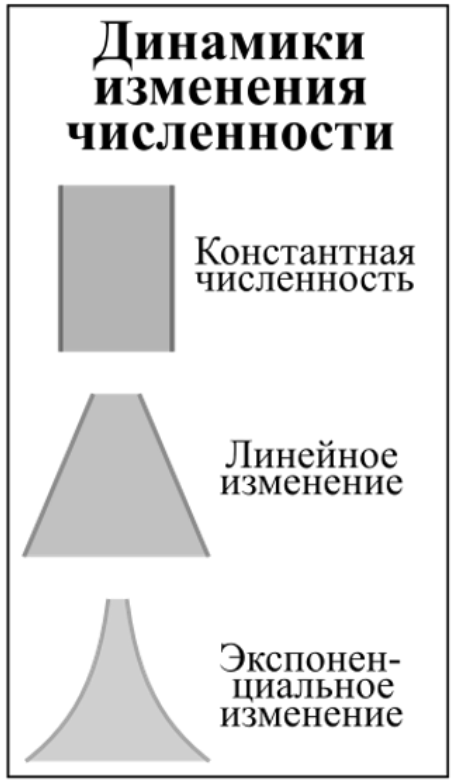


Метрическое дерево с функциями на ребрах

=



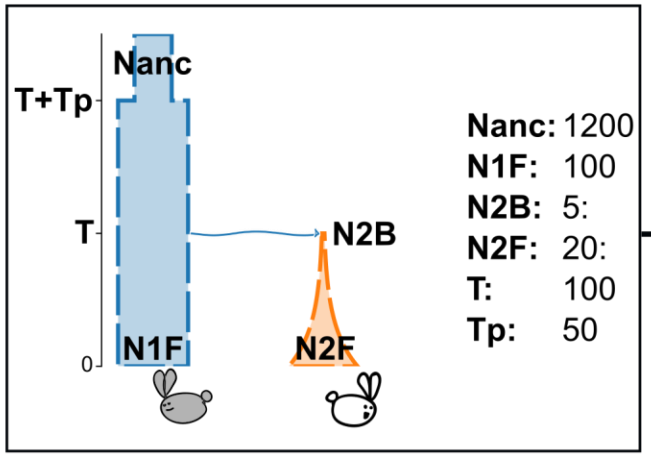
Модель демографической истории



Функция правдоподобия для модели

Методы вычисления правдоподобия генетических данных при условии заданной модели демографической истории популяций — это **методы численного имитационного моделирования**

Модель демографической истории со значениями параметров

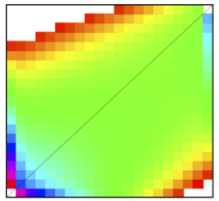


Моделирование процесса эволюции

Численное решение дифференциальных уравнений

$$\frac{\partial f(x;t)}{\partial \tau} = -\sum_{i=1,..,P} \frac{\partial}{\partial x}\left(\sum_{j=1,..P} M_{ij}(x_i - x_j)\right) f(x;t) + \frac{1}{2}\sum_{i=1,..,P} \frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{\nu_i} f(x;t).$$

Ожидаемая статистика



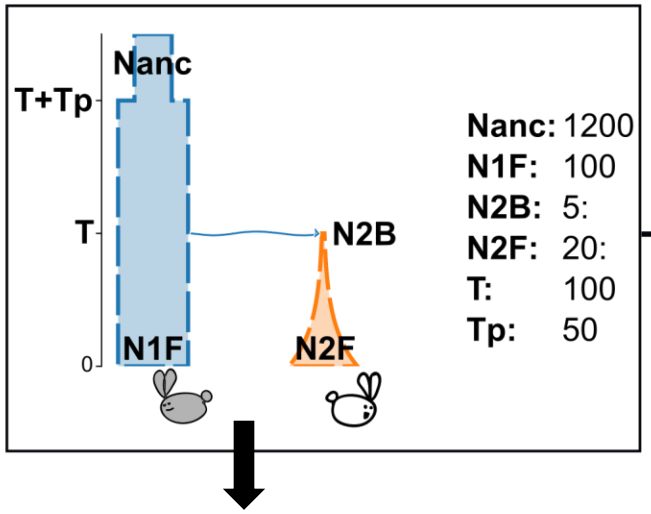
Генетические данные

Логарифм правдоподобия

Функция правдоподобия для модели

Методы вычисления правдоподобия генетических данных при условии заданной модели демографической истории популяций — это **методы численного имитационного моделирования**

Модель демографической истории со значениями параметров



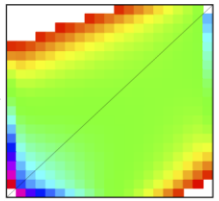

```
# momi model
model = momi.DemographicModel()
model.add_leaf("1", N="N1F", g=0)
model.add_leaf("2", N="N2F", g="r2")
model.move_lineages("2", "1", t="T")
model.set_size("1", N="Nanc", t="Tp")
```

Моделирование процесса эволюции

Численное решение дифференциальных уравнений

$$\frac{\partial f(x;t)}{\partial \tau} = -\sum_{i=1,..,P} \frac{\partial}{\partial x}\left(\sum_{j=1,..P} M_{ij}(x_i - x_j)\right) f(x;t) + \frac{1}{2}\sum_{i=1,..,P} \frac{\partial^2}{\partial x^2}\frac{x_i(1-x_i)}{\nu_i} f(x;t).$$

Ожидаемая статистика



Генетические данные

Логарифм правдоподобия

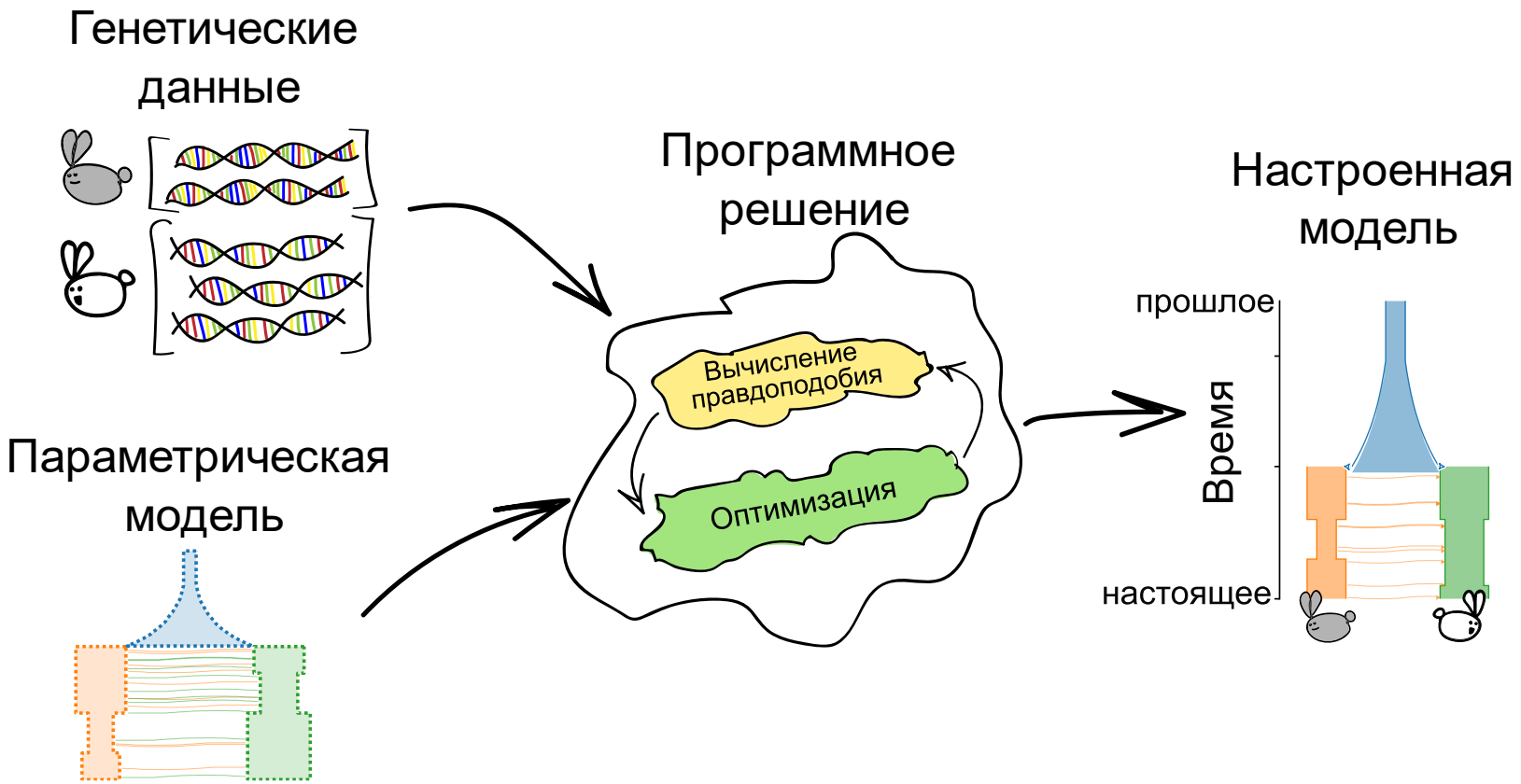
Вывод демографической истории популяций

Вход:

- Генетические данные
- Параметрическая модель

Выход:

- Демографическая история популяций как модель с настроенными параметрами



Степень разработанности проблемы

- Существуют следующие методы и программные решения для компьютерного моделирования демографических историй популяций:

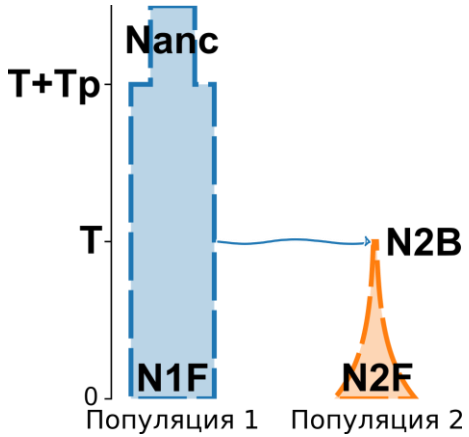
Библиотека	**Год**	**Метод вычисления правдоподобия**	**Метод оптимизации**	**Перебор моделей**
dadi	2009	Метод аппроксимации диффузией	Локальная оптимизация (BFGS, Powell, N-M, BOBYQA)	Ручной
moments	2017	Метод моментов для статистик AFS	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momentsLD	2019	Метод моментов для статистик LD	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momi2	2020	Метод непрерывной модели Морана	Локальная оптимизация (TNC)	Ручной
dadi-pipeline	2017	Метод из *dadi*	Множественный запуск локальной оптимизации N-M	Ручной
moments-pipeline	2019	Метод из *moments*	Множественный запуск локальной оптимизации N-M	Ручной

Степень разработанности проблемы

- Существуют следующие методы и программные решения для компьютерного моделирования демографических историй популяций:

Библиотека	Год	Метод вычисления правдоподобия	Метод оптимизации	Перебор моделей
dadi	2009	Метод аппроксимации диффузией	Локальная оптимизация (BFGS, Powell, N-M, BOBYQA)	Ручной
moments	2017	Метод моментов для статистик AFS	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momentsLD	2019	Метод моментов для статистик LD	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momi2	2020	Метод непрерывной модели Морана	Локальная оптимизация (TNC)	Ручной

- Имеют **собственные интерфейсы** для **ручной** спецификации параметрических моделей




```python
def dadi_model(params, ns, pts):
    Nanc, N1F, N2B, N2F, Tp, T = params

    xx = yy = dadi.Numerics.default_grid(pts)
    phi = dadi.PhiManip.phi_1D(xx, nu=Nanc)
    phi = dadi.Integration.one_pop(phi, xx, T=Tp, nu=N1F)
    phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)
    n2_func = lambda t: N2B * (N2F / N2B) ** (t / T)
    phi = dadi.Integration.two_pops(phi, xx, T=T,
                                    nu1=N1F, nu2=n2_func)
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,yy))
    return sfs
```

```python
# momi model
model = momi.DemographicModel()
model.add_leaf("1", N="N1F", g=0)
model.add_leaf("2", N="N2F", g="r2")
model.move_lineages("2", "1", t="T")
model.set_size("1", N="Nanc", t="Tp")
```

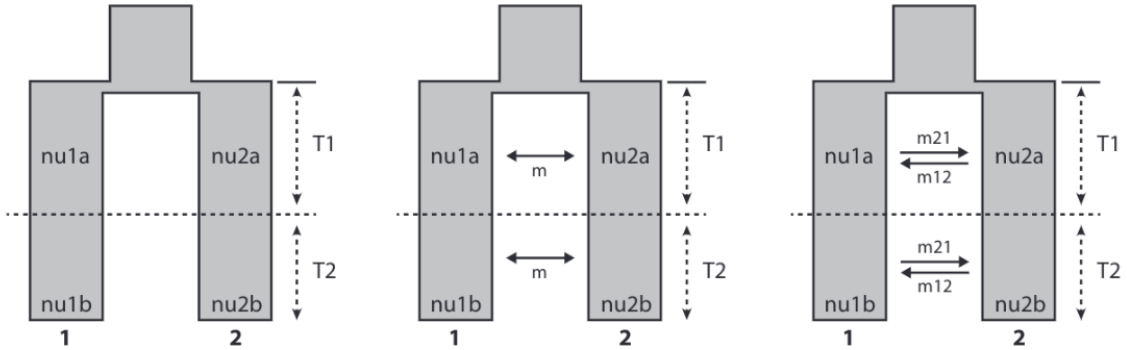
Степень разработанности проблемы

- Существуют следующие методы и программные решения для компьютерного моделирования демографических историй популяций:

Библиотека	**Год**	**Метод вычисления правдоподобия**	**Метод оптимизации**	**Перебор моделей**
dadi	2009	Метод аппроксимации диффузией	Локальная оптимизация (BFGS, Powell, N-M, BOBYQA)	Ручной
moments	2017	Метод моментов для статистик AFS	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momentsLD	2019	Метод моментов для статистик LD	Локальная оптимизация (BFGS, Powell, N-M)	Ручной
momi2	2020	Метод непрерывной модели Морана	Локальная оптимизация (TNC)	Ручной
dadi-pipeline	2017	Метод из *dadi*	Множественный запуск локальной оптимизации N-M	Ручной
moments-pipeline	2019	Метод из *moments*	Множественный запуск локальной оптимизации N-M	Ручной

Степень разработанности проблемы

- Существуют следующие методы и программные решения для компьютерного моделирования демографических историй популяций:



- Имеют **каталог** заранее-специфицированных моделей для выбора пользователя

dadi-pipeline	2017	Метод из *dadi*	Множественный запуск локальной оптимизации N-M	Ручной
moments-pipeline	2019	Метод из *moments*	Множественный запуск локальной оптимизации N-M	Ручной

Недостатки существующих решений

1. Требуется **вручную перебирать** множество похожих моделей, отличающихся общим видом функций и числом временных интервалов
2. Требуется **специфицировать** одну и ту же модель каждый раз заново для каждого используемого программного средства, их нельзя переиспользовать
3. Для настройки параметров моделей используются **алгоритмы локального поиска**, которые не гарантируют нахождения глобального оптимума
4. Г**иперпараметры** методов оптимизации **не настроены**
5. Не существует **метода автоматического перебора моделей**

Недостатки существующих решений

1. Требуется **вручную перебирать** множество похожих моделей, отличающихся общим видом функций и числом временных интервалов
2. Требуется **специфицировать** одну и ту же модель каждый раз заново для каждого используемого программного средства, их нельзя переиспользовать
3. Для настройки параметров моделей используются **алгоритмы локального поиска**, которые не гарантируют нахождения глобального оптимума
4. Г**иперпараметры** методов оптимизации **не настроены**
5. Не существует **метода автоматического перебора моделей**

Актуальна разработка специализированных моделей и методов для автоматического построения и настройки моделей метрических деревьев с функциями на ребрах

Цель диссертационной работы

- Повышение **качества**[1] компьютерного моделирования явлений реального мира **за счет автоматизации построения и настройки** моделей метрических деревьев с функциями на ребрах.

[1]**Качество моделей** определяется степенью соответствия настроенной модели данным натурного эксперимента. В случае задачи вывода демографических историй популяций качество определяется значением функции правдоподобия, полученным численными методами за фиксированное время настройки модели.

Задачи диссертационной работы

- исследование текущего состояния предметной области, уточнение проблемы и способов оценки результатов;
- формализация постановки задачи построения и настройки моделей метрического дерева с функциями на ребрах;
- **разработка метода автоматической настройки моделей** метрического дерева с функциями на ребрах на основе комбинации методов глобальной и локальной оптимизации;
- **разработка метода автоматического перебора моделей** метрического дерева с кусочно-заданными функциями на ребрах;
- **проектирование и реализация программного комплекса**, включающего разработанные модели и методы для вывода демографической истории популяций по генетическим данным;
- **проведение экспериментальных исследований**, подтверждающих эффективность разработанных моделей и методов, а также их применимость для вывода демографической истории популяций по генетическим данным, анализ результатов экспериментов.

Основные положения, выносимые на защиту:

1. **Метод моделирования и настройки параметров моделей метрических деревьев с функциями на ребрах** по данным натурного эксперимента, содержащий модели с непрерывными функциональными параметрами, отличающийся тем, что с целью **автоматической настройки без привлечения экспертных данных** в нем используются модели с дискретными параметрами, определяющими семейства функций, а также методы глобальной оптимизации — генетический алгоритм и байесовская оптимизация, и реализующий его комплекс программ.
2. **Метод автоматического перебора моделей** метрических деревьев с функциями на ребрах с разным числом параметров и настройки этих параметров по данным натурного эксперимента, содержащий сравнение моделей с использованием информационного критерия Акаике, отличающийся тем, что с целью повышения уровня автоматизации и обеспечения возможности **настраивать не только параметры модели, но и саму модель**, он включает метод увеличения числа временных интервалов для кусочно-заданных функций на ребрах дерева, а также реализующий его комплекс программ.

Соответствие паспорту специальности 1.2.2

- Пункт 2 паспорта специальности
  - Разработка, обоснование и тестирование эффективных вычислительных методов с применением современных компьютерных технологий
- Пункт 4 паспорта специальности
  - Разработка новых математических методов и алгоритмов интерпретации натурного эксперимента на основе его математической модели

Структура диссертационной работы

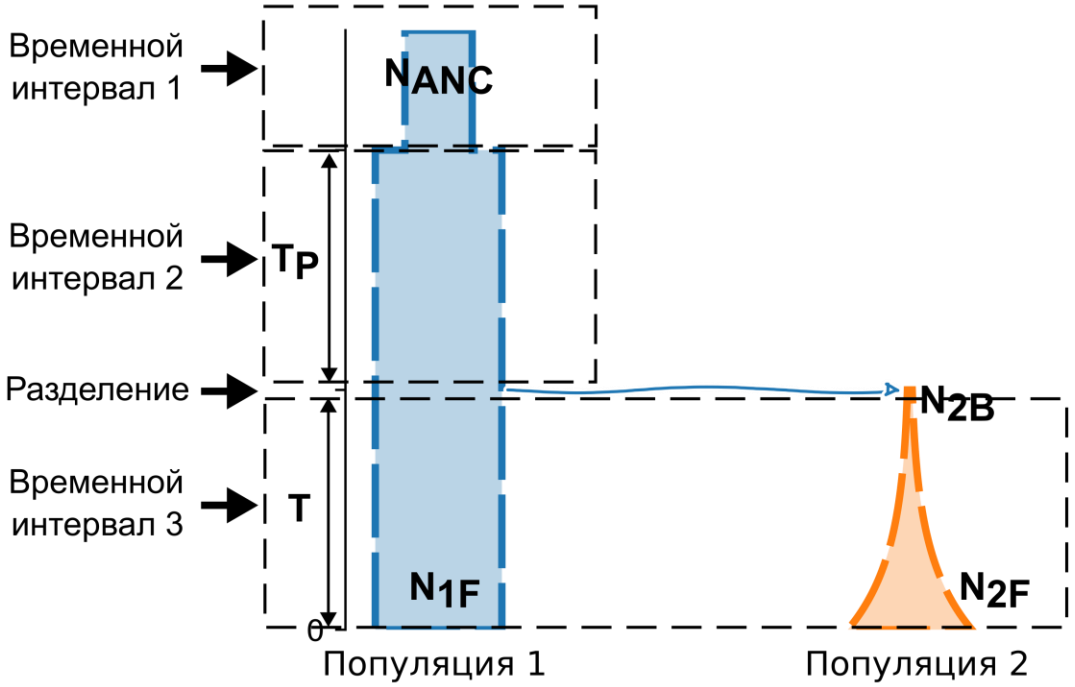
- Глава 1. Обзор предметной области
- Глава 2. Расширенный класс моделей демографической истории популяций и методы настройки параметров моделей по генетическим данным
- Глава 3. Метод автоматического перебора расширенных моделей с разным числом параметров и настройки параметров по генетическим данным одной, двух и трех популяций
- Глава 4. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным и расширение библиотек *stdpopsim* и *demes*

Глава 1. Обзор предметной области

- Демографическая история популяций
- Методы вывода демографической истории популяций по генетическим данным
- Методы моделирования демографической истории популяций
- Методы и программные комплексы вычисления правдоподобия
- Методы оптимизации для настройки параметров моделей
- Методы перебора моделей демографической истории

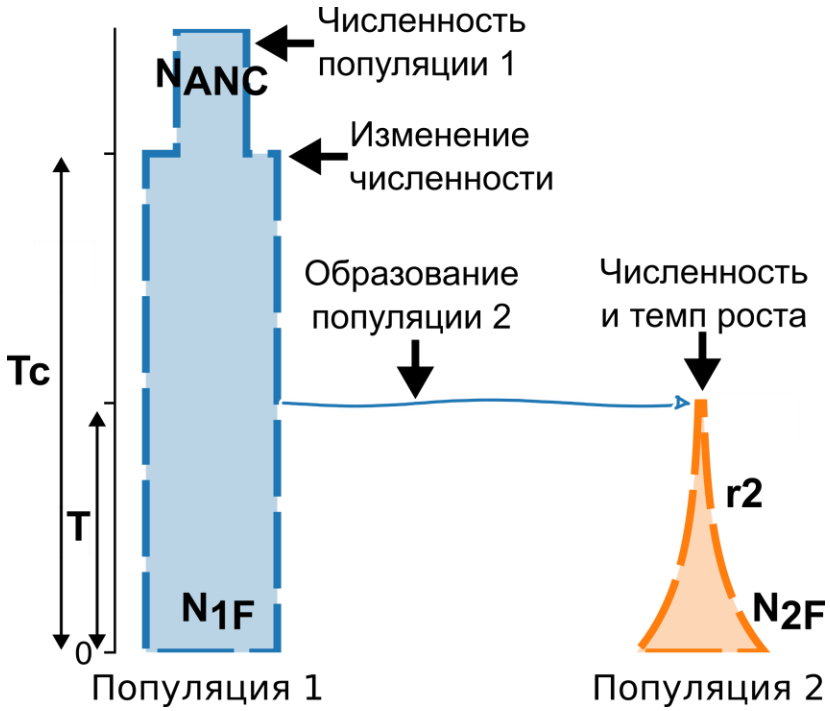
Методы моделирования демографической истории (1/4)

- Существующие программные решения используют модели двух классов
- Все эти модели имеют непрерывные параметры и фиксированные динамики изменения численности (константная численность или экспоненциальное изменение)



Модель I класса

∂a∂i, moments, momentsLD



Модель II класса

momi

Методы моделирования демографической истории (2/4)

Определение 2. Элемент временного интервала $\mathcal{I}$ — это шестерка $\langle p, T, \mathfrak{N}^{\text{start}}, \mathfrak{N}^{\text{end}}, \mathfrak{M}, \mathfrak{d} \rangle$, где
$p \in \mathbb{N}$ — число популяций, T — время продолжительности временного интервала,
$\mathfrak{N}^{\text{start}} = \{N_1^s, \ldots, N_p^s\}$ — численности каждой из популяций в начале временного интервала,
$\mathfrak{N}^{\text{end}} = \{N_1^e, \ldots, N_p^e\}$ — численности каждой из популяций в конце,
$\mathfrak{M} = \{m_{1,2}, m_{1,3}, \ldots, m_{p,(p-1)}\}$ — миграции между популяциями,
$\mathfrak{d} = \{d_1, \ldots, d_p\},\ d_i \in \{0, 1, 2\}$ — законы изменения численности.

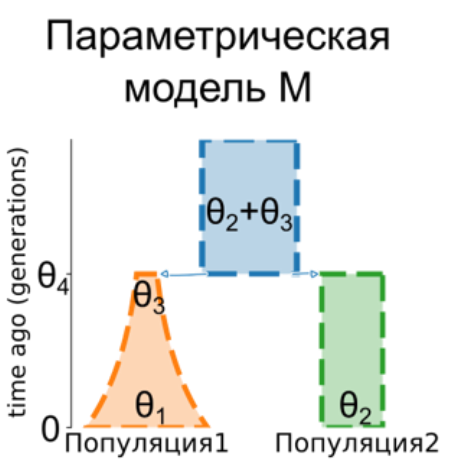
Определение 3. Элемент разделения $\mathcal{S}$ — это двойка $\langle p, i \rangle$, где p — число популяций до разделения, $i \in \{1, \ldots, p\}$ — индекс разделившейся популяции.

Определение 4. Элемент единичной миграции $\mathcal{A}$ — это тройка $\langle i^{\text{from}}, i^{\text{to}}, m \rangle$, где i^{from} — популяция-исток , i^{to} — популяция-сток, m — интенсивность единичной миграции.

Методы моделирования демографической истории (3/4)

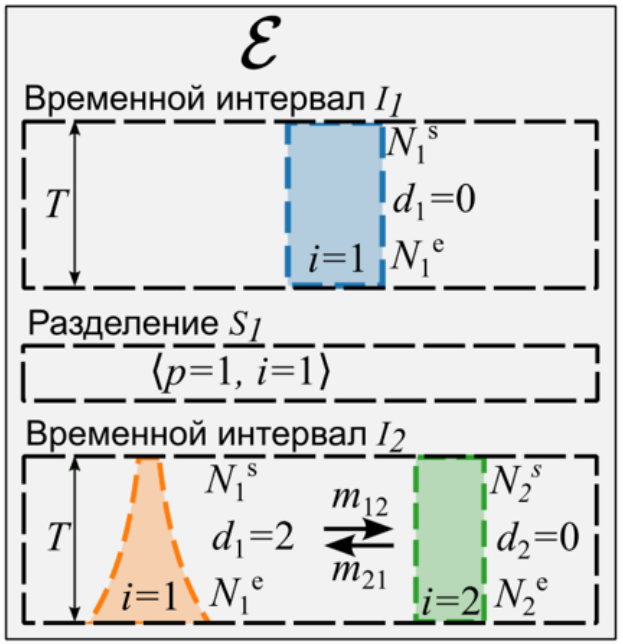
Определение 5. Характеристиками $\chi(\mathcal{I})$ временного интервала $\mathcal{I}$ называется множество $\{T, N_1^s, \ldots, N_p^s, N_1^e, \ldots, N_p^e, m_{1,2}, \ldots, m_{p,p-1}\}$.

Определение 6. Модель первого класса — параметрическая модель, которая представляется в виде тройки $\langle \Theta, \mathcal{E}, \mathfrak{F} \rangle$, где $\Theta \subset \mathbb{R}_+^d$ — множество значений непрерывных параметров модели, $\mathcal{E} = \{E_i\}_{i=1}^K$, $E_i \in \mathcal{I} \cup \mathcal{S}$ — последовательность элементов временных интервалов и разделений, $\mathfrak{F} : \Theta \to \cup\chi(E_i)$ — отображение параметров модели в характеристики элементов.



$$\Theta$$

$$\theta = (\theta_1, \theta_2, \theta_3, \theta_4) \in R_+^4$$



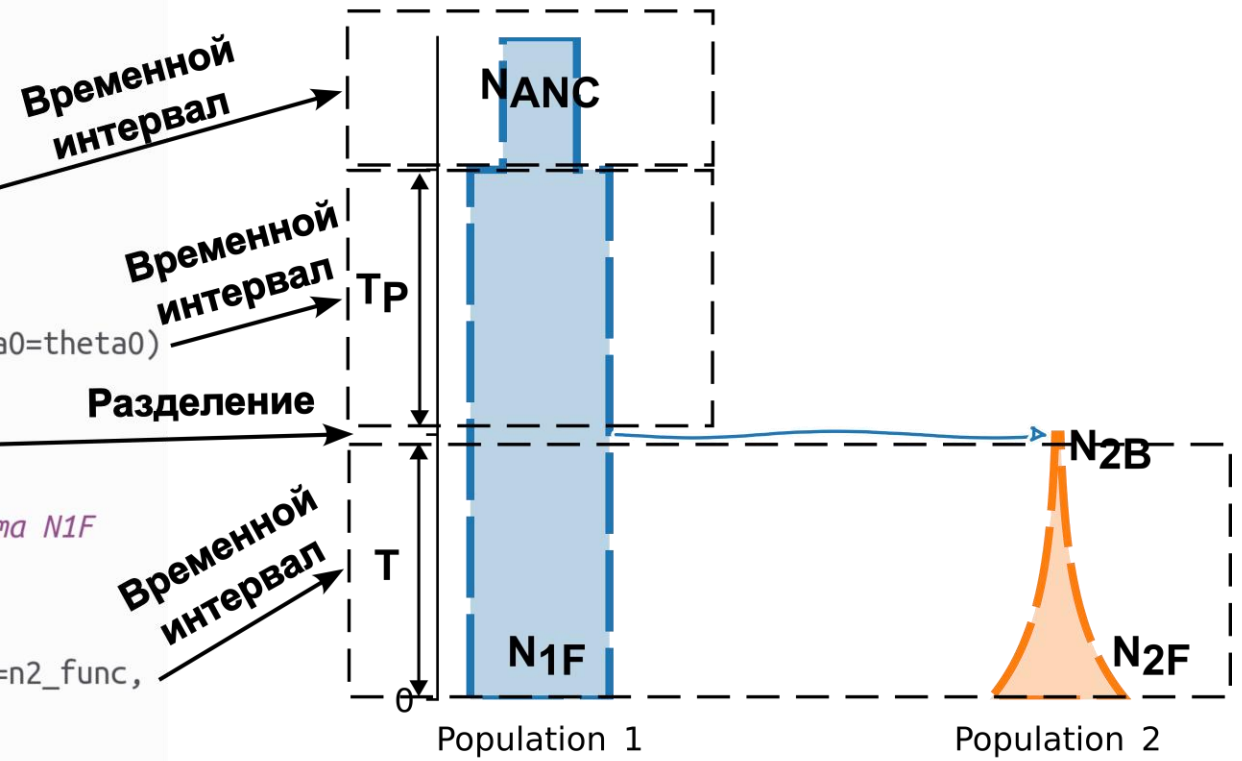
$$\mathfrak{F}$$

$$N_1^s(I_1) = N_1^e(I_1) = \theta_2 + \theta_3$$
$$T(I_1) = \infty$$
$$N_1^s(I_2) = \theta_3$$
$$N_1^s(I_2) = \theta_1$$
$$N_2^s(I_2) = N_2^e(I_2) = \theta_2$$
$$T(I_2) = \theta_4$$
$$m_{12}(I_2) = 0$$
$$m_{21}(I_2) = 0$$
$$F_1 = 0$$
$$F_2 = 0$$

Методы моделирования демографической истории (4/4)

Спецификация моделей I класса с использованием *dadi*

```python
import dadi

def model(params, ns, theta0, pts):
    Nanc, N1F, N2B, N2F, Tp, T = params

    # Задание сетки для численных вычислений
    xx = yy = dadi.Numerics.default_grid(pts)

    # Инициализация модели начальным размером популяции
    phi = dadi.PhiManip.phi_1D(xx, nu=Nanc, theta0=theta0)

    # Первый временной интервал
    # Функция изменения численности - консатнта N1F
    phi = dadi.Integration.one_pop(phi, xx, T=Tp, nu=N1F, theta0=theta0)

    # Второй элемент модели - разделение популяции
    phi = dadi.PhiManip.phi_1D_to_2D(xx, phi)

    # Функция изменения численности первой популяции - консатнта N1F
    # Задание функции изменения численности второй популяции
    n2_func = lambda t: N2B * (N2F / N2B) ** (t / T)
    # Третий элемент - второй временной интервал
    phi = dadi.Integration.two_pops(phi, xx, T=T, nu1=N1F, nu2=n2_func,
                                    theta0=theta0)

    # Вычисляем численными методами ожидаемую статистику данных
    sfs = dadi.Spectrum.from_phi(phi, ns, (xx,yy))
    return sfs
```



Методы сравнения моделей с разным числом параметров

- **Информационный критерий Акаике (AIC)**

$$\mathrm{AIC}(\mathcal{M}, \mathfrak{D}) = 2 \cdot k - 2 \cdot \log \mathcal{L}(\theta^* | \mathfrak{D}),$$

где k — количество параметров θ модели $\mathcal{M}$, $\mathcal{L}(\theta^* | \mathfrak{D})$ — максимальное значение функции правдоподобия модели $\mathcal{M}$ для данных $\mathfrak{D}$

- Байесовский информационный критерий (BIC)

$$\mathrm{BIC}(\mathcal{M}, \mathfrak{D}) = k \cdot \log(n) - 2 \cdot \log \mathcal{L}(\theta^* | \mathfrak{D}),$$

где n — размер выборки данных $\mathfrak{D}$

- Тест отношения правдоподобия (в случае вложенных моделей)

$$\lambda_{LRT} = 2(\log \mathcal{L}(\theta^*_{full}) - \log \mathcal{L}(\theta^*_{nested}))$$

- Модификация критерия Акаике в случае зависимых данных

$$\mathrm{CLAIC}(\mathcal{M}, \mathfrak{D}) = 2 \cdot tr(J(\theta^*) H^{-1}(\theta^*)) - 2 \cdot \log(\mathcal{L}(\theta^* | \mathfrak{D}))$$

Выводы по главе 1

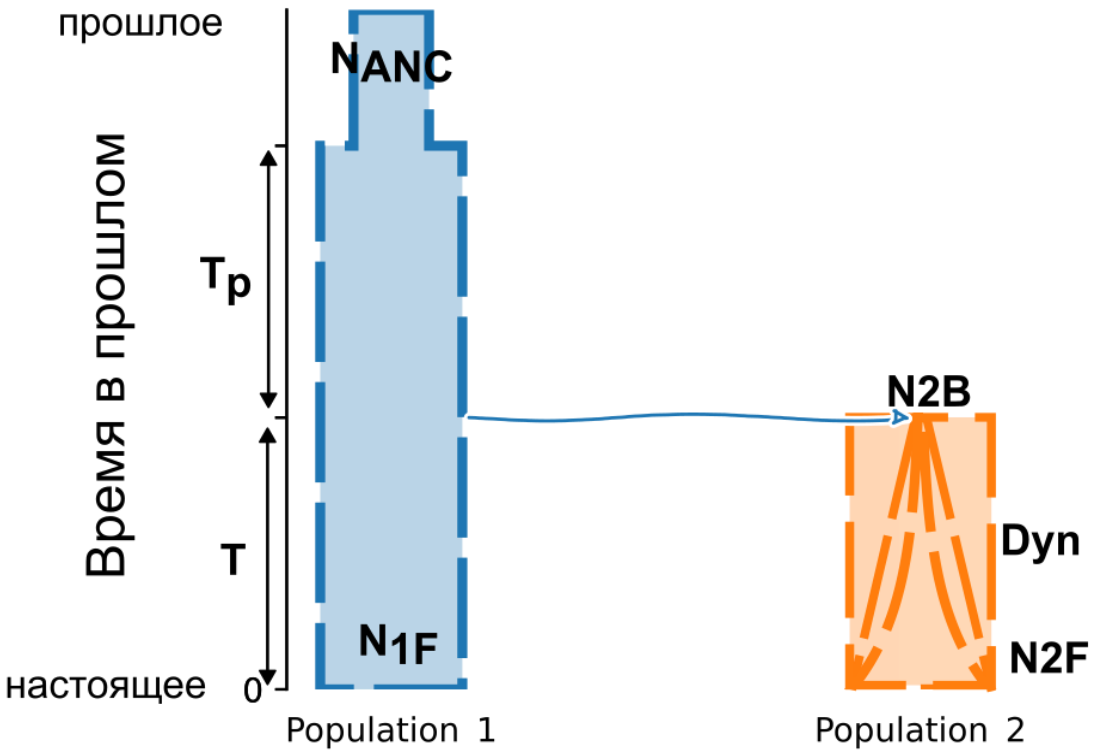
1. Проведен обзор существующих методов моделирования, вычисления правдоподобия и оптимизации для вывода демографических историй популяций
2. Сформулированы недостатки существующих методов:
   - применение экспертных данных при построении моделей
   - ограниченность методов локального поиска
   - ручной перебор моделей

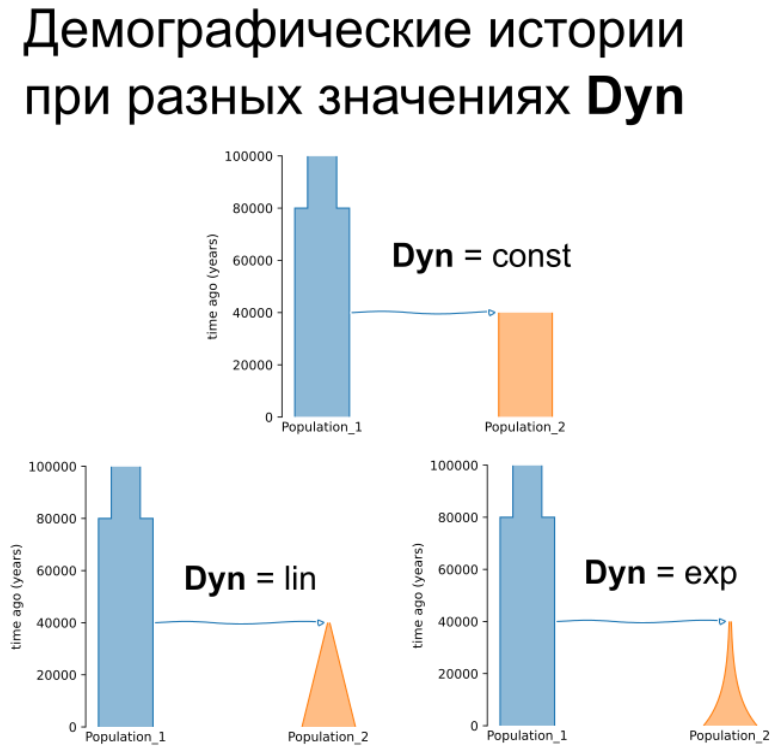
Глава 2. Модели расширенного класса и методы настройки параметров моделей по генетическим данным

- Результат 1. Расширенный класс моделей
- Результат 2. Методы настройки параметров моделей расширенного класса:
  - Результат 2.1. на основе генетического алгоритма
  - Результат 2.2. на основе байесовской оптимизации
- Экспериментальные исследования по выявлению эффективности разработанных моделей и методов на симулированных и реальных данных

Результат 1. Расширенный класс моделей (1/3)

- Расширенная модель включает динамики изменения численности (константная, линейная, экспоненциальная) как параметры для настройки
- Не требуется перебирать модели, отличающиеся динамиками вручную

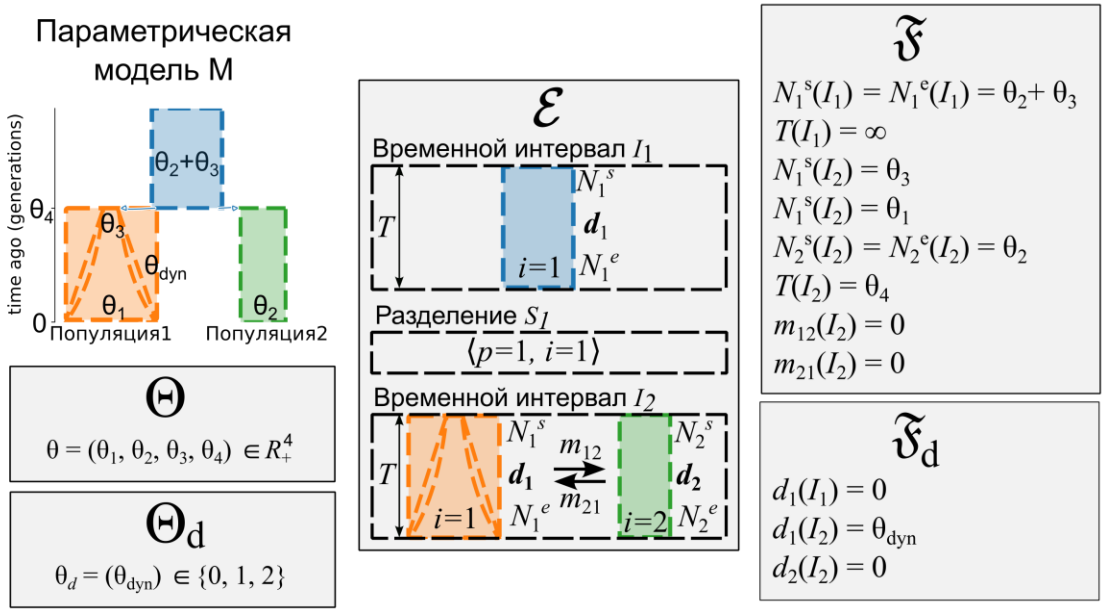




Результат 1. Расширенный класс моделей (2/3)

Определение 7. Динамическими характеристиками $\chi_{dyn}(\mathcal{I})$ интервала $\mathcal{I}$ называется множество $\{d_1, \ldots, d_p\}$.

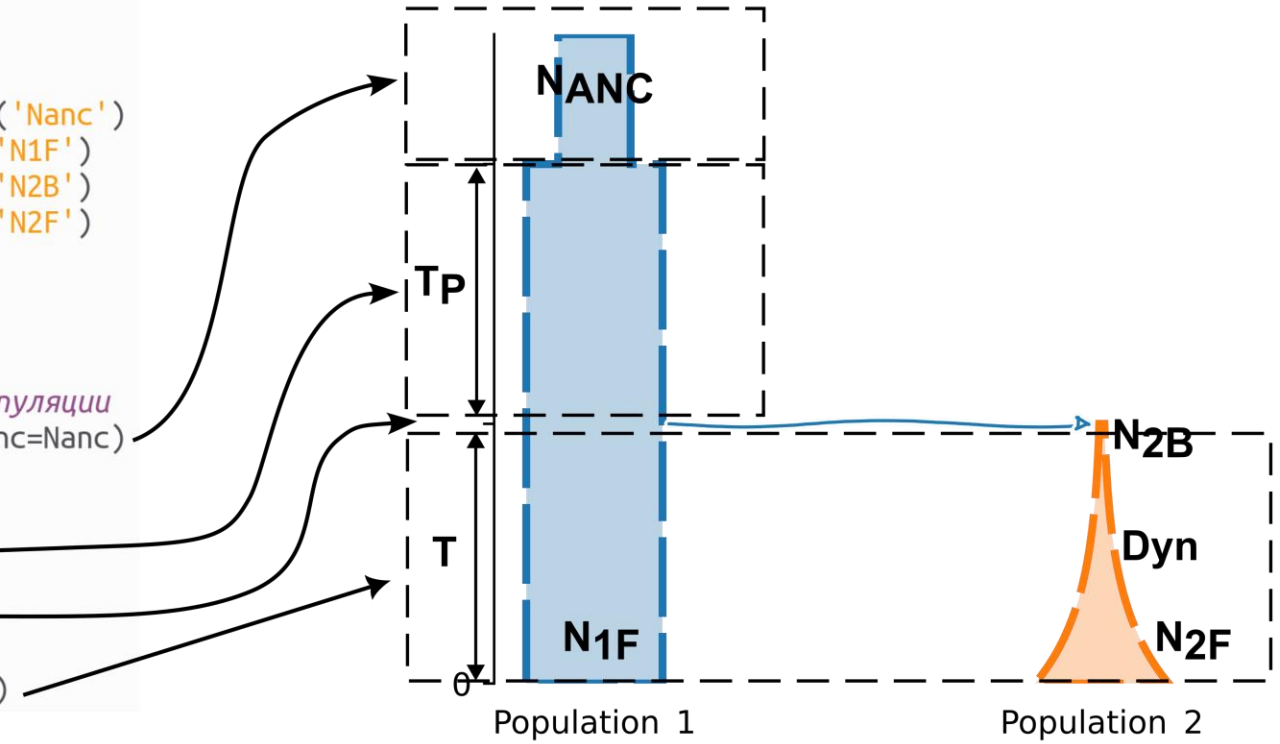
Определение 8. Модель расширенного класса — параметрическая модель, которая представляется в виде пятерки $\langle \Theta, \Theta_d, \mathcal{E}, \mathfrak{F}, \mathfrak{F}_d \rangle$, где $\Theta \subset \mathbb{R}_+^d$ — множество значений непрерывных параметров модели,
$\Theta_d \subset \{0, 1, 2\}^{k_2}$ — множество значений дискретных параметров динамики,
$\mathcal{E} = \{E_i\}_{i=1}^K$, $E_i \in \mathcal{I} \cup \mathcal{S}$ — последовательность элементов временных интервалов и разделений,
$\mathfrak{F} : \Theta \to \cup\chi(E_i)$ — отображение параметров модели в характеристики элементов,
$\mathfrak{F}_d : \Theta_d \to \cup\chi_{dyn}(E_i)$ — отображение дискретных параметров динамики в динамические характеристики элементов временных интервалов.



Результат 1. Расширенный класс моделей (3/3)

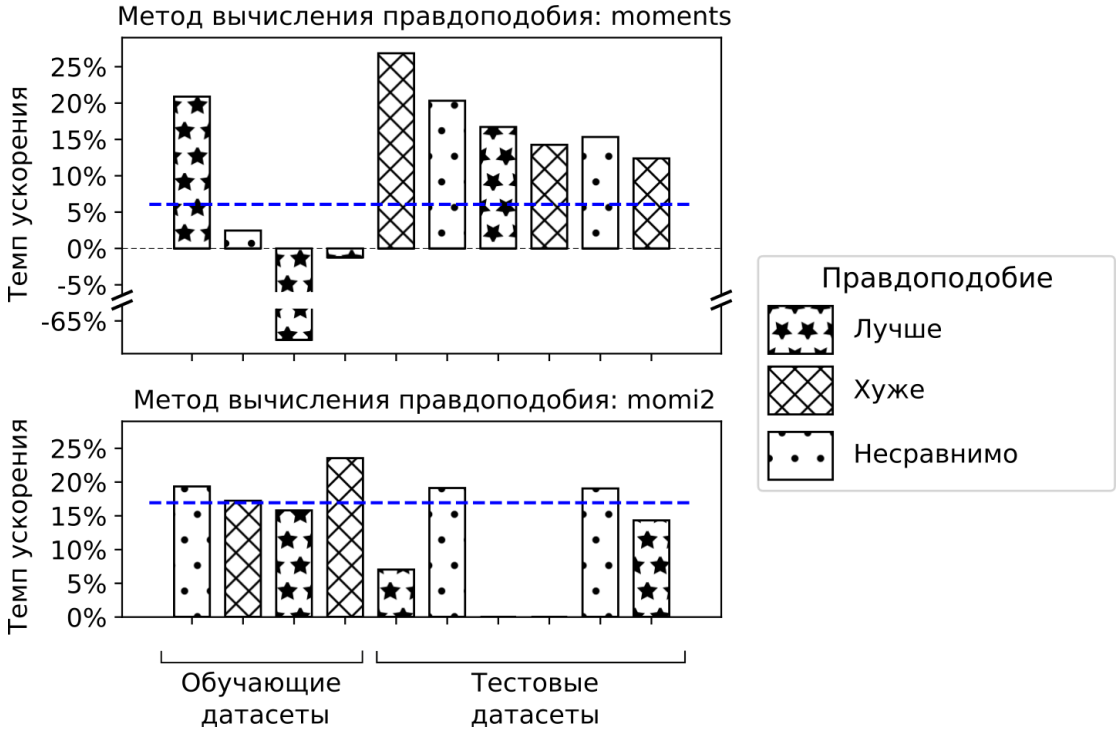
- Спецификация универсальна: можно один раз задать модель, а затем настроить параметры сначала для *dadi*, потом для *moments*, затем для *momentsLD*, а потом и для *momi2*

```python
import gadma

# Спецификация параметров модели
Nanc = gadma.variables.PopulationSizeVariable('Nanc')
N1F = gadma.variables.PopulationSizeVariable('N1F')
N2B = gadma.variables.PopulationSizeVariable('N2B')
N2F = gadma.variables.PopulationSizeVariable('N2F')
Tp = gadma.variables.TimeVariable('Tp')
T = gadma.variables.TimeVariable('T')
Dyn = gadma.variables.DynamicVariable('Dyn')

# Инициализация модели и размера предковой популяции
model = gadma.models.EpochDemographicModel(Nanc=Nanc)

# Добавление первого временного интервала
model.add_epoch(Tp, [nu1F])
# Добавление разделения популяции
model.add_split(0, [nu1F, nu2B])
# Добавление второго временного интервала
model.add_epoch(T, [nu1F, nu2F], ['Sud', Dyn])
```



Результат 2.1. Метод на основе генетического алгоритма
Особенности реализации

- Комбинация генетического алгоритма и метода локального поиска
- **Адаптивные сила и степень мутации**: при приближении к оптимуму более слабые изменения меньшего числа параметров
- **Гиперпараметры** генетического алгоритма были автоматически настроены (SMAC) для более эффективного решения поставленной задачи (ускорение 10%)



Результат 2.2. Метод на основе байесовской оптимизации

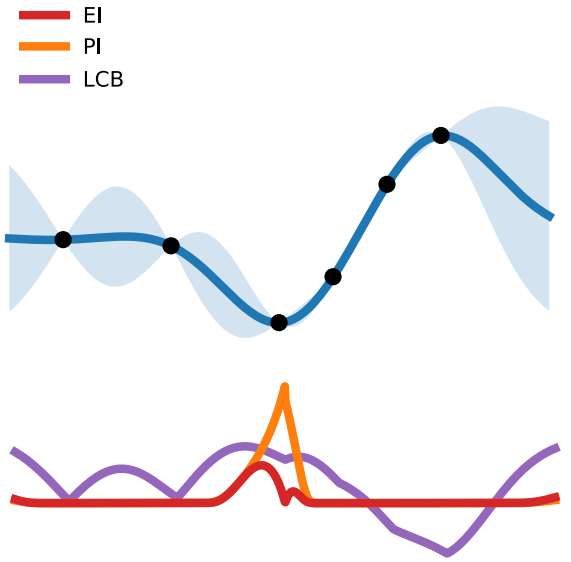
- Цель байесовской оптимизации: найти оптимум за минимальное число итераций
- Эффективна для оптимизации сложновычислимых функций
- Использует суррогатную модель для аппроксимации целевой функции
- Функция выбора позволяет выбрать новую точку для вычисления целевой функции

Пример работы байесовской оптимизации

Суррогатная модель
(гауссовский процесс)

Функция выбора

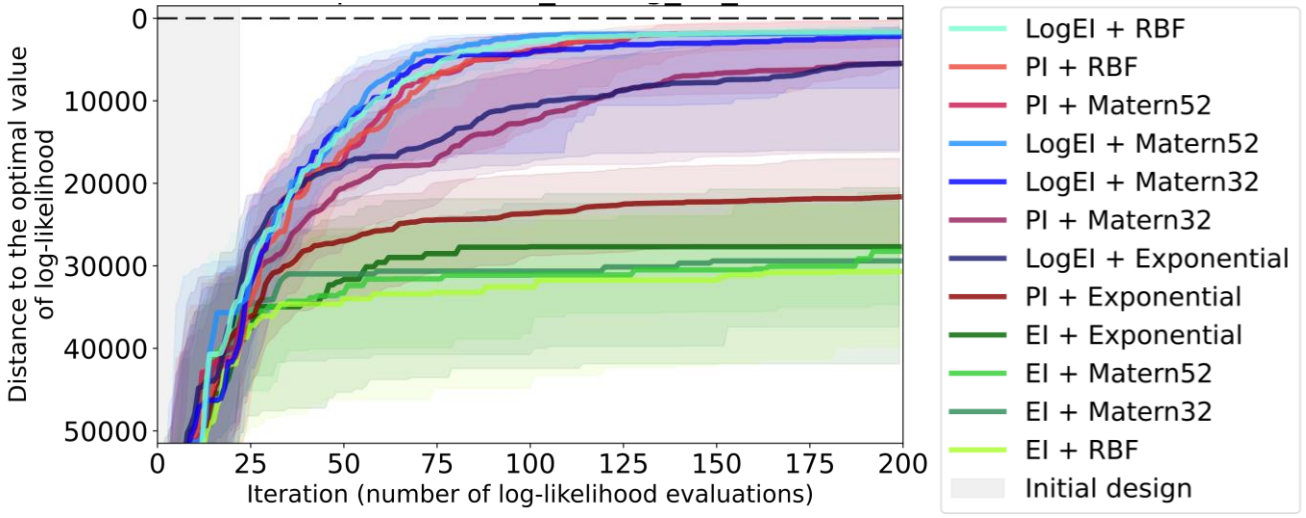
Примеры функций выбора



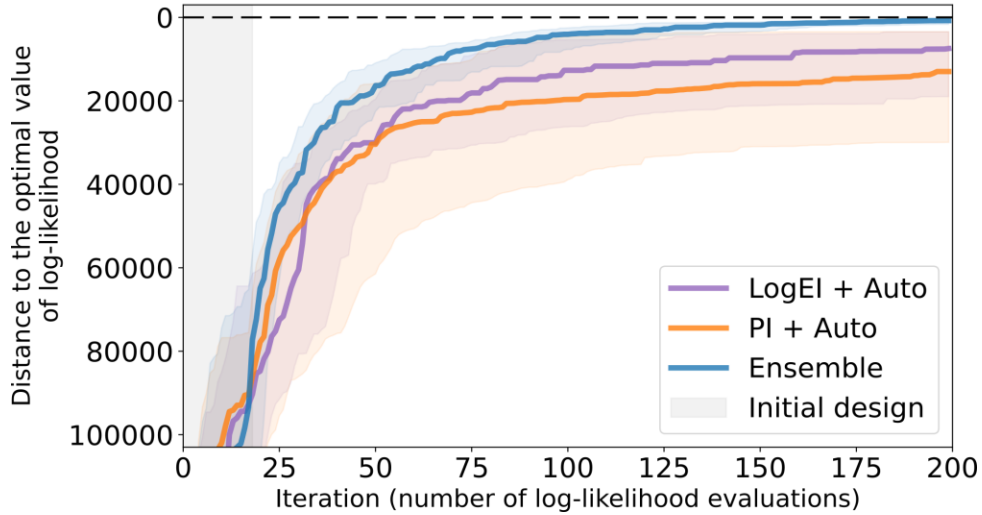
Результат 2.2. Метод на основе байесовской оптимизации
Особенности реализации

- Комбинация байесовской оптимизации и метода локального поиска
- **Гиперпараметры** байесовской оптимизации (функция выбора, функция ковариации) были вручную настроены для более эффективного решения поставленной задачи
- Разработан **ансамблевый метод** байесовской оптимизации (logEI, PI + Matern52, RBF)

Сравнение 12 вариантов байесовской оптимизации



Сравнение лучших вариантов с ансамблевым методом



Экспериментальные исследования

Метод на основе **генетического алгоритма**:

- Сравнение с существующими методами (*dadi, moments, dadi-pipeline, moments-pipeline*) на различных наборах данных:
  - Симулированные данные
  - Данные кошачьей лягушки
  - Данные американской пумы
  - Данные огородной капусты
- Вывод демографической истории с использованием модели расширенного класса
- Вывод демографической истории трех популяций современного человека на территории России с использованием модели расширенного класса

Метод на основе **байесовской оптимизации**:

- Сравнение с разработанным методом на основе генетического алгоритма
- Сравнение с *moments* на данных четырех и пяти популяций современного человека

Сравнение **генетического алгоритма** с существующими методами

- Сравнение с *moments* и *moments-pipeline* на симулированных данных:
  на 97% ближе к оптимуму в случае одной популяции, **на 60% ближе к оптимуму** в случае трех популяций
- Сравнение с *dadi-pipeline* на данных кошачьей лягушки (36 сценариев):
  92% логарифм правдоподобия лучше (33/36), 5% логарифм правдоподобия совпал (2/36)

Модель \ лучший logLL*	*dadi-pipeline*	**Метод GA**
d1_sec_contact_asym_mig_size	−445	**−407**
d1_anc_asym_mig_size	−522	**−499**
d2_anc_sym_mig_size	−600	**−550**

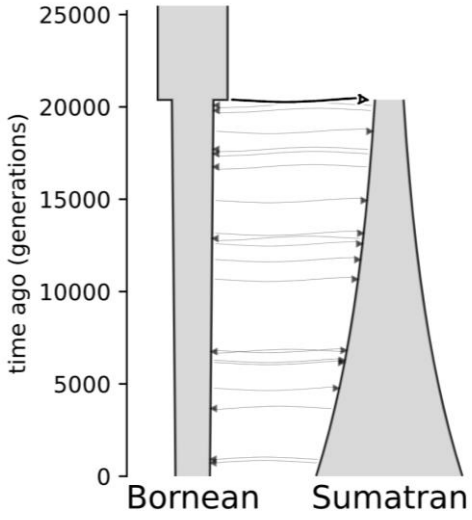
- Сравнение с *dadi* (методы BFGS и BOBYQA) на данных американской пумы: **в среднем лучше**

	Метод BFGS		Метод BOBYQA		**Метод GA 1 запуск**
	1 запуск	N запусков	1 запуск	N запусков	
Модель 1	$-1\,418\,712 \pm 2\cdot10^6$	$-455\,783 \pm 2\cdot10^4$	$-1\,015\,406 \pm 2\cdot10^6$	$-455\,684 \pm 2\cdot10^4$	**$-453\,000 \pm 53$**
Модель 2	$-1\,729\,870 \pm 4\cdot10^6$	**$-320\,947 \pm 5\cdot10^3$**	$-381\,979 \pm 1\cdot10^5$	$-320\,503 \pm 8\cdot10^3$	**$-319\,451 \pm 7\cdot10^3$**

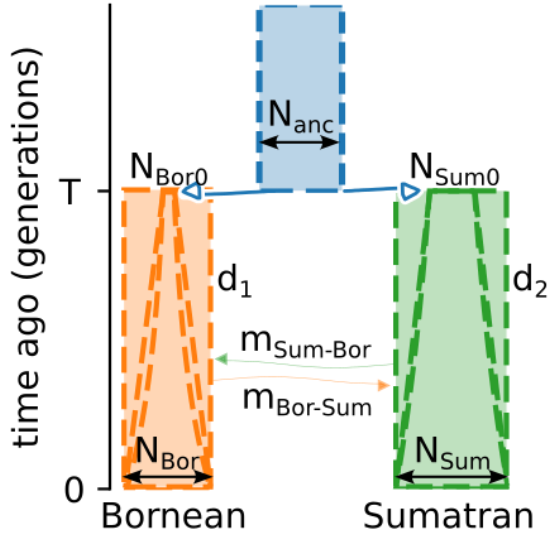
*logLL — логарифм правдоподобия

Вывод демографической истории с использованием модели расширенного класса

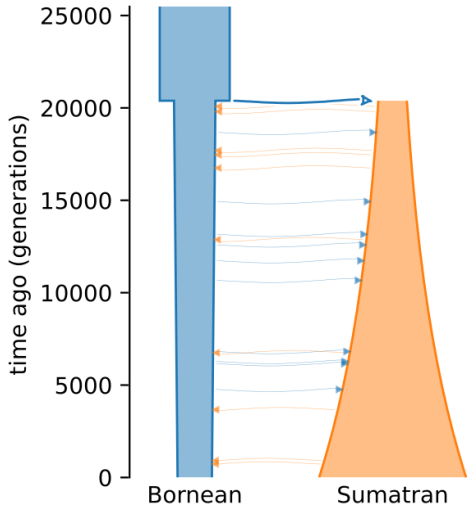
- Динамики в моделях расширенного класса настраиваются корректно



Истинная дем. история



Модель расширенного класса

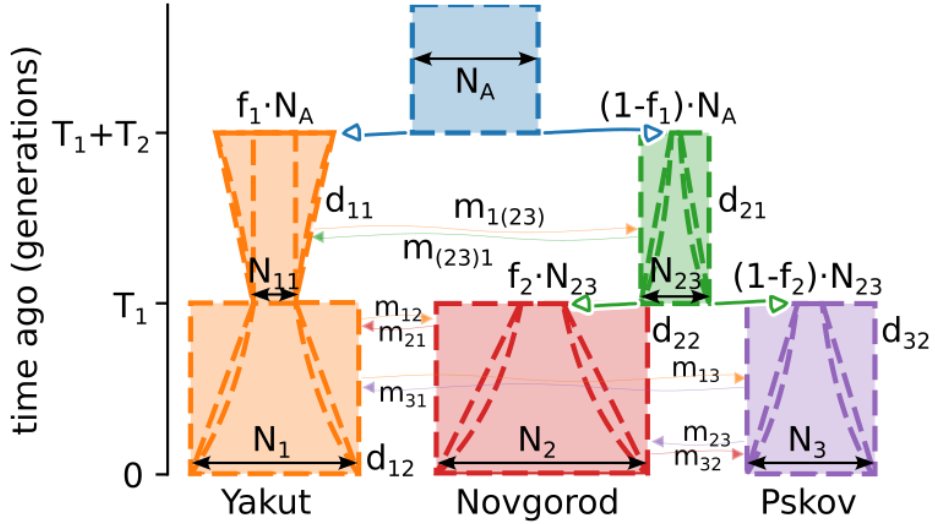


Настроенная модель

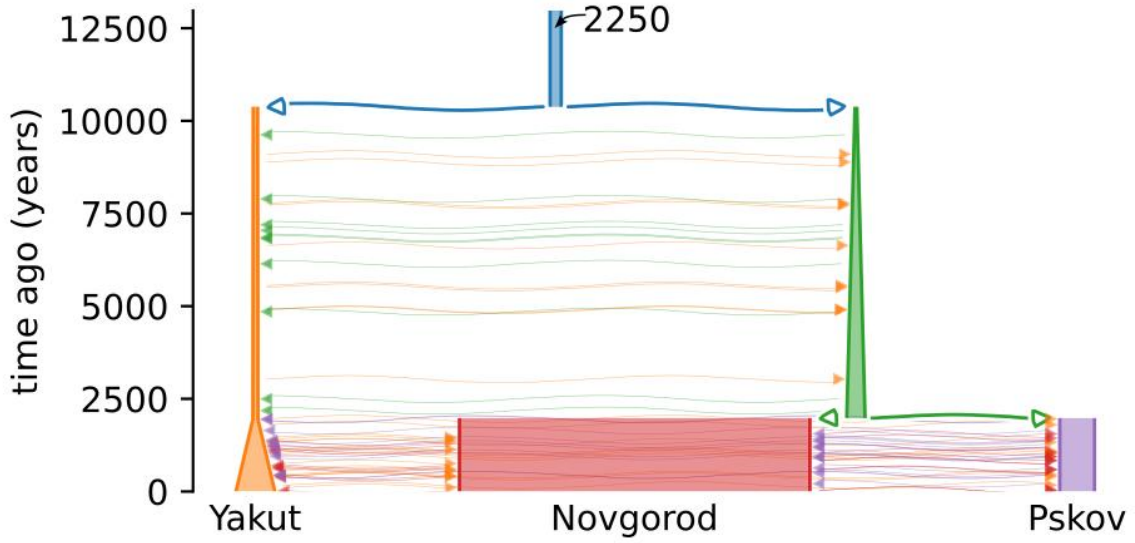
Демографическая история популяций современного человека на территории России

- Были проанализированы генетические данные популяций современного человека на территории Российской федерации:
  - жители территории Якутии (Yakut)
  - жители территории Новгорода (Novgorod)
  - жители территории Пскова (Pskov)

Zhernakova et al. 2020



Модель расширенного класса

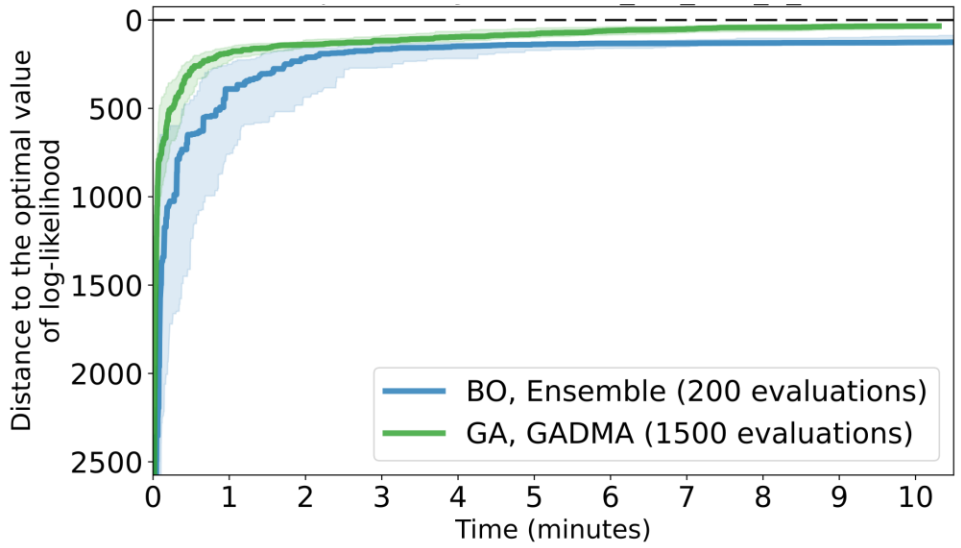


Полученная демографическая история

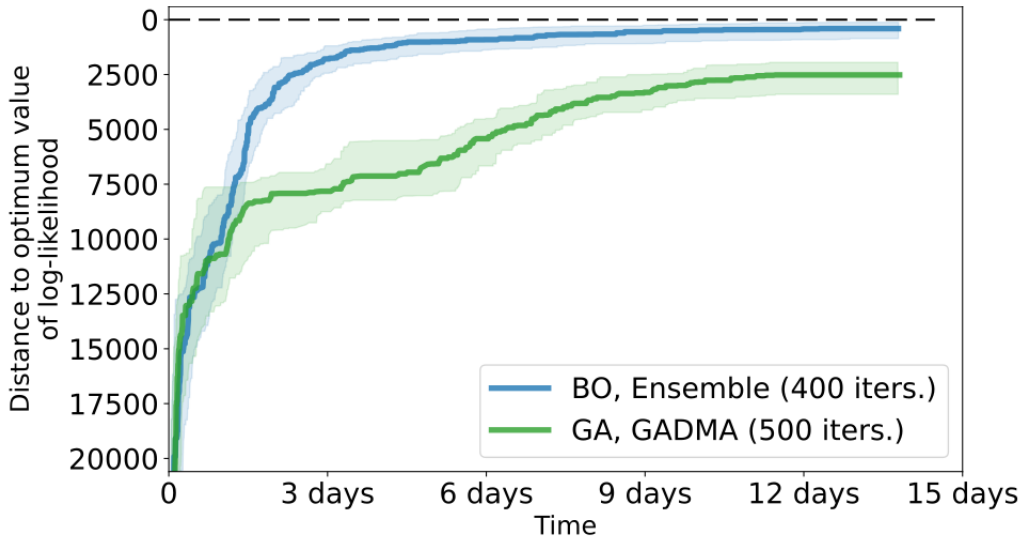
Сравнение **байесовской оптимизации** с генетическим алгоритмом

- Были построены графики сходимости генетического алгоритма и байесовской оптимизации на 13 наборах данных от одной до пяти популяций
- **В случае четырех и пяти популяций**, байесовская оптимизации показывает более быструю сходимость (50–80%), чем генетический алгоритм

Сравнение на данных двух популяций

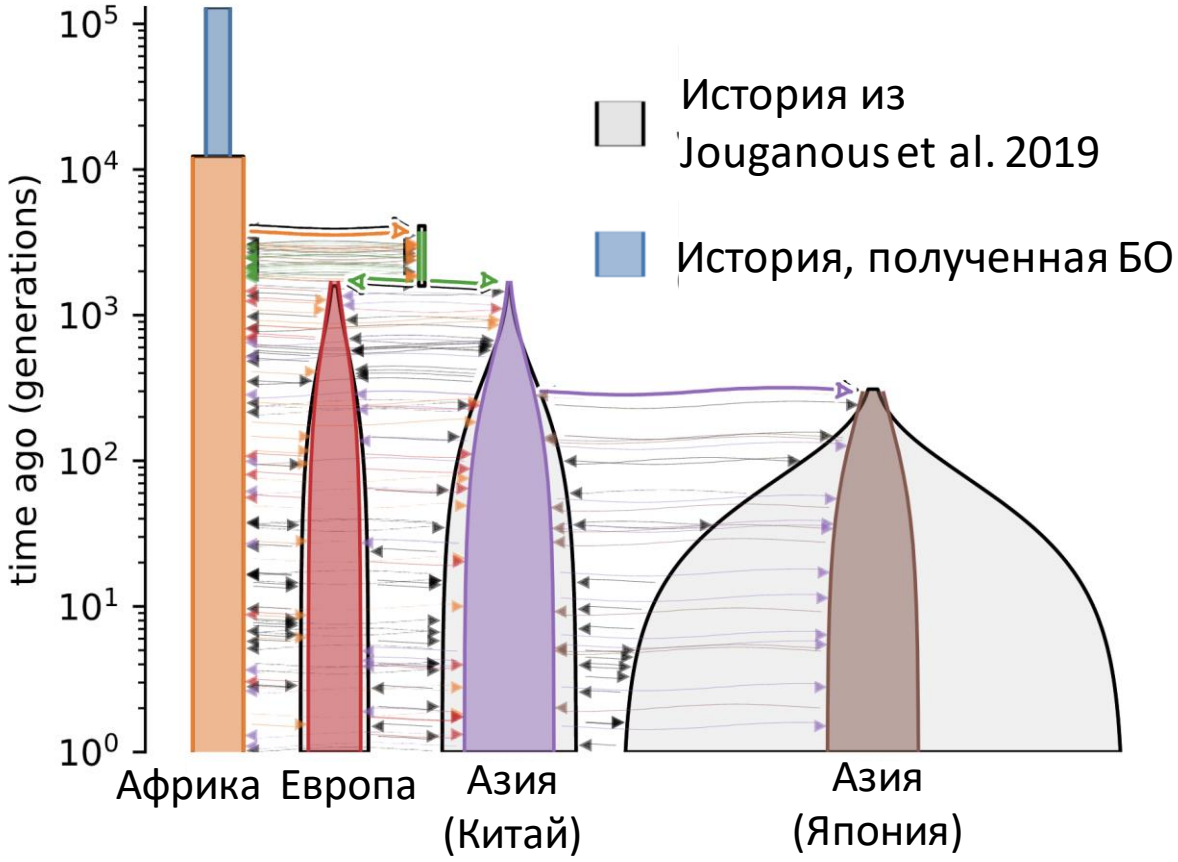


Сравнение на данных пяти популяций



Обновленная модель для популяций современного человека

- Сравнение с *moments* на данных четырех и пяти популяций современного человека: **лучшее значение правдоподобия**



Выводы по главе 2

- Разработан **расширенный класс моделей** демографических историй, который включает модели с дискретными параметрами динамики изменения численности популяции для настройки.
- Разработаны **два метода настройки моделей на основе комбинации глобального и локального поиска**. Гиперпараметры разработанных методов были настроены (ускорение на 10%).
- Экспериментальные исследования показали, что разработанные комбинированные методы настройки являются **более эффективными, чем существующие методы** (в >90% случаев).
- Генетический алгоритм демонстрирует быструю сходимость, чем байесовская оптимизация, при настройке параметров моделей **одной, двух и трех популяций**. Байесовская оптимизация оказывается более эффективной (50-80% быстрее) **в случае более трех популяций**.
- Полученные результаты подтверждают **положение 1**, выносимое на защиту.
- Результаты представлены в публикациях:

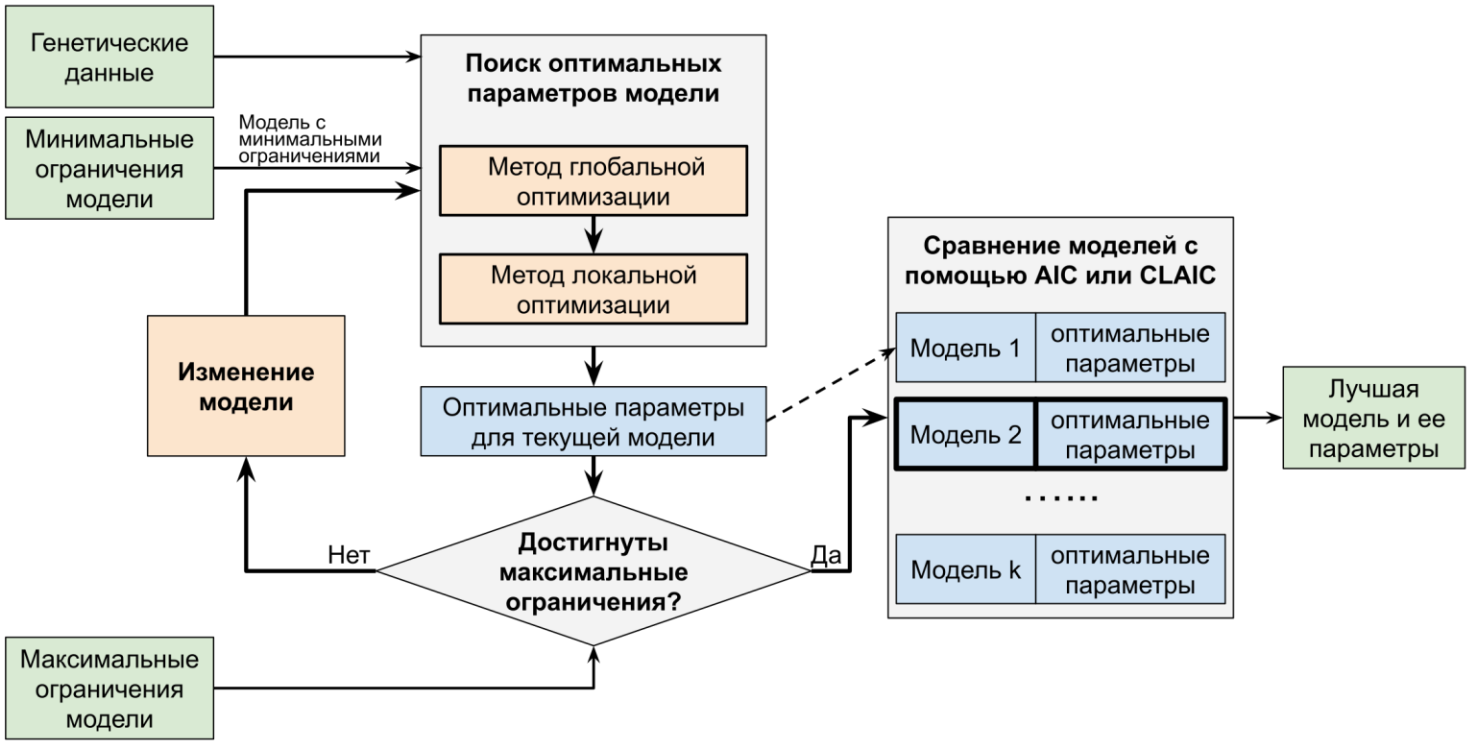
1. ***Noskova E.**, Ulyantsev V., Koepfli K.-P., O'Brien S.J., Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience **(Q1)**. — 2020.
2. ***Noskova E.**, Borovitskiy V.* Bayesian optimization for demographic inference // G3. — 2023.
3. ***Noskova E.**, Abramov N., Iliutkin S., Sidorin A., Dobrynin P., Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience **(Q1)**, — 2023.
4. *Zhernakova D. V., ..., Ulantsev V., **Noskova E.**, ..., O'Brien S. J.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. — 2020. — Т. 112, № 1. — С. 442–458.

Глава 3. Метод автоматического перебора моделей с разным числом параметров

- Результат 3. Метод автоматического перебора расширенных моделей с разным числом параметров для демографической истории одной, двух и трех популяций по генетическим данным
- Экспериментальные исследования по поиску демографической истории популяций с использованием разработанного метода

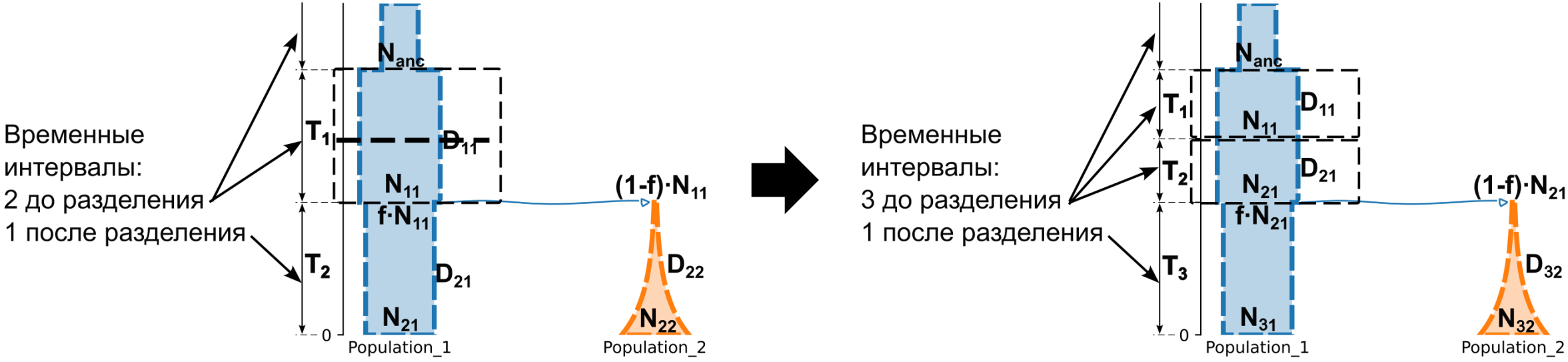
Метод автоматического перебора моделей с разным числом параметров (1/2)

- На вход метод принимает минимальные и максимальные ограничения на модель
- На каждой итерации происходит изменение модели, увеличение числа ее параметров и настройка модели разработанным методом на основе генетического алгоритма
- В результате работы выбирается модель с наилучшим значением метрики AIC



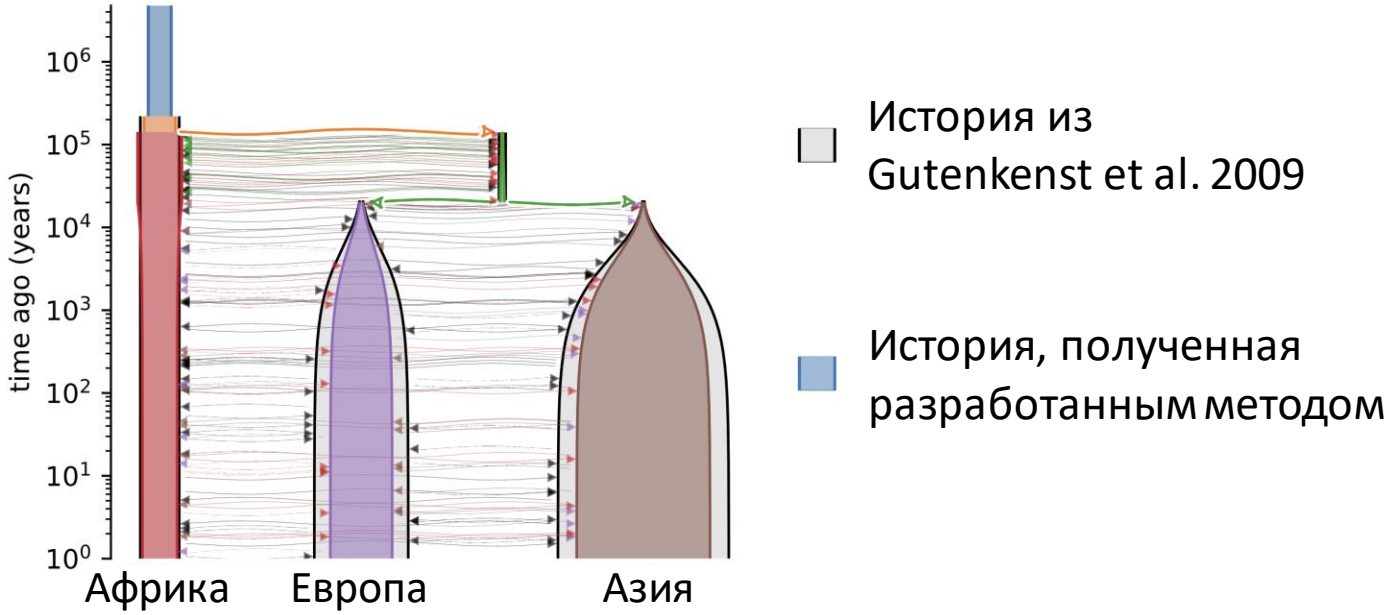
Метод автоматического перебора моделей с разным числом параметров (2/2)

- Минимальные и максимальные ограничения задаются числом временных интервалов в модели
- Процесс изменения модели: разделение выбранного интервала на два



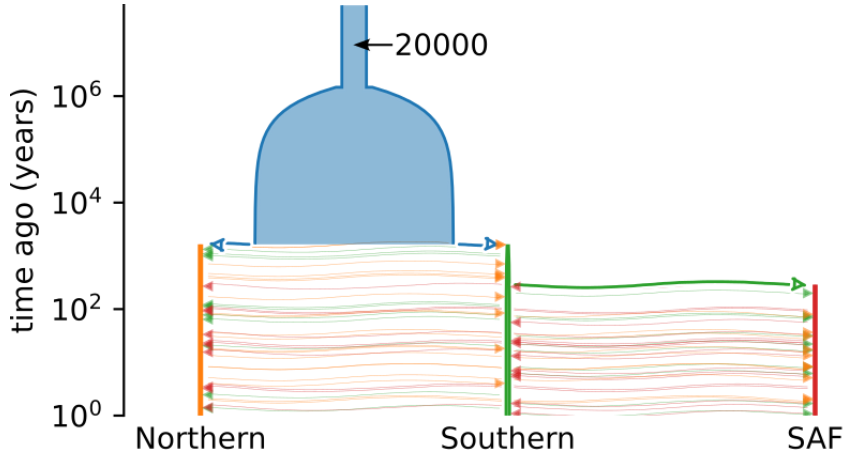
Результаты (1/2)

- Полученный метод позволил автоматически построить и настроить модели демографических историй для:
  - одного набора данных популяций современного человека (сравнение с *dadi*)
  - трех наборов данных популяций кошачьей лягушки (сравнение с *dadi-pipeline*)
- Полученные модели имеют **наилучшие значения AIC**, чем модели, построенные по экспертным данным или полученные существующими методами ручного перебора

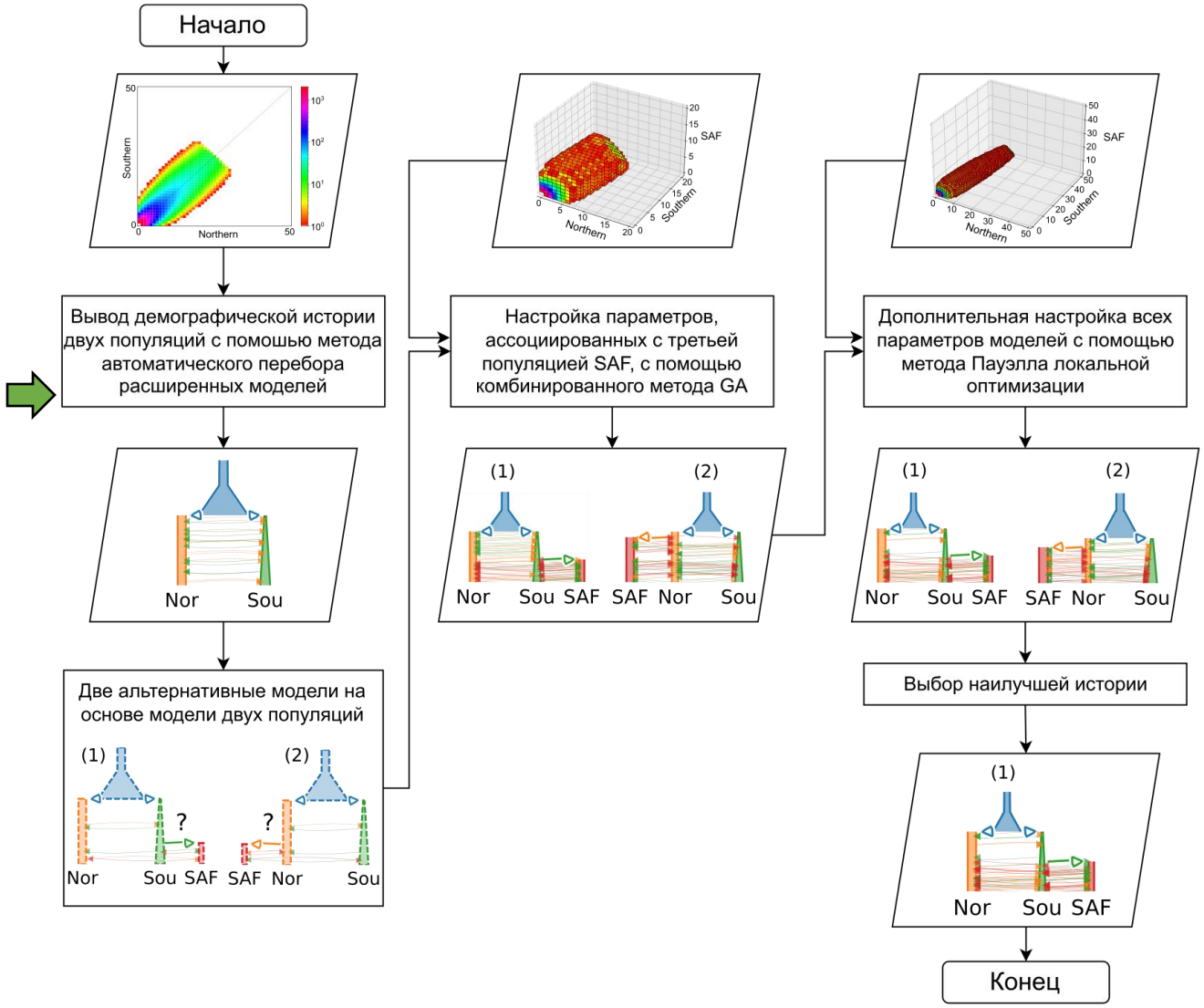


Результаты (2/2)

- Разработанный метод автоматического перебора моделей был применен для вывода демографической истории двух и трех популяций голубой акулы по данным, которые ранее не были проанализированы



Полученная демографическая история трех популяций



Выводы по главе 3

- Разработан **метод автоматического перебора** расширенных моделей демографической истории одной, двух и трех популяций по генетическим данным.
- Экспериментальные исследования подтвердили, что разработанный метод позволяет **построить модели с лучшим значением информационного критерия Акаике**, чем существующие методы ручного перебора.
- Выведена демографическая история трех популяций голубой акулы по данным, которые ранее не были проанализированы.
- Полученные результаты подтверждают **положение 2**, выносимое на защиту
- Результаты представлены в публикациях:

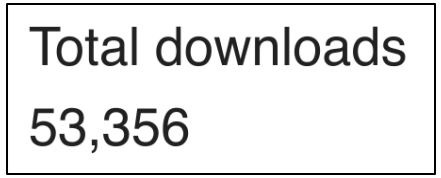
1. ***Noskova E.**, Ulyantsev V., Koepfli K.-P., O'Brien S.J., Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience **(Q1)**. — 2020.
2. *Nikolic N., ..., **Noskova E.**,* [*et al.*] Stepping up to genome scan allows stock differentiation in the worldwide distributed blue shark Prionace glauca // Molecular Ecology **(Q1)**. — 2023.

Глава 4. Программный комплекс GADMA для вывода демографической истории популяций по генетическим данным и расширение библиотек *stdpopsim* и *demes*

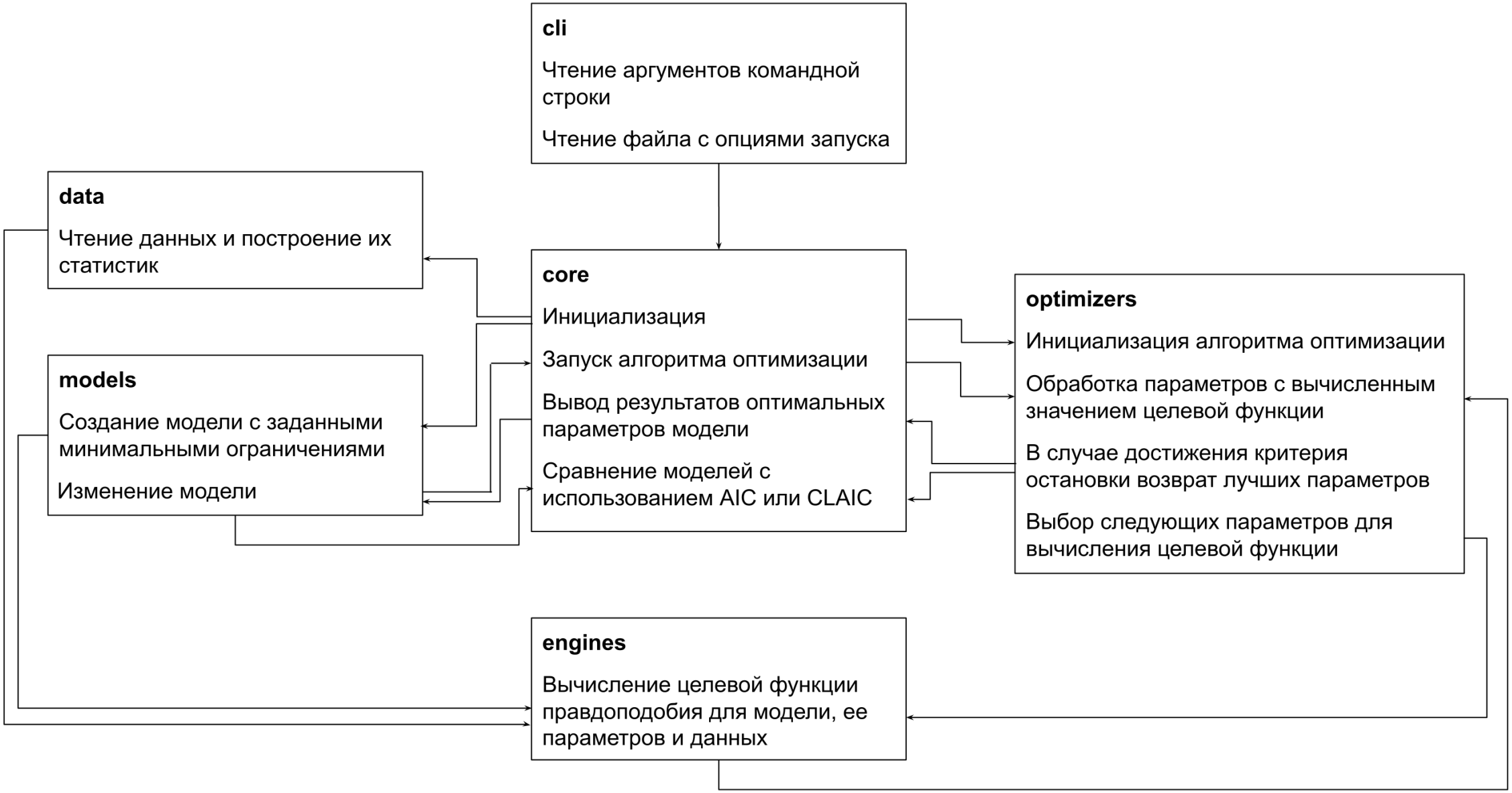
- Результат 4. Программный комплекс GADMA, реализующий все разработанные модели и методы
- Расширение библиотеки *stdpopsim* для симулирования генетических данных
- Расширение библиотеки *demes* для текстового и визуального представления демографических историй

Программный комплекс GADMA (1/3)

- GADMA (Global search Algorithm for Demographic Model Analysis)
- Выбор из четырех «движков» для вычисления функции правдоподобия: *dadi, moments, momentsLD, momi2*
- Включает интерфейс для спецификации расширенных моделей
- Выбор методов настройки параметров модели на основе:
  - Генетического алгоритма
  - Байесовской оптимизации
- Метод автоматического перебора моделей демографической истории
- Около 170 скачиваний в месяц

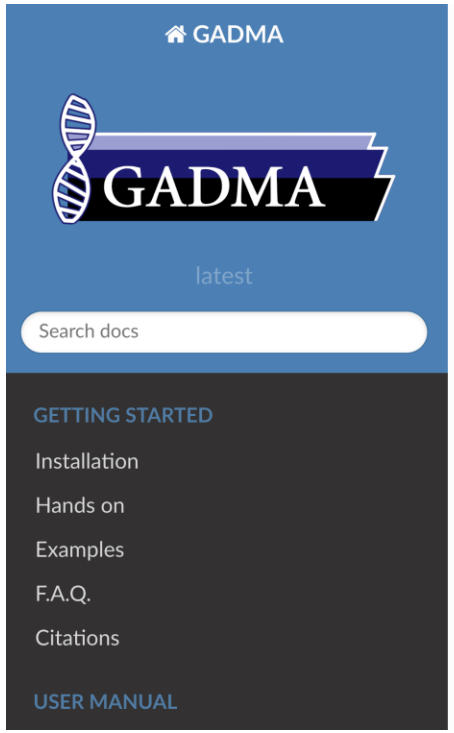


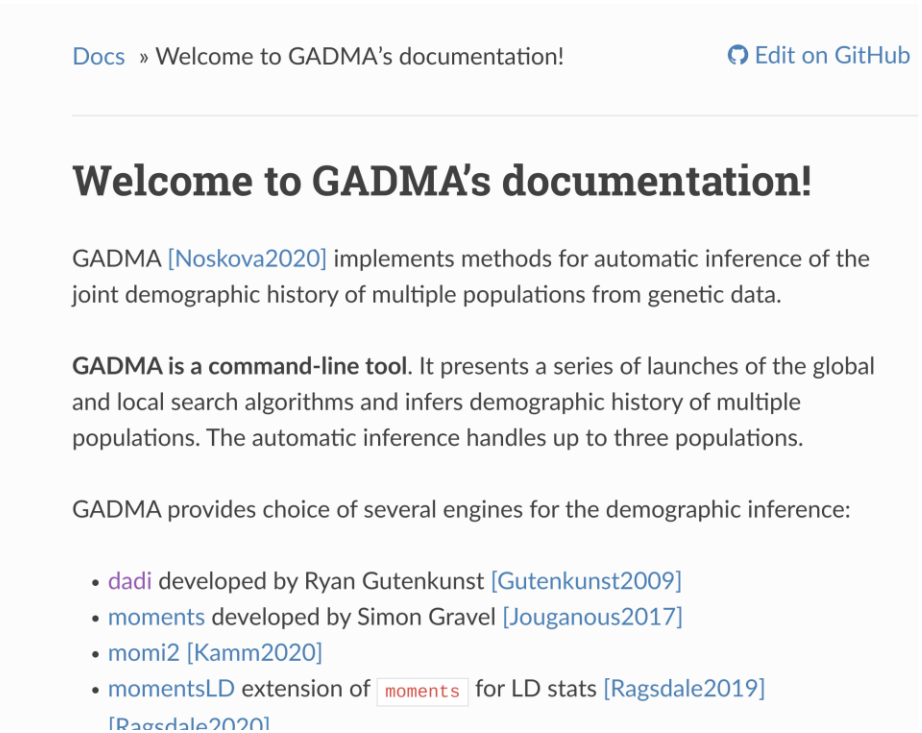
Программный комплекс GADMA (2/3)



Программный комплекс GADMA (3/3)

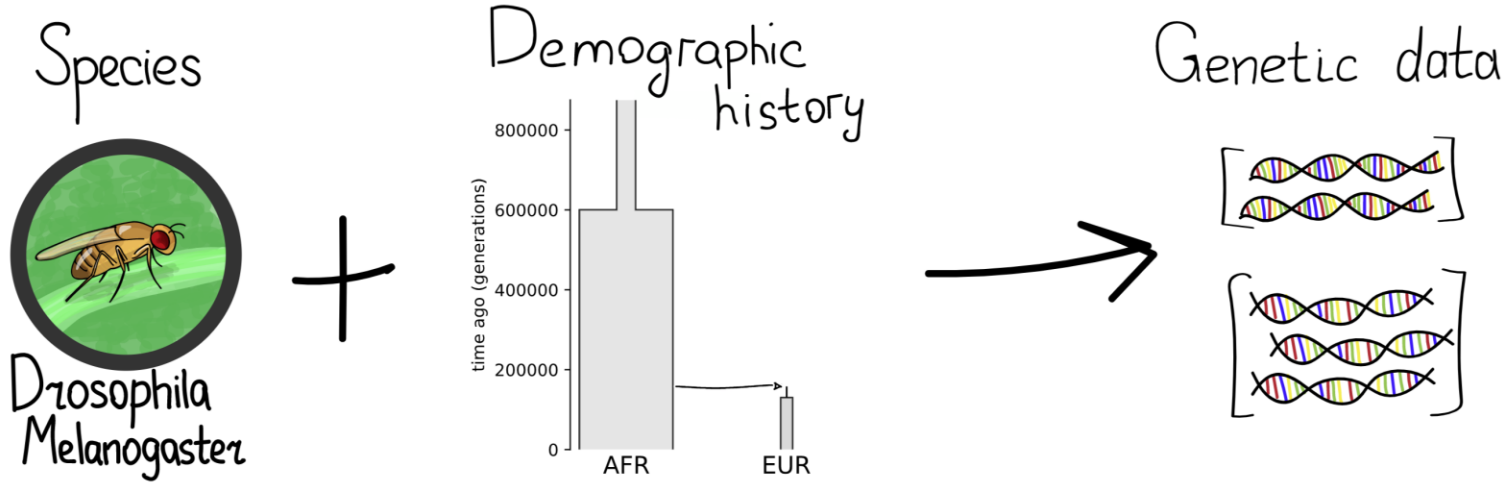
- Исходный код GADMA находится в открытом доступе на GitHub под лицензией GPLv3: https://github.com/ctlab/GADMA
- Общедоступная документация: https://gadma.readthedocs.io
- Система непрерывной интеграции GitHub Actions
- Автоматическое тестирование на различных платформах: Linux, Windows и macOS





Расширение библиотеки *stdpopsim*

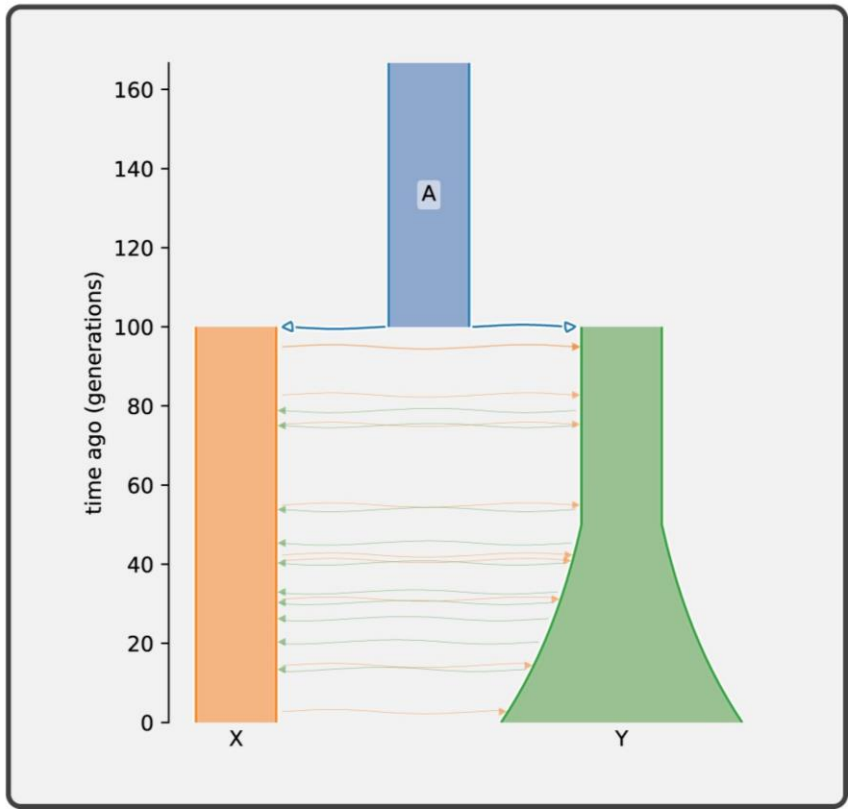
- Библиотека для симулирования генетических данных существующих видов
- Включает каталог видов, для каждого вида:
  - Информация о геноме (кол-во хромосом, длина хромосом)
  - Информация о виде (вероятность мутации, вероятность рекомбинации)
  - Набор опубликованных демографических историй
- **Расширение**: добавление видов, демографических историй, тестирование



Расширение библиотеки *demes*

- Библиотека текстового и визуального изображения демографической истории
- Все изображения демографических историй в работе получены с помощью *demes*
- **Расширение**: добавление линейной динамики, интеграция с GADMA

```yaml
# Comments start with a hash.
description:
  Two-deme isolation-with-migration model.
time_units: generations
defaults:
  epoch:
    start_size: 1000
demes:
  - name: A
    description: The ancestral deme
    epochs:
      - end_time: 100
  - name: X
    description: First descendant deme.
    ancestors: [A]
  - name: Y
    description: Second descendant deme.
    ancestors: [A]
    epochs:
      - end_time: 50
      - end_size: 3000
migrations:
  - demes: [X, Y]
    rate: 1e-4
```



Выводы по главе 4

- Описан разработанный программный комплекс GADMA для вывода демографической истории, реализующий разработанные модели и методы. Программный комплекс имеет репозиторий с открытым исходным кодом (https://github.com/ctlab/GADMA), общедоступную документацию и систему автоматического тестирования программного кода. Общее число скачиваний: >53 000
- Библиотека *stdpopsim* была расширена, протестирована и использована при проведении экспериментальных исследований.
- Библиотека *demes* была расширена добавлением линейной динамики изменения численности популяций и была интегрирована в программный комплекс GADMA. Все визуальные представления демографических историй, представленные в работе, получены с применением библиотеки *demes*.
- Результаты представлены в публикациях:

1. ***Noskova E.**, Ulyantsev V., Koepfli K.-P., O'Brien S. J., Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience **(Q1)**. — 2020.
2. ***Noskova E.**, Abramov N., Iliutkin S., Sidorin A., Dobrynin P., Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience **(Q1)**, — 2023.
3. *Adrion J. R., ..., Noskova E.,* [*et al.*] A community-maintained standard library of population genetic models // eLife **(Q1)**. — 2020.
4. *Lauterbur M. E., ..., Noskova E.* [*et al.*] Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations // eLife **(Q1)**, — 2023.
5. *Gower G., Ragsdale A. P., Bisschop G., Gutenkunst R. N., Hartfield M., Noskova E., Schiffels S., Struck T. J., Kelleher J., Thornton K. R.* Demes: a standard format for demographic models // Genetics **(Q1)**. — 2022.

Заключение (1/4)

На примере задачи вывода демографической истории популяций по генетическим данным:

- Разработан расширенный класс моделей метрических деревьев с функциями на ребрах, которые включают вид функций, как параметры для настройки
- Разработаны методы настройки параметров моделей метрических деревьев с функциями на ребрах на основе комбинации глобальной оптимизации и локального поиска
- Разработан метод автоматического перебора моделей метрических деревьев с функциями на ребрах
- Спроектирован и реализован программный комплекс GADMA, включающий разработанные модели и методы

Заключение (2/4): опубликованные работы

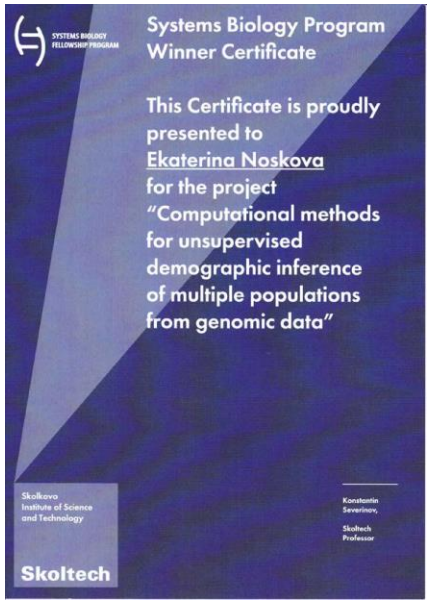
Scopus/WoS:

1. ***Noskova E.**, Ulyantsev V., Koepfli K.-P., O'Brien S. J., Dobrynin P.* GADMA: Genetic algorithm for inferring demographic history of multiple populations from allele frequency spectrum data // GigaScience **(Q1)**. — 2020.
2. ***Noskova E.**, Borovitskiy V.* Bayesian optimization for demographic inference // G3. — 2023.
3. ***Noskova E.**, Abramov N., Iliutkin S., Sidorin A., Dobrynin P., Ulyantsev V.* GADMA2: more efficient and flexible demographic inference from genetic data // GigaScience **(Q1)**, — 2023.
4. *Zhernakova D. V., ..., Ulantsev V.,* ***Noskova E.**, ..., O'Brien S. J.* Genome-wide sequence analyses of ethnic populations across Russia // Genomics. — 2020.
5. *Nikolic N., ...,* ***Noskova E.**,* [*et al.*] Stepping up to genome scan allows stock differentiation in the worldwide distributed blue shark Prionace glauca // Molecular Ecology **(Q1)**. — 2023.
6. *Adrion J. R., ...,* ***Noskova E.**,* [*et al.*] A community-maintained standard library of population genetic models // eLife **(Q1)**. — 2020.
7. *Lauterbur M. E., …,* ***Noskova E.*** [*et al.*] Expanding the stdpopsim species catalog, and lessons learned for realistic genome simulations // eLife **(Q1)**, — 2023.
8. *Gower G., Ragsdale A. P., Bisschop G., Gutenkunst R. N., Hartfield M.,* ***Noskova E.**, Schiffels S., Struck T. J., Kelleher J., Thornton K. R.* Demes: a standard format for demographic models // Genetics **(Q1)**. — 2022.

Заключение (3/4): награды

- Бронзовая награда в номинации 17th Human-Competitive Awards на онлайн конференции The Genetic and Evolutionary Computation Conference (GECCO) в 2020 году

- Победитель конкурсной программы поддержки исследовательских проектов System Biology Fellowship от Сколковского института науки и технологий по проекту «Computational methods for unsupervised demographic inference of multiple populations from genomic data» в 2021 году
  Число победителей — пять на всю страну в год



Заключение (4/4): аппробация результатов

1. Международный конгресс «VII съезд Вавиловского общества генетиков и селекционеров, посвященный 100-летию кафедры генетики СПбГУ, и ассоциированные симпозиумы», 2019, Санкт-Петербург, Россия
2. Moscow Conference on Computational Molecular Biology, 2019, Москва, Россия
3. Probabilistic Modeling in Genomics, 2019, Осуа, Франция
4. Probabilistic Modeling in Genomics, 2021, онлайн
5. Moscow Conference on Computational Molecular Biology, 2021, Москва, Россия
6. Probabilistic Techniques in Analysis, 2021, Сочи, Россия
7. Conservation Genomics at the Population Level, 2022, Кембридж, Великобритания
8. LI Научная и учебно-методическая конференция Университета ИТМО, 2022, Университет ИТМО, Санкт-Петербург, Россия
9. XI Конгресс молодых ученых, 2022, Университет ИТМО, Санкт-Петербург, Россия
10. Probabilistic Modeling in Genomics, 2022, Окфорд, Великобритация
11. XII Конгресс молодых ученых, 2023, Университет ИТМО, Санкт-Петербург, Россия
12. Probabilistic Modeling in Genomics, 2023, Колд Спринг Харбор, США
13. Society for Molecular Biology and Evolution Meeting (SMBE23), 2023, Феррара, Италия

Спасибо за внимание!

ITMO more than a UNIVERSITY