

Федеральное государственное бюджетное научное учреждение «Всероссийский научно-исследовательский институт сельскохозяйственной микробиологии»

На правах рукописи

Иголкина Анна Андреевна

**Реконструкция эволюционной истории нута с применением
моделирования сложных событий смещений и
композиционного анализа данных**

Специальность 1.5.8 —
«Математическая биология, биоинформатика»

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
доктор биологических наук
Андронов Евгений Евгеньевич

Санкт-Петербург — 2025

Оглавление

	Стр.
Введение	5
Глава 1. Обзор литературы	16
1.1 Генетическая структура и географическая организация популяций .	16
1.1.1 Проблема оценки частот аллелей в популяции	16
1.1.2 Методы анализа популяционной структуры без учета географического сигнала	19
1.1.3 Методы, связывающие пространственную и генетическую информацию	22
1.1.4 Методы оценки географических расстояний	24
1.2 Методы реконструкции эволюционной истории популяций	25
1.2.1 Описание задачи	25
1.2.2 Варианты параметризации событий смешения	27
1.3 Модели генетического дрейфа	32
1.3.1 Модель Райта-Фишера	32
1.3.2 Аппроксимация модели Райта-Фишера с помощью Бета-распределения	35
1.3.3 Аппроксимация модели Райта-Фишера с помощью нормального распределения	37
1.4 Анализ композиционных данных	38
1.5 Байесовский подход	44
1.6 Выводы к главе 1	48
Глава 2. Разработка моделей для реконструкции эволюционной истории популяций	50
2.1 Композиционная аппроксимация модели Райта-Фишера	50
2.1.1 Формальное описание входных данных	50
2.1.2 Вывод аппроксимации	51
2.2 Оценка характеристических частот аллелей и помощью модели popdisp	53
2.2.1 Байесовская иерархическая модель	54
2.2.2 Байесовский вывод апостериорных распределений	55
2.2.3 Производные апостериорных распределений по параметрам	58

2.2.4	Модификация popdisp модели на случай нескольких аллелей в локусе	60
2.2.5	Модификация popdisp модели с учетом неравновесия по сцеплению	61
2.2.6	Расчет ковариации при распространении популяций по географии	63
2.3	Реконструкция событий смешений с помощью модели migadmi . .	65
2.3.1	Формальное описание модели	67
2.3.2	Параметры в модели migadmi и их идентифицируемость . .	69
2.3.3	Оценка вкладов популяций-источников в смешанную популяцию с точки зрения дисперсии	71
2.3.4	Перебор топологий графов смешений	73
2.3.5	Сравнение migadmi с другими методами	74
2.4	Выводы к главе 2	75
Глава 3.	Исследования по нуту и исходные данные	78
3.1	Общие сведения о нуте	78
3.2	Обзор популяционных исследований нута	80
3.3	Исходные данные и обзор ранее проведенных на них исследований	82
Глава 4.	Результаты применения разработанных моделей	86
4.1	Популяционный анализ исходных данных	86
4.2	Модель popdisp	92
4.2.1	Тестирование на симуляциях	92
4.2.2	Результаты на реальных данных по нуту	96
4.3	Модель migadmi	100
4.3.1	Тестирование на симуляциях	100
4.3.2	Реконструкция эволюционной истории популяций нута дези	101
4.3.3	Реконструкция эволюционной истории популяций нута кабули	107
4.4	Сравнение результатов программы migadmi с другими методами . .	109
4.4.1	Результаты программы TreeMix	111
4.4.2	Результаты программы MixMapper	111
4.4.3	Результаты программы admixtools	113

	Стр.
4.5 Выводы к главе 4	115
Заключение	121
Список литературы	124
Словарь терминов	133
Список сокращений и условных обозначений	135
Список рисунков	136
Список таблиц	138

Введение

Актуальность

Доместикация (введение в культуру) растений представляет собой биологическую коэволюцию с участием человека, направленную на создание новых форм с улучшенными и полезными свойствами. Активная доместикация началась примерно 10–12 тысяч лет назад и продолжалась до 1-2-го тысячелетия до н.э. [1]. После первичного этапа окультуривания растений продолжалось их дальнейшее поддержание и улучшение в соответствии с потребностями человека: адаптация к разнообразным условиям окружающей среды, повышение урожайности и улучшение вкусовых качеств, неприхотливость в хранении и приспособленность к методам выращивания. Все это осуществлялось методами народной селекции. Исследования в области доместикации в первую очередь мотивированы экономическими интересами. Однако, помимо этого, они предоставляют ответы на фундаментальные вопросы культурно-исторического значения, такие как определение местоположения и времени начала доместикации. Научная ценность таких исследований заключается в том, что знания, полученные о введении в культуру одного вида растений, могут быть применены к другим благодаря сходству процессов.

На протяжении тысячелетий оседлые сообщества, а затем крестьяне и местные фермеры занимались доместикацией и улучшением растений. Такая примитивная или народная селекция сохранялась до начала XIX века, точки начала интенсивного развития методов научной селекции и системного подхода к улучшению растений. Эти достижения к середине XX века привели к так называемой "Зелёной революции"— периоду разработки и применения централизованных сельскохозяйственных программ, направленных на повышение производительности и глобальное обеспечение существования больших масс населения. Последствия этого масштабного перехода оказались неоднозначными: произошло сокращение генетического разнообразия сельскохозяйственных культур и массовое использование современных сортов, которые не редко экологически неустойчивы, уязвимы к изменениям климата и бедны питательными веществами [2]. Восстановлению разнообразия может помочь обращение к генетическому материалу, не затронутому научной селекцией и Зелёной революцией, например, к так называемым староместным сортам или сортам народной се-

лекции (англ. landraces) Такой генетический материал хранится в банках семян, крупнейшим из которых в России является коллекция генетических ресурсов растений, расположенная в Федеральном исследовательском центре — Всероссийском институте генетических ресурсов растений имени Н. И. Вавилова (ВИР). Основным источником староместных сортов стали многочисленные ботанико-агрономические экспедиции, проведённые в первой половине XX века. Для правильного внедрения этого материала в селекционную разработку при создании новых сортов важно знать их генетическое родство, а именно эволюционную историю формирования генетических пулов в регионах доместикации.

Одной из сельскохозяйственных культур в коллекции ВИР является нут, бобовое растение, служащее незаменимым источником высококачественного белка [3], и занимающее второе место в мире среди бобовых по производству зерна [4]. Ввиду значимости нута в сельском хозяйстве исследования его глобального генетического разнообразия сохраняют свою актуальность. Несмотря на то, что история окультуривания нута частично установлена с помощью археологических данных, в этом вопросе все еще присутствуют значительные пробелы. Во-первых, генетико-географические группы разнообразия нута описаны не полно. Во-вторых, остается неизвестной эволюционная история взаимосвязи между этими группами, включая потенциальные события смешения.

Для ответа на поставленные вопросы необходимо составить репрезентативную выборку мирового разнообразия нута. В нее должны войти староместные сорта собранные по всему ареалу распространения нута. Соответствующая этим требованиям выборка из 421 образца староместных сортов была ранее сформирована из коллекции ВИР [5]. Хотя географическое распределение этой выборки ожидаемо неравномерно, она достаточно полна для выделения генетико-географических групп нута.

Реконструкция эволюционной истории возникновения этих генетико-географических групп подразумевает построение графа смешений (англ. admixture graph). В то время как формирование новых групп в результате миграций обычно представимо в виде бинарного дерева, граф смешений является направленным ациклическим графом более сложной сетевидной структуры, которая учитывает дополнительные сценарии: события смешения могут быть множественными, вложенными и включать более двух популяций-источников. Существующие модели реконструкции графов смешений не охватывают все из

перечисленных сценариев, что делает актуальной необходимость разработки новых, более полных моделей.

Дополнительной сложностью в задачах реконструкции является необходимость индивидуального подхода к хромосомным блокам. В то время как на уровне общих закономерностей в рамках всего генома эволюционная история популяций может быть весьма однозначной, на уровне отдельных хромосомных блоков она может иметь отклонения. Это происходит из-за того, что разные участки ДНК могут подвергаться различным независимым воздействиям — например, естественному отбору или генетическому дрейфу — и соединяться или перестраиваться в процессе рекомбинации. В данном контексте особую важность приобретает концепция графа предковых рекомбинаций (англ. *ancestral recombination graph*). Эта структура позволяет проследить уникальную эволюционную историю каждого конкретного участка хромосомы, представляя её в виде бинарного дерева [6]. По аналогии с этим можно предположить, что каждому хромосомному блоку также может соответствовать свой уникальный граф смещений. Таким образом, имеет место задача реконструкции графов смещений в режиме скользящего окна вдоль хромосом.

Большинство методов реконструкции эволюционной истории популяций базируются на модели Райта-Фишера. Эта модель описывает изменения частот аллелей во времени в популяции фиксированного размера и основывается на предположениях о случайном скрещивании в отсутствие мутаций, миграций и естественного отбора. Модель Райта-Фишера является дискретной по времени, поэтому для в методах реконструкции она используется в непрерывной аппроксимации. Все известные аппроксимации предполагают, что каждый генетический локус имеет два аллельных варианта, и описывают изменение частоты только одного аллеля, считая второй зависимым. Во-первых, такой подход является несимметричным. Во-вторых, он не является масштабируемым на случай мультиаллельных локусов. Поэтому необходимо разработать аппроксимации модели Райта-Фишера, которые обходят эти ограничения.

Частоты аллелей в популяции принадлежат к композиционному типу данных. Композиционные данные представляют собой наборы значений определенной длины (так называемые композиции), которые отражают относительные частоты, доли или проценты компонентов внутри фиксированного общего объема или суммы. Например, композициями являются все наборы длины 3 следующего вида: $x = (x_1, x_2, x_3) : x_1 + x_2 + x_3 = 1, x_i > 0$. Все элементы в ком-

позиции не являются независимыми: увеличение одного элемента композиции влечет уменьшение всех остальных. В связи с этим многие стандартные методы статистического анализа (например, выборочная корреляция Пирсона) являются неприменимыми и дают ложные результаты. Поэтому для работы с композиционными данными была предложена теория композиционного анализа данных [7; 8]. Эту теорию начали разрабатывать в середине XX века, и она хорошо себя зарекомендовала в геологии, экологии, химии, микробиологии и других дисциплинах. Использование композиционного анализа данных в популяционной генетике пока не распространено, несмотря на то, что эта область оперирует частотами аллелей, что напрямую относится к композиционным данным.

Степень разработанности темы исследования

Существует ряд исследований популяционной структуры нута, крупнейшим из которых является работа на 3366 образцах, включая 2436 староместных сортов и 197 дикорастущих образцов [9]. В этом исследовании все образцы были генотипированы по однонуклеотидным полиморфизмам (ОНП, англ. single nucleotide polymorphism, SNP) и распределены на семь групп. По части анализа этих групп работа имеет ряд недостатков. Во-первых, группирование было сделано исключительно по географическому принципу, без учета генетических сходств и различий. Во-вторых, не были получены генетические характеристики для групп - вектора характеристических частот аллелей ОНП. В-третьих, гипотезы о распространении популяций нута были сформулированы только на основании визуального анализа представления генетических данных в двух первых главных компонентах. Следовательно, более детальный анализ популяций нута необходим. Наилучшим способом подтверждения или опровержения наличия выделенных групп и сформулированных гипотез распространения является независимый анализ на отдельной, но сравнимой по структуре выборке. Отобранная из коллекции ВИР и генотипированная по ОНП панель из 421 староместного сорта нута как раз является таковой. Серьезным улучшением в понимании популяционной структуры нута станет детальный анализ популяций и получение их характеристик на базе отобранной панели как на независимом источнике, а предыдущие исследования будут использоваться для сравнения результатов на качественном уровне.

В отсутствие четко определенных генетико-географических групп нута ряд вопросов, касающихся их происхождения, может оставаться нераскрытым и даже несформулированным. Несмотря на неполноту описания популяций нута, для

одной из них была продемонстрирована географическая и генетическая обособленность — это относится к нуту дези из Эфиопского региона. В исследовании [9] Эфиопская популяция занимает изолированное положение на графике первых двух главных компонент, располагаясь как бы в центре между всеми остальными популяциями (Рисунок 3.1). Так как Эфиопская популяция не является центром первичной доместики [10], такое ее положение потенциально может свидетельствовать о сложном и смешанном происхождении. Этот результат был интерпретирован как то, что Эфиопская популяция произошла из Ближневосточного или Среднеазиатского региона. Однако вопрос о том, произошла ли эта популяция в результате событий смешения или просто в результате миграции, остается открытым.

Существуют две основных культивируемых группы сортов нута, отличающихся по внешнему виду, вкусу и другим характеристикам: дези и кабули. Место первичной доместики дези было установлено с помощью археологических находок и подтверждено совпадением с ареалом произрастания дикого типа, который фенотипически схож с дези. Регион первичной доместики кабули остается неизвестным. В литературе были выдвинуты предположения, что он связан с Кабулом, исходя из лингвистического сходства между словами "кабули" и "Кабул" [11]. Однако попыток количественных подтверждений или опровержений этой гипотезы не было предпринято.

Существуют более точные методы выделения генетических групп в наборе образцов, чем простой анализ главных компонент. К таким методам относятся программы типа STRUCTURE и ADMIXTURE [12; 13]. Принцип их работы похож на метод кластеризации k-средних, поэтому эти методы косвенно определяют векторы частот аллелей ОНП для каждой группы. Однако ни один из существующих методов выделения генетических групп в данных не учитывает их географическую неравномерность. Следовательно, требуется разработка метода оценки генетических характеристик групп образцов с учетом их географического распределения.

Для решения задачи реконструкции графа смешений для выделенных групп существует несколько методов. Различия между ними заключаются в подходе к моделированию событий смешения: асимметричная для популяций-источников модель [14], модель с ограниченным числом вложенных событий смешения [15] и модель смешения с не более чем двумя источниками [16]. Не существует единого метода, который бы использовал симметричную модель с более сложными собы-

тиями смещения, допускающий участие более чем двух популяций-источников. Таким образом, существует потребность в разработке нового усовершенствованного метода.

Существует еще несколько возможных усовершенствований, которые ранее не применялись в методах реконструкции эволюционной истории популяций. Во-первых, как уже упоминалось выше, целесообразно оценивать граф смещений в режиме скользящего окна. Во-вторых, использование композиционного подхода к аппроксимации модели Райта-Фишера вместо стандартной нормальной аппроксимации позволит не только более точно учитывать природу частот, но и перейти от моделирования биаллельных локусов к мультиаллельным.

Цель данной работы заключается в установлении происхождения популяций нута в результате разработки и применения новых методов для реконструкции эволюционной истории.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Предложить аппроксимацию модели Райта-Фишера с учетом композиционной природы частот аллелей.
2. Разработать модель для оценки характеристических частот аллелей в популяции с учетом ковариационных связей между образцами, которые определяются их географическим местоположением.
3. Разработать модель для реконструкции эволюционной истории популяций, учитывающую события смещения сложной структуры и неравномерность сигнала смещения вдоль хромосом.
4. Разработать программные пакеты на основе этих моделей и провести их тестирование на симулированных данных.
5. Исследовать генетическую структуру панели образцов нута из коллекции ВИР и сгруппировать образцы в отдельные популяции по географическому и генетическому сходству.
6. Применить разработанные модели к выделенным генетико-географическим группам разнообразия нута.
7. Установить происхождение Эфиопской и других популяций нута дези.
8. Установить регион происхождения нута кабули.

Научная новизна

1. Идентифицированы новые генетико-географические популяции нута: Ливанская, западно-Среднеазиатская, а также две отдельные Средиземноморские популяции – северобережная и южнобережная.
2. Впервые разработана модель – `popdisp` – для оценки частот аллелей в популяциях, использующая теорию анализа композиционных данных и позволяющая учитывать ковариационную связь образцов в популяции.
3. Впервые разработана модель – `migadmi` – для реконструкции эволюционной истории популяций, способная моделировать множественные и вложенные события смешения, а также учитывающая композиционную природу частот аллелей.
4. Впервые продемонстрирована связь между генетическим сходством образцов внутри популяций и географически оптимальным путем перемещения популяций по ареалу.
5. Впервые определено происхождение двух смешанных популяций нута дези (Эфиопской и западно-Среднеазиатской), а также установлено происхождение нута кабули. Кроме того, предложены гипотезы о торговых путях, которые могли способствовать формированию этих смешанных популяций.

Теоретическую значимость работы определяют:

- Аппроксимация модели Райта-Фишера в терминах композиционного анализа данных.
- Байесовская иерархическая модель для оценки характеристических частот аллелей в популяции, учитывающая ковариационную матрицу связей между образцами. Расширение модели на случай многоаллельных генетических маркеров.
- Формула преобразования матрицы расстояний между образцами в ковариационную матрицу с учетом репрезентативного центра популяции.
- Параметризация события смешения нескольких популяций, позволяющая смешанной популяции обладать собственной дисперсией после события смешения.
- Модели смешения популяций с использованием композиционного представления частот аллелей.

Практическую значимость работы определяют:

- Программный код модели `popdisp`, предоставленный в открытый доступ на платформе GitHub: github.com/iganna/popdisp.

- Программный код модели `migadmi`, предоставленный в открытый доступ на платформе GitHub: github.com/iganna/migadmi.
- Более детальное описание популяционной структуры коллекции нута ВИР, имеющее сельскохозяйственное значение для поддержания генетического разнообразия при разработке новых сортов нута.
- Характеристические частоты аллелей популяций нута, которые могут быть использованы для определения происхождения образцов в рамках селекционных программ.
- Установленное смешанное происхождение некоторых популяций дези, что указывает на существование двух древних, независимых торговых путей из Ближневосточного региона (включая Ливан, Турцию, Южно-Черноморский регион и Плодородный полумесяц) в Эфиопию. Этот результат позволяет выдвигать гипотезы, которые могут быть применены при анализе разнообразия и определении центров доместикации других культурных растений.

Методология и методы исследования

Программный код методов `popdisp` и `migadmi` выполнен на языке программирования Python. Для оптимизации использовались методы из существующих библиотек: метод `hmc`, осуществляющий Гамильтониан Монте Карло, из библиотеки `ruhmc` и метод `SLSQP`, последовательного квадратичного программирования, из библиотеки `scipy`. Анализ координат точек сбора образцов, а так же построение матриц географических расстояний были проведены с помощью инструментов пакетов `leastcostpath`, `elevatr`, `movecost`, `sp` на языке программирования R. Анализ генетической структуры данных был выполнен с помощью программы `ADMIXTURE` (запуск из интерфейса командной строки). Подтверждение результатов `migadmi` было выполнено путем сравнения с результатами работы программ `TreeMix` (запуск из интерфейса командной строки), `MiaMapper` (язык программирования Matlab) и `ADMIXTOOLS` (язык программирования R). Программный код для визуализации был выполнен на языках программирования R и Matlab.

Основные результаты и положения, выносимые на защиту:

1. Разработан и протестирован метод `popdisp` для оценки характеристических частот аллелей в популяциях, отличающийся тем, что использует знание о ковариационной структуре популяции, а также использует анализ композиционных данных для работы с частотами аллелей в популяции.

2. Разработан и протестирован метод *migadmi*, реконструирующий эволюционную историю популяций, отличающийся тем, что позволяет учитывать сложные события смешения (больше двух исходных популяций и множественные вложенные события), а также использует анализ композиционных данных для работы с частотами аллелей в популяции.
3. В результате реконструкции эволюционной истории популяций нута с помощью методов *popdisp* и *migadmi* было показано, происхождение Эфиопской популяции нута дези является смешанным, при этом исходными популяциями выступают Черноморская, Индийская и Ливанская популяции нута дези.
4. Происхождение западной-Среднеазиатской популяции нута дези является смешанным, при этом исходными популяциями выступают Ливанская и восточная-Среднеазиатская популяции нута дези.
5. Происхождение нута кабули относится к региону современной Турции, а не к азиатскому региону, как можно было бы предположить из-за сходства названий 'кабули' и 'Кабул'.
6. Генетические сходства между образцами нута внутри каждой популяции больше соответствует оптимальному перемещению по ареалу с учетом географических особенностей, чем прямолинейному пути распространения.

Достоверность полученных результатов для популяций нута подтверждается их согласованностью с ранее опубликованными археологическими находками и не противоречит, а скорее уточняет данные генетического анализа, проведенного на других панелях образцов нута. Несмотря на различия в результатах программ *TreeMix*, *MixMapper* и *ADMIXTOOLS*, программа *migadmi* показала результаты, согласующиеся с большинством из них. Надежность работы программ *popdisp* и *migadmi* обеспечивается детальным тестированием на симулированных данных.

Личный вклад автора заключался в участии в процессе формулирования цели исследования, а также в определении конкретных задач, необходимых для ее достижения. Автор лично предложил и аналитически построил композиционную аппроксимацию модели Райта-Фишера, разработал и программно реализовал новые методы *popdisp* и *migadmi*, разработал схему для симуляции данных и провел тестирование разработанных программных пакетов. Автор самостоятельно провел анализ генетических данных панели нута из коллекции ВИР и определил

генетически однородные и географически связанные популяции нута. Автором единолично были составлены матрицы географических расстояний между координатами точек сбора образцов, применены как новые разработанные, так и существующие модели к данным по нуту, а также проанализированы полученные пути происхождения популяций нута. Все рисунки, не заимствованные из опубликованных источников, выполнены автором лично.

Публикации

В рамках темы диссертации было опубликовано 8 статей в научных журналах, индексируемых в Web of Science и Scopus:

1. Igolkina A. A., Noujdina N. V., Vishnyakova M., Longcore T., von Wettberg E., Nuzhdin S. V., Samsonova M. G. Historical routes for diversification of domesticated chickpea inferred from landrace genomics // *Molecular Biology and Evolution*. 2023. Vol. 40, No 6. P. msad110.
2. Igolkina A. A., Meshcheryakov G., Gretsova M. V., Nuzhdin S. V., Samsonova M. G. Multi-trait multi-locus SEM model discriminates SNPs of different effects // *BMC Genomics*. 2020. Vol. 21, No S8. P. 490.
3. Igolkina A. A., Meshcheryakov G. semopy: A Python package for structural equation modeling // *Structural Equation Modeling: A Multidisciplinary Journal*. 2020. Vol. 27, No 6. P. 952–963.
4. Korvigo I., Igolkina A. A., Kichko A. A., Aksenova T., Andronov E. E. Be aware of the allele-specific bias and compositional effects in multi-template PCR // *PeerJ*. 2022. Vol. 10. P. e13888. doi:10.7717/peerj.13888.
5. Sokolkova A., Bulyntsev S. V., Chang P. L., Carrasquilla-Garcia N., Igolkina A. A., Noujdina N. V., von Wettberg E., Vishnyakova M. A., Cook D. R., Nuzhdin S. V., Samsonova M. G. Genomic analysis of Vavilov's historic chickpea landraces reveals footprints of environmental and human selection // *International Journal of Molecular Sciences*. 2020. Vol. 21, No 11. P. 3952.
6. Kovalev M. S., Igolkina A. A., Samsonova M. G., Nuzhdin S. V. A pipeline for classifying deleterious coding mutations in agricultural plants // *Frontiers in Plant Science*. 2018. Vol. 9. P. 1734.
7. Igolkina A. A., Bazykin G. A., Chizhevskaya E. P., Provorov N. A., Andronov E. E. Matching population diversity of rhizobial nodA and legume NFR5 genes in plant-microbe symbiosis // *Ecology and Evolution*. 2019. Vol. 9, No 18. P. 10377-10386. doi:10.1002/ece3.5556.

8. Igolkina A. A., Porozov Yu. B., Chizhevskaya E. P., Andronov E. E. Structural insight into the role of mutual polymorphism and conservatism in the contact zone of the NFR5–K1 heterodimer with the Nod factor // *Frontiers in Plant Science*. 2018. Vol. 9. doi:10.3389/fpls.2018.00344.

Апробация работы

Результаты работы опубликованы в сборниках тезисов 2 конференций:

1. Igolkina A., Samsonova M., Nuzhdin S. Compositional approximation of the Wright-Fisher model in estimating genetic centroids and admixture graphs // *ProbGen2024*, Вена, Австрия, 10-12 апреля 2024.
2. Igolkina A., Noujdina N., Longcore T., von Wettberg E., Samsonova M., Nuzhdin S. Models for crop species diversification within and between centers of domestication // *MCCMB2021*, Москва, Россия, 30 июля - 2 августа 2021.

Объем и структура работы.

Диссертация состоит из введения, 4 глав и заключения. Полный объем диссертации составляет 138 страниц, включая 32 рисунка и 4 таблицы. Список литературы содержит 98 наименований.

Глава 1. Обзор литературы

1.1 Генетическая структура и географическая организация популяций

1.1.1 Проблема оценки частот аллелей в популяции

Популяции живых организмов представляют собой совокупности индивидуумов одного вида, проживающих на ограниченной географической территории и частично или полностью изолированных от других таких же групп. На генетическом уровне популяции представляют собой изменчивые группы, где каждый индивидуум обладает уникальным набором генетических аллелей. В условиях панмиксии (свободное скрещивание) эти генетические аллели будут равномерно распределены в популяции. Однако на практике это не наблюдается из-за неизбежных отклонений от условий панмиксии. Например, особи, находящиеся географически ближе друг к другу, будут с большей вероятностью скрещиваться и иметь большее генетическое сходство, чем географически удаленные.

Такая положительная связь между генетическим и географическим расстоянием наблюдается и между популяциями и называется изоляцией расстоянием (isolation by distance, IBD) [17]. В ряде исследований это можно продемонстрировать простым анализом главных компонент: распределение в первых двух главных компонентах, построенное на генетических данных человеческих популяций, является "зеркалом" широты и долготы точек сбора образцов на обширных географических ареалах [18—20]. Однако на более мелких масштабах (на уровне одной популяции) влияние гетерогенности ландшафта может стать существенным, и наблюдаемые генетические расстояния между особями будут отражать не прямолинейные, а скорее оптимальные географические пути. Поэтому фактор гетерогенности ландшафта должен быть как минимум рассмотрен и, в конечном итоге, учтен при оценке характеристических частот аллелей в популяции.

Одной из характеристик популяции организмов являются частоты аллелей в полиморфных однонуклеотидных локусах (однонуклеотидные полиморфизмы, ОНП, англ. single nucleotide polymorphism, SNP), которые представляют собой самые простые и наиболее распространенные типы генетических вариаций.

Частоты аллелей ОНП (далее аллелей, опуская ОНП) позволяют оценить генетическое разнообразие популяции, реконструировать процессы и события, которые определили генетические изменения с течением времени: миграции, генетический дрейф, смешение популяций. Однако оценка частот аллелей не является простой задачей.

Предположим, что для некоторого вида была произведена выборка образцов на некоторой территории (будем считать ее полным ареалом для простоты) и каждый образец был генотипирован (т.е. определены значения аллелей) (Рисунок 1.1А). Простейшей оценкой частот аллелей в популяции является вектор усредненных значений по каждому локусу ОНП без учета координат точек сбора образцов. Заметим, во-первых, что выборочное среднее по своей сути является неробастной оценкой (чувствительной к выбросам), а использование выборочной медианы вместо среднего в данном случае невозможно, так как ОНП принимают дискретные значения из множества $\{0,1,2\}$. Во-вторых, неравное количество образцов в локациях сбора может приводить к неконтролируемому смещению выборочного среднего, особенно если размеры выборок значительно различаются (Рисунок 1.1Б). Например, если одна локация содержит значительно больше образцов, чем другие локации, и эти образцы имеют уникальные аллели, то выборочное среднее будет смещено в пользу этой локации. Если несколько образцов взяты слишком близко друг к другу, то они могут быть генетически более схожи, что может привести к тем же искажениям, что и в случае разного количества образцов в локациях. Поэтому, в-третьих, при оценке частот аллелей важно учитывать взаимное расположение точек сбора образцов (Рисунок 1.1В).

Близость двух точек в географических координатах еще не указывает на то, что миграция между этими точками происходит чаще, чем между более удаленными областями. Некоторые зоны более благоприятны для миграции, в то время как другие могут служить барьерами и быть непригодными для обитания, например, горы или водоемы. На Рисунке 1.1Г показано, что наличие таких барьеров может сделать даже очень близкие друг к другу локации фактически удаленными.

При исследовании связи между генетическими и географическими расстояниями, понятие частоты аллелей в популяции является абстрактным конструктом и результатом подгонки моделей к данным без учета процессов, происходивших в популяции при ее распространении по ареалу [21]. При этом подразумевается, что в каждом месте сбора существует локальная подпопуляция, и ее частоты аллелей некоторым образом отклоняются от глобальной частоты аллелей для всей попу-

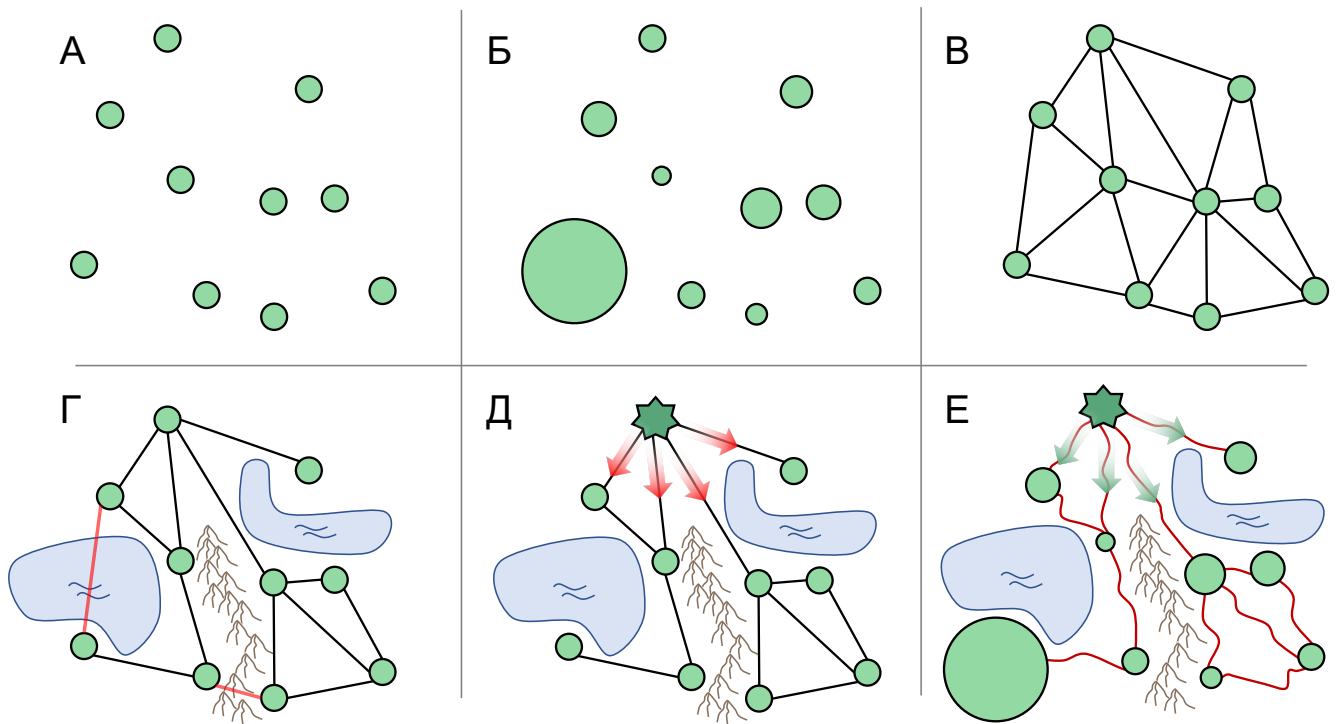


Рисунок 1.1 — Примеры расположения выборки образцов на гипотетическом географическом ландшафте с учетом различных аспектов.

(А) Исходное расположение точек сбора образцов без учета каких-либо факторов. Каждый круг представляет точку сбора. (Б) Расположение точек сбора образцов с указанием количества образцов в каждой точке; размер круга соответствует количеству образцов. (В) Линии между точками представляют пути потенциального обмена генетической информацией. (Г) Не все прямолинейные пути обмена возможны из-за географических барьеров. Красным показаны невозможные или маловероятные пути. (Д) Генетический центроид популяции задает направление дрейфа. (Е) Полное схематичное представление географической гетерогенности популяции. Пути перемещения являются оптимальными географическими, а не прямолинейными.

ляции. Однако естественно предположить, что в ареале существует определенное географическое местоположение (Рисунок 1.1Д), в котором частота аллелей локальной подпопуляции соответствует частоте аллелей для всей популяции. Это местоположение будем называть генетическим центроидом, а частоты аллелей в нем – характеристическими для всей популяции. Важно отметить, генетический центроид может не совпадать с геометрическим центром ареала популяции.

И наконец, необходимо учитывать, что реальное расстояние между локациями двух образцов определяется длиной оптимального пути перемещения по

конкретному географическому ландшафту, а не линейным расстоянием между локациями (Рисунок 1.1Е). В то время как одни организмы распространяются по территории самостоятельно, для других организмов, основной движущей силой в распространении по территории могут быть различные биотические и абиотические факторы. Это должно быть учтено при выборе функции затрат для определения оптимального пути. В частности, для агрокультур основной движущей силой в распространении по территории является человек. В этом случае пути распространения популяции по ареалу являются оптимальными для передвижения человека. Можно предположить, что генетический центроид популяции в этом случае находится вблизи центра доместикации.

В результате, для оценки характеристических частот аллелей в популяции, необходимо учитывать количество образцов в локациях сбора, генетический центроид и оптимальные пути перемещения между локациями (Рисунок 1.1Е).

1.1.2 Методы анализа популяционной структуры без учета географического сигнала

Для оценки генетического разнообразия и популяционной структуры в исследуемых образцах могут быть использованы прямые методы и методы с использованием моделирования.

Прямые методы заключаются в представлении генетических данных (таблица данных с большим количеством генетических маркеров для различных образцов) в пространстве меньшей размерности, удобном для визуального или кластерного анализа. В качестве наиболее широко используемого метода сокращения размерности можно назвать метод главных компонент.

Этот метод представляет образцы в ортогональном пространстве, размерность которого равна числу образцов. Оси в этом пространстве, называемые главными компонентами, определяются как линейные комбинации генетических маркеров, чтобы каждая из них описывала как можно больший процент общей дисперсии данных. Современные методы сокращения размерности, такие как tSNE (t-distributed Stochastic Neighbor Embedding) [22] и UMAP (Uniform Manifold Approximation and Projection) [23], используют нелинейные преобразования дан-

ных. Эти методы в первую очередь сохраняют локальную структуру данных, то есть генетически близкие образцы отображаются близко друг к другу.

На основании визуального представления образцов в пространстве меньшей размерности можно провести качественный анализ популяционной структуры. Классическим же количественным методом группировки образцов в определенном пространстве является метод k -средних, направленный на выявление заранее заданного количества кластеров - k . Этот метод распределяет образцы по кластерам на основе близости к центрам кластеров, стремясь минимизировать сумму квадратов расстояний от образцов до центров. Современным альтернативным методом является DBSCAN (Density-Based Spatial Clustering of Applications with Noise), который учитывает локальные плотности распределения точек и способен обнаруживать кластеры различной нерегулярной формы и размеров. Общим недостатком описанных прямых подходов является сложность интерпретации кластеризации образцов с точки зрения общей эволюционной истории и событий смешения популяций.

Другим подходом для анализа популяционной структуры являются методы с использованием моделирования и оптимизации. Характерными методами этого типа являются STRUCTURE [13] и ADMIXTURE [12]. Эти методы основаны на предположении о том, что образцы могут быть сформированы в результате смешения нескольких генетически отличимых базовых популяций. Сами базовые популяции изначально недетерминированы и их характеристики (вектор частот аллелей) оценивается в результате оптимизации. Программы STRUCTURE/ADMIXTURE принимают на вход только число предполагаемых базовых популяций, а на выходе выдают доли принадлежности каждого образца к каждой базовой популяции. Разница между методами заключается в деталях моделирования композиций и методах оптимизации. Программа ADMIXTURE является вычислительно более эффективной, а так же способна справляться с сотнями тысяч образцов [24]. Существует и вычислительно ускоренная версия программы STRUCTURE - fastSTRUCTURE [25], однако по результатам симуляций она по точности уступает другим программам [26]. Еще одной программой с аналогичным STRUCTURE/ADMIXTURE подходом является sNMF, которая в 10-30 раз быстрее ADMIXTURE и использует разреженное неотрицательное матричное разложение для нахождения коэффициентов смешения [27].

Несмотря на то, что подходы STRUCTURE/ADMIXTURE/sNMF дают большее представление об эволюционной истории популяций, чем прямые методы

кластеризации, интерпретация получаемых композиций базовых популяций не является тривиальной задачей. Было показано, что один и тот же композиционный паттерн для четырех популяций можно получить в при разных эволюционных сценариях [24]. Оказывается, что события недавнего смешения трех популяций с формированием четвертой будет давать тот же результат программы ADMIXTURE, что и сценарий недавнего события ”бутылочного горлышка” или сценарий смеси двух популяций, одна из которых известна, а вторая не является одной из наблюдаемых. То есть композиционные паттерны могут быть неоднозначно интерпретированы и требуют дополнительных методов анализа эволюционной истории образцов.

Можно назвать несколько факторов, которые могут усложнять интерпретацию результатов программ STRUCTURE/ADMIXTURE/sNFT. Одно из предположений программы ADMIXTURE заключается в том, что после события смешения популяций не произошло никакого дрейфа [28]. Если событие смешения популяций произошло некоторое время назад, то и смешанная популяция, и современные популяции, произошедшие от участвовавших в смешении, приобрели свое собственное разнообразие как результат независимого генетического дрейфа. В результате, ни один образец смешанной популяции не является точной композицией современных популяций. Таким образом, представление смешанных образцов в виде точной композиции является неполным, что может служить потенциальной причиной неоднозначности в интерпретации результатов.

Другим фактором, потенциально мешающим интерпретации результатов программ STRUCTURE/ADMIXTURE/sNFT, является отсутствие учета географического сигнала в данных. Когда в выборке присутствует группа генетически близких образцов, эти программы склонны к формированию из них отдельной базовой популяции. Однако близкое генетическое сходство может быть обусловлено лишь географической близостью, например, когда образцы собраны в одном месте. В данном случае нет смысла выделять эти образцы в отдельную популяцию, так как они не представляют собой отдельную группу, а скорее указывают на отсутствие равномерности выборки. Таким образом, возможное улучшение моделей типа STRUCTURE/ADMIXTURE/sNFT заключается в учете географического положения образцов. То есть вместо того чтобы считать ошибки в моделях независимыми, их можно рассматривать как связанные через ковариационную матрицу, некоторым образом отражающую географическую близость.

1.1.3 Методы, связывающие пространственную и генетическую информацию

Все естественные популяции имеют популяционную структуру, то есть неравномерное распределение аллелей в популяции, которое может приводить к выделению локальных подгрупп или подпопуляций. Поскольку одним из факторов, способствующих формированию популяционной структуры, являются географические условия, в естественных популяциях генетическое расстояние часто коррелирует с географическим. Именно поэтому для генетического анализа популяций были разработаны методы, которые используют информацию о координатах точек сбора образцов.

В качестве методов, которые используют подход программ STRUCTURE/ADMIX вместе с информацией о точках сбора образцов, можно выделить TESS3 (Tessellation Engine for Spatial Simulation) [29; 30] и conStruct (continuous STRUCTURE) [31]. Метод TESS3 применяет байесовскую оптимизацию для оценки композиционных вкладов базовых популяций в каждый из образцов. При этом географические координаты точек сбора образцов учитываются в моделировании как переменные, влияющие на параметры распределения Дирихле, из которого производится сэмплинг интересующих композиционных вкладов при оптимизации [29]. В качестве результатов программа предоставляет не только оценки композиционных вкладов для каждого образца, но и плотность распределения вкладов каждой базовой популяции на участках географической карты, где расположены анализируемые образцы.

В методе conStruct данные о географическом местоположении образцов влияют не на параметры композиционных вкладов базовых популяций, а на плотность распределения базовых популяций в пространстве [31]. Для каждой i -ой базовой популяции отдельно на основании плотностей можно вычислить матрицу расстояний между D_i образцами и связать ее с ковариационной матрицей, G_i с помощью преобразования Матерна (Matérn covariance function). Метод conStruct предполагает, что выборочная матрица ковариации между образцами, Ω , рассчитанная на основе генетических данных, следует распределению Вишарта. Параметр в этом распределении — квадратная симметричная матрица масштаба — представляет в модели взвешенную сумму ковариационных матриц G_i , соответствующих слоям базовых популяций, а веса выражаются через параметры

искомых композиционных вкладов. В отличие от программы TESS3, программа conStruct не выдает распределение композиционных вкладов на всем географическом пространстве, а только в точках сбора образцов.

Другая группа методов, которая учитывает географическое положение образцов, включает методы анализа ассоциаций между генетической и географической информацией. Метод BEDASSLE, аналогично методу conStruct, учитывает координаты точек сбора образцов, но направлен на количественную оценку вклада географических и экологических сигналов в генетическое различие между локальными популяциями [21]. Согласно этой модели, частоты аллелей в локальных популяциях отклоняются не независимо от глобальной средней частоты аллелей, и матрица ковариаций этих отклонений определяется как преобразование Матерна взвешенной суммы географических и экологических расстояний. Модель BEDASSLE представляет собой подгонку пространственного распределения к генетической изменчивости, без моделирования процессов, которые ее породили.

Недостатком перевода матрицы географических расстояний в ковариационную матрицу с помощью преобразования Матерна, предложенного в работах [21; 31], заключается в том, что это нелинейное преобразование требует дополнительной оценки параметров. В исходной формулировке для функции Матерна участвуют Гамма-функция и функция Бесселя. Однако в простейшей параметризации она представлена как экспоненциальное убывание: $\rho = \exp(-d)$, где d - расстояние, а ρ - ковариация. В работах [21; 31] были введены дополнительные скалирующие параметры для этой функции:

$$\rho = \alpha_0 \exp(-(\alpha_D d)^{\alpha_2}) + \varphi. \quad (1.1)$$

Параметр α_0 определяет масштабирование или амплитуду ковариации. Параметр α_D действует как обратное значение длины масштаба и определяет, насколько быстро корреляция уменьшается с расстоянием. Степенной параметр α_2 влияет на форму убывания функции. Параметр φ представляет базовый уровень ковариации, который сохраняется даже на больших расстояниях. Таким образом, преобразование расстояний в ковариации при таком подходе требует четырех дополнительных параметров, что может привести к ряду проблем, связанных с переобучением, вычислительной эффективностью и интерпретацией результатов.

1.1.4 Методы оценки географических расстояний

Одним из факторов, влияющих на формирование структуры популяции, являются естественные географические барьеры для потока генов. Это позволяет, анализируя генетическое разнообразие популяции на территории, выявлять зоны с различной интенсивностью потока генов. Одной из программ, оценивающей барьеры для потока генов, является EEMS [32]. Эта программа создает карту, которая показывает, где генетические различия между образцами увеличиваются или уменьшаются в сравнении с ожидаемым на основе географического расстояния. В результате карта миграционных путей показывает зоны, препятствующие потоку генов или, наоборот, коридоры, способствующие миграции.

В ряде исследований с использованием этого подхода было показано, что барьеры для потока генов совпадают с реальными географическими преградами [32; 33]. И образцы, находящиеся близко друг к другу с точки зрения прямых путей, на самом деле находятся далеко друг от друга с точки зрения реальных миграционных путей. Таким образом, использование в моделях из параграфа 1.1.3 матриц линейных расстояний между образцами без учета реального географического ландшафта может вносить искажения в результаты применения этих моделей.

Вместо линейных расстояний между образцами, можно использовать так называемые длины путей наименьшей стоимости (least-cost paths, least-cost distance). Этот подход широко используется в экологических исследованиях для моделирования миграционных путей животных, планирования транспортных путей, а также в других областях, где требуется учет сложных условий местности или других препятствий. Для определения пути наименьшей стоимости между двумя точками необходимо наложить на область ортогональную сетку и задать функцию затрат (cost function) перехода между соседними ячейками. Затраты будут зависеть от географических характеристик каждой ячейки. Поиск пути наименьшей стоимости схож с поиском кратчайшего пути в графе, где вершины представляют собой ячейки, а ребра взвешены на основе стоимости перехода между ячейками.

Существует множество типов функций затрат, созданных для оценки перемещения людей, автомобилей и других объектов по ландшафту. Наиболее широко используемая функция затрат, оценивающая стоимость перемещения человека

по ландшафту, — это ”Функция пеших прогулок” Уолдо Тоблера. Эта функция позволяет оценить время, необходимое для пересечения поверхности, учитывая отдельно подъемы и спуски [34]. В качестве альтернативы, не зависящей от направления движения (вверх или вниз), можно использовать функцию Герцога [35]. Благодаря своей симметричности, эта функция особенно удобна для оценки путей, когда направление неизвестно и движение возможно в обе стороны.

1.2 Методы реконструкции эволюционной истории популяций

1.2.1 Описание задачи

Реконструкция эволюционной истории вида играет важную роль в понимании процессов, которые формируют генетическое разнообразие данного вида. При реконструкции эволюционной истории исследуемые данные удобно разбивать на группы генетически сходных образцов по ряду причин. Во-первых, такое группирование позволяет сосредоточиться на более общих шаблонах и динамике между группами, нежели на сложно интерпретируемой динамике на уровне индивидов. Во-вторых, группировка образцов снижает влияние случайных вариаций отдельных индивидов и усиливает общие сигналы, что приводит к более четкому пониманию исторической динамики и позволяет делать более уверенные выводы.

Существует два основных подхода к терминологии при именовании групп генетически сходных образцов. Первый подход предлагает использовать термин ”популяции”. В этом случае предполагается, что каждая группа рассматривается как отдельное и относительно изолированное сообщество индивидов, обладающих общими генетическими характеристиками. Второй подход предлагает использовать термин ”подпопуляции”. Этот термин указывает на то, что группы являются частью большего генетического континуума и могут иметь меньшие различия между собой по каким-то признакам. В этом случае подпопуляции могут представлять собой подгруппы внутри более крупной популяции, и границы между ними могут быть менее строгими. Таким образом, выбор между терминами ”популяция” и ”подпопуляция” зависит от того, как выделенные группы рассматриваются в анализе: как разделенные и изолированные или как взаимосвязанные

и переплетенные. Поскольку строгих правил разделения нет, то в настоящей работе во избежание избыточности, будем использовать термин ”популяция”.

Одним из наиболее базовых методов реконструкции эволюционной истории популяций является построение бинарного филогенетического дерева методом иерархической кластеризации, например, методом присоединения соседей (англ. ”Neighbor joining”). Для работы этого метода необходима матрица попарных генетических расстояний между популяциями, каковой может служить, например, матрица попарных значений индекса фиксации F_{ST} . Индекс фиксации F_{ST} является одним из трех индексов, предложенных Стюаллом Райтом [36]. В основе этого показателя лежит сравнение генетического разнообразия внутри популяций с разнообразием между популяциями. Значение F_{ST} , посчитанное для двух популяций, всегда лежит между 0 и 1 и равно 0 только в отсутствие структуры в объединенной популяции. В том случае, когда в наборе рассматриваемых популяций есть как смешанные популяции, так и их популяции-источники, восстановить историю смешения только на основе значения индекса F_{ST} невозможно, несмотря на определенную связь между значениями F_{ST} и коэффициентами смешивания [37].

Существует ряд программ, которые реконструируют эволюционную историю популяций с учетом событий смешения: TreeMix [14], Mixmapper [15], инструменты qpGraph и findGraph из пакета ADMIXTOOLS [16], OrientAGraph [38]. Все эти программы оценивают граф смешений, который представляет собой ориентированный ациклический граф, где все узлы-стоки и возможно некоторые внутренние узлы представляют собой анализируемые популяции (примеры графов смешения показаны на Рисунке 1.2 Б,Г-З). Ребра в графах смешения бывают двух типов: ребра дрейфа и ребра смешения. Ребра дрейфа отражают независимую историческую траекторию, пройденную определенной популяцией между узлами, соответствующими началу и концу данного ребра. Каждому ребру дрейфа соответствует положительное число, отражающее объем генетического дрейфа или изменений, произошедших в популяции на протяжении времени, которое покрывает данное ребро. Ребра смешения отражают исторические события смешения между двумя или более популяциями и организованы в группы по каждому событию смешения. Каждому ребру смешения соответствует вес, число от 0 до 1, так чтобы сумма весов ребер в каждой группе равнялась 1.

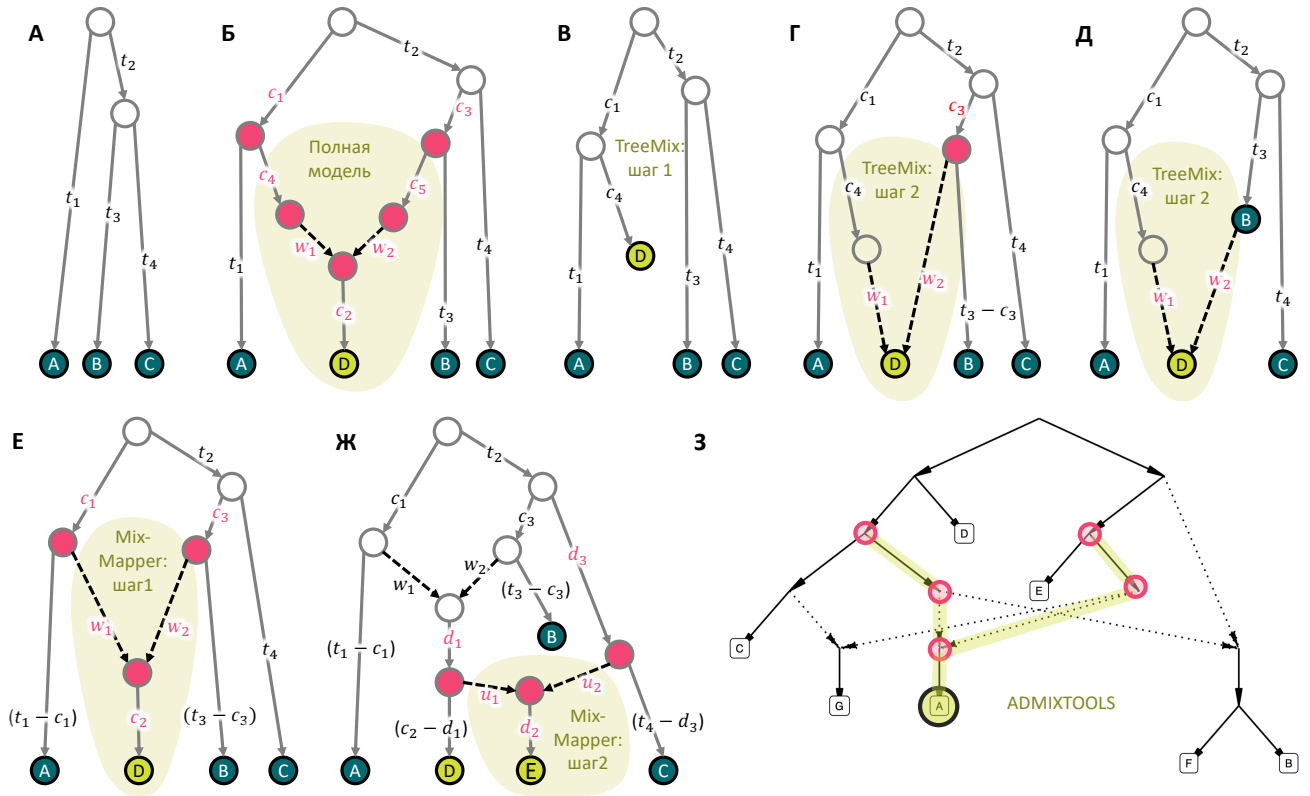


Рисунок 1.2 — Параметризация событий смешения в различных программах.

(А) Бинарное дерево связей между популяциями. Синие кружки — наблюдаемые популяции, пустые кружки — предковые. Стрелки показывают эволюционные связи. (Б) Модель смешения для популяции D: популяции A и B не напрямую вносят вклад в D. (В) Шаг 1 программы TreeMix: построение бинарного дерева для всех популяций, включая смешанную. (Г) Шаг 2 программы TreeMix: добавление в бинарное дерево события смешения. (Д) Результат TreeMix: наблюдаемая популяция вносит прямой вклад в смешанную без учета популяций-предшественников. (Е) Модель одного смешения в MixMapper. (Ж) Модель двух событий смешения в MixMapper. (З) Модель смешения в ADMIXTOOLS, полученная с помощью команды `rand_graph = random_admixturegraph(7, numadmix = 3)`.

1.2.2 Варианты параметризации событий смешения

Рассмотрим бинарное дерево, построенное для трех популяций (Рисунок 1.2 А). Каждое ребро в этом дереве является ребром дрейфа, и для описания такого дерева необходимо четыре параметра, соответствующих длинам ребер. Если

существует четвертая популяция, являющаяся результатом смешения двух других, то включение ее в бинарное дерево приведет к созданию графа смешения. Полная модель события смешения потребует добавления пяти дополнительных внутренних узлов к графу и одного узла-стока в сравнении с бинарным деревом для трех популяций (Рисунок 1.2 Б). И для описания такого графа смешения потребуется ввести шесть дополнительных параметров. Параметры весов смешения связаны соотношением $w_1 + w_2 = 1$, поэтому вместо двух весов можно рассмотреть только какой-то один. Однако полученная полная модель смешения вероятно будет перепараметризованной, поэтому в моделях, используемых для реконструкции истории смешения популяций, зачастую применяются упрощенные модели.

Чтобы избежать переобучения параметров, программа TreeMix [14] оценивает их в два этапа и использует усеченную модель события смешения. Сначала программа оценивает оптимальное бинарное дерево без событий смешения, основанное на всех популяциях, которые располагаются только в листьях (Рисунок 1.2 В). Затем, события смешения добавляются в дерево последовательно в заранее определенном количестве. Параметризация события смешения при таком подходе не является симметричной (Рисунок 1.2 Г). В дополнительных материалах к статье, авторы анализируют этот тонкий момент, говоря, что если истинные вклады c_5 или c_2 велики, то это будет приводить к смещению оценки w_2 в сторону уменьшения (Рисунок 1.2 Б,Г). Таким образом, параметры весов w_1 и w_2 не отражают доли при смешении популяций, а скорее являются некоторыми параметрами, которые обеспечивают наилучшее соответствие несимметричной модели и данных. Другим недостатком модели TreeMix можно назвать отсутствие параметров, характеризующих независимый дрейф у смешанной популяции (параметр c_2 , Рисунок 1.2 Б). Кроме того, на практике часто встречаются ситуации, когда ребра смешения в графе напрямую соединяют одну популяцию с другой (Рисунок 1.2Д). Для оптимизации параметров, TreeMix использует нормальную аппроксимацию модели генетического дрейфа, которая будет подробнее описана в Параграфе 1.3.3.

При добавлении события смешения программа TreeMix перебирает возможные положения для этого события на графе смешений и выбирает наилучший вариант на основе значений функции правдоподобия. Поскольку в общем случае такой подход требует полного перебора, в программе TreeMix реализована определенная эвристика, чтобы этого избежать. Как и любая эвристика, TreeMix не всегда находит наилучшие возможные ребра смешения. Поэтому на базе модели

смешения программы TreeMix была создана программа OrientAGraph [38], которая в половине характерных типов моделей (4 из 8) выбирает ребра, дающие более высокие значения правдоподобия. Стоит также отметить, что программа TreeMix не уделяет особого внимания вложенным событиям смешения, однако они могут появиться в результате последовательного наращивания графа смешений.

Программа MixMapper [15] использует симметричную модель смешения (Рисунок 1.2 Е), и не оценивает больше двух событий смешения в одном графе. При этом эти два события смешения обязательно должны быть вложенными (Рисунок 1.2 Ж), что является большим ограничением программы MixMapper. Каждое из событий смешения добавляет в граф три внутренних узла, один узел-сток (смешанная популяция), и четыре дополнительных параметра по сравнению с бинарным деревом для трех несмешанных популяций (Рисунок 1.2 А,Е). Если сравнивать модели смешения MixMapper и TreeMix для одного события смешения, то они равны по количеству дополнительных параметров и количеству дополнительных элементов в графе смешения по сравнению с бинарным деревом для трех несмешанных популяций. Для оптимизации параметров MixMapper решает систему уравнений, основанных на f -статистиках [39—41].

По сравнению с другими программами, пакет ADMIXTOOLS обладает наиболее широким спектром инструментов для построения и анализа графов смешения [16]. Кроме того, инструменты ADMIXTOOLS предоставляют возможность работы с топологиями графов смешения, недоступными в других существующих на сегодняшний день программах (Рисунок 1.2 З). Например, ADMIXTOOLS предоставляет возможность работы с событиями смешения, которые соответствуют полной модели (Рисунок 1.2 Б). Во избежание перепараметризации в этой ситуации дополнительные ребра, которые возникают в одном событии смешения, переиспользуются повторно в другом событии смешения (Рисунок 1.2 З).

Все инструменты ADMIXTOOLS и программа MixMapper основаны на анализе f -статистик [39—41]. Три статистики — f_2 , f_3 , f_4 — позволяют определить степень родства между двумя, тремя и четырьмя популяциями соответственно. Визуально значения этих статистик для несмешанных популяций представлены на Рисунке 1.3. Статистика f_2 используется для оценки генетического дрейфа между двумя популяциями (Рисунок 1.3А). Вычислительно, эта статистика представляет собой среднее квадратов разностей между частотами аллелей в двух популяциях:

$$f_2(A, B) = \frac{\langle p_A - p_B, p_A - p_B \rangle}{N},$$

где p_A и p_B - вектора частот аллелей в популяциях A и B , N - длина векторов (количество генетических маркеров), $\langle \cdot, \cdot \rangle$ - скалярное произведение. Статистика f_3 показывает объем общего генетического дрейфа между кладой из двух популяций и целевой внешней популяцией (Рисунок 1.3Б). Вычислительно эта статистика представляет собой скалярное произведение векторов разностей частот между популяциями в кладе и целевой популяцией, деленное на длину вектора:

$$f_3(A, B;) = \frac{\langle p_A - p, p_B - p_C \rangle}{N}.$$

Статистика f_3 может быть использована как тест на смешение: если значение $f_3 < 0$, то целевая популяция является смешанной, но если $f_3 > 0$, определить событие смешения однозначно нельзя [41].

Статистика $f_4(A, B; C, D)$ показывает объем общего генетического дрейфа между популяциями A и B в сравнении с популяциями C и D (Рисунок 1.3В). Эта статистика вычисляется как ковариация различий частот аллелей между двумя парами популяций:

$$f_4(A, B; , D) = \frac{\langle p_A - p_B, p_C - p_D \rangle}{N}.$$

В отсутствие событий смешения, если значение статистики $f_4(A, B; , D)$ больше нуля, то это означает, что популяции A и C находятся в кладе, отделенной от популяций B и D (Рисунок 1.3В). Если значение меньше нуля - то популяции A и D находятся в кладе, отдельной от популяций B и C . Если популяции A и B образуют независимую кладу с популяциями C и D , то $f_4(A, B; , D) = 0$ (Рисунок 1.3Г). То есть дрейфт, разделяющий A и B не имеет ничего общего с дрейфтом, разделяющим C и D . Статистика f_4 чувствительна к порядку популяций в парах. Тестом на отсутствие событий смешения между четырьмя популяциям может быть следующее: существует хотя бы одна перестановка параметров в подсчете f_4 статистики, выдающая значение 0. Или иначе: если значение f_4 на всех перестановках для определенного набора из четырех популяций являются статистически значимо ненулевыми, то по крайней мере одна из популяций должна являться смешанной [42].

Инструмент `qrGraph` пакета `ADMIXTOOLS` оценивает параметры заранее заданного графа смешения, а инструмент `findGraph` находит и оптимизирует граф смешений для заданных популяций с заданным числом событий смешения. Программа `qrGraph` оценивает параметры, опираясь на минимизацию разницы между

истинными и оцененными значениями f_3 статистики. Программа findGraph использует f_4 статистику при переборе возможных топологий графов смешения. Расчет статистик f_3 и f_4 может быть весьма вычислительно затратным из-за количества возможных комбинаций параметров. Однако, можно упростить задачу, предварительно вычислив попарные статистики f_2 и затем переиспользуя их для вычисления f_3 и f_4 , опираясь на свойства скалярного произведения:

$$2f_3(A, B; C) = f_2(A, C) + f_2(B, C) - f_2(A, B),$$

$$2f_4(A, B; C, D) = f_2(A, D) + f_2(C, B) - f_2(A, C) - f_2(B, D).$$

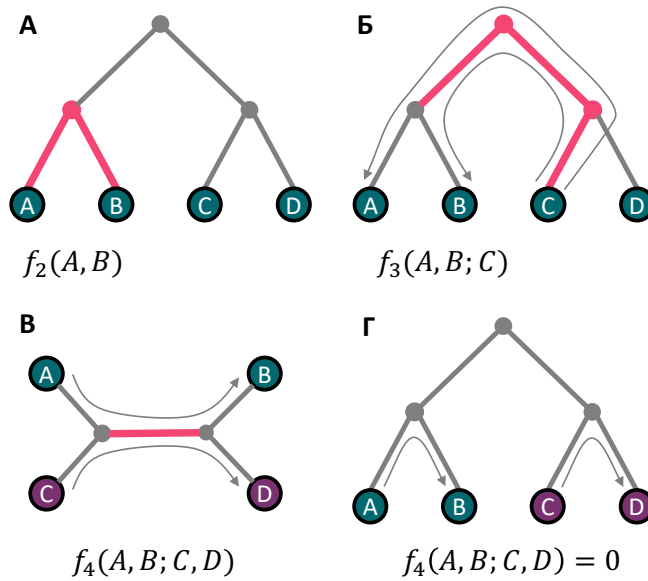


Рисунок 1.3 — Визуальное представление статистик f_2 , f_3 , f_4 .

(А) Статистика f_2 соответствует полному эволюционному расстоянию между двумя популяциями. (Б) Статистика f_3 соответствует полному эволюционному расстоянию от одной популяции до общего предкового состояния двух других. (В) Ситуация родства между популяциями, при которой статистика f_4 ненулевая. (Г) Ситуация родства между популяциями, при которой статистика f_4 равна нулю.

В программах TreeMix и MixMapper используется модель смешения, в которой формирование популяции происходит на основе двух популяций-источников. Однако в реальных сценариях это не всегда так. Инструмент qpGraph способен оценивать параметры графа смешения, включающего события с участием более чем двух популяций-источников. В то же время инструмент findGraph, предназначенный для поиска оптимальной топологии, работает только с графами, в

которых смешение происходит между двумя популяциями-источниками. Поэтому существует необходимость улучшить метод поиска графа смешений, добавив возможность работы с более сложными схемами смешения.

Общим недостатком существующих методов является их ограниченность в идентификации параметров и неспособность моделировать смешение с использованием полной модели (Рисунок 1.3Б). Чтобы избежать чрезмерного количества параметров и перепараметризации, модели упрощаются, но при этом параметры, предполагаемые как веса в упрощенной модели, фактически отражают комбинации параметров более сложной полной модели. В связи с этим, утверждение о том, что полученные параметры отражают веса в процессе смешения, не совсем корректно. Следовательно, необходимо разработать альтернативный метод оценки вклада популяций-источников в смешанную популяцию.

1.3 Модели генетического дрейфа

Дрейф генов — это случайное изменение частот аллелей в популяции конечного размера, происходящее в результате произвольного выбора родительских индивидуумов при смене поколений [43]. Для описания динамики случайного генетического дрейфа были предложены различные математические модели (обзор [44]), например, модель Райта-Фишера [45; 46], модель Морана [47] и модель Кимуры [48]. Наиболее популярной для описания изменения частот аллелей в популяциях является модель Райта-Фишера в форме аппроксимаций.

1.3.1 Модель Райта-Фишера

Модель Райта-Фишера лежит в основе большинства теоретических и прикладных исследований в популяционной генетике. Она рассматривает изменение частоты аллелей в одном локусе в популяции постоянного размера, состоящей из диплоидных бесполой индивидуумов, скрещивающихся случайным образом (панмиксия) для формирования нового поколения (скрещивания индивидуумов из разных поколений запрещены) [43]. В модели Райта-Фишера для биаллельного

локуса с вариантами a (референсным) и A (альтернативным) состояние популяции математически описывается с помощью случайной величины X_t , которая характеризует количество аллелей A в популяции фиксированного размера N в поколении t . В каждом поколении величина X_t может принимать значение из конечного набора $0, 1, \dots, 2N$, а вероятности перехода из состояния $X_t = i$ в состояние $X_{t+1} = j$ описываются с помощью вероятности:

$$P_{ij} = \binom{2N}{j} \left(\frac{i}{2N} \right)^j \left(\frac{2N-i}{2N} \right)^{2N-j}. \quad (1.2)$$

Таким образом, изменение количества аллелей A в модели Райта-Фишера описывается Марковской цепью с матрицей переходов $\{P_{ij}\}$ и начальным состоянием X_0 . Другими словами, случайная величина X_t подчинена условному биномиальному распределению:

$$X_t | (X_{t-1} = x_{t-1}) \sim \text{Bin}(2N, \frac{x_{t-1}}{2N}),$$

а математическое ожидание и дисперсия X_t вычисляются следующим образом:

$$\begin{aligned} E[X_t] &= E[E[X_t | X_{t-1}]] = E\left[2N \frac{X_{t-1}}{2N}\right] \\ &= E[X_{t-1}] = \dots = E[X_0] = X_0. \end{aligned} \quad (1.3)$$

$$\begin{aligned}
D[X_t] &= E[X_t^2] - E[X_t]^2 = E[X_t^2] - X_0^2 = \\
&= E[E[X_t^2|X_{t-1}]] - X_0^2 = \\
&= E[D[X_t|X_{t-1}] + E[X_t|X_{t-1}]^2] - X_0^2 = \\
&= E\left[2N \frac{X_{t-1}(2N - X_{t-1})}{(2N)^2} + X_{t-1}^2\right] - X_0^2 = \\
&= E\left[X_{t-1}^2 \left(1 - \frac{1}{2N}\right) + X_{t-1}\right] - X_0^2 = \\
&= \left(1 - \frac{1}{2N}\right) E[X_{t-1}^2] + X_0 - X_0^2 = \\
&= \left(1 - \frac{1}{2N}\right) (D[X_{t-1}] + X_0^2) + X_0 - X_0^2 = \\
&= \left(1 - \frac{1}{2N}\right) D[X_{t-1}] + \left(1 - \frac{X_0}{2N}\right) X_0 = \dots = \\
&= \left(1 - \frac{1}{2N}\right)^t D[X_0] + \left(1 - \frac{X_0}{2N}\right) X_0 \cdot 2N \left(1 - \left(1 - \frac{1}{2N}\right)^t\right) = \\
&= X_0(2N - X_0) \left(1 - \left(1 - \frac{1}{2N}\right)^t\right).
\end{aligned} \tag{1.4}$$

Так как $1/2N$ можно считать малой величиной, то для выражения 1.4 применимо биномиальное приближение:

$$D[X_t] \approx X_0(2N - X_0) \frac{t}{2N}.$$

Пусть f_t - частота аллеля A в популяции в поколении t , $f_t = X_t/2N$, тогда математическое ожидание и дисперсия для f_t принимают следующий вид:

$$\begin{aligned}
E[f_t] &= f_0, \\
D[f_t] &= f_0(1 - f_0) \frac{t}{2N}.
\end{aligned} \tag{1.5}$$

Несмотря на простую математическую формулировку модели Райта-Фишера, не существует явного аналитического представления плотности распределения f_t в зависимости от времени [49; 50]. Кроме того, время в модели Райта-Фишера является дискретным и измеряется в поколениях, что не является реалистичным. Поэтому, были разработаны аппроксимации, которые можно разделить на две

группы: (1) использующие уравнения диффузии, и (2) использующие вероятностные распределения, моменты которых удовлетворяют выражениям 1.5. Для использования аппроксимаций первой группы необходимо получить аналитическое или численное решения диффузионных уравнений [51—53], в то время как аппроксимации второй группы являются более удобными в применении, и требуют лишь соответствия первых моментов 1.5 [14; 50; 54—59].

1.3.2 Аппроксимация модели Райта-Фишера с помощью Бета-распределения

Так как частота альтернативного аллеля, f_t , является случайной величиной, значения которой ограничены интервалом $[0,1]$, то естественным способом описания f_t является Бета-распределение [57—59]. Бета-распределение является двухпараметрическим с плотностью распределения следующего вида:

$$p(x|\alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1},$$

где α и β являются параметрами формы (англ. shape parameters) и принимают строго положительные значения. При такой параметризации математическое ожидание (μ) и дисперсия (v) случайной величины, имеющей Бета-распределение, принимают вид:

$$\mu = \frac{\alpha}{\alpha + \beta}; v = \frac{\alpha\beta}{(\alpha + \beta)^2(\alpha + \beta + 1)}. \quad (1.6)$$

Важным свойством Бета-распределения является связь между μ и v , вытекающая из 1.6:

$$v = \mu(1 - \mu) \frac{1}{(\alpha + \beta + 1)},$$

то есть

$$v < \mu(1 - \mu). \quad (1.7)$$

В качестве альтернативной параметризации Бета-распределения можно предложить использование пары (μ, v) вместо пары (α, β) с помощью замены переменных, получаемых после преобразования 1.6:

$$\alpha = \mu \left(\frac{\mu(1 - \mu)}{v} - 1 \right); \beta = (1 - \mu) \left(\frac{\mu(1 - \mu)}{v} - 1 \right). \quad (1.8)$$

Для того, чтобы Бета-распределение аппроксимировало распределение f_t в Модели Райта-Фишера, то есть выполнялось 1.5, необходимо положить $\mu = f_0$, а $v = \mu(1 - \mu)t/2N$. Тогда распределение f_t можно аппроксимировать с помощью Бета-распределения с параметрами:

$$\alpha = f_0 \left(\frac{2N}{t} - 1 \right); \beta = (1 - f_0) \left(\frac{2N}{t} - 1 \right). \quad (1.9)$$

С учетом 1.7, параметризация 1.9 имеет место в случае ограничения $t < 2N$, которое отсутствует в модели Райта-Фишера. Кроме того, Марковская цепь в модели Райта-Фишера имеет два поглощающих состояния – $f_t = 0$ и $f_t = 1$. Однако, плотность вероятности Бета-распределения на краях интервала $[0, 1]$ равна нулю. Для решения этой проблемы было предложено использовать модель Бета-распределения со спайками (англ. spikes), в которой плотность распределения для f_t представляет собой взвешенную сумму плотности Бета-распределения и двух функций Дирака, расположенных в 0 и 1 [50].

Еще одним недостатком является трудность в описании совместного распределения нескольких не независимых случайных величин, каждая из которых имеет Бета-распределение. Необходимость в таком совместном распределении возникает в задачах описания изменения частот аллелей в популяциях, которые имели частично общую эволюционную историю. Для того, чтобы описать совместное распределение двух не независимых случайных величин, имеющих Бета-распределение, были предложены подходы со сложной параметризацией [60; 61] и с использованием копула-функций [62]. Однако интерпретация предложенных подходов в терминах популяционной генетики неочевидна и требует дополнительных исследований. Кроме того, в реальных задачах популяционной генетики необходимо рассмотрение совместного распределения большего количества случайных величин, чем двух.

1.3.3 Аппроксимация модели Райта-Фишера с помощью нормального распределения

Наиболее распространенной аппроксимацией модели Райта-Фишера является аппроксимация нормальным распределением. В таком подходе предполагается, что частота альтернативного аллеля имеет нормальное распределение со средним и дисперсией, соответствующим выражению 1.5:

$$f_t \sim \mathcal{N} \left(f_0, f_0(1 - f_0) \frac{t}{2N} \right). \quad (1.10)$$

Недостатком такой аппроксимации является то, что нормальное распределение определено для всех действительных чисел, а значения f_t не могут быть отрицательными и превышать 1. Таким образом, нормальное приближение целесообразно использовать, при моделировании частот аллелей с промежуточными значениями и когда объем дрейфа невелик. Чтобы значения f_t не выходили за допустимые границы, нормальное распределение заменяют усеченным нормальным:

$$f_t \sim \text{trunc}\mathcal{N} \left(f_0, f_0(1 - f_0) \frac{t}{2N} \right).$$

Преимущество использования нормальной аппроксимации заключается в том, что она легко расширяется на многомерный случай [14]. Предположим, что есть четыре популяции с общей историей (Рисунок 1.4А), и частоты альтернативного аллеля в одном локусе равны f_1, f_2, f_3 и f_4 , соответственно. Пусть частота аллеля у общей предковой популяции была f_0 , тогда:

$$\forall i \in \overline{1,4} : f_i \sim \mathcal{N} \left(f_0, f_0(1 - f_0) \frac{t_{0,i}}{2N} \right),$$

где $t_{0,i}$ - объем генетического дрейфа, накопившегося между предковой популяцией и i -ой популяцией.

Так как, например, популяции f_1 и f_3 имели общую историю между предковым состоянием 0 и состоянием y , то объем общего генетического дрейфа между этими популяциями будет равно $t_{0,y}$. Аналогичным образом можно получить объем дрейфа между каждой парой популяций, и совместное распределение частот аллелей данных четырех популяций будет выражаться следующим образом:

$$\begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ f_4 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} f_0 \\ f_0 \\ f_0 \\ f_0 \end{pmatrix}, f_0(1 - f_0)V \right), \quad (1.11)$$

где матрица V представлена на Рисунке 1.4Б.

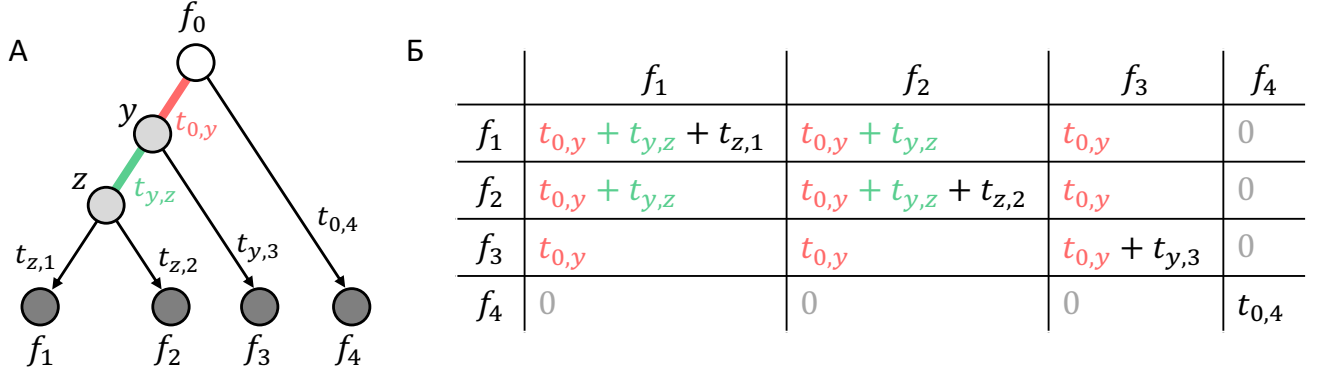


Рисунок 1.4 — Пример расчета генетического дрейфа между популяциями.

(А) Пример эволюционной истории, связывающей четыре популяции. Переменные на ветвях соответствуют объему генетического дрейфа, накопленному на соответствующем промежутке истории. **(Б)** Матрица ковариаций между популяциями, выраженная через введенные переменные.

Чтобы значения $f_i, i \in \overline{1,4}$ не выходили за границы отрезка $[0,1]$, многомерное нормальное распределение 1.11 необходимо заменить на усеченное нормальное. Обобщение этой модели на случай произвольной матрицы V приведено в работе [63].

1.4 Анализ композиционных данных

Прежде чем проводить анализ данных, необходимо исследовать свойства пространства, к которому эти данные принадлежат. Это позволяет адекватно подобрать распределения, наиболее точно отражающие характеристики исследуемых данных, строить и проверять статистические гипотезы с учетом этих

распределений, а также аккуратно разрабатывать математические модели. В традиционной иерархии типов данных первый уровень представляет собой разделение на числовые и категориальные. Затем числовые данные подразделяются на непрерывные и дискретные, а категориальные - на номинальные и порядковые. При работе с числовыми непрерывными данными часто возникает искушение не углубляться в иерархию и предположить их нормальное распределение, поскольку существует множество эффективных и удобных методов статистического анализа, разработанных специально для нормально распределенных данных. Однако, в зависимости от того, какие значения могут принимать числовые непрерывные данные - числа из всего вещественного пространства, положительного ортанта, гиперкуба, гиперсферы или симплекса - допустимые операции над данными, подходы к их анализу, обработке, допустимость применения статистических методов могут существенно отличаться.

В Главе 2 настоящей диссертации приведено описание двух разработанных автором моделей популяционной генетики, каждая из которых оперирует с частотами аллелей в популяциях. Частоты представляют собой числа от 0 до 1, то есть они строго ограничены на вещественной оси. Кроме того, частоты всех аллелей одного генетического локуса суммируются к единице, то есть являются не независимыми и представляют собой так называемую композицию. Поэтому описание частот с помощью нормального распределения в моделях генетического дрейфа, как, например, в параграфе 1.3.3, является концептуально неточным. Для работы с композициями в 1982 году Джоном Эйтчисоном (англ. John Aitchison) была предложена теоретическая основа [7], которая впоследствии стала фундаментом для разработки множества методов и подходов для анализа таких данных. Это позволило значительно повысить точность и надежность статистических выводов в различных областях применения, от геологии и экологии до социальных наук и экономики [64]. В частности, автор настоящей диссертации применял методы анализа композиционных данных для исследования метагеномных данных [65]. Разработанные методы для анализа композиций лежат и в основе моделей, представленных в настоящей диссертации.

Композиционные данные представляют собой доли, частоты или проценты отдельных компонент, которые составляют нечто целое. Эти данные обычно получаются путем деления наблюдаемых положительных значений на число, пропорциональное их сумме. Не умаляя общности, будем считать, что одна наблюдаемая композиция представляет собой вектор из строго положительных

значений, сумма которых равна единице. В этом случае говорят, что к наблюдаемым значениям $\mathbf{z} = [z_1, \dots, z_D] \in \mathbb{R}_+^D$ применена операция замыкания (англ. closure):

$$\mathbf{x} = \mathcal{C}(\mathbf{z}) = \left[\frac{z_1}{\sum_{i=1}^D \hat{z}_i}, \frac{z_2}{\sum_{i=1}^D \hat{z}_i}, \dots, \frac{z_D}{\sum_{i=1}^D \hat{z}_i} \right].$$

Все такие композиции $\mathbf{x}$ длины D располагаются в открытом единичном симплексе S^D :

$$S^D = \left\{ \mathbf{x} = [x_1, \dots, x_D] \in \mathbb{R}^D \mid x_i > 0, \sum_{i=1}^D x_i = 1 \right\}.$$

На симплексе может быть задано векторное пространство, например, как это предложил Эйтчисон:

$$\begin{aligned} \mathbf{x} \oplus \mathbf{y} &= \mathcal{C} [x_1 y_1, x_2 y_2, \dots, x_D y_D], \\ \alpha \otimes \mathbf{x} &= \mathcal{C} [x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha], \\ \langle \mathbf{x}, \mathbf{y} \rangle_A &= \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \end{aligned} \tag{1.12}$$

где $\mathbf{x}$ и $\mathbf{y}$ - композиции, α - скаляр, $\langle \cdot, \cdot \rangle_A$ - скалярное произведение Эйтчисона.

Визуально можно представить симплекс размерности 3 в виде треугольника (Рисунок 1.5 А), каждой внутренней точке которого однозначно соответствует композиция длины 3. Границам треугольника не соответствуют композиции в предложенном определении, так как иначе некоторые элементы соответствующих композиций были бы равны нулю. Выборочное среднее для набора композиций $\{\mathbf{x}_k\}_{k \in \overline{1..N}}$ из S^D в смысле Эйтчисона имеет следующий вид:

$$\mu = \frac{1}{N} \otimes \bigoplus_{k=1}^N \mathbf{x}_k.$$

Можно показать, что μ соответствует среднему геометрическому, а не среднему арифметическому в привычном представлении. Среднее арифметическое набора композиций нередко действительно находится вне "облака" этих композиций, что можно подкрепить примером на Рисунке 1.5. Поэтому в качестве усреднения композиционных данных рекомендуется использовать именно геометрическое среднее, а не арифметическое.

Распределений, имеющих своим доменом симплекс, не так много. Самым известным, пожалуй, является распределение Дирихле. Однако, это распределение не подходит для описания композиционных данных со сложной структурой

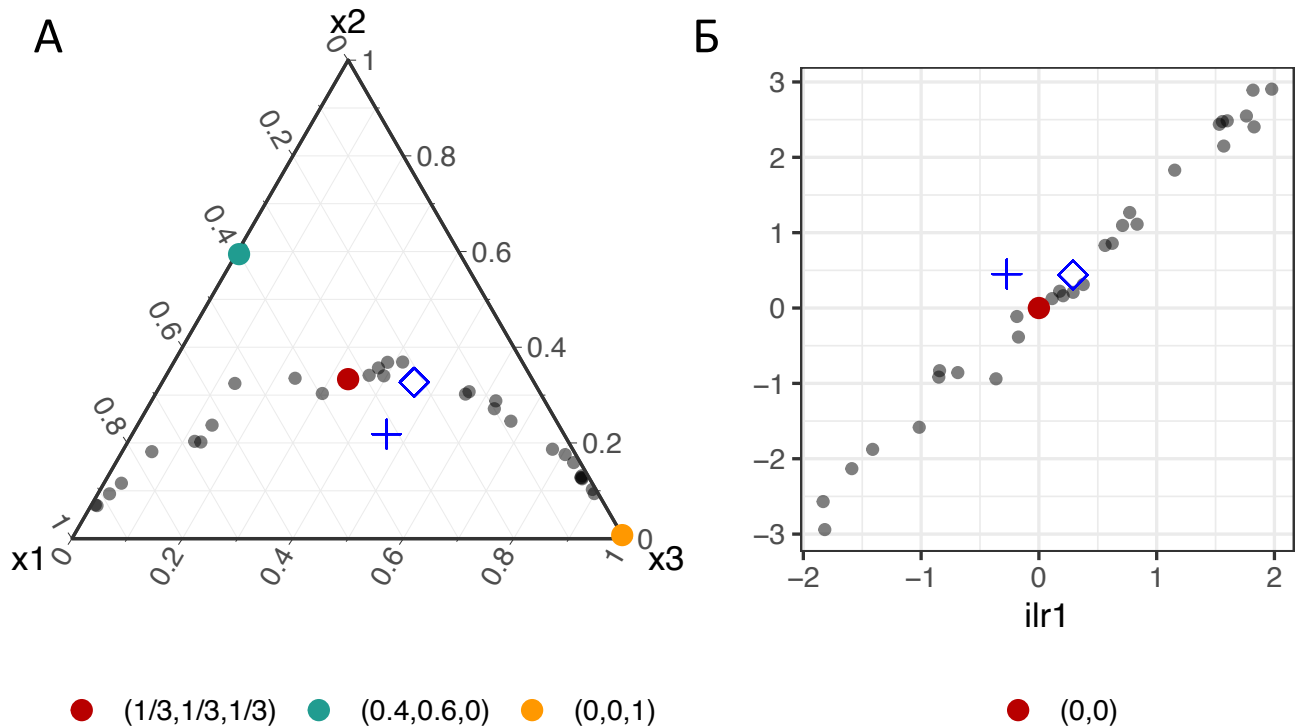


Рисунок 1.5 — **Различные представления композиционных данных.**

(А) Представление 30 случайных композиций длины три (серые полупрозрачные точки) в барицентрических координатах. Ромб показывает геометрическое среднее, крест - среднее арифметическое. Красная точка - равномерная композиция, соответствующая центру симплекса. Зеленая и желтая точки не соответствуют композициям, так как некоторые координаты равны нулю. **(Б)** Представление преобразованных композиций из панели А в ilr координатах.

зависимостей [66]. Другим классом распределений является логистически-нормальное. Было показано, что логистически-нормальное распределение близко к любому распределению Дирихле с точки зрения дивергенции Кульбака-Лейблера [64; 67]. Кроме того, использование логистически-нормального распределения открывает полный спектр линейного моделирования, доступного для многомерного нормального распределения в $\mathbb{R}^D$ [64].

Для того, чтобы использовать нормальное распределение при моделировании процессов, происходящих с композиционными данными, необходимо провести специфическое преобразование этих данных. Так как элементы композиции суммируются к единице, то они не являются независимыми, то есть из композиции длины D необходимо некоторым образом составить набор из незави-

симых компонент. В попытках сделать это были предложены три преобразования: alr , clr и ilr (англ. additive, center и isometric logratio transforms, соответственно). Преобразование alr переводит композиции из симплекса S^D в пространство $\mathbb{R}^{D-1}$, принимая один из элементов композиции в качестве нормализующего:

$$\begin{aligned}\text{alr}(\mathbf{x}) &= \left[\log \frac{x_1}{x_D} \quad \dots \quad \log \frac{x_{D-1}}{x_D} \right] = \varphi, \\ \text{alr}^{-1}(\varphi) &= \mathcal{C} [\exp(\varphi_1), \dots, \exp(\varphi_{D-1}), 1].\end{aligned}$$

Преобразование симплекса S^D в $\mathbb{R}^{D-1}$ открывает возможность для применения полного спектра статистических инструментов классической статистики и методов многомерного анализа в евклидовом пространстве. Недостатками такого преобразования являются отсутствие симметрии и изменение расстояния между композициями при различных выборах нормализующего элемента [8].

Преобразование clr является симметричным, и переводит композиции из симплекса S^D в пространство $\mathbb{R}^D$:

$$\begin{aligned}\text{clr}(\mathbf{x}) &= \left[\log \frac{x_1}{g(\mathbf{x})} \quad \dots \quad \log \frac{x_D}{g(\mathbf{x})} \right] = \xi, \\ \text{clr}^{-1}(\xi) &= \mathcal{C} [\exp(\xi_1), \dots, \exp(\xi_D)],\end{aligned}$$

где $g(\mathbf{x})$ - геометрическое среднее элементов композиции $\mathbf{x}$. Так как D исходных элементов композиции не являются независимыми (степень свободы $D - 1$), то и D полученных элементов ξ обладают тем же свойством. В результате ковариационная матрица, рассчитанная на векторах clr -преобразования, будет вырожденной. Другим недостатком clr -преобразования является отсутствие субкомпозиционной согласованности (англ. subcompositional coherence) [8]. В то время как в Евклидовой геометрии длина проекции некоторого сегмента всегда меньше или равна длине исходного сегмента, субкомпозиционная согласованность предполагает, что расстояние между двумя композициями должно быть больше либо равно расстоянию между их соответствующими субкомпозициями:

$$\begin{aligned}d((x_1, x_2, \dots, x_D), (y_1, y_2, \dots, y_D)) &\geq \\ d((x_1, x_2, \dots, x_{i-1}, x_{i+1}, \dots, x_D), (y_1, y_2, \dots, y_{i-1}, y_{i+1}, \dots, y_D)).\end{aligned}$$

Преобразование ilr было предложено в 2003 году, и оно не обладает недостатками, присущими alr и clr [68]. Кроме всего прочего это преобразование позволяет ввести на симплексе ортогональную систему координат. Пусть

$\{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_{D-1}\}$ ортонормированный базис на симплексе S^D , то есть $\langle \mathbf{e}_i, \mathbf{e}_j \rangle_A = 0$, если $i \neq j$, и $\langle \mathbf{e}_i, \mathbf{e}_i \rangle_A = 1$. Тогда любую композицию $\mathbf{x}$ можно представить в этом базисе следующим образом:

$$\mathbf{x} = \bigoplus_{i=1}^{D-1} x_i^* \odot \mathbf{e}_i, \quad (1.13)$$

где x_i^* - вещественные числа. Тогда композиция $\mathbf{x}$ имеет координаты $(x_1^*, \dots, x_{D-1}^*)$ в введенном ортонормированном базисе. По ортогональному базису можно построить так называемую базисную матрицу Ψ размера $(D-1 \times D)$, строки которой представляют собой clr-преобразования элементов этого базиса. А само преобразование ilr выглядит следующим образом:

$$\text{ilr}(\mathbf{x}) = \text{clr}(\mathbf{x}) \Psi^T.$$

Учитывая 1.12, и 1.13 можно получить, что $\text{ilt}(\mathbf{x}) = [x_1^*, \dots, x_{D-1}^*] \in \mathbb{R}^{D-1}$. Таким образом все преобразования в ортогональном базисе в векторном пространстве $\mathbb{R}^{D-1}$ соответствуют преобразованиям в ортогональном базисе на симплексе S^D [8].

В качестве иллюстрации на Рисунке 1.5Б показано преобразование ilr для композиций из Рисунка 1.5А. Видно, что композиции, которые располагаются в барицентрических координатах как бы вдоль дуги, лежат вдоль прямой, что значительно упрощает работу с данными.

Несмотря на то, что ilr -преобразование кажется зависимым от выбранного ортогонального базиса, на самом деле любое ilr -преобразование при одном базисе может быть приведено к ilr -преобразованию при другом базисе с помощью матрицы поворота. Это означает, что все ilr -преобразования суть одинаковы с точностью до поворота. Однако, при такой произвольности в выборе базиса, хотелось бы его составить так, чтобы построенные оси в $\mathbb{R}^{D-1}$ были осмысленными, а получаемые результаты легко интерпретируемыми. Для этого был разработан подход последовательного бинарного разбиения (англ. sequential binary partition). В этом подходе строится бинарное укорененное дерево с D листьями и, соответственно, $(D-1)$ внутренними узлами. Построение дерева и назначение каждому листу дерева одного элемента композиции происходит либо случайно, либо исходя из каких-нибудь априорных соображений, например, смыслового родства элементов. Каждому внутреннему узлу ставится в соответствие базисный вектор таким образом, что координаты на соответствующей оси в $\mathbb{R}^{D-1}$ пропорциональны логарифму отношения средних геометрических значений в листьях в правой и левой

клад этого внутреннего узла. Смысл каждой такой координаты - баланс между двумя наборами элементов композиции. Если композицию составляют всего два элемента, то ось будет соответствовать логарифму отношения между элементами, то есть логистическому преобразованию. Если композицию составляют три элемента, то первой осью будет баланс между, например, первым и вторым элементом, а второй осью - баланс между третьим и первыми двумя.

При случайном выборе бинарного дерева реальные и значимые контрасты в данных могут размываться или почти исчезать, поэтому рекомендуется подходить к выбору дерева осмысленно. В качестве примерной схемы анализа композиционных данных можно предложить следующую последовательность шагов. Рекомендуется выполнить ilr -преобразование композиционных данных, используя осмысленные балансы, провести анализ в соответствующем $\mathbb{R}^{D-1}$ пространстве и затем трактовать результаты анализа в контексте элементов исходной композиции.

Проблемой работы с композиционными данными являются нулевые значения элементов композиции. Нули могут означать отсутствие наблюдения из-за нечувствительности измерения, тогда их можно заменить на малое положительное число. Либо нули могут означать отсутствующие данные, тогда их можно заменить с помощью ЕМ-алгоритма (англ. Expectation-maximization (EM) algorithm) [64] или выбросить.

1.5 Байесовский подход

В данной диссертации, в Главе 2, представлена новая модель для оценки характеристических частот аллелей в популяции. Эта модель включает в себя множество параметров, что обусловило необходимость применения байесовского метода для их оценки. Байесовский подход к оценке параметров модели можно рассматривать с точки зрения формул как надстройку над методом максимального правдоподобия. Концептуальное отличие заключается в том, как в этих методах воспринимаются параметры модели. В результате в байесовском подходе снимается условие на необходимый размер выборки для оценки параметров. Детальное объяснение этого подхода приведено ниже.

В байесовском подходе все параметры модели считаются случайными величинами с заданными априорными распределениями, которые играют роль регуляризации параметров. Пусть θ - набор всех параметров модели, D - наблюдаемая выборка. Так как в Байесовском подходе, в отличие от метода максимального правдоподобия, параметры θ являются случайными величинами, то в качестве целевой функции для максимизации рассматривается не плотность условного распределения $p(D|\theta)$, а совместная плотность распределения параметров и данных $p(\theta, D)$. Используя формулу для условной вероятности, справедливо следующее преобразование:

$$p(\theta, D) = p(\theta|D)p(D) \propto p(\theta|D)$$

Выражение $p(\theta|D)$ называется апостериорным распределением параметров θ , и максимизация $p(\theta, D)$ эквивалентна максимизации апостериорного распределения параметров.

Используя теорему Байеса, справедлив следующий вывод:

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \propto p(D|\theta)p(\theta),$$

где плотность распределения $p(\theta)$ называется плотностью априорного распределения параметров. В результате задача максимизации апостериорного распределения параметров соответствует задаче максимизации произведения правдоподобия данных и априорного распределения параметров.

Оценка параметров модели в байесовском подходе производится обычно с помощью методов Монте Карло по схеме Марковской цепи (англ. Markov Chain Monte Carlo, MCMC). В процессе работы этих методов происходит запуск Марковской цепи с последовательным наращиванием выборки значений параметров [69]. Окончательная оценка параметров модели определяется на основе среднего значения, вычисленного по этой выборке. Процесс создания выборки включает в себя повторяющиеся шаги, на каждом из которых генерируются значения для всех параметров модели, основываясь на значениях, полученных на предыдущем шаге. Пусть на шаге $t - 1$ были получены значения вектора параметров $\theta^{(t-1)} = (\theta_1^{(t-1)}, \theta_2^{(t-1)}, \dots, \theta_k^{(t-1)})$, тогда на шаге t генерация значений параметров выглядит следующим образом:

$$\begin{aligned}
\theta_1^{(t)} &\sim p\left(\theta_1|\theta_2^{(t-1)}, \theta_3^{(t-1)} \dots, \theta_k^{(t-1)}\right) \\
\theta_2^{(t)} &\sim p\left(\theta_2|\theta_1^{(t)}, \theta_3^{(t-1)} \dots, \theta_k^{(t-1)}\right) \\
&\dots \\
\theta_k^{(t)} &\sim p\left(\theta_k|\theta_1^{(t)}, \theta_2^{(t)} \dots, \theta_{k-1}^{(t)}\right)
\end{aligned}$$

Если апостериорное распределение параметра θ_i принадлежит известному классу распределений, и для этого класса распределений существует алгоритм генерации выборки, то новое значение для параметра $\theta_i^{(t)}$ получается с помощью этого алгоритма. Такой способ получения параметра называется схемой Гиббса. В том случае, если алгоритм генерации выборки из апостериорного распределения параметра θ неизвестен, то получение нового значения $\theta_i^{(t)}$ реализуется согласно схеме Метрополиса-Гастингса [69]. В этой схеме для каждого параметра определяется вспомогательная зависящая от $\theta_i^{(t-1)}$ функция распределения $q(\theta_i^*|\theta_i^{(t-1)})$, для которой существует алгоритм генерации нового значения θ_i^* , учитывая старое. Грубо говоря, с помощью этой функции можно сделать скачок к следующему пробному значению вокруг предыдущего. После того, как получено новое значение θ_i^* производится вычисление следующей величины:

$$u = \frac{p\left(\theta_i^*|\theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)} \dots, \theta_k^{(t-1)}\right) q\left(\theta_i^{(t-1)}|\theta_i^*\right)}{p\left(\theta_i^{(t-1)}|\theta_1^{(t)}, \dots, \theta_{i-1}^{(t)}, \theta_{i+1}^{(t-1)} \dots, \theta_k^{(t-1)}\right) q\left(\theta_i^*|\theta_i^{(t-1)}\right)}.$$

С вероятностью $\min(u, 1)$ значение θ_i^* принимается за $\theta_i^{(t)}$, иначе $\theta_i^{(t)}$ полагается равным старому значению $\theta_i^{(t-1)}$. Так как при реализации схемы Метрополиса-Гастингса значение параметра не всегда обновляется, то получение оценок параметров по этой схеме зачастую занимает больше времени, чем при реализации схемы Гиббса. Поэтому последняя схема является предпочтительной, однако не всегда возможной.

Усовершенствование схемы Метрополиса-Гастингса является схема Гамильтониан Монте Карло (англ. Hamiltonian Monte Carlo), которая использует уравнения Гамильтона для генерации пробных значений ([70]). В классической механике, уравнения Гамильтона - это система дифференциальных уравнений, описывающих движение системы в обобщенных координатах (положение и импульс). Чтобы сгенерировать пробное значение по схеме Гамильтониан Монте

Карло, сначала, случайно выбирается значение импульса, затем, производится симуляция динамики с помощью численного решения уравнений Гамильтона. Пробное значение по результатам симуляции принимается или отвергается по тому же принципу, что и в схеме Метрополиса-Гастингса. При симуляции динамики системы используются аналитические выражения для производных апостериорных вероятностей, чтобы избежать численного дифференцирования и ускорить численную схему. То есть схема Гамильтониан Монте Карло принимает в качестве параметров аналитические выражения для производных $p(\theta|D)$ по параметрам. Преимущество схемы Гамильтониан Монте Карло заключается в том, что она может преодолевать проблему автокорреляции, которая часто возникает при использовании других схем.

Для получения оценок параметров с помощью метода Монте Карло по схеме марковской цепи производится запуск нескольких цепи одинаковой длины. Анализ различий в значениях полученных цепей проводится с помощью многомерной диагностики Гельмана-Рубина [71] следующим образом. Пусть был произведен запуск M цепей длины N и для каждого параметра θ_i был получен набор значений в каждой цепи $\{\theta_i^{(t,m)}\}_{t=1,N}, \forall m = \overline{1,M}$. Пусть $\hat{\theta}_i^{(m)}$ и $\hat{\sigma}_i^{(m)}$ - среднее и дисперсия параметра θ_i в m -ой цепи, а среднее значение $\hat{\theta} = \frac{1}{M} \sum_{m=1,M} \hat{\theta}_i^{(m)}$. Тогда вариабельность между цепями, B (от англ. between), и вариабельность внутри цепей, W (от англ. within), можно вычислить следующим образом:

$$B = \frac{N}{M-1} \sum_{m=1}^M (\hat{\theta}_i^{(m)} - \hat{\theta})^2$$

$$W = \frac{1}{M} \sum_{m=1}^M \hat{\sigma}_i^{(m)}$$

Суммарная вариабельность V рассчитывается как взвешенная сумма B и W :

$$V = \frac{N-1}{N} W + \frac{M+1}{MN} B$$

Критерием того, что все цепи сходятся к одному распределению оценивается близостью отношения V/W к единице или с помощью t-критерия Стьюдента [72].

1.6 Выводы к главе 1

Задача реконструкции эволюционных и исторических связей между различными популяциями не является новой, а даже существует ряд программ, которые решают эту задачу с определенными допущениями. Тем не менее, чем глубже рассматривать эту задачу, тем более очевидным становится, что существующие методы могут быть не всегда применимы, они могут давать неточные результаты в своих аппроксимациях, и, кроме того, исходная задача может быть разделена на самостоятельные подзадачи, каждая из которых требует своего уникального подхода.

Известные методы для реконструкции эволюционных связей между популяциями берут в качестве входных данных для каждой популяции генотипированные индивидумы, предполагая, что это они являются репрезентативной выборкой из панмиксичной популяции. Однако, в реальности отклонение от этих условий является более вероятным, чем их строгое соблюдение. В реальных условиях популяции часто сталкиваются с географической гетерогенностью ареала, которая формирует неравномерные паттерны миграции. Кроме того, в реальных выборках может наблюдаться неравномерность местоположений точек сбора образцов. Эти аспекты могут оказывать значительное влияние на структуру выборки, и их необходимо учитывать при реконструкции эволюционной истории.

В связи с этим возникает подзадача оценки характеристических частот аллелей в популяции с учетом географического местоположения или любых других факторов, связывающих образцы. Простое выборочное среднее очевидно не является характеристической оценкой частот аллелей в популяциях, а существующие методы, опосредованно оценивающие частоты аллелей, так же не используют дополнительную информацию (например, координаты точек сбора образцов). Поэтому существует запрос на метод, который бы справлялся с поставленной подзадачей.

При оценке частот аллелей в репрезентативных выборках важно учитывать взаимосвязь между образцами. В зависимости от природы исследуемой выборки, эта связь может быть обусловлена различными факторами: географическим положением, общими историческими миграционными путями, а также экологическими и культурными особенностями. По каждому из этих факторов можно составить матрицу расстояний между образцами и учитывать ее в модели. Од-

нако, матрица расстояний не фигурирует в распределениях, и зачастую требует перевода в ковариационную матрицу. Один из способов это сделать предложен в работах [21; 31], однако, это нелинейное преобразование требует дополнительной оценки параметров. Поэтому возникает необходимость пересмотра концепта перевода матрицы расстояний в ковариационную матрицу, в особенности в контексте географических расстояний.

Реконструкция эволюционной истории популяций заключается в построении графа смещений. Однако модель события смещения в этом графе отличается в различных программах. Например, некоторые программы рассматривают события смещения только между двумя популяциями (TreeMix, MixMapper, findGraph), или они ограничены по вложенным смещениям (TreeMix) и по глубине этих вложений (MixMapper), а также отсутствует возможность перебора событий смещения с несколькими источниками (qpGraph). В связи с этим возникает необходимость пересмотра модели смещения для преодоления указанных ограничений.

Перечисленные выше усовершенствования методов связаны в основном с углублением в понимание многогранности задачи реконструкции исторических связей между популяциями. Однако ключевым усовершенствованием текущих методов, которое может значительно продвинуть область вперед, является анализ композиционных данных. Поскольку частоты всех аллелей в каком-либо локусе в сумме дают единицу, то есть представляют собой композицию, возникает необходимость использования композиционного анализа данных. А именно, необходимо пересмотреть модель генетического дрейфа с точки зрения композиционного анализа данных. Это позволит перейти от биаллельных локусов к мульталлельным, что представляет особый интерес при анализе, например, структурных вариантов или гаплоблоков.

Глава 2. Разработка моделей для реконструкции эволюционной истории популяций

Для реконструкции эволюционной истории распространения и смещений популяций некоторого вида был разработан двухступенчатый подход. Сначала для каждой из популяций проводилась оценка характеристических частот аллелей в географическом регионе с помощью модели `popdisp`. Затем на основе полученных частот осуществлялась реконструкция эволюционной истории популяций с помощью модели `migadmi`. Обе модели могут быть применены в более широком контексте и не ограничиваются исследованием географических перемещений. Поэтому приведенное в этой главе описание моделей не опирается на географические местоположения индивидуумов в популяции.

2.1 Композиционная аппроксимация модели Райта-Фишера

2.1.1 Формальное описание входных данных

Пусть M - количество популяций некоторого вида. Для каждой из популяций была произведена выборка из индивидуумов, которые были генотипированы по единой панели из N несвязных друг с другом (отсутствие неравновесия по сцеплению) биаллельных локусов (например, ОНП). Для определенности назовем один из аллелей каждого локуса референсным, а второй - альтернативным. Каждый из биаллельных локусов может принимать одно из четырех значений: 0 (гомозигота по референсной аллели), 1 (гетерозигота), 2 (гомозигота по альтернативной аллели) и NA (отсутствие данных). Из-за отсутствия данных по некоторым локусам у некоторых индивидуумов, общее количество известных аллелей в локусах в каждой j -ой популяции неодинаково. Таким образом для каждого i -ого локуса и j -ой популяции даны два числа: n_j^i - общее количество аллелей, и y_j^i - количество альтернативных аллелей.

Предположим, что в каждой популяции для каждого локуса существует своя частота альтернативного аллеля, f_j^i . В силу отсутствия сцепления наблюдаемое

значение y_j^i является результатом случайной выборки из n_j^i независимых повторений, тогда y_j^i следует биномиальному распределению: $y_j^i \sim \text{Bin}(n_j^i, f_j^i)$. Все рассматриваемые популяции произошли от некоторой одной, которая в свою очередь тоже характеризуется частотами аллелей в каждом i -ом локусе, f_A^i . При этом связь между f_j^i и f_A^i описывается с помощью ковариационной матрицы V , пропорциональной объему генетического дрейфа, накопленного в f_j^i относительно f_A^i . Согласно модели Райта-Фишера необходимо, чтобы два первых момента для f_j^i соответствовали выражениям 1.5, то есть:

$$\begin{aligned}\mathbb{E}[f_j^i] &= f_A^i, \\ \text{var}[f_j^i] &= f_A^i(1 - f_A^i)s^i V_{jj},\end{aligned}\tag{2.1}$$

где s^i - некоторая константа пропорциональности. Константа s^i здесь используется, потому что (i) эффективный размер популяции неизвестен, и (ii) матрица V отражает время (количество поколений) в модели Райта-Фишера не напрямую, а предполагается только пропорциональной объему генетического дрейфа.

2.1.2 Вывод аппроксимации

При рассмотрении частот, всегда нужно иметь в виду, что они образуют так называемую композицию, то есть набор относительных значений или долей, которые представляют собой пропорциональные части целого. Композиционные данные характеризуются тем, что сумма их компонентов равна константе. В нашем случае частота референсного и альтернативного аллеля суммируются к единице. Поэтому, чтобы не строить модель вокруг какого-то одного компонента композиции (частоты альтернативного или референсного аллеля), можно воспользоваться теорией композиционного анализа данных и преобразовать зависимые частоты в балансы с помощью ilr -преобразования (см. Параграф 1.4). Для случая биаллельных локусов это преобразование является просто logit функцией:

$$x_j^i = g(f_j^i) = \ln \left(\frac{1 - f_j^i}{f_j^i} \right),$$

Полученные балансы $\{x_j^i\}_{i \in \overline{1, M}}$ и x_A^i могут принимать значения от $-\infty$ до $+\infty$, что позволяет моделировать корреляции между частотами аллелей, ис-

пользуя многомерные нормальные распределения без искусственного усечения, которое необходимо при работе с частотами (см. Параграф 1.3.3):

$$\begin{pmatrix} x_1^i \\ \dots \\ x_M^i \end{pmatrix} \sim mv\mathcal{N} \left(\begin{pmatrix} x_A^i \\ \dots \\ x_A^i \end{pmatrix}, \sigma_i V \right), \quad (2.2)$$

где x_A^i и σ_i - параметры, а матрица V отражает заданную связь между популяциями. Маргинализация распределения 2.2 выглядит следующим образом: $x_j^i \sim \mathcal{N}(x_A^i, \sigma_i V_{jj})$. Так как x_j^i есть функция от f_j^i , и должно выполняться условие 2.1, то можно получить приблизительные оценки для x_A^i и σ_i используя разложения в ряд Тейлора моментов функций случайных величин ([73]):

$$\begin{aligned} \sigma_i V_{jj} &= \text{var}[x_j^i] = \text{var}[g(f_j^i)] \approx \text{var}[f_j^i] (g'(\mathbb{E}[f_j^i]))^2 = \\ &= f_A^i (1 - f_A^i) s^i V_{jj} ((\ln(1 - f_A^i) - \ln(f_A^i)))'^2 = \\ &= f_A^i (1 - f_A^i) \left(\frac{1}{1 - f_A^i} - \frac{1}{f_A^i} \right)^2 s^i V_{jj} = \\ &= \frac{s^i V_{jj}}{(f_A^i (1 - f_A^i))}. \end{aligned}$$

Таким образом, $\sigma_i \approx s^i / (f_A^i (1 - f_A^i))$.

Разложение в ряд Тейлора для математического ожидания x_j^i выглядит следующим образом:

$$\begin{aligned} x_A^i &= \mathbb{E}[x_j^i] \approx g(\mathbb{E}[f_j^i]) + \frac{\text{var}[f_j^i]}{2} g''(\mathbb{E}[f_j^i]) = \\ &= \ln \left(\frac{1 - f_A^i}{f_A^i} \right) + \frac{f_A^i (1 - f_A^i) s^i V_{jj}}{2} \left(\frac{1}{(f_A^i)^2} - \frac{1}{(1 - f_A^i)^2} \right) \\ &= \ln \left(\frac{1 - f_A^i}{f_A^i} \right) + \frac{f_A^i (1 - f_A^i) s^i V_{jj}}{2} \frac{1 - 2f_A^i}{(f_A^i (1 - f_A^i))^2} = \\ &= \ln \left(\frac{1 - f_A^i}{f_A^i} \right) + (0.5 - f_A^i) \text{var}[x_j^i]. \end{aligned}$$

Приведенный вывод показывает, что $x_A^i \approx g(f_A^i)$ только при линейной аппроксимации. Однако, так как в модели присутствует два перехода - из пространства частот в пространство балансов и обратно - будем использовать эту аппроксимацию для простоты. В будущем это место в модели может быть использовано для ее улучшения.

В результате для описания дрейфа в пространстве балансов предлагается следующая многомерная нормальная аппроксимирующая модели Райта-Фишера:

$$\vec{x}^i \sim mv\mathcal{N}\left(\mathbf{1}_M x_A^i, \frac{s^i V}{(f_A^i(1 - f_A^i))}\right), \quad (2.3)$$

где $\vec{x}^i = (x_1^i, x_2^i, \dots, x_M^i)$, $\mathbf{1}_M$ - единичный вектор длины M .

2.2 Оценка характеристических частот аллелей и помощью модели popdisp

Модель popdisp является Байесовской иерархической моделью для оценки частот аллелей в популяции с учетом ковариационной матрицы между индивидами. Предполагается, что ковариационная матрица отражает генетический дрейф из некоторого генетического центроида, частоты аллелей в котором будут характеристическими для популяции. Для моделирования генетического дрейфа, popdisp использует аппроксимацию модели Райта-Фишера согласно композиционному анализу данных (см. Параграф 1.4), что позволяет расширить popdisp на случай работы с многоаллельными генетическими локусами. В качестве входных данных модель использует генотипированные индивиды и матрицу ковариаций между ними.

Модель popdisp может быть применена не только для оценки характеристических частот аллелей. Например, эта модель может быть использована для исследования распределения популяции по географическому региону, позволяя сравнивать различные гипотезы о распространения популяции в этом регионе. В качестве генетического центроида популяции могут быть рассмотрены различные предполагаемые географические точки в регионе, а ковариационная матрица построена в соответствии с интересующими гипотезами распространения из центроида. На основании критериев оценки качества моделей (например, BIC, Bayesian Information Criterion), можно оценить, какой из центроидов и гипотез распространения наилучшим образом соответствует наблюдаемым генетическим данным.

2.2.1 Байесовская иерархическая модель

Построим байесовскую иерархическую модель для оценки частоты f_A^i при заданных y_j^i , n_j^i и V , где индексы i и j соответствуют генетическому локусу и популяции. Как было введено ранее, каждой популяции соответствует частота аллели f_j^i , так что:

$$y_j^i \sim \text{Bin}(n_j^i, f_j^i),$$

а частоты f_j^i были получены в результате дрейфа от частоты популяции f_A^i . Случайный дрейф будет смоделирован не в пространстве частот f_j^i , а в пространстве балансов x_j^i , при этом связь между этими пространствами описывается следующим образом:

$$f_j^i = \frac{1}{1 + \exp(x_j^i)}.$$

Совместное распределение $(x_1^i, x_2^i, \dots, x_M^i)$ 2.3 является многомерным нормальным согласно 2.3. Среднее этого распределения, x_A^i , является балансом для частоты f_A^i , поэтому:

$$x_A^i = \ln \left(\frac{1 - f_A^i}{f_A^i} \right).$$

Априорное распределение для частоты f_A^i является Бета-распределением, что естественно для величин, варьирующих от 0 до 1:

$$f_A^i \sim \text{Beta}(\alpha^i, \beta^i).$$

Мы будем использовать неинформативное распределение [63], полагая равномерное априорное распределение для среднего значения:

$$\alpha^i / (\alpha^i + \beta^i) \sim \text{Uniform}(0,1),$$

и экспоненциальное априорное распределение для так называемой ”концентрации” или ”размера выборки” (”sample size”):

$$\alpha^i + \beta^i \sim \text{Expon}(1).$$

Мы также предполагаем экспоненциальное априорное распределение для постоянной пропорциональности s^i :

$$s^i \sim \text{Expon}(1).$$

Построенная модель изображена на Рисунке 2.1.

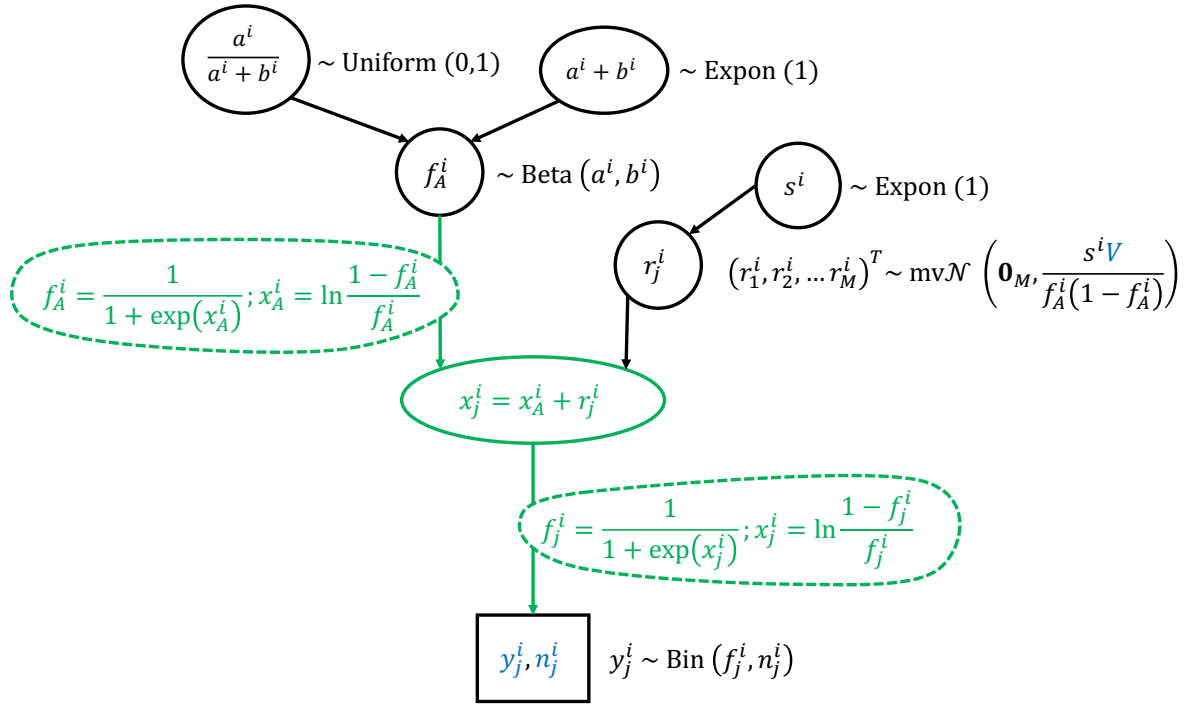


Рисунок 2.1 — Схема иерархической Байесовской модели **pordisp** для случая биаллельных генетических маркеров. Наблюдаемые переменные выделены синим цветом. Те из них, которые подчиняются определенному распределению, представлены в прямоугольном блоке, а параметры модели отображены в овальных блоках. Стрелка от одного блока к другому показывает, что переменная в первом блоке является параметром распределения переменной во втором блоке. Зеленым выделены преобразования, связанные с обработкой частот аллелей в популяциях, как композиционных данных. Подробное описание значения переменных описано в Параграфе 2.2.1.

2.2.2 Байесовский вывод апостериорных распределений

В Байесовском подходе, используемом для получения оценок параметров, предполагается, что все неизвестные параметры модели являются случайными величинами с некоторыми априорными распределениями, и наилучшие оценки параметров соответствуют наибольшему значению плотности апостериорного распределения (Обзор в параграфе 1.5). Так как все генетические локусы считаются независимыми, то верхний индекс i будет опущен при выводе. Введем дополнительные обозначения для удобства: $x = (x_1, x_2, \dots, x_M)^T$, $\underline{x}_A$ - вектор длины M со всеми элементами, равными x_A , $r = x - \underline{x}_A$ - вектор отклонений x_j

от x_A , $\theta = \{x_A, r_1, \dots, r_M, \alpha, \beta, s\}$ - совокупный вектор всех параметров, $D = \{y_1, \dots, y_M, n_1, \dots, n_M, V\}$ - совокупность всех данных. Согласно построенной модели (Рисунок 2.1), выпишем плотность апостериорного распределения всех неизвестных параметров:

$$\begin{aligned}
 p(\theta|D) &\propto p(D|\theta)p(\theta) = \prod_{j=1}^M f_j^{y_j} (1 - f_j)^{(n_j - y_j)} \times \\
 &\quad \det \left(\frac{sV}{f_A(1 - f_A)} \right)^{-1/2} \exp \left(-\frac{r^\top V^{-1} r}{2 \frac{s}{f_A(1 - f_A)}} \right) \times \\
 &\quad f_A^{\alpha-1} (1 - f_A)^{\beta-1} \exp(-(\alpha + \beta)) \exp(-s) \propto \\
 &\propto \prod_{j=1}^M \left(\frac{1}{1 + \exp(x_j)} \right)^{y_j} \left(\frac{\exp(x_j)}{1 + \exp(x_j)} \right)^{n_j - y_j} \\
 &\quad \left(\frac{f_A(1 - f_A)}{2s} \right)^{M/2} \\
 &\quad \exp \left(-\frac{f_A(1 - f_A)}{2s} r^\top V^{-1} r \right) \\
 &\quad f_A^{\alpha-1} (1 - f_A)^{\beta-1} \exp(-(\alpha + \beta)) \exp(-s) = \\
 &= \prod_{j=1}^M \frac{\exp(x_A + r_j)^{n_j - y_j}}{(1 + \exp(x_A + r_j))^{n_j}} \\
 &\quad \frac{\exp(x_A)^{\beta-1+M/2}}{(1 + \exp(x_A))^{\alpha+\beta-2+M}} s^{-M/2} \\
 &\quad \exp \left(-\frac{1}{2s} \frac{\exp(x_A)}{(1 + \exp(x_A))^2} r^\top V^{-1} r \right) \\
 &\quad \exp(-(\alpha + \beta)) \exp(-s).
 \end{aligned}$$

Максимум вероятности соответствует максимуму логарифма вероятности, поэтому возьмем логарифм от получившегося выражения:

$$\begin{aligned}
\ln(p(D|\theta)p(\theta)) &= \sum_{j=1}^M (x_A + r_j)(n_j - k_j) - \sum_{j=1}^M n_j \ln(1 + \exp(x_A + r_j)) + \\
&\quad + x_A(\beta - 1 + M/2) - (\alpha + \beta - 2 + M) \ln(1 + \exp(x_A)) - \\
&\quad - \frac{M}{2} \ln(s) - \frac{1}{2s} \frac{\exp(x_A)}{(1 + \exp(x_A))^2} r^\top V^{-1} r - \\
&\quad - s - \alpha - \beta = \\
&= \sum_{j=1}^M (n_j - k_j) r_j - \sum_{j=1}^M n_j \ln(1 + \exp(x_A + r_j)) + \\
&\quad + \left(\beta - 1 + M/2 + \sum_{j=1}^M (n_j - k_j) \right) x_A - \\
&\quad - (\alpha + \beta - 2 + M) \ln(1 + \exp(x_A)) - \\
&\quad - \frac{1}{2s} \frac{r^\top V^{-1} r}{\exp(x_A) + 2 + \exp(-x_A)} - \\
&\quad - \frac{M}{2} \ln(s) - s - \alpha - \beta.
\end{aligned} \tag{2.4}$$

Для получения логарифмов апостериорных распределений для каждого из параметров необходимо выбрать только те слагаемые из уравнения 2.4, которые содержат интересующий параметр.

$$\begin{aligned}
\ln(p(x_A|D, \theta \setminus x_A)) &\propto - \sum_{j=1}^M n_j \ln(1 + \exp(x_A + r_j)) + \\
&\quad + (\beta - 1 + M/2 + \sum_{j=1}^M (n_j - y_j)) x_A - \\
&\quad - (\alpha + \beta - 2 + M) \ln(1 + \exp(x_A)) - \\
&\quad - \frac{1}{2s} \frac{r^\top V^{-1} r}{\exp(x_A) + 2 + \exp(-x_A)}.
\end{aligned}$$

$$\begin{aligned}
\ln(p(r_j|D, \theta \setminus r_j)) &\propto \sum_{j=1}^M (n_j - y_j) r_j - \\
&\quad - \sum_{j=1}^M n_j \ln(1 + \exp(x_A + r_j)) - \\
&\quad - \frac{1}{2s} \frac{r^\top V^{-1} r}{\exp(x_A) + 2 + \exp(-x_A)}. \\
\ln(p(s|D, \theta \setminus s)) &\propto - \frac{1}{2s} \frac{r^\top V^{-1} r}{\exp(x_A) + 2 + \exp(-x_A)} - \\
&\quad - \frac{M}{2} \ln(s) - s.
\end{aligned}$$

$$\ln(p(\alpha|D, \theta \setminus \alpha)) \propto - \alpha \ln(1 + \exp(x_A)) - \alpha.$$

$$\ln(p(\beta|D, \theta \setminus \beta)) \propto \beta x_A - \beta \ln(1 + \exp(x_A)) - \beta.$$

2.2.3 Производные апостериорных распределений по параметрам

Модель `popdisp` имеет следующий набор параметров: x_A , $\{r_j\}_{j=\overline{1, M}}$, s , a , b . Апостериорные распределения для каждого из этих параметров не имеют стандартного аналитического вида, из которого можно получить выборку напрямую. Следовательно, схема сэмплирования по Гиббсу не подходит для оптимизации параметров в данном случае. Эффективной альтернативой является Гамильтониан Монте-Карло, который требует вычисления производных $\ln(p(D|\theta)p(\theta))$ по каждому из этих параметров (1.5). Произведем вывод требуемых производных:

$$\begin{aligned}
\frac{\partial \ln(p(x_A|D, \theta \setminus x_A))}{\partial x_A} &= - \sum_{j=1}^M \left(n_j \frac{\exp(x_A + r_j)}{1 + \exp(x_A + r_j)} \right) + \\
&+ \sum_{j=1}^M (n_j - y_j) + \beta - 1 + \frac{M}{2} - \\
&- (\alpha + \beta - 2 + M) \frac{\exp(x_A)}{1 + \exp(x_A)} - \\
&- \frac{r^\top V^{-1} r}{2s} \frac{(-1)(\exp(x_A) - \exp(-x_A))}{(\exp(x_A) + 2 + \exp(-x_A))^2} = \\
&= \sum_{j=1}^M \left(n_j \frac{1}{1 + \exp(x_A + r_j)} - y_j \right) + \\
&+ \beta - 1 + \frac{M}{2} - \\
&- (\alpha + \beta - 2 + M) \frac{\exp(x_A)}{1 + \exp(x_A)} + \\
&+ \frac{r^\top V^{-1} r}{2s} \frac{\exp(x_A) - \exp(-x_A)}{(\exp(x_A) + 2 + \exp(-x_A))^2}.
\end{aligned}$$

$$\begin{aligned}
\frac{\ln(p(r_j|D, \theta \setminus r_j))}{\partial r_j} &= \sum_{j=1}^M \left(n_j \frac{1}{1 + \exp(x_A + r_j)} - y_j \right) - \\
&- \frac{1}{2s} \frac{2 [V^{-1} r]_j}{\exp(x_A) + 2 + \exp(-x_A)}.
\end{aligned}$$

$$\frac{\ln(p(s|D, \theta \setminus s))}{\partial s} = \frac{1}{2s^2} \frac{r^\top V^{-1} r}{\exp(x_A) + 2 + \exp(-x_A)} - \frac{M}{2s} - 1.$$

$$\frac{\ln(p(\alpha|D, \theta \setminus \alpha))}{\partial s} = -\ln(1 + \exp(x_A)) - 1.$$

$$\frac{\ln(p(\beta|D, \theta \setminus \beta))}{\partial s} = x_A - \ln(1 + \exp(x_A)) - 1.$$

Оптимизация методом Гамильтониан Монте-Карло в пакете `popdisp` осуществляется с использованием инструментов Python пакета `pyhmc` (<https://pythonhosted.org/pyhmc/>). Полученные автором диссертации в аналитическом виде логарифмы из параграфа 2.2.2 производные реализованы в пакете `popdisp` и подаются на вход методу ”hm” для оптимизации параметров.

2.2.4 Модификация popdisp модели на случай нескольких аллелей в локусе

Использование в модели popdisp перехода от частот к балансам (см. Параграф 2.1) делает ее легко расширяемой для случая нескольких аллелей в одном генетическом локусе.

Пусть M - количество популяций некоторого вида. Для каждой из популяций была произведена выборка из индивидуумов, которые были генотипированы по единой панели из N несвязных друг с другом многоаллельных локусов. Так как все генетические локусы в модели считаются независимыми, то рассмотрим модель для случая одного локуса, чтобы не загромождать вывод индексами. Пусть в рассматриваемом локусе может существовать h различных состояний (аллелей). Тогда для этого локуса для j -ой популяции даны следующие значения: n_j - общее количество аллелей и вектор $y_j = (y_{j1}, \dots, y_{jh})$, показывающий, сколько аллелей каждого из h вариантов было генотипировано в j -ой популяции. Предположим, что каждая популяция характеризуется своим вектором частот аллелей рассматриваемого генетического локуса, $f_j = (f_{j1}, \dots, f_{jh})$, где $\sum_{k=1}^h f_{jk} = 1$. Тогда наблюдаемый вектор y_j является выборкой из мультиномиального распределения с параметрами вероятностей событий f_j и количеством испытаний n_j :

$$(y_{j1}, \dots, y_{jh}) \sim Multinomial((f_{j1}, \dots, f_{jh}), n_j).$$

Так как вектор f_j представляет собой композицию, то ее необходимо перевести с помощью илг-преобразования в $\eta = (h - 1)$ балансов (см. обзор в Параграфе 1.4). Обозначим эту функцию как $g(\cdot)$:

$$x_j = (x_{j1}, \dots, x_{j\eta}) = g(f_j).$$

Предположим, что каждый баланс раскладывается на сумму из среднего значения, соответствующего балансу для всей популяции, и добавки, соответствующей отклонению для подпопуляции:

$$x_{jk} = x_{Ak} + r_{jk}$$

Будем предполагать, что добавочные члены r_{jk} для M популяций связаны через матрицу V :

$$(r_{1k}, r_{2k}, \dots, r_{Mk}) \sim mv\mathcal{N}(0, s_k V),$$

где $s_k \sim \text{Exp}(1)$. Балансы, соответствующие всей популяции $(x_{A1}, \dots, x_{A\eta})$, можно обратным преобразованием перевести в частоты аллелей локуса, характерные для популяции:

$$(f_{A1}, \dots, f_{Ah}) = g^{-1}(x_{A1}, \dots, x_{A\eta}).$$

В свою очередь, частоты $(f_{A1}, \dots, f_{Ah})$ имеют априорное распределение Дирихле:

$$(f_{A1}, \dots, f_{Ah}) \sim \text{Dirichlet}(\lambda_1, \dots, \lambda_h),$$

где априорное распределение для параметров концентраций: $\sum_{k=1}^h \lambda_k \sim \text{Exp}(1)$.

В результате получаем Байесовскую иерархическую модель `popdisp` для случая нескольких аллелей (см. Рисунок 2.2).

2.2.5 Модификация `popdisp` модели с учетом неравновесия по сцеплению

Одним из основных предположений в модели `popdisp` является предположение о независимости локусов. Такое предположение не является уникальным для `popdisp` и используется во всех основных моделях в области. Однако, существует такое явление как неравновесие по сцеплению (linkage disequilibrium, LD), когда наблюдаемое распределение аллелей на двух или более локусах генома отклоняется от ожидаемого распределения, основанного на независимости локусов. Неравновесие по сцеплению может быть вызвано различными факторами, включая частоту рекомбинации. Чем выше частота рекомбинации, тем более вероятно, что аллели на разных локусах будут независимо перетасовываться и распределяться в популяции, в результате будет наблюдаться более низкий уровень неравновесия по сцеплению.

В рамках модели на Рисунке 2.2, неравновесие по сцеплению будет означать, что $\{f_A^i\}$ для разных i не независимы. Пусть для каждого локуса i известен набор локусов, с которыми рассматриваемый локус показывает значимое неравновесие по сцеплению, и пусть этот набор определен и зафиксирован заранее. Другими словами, для любого локуса i определено множество индексов локусов Ω^i , таких что любой локус k из Ω^i сцеплен с i -м локусом и значение неравновесия по сцеплению между ними, D^{ik} , заранее определено. В этом случае среднее

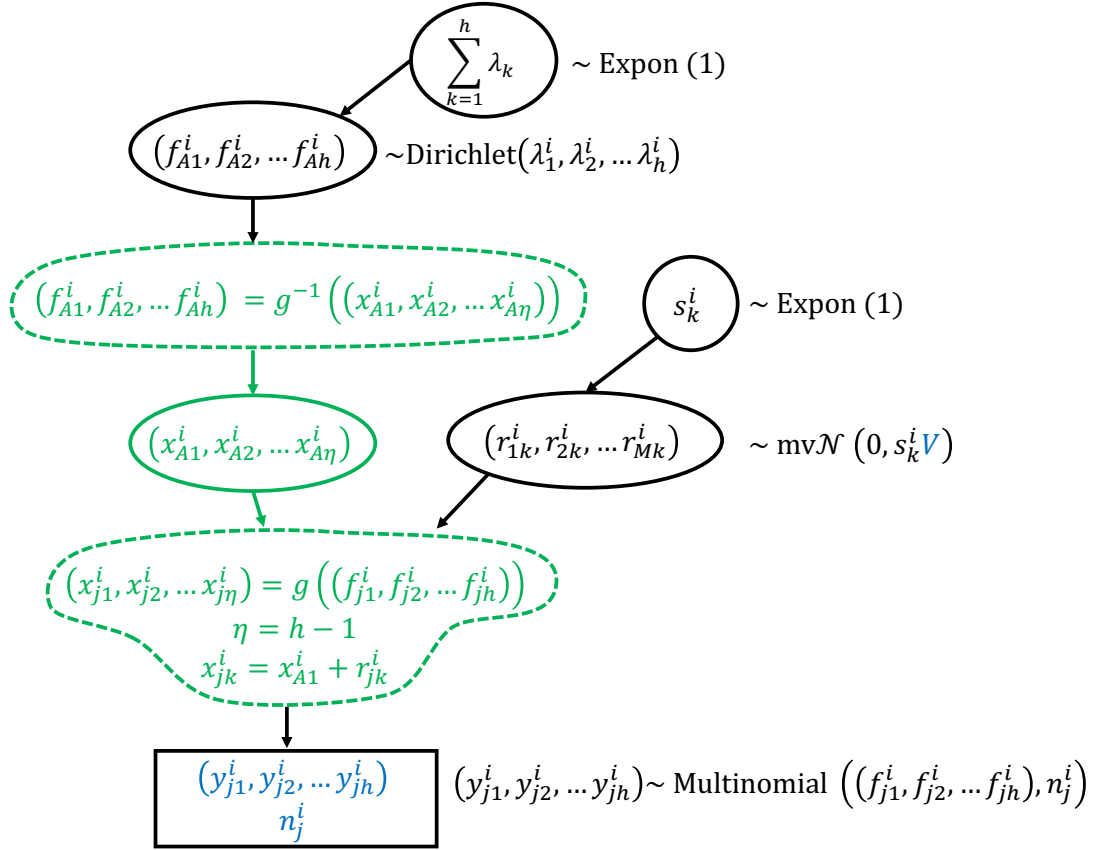


Рисунок 2.2 — Схема иерархической Байесовской модели **popdisp** для случая генетических локусов с несколькими аллелями. Наблюдаемые переменные выделены синим цветом. Те из них, которые подчиняются определенному распределению, представлены в прямоугольном блоке, а параметры модели отображены в овальных блоках. Стрелка от одного блока к другому показывает, что переменная в первом блоке является параметром распределения переменной во втором блоке. Зеленым выделены преобразования, связанные с обработкой частот аллелей в популяциях, как композиционных данных. Подробное описание значения переменных описано в Параграфе 2.2.4.

для априорного Бета-распределение для f_A^i можно записать как взвешенное значение $\{f_A^k\}_{k \in \Omega^i}$:

$$\frac{\alpha^i}{\alpha^i + \beta^i} = \frac{\sum_{k \in \Omega^i} D^{ik} (f_A^k + (-1)^{[D < 0]})}{\sum_{k \in \Omega^i} |D^{ik}|},$$

а параметр концентрации пропорциональным среднему значению сцепления по множеству локусов Ω^i :

$$(\alpha^i + \beta^i) = \gamma \frac{\sum_{k \in \Omega^i} D^{ik}}{|\Omega^i|},$$

где коэффициент пропорциональности γ единый для всех локусов, и имеет следующее априорное распределение:

$$\gamma \sim Expon(1).$$

Возникает вопрос, как проводить генерацию значений параметров в такой сложно связной модели (Рисунок 2.3). В модели `popdisp` для несвязных локусов оптимизация параметров проводится независимо для каждого локуса, что позволяет распараллеливать процесс по данным. В модели с взаимосвязанными локусами необходимо синхронизировать процесс оптимизации для всех локусов. Процесс можно представить следующим образом:

1. Зафиксировать значения параметров $\{f_A^i\}$ с предыдущего шага MCMC.
2. Сгенерировать одно новое значение для всех параметров каждого локуса.
3. Обновить общий для всех локусов параметр γ .

Оптимизация этой модели занимает больше времени по сравнению с моделью с независимыми локусами, так как синхронизация требует завершения шага (2) для всех локусов, что не требуется при оптимизации независимой оптимизации.

>

2.2.6 Расчет ковариации при распространении популяций по географии

Модель `popisd` может использоваться для оценки характеристических частот аллелей в популяции в контексте географического распределения образцов, когда матрица V отражает географические расстояния как между анализируемыми образцами, так и между образцами и генетическим центроидом. В этом

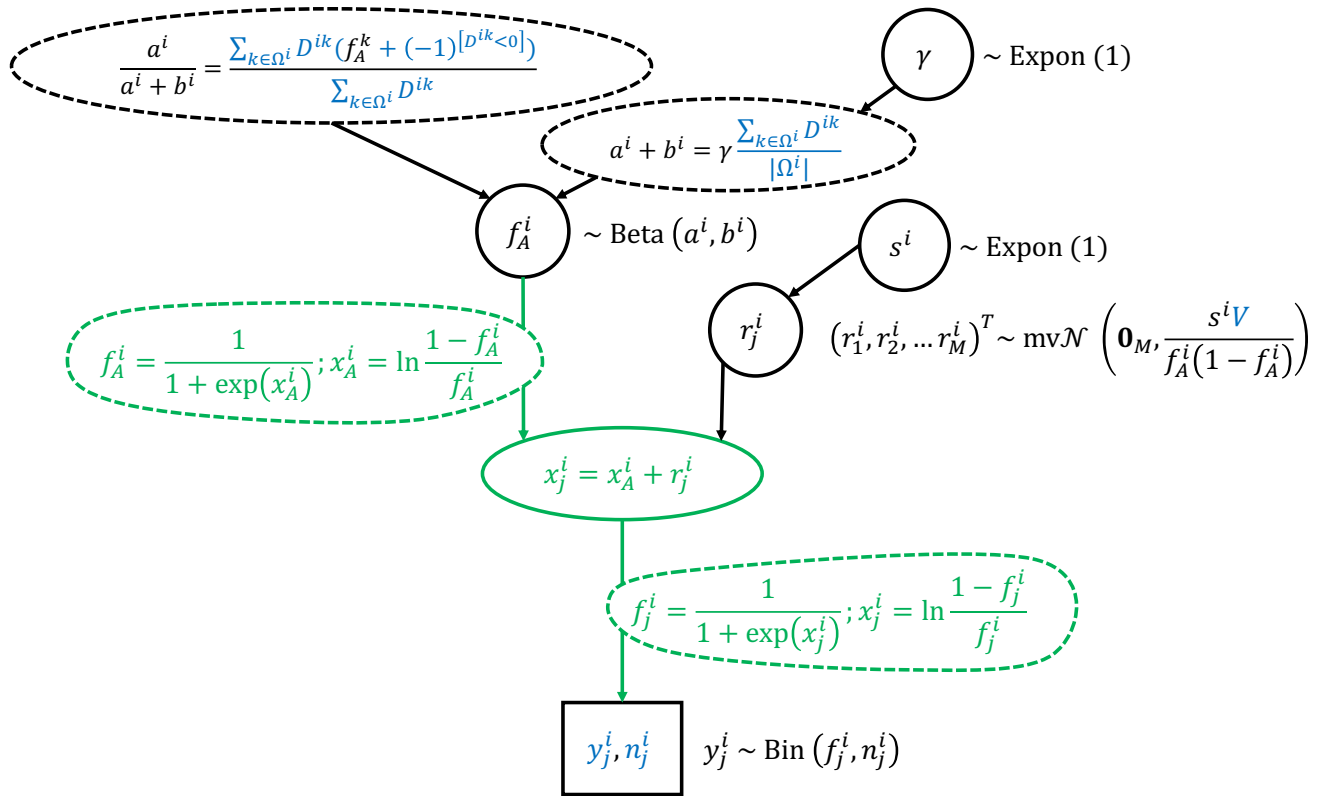


Рисунок 2.3 — Схема иерархической Байесовской модели **popdisp** для случая биаллельных генетических маркеров неравновесия по сцеплению. Наблюдаемые переменные выделены синим цветом. Те из них, которые подчиняются определенному распределению, представлены в прямоугольном блоке, а параметры модели отображены в овальных блоках. Стрелка от одного блока к другому показывает, что переменная в первом блоке является параметром распределения переменной во втором блоке. Зеленым выделены преобразования, связанные с обработкой частот аллелей в популяциях, как композиционных данных.

контексте можно построить несколько гипотез о перемещениях образцов по географическому региону, для каждой из которых можно построить свою матрицу V и сравнить эти гипотезы с использованием критериев выбора моделей (например, BIC, Bayesian Information Criterion). При построении таких гипотез возникает два вопроса: (i) какую функцию использовать для оценки расстояния между двумя географическими точками, и (ii) как преобразовать расстояния в значения ковариационной матрицы.

Простейшей мерой расстояния между точками сбора образцов является линейное расстояние, однако оно не всегда применимо на практике, поскольку прямой путь между точками сбора может быть прегражден естественными пре-

пятствиями, такими как горы или водоемы (см. Рисунок 1.1). Более реалистичной мерой расстояния являются так называемые ”расстояния с наименьшей стоимостью” (Least Cost Distance, LCD, 1.1.4). LCD расстояние определяется как длина пути между двумя точками, который минимизирует общие затраты на перемещение с учетом географических условий.

После того, как матрица расстояния между всеми точками сбора образцов и географическим положением генетического центроида получена, оценку ковариационной матрицы можно выполнить согласно следующей логике. Пусть бинарное дерево представляет собой пути распространения от положения центроида к точкам сбора двух образцов. Введем на этом дереве длины трех ветвей ρ , x_1 и x_2 , как показано на Рисунке 2.4А. Тогда ковариация между двумя точками сбора пропорциональна ρ , то есть их общему пути на дереве. Значение ρ может быть выражено как линейная комбинация трех расстояний: от каждой точки сбора до положения центроида, d_1 и d_2 , и между двумя точками сбора, x (см. формулу на Рисунке 2.4 А). Построение оптимального LCD пути в виде бинарного ветвления от центра к двум точкам сбора образцов требует значительных вычислительных затрат, так как подразумевает рассмотрение всех возможных развилок на поверхности (Рисунок 2.4Б). Вместо этого предлагается независимо найти LCD пути между положением центроида и точками сбора образцов и использовать длины этих путей для оценки ковариации ρ (Рисунок 2.4В). Будем использовать формулу $\rho = (d_1 + d_2 - x)/2$ и для более общего случая, когда конкретный путь неизвестен (Рисунок 2.4В, справа). Такой подход может применяться для получения ковариационных матриц как для случая прямолинейных путей, так и для оптимальных LCD путей.

2.3 Реконструкция событий смешений с помощью модели migadmi

Модель migadmi была создана для реконструкции потенциальных путей миграции и событий смешений популяций одного вида, каждая из которых характеризуется вектором частот аллелей. Кроме векторов частот аллелей, модель принимает на вход информацию о том, какие популяции являются исходными (чистыми, не смешанными), а какие - результатом смешений. Migadmi, основываясь на предоставленных данных, начинает с построения бинарного дерева,

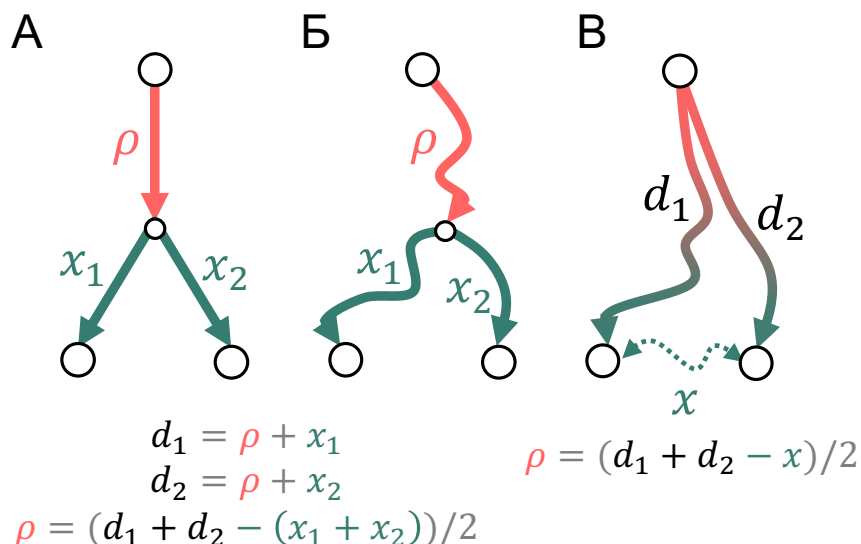


Рисунок 2.4 — **Метод расчета ковариации между двумя географическими точками относительно положения генетического центроида, откуда произошло распространение.** (А) Расчет ковариации с предположением о бинарном ветвящемся распространении от центра к точкам. (Б) Расчет ковариации с предположением о бинарном ветвящемся нелинейном распространении от центра к точкам. (В) Аппроксимация ковариации, когда точка бинарного ветвления неизвестна.

которое отображает эволюционные связи между исходными популяциями. Затем модель наращивает на это дерево события смещений, согласно принципу максимума правдоподобия, используя метод полного перебора. Для осуществления полного перебора требуется задать глубину перебора, то есть максимальное число популяций, которые могут участвовать как компоненты в одной смеси. Модель *migadmi* позволяет оценивать смещения с более чем двумя компонентами и рассматривать множественные смещения и вложенные события (когда смешанная популяция становится исходной для другой смешанной популяции), что является ее преимуществом по сравнению с другими моделями.

В *migadmi* также существует возможность использования априорной информации: (i) невзвешенное бинарное дерево, отображающее эволюционные связи между исходными популяциями, (ii) последовательность и группировка событий смещений, которые требуется оценить. Кроме того, *migadmi* оценивает параметры для одного события смещения в режиме скользящего окна, а общая статистика по композиционным вкладам популяций-источников в смешанную усредняется по всем окнам. Такой подход к скринингу смещений вдоль генома позволяет визуально оценить блоки в геноме со сходным паттерном смещений.

Например, какие блоки произошли от одной исходной популяции, а какие содержат значительное количество независимой вариации. Последнее может указывать на то, что не все компоненты смеси доступны для исследования.

Модель *migadmi* имеет следующие отличия от других моделей реконструкции графа смещений для популяций:

- Модель *migadmi* основана на тех же предположениях, что и модель *popdisp*, и использует композиционную аппроксимацию модели дрейфа Райта-Фишера.
- В модели смещения предполагается, что событие смещения произошло какое-то время назад, так что и исходные популяции, и смешанная популяция уже накопили собственную вариабельность после смещения.
- Для параметризации событий смещения *migadmi* использует оригинальный алгоритм, основанный на символьных вычислениях, так что метод оптимизации параметров получает функцию для оптимизации в аналитическом виде.

2.3.1 Формальное описание модели

Рассмотрим P популяций, каждая из которых характеризуется вектором частот аллелей в независимых биаллельных локусах. С помощью *ilr*-преобразования переведем частоты в балансы (см. Параграф 1.4) и обозначим балансы частот аллелей в i -м локусе j -й популяции за x_j^i . Так как в отличие от модели *popdisp*, *migadmi* не ставит целью оценку частот аллелей в общей предковой популяции, то будем строить модель, избегая необходимости использования x_A^i (Рисунок 2.1).

На основании бинарного дерева, связывающего исходные (несмешанные) популяции, сконструируем параметризованную матрицу ковариаций V аналогично тому, как это сделано в модели *Treemix* [14] или в методе Фельзенштейна [74]. Тогда, как было показано в *popdisp*:

$$(x_1^i, x_2^i, \dots, x_P^i)^\top \sim mv\mathcal{N}((x_A^i, x_A^i, \dots, x_A^i)^\top, c^i V),$$

где c - константа пропорциональности. Чтобы избежать использования x_A^i , рассмотрим распределения для разностей между балансами частот для каждой пары

популяций. Для популяций j и k справедливо следующее:

$$x_j^i \sim \mathcal{N}(x_A^i, c^i V_{jj}),$$

$$x_k^i \sim \mathcal{N}(x_A^i, c^i V_{kk}),$$

$$\text{cov}(x_j^i, x_k^i) = c^i V_{jk},$$

тогда:

$$(x_j^i - x_k^i) \sim \mathcal{N}(0, c^i (V_{jj} + V_{kk} - 2V_{jk})).$$

Введем матрицу D :

$$D_{jk} = V_{jj} + V_{kk} - 2V_{jk} \quad (2.5)$$

и запишем функцию правдоподобия для всех разностей между балансами:

$$\mathcal{L}_i = \prod_{j=1}^{P-1} \prod_{k=j+1}^P p_{\mathcal{N}}(x_j^i - x_k^i | 0, \frac{s_i D_{jk}}{f_A^i (1 - f_A^i)}), \quad (2.6)$$

где f_A^i - это частота альтернативного аллеля в предповом состоянии i -ого ОНП. Значение этой величины не является параметром и играет роль масштаба. В соответствии с композиционным анализом данных 1.4, $f_A^i = 1/(1 + \exp(x_A^i))$, где x_A^i положим равным среднему, $x_A^i = 1/P \sum_{j=1}^P x_j^i$. Введем дополнительное обозначение, которое будем использовать в дальнейшем: $q_i^2 = f_A^i (1 - f_A^i)$.

Функция правдоподобия 2.6, кроме параметризованной матрицы D , содержит параметр s_i уникальный для каждого ОНП, что делает модель перепараметризованной. Чтобы уменьшить количество параметров, был применен метод скользящего окна. Будем считать, что в каждом окне ОНП связаны неравновесием по сцеплению, поэтому значения s_i для ОНП в окне являются одинаковыми. На практике рекомендуется брать окна размером, равным оценке величины неравновесия по сцеплению, с шагом равным трети размера окна. Таким образом предполагая, что изменения частот аллелей ОНП в окне происходило приблизительно с одним темпом. Пусть G_w - это набор ОНП, соответствующих w -му окну, а s^w - это темп, специфичный для этого окна. Тогда функция правдоподобия для логарифмических балансов частот аллелей в w -м окне выглядит следующим образом:

$$\begin{aligned}
\mathcal{L}^w &= \prod_{i \in G_w} \prod_{j=1}^{P-1} \prod_{k=j+1}^P p_{\mathcal{N}} \left(x_j^i - x_k^i | 0, \frac{s^w D_{jk}}{q_i^2} \right) \propto \\
&\prod_{i \in G_w} \prod_{j=1}^{P-1} \prod_{k=j+1}^P \frac{1}{\sqrt{2\pi s^w D_{jk}}} \exp \left(-\frac{(x_j^i - x_k^i)^2}{s^w D_{jk}/(q_i^2)} \right) = \\
&\prod_{i \in G_w} \prod_{j=1}^{P-1} \prod_{k=j+1}^P \frac{1}{\sqrt{2\pi s^w D_{jk}}} \exp \left(-\frac{((x_j^i - x_k^i) q_i)^2}{s^w D_{jk}} \right) = \\
&\prod_{j=1}^{P-1} \prod_{k=j+1}^P \frac{1}{(2\pi s^w D_{jk})^{|G_w|/2}} \exp \left(-\sum_{i \in G_w} \frac{((x_j^i - x_k^i) q_i)^2}{s^w D_{jk}} \right) = \\
&\prod_{j=1}^{P-1} \prod_{k=j+1}^P \left[\frac{1}{\sqrt{2\pi s^w D_{jk}}} \exp \left(-\frac{1}{|G_w|} \sum_{i \in G_w} \frac{((x_j^i - x_k^i) q_i)^2}{s^w D_{jk}} \right) \right]^{|G_w|} = \\
&\left[\prod_{j=1}^{P-1} \prod_{k=j+1}^P p_{\mathcal{N}} \left(d_w(j, k) \mid 0, s^w D_{jk} \right) \right]^{|G_w|},
\end{aligned} \tag{2.7}$$

где

$$d_w(j, k) = \sqrt{\frac{1}{|G_w|} \sum_{i \in G_w} ((x_j^i - x_k^i) q_i)^2}$$

играет роль некоторого взвешенного расстояния между популяциями i и j в окне w . Для оценки параметров можно избавиться от показателя степени и максимизировать следующую функцию логарифма правдоподобия:

$$\log \mathcal{L}^w \propto \sum_{j=1}^{P-1} \sum_{k=j+1}^P \log (p_{\mathcal{N}} (d_w(x_{j,k}) \mid 0, D_{jk})).$$

2.3.2 Параметры в модели `migadmi` и их идентифицируемость

Модель `migadmi` предлагает новую модель смешения, согласно которой событие смешения произошло в прошлом, то есть все популяции, участвовавшие в смешении, имеют собственную независимую эволюционную историю после этого события. Как и в других методах реконструкции графов смешений, новая параметризация события смешения является неполной, чтобы обеспечить выполнение условий идентифицируемости параметров.

Полный набор параметров в графе смешений состоит из длин рёбер и весов событий смешений. Рассмотрим параметризацию графа смешений более подробно. Пусть для набора исходных несмешанных популяций дано бинарное дерево их совместной эволюционной истории. В этом случае, каждому ребру в графе ставится в соответствие параметр длины (2.5a). Пусть одна из популяций, которая пока еще не включена в граф смешений, предполагается смесью нескольких популяций, которые уже присутствуют в этом графе. В *miGAdmI* предполагается, что в графе смешений каждая из популяций имеет свою индивидуальную ветвь, при этом в событии смешения участвуют не популяции, располагающиеся на листьях графа, а некоторые предшествующие им, находящиеся на в каком-то месте на индивидуальных ветвях (2.5b). Положение такой предшествующей популяции на индивидуальном ребре определяется с помощью одного параметра в диапазоне $(0, 1)$ (α_i на Рисунке 2.5b). Все популяции, участвующие в смешении, формируют новую популяцию с помощью весовых коэффициентов, сумма которых равна единице. После события смешения новая популяция получает свою индивидуальную ветвь, длина которой также является параметром. Введение дополнительной индивидуальной ветви позволяет смешанной популяции тоже участвовать в некоторой смеси.

В таком представлении, событие смеси из m популяций будет добавлять в модель $2m$ новых параметров (m положений предшествующих популяций на индивидуальных ребрах, $(m - 1)$ независимых весов и 1 новую ветвь). Чтобы все они были идентифицируемы, необходимо, чтобы $2m$ было меньше либо равно количеству популяций, которые уже соединены графом смешений. Однако, при малом количестве рассматриваемых популяций, такое представление событий смеси будет сильно ограничивать количество популяций, которые могут участвовать в смеси. Поэтому, во избежание перепараметризации, были введены несколько ограничений.

Во-первых, все α_i предполагаются равными друг другу и обозначаются как α , что сокращает количество свободных параметров до $(m + 1)$ (2.5c). Во-вторых, дополнительно по желанию можно добавить при оптимизации *miGAdmI* регуляризацию на параметры весов с помощью априорного распределения. В качестве априорного распределения предлагается использовать распределение Дирихле с одинаковыми значениями концентраций для всех компонент (λ). Значение $\lambda = 1$ соответствует отсутствию регуляризации, а значения меньше 1 будут приближать композицию весов к границе пространства возможных весов (т.е., некоторые из

параметров весов к нулю). Такое регуляризация моделирует эффект поглощающих состояний в генетическом дрейфе.

Все элементы матрицы ковариаций между всеми популяциями, обозначенной ранее как V , являются функциями от параметров графа смещений (пример на Рисунке 2.5d). А каждый элемент матрицы D в свою очередь, согласно 2.5, является функцией от элементов матрицы V . Для того, чтобы получить аналитический вид функции для элементов матрицы D , в `migadmi` используются символьные операции с применением библиотеки `sympy` [75].

2.3.3 Оценка вкладов популяций-источников в смешанную популяцию с точки зрения дисперсии

Как было обсуждено в Параграфе 1.2.2, полная модель смешения является неидентифицируемой. Поэтому при использовании усеченной модели параметры не полностью отражают вкладываемый смысл, а представляют собой комбинацию параметров полной неидентифицируемой модели. Для точной оценки вкладов популяций-источников в смешанную популяцию `migadmi` использует не веса смешения, а разбиение полной дисперсии смешанной популяции на составляющие, соответствующие предшествующим популяциям и собственной дисперсии.

Для оценки относительных вкладов предшествующих популяций в смешанную популяцию, с учетом ее собственной дисперсии, `migadmi` оценивает разложение дисперсии, аналогично представленному в Таблице 1 для примера на Рисунке 2.5, и нормализует полученные вклады до единицы.

Таблица 1 — Разложение дисперсии смешанной популяции y на индивидуальные вклады.

Источник	Вклад
x_1	$w_1^2(a_{t_1} + t_{12}) + w_1(w_1 + w_3)t_{14}$
x_3	$w_3^2(a_{t_3} + t_{34}) + w_3(w_1 + w_3)t_{14}$
x_6	$w_6^2(a_{t_6} + t_{56})$
Собственная	t_y

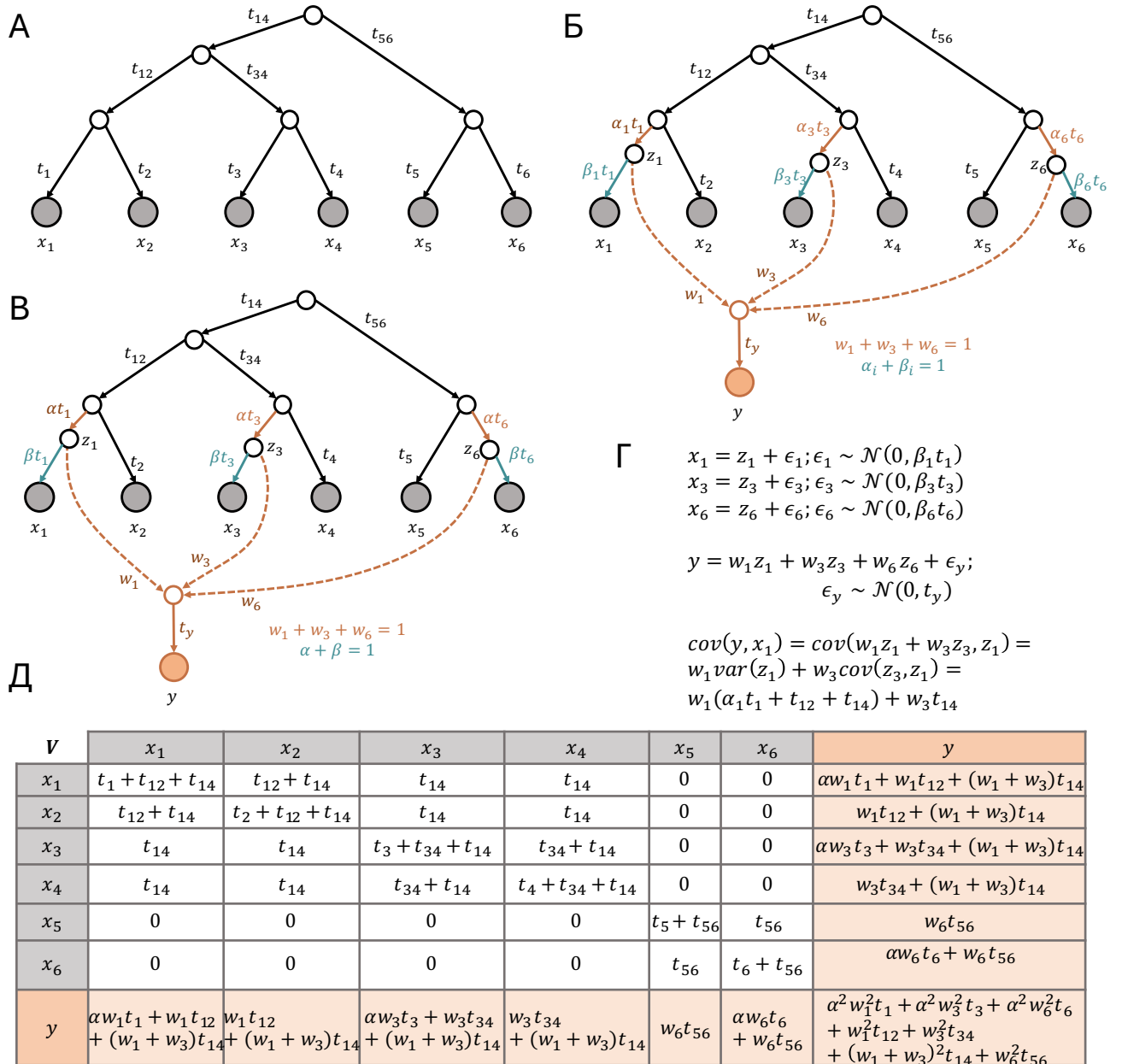


Рисунок 2.5 — Пример параметризации события смешения в модели **migadmi**. (А) Исходное эволюционное дерево, соединяющее шесть популяций x_1 до x_6 . Параметры, соответствующие объему накопленного дрейфа, указаны на соответствующих ветвях дерева. (Б) Полная параметризация события смешения. Популяция y является смесью, которая произошла в прошлом. В качестве популяций-источников выступали предковые состояния популяций x_1 , x_3 и x_6 . (В) Параметризация события смешения с меньшим количеством параметров, которая используется в модели **migadmi**. (Г) Пример вывода ковариации между смешанной популяцией и несмешанной популяцией. (Д) Матрица ковариаций между всеми парами популяций для примера (С).

2.3.4 Перебор топологий графов смешений

Программа `migadmi` может оценивать вклады как заранее заданных событий смешений в наборе популяций, так и наращивать их с помощью метода полного перебора. Допустим, в качестве исходных данных для модели `migadmi` задана топология эволюционного бинарного дерева, соединяющего все несмешанные популяции. Также предположим, что известен набор популяций, которые являются смешанными, и глубина перебора - то есть количество популяций, которые могут быть использованы в смеси для формирования смешанных популяций.

Пусть количество исходных несмешанных популяций - u , количество смешанных популяций v , глубина перебора h (где $h \geq 2$). Тогда, для добавления одной популяции к графу смешений, происходит перебор следующего количества моделей:

$$v \sum_{i=2}^h \binom{u}{i}.$$

После того, как к модели была добавлена одна популяция, следующий раунд перебора будет содержать следующее количество моделей:

$$(v-1) \sum_{i=2}^h \binom{u+1}{i}.$$

В результате, при добавлении всех смешанных популяций в граф смешений, количество перебранных моделей будет:

$$\sum_{j=0}^{v-1} (v-j) \sum_{i=2}^h \binom{u+j}{i}.$$

Предложенный подход соответствует жадному алгоритму, а выбор наилучшей модели на каждом этапе предлагается делать с помощью критерия BIC.

2.3.5 Сравнение migadmi с другими методами

Чтобы поставить новый метод migadmi в ряд уже существующих, приведем сравнительную таблицу по ключевым параметрам и возможностям, которые важны в задаче реконструкции эволюционной истории популяций с учетом событий смещений (Таблица 2).

Таблица 2 — Сравнение методов для моделирования событий смещения.

	migadmi	TreeMix	MixMapper	admixtools: qpGraph(), find_graph()
Количество событий смещений ≥ 2	+	—	—	—
Несколько вложенных смещений	+	+	—	+
Количество вложенных смещений	неограничено	0	2	неограничено
Получение паттерна смещений вдоль хромосом	+	-	-	-
Оценка собственной эволюционной истории у смешанных популяций	+	-	-	+
Композиционный анализ данных	+	-	-	-

2.4 Выводы к главе 2

В Главе 2 представлено описание двух разработанных моделей: *popdisp* и *migadmi*. Модель *popdisp* используется для оценки характеристических частот аллелей в популяции, учитывая ковариационную матрицу между индивидуумами, составляющими эту популяцию. В дополнение, предложен новый подход к построению ковариационной матрицы на основе географического положения точек сбора образцов. Модель *migadmi*, в свою очередь, принимает на вход частоты аллелей в популяции, которые, например, могут быть оценены с помощью *popdisp*, и позволяет реконструировать граф смещений, который отражает миграции и события смещений между этими популяциями.

Для работы с частотами аллелей в популяциях обе модели используют композиционный анализ данных [64]. Частоты аллелей в популяциях подчиняются определённым ограничениям: они всегда должны быть неотрицательными, а их сумма для одного генетического локуса в популяции должна равняться единице. Эти ограничения создают проблемы при проведении стандартного статистического анализа. Например, некоторые методы статистического анализа могут предполагать нормальное распределение данных, что не соответствует действительности для частот аллелей. Композиционный анализ данных позволяет избежать этих ограничений путем перевода частот аллелей в пространство балансов между частотами. Полученные балансы могут принимать любые значения на вещественной оси и рассматриваться независимо друг от друга. В настоящей работе предложена новая аппроксимация изменения частот аллелей в популяции Райта-Фишера, основанная на балансах частот, и использована в *popdisp* и *migadmi* моделях.

Модели *popdisp* и *migadmi* изначально разрабатывались для работы с биаллельными генетическими локусами. Однако, мультиаллельные генетические локусы не являются редким явлением, поэтому имеет смысл распространить модели *popdisp* и *migadmi* на такой случай. Автором статьи были предложена модификации модели *popdisp* для такого случая. Однако, разработка модели *migadmi* для мультиаллельного случая еще предстоит в будущем.

Как упоминалось ранее, модель *popdisp* оценивает характеристические частоты аллелей в популяции в географической точке, которая является генетическим центроидом. Понимание значения местоположения генетического

центроида не является тривиальным и однозначным. Это местоположение может соответствовать положению предковой популяции, и тогда оно вероятно является центром доместикации. С другой стороны, местоположение генетического центроида может рассматриваться как точка наибольшего потока или обмена генами. Таким образом, можно предположить, что найденное местоположение может быть близким к какому-либо (возможно древнему) торговому центру. Точная интерпретация положения генетического центроида в каждом отдельном случае требует отдельного анализа.

В начальной постановке задачи модель `popdisp` рассматривает все генетические локусы как независимые. В реальности же в популяции может существовать неравновесие сцепления (LD, Linkage Disequilibrium) между генетическими локусами. Неравновесие сцепления означает, что определенные комбинации аллелей на разных локусах встречаются чаще, чем ожидалось бы при случайном сочетании. Учет неравновесия по сцеплению является важным аспектом анализа генетических данных, поскольку связанные генетические локусы могут содержать информацию о совместном наследовании и эволюционных процессах. В параграфе 2.2.5 приведено дополнение модели `popdisp` для учета неравновесия по сцеплению. Это позволяет более полно учесть сложную генетическую структуру популяции и получить более точные оценки характеристических частот аллелей. Однако, следует отметить, что в такой формулировке модель `popdisp` будет работать медленнее, так как требует параллельного выполнения типа "ветвление-слияние" вместо параллельного выполнения независимых потоков.

Один из режимов работы программы `migadmi` - жадное построение графа смешений. В этом случае на вход подается набор несмешанных популяций и набор популяций, которые предполагаются смешанными. `Migadmi` по очереди добавляет в модель события смешений по принципу наименьшего значения критерия BIC. В качестве будущего улучшения алгоритма оптимизации графа смешений можно предложить разработку метода для поиска в полном пространстве графов смешений. Однако, такая задача является аналогичной NP-полной задаче поиска филогенетического дерева в пространстве всех возможных бинарных деревьев.

При интерпретации картины смешений вдоль хромосом, предсказанной с помощью скользящего окна в `migadmi`, возникает вопрос: является ли это действительным результатом смешения нескольких популяций или неполной сортировки линий (incomplete lineage sorting, ILS) [76]. Действительно, проблема

различия между неполной сортировкой линий и смешением может быть сложной, и инструмент *migadmi* не тестирует это напрямую. Одним из способов различить эти два процесса является анализ распределения паттерна вкладов популяций-источников вдоль хромосом. В случае чистого смешения следует ожидать набор визуально отчетливых хромосомных блоков, внутри каждого из которых будет наблюдаться более-менее постоянная картина вкладов популяций-предшественников. В отличие от этого, неполная сортировка линий приведет к более разнообразному паттерну смешений вдоль генома, без каких-либо явных участков одинакового паттерна смешений. Отсутствие блоков при неполной сортировке линий происходит потому, что деревья генов не соответствуют дереву видов, и это может произойти для каждого локуса независимо. В случае события смешения, гены в соседних локусах, скорее всего, соответствуют одинаковым с точки зрения топологии эволюционным деревьям. Однако, для точного различия между ILS и смешением с использованием *migadmi*, вероятно, потребуется дополнительный анализ распределения длин хромосомных блоков.

Обе программы, *popdisp* и *migadmi*, представляют собой универсальные инструменты, которые могут быть использованы как совместно, так и отдельно. Они могут быть применены к широкому спектру задач, не ограничиваясь только географическим распространением популяций. Так, например, было показано, что разнообразие в симбиотических бактериальных популяциях следует за разнообразием в симбиотических генах растений-хозяев [77; 78]. В подобных случаях в качестве матрицы ковариаций, связывающей образцы, может быть использована ковариационная матрица, построенная на основе генетических расстояний между растериями-хозяевами. Это лишь один из примеров того, как разработанные программные инструменты могут быть адаптированы для задач, не связанных напрямую с географическим распределением.

Глава 3. Исследования по нуту и исходные данные

Нут или бараний горох (*Cicer arietinum* L.) занимает второе место среди бобовых культур по мировому производству зерна и культивируется в более чем 50 странах. Он играет важную роль в пищевом обеспечении, предоставляя ценный источник белка, пищевых волокон и микроэлементов. Как и у большинства культурных растений, генетическое разнообразие окультуренных форм нута меньше, чем у их диких предшественников из-за прохождения через два "бутылочных горлышка": доместикацию и научную селекцию [79]. Это затрудняет работу селекционеров, которые имеют дело лишь с небольшой частью генетического разнообразия вида и пытаются создать сорта, способные эффективно справляться с биотическими и абиотическими стрессами.

Используя дикорастущие формы нута и коллекции староместных сортов, современные селекционеры получают доступ к естественному генетическому разнообразию, которое может быть использовано для улучшения культурных сортов. Популяционный анализ традиционных староместных сортов с учетом их географического распределения играет важную роль в понимании генетической структуры популяции нута на различных уровнях (глобальном и региональном) и в изучении эволюционных путей нута, а также его адаптации к местным условиям. Все это используется при создании современных селекционных программ. Таким образом, анализ староместных сортов играет важную роль в селекции и открывает новые перспективы для разработки более устойчивых и высокоурожайных сортов нута.

3.1 Общие сведения о нуте

Cicer arietinum L. - это однолетнее травянистое растение, относящееся к семейству бобовые (Leguminosae или Fabaceae), подсемейству мотыльковые (Papilionoideae или Faboideae) и трибе *Cicereae* Alef, которая содержит только один род *Cicer*. Вид *Cicer arietinum* L. - единственный культивируемый вид рода *Cicer*, а все остальные виды рода относятся к различным генетическим пулам: в зависимости от их способности к скрещиванию с *C. arietinum* L. [80]. Скрещивание

с представителями первичного генпула приводит к образованию жизнеспособных гибридов, при скрещивании с представителями вторичного генпула первое поколение проявляет частичную стерильность, а скрещивание с представителями третичного крайне затруднительно и в большинстве случаев приводит к образованию бесплодных гибридов. В первичный генпул входит дикорастущий однолетний предковый вид *C. reticulatum*, что указывает на него как на непосредственного предшественника культурного нута.

C. reticulatum имеет узкое экогеографическое распространение на юго-востоке региона современной Турции [81], что указывает на ограниченность генетической основы культивируемого нута и, следовательно, ограниченность адаптивной изменчивости. Изначально низкое разнообразие заставляет современные сорта сталкиваться с проблемой разнообразия особенно остро.

В регионах распространения дикого вида (современных Турции и Сирии) также были найдены несколько самых ранних археологических находок с образцами культивируемого нута (около 9500 лет назад), поэтому эта область считается местом первичной доместики нута. Согласно археологическим находкам, окультуренный нут затем распространился по всему древнему миру, достигая западной границы Средней Азии и долины реки Инд около 6000 лет назад, Средиземноморского бассейна около 5500 лет назад и Эфиопии около 3500 лет назад [10; 82]. В то время как миграционные связи нута между Ближневосточным, Индийским и Среднеазиатским регионами подтверждаются археологическими находками, точная история распространения нута с возможными событиями смешений в Средиземноморском бассейне и в Эфиопском регионе остается загадкой.

У нута (*C. arietinum*) существует две основных группы сортов: дези и кабули. Эти группы различаются по фенотипу: окраске лепестков, размеру и форме семян, а также по вкусовым качествам. Кабули обычно имеет светлые лепестки, большие, округлые и светлые семена, в то время как дези имеет сиреневые лепестки, меньшие, узкие и более темные семена. Кабули более распространен в международной торговле, в то время как дези чаще используется в региональных кулинарных традициях. Нут дези считается первым доместичированным нутом, так как напоминает дикий предшественник *C. reticulatum*, а кабули вероятно был отобран из местных дези, а затем распространился, однако точное место доместикации кабули неизвестно.

Нут является преимущественно самоопыляемым видом, у которого опыление возможно уже за 1-2 дня до раскрытия цветков (буточная автогамия). Такая

форма опыления приводит к низкой вероятности скрещивания с другими растениями вида, так что коэффициент скрещивания (англ. outcrossing rate) у нута составляет менее 2% [83]. Кариотип у нута состоит из 8 пар хромосом, общая длина генома нута оценивается примерно в 738 миллионов пар оснований, количество генов - приблизительно 28 тысяч. Референсным геномом *C. arietinum* L. является сборка для нута дези [84].

3.2 Обзор популяционных исследований нута

Для *Cicer arietinum* L. были проведены несколько масштабных популяционных исследований. В одном из них авторы генотипировали 429 образцов нута из 45 стран и выявили 122 геномных региона, находящихся под действием отбора, а также 262 генетических маркера, связанных с 13 фенотипическими признаками [85]. Среди образцов в наборе присутствовали староместные сорта в количестве 268, а так же элитные культивары и селекционные линии. На основании анализа F_{ST} , авторы подтверждают регион плодородного полумесяца как центр первичной доместики нута и называют Южную Азию и Эфиопию как вторичные центры.

В 2021 году это исследование было расширено до 3366 образцов, включая 2436 староместных сортов и 197 дикорастущих образцов [9]. В ходе этого исследования было показано, что первое разделение рода *Cicer* на виды произошло около 21,4 (19,6–22,8) миллионов лет назад, а разделение между *C. arietinum* и *C. reticulatum* является самым молодым в эволюционном дереве рода и произошло менее 150 тысяч лет назад. Также было обнаружено, что нут пережил первое "бутылочное горлышко" около 10 000 лет назад в ходе доместики, и наименьшая эффективная численность популяции была достигнута примерно 1 000 лет назад. Затем, произошло сильное увеличение численности популяции в течение последних 400 лет из-за широкого использования нута в сельском хозяйстве. Кластеризация образцов в координатах двух первых главных компонент показала два возможных независимых пути распространения нута: (i) из Ближневосточного региона в Южную Азию и (ii) из Ближневосточного региона в Восточную Африку и Средиземноморский регион. Однако такие результаты не раскрывают в полной

мере эволюционные пути нута, в частности не проверяют смещения популяций и происхождение нута кабули.

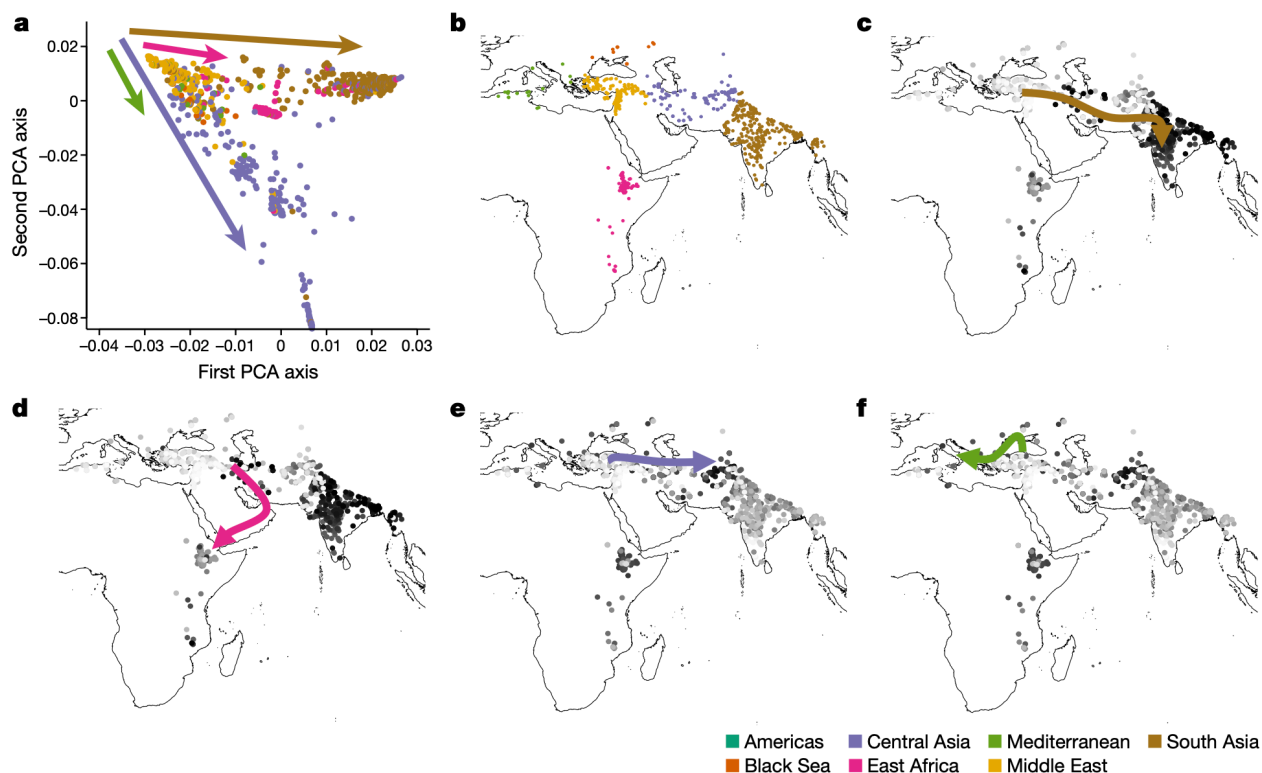


Рисунок 3.1 — Гипотетические пути распространения нута на основании анализа главных компонент. Рисунок заимствован из работы [9], Fig. 2. (a) Представление генетического разнообразия выборки в координатах двух первых главных компонент. (b) Географические точки сбора образцов. Популяции раскрашены согласно географическим кластерам. (c) Путь распространения нута в Южную Азию. (d) Путь распространения нута в Восточную Африку. (e) Путь распространения нута в Среднюю Азию. (f) Путь распространения нута в Средиземноморский регион.

Основным источником для составления набора образцов в этих двух работ была коллекция генетических ресурсов ICRISAT (International Crops Research Institute for the Semi-Arid Tropics). Однако ICRISAT не является единственным хранилищем, предлагающим широкий спектр семян нута. В России также существует обширная и независимая коллекция семян, известная как коллекция ВИР имени Н.И. Вавилова.

Начатые в 1916 г. экспедиции Н.И. Вавилова и его коллег, и продолженные его последователями, вылились в создание уникальной коллекции культурных растений и их диких форм со всего мира. По результатам экспедиций, Вавилов

выдвинул идею о "центрах происхождения культурных растений", то есть местах на Земном шаре, где большинство сельскохозяйственных культур было впервые domesticiровано человеком (см. Рисунок 3.2). В настоящее время коллекция хранится во Всероссийском институте генетических ресурсов растений имени Н. И. Вавилова (ВИР) и насчитывает более 300 000 образцов растений, в том числе 3380 образцов нута, почти половина из которых представляет староместные сорта с известными географическими координатами точек сбора. Эти сорта можно рассматривать как окончательный результат диверсификации культурных линий нута, локально адаптированных к выращиванию в конкретных географических зонах. Таким образом, эта коллекция нута является подходящим материалом для изучения его доместикиции, так как сохраняет больше следов истории, чем современные рыночные сорта научной селекции. Из имеющихся образцов нута в коллекции ВИР лабораторией М.Г. Самсоновой была отобрана панель из 421 образца, охватывающая все центры, предположенные Вавиловым, фенотипирована и генотипирована по 2759 однонуклеотидным локусам [5].

3.3 Исходные данные и обзор ранее проведенных на них исследований

Исходными данными для настоящей работы была панель из 421 образцов нута, отобранная в лаборатории М.Г. Самсоновой из коллекции ВИР. Фенотипирование панели было проведено в 2016 году на экспериментальной станции Кубанской ВИР, где климатические условия отлично подходят для выращивания нута (детальное описание фенотипирования приведено в дополнительных материалах к статье [5]). В течение вегетационного периода было измерено тридцать шесть фенологических, морфологических, агрономических и биологических характеристик.

Генотипирование образцов было выполнено с помощью технологии "генотипирование секвенированием" (genotype by sequencing, GBS) согласно протоколу из статьи [86]. GBS сочетает в себе технологию высокопроизводительного секвенирования (high-throughput sequencing) и технику усеченной репрезентации генома. Усеченная репрезентация происходит благодаря тому, что фрагментация генома перед ПЦР происходит не случайным образом, а с помощью сайт-специфичных рестриктаз. Таким образом, информация о геноме оказывается



Рисунок 3.2 — Карта центров происхождения культурных растений, предложенных Н.И. Вавиловым. Оригинал находится во Всероссийском институте растениеводства имени Н.И. Вавилова. Фото выполнено Иголкиной А.А.

ограниченной, но достаточной для характеристики вариабельности за счёт высокого значения величины неравновесия по сцеплению [5]. Данные секвенирования выбранных образцов нута доступны в базе данных NCBI (National Center for Biotechnology Information) под номером проекта PRJNA388691. После секвенирования были определены вариабельные ОНП для панели с использованием программы Genome Analysis Tool Kit (GATK) [87], а затем они были дополнительно отфильтрованы с помощью инструмента VCFtools [88]. Количество ОНП, прошедших все фильтрации оказалось 2579; их положение вдоль хромосом показано на Рисунке 3.3.

На основании полученных данных было проведено два полногеномных анализа ассоциаций для фенотипических признаков (Genome-wide association studies, GWAS): [5] и [89]. В первой работе, [5], были использованы модели, тестирующие ассоциации между каждым фенотипом и каждым ОНП маркером отдельно. Для этой работы, автором настоящей диссертации был разработан метод поиска

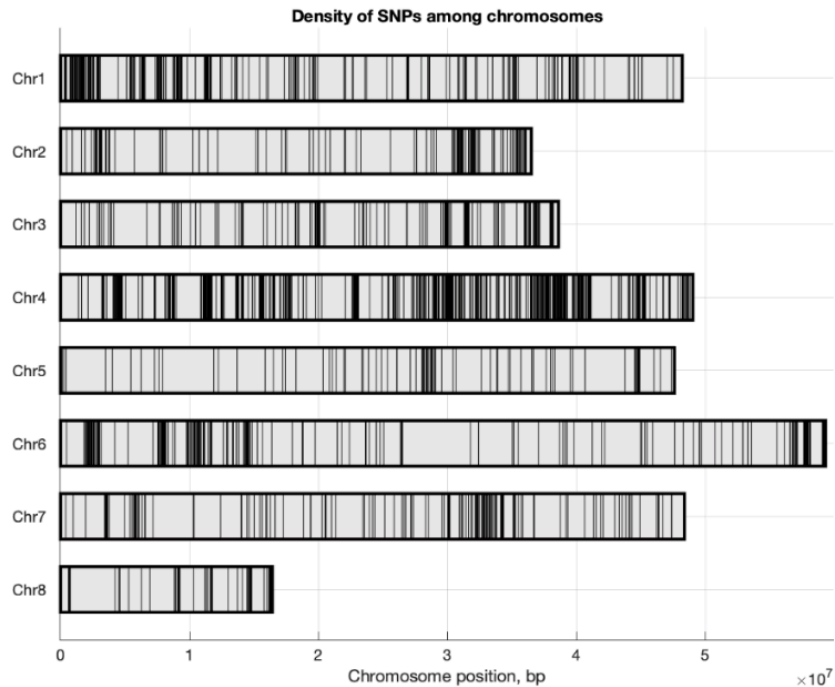


Рисунок 3.3 — Плотность ОНП вдоль хромосом генома *Cicer arietinum*
 L. Положение каждого ОНП показано вертикальной линией.

участков в геноме, обогащенных ОНП, ассоциированными с каждым конкретным фенотипическим признаком. По результатам этого анализа было выявлено, что некоторые фенотипические признаки имеют общие обогащенные участки генома, что свидетельствует о присутствии плеiotропности, то есть наличии генов, оказывающих эффект на несколько фенотипических признаков.

Во второй работе, [89], был проведен анализ ассоциаций с учетом плеiotропности. Автором настоящей диссертации был разработан метод mtmlSEM, позволяющий объединить все фенотипические признаки в модель структурных уравнений, которая позволяет искать ОНП плеiotропного и однофенотипного эффектов. Построение модели mtmlSEM состоит из двух этапов. На первом этапе происходит формирование латентных переменных, описывающих общую дисперсию сразу нескольких фенотипов, и объединение их в направленный граф. На втором этапе, согласно жадному подходу и принципу максимума правдоподобия, в модель сначала добавляются плеiotропные ОНП, то есть ассоциированные с введенными латентными переменными, а затем ОНП однофенотипного эффекта. Тестирование ОНП перед их добавлением осуществляется с помощью методов пакета *sem*тору, разработанного автором настоящей диссертации [90]. После того, как модель построена, запускается байесовская оптимизация параметров.

Еще одним исследованием, которое было проведено с участием автора настоящей диссертации и применено к генетическим данным панели из 421 образца, была разработка метода детектирования вредных мутаций в белок-кодирующих областях [Kovalev2018]. Модель на основе случайного леса (Random Forest) показала, что, будучи обученной на данных о вредных мутациях *Arabidopsis thaliana*, она способна с высокой точностью предсказывать вредные мутации у других растений. В результате применения модели к несинонимичным мутациям в данных 421 образцов нута была получена карта вредных мутаций вдоль генома.

Глава 4. Результаты применения разработанных моделей

Результаты, представленные в данной главе, были опубликованы в научном журнале *Molecular Biology and Evolution* в 2023 году [91].

4.1 Популяционный анализ исходных данных

Для выявления популяционной структуры в генотипированной панели из 421 образца нута была использована программа ADMIXTURE [12]. Эта программа в качестве параметра принимает количество популяций, K , которые предполагаются базовыми (исходными или предковыми) в модели. Программа ADMIXTURE была последовательно запущена со значениями параметра K от 2 до 20. При этом ошибка кросс-валидации монотонно уменьшалась и не достигала минимума (Рисунок 4.1). То есть никакое значение K нельзя было назначить оптимальным. Такое поведение возможно связано с тем, что в наборе данных присутствуют очень похожие образцы, например, собранные в одной локации, и программа ADMIXTURE стремится назначить таким образцам отдельную базовую популяцию. Чтобы получить приемлемое количество базовых популяций, все значения K были последовательно проанализированы с одновременным сопоставлением расположения композиционного состава образцов на географической карте (Рисунок 4.2).

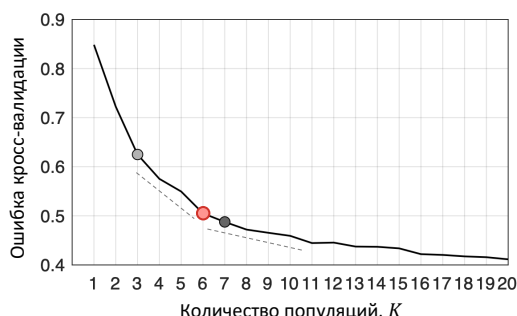


Рисунок 4.1 — **Ошибка кросс-валидации классификации программы ADMIXTURE при варьировании параметра количества базовых популяций. Красной точкой обозначено количество популяций, которое было взято за оптимальное.**

Анализ ADMIXTURE при $K = 2$ выделяет в отдельную группу Азиатские регионы от остальных. Анализ при $K = 3$ выделяет в отдельную группу Эфиопский регион вместе с частью Ближнего Востока и Средиземноморья. Анализ с $K = 4$ разделяет Среднеазиатские и Индийские регионы как отдельные группы. Анализ с $K = 5$ делит Среднеазиатский регион на две группы. Анализ с $K = 6$ изолирует Эфиопские образцы (Рисунок 4.2). Более высокие значения K не вносят новой географически обособленной группы образцов, поэтому значение $K = 6$ было принято за оптимальное. Полученные шесть групп образцов практически полностью географически изолированы, что позволило установить однозначное соответствие между генетическими группами и географическими регионами, присвоив им названия ближайших современных стран: Эфиопская, Индийская, восточно-Среднеазиатская, западно-Среднеазиатская, северная Ближневосточная (включая северную Средиземноморскую), южная Ближневосточная (включая южную Средиземноморскую).

Компактное расположение групп смешения на географической карте подтверждает наличие сильного географического сигнала в генетических данных. Незначительные несоответствия свидетельствуют о возможных малых перемещениях семян, не связанных с основной историей формирования популяций нута в регионах, или следы экспериментального загрязнения. Поэтому образцы, которые показывали несоответствие между географическим регионом и группой смешения, были исключены из дальнейшего анализа; географическое положение исключенных образцов показано на Рисунке 4.3. Кроме того, из анализа были исключены сильно смешанные образцы, то есть те, в которых вклад наибольшей из базовых популяций составлял менее 60% (т.е. ни одна из базовых популяций не имела доминирующего эффекта). В результате после фильтрации осталось 294 образца. (Таблица 3).

Оставшиеся после фильтрации образцы были разделены на 9 географических регионов, каждый из которых был представлен одной группой смешения. при этом, существовали группы смешения, которые охватывали несколько географических регионов (Рисунок 4.4А). Сравнение композиций групп смешения в образцах после фильтрации представлено на Рисунке 4.4Б. Больше всего образцов было потеряно в результате фильтрации в Среднеазиатском регионе, так как изначально он содержал большее количество образцов по сравнению с другими регионами.

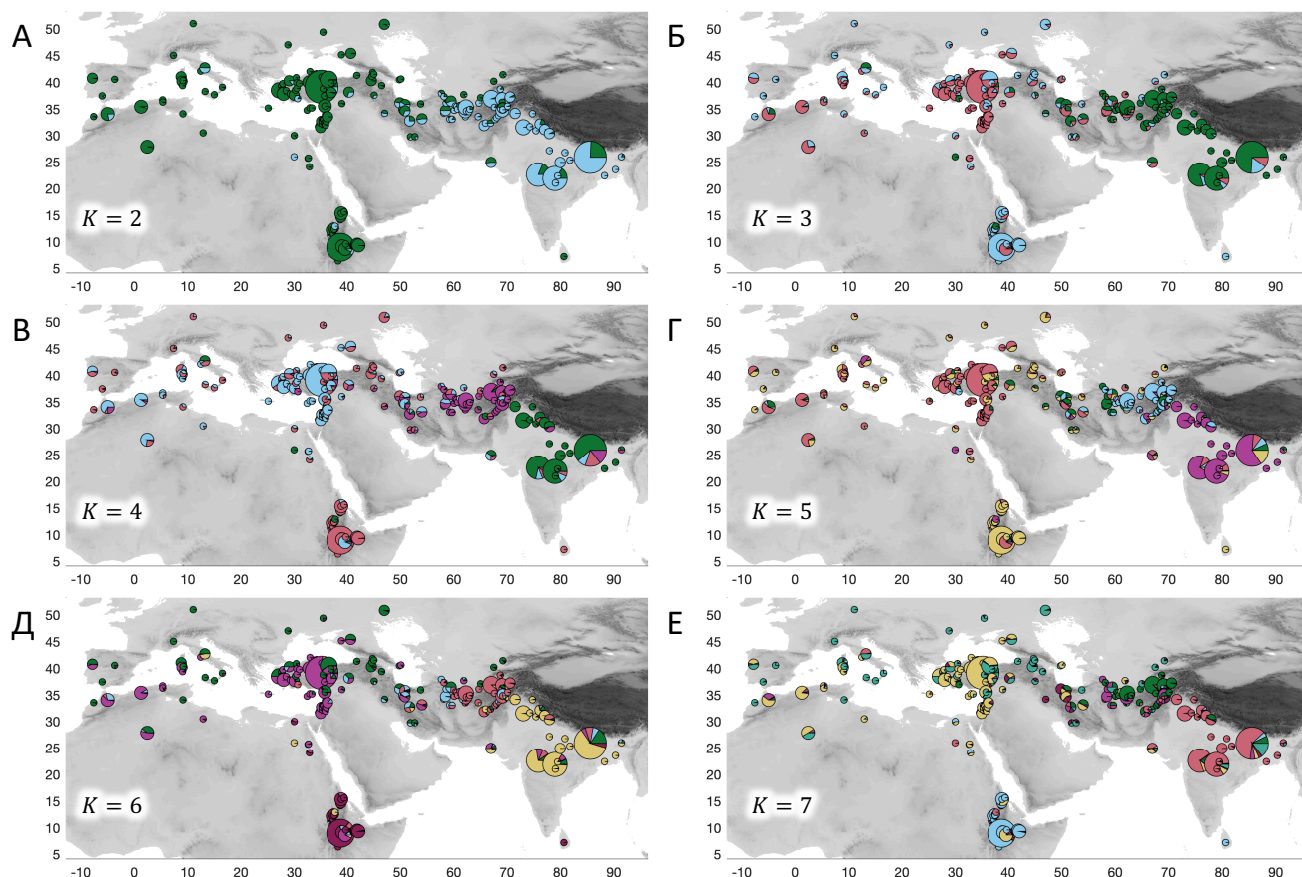


Рисунок 4.2 — Результаты работы программы ADMIXTURE при разных значениях параметра K . При каждом K каждому образцу была сопоставлена популяция с доминирующим вкладом. Положения круговых диаграмм соответствуют точкам сбора образцов. Поскольку в одной точке может быть отобрано несколько образцов, каждая диаграмма соответствует распределению K исходных популяций для образцов в этой точке. Цветовая палитра на каждом подрисунке независима и состоит из K случайных цветов. (А) Результаты при $K = 2$. (Б) Результаты при $K = 3$. (В) Результаты при $K = 4$. (Г) Результаты при $K = 5$. (Д) Результаты при $K = 6$. (Е) Результаты при $K = 7$.

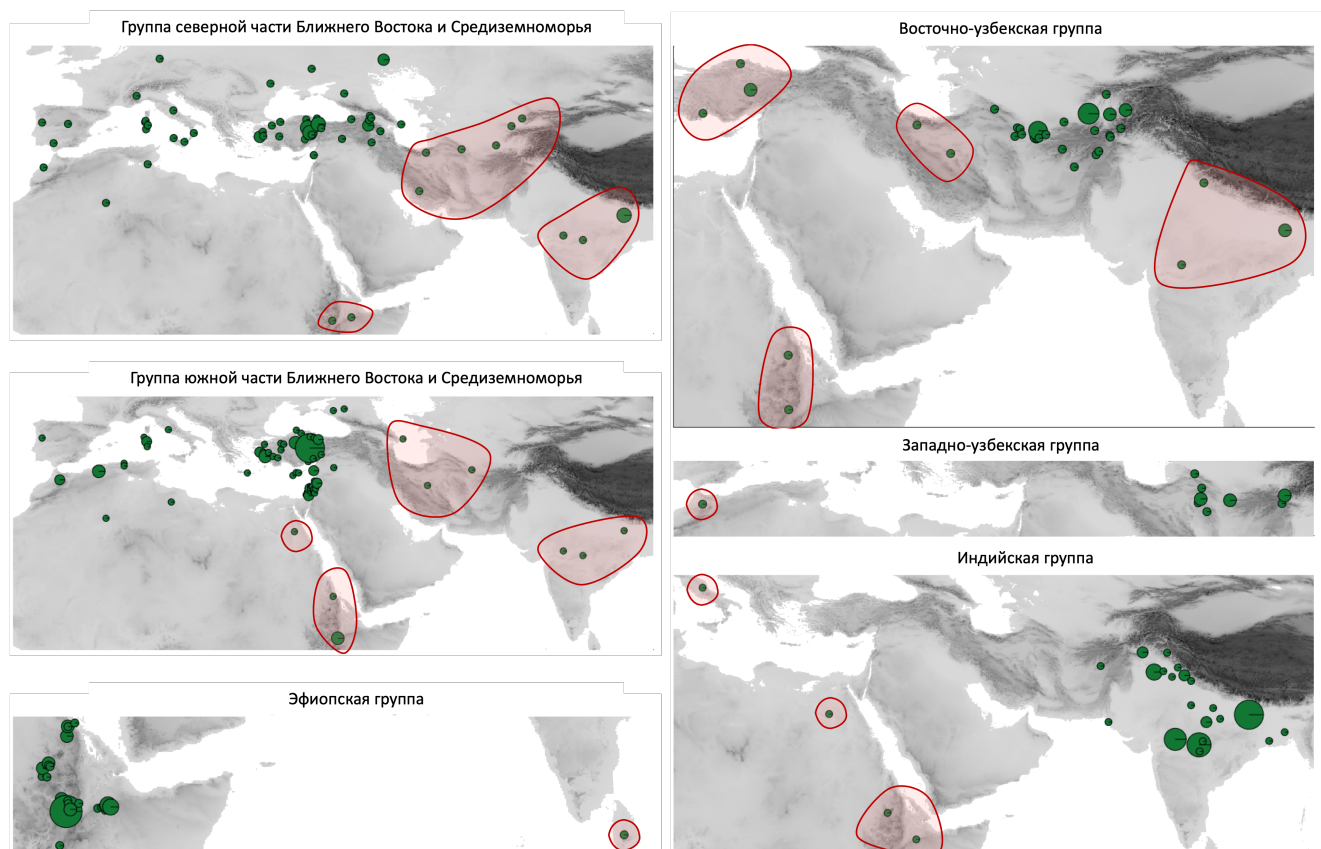


Рисунок 4.3 — Точки сбора образцов, соответствующих каждой из шести исходных популяций. Образцы, находящиеся за пределами основного ареала распространения популяции, были исключены из анализа. Эти образцы обведены красными линиями.

В результате анализа ADMIXTURE группы сортов дези и кабули не были выделены в отдельные исходные популяции, то есть географический сигнал в данных превышает сигнал дези-кабули. Анализ главных компонент на генетических данных показывает, что первые две главные компоненты объясняют 72% дисперсии, а первые 6 главных компонент объясняют 90% дисперсии (Рисунок 4.5 А). В координатах первых двух главных компонент образцы группируются в кластеры, соответствующие географическим регионам, причем первая главная компонента отражает вариацию дези-кабули в кластерах. Тем не менее, разделение на дези и кабули в первых двух главных компонентах не является строгим, что указывает на отсутствие репродуктивной изоляции между этими группами сортов. Несмотря на отсутствие репродуктивной изоляции, дези и кабули являются изолированными, так как культивируются отдельно. Поскольку географический сигнал в данных оказывается сильнее сигнала групп сортов, можно заключить, что географическая изоляция произошла раньше, чем изоляция групп. Иными словами, domestикация кабули произошла после того, как нут распространился

Таблица 3 — **Количество образцов в популяциях до и после фильтрации.** Зеленым цветом подсвечены образцы, которые были взяты для анализа.

Регион	До фильтрации			После фильтрации		
	дези	кабули	неопр.	дези	кабули	остальное
Эфиопия	61	3	0	50	0	14
Индия	64	5	1	57	0	32
Ливан	18	16	0	10	12	12
Турция	40	73	4	13	39	29
Черноморский				18	18	
западно-Среднеаз.	70	25	2	10	0	54
восточно-Среднеаз.				20	13	
северно-Средиземн.	19	19	1	5	11	8
южно-Средиземн.				9	6	

по регионам. Поэтому при анализе образцов нута, помимо разбиения на регионы, было применено дальнейшее деление региональных популяций на субпопуляции дези и кабули.

Используя статистику F_{ST} , иерархическая кластеризация популяций показала, что субпопуляции дези и кабули из одного региона всегда располагаются в смежных ветвях дерева (Рисунок 4.5 Б). Наибольшее генетическое расстояние между популяциями соответствует географическому разделению между Азией и Ближним Востоком. Однако две популяции, занимающие необычное положение на дереве, — это Эфиопские и западно-Среднеазиатские дези. Такое расположение на филогенетическом дереве, особенно у западно-Среднеазиатских дези без явной географической изоляции, свидетельствует скорее о смешанном происхождении. Из-за сходства в необычном расположении можно допустить, что Эфиопская популяция также имеет смешанное происхождение.

Исходя из того, что две популяции Средиземноморского региона принадлежат разным группам смешения, при этом северная популяция находится в одной группе с Черноморской, а южная — с Турецкой и Ливанской, можно предположить, что Средиземноморские популяции возникли в результате двух независимых путей распространения нута из Ближнего Востока на запад вдоль южного и северного побережий Средиземного моря соответственно. В настоящей работе одним из методов анализа нута в регионах является оценка частот аллелей, которая требует построения географических путей перемещения внутри региона и их взвешенной оценки. Однако, поскольку в Средиземноморских популяциях такие пути проходят частично по воде, получение взвешенной оценки

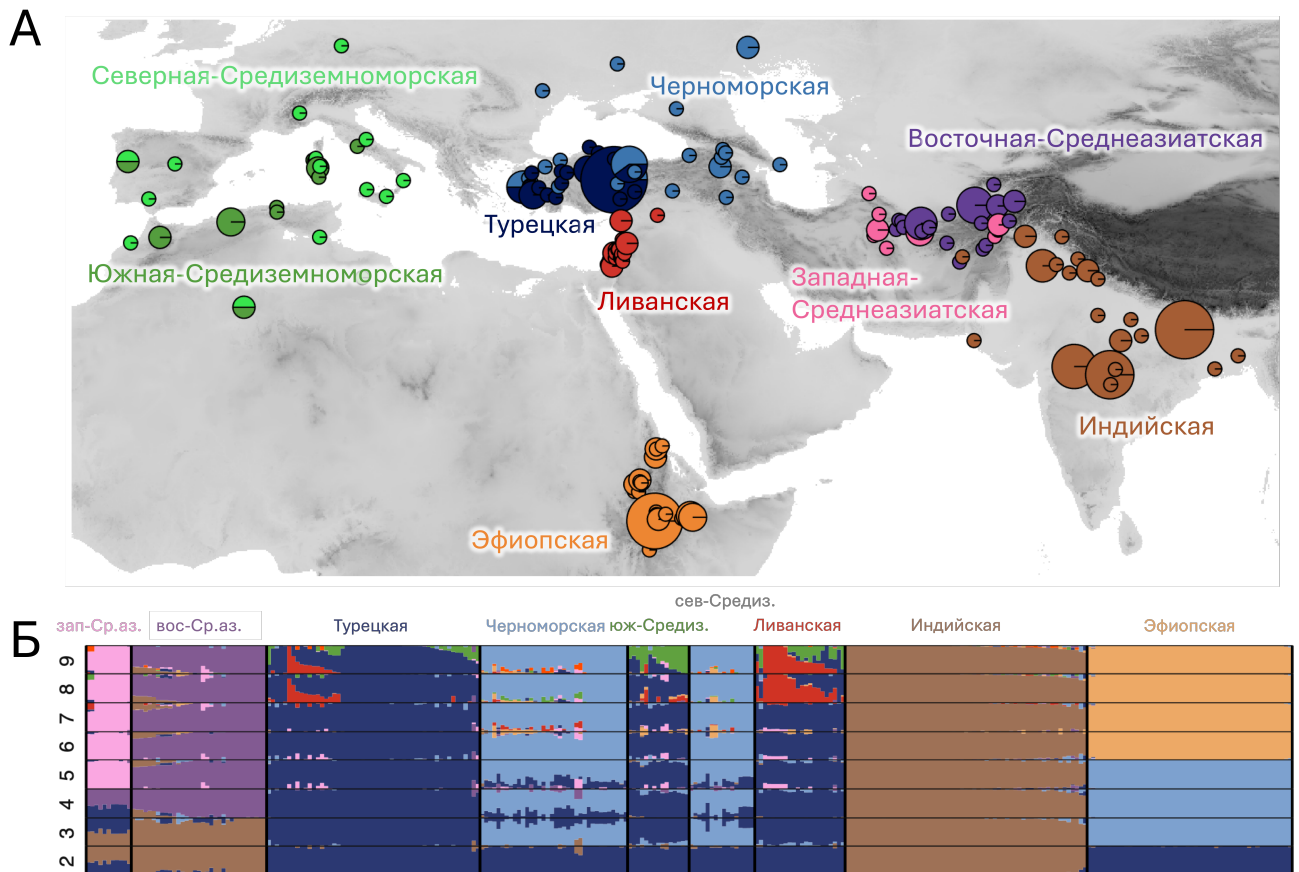


Рисунок 4.4 — Характеристика образцов в выделенных девяти популяциях после фильтрации. (А) Географическое разделение популяций, каждая из которых обозначена своим цветом. Размер круговой диаграммы показывает относительное количество образцов в точке сбора. **(Б)** Результаты программы ADMIXTURE для параметра K от 2 до 9. Результаты демонстрируют генетическую чистоту популяций.

таких путей является нетривиальной задачей, так как требует экспертного анализа относительной сложности перемещения по воде и по суше. Количества образцов в Средиземноморских популяциях в основном составляют менее 10, то есть эти популяции относительно слабо представлены в данных. Из-за сложности построения путей и малой представленности Средиземноморские популяции были исключены из дальнейшего анализа. Относительно формирования этих популяций можно сделать следующие выводы: северная Средиземноморская популяция была получена как дочерняя популяция от Черноморской, а южная Средиземноморская — как дочерняя от Турецкой, но не от Ливанской. Последнее следует из композиционного графика ADMIXTURE анализа при $K = 7$, где Ливанская популяция образует отдельный кластер, не совпадающий по составу с южной Средиземноморской популяцией.



Рисунок 4.5 — Сходство и различия между образцами популяций нута. (А) Представление образцов в первых двух главных компонентах. (Б) Иерархическая кластеризация популяций на основании значений статистики F_{ST} .

4.2 Модель popdisp

Метод popdisp оценивает характеристические частоты аллелей в популяции с учетом ковариационной матрицы, связывающей образцы этой популяции. Для метода popdisp была разработана специальная схема симуляции, которая моделирует распространение популяции по географически неоднородной территории. Сначала метод popdisp был протестирован на симулированных данных, после чего он был применен к данным по нуту для определения характеристических частот аллелей в популяциях.

4.2.1 Тестирование на симуляциях

Для тестирования программы popdisp была разработана специальная схема симуляции данных. Каждая симуляция представляла собой распространение популяции из некоторого центра по географически неоднородному региону вдоль бинарного ветвящегося пути. В начале симуляции задавался вектор частот аллелей для популяции, затем этот вектор плавно изменялся вдоль пути, так что каждой конечной точке пути соответствовал свой вектор частот аллелей. В каж-

дой конечной точке генерировалась выборка образцов по полученным частотам. Эта выборка образцов и матрица связей между образцами на основании пути подавались на вход программе `popdisp`. Точность работы программы определялась как отклонение оцениваемых частот аллелей в центральной точке от частот аллелей, заданных в начале симуляции. Симуляция имела следующие основные параметры: начальные частоты аллелей в центре, количество точек сбора образцов в регионе, возможное количество образцов в каждой точке.

Симуляция бинарно ветвящегося распространения популяции по региону происходила следующим образом. Сначала было сгенерировано случайное взвешенное укорененное бинарное дерево с числом листьев, равным заданному количеству точек сбора образцов. Корень дерева предполагался соответствующим центру, из которого происходило распространение популяции. Затем это дерево было размещено на двумерной плоскости согласно алгоритму `equal-daylight` [92] (Рисунок 4.6А). Затем ветви были разделены на части приблизительно одинакового размера, и каждая точка соединения частей была незначительно и случайным образом отклонена (Рисунок 4.6Б). Таким образом, путь из центра в конечные местоположения определялся набором точек и соединяющими их прямыми линиями. После этого в прямоугольной области, занимаемой деревом, случайным образом были размещены пять неблагоприятных зон, которые имитировали, например, горы, то есть зоны, через которые трудно перемещаться. Каждая зона представляла собой симметричное двумерное нормальное распределение вокруг центра зоны, а все зоны вместе формировали ландшафт (Рисунок 4.6В). Для того, чтобы путь из центра в конечные местоположения соответствовал оптимальному перемещению по ландшафту, положение каждой точки было изменено согласно нескольким итерациям алгоритма градиентного спуска по полученному ландшафту. В результате получались пути, которые выглядели естественно для распространения по географическому ландшафту популяции из центра в конечные локации (Рисунок 4.6В).

После создания пути распространения популяции были сгенерированы генетические данные для образцов в конечных точках этого пути. Для центральной популяции было определено 99 ОНП с частотами аллелей, изменяющимися от 0,01 до 0,99 с интервалом 0,01. Затем частоты были преобразованы в пространство балансов, где с некоторой дисперсией была проведена симуляция изменения частот аллелей вдоль пути, аналогичного броуновскому движению. В результате каждой конечной точке соответствовал свой вектор частот аллелей (Рисунок

4.6Г). В каждой конечной точке было определено случайное количество образцов от 1 до некоторого заданного для симуляции значения, и для каждого образца был сгенерирован вектор аллелей ОНП (значения 0 или 1) в соответствии с биномиальным распределением, основанным на частотах аллелей, полученных для данной конечной точки пути.

Для создания тестового набора была запущена 100 симуляций, как описано выше, при этом количество образцов в конечных точках пути было случайно выбрано из равномерного распределения на отрезке от 1 до 30 образцов. На полученных данных были протестированы три характеристики: чувствительность `popdisp` к матрице связей между образцами, точность при оценке частот аллелей в центральной популяции, устойчивость метода к разной степени искажения выборки. При первом тестировании для каждой симуляции были созданы три матрицы связей между образцами: (i) матрица, соответствующая пути распространения (то есть корректная матрица), (ii) матрица, основанная на линейных расстояниях между образцами, и (iii) случайно сформированная квадратная матрица с диагональным преобладанием. Во всех случаях наименьший показатель ВИС наблюдался при запуске `popdisp` с использованием корректной матрицы, в то время как наибольший — при запуске со случайной матрицей (4.7А). Иными словами, `popdisp` правильно идентифицировал модель распространения, использованную в симуляции. С точки зрения точности, `popdisp` более аккуратно оценивал вектор частот аллелей в центральной популяции при использовании корректных матриц, чем при использовании любых других. Кроме того, эта оценка была статистически значимо лучше, чем вектор выборочных средних частот аллелей в популяции (4.7Б).

Для тестирования устойчивости оценки вектора частот аллелей, полученной с помощью `popdisp`, к размеру выборки, был произведен ряд сокращений образцов в конечных точках симуляционных данных. Максимально допустимое количество образцов в конечных точках было сначала снижено до 10, затем до 5, и, наконец, до 2 образцов. В результате была сформирована панель из трех наборов симулированных данных для тестирования. С уменьшением допустимого количества образцов `popdisp` демонстрировал увеличение ошибки в оценке вектора частот аллелей; при этом, даже в случае с 1-2 образцами на точку сбора образцов, оценка вектора частот аллелей оказалась лучше, чем при использовании простого выборочного среднего (4.7В). Стоит отметить, что в каждом из этих трех случаев на вход в `popdisp` подавалась корректная матрица связей между образцами. С це-

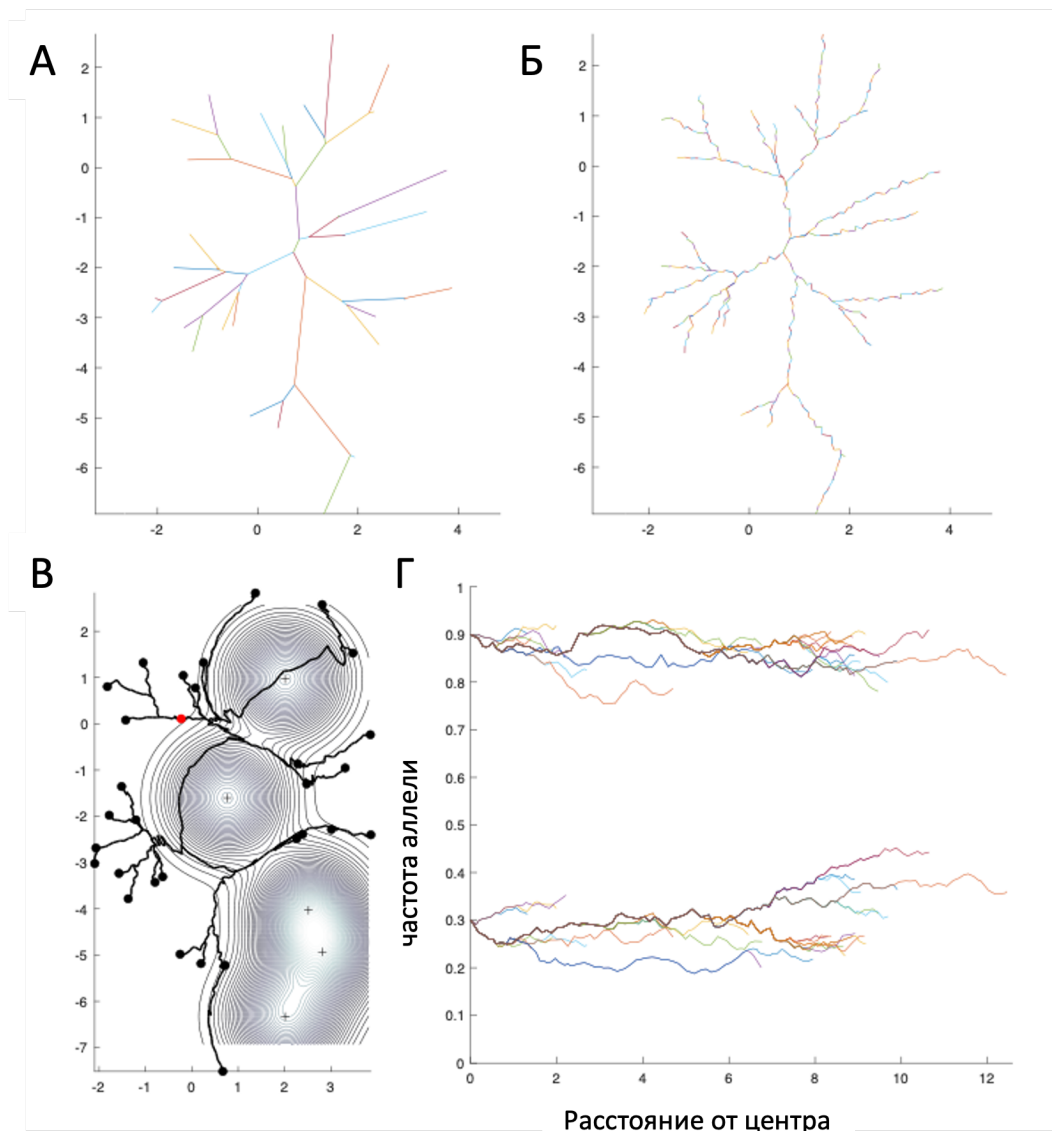


Рисунок 4.6 — Пример симуляции одного набора данных для тестирования программы **popdisp**. (А) Визуализация случайного дерева на 2D поверхности согласно алгоритму equal-daylight. Цвета назначены ветвям случайным образом. (Б) Ветви дерева разделены на фрагменты приблизительно одинаковой длины, соединительные точки между которыми подверглись отклонениям. Цвета назначены фрагментам случайным образом. (В) Ландшафт с пятью неблагоприятными зонами, центры которых обозначены крестами. Контурные линии, соединяющие точки с одинаковой ”неблагоприятностью”. Черные точки показывают точки сбора образцов (соответствуют положениям листьев на предыдущих деревьях), красная точка — расположение центра региона. Ломаные линии — оптимальные пути из центра к местам отбора образцов. (Г) Два примера симуляции частот аллелей вдоль полученной структуры распространения (начальные частоты в центре предполагаются быть 0.3 и 0.9). Цвета представляют пути из центра к различным точкам сбора образцов.

лью дальнейшей проверки устойчивости `popdisp` к размеру выборки, количество конечных точек в симулированных данных было уменьшено до 20 и до 10. По результатам работы `popdisp` на двух новых набора данных, ошибка оценки вектора частот аллелей в центральной популяции вела себя предсказуемым образом, при этом оставаясь ниже, чем ошибка выборочного среднего (4.7B).

Устойчивость оценки `popdisp` к искажению в данных была продемонстрирована на измененных симулированных данных, в которых выборки образцов в пяти, а затем в десяти точках сбора, были случайным образом поменяны местами. При увеличении искажения ошибка оценки вектора частот аллелей в центральной популяции ожидаемо росла, но все же оставалась ниже, чем ошибка выборочного среднего (4.7B). Исходя из проведенных экспериментов, можно заключить, что `popdisp` является методом, чувствительным к матрице ковариаций между образцами, точным в оценке частот и устойчивым к различного рода шуму.

4.2.2 Результаты на реальных данных по нуту

По результатам анализа популяционной структуры панели образцов нута было выделено 11 популяций (7 популяций дези и 4 популяции кабули). Каждой популяции соответствовал отдельный географический регион и одна доминантная генетическая группа смешения. Для каждой популяции была построена сеть попарных оптимальных географических путей между точками сбора образцов согласно функции наименьшей стоимости Герцог [35], реализованной в R пакете `leastcostpath` [93]. На основании путей были получена матрица попарных расстояний между образцами.

Помимо точек сбора образцов, в регионе каждой популяции дополнительно были назначены локации 16-ти гипотетических репрезентативных центров. Для этого сначала были определены крайние положения образцов в рассматриваемой популяции по широте и долготе, что определило прямоугольник региона. Затем, внутри этого прямоугольника, был построен второй, с отступом в 10% от длин сторон. Получившийся внутренний прямоугольник послужил рамкой для сетки 4×4 , узлы которой были взяты за гипотетические центры. Центры, которые попадали на территории водоемов, были исключены из рассмотрения как нереалистичные положения центров доместикации. От центров до всех точек сбора образцов в

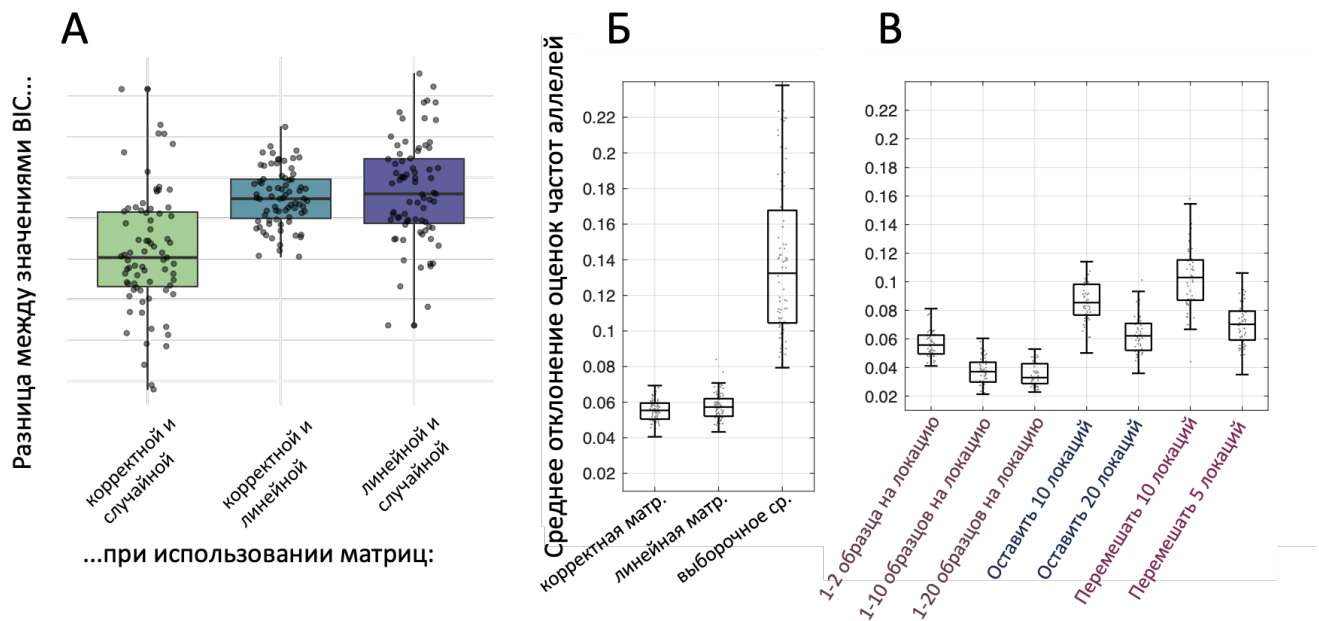


Рисунок 4.7 — **Результаты работы программы popdisp на симулированных данных.** (А) Разница между BIC критериями для запусков popdisp с тремя матрицами: корректной, линейными расстояниями и случайной. Чем меньше значение BIC, чем лучше модель соответствует данным. То есть чем меньше разница, тем лучше первый вариант матрицы подходит для описания данных. (Б) Распределение средних абсолютных отклонений оценок частот аллелей в центрах популяций от их фактических значений по результатам работы программы popdisp на всех симуляциях. (В) Средние абсолютные значения различий между оценками частот аллелей в центрах популяций и фактическими значениями для данных, полученных в различных сценариях искажения.

свою очередь также были построены оптимальные пути, что позволило получить из матриц расстояний ковариационные матрицы между точками образцов для каждого центра согласно формуле на Рисунке 2.4. Примеры путей из центров до точек образцов представлены на Рисунке 4.8. В результате для каждого гипотетического центра были построены две альтернативные ковариационные матрицы: (1) на основании стоимости оптимальных путей (2) на основании линейных расстояний между образцами.

Модель popdisp была применена к каждой из 11 популяций, 16 гипотетическим центрам в каждой и 2 матрицам ковариации для каждого центра, что в совокупности составило 352 запуска. Сначала, для каждого из центров было определено, какая модель распространения является наилучшей с точки зрения

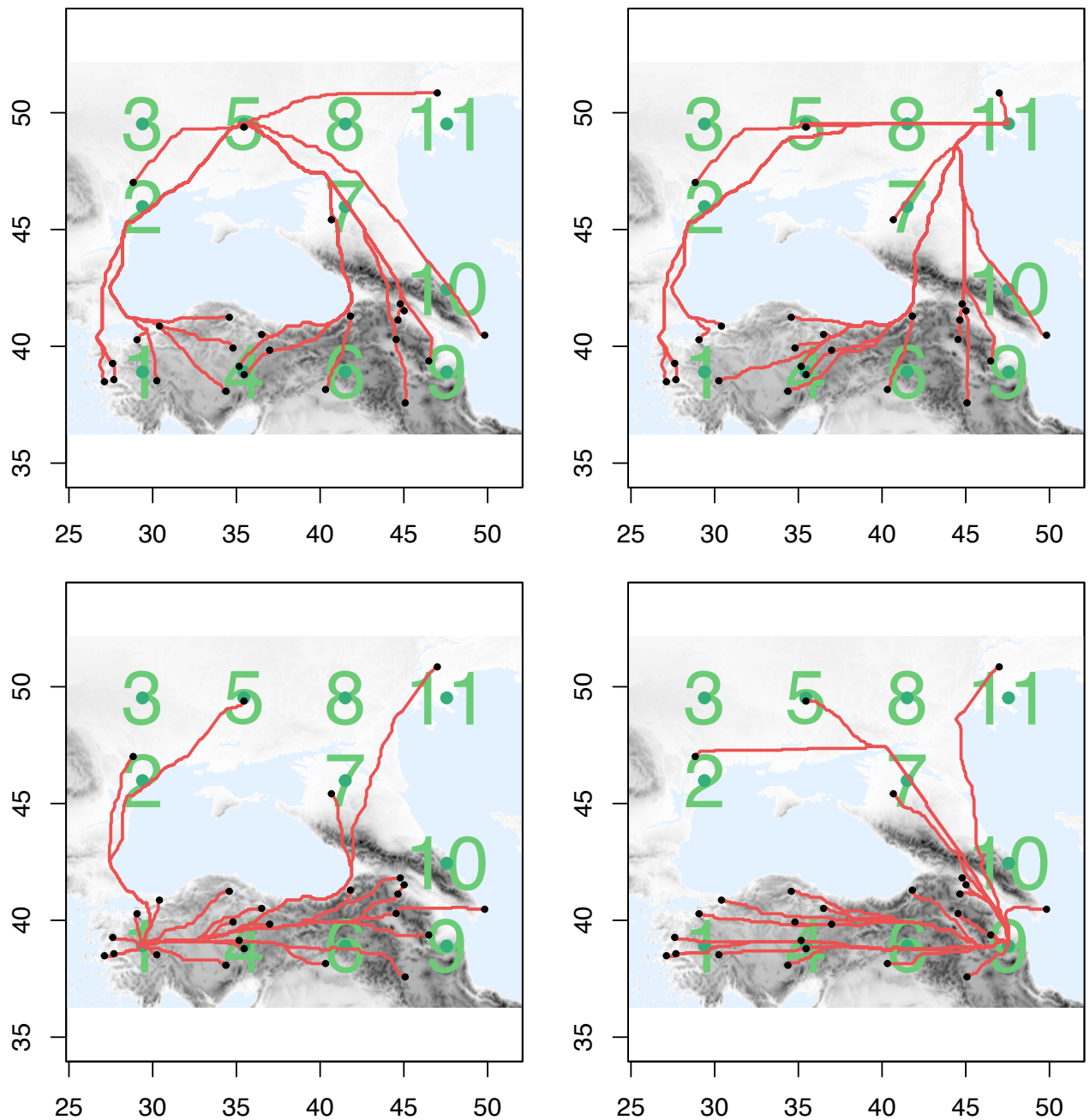


Рисунок 4.8 — Примеры построенных оптимальных путей для Черноморской популяции. Зелеными цифрами обозначены узлы на сетке 4×4 гипотетических репрезентативных центров. Центры, расположенные на водных объектах, были удалены. Черными точками указаны точки сбора образцов. Красными ломаными линиями изображены оптимальные пути перемещения от центров к точкам сбора с учетом географического ландшафта.

критерия ВИС. Оказалось, что модель распространения, учитывающая стоимость перемещений, являлась лидирующей почти во всех регионах (Таблица 4). Полученный результат указывает на то, что генетическое разнообразие образцов в популяциях больше соответствует модели распространения, учитывающей географические особенности региона, нежели модели без такого учета. В качестве репрезентативного центра для каждой популяции было взято местоположение с наибольшим значением ВИС. В регионах, где присутствовали обе популяции (дези и кабули), полученные репрезентативные центры либо были одинаковыми, либо располагались в соседних положениях. Этот факт подтверждает, что полученные центры являются репрезентативными для регионов.

Таблица 4 — Доминирование взвешенных моделей распространения в популяциях.

Номер популяции	Название популяции	Группа сортов	Доля центров с лидированием взвешенной модели распространения
1	Турецкая	дези	0.93
2	Турецкая	кабули	1.00
3	Черноморская	дези	1.00
4	Черноморская	кабули	0.09
5	Ливанская	дези	1.00
6	Ливанская	кабули	0.62
7	восточно-Среднеазиатская	дези	1.00
8	восточно-Среднеазиатская	кабули	1.00
9	западно-Среднеазиатская	дези	1.00
10	Эфиопская	дези	0.20
11	Индийская	дези	0.67

Для двух из 11 популяций — Черноморские кабули и Эфиопские дези — наилучшей оказалась модель распространения на основе линейных расстояний. Поскольку эти два региона в целом не показали преимуществ взвешенной модели для совокупности гипотетических центров (Таблица 4), вероятно, функция Герцог не подходит для описания перемещения в этих географических регионах. Однако в этих двух популяциях центры, соответствующие наилучшим значениям ВИС для моделей прямолинейных путей, также были лучшими по значениям ВИС для

моделей оптимальных географических путей, и различия в значениях ВИС между двумя моделями были менее выраженными, чем между разными центрами. Таким образом, модель с учетом стоимости перемещения, хоть и не является оптимальной для этих двух регионов, остается приемлемой. Поэтому, во избежание неоднородности подходов, для оценок частот аллелей в всех популяциях использовались оценки, полученные с помощью popdisp на основе моделей с учетом стоимости перемещения.

4.3 Модель migadmi

В качестве входных данных метод migadmi использует векторы частот аллелей в популяциях, например, полученные с помощью popdisp. На их основе он рассчитывает матрицу расстояний между популяциями и затем оценивает параметры миграций и событий смещений. Сначала точность метода migadmi была продемонстрирована на симулированных данных. Затем migadmi был применен для анализа гипотез о смещениях популяций нута. При формировании гипотез учитывались два утверждения: (i) доместикация нута дези и его распространение произошли до доместикации нута кабули, (ii) в каждом географическом регионе кабули является результатом смещений первых доместичированных кабули и дези из того же географического региона. Первое утверждение является общепринятым и было сформулировано на основании археологических данных и ряда популяционных исследований с генотипированием образцов. Второе - на основании анализа сходства векторов частот аллелей в популяциях: наиболее близкой для каждой популяции кабули является популяция дези из того же географического региона (Рисунок 4.5Б).

4.3.1 Тестирование на симуляциях

Для проверки функциональности пакета migadmi были создан набор из 100 симулированных данных. На первом этапе симуляции формировалась случайная топологии событий миграции и смещений, после чего каждому событию

смешения назначались веса. В начале каждой симуляции определялись четыре несмешанные популяции, объединенные симметричным бинарным деревом. Затем к дереву последовательно были добавлены две популяции, каждая из которых являлась результатом отдельного события смешения. Каждое из этих событий с равной вероятностью основывалось на двух или трех случайно выбранных популяциях-источниках. Далее каждому ребру в созданном графе смешений и каждому вкладу популяций-источников в смесь присваивались случайные положительные значения. По полученной истории смешений была построена точная матрица расстояний между популяциями. Оптимизация параметров с этой матрицей в качестве входных данных использовалась для демонстрации идентифицируемости параметров и возможных ограничений метода оптимизации.

Поскольку метод *migadmi* применялся для определения весов событий смешений, точность оценок параметров оценивалась исключительно на основе этих значений. По результатам 100 симуляций среднее абсолютное отклонение оценок весов от исходных значений при использовании точной матрицы расстояний составило менее 0.001. Такой результат указывает на то, что метод *migadmi* точен и ограничен методом оптимизации (в данном случае SLSQP), но не неидентифицируемостью параметров.

4.3.2 Реконструкция эволюционной истории популяций нута дези

Как было описано в разделе 3.1, у domesticiрованного нута существуют две основные группы сортов: дези и кабули. Нут дези считается первым domesticiрованным из дикого вида *C. reticulatum* и послужил источником для нута кабули. Поэтому в анализе популяций с помощью *migadmi* сначала была восстановлена эволюционная история популяции дези, а затем кабули. Входными данными для *migadmi* были векторы характеристических частот аллелей в популяциях, полученные с помощью *popdisp*.

В качестве основы для построения графа смешений использовалось бинарное дерево эволюционных связей между пятью популяциями дези, считающимися несмешанными: Турецкой, Черноземноморской, Ливанской, восточно-Среднеазиатской и Индийской (Рисунок 4.9). Топология бинарного дерева была определена как подструктура иерархической кластеризации, по-

строенной на основании F_{ST} статистики (Рисунок 4.5Б). Корень этого дерева был положен согласно разделению популяций по результатам ADMIXTURE при $K = 2$. Оставшиеся популяции дези были предположены как смешанные: Эфиопская и западная-Среднеазиатская популяции. Стоит заметить, что эти две популяции имеют необычно длинные ветви на дереве, построенном по F_{ST} -статистике, что косвенно может указывать на их смешанное происхождение.

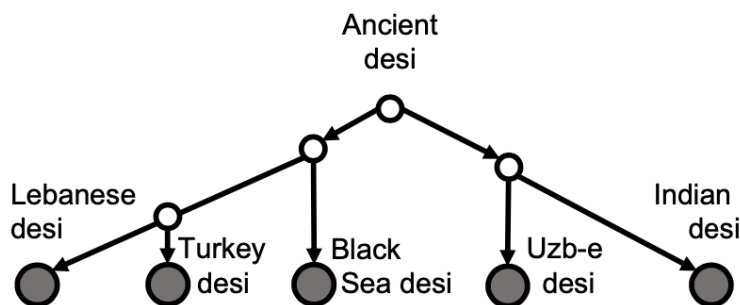


Рисунок 4.9 — Исходное эволюционное дерево, соединяющее пять несмешанных популяций нута дези.

Сначала было протестировано происхождение Эфиопской дези популяции. Были рассмотрены все возможные варианты, в которых Эфиопская популяция могла бы произойти от двух или трёх несмешанных популяций. Наименьшее значение BIC соответствовало случаю, при котором Эфиопская популяция имеет две предковые популяции: Черноморскую и Индийскую. Второе по меньшинству значение BIC соответствовало предположению, что Эфиопская популяция - это смесь Ливанской и Индийской популяций (Рисунок 4.10). То есть Эфиопская популяция является результатом смешения из двух основных групп, полученных в результате анализа ADMIXTURE. Если рассматривать Эфиопскую популяцию как результат смешения трех Ближневосточных популяций, то популяции Черноморских дези действительно показывает наибольший вклад в Эфиопский дези, что было выдвинуто ранее как гипотеза в статье [9]. Если рассматривать Эфиопскую популяцию как смесь Индийской и восточно-Среднеазиатской, то наибольший вклад соответствует Индийской популяции, как и в наилучшей модели.

Для более детального анализа участков хромосом, которые были унаследованы от несмешанных предковых популяций, были проанализированы распределения весов смешения вдоль хромосом. Программа *mgadm* была применена в технике скользящего окна, где все параметры смеси оценивались отдельно

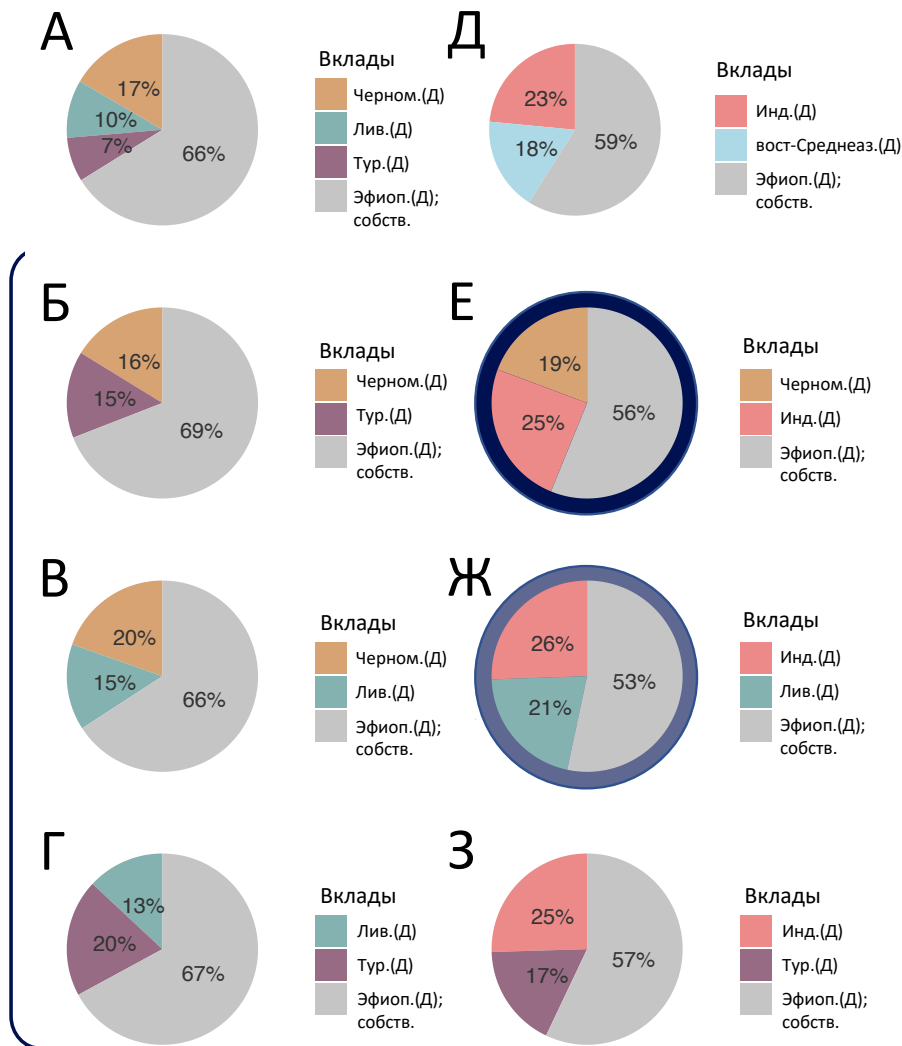


Рисунок 4.10 — **Результаты моделирования Эфиопской дези популяции как смеси.** Наиболее правдоподобная модель (Подрисунок Е) выделена темно-синим цветом. Следующая по значению правдоподобия модель выделена светло-синим цветом (Подрисунок Ж). В первом столбце представлены результаты моделирования Эфиопской дези популяции как смеси только Ближневосточных популяций. Во втором столбце представлены результаты моделирования Эфиопской дези популяции как смеси с участием Индийской дези популяции. Вертикальной скобкой отмечены модели всех попарных вкладов Ближневосточных популяций. Серым цветом обозначен вклад собственной накопленной дисперсии Эфиопской популяции, независимый от вкладов других рассматриваемых популяций. Эфиопская дези популяция как смесь, в которой участвуют следующие дези популяции: **(А)** все Ближневосточные; **(Б)** Черноморская и Турецкая; **(В)** Черноморская и Ливанская; **(Г)** Ливанская и Турецкая; **(Д)** Индийская и восточно-Среднеазиатская; **(Е)** Индийская и Черноморская; **(Ж)** Индийская и Ливанская; **(З)** Индийская и Турецкая.

для каждого хромосомного участка. Размер окна был взят равным 300 тыс. пар оснований, что соответствует значению LD для этой рассматриваемой панели образцов нута [5]. Размер шага в технике скользящего окна был взят равным 100 тыс. пар оснований. Окна с количеством ОНП менее 10 были исключены из рассмотрения. Далее в тексте дези популяции будут обозначаться по первым буквам названий (например, И - Индийская).

Для Эфиопской дези популяции были сопоставлены хромосомные профили смешения для двух гипотез происхождения этой популяции: [И + Ч] (Рисунок 4.11 А) и [И + Л] (Рисунок 4.11 Б). Было обнаружено, что некоторые хромосомные регионы происходят либо только от Черноморской популяции (Рисунок 4.11 А, указатель 1, либо только от Ливанской популяции (Рисунок 4.11 Б, указатель 2). Тем не менее, большинство регионов, приписываемых Черноморской популяции в гипотезе [И + Ч], классифицируются как вклад Ливанской при оценке вкладов под гипотезой [И + Л] (Рисунок 4.11 указатель 3). То есть, эти регионы не связаны с какой-то конкретной популяцией, а скорее относятся в целом к ветви Ближнего Востока. Сравнение хромосомных профилей смешения для гипотез [И + Ч], [И + Л] и [Ч + Л + Т] показывает, что некоторые блоки в Эфиопской популяции имеют исключительно Индийское происхождение (Рисунок 4.11Б, Г, указатель 4). При рассмотрении гипотезы [Ч + Л + Т], то есть без участия Индийской популяции, эти блоки почти полностью характеризуются как собственная дисперсия Эфиопской популяции, не связанная с предполагаемыми популяциями-источниками (Рисунок 4.11 Г, указатель 4).

Вторая дези популяция с предполагаемым смешанным происхождением - западно-Среднеазиатская популяция. Эта популяция демонстрирует отдельную группу смешения (Рисунок 4.2) и отдельную длинную ветвь в иерархической кластеризации (Рисунок 4.5Б). С помощью *miGAdmi* эта популяция была протестирована как смешанная из двух или трех исходных несмешанных популяций дези. Среди азиатских популяций восточно-Среднеазиатская популяция дези показала доминирующий эффект, что выглядит логичным, исходя из географической близости Среднеазиатских популяций. Среди популяций Ближнего Востока доминирующий вклад показала популяция Ливанских дези (Рисунок 4.12 А-В,Е). Наименьшее значение ВИС среди всех гипотез соответствует происхождению от восточно-Среднеазиатской и Ливанской популяций (Рисунок 4.12Г, отмечено темно-синим кругом), а второе по величине значение ВИС снизу указывает на

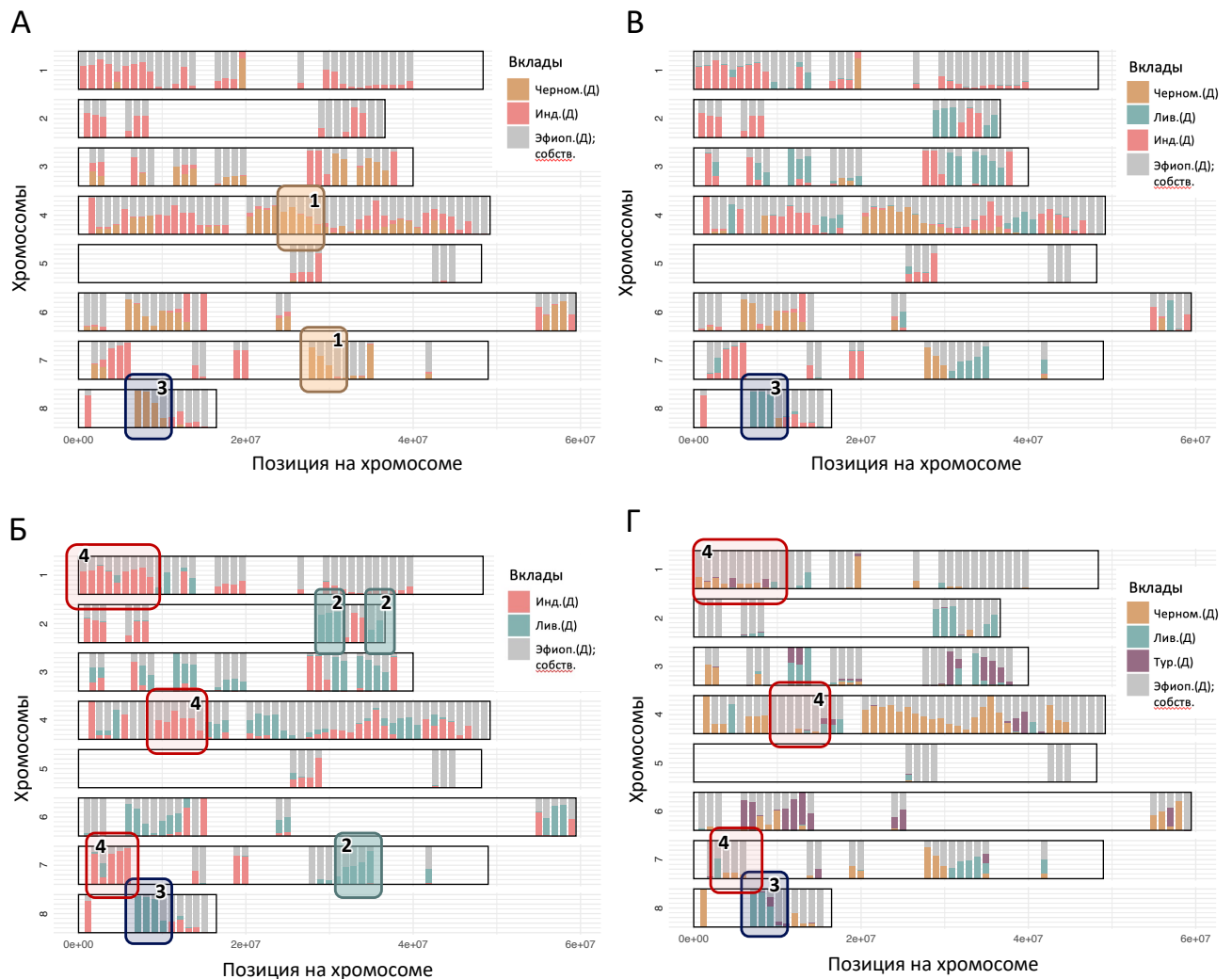


Рисунок 4.11 — Распределение вдоль хромосом вкладов различных популяций в Эфиопскую. Цифрами обозначены локусы для особого внимания. Эфиопская дези популяция рассматривается как смесь, в которой участвуют следующие дези популяции: **(А)** Индийская и Черноморская; **(Б)** Индийская и Ливанская; **(В)** Индийская, Черноморская и Ливанская; **(Г)** все Ближневосточные.

смешение восточно-Среднеазиатской и Турецкой популяций (Рисунок 4.12 Д, отмечено светло-синим кругом).

Сравнение хромосомных профилей смешения для гипотез [вС+Т] и [вС+Л] показало, что все блоки с происхождением из Турецкой популяции совпадают с блоками, предполагаемыми как происходящие из Ливанской популяции, однако есть ряд блоков с происхождением только из Ливанской популяции (Рисунки 4.12 Ж, З, третья хромосома). Несмотря на то, что Ливанская и Турецкая популяции находятся в одной кладе исходного дерева несмешанных популяций, вклад Ливанской популяции в западно-Среднеазиатскую все же больше не только на уровне ВИС критерия, но и на хромосомном уровне. Таким образом, можно заключить, что западно-Среднеазиатская популяция является результатом смешения восточно-Среднеазиатской и Ливанской популяций.

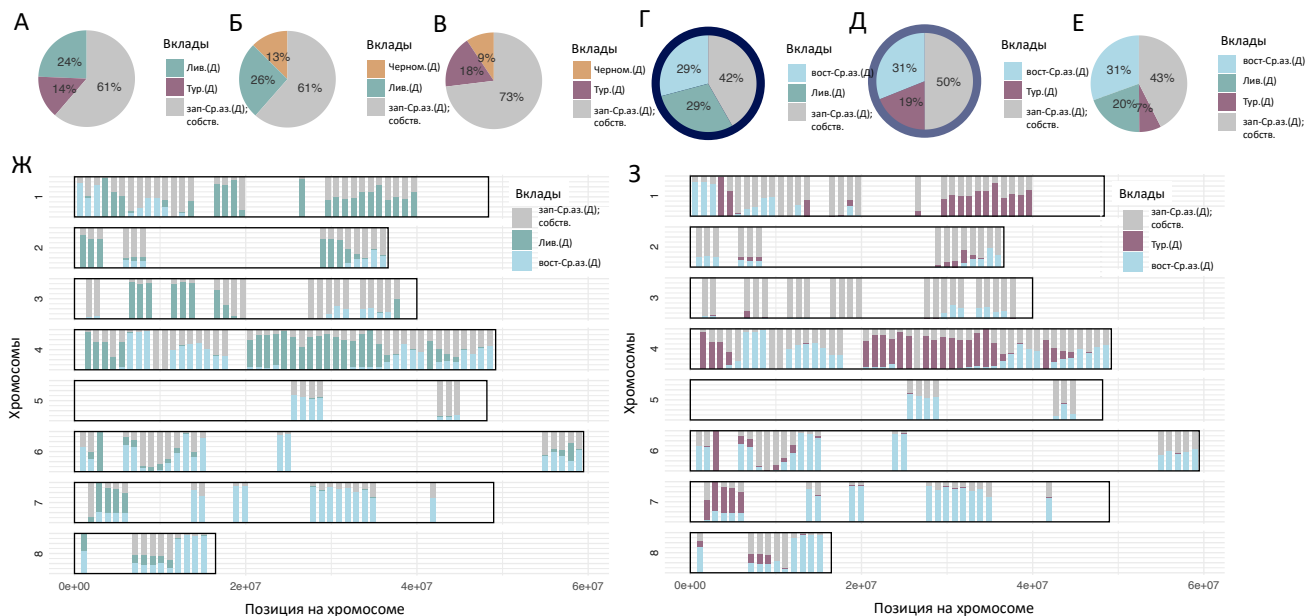


Рисунок 4.12 — Результаты моделирования западно-Среднеазиатской деми популяции как смешанной. Западно-Среднеазиатской деми популяция рассматривается как смешанная, в которой участвуют следующие деми популяции: **(А)** Ливанская и Турецкая; **(Б)** Черноморская и Ливанская; **(В)** Черноморская и Турецкая; **(Г)** восточно-Среднеазиатской и Ливанская; **(Д)** восточно-Среднеазиатской и Турецкая; **(Е)** восточно-Среднеазиатской, Ливанская и Турецкая. Распределение вдоль хромосом вкладов: **(Ж)** Ливанской и восточной-Среднеазиатской деми популяций; **(З)** Турецкой и восточной-Среднеазиатской деми популяций.

4.3.3 Реконструкция эволюционной истории популяций нута кабули

На основе лингвистических данных можно предположить, что кабули возникли в Средней Азии и названы в честь города Кабул (территория современного Афганистане). С другой стороны, на Ближнем Востоке кабули встречаются чаще, поэтому можно предположить, что появление кабули в Средней Азии - это относительно более новое явление [11]. Хотя у групп дези и кабули много общего, селекционные программы, как правило, держат их отдельно, вероятно, из-за различий в адаптивных требованиях и потребительских предпочтениях [85; 94; 95]

На основании этих противоречивых предположений были построены 4 гипотезы смешения. Три из них предполагали происхождение нута кабули из Ближневосточных дези популяций: Турецкой, Черноморской или Ливанской (Рисунок 4.13 А,В,Г) Четвертая гипотеза отражала его происхождение из Средней Азии с последующим внедрением на Ближний Восток (Рисунок 4.13Б). Значение ВИС для четвертой модели было наибольшим, что говорит о ее наименьшей предпочтительности (Рисунок 4.13Б). Модель с наименьшим значением ВИС соответствовала происхождению кабули от Турецкой популяции, следующее по величине значение ВИС снизу — происхождение от Ливанской популяции, третье — происхождение от Черноморской популяции. Кроме того, анализ главных компонент характеристических частот аллелей всех рассматриваемых популяций (Рисунок 4.14) показал смещение всех популяций кабули вдоль оси первой главной компоненты. Важно отметить, что это смещение направлено не в сторону восточно-Среднеазиатской дези популяции, как можно было бы ожидать, а в сторону Турецкой дези популяции. То есть модель происхождения кабули из Турецкого региона является предпочтительной по сравнению с происхождением кабули из восточно-Среднеазиатских дези.

Наиболее выраженное фенотипическое различие между группами сортов дези и кабули - это окраска семян. У бобовых растений этот признак является менделевским и контролируется так называемым геном А [96]. Для гороха (*Pisum sativum*) и люцерны (*Medicago truncatula*) последовательности этого гена можно найти в базе данных GenBank по номерам: GU132940 (MtbHLH) и GU132941 (PsbHLH). Чтобы найти регион с гомологичными генами в геноме *Cicer arietinum*, был проведен BLAST-поиск [97], и было найдено соответствие с геном LOC101506726 (позиции 2149255-2158629bp, начало 4-й хромосомы). Был

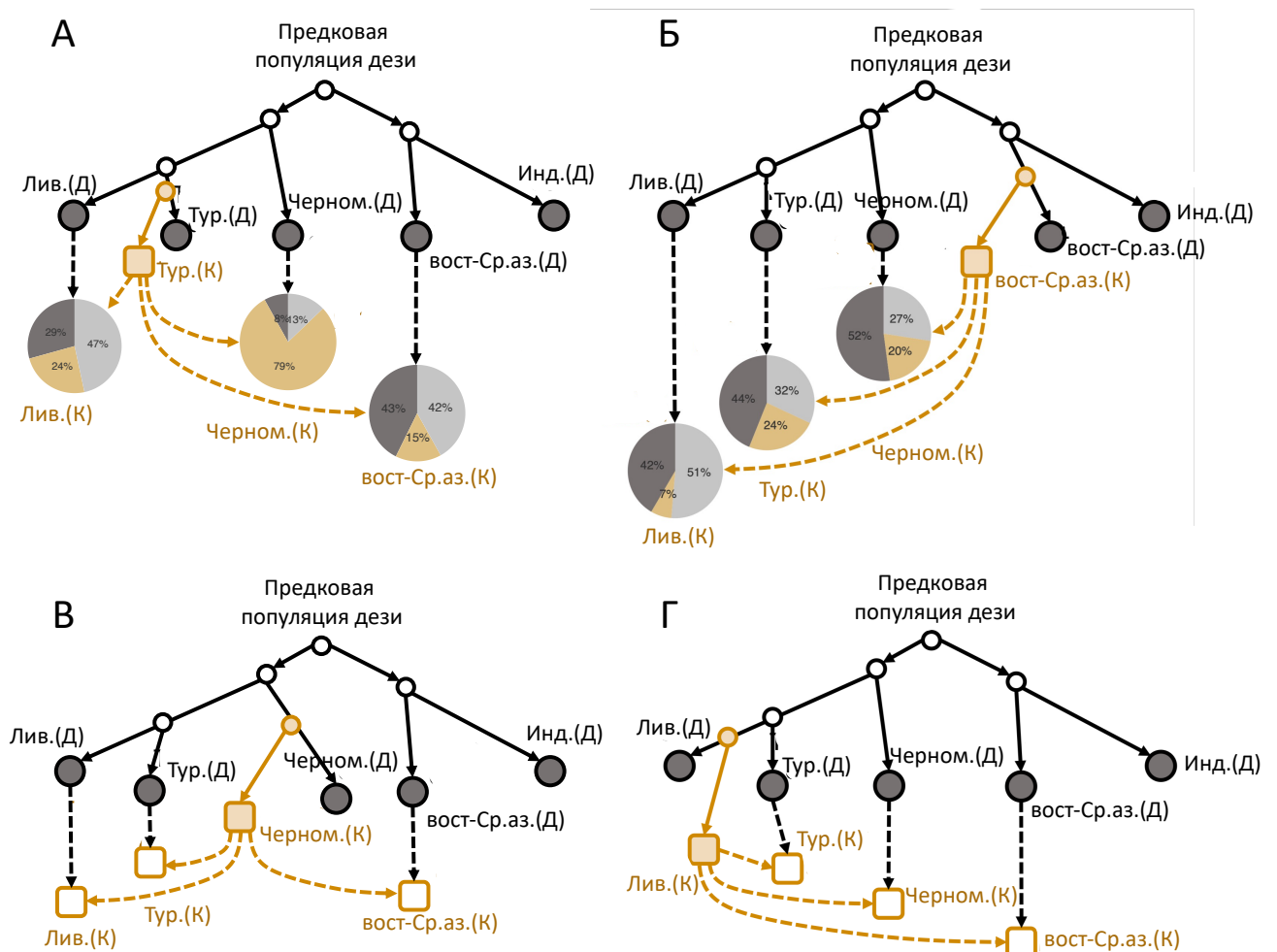


Рисунок 4.13 — **Четыре альтернативные гипотезы происхождения популяций кабули.** Серыми закрашенными кружками обозначены наблюдаемые популяции, серыми незакрашенными — предковые состояния. Желтым закрашенным квадратом показана кабули популяция, впервые возникшая при доместикации. Желтыми незакрашенными квадратами — популяции кабули, образовавшиеся в результате смешения местных дези и первых возникших кабули. Модели предполагают, что доместикация кабули произошла от: **(А)** Турецких дези. **(Б)** Восточно-Среднеазиатских дези. **(В)** Черноморских дези. **(Г)** Ливанских дези.

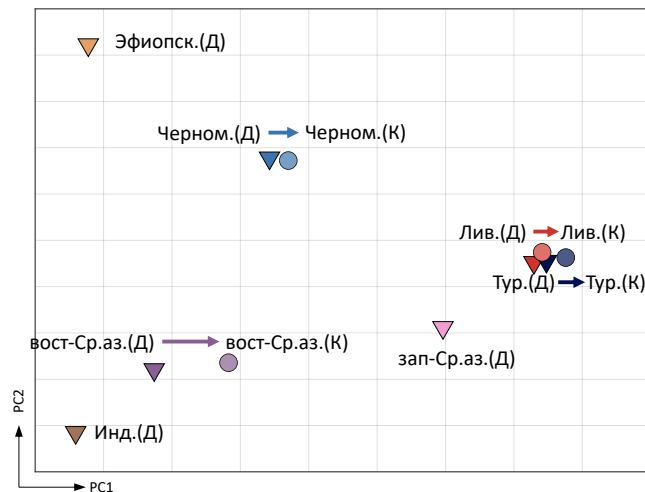


Рисунок 4.14 — **Представление характеристических частот аллелей в осях первых двух главных компонент.** Стрелками обозначены направления от популяции дези к популяции кабули в каждом регионе.

проведен полногеномный поиск ассоциаций для признака ”группа сортов” (дези или кабули в зависимости от окраски семян) с использованием метода rrBLUP. Был найден один значимый ОНП, который расположен на 4173 нуклеотида выше по хромосоме от гена LOC101506726 (Рисунок 4.15 А). Анализ хромосомных профилей смещений в популяциях Ливанских и Черноморских кабули в конкурирующих гипотезах (Рисунки 4.15 Б,В) показал, что этот регион унаследован от Турецких кабули, а не от восточно-Среднеазиатских кабули (отмечено треугольником на Рисунках 4.15 Б,В).

4.4 Сравнение результатов программы migadmi с другими методами

Хотя существующие методы реконструкции истории популяций не полностью соответствуют подходу, используемому в методе migadmi (см. Таблицу 2), и входные данные для migadmi учитывали географическое расположение образцов, чего не предполагали другие программы, сравнительный анализ все же был выполнен и результаты представлены ниже.

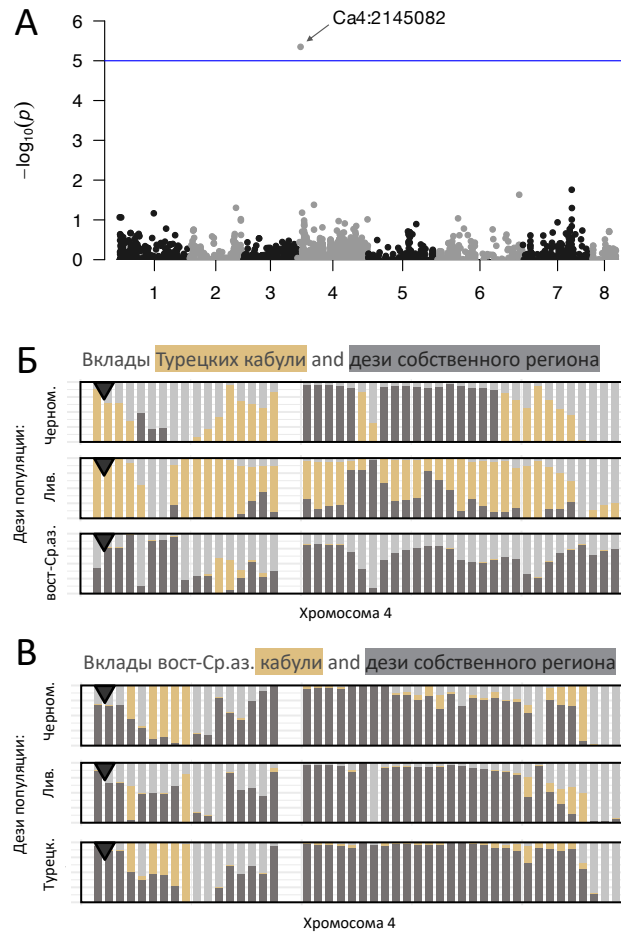


Рисунок 4.15 — **Хромосомный локус, ассоциированный с признаком принадлежности к дези/кабули.** (А) Манхэттенский график результата полногеномного поиска ассоциаций с признаком принадлежности к дези/кабули. Каждая точка соответствует одному генетическому маркеру ОНП. Положение по оси X соответствует порядковому номеру маркера в упорядоченном по позиции и по хромосомам списке всех маркеров. Оттенки серого выбраны для визуального разделения маркеров, соответствующих разным хромосомам. Номера хромосом подписаны по оси X. Синяя горизонтальная линия показывает порог статистической значимости ассоциации. Единственный маркер, статистически значимо ассоциированный с исследуемым признаком, обозначен стрелочкой. (Б) Распределение вдоль хромосом вкладов популяций Турецких кабули и дези популяций из соответствующего региона. Треугольником обозначено положение маркера, статистически значимо ассоциированного с признаком дези/кабули. (В) Распределение вдоль хромосом вкладов популяций восточно-Среднеазиатских кабули и дези популяций из соответствующего региона. Треугольником обозначено положение маркера, статистически значимо ассоциированного с признаком дези/кабули.

4.4.1 Результаты программы TreeMix

С целью выяснения происхождения Эфиопской и западно-Среднеазиатской популяций дези, программа TreeMix была запущена на семи дези-популяциях. При запуске TreeMix внешняя популяция для укоренения дерева не указывалась, параметр бутстрепа был задан равным 35, что соответствует среднему числу ОНП в методе скользящего окна для migadmi. Были рассмотрены варианты запуска программы с количеством событий смешений, равным 1, 2, 3 и 4 (см. Рисунок 4.16). Из результатов запусков программы становится ясно, что среди семи дези-популяций западно-Среднеазиатская и Эфиопская дези-популяции являются двумя наиболее вероятно смешанными популяциями. Это результаты соответствуют результатам, полученным в Параграфе 4.1, однако исходные популяций для смесей не совсем совпадают с результатами migadmi. По результатам работы TreeMix Эфиопская популяция дези была определена как смесь Индийской и Турецкой популяций дези (в отличие от Индийской и Черноморской, как было в migadmi), а западно-Среднеазиатская популяция дези - как смесь Индийской и Турецкой популяций дези (а не восточно-Среднеазиатской и Ливанской, как в migadmi). Несмотря на несовпадения, можно заключить, что на уровне глобальных паттернов между результатами migadmi и TreeMix наблюдается согласованность: каждая из смешанных популяций имеет в качестве исходных одну популяцию из Ближневосточного региона и одну из Азиатского. Однако результаты migadmi выглядят более логичными с учетом географической близости западно-Среднеазиатской и восточно-Среднеазиатской популяций.

4.4.2 Результаты программы MixMapper

Происхождение Эфиопской и западно-Среднеазиатской популяций дези было также исследовано с помощью программы MixMapper. MixMapper принимает в качестве входных данных несмешанные популяции, затем создает дерево из них и проверяет смешанную популяцию, добавляя ее к дереву. Для оценки происхождения Эфиопской популяции дези программа принимала в качестве несмешанных 5 дези-популяций: Турецкую, Ливанскую, Черноморскую, Индий-

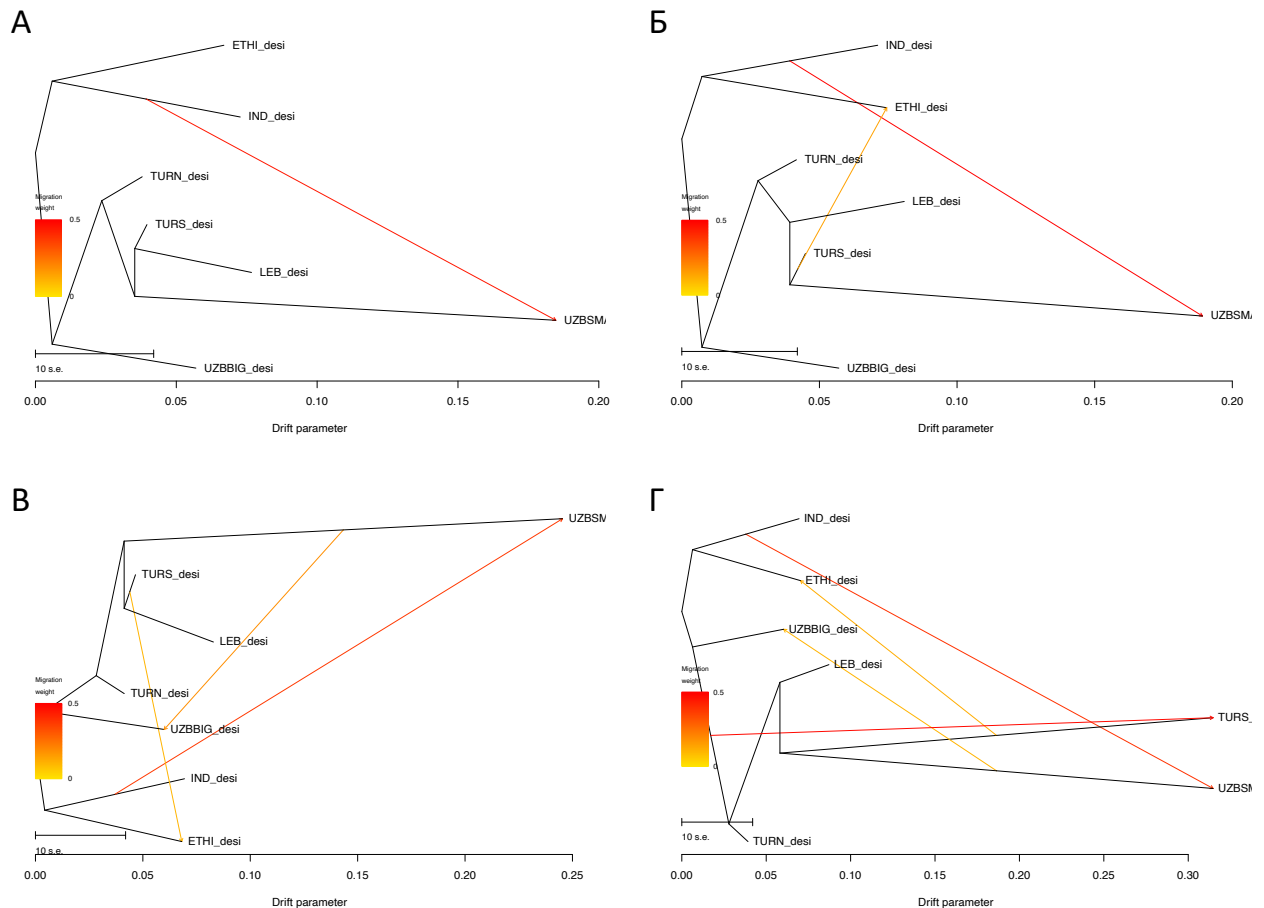


Рисунок 4.16 — Результаты работы программы TreeMix. Параметр количества событий смешений был от 1 до 4, результаты показаны на подрисунках (А-Г), соответственно.

скую и восточно-Среднеазиатскую. MixMapper выявил два источника Эфиопской дези: Черноморскую дези (74%) и Индийскую дези (26%), что качественно соответствует результатам migadmi. Тестирование западно-Среднеазиатской дези популяций как смешанной тех же пяти несмешанных показало, что она представляет собой смесь Ливанской дези (68%) и восточно-Среднеазиатской дези популяций (32%), что опять-таки качественно совпадает с результатами migadmi.

Для проверки происхождения нута кабули были протестированы две контрастные гипотезы с помощью MixMapper, сравнение которых происходило на основании величины коэффициентов смешения. В первой гипотезе ("кабули из Турецкого региона") предполагалось, что шесть популяций — Турецкие, Ливанские, Черноморские, Индийские, восточно-Среднеазиатские дези и Турецкие кабули — являются несмешанными, в то время как восточно-Среднеазиатские, Ливанские и Черноморские кабули представляют собой смеси. При запуске программы MixMapper было установлено, что восточно-Среднеазиатские кабули

состоят на 69% из восточно-Среднеазиатских дези и на 31% из генетического пула клады, включающей Турецкие кабули, Турецкие дези и Ливанские дези. Аналогично, с использованием тех же шести исходных популяций было определено, что Ливанские кабули на 61% происходят от Ливанских дези и на 39% — от Турецких кабули, а Черноморские кабули — на 90% от Черноморских дези и на 10% от Турецких кабули. Таким образом, если считать Турецкие кабули несмешанной популяцией, то остальные кабули популяции частично происходят из неё.

Во второй гипотезе (”кабули из Среднеазиатского региона”) предполагалось, что шесть популяций — Турецкие, Ливанские, Черноморские, Индийские, восточно-Среднеазиатские дези и восточно-Среднеазиатские кабули — являются несмешанными, в то время как Турецкие, Ливанские и Черноморские кабули представляют собой смеси. Результаты показали, что: (i) Турецкие и Ливанские кабули являются смесями Турецких и Ливанских дези, а (ii) Черноморские кабули — смесью Турецких и Черноморских дези. Это означает, что ни одна из предполагаемых смешанных популяций не имеет общего с восточно-Среднеазиатскими популяциями. Таким образом, можно заключить, что происхождение кабули, вероятно, связано с регионом Турции, а не со Среднеазиатским регионом. Полученные результаты полностью соответствуют качественным результатам анализа *migadmi*.

4.4.3 Результаты программы *admixtools*

Для сравнения *migadmi* с *admixtools* были запущены два инструмента: *qrgraph* (для оценки заданного графа смешений) и *find_graph* (для поиска оптимального графа смешений в пространстве топологий всех графов смешений). Сначала для каждой топологии графа смешений, созданного *migadmi*, была проведена оценка с помощью *qrgraph*. Затем был запущен *find_graph*, который использовал в качестве начальных значений либо графы, полученные с помощью *migadmi*, либо случайные начальные значения.

Сначала был рассмотрен граф смешений для пяти несмешанных популяций дези с одним событием смешения для Эфиопской дези популяции (Рисунок 4.17A). Оценка этого графа с помощью *qrgraph* совпала с оценкой оптимальной топологии, полученной с помощью *find_graph*, когда исходный граф был

использован в качестве начальной точки для поиска (Рисунок 4.17Б). Однако два случайных запуска `find_graph` привели к созданию топологий с менее оптимальными оценками (чем ниже оценка, тем лучше) и сомнительными структурами: в одном случае Индийская популяция, а в другом Черноморская популяция были предсказаны как смешанные (Рисунок 4.17В, Г).

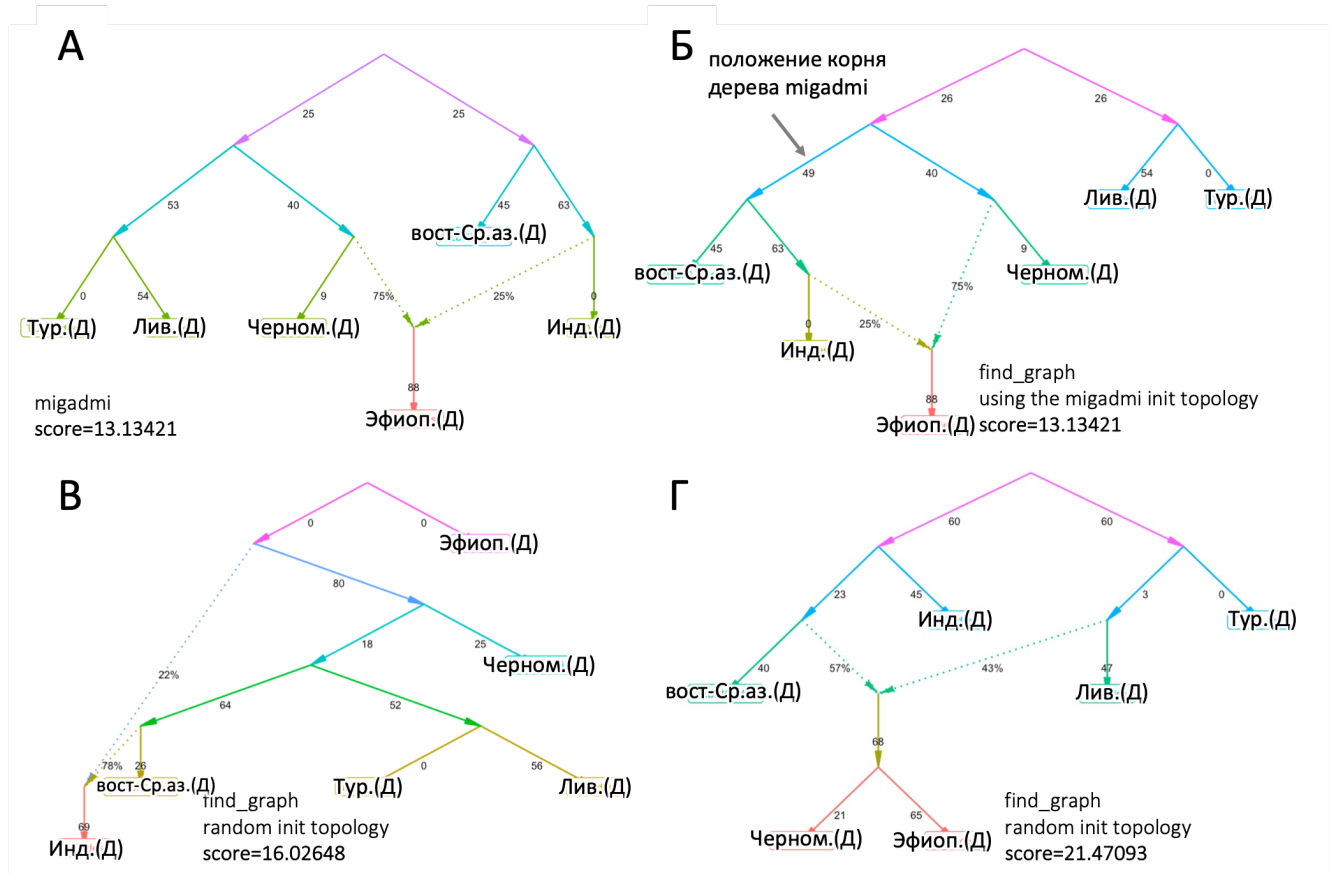


Рисунок 4.17 — Топологии графов смешений для шести популяций дези и одного события смешения. Рядом с каждым графом представлено значение score, оценка графа, выдаваемая инструментами `admixtools`. **(А)** Топология графа смешений, полученная из анализа `migadmi`. **(Б)** Оптимальный граф смешений, полученный с помощью инструмента `find_graph`, начальной топологией для которого был граф с рисунка А. **(С)** Оптимальный граф смешений, полученный с помощью инструмента `find_graph`, начальная топология была выбрана случайно (начальное значение генератора случайных чисел равнялось 30). **(Д)** Оптимальный граф смешений, полученный с помощью инструмента `find_graph`, начальная топология была выбрана случайно (начальное значение генератора случайных чисел равнялось 239).

Затем та же методика была применена к расширенному графу, который включал западно-Среднеазиатские дези популяции как второе событие смешения. В этом случае топология на рисунке (Рисунок 4.18А) была взята за исходную. По результатам запусков `find_graph` из случайных начальных топологий, топология графа `migadmi` не оказалась оптимальной. Однако более оптимальные топологии, полученные `find_graph` (Рисунок 4.18В,С) предполагали, что западно-Среднеазиатская популяция является смесью Индийской и Ливанской популяций, что менее вероятно, чем заключение о том, что западно-Среднеазиатская популяция имеет в своем происхождении корни из соседней восточно-Среднеазиатской популяции. Кроме того, некоторые оптимальные топологии от `find_graph` были хуже топологии `migadmi` с точки зрения как оценки, так и логики (Рисунок 4.18D).

В вопросе происхождения кабули популяций не удалось напрямую сравнить топологии, построенные с использованием `migadmi` и `find_graph`, поскольку `migadmi` поддерживает работу с топологиями, которые недоступны для `find_graph`. Вместо прямого сравнения, был произведен запуск `find_graph` со случайных топологий для 5 популяций дези и 4 популяций кабули в режимах поиска одного, двух или трех смешений. По результатам можно заключить, что оптимальные графы довольно часто показывали восточно-Среднеазиатскую популяцию кабули как смешанную, что соответствует результатам, полученным с помощью `migadmi` (Рисунок 4.19).

4.5 Выводы к главе 4

Анализ популяционной структуры генотипированной панели из 421 образца нута позволил выделить четкие генетически однородные и географически связанные популяции (так называемые генетико-географические группы). Эти популяции не противоречат кластерам из популяционных анализов других коллекций нута, но оказались более детализированными, так как некоторые ранее определенные популяции были разделены на несколько в новой группировке. Так, в Средиземноморском регионе были определены такие популяции, как северная-Средиземноморская и южная-Средиземноморская, и предложены их происхождения на основании статистики F_{ST} . А образцы в Ближневосточном регионе были подразделены на Турецкую, Ливанскую и Черноморскую группы.

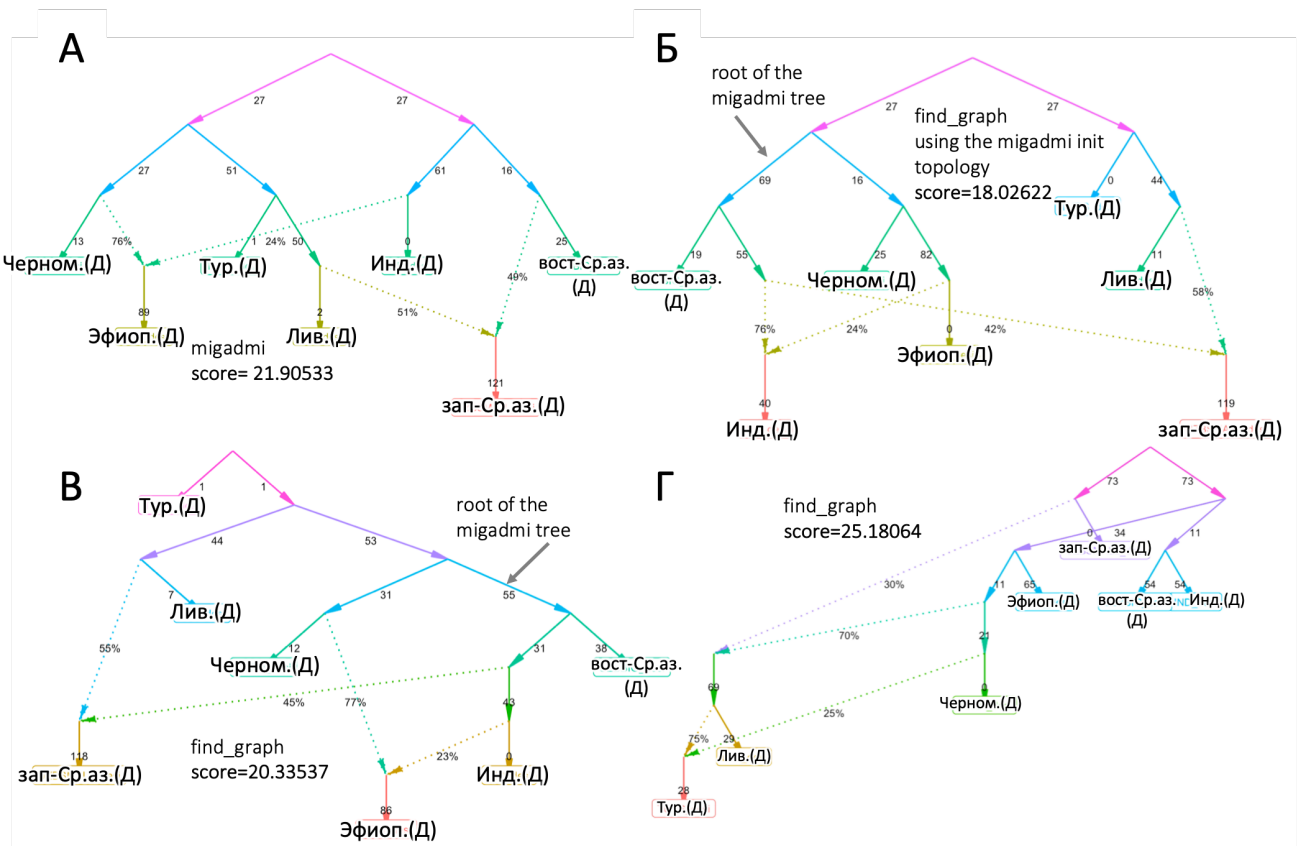


Рисунок 4.18 — Топологии графов смешений для шести популяций дези и двух событий смешения. Рядом с каждым графом представлено значение score, оценка графа, выдаваемая инструментами admixtools. (А) Топология графа смешений, полученная из анализа migadmi. (Б) Оптимальный граф смешений, полученный с помощью инструмента find_graph, начальной топологией для которого был граф с подрисунка А. (С) Оптимальный граф смешений, полученный с помощью инструмента find_graph, начальная топология была выбрана случайно (начальное значение генератора случайных чисел равнялось 30). (Д) Оптимальный граф смешений, полученный с помощью инструмента find_graph, начальная топология была выбрана случайно (начальное значение генератора случайных чисел равнялось 239).

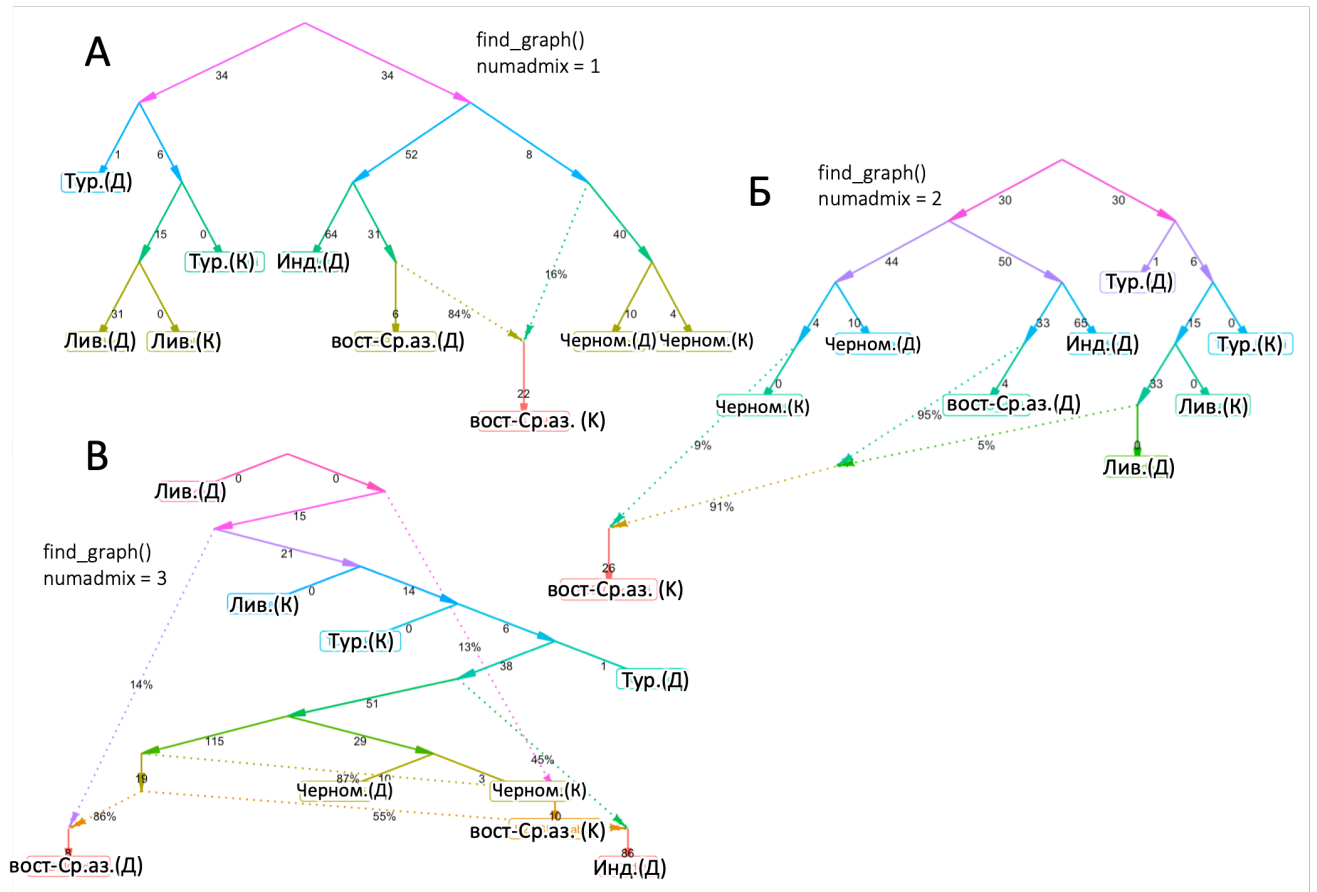


Рисунок 4.19 — Топологии графов смешений для пяти популяций дези и четырех популяций кабули. Рядом с каждым графом представлено значение score, оценка графа, выдаваемая инструментами admixtools. (А) Оптимальный граф смешений, полученный с помощью find_graph с параметром количества событий смешений, равным одному. (Б) Оптимальный граф смешений, полученный с помощью find_graph с параметром количества событий смешений, равным двум. (В) Оптимальный граф смешений, полученный с помощью find_graph с параметром количества событий смешений, равным трем.

Кроме того, две популяции - Эфиопская и западно-Среднеазиатская дези популяции - были идентифицированы как потенциально смешанные, что поставило дополнительные вопросы, на которые следует ответить через реконструкцию эволюционной истории популяций нута.

Две разработанные модели популяционной генетики — *popdisp* и *migadmi* — были протестированы сначала на симулированных данных, а затем на реальных данных генотипированной панели из 421 образца нута. В результате были получены важные исторические данные о путях передвижения нута между географическими регионами, а также основание для понимания путей его перемещения внутри регионов.

Программа *popdisp* оценивает характеристические частоты аллелей в популяции, используя генетическую информацию образцов популяции и матрицу ковариаций между ними, рассчитанную, например, на основе географической близости точек сбора образцов. По результатам работы на симулированных данных программа *popdisp* продемонстрировала хорошую устойчивость к выбросам, зашумлению и искажениям в данных. Такая устойчивость *popdisp* достигается за счет связности всех образцов в единую структуру с использованием матрицы ковариаций. Важно подчеркнуть, что матрица ковариаций должна в действительности отражать распределение аллелей в популяции, а не быть случайной. Для проверки чувствительности программы *popdisp* к матрицам ковариации для каждой симуляции проводились запуски с разными матрицами, чтобы определить, какая из них дает наиболее достоверные оценки. Оказалось, что лучшие результаты достигаются при использовании исходной матрицы ковариации, использованной в симуляциях.

Способность *popdisp* различать матрицы ковариации побудила провести следующий эксперимент на данных по нуту. Для каждой популяции нута были построены две ковариационные матрицы: одна построена на основе линейных расстояний между точками сбора образцов в регионе, а другая — на основе расстояний согласно функции перемещения Герцог. Эта функция учитывает ландшафт и сложность перемещения по нему. Оказалось, что матрицы расстояний, построенные на оптимальных географических расстояниях, соответствуют более высокому значению функции правдоподобия *popdisp*. Таким образом, генетическое родство образцов в регионах связано не с линейными расстояниями, как в если бы это было в силу случайного дрейфа, а с реальным перемещением по ландшафту, что, вероятно, отражает то, что нут распространялся вместе с человеком. В

способности `popdisp` сравнивать географические пути перемещения видится еще одно применение пакета `popdisp`, кроме оценки характеристических частот аллелей.

Программа `migadmi` реконструирует эволюционную историю популяций с возможностью учета сложных событий смешения. По результатам тестирования она показала свою точность, а на реальных данных – доминирующую согласованность с другими методами (`TreeMix`, `MixMapper` и `admixtools`). То есть результаты, представленные программой `migadmi`, согласуются с результатами большинства других программ по всем поставленным в исследовании вопросам. С одной стороны, это свидетельствует о правильности полученных результатов, с другой стороны, это, скорее всего, указывает на более высокую точность параметров смешений, что имеет большое значение при анализе, например, селекции с помощью программы `Gross` [98].

По результатам применения к реальным данным по нуту, программа `migadmi` раскрыла происхождение двух дези популяций, тем самым указывая на потенциальные исторические торговые пути. Можно заключить, что существовали три основных торговых пути в Эфиопию. Два из них были морскими: один шел из Индии, а другой, по-видимому, проходил из Черноморского региона, вероятно через Аравийский полуостров, минуя Ливан. Из Ливана, вероятно, существовал отдельный сухопутный торговый путь в Эфиопию.

Другой пример популяции со смешанным происхождением — это дези популяция из западно-Среднеазиатского региона. По результатам, полученным с помощью программы `migadmi`, выяснилось, что эта популяция образовалась в результате смешения Ливанской дези популяции и восточно-Среднеазиатской дези популяции. Это смешение можно описать как результат встречного движения в одних и тех же широтах. Стоит отметить, что формирование западно-Среднеазиатской популяции из восточно-Среднеазиатской популяции является как бы движением в обратную сторону относительно первоначального направления при формировании западно-Среднеазиатской популяции.

Другим вопросом является происхождение нута кабули. С лингвистической точки зрения, можно предположить, что кабули происходит от названия города Кабул, что говорит о возможном происхождении из восточно-Среднеазиатской популяции, однако археологические доказательства этого не были обнаружены. По генетическим данным, полученным с помощью программы `migadmi`, dome-

стикация нута кабули из местных дези произошла в Турецком регионе, а затем он распространился на другие регионы, включая восточно-Среднеазиатский регион.

В отличие от программ TreeMix и MixMapper, модель migadmi напрямую учитывает собственную дисперсию смешанной популяции после события смешения. Из результатов работы migadmi можно сделать вывод, что в некоторых хромосомных локусах у определенных популяций наблюдается значительный вклад именно этой компоненты. То есть, популяции, предполагаемые как исходные для рассматриваемой смеси, не оказывают значительного влияния. Это может свидетельствовать о том, что в таком локусе генома, возможно, происходила интенсивная селекция, уменьшающая вариабельность от исходных популяций, или в смешенной популяции присутствует еще какой-то неизвестный в рамках текущего анализа источник. Такие локусы могут представлять особый интерес в последующем анализе.

Программа migadmi не проводит анализ паттернов вкладов исходных популяций вдоль хромосомы и, в частности, не тестирует природу паттерна: является ли он результатом смешения популяций или неполной сортировки линий (англ. incomplete lineage sorting). В случае смешения мы ожидаем увидеть мозаику из блоков, где одни блоки более близко связаны с одной исходной популяцией, чем с другой. В отличие от этого, неполная сортировка линий приводит к более переменному паттерну смешения без четких блоков или кластеров. Поэтому, чтобы точно различить между неполной сортировкой линий и смешением с использованием migadmi, возможно, потребуется дополнительный анализ распределения длин хромосомных блоков с похожими вкладами исходных популяций. Однако, основываясь на визуальном анализе полученных для рассматриваемых популяций нута паттернов, скорее всего, речь идет о смешении, а не о неполной сортировке линий.

Заключение

Реконструкция эволюционной истории популяций на основе генетических данных представляет собой задачу, богатую нюансами, которые еще не в полной мере учтены в существующих решениях. В настоящей диссертации был предложен двухступенчатый подход к решению этой задачи, учитывающий целый ряд важных факторов. Первый этап заключается в применении новой разработанной программы *popdisp*, которая учитывает географический сигнал при оценке характеристических частот аллелей популяций и способна сравнивать различные гипотезы о перемещениях популяции в ареале. Второй этап заключается в применении новой разработанной программы *migadmi*, которая учитывает события смешения сложной структуры, предлагая новый подход к оценке вкладов популяций-источников в смешанную популяцию, а также демонстрируя распределение этого сигнала вдоль хромосом. В качестве основы для моделирования случайного дрейфа в обеих программах была разработана и использована композиционная аппроксимация модели Райта-Фишера.

Предложенный подход был применен для реконструкции истории популяций нута и позволил ответить на интригующие вопросы о происхождении некоторых из них. В частности, было показано смешанное происхождение Эфиопской и восточно-Среднеазиатской популяций нута дези. Эти результаты позволяют проследить древние торговые пути, в частности, морской путь из Индии в Эфиопию через Аравийское море. Кроме того, настоящая работа разрешает вопрос о происхождении нута кабули. Вопреки лингвистически обоснованной гипотезе, кабули произошли не из Кабула, а из Турецкого региона, что подтверждается не только результатами программы *migadmi*, но и полногеномным анализом ассоциаций.

Выводы

Выводы по разработке вычислительных средств:

1. Предложенная композиционная аппроксимация модели Райта-Фишера может быть эффективно использована в моделях популяционной генетики.
2. Разработанный метод `popdisp` может быть использован для устойчивой оценки характеристических частот аллелей в популяциях с заданной ковариационной матрицей между образцами.
3. Разработанный метод `migadmi` может быть использован для детальной реконструкции сложных эволюционных истории популяций с множественными и вложенные события смешения.

Выводы, полученные по применению разработанных вычислительных средств:

1. Генетические сходства между образцами в 9 из 11 популяций нута соответствуют длинам оптимальных путей перемещения по ландшафту больше, чем линейным расстояниям между образцами.
2. Северная и южная Средиземноморские популяции нута сформировались в результате продвижения на запад по северному и южному берегам Средиземного моря Черноморской и Турецкой популяций соответственно.
3. Эфиопская популяция нута дези сформировалась в результате смешения Черноморской, Индийской и Ливанской популяций.
4. Западная-Среднеазиатская популяция нута дези возникла в результате смешения Ливанской и восточной-Среднеазиатской популяций.
5. Нут кабули произошел из региона современной Турции, а не из азиатского региона.

Благодарности

Автор выражает благодарность Андронову Евгению Евгеньевичу за огромную поддержку в научной деятельности и за руководство в проведении ряда исследований. Самсоновой Марии Георгиевне – за предоставленную возможность работать с уникальными данными и исследовать интересные биологические вопросы. Маргарите Афанасьевне Вишняковой, Сергею Нуждину и Эрику фон Веттбергу – за консультации по эволюционной истории нута; Трэвису Лонгкору и Нине Нуждиной – за обсуждение способов построения географических путей по ландшафту. Надежде Иголкиной – за психологическую поддержку и мотивацию при работе с текстом диссертации.

Список литературы

1. *Meyer, R. S.* Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops / R. S. Meyer, A. E. DuVal, H. R. Jensen // *New Phytologist*. — 2012. — Oct. — Vol. 196, no. 1. — P. 29—48.
2. *Pingali, P. L.* The Green Revolution and Crop Biodiversity / P. L. Pingali // *Biological Extinction: New Perspectives* / ed. by P. Dasgupta, P. Raven, A. McIvor. — Cambridge University Press, 2019. — P. 175—192.
3. *Abbo, S.* Viewpoint: Evolution of cultivated chickpea: four bottlenecks limit diversity and constrain adaptation / S. Abbo, J. Berger, N. C. Turner // *Functional Plant Biology*. — 2003. — Vol. 30, issue 10. — P. 1081.
4. A comprehensive review on grain legumes as climate-smart crops: Challenges and prospects / A. Dutta [et al.] // *Environmental Challenges*. — 2022. — Vol. 7. — P. 100479.
5. Genomic Analysis of Vavilov's Historic Chickpea Landraces Reveals Footprints of Environmental and Human Selection / A. Sokolkova [et al.] // *International Journal of Molecular Sciences*. — 2020. — May. — Vol. 21, no. 11. — P. 3952.
6. *Lewanski, A. L.* The era of the ARG: An introduction to ancestral recombination graphs and their significance in empirical evolutionary genomics / A. L. Lewanski, M. C. Grudler, G. S. Bradburd // *PLOS Genetics*. — 2024. — Jan. — Vol. 20, issue 1. — e1011110.
7. *Aitchison, J.* The Statistical Analysis of Compositional Data / J. Aitchison // *Journal of the Royal Statistical Society. Series B (Methodological)*. — 1982. — Vol. 44, no. 2. — P. 139—177.
8. *Pawlowsky-Glahn, V.* / V. Pawlowsky-Glahn, J. Egozcue, R. Tolosana-Delgado. — 2007. — Girona: Universitat.
9. A chickpea genetic variation map based on the sequencing of 3,366 genomes / R. K. Varshney [et al.] // *Nature*. — 2021. — Nov. — Vol. 599, no. 7886. — P. 622—627.
10. *Redden, R. J.* History and origin of chickpea. / R. J. Redden, J. D. Berger // *Chickpea breeding and management*. — UK : CABI, 01/2007. — P. 1—13.

11. *Maesen, L. van der*. Origin, history and taxonomy of Chickpea / L. van der Maesen // — 1987.
12. *Alexander, D. H.* Fast model-based estimation of ancestry in unrelated individuals / D. H. Alexander, J. Novembre, K. Lange // *Genome Research*. — 2009. — Sept. — Vol. 19, no. 9. — P. 1655—1664.
13. *Pritchard, J. K.* Inference of Population Structure Using Multilocus Genotype Data / J. K. Pritchard, M. Stephens, P. Donnelly // *Genetics*. — 2000. — June. — Vol. 155, no. 2. — P. 945—959.
14. *Pickrell, J. K.* Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data / J. K. Pickrell, J. K. Pritchard // *PLoS Genetics*. — 2012. — Nov. — Vol. 8, no. 11. — e1002967.
15. Efficient Moment-Based Inference of Admixture Parameters and Sources of Gene Flow / M. Lipson [et al.] // *Molecular Biology and Evolution*. — 2013. — Aug. — Vol. 30, no. 8. — P. 1788—1802.
16. Ancient Admixture in Human History / N. Patterson [et al.] // *Genetics*. — 2012. — Nov. — Vol. 192, no. 3. — P. 1065—1093.
17. *Wright, S.* ISOLATION BY DISTANCE / S. Wright // *Genetics*. — 1943. — Mar. — Vol. 28, no. 2. — P. 114—138.
18. Correlation between Genetic and Geographic Structure in Europe / O. Lao [et al.] // *Current Biology*. — 2008. — Aug. — Vol. 18, no. 16. — P. 1241—1248.
19. Genes mirror geography within Europe / J. Novembre [et al.] // *Nature*. — 2008. — Nov. — Vol. 456, no. 7218. — P. 98—101.
20. *Wang, C.* A Quantitative Comparison of the Similarity between Genes and Geography in Worldwide Human Populations / C. Wang, S. Zöllner, N. A. Rosenberg // *PLoS Genetics*. — 2012. — Aug. — Vol. 8, no. 8. — e1002886.
21. *Bradburd, G. S.* Disentangling the effects of geographic and ecological isolation on genetic differentiation / G. S. Bradburd, P. L. Ralph, G. M. Coop // *Evolution*. — 2013. — Nov. — Vol. 67, no. 11. — P. 3258—3273.
22. *Maaten, L. van der*. Visualizing High-Dimensional Data Using t-SNE / L. van der Maaten, G. Hinton // *Journal of Machine Learning Research*. — 2008. — Vol. 9. — P. 2579—2605.

23. UMAP: Uniform Manifold Approximation and Projection / L. McInnes [et al.] // Journal of Open Source Software. — 2018. — Vol. 3, no. 29. — P. 861.
24. *Lawson, D. J.* A tutorial on how not to over-interpret STRUCTURE and ADMIXTURE bar plots / D. J. Lawson, L. van Dorp, D. Falush // Nature Communications. — 2018. — Aug. — Vol. 9, no. 1. — P. 3258.
25. *Raj, A.* fastSTRUCTURE: variational inference of population structure in large SNP data sets / A. Raj, M. Stephens, J. K. Pritchard // Genetics. — 2014. — Vol. 197, no. 2. — P. 573—589.
26. Inferring population structure in biobank-scale genomic data / A. Chiu [et al.] // Am J Hum Genet. — 2022. — Apr. — Vol. 109, no. 4. — P. 727—737.
27. Fast and Efficient Estimation of Individual Ancestry Coefficients / E. Frichot [et al.] // Genetics. — 2014. — Vol. 196, no. 4. — P. 973—983.
28. *Garcia-Erill, G.* Evaluation of model fit of inferred admixture proportions / G. Garcia-Erill, A. Albrechtsen // Molecular Ecology Resources. — 2020. — July. — Vol. 20, no. 4. — P. 936—949.
29. Spatial Inference of Admixture Proportions and Secondary Contact Zones / E. Durand [et al.] // Molecular Biology and Evolution. — 2009. — Vol. 26, no. 9. — P. 1963—1973.
30. TESS3: fast inference of spatial population structure and genome scans for selection / K. Caye [et al.] // Molecular ecology resources. — 2016. — Vol. 16, no. 2. — P. 540—548.
31. *Bradburd, G. S.* Inferring Continuous and Discrete Population Genetic Structure Across Space / G. S. Bradburd, G. M. Coop, P. L. Ralph // Genetics. — 2018. — Sept. — Vol. 210, no. 1. — P. 33—52.
32. *Petkova, D.* Visualizing spatial population structure with estimated effective migration surfaces / D. Petkova, J. Novembre, M. Stephens // Nat Genet. — 2016. — Vol. 48. — P. 94—100.
33. *Peter, B. M.* Genetic Landscapes Reveal How Human Genetic Diversity Aligns with Geography / B. M. Peter, D. Petkova, J. Novembre // Molecular Biology and Evolution. — 2020. — Vol. 37, issue 4. — P. 943—951.

34. *Tobler, W.* Three presentations on geographical analysis and modeling: Non-isotropic geographic modeling speculations on the geometry of geography global spatial analysis : Technical Report / W. Tobler ; National center for geographic information ; analysis. — 02/1993. — 93 (1).
35. *Herzog, I.* Theory and Practice of Cost Functions / I. Herzog // — 2013.
36. *Wright, S.* The genetical structure of populations / S. Wright // *Annals of Eugenics*. — 1949. — Jan. — Vol. 15, no. 1. — P. 323—354.
37. *Boca, S.* Mathematical properties of F_{st} between admixed populations and their parental source populations / S. Boca, N. Rosenberg // *Theoretical Population Biology*. — 2011. — Vol. 80, no. 3. — P. 208—216.
38. *Molloy, E. K.* Advancing admixture graph estimation via maximum likelihood network orientation / E. K. Molloy, A. Durvasula, S. Sankararaman // *Bioinformatics*. — 2021. — July. — Vol. 37, Supplement 1. — P. i142—i150.
39. *Peter, B.* Admixture, Population Structure, and F-Statistics / B. Peter // *Genetics*. — 2016. — Apr. — Vol. 202, no. 4. — P. 1485—1501.
40. *Peter, B. M.* A geometric relationship of F_2 , F_3 and F_4 -statistics with principal component analysis / B. M. Peter // *Philosophical Transactions of the Royal Society B: Biological Sciences*. — 2022. — June. — Vol. 377, no. 1852.
41. *Mughal, M. R.* Properties and unbiased estimation of F - and D -statistics in samples containing related and inbred individuals / M. R. Mughal, M. DeGiorgio // *Genetics*. — 2022. — Jan. — Vol. 220, no. 1.
42. *Lipson, M.* Applying f_4 -statistics and admixture graphs: Theory and examples / M. Lipson // *Molecular Ecology Resources*. — 2020. — Nov. — Vol. 20, no. 6. — P. 1658—1667.
43. *Muirhead, C.* Genetic Drift, Models of Random / C. Muirhead // *Encyclopedia of Evolutionary Biology*. — Elsevier, 2016. — P. 136—143.
44. *Der, R.* Generalized population models and the nature of genetic drift / R. Der, C. L. Epstein, J. B. Plotkin // *Theoretical Population Biology*. — 2011. — Sept. — Vol. 80, no. 2. — P. 80—99.
45. *Wright, S.* Evolution in Mendelian Populations. / S. Wright // *Genetics*. — 1931. — Mar. — Vol. 16, no. 2. — P. 97—159.

46. *Fisher, R. A.* On the Dominance Ratio / R. A. Fisher // Proceedings of the Royal Society of Edinburgh. — 1923. — Sept. — Vol. 42. — P. 321—341.
47. *Moran, P. A. P.* Random processes in genetics / P. A. P. Moran // Mathematical Proceedings of the Cambridge Philosophical Society. — 1958. — Jan. — Vol. 54, no. 1. — P. 60—71.
48. *Kimura, M.* Solution of a process of random genetic drift with a continuous model / M. Kimura // Proceedings of the National Academy of Sciences. — 1955. — Mar. — Vol. 41, no. 3. — P. 144—150.
49. *Ewens, W. J.* Mathematical Population Genetics. Vol. 27 / W. J. Ewens. — New York, NY : Springer New York, 2004.
50. *Tataru, P.* Inference Under a Wright-Fisher Model Using an Accurate Beta Approximation / P. Tataru, T. Bataillon, A. Hobolth // Genetics. — 2015. — Nov. — Vol. 201, no. 3. — P. 1133—1141.
51. *McKane, A.* Singular solutions of the diffusion equation of population genetics / A. McKane, D. Waxman // Journal of Theoretical Biology. — 2007. — Aug. — Vol. 247, no. 4. — P. 849—858.
52. *Waxman, D.* A compact result for the time-dependent probability of fixation at a neutral locus / D. Waxman // Journal of Theoretical Biology. — 2011. — Apr. — Vol. 274, no. 1. — P. 131—135.
53. *Song, Y. S.* A Simple Method for Finding Explicit Analytic Transition Densities of Diffusion Processes with General Diploid Selection / Y. S. Song, M. Steinrücken // Genetics. — 2012. — Mar. — Vol. 190, no. 3. — P. 1117—1129.
54. Assessing population differentiation and isolation from single-nucleotide polymorphism data / G. Nicholson [et al.] // Journal of the Royal Statistical Society: Series B (Statistical Methodology). — 2002. — Oct. — Vol. 64, no. 4. — P. 695—715.
55. Using Environmental Correlations to Identify Loci Underlying Local Adaptation / G. Coop [et al.] // Genetics. — 2010. — Aug. — Vol. 185, no. 4. — P. 1411—1423.
56. *Gautier, M.* A Bayesian Outlier Criterion to Detect SNPs under Selection in Large Data Sets / M. Gautier, T. D. Hocking, J.-L. Foulley // PLoS ONE. — 2010. — Aug. — Vol. 5, no. 8. — e11913.

57. *Terhorst, J.* Multi-locus Analysis of Genomic Time Series Data from Experimental Evolution / J. Terhorst, C. Schlötterer, Y. S. Song // PLOS Genetics. — 2015. — Apr. — Vol. 11, no. 4. — e1005069.
58. *Balding, D. J.* A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity / D. J. Balding, R. A. Nichols // Genetica. — 1995. — June. — Vol. 96, no. 1/2. — P. 3—12.
59. *Siren, J.* Reconstructing Population Histories from Single Nucleotide Polymorphism Data / J. Siren, P. Marttinen, J. Corander // Molecular Biology and Evolution. — 2011. — Jan. — Vol. 28, no. 1. — P. 673—683.
60. *Nadarajah, S.* A new bivariate beta distribution / S. Nadarajah, S. H. Shih, D. K. Nagar // Statistics. — 2017. — Mar. — Vol. 51, no. 2. — P. 455—474.
61. *Arnold, B. C.* Flexible bivariate beta distributions / B. C. Arnold, H. K. Tony Ng // Journal of Multivariate Analysis. — 2011. — Sept. — Vol. 102, no. 8. — P. 1194—1202.
62. *Nyaga, V. N.* CopulaDTA : An R Package for Copula-Based Bivariate Beta-Binomial Models for Diagnostic Test Accuracy Studies in a Bayesian Framework / V. N. Nyaga, M. Arbyn, M. Aerts // Journal of Statistical Software. — 2017. — Vol. 82.
63. *Gautier, M.* Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates / M. Gautier // Genetics. — 2015. — Dec. — Vol. 201, no. 4. — P. 1555—1579.
64. *Pawlowsky-Glahn, V.* Compositional Data Analysis / V. Pawlowsky-Glahn, A. Buccianti. — Chichester, UK : John Wiley & Sons, Ltd, 09/2011.
65. Be aware of the allele-specific bias and compositional effects in multi-template PCR / I. Korvigo [et al.] // PeerJ. — 2022. — Vol. 10. — e13888.
66. *Aitchison, J.* A General Class of Distributions on the Simplex / J. Aitchison // Journal of the Royal Statistical Society. Series B (Methodological). — 1985. — Vol. 47, no. 1. — P. 136—146.
67. *Aitchison, J.* Logistic-Normal Distributions: Some Properties and Uses / J. Aitchison, S. M. Shen // Biometrika. — 1980. — Vol. 67, no. 2. — P. 261—272.

68. Isometric Logratio Transformations for Compositional Data Analysis / J. J. Egozcue [et al.] // *Mathematical Geology*. — 2003. — Apr. — Vol. 35, no. 3. — P. 279—300.
69. *Dellaportas, P.* An Introduction to MCMC / P. Dellaportas, G. O. Roberts // *Spatial Statistics and Computational Methods* / ed. by J. Møller. — New York, NY : Springer New York, 2003. — P. 1—41.
70. *Handbook of Markov Chain Monte Carlo* / ed. by S. Brooks [et al.]. — Chapman, Hall/CRC, 05/2011.
71. *Gelman, A.* Inference from Iterative Simulation Using Multiple Sequences / A. Gelman, D. B. Rubin // *Statistical Science*. — 1992. — Vol. 7, no. 4. — P. 457—472.
72. *Brooks, S. P.* General Methods for Monitoring Convergence of Iterative Simulations / S. P. Brooks, A. Gelman // *Journal of Computational and Graphical Statistics*. — 1998. — Vol. 7, no. 4. — P. 434—455.
73. *Benaroya, H.* Probability Models in Engineering and Science / H. Benaroya, S. Han, M. Nagurka. — Taylor & Francis, 2005. — P. 166.
74. *Felsenstein, J.* Maximum-likelihood estimation of evolutionary trees from continuous characters / J. Felsenstein // *American journal of human genetics*. — 1973. — Sept. — Vol. 25, no. 5. — P. 471—492.
75. SymPy: symbolic computing in Python / A. Meurer [et al.] // *PeerJ Computer Science*. — 2017. — Jan. — Vol. 3. — e103.
76. Incomplete lineage sorting or secondary admixture: disentangling historical divergence from recent gene flow in the Vinous-throated parrotbill (*Paradoxornis webbianus*) / Y. Qu [et al.] // *Molecular Ecology*. — 2012. — Vol. 21, no. 24. — P. 6117—6133.
77. Matching population diversity of rhizobial nodA and legume NFR5 genes in plant-microbe symbiosis / A. A. Igolkina [et al.] // *Ecology and Evolution*. — 2019. — Aug. — Vol. 9, no. 18. — P. 10377—10386.
78. Structural Insight Into the Role of Mutual Polymorphism and Conservatism in the Contact Zone of the NFR5–K1 Heterodimer With the Nod Factor / A. A. Igolkina [et al.] // *Frontiers in Plant Science*. — 2018. — Vol. 9.

79. *Tanksley, S. D.* Seed banks and molecular maps: unlocking genetic potential from the wild / S. D. Tanksley, S. R. McCouch // *Science*. — 1997. — Aug. — Vol. 277, no. 5329. — P. 1063—1066.
80. Draft genome sequence of *Cicer reticulatum* L., the wild progenitor of chickpea provides a resource for agronomic trait improvement / S. Gupta [et al.] // *DNA Research*. — 2016. — Aug.
81. *Berger, J.* Ecogeography of Annual Wild Cicer Species / J. Berger, S. Abbo, N. C. Turner // *Crop Science*. — 2003. — May. — Vol. 43, no. 3. — P. 1076—1090.
82. *Zohary, D.* Domestication of Plants in the Old World: The Origin and Spread of Cultivated Plants in West Africa, Europe, and the Nile Valley / D. Zohary, M. Hopf. — Oxford: Clarendon Press, 2000.
83. *Toker, C.* Estimation of Outcrossing Rate in Chickpea (*Cicer Arietinum* L.) Sown in Autumn / C. Toker, H. Canci, F. O. Ceylan // *Euphytica*. — 2006. — Sept. — Vol. 151, no. 2. — P. 201—205.
84. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement / R. K. Varshney [et al.] // *Nature Biotechnology*. — 2013. — Mar. — Vol. 31, no. 3. — P. 240—246.
85. Resequencing of 429 chickpea accessions from 45 countries provides insights into genome diversity, domestication and agronomic traits / R. Varshney, M. Thudi, M. Roorkiwal, [et al.] // *Nature Genetics*. — 2019. — Vol. 51. — P. 857—864.
86. Ecology and genomics of an important crop wild relative as a prelude to agricultural innovation / E. J. von Wettberg [et al.] // *Nature Communications*. — 2018. — Feb. — Vol. 9, no. 1. — P. 649.
87. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data / A. McKenna [et al.] // *Genome Research*. — 2010. — Sept. — Vol. 20, no. 9. — P. 1297—1303.
88. The variant call format and VCFtools / P. Danecek [et al.] // *Bioinformatics*. — 2011. — Aug. — Vol. 27, no. 15. — P. 2156—2158.
89. Multi-trait multi-locus SEM model discriminates SNPs of different effects / A. A. Igolkina [et al.] // *BMC Genomics*. — 2020. — July. — Vol. 21, S8. — P. 490.

90. *Igolkina, A. A. semopy: A Python Package for Structural Equation Modeling / A. A. Igolkina, G. Meshcheryakov // Structural Equation Modeling: A Multidisciplinary Journal. — 2020. — Nov. — Vol. 27, no. 6. — P. 952—963.*
91. Historical Routes for Diversification of Domesticated Chickpea Inferred from Landrace Genomics / A. A. Igolkina [et al.] // *Molecular Biology and Evolution.* — 2023. — May. — Vol. 40, no. 6.
92. *Wilgenbusch, J. C. Inferring Evolutionary Trees with PAUP / J. C. Wilgenbusch, D. Swofford // Current Protocols in Bioinformatics. — 2003. — Jan. — Vol. 00, no. 1.*
93. *Lewis, J. leastcostpath: Modelling Pathways and Movement Potential Within a Landscape / J. Lewis. — 2023.*
94. Kabuli and desi chickpeas differ in their requirement for reproductive duration / R. Purushothaman [et al.] // *Field Crops Research.* — 2014. — Vol. 163. — P. 24—31.
95. Exploring germplasm diversity to understand the domestication process in *Cicer* spp. using SNP and DArT markers / M. Roorkiwal [et al.] // *PloS one.* — 2014. — Vol. 9, no. 7. — e102016.
96. Identification of Mendel's White Flower Character / R. Hellens [et al.] // *PLoS One.* — 2010. — Vol. 5. — e13230.
97. BLAST+: architecture and applications / C. Camacho [et al.] // *BMC Bioinformatics.* — 2009. — Vol. 10, no. 1. — P. 421.
98. Identifying loci under positive selection in complex population histories / A. Refoyo-Martínez [et al.] // *Genome Research.* — 2019. — Sept. — Vol. 29, no. 9. — P. 1506—1520.

Словарь терминов

Аллель - один из двух или более версий последовательности ДНК (одно основание или сегмент оснований) в данном геномном локусе.

Генетический дрейф или дрейф генов - эволюционный процесс, при котором частоты аллелей в популяции изменяются случайным образом от поколения к поколению.

Граф предковых рекомбинаций (англ. ancestral recombination graph, ARG) - это структура, которая отражает все события рекомбинаций и коалесценций для набора генетических последовательностей.

Граф смешения (англ. admixture graph) - представление демографической истории группы популяций в виде направленного ациклического графа, отображающего разделение и слияния популяций.

Доместикация - это процесс изменения диких растений или животных, при котором они на протяжении многих поколений содержатся человеком в условиях генетической изоляции от своей дикой формы и подвергаются искусственному отбору с целью получения культурных или домашних форм.

Изоляция расстоянием (англ. isolation by distance, IBD) - эволюционный процесс, при котором генетическое различие между популяциями возрастает с увеличением географического расстояния между ними.

Композиция - набор относительных значений или долей, которые представляют собой пропорциональные части целого.

Модель Райта-Фишера - это модель для описания изменений в частотах аллелей с течением времени в популяции фиксированного размера. Она включает следующие ключевые предположения: (1) популяция характеризуется текущим количеством каждого аллеля; (2) поколения не перекрываются: все особи размножаются и умирают одновременно; (3) новые поколения формируются путем биномиальной выборки аллелей из родительского поколения в случае двух аллелей и мультиномиальной выборки, если аллелей больше двух.

Неполная сортировка линий (англ. incomplete lineage sorting, ILS) - процесс, при котором аллели из разных локусов в популяциях не группируются в соответствии с филогенетическим деревом популяций.

Однонуклеотидный полиморфизм (ОНП, англ. single nucleotide polymorphism, SNP) - локус в геноме, где у разных особей одной популяции различается один нуклеотид.

Панмиксия - случайное скрещивание особей внутри популяции, при котором все индивиды имеют равные шансы на репродукцию.

Поток генов - перемещение генетического материала между популяциями через миграцию особей.

Событие смешения (англ. admixture event) - генетическое событие, при котором происходит скрещивание между двумя или более генетически различными популяциями, что приводит к появлению потомков с новыми комбинациями генов.

Староместный сорт — это продукт народной селекции, главным образом длительного массового отбора, который выращивали на определённой территории на протяжении долгого времени, в результате чего он адаптировался к местным условиям.

Список сокращений и условных обозначений

ОНП: Однонуклеотидный полиморфизм

LD: Linkage Disequilibrium, Неравновесие по сцеплению

LCD: Least Cost Distance, Пути наименьшей стоимости

Список рисунков

1.1	Примеры расположения выборки образцов на гипотетическом географическом ландшафте с учетом различных аспектов.	18
1.1		18
1.2	Параметризация событий смещения в различных программах. . .	27
1.2		27
1.3	Визуальное представление статистик f_2, f_3, f_4	31
1.3		31
1.4	Пример расчета генетического дрейфа между популяциями. . . .	38
1.4		38
1.5	Различные представления композиционных данных.	41
1.5		41
2.1	Схема иерархической Байесовской модели <i>popdisp</i> для случая биаллельных генетических маркеров.	55
2.2	Схема иерархической Байесовской модели <i>popdisp</i> для случая генетических локусов с несколькими аллелями.	62
2.3	Схема иерархической Байесовской модели <i>popdisp</i> для случая биаллельных генетических маркеров неравновесия по сцеплению.	64
2.4	Метод расчета ковариации между двумя географическими точками относительно положения генетического центроида, откуда произошло распространение.	66
2.5	Пример параметризации события смещения в модели <i>migadmi</i> . . .	72
3.1	Гипотетические пути распространения нута на основании анализа главных компонент.	81
3.2	Карта центров происхождения культурных растений, предложенных Н.И. Вавиловым.	83
3.3	Плотность ОНП вдоль хромосом генома <i>Cicer arietinum</i> L.	84
4.1	Ошибка кросс-валидации классификации программы ADMIXTURE при варьировании параметра количества базовых популяций.	86

4.2	Результаты работы программы ADMIXTURE при разных значениях параметра K	88
4.3	Точки сбора образцов, соответствующих каждой из шести исходных популяций.	89
4.4	Характеристика образцов в выделенных девяти популяциях после фильтрации.	91
4.5	Сходство и различия между образцами популяций нута.	92
4.6	Пример симуляции одного набора данных для тестирования программы popdisp.	95
4.7	Результаты работы программы popdisp на симулированных данных.	97
4.8	Примеры построенных оптимальных путей для Черноморской популяции.	98
4.9	Исходное эволюционное дерево, соединяющее пять несмешанных популяций нута дези.	102
4.10	Результаты моделирования Эфиопской дези популяции как смеси.	103
4.11	Распределение вдоль хромосом вкладов различных популяций в Эфиопскую.	105
4.12	Результаты моделирования западно-Среднеазиатской дези популяции как смешанной.	106
4.13	Четыре альтернативные гипотезы происхождения популяций кабули.	108
4.14	Представление характеристических частот аллелей в осях первых двух главных компонент.	109
4.15	Хромосомный локус, ассоциированный с признаком принадлежности к дези/кабули.	110
4.16	Результаты работы программы TreeMix.	112
4.17	Топологии графов смешений для шести популяций дези и одного события смешения.	114
4.18	Топологии графов смешений для шести популяций дези и двух событий смешения.	116
4.19	Топологии графов смешений для пяти популяций дези и четырех популяций кабули.	117

Список таблиц

1	Разложение дисперсии смешанной популяции y на индивидуальные вклады.	71
2	Сравнение методов для моделирования событий смешения.	74
3	Количество образцов в популяциях до и после фильтрации.	90
4	Доминирование взвешенных моделей распространения нута в популяциях.	99