Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи

Вахрушева Ольга Александровна

Эволюционно-генетический сигнал отрицательного отбора и рекомбинации в полногеномных данных

1.5.8. – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата биологических наук

Научный руководитель:
доктор биологических наук
Георгий Александрович Базыкин

Москва – 2022

Оглавление

Введение **4**

Глава 1. Обзор литературы **15**

1.1 Подходы к поиску консервативных некодирующих элементов в геномах эукариот и функциональное значение консервативных некодирующих элементов **15**

1.2 Синергический эпистаз как возможное объяснение парадокса мутационного груза и преимущества полового размножения перед бесполым **18**

1.3 Другие гипотезы, объясняющие преимущество полового размножения **26**

1.4 «Эволюционные скандалы» – предположительно древние группы бесполых организмов **28**

Глава 2. Сигнал действия отрицательного отбора на ортологичные интроны в далеких видах **36**

2.1 Материалы и методы **37**

2.1.1 Геномные данные 37

2.1.2 Определение и анализ ортологичных интронов 40

2.1.3 Оценка ожидаемого числа случаев, в которых интроны в обеих парах внутри четверки несут сегмент сходства 41

2.1.4 Анализ данных по модификациям хроматина 43

2.1.5 Определение потерь интронов 44

2.2 Результаты и обсуждение **45**

2.2.1 Анализ давления отбора на интроны, несущие сегменты сходства в далеких парах видов 45

2.2.2 Анализ давления отбора на интроны, несущие в далеких видах регуляторный элемент 53

2.2.3 Анализ потерь интронов, несущих сегмент сходства в далекой паре видов 56

2.2.4 Обсуждение 60

Глава 3. Сигнал действия синергического эпистатического отрицательного отбора на вредные аллели в популяциях *D. melanogaster* **62**

3.1 Материалы и методы **63**

3.1.1 Наборы данных по полногеномному полиморфизму *D. melanogaster* 63

3.1.2 Референтный геном и данные по аннотации белок-кодирующих генов 64

3.1.3 Контроль качества данных 65

3.1.4 Идентификация и аннотация минорных аллелей 67

3.1.5 Анализ свойств распределения мутационной нагрузки 70

3.1.6 Анализ выборок аллелей, попадающих в «необходимые» гены *D. melanogaster* 70

3.1.7 Данные по отношению скоростей несинонимической и синонимической эволюции для генов *D. melanogaster* 71

3.2 Результаты и обсуждение **72**

3.2.1 Анализ дисперсии распределения мутационной нагрузки для минорных аллелей разных типов в популяциях *D. melanogaster* 73

3.2.2 Анализ распределения мутационной нагрузки для несинонимических аллелей, попадающих в «необходимые» гены 90

3.2.3 Анализ распределения мутационной нагрузки для аллелей, попадающих в гены с разным отношением скоростей несинонимической и синонимической эволюции 92

Глава 4. Подписи рекомбинации и обмена генетическим материалом в популяции бделлоидных коловраток вида *A. vaga* **96**

4.1 Материалы и методы **98**

4.1.1 Получение клональных линий *A. vaga* 98

4.1.2 Выделение ДНК и подготовка библиотек для секвенирования 99

4.1.3 Секвенирование геномной ДНК 99

4.1.4 Первичная обработка и фильтрация парно-концевых прочтений 99

4.1.5 Получение референтной сборки генома для *A. vaga* (L1) 100

4.1.6 Фильтрация контигов, входящих в первоначальную сборку генома референтной линии *A. vaga* (L1) 102

4.1.7 Выделение неизбыточного гаплоидного набора сегментов 105

4.1.8 Аннотация белок-кодирующих генов ... 107
4.1.9 Разбиение генома на аллельные блоки ... 107
4.1.10 Картирование парно-концевых прочтений и фильтрация картирований ... 109
4.1.11 Определение и фильтрация однонуклеотидных полиморфизмов ... 111
4.1.12 Реконструкция гаплотипов для индивидуумов L1-L11 ... 114
4.1.13 Получение и обработка данных, использовавшихся для оценки частоты ошибок при реконструкции гаплотипов ... 119
4.1.14 Анализ неравновесия по сцеплению с использованием реконструированных гаплотипов (фазированных данных) ... 120
4.1.15 Оценка корреляции зиготности (Δ) ... 122
4.1.16 Поиск сигнала рекомбинации ... 123
4.1.17 Анализ распределения значений коэффициента инбридинга ... 123
4.1.18 Симуляции популяций с разной частотой бесполого размножения ... 124
4.1.19 Анализ филогений гаплотипов ... 125
4.1.20 Филогенетический анализ гаплотипов L1-L11 ... 128
4.1.21 Проведение SOWH тестов ... 129
4.1.22 Построение митохондриальной филогении ... 129
4.1.23 Оценка популяционной скорости возникновения мутаций θ ... 130
4.1.24 Оценка популяционной скорости рекомбинации ... 131
4.1.25 Оценка гипотетической частоты мейоза в популяции *A. vaga* ... 132
4.2 Результаты и обсуждение ... 134
4.2.1 Популяционная геномика *A. vaga* ... 134
4.2.2 Анализ митохондриальной изменчивости L1-L11 ... 142
4.2.3 Подписи рекомбинации в геномах *A. vaga* ... 144
4.2.4 Подписи реципрокной рекомбинации у *A. vaga* ... 153
4.2.5 Анализ частоты ошибок реконструкции гаплотипов ... 156
4.2.6 Подписи обмена генетическим материалом между индивидуумами ... 162
4.2.7 Поиск потенциальных признаков контаминации между культурами *A. vaga* в данных по митохондриальной изменчивости ... 168
4.2.8 Исследование возможных сценариев обмена генетическим материалом у *A. vaga* ... 176
4.2.9 Оценка гипотетической частоты мейоза ... 190
4.2.10 Обсуждение ... 191

Заключение ... 193

Выводы ... 196

Благодарности ... 197

Список сокращений и условных обозначений ... 198

Список литературы ... 199

Приложения ... 208

Введение

Актуальность темы исследования

Отрицательный отбор и рекомбинация – важные факторы эволюции геномных последовательностей. Накопленные за последние годы данные секвенирования геномов для большого количества видов позволили понять многое про то, как последовательности изменяются в результате мутаций, отбора и рекомбинации. Развитие технологий секвенирования дает возможность определять и сравнивать не только последовательности геномов особей из разных видов, но и анализировать множественные последовательности геномов особей, принадлежащих к одному и тому же виду. В результате появилась возможность изучать действие отбора и других эволюционных процессов на различных эволюционных масштабах.

Методы сравнительной геномики позволяют проводить поиск участков генома, находящихся под действием отбора, а также изучать зависимость эффективности отбора от разных факторов, в том числе от частоты рекомбинации. Одним из наиболее интересных направлений исследований, которые могут быть проведены с использованием данных по внутривидовой изменчивости, накопленных за последние годы, является изучение эпистаза – явления зависимости эффектов мутации от геномного контекста, в котором она произошла.

Другое направление исследований, ставшее доступным с распространением технологий секвенирования, заключается в поиске «подписей» рекомбинации в геномах видов, считающихся бесполыми. В случае микроскопических организмов убедительно доказать отсутствие полового размножения или выявить свидетельства криптического обмена генетическим материалом с применением классических подходов может быть чрезвычайно сложно. В то же время с помощью сравнения последовательностей геномов можно провести анализ совместимости структуры внутривидовой изменчивости с тем, что ожидается в случае отсутствия полового размножения и рекомбинации. Так, с использованием такого подхода признаки «криптической» рекомбинации были выявлены в геномах паразита *Giardia lamblia* [1], считавшегося бесполым, и геномах вида из группы *Placozoa* [2], группы, вопрос о существовании полового размножения в которой длительное время оставался без ответа. Таким образом, выявление и изучение сигналов отбора и рекомбинации в полногеномных данных позволяет отвечать на различные важные биологические вопросы, многие из которых долгое время оставались открытыми.

Степень разработанности темы

К настоящему моменту опубликовано множество работ, посвященных поиску сигнала отрицательного отбора и рекомбинации в геномных данных разных типов. Выявление сигнала отрицательного отбора по данным дивергенции обычно проводится на основе оценки степени консервативности рассматриваемого участка в геномах далеких видов. Несмотря на большое число методов, направленных на поиск сигнала отрицательного отбора, в некоторых случаях выявление сигнала отрицательного отбора или определение типа отбора является сложной задачей. В данной работе рассмотрено два таких случая. Первый случай соответствует ситуации, в которой рассматриваются ортологичные последовательности из далеких видов, которые могли эволюционировать под действием продолжающегося отрицательного отбора, направленного на сохранение функции, но не обязательно направленного на сохранение последовательности. Такая ситуация может возникнуть, например, для регуляторных элементов, если происходит относительно быстрая эволюция набора сайтов связывания или их взаимного расположения. В этом случае методы, основанные на поиске консервативных участков генома, не выявят сигнал отрицательного отбора. Другой случай, в котором детекция отрицательного отбора может быть осложнена, соответствует ситуации поиска сигнала отбора (в первую очередь в этом контексте интересен эпистатический отбор) на основании данных по внутривидовому полиморфизму, сигнал отбора или эпистатических взаимодействий в которых может быть очень слабым.

В области исследований сигнала отрицательного отбора отдельное направление посвящено изучению эволюции консервативных некодирующих последовательностей. Такие последовательности были описаны как для геномов позвоночных, так и для геномов двукрылых [3–5]. Согласно оценкам, приведенным в одной из ранних работ в этой области, от 0.3 до 1% генома человека соответствует консервативным некодирующим областям, находящимся под давлением сильного отбора у большинства млекопитающих [3]. Несмотря на то, что функциональное значение большинства консервативных некодирующих элементов неизвестно, результаты значительного числа исследований указывают на то, что такие элементы, по всей видимости, часто выполняют регуляторную функцию [4,6,7], в частности играют роль энхансеров или инсуляторов. Однако оценки доли генома, находящейся под действием отрицательного отбора, полученные на основе оценки консервативности, вероятно, являются заниженными, поскольку не все функциональные элементы сохраняют сходство последовательностей на больших эволюционных расстояниях. Вопрос о возможном сохранении функции участка генома без сохранения сходства последовательностей особенно интересен в контексте эволюции регуляторных последовательностей. Так, для функциональных

некодирующих последовательностей ДНК описан ряд случаев, в которых некодирующие последовательности из разных организмов могут выполнять похожие функции и, скорее всего, имеют общее происхождение, несмотря на отсутствие между ними осмысленного выравнивания [8]. Например, энхансер гена человека может обеспечивать нормальную экспрессию ортологичного гена в трансгенных *Danio rerio*, притом что сходство последовательностей между энхансерами человека и *D. rerio* отсутствует [9]. Однако насколько нам известно, до нашей работы на уровне полного генома не проводилось изучения явления, при котором отрицательный отбор может продолжать действовать на ортологичные участки генома, утратившие в далеких видах сходство последовательностей.

Помимо вопроса о действии отбора на определенные участки генома, большое значение имеет вопрос о типе отбора, в частности, действует ли отбор на каждую мутацию независимо от геномного контекста или является эпистатическим. Изучение распространенности и типа эпистатического отбора на вредные мутации в естественных популяциях важно для понимания того, как популяциям человека и других живых существ удается противостоять постоянному притоку вредных мутаций. Большой объем теоретических работ посвящен влиянию эпистатического отбора на мутационный груз в популяциях с половым и бесполым размножением [10–12]. Основной результат этих работ заключается в том, что в случае существования синергического (усиливающего) эпистаза между вредными мутациями мутационный груз в популяции с половым размножением ниже, чем в том случае, если отбор действует на каждую мутацию по отдельности. При этом при бесполом размножении мутационный груз не зависит от типа отбора [10]. Таким образом, с одной стороны, синергические эпистатические взаимодействия между вредными мутациями являются возможным объяснением парадокса мутационного груза [13], а, с другой стороны, могут быть одним из ключевых факторов, определяющих преимущество полового размножения над бесполым [11–13]. Несмотря на то, что эпистатические взаимодействия разных типов были описаны для большого числа мутаций [14,15], до недавнего времени вопрос о том, насколько распространен синергический эпистаз между вредными мутациями на уровне всего генома, оставался открытым.

Помимо гипотезы, объясняющей преимущество полового размножения и рекомбинации более низким мутационным грузом (в случае присутствия синергических взаимодействий), существует целый ряд теорий, предлагающих другие объяснения тому факту, что половое размножение преобладает среди эукариот. Так, существуют гипотезы, объясняющие преимущество полового размножения над бесполым эффектами, связанными с дрейфом генов [16,17], более высокой эффективностью положительного отбора [18–21], и более высокой скоростью приспособления к изменяющимся условиям среды [22].

Какие именно из этих факторов в действительности создают преимущество для полового размножения, неизвестно, но, по всей видимости, это преимущество является значительным, т.к. переход к бесполому размножению обычно заканчивается относительно быстрым вымиранием [23]. На этом основании бесполое размножение часто рассматривается как «эволюционный тупик». В связи с этим внимание исследователей привлекли немногочисленные исключения из этого правила – предположительно древние группы бесполых организмов. В качестве наиболее яркого примера такой группы обычно приводили класс бделлоидных коловраток, группу микроскопических беспозвоночных, как считалось, отказавшихся от полового размножения десятки миллионов лет назад. Основным аргументом в пользу строго бесполого размножения у видов этой группы служило отсутствие самцов среди сотен тысяч особей бделлоидных коловраток, проанализированных разными исследователями [24]. Однако молекулярно-генетические и геномные данные, полученные в последние годы, не позволяли сделать убедительный вывод о существовании или отсутствии полового размножения и рекомбинации у видов этой группы. Так, анализ первого опубликованного генома бделлоидной коловратки *Adineta vaga*, вышедший в 2013, дал основания предполагать, что структура этого генома несовместима с классическим мейозом [25]. Однако эти данные не нашли подтверждения при анализе генома другой коловратки из рода *Adineta*, *A*. *ricciae* [26], а также при анализе новой сборки генома хромосомного уровня, недавно полученной для *A*. *vaga* с использованием комбинации различных технологий секвенирования [27]. Первая работа в области популяционной геномики бделлоидных коловраток была опубликована в 2015 году: в этой работе был проведен анализ нескольких участков ядерного генома у 6 особей вида *Macrotrachela quadricornifera* [28]. Результаты этого анализа были интерпретированы как вероятное свидетельство чрезвычайно редкого типа мейоза, описанного ранее у растений из рода *Oenothera*, происходящего таким образом, что рекомбинация затрагивает только теломерные области хромосом, без выравнивания гомологичных хромосом по длине относительно друг друга [28]. Работ, в которых на полногеномных данных проводили бы поиск рекомбинации и «подписей» генетического обмена у бделлоидных коловраток, до последнего времени опубликовано не было.

Цели и задачи исследования

Целью данной работы являлся поиск сигнала отрицательного отбора и рекомбинации в данных разных типов, в том числе поиск возможных свидетельств рекомбинации в геномах бделлоидных коловраток, которых ранее рассматривали как группу предположительно древних бесполых видов.

Для достижения этой цели были поставлены следующие задачи:

1) изучение возможного продолжения действия отрицательного отбора на ортологичные некодирующие участки генома, потерявшие сходство последовательностей в далеких видах;
2) поиск сигнала эпистатического отбора, действующего на вредные аллели в белок-кодирующих генах, по данным внутрипопуляционной изменчивости *Drosophila melanogaster*;
3) поиск сигнала рекомбинации и обмена генетическим материалом в данных по внутрипопуляционной изменчивости бделлоидной коловратки вида *Adineta vaga* и исследование совместимости данных по внутрипопуляционной изменчивости *A. vaga* с различными эволюционными сценариями.

Научная новизна

В данной работе был получен ответ на ряд новых и сформулированных ранее, но остававшихся открытыми, вопросов в области эволюционной геномики. Так, нами на полногеномных данных было получено свидетельство о продолжении действия отрицательного отбора на ортологичные некодирующие участки генома, потерявшие в далеких видах сходство последовательностей. Примеры этого явления были описаны ранее на основании экспериментальных данных, но, насколько нам известно, до нашей работы на уровне всего генома это явление исследовано не было.

Кроме того, в данной работе впервые на уровне всего генома был выявлен сигнал синергического эпистатического отбора, действующего на вредные аллели в популяции *D. melanogaster*. Возможное существование синергических эпистатических взаимодействий между вредными мутациями ранее широко обсуждалось в литературе, в первую очередь, в теоретических работах, в которых исследовалась зависимость мутационного груза от присутствия и типа эпистатических взаимодействий. Тем не менее вопрос о существовании и

преобладающем типе эпистатических взаимодействий на уровне всего генома у эукариотических видов оставался малоизученным.

В работе впервые на полногеномных данных выявлен сигнал рекомбинации и обмена генетическим материалом для вида, относящегося к группе бделлоидных коловраток. Вопрос о возможном существовании обмена генетическим материалом у бделлоидных коловраток исследовался ранее, но в предыдущих работах, посвященных изучению этого вопроса, использовались или полногеномные последовательности одного индивидуума, или последовательности небольшого числа геномных локусов. Поиск сигнала рекомбинации у бделлоидных коловраток на полногеномных данных ранее также не проводился. Таким образом, в данной работе впервые получены полногеномные свидетельства рекомбинации у бделлоидных коловраток. В предыдущих исследованиях выдвигались различные гипотезы относительно возможного механизма обмена генетическим материалом у бделлоидных коловраток. В частности, обсуждалась возможность существования горизонтального переноса генов внутри популяций видов из этой группы. В рамках выполнения исследования получены указания на то, что половое размножение является более вероятным объяснением обмена генетическим материалом и рекомбинации в популяциях бделлоидных коловраток.

Теоретическая и практическая значимость

Результаты, полученные в данной работе, позволяют лучше понять некоторые фундаментальные эволюционные закономерности и имеют в первую очередь теоретическую значимость. Исследование сигнала отрицательного отбора, продолжающего действовать на ортологичные некодирующие участки генома после утраты очевидного сходства последовательностей, представляет интерес с точки зрения изучения принципов эволюции функциональных некодирующих последовательностей. Результаты, указывающие на вероятное существование синергического эпистаза, сигнал которого был выявлен в данной работе у *D. melanogaster* для аллелей, вызывающих потерю функции гена, являются важными с точки зрения исследований парадокса мутационного груза и возможных причин преобладания полового размножения. Свидетельства рекомбинации и обмена генетическим материалом у бделлоидной коловратки *A. vaga*, выявленные в данной работе, имеют фундаментальное значение, поскольку потенциально позволяют объяснить эволюционный парадокс, которым считалось существование видов этой группы. Кроме того, обнаружение сигнала рекомбинации в геномах бделлоидных коловраток, которых в течение длительного времени относили к

группам древних бесполых видов, по всей видимости, является важным аргументом, говорящим о важности рекомбинации для долгосрочного эволюционного успеха.

Некоторые полученные результаты и подходы, примененные в рамках данной работы, могут иметь практическую значимость. Так, результаты, указывающие на то, что отрицательный отбор может продолжать действовать на имеющие общее происхождение участки генома даже после того, как они разошлись в далеких видах до неузнаваемости, могут иметь значение с точки зрения разработки методов выявления функциональных элементов, методов поиска последовательностей, находящихся под действием отрицательного отбора, и поиска гомологичных последовательностей на далеких эволюционных расстояниях. Подход к поиску сигнала обмена генетическим материалом, основанный на анализе трехаллельных сайтов и использовавшийся в данной работе при анализе данных *A. vaga*, может быть применен и в других исследованиях с похожей проблематикой.

Методология и методы исследования

Для решения поставленных задач использовалось множество методов биоинформатики, сравнительной и популяционной геномики. Использованные методы включают выравнивание ортологичных белковых и нуклеотидных последовательностей, определение ортологичных последовательностей в разных геномах, построение филогенетических деревьев (с помощью методов максимального правдоподобия и метода ближайших соседей). Часть задач осуществлялась с использованием открытых геномных данных. Кроме того, в рамках реализации работы были получены собственные данные секвенирования 11 геномов бделлоидных коловраток вида *A. vaga*. Для последующего анализа полученных данных секвенирования использовались стандартные методы первичной обработки и картирования прочтений, определения однонуклеотидных полиморфизмов. Определение однонуклеотидных полиморфизмов для особей *A. vaga* подразумевало использование референтного генома. В связи с этим для одной из анализируемых особей была получена геномная сборка, далее использовавшаяся в качестве референтной последовательности. Кроме того, для локальной реконструкции гаплотипов в работе был использован метод вычислительного фазирования полиморфизмов. Поиск сигнала рекомбинации проводили с помощью широко применяемых методов для выявления рекомбинации и с применением модифицированного нами варианта четырехгаметного теста. Исследование внутрипопуляционной изменчивости *A. vaga* включало определение коэффициента инбридинга для биаллельных полиморфных сайтов и анализ трехаллельных сайтов.

Основные положения, выносимые на защиту

1. Действие отрицательного отбора на ортологичные интроны может продолжаться даже тогда, когда между последовательностями имеющих общее происхождение интронов в далеких видах уже не существует осмысленного выравнивания. Сигнал, свидетельствующий о том, что давление отрицательного отбора на ортологичные, но разошедшиеся до неузнаваемости последовательности, может сохраняться, присутствует в геномных данных как для позвоночных, так и для двукрылых. Это явление может быть связано с сохранением предковой функции некодирующих участков генома, утративших в далеких видах сходство последовательностей.

2. Понижение дисперсии мутационной нагрузки по сравнению с аддитивной дисперсией для аллелей, вызывающих потерю функции гена, в двух популяциях *D. melanogaster* указывает на существование синергических эпистатических взаимодействий между мутациями данного типа. Сигнал синергического эпистаза также присутствует в подмножестве несинонимических аллелей, попадающих в гены *D. melanogaster*, находящиеся под сильным давлением отрицательного отбора.

3. В геномах бделлоидных коловраток вида *A. vaga* выявлен сигнал рекомбинации, который не может быть объяснен исключительно действием генной конверсии и, вероятно, связан с реципрокной рекомбинацией. Данные по внутрипопуляционной изменчивости *A. vaga* свидетельствуют об обмене генетическим материалом, происходящем в популяциях этого вида. Некоторые закономерности, выявленные при анализе филогений гаплотипов разных особей, указывают на то, что половое размножение является более вероятным механизмом обмена генетическим материалом внутри популяций бделлоидных коловраток, чем горизонтальный перенос генов.

Личный вклад автора в исследование

Все результаты, представленные в диссертации, получены лично соискателем или при его непосредственном участии, за исключением результатов, соответствующих перечисленным ниже частям работы. Определение видовой принадлежности бделлоидных коловраток и получение первичных клональных культур *A. vaga* было выполнено Е. А. Мнацакановой и Я. Р. Галимовым. Выделение ДНК для части клональных культур *A. vaga* было проведено Т. В. Неретиной. Секвенирование ДНК клональных культур *A. vaga* на инструментах Illumina было выполнено М. Д. Логачёвой и А. А. Пениным. Аннотация белок-кодирующих генов в геноме *A. vaga* L1 (раздел 4.1.8; Таблица 4.4) и анализ «полноты» геномной сборки *A. vaga* L1 были выполнены Е. С. Герасимовым. Построение филогенетических деревьев для особей *A. vaga* L1-L11 и референтных изолятов бделлоидных коловраток по митохондриальным данным (раздел 4.1.22; Рисунки 4.3, 4.4, 4.22 и 4.23) и подготовка данных для этого анализа были выполнены С. А. Науменко. Кроме того, в диссертации коротко обсуждаются результаты симуляций и анализа мутационной нагрузки, выполненного на наборах данных по полиморфизму человека, полученные М. Сохаил. Обсуждение и интерпретация результатов осуществлялись автором совместно с научным руководителем и соавторами публикаций.

Степень достоверности и апробация результатов

По материалам диссертации опубликовано три статьи в рецензируемых научных журналах. Результаты работы были представлены на встречах Общества молекулярной биологии и эволюции (Society for Molecular Biology and Evolution) в 2012, 2017 и 2019 годах (SMBE 2012 – Дублин, Ирландия; SMBE 2017 – Остин, Техас, США; SMBE 2019 – Манчестер, Англия) и Московской международной конференции по вычислительной молекулярной биологии в 2019 году (Moscow Conference on Computational Molecular Biology, MCCMB'19 – Москва, Россия), а также на конференциях «Информационные технологии и системы» в 2012, 2016 и 2018 годах (ИТиС 2012 – Петрозаводск, Россия; ИТиС 2016 – Репино, Санкт-Петербург, Россия; ИТиС 2018 – Казань, Россия).

Структура и объем диссертации

Диссертация изложена на 218 страницах машинописного текста и содержит следующие разделы: введение, обзор литературы, результаты и обсуждение в трех главах, заключение и выводы. В конце приведён список литературы. Материал включает 42 рисунка, 31 таблицу, 4 таблицы в приложении, а также список литературы, содержащий 216 ссылок.

Список публикаций по теме диссертации

По теме диссертации опубликовано три статьи в рецензируемых международных научных журналах, входящих в основные библиометрические базы данных (PubMed, WoS и Scopus):

1. **Vakhrusheva O. A.**, Bazykin G. A., Kondrashov A. S. Genome-Level Analysis of Selective Constraint without Apparent Sequence Conservation // Genome Biology and Evolution. 2013. Vol. 5, № 3. P. 532–541.

2. Sohail M., **Vakhrusheva O. A.**, Sul J. H., Pulit S. L., Francioli L. C., Genome of the Netherlands Consortium, Alzheimer's Disease Neuroimaging Initiative, van den Berg L. H., Veldink J. H., de Bakker P. I. W., Bazykin G. A., Kondrashov A. S., Sunyaev S. R. Negative selection in humans and fruit flies involves synergistic epistasis // Science. 2017. Vol. 356, № 6337. P. 539–542.

3. **Vakhrusheva O. A.**, Mnatsakanova E. A., Galimov Y. R., Neretina T. V., Gerasimov E. S., Naumenko S. A., Ozerova S. G., Zalevsky A. O., Yushenova I. A., Rodriguez F., Arkhipova I. R., Penin A. A., Logacheva M. D., Bazykin G. A., Kondrashov A. S. Genomic signatures of recombination in a natural population of the bdelloid rotifer *Adineta vaga* // Nature Communications. 2020. Vol. 11, № 1:6421. doi: 10.1038/s41467-020-19614-y.

Кроме того, результаты работы опубликованы в сборниках тезисов международных и российских конференций:

1. Vakhrusheva O. A., Bazykin G. A., Kondrashov A. S. Selective constraint beyond apparent sequence conservation // Информационные технологии и системы 2012 (ИТиС 2012), Петрозаводск, Россия, 19–25 августа 2012. http://www.itas2012.iitp.ru/pdf/1569601189.pdf

2. Vakhrusheva O. A., Mnatsakanova E. A., Galimov Y., Gerasimov E. S., Neretina T. V., Penin A. A., Logacheva M. D., Bazykin G. A., Kondrashov A. S. Population genomic data reveal signatures of genetic exchange in the bdelloid rotifer *Adineta vaga* // Информационные технологии и системы 2018 (ИТиС 2018), Казань, Россия, 25–30 сентября 2018. http://itas2018.iitp.ru/media/papers/1570471700.pdf

3. Vakhrusheva O. A., Mnatsakanova E. A., Galimov Y., Neretina T. V., Gerasimov E. S., Ozerova S. G., Zalevsky A. O., Yushenova I. A., Arkhipova I. R., Penin A. A., Logacheva M. D., Bazykin G. A., Kondrashov A. S. Signatures of genetic exchange in a natural population of the bdelloid rotifer *Adineta vaga* inferred from whole-genome data // Proceedings of the International Moscow Conference on Computational Molecular Biology 2019 (MCCMB'19), Moscow, Russia, July 27–30, 2019. http://mccmb.belozersky.msu.ru/2019/thesis/MCCMB2019/abstracts/64.pdf

Глава 1. Обзор литературы

1.1 Подходы к поиску консервативных некодирующих элементов в геномах эукариот и функциональное значение консервативных некодирующих элементов

В основе значительного числа методов для поиска подписей отрицательного отбора на геномных данных лежит анализ сопоставленных друг другу ортологичных участков из геномов разных видов. При этом степень консервативности последовательности в геномах из далеких видов обычно рассматривается как показатель того, насколько данный участок функционально важен.

Безусловно, действие отрицательного отбора, направленное на сохранение функции, часто выражается в консервативности последовательности соответствующего участка генома на значительных эволюционных расстояниях. Так, многие важные эукариотические белки сохраняют сходство последовательностей с бактериальными ортологами. В число этих белков входят, например, белки, вовлеченные в репарацию мисмэтчей, рибосомные белки, аминоацил-тРНК-синтетазы и ДНК-хеликаза [29]. Примеры такой ультраконсервативности существуют и среди некодирующих участков генома и представлены, например, элементами, последовательность которых идентична или практически идентична среди всех млекопитающих или даже позвоночных [3,4]. В одной из первых работ, описывающих консервативные некодирующие участки в человеческом геноме, приводились оценки, согласно которым от 0.3 до 1% генома человека соответствуют консервативным некодирующим областям, находящимся под давлением сильного отбора у большинства млекопитающих [3]. Эти участки в среднем характеризуются более высокой консервативностью, чем белок-кодирующие гены. Распределение немногочисленных замен в ультраконсервативных областях позволило сделать предположение о том, что эти участки, вероятно, имеют регуляторную функцию [3]. В более поздней работе было показано, что ультраконсервативные элементы, которые определяли как некодирующие участки длиной хотя бы 200 нуклеотидов, последовательность которых идентична в геномах мыши и человека, чаще, чем ожидается по случайным причинам, находятся рядом с генами, вовлеченными в процессы регуляции транскрипции и процессы, связанные с эмбриональным развитием, в частности, с развитием нервной системы [6]. В той же работе было показано, что ~50% из проанализированных ультраконсервативных элементов обладают активностью эмбриональных энхансеров в опытах, проведенных на трансгенных эмбрионах мыши. Энхансерная активность ультраконсервативных некодирующих участков наиболее часто проявлялась в нервных тканях

эмбриона [6]. Кроме того, было показано, что консервативные элементы, находящиеся вне экзонов, обогащены мотивами, который распознает белок CTCF, связывающийся с инсуляторами [7]. Вероятно, это означает, что часть консервативных последовательностей, находящихся вне экзонов, обладают активностью инсуляторов. В целом было обнаружено 233 мотива, которыми обогащены консервативные элементы, находящиеся вне экзонов [7], однако функциональное значение для большинства из этих мотивов неизвестно.

Другие свидетельства в пользу регуляторной роли консервативных некодирующих элементов приведены в работе 2011 года, в которой было показано, что эти элементы обогащены сайтами гиперчувствительности к ДНКазе I и, следовательно, часто находятся в участках открытого хроматина [4]. По данным ChiP-Seq та же выборка элементов обогащена и другими функциональными мотивами, включающими сайты связывания транскрипционных факторов SRF (сокращение от англ. serum response factor) и GABP (сокращение от англ. growth associated binding protein) [4].

В эволюции позвоночных было выделено три основных периода возникновения новых регуляторных элементов, которые затем оказались под действием сильного отрицательного отбора, и составляют значительную долю консервативных некодирующих элементов в геномах позвоночных [4]. При этом для элементов, возникших в разные периоды «регуляторных инноваций», характерна ассоциация с разными функциональными группами генов [4]. Регуляторные элементы, возникшие в первый период, пришедшийся на раннюю эволюцию позвоночных (время от возникновения общего предка всех позвоночных до разделения ветви млекопитающих с ветвью, ведущей к рептилиям и птицам), обогащены рядом с генами транскрипционных факторов и генами, участвующими в процессах развития [4]. Для регуляторных элементов, ставших мишенями отрицательного отбора в два следующих периода, характерна ассоциация с генами, вовлеченными в передачу сигнала между клетками, и генами, участвующими в процессах посттрансляционных модификаций белков, соответственно [4].

Консервативные некодирующие элементы присутствуют не только в геномах позвоночных, но и в геномах двукрылых. В одной из первых статей, посвященных изучению этого явления у *Drosophila*, был проведен поиск некодирующих последовательностей, консервативных между видами *D. melanogaster* и *D. virilis*, разошедшихся ~40 миллионов лет назад [5]. Уровень геномной дивергенции между этими видами сопоставим с уровнем дивергенции между видами в паре человек – мышь [5]. Медианная длина консервативного блока в некодирующих участках генома между *D. melanogaster* и *D. virilis* составляет всего 19 нуклеотидов, причем распределение длин и плотность консервативных блоков не отличается между интронами и межгенными интервалами [5]. В одной из последующих работ было сделано предположение, что выявленное распределение длин консервативных некодирующих

блоков у *Drosophila* [5] может быть совместимо с моделью, в которой их существование объясняется не действием отрицательного отбора, сохраняющего их последовательность, а пониженной скоростью мутирования [30]. Однако сравнение данных по дивергенции и полиморфизму показало, что некодирующие консервативные участки в геномах видов *Drosophila* действительно находятся под сильным отрицательным отбором [31]. Паттерны эволюции некодирующих консервативных участков у *Drosophila* не могут быть объяснены в рамках нейтральной модели. Так, например, спектр аллельных частот для производных полиморфных вариантов в таких участках смещен в сторону редких производных аллелей [31]. Интересно, что оценки коэффициентов отбора, действующего на мутации, возникающие в некодирующих консервативных элементах у *Drosophila* ($N_e s$ в диапазоне 10–100), указывают на более сильный отрицательный отбор на элементы этого типа у *Drosophila* по сравнению с человеком [31].

Проведенный в работе [5] анализ показал, что 22–26% некодирующих сайтов находятся под отбором в видах *Drosophila* (здесь речь идет не об ультраконсервативных элементах, а о любой значимой консервативности последовательностей). Схожие оценки были получены ранее и для других эукариотических геномов. Так, с помощью применения метода, основанного на скрытой марковской модели, было предсказано, что в геноме мыши 36% и 23% сайтов в апстримных областях генов и в интронах соответственно находятся в консервативных некодирующих блоках [32]. В другой работе было проведено выравнивание некодирущих областей геномов *Caenorhabditis elegans* и *C. briggsae* с помощью алгоритмов динамического программирования [33]. Это позволило оценить, что по крайней мере ~17–18% сайтов как в межгенных областях, так и в интронах в геномах *Caenorhabditis* эволюционируют под давлением отрицательного отбора.

По другим оценкам, полученным на основе более позднего метода, использующего скрытую марковскую модель, от 3% до 8% последовательности генома человека входит в состав консервативных элементов [34]. При этом 70% выявленных консервативных элементов приходится на некодирующие участки генома [34]. У более просто устроенных эукариот, таких как двукрылые или черви, консервативные элементы занимают более значительную долю генома (37–53% у *D. melanogaster* и 18–37% у *C. elegans*) [34]. Однако процент некодирующих участков среди консервативных элементов растет с увеличением размера генома и общей сложности организма (от дрожжей к позвоночным) [34].

Все приведенные выше оценки доли некодирующих позиций, находящихся под отбором в разных видах, получены на основании поиска сигнала консервативности и, вероятно, являются заниженными, поскольку не все функциональные элементы сохраняют сходство последовательностей на больших филогенетических расстояниях. Например, 32–40% сайтов

связывания транскрипционных факторов, для которых выявлена функциональная активность у человека, не обладают соответствующей функцией у грызунов [35], что, по всей видимости, указывает на быструю эволюцию сайтов связывания.

Кроме того, сохранение первичной последовательности, на основании которой можно реконструировать осмысленное выравнивание, не является обязательным условием для сохранения других свойств молекулы. Так, белки, последовательности которых не могут быть выровнены, тем не менее могут иметь похожие трехмерные структуры [36].

Одноцепочечные РНК, имеющие несхожие последовательности, могут обладать одинаковой вторичной структурой [37] (в качестве очевидного примера можно привести последовательности AAAAAGGGTTTTT и GGGGGTTTCCCCC). Для функциональных некодирующих последовательностей ДНК также описан ряд случаев, в которых некодирующие последовательности из разных организмов могут выполнять похожие функции и, скорее всего, имеют общее происхождение, несмотря на отсутствие между ними осмысленного выравнивания [8]. Например, энхансер гена человека может обеспечивать нормальную экспрессию ортологичного гена в трансгенных *D. rerio*, притом что сходство последовательностей между энхансерами человека и *D. rerio* отсутствует [9]. В другом исследовании, проведенном для *D. rerio*, было показано, что использование существовавших в тот момент методов построения выравниваний в значительном проценте случаев не позволяет выявить функциональные последовательности [38]. Некодирующие элементы, регулирующие гомологичные гены в нематодах и позвоночных, обладают сходством последовательностей внутри обеих клад, но не между организмами из разных клад, что подразумевает «переключение» регуляторных механизмов [39]. Однако насколько нам известно, до нашей работы (см. Главу 2) не проводилось полногеномных исследований возможного действия отрицательного отбора на ортологичные участки генома, разошедшиеся в далеких видах до неузнаваемости.

1.2 Синергический эпистаз как возможное объяснение парадокса мутационного груза и преимущества полового размножения перед бесполым

Основная часть методов, направленных на выявление подписей отрицательного отбора на данных дивергенции или полиморфизма, не позволяют определить, действует ли отбор на каждую мутацию независимо от других или эффект мутации на приспособленность зависит от геномного контекста, в котором она произошла. Тем временем вопрос о распространенности и типе эпистатического отбора на вредные мутации в естественных популяциях представляет

большой интерес с точки зрения того, как популяциям человека и других живых существ удается противостоять постоянному притоку вредных мутаций.

В том случае, если отрицательный отбор действует на вредные мутации по отдельности, скорость падения относительной приспособленности с каждой новой мутацией не зависит от количества мутаций в геноме [10,40]. Популяционно-генетическая теория позволяет соотнести скорость возникновения вредных мутаций на геном с долей особей в каждом поколении, которая не должна оставить потомков в результате действия отрицательного отбора [10]. Однако соответствующие вычисления показывают, что в том случае, если эпистатические взаимодействия между вредными мутациями отсутствуют, в видах, у которых на поколение на геном происходит более одной вредной мутации, >60% особей не должны оставлять жизнеспособное потомство («генетическая смерть») [41,42]. Очевидно, это противоречит тому, что мы в действительности наблюдаем у человека и других видов, оставляющих небольшое количество потомков. Это так называемый парадокс мутационного груза [43].

Термином мутационный груз обычно обозначают относительное понижение средней приспособленности популяции из-за вредных мутаций относительно приспособленности оптимального генотипа [10,40,44]. В случае отсутствия эпистатических взаимодействий между вредными мутациями в равновесии между мутациями и отбором средняя приспособленность как половой, так и бесполой популяции равна $\overline{w} = e^{-U}$, а мутационный груз L и доля «генетических смертей» в популяции равны $L \approx 1 - e^{-U}$, где U – общее число вредных мутаций, происходящих на диплоидный геном на поколение [10,40].

Однако в том случае, если вклад отдельной мутации в приспособленность генотипа зависит от присутствия других мутаций (эпистаз), при половом размножении мутационный груз может быть значительно меньше или больше, чем при бесполом размножении. С точки зрения взаимных эффектов мутаций на приспособленность обычно выделяют два основных типа эпистаза – синергический (усиливающий) и антагонистический (ослабляющий). При синергическом эпистазе эффект от одновременного присутствия нескольких вредных мутаций оказывается больше, чем мы бы ожидали, исходя из влияния на приспособленность каждой мутации по отдельности. В этом случае логарифм приспособленности с ростом числа вредных мутаций в геноме падает быстрее, чем линейно (Рисунок 1.1) [12]. И, напротив, при антагонистическом эпистазе эффект нескольких вредных мутаций меньше, чем можно было бы ожидать из независимых эффектов этих мутаций на приспособленность. При этом типе эпистаза падение приспособленности с ростом числа вредных мутаций в геноме происходит медленнее, чем в отсутствие эпистаза (Рисунок 1.1) [12].

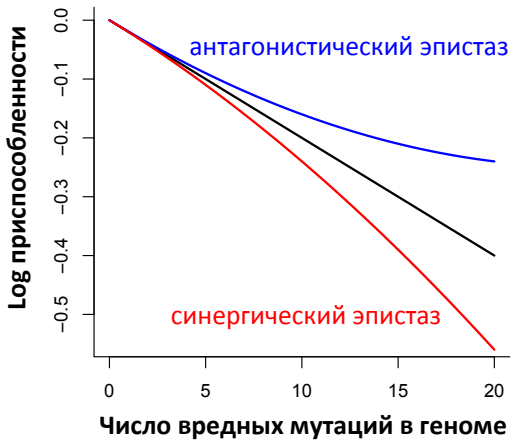

Рисунок 1.1. Зависимость приспособленности от числа вредных мутаций в геноме в отсутствие взаимодействий между мутациями и при эпистатических взаимодействиях двух типов. Черной сплошной линией показано падение натурального логарифма приспособленности с увеличением числа вредных мутаций в геноме в отсутствие эпистаза, красной – при синергических эпистатических взаимодействиях, синей – при антагонистических эпистатических взаимодействиях. Для построения графика использована формула, описывающая зависимость приспособленности от числа вредных мутаций, из статьи B. Charlesworth (1990) [12].

Явление эпистаза на уровне всего генома привлекло большой интерес теоретических биологов во второй половине двадцатого века [10–13]. Связано это в первую очередь с тем, что вероятное существование синергического эпистаза рассматривают как возможное объяснение преобладания полового размножения среди эукариот. Тот факт, что половое размножение доминирует среди эукариотических организмов, является парадоксальным из-за «двукратной цены» полового размножения [45]. Впервые парадокс «двукратной цены» полового размножения был сформулирован Мэйнардом Смитом в его книге «Эволюция полового размножения» [45]. Парадокс заключается в том, что в половой популяции самка «вкладывает» половину ресурсов в самцов, которые сами не могут производить потомство. В то же время в бесполой популяции самка не тратит ресурсы на самцов. При прочих равных это должно приводить к тому, что рост половой популяции происходит в два раза медленнее, чем бесполой [45]. Соответственно, ожидается, что бесполый «клон», возникший в популяции, размножающейся половым способом, должен быстро вытеснить половых особей, т.к. частота соответствующей мутации будет удваиваться каждое поколение.

Попытки объяснить, почему половое размножение преобладает среди многоклеточных организмов, несмотря на его высокую цену, привели к созданию множества разных теорий [46]. Согласно существующим классификациям эти теории условно делятся на две большие группы:

теории «немедленного преимущества» (или «физиологические» теории) и «генетические» теории [22,47]. Согласно одной из основных теорий из группы «немедленного преимущества» половое размножение возникло как побочный результат механизма репарации двухцепочечных разрывов и поддерживается в связи с тем, что мейоз и рекомбинация необходимы для репарации ДНК [48]. «Генетические» теории исходят из того, что половое размножение и рекомбинация обладают эволюционным преимуществом, которое связано с эффектами этих процессов на изменчивость в популяции, изменяющими ответ на отбор [22].

Половое размножение может иметь преимущество перед бесполым, например, в том случае, если существуют факторы, благодаря которым распределение аллелей по гаплотипам отличается от случайного так, что некоторые комбинации аллелей систематически недопредставлены в популяции [46]. Такими факторами могут быть дрейф генов [18,19] или синергический эпистаз между вредными мутациями [11,12].

Так, Кимура и Маруяма в 1966 опубликовали работу, посвященную изучению мутационного груза при эпистатических взаимодействиях [10]. Они показали, что в том случае, если эффект вредных мутаций на приспособленность пропорционален квадрату числа вредных мутаций (эта ситуация соответствует синергическому эпистазу), то при половом размножении мутационный груз может быть приблизительно в два раза меньше ($L \approx M$, где M – скорость вредных мутаций на гамету на поколение), чем в отсутствие эпистаза (в этом случае $L \approx 2M$). И, напротив, при половом размножении в случае антагонистического эпистаза (суммарный эффект нескольких вредных мутаций слабее, чем ожидалось бы исходя из их независимых эффектов на приспособленность), мутационный груз выше, чем в отсутствие эпистаза [10].

Важно отметить, что в случае бесполого размножения мутационный груз не зависит от присутствия эпистатических взаимодействий [10]. Если эпистатические взаимодействия отсутствуют, то при низкой скорости мутаций мутационный груз и приспособленность популяции не зависит от типа размножения (средняя приспособленность как половой, так и бесполой популяции бесконечного размера в равновесии равна $\overline{w} = e^{-U}$) [10,40].

Кондрашов рассмотрел специальный случай синергического эпистаза – отсекающий отбор, соответствующий ситуации, при которой все особи, несущие менее чем k мутаций, имеют одинаковую приспособленность, а особи с большим числом мутаций – нежизнеспособны [11]. В этой работе было показано, что в случае отсекающего отбора половое размножение может иметь значительное преимущество перед бесполым (при достаточно высокой геномной скорости вредных мутаций). Кроме того, была рассмотрена более общая функция приспособленности вида $w_n = 1 - \frac{n}{k^\alpha}$, где w_n – приспособленность особи, несущей n вредных мутаций. При $\alpha = 0$ приспособленность падает линейно, при $\alpha \to \infty$ ситуация соответствует отсекающему отбору. Помимо этих двух сценариев, был рассмотрен

промежуточный случай с $\alpha = 2$. Сравнение этих трех сценариев показало, что преимущество полового размножения тем сильнее, чем больше α: минимальное повышение равновесной приспособленности половой популяции относительно бесполой наблюдается в случае линейного падения приспособленности с ростом мутационной нагрузки ($\alpha = 0$), а максимальное преимущество достигается в случае отсекающего отбора ($\alpha \to \infty$). Тем не менее согласно результатам этой работы, если скорость вредных мутаций достаточна высока, половое размножение выигрывает у бесполого даже в случае линейного падения приспособленности с числом вредных мутаций [11]. При этом половое размножение тем выгоднее, чем выше геномная скорость возникновения вредных мутаций на поколение (U) [11]. В том случае, если $U \geq 2$, преимущество полового размножения перед бесполым более чем двукратное [11].

Чарльзворт исследовал среднюю приспособленность популяции при разных типах отбора и отбор на модификаторы скорости рекомбинации в случае мутационно-отборного равновесия в многолокусной системе в бесконечно большой популяции [12]. В отличие от работ Кимуры и Маруямы (рассматривавших ситуацию, в которой зависимость приспособленности от числа вредных мутаций на геном описывается квадратичной функцией) [10] и Кондрашова [11], Чарльзворт исследовал функцию приспособленности, устроенную таким образом, что логарифм приспособленности является квадратичной функцией от числа вредных мутаций. Такой подход позволяет сохранить нормальность распределения числа вредных мутаций после отбора [12], при этом в случае мутаций малого эффекта (и если число вредных мутаций на геном достаточно низкое) данная функция ведет себя похожим образом на функцию, использовавшуюся в работе Кимуры и Маруямы [10].

Функция, определяющая зависимость приспособленности от числа вредных мутаций, которую использует Чарльзворт, имеет вид $w_n = e^{-(\alpha n + \frac{1}{2}\beta n^2)}$. Параметр β отвечает за существование и тип эпистатических взаимодействий: при $\beta = 0$ эпистаз отсутствует (мультипликативный отбор), синергический эпистаз наблюдается, если $\beta > 0$, а антагонистический, если $\beta < 0$.

Результаты, полученные для бесполой популяции на основе таким образом определенной функции приспособленности, согласуются с результатами Кимуры и Маруямы: в случае бесполого размножения средняя приспособленность популяции зависит только от скорости вредных мутаций [12]. На среднюю приспособленность бесполой популяции не влияют ни существование, ни тип эпистатических взаимодействий, ни абсолютные значения параметров отбора (α и β). Ситуация в случае полового размножения обстоит иначе: также в соответствии с предыдущими результатами [10,11] в работе Чарльзворта показано, что средняя приспособленность половой популяции может быть значительно выше, чем у бесполой ($\overline{w} = e^{-U}$), если существуют синергические взаимодействия между вредными мутациями, и

значительно ниже, чем у бесполой, если существуют антагонистические взаимодействия [12]. Таким образом, синергический эпистаз может давать половому размножению преимущество перед бесполым (в случае, если скорость возникновения вредных мутаций на геном достаточно высока). Кроме того, в отличие от ситуации бесполого размножения, средняя приспособленность половой популяции находится в зависимости от значений параметров отбора: с одной стороны, при фиксированном α приспособленность растет с ростом β (т.е. с увеличением силы синергических взаимодействий), с другой стороны, при фиксированном положительном β приспособленность падает с ростом α (т.е. с увеличением «вредности» мутаций). Таким образом, в случае синергического эпистаза ситуация отличается от того, что наблюдается в случае независимых эффектов мутаций на приспособленность, в котором коэффициент отбора не оказывает влияния на равновесную приспособленность [10,12]. Чарльзворт также исследовал отбор на модификаторы скорости рекомбинации и показал, что при малых положительных значениях β средняя равновесная приспособленность популяции растет с ростом частоты рекомбинации и с увеличением числа хромосом, в то время как в случае существования антагонистического эпистаза рекомбинация вредна и отбор будет действовать против модификаторов, повышающих частоту рекомбинации [12].

Помимо нормального полового размножения, включающего рекомбинацию, Чарльзворт рассмотрел специальный случай полового размножения, при котором происходит сегрегация, но не происходит рекомбинации. Интересно, что даже в отсутствие рекомбинации половое размножение при условии синергического эпистаза получает преимущество перед бесполым [12]. Добавление в эту систему рекомбинации приводит к дальнейшему повышению средней равновесной приспособленности, однако прирост приспособленности оказывается не таким значительным, как при переходе от бесполого размножения к половому без рекомбинации [12].

Чарльзворт исследовал распределение мутационной нагрузки в популяции, опираясь на результат из [49]: дисперсия мутационной нагрузки (σ^2) в популяции в начале каждого поколения может быть представлена как сумма аддитивной дисперсии (V_A) и остаточного члена (C_L) [14]. Аддитивная дисперсия соответствует вкладу в дисперсию отдельных геномных локусов и может быть рассчитана как сумма дисперсий для отдельных локусов. Остаточный член отражает зависимость между аллелями в различных локусах, т.е. неравновесие по сцеплению [49]. В случае независимости аллелей в разных локусах друг от друга, т.е. если неравновесие по сцеплению отсутствует ($C_L = 0$), и если частоты вредных производных аллелей низкие, ожидается, что распределение числа вредных мутаций на геном будет иметь форму распределения Пуассона. Дисперсия распределения числа вредных мутаций на геном (σ^2) в таком случае будет равна среднему и равна аддитивной дисперсии (V_A) [12]. Чарльзворт показывает, что при синергическом эпистазе ($\beta > 0$) дисперсия распределения числа вредных

мутаций на геном становится меньше аддитивной дисперсии ($\sigma^2 < V_A$), поскольку C_L принимает отрицательные значения. В то же время при антагонистическом эпистазе ($\beta < 0$) дисперсия распределения числа вредных мутаций на геном выше аддитивной дисперсии ($\sigma^2 > V_A$) [12].

Наглядное объяснение того, почему половое размножение и рекомбинация при условии синергического эпистаза снижают мутационный груз и делают отрицательный отбор более эффективным, приводится в обзорной статье Кондрашова 1988 года [13]. В случае бесполого размножения для элиминации одной мутации должна произойти одна «генетическая смерть» [42], что при достаточно высокой геномной скорости вредных мутаций приводит к тому, что выживает только маленькая часть популяции (большой мутационный груз). Однако если происходит рекомбинация, а вредные мутации взаимодействуют синергически, мутационный груз может быть значительно меньше, чем в случае бесполого размножения, поскольку в таком случае восстановить среднюю приспособленность можно за счет меньшего числа генетических смертей. В результате отбора с синергическим эпистазом происходит снижение дисперсии (σ^2) распределения мутационной нагрузки [13]. Однако эффективность отбора в следующем поколении зависит от стандартного отклонения этого распределения, σ [13]. Рекомбинация позволяет заново «расширить» это распределение, т.е. увеличить его дисперсию. Поскольку доля генетических смертей тем меньше, чем больше σ, то рекомбинация в случае синергического эпистаза обеспечивает более низкий мутационный груз по сравнению с бесполым размножением [13]. Синергический эпистаз при бесполом размножении не позволяет понизить мутационный груз, поскольку отсутствуют факторы, позволяющие восстановить дисперсию распределения мутационной нагрузки. В то же время рекомбинация в отсутствие эпистатических взаимодействий не оказывает эффекта на мутационный груз, поскольку не существует неравновесия по сцеплению, которое могла бы разрушить рекомбинация, т.к. мутации в разных локусах уже независимы друг от друга.

Редфилд в работе 1988 года исследовала влияние трансформации на равновесную приспособленность бактериальных популяций в компьютерных симуляциях [50]. В этой работе было показано, что в том случае, если источником ДНК для трансформации являются живые бактерии, трансформация имеет эффект схожий с обычным половым размножением [50]. В отсутствие эпистаза трансформация не оказывает влияния на равновесную приспособленность бактериальной популяции. При синергических эпистатических взаимодействиях равновесная приспособленность бактериальной популяции, в которой происходит трансформация, становится выше, а при антагонистических ниже, чем у нерекомбинирующей популяции. Повышение приспособленности от трансформации при синергическом эпистазе (или понижение приспособленности при антагонистическом эпистазе) по сравнению с бесполой ситуацией тем сильнее, чем сильнее выражены эпистатические взаимодействия. Кроме того,

как и при половом размножении, эффект от трансформации увеличивается с ростом геномной скорости вредных мутаций.

Другой вопрос, который исследует Редфилд, заключается в том, как влияет на приспособленность бактериальной популяции трансформация, источником ДНК для которой являются погибшие бактериальные клетки [50]? Поскольку геномы погибших бактерий будут в среднем содержать больше вредных мутаций, чем геномы живых бактерий, ожидается, что трансформация с ДНК из погибших бактерий будет приводить к увеличению мутационной нагрузки [50]. Редфилд показывает, что в присутствии синергического эпистаза даже трансформация с ДНК из погибших бактерий может увеличивать равновесную приспособленность популяции и быть выгодной, в том случае если интенсивность трансформации достаточно низкая. Однако если интенсивность трансформации становится выше определенного порога (зависящего от силы эпистатических взаимодействий), приспособленность популяции опускается ниже уровня бесполой популяции. Кроме того, при прочих равных равновесная приспособленность популяции, в которой происходит трансформация с ДНК из живых бактерий, будет выше, чем у популяции, в которой происходит трансформация с ДНК из погибших бактерий [50].

Редфилд также приводит результаты для сценария «регулируемой» трансформации, в котором трансформация происходит только, если клетка несет не менее чем какое-то пороговое число мутаций (k). Интересно, что в том случае, если трансформация регулируется и трансформируются только клетки, несущие хотя бы одну мутацию ($k \geq 1$), а клетки свободные от мутаций не подвергаются трансформации, равновесная приспособленность такой популяции при любом сценарии выше, чем приспособленность строго бесполой популяции. Это справедливо при любом типе отбора, даже в присутствии антагонистических эпистатических взаимодействий [50].

Еще в конце 1980-х годов были получены оценки скорости вредных мутаций на геном на поколение, U, указывающие на то, что у многих организмов U, по всей видимости, больше 1 [13]. По оценке, сделанной в работе 1999 года и основанной на данных межвидовой дивергенции у человекообразных, у человека $U \geq 1.6$ [41]. Последние оценки скорости мутирования, произведенные на основе поиска *de novo* мутаций в тройках геномов родители–ребенок, говорят о том, что U у человека может быть больше 7 [51,52]. В отсутствие синергических эпистатических взаимодействий мутационный груз, соответствующий этим оценкам, является плохо совместимым с существованием популяции человека [13]. По всей видимости, парадокс мутационного груза может быть объяснен за счет синергических эпистатических взаимодействий [13]. При такой геномной скорости вредных мутаций в случае синергического эпистаза половое размножение получает преимущество перед бесполым [13],

однако до последнего времени не было работ, в которых бы приводились прямые свидетельства в пользу существования распространенных синергических взаимодействий между вредными мутациями.

1.3 Другие гипотезы, объясняющие преимущество полового размножения

Половое размножение может получать преимущество перед бесполым и в отсутствие синергических эпистатических взаимодействий из-за эффектов, которые может создавать генетический дрейф [22,47]. Одним из гипотетических результатов действия генетического дрейфа в бесполой популяции ограниченного размера может быть необратимая потеря класса генотипов свободных от мутаций – так называемый храповик Мёллера [16,17]. Если в популяции из-за генетического дрейфа теряется класс генотипов свободных от вредных мутаций, в бесполой популяции не существует механизмов (кроме обратных мутаций), с помощью которых этот класс генотипов мог бы быть восстановлен. Таким образом, потеря «лучшего класса» генотипов в бесполой популяции является необратимой, что соответствует одному повороту «храповика». Следующий поворот «храповика» может привести к потере следующего класса генотипов с минимальным количеством вредных мутаций среди оставшихся. Эта идея была выдвинута Германом Мёллером, предположившим, что бесполые популяции должны быть уязвимы к необратимому накоплению вредных мутаций из-за необратимости «храповика» [16,17]. В то же время в половой популяции класс генотипов свободных от вредных мутаций, потерянный из-за дрейфа генов, может быть восстановлен за счет рекомбинации, что может в долгосрочной перспективе давать преимущество половым популяциям перед бесполыми. Однако другие исследователи приводили аргументы в пользу того, что необратимая потеря свободных от мутаций генотипов должна представлять значительную угрозу только для бесполых популяций маленького размера, поскольку в достаточно большой популяции вероятность потери этого класса генотипов очень низкая [13,45]. Насколько распространено действие храповика Мёллера в бесполых популяциях и насколько сильно из-за этого механизма снижается их приспособленность, в настоящий момент до конца непонятно [46].

Кроме того, преимущество полового размножения может заключаться в более высокой эффективности положительного отбора и соответственно более высокой скорости адаптивной эволюции в половых популяциях. Здесь можно выделить две основные гипотезы. Первая гипотеза рассматривает явление клональной интерференции как процесс, замедляющий скорость адаптации, в бесполых популяциях [18,19]. Клональная интерференция происходит в

том случае, если в разных клонах в разных локусах в течение короткого промежутка времени происходят полезные мутации, частота которых начинает повышаться под действием положительного отбора. Однако поскольку в популяции одновременно присутствуют разные положительно отбираемые клоны, соответствующие разным мутациям, между ними происходит «конкуренция» за закрепление [18,19]. В результате часть полезных мутаций может быть потеряна. В половой популяции две возникшие в разных генотипах полезные мутации могут быть «собраны» за счет рекомбинации в одном геноме и одновременно закрепиться. Та же идея о том, что в бесполой популяции скорость адаптации ограничена из-за того, что полезные мутации остаются сцепленными с геномным контекстом, в котором они произошли, лежит и в основе второй гипотезы [20,21]. В данной гипотезе в качестве основного фактора, ограничивающего положительный отбор в бесполых популяциях, рассматривается сцепление полезных мутаций с вредными мутациями, которое в отсутствие рекомбинации не может быть разрушено. Если полезная мутация произошла в генотипе, в котором присутствуют вредные мутации, отрицательный отбор, действующий против этих вредных мутаций, будет препятствовать закреплению полезного варианта (фоновый отбор) [20,21]. В то же время в половой популяции рекомбинация позволяет разрушить подобное сцепление, что дает полезным мутациям возможность закрепиться вне зависимости от генотипа, в котором они изначально произошли.

Как отмечается в работах, посвященных классификации гипотез, объясняющих преобладание полового размножения, преимущество рекомбинации в случае действия храповика Мёллера, клональной интерференции, и фонового отбора связано с присутствием того же фактора, который создает преимущество для полового размножения в рамках гипотезы о синергических эпистатических взаимодействиях [22,47]. Этот фактор – это отрицательное неравновесие по сцеплению между мутациями, которое рекомбинация каждое поколение заново разрушает, расширяя распределение числа мутаций на геном [22,47]. Отрицательное неравновесие по сцеплению и сужение распределения числа мутаций на геном может создаваться в результате отбора за счет эпистатических взаимодействий [12] или за счет дрейфа генов [22,47]. Рекомбинация позволяет восстановить или создать недопредставленные генотипы – например, генотипы свободные от вредных мутаций (храповик Мёллера) [16,17] или генотипы, несущие несколько полезных мутаций (клональная интерференция) [18,19]. В том случае, если в популяции отсутствует неравновесие по сцеплению (отсутствуют ассоциации между аллелями), рекомбинация не может изменить дисперсию распределения числа мутаций на геном в популяции [22].

Отдельная группа гипотез объясняет преобладание полового размножения тем, что оно дает значительное преимущество в условиях частых изменений среды, то есть в ситуации,

когда часто меняется ландшафт приспособленности. Например, если гаплотип, который имел низкую приспособленность на протяжении многих поколений, в какой-то момент оказывается гаплотипом, обеспечивающим максимальную приспособленность, его частота за счет рекомбинации в начале следующего поколения в половой популяции будет больше, чем в бесполой [22]. В результате половая популяция сможет более эффективно ответить на изменение отбора. Более подробно подобные сценарии, а также другие гипотезы происхождения и поддержания полового размножения у эукариот освещены, например, в обзорной статье Кондрашова 1993 года [22].

Вне зависимости от конкретных причин преобладания полового размножения среди эукариот, оно должно давать значительные преимущества, поскольку отказ от полового размножения обычно приводит к быстрому вымиранию [23]. В этом контексте особенно интересны редкие исключения из этого правила – группы древних бесполых организмов, существование которых многими биологами рассматривается как «эволюционный скандал» [53].

1.4 «Эволюционные скандалы» – предположительно древние группы бесполых организмов

В качестве одного из главных аргументов в пользу необходимости полового размножения для эволюционного успеха вида обычно рассматривают распределение бесполых таксонов на филогенетических деревьях [45,54–56]. Несмотря на то, что переходы к бесполому размножению случались многократно в эволюции эукариот, бесполые группы обычно включают небольшое количество видов и располагаются близко к концу веток филогенетических деревьев, то есть имеют небольшой эволюционный возраст [45,54–56]. По всей видимости, это указывает на то, что переход к половому размножению приводит к быстрому вымиранию и является тупиковой эволюционной стратегией [45,54–56].

Однако существует небольшое число возможных исключений из этого правила – это группы предположительно древних бесполых организмов. Их существование противоречит гипотезе о том, что половое размножение необходимо для долгосрочного выживания. На этом основании Мэйнард Смит назвал факт существования одной из таких групп «эволюционным скандалом» [53]. В качестве наиболее ярких примеров предположительно древних групп партеногенетических организмов обычно приводят дарвинулид (семейство *Darwinulidae* из класса *Ostracoda* подтипа *Crustacea*, тип *Arthropoda*) [57,58], часть видов орибатидных клещей (отряд *Oribatida* подкласса *Acari* класса *Arachnida*, тип *Arthropoda*) [59,60] и бделлоидных

коловраток (класс *Bdelloidea*, тип *Rotifera*) [24,25,61–63]. Считается, что представители этих трех групп полностью отказались от полового размножения и перешли к облигатному партеногенезу миллионы лет назад [64,65]. Если это действительно так, то изучение стратегий, которые используются этими организмами для того, чтобы противостоять последствиям отказа от полового размножения, было бы чрезвычайно интересно для понимания эволюции полового размножения и причин, по которым большинство животных не могут длительное время размножаться исключительно партеногенетически [64].

Однако вопрос о том, является ли партеногенез единственной формой размножения в этих трех и других предположительно древних группах партеногенетических организмов, остается открытым [64,66]. Основным аргументом в пользу того, что группа организмов размножается исключительно партеногенетически, обычно является отсутствие самцов среди большого количества проанализированных особей из естественных популяций. Палеонтологические данные могут позволить примерно оценить временной промежуток, в который произошел предполагаемый отказ от полового размножения [57]. В то же время данные об отсутствии самцов среди больших экспериментальных выборок не обязательно указывают на полный отказ от обмена генетическим материалом [24,28]. В случае очень редкого полового размножения или полового размножения, происходящего только в определенные промежутки времени или при определенных условиях, самцы могут не попасть в экспериментальную выборку [24,28]. Кроме того, даже если половое размножение в классическом понимании не происходит, в популяции могут существовать другие формы обмена генетическим материалом, например, горизонтальный перенос генов [67]. В связи с этим для того, чтобы убедительно доказать строгую клональность популяции, в дополнение к экспериментальным данным необходимо получить популяционно-генетические данные [68].

Исключительно клональный способ наследования ДНК должен оставлять несколько разных типов геномных «подписей» в популяционно-генетических данных [68]. Соответствующие подписи ожидаются как при анализе индивидуальных однонуклеотидных сайтов генома, так и при анализе разных локусов.

Так, полный отказ от полового размножения должен приводить к отклонениям частот генотипов от частот, ожидаемых при равновесии Харди-Вайнберга [69]. Оценить, находится ли популяция в равновесии Харди-Вайнберга, можно с помощью коэффициента инбридинга F_{IS}. Ожидаемое среднее значение F_{IS} в равновесии в половой популяции – 0, положительные значения F_{IS} соответствуют случаю избытка гомозиготных генотипов, отрицательные значения – случаю избытка гетерозиготных генотипов. Поскольку при бесполом размножении два гаплотипа накапливают замены независимо друг от друга, в строго клональной популяции ожидается значительный избыток гетерозигот. Так, в теоретической работе Balloux *et al*. 2003

года аналитически и в симуляциях было показано, что ожидаемое значение коэффициента инбридинга F_{IS} в строго клональной популяции близко к −1 [69].

В том случае, если популяция строго клональная и отсутствуют какие-либо формы рекомбинации, все сайты генома оказываются сцепленными друг с другом [68,70,71]. Это означает, что все полиморфные сайты должны находиться в неравновесии по сцеплению и степень неравновесия по сцеплению не должна зависеть от расстояния между сайтами. Стоит отметить, что сигнал распада неравновесия по сцеплению с расстоянием может быть связан не только с реципрокной мейотической рекомбинацией, но и с другими формами рекомбинации, такими, как генная конверсия [72,73] и митотическая рекомбинация [74,75].

Еще один подход к поиску возможных «подписей» бесполого размножения основывается на ожидаемом при строгой клональности паттерне кластеризации гаплотипов. Впервые эта идея была выдвинута М. Мезельсоном, в связи с чем ожидаемый паттерн обычно называют «эффект Мезельсона» [76]. Эффект Мезельсона отражает независимую эволюцию двух гаплотипов, присутствующих в одной особи, при бесполом размножении в отсутствии рекомбинации. Предположим, что некоторое множество бесполых индивидуумов ведут происхождение от особи, в которой в момент перехода к бесполому размножению присутствовали два гаплотипа А и Б. После отказа от полового размножения и рекомбинации гаплотипы А и Б начинают независимо накапливать замены, что приводит к постепенному увеличению дивергенции между ними и соответственно высокому уровню гетерозиготности внутри особи [76]. Это должно приводить к существованию в популяции двух клад гаплотипов, соответствующих первоначальным гаплотипам А и Б. При этом генетическое расстояние между гаплотипами из разных индивидуумов, принадлежащими к одной и той же кладе, должно быть меньше, чем расстояние между двумя гаплотипами внутри одного индивидуума [76]. Стоит отметить, что процессы, обеспечивающие рекомбинацию между гаплотипами в отсутствие обмена генетическим материалом между индивидуумами (генная конверсия или митотическая рекомбинация), могут приводить к «перемешиванию» между гаплотипами из разных клад и «размыванию» клад [77,78]. В связи с этим отсутствие эффекта Мезельсона может наблюдаться и в бесполой популяции и не является доказательством обмена генетическим материалом между индивидуумами [77,78].

Подтвердить теоретические предсказания отсутствия полового размножения удалось только для небольшого количества организмов, которых длительное время считали бесполыми. Одним из примеров видов, для которых отсутствие полового размножения было убедительно показано с помощью популяционных данных, является паразит человека *Trypanosoma brucei gambiense*, вызывающий сонную болезнь [79]. Секвенирование большого количества изолятов *T. brucei gambiense* из трех стран Африки выявило, что популяция *T. brucei gambiense*

эволюционирует в соответствии с ожиданиями строгой клональности. Так, в популяции этого вида наблюдается избыток гетерозигот: медианное значение F_{IS} для полиморфных локусов равно −1. Отсутствие рекомбинации выражается в том, что полиморфные сайты находятся в неравновесии по сцеплению и весь геном представляет собой единую группу сцепления. Кроме того, в данной работе на уровне всего генома был продемонстрирован эффект Мезельсона: филогении, построенные на основе гаплотипов секвенированных индивидуумов *T. brucei gambiense*, указывают на независимую эволюцию двух гаплотипов одного индивидуума. Гаплотипы секвенированных изолятов распадаются на два независимо эволюционирующих кластера так, что каждый изолят несет по одному гаплотипу из обоих кластеров. Более того, филогенетическая история гаплотипов из первого кластера соответствует филогенетической истории гаплотипов из второго кластера, что отражает их параллельную эволюцию [79]. Интересно, что возраст бесполой линии *T. brucei gambiense*, к которой относятся секвенированные изоляты, составляет менее 10,000 лет [79].

Паттерны, совместимые с эффектом Мезельсона, были выявлены в работе 2011 года и для некоторых насекомых из рода *Timema* [80]. Род *Timema* включает как виды, размножающиеся половым путем, так и виды, представляющие собой достаточно древние бесполые линии. Переходы к бесполому размножению в роде *Timema* случались независимо по крайней мере 5 раз, а возраст бесполых линий значительно отличается: самая «молодая» среди проанализированных в данной работе линия перешла к партеногенезу около 500,000 лет назад, а самая «старая» – почти 2,000,000 лет назад [80]. В соответствии с тем, что ожидается в результате длительной эволюции без рекомбинации, степень дивергенции аллелей внутри одного индивидуума значительно выше в партеногенетических линиях *Timema* по сравнению с родственными видами, у которых сохранилось половое размножение [80]. Однако четкое разделение гаплотипов на две независимо эволюционирующие клады (эффект Мезельсона) у насекомых из рода *Timema* было показано только для части индивидуумов и только для двух из трех предположительно «старых» партеногенетических линий [80].

Относительно древние бесполые линии встречаются и среди позвоночных. Примером такой древней линии у рыб является амазонская моллинезия *Poecilia formosa*, являющаяся гибридом двух достаточно далеких видов из рода *Poecilia* с нормальным половым размножением [81]. Гибридизация, которая привела к появлению *P. formosa*, вида, размножающегося исключительно путем гиногенеза, произошла не менее чем 100,000 лет назад. Данные популяционного секвенирования выявили высокий уровень гетерозиготности (превышающий гетерозиготность в «родительских» видах в ~10 раз) и существование двух выраженных клад гаплотипов [81]. Однако в данном случае высокая гетерозиготность и присутствие в каждом индивидууме гаплотипов из двух разных клад являются не проявлениями

эффекта Мезельсона, а в первую очередь результатом гибридизации: гаплотипы в каждой из двух клад происходят из гаплотипов соответствующих «родительских» видов и кластеризуются с ними [81].

Ситуация с «бесполым статусом» трех наиболее древних групп предположительно бесполых видов – дарвинулид, орибатидных клещей и бделлоидных коловраток – значительно более запутанная. Долгое время считалось, что дарвинулиды отказались от полового размножения более 200 миллионов лет назад [57]. Это заключение было сделано на основе палеонтологических данных: раковины, оставшиеся от предположительных самцов дарвинулид, перестают встречаться в позднем триасе [57]. Но в 2006 году самцы были найдены в трех естественных популяциях одного из видов дарвинулид, принадлежащего к роду *Vestalenula* [65]. Всего среди нескольких сотен проанализированных особей было найдено три самца. Интерпретация этой находки не является однозначной: согласно одной из точек зрения найденные самцы могут являться «нефункциональным атавизмом» [82,83]. Интересно, что эффект Мезельсона в данных по дивергенции между аллеллями для трех ядерных локусов у дарвинулиды *Darwinula stevensoni* выявлен не был [58]. Таким образом, молекулярные данные тоже не позволили получить ответ на вопрос о существовании полового размножения у дарвинулид [82]. Эффект Мезельсона отсутствует и у предположительно древних бесполых линий орибатидных клещей: по крайней мере на это указывает анализ последовательностей двух генов (*ef-1alpha* и *hsp82*) [77]. Тем не менее в соответствии с тем, что ожидается в случае бесполого размножения, анализ последовательностей гена *ef-1alpha* не выявил подписей рекомбинации и генной конверсии у этих предположительно древних партеногенетических линий [77]. В то же время подписи этих двух процессов были найдены у тех видов орибатидных клещей, у которых сохранилось обычное частое половое размножение [77]. Считается, что ~10% видов орибатидных клещей размножаются партеногенетически [83,84]. Даже в этих вероятно партеногенетических видах орибатидных клещей встречаются самцы, однако было показано, что эти самцы являются стерильными, а сперматогенез не проходит до конца [83,85].

Особенно интересен вопрос о том, есть ли какие-то формы обмена генетическим материалом у бделлоидных коловраток (лат. *Bdelloidea*) – группы коловраток (лат. *Rotifera*), предположительно перешедшей к бесполому размножению по меньшей мере 35 миллионов лет назад [61]. Партеногенетическое развитие из диплоидных яиц, образующихся в результате митотических делений, − единственный описанный на сегодняшний день способ размножения бделлоидных коловраток [86]. Ни мейоз, ни половое размножение у бделлоидных коловраток не известны. Бесполый статус класса *Bdelloidea*, в котором описано более 460 видов [87], основывается в первую очередь на том, что за более чем вековую историю его изучения, за

исключением одного случая, исследователи не обнаружили в популяциях бделлоидных коловраток ни самцов, ни гермафродитов [24]. Единственное наблюдение гипотетического самца бделлоидной коловратки относится к 1930 году: Вайзенберг-Лунд пишет, что видел вероятного самца бделлоидной коловратки среди тысяч самок бделлоидных коловраток из семейства *Philodinidae* [88]. Однако из-за того, что самец передвигался с большой скоростью, Вайзенберг-Лунд не смог ни поймать, ни зарисовать его [88]. Морфологически этот предполагаемый самец *Bdelloidea* напоминал самца моногононтной коловратки (класс коловраток, для которых характерно чередование партеногенеза и эпизодов полового размножения [89]). Впоследствии было сделано предположение, что самец, описанный в монографии Вайзенберг-Лунда, мог быть самцом моногононтной коловратки, оказавшимся в одном образце с большим количеством бделлоидных [24]. Такой скептицизм связан в первую очередь с тем, что среди сотен тысяч особей *Bdelloidea*, проанализированных разными исследователями, других самцов зафиксировано не было [24]. Так, Вильям Бирки оценил, что разные исследователи в совокупности визуально проанализировали не менее 458,515 особей бделлоидных коловраток [24], среди которых не было найдено ни одного самца или гермафродита. На этом основании Бирки делает заключение, что в том случае, если самцы бделлоидных коловраток все-таки существуют, их частота в популяции не должна превышать 8.1×10^{-6} [24]. Однако это заключение сделано в предположении того, что определить особь как самца бделлоидной коловратки в случае существования самцов не составило бы труда.

Помимо эмпирических наблюдений об отсутствии самцов в популяциях бделлоидных коловраток, приводились и другие данные, указывающие на то, что облигатный партеногенез является наиболее вероятным способом размножения организмов из этой группы. Так, опубликованный в 2013 году биоинформатический анализ генома бделлоидной коловратки *Adineta vaga* показал, что геном *A. vaga* обладает рядом характеристик, несовместимых с классическим мейозом [25]. В частности, данные, приведенные в этой работе, указывали на то, что в геноме *A. vaga* отсутствуют гомологичные хромосомы. Модель устройства генома *A. vaga*, предложенная в этой работе, такова: на уровне генов геном *A. vaga* является диплоидным и большинство генов в геноме *A. vaga* представлены двумя аллелями, однако гены перетасованы между хромосомами таким образом, что не существует двух хромосом с одинаковым набором и порядком генов. Так, аллели генов, находящихся рядом на одной хромосоме, могут быть разбросаны по нескольким хромосомам [25]. Более того, в геноме *A. vaga* были найдены случаи, в которых гомологичные блоки, включающие значительное количество аллельных генов, находятся на одной хромосоме и организованы как палиндромы [25]. Максимальная длина такого палиндрома составила 705 тысяч нуклеотидов [25].

Первоначально эти данные рассматривали как доказательство того, что устройство генома *A. vaga* несовместимо с классическим мейозом.

Однако в 2018 году был опубликован анализ генома близкого вида из того же рода – *Adineta ricciae*, в котором не было найдено геномных перестроек, несовместимых с мейозом: в частности в геноме *A. ricciae* не было найдено ни одного палиндрома [26]. В этой же работе были опубликованы геномы двух видов бделлоидных коловраток из другого рода – *Rotaria* [26]. Интересно, что уровень гетерозиготности, характерный для геномов бделлоидных коловраток из родов *Adineta* и *Rotaria*, очень сильно отличается: в то время как в геномах коловраток из рода *Adineta* была выявлена высокая гетерозиготность (1.42% – *A. vaga*, 4.55% – *A. ricciae*), гетерозиготность в роде *Rotaria* ниже более чем на порядок (0.026% – *R. macrura* и 0.104% – *R. magnacalcarata*) [26]. Таким образом, геномные данные не позволяли сделать убедительное заключение о способе размножения бделлоидных коловраток.

Отдельно стоит отметить, что в 2021 году вышла работа [27], авторы которой с использованием комбинации различных технологий секвенирования (включая данные Hi-C) заново собрали геном линии *A. vaga*, впервые опубликованный в 2013 году. Сочетание разных типов данных позволило собрать геном до уровня хромосом. Анализ новых данных позволил заключить, что значительное количество геномных перестроек, выявленных в работе 2013 года, и отсутствие гомологичных хромосом в первой сборке, по всей видимости, объясняется техническими артефактами [27]. Структура генома *A. vaga* согласно новой модели полностью совместима с классическим мейозом: геном *A. vaga* состоит из 6 пар гомологичных хромосом и не содержит каких-либо значительных перестроек, которые могли бы затруднить протекание мейоза [27]. Таким образом, первоначальное заключение о несовместимости генома *A. vaga* с классическим мейозом было сделано на основании ошибочных данных.

Первые данные по популяционной изменчивости у бделлоидных коловраток были опубликованы в работе 2015 года, в которой секвенировали несколько участков ядерного генома у 6 особей вида *Macrotrachela quadricornifera* [28]. Паттерны, найденные в этой работе в филогениях, построенных на основе последовательностей гаплотипов, указывали на вероятные события обмена генетическим материалом [28]. Эти данные были интерпретированы как вероятное свидетельство чрезвычайно редкого типа мейоза, описанного у растений из рода *Oenothera* [28]. При таком атипичном мейозе при образовании гамет большая часть генома не вовлечена в рекомбинацию, затрагивающую только концы хромосом [28]. То есть в такой модели происходит сегрегация, но практически не происходит рекомбинации [28]. Однако в работе 2016 года, в которой были секвенированы последовательности пяти маркеров у более чем 500 бделлоидных коловраток из вида *A. vaga*, были получены данные, плохо с согласующиеся как с моделью атипичного мейоза, так и с моделью строго бесполого

размножения, но совместимые со сценарием, при котором коловратки обмениваются генетическим материалом внутри популяции за счет горизонтального переноса генов [90]. Эти данные были затем проанализированы более аккуратно другими авторами в статье 2018 года, в которой было показано, что случаи, изначально интерпретированные как свидетельство горизонтального переноса генов между бделлоидными коловратками, являются результатом экспериментальной ошибки [91]. Таким образом, до последнего времени консенсус по поводу присутствия обмена генетическим материалом у бделлоидных коловраток достигнут не был [28,90–92].

Глава 2. Сигнал действия отрицательного отбора на ортологичные интроны в далеких видах

Отбор на сохранение функции гомологичных участков генома может сопровождаться сохранением очевидного сходства на уровне последовательностей, которое может поддерживаться в течение миллиардов лет [93]. Так, например, последовательность многих бактериальных белков на 50% и более идентична последовательности их эукариотических ортологов. Более того, анализ реконструированного генома последнего общего предка всех живых организмов (англ. Last Universal Common Ancestor, сокращенно LUCA) выявил гены, возникшие в результате дупликации еще в линии предковой LUCA. Сходство последовательностей таких паралогичных белков до сих пор статистически значимо [93]. Даже в некодирующих сегментах генома присутствуют ультраконсервативные участки, сходство последовательностей между которыми сохраняется, например, среди всех позвоночных [3,4].

Однако сохранение первичной последовательности, на основании которой можно реконструировать осмысленное выравнивание, не является обязательным условием для сохранения других свойств молекулы. Так, белки, последовательности которых не могут быть выровнены, тем не менее могут иметь похожие трехмерные структуры [36]. Одноцепочечные РНК, имеющие несхожие последовательности, могут обладать одинаковой вторичной структурой [37] (в качестве очевидного примера можно привести последовательности AAAAAGGGTTTTT и GGGGGTTTCCCCC). Для функциональных некодирующих последовательностей ДНК также описан ряд случаев, в которых некодирующие последовательности из разных организмов могут выполнять похожие функции и, скорее всего, имеют общее происхождение, несмотря на отсутствие между ними осмысленного выравнивания [8]. Так, например, энхансер гена человека может обеспечивать нормальную экспрессию ортологичного гена в трансгенных *D. rerio*, притом что сходство последовательностей между энхансерами человека и *D. rerio* отсутствует [9]. В другом исследовании, проведенном для *D. rerio*, было показано, что использование существовавших в тот момент методов построения выравниваний в значительном проценте случаев не позволяет выявить функциональные последовательности [38]. Некодирующие элементы, регулирующие гомологичные гены в нематодах и позвоночных, обладают сходством последовательностей внутри обеих клад, но не между организмами из разных клад, что подразумевает «переключение» регуляторных механизмов [39]. В связи с этим интересным представляется вопрос о том, насколько феномен вероятного сохранения функции участка генома и действия отрицательного отбора на него без сохранения первичной последовательности распространен на уровне генома. Для того, чтобы изучить возможное существование этого явления на уровне

всего генома, мы проанализировали гомологичные интроны в двух четверках видов – в четверке видов насекомых и в четверке видов позвоночных. Каждая четверка состоит из двух пар геномов, подобранных так, что геномы внутри пары находятся на умеренно большом филогенетическом расстоянии друг от друга, а расстояние, разделяющее разные пары, значительно превышает расстояния внутри пар. Мы показали, что в обеих четверках видов интроны, несущие регуляторный или консервативный элемент в первой паре, с повышенной частотой несут консервативный сегмент во второй паре, несмотря на то, что сходство последовательностей между интронами из разных пар отсутствует. Более того, интроны из одной пары, сохранившиеся в ходе эволюции во второй паре, чаще несут консервативный сегмент в первой паре, несмотря на то, что сходство последовательностей между такими сохранившимися в обеих парах интронами отсутствует. Результаты, представленные в данной главе, указывают на то, что давление отбора, вероятно, связанного с сохранением предковой функции, часто продолжает существовать даже после того, как гомологичные сегменты ДНК полностью теряют сходство на уровне последовательностей.

2.1 Материалы и методы

2.1.1 Геномные данные

Мы провели анализ для двух четверок видов, рассматривая их совместно с соответствующими видами-аутгруппами (Рисунок 2.1). В каждой четверке число синонимических замен на сайт (стандартное обозначение Ks) между видами из разных пар превышает 1 и в связи с этим не может быть оценено точно.

Приблизительные значения Ks для видов из разных пар в четверке позвоночных были оценены в опубликованной работе [94]. Приблизительные значения Ks для четверки двукрылых, показанные на Рисунке 2.1, были получены следующим образом. Для того, чтобы приблизительно оценить Ks между парами, мы шкалировали расстояния на филогенетическом дереве, построенном на основе белковых последовательностей [95], в соответствии с известным значением Ks [96], полученным для более близких видов, относящихся к паре 1.

Первая четверка – четверка двукрылых – включает в себя два вида из рода *Drosophila, D. melanogaster* и *D. mojavensis* (пара 1: $Ks_1 \sim 2.37$) [96], и два вида комаров, *Culex quinquefasciatus* и *Aedes aegypti* (пара 2: $Ks_2 \sim 2.6$); приблизительная оценка числа синонимических замен на сайт между видами из разных пар, $Ks_3 \sim 6.5$. Для того, чтобы оценить ожидаемые значения Ks между *C. quinquefasciatus* и *Aed. aegypti,* а также между видами *Drosophila* и комарами, мы откалибровали филогенетические расстояния на опубликованном

дереве членистоногих, построенном на основе конкатенированных последовательностей моторных белков [95], используя известное значение Ks в паре *D. melanogaster – D. mojavensis* [96].

Вторая четверка – четверка позвоночных – состоит из двух видов млекопитающих, *Homo sapiens* и *Mus musculus* (пара 1: $Ks_1 \sim 0.43$) [94], и двух видов рыб, *Tetraodon nigroviridis* и *Takifugu rubripes* (пара 2: $Ks_2 \sim 0.35$) [94]. Значение Ks между видами млекопитающих и рыб, Ks_3, примерно равно 1.5 [94]. В случае четверки позвоночных оценки Ks для видов, принадлежащих к одной и той же и к разным парам, были взяты из опубликованной работы [94].

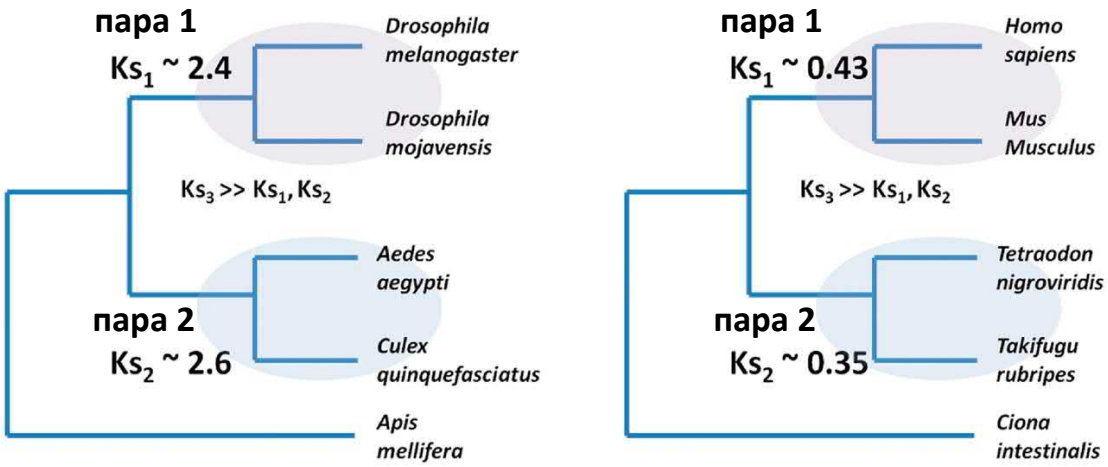

Рисунок 2.1. Две четверки видов, для которых проводился анализ. Четверка двукрылых (слева) и четверка позвоночных (справа). Для каждой четверки отображена соответствующая аутгруппа. Показаны эволюционные расстояния для видов внутри каждой пары, выраженные через ожидаемое число синонимических замен на сайт, Ks.

Списки ортологичных белков для каждой попарной комбинации видов внутри четверки были получены из базы данных INPARANOID [97] (http://inparanoid.sbc.su.se/cgi-bin/index.cgi, дата последнего доступа 27 февраля 2013). Для каждой четверки мы сначала рассматривали все 6 возможных попарных сравнений 4 видов (10 попарных сравнений 5 видов в случае анализа с привлечением вида-аутгруппы). Для каждого попарного сравнения видов мы выделили множество пар однозначно идентифицируемых ортологов, являющихся взаимными лучшими находками друг для друга. Для последующего анализа мы оставили только те группы ортологов, в которых ортологи из 4 (5) видов образовывали клику.

В результате этой процедуры были определены наборы из 5189 (3565 при включении аутгруппы) и 8179 (2522 при включении аутгруппы) однозначно идентифицируемых групп ортологов для четверки двукрылых и позвоночных соответственно. После исключения кодирующих последовательностей, несущих в каком-либо из видов преждевременный стоп-кодон, доступными для анализа остались 5183 (3541) и 8159 (2518) групп ортологов для четверки двукрылых и позвоночных соответственно. Для того, чтобы исключить вероятность неверного определения ортологов, связанного с несоответствиями между аннотациями разных версий геномных сборок, мы использовали те же версии сборок, которые использовались при создании базы данных INPARANOID. Для видов *H. sapiens*, *M. musculus*, *C. intestinalis*, *T. nigroviridis*, *Tak. rubripes,* и *Aed. aegypti* мы использовали версии сборок генома NCBI 36 [98,99], NCBI m37 [100], JGI 2 [101], TETRAODON 8.0 [94], FUGU 4.0 [102], и AaegL1 [103] соответственно. Перечисленные версии геномных сборок входили в состав выпуска 52 базы данных ENSEMBL. Для видов *D. melanogaster* и *D. mojavensis* мы использовали версии геномных сборок r5.13 [104] и r1.3 [105], входящие в состав выпусков ENSEMBL 58 и 63 соответственно. Для *C. quinquefasciatus* мы использовали версию сборки генома CpipJ1.2 [106], а для генома *A. mellifera* [107] – версию NCBI 4.1.

Геномные последовательности и данные по аннотации белок-кодирующих генов для всех видов, за исключением *A. mellifera* и *C. quinquefasciatus,* были загружены из базы данных ENSEMBL [108] с использованием интерфейса ENSEMBL PERL API для языка программирования Perl. Геномные последовательности и данные по аннотации для *A. mellifera* и *C. quinquefasciatus* были загружены из баз данных NCBI (http://www.ncbi.nlm.nih.gov) [109] и VectorBase (http://cquinquefasciatus.vectorbase.org) [110] соответственно.

Выравнивания ортологичных белков из разных видов были выполнены с использованием программы MUSCLE [111] с параметрами по умолчанию.

2.1.2 Определение и анализ ортологичных интронов

Мы выделили множество интронов ортологичных во всех видах, принадлежащих к рассматриваемой четверке, то есть, таких интронов, которые с большой вероятностью имеют общее происхождение. Мы определяли ортологичные интроны как интроны, «попадающие» в ортологичные позиции кодирующих последовательностей в ортологичных генах, и, кроме того, имеющие одинаковую фазу. Для этого координаты интронов были отображены на выравнивания ортологичных белков. Для того, чтобы уменьшить вероятность включения в анализ интронов, ошибочно определенных как ортологичные, мы рассматривали только те интроны, отображение координат которых в белковой последовательности («тень интрона») попадало в область выравнивания с высоким качеством. При отображении координат интронов в белковую последовательность интронам с фазой 0 (т.е. находящимся между двумя кодонами) соответствует две аминокислотных позиции, а интронам с фазами 1 и 2 (разделяющим кодон между первой и второй или второй и третьей позицией) соответствует одна аминокислотная позиция. В качестве первичного набора ортологичных интронов мы выбирали такие интроны из разных видов, отображения координат которых в выравнивании совпадали. При этом мы требовали, чтобы в колонках выравнивания, соответствующих координатам интронов, и двух колонках слева и справа отсутствовали пропуски (гэпы). Кроме того, мы требовали, чтобы в 10 колонках белкового выравнивания, фланкирующих «тень интрона» слева и справа, было как минимум пять колонок, аминокислотные остатки в которых похожи во всех рассматриваемых видах согласно матрице BLOSUM62. Если хотя бы в одном виде в 10 колонках белкового выравнивания, фланкирующих «тень интрона» слева или справа, было более двух пропусков, интрон исключали из рассмотрения. Чтобы сосредоточиться на анализе истинно некодирующих последовательностей, мы также исключали из рассмотрения интроны, пересекающиеся с белок-кодирующими экзонами, входящими в состав альтернативных транскриптов для рассматриваемого гена.

В набор ортологичных интронов, оставшихся после описанных выше шагов фильтрации и использованных для дальнейшего анализа, вошло 5367 и 51,844 групп интронов в четверках двукрылых и позвоночных соответственно.

При поиске консервативных элементов мы не рассматривали первые 6 и последние 16 нуклеотидов интрона, поскольку эти позиции с высокой вероятностью находятся под давлением отрицательного отбора, связанного с присутствием элементов, необходимых для правильного сплайсирования интрона [112,113]. Оставшиеся центральные части ортологичных интронов из четырех видов выравнивались с применением программы bl2seq [114] с длиной якоря 7 нуклеотидов и включенной фильтрацией участков низкой сложности.

Выравнивание строили отдельно для каждого попарного сочетания видов из четверки. Поскольку мы были заинтересованы в случаях, в которых сходство между последовательностями из видов, принадлежащих к разным парам, отсутствует, мы исключали из рассмотрения группы ортологичных интронов со значимым «сквозным» сходством (bl2seq E-value между последовательностями из разных пар видов внутри четверки ≤0.0001).

2.1.3 Оценка ожидаемого числа случаев, в которых интроны в обеих парах внутри четверки несут сегмент сходства

Для каждого порогового значения E-value мы сначала определяли число интронов, $N_1(E)$ и $N_2(E)$, несущих сегмент (участок) локального сходства последовательностей в парах 1 и 2 соответственно. Ожидаемое число интронов, несущих сегмент локального сходства в обеих парах одновременно, было рассчитано с помощью четырех разных типов рандомизации: 1) без учета влияния возможных искажающих факторов, 2) с учетом длин интронов, 3) с учетом идентичности гена, 4) с учетом идентичности гена и того, является ли интрон первым в гене или последующим.

Для проведения первого типа рандомизации мы случайным образом выбирали отдельно $N_1(E)$ и $N_2(E)$ интронов среди всех интронов в первой и второй паре соответственно и определяли число интронов, попавших одновременно в первое и второе множество. Данное число соответствует числу интронов, одновременно несущих в перетасованной выборке «участок сходства» в обеих парах. Данную процедуру повторяли 10,000 раз. Распределения, основанные на 10,000 перетасованных выборок, использовали для того, чтобы получить среднее значение и доверительные интервалы для ожидаемого числа интронов, несущих участок сходства в обеих парах одновременно. Полученная с помощью данной процедуры оценка ожидаемой доли интронов, несущих участок сходства в обеих парах видов одновременно, грубо соответствует оценке, полученной умножением частоты интронов, несущих участок сходства в первой паре, на соответствующую частоту для второй пары.

Однако в описанной выше процедуре не учитывается то, что длины интронов скоррелированы между видами из разных пар (так, например, коэффициент корреляции Спирмена для длин интронов между *D. melanogaster* и *C. quinquefasciatus* равен 0.29, P-значение $< 2.2 \times 10^{-16}$; для *H. sapiens* и *Tak. rubripes* коэффициент корреляции Спирмена равен 0.215, P-значение $< 2.2 \times 10^{-16}$). Поскольку более длинные интроны в целом более консервативны [112] (см. также Рисунок 2.2), повышенное число интронов, несущих участок сходства в обеих парах видов, может быть связано с корреляцией длин. Для того, чтобы учесть этот эффект при расчёте ожидаемого числа интронов, несущих сегмент сходства в обеих парах видов в четверке, мы использовали второй тип рандомизации. При проведении рандомизации

второго типа все интроны в каждой паре видов были разделены на 10 групп схожего размера в соответствии с длиной интрона в данной паре видов. Разделить интроны на 10 групп так, чтобы в каждую группу попало строго одно и то же число интронов, было невозможно в связи с тем, что длины заметного числа интронов попадают на границы между группами. Особенно этот эффект выражен для двукрылых, характерная длина интронов у которых значительно меньше характерной длины интронов у позвоночных.

Группы имели следующие пороговые значения для длин интронов: пара *Homo – Mus* (125, 279, 499, 757, 1077, 1479, 2046, 3036 и 5620 нуклеотидов); пара *Tetraodon – Takifugu* (72, 77, 82, 89, 102, 130, 192, 337 и 709 нуклеотидов); пара *D. melanogaster – D. mojavensis* (56, 58, 60, 62, 64, 67, 73, 124 и 573 нуклеотидов); пара *Aedes – Culex* (57, 59, 61, 63, 66, 71, 125, 666 и 4476 нуклеотидов).

Затем каждому интрону из $N_1(E)$ интронов в первой паре и $N_2(E)$ интронов во второй паре мы приписывали группу длины в соответствующей паре видов. На основании этого для каждой пары мы получали распределение интронов, имеющих участок сходства между видами из данной пары, по группам длины. Затем мы получали 10,000 перетасованных выборок, распределяя «участки сходства» по $N_1(E)$ и $N_2(E)$ случайным интронам так, что в каждой выборке сохранялось распределение $N_1(E)$ и $N_2(E)$ по группам длин интронов. Для каждой такой выборки мы определяли число интронов, несущих «участок сходства» одновременно в обеих парах видов. Распределения, основанные на 10,000 перетасованных выборок, использовали для того, чтобы получить среднее значение и доверительные интервалы для ожидаемого числа интронов, несущих участок сходства в обеих парах одновременно.

Третий тип рандомизации аналогичен первому типу, однако при проведении третьего типа рандомизации мы контролировали идентичность гена, сохраняя распределение $N_1(E)$ и $N_2(E)$ по генам.

Четвертый тип рандомизации аналогичен третьему типу, за тем исключением, что первые по порядку интроны в генах были исключены из анализа.

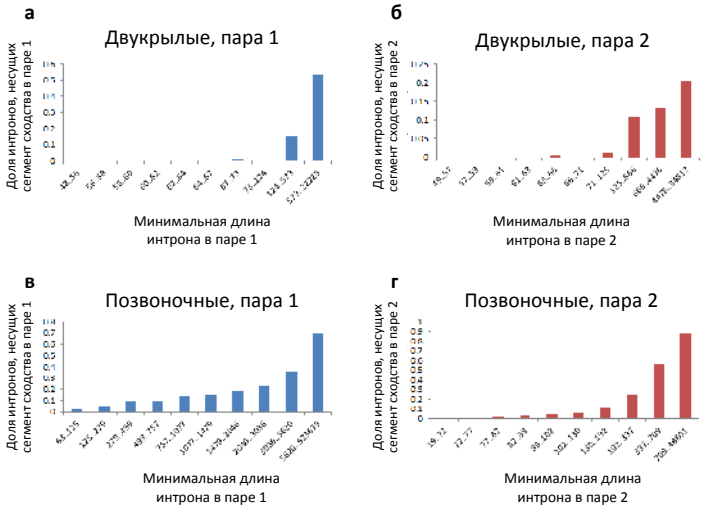

Рисунок 2.2. Длинные интроны чаще несут консервативные участки. Интроны в каждой паре видов разбиты на 10 групп по длине так, чтобы в каждую группу попадало примерно равное количество интронов. Для разбиения на группы использована длина более короткого интрона в паре видов. Для разных групп длин интронов показана доля интронов в группе, имеющих участок локального сходства последовательностей между геномами в паре 1 (**а**, **в**) и паре 2 (**б**, **г**), в четверке двукрылых (**а**, **б**) и позвоночных (**в**, **г**). Поиск интронов, несущих консервативный участок в паре видов, проводили на основании выравнивания последовательностей интронов с применением программы bl2seq, используя пороговое значение E-value $\leq 1.00 \times 10^{-8}$.

2.1.4 Анализ данных по модификациям хроматина

Мы использовали данные по модификациям хроматина для 51,541 интрона *H. sapiens* и 5367 интронов *D. melanogaster* из баз данных ENCODE [115,116] (http://genome.ucsc.edu/ENCODE/) и modENCODE [117,118] (http://www.modencode.org/) соответственно.

В базе данных ENCODE доступна аннотация сегментов генома по хроматиновым состояниям и основанная на этом функциональная аннотация для девяти клеточных линий. Мы исключили из рассмотрения две линии раковых клеток (K562 и HepG2).

Оставшиеся клеточные линии были разделены на две группы – линии клеток, полученные из тканей взрослого организма (GM12878, HMEC, HSMM, NHEK и NHLF), и эмбриональные клеточные линии, включающие эмбриональные стволовые клетки (H1-hESC) и эндотелиальные клетки пупочной вены человека (HUVEC). Мы использовали данные по сегментации генома для версии сборки генома человека Human Genome Build 36 (hg18). В качестве подмножества интронов, вероятно, выполняющих регуляторную роль, мы выбирали интроны, пересекающиеся с сегментами, проаннотированными как сильные энхансеры

(состояния 4 и 5) или инсуляторы (состояние 8) во всех линиях, полученных из тканей взрослого организма, или в обеих эмбриональных клеточных линиях.

Модели сегментации генома для *D. melanogaster* в ModENCODE были доступны для двух клеточных линий (BG3 и S2). Мы использовали аннотацию сегментов генома *D. melanogaster*, соответствующую 5-му выпуску базы данных FlyBase. В качестве подмножества интронов, вероятно, имеющих регуляторную функцию, мы выбирали интроны, пересекающиеся с сегментами, проаннотированными как регуляторные участки (энхансеры, состояние 3) или активными интронами (состояние 4) по крайней мере в одной из клеточных линий.

Значимость превышения наблюдаемого числа интронов, одновременно пересекающихся с регуляторным сегментом в паре 1 и несущих консервативный участок в паре 2, над ожидаемым числом, была оценена на основе 10,000 перетасованных выборок, полученных с учетом длин интронов (см. описание процедуры рандомизации второго типа в предыдущем разделе).

2.1.5 Определение потерь интронов

Для определения потерь интронов на филогении мы использовали информацию о присутствии интронов в геномах видов-аутгрупп (Рисунок 2.1). В качестве интронов, вероятно, присутствовавших в общем предке всех видов из четверки, мы рассматривали интроны, присутствующие в двух видах в паре 1 и в виде-аутгруппе. Такие интроны были разделены на две группы: (i) интроны, также присутствующие в двух видах из пары 2, и (ii) интроны, потерянные по крайней мере в одном из видов в паре 2. Стоит отметить, что в большинстве случаев интроны, отсутствовавшие в одном из видов в паре 2, также отсутствовали во втором виде, что указывает на вероятную потерю таких интронов на ветке, ведущей к общему предку видов из пары 2. Случаи, соответствующие событиям потерь интронов на внешних ветках, сравнительно редки.

Затем для двух групп интронов, обозначенных выше как (i) и (ii), мы сравнивали распределения значений E-value в паре 1 и распределения длин интронов в паре 1. Как и при проведении предыдущих типов анализа, мы исключали из рассмотрения интроны со значительным сходством последовательностей (bl2seq E-value ≤ 0.0001) между видами из разных пар. Аналогичным образом мы анализировали симметричную ситуацию, рассматривая интроны, присутствующие в обоих видах в паре 2 и в виде-аутгруппе.

2.2 Результаты и обсуждение

В этой главе мы на уровне генома исследовали явление сохранения давления отрицательного отбора на гомологичные, но значительно разошедшиеся некодирующие последовательности. Вероятно, что давление отрицательного отбора в таких случаях направлено на сохранение предковой функции. При проведении анализа мы сосредоточились на участках генома, которые разошлись от последовательности общего предка до такой степени, что сходство первичной последовательности было полностью утеряно. Мы предполагаем, что сохранение некоторых характеристик ортологичных последовательностей, для которых тем не менее не может быть построено выравнивание, указывает на продолжение действия отбора на некоторые общие характеристики ортологичных участков. Стоит отметить, что это может быть неверно в том случае, если общие свойства невыравниваемых участков могут быть объяснены как-то еще. В качестве выборки участков генома, ортологичных в далеких видах, мы использовали интроны из ортологичных генов, поскольку общее происхождение таких интронов может быть установлено на основании выравнивания фланкирующих экзонов даже на таких филогенетических расстояниях, на которых сходство последовательностей между интронами полностью потеряно.

Мы использовали четыре типа анализов, результаты которых могут свидетельствовать о продолжении действия отрицательного отбора на участки генома, разошедшиеся до неузнаваемости. Каждый анализ проводили на четверке видов. Четверка состоит из двух пар геномов, подобранных так, что эволюционные расстояния по нейтральным заменам внутри первой (Ks_1) и второй (Ks_2) пары значительно меньше, чем расстояние между парами (Ks_3), но достаточно велики для того, чтобы консервативность последовательностей между видами внутри пары означала действие отрицательного отбора на данный участок генома. Мы рассмотрели две четверки видов, удовлетворяющие таким условиям, – четверку двукрылых и четверку позвоночных (Рисунок 2.1). В обеих четверках значительное число интронов имеют участки значимого локального сходства последовательностей между видами из одной пары, вероятно связанного с давлением отрицательного отбора. В противоположность этому, лишь в незначительном числе интронов наблюдается сходство последовательностей между интронами из разных пар, принадлежащих к одной четверке.

2.2.1 Анализ давления отбора на интроны, несущие сегменты сходства в далеких парах видов

Первый вопрос, ответ на который может позволить выявить сохранение давления отрицательного отбора на ортологичные участки генома, разошедшиеся до неузнаваемости,

можно сформулировать следующим образом: является ли наличие консервативного (и следственно, вероятно, функционального) сегмента в одной паре видов, входящей в четверку, значимым предиктором наличия консервативного сегмента в ортологичном интроне между видами во второй паре. Для этого сначала рассмотрим ортологичные некодирующие участки генома, сохранившие одну и ту же функцию во всех видах, принадлежащих к одной четверке. Сохранение функции должно приводить к более высокому уровню сохранения сходства последовательностей между видами внутри каждой пары по сравнению с ожидаемым в предположении нейтральной скорости эволюции. Кроме того, сильный отбор на сохранение функции может приводить к сохранению существенного сходства последовательностей и на значительно больших эволюционных расстояниях, разделяющих виды из разных пар. В этом случае консервативность последовательности может присутствовать между геномами из разных пар. Поскольку мы были заинтересованы в изучении возможных случаев сохранения функции без сохранения сходства последовательностей, мы исключили из рассмотрения 34 и 303 интрона, несущие участки значимого сходства между видами из разных пар, для четверки двукрылых и позвоночных соответственно. Таким образом, оставшиеся 5333 интрона в четверке двукрылых и 51,541 интрон в четверке позвоночных включают в себя только такие интроны, для которых осмысленное выравнивание между видами из разных пар отсутствует. Тем не менее интрон, имеющий участок локального сходства между видами в одной из пар внутри четверки, несет участок локального сходства в другой паре чаще, чем ожидалось бы в том случае, если бы участки локального сходства в каждой паре были распределены независимо друг от друга (Рисунок 2.3).

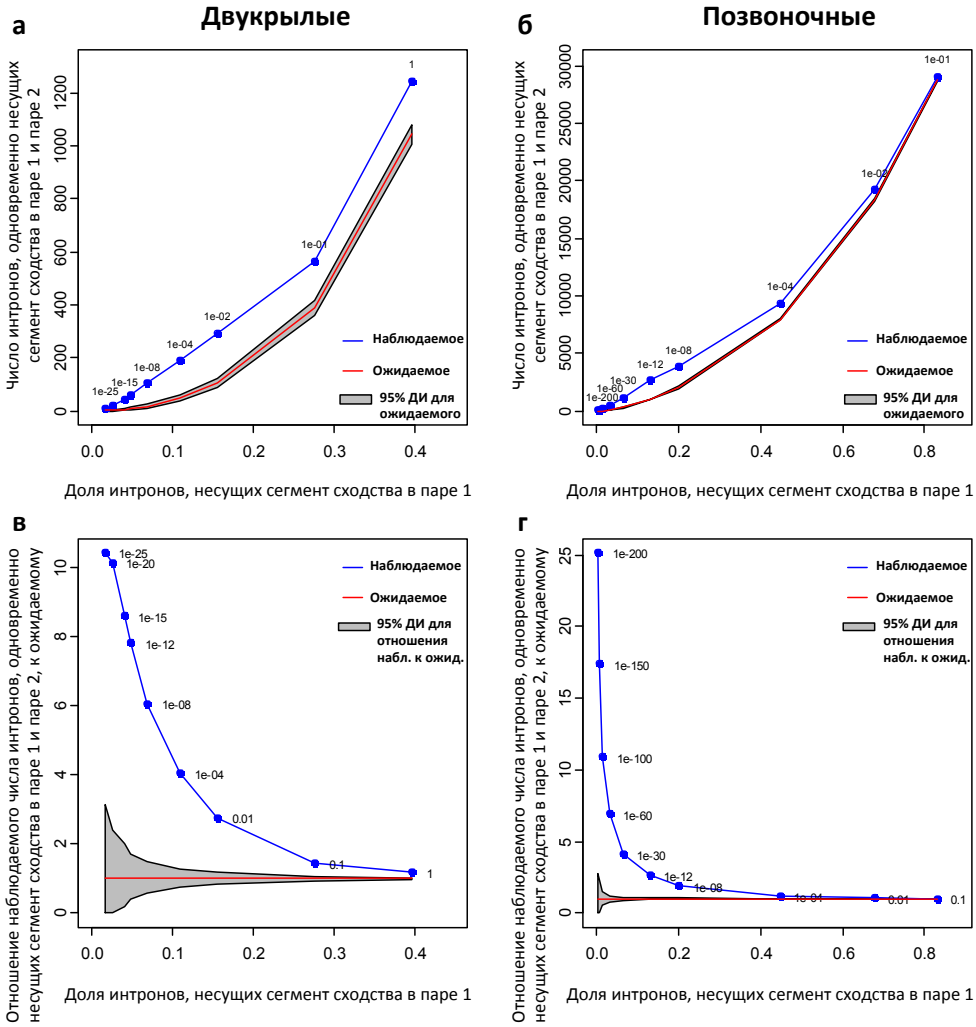

Рисунок 2.3. Интроны, несущие сегмент локального сходства последовательностей между видами из одной пары в четверке, с повышенной частотой также несут сегмент локального сходства в другой паре видов в четверке (без поправки на длину интронов). Синие точки соответствуют различным значениям E-value bl2seq (указаны рядом с точками), использованным как пороговые значения для поиска участков локального сходства между ортологичными интронами в паре 1 и паре 2. Более низкие значения E-value соответствуют более строгим порогам для выявления сходства последовательностей. На горизонтальной оси для данного значения E-value отмечена доля интронов, несущих сегмент локального сходства между ортологичными интронами в паре 1, среди интронов, присутствующих во всех видах в четверке. На вертикальной оси отмечено число (**а**, **б**) или отношение наблюдаемого к ожидаемому (**в**, **г**) для числа интронов, одновременно несущих сегмент сходства в паре 1 и паре 2. (**а**, **в**) – двукрылые; (**б**, **г**) – позвоночные. Отношение наблюдаемого числа к ожидаемому рассчитывали как отношение наблюдаемого числа интронов с участком сходства в обеих парах в четверке к ожидаемому в том случае, если бы участки сходства были случайно распределены по интронам в каждой паре видов, без учета длин интронов (первый тип рандомизации, см.

Материалы и методы). Красная линия и серая область соответствуют среднему значению и 95% доверительному интервалу (ДИ) для ожидаемых значений, рассчитанным на основе 10,000 случайных выборок интронов.

Однако результаты этого анализа могут быть искажены неоднородностью длин интронов. Дело в том, что более длинные интроны имеют большую вероятность нести участки со значимым сходством последовательностей между видами из одной пары (Рисунок 2.2), что согласуется с данными о том, что длинные интроны характеризуются повышенной консервативностью по крайней мере у видов *Drosophila* [112]. В сочетании с тем, что длины интронов положительно скоррелированы между видами (см. Материалы и методы), это само по себе может приводить к превышению над ожиданием числа интронов, несущих консервативные участки в обеих парах в четверке. В связи с этим при оценке ожидаемого по случайным причинам числа интронов, несущих участки локального сходства в обеих парах видов в четверке, необходимо учитывать неоднородность длин интронов.

Для этого мы использовали следующую процедуру. Для разных пороговых значения E-value мы сначала определяли число интронов, $N_1(E)$ и $N_2(E)$, несущих участок локального сходства последовательностей в парах 1 и 2 соответственно. Все интроны в каждой паре видов были разделены на 10 групп схожего размера в соответствии с длиной интрона в данной паре видов. Затем для каждого рассматриваемого значения E-value мы получали 10,000 перетасованных выборок, распределяя сегменты сходства по $N_1(E)$ и $N_2(E)$ случайным интронам так, что в каждой выборке сохранялось распределение $N_1(E)$ и $N_2(E)$ по группам длин интронов. Для каждой перестановки мы определяли число интронов, несущих сегмент сходства одновременно в обеих парах видов. Наблюдаемое в настоящих данных число интронов, несущих сегмент локального сходства в обеих парах видов одновременно, сравнивали с распределением числа таких интронов в 10,000 перетасованных выборок (Рисунок 2.4).

Как видно из Рисунка 2.4, наблюдаемое число интронов, одновременно имеющих участок сходства в видах из первой и второй пары, остается значительно выше ожидаемого и при оценке ожидаемого числа, полученной с учетом неоднородности длин интронов.

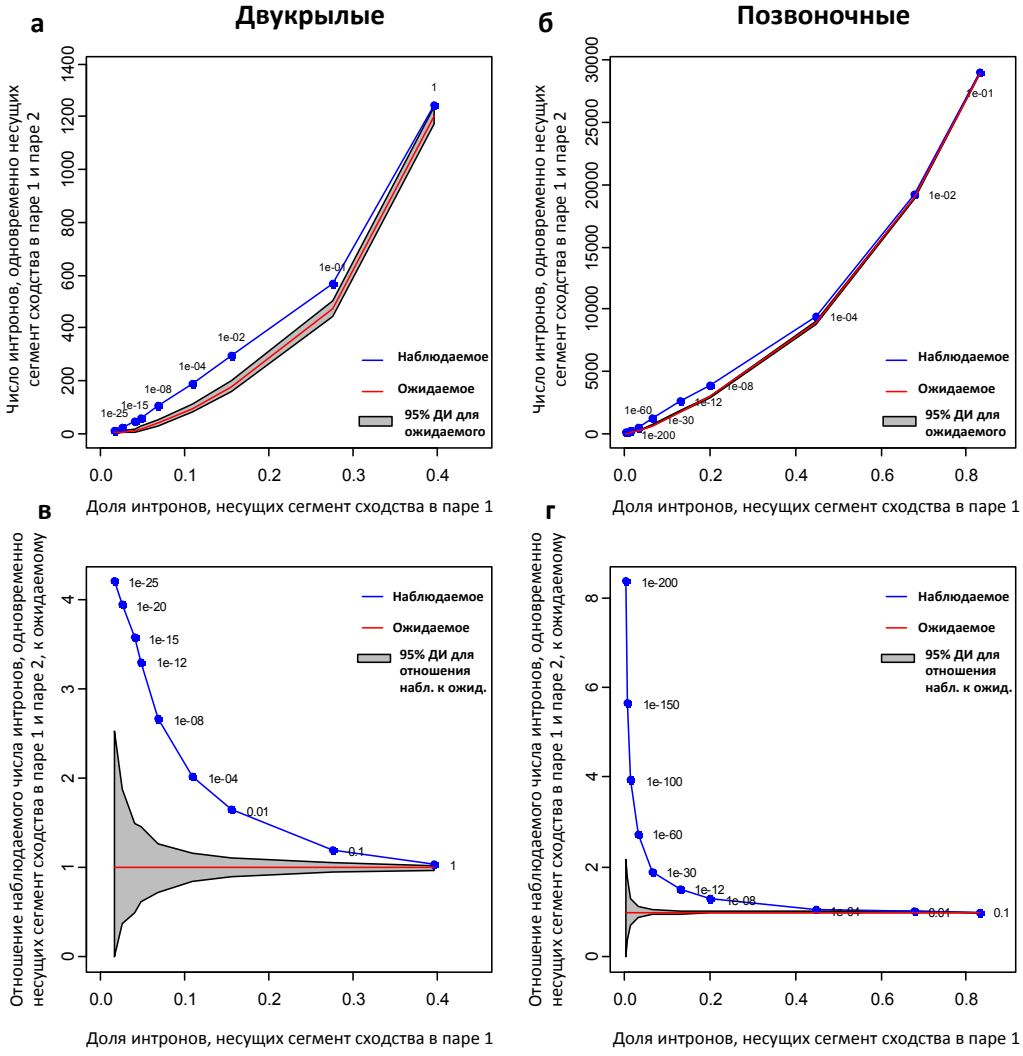

Рисунок 2.4. Интроны, несущие сегмент локального сходства последовательностей между видами из одной пары в четверке, с повышенной частотой также несут сегмент локального сходства в другой паре видов в четверке (с поправкой на неоднородность длин интронов). Синие точки соответствуют различным значениям E-value bl2seq (указаны рядом с точками), использованным как пороговые значения для поиска участков локального сходства между ортологичными интронами в паре 1 и паре 2. Более низкие значения E-value соответствуют более строгим порогам для выявления сходства последовательностей. На горизонтальной оси для данного значения E-value отмечена доля интронов, несущих сегмент локального сходства между ортологичными интронами в паре 1, среди интронов, присутствующих во всех видах в четверке. На вертикальной оси отмечено число (**а**, **б**) или отношение наблюдаемого к ожидаемому (**в**, **г**) для числа интронов, одновременно несущих сегмент сходства в паре 1 и паре 2. (**а**, **в**) – двукрылые; (**б**, **г**) – позвоночные. Отношение наблюдаемого числа к ожидаемому рассчитывали как отношение наблюдаемого числа интронов с участком сходства в обеих парах в четверке к ожидаемому в том случае, если бы участки сходства были случайно распределены по интронам в каждой паре видов, контролируя на длину интронов (второй тип рандомизации,

см. Материалы и методы). Красная линия и серая область соответствуют среднему значению и 95% доверительному интервалу (ДИ) для ожидаемых значений, рассчитанным на основе 10,000 случайных выборок интронов.

Таким образом, среди интронов, для которых отсутствует осмысленное выравнивание между видами из разных пар, тем не менее существует избыток интронов, одновременно несущих консервативный сегмент внутри каждой пары. Данный избыток свидетельствует о продолжении давления отрицательного отбора на некодирующие участки генома в далеких видах, которое тем не менее не может быть выявлено путем непосредственного сравнения ортологичных последовательностей из филогенетически далеких видов, принадлежащих к разным парам. Превышение наблюдаемого числа над ожидаемым тем сильнее, чем более строгий порог использовался для поиска участков локального сходства последовательностей между видами из одной пары. Другими словами, наиболее сильный эффект наблюдается в тех случаях, когда параметры на выявление участков сходства выбраны таким образом, что только небольшая доля интронов имеет участок локального сходства, отвечающий заданным параметрам, – то есть, когда рассматриваются интроны, несущие наиболее консервативные элементы. Так, в том случае, если параметры были выбраны таким образом, чтобы участок значимого локального сходства определялся менее чем в 5% интронов в четверке двукрылых (E-value $\leq 1 \times 10^{-12}$, Рисунок 2.4в) или менее чем в 3% интронов в четверке позвоночных (E-value $\leq 1 \times 10^{-100}$, Рисунок 2.4г), наблюдаемое число интронов, несущих участок сходства одновременно в видах из обеих пар, превышало ожидаемое по случайным причинам в 3 раза. В четверке позвоночных, если рассматривались только интроны, несущие наиболее консервативные сегменты, встречающиеся только в 0.4% интронов, наблюдалось восьмикратное превышение наблюдаемого числа интронов, имеющих участок сходства в обеих парах, над ожидаемым (E-value $\leq 1 \times 10^{-200}$, Рисунок 2.4г).

Однако неслучайное распределение консервативных элементов между интронами в далеких видах может быть вызвано не только сохранением предковой функции, но и тем, что интроны имеют общие характеристики, не обязательно связанные с общим происхождением. В частности, мы ожидаем, что подобная картина будет наблюдаться в том случае, если консервативные участки чаще встречаются в интронах из определенной подвыборки генов (например, генов с высоким уровнем экспрессии), или, если первые по порядку в гене интроны в среднем более консервативны. Чтобы проверить, могут ли полученные нами результаты быть объяснены общими характеристиками интронов из одних и тех же генов, не обязательно связанными с общим происхождением интронов, мы провели следующую процедуру: для каждой пары видов мы пермутировали интроны внутри каждого гена, затем для полученной

таким образом выборки мы определяли число интронов, одновременно несущих участок сходства последовательностей в обеих парах видов. Данное число использовали как контроль, с которым сравнивали число интронов, одновременно несущих участок сходства последовательностей в обеих парах видов, в настоящих данных.

Такой подход позволяет выявить, существует ли превышение наблюдаемого числа интронов, одновременно консервативных в видах из далеких пар, над ожидаемым, которое не может быть объяснено общими свойствами интронов из одного и того же гена. Данный подход очень консервативен, поскольку ожидается, что он должен давать значительное количество ложно-отрицательных результатов, если число интронов на ген низкое и, особенно, если большое число генов имеют всего один интрон. В нашем наборе данных среднее число интронов, для которых удалось установить ортологические соответствия в разных видах, на ген было равно 1.95 (медиана = 1) в четверке двукрылых и 7.00 (медиана = 5) в четверке позвоночных. Тем не менее избыток интронов, несущих консервативные участки в видах из обеих пар, был выявлен и при сравнении с контрольной выборкой, полученной пермутированием интронов внутри каждого гена (Рисунок 2.5). Превышение наблюдаемого над ожидаемым сохранилось и после исключения из анализа первых по порядку в гене интронов (Рисунок 2.6), что указывает на то, что описанный нами эффект не может быть объяснен более высокой консервативностью первых интронов. Таким образом, одновременное присутствие консервативных участков в ортологичных интронах в далеких парах видов при отсутствии сквозного выравнивания между видами из разных пар, по всей видимости, не может быть объяснено общими свойствами интронов, не связанных с их общим происхождением (например, тем, что интрон является первым в гене), или генов, в которых они находятся.

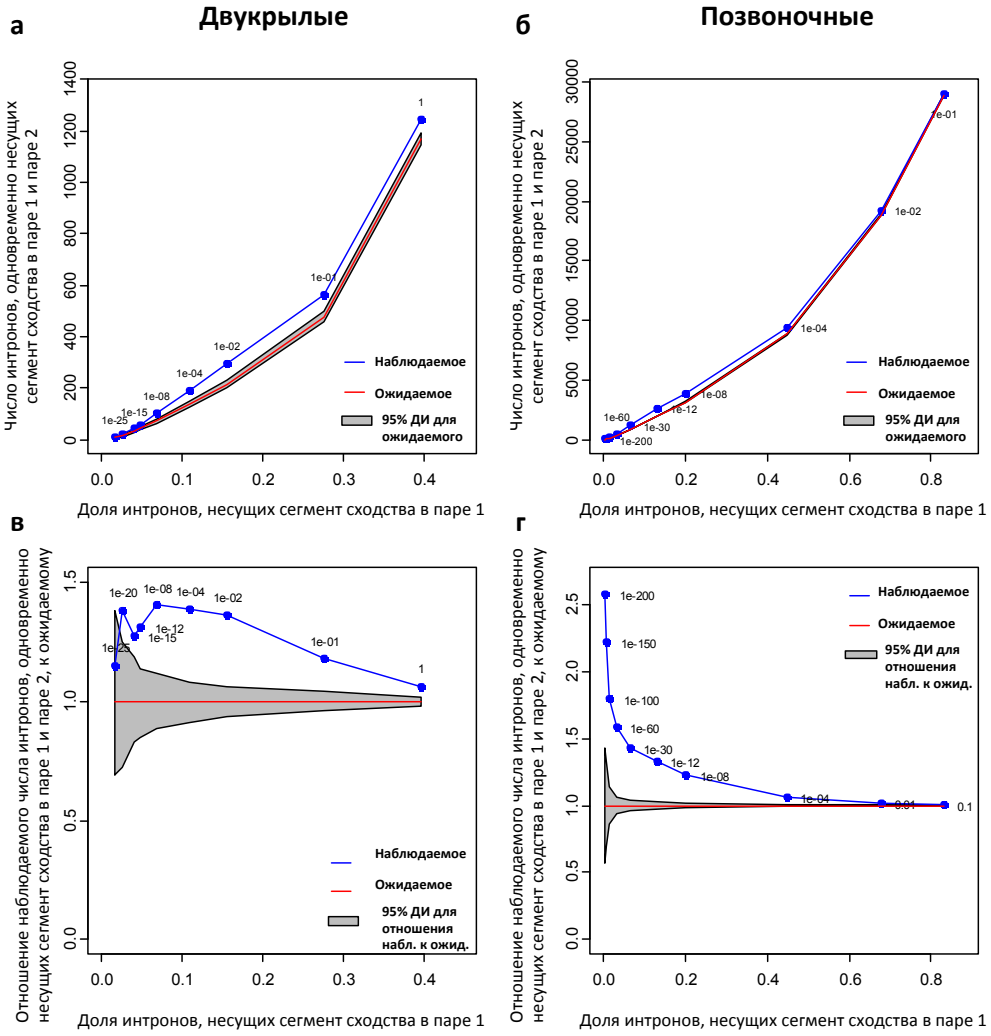

Рисунок 2.5. Избыток интронов, одновременно несущих сегмент локального сходства последовательностей в обеих парах видов в четверке, не объясняется свойствами гена. Использованные условные обозначения такие же, как на Рисунках 2.3 и 2.4. Ожидаемое число интронов с участком сходства в обеих парах видов в четверке рассчитывали на основе пермутаций, в которых «участки сходства» случайно переставляли между интронами каждого гена (третий тип рандомизации, см. Материалы и методы).

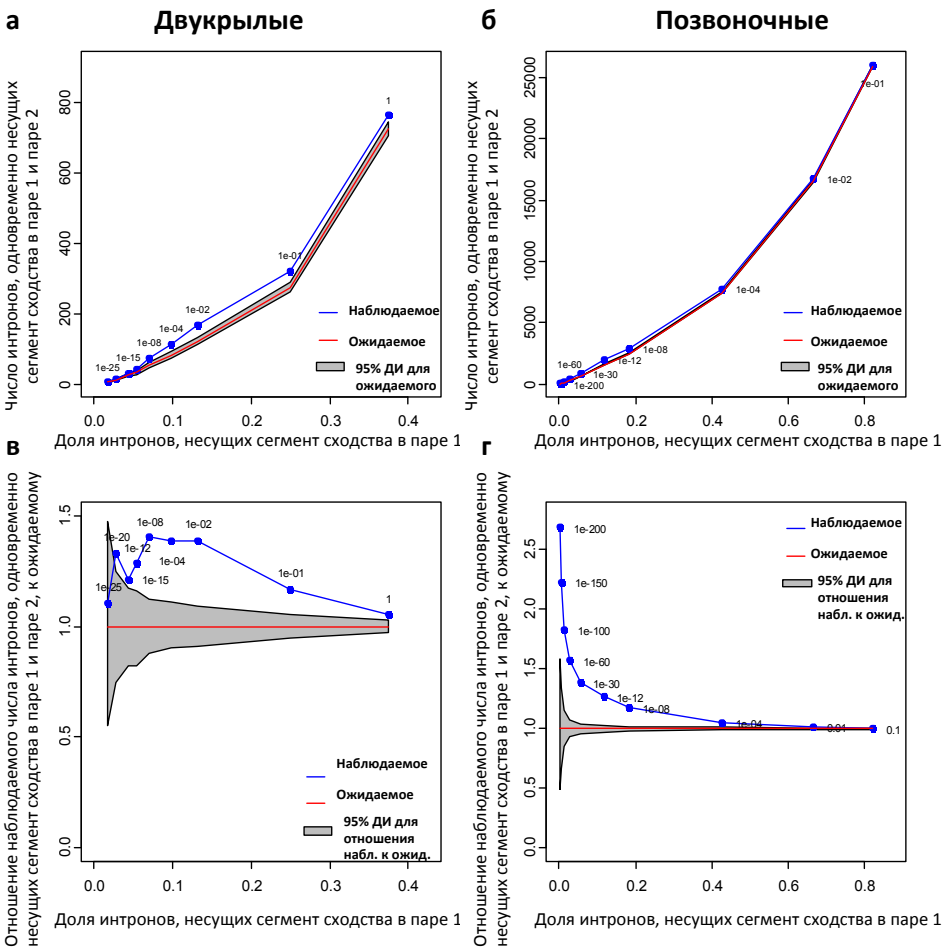

Рисунок 2.6. Избыток интронов, одновременно несущих сегмент значительного локального сходства последовательностей в обеих парах видов в четверке, не объясняется свойствами гена и не определяется присутствием в выборке первых по порядку интронов. Рисунок аналогичен Рисунку 2.5 за тем исключением, что из анализа исключены первые по порядку в гене интроны. Использованные условные обозначения такие же, как на Рисунках 2.3 и 2.4. Ожидаемое число интронов с участком сходства в обеих парах видов в четверке рассчитывали на основе пермутаций, в которых «участки сходства» случайно переставляли между интронами каждого гена после исключения первых интронов (четвертый тип рандомизации, см. Материалы и методы).

2.2.2 Анализ давления отбора на интроны, несущие в далеких видах регуляторный элемент

Затем мы сформулировали и проверили еще одно предсказание гипотезы о сохранении давления отрицательного отбора на ортологичные, но разошедшиеся до неузнаваемости последовательности. Мы предположили, что если некоторые регуляторные элементы существуют дольше, чем сходство последовательностей соответствующих участков генома, мы

ожидали бы, что интроны, несущие регуляторный сегмент в паре 1, чаще, чем ожидается по случайным причинам, несут консервативный сегмент в паре 2. Для того, чтобы проверить эту гипотезу, мы использовали полногеномную разметку генома по регуляторным элементам, предсказанным на основе данных по модификациям хроматина [116–118]. Как и при проведении предыдущего анализа, мы исключили из рассмотрения 34 и 303 интрона с участками локального сходства между видами из разных пар в четверке двукрылых и позвоночных соответственно. Тем не менее мы обнаружили, что интроны *D. melanogaster*, пересекающиеся с областями генома, обогащенными активными модификациями хроматина (и, следовательно, вероятно вовлеченными в регуляторные процессы), до ~3-х раз чаще несут участок значимого сходства последовательностей во второй паре из четверки двукрылых (комары *C. quinquefasciatus – A. aegypti*, Рисунок 2.7а, б). Аналогичным образом, интроны, пересекающиеся с инсуляторами или энхансерами у человека, до 2.6 и 1.4 раз соответственно чаще по сравнению с ожиданием несут участок значимого сходства последовательностей во второй паре из четверки позвоночных (рыбы *T. nigroviridis – T. rubripes*, Рисунок 2.7в–е). При проведении данного анализа, как и при проведении предыдущего, учитывали неоднородность длин интронов (см. Материалы и методы). В случае энхансеров более сильный эффект наблюдался при использовании разметки регуляторных элементов, полученной для эмбриональных клеточных линий (линии эмбриональных стволовых клеток, H1-hESC, и эндотелиальных клеток пупочной вены человека, HUVEC, Рисунок 2.7д–е). Это результат находится в соответствии с наблюдениями о том, что консервативные некодирующие участки часто ассоциированы с генами, вовлеченными в процессы развития [119].

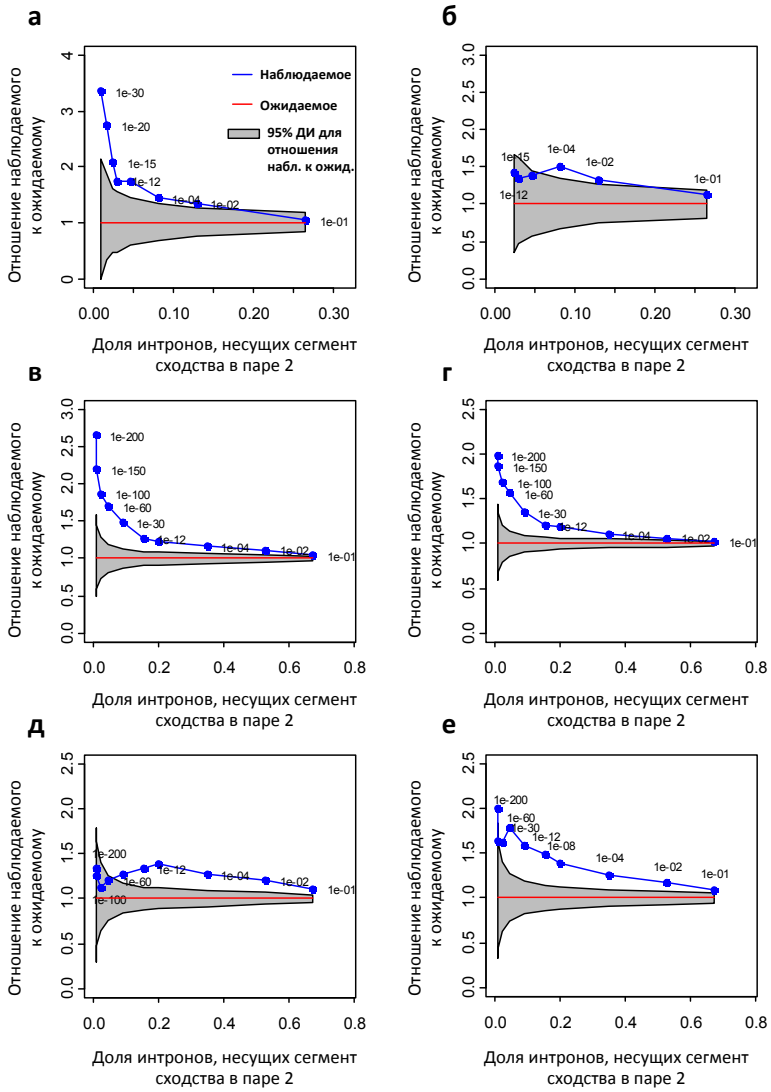

Рисунок 2.7. Интроны, несущие сегмент локального сходства последовательностей в паре 2 в четверке видов, с повышенной частотой пересекаются с регуляторными элементами в паре 1. Синие точки соответствуют различным значениям E-value bl2seq (указаны рядом с точками), использованным как пороговые значения для поиска участков локального сходства между ортологичными интронами в паре 2. Более низкие значения E-value соответствуют более строгим порогам для выявления сходства последовательностей. На горизонтальной оси для данного значения E-value отмечена доля интронов, несущих сегмент локального сходства между ортологичными интронами в паре 2, среди интронов, присутствующих во всех видах в четверке. На вертикальной оси отмечено отношение наблюдаемого к ожидаемому для числа интронов, одновременно пересекающихся с регуляторным элементом в паре 1 и несущих сегмент сходства в паре 2. Для четверки двукрылых (**а**, **б**) использована аннотация регуляторных элементов *D. melanogaster* согласно modENCODE [117,118]; для четверки позвоночных (**в**–**е**) – аннотация регуляторных элементов *H. sapiens* согласно ENCODE

[115,116]. (**а**, **б**) – двукрылые; (**в**–**е**) – позвоночные. (**а**) – активные интроны, любая клеточная линия; (**б**) – энхансеры, любая клеточная линия; (**в**) – инсуляторы, все клеточные линии, полученные из клеток взрослого организма; (**г**) – инсуляторы, обе эмбриональные клеточные линии; (**д**) – сильные энхансеры, все клеточные линии, полученные из клеток взрослого организма; (**е**) – сильные энхансеры, обе эмбриональные клеточные линии. Отношение наблюдаемого числа к ожидаемому рассчитывали как отношение наблюдаемого числа интронов, одновременно пересекающихся с регуляторным элементом в паре 1 и несущих сегмент сходства в паре 2, к ожидаемому в том случае, если бы регуляторные элементы в паре 1 и сегменты сходства в паре 2 были бы случайно распределены по интронам, контролируя на длину интронов (см. Материалы и методы). Серая область соответствует 95% доверительному интервалу (ДИ) для ожидаемых значений, рассчитанному на основе 10,000 случайных выборок интронов.

2.2.3 Анализ потерь интронов, несущих сегмент сходства в далекой паре видов

Затем мы предположили, что присутствие функционального сегмента в интронах может выражаться не только в сохранении последовательности интронов в далеких видах, но и в низкой скорости потерь таких интронов в эволюции. Чтобы проверить эту гипотезу, мы проанализировали, чаще ли те интроны из пары 1, которые сохранились в паре 2, несут консервативный участок в паре 1 по сравнению с интронами, которые были потеряны в паре 2. Поскольку при проведении соответствующего анализа необходимо отличать потери интронов от приобретений интронов, мы использовали информацию из вида-аутгруппы для определения предкового состояния (присутствие или отсутствие интрона в общем предке четверки видов). Как и ранее, чтобы не рассматривать случаи, в которых присутствовало сквозное сходство последовательностей между видами из разных пар, мы исключили из анализа интроны, последовательности которых выравнивались между видами из разных пар. По этому критерию мы исключили 14 из 3073 и 54 из 6609 интронов в четверках двукрылых и позвоночных соответственно. Общее число рассматриваемых интронов в данном анализе меньше, чем в предыдущих, поскольку в выборку включены только те интроны, ортологи которых присутствовали в виде-аутгруппе (*A. mellifera* – для четверки двукрылых и *C. intestinalis* – для четверки позвоночных).

Для интронов в оставшейся подвыборке более высокая доля интронов, консервативных между видами из одной пары, среди интронов, сохранившихся в другой паре, по сравнению с интронами, потерянными в видах из второй пары, может указывать на сохранение предковой функции в видах из разных пар в отсутствие осмысленного выравнивания между интронами из этих видов.

Среди всех интронов, присутствующих в обоих видах *Drosophila* и в виде аутгруппе *(Apis mellifera),* 69% также присутствуют у обоих видов комаров. Среди этих 69% интронов, доля интронов, несущих участок локального сходства между видами *Drosophila*, значительно выше, чем среди оставшихся 31% интронов, потерянных у комаров (P-значение = 5.71×10^{-7}, точный критерий Фишера; Рисунок 2.8а, врезка).

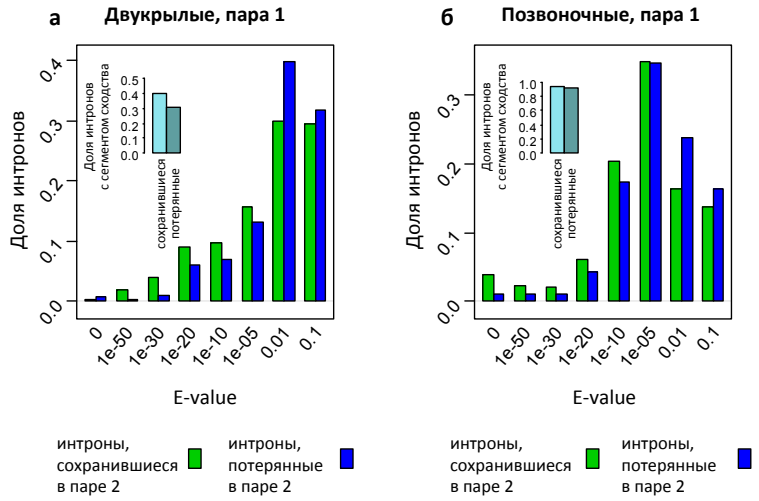

Рисунок 2.8. Последовательность интронов, сохранившихся в видах из пары 2, более консервативна в видах из пары 1. Для двух четверок видов (**а** – двукрылые; **б** – позвоночные) показаны распределения значений E-value для сегментов локального сходства между ортологичными интронами в паре 1. Распределения показаны для двух групп интронов – для интронов, сохранившихся в двух видах в паре 2 (зелёный) и потерянных хотя бы в одном виде из пары 2 (синий). Значения E-value, отмеченные на горизонтальной оси, соответствуют нижнему пороговому значению. На врезках для тех же двух групп показаны доли интронов, для которых в паре 1 находится хотя бы какое-то локальное выравнивание (bl2seq E-value ≤1).

Интрон, присутствующий у обоих видов млекопитающих и виде-аутгруппе *Ciona intestinalis*, также присутствует у обоих видов рыб в 98%. Более высокая доля интронов, сохранившихся одновременно во всех видах, в четверке позвоночных по сравнению с четверкой двукрылых объясняется тем, что анализируемые виды позвоночных филогенетически ближе друг к другу. Кроме того, потери интронов в целом у позвоночных происходят реже [120]. Среди 98% интронов, сохранившихся во второй паре из четверки позвоночных, доля интронов, имеющих участок сходства в паре человек–мышь, немного выше, чем среди оставшихся 2% интронов, однако разница оказалась не значима (P-значение = 0.167, точный критерий Фишера; Рисунок 2.8б, врезка). Вероятно, это связано с тем, что при

использованных порогах для поиска участков локального сходства почти во всех интронах между человеком и мышью определяется сходство последовательностей.

Как в четверке двукрылых, так и в четверке позвоночных участки локального сходства в паре 1 оказались более консервативными среди интронов, сохранившихся в паре 2, по сравнению с интронами, потерянными в паре 2 (Рисунок 2.8). Консервативность оценивали как E-value локального выравнивания, построенного для пары интронов с применением программы bl2seq (более низкие значения E-value соответствуют более строгим порогам на качество и/или длину выравнивания). Распределение значений E-value участков локального сходства в паре 1 для интронов, сохранившихся в паре 2, смещено в сторону более низких значений по сравнению с распределением E-value для участков локального сходства в паре 1 для интронов, потерянных в паре 2 (критерий Манна-Уитни, P-значение = 0.00103 для четверки двукрылых, P-значение = 0.0203 для четверки позвоночных; Рисунок 2.8). Похожие результаты были получены при аналогичном сравнении интронов из пары 2, сохранившихся в паре 1, с интронами из пары 2, потерянными в паре 1 (Рисунок 2.9).

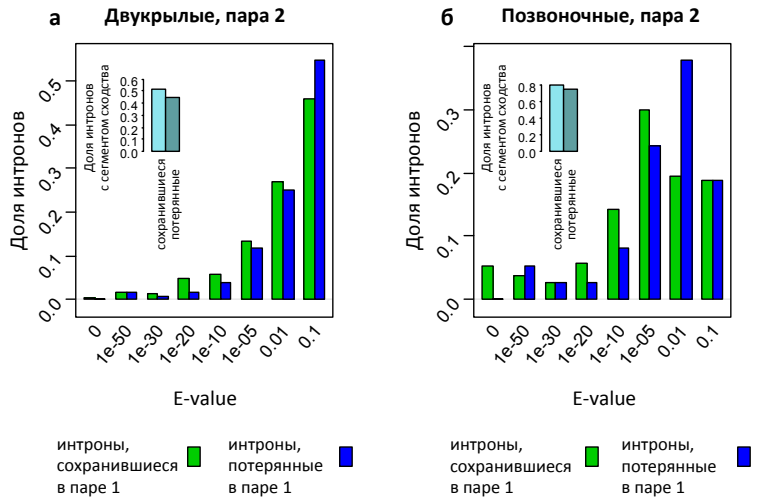

Рисунок 2.9. Последовательность интронов, сохранившихся в видах из пары 1, более консервативна в видах из пары 2. Для двух четверок видов (**а** – двукрылые; **б** – позвоночные) показаны распределения значений E-value для сегментов локального сходства между ортологичными интронами в паре 2. Распределения показаны для двух групп интронов – для интронов, сохранившихся в двух видах в паре 1 (зелёный) и потерянных хотя бы в одном виде из пары 1 (синий). Значения E-value, отмеченные на горизонтальной оси, соответствуют нижнему пороговому значению. На врезках для тех же двух групп показаны доли интронов, для которых в паре 1 находится хотя бы какое-то локальное выравнивание (bl2seq E-value ≤1). Распределения значений E-value в паре 2 для сохранившихся и потерянных в паре 1 интронов значимо отличались в четверке двукрылых (P-значение = 0. 000627, критерий Манна-Уитни), но

не в четверке позвоночных (P-значение = 0.106). Аналогичным образом, доли интронов, имеющих хотя бы какое-то сходство в паре 2, значимо отличались между сохранившимися и потерянными в паре 1 интронами в четверке двукрылых (P-значение = 0.00142, точный критерий Фишера), но не в четверке позвоночных (P-значение = 0.471). Вероятно, это связано с низким числом интронов из пары 2, потерянных в паре 1 в четверке позвоночных (n = 49).

В предыдущих исследованиях было показано, что у *Drosophila* последовательность длинных интронов более консервативна по сравнению с короткими интронами, что, вероятно, указывает на то, что длинные интроны чаще несут регуляторные элементы [112]. В связи с этим сохранение общей функции интрона во всех видах в четверке может проявляться как повышенная в паре 1 длина интронов, сохранившихся в паре 2, по сравнению с интронами, потерянными в паре 2. В соответствии с ожиданием среди 3059/6555 интронов, присутствующих в паре 1 и виде аутгруппе в четверке двукрылых/позвоночных, распределение длин интронов, присутствующих в обоих видах пары 2, смещено в сторону больших длин по сравнению с распределением длин интронов, потерянных в паре 2 (двукрылые: P-значение = 2.64×10^{-18}; позвоночные: P-значение = 0.00111; критерий Манна-Уитни; Рисунок 2.10). Похожие результаты были получены и при аналогичном анализе, проведенном в обратном направлении для интронов, присутствующих в обоих видах пары 2 и виде-аутгруппе (Рисунок 2.11). Этот анализ показывает, что интроны, имеющие более длинных ортологов в филогенетически далеких видах (и, вероятно, несущих функциональный участок ДНК), имеют большую вероятность сохраниться в эволюции, по сравнению с интронами, имеющими коротких ортологов.

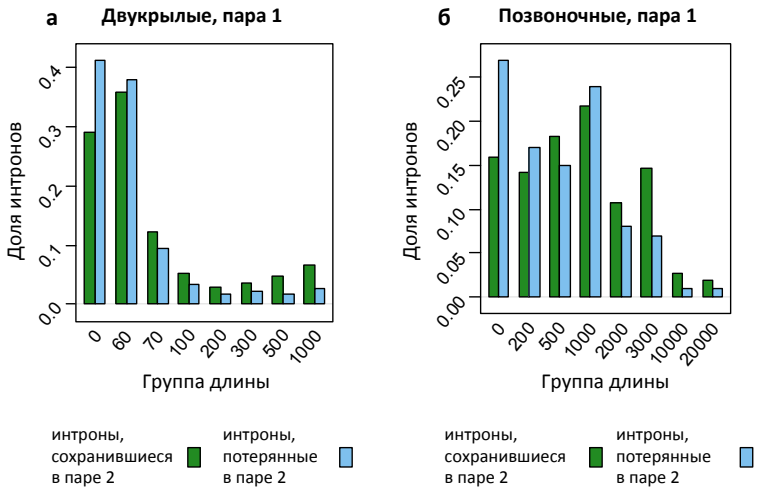

Рисунок 2.10. Интроны, сохранившиеся в видах из пары 2, имеют тенденцию быть длиннее в паре 1 по сравнению с интронами, потерянными в паре 2. Для двух четверок видов (**а** – двукрылые; **б** – позвоночные) показаны распределения длин интронов в паре 1. Распределения показаны для двух групп интронов – для интронов, сохранившихся в двух видах в паре 2 (зелёный) и потерянных хотя бы в одном виде из пары 2 (голубой). Для построения распределения использовали длину более короткого интрона в паре 1. Числа, отмеченные на горизонтальной оси, соответствуют нижнему пороговому значению для данной группы длины.

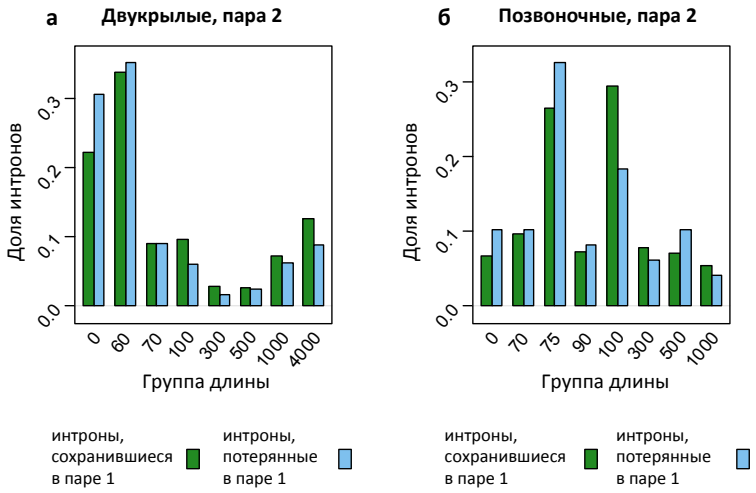

Рисунок 2.11. Интроны, сохранившиеся в видах из пары 1, имеют тенденцию быть длиннее в паре 2 по сравнению с интронами, потерянными в паре 1. Для двух четверок видов (**а** – двукрылые; **б** – позвоночные) показаны распределения длин интронов в паре 2. Распределения показаны для двух групп интронов – для интронов, сохранившихся в двух видах в паре 1 (зелёный) и потерянных хотя бы в одном виде из пары 1 (голубой). Для построения распределения использовали длину более короткого интрона в паре 2. Числа, отмеченные на горизонтальной оси, соответствуют нижнему пороговому значению для данной группы длины. Распределения длин для сохранившихся и потерянных в паре 1 интронов значимо отличались в четверке двукрылых (P-значение = 5.9991×10^{-10}, критерий Манна-Уитни), но не в четверке позвоночных (P-значение = 0.186).

2.2.4 Обсуждение

Таким образом, четыре типа анализов свидетельствуют о существовании в эволюции двукрылых и позвоночных значительного количества случаев, в которых действие отрицательного отбора на ортологичные участки генома продолжается даже тогда, когда между

последовательностями этих участков в далеких видах уже не существует осмысленного выравнивания. Наши результаты показывают, что ситуации, в которых функциональный элемент существует дольше, чем сохраняется сходство предковой последовательности, достаточно частые. Избыток интронов с участками одновременно консервативными в далеких парах видов (Рисунок 2.4) указывает на то, что такие функциональные участки, существующие дольше, чем сходство первичных последовательностей, присутствуют приблизительно в 5% интронов у двукрылых и 3% интронов у позвоночных.

Вероятно, что описанное нами явление существует и в межгенных областях генома, в которых присутствуют множественные регуляторные элементы [121]. Однако в отличие от интронов, для разошедшихся до неузнаваемости межгенных участков сложно достоверно установить ортологические соответствия. Отдельный интерес представляет вопрос о том, с чем именно связано действие отрицательного отбора на разошедшиеся до неузнаваемости ортологичные участки генома, – это может стать направлением для дальнейших исследований в области изучения консервативных некодирующих элементов.

В работе 2015 года, в которой использовался подход близкий к нашему (опубликованному в 2013 году), экспериментально было показано, что предполагаемые энхансеры, находящиеся в ортологичных интронах у мыши и рыб, но не имеющие сходства последовательностей, относительно часто сохраняют гомологичную энхансерную функцию в далеких видах [122]. Как обсуждают авторы, по всей видимости, это объясняется сохранением в энхансерах из далеких видов общего пула коротких сайтов связывания транскрипционных факторов, которые при этом расположены в разных видах в разном порядке [122]. Такое «перемешивание» коротких сайтов связывания приводит к тому, что отдельные функциональные модули, обеспечивающие предковую функцию энхансера, сохраняются, притом что последовательности энхансеров целиком теряют сходство.

Глава 3. Сигнал действия синергического эпистатического отрицательного отбора на вредные аллели в популяциях *D. melanogaster*

Отрицательный отбор, действующий на вредные аллели, появляющиеся в популяциях в результате мутаций, является преобладающим видом естественного отбора. В связи с этим в качестве упрощенной меры приспособленности особи может служить число вредных аллелей в геноме данной особи, или «мутационная нагрузка» вредных аллелей. Однако до последнего времени не было известно, оказывают ли вредные аллели эффект на приспособленность независимо друг от друга или участвуют в эпистатических взаимодействиях. В том случае, если эффекты мутаций на приспособленность независимы друг от друга, приспособленность падает экспоненциально с ростом числа вредных мутаций в геноме. В ситуации мутационно-отборного равновесия это означает, что доля особей в популяции, которая не должна оставить потомство в результате действия отбора («мутационный груз»), равна $1-e^{-U}$, где U – общее количество вредных мутаций, возникающих на геном за поколение [10]. Отсюда проистекает парадокс мутационного груза в видах с низкими темпами размножения и высокой скоростью возникновения вредных мутаций ($U > 1$). Так, в соответствии с предсказанием этой модели более 80% людей не должны оставлять жизнеспособное потомство [123], что, очевидно, расходится с тем, что мы наблюдаем в действительности. Однако в том случае, если вредные мутации находятся в синергических эпистатических взаимодействиях, то есть усиливают влияние друг друга на приспособленность, мутационный груз при условии полового размножения может быть значительно ниже [10,11]. Таким образом, одним из возможных объяснений парадокса мутационного груза может быть синергический эпистатический отбор, действующий на вредные мутации. В связи с этим вопрос о существовании систематических эпистатических взаимодействий на уровне генома имеет важное значение для понимания того, как популяции живых существ противостоят постоянному притоку вредных мутаций. До последнего времени работы по систематическому изучению взаимного влияния вредных мутаций отсутствовали. Ожидается, что синергический эпистаз между вредными мутациями будет выражаться в понижении дисперсии распределения числа вредных аллелей на геном по сравнению с ситуацией, в которой отсутствуют эпистатические взаимодействия [12,124]. В данной главе мы изучили форму распределения числа вредных аллелей на геном в двух популяциях *D. melanogaster* и показали, что наблюдаемые закономерности указывают на эпистатический отрицательный отбор, действующий на вредные аллели в популяциях этого вида.

3.1 Материалы и методы

3.1.1 Наборы данных по полногеномному полиморфизму *D. melanogaster*

Анализ мутационной нагрузки проводили для двух независимых наборов данных по полногеномному секвенированию популяций плодовой мушки *D. melanogaster*. Нами были проанализированы данные по полногеномному полиморфизму, полученные для популяции африканских мух, геномы которых были секвенированы в рамках фазы 3 проекта по популяционной геномике дрозофилы (англ. the Drosophila Population Genomics Project, далее – DPGP3) [125], и данные по полногеномному полиморфизму популяции североамериканских мух (англ. the Drosophila Genetic Reference Panel, далее – DGRP) [126,127].

Отличительной особенностью африканской популяции *D. melanogaster* DPGP3 является то, что она характеризуется наиболее низким из описанных на сегодняшний день уровнем геномной примеси из геномов неафриканских популяций [128]. Еще одной сильной стороной геномов DPGP3 с точки зрения проведения популяционно-генетических анализов является высокий уровень генетической изменчивости, обусловленный тем, что в отличие от североамериканской популяции, африканские популяции не проходили через значительное сокращение численности («бутылочное горлышко») [129].

Для обеих популяций нами были использованы полногеномные данные в формате последовательностей «псевдохромосом» доступные на сайте проекта «Drosophila Genome Nexus project» http://johnpool.net/genomes.html [125]. В рамках данного проекта к наборам данных по полногеномному секвенированию африканской и североамериканской популяций *D. melanogaster* был применен один и тот же алгоритм выравнивания прочтений и поиска однонуклеотидных вариантов [125]. Таким образом, методы определения однонуклеотидных вариантов не отличаются между двумя популяциями, что позволяет в обоих случаях использовать один и тот же подход к анализу данных.

Геномы африканских мух, опубликованные в рамках проекта DPGP3, получены путем секвенирования гаплоидных эмбрионов [125,130], в то время как геномы североамериканских мух DGRP получены путем секвенирования инбредных линий [126]. Для каждого хромосомного плеча каждой особи *D. melanogaster* как из африканской, так и из североамериканской популяции доступна одна «консенсусная» последовательность. То есть каждое хромосомное плечо представлено единственным гаплотипом. В случае африканской популяции это связано с тем, что был использован метод [130], позволяющий определить последовательность гаплоидных геномов. В случае североамериканской популяции значительная доля геномов находится в гомозиготном состоянии в результате близкородственного скрещивания в течение 20 поколений. Поскольку два гаплотипа в

гомозиготных участках опубликованных геномов североамериканских мух не отличаются друг от друга, они могут быть представлены в виде последовательности одного гаплотипа. Гетерозиготные участки, оставшиеся в геномах североамериканских мух, исключены из числа позиций доступных для анализа. Таким образом, геномы африканских мух являются истинно гаплоидными, в то время как геномы североамериканских мух в целях анализа могут рассматриваться как гаплоидные.

Геномные последовательности африканских и североамериканских мух отличаются по качеству и полноте покрытия. Геномы североамериканских мух DGRP по сравнению с африканскими мухами DPGP3 характеризуются значительно большим разбросом значений среднего геномного покрытия [125]. Кроме того, доля геномных позиций, недоступных для анализа, в наборе геномов DGRP значительно выше соответствующей доли в наборе геномов DPGP3. В геномах *D. melanogaster* из DGRP в среднем 22.57% геномных позиций недоступны для анализа (медианное значение 16.11%), в то время как в геномах *D. melanogaster* из DPGP3 в среднем только 6.91% позиций недоступны для анализа (медианное значение 6.87%) (см. Таблицы П3.1 и П3.2). Позиции могут быть недоступны для анализа по различным причинам, таким как отсутствие покрытия в конкретном геноме *D. melanogaster*, гетерозиготность и различные технические артефакты [125]. Гетерозиготность может быть как истинной (в случае инбредных *D. melanogaster* из DGRP), так и ложной, обусловленной техническими ошибками (как в случае диплоидных геномов DGRP, так и в случае гаплоидных геномов DPGP3) [125].

В связи с описанными выше особенностями популяций и секвенированных геномов из DPGP3 и DGRP, набор данных DPGP3 использовали в качестве основного, а DGRP в качестве вспомогательного.

3.1.2 Референтный геном и данные по аннотации белок-кодирующих генов

Референтный геном для *D. melanogaster* и файлы, содержащие аннотацию белок-кодирующих генов, были получены с сайта геномного браузера UCSC Genome Browser (https://genome.ucsc.edu) [131]. В работе была использована сборка генома *D. melanogaster* версии dm3 по классификации базы данных UCSC. Анализ основан на аннотации белок-кодирующих генов *D. melanogaster* из базы данных FlyBase (версия 5.12) [132]. Рассматривались только канонические изоформы генов (n = 13,300). В анализ были включены только гены, находящиеся на преимущественно эухроматических плечах хромосом (2L, 2R, 3L, 3R и X).

3.1.3 Контроль качества данных

Исключение возможных ошибок аннотации белок-кодирующих генов

Для того, чтобы уменьшить влияние на анализ мутаций в нефункциональных участках генома и ошибок аннотации, мы исключили из рассмотрения модели генов, предположительно соответствующие псевдогенам или содержащие ошибки аннотации. На этом этапе было исключено 48 моделей генов (Таблица 3.1).

Модели генов исключали из анализа в следующих случаях:

1. Кодирующая область гена в референтном геноме содержала преждевременный стоп-кодон.
2. В гене отсутствовал канонический терминаторный кодон (UAG, UGA или UAA).
3. Длина кодирующей области гена была не кратна трем.

Исключение генов, кодирующих хеморецепторы

Дополнительно мы исключили из анализа гены хеморецепторов и гены одорант-связывающих белков. Основанием для этого послужили данные о том, что гены этих групп часто подвергаются псевдогенизации [133,134] и обогащены нонсенс-аллелями [127,135]. Гены этих двух крупных семейств включают большое количество быстро эволюционирующих паралогов, которые часто теряются в ходе эволюции [136].

В связи с этим вероятно, что мутации, нарушающие функции генов из этих семейств, не будут оказывать значительного влияния на приспособленность и соответственно не будут находиться под действием сильного отрицательного отбора. Списки генов хеморецепторов и одорант-связывающих белков были получены из базы данных FlyBase (http://flybase.org) [132]. Согласно классификации FlyBase к семейству генов хемероцепторов и семейству генов одорант-связывающих белков для вида *D. melanogaster* отнесено 121 и 52 гена соответственно. Для того, чтобы сопоставить названия генов из FlyBase с идентификаторами генов, указанными в аннотации, была использована таблица flyBaseToCG.txt, полученная на сайте геномного браузера UCSC [131]. Суммарно на этом этапе из рассмотрения были исключены еще 164 модели генов.

Исключение областей, соответствующих полиморфным инверсиям

Ранее было показано, что полиморфные инверсии определяют значительную часть популяционной структуры в популяциях *D. melanogaster* [127,137]. Поскольку ожидается, что присутствие популяционной структуры в данных может приводить к повышению дисперсии распределения мутационной нагрузки [138], мы отдельно проанализировали однонуклеотидные

варианты, находящиеся в участках генома *D. melanogaster*, свободных от инверсий. Для этого мы использовали геномные координаты известных инверсий, присутствующих в популяциях *D. melanogaster*, из опубликованных работ [127,137] и получили список генов, не пересекающихся с соответствующими участками генома. Первоначальное количество канонических моделей генов согласно базе данных FlyBase, а также статистика по количеству моделей генов, оставшихся после вышеописанных этапов фильтрации приведены в Таблице 3.1.

Этап фильтрации	**Количество канонических моделей генов *D. melanogaster* (FlyBase), доступных для анализа**
До фильтрации	13,300
Исключение псевдогенов и ошибок аннотации	13,252
Исключение генов хеморецепторов и одорант-связывающих белков	13,088
Исключение генов, находящихся в участках генома, соответствующих полиморфным инверсиям	4,881

Таблица 3.1. Статистика по количеству канонических моделей генов *D. melanogaster*, доступных для анализа после нескольких последовательных этапов фильтрации. В анализ были включены только гены, находящиеся на 5 преимущественно эухроматических плечах хромосом. Этапы фильтрации были проведены в той последовательности, в которой они перечислены в таблице.

Контроль качества образцов

Прежде чем проводить анализ мутационной нагрузки в популяциях *D. melanogaster*, в каждой популяции мы исключили из рассмотрения образцы, соответствующие выбросам по количеству нереферентных вариантов или геномных позиций недоступных для анализа (замененных на знак N в геномной последовательности данного образца). Образец исключали из анализа, если число однонуклеотидных вариантов или число недоступных для анализа позиций в геномной последовательности данного образца отличалось от среднего значения в соответствующей популяции более чем на три среднеквадратических отклонения. На основании данного критерия из популяций DGRP и DPGP3 было исключено 7 и 6 образцов соответственно. Кроме того, из набора данных DPGP3 был исключен образец с идентификатором ZI382 в связи с тем, что для этого образца отсутствовала последовательность X хромосомы. В случае популяции DGRP мы дополнительно исключили из рассмотрения все

образцы, в геномной последовательности которых более 20% позиций были недоступны для анализа (заменены на знак N). В результате в наборе данных DPGP3 для проведения анализа остался доступным 191 из 197 образцов, а в наборе данных DGRP – 125 из 205 образцов. Списки образцов, использовавшихся при проведении анализа, приведены в Таблицах П3.1 и П3.2.

3.1.4 Идентификация и аннотация минорных аллелей

Анализ мутационной нагрузки в популяциях *D. melanogaster* проводили для однонуклеотидных полиморфизмов (далее в работе также используется сокращенное обозначение SNP, от англ. single nucleotide polymorphism), находящихся в участках генома, соответствующих белок-кодирующим генам. Среди однонуклеотидных полиморфизмов, попадающих в белок-кодирующие гены, мы рассматривали только полиморфизмы, затрагивающие экзоны или сайты сплайсинга. Другие типы геномных вариантов, сегрегирующих в популяциях *D. melanogaster*, в рамках данной работы не анализировали. Данный подход обусловлен тем, что для выбранных типов полиморфизмов можно относительно легко предсказать их ожидаемый эффект на функцию белка. Действительно, полиморфизмы, попадающие в экзоны, можно разделить на три большие группы: синонимические полиморфизмы, несинонимические полиморфизмы и нонсенс-мутации. Синонимические мутации не вызывают замену аминокислотного остатка в последовательности белка и в силу этого являются наиболее нейтральным классом мутаций. Приводящие к изменению аминокислотной последовательности несинонимические мутации гораздо чаще вызывают изменения в структуре и оказывают влияние на функции белка [139,140]. Но наиболее вредным типом мутаций являются нонсенс-мутации, наиболее редкие среди всех мутаций, затрагивающих экзоны [141]. Мутации этого типа приводят к появлению преждевременного стоп-кодона в последовательности гена, что вызывает досрочное прекращение трансляции белка и чаще всего выражается в отсутствии функционального белкового продукта. Значительная часть мРНК, содержащих преждевременные стоп-кодоны, подвергаются деградации по пути нонсенс-опосредованного распада [142,143].

Наряду с нонсенс-мутациями к потери функции белка могут приводить и мутации, вызывающие поломку сайтов сплайсинга [144]. Возможные последствия мутаций в сайтах сплайсинга включают в себя пропуск экзона и включение последовательности интрона. Нарушение функции белка и в том, и в другом случае может быть обусловлено нарушением рамки считывания или изменением аминокислотной последовательности.

Для того, чтобы определить минорные однонуклеотидные варианты, присутствующие в популяциях *D. melanogaster*, для кодирующей области каждого рассматриваемого гена (англ.

coding sequence, CDS) сначала была получена консенсусная последовательность. Консенсусная последовательность была построена независимо для каждой популяции. Для этого в каждой популяции в каждой позиции кодирующей последовательности выбирали наиболее часто встречающийся нуклеотид. После этого мы рассматривали по отдельности каждый кодон полученной консенсусной последовательности и исключали из анализа кодоны, содержащие позиции, недоступные для анализа в большей части особей в данной популяции (содержащие знаки N). Аналогичным образом была получена консенсусная последовательность для всех сайтов сплайсинга в рассматриваемых генах. В дальнейший анализ включали только сайты сплайсинга, имеющие каноническую последовательность (GT для донорных сайтов сплайсинга и AG для акцепторных сайтов сплайсинга) как в референтном геноме *D. melanogaster*, так и в консенсусе.

Затем для каждой популяции *D. melanogaster* был проведен поиск вариабельных позиций в кодирующих областях генов и в сайтах сплайсинга. Для каждого аллеля в каждом вариабельном сайте была определена популяционная частота. Мы исключили из анализа SNPs, попадающие в кодоны, несущие одновременно более одного SNP в одной и той же особи *D. melanogaster*, поскольку эффекты таких SNPs на последовательность белка нельзя оценивать независимо друг от друга. Кроме того, такие случаи, вероятно, обогащены двойными мутациями, что делает предположение о том, что соответствующие SNPs произошли в результате независимых событий, неверным (Таблица 3.2). Следуя той же логике, мы не рассматривали при проведении дальнейшего анализа сайты сплайсинга, несущие более одного SNP в одной и той же особи (Таблица 3.3). SNPs классифицировали как синонимические, несинонимические, нонсенс (приводящие к появлению преждевременного стоп-кодона в последовательности мРНК) или приводящие к поломке сайтов сплайсинга. Классификация проводилась согласно аннотации белок-кодирующих генов из базы данных Flybase относительно наиболее распространенного в рассматриваемой популяции (консенсусного) аллеля. Мутации, вызывающие потерю терминаторного кодона, в данной работе не анализировали, т.к. мутации этого типа очень редки и находятся под значительно более слабым отрицательным отбором по сравнению с нонсенс-мутациями [135].

Все SNPs, затрагивающие сайты сплайсинга, рассматривали как приводящие к поломке сайтов сплайсинга, за исключением SNPs в донорных сайтах, соответствующих переходу из канонического донорного сайта (GT) в слабый вариант донорного сайта (GC) [145], поскольку такой вариант тоже может обеспечивать нормальный сплайсинг [146]. Дополнительно мы ввели определение общей категории аллелей, вызывающих потерю функции гена. Данная категория является объединением множества нонсенс-аллелей и множества аллелей, вызывающих поломку сайтов сплайсинга.

Популяция	**Общее число вариабельных кодонов**	**Общее число вариабельных кодонов, оставшихся после исключения кодонов, несущих вероятные двойные мутации**	**Общее число вариабельных кодонов, оставшихся после исключения кодонов, несущих вероятные двойные мутации, и кодонов доступных для анализа не во всех образцах**
DPGP3	1,147,021	1,128,927	858,135
DGRP	471,696	463,690	51,608

Таблица 3.2. Статистика по количеству вариабельных кодонов в двух рассмотренных популяциях *D. melanogaster*.

Популяция	**Общее число вариабельных сайтов сплайсинга**	**Общее число вариабельных сайтов сплайсинга, оставшихся после исключения сайтов, несущих вероятные двойные мутации**	**Общее число вариабельных сайтов сплайсинга, оставшихся после исключения сайтов, несущих вероятные двойные мутации, и сайтов доступных для анализа не во всех образцах**
DPGP3	712	703	516
DGRP	297	294	38

Таблица 3.3. Статистика по количеству вариабельных сайтов сплайсинга в двух рассмотренных популяциях *D. melanogaster*.

Для каждой особи *D. melanogaster* мутационную нагрузку рассчитывали как количество минорных (неконсенсусных) аллелей в геноме данной особи. Мутационную нагрузку рассчитывали отдельно для каждого функционального класса полиморфизмов для всех минорных аллелей (частота аллеля в популяции <50%) и для редких минорных аллелей (встречающихся не более чем в 5 особях в рассматриваемой популяции).

Использование для расчета мутационной нагрузки позиций, информация о последовательности которых в части особей не определена, может искусственно завышать дисперсию распределения мутационной нагрузки. В связи с этим в случае популяции африканских мух DPGP3 такие позиции не включали в анализ. В случае североамериканской популяции DGRP для расчета мутационной нагрузки использовали все вариабельные позиции,

поскольку только 11% и 13% вариабельных кодонов и сайтов сплайсинга соответственно оставались доступны для анализа после исключения позиций с аллелем, последовательность которого в части особей не определена (Таблица 3.2, Таблица 3.3). Это связано с высокой долей позиций генома замаскированных в геномах *D. melanogaster* в популяции DGRP.

3.1.5 Анализ свойств распределения мутационной нагрузки

Мы рассматривали распределение мутационной нагрузки для разных типов аллелей (синонимических, несинонимических, нонсенс-аллелей и общей категории аллелей потери функции) и для разных пороговых значений, ограничивающих сверху частоту минорного аллеля. Для каждого такого распределения, представляющего собой распределение количества минорных аллелей данного типа на особь в популяции, мы рассчитывали среднее число аллелей на особь, дисперсию (σ^2) числа аллелей на особь и аддитивную дисперсию (V_A).

Аддитивная дисперсия определяется как сумма дисперсий для отдельных полиморфных локусов:

$$V_A = \sum_{i=1}^{N} \mathrm{Var}_i \qquad (3.1)$$

где Var_i – дисперсия мутационной нагрузки для полиморфного сайта *i, N* – общее число полиморфных сайтов данного типа.

Далее мы определяли отношение дисперсии к аддитивной дисперсии σ^2/V_A для распределения мутационной нагрузки минорных аллелей данного типа. Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

3.1.6 Анализ выборок аллелей, попадающих в «необходимые» гены *D. melanogaster*

Список генов, функции которых жизненно важны для *D. melanogaster* (далее – «необходимые» гены), был получен в базе данных DEG (англ. аббревиатура расшифровывается как Database of Essential Genes, http://tubic.tju.edu.cn/deg/) [147].

Согласно базе данных DEG, 339 генов *D. melanogaster* проаннотированы как «необходимые». Для 267 из 339 «необходимых» генов было определено точное соответствие с генами из профильтрованного списка, использованного в данной работе. Мы рассчитали σ^2/V_A для распределения количества несинонимических и синонимических минорных аллелей,

попадающих в «необходимые» гены, для обеих популяций *D. melanogaster*. В рамках данного анализа рассматривали все минорные несинонимические и синонимические аллели с частотой в популяции до 50%. Ожидаемо низкое число нонсенс-аллелей, затрагивающих «необходимые» гены, сделало отдельный анализ этой категории полиморфизмов невозможным (мы выявили 6 и 3 нонсенс-аллеля, затрагивающих «необходимые» гены, в популяциях DPGP3 и DGRP соответственно).

Для того, чтобы оценить значимость снижения дисперсии распределения числа несинонимических аллелей, попадающих в «необходимые» гены в популяции DPGP3, мы генерировали 1000 случайных выборок синонимических аллелей, попадающих в любые гены, с популяционными частотами, соответствующими частотам несинонимических аллелей, попадающих в «необходимые» гены. Затем мы рассчитывали отношение σ^2/V_A для каждой случайной выборки. Аналогичным образом мы сгенерировали 1000 случайных выборок синонимических аллелей, попадающих в «необходимые» гены. Отношение σ^2/V_A для числа несинонимических аллелей, попадающих в «необходимые» гены («опыт»), сравнивали с распределением σ^2/V_A в 1000 случайных «контрольных» выборок и рассчитывали одностороннее P-значение. P-значение рассчитывали как долю случайных выборок, в которых σ^2/V_A было ниже или равно значению σ^2/V_A для числа несинонимических аллелей, попадающих в «необходимые» гены. Данная процедура была проведена для двух наборов случайных выборок (синонимических аллелей, попадающих в любые гены, и синонимических аллелей, попадающих в «необходимые» гены).

3.1.7 Данные по отношению скоростей несинонимической и синонимической эволюции для генов *D. melanogaster*

В качестве показателя интенсивности отрицательного отбора, действующего на ген, мы использовали отношение числа несинонимических замен к синонимическим заменам на сайт. Далее в тексте используется стандартное обозначение данного отношения dN/dS.

Мы использовали значения dN/dS для генов *D. melanogaster*, присутствующих строго в одной копии в шести геномах из подгруппы *melanogaster* рода *Drosophila* (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*), из опубликованной работы [148]. Использованные значения dN/dS доступны в сети Интернет по адресу: ftp://ftp.flybase.net/genomes/12_species_analysis/clark_eisen/paml. Использованный набор данных содержит значения dN/dS для 8510 генов, рассчитанные на основании выравниваний последовательностей из шести видов, принадлежащих к подгруппе *melanogaster*. Для 7888 из 8510 генов с имеющимся значениями dN/dS было определено точное соответствие с генами из профильтрованного списка, использованного в данной работе. Для каждого гена мы

использовали единственное значение dN/dS, определенное для подгруппы *melanogaster* в базовой модели M0 с помощью программы CODEML [148,149]. Гены со значениями dN/dS >0.7 были исключены из рассмотрения. В результате для анализа осталось доступно 7836 генов.

Оставшиеся гены были разделены в соответствии со скоростью белковой эволюции (dN/dS) в подгруппе *melanogaster* на 5 групп одинакового размера. Группы пронумерованы по возрастанию скорости белковой эволюции так, что в группу 1 входят наиболее медленно эволюционирующие гены, а в группу 5 – наиболее быстро эволюционирующие гены. Медианные значения отношения dN/dS для полученных 5 групп равны 0.018, 0.041, 0.063, 0.097 и 0.184 соответственно. Отношение σ^2/V_A рассчитывали отдельно для каждой группы dN/dS для всех минорных (популяционная частота <50%) нонсенс-полиморфизмов, несинонимических и синонимических полиморфизмов.

3.2 Результаты и обсуждение

Ранее в теоретических работах было показано, что присутствие эпистатических взаимодействий между вредными мутациями на уровне генома должно менять форму распределения числа вредных аллелей (мутационной нагрузки) на геном в популяции [12,49,124]. Так, дисперсия числа мутаций на геном (σ^2) в популяции в начале каждого поколения может быть представлена как сумма аддитивной дисперсии (V_A) и остаточного члена (C_L) [12,49]. Аддитивная дисперсия соответствует вкладу в дисперсию отдельных локусов и рассчитывается как сумма дисперсий по отдельным локусам. Остаточный член C_L отражает зависимость в распределении мутаций в отдельных локусах друг от друга или неравновесие по сцеплению между локусами [12,49]. В случае отсутствия эпистаза мутации в разных локусах должны оказывать влияние на приспособленность независимо друг от друга и, соответственно, вносить независимый вклад в мутационную нагрузку [10], а дисперсия распределения числа мутаций на геном должна быть равна аддитивной дисперсии ($\sigma^2 = V_A$). Если рассматриваются редкие аллели, то распределение числа таких аллелей на геном должно иметь форму распределения Пуассона, а его дисперсия быть равной среднему числу аллелей на геном в популяции, M, что можно записать как: $\sigma^2 = V_A = M$ [12,124].

Равенство между дисперсией и аддитивной дисперсией нарушается при эпистатическом отборе, поскольку эпистатический отбор создает зависимость в распределении вредных аллелей в геномах [12,124]. Так, синергический эпистатический отбор, при котором логарифм приспособленности падает быстрее, чем линейно (см. Рисунок 1.1), должен создавать недостаток особей с большим числом вредных аллелей в геноме или «расталкивание» между аллелями, соответствующее отрицательному неравновесию по сцеплению. В этом случае

ожидается, что дисперсия распределения числа вредных аллелей на геном будет ниже аддитивной дисперсии ($\sigma^2 < V_A$), поскольку C_L принимает отрицательные значения [12,124]. И, напротив, антагонистический эпистаз, при котором логарифм приспособленности падает медленнее, чем линейно, создает «притяжение» или положительное неравновесие по сцеплению между вредными аллелями, что должно приводить к «сверхдисперсии» ($\sigma^2 > V_A$) распределения числа вредных аллелей на геном [12,124]. Мы использовали эти теоретические предсказания об отношении дисперсии (σ^2) к аддитивной дисперсии (V_A) для поиска подписей эпистатического отрицательного отбора на открытых данных по полногеномному полиморфизму *D. melanogaster*.

3.2.1 Анализ дисперсии распределения мутационной нагрузки для минорных аллелей разных типов в популяциях *D. melanogaster*

Во время проведения исследования было доступно два крупных набора данных по полногеномному полиморфизму у *D. melanogaster* – данные полногеномного секвенирования для популяции африканских мух из Замбии (фаза 3 проекта по популяционной геномике дрозофилы, англ. – the Drosophila Population Genomics Project, сокращенно DPGP3, 197 особей) [125] и данные для популяции мух из Северной Америки (англ. – the Drosophila Genetic Reference Panel, сокращенно DGRP, 205 особей) [126,127].

Стоит отметить, что присутствие популяционной структуры в данных, например, примеси из генетически далеких популяций может приводить к повышению дисперсии распределения числа аллелей на геном в отсутствие эпистаза [138]. В связи с этим анализ дисперсии распределения мутационной нагрузки предпочтительно проводить в популяциях с минимальным уровнем примеси. Значительный уровень африканской и европейской примеси в геномах североамериканских *D. melanogaster* [129] осложняет использование данных DGRP для анализа мутационной нагрузки. Кроме того, в геномах из популяции DGRP высокая доля сайтов недоступна для анализа из-за остаточной гетерезогитности и технических артефактов (см. Материалы и методы). В то же время низкие уровни примесей из геномов неафриканских *D. melanogaster* в популяции африканских мух DPGP3 из Замбии [125,128] и другие свойства этих геномов (см. Материалы и методы) делают этот набор данных по полиморфизму наиболее подходящим для цели нашей работы. В связи с этим набор данных DPGP3 использовали в качестве основного, а DGRP – в качестве дополнительного.

Мы провели анализ распределения мутационной нагрузки в двух популяциях *D. melanogaster*, сосредоточившись на однонуклеотидных полиморфизмах, попадающих в белок-кодирующие гены (см. Материалы и методы). Использование данного подмножества полиморфизмов позволяет относительно просто классифицировать аллели по их

предполагаемому эффекту на функции гена. В качестве вредных аллелей, которые вероятнее всего находятся под давлением сильного отрицательного отбора, мы рассматривали нонсенс-мутации, приводящие к появлению преждевременного стоп-кодона в последовательности гена, и мутации, вызывающие поломку сайта сплайсинга. Минорные аллели этих двух типов были объединены в общую категорию аллелей, вызывающих потерю функции гена. Мы проводили анализ распределения мутационной нагрузки для общей категории аллелей, вызывающих потерю функции гена, а также отдельно для нонсенс-аллелей.

Прежде чем использовать данные по полиморфизму для анализа, мы провели фильтрацию данных, применив жесткие пороги на качество (см. Материалы и методы). В частности, из рассмотрения были исключены геномы, соответствующие выбросам по количеству нереферентных вариантов, и полиморфные сайты, последовательность которых была определена не во всех образцах (в случае популяции DPGP3).

В качестве характеристики мутационной нагрузки для каждого генома мы использовали число минорных аллелей, то есть таких аллельных вариантов в полиморфных позициях генома, которые присутствуют менее чем в половине особей в данной популяции.

Поскольку более вредные полиморфизмы в среднем имеют более низкую аллельную частоту в популяциях [150], ожидается, что подмножество редких полиморфизмов будет обогащено вредными аллелями. В связи с этим мы рассчитывали мутационную нагрузку на основе двух наборов данных – на основе всех минорных полиморфных вариантов (популяционная частота <50%) и на основе подмножества редких минорных полиморфных вариантов. Подмножество редких минорных полиморфных вариантов включает в себя варианты, встречающиеся не более чем в 5 особях в данной популяции.

Описанные выше теоретические ожидания того, как эпистатические взаимодействия влияют на форму распределения мутационной нагрузки в популяции, получены для вредных мутаций [12,49,124]. В связи с этим мы сосредоточились на анализе минорных аллельных вариантов, предположительно вызывающих поломку гена, поскольку такие аллели с наибольшей вероятностью находятся под действием отрицательного отбора. Однако ошибки аннотации генов, например, включение псевдогенов в анализ, могут привести к некорректной классификации аллелей как вредных. На этом основании мы провели фильтрацию моделей генов, исключив из анализа модели, вероятно, являющиеся результатами ошибочной аннотации (см. Материалы и методы). Кроме того, мы не рассматривали полиморфные варианты, попадающие в гены хеморецепторов и одорант-связывающих белков, поскольку гены этих групп часто подвергаются псевдогенизации [133,134] и обогащены нонсенс-аллелями [127,135]. В связи с этим вероятно, что мутации, нарушающие функции генов из этих семейств, как и мутации, попадающие в ошибочно проаннотированные участки генома, не будут оказывать

значительного влияния на приспособленность и, соответственно, не будут находиться под действием сильного отрицательного отбора.

Полученные наборы однонуклеотидных полиморфизмов, оставшиеся после применения описанных выше шагов фильтрации, были использованы для анализа распределения мутационной нагрузки. Анализ проводили отдельно для синонимических и несинонимических минорных вариантов, а также для вариантов, вероятно, вызывающих потерю функции гена (нонсенс-аллелей и аллелей, приводящих к поломке сайта сплайсинга).

В обеих популяциях для каждой особи определяли число минорных аллельных вариантов каждого типа. На основании полученных чисел рассчитывали среднее, дисперсию (σ^2) и аддитивную дисперсию (V_A) распределения числа минорных вариантов каждого типа на особь в рассматриваемой популяции.

В популяции DPGP3 распределение числа на геном вредных аллелей, вызывающих поломку гена, имеет пониженную по сравнению с ожиданием дисперсию ($\sigma^2 < V_A$, Таблица 3.4). Данный эффект присутствует, если рассчитывать мутационную нагрузку по отдельности для нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, и если рассчитывать мутационную нагрузку для этих двух типов аллелей одновременно (общая категория аллелей, вызывающих потерю функции гена). Эффект сохраняется как в том случае, если в анализ включены все минорные полиморфизмы (аллельная частота <50%), так и в том случае, если рассматриваются только редкие минорные полиморфизмы (аллельная частота ≤5; Таблица 3.4).

Так, для распределения числа редких минорных полиморфизмов, вызывающих потерю функции, $\sigma^2/V_A = 0.929$, если же рассматривать все минорные полиморфизмы, вызывающие потерю функции, $\sigma^2/V_A = 0.851$. В то же время значения σ^2/V_A для синонимических и несинонимических минорных вариантов значительно превышают 1 (Таблица 3.4). Мы сравнили наблюдаемое распределение числа минорных аллелей на геном особи в популяции DPGP3 с ожидаемым в случае независимого распределения аллелей по геномам распределением Пуассона с таким же средним (Рисунок 3.1). Данное сравнение показало, что понижение дисперсии по сравнению с аддитивной дисперсией для аллелей, вызывающих поломку белка, вызвано недопредставленностью особей с большим числом таких аллелей в геноме (Рисунок 3.1). Это наблюдение указывает на то, что понижение σ^2 по отношению к V_A может быть вызвано синергическими эпистатическими взаимодействиями между вредными аллелями на уровне генома, определяющими эффективную элиминацию из популяции особей, несущих большое число вредных аллелей.

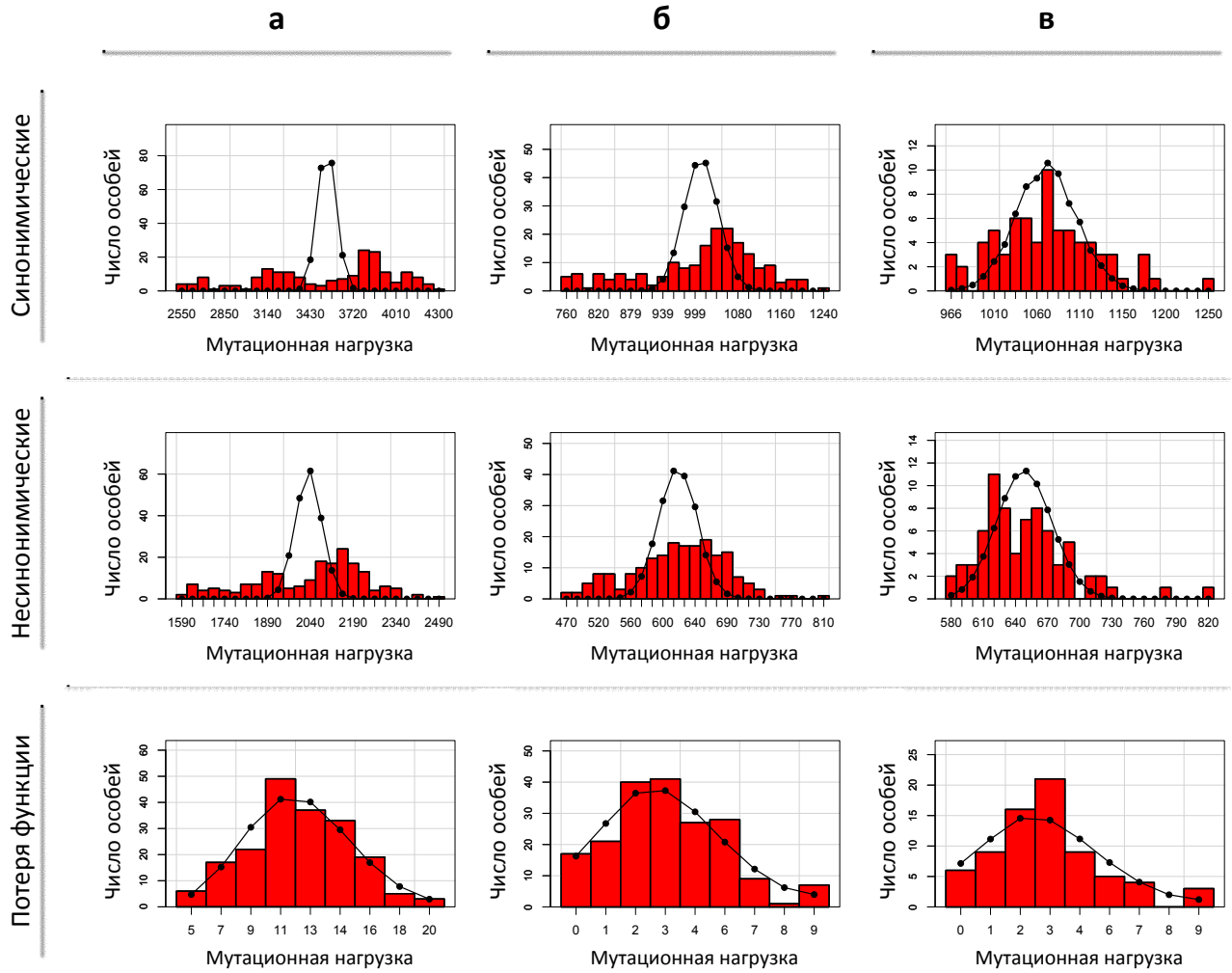

Рисунок 3.1. Мутационная нагрузка в популяции африканских мух DPGP3. Показано распределение числа редких минорных аллелей на особь в популяции DPGP3 для синонимических, несинонимических аллелей и аллелей, вызывающих потерю функции гена. Тип минорных аллелей обозначен слева. В анализ включены аллели, встречающиеся не более чем в 5 особях в популяции DPGP3. Черная линия соответствует ожидаемому числу особей, имеющему заданное число аллелей в геноме в случае, если бы число аллелей на геном особи имело распределение Пуассона с таким же средним, как в данных. (**а**) Мутационная нагрузка была рассчитана на основе всех доступных для анализа редких минорных аллелей. (**б**) Мутационная нагрузка была рассчитана на основе редких минорных аллелей, оставшихся после исключения аллелей, попадающих в участки известных инверсионных полиморфизмов. (**в**) Мутационная нагрузка была рассчитана для особей, в геномах которых отсутствуют известные инверсионные полиморфизмы, на основе редких минорных аллелей, оставшихся после исключения аллелей, попадающих в участки известных инверсионных полиморфизмов.

Для того, чтобы оценить значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей в популяции DPGP3, мы сравнили наблюдаемые значения

σ^2/V_A со значениями σ^2/V_A в 1000 контрольных выборок синонимических аллелей с такими же популяционными частотами, как у вредных аллелей (Таблица 3.4). Контрольные выборки получали отдельно для каждой группы вредных аллелей – нонсенс аллелей, аллелей, вызывающих поломку сайта сплайсинга, и общей категории аллелей, вызывающих потерю функции гена. Аналогичным образом, для каждой группы вредных аллелей мы получили 1000 контрольных выборок несинонимических аллелей. При проведении этой процедуры мы руководствовались тем, что эффекты, связанные с популяционной структурой и техническими артефактами, могут вносить сравнимый вклад в σ^2/V_A для всех типов аллельных вариантов. При этом ожидается, что эффекты, связанные с действием отбора и эпистазом, должны быть наиболее сильно выражены для наиболее вредных аллелей и наименее сильно – для нейтральных вариантов. Таким образом, сопоставление σ^2/V_A для вредных аллелей и нейтральных алеллей можно использовать для того, чтобы определить, связано ли понижение σ^2/V_A с действием отбора и эпистатическими взаимодействиями. В использованной нами классификации в качестве наиболее вредных рассматриваются аллели, вызывающие потерю функции гена, а в качестве наиболее нейтральных – синонимические аллели. Несинонимические аллели занимают промежуточное положение, поскольку в отличие от синонимических аллелей, изменяют последовательность белка, но в среднем должны оказывать меньший эффект на приспособленность по сравнению с аллелями, приводящими к возникновению преждевременного стоп-кодона в мРНК [139,140].

Популяция DPGP3, весь геном					
Частота минорного аллеля ≤5					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	3,577.06	3,544.10	203,688.98	57.473	
Несинонимические	2,051.52	2,036.36	37,746.26	18.536	
Нонсенс	10.21	10.15	9.42	0.928	0.002
Поломка сайта сплайсинга	2.60	2.59	2.45	0.948	0.177
Потеря функции	12.81	12.74	11.84	0.929	<0.001
Частота минорного аллеля <50%					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	50,666.23	37,671.02	1,418,266.56	37.649	
Несинонимические	11,980.92	9,520.01	25,950.18	2.726	
Нонсенс	25.84	22.89	20.44	0.893	0.010
Поломка сайта сплайсинга	9.62	7.96	6.94	0.872	0.070
Потеря функции	35.46	30.85	26.26	0.851	0.001

Таблица 3.4. Мутационная нагрузка минорных аллелей разных типов в африканской

популяции *D. melanogaster* DPGP3. Мутационную нагрузку рассчитывали для всего генома для минорных синонимических, несинонимических, нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, а также для общей категории аллелей, вызывающих потерю функции гена. Анализ проводили для редких минорных алеллей (встречающихся не более чем в 5 особях из 191 особи в популяции DPGP3) и для всех минорных аллелей (встречающихся менее чем в 50% особей в данной популяции). Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A). Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей (нонсенс-аллелей, аллелей, вызывающих поломку сайтов сплайсинга, и аллелей, вызывающих потерю функции гена) оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

Как обсуждалось выше, дисперсия числа мутаций на геном (σ^2) в популяции в начале каждого поколения может быть представлена как сумма аддитивной дисперсии (V_A) и остаточного члена (C_L) [12,49]:

$$\sigma^2 = V_A + C_L \qquad (3.2)$$

Остаточный член, C_L, соответствует сумме попарных ковариаций C_{ij} для каждой пары локусов i и j, отражающих неравновесие по сцеплению для данной пары локусов (см. Bulmer, 1980 [49] и Charlesworth, 1990 [12]). Так, в том случае если наблюдается положительное неравновесие по сцеплению между аллелями в разных локусах ($C_{ij} > 0$), C_L будет увеличиваться с добавлением каждой новой пары локусов, поскольку каждая пара локусов вносит вклад в C_L. Поэтому разница между σ^2 и V_A для распределения числа аллелей на геном зависит от общего числа рассматриваемых локусов (см. Bulmer, 1980 [49] и Charlesworth, 1990 [12]). В связи с этим напрямую сравнивать значения σ^2/V_A для разных типов аллелей некорректно, т.к. каждая особь несет гораздо больше несинонимических и синонимических аллелей, чем нонсенс-аллелей. Получение случайных выборок несинонимических и синонимических аллелей такого же размера как набор аллелей, вызывающих потерю функции гена, и с популяционными частотами, соответствующими частотам аллелей, вызывающих потерю функции гена, позволяет решить эту проблему.

С помощью процедуры получения случайных контрольных выборок мы показали, что понижение дисперсии ($\sigma^2/V_A < 1$) значимо для нонсенс-аллелей и общей категории аллелей,

вызывающих потерю функции гена (P-значение < 0.05; Рисунок 3.2, Таблица 3.4). Это позволяет сделать заключение о том, что понижение дисперсии по сравнению с ожиданием для вредных аллелей имеет отборную природу и, по всей видимости, объясняется синергическими эпистатическими взаимодействиями между вредными аллелями.

P-значение, рассчитанное отдельно для аллелей, вызывающих поломку сайтов сплайсинга, оказалось не значимо (P-значение > 0.05), что может быть связано со слишком маленьким числом таких аллелей в индивидуальных геномах.

Значения σ^2/V_A для большинства выборок несинонимических аллелей выше σ^2/V_A для вредных аллелей, но в среднем ниже, чем значения σ^2/V_A для синонимических аллелей (Рисунок 3.2, Таблица 3.4). Смещение значений σ^2/V_A в сторону более низких значений для несинонимических аллелей по сравнению с синонимическими аллелями может указывать на существование синергических эпистатических взаимодействий не только между наиболее вредными аллелями, приводящими к потере функции гена, но и между менее вредными несинонимическими аллелями. Действительно, данное наблюдение может соответствовать ситуации, в которой синергический эпистаз наиболее распространен среди наиболее вредных мутаций, но вносит вклад и в отбор против менее вредных несинонимических мутаций. Однако поскольку распределение числа несинонимических аллелей на геном в целом характеризуется сверхдисперсией ($\sigma^2/V_A > 1$, Таблица 3.4), более низкие значения σ^2/V_A для несинонимических аллелей по сравнению с синонимическими не являются достаточным основанием для того, чтобы утверждать о существовании синергических эпистатических взаимодействий между несинонимическими аллелями.

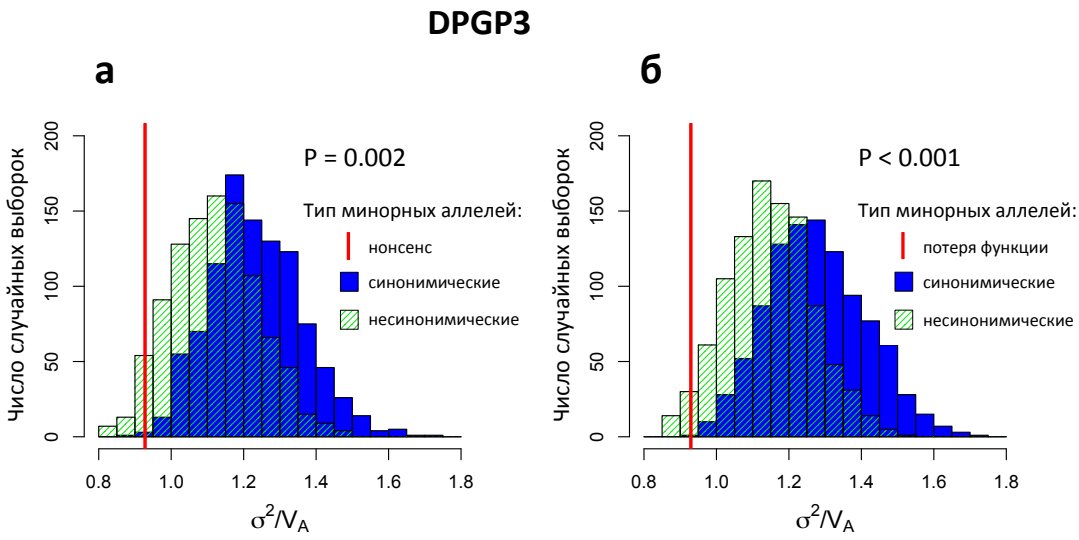

Рисунок 3.2. Распределение значений σ^2/V_A в 1000 случайных контрольных выборок синонимических (синий) и несинонимических (зеленый) минорных аллелей с популяционными частотами, соответствующими популяционным частотам нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**), полученное для популяции африканских мух DPGP3. Значение σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**), отмечено красной вертикальной линией. Одностороннее P-значение рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**). В анализ включены аллели, встречающиеся не более чем в 5 особях из 191.

Анализ мутационной нагрузки в популяции DGRP, проведенный на полногеномных данных для аллелей, попадающих в любые участки генома, показал, что распределение мутационной нагрузки в этой популяции характеризуется сверхдисперсией ($\sigma^2/V_A > 1$) для всех основных типов аллелей, включая нонсенс-аллели и аллели, вызывающие потерю функции гена (Рисунок 3.3, Таблица 3.5). Однако значения σ^2/V_A для нонсенс-аллелей и аллелей, вызывающих потерю функции гена, были лишь незначительно выше 1 ($\sigma^2/V_A = 1.097$ и 1.124 для редких минорных нонсенс-аллелей и редких минорных аллелей, вызывающих потерю функции гена, соответственно). В то же время значения σ^2 для несинонимических и синонимических минорных аллелей, сегрегирующих среди особей в популяции DGRP, оказались выше V_A более чем в 40 раз и более чем в 100 раз соответственно ($\sigma^2/V_A = 46.330$ и 153.976 для редких минорных несинонимических и синонимических аллелей соответственно).

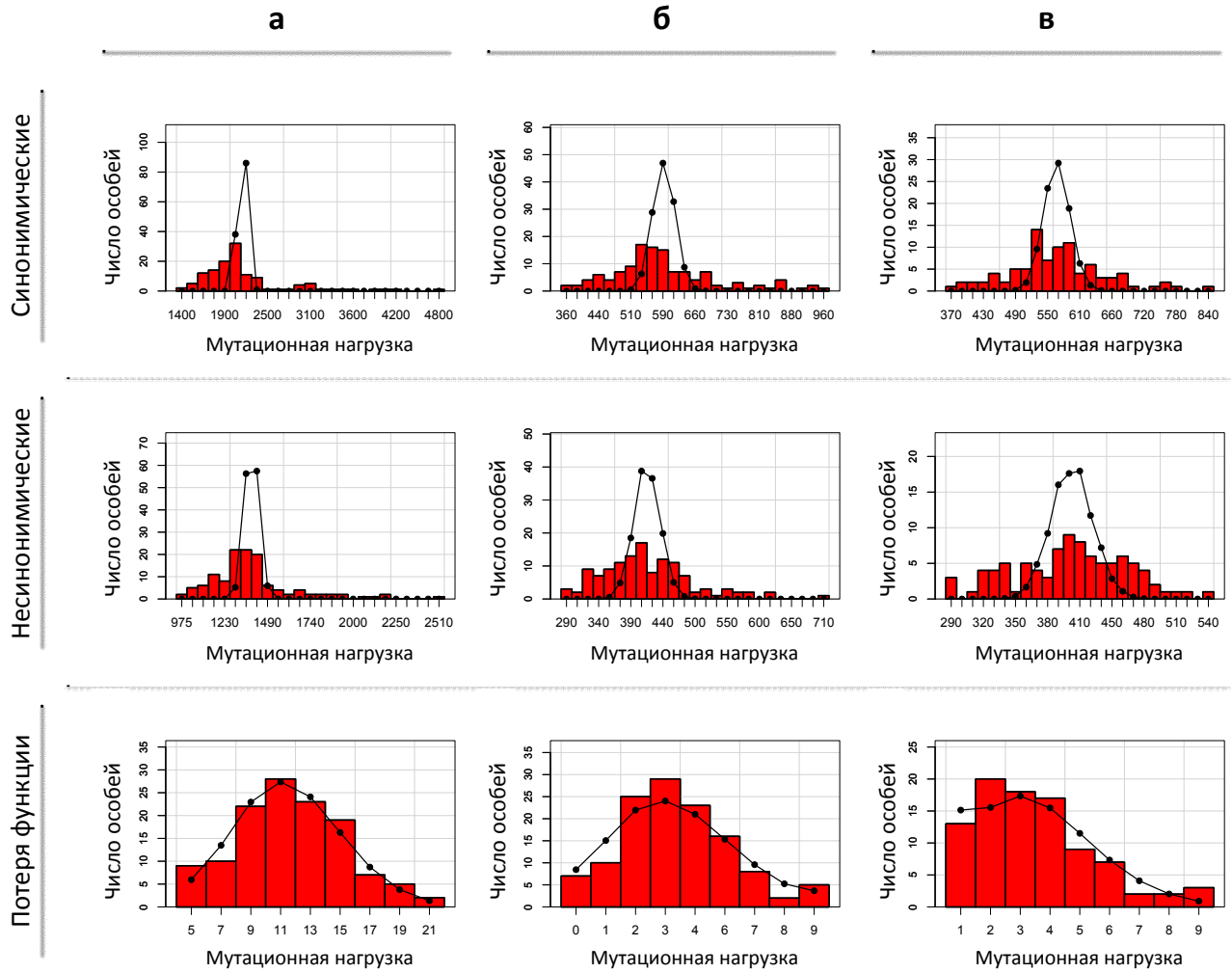

Рисунок 3.3. Мутационная нагрузка в популяции североамериканских мух DGRP. Показано распределение числа редких минорных аллелей на особь в популяции DGRP для синонимических, несинонимических аллелей и аллелей, вызывающих потерю функции гена. Тип аллелей обозначен слева. В анализ включены аллели, встречающиеся не более чем в 5 особях в популяции DGRP. Черная линия соответствует ожидаемому числу особей, имеющему заданное число аллелей в геноме в случае, если бы число аллелей на геном особи имело распределение Пуассона с таким же средним, как в данных. (**а**) Мутационная нагрузка была рассчитана на основе всех доступных для анализа редких минорных аллелей. (**б**) Мутационная нагрузка была рассчитана на основе редких минорных аллелей, оставшихся после исключения аллелей, попадающих в участки известных инверсионных полиморфизмов. (**в**) Мутационная нагрузка была рассчитана для особей, в геномах которых отсутствуют известные инверсионные полиморфизмы, на основе редких минорных аллелей, оставшихся после исключения аллелей, попадающих в участки известных инверсионных полиморфизмов.

Мы показали, что σ^2/V_A для вредных аллелей в популяции DGRP действительно значимо ниже, чем эмпирическое ожидание σ^2/V_A, полученное на основе менее вредных типов аллелей с

таким же популяционными частотами (Рисунок 3.4, Таблица 3.5). Для этого мы получили 1000 контрольных выборок синонимических и несинонимических аллелей, сегрегирующих в популяции DGRP с популяционными частотами, соответствующими популяционным частотам вредных аллелей. Данная процедура была проведена аналогично тому, как это было выполнено для популяции DPGP3. Сравнение σ^2/V_A для аллелей разных типов показало, что в популяции DGRP σ^2/V_A для нонсенс-аллелей и аллелей, вызывающих потерю функции, значимо ниже по сравнению с σ^2/V_A в контрольных выборках синонимических аллелей (Рисунок 3.4, Таблица 3.5).

Популяция DGRP, весь геном					
Частота минорного аллеля ≤5					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	2,235.72	2,199.90	338,732.36	153.976	
Несинонимические	1,425.38	1,406.40	65,158.54	46.330	
Нонсенс	10.50	10.38	11.38	1.097	<0.001
Поломка сайта сплайсинга	2.61	2.58	2.85	1.108	0.258
Потеря функции	13.10	12.95	14.56	1.124	<0.001
Частота минорного аллеля <50%					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	43,834.42	31,725.65	6,867,351.96	216.461	
Несинонимические	12,048.18	9,160.84	441,121.62	48.153	
Нонсенс	38.34	31.36	32.60	1.039	0.002
Поломка сайта сплайсинга	10.22	8.39	7.08	0.843	0.012
Потеря функции	48.57	39.75	44.57	1.121	<0.001

Таблица 3.5. Мутационная нагрузка минорных аллелей разных типов в североамериканской популяции *D. melanogaster* DGRP. Мутационную нагрузку рассчитывали для всего генома для минорных синонимических, несинонимических, нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, а также для общей категории аллелей, вызывающих потерю функции гена. Анализ проводили для редких минорных алеллей (встречающихся не более чем в 5 особях из 125 особей в популяции DGRP) и для всех минорных аллелей (встречающихся менее чем в 50% особей в данной популяции). Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A). Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей (нонсенс-аллелей, аллелей, вызывающих поломку сайтов сплайсинга, и аллелей, вызывающих потерю функции гена) оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со

значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

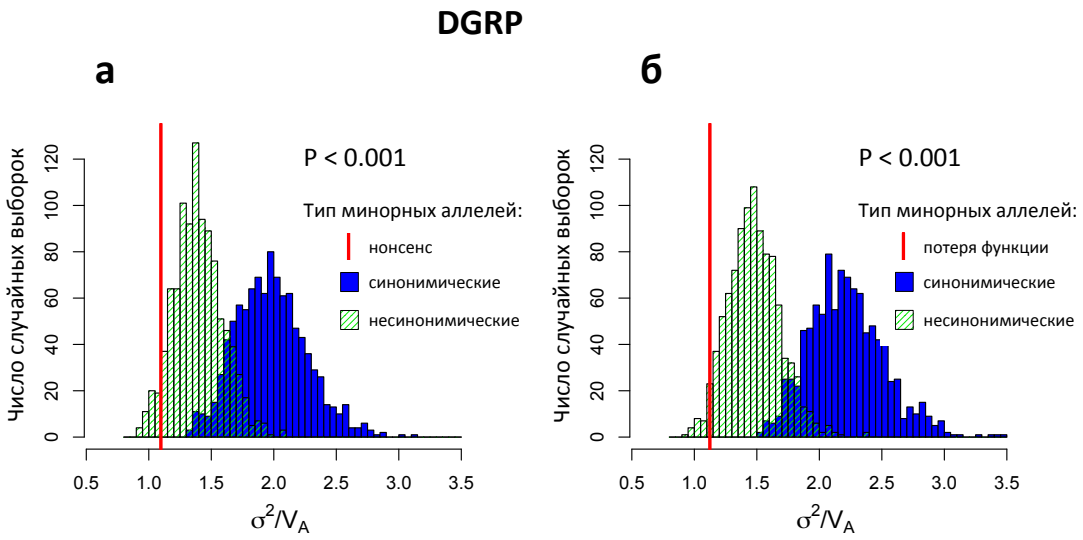

Рисунок 3.4. Распределение значений σ^2/V_A в 1000 случайных контрольных выборок синонимических (синий) и несинонимических (зеленый) минорных аллелей с популяционными частотами, соответствующими популяционным частотам нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**), полученное для популяции североамериканских мух DGRP. Значение σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**), отмечено красной вертикальной линией. Одностороннее P-значение рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**). В анализ включены аллели, встречающиеся не более чем в 5 особях из 125.

Так, для редких минорных нонсенс-аллелей и редких минорных аллелей, вызывающих потерю функции, ни в одной контрольной выборке синонимических аллелей из 1000 не наблюдалось значение σ^2/V_A ниже или равное значению σ^2/V_A для соответствующего типа вредных аллелей (P-значение < 0.001; Рисунок 3.4, Таблица 3.5). Кроме того, в популяции DGRP, как и в популяции DPGP3, распределение отношений σ^2/V_A в 1000 выборок несинонимических аллелей было сдвинуто в сторону более низких значений по сравнению с выборками синонимических аллелей (Рисунок 3.4). Эти результаты указывали на то, что в данных присутствует сигнал, указывающий на сужение распределения числа мутаций на геном для мутаций, понижающих приспособленность. Но, вероятно, что в популяции DGRP

существуют факторы, систематически повышающие дисперсию распределения числа аллелей на геном и препятствующие выявлению сигнала синергического эпистатического отбора ($\sigma^2/V_A < 1$).

Мы предположили, что исключение основных факторов, вносящих вклад в сверхдисперсию числа аллелей на геном в популяции DGRP, позволит выявить подпись синергического эпистатического отбора ($\sigma^2/V_A < 1$) для вредных аллелей и в этой популяции.

К повышению дисперсии распределения числа аллелей на геном в отсутствие эпистаза может приводить присутствие популяционной структуры, связанной, например, с примесью из генетически далеких популяций [138]. Ранее было показано, что полиморфные инверсии определяют значительную часть популяционной структуры в популяциях *D. melanogaster* [127,137]. Особенно сильно этот эффект выражен именно в североамериканской популяции *D. melanogaster*, в которой генетическая дифференциация между инвертированными и стандартными гаплотипами гораздо выше, чем в африканских популяциях [137]. Одним из возможных объяснений является интрогрессия инвертированных гаплотипов из генетически далеких популяций [137]. Кроме того, пониженная частота рекомбинации в области инверсии в особях гетерозиготных по присутствию инверсии приводит к накоплению замен между инвертированным и стандартным гаплотипом [151] и поддерживает неравновесие по сцеплению между полиморфными сайтами внутри инверсий, что может выражаться в сверхдисперсии распределения мутационной нагрузки.

Мы предположили, что сверхдисперсия распределения числа минорных аллелей на геном в популяции североамериканских мух DGRP может быть частично объяснена вкладом неравновесия по сцеплению между аллелями, попадающими в участки полиморфных инверсий. Действительно, после исключения областей инверсионных полиморфизмов распределение числа минорных аллелей на геном становится значительно уже (Рисунок 3.3). Стоит отметить, что этот эффект наблюдается не только в популяции DGRP (Рисунок 3.3), но и в популяции DPGP3 (Рисунок 3.1). Это верно для всех типов минорных аллелей, но особенно сильно выражено для синонимических и несинонимических минорных аллелей (Рисунок 3.3, Рисунок 3.1). Т.к. многие инверсии присутствуют в популяциях *D. melanogaster* на достаточно высокой частоте, геномы более половины особей, как в наборе данных DGRP, так и в наборе данных DPGP3, имеют хотя бы один инверсионный полиморфизм. В связи с этим ограничивать анализ только особями *D. melanogaster*, в геномах которых отсутствуют инвертированные участки, представляется нецелесообразным, т.к. в этом случае для анализа остаются доступны всего 73 особи в популяции DPGP3 и 91 особь в популяции DGRP. Однако стоит отметить, что если проводить анализ только на особях, в геномах которых отсутствуют известные инверсионные полиморфизмы, отклонения распределения мутационной нагрузки от распределения Пуассона

становятся еще менее выраженными (Рисунок 3.3, Рисунок 3.1). Ранее было показано, что инвертированные гаплотипы у *D. melanogaster*, по всей видимости, часто привнесены из генетически неродственных популяций [137]. Поэтому, вероятно, что присутствие инвертированного гаплотипа может указывать на более высокий уровень примеси из других популяций не только в области инвертированного участка, но и в других областях генома. Это могло бы объяснить сниженную дисперсию распределения мутационной нагрузки в участках генома свободных от известных инверсионных полиморфизмов среди особей без инверсионных полиморфизмов по сравнению с особями, имеющими такие полиморфизмы.

После исключения из рассмотрения в популяции DGRP аллелей, попадающих в участки инверсионных полиморфизмов, значение σ^2/V_A для нонсенс-аллелей и общей категории аллелей, вызывающих потерю функции гена, опустилось ниже 1 (Таблица 3.6, Рисунок 3.5). Данный эффект присутствует как в том случае, если в анализ включены все минорные полиморфизмы (аллельная частота <50%), так и в том случае, если рассматриваются только редкие минорные полиморфизмы (аллельная частота ≤5; Таблица 3.6). Так, для распределения числа редких минорных полиморфизмов, вызывающих потерю функции, $\sigma^2/V_A = 0.883$, если же рассматривать все минорные полиморфизмы, вызывающие потерю функции, $\sigma^2/V_A = 0.965$. Значимость этих результатов была подтверждена с помощью получения контрольных выборок синонимических аллелей с сопоставленными частотами (Таблица 3.6, Рисунок 3.5).

Популяция DGRP, участки генома свободные от известных инверсионных полиморфизмов					
Частота минорного аллеля ≤5					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	599.53	590.02	15,458.41	26.200	
Несинонимические	427.58	421.92	5,701.12	13.512	
Нонсенс	3.52	3.48	2.93	0.842	0.001
Поломка сайта сплайсинга	0.86	0.85	1.04	1.235	0.918
Потеря функции	4.38	4.32	3.82	0.883	0.014
Частота минорного аллеля <50%					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	10,720.41	7,857.28	519,790.61	66.154	
Несинонимические	3,381.59	2,601.40	47,844.78	18.392	
Нонсенс	13.21	10.68	10.62	0.995	0.132
Поломка сайта сплайсинга	3.70	3.12	3.42	1.094	0.600
Потеря функции	16.91	13.80	13.32	0.965	0.040

Таблица 3.6. Мутационная нагрузка минорных аллелей разных типов в североамериканской

популяции *D. melanogaster* DGRP после исключения участков известных инверсионных полиморфизмов. Мутационную нагрузку рассчитывали для участков генома *D. melanogaster* свободных от известных инверсионных полиморфизмов для минорных синонимических, несинонимических, нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, а также для общей категории аллелей, вызывающих потерю функции гена. Анализ проводили для редких минорных алеллей (встречающихся не более чем в 5 особях из 125 особей в популяции DGRP) и для всех минорных аллелей (встречающихся менее чем в 50% особей в данной популяции). Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A). Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей (нонсенс-аллелей, аллелей, вызывающих поломку сайтов сплайсинга, и аллелей, вызывающих потерю функции гена) оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

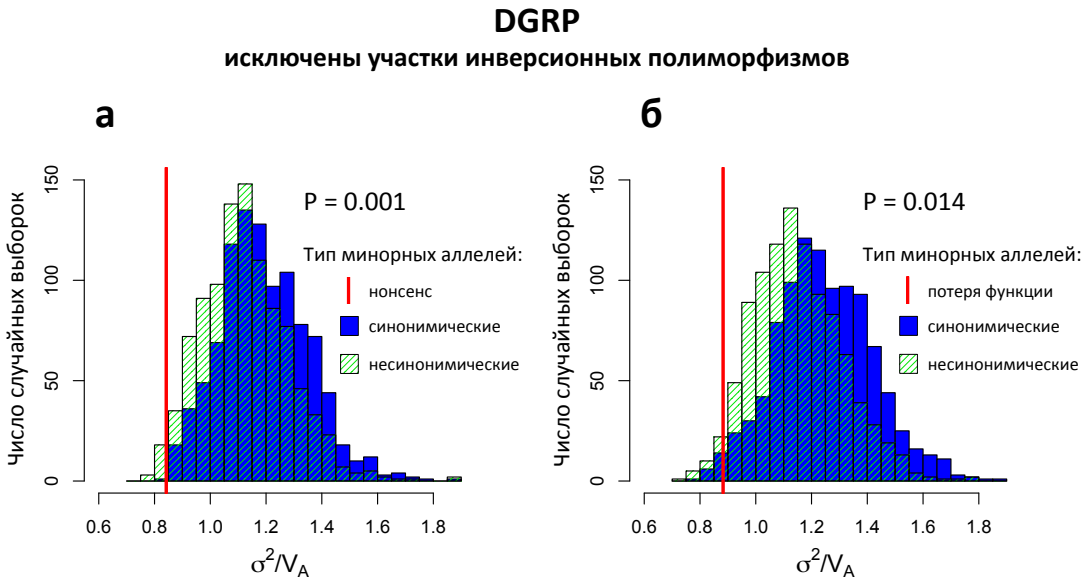

Рисунок 3.5. Распределение значений σ^2/V_A в 1000 случайных контрольных выборок синонимических (синий) и несинонимических (зеленый) минорных аллелей с популяционными частотами, соответствующими популяционным частотам нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**), полученное для популяции североамериканских мух DGRP после исключения аллелей, попадающих в участки инверсионных полиморфизмов. Значение σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**),

отмечено красной вертикальной линией. Одностороннее P-значение рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для нонсенс-аллелей (**а**) или аллелей, вызывающих потерю функции гена (**б**). В анализ включены аллели, встречающиеся не более чем в 5 особях из 125.

Таким образом, сигнал синергического эпистатического отрицательного отбора ($\sigma^2/V_A < 1$) против вредных мутаций был выявлен в обоих наборах по полногеномному полиморфизму *D. melanogaster* – наборе данных для африканской популяции *D. melanogaster* из Замбии (DPGP3) и наборе данных для североамериканской популяции (DGRP). В африканской популяции, характеризующейся низким уровнем примеси из неродственных геномов [128], сужение распределения мутационной нагрузки по сравнению с ожиданием наблюдается на полногеномных данных без дополнительных шагов, направленных на понижение влияния факторов, систематически завышающих дисперсию. В североамериканской популяции полногеномное распределение мутационной нагрузки для всех типов аллелей характеризуется сверхдисперсией ($\sigma^2/V_A > 1$), предположительно, связанной с примесью из неродственных популяций [129]. Исключение участков генома, находящихся внутри участков известных инверсионных полиморфизмов, приводит к значительному снижению отношения дисперсии к аддитивной дисперсии практически для всех типов аллелей (Таблицы 3.5 и 3.6). Вероятно, это связано с тем, что инвертированные гаплотипы привнесены в североамериканскую популяцию из генетически далеких популяций [137] и обогащены аллелями в сильном неравновесии по сцеплению друг с другом. Проведение анализа на областях генома, свободных от известных инверсионных полиморфизмов, позволяет снизить влияние факторов, завышающих дисперсию мутационной нагрузки, и выявить сигнал синергических эпистатических взаимодействий ($\sigma^2/V_A < 1$), для вредных аллелей.

В связи с высоким уровнем генетической изменчивости в африканской популяции *D. melanogaster* заметная часть кодонов имеет более чем два варианта алеллей, присутствующих среди особей DPGP3 (8% вариабельных кодонов являются мультиаллельными). Для того, чтобы учесть возможное влияние присутствия сайтов с множественными аллелями на результаты анализа, мы дополнительно рассчитали мутационную нагрузку в обеих популяциях для каждого функционального класса минорных аллелей (нонсенс-аллели, несинонимические, синонимические аллели) после исключения кодонов, для которых было найдено более двух аллелей, принадлежащих к данному функциональному классу (Таблица 3.7, Таблица 3.8). Аналогичным образом для этого анализа исключали мультиаллельные сайты сплайсинга (Таблица 3.7, Таблица 3.8). Исключение сайтов с множественными аллелями не повлияло на направление наблюдаемого эффекта.

Популяция DPGP3, весь геном, исключены множественные аллели					
Частота минорного аллеля ≤5					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	3,291.65	3,261.33	171,627.86	52.625	
Несинонимические	1,829.30	1,815.92	28,979.27	15.958	
Нонсенс	10.14	10.08	9.49	0.941	0.008
Поломка сайта сплайсинга	2.28	2.26	2.20	0.972	0.247
Потеря функции	12.41	12.34	11.89	0.963	0.002
Частота минорного аллеля <50%					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	47,075.73	34,921.77	1,214,507.09	34.778	
Несинонимические	10,555.57	8,374.02	22,616.09	2.701	
Нонсенс	25.15	22.37	19.87	0.889	0.012
Поломка сайта сплайсинга	9.14	7.50	6.56	0.875	0.070
Потеря функции	34.29	29.86	25.87	0.866	0.002

Таблица 3.7. Мутационная нагрузка минорных аллелей разных типов в африканской популяции *D. melanogaster* DPGP3 после исключения сайтов с множественными аллелями. Мутационную нагрузку рассчитывали для всего генома для минорных синонимических, несинонимических, нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, а также для общей категории аллелей, вызывающих потерю функции гена. Анализ проводили для редких минорных алеллей (встречающихся не более чем в 5 особях из 191 особи в популяции DPGP3) и для всех минорных аллелей (встречающихся менее чем в 50% особей в данной популяции). Из рассмотрения были исключены кодоны, представленные более чем двумя аллелями, принадлежащими к одному функциональному классу, и сайты сплайсинга с множественными аллелями. Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A). Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей (нонсенс-аллелей, аллелей, вызывающих поломку сайтов сплайсинга, и аллелей, вызывающих потерю функции гена) оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

Популяция DGRP, участки генома свободные от известных инверсионных полиморфизмов, исключены множественные аллели					
Частота минорного аллеля ≤5					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	585.22	575.94	14,801.49	25.700	
Несинонимические	413.12	407.67	5,301.07	13.003	
Нонсенс	3.52	3.48	2.93	0.842	0.005
Поломка сайта сплайсинга	0.86	0.85	1.04	1.235	0.920
Потеря функции	4.38	4.32	3.82	0.883	0.003
Частота минорного аллеля <50%					
Тип минорных аллелей	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**P-значение**
Синонимические	10,499.92	7,689.47	501,206.46	65.181	
Несинонимические	3,279.46	2,519.34	45,407.98	18.024	
Нонсенс	13.21	10.68	10.62	0.995	0.136
Поломка сайта сплайсинга	3.70	3.12	3.42	1.094	0.648
Потеря функции	16.91	13.80	13.32	0.965	0.038

Таблица 3.8. Мутационная нагрузка минорных аллелей разных типов в североамериканской популяции *D. melanogaster* DGRP после исключения участков известных инверсионных полиморфизмов и сайтов с множественными аллелями. Мутационную нагрузку рассчитывали для участков генома *D. melanogaster* свободных от известных инверсионных полиморфизмов для минорных синонимических, несинонимических, нонсенс-аллелей и аллелей, вызывающих поломку сайтов сплайсинга, а также для общей категории аллелей, вызывающих потерю функции гена. Анализ проводили для редких минорных алеллей (встречающихся не более чем в 5 особях из 125 особей в популяции DGRP) и для всех минорных аллелей (встречающихся менее чем в 50% особей в данной популяции). Из рассмотрения были исключены кодоны, представленные более чем двумя аллелями, принадлежащими к одному функциональному классу, и сайты сплайсинга с множественными аллелями. Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A). Значимость понижения дисперсии по отношению к аддитивной дисперсии для вредных аллелей (нонсенс-аллелей, аллелей, вызывающих поломку сайтов сплайсинга, и аллелей, вызывающих потерю функции гена) оценивали путем сравнения значения σ^2/V_A для данной категории вредных аллелей со значениями σ^2/V_A в 1000 случайных выборок синонимических аллелей с таким же

распределением популяционных частот. Односторонние P-значения для вредных аллелей рассчитывали как долю 1000 случайных выборок синонимических аллелей, в которых значение σ^2/V_A было ниже или равно значению σ^2/V_A для соответствующей категории вредных аллелей.

3.2.2 Анализ распределения мутационной нагрузки для несинонимических аллелей, попадающих в «необходимые» гены

Показав присутствие сигнала синергических эпистатических взаимодействий для наиболее вредного класса аллелей – аллелей, приводящих к потере функции гена, мы перешли к поиску возможного сигнала синергического эпистаза между менее вредными несинонимическими аллелями. На предыдущих этапах анализа мы отметили, что значения σ^2/V_A для несинонимических аллелей смещены в сторону более низких значений по сравнению с синонимическими аллелями в обеих популяциях *D. melanogaster* (Рисунки 3.2, 3.4 и 3.5). Мы предположили, что это явление может объясняться действием отрицательного отбора. Основанием для данного предположения является то, что факторы, связанные с популяционной структурой и техническими артефактами, должны оказывать схожее влияние на синонимические и несинонимические аллели. Однако поскольку отбор против несинонимических мутаций значительно сильнее отбора против синонимических мутаций, которые часто рассматривают как нейтральные, более низкая дисперсия распределения мутационной нагрузки для несинонимических аллелей скорее всего имеет отборную природу.

Наши коллеги в симуляциях показали, что даже в отсутствие эпистаза при наличии факторов, завышающих дисперсию распределения мутационной нагрузки, сверхдисперсия этого распределения падает с увеличением силы отрицательного отбора против мутаций (см. Sohail M., Vakhrusheva O. *et al*., 2017 [152]). Учитывая это, мы проанализировали распределение мутационной нагрузки для аллелей, попадающих в «необходимые» гены *D. melanogaster*, предположив, что несинонимические мутации в таких генах будет обогащены мутациями под сильным отрицательным отбором. В набор генов, использованный в данной работе, входит 267 генов, проаннотированных как «необходимые» согласно базе данных DEG (см. Материалы и методы).

Значительная сверхдисперсия распределения мутационной нагрузки, наблюдаемая в популяции DGRP даже для наиболее вредных аллелей, делает целесообразным проведение анализа в этой популяции только в областях генома свободных от инверсионных полиморфизмов. Однако из 267 «необходимых» генов только 88 не пересекаются с областями известных инверсионных полиморфизмов. В связи с этим анализ дисперсии распределения мутационной нагрузки в «необходимых» генах в первую очередь проводили для основного набора данных – DPGP3, для которого возможно целиком рассматривать множество

полиморфизмов, попадающих в «необходимые» гены (см. Материалы и методы).

Мы показали, что в популяции DPGP3 дисперсия распределения числа несинонимических аллелей, попадающих в «необходимые» гены, ниже аддитивной дисперсии (σ^2/V_A = 0.947, Рисунок 3.6). В то же время для синонимических аллелей в «необходимых» генах в популяции DPGP3 по-прежнему наблюдается сверхдисперсия (σ^2/V_A = 2.125, Рисунок 3.6). В популяции DGRP понижения дисперсии по сравнению с аддитивной дисперсией для аллелей, попадающих в «необходимые» гены, выявлено не было (σ^2/V_A = 2.729 и σ^2/V_A = 13.445 для несинонимических и синонимических аллелей соответственно).

Мы оценили значимость понижения дисперсии по сравнению с ожиданием для несинонимических аллелей, попадающих в «необходимые» гены, в популяции DPGP3 с помощью сравнения σ^2/V_A для аллелей этого класса с распределением значений σ^2/V_A в контрольных выборках синонимических аллелей из случайных генов и в контрольных выборках синонимических аллелей из «необходимых» генов (см. Материалы и методы). Данный анализ показал, что распределение числа несинонимических аллелей, попадающих в «необходимые» гены, на геном имеет значимо более низкую дисперсию как по сравнению с выборками синонимических аллелей из любого участка генома (P = 0.002), так и по сравнению с выборками синонимических аллелей, попадающих в «необходимые» гены ($P < 10^{-3}$).

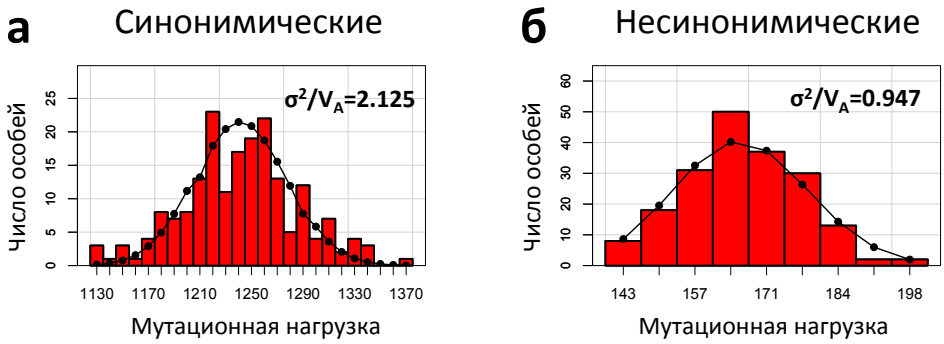

Рисунок 3.6. Мутационная нагрузка в «необходимых» генах в популяции африканских мух DPGP3. Мутационную нагрузку рассчитывали на основе всех минорных аллелей (популяционная частота до 50%), попадающих в гены, проаннотированные как «необходимые» согласно базе данных DEG. Мутационная нагрузка синонимических (**а**) и несинонимических (**б**) аллелей в «необходимых» генах *D. melanogaster* в популяции африканских мух DPGP3. Наблюдаемое распределение числа аллелей на геном (красный) и ожидаемое в случае независимости аллелей в разных локусах (черная линия, соответствующая распределению Пуассона с таким же средним, как в данных).

3.2.3 Анализ распределения мутационной нагрузки для аллелей, попадающих в гены с разным отношением скоростей несинонимической и синонимической эволюции

Затем мы провели более общий анализ, изучив, как дисперсия распределения мутационной нагрузки для аллелей разных типов зависит от силы отрицательного отбора, действующего на ген. Для этого мы разделили гены в соответствии со значением отношения скоростей несинонимической и синонимической эволюции (dN/dS) в шести видах из подгруппы *melanogaster* рода *Drosophila* (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba*, *D. erecta*, *D. ananassae*) [148] на 5 групп одинакового размера (см. Материалы и методы). Отношение σ^2/V_A рассчитывали отдельно для каждой группы dN/dS для всех минорных (популяционная частота <50%) нонсенс-полиморфизмов, несинонимических и синонимических полиморфизмов.

Полученные значения σ^2/V_A приведены в Таблице 3.9. Мы показали, что в популяции DPGP3 дисперсия распределения мутационной нагрузки для несинонимических аллелей, попадающих в гены, белковая последовательность которых находится под наиболее сильным отрицательным отбором (группа dN/dS 1), ниже аддитивной дисперсии (σ^2/V_A = 0.98). Этот результат находится в соответствии с пониженной дисперсией для несинонимических аллелей, попадающих в «необходимые» гены, в этой популяции (σ^2/V_A = 0.947). Понижение дисперсии относительно аддитивной дисперсии для аллелей, вызывающих потерю функции, наблюдается в обеих популяциях для генов с низкой скоростью белковой эволюции (группы dN/dS 1–3 в случае популяции DPGP3 и группы dN/dS 1–2 в популяции DGRP).

		Тип минорных аллелей											
		Синонимические				Несинонимические				Потеря функции			
Популяция	**группа dN/dS**	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**	**Среднее число аллелей на особь**	**V_A**	**Дисперсия (σ^2)**	**σ^2/V_A**
Африканская популяция *D. melanogaster* (DPGP3)													
DPGP3	1	5,839.39	4,338.99	30,881.30	7.12	308.33	257.52	251.34	0.98	1.07	0.98	0.87	0.89
	2	7,056.41	5,232.02	43,371.70	8.29	826.86	677.29	965.88	1.43	1.74	1.51	1.40	0.93
	3	6,699.23	4,980.20	33,678.08	6.76	1,229.52	995.28	1,361.40	1.37	1.77	1.66	1.58	0.95
	4	5,874.95	4,367.07	24,783.37	5.68	1,651.42	1,313.24	1,654.44	1.26	3.07	2.74	3.05	1.11
	5	4,716.83	3,516.84	15,961.11	4.54	2,471.01	1,926.31	3,515.06	1.82	5.61	4.58	4.87	1.06
Североамериканская популяция *D. melanogaster* (DGRP)													
DGRP	1	5,056.05	3,649.62	113,375.16	31.06	369.14	292.78	639.58	2.18	0.97	0.87	0.79	0.91
	2	6,313.25	4,546.81	160,073.53	35.21	935.93	722.93	3,209.08	4.44	1.76	1.41	1.38	0.97
	3	5,759.19	4,173.72	154,200.54	36.95	1,297.98	998.44	7,034.30	7.05	2.49	2.23	2.27	1.02
	4	5,140.92	3,719.95	97,228.15	26.14	1,661.77	1,262.46	11,103.63	8.80	3.93	3.37	4.20	1.25
	5	3,855.64	2,789.34	46,692.78	16.74	2,223.78	1,663.24	16,708.90	10.05	5.89	4.64	3.84	0.83

Таблица 3.9. Мутационная нагрузка минорных аллелей в двух популяциях *D. melanogaster* в генах с разной скоростью эволюции белковой последовательности. Гены были разделены на 5 групп одинакового размера в соответствии со значением dN/dS, определенным для подгруппы *melanogaster* рода *Drosophila*, таким образом, что группа 1 содержит гены с наиболее низкой скоростью белковой эволюции, а группа 5 – с наиболее высокой скоростью белковой эволюции.

Для каждой группы в обеих популяциях рассчитывали нагрузку минорных синонимических аллелей, несинонимических аллелей и аллелей, вызывающих потерю функции гена. В анализ включены все минорные аллели, встречающиеся менее чем в 50% особей в данной популяции. Показано среднее число аллелей на особь, аддитивная дисперсия (V_A) и дисперсия (σ^2) распределения числа аллелей на особь, а также отношение дисперсии к аддитивной дисперсии (σ^2/V_A).

В целом, отношение σ^2/V_A для распределения числа на геном аллелей, вызывающих потерю функции, и несинонимических аллелей имеет тенденцию увеличиваться с ростом dN/dS, то есть с ослаблением интенсивности действия отрицательного отбора на ген (Таблица 3.9). В то же время для синонимических аллелей подобной зависимости не наблюдается (Таблица 3.9). Однако данная зависимость может объясняться не только эффектами отбора, но и ростом числа локусов с несинонимическими полиморфизмами, наблюдаемым при переходе от группы генов с наиболее низким значением dN/dS к группам с более высокими значениями dN/dS (Таблица 3.9).

Для того, чтобы явным образом учитывать влияние числа рассматриваемых полиморфных локусов и популяционных частот аллелей на дисперсию распределения, мы получили контрольные выборки несинонимических и синонимических аллелей из генов, эволюционирующих с разной скоростью. Все выборки имели такое же распределение частот и, соответственно, такое же среднее число аллелей на геном, как подмножество несинонимических аллелей из группы с наиболее низкими значениями dN/dS (группа dN/dS 1, Рисунок 3.7). Контрольные выборки несинонимических аллелей были получены для генов, принадлежащих к группам dN/dS с номерами 2, 3, 4 и 5, а контрольные выборки синонимических аллелей – для генов, принадлежащих к группам dN/dS с номерами 1, 2, 3, 4 и 5. Для каждой группы dN/dS были получены 1000 контрольных выборок, на основании которых рассчитывали медианное значение и 95% доверительный интервал для σ^2/V_A. Данную процедуру проводили независимо для синонимических и несинонимических аллелей.

Сравнение значений σ^2/V_A для сопоставленных по популяционным частотам выборок аллелей из генов, эволюционирующих с разными скоростями, показало, что в случае несинонимических аллелей σ^2/V_A падает с уменьшением dN/dS, то есть с увеличением силы отрицательного отбора на ген в обеих популяциях (Рисунок 3.7). В то же время в случае синонимических аллелей подобная зависимость отсутствует.

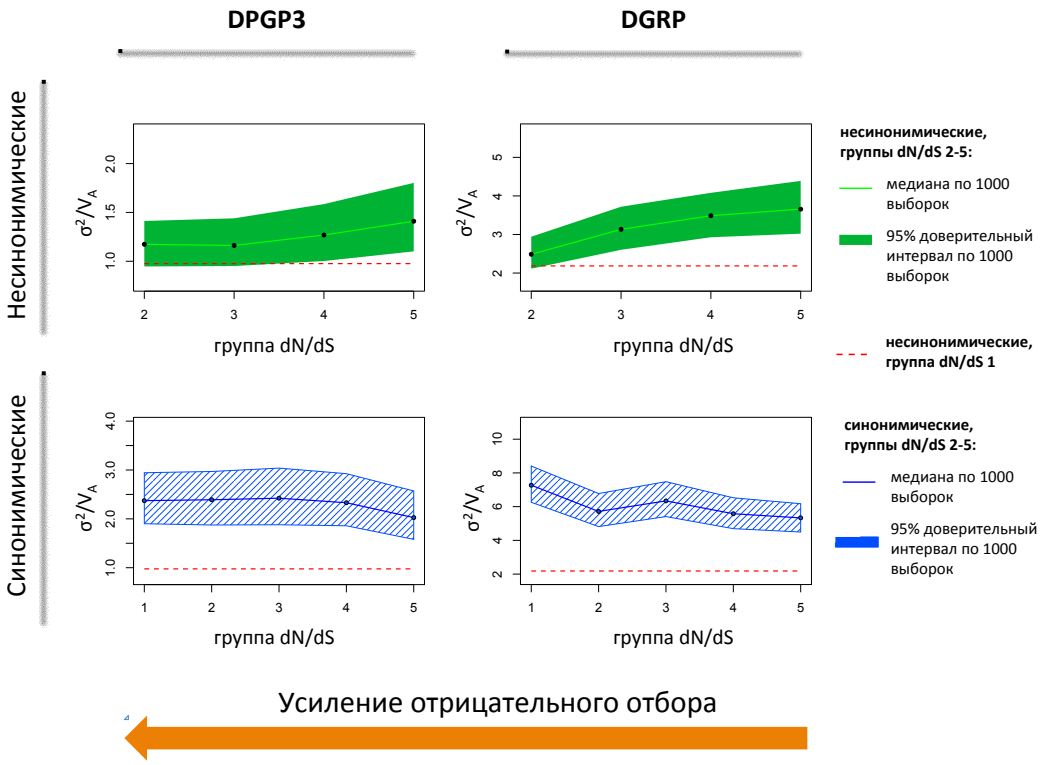

Рисунок 3.7. σ^2/V_A для мутационной нагрузки несинонимических и синонимических аллелей в двух популяциях *D. melanogaster* в генах с разной скоростью эволюции белковой последовательности. Мутационную нагрузку рассчитывали на основе всех минорных аллелей (популяционная частота до 50%). Гены были разделены на 5 групп одинакового размера в соответствии со значением dN/dS, определенным для подгруппы *melanogaster* рода *Drosophila*, таким образом, что группа 1 содержит гены с наиболее низкой скоростью белковой эволюции, а группа 5 – с наиболее высокой скоростью белковой эволюции. Несинонимические аллели в группах 2, 3, 4, 5 (верхняя часть рисунка) и синонимические аллели в группах 1, 2, 3, 4, 5 (нижняя часть рисунка) были использованы для получения случайных выборок с популяционными частотами, соответствующими популяционным частотам множества несинонимических аллелей в группе 1. Значение σ^2/V_A для несинонимических аллелей в группе dN/dS 1 показано с помощью красной пунктирной линии. Для каждой группы dN/dS черной точкой отмечено медианное значение σ^2/V_A в 1000 случайных контрольных выборок аллелей из этой группы. 95% доверительные интервалы, построенные по случайным выборкам, отображены с помощью многоугольников зеленого (для несинонимических аллелей) и синего (для синонимических аллелей) цвета.

Данный результат свидетельствует об отборной природе понижения дисперсии для несинонимических аллелей в генах, находящихся под действием сильного отрицательного

отбора. Как обсуждалось выше, ожидается, что факторы, завышающие дисперсию распределения мутационной нагрузки, в том числе популяционная структура, будут вносить наибольший вклад в дисперсию для мутаций, находящихся под наиболее слабым отбором. Это может объясняться тем, что нейтральные аллели и аллели, находящиеся под слабым отбором, дольше сегрегируют в популяциях по сравнению с аллелями, находящимися под сильным отбором [153]. В связи с этим различные демографические процессы в истории популяции будут дольше заметны при анализе аллелей под слабым отбором. И, напротив, более молодые вредные аллели в меньшей степени отражают историю популяции.

Полученный нами результат находится в соответствии с данным предсказанием. Кроме того, возможно, что наиболее сильно эпистатические взаимодействия выражены между наиболее вредными аллелями – в случае несинонимических полиморфизмов, это полиморфизмы в генах с наиболее низкой скоростью белковой эволюции. Однако имеющиеся данные не позволяют ответить на вопрос о том, существуют ли эпистатические взаимодействия между несинонимическими аллелями в не самых консервативных генах (группы dN/dS 2–5).

Таким образом, в данной главе мы провели поиск подписей синергического эпистатического отбора на вредные аллели в двух популяциях *D. melanogaster*. Мы показали присутствие подписи синергических эпистатических взаимодействий у *D. melanogaster* для аллелей, вызывающих потерю функции гена, а также для подмножества несинонимических аллелей, попадающих в гены, аминокислотная последовательность которых находится под сильным отрицательным отбором. Одновременно наши коллеги показали присутствие сигнала синергического эпистаза ($\sigma^2/V_A < 1$) для редких вредных аллелей на 6 наборах данных по полиморфизму человека (см. Sohail M., Vakhrusheva O. *et al*., 2017 [152]). Использованный нами подход основан на ожидаемом снижении дисперсии распределения мутационной нагрузки в присутствии синергических эпистатических взаимодействий по сравнению с ситуацией, в которой вредные мутации оказывают влияние на приспособленность независимо друг от друга. В отличие от традиционных подходов к поиску эпистатических взаимодействий, примененная нами методика не требует экспериментального измерения приспособленности, что значительно расширяет множество организмов, для которых она может быть использована. Присутствие синергических эпистатических взаимодействий между вредными аллелями у таких разных с точки зрения популяционной генетики видов как *D. melanogaster* и *H. sapiens* указывает на вероятную распространенность этого явления среди живых существ. Свидетельство в пользу существования синергического эпистаза на полногеномном уровне представляет большой интерес, поскольку может служить одним из объяснений парадокса мутационного груза.

Глава 4. Подписи рекомбинации и обмена генетическим материалом в популяции бделлоидных коловраток вида *A. vaga*

Половое размножение, включающее в себя чередование мейоза и слияния гамет, является преобладающим типом размножения среди эукариот. Несмотря на то, что переходы к бесполому размножению в эволюции эукариот случаются часто, бесполые линии обычно быстро вымирают [54,154]. В связи с этим отказ от полового размножения часто рассматривается как «эволюционный тупик». Однако существование предположительно древних групп бесполых видов ставит под вопрос эту точку зрения [53,64]. Список таких групп включает дарвинулид [57,58], орибатидных клещей [66], некоторые виды палочников [155] и бделлоидных коловраток [24,25,61,62], наиболее примечательную группу из этого списка. Бделлоидные коловратки – микроскопические пресноводные беспозвоночные, которые, как считалось в течение длительного времени, перешли к облигатному партеногенезу десятки миллионов лет назад.

В качестве основного доказательства того, что бделлоидные коловратки размножаются исключительно бесполым путем, обычно приводят свидетельства о том, что, несмотря на многолетние наблюдения, среди сотен тысяч проанализированных особей не было обнаружено ни одного самца [24]. В то же время геномные данные, опубликованные до недавнего времени, допускали различные интерпретации и не позволяли уверенно утверждать, что обмен генетическим материалом у видов этой группы никогда не происходит. Первоначальный анализ опубликованного в 2013 году генома бделлоидной коловратки *A. vaga* выявил признаки, указывающие на отсутствие гомологичных хромосом в геноме этого вида [25]. Это наблюдение интерпретировали как свидетельство в пользу того, что протекание мейоза по классическому типу у *A. vaga*, по всей видимости, невозможно [25]. Однако данные об атипичной структуре генома бделлоидных коловраток не нашли дальнейшего подтверждения: секвенирование генома близкого к *A. vaga* вида – *A. riccae* – не выявило аномальной структуры генома и перестроек, которые бы делали мейоз невозможным [26]. Отдельно стоит отметить, что анализ новых геномных данных для *A. vaga*, опубликованный в 2021 году, показал, что значительное количество геномных перестроек, выявленных в работе 2013 года, и отсутствие гомологичных хромосом в первой геномной сборке, по всей видимости, объясняется техническими артефактами [27]. Согласно новой модели геном *A. vaga* состоит из 6 пар гомологичных хромосом и не содержит каких-либо значительных перестроек, которые могли бы затруднить протекание мейоза [27].

В качестве свидетельства за или против древнего перехода к бесполому размножению могут в принципе использоваться данные о степени дивергенции между аллелями. Ожидается,

что отсутствие рекомбинации между гомологичными хромосомами в бесполых видах должно приводить к постепенному накоплению различий между двумя аллелями из одного локуса [76] и соответственно высокой дивергенции между аллелями (так называемый «эффект Мезельсона»). Однако оказалось, что геномы разных видов бделлоидных коловраток значительно отличаются по уровню такой дивергенции: оценки доли нуклеотидных различий между аллелями для разных видов находятся в широком диапазоне от ~0.03% до ~5% [26]. Таким образом, данные об организации генома бделлоидных коловраток не позволяют сделать вывод о существовании полового размножения и обмена генетическим материалом у видов этой группы.

Отсутствие самцов в проанализированных выборках бделлоидных коловраток не исключает возможности существования криптических форм обмена генетическим материалом в их популяциях [28,90–92]. Анализ последовательностей нескольких геномных локусов, представленный в двух работах, опубликованных в последние несколько лет, указывал на возможные события обмена генетическим материалом у бделлоидных коловраток [28,90]. Однако полногеномный анализ с целью поиска событий рекомбинации в этих работах проведен не был.

В этой главе мы проанализировали последовательности геномов 11 особей, принадлежащих к виду бделлоидных коловраток *Adineta vaga*, и показали, что структура популяционной изменчивости *A. vaga* несовместима с облигатным бесполым размножением. На вероятное существование рекомбинации у *A. vaga* указывает распад неравновесия по сцеплению между полиморфными сайтами с увеличением физического расстояния между ними. Свидетельства в пользу обмена генетическим материалом были получены в ходе анализа трехаллельных однонуклеотидных полиморфизмов и филогенетических деревьев гаплотипов, построенных для разных локусов генома. Кроме того, полученные данные полногеномного полиморфизма позволили заключить, что в популяции *A. vaga* не наблюдается значительных отклонений от равновесия Харди-Вайнберга, которые ожидаются в случае исключительно бесполого размножения. Результаты, представленные в этой главе, указывают на вероятное существование рекомбинации и обмена генетическим материалом в популяции бделлоидной коловратки *A. vaga*.

4.1 Материалы и методы

4.1.1 Получение клональных линий *A. vaga*

Для получения клональных линий *A. vaga* мы отбирали отдельных коловраток из образцов мха, собранных на стволах осины *Populus tremula* на высоте 120–170 см. Образцы мха собирали в двух географически удаленных точках. Первое место сбора находилось рядом с гидробиологической станцией «Глубокое озеро» в Рузском районе Московской области (здесь были собраны образцы, из которых впоследствии были получены девять линий: L1-L4 и L6-L10). Второе место сбора находилось поблизости от деревни Шилово в Костромской области (здесь были собраны образцы, из которых впоследствии были получены две линии: L5 и L11). Все линии, происходящие из одной географической локации, были выделены из образцов мха, собранных с деревьев, находящихся на расстоянии не менее 20 метров друг от друга.

Мы выделяли индивидуальных коловраток из образцов мха и определяли видовую принадлежность на основании морфологических критериев [156]. Затем мы отбирали особей, соответствующих по морфологическим критериям виду *A. vaga*. Такие особи использовались для получения клональных линий. Для этого индивидуальных коловраток промывали в воде Milli-Q и переносили в индивидуальные лунки 96-луночных планшетов для культуры клеток. Для того, чтобы убедиться в том, что в лунку действительно была перенесена только одна коловратка, мы просматривали лунки с помощью бинокуляра в течение трех дней после пересадки. Когда количество особей в культуре достигало ~30, культуру переносили в отдельную чашку Петри, содержащую воду Milli-Q. Культуры инкубировали при температуре 15–20 °C. В качестве корма для коловраток использовали бактерий *E. coli* (штамм DH5α), которых предварительно выращивали на среде LB при температуре 37°C на протяжении ночи. В конечном итоге мы получили 11 клональных линий *A. vaga*, которые в тексте работы обозначаются как L1-L11. Каждая из этих линий является потомством одной особи, соответствующей по морфологическим критериям виду *A. vaga*. Все 11 линий первоначально выделены из образцов мха, собранных с разных деревьев. Для того, чтобы подтвердить видовую принадлежность линий L1-L11, мы проверили число глоточных зубов у индивидуумов из каждой линии и убедились, что оно соответствует описанию для *A. vaga* (четыре крюка в форме буквы U с каждой стороны рта – отличительный признак *A. vaga* [156]). Когда общее число коловраток в чашке Петри достигало ~1000, из культуры выделяли ДНК. Определение видовой принадлежности бделлоидных коловраток и получение первичных клональных культур *A. vaga* было выполнено Е. А. Мнацакановой, а также Я. Р. Галимовым. Поскольку каждая из линий L1-L11 является потомством одной особи, соответствующей по

морфологическим критериям виду *A. vaga*, понятие «линия» в этой главе используется взаимозаменяемо с понятием «индивидуум» или «особь».

4.1.2 Выделение ДНК и подготовка библиотек для секвенирования

Геномную ДНК выделяли с использованием наборов Promega Wizard SV Genomic DNA Purification Kit (Promega, США) по протоколу производителя. ДНК фрагментировали с помощью соникатора Covaris S2. Библиотеки конструировали с использованием набора TruSeq DNA sample preparation kit (Illumina) с соблюдением протокола производителя. Выделение ДНК для части клональных культур *A. vaga* было проведено Т. В. Неретиной.

4.1.3 Секвенирование геномной ДНК

Геномную ДНК для 11 клональных линий *A. vaga* секвенировали с получением парно-концевых прочтений на инструментах Illumina HiSeq 2000 или 2500 с покрытием ~40–100×. Точное число парно-концевых прочтений и покрытие для каждой линии приведены в Таблице 4.1.

Кроме того, для того, чтобы получить *de novo* сборку генома *A. vaga*, мы секвенировали одну из линий, L1, на инструменте MiSeq (см. раздел 4.1.5). Исходя из необходимости минимизировать отличия в обработке и анализе данных для разных линий, мы использовали парно-концевые прочтения MiSeq, полученные для L1, только для сборки референтного генома *A. vaga*. Определение однонуклеотидных полиморфизмов для L1 проводилось так же, как и для остальных линий, с использованием выравниваний прочтений HiSeq. Получение геномных библиотек и секвенирование ДНК клональных культур *A. vaga* на инструментах Illumina было выполнено М. Д. Логачёвой и А. А. Пениным.

4.1.4 Первичная обработка и фильтрация парно-концевых прочтений

Первичная обработка и фильтрация парно-концевых прочтений была проведена с использованием программы Trimmomatic (V0.33) [157]. В частности, были удалены последовательности адаптеров, использованных при секвенировании, участки низкого качества прочтений и нуклеотиды с низким качеством на концах парно-концевых прочтений (параметры для фильтрации по качеству, переданные Trimmomatic: “LEADING:20 TRAILING:20 SLIDINGWINDOW:5:20”). Кроме того, была проведена фильтрация слишком коротких прочтений (короче 50 нуклеотидов). Данные по числу прочтений, оставшихся после этих процедур для каждой особи L1-L11, приведены в Таблице 4.1.

Особь	Общее число полученных парно-концевых прочтений	Общее число парно-концевых прочтений, оставшихся после фильтрации	% парно-концевых прочтений, оставшихся после фильтрации	Покрытие диплоидной сборки L1			Покрытие гаплоидной сборки			% нуклеотидов в гаплоидной сборке с нулевым покрытием
				Среднее покрытие на нуклеотид	Среднее покрытие на контиг	Медианное покрытие на контиг	Среднее покрытие на нуклеотид	Среднее покрытие на контиг	Медианное покрытие на контиг	
L1	45,954,860	43,213,727	94%	40.74	38.81	44.11	75.90	61.27	73.60	0.9%
L2	96,143,133	89,788,445	93%	78.57	71.76	80.17	149.93	116.40	135.94	1.7%
L3	70,332,648	62,339,933	89%	53.30	48.52	55.40	102.85	79.93	96.69	1.7%
L4	57,693,186	50,554,308	88%	43.40	40.00	23.32	83.07	64.19	76.18	3.8%
L5	123,544,337	113,775,664	92%	100.33	94.07	59.73	188.50	148.41	173.94	3.1%
L6	70,235,552	64,059,313	91%	56.18	52.00	31.62	107.25	83.46	97.39	3.3%
L7	91,883,293	82,583,941	90%	62.40	58.88	36.61	117.42	92.76	108.26	3.4%
L8	69,776,253	66,175,390	95%	57.27	52.29	30.80	109.84	83.92	95.89	3.4%
L9	72,694,352	63,994,638	88%	53.80	51.22	39.41	98.00	79.41	89.17	3.2%
L10	83,452,771	74,839,707	90%	55.54	53.30	36.35	102.08	81.96	93.51	3.6%
L11	88,716,715	80,891,704	91%	68.67	62.25	38.44	129.61	99.40	111.12	3.4%

Таблица 4.1. Статистика по полногеномному секвенированию одиннадцати особей *A. vaga.* Покрытие для каждой особи определяли на основе картирований парно-концевых прочтений Illumina HiSeq на диплоидную сборку *A. vaga* L1 и гаплоидную сборку. Статистика по покрытию основана на контигах с минимальной длиной 1000 п.н. Прочтения картировали с помощью Bowtie 2. Покрытие для диплоидной сборки определяли на основе лучших картирований парно-концевых прочтений согласно Bowtie 2. Парно-концевые прочтения, для которых с помощью Bowtie 2 было найдено одно или два выравнивания, затем картировали на гаплоидную сборку. Полученные выравнивания были дополнительно профильтрованы, и для дальнейших анализов использовали только прочтения с уникальным выравниванием с гаплоидной сборкой. Статистика по покрытию для гаплоидной сборки основана на профильтрованных выравниваниях.

4.1.5 Получение референтной сборки генома для *A. vaga* (L1)

Первый геном бделлоидной коловратки вида *A. vaga* был опубликован в 2013 [25]. Однако в процессе предварительного анализа данных была выявлена высокая степень дивергенции между секвенированными нами образцами и первым опубликованным геномом *A. vaga*. Так, среднее значение нуклеотидного сходства (определенного на основе находок BLAST) между прочтениями HiSeq из L1-L11 и геномом *A. vaga*, опубликованным в 2013 году, составляет всего ~87–88% (Таблица 4.2). Это наблюдение находится в соответствии с результатами предыдущих работ, где было показано, что выделенные на основании морфологических признаков виды бделлоидных коловраток часто представляют из себя комплексы дивергентных криптических видов [158]. Из-за значительного генетического

расстояния между полученными нами клональными линиями *A. vaga* и первым опубликованным геномом *A. vaga* использовать этот геном для картирования парно-концевых прочтений из L1-L11 не представлялось возможным. В связи с этим для одной из секвенированных линий (L1) была получена *de novo* сборка генома (далее – геном референтного клона *A. vaga*).

Особь	Среднее значение нуклеотидного сходства, %	Медианное значение нуклеотидного сходства, %
L1	87.35	87.76
L2	87.38	87.76
L3	87.52	87.76
L4	87.43	87.76
L5	87.50	87.76
L6	87.47	87.76
L7	87.49	87.76
L8	87.40	87.76
L9	87.17	87.36
L10	87.45	87.76
L11	87.33	87.64

Таблица 4.2. Оценки процентного сходства между геномами особей *A. vaga* L1-L11 и первым опубликованным геномом *A. vaga*. Для каждой особи 1,000,000 выбранных случайным образом прочтений Illumina HiSeq были использованы для поиска blastn против первой опубликованной геномной сборки *A. vaga* (Flot *et al.*, 2013) [25]. Для каждого прочтения вывод blastn ограничивался одним выравниванием. Затем среднее (или медианное) значение нуклеотидного сходства было рассчитано на основании выравниваний для «находок» blastn в геномной сборке *A. vaga* 2013 года (в анализ были включены только те прочтения, для которых было найдено выравнивание с длиной не менее 70 п.н.).

Для получения сборки генома, который мог бы быть использован как референтный, независимо сконструированная геномная библиотека для линии L1 была секвенирована на платформе MiSeq. Всего было проведено три запуска, в результате которых было получено 10,172,970 (2 × 301 п.н.), 15,397,651 (2 × 251 п.н.) и 20,061,190 (2 × 261 п.н.) прочтений соответственно. Первичная обработка и фильтрация парно-концевых прочтений MiSeq была проведена с использованием программы Trimmomatic V0.33 [157] так же, как для прочтений HiSeq (см. выше). Первичная *de novo* геномная сборка для референтного клона была получена на основе оставшихся после фильтрации 34,403,183 парно-концевых прочтений MiSeq с использованием программы SPAdes (версия 3.6.0) [159]. Сборку проводили с параметрами --diploid и --only-assembler с использованем k-меров следующих размеров: -k 21,33,55,77,99,127. Значение метрики N50 для полученной сборки составляет ~18 кб. В состав этой

первоначальной сборки вошло 51,852 контига ≥500 п.н. с суммарной длиной 233.8 Мб (Таблица 4.3).

	Первоначальная сборка	**Сборка после удаления контигов, являющихся результатом контаминации, и контигов, для которых не было найдено выравнивания с первой опубликованной сборкой генома *A. vaga***	**Гаплоидная сборка**
Число контигов (>0 п.н.)	78,825	19,202	12,034
Число контигов (≥500 п.н.)	51,852	19,068	8,999
Число контигов (≥1,000 п.н.)	25,404	15,929	7,771
Число контигов (≥10,000 п.н.)	6,530	6,383	2,373
Число контигов (≥100,000 п.н.)	27	27	1
Общая длина (>0 п.н.)	243,756,304	197,096,676	76,679,421
Общая длина (≥500 п.н.)	233,752,219	197,031,160	76,098,573
Общая длина (≥1,000 п.н.)	215,896,885	194,878,077	75,214,106
Общая длина (≥10,000 п.н.)	158,370,356	154,705,106	54,102,273
Общая длина (≥100,000 п.н.)	3,101,138	3,101,138	103,250
Число контигов	51,852	19,068	8,999
Максимальная длина контига	167,368	167,368	103,250
GC (%)	33.47	29.92	29.52
N50	18,125	22,073	18,007
N75	6,789	11,331	8,568
L50	3,473	2,620	1,182
L75	8,572	5,731	2,695
Число символов N на 100 кб	0	0	0

Таблица 4.3. Характеристики геномной сборки, полученной для линии *A. vaga* L1 и использовавшейся в качестве референтного генома в данной работе. Показатели, представленные в таблице, были рассчитаны с помощью программы QUAST (v5.0.0) [160] и, если не указано иное, основаны на контигах (или гаплоидных сегментах в случае гаплоидной сборки) с длиной не менее 500 п.н.

4.1.6 Фильтрация контигов, входящих в первоначальную сборку генома референтной линии *A. vaga* (L1)

Бимодальное распределение GC-состава в полученных контигах указывало на присутствие контигов, предположительно соответствующих бактериальным контаминантам

(Рисунок 4.1). Для фильтрации бактериальных контигов мы использовали метод, реализованный в пакете Blobology [161] (https://github.com/blaxterlab/blobology), основанный на анализе распределения GC-состава контигов и покрытия контигов парно-концевым прочтениями.

Покрытие контигов, входящих в первоначальную сборку, было определено отдельно по парно-концевым прочтениям MiSeq (на основе которых была получена сборка) и отдельно по парно-концевым прочтениям HiSeq (Рисунок 4.1), которые не использовались для получения сборки. Для каждого контига определяли покрытие парно-концевыми прочтениями, GC-состав, а также таксономическую принадлежность лучшей находки программы blastn [114] для данного контига в базе данных nt (с максимальным пороговым значением E-value 1×10^{-5}).

Как видно из Рисунка 4.1, контиги, для которых лучшая находка в базе данных nt соответствует бактериальной последовательности, характеризуются значительно более высоким GC-составом и низким покрытием по сравнению с контигами, для которых лучшая находка в базе данных nt соответствует последовательности, таксономически относящейся к бделлоидным коловраткам (порядок *Adinetida*). На этом основании контиги, вероятно являющиеся результатом бактериальной контаминации, могут быть достаточно просто отделены от контигов, происходящих из генома *A. vaga* (Рисунок 4.1). Таким образом, контиги могут быть разделены по покрытию и GC-составу на два подмножества – (i) контиги, предположительно являющиеся результатом бактериальной контаминации, и (ii) остальные контиги, последовательность которых, вероятно, присутствует в геноме *A. vaga*. Определение покрытия контигов по парно-концевым прочтениям HiSeq (Рисунок 4.1), которые не были включены в первоначальную сборку генома (по сравнению с определением покрытия по использованным в сборке прочтениям MiSeq; данные не показаны), позволяет лучше разделить подмножество контигов, соответствующих геному *A. vaga*, и подмножество контигов-контаминантов.

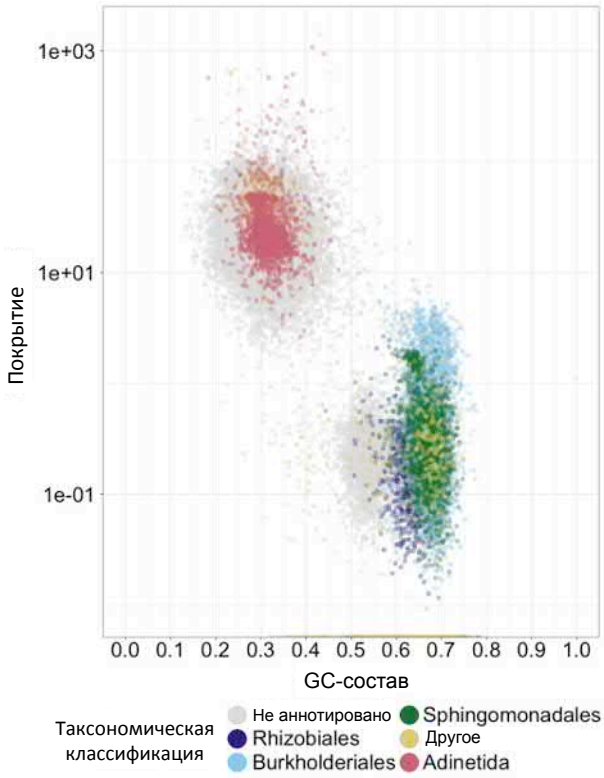

Рисунок 4.1. Таксономическая классификация контигов из первоначальной сборки генома *A. vaga* линии L1. GC-состав (ось X) и покрытие прочтениями (ось Y) для контигов из первоначальной сборки генома *A. vaga* L1 из прочтений Illumina MiSeq. Покрытие контигов определяли на основании прочтений Illumina HiSeq, полученных в результате секвенирования отдельной библиотеки и не использовавшихся для сборки генома. Точки (соответствующие отдельным контигам) покрашены в соответствии с таксономической классификацией [161] находок BLAST в базе данных nt (с пороговым значением E-value 1×10^{-5}). Представлена информация только о тех категориях таксономической классификации, к которым было отнесено не менее 6% контигов. Контиги без таксономической аннотации и контиги, представляющие менее частые категории, показаны в сером и желтом цветах соответственно.

В качестве пороговых значений для фильтрации контигов был выбран GC-состав ≥50% и среднее покрытие парно-концевыми прочтениями HiSeq ≤3× (контиги, удовлетворяющие хотя бы одному из этих условий, были удалены из сборки).

Затем была проведена дополнительная процедура фильтрации контигов, основанная на поиске выравниваний контигов из полученной нами сборки с опубликованным геномом *A. vaga* [25] с помощью программы blastn со следующими параметрами: -evalue 1e-10 -outfmt “6 qseqid sseqid pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen” -task dc-megablast. В финальную версию сборки вошли только те контиги, которые имели хотя бы одно

выравнивание с опубликованным геномом *A. vaga* с E-value ≤1e-100 и длиной не менее 500 нуклеотидов.

Оставшиеся после двух описанных выше ступеней фильтрации контиги (n = 19,068) имели суммарную длину ~197 Мб, что сопоставимо с длиной сборки опубликованного в 2013 году генома *A. vaga* (~218 Мб). Значение метрики N50 для финальной профильтрованной сборки равно ~22 кб. Эта сборка использовалась в качестве референтного генома *A. vaga* для последующих анализов. Более подробная информация о параметрах сборки приведена в Таблице 4.3.

4.1.7 Выделение неизбыточного гаплоидного набора сегментов

В связи с высоким уровнем гетерозиготности два гаплотипа в значительной части локусов генома *A. vaga* собираются в отдельные контиги [25]. В то же время в части генома при сборке происходит «схлопывание» гаплотипов в один контиг. В результате сборка генома *A. vaga* имеет мозаичную структуру с переменной плоидностью. Это осложняет применение стандартных методов для определения полиморфизмов и популяционных анализов, поскольку эти методы предназначены для работы с гаплоидным референтным геномом. В этом случае генотипы в полиморфных сайтах обычно определяются в диплоидном формате. При определении полиморфизмов на основании картирования прочтений на референтный геном с переменной плоидностью могут возникать различные нежелательные артефакты. Для того, чтобы минимизировать их возможное влияние на последующий анализ, мы выделили неизбыточный гаплоидный поднабор сегментов «диплоидной» сборки, который использовали в качестве гаплоидного референтного генома *A. vaga* (далее именуется сокращенно как «гаплоидная сборка»).

На первом этапе мы провели поиск пар геномных сегментов с высокой степенью сходства последовательностей – такие пары скорее всего соответствуют двум гаплотипам. После этого из каждой такой найденной пары сегментов мы оставляли только один сегмент, а второй исключали из анализа. Кроме того, из анализа были исключены сегменты, которые не удалось отнести к гаплотипическим парам. Данная процедура позволяет (i) уменьшить избыточность сборки за счет исключения одного из гаплотипов и (ii) исключить из анализа значительную часть сегментов с недостоверной плоидностью в первоначальной сборке.

Мы начали с определения для каждого контига подмножества других контигов, вероятно содержащих сегменты, составляющие гаплотипические пары с участками данного контига. Для этого для каждого контига, входящего в состав профильтрованной сборки, мы провели поиск BLAST [114] против остальных контигов сборки (использовалась программа blastn, входящая в состав BLAST+ 2.2.31, со следующими параметрами: -evalue 1e-10 -outfmt “6 qseqid sseqid

pident length mismatch gapopen qstart qend sstart send evalue bitscore qlen slen" -task dc-megablast -max_hsps 1). Затем для каждого контига с помощью скрипта на языке bash отбирали контиги среди остальных контигов с участками высокого сходства с данным (требовалось, чтобы было найдено хотя бы одно выравнивание с данным контигом с E-value ≤1e-50 и сходством последовательностей в области выравнивания ≥90%). На следующем шаге мы использовали программу all_bz (v.15) [162] для создания попарных blastz [163] выравниваний между каждым контигом и набором отобранных для него на предыдущем этапе контигов. Первичные результаты blastz были далее обработаны с помощью утилиты single_cov2 для того, чтобы отфильтровать вторичные выравнивания в участках контига, для которых было найдено более чем одно выравнивание. В результате данной процедуры для каждого контига был получен набор попарных выравниваний таких, что каждый сайт контига был сопоставлен не более чем с одним сайтом другого контига. Таким образом, полученный набор выравниваний для данного контига можно рассматривать как набор лучших находок BLAST между этим контигом и остальными контигами, входящими в состав сборки.

Для того, чтобы определить пары геномных сегментов, являющихся реципрокными лучшими находками друг для друга внутри сборки, мы применили процедуру аналогичную той, которая используется при поиске лучших реципрокных «хитов» BLAST между генами. Для каждого выравнивания между контигом A и контигом B, входящего в число лучших выравниваний между контигом A и остальным геномом (обозначим его как «прямое выравнивание»), мы проверяли, существует ли соответствующее выравнивание между контигом B и контигом A среди лучших выравниваний между контигом B и остальным геномом («обратное выравнивание»). Мы оставляли пару выровненных сегментов, если координаты прямого и обратного выравниваний совпадали или если область пересечения прямого и обратного выравниваний покрывала ≥80% более длинного из двух выравниваний. Во втором случае границы сегментов, являющихся лучшими реципрокными «хитами» друг для друга, определяли по позициям пересечения прямого и обратного выравниваний.

Полученный таким образом набор сопоставленных друг другу геномных сегментов с большой вероятностью представляет собой пары гаплотипов. Для того, чтобы получить неизбыточный гаплоидный поднабор сборки, из каждой пары сопоставленных сегментов мы выбирали и включали в состав гаплоидной сборки только один сегмент из пары – тот, который относился к более длинному контигу. Для увеличения непрерывности сборки в том случае, если два неизбыточных гаплоидных сегмента находились на одном контиге и были разделены ≤200 п.н., в состав гаплоидной сборки включали участок, содержащий оба сегмента. Информация о параметрах полученной в результате данных шагов гаплоидной сборки приведена в Таблице 4.3.

Суммарная длина гаплоидной сборки составляет 76,098,573 п.н (учитывались только сегменты с минимальной длиной 500 п.н.). Из этого можно заключить, что по крайней мере ~77% первоначальной сборки представлено двумя гаплотипами.

4.1.8 Аннотация белок-кодирующих генов

Для полученных контигов, входящих в профильтрованную сборку генома референтного клона *A. vaga* L1, была проведена *ab initio* аннотация белок-кодирующих генов с помощью программ AUGUSTUS [164] и GeneMark.ES [165]. Этот анализ был выполнен Е. С. Герасимовым. Некоторые характеристики полученного набора генных моделей приведены в Таблице 4.4. Первоначальный набор предсказаний включал 78,303 генных модели, происходящие из 75,877 локусов (генов). Мы провели фильтрацию первоначального набора генных моделей, исключив из рассмотрения модели с вероятными ошибками аннотации, а также модели, вероятно соответствующие фрагментам генов. Так, мы исключили генные модели, попадающие на самые края контигов, в результате чего осталось 72,406 из 78,303 генных моделей. Затем мы проверили кодирующие участки каждой генной модели и исключили те модели, длина кодирующей области для которых была не кратна трем (n = 6919). На следующем шаге были исключены модели, несущие преждевременный стоп-кодон (n = 691), а также модели, не имеющие канонического терминаторного кодона (n = 1426). В результате данных шагов фильтрации осталось доступно 63,370 генных моделей, происходящих из 61,531 локуса. Для каждого локуса (гена) выбирали наиболее длинную генную модель (предсказанный транскрипт) среди прошедших все шаги фильтрации. Оставшиеся после этого генные модели (n = 61,531) были использованы для последующих анализов.

Средняя (медианная) длина CDS, п.н.	1278.3 (990)
Средняя (медианная) длина интронов, п.н.	93.3 (55)
Среднее число интронов в гене	4.1
GC-состав транскриптов, %	31.9

Таблица 4.4. Некоторые параметры набора моделей белок-кодирующих генов, предсказанных для диплоидной сборки генома *A. vaga* L1. Данный анализ был выполнен Е. С. Герасимовым.

4.1.9 Разбиение генома на аллельные блоки

Для того, чтобы проверить устойчивость наших результатов к ошибочному определению гаплотипических пар, на основе которых была сконструирована гаплоидная сборка, мы отдельно провели анализ для части гаплоидной сборки, покрытой длинными блоками генов

коллинеарных между двумя гаплотипами в диплоидном геноме («аллельные блоки»). Кроме того, были идентифицированы аллели (гены, относящиеся к парам генов коллинеарных между аллельными блоками).

Поиск аллелей и разбиение генома на аллельные блоки был выполнен с использованием полученной на предыдущем этапе аннотации белок-кодирующих генов в геноме L1. Для того, чтобы выделить в сборке референтного генома *A. vaga* аллельные блоки, вначале был проведен поиск всех пар геномных сегментов, в которых гомологичные гены идут в одном и том же порядке (так называемые коллинеарные блоки) с помощью алгоритма McScanX [166]. В качестве входных данных для McScanX использовали результаты поиска с blastp (версия BLAST+ 2.2.31), выполненного для каждого из белков, предсказанных в геноме *A. vaga* L1, против всех предсказанных белков.

Результаты BLAST ограничивали «хитами» с E-value ≤1e-10, кроме того, для каждого белка вывод была ограничен хитами к пяти последовательностям. McScanX запускали с пороговым значением E-value ≤1e-5. Всего было идентифицировано 1,770 коллинеарных блоков. Для каждого такого блока определяли процент коллинеарных генов (процент гомологичных генов, идущих в одном и том же порядке, среди всех генов, входящих в блок) и среднее значение Ks (число синонимических замен на сайт – мера дивергенции гомологичных генов в синонимических сайтах). Значение Ks для каждой пары коллинеарных генов было рассчитано с помощью скрипта add_ka_and_ks_to_collinearity.pl, входящего в состав MCScanX.

Было показано, что, как и в первом опубликованном геноме *A. vaga*, в геноме *A. vaga* L1 коллинеарные блоки могут быть разделены на две достаточно четкие группы (см. раздел 4.2.1): группу с высоким процентом коллинеарных генов и низким уровнем синонимической дивергенции (аллельные блоки) и группу со значительно более высоким уровнем синонимической дивергенции и меньшей долей коллинеарных генов (паралогичные блоки, вероятно, являющиеся результатом полногеномной дупликации) [25].

Для использования в последующих анализах в качестве участков с подтвержденной плоидностью мы определяли аллельные блоки следующим образом: коллинеарный блок относили к аллельным блокам, если доля коллинеарных генов в блоке была ≥0.7, а среднее значение синонимической дивергенции Ks ≤0.2. Всего данным критериям соответствовало 1,387 коллинеарных блоков, из которых в финальный список вошло 1,354 блока, оставшихся после удаления «конфликтующих» блоков с сегментами синтении, перекрывающимися друг с другом.

Кроме того, данная процедура позволила определить пары аллелей: для этого пары гомологичных генов, входящих в состав аллельных блоков, были дополнительно профильтрованы на основании индивидуальных значений синонимической дивергенции

(требовалось, чтобы значение Ks было ≤0.2). В первоначальный набор аллелей вошло 12,489 пар коллинеарных генов. Подмножества генома, соответствующие аллельным блокам и аллелям, с высокой вероятностью представлены в первоначальной сборке двумя гаплотипами.

Для того, чтобы выделить гаплоидный эквивалент аллельных блоков, для каждой из 1354 пар коллинеарных геномных сегментов, составляющих блок, мы выбирали сегмент, принадлежащий к более длинному контигу. Второй сегмент из пары исключали из анализа. Суммарная длина уникальных непересекающихся аллельных участков, входящих в получившийся набор, составила 34,691,452 п.н.

Получив таким образом неизбыточное гаплоидное представление участков генома с достоверной плоидностью, мы отобразили его в систему координат гаплоидных сегментов, определенных ранее (см. раздел 4.1.7). Далее мы рассматривали только те аллельные участки, которые полностью содержались внутри первоначальных гаплоидных сегментов (аллельные участки, частично перекрывающиеся с гаплоидными сегментами, были исключены). Всего в результате этой процедуры для анализа осталось доступно 833 аллельных участка с суммарной длиной 19,300,566 п.н., в состав которых входило 7245 аллельных генов. Именно эти финальные подмножества гаплоидных сегментов, именуемые как «аллельные блоки» (n = 833) и «аллельные гены» (n = 7245), были использованы далее как области генома с достоверной плоидностью.

4.1.10 Картирование парно-концевых прочтений и фильтрация картирований

Для каждой линии, L1-L11, было проведено картирование оставшихся после первичной обработки парно-концевых прочтений HiSeq (см. раздел 4.1.4) на геном референтной линии L1. Картирование проводилось с помощью программы Bowtie 2 (v. 2.3.2) [167]. Выбор данной программы был мотивирован тем, что данная программа позволяет получать глобальные выравнивания парно-концевых прочтений на референтный геном (то есть такие выравнивания, в которых парно-концевое прочтение целиком картируется на геном). В случае геномов бделлоидных коловраток, несущих следы полногеномной дупликации [25], и вследствие этого богатых повторяющимися участками, использование глобальных выравниваний является предпочтительным по сравнению с локальными выравниваниями, поскольку позволяет уменьшить количество ошибочных картирований.

Тем не менее даже при поиске глобальных выравниваний с геномом *A. vaga* ожидается значительное количество ошибочных картирований, связанных со сложной структурой генома *A. vaga*, включающего большое количество повторяющихся последовательностей разной степени вырожденности.

Для того, чтобы уменьшить число ложных картирований парно-концевых прочтений, вероятно соответствующих выравниванию парно-концевого прочтения с паралогичным локусом, была разработана двуступенчатая процедура фильтрации парно-концевых прочтений. Эта процедура основана на сравнении наблюдаемого числа картирований для прочтения с ожидаемым числом истинных картирований при выравнивании с (i) первоначальной (диплоидной) сборкой генома референтной линии L1 и (ii) с гаплоидной сборкой генома этой линии.

Сначала парно-концевые прочтения для каждой линии были картированы на первоначальную (диплоидную) сборку генома референтной линии L1 с параметрами Bowtie 2 “--no-mixed --no-discordant” и максимальным разрешенным размером вставки 800 п.н. Вывод Bowtie 2 был ограничен пятью выравниваниями для каждого прочтения. Процент прочтений, которые удалось закартировать, для разных индивидуумов находился диапазоне 74.29–93.43%. Поскольку в диплоидной сборке большинство геномных локусов представлено двумя копиями (гаплотипами), у большинства парно-концевых прочтений может быть два «истинных» картирования, соответствующих выравниванию с двумя аллельными локусами. Соответственно, парно-концевые прочтения, имеющие более чем два картирования, вероятно, ошибочно помещены программой-картировщиком в паралогичные участки генома. Опираясь на эту логику, мы исключили из дальнейшего рассмотрения парно-концевые прочтения, имеющие более чем два выравнивания с диплоидной сборкой L1. Эта процедура была проведена с применением seqtk (v.1.2-r94) (https://github.com/lh3/seqtk) и набора собственных скриптов на Perl и bash.

Оставшиеся после первого шага фильтрации парно-концевые прочтения были в свою очередь картированы на гаплоидную сборку L1 (выравнивание с Bowtie 2 против гаплоидной сборки проводилось с такими же параметрами, как против диплоидной). Поскольку большая часть локусов должна быть представлена в гаплоидной сборке одной копией, мы ожидаем, что значительная доля парно-концевых прочтений должна иметь только одно истинное картирование.

В связи с этим парно-концевые прочтения, имеющие более чем одно выравнивание с гаплоидной сборкой (метка XS), не использовались для последующего определения однонуклеотидных полиморфизмов. Пара прочтений исключалась из дальнейшего анализа как в случае неуникального картирования обоих прочтений из пары, так и в случае неуникального картирования одного из прочтений.

Затем с помощью SAMtools (v.1.4.1) [168] (http://samtools.sourceforge.net) профильтрованные файлы, содержащие информацию о выравнивании прочтений, в формате SAM были конвертированы в формат BAM и дополнительно профильтрованы по качеству

картирования (исключались прочтения с MAPQ <20). Наборы выравниваний, полученные в результате применения описанной в данном разделе процедуры к картированиям прочтений для каждой линии, использовались далее для определения однонуклеотидных полиморфизмов.

4.1.11 Определение и фильтрация однонуклеотидных полиморфизмов

Для анализа популяционной изменчивости у *A. vaga*, представленного в данной главе, использовались два основных набора однонуклеотидных полиморфизмов. Набор однонуклеотидных полиморфизмов I включает генотипы, определенные для сайтов, являющихся вариабельными среди L1-L11. Набор однонуклеотидных полиморфизмов II включает генотипы как для вариабельных, так и для мономорфных сайтов.

Прошедшие жесткую фильтрацию полиморфизмы из набора I использовались для локальной реконструкции гаплотипов (см. раздел 4.1.12). Подход к фильтрации полиморфизмов из этого набора был разработан с целью максимально снизить процент сайтов, ошибочно определенных как полиморфные. Для этого перед проведением реконструкции гаплотипов из данного набора были исключены все сайты, для которых в картированных прочтениях было найдено более чем два различных нуклеотида, даже если некоторые нуклеотиды не входили в состав генотипов, определенных для этих сайтов. При определении генотипов для набора I все нуклеотиды, встречающиеся в картированных прочтениях, рассматривались как аллели вне зависимости от того, были ли они включены в состав генотипов для каких-то индивидуумов.

В то же время набор однонуклеотидных полиморфизмов II предназначался в первую очередь для изучения трехаллельных сайтов и определения попарных генотипических расстояний между индивидуумами. В данном контексте подход, при котором все нуклеотиды, присутствующие в прочтениях, рассматриваются как аллели, может приводить к ошибочной классификации сайтов с точки зрения числа аллелей, сегрегирующих в популяции. В связи с этим при определении генотипов для набора II для того, чтобы минимизировать число сайтов, определенных как трехаллельные из-за ошибок секвенирования, как аллели рассматривались только те нуклеотиды, которые были включены в состав хотя бы одного генотипа.

Определение однонуклеотидных полиморфизмов осуществлялось для всех индивидуумов L1-L11 одновременно на основании профильтрованных картирований парно-концевых прочтений на гаплоидную сборку L1. Как следствие генотипы определялись в диплоидном формате, поскольку гомологичные сайты из обоих гаплотипов выравнивались с одним и тем же сайтом гаплоидной сборки. Определение однонуклеотидных полиморфизмов как для набора I, так и для набора II проводилось с использованием утилиты mpileup из пакета SAMtools (v.1.4.1) [168] с параметрами “-aa -u -t DP,AD,ADF,ADR”, вывод которой обрабатывался дальше с применением команды “bcftools call” с опцией “-m”. Для детекции всех

возможных аллелей, присутствующих в картированных прочтениях, включая те, которые потенциально могли отсутствовать в определенных генотипах, и для того, чтобы в выводе присутствовали только вариабельные сайты, команду “bcftools call” запускали с дополнительными опциями “-A” и “-v”. Эти дополнительные опции использовались при определении генотипов для набора однонуклеотидных полиморфизмов I.

Затем мы провели фильтрацию полученных наборов полиморфизмов. Фильтрация определенных однонуклеотидных полиморфизмов включала следующие последовательные стадии:

1. Исключение из рассмотрения сайтов с полиморфными вариантами, находящихся на расстоянии меньшем или равном 10 п.н. от полиморфной инсерции или делеции.
2. Исключение из рассмотрения сайтов со значением QUAL <50 или с генотипами, не определенными для части индивидуумов.
3. Исключение из рассмотрения вариантов, находящихся на гаплоидных контигах короче 1000 нуклеотидов.
4. Исключение из рассмотрения вариантов, попадающих в участки генома низкой сложности (определенные с помощью программы RepeatMasker v. open-4.0.7).
5. Исключение из рассмотрения сайтов с низким покрытием (сайт исключали из последующего анализа, если хотя бы в одном индивидууме покрытие данной позиции парно-концевыми прочтениями было меньше 10×).
6. Исключение из рассмотрения вариантов с аномально высоким покрытием хотя бы в одном из индивидуумов*.
7. Исключение из рассмотрения вариантов, попадающих в геномные участки с аномально высокой плотностью определенных однонуклеотидных полиморфизмов**.

*В случае набора однонуклеотидных полиморфизмов I (содержащего только вариабельные сайты) мы исключали из дальнейшего рассмотрения сайты, значения покрытия для которых являлись статистическими выбросами (при анализе среднего покрытия или суммарного покрытия для 11 индивидуумов, или в случае анализа покрытия для каждого индивидуума по отдельности). Поиск статистических выбросов был проведен в R (v. 3.3.2) с использованием метода, основанного на расчете межквартильного диапазона. В случае набора однонуклеотидных полиморфизмов II сайт исключали из последующего анализа, если хотя бы в одном индивидууме покрытие данной позиции прочтениями было выше 300×.

**Окна с аномально высокой плотностью полиморфных сайтов с большой вероятностью соответствуют участкам, в которые картируются парно-концевые прочтения из паралогичных

областей генома. Для того, чтобы исключить из анализа сайты, ошибочно определенные как полиморфные из-за неправильного картирования прочтений, мы провели поиск участков генома, являющихся выбросами с точки зрения плотности однонуклеотидных полиморфизмов. Для этого сначала в каждом геномном окне длиной 1000 п.н. со сдвигом 500 п.н. была определена плотность однонуклеотидных полиморфизмов (на основании набора однонуклеотидных полиморфизмов I). На основании полученного распределения значений плотностей полиморфизмов были определены окна, соответствующие статистическим выбросам по плотности полиморфизмов. Поиск статистических выбросов был проведен в R (v. 3.3.2) с использованием метода, основанного на расчете межквартильного диапазона.

Фильтрация проводилась с применением комбинации утилит BCFtools (v.1.4.1), VCFtools (v. 0.1.15) [169], bedtools (v2.26.0) [170], SnpSift (v.4.3s) [171] и команд на языке awk. Количество сайтов в сырых наборах однонуклеотидных полиморфизмов I и II и сайтов, оставшихся доступными для анализа после последовательного применения описанных выше шагов фильтрации, приведено в Таблице 4.5.

	Набор однонуклеотидных полиморфизмов	
	Набор полиморфизмов I (только вариабельные сайты)	**Набор полиморфизмов II (вариабельные и мономорфные сайты)**
До фильтрации	3,318,352	76,306,143
Критерий исключения сайтов:		
Исключены полиморфные сайты на расстоянии меньшем или равном 10 п.н. от инсерции или делеции	2,979,193	75,968,700
Исключены сайты со значением QUAL <50 или с генотипами, не определенными для части образцов	2,655,917	49,222,726
Исключены сайты на гаплоидных контигах короче 1,000 п.н.	2,634,341	48,730,324
Исключены сайты, попадающие в участки генома низкой сложности	2,596,490	47,531,499
Исключены сайты с покрытием <10× (в одном или более образцах)	2,409,323	44,541,675
Исключены сайты с аномально высоким покрытием в одном или более образцах	2,391,710	43,924,505
Исключены сайты, попадающие в геномные участки с аномально высокой плотностью однонуклеотидных полиморфизмов	**2,282,099**	**42,850,155**

Таблица 4.5. Количество сайтов, включенных в первоначальные наборы однонуклеотидных полиморфизмов, и количество сайтов доступных для анализа после последовательного применения разных шагов фильтрации. Числа, соответствующие финальным профильтрованным наборам полиморфизмов (профильтрованные наборы однонуклеотидных полиморфизмов I и II), выделены жирным шрифтом.

Многомерное шкалирование попарных полногеномных IBS (сокр. от англ. identical by state) расстояний между индивидуумами проводилось с помощью программы PLINK (v1.90b5.4) [172]. Для этого анализа был использован «прореженный» набор однонуклеотидных

полиморфизмов из набора I. На первом шаге были исключены синглтоны (полиморфные варианты, встречающиеся только в одной особи). Затем оставшиеся полиморфизмы были профильтрованы так, чтобы в наборе данных не было сайтов, находящихся на расстоянии ≤1000 п.н. друг от друга. В результате этой процедуры осталось 66,483 полиморфных сайта, которые использовали для многомерного шкалирования.

В случае набора однонуклеотидных полиморфизмов II после фильтрации сайтов со значениями QUAL <50 доля отфильтрованных мономорфных сайтов оказалась значительно выше, чем доля отфильтрованных полиморфных сайтов: после данного шага фильтрации для анализа осталось доступно 97.8% полиморфных и только 68.8% мономорфных сайтов. Из-за такой «несимметричной» фильтрации оценки доли мономорфных сайтов в геномах анализируемых особей могут быть значительно занижены, что, в свою очередь, может приводить к завышенным оценкам гетерозиготности и генетических расстояний. Для того, чтобы избежать подобного смещения при проведении анализов, в которых используются мономорфные сайты, мы получили дополнительный набор однонуклеотидных полиморфизмов, далее обозначаемый как набор однонуклеотидных полиморфизмов III. Набор однонуклеотидных полиморфизмов III был получен из «сырого» набора однонуклеотидных полиморфизмов II аналогично тому, как был получен профильтрованный набор однонуклеотидных полиморфизмов II, но с двумя исключениями: (i) для фильтрации по полю QUAL использовалось пороговое значение 20 (в результате такой фильтрации для анализа остались доступны похожие доли полиморфных [99.1%] и мономорфных [99.7%] сайтов), (ii) из анализа исключались не только полиморфные сайты на расстоянии ≤10 п.н. от инсерции или делеции, но и мономорфные сайты на таком же расстоянии от инсерции или делеции. В состав набора однонуклеотидных полиморфизмов III вошло 58,163,647 сайтов, 3.93% (n = 2,285,700) и 96.07% (n = 55,877,947) из которых были определены как полиморфные и мономорфные среди особей L1-L11 соответственно. Это соотношение близко к соотношению, полученному в том случае, если фильтрация по полю QUAL не проводилась (3.94% и 96.06% соответственно). Это позволило заключить, что фильтрация с пороговым значением QUAL 20 не приводит к несимметричному отсеиванию мономорфных сайтов и следовательно подходит для проведения анализов с одновременным использованием и полиморфных, и мономорфных сайтов.

4.1.12 Реконструкция гаплотипов для индивидуумов L1-L11

Для каждого из индивидуумов L1-L11 была проведена реконструкция гаплотипов (вычислительное фазирование генотипов) с применением алгоритма HapCUT2 [173]. Для вычислительного фазирования использовались биаллельные полиморфные варианты из профильтрованного набора однонуклеотидных полиморфизмов I (n = 1,774,991) и

профильтрованные картирования прочтений на гаплоидную сборку (см. разделы 4.1.10 и 4.1.11). Для уменьшения влияния ошибок секвенирования на реконструкцию гаплотипов сайты, в которых встречалось более чем два нуклеотида в L1-L11, не были включены в набор сайтов, подвергнутый фазированию. Для локальной реконструкции гаплотипов HapCUT2 [173] запускали с опцией “--error_analysis_mode 1”, позволяющей рассчитывать вероятности ошибок фазирования. Фазирование проводилось для каждого из индивидуумов L1-L11 независимо.

Перед проведением дальнейших анализов участки генома с восстановленными гаплотипами (далее коротко именуются как «фазированные блоки») были подвергнуты агрессивной фильтрации. При проведении основного шага фильтрации мы опирались на следующую логику: поскольку пара полиморфных сайтов в каждом индивидууме может быть представлена не более чем двумя разными гаплотипами, те пары сайтов, для которых в прочтениях из одного индивидуума найдено более чем два разных «гаплотипа», с большой вероятностью ассоциированы с ошибками фазирования. Такие пары сайтов могут возникать, например, из-за переключения матриц в ходе полимеразной цепной реакции [174] или из-за выравнивания прочтений с паралогичными областями генома. Мы исключали фазированные блоки, в составе которых были найдены такие пары «конфликтующих» сайтов из дальнейщего анализа, поскольку их присутствие может потенциально создать ошибочную картину распада LD с расстоянием.

Для этого мы процессировали файлы, сгенерированные HapCUT2 для каждого индивидуума, L1-L11, и содержащие информацию о поддержке «гаплотипов» прочтениями. Мы извлекли такую информацию для каждой пары полиморфных вариантов, фаза для которых была определена HapCUT2 в данном индивидууме. Далее те пары сайтов, для которых в картированных прочтениях из одного индивидуума было найдено более чем два «гаплотипа», рассматривались как «конфликтующие». В большинстве таких случаев третий (наиболее редкий) «гаплотип» был поддержан всего лишь одним прочтением. По всей видимости, значительная часть таких ситуаций является результатом ошибок секвенирования, затрагивающих один нуклеотид. Поскольку такие случаи не должны оказывать серьезного влияния на качество фазирования, мы сузили список пар «конфликтующих» сайтов, оставив в нем только те пары, для которых в прочтениях из одного индивидуума было найдено три различных «гаплотипа» и каждый из этих «гаплотипов» был поддержан хотя бы двумя прочтениями. Кроме того, в список пар «конфликтующих» сайтов были включены пары сайтов, для которых в прочтениях из одного индивидуума были найдены все четыре возможных «гаплотипа» вне зависимости от числа прочтений, поддерживающих эти гаплотипы. Составив таким образом список пар «конфликтующих» сайтов, мы определили фазированные блоки, в

составе которых присутствовали такие сайты и исключили эти блоки из дальнейщего рассмотрения.

Фазированные блоки, оставшиеся после этого шага фильтрации, использовались в последующих анализах как основной набор фазированных данных (далее именуется как «набор фазированных данных 1»).

Для того, чтобы проверить, зависит ли распад LD с расстоянием от жесткости фильтрации гаплотипов, и убедиться в том, что наблюдаемая картина не может быть объяснена исключительно за счет ошибок фазирования, мы получили еще один вариант набора фазированных данных, подвергнутый дополнительной ступени фильтрации. Этот более жестко профильтрованный набор фазированных данных далее именуется как «набор фазированных данных 2». Для этого, помимо исключения фазированных блоков с парами «конфликтующих» сайтов, мы провели дополнительную фильтрацию, используя вероятности ошибок фазирования в шкале Phred, определенные HapCUT2. Для каждого фазированного блока, оставшегося после исключения блоков с парами «конфликтующих» сайтов, мы рассматривали сайты с вероятностью ошибок фазирования в шкале Phred <100 как «проблематичные». Блоки, несущие более одного такого сайта, исключались из дальнейшего анализа. Блоки, несущие один такой сайт, разбивались на два отдельных блока по координате данного сайта. Блоки, получившиеся в результате такого разбиения, далее анализировались отдельно.

Поиск пар «конфликтующих» сайтов и последующая обработка файлов HapCUT2 проводились с помощью собственных скриптов на языке Perl. Профильтрованные файлы, содержащие информацию о гаплотипах в формате HapCUT2, далее были конвертированы в формат VCF с применением утилиты HapCutToVcf из пакета fgbio (http://fulcrumgenomics.github.io/fgbio/).

Поскольку алгоритм HapCUT2 проводит фазирование только для гетерозиготных в данном индивидууме однонуклеотидных вариантов, мы использовали собственный скрипт на языке Perl, позволяющий приписать часть гомозиготных вариантов к фазированным участкам генома. Мы относили гомозиготный вариант к фазированному блоку, если он лежал внутри фазированного сегмента генома, то есть в том случае, если ближайшие к гомозиготному варианту гетерозиготные варианты принадлежали к одному фазированному сегменту генома. В этом случае полиморфный однонуклеотидный вариант приписывали одновременно к обоим гаплотипам.

Это позволило значительно увеличить количество доступных для анализа фазированных однонуклеотидных полиморфизмов, поскольку вариабельная позиция генома, гомозиготная в данном индивидууме, в большинстве случаев представлена в виде гетерозиготы в одном из других индивидуумов. Таким образом, включение гомозиготных однонуклеотидных вариантов

в анализ позволяет значительно расширить набор однонуклеотидных вариантов, для которых фаза определена сразу в нескольких индивидуумах.

Данные по длине участков генома с восстановленными гаплотипами, входящих в состав набора фазированных данных 1, для разных индивидуумов приведены в Таблице 4.6. Данные по общему числу полиморфных вариантов, попадающих в участки генома с восстановленными гаплотипами, для разных индивидуумов приведены в Таблице 4.7.

Особь	Медианное число гетерозиготных сайтов в фазированном блоке	Медианное число гетерозиготных сайтов в фазированном блоке	Медианное число полиморфных сайтов в фазированном блоке (после добавления гомозиготных сайтов)	Среднее число полиморфных сайтов в фазированном блоке (после добавления гомозиготных сайтов)	Медианная длина фазированного блока (п.н.)	Средняя длина фазированного блока (п.н.)
L1	26	47.9	50	90.5	1,720	3,196
L2	25	45.4	49	86.4	1,691	3,087
L3	32	59.8	65	114.7	2,205	4,143
L4	5	7.9	15	24.3	461	697
L5	5	8.5	17	27.5	541	813
L6	6	9.9	21	34.4	694	1,060
L7	6	9.5	19	32.2	639	980
L8	5	8.9	17	29.1	562	865
L9	6	9.8	20	33.5	680	1,027
L10	5	9.0	18	30.1	600	917
L11	6	9.1	18	30.1	596	906

Таблица 4.6. Статистика по количеству полиморфных сайтов, входящих в состав фазированных блоков, и длине фазированных блоков, восстановленных с использованием HapCUT2 и включенных в набор фазированных данных 1. Гомозиготные сайты были приписаны к фазированным блокам на основании данных о принадлежности к фазированным блокам ближайших фланкирующих гетерозиготных сайтов.

Особь	Общее число фазированных блоков	Общее число полиморфных сайтов, для которых проводили фазирование	Общее число фазированных сайтов			Сайты вне фазированных блоков		
			Фазированные сайты	Гетерозиготные фазированные сайты	Гомозиготные фазированные сайты	Сайты вне фазированных блоков	Гетерозиготные сайты вне фазированных блоков	Гомозиготные сайты вне фазированных блоков
L1	15,281	1,774,991	1,382,306	731,887	650,419	392,685	164,023	228,662
L2	11,503	1,774,991	994,252	522,769	471,483	780,739	359,381	421,358
L3	9,337	1,774,991	1,070,549	558,405	512,144	704,442	321,524	382,918
L4	31,777	1,774,991	771,705	251,366	520,339	1,003,286	33,706	969,580
L5	30,092	1,774,991	826,656	255,968	570,688	948,335	37,929	910,406
L6	26,488	1,774,991	911,315	262,482	648,833	863,676	31,445	832,231
L7	27,856	1,774,991	896,656	263,385	633,271	878,335	29,862	848,473
L8	29,752	1,774,991	865,236	264,824	600,412	909,755	30,324	879,431
L9	26,933	1,774,991	901,716	263,112	638,604	873,275	31,158	842,117
L10	24,135	1,774,991	727,669	217,007	510,662	1,047,322	23,005	1,024,317
L11	28,312	1,774,991	852,563	257,000	595,563	922,428	32,589	889,839

Таблица 4.7. Статистика по общему числу полиморфных сайтов, входящих в состав фазированных блоков, восстановленных с использованием HapCUT2 и включенных в набор фазированных данных 1. Общее число фазированных полиморфных сайтов было рассчитано как число полиморфных сайтов, принадлежащих к какому-либо фазированному блоку, включающему по крайней мере два гетерозиготных сайта (независимо от длины блока). Гомозиготные сайты были приписаны к фазированным блокам на основании данных о принадлежности к фазированным блокам ближайших фланкирующих гетерозиготных сайтов.

Для исследования зависимости неравновесия по сцеплению от расстояния между полиморфными сайтами и других анализов были выделены блоки однонуклеотидных вариантов, для которых во всех исследуемых индивидуумах (L4-L11 или L1-L11) удалось одновременно восстановить гаплотипы. Поиск таких блоков однонуклеотидных вариантов, фазированных одновременно в восьми или одиннадцати индивидуумах, осуществлялся с помощью собственных скриптов на языке Perl. Далее такие участки генома именуются как «фазированные сегменты». Для каждого фазированного сегмента мы извлекли соответствующие строчки из общего VCF-файла в отдельный файл в том же формате с применением команд awk. Кроме того, с использованием BCFtools (v.1.4.1) мы получили поднаборы однонуклеотидных вариантов, входящих в состав фазированных сегментов, оставшиеся после применения разных порогов на минимальную частоту минорного аллеля.

Для расчета значений r^2 и других анализов группы вариантов, входящие в состав разных фазированных сегментов, обрабатывались по отдельности.

4.1.13 Получение и обработка данных, использовавшихся для оценки частоты ошибок при реконструкции гаплотипов

Для того, чтобы оценить частоту ошибок при реконструкции гаплотипов, мы сравнили результаты фазирования для трех индивидуумов (L1, L2 и L11), которые были секвенированы с использованием разных инструментов. В частности, три независимо сконструированные библиотеки для L1 были секвенированы на платформах Illumina HiSeq, Illumina MiSeq и PacBio. Прочтения MiSeq были использованы для сборки референтного генома *A. vaga*, но не для определения однонуклеотидных вариантов; прочтения HiSeq, напротив, использовались только для определения однонуклеотидных вариантов и не были включены в сборку. Прочтения PacBio, в свою очередь, не были включены в основные анализы и использовались исключительно для оценки качества фазирования.

В случае L11 две независимые библиотеки секвенировали на платформах Illumina HiSeq и Illumina MiSeq. L2 тоже секвенировали на платформах Illumina HiSeq и Illumina MiSeq, но в случае L2 секвенирована была одна и та же библиотека. В то время как прочтения HiSeq, полученные для L2 и L11, использовались для определения однонуклеотидных вариантов и последующих анализов, прочтения MiSeq для этих двух культур использовались только для оценки частоты ошибок фазирования.

Несоответствие между фазами гаплотипов, восстановленными для одного и того же индивидуума с использованием разных наборов данных, может иметь разные источники. Во-первых, такие случаи могут возникать из-за переключения матрицы в ходе полимеразной цепной реакции [174]. Поскольку библиотеки для L1 и L11, секвенированные на инструментах HiSeq и MiSeq, конструировались независимо, каждая библиотека скорее всего содержала свой собственный набор молекул, возникших в результате переключения матрицы. Во-вторых, несоответствия между фазами гаплотипов могут возникать из-за ошибочного картирования прочтений в паралогичные области генома. Можно ожидать, что более длинные прочтения MiSeq (~250–300 п.н.) будут реже картироваться ошибочно по сравнению с короткими прочтениями HiSeq (~100 п.н.). В связи с этим, несмотря на то, что библиотеки, секвенированные на HiSeq и MiSeq, вероятно, содержали схожие доли молекул, возникших в результате переключения матрицы, гаплотипы, восстановленные на основе более длинных прочтений MiSeq, скорее всего будут более аккуратными.

В случае L2 одна и та же библиотека была секвенирована сначала на платформе HiSeq, а затем на платформе MiSeq. В связи с этим для L2 среди фрагментов, секвенированных на двух разных инструментах, могли присутствовать продукты одних и тех же событий переключения матрицы. Таким образом, можно предположить, что для L2 несоответствие в результатах фазирования, полученных на основе прочтений HiSeq и MiSeq, вероятно, в меньшей доле

случаев связано с переключением матрицы в ходе полимеразной цепной реакции. Однако ранее было показано, что переключение матрицы преимущественно происходит на поздних циклах полимеразной цепной реакции и продукты таких событий обычно представлены маленьким числом копий [174]. В таком случае возможно, что продукты события переключения матрицы могут быть представлены только в прочтениях, полученных на каком-то одном из двух инструментов, и какая-то часть таких случаев может вносить свой вклад в несоответствия между гаплотипами, восстановленными для L2 на основе двух наборов прочтений. Стоит отметить, что мы не пытались оценить, как часто продукты одного и того же события переключения матрицы могут быть выявлены и в HiSeq-, и в MiSeq- прочтениях, полученных для L2.

Мы сравнили фазированные блоки, восстановленные на основании картирований прочтений HiSeq и использовавшиеся в последующих анализах, с фазированными блоками, восстановленными для того же индивидуума на основе прочтений MiSeq (в случае L1, L2 и L11) или PacBio (в случае L1). Для этого мы провели вычислительное фазирование генотипов с применением алгоритма HapCUT2 [173], используя тот же набор биаллельных полиморфных вариантов (n = 1,774,991), который использовался при проведении фазирования на основании прочтений HiSeq (см. раздел 4.1.12). Однако в этот раз фазирование проводилось на основании картирований прочтений MiSeq или PacBio.

В случае реконструкции гаплотипов, основанной на картировании прочтений MiSeq, в дополнение к набору «сырых» фазированных блоков, непосредственно сгенерированных HapCut2, мы получили два набора профильтрованных блоков, к которым при фильтрации были применены те же критерии, что и к основному набору фазированных данных, восстановленному на основе прочтений HiSeq и использовавшемуся в основных анализах. Первый и второй наборы блоков, восстановленные на основе прочтений MiSeq, были подвергнуты фильтрации аналогичной той, что применялась для получения набора фазированных данных 1 и набора фазированных данных 2 соответственно (см. раздел 4.1.12).

Мы не подвергали фильтрации фазированные блоки, восстановленные на основе прочтений PacBio, поскольку высокая частота ошибок при секвенировании (~14%) в этом случае делает нецелесообразным основной шаг фильтрации: исключение блоков, несущих пары вариантов, представленных более чем двумя гаплотипами в одном индивидууме.

4.1.14 Анализ неравновесия по сцеплению с использованием реконструированных гаплотипов (фазированных данных)

Для того, чтобы изучить паттерны неравновесия по сцеплению (далее – коротко LD от англ. термина linkage disequilibrium) у *A. vaga* на основании реконструированных гаплотипов,

мы рассчитали значения r^2 для пар полиморфных вариантов, попадающих в один и тот же фазированный сегмент с помощью VCFtools (v. 0.1.15) [169]. Данный анализ проводили для каждого фазированного сегмента по отдельности. Если не указано иное, описываемые результаты основаны на анализе полиморфных вариантов из набора фазированных данных 1. Для этого анализа мы дополнительно исключали сайты, которые могли быть ложно определены как гомозиготные в некоторых индивидуумах. С этой целью для каждого индивидуума мы провели поиск сайтов, которые были определены в данном индивидууме как гомозиготные, но при этом были представлены в картированных прочтениях (из данного индивидуума) двумя нуклеотидами, каждый из которых был поддержан хотя бы двумя прочтениями. Такие сайты были исключены из дальнейшего анализа во всех индивидуумах. Приведенные результаты основаны на полиморфных вариантах с частотой минорного варианта ≥4 среди восьми индивидуумов большого кластера, L4-L11. Результаты, полученные с применением других порогов на минимальную частоту минорного варианта, качественно похожи (данные не приведены). Результаты, полученные на основании полиморфных вариантов из более строго профильтрованного набора фазированных данных 2, также качественно похожи (см. раздел 4.2.3). Кроме того, мы воспроизвели основные результаты на подмножестве полиморфных вариантов из набора фазированных данных 1, принадлежащих к аллельным участкам генома *A. vaga* (см. разделы 4.1.9 и 4.2.3).

Для того, чтобы определить фоновый уровень r^2, мы рассчитали значения r^2 для пар сайтов, находящихся на разных контигах в первоначальной сборке. Если общее число пар сайтов с разных контигов в рассматриваемом подмножестве сайтов превышало 10,000,000, мы уменьшали объем данных, случайным образом выбирая 10,000,000 пар сайтов. В таких случаях показанные распределения значений r^2 для разных контигов и соответствующие средние и медианные значения r^2 основаны на таких выборках пар сайтов.

Распад LD (выраженного через r^2) с физическим расстоянием (в нуклеотидах) визуализировали с помощью локальной регресии LOESS (имплементированной в R [v.3.5.1]) с параметром сглаживания 0.4.

Чтобы дополнительно подтвердить, что распад LD не вызван исключительно ошибками фазирования, мы независимо оценили LD, используя нефазированные данные по генотипам L4-L11 с использованием двух подходов. Первый подход основан на «восстановлении» гаплотипов с использованием полиморфных сайтов гомозиготных во всех индивидуумах L4-L11. Этот подход использует следующую логику: «фаза» находящихся на одном контиге полиморфных вариантов гомозиготных в данном индивидууме уже известна, поэтому для анализа LD можно использовать полиморфные сайты, гомозиготные во всех восьми индивидуумах (т.е. сайты, для которых в части индивидуумов предсказано наличие в геноме двух вариантов идентичных

варианту, присутствующему в референтном геноме, а в остальных индивидуумах предсказано наличие в геноме двух вариантов, отличающихся от референтного). Такие отличающиеся между разными индивидуумами, но не несущие отличий между гаплотипами в каждом отдельно взятом индивидууме сайты позволяют проанализировать неравновесие по сцеплению без формального фазирования генотипов. Для L4-L11 мы нашли 18,995 таких полиморфных сайтов одновременно гомозиготных во всех восьми индивидуумах. Далее мы исключили сайты, в которых менее частый генотип присутствовал только в одном индивидууме, оставив только такие сайты, в которых оба гомозиготных генотипа (0/0 и 1/1) встречались хотя бы в двух индивидуумах (n = 3410). Чтобы анализировать только истинно гомозиготные сайты, мы исключили сайты, в которых в прочтениях из одного индивидуума встречалось более одного нуклеотида. Кроме того, мы требовали, чтобы генотипы были поддержаны и прямыми, и обратными прочтениями во всех индивидуумах. В результате данной фильтрации для анализа осталось доступно 2573 полиморфных сайта. Мы конвертировали поле GT таких сайтов в VCF-файле в формат фазированного генотипа (0/0 ➔ 0|0; 1/1 ➔ 1|1) и использовали получившийся файл для того, чтобы рассчитать значения r^2 с применением VCFtools (см. раздел 4.2.3).

Второй подход к оценке LD из нефазированных данных основан на использовании в качестве меры LD коэффициентов корреляции между генотипами с использованием команды --geno-r2 программы VCFtools [169]. Эта команда в качестве меры LD использует ту же метрику, что и программа PLINK [172]: для каждой пары полиморфных вариантов рассчитывается квадрат коээфициента корреляции между числом нереферентных вариантов в генотипах (генотип может нести 0, 1 или 2 нереферентных варианта) в двух соответствующих сайтах у рассматриваемых индивидуумов. Как и раньше, мы исключали из анализа сайты, которые могли быть ошибочно определены как гомозиготные, и использовали полиморфные варианты с частотой минорного варианта ≥4 среди восьми индивидуумов большого кластера, L4-L11. Представленные в работе результаты основаны на попарных сравнениях для 10,000 случайно выбранных биаллельных сайтов со всеми остальными биаллельными сайтами. Пары полиморфных вариантов были разбиты на группы по расстоянию, разделяющему сайты в паре, с шагом в 200 п.н. Далее для каждой такой группы мы определяли среднее значение квадрата коэффициента корреляции среди всех пар вариантов, попадающих в данную группу. Доверительные интервалы были рассчитаны на основании 1000 бутстреп-реплик.

4.1.15 Оценка корреляции зиготности (Δ)

В качестве альтернативного подхода к оценке зависимости LD от физического расстояния между сайтами мы сравнили корреляцию зиготности (Δ) для пар сайтов, находящихся на разных расстояниях. Δ является мерой зависимости между разными сайтами в

геноме [175,176], которая рассчитывается для пар сайтов внутри генома одного индивидуума. Мы оценили корреляцию зиготности для пар сайтов внутри геномов отдельных индивидуумов, использовав подход, опирающийся на метод максимального правдоподобия, реализованный в программе mlRho [175,176]. Оценки Δ для каждого инидивидуума L1-L11 были получены с использованием mlRho для сайтов с покрытием не менее 20 прочтений в данном индивидууме. Графики, показывающие поведение Δ с увеличением расстояния между сайтами для четырех индивидуумов (L1, L4, L7 и L11), приведены в разделе 4.2.3. Картина падения Δ с расстоянием для остальных индивидуумов из маленького и большого кластеров схожа с картиной, наблюдаемой для L1 и L4/L7/L11 соответственно.

4.1.16 Поиск сигнала рекомбинации

Мы провели поиск сигнала рекомбинации, применив ранее разработанные тесты на рекомбинацию к 434 геномным сегментам, несущим по меньшей мере 15 полиморфных сайтов с частотой минорного варианта ≥2, фаза которых была одновременно определена для всех особей L4-L11. Эти сегменты были распределены между 352 контигами диплоидной сборки L1 и суммарно покрывали 364,361 п.н. Для каждого сегмента мы восстановили последовательности двух гаплотипов в каждой особи L4-L11 с помощью BCFtools (v.1.4.1), использовав соответствующую референтную последовательность из гаплоидной сборки и набор фазированных полиморфных вариантов. Как и для анализа LD, сайты, которые могли быть ложно определены как гомозиготные, не рассматривались. Для реконструкции гаплотипов были использованы только те полиморфные сайты, которые прошли все шаги фильтрации и были одновременно фазированы у индивидуумов L4-L11; остальные сайты рассматривались как мономорфные. Для каждого сегмента мы провели анализ корреляции r^2 с физическим расстоянием [177], тест на «сумму расстояний» [178] (англ. sum of the distances), а также PHI тест [179]. PHI тесты проводили с помощью PhiPack (программа доступна по адресу http://www.maths.otago.ac.nz/~dbryant/software/PhiPack.tar.gz), тесты двух других типов проводили с использованием LDhat (версия 2.2) [180].

4.1.17 Анализ распределения значений коэффициента инбридинга

Значения коэффициента инбридинга F_{IS} рассчитывали для биаллельных полиморфных сайтов, принадлежащих к профильтрованному набору однонуклеотидных полиморфизмов II, с частотой минорного варианта среди особей большого кластера L4-L11 ≥4. Для каждого сайта значение F_{IS} рассчитывали как $1 - \frac{H_{\mathrm{o}}}{H_{\mathrm{e}}}$, где H_{o} и H_{e} обозначают наблюдаемое и ожидаемое число гетерозиготных генотипов соответственно. Значения, ожидаемые при равновесии Харди-Вайнберга, были получены с помощью программы populations из пакета Stacks (версия 2.4)

[181]. Нецелые ожидаемые значения были округлены до ближайшего целого числа. Анализ проводили для полногеномных вариантов, а также для вариантов, попадающих в участки генома *A. vaga* с достоверной плоидностью (аллельные блоки и аллельные гены).

4.1.18 Симуляции популяций с разной частотой бесполого размножения

Для того, чтобы напрямую оценить, какие значения F_{IS} ожидаются при разной частоте бесполого размножения в выборке такого же размера как анализируемая (8 особей), мы симулировали популяции в SLiM (версия 3.2) [182], варьируя частоту бесполого размножения от 0 (строго половое размножение) до 1 (строго бесполое разможение). Затем мы получили случайные подвыборки «особей» из симулированных популяций, включающие как и анализируемая популяция 8 индивидуумов, и оценили распределение значений F_{IS} в таких подвыборках. Перед проведением симуляций мы получили оценки популяционной скорости возникновения мутаций θ ($4N_e\mu$) с использованием метода максимального правдоподобия, реализованного в программе mlRho [176] (см. раздел 4.1.23). Оценки θ для индивидуумов из большого кластера имели порядок 10^{-2} (оценки находились в диапазоне от 0.0072 до 0.0094; см. раздел 4.1.23). Параметры симуляций были подобраны так, чтобы популяционная скорость возникновения мутаций $4N_e\mu$ в симуляциях была близка к значению этого параметра, оцененного из данных: $N_e = 2500$, $\mu = 10^{-6}$. Во всех случаях геном симулированных особей состоял из единственной хромосомы длиной 10^6 п.н. В симуляциях популяций с половым размножением скорость рекомбинации была равна 10^{-6} на нуклеотид на поколение с половым размножением. Соотношение частоты вредных мутаций к частоте нейтральных мутаций было принято равным 1:0.4. Распределение коэффициентов отбора для вредных мутаций имело форму Гамма-распределения (средний коэффициент отбора −0.01, параметр формы 0.1). Предполагалось, что эффекты вредных мутаций аддитивны. Симуляции проводили, варьируя частоту бесполого размножения от 0 (строго половое размножение) до 1 (строго бесполое разможение). Для каждой рассмотренной частоты бесполого размножения (0, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999 и 1) было проведено 100 реплик симуляции. Через 200,000 поколений из каждой реплики каждой симуляции случайно выбирали 8 особей (что соответствует числу анализируемых особей L4-L11), оставляли только полиморфные среди этих особей сайты с частотой минорного варианта ≥4 и затем случайно отбирали 200 полиморфных сайтов. В результате этой процедуры для каждой симуляции было получено 100 выборок полиморфных сайтов. Для каждой выборки мы определяли среднее значение F_{IS}. Для сравнения с данными мы аналогичным образом определили средние значения F_{IS} для 100 случайных выборок сайтов (n = 200) биаллельных среди L4-L11 (частота минорного варианта ≥4). Значения F_{IS} для выборок сайтов полиморфных среди L4-L11 оказались значительно выше, чем ожидаемые при строго

бесполом размножении (P < 0.01) и частоте бесполого размножения 0.999 (P = 0.03). P-значения для всех остальных рассмотренных частот бесполого размножения ниже чем 0.999 (частоты бесполого размножения 0, 0.5, 0.7, 0.9, 0.95, 0.99) > 0.05. Для каждой анализируемой частоты бесполого размножения одностороннее P-значение рассчитывали как долю среди 100 выборок полиморфных сайтов, полученных для данной симуляции, со средним значением F_{IS} равным или более высоким чем минимальное среднее значение F_{IS} (−0.0917) среди 100 выборок полиморфных сайтов, полученных для L4-L11.

4.1.19 Анализ филогений гаплотипов

Мы провели поиск инконгруэнтных филогений для гаплотипов для трех наборов фазированных сегментов. Наборы A и B включают сегменты, в которых удалось фазировать полиморфные варианты для всех индивидуумов большого кластера (L4-L11). Набор C включает сегменты, в которых удалось фазировать полиморфные варианты для всех одиннадцати индивидуумов из большого (L4-L11) и малого (L1-L3) кластеров.

На первом этапе для поиска инконгруэнтных филогений гаплотипов, которые могли бы указывать на генетический обмен у L4-L11, мы начали с того же набора из 434 геномных сегментов, который использовался для проведения тестов на рекомбинацию (это сегменты, несущие по меньшей мере 15 полиморфных сайтов с частотой минорного варианта ≥2, одновременно фазированных у L4-L11). Как и при анализе LD, мы дополнительно исключали сайты, которые могли быть ложно определены как гомозиготные в некоторых индивидуумах.

Чтобы уменьшить возможное влияние повторяющихся участков генома и/или близких паралогов, которые могут быть ошибочно приняты за гаплотипы (аллели), на филогенетический анализ гаплотипов, мы провели дополнительную фильтрацию фазированных сегментов, опираясь на результаты поиска BLAST против диплоидной сборки генома L1. Для каждого гаплотипа (n = 16) в каждом рассматриваемом сегменте (n = 434) мы определили число уникальных «хитов» в диплоидной сборке генома L1 (для каждого гаплотипа учитывали только одно выравнивание с одним и тем же контигом L1, опция BLAST “-max_hsps 1”). Медианное число хитов на гаплотип составляет 4; в значительном числе случаев для гаплотипа было найдено два «хита» с высокой степенью нуклеотидного сходства (≥90%), соответствующие двум гаплотипам, присутствующим в сборке L1, и два «хита» с более низким нуклеотидным сходством (~65–85%), вероятно, соответствующие паралогам. Однако для гаплотипов из некоторых сегментов были найдены «хиты» к значительно большему числу контигов L1. В таких случаях в результатах поиска BLAST обычно присутствовало два выравнивания с высоким нуклеотидным сходством, покрывающих сегмент целиком или практически целиком, и некоторое число коротких выравниваний с низким нуклеотидным сходством, покрывающих

небольшую часть сегмента. Чтобы уменьшить вероятность того, что геномные сегменты, содержащие последовательности с большим количеств дивергентных копий в геноме *A. vaga*, создают ложный сигнал инконгруэнтности, мы исключали сегменты с «хитами» к более чем 10 контигам в диплоидной сборке вне зависимости от нуклеотидного сходства «хита». В оставшемся наборе сегментов среднее число хитов с высоким нуклеотидным сходством (≥90%) для гаплотипа составило 2.07, а среднее число хитов с низким нуклеотидным сходством (<90%) составило 3.02. Мы дополнительно исключили сегменты с гаплотипами, для которых было найдено более двух хитов с высоким нуклеотидным сходством. После этих шагов из первоначальных 434 сегментов для анализа осталось доступно 303 сегмента, покрывающих 243,455 п.н. и распределенных между 263 контигами диплоидной сборки L1. Эти 303 сегмента были использованы для поиска инконгруэнтных филогений гаплотипов (набор A). Аналогичная процедура фильтрации была применена к наборам фазированных сегментов B и C (см. ниже).

Для каждого из двух гаплотипов каждого индивидуума L4-L11 мы определили нуклеотидное расстояние (долю нуклеотидных отличий) до другого гаплотипа в том же индивидууме и до гаплотипов остальных индивидуумов. Затем для каждого гаплотипа мы определили ближайший (наиболее похожий) гаплотип в других индивидуумах. Для того, чтобы проверить устойчивость такого сопоставления, мы сравнили число нуклеотидных отличий между рассматриваемым гаплотипом и ближайшим к нему гаплотипом (N1) из других индивидуумов и следующим по расстоянию гаплотипом (N2) из других индивидуумов. Считалось, что для гаплотипа однозначно определен ближайший гаплотип в другом индивидууме («ближайший сосед»), если разница между N2 и N1 составляла 3 или более нуклеотидных отличия (N2–N1 ≥ 3). Чтобы проанализировать то, как гаплотипы группируются в филогениях, для каждого индивидуума мы отобрали те фазированные сегменты, в которых «ближайшие соседи» были однозначно определены для обоих гаплотипов этого индивидуума (обозначим их как H1 и H2). Такие случаи могут быть разделены на две группы в зависимости от того, найдены ли «ближайшие соседи» двух гаплотипов (H1′ и H2′) в одном индивидууме или в разных. Чтобы исключить из рассмотрения случаи, в которых то, как гаплотипы группируются, могло быть обусловлено действием генной конверсии, мы требовали, чтобы расстояния, разделяющие пары «ближайших соседей» из разных индивидуумов (H1-H1′ и H2-H2′), были короче, чем расстояния между двумя гаплотипами одного индивидуума. Затем мы оставляли только случаи, в которых определенные пары «ближайших соседей» соответствовали «лучшим реципрокным хитам». Более точно, мы требовали, чтобы гаплотиы H1 и H2 также являлись однозначно определенными «ближайшими соседями» H1′ и H2′ соответственно. В общей сложности нам удалось определить «ближайших соседей» для двух гаплотипов хотя бы одного индивидуума в 90 из 303 рассмотренных фазированных геномных сегментов.

Среди этих 90 сегментов только в 12 сегментах мы нашли пару индивидуумов такую, что их гаплотипы представляли собой две пары «ближайших соседей» («конгруэнтная группа»). В противоположность этому, в 79 сегментах «ближайшие соседи» двух гаплотипов по крайней мере одного индивидуума находились в двух разных индивидуумах («инконгруэнтная группа»; в одном сегменте была обнаружена как конгруэнтная, так и инконгруэнтная группа гаплотипов, включающие гаплотипы разных индивидуумов).

Чтобы убедиться в надежности найденных групп гаплотипов, для каждого из 90 сегментов мы построили неукорененное филогенетическое дерево с применением метода максимального правдоподобия, реализованного в программе PhyML (версия 3.1) [183]. Филогенетические деревья были построены с использованием модели GTR+G (4 категории скоростей замен, параметр формы Гамма-распределения оценивался из данных, 1000 бутстреп-реплик). Деревья были укоренены посередине с помощью пакета ETE 3 [184]. Затем для найденных на предыдущем этапе групп гаплотипов были извлечены значения бутстреп-поддержки. В совокупности, в 10 из 12 сегментов с «конгруэнтной группой» такая группа получила достойную бутстреп-поддержку (≥70%). В 52 из 79 сегментов с «инконгруэнтной группой/группами» хотя бы одна из этих групп была достойно поддержана (бутстреп-поддержка ≥70%).

Затем, чтобы проверить, не связан ли сигнал инконгруэнтности с ошибками при определении полиморфных вариантов, включая те, которые могут быть вызваны аберрантным отнесением индексов к прочтениям, а также ошибочным определением гетерозиготных генотипов как гомозиготных, мы повторили анализ для L4-L11, применив дополнительные фильтры к полиморфным вариантам и используя гаплотипы, восстановленные на основании только тех полиморфных вариантов, которые прошли эти дополнительные фильтры. А именно, для этого мы исключили (1) сайты, определенные как гомозиготные в части индивидуумов, но тем не менее представленные в прочтениях из этих индивидуумов более чем одним нуклеотидом (даже если второй нуклеотид был поддержан только одним прочтением), (2) сайты, определенные как гетерозиготные в части индивидуумов, в случае если один из двух аллелей в гетерозиготном генотипе был поддержан менее чем 30% прочтений, и (3) сайты с аллелями, поддержанными в части индивидуумов только прямыми или только обратными прочтениями. Дополнительная фильтрация полиморфных вариантов привела к падению доступного для анализа набора фазированных сегментов с 303 до 190 (набор B). В совокупности, среди 190 сегментов из набора B в 40 была найдена инконгруэнтность двух гаплотипов хотя бы одного индивидуума, и в 25 сегментах соответствующие ветви в филогенетическом дереве имели бутстреп-поддержку ≥70% (см. раздел 4.2.8).

Кроме того, чтобы проверить, могут ли подписи генетического обмена потенциально

затрагивать индивидуумов из разных кластеров, мы провели поиск случаев инконгруэнтности среди всех одиннадцати индивидуумов L1-L11. Для этого мы использовали набор из 152 сегментов, несущих как минимум 15 полиморфных вариантов (синглтоны не рассматривались), одновременно фазированных во всех индивидуумах L1-L11 (набор C). Эти 152 сегмента были распределены между 138 контигами диплоидной сборки L1 и суммарно покрывали 114,592 п.н. Последовательности фазированных сегментов, включенных в набор C, были получены с применением дополнительных строгих критериев, направленных на уменьшение потенциальных эффектов ошибок при определении однонуклеотидных полиморфизмов и аналогичных тем, которые были применены к набору B (см. выше).

4.1.20 Филогенетический анализ гаплотипов L1-L11

Мы дополнительно проанализировали филогенетические деревья гаплотипов, полученные методом максимального правдоподобия, для всех 152 сегментов, входящих в набор C (см. выше), без предварительного поиска инконгруэнтности. Для каждого сегмента из набора C мы получили неукорененное филогенетическое дерево, построенное на основании реконструированных последовательностей двух гаплотипов каждого индивидуума L1-L11. Как и ранее, филогенетические деревья были построены методом максимального правдоподобия, реализованного в программе PhyML (версия 3.1) [183], с использованием модели GTR+G (4 категории скоростей замен, параметр формы Гамма-распределения оценивался из данных, 1000 бутстреп-реплик) и укоренены посередине с помощью пакета ETE 3 [184]. Затем мы провели поиск монофилетических групп, включающих исключительно гаплотипы индивидуумов L1, L2, и L3 и имеющих достойную бутстреп-поддержку (≥70%). Для каждой такой группы мы определили число входящих в нее гаплотипов. В целом, из 152 проанализированных сегментов (распределенных между 138 контигами диплоидной сборки L1) в 119 была найдена монофилетическая группа, сформированная тремя гаплотипами, по одному гаплотипу каждого индивидуума L1, L2 и L3. В 100 сегментах (распределенных между 93 контигами) такая группа имела бутстреп-поддержку ≥70%. Среди этих 100 сегментов в 6 случаев такая группа содержалась внутри плохо поддержанной группы, включающей 4 гаплотипа индивидуумов L1, L2, и L3. Суммарно, было найдено 8 сегментов, в которых существовала монофилетическая группа, включающая 4 из 6 гаплотипов L1, L2, и L3, однако ни одна из этих групп не получила достойную бутстреп-поддержку. Ни для одного сегмента не было найдено монофилетической группы, включающей 5 или 6 гаплотипов индивидуумов L1, L2, и L3. Кроме того, не было идентифицировано ни одного случая, в котором L1, L2, и L3 одновременно кластеризовались сразу по двум гаплотипам. Иными словами, если существовала монофилетическая группа, включающая три гаплотипа из L1, L2, и L3, оставшиеся три гаплотипа этих индивидуумов

никогда не формировали монофилетическую группу. Мы провели поиск случаев, в которых корень дерева (помещенный посередине) отделял бы монофилетическую группу из трех гаплотипов L1, L2, и L3 от остальной части дерева. Такой паттерн был найден для 36 из 152 сегментов. Эти 36 сегментов были распределены между 35 контигами диплоидной сборки L1.

4.1.21 Проведение SOWH тестов

Чтобы убедиться в том, что присутствие двух групп среди гаплотипов L1-L3 имеет статистическую поддержку, для каждого из 152 сегментов мы провели SOWH [185] (сокращение от англ. Swofford–Olsen–Waddell–Hillis) тест, в котором дерево, полученное методом максимального правдоподобия без наложения ограничений на топологию дерева, сравнивается с деревом, полученным методом максимального правдоподобия, при построении которого требуется, чтобы 6 гаплотипов индивидуумов L1-L3 формировали монофилетическую группу. SOWH тесты проводили с использованием программы SOWHAT [186] (ревизия 907c289, https://github.com/josephryan/sowhat); односторонние P-значения были вычислены на основе 10,000 бутстреп-реплик, полученных в результате симуляций. Из 152 сегментов в 148 дерево, полученное методом максимального правдоподобия без наложения ограничений на топологию дерева, значимо отличалось от дерева, при построении которого требовалось, чтобы 6 гаплотипов индивидуумов L1-L3 формировали монофилетическую группу, на уровне значимости 0.05. После поправки Бонферрони разница между деревьями осталась статистически значимой для 117 сегментов.

4.1.22 Построение митохондриальной филогении

Для того, чтобы проанализировать то, как митохондриальные гаплотипы разных особей соотносятся друг с другом, мы построили митохондриальное филогенетическое дерево для десяти особей L2-L11, имеющих близкие митохондриальные гаплотипы, а также для всех одиннадцати особей L1-L11.

Для этого мы восстановили митохондриальные гаплотипы для каждой из десяти особей L2-L11, использовав митохондриальные генотипы, определенные на основании картирований прочтений L2-L11 на митохондриальный контиг L4 (длина 14,052 п.н.; см. раздел 4.2.8). Для реконструкции митохондриальных гаплотипов генотипы были подвергнуты жесткой фильтрации: мы не рассматривали (1) сайты, определенные как «гетерозиготные» (3 сайта), (2) полиморфные сайты на расстоянии менее или равном десяти нуклеотидов от инсерции или делеции, (3) инсерции и делеции, (4) сайты, в которых генотипы для части особей не были определены, и сайты с покрытием DP <50 в одной или более особях L2-L11, (5) сайты с ненадежно определенными генотипами (QUAL <15). В результате данной процедуры

фильтрации для анализа осталось доступно 13,765 надежных сайтов, что соответствует 98% референтного митохондриального контига (L4). Это позволило восстановить почти полные митохондриальные последовательности для особей L2-L11 с помощью команды “consensus” из программы BCFtools. Митохондриальная филогения L2-L11 (см. раздел 4.2.8) была построена на основе полученных последовательностей с применением программы RAxML (версия 8.2.12) [187] (модель GTR+G, 1000 бутстреп-реплик).

Кроме того, мы дополнительно восстановили митохондриальную филогению для всех 11 особей L1-L11. Значительная степень дивергенции между митохондриальным гаплотипом L1 и гаплотипами L2-L11 осложняет одновременное определение митохондриальных полиморфизмов для всех 11 особей (см. раздел 4.2.7). В связи с этим для того, чтобы получить выравнивание достаточно длинного митохондриального фрагмента для всех 11 особей, мы построили выравнивание самого длинного митохондриального контига L1 (contig8072, длина = 7124 п.н.; см. раздел 4.2.8) с митохондриальными гаплотипами L2-L11, реконструированными на предыдущем этапе. Выравнивание было построено с помощью программы MUSCLE (версия 3.8.31) [111]. Для восстановления митохондриальной филогении L1-L11 мы использовали участок выравнивания, соответствующий области митохондриального генома, присутствующей в контиге contig8072 особи L1 (общая длина выравнивания = 7,126 п.н.). Фланкирующие участки выравнивания, не включающие contig8072, были удалены до построения дерева. Построение митохондриальной филогении для L1-L11 было проведено аналогично тому, как это было сделано для L2-L11 (см. выше). Построение филогенетических деревьев для особей *A. vaga*, описанное в данном разделе, и подготовка данных были выполнены С. А. Науменко.

4.1.23 Оценка популяционной скорости возникновения мутаций θ

Мы получили оценки популяционной скорости возникновения мутаций θ ($4N_e\mu$) с использованием метода максимального правдоподобия, реализованного в программе mlRho [176]. Оценки были получены для каждого индивидуума независимо на основе геномных сайтов с покрытием не менее 20 прочтений в данном индивидууме. Оценки θ для индивидуумов из большого кластера находятся в диапазоне от 0.0072 до 0.0094 (Таблица 4.8) со средним значением равным 0.0086. Оценки θ для индивидуумов из маленького кластера выше (от 0.0221 до 0.0226; Таблица 4.8), что является отражением более высокого уровня гетерозиготности в маленьком кластере. 95% доверительные интервалы для оценок θ приведены в Таблице 4.8.

Особь	Число сайтов (≥20×)	$4N_e\mu$	95% доверительный интервал для $4N_e\mu$		Частота ошибок секвенирования
L1	71,024,512	2.21E-02	2.21E-02	2.21E-02	4.40E-04
L2	72,189,265	2.26E-02	2.26E-02	2.27E-02	5.44E-04
L3	70,758,489	2.23E-02	2.23E-02	2.24E-02	4.19E-04
L4	67,943,878	8.15E-03	8.12E-03	8.17E-03	8.30E-04
L5	70,805,730	9.40E-03	9.37E-03	9.42E-03	4.53E-04
L6	69,056,642	8.63E-03	8.61E-03	8.65E-03	5.17E-04
L7	69,092,259	8.72E-03	8.70E-03	8.75E-03	4.81E-04
L8	69,160,624	8.77E-03	8.74E-03	8.79E-03	5.58E-04
L9	69,951,042	8.80E-03	8.78E-03	8.82E-03	5.86E-04
L10	68,708,316	7.22E-03	7.20E-03	7.24E-03	4.79E-04
L11	69,943,870	8.93E-03	8.91E-03	8.95E-03	5.48E-04

Таблица 4.8. Оценки $4N_e\mu$ и частоты ошибок секвенирования, полученные для особей *A. vaga* L1-L11 методом максимального правдоподобия. Оценки $4N_e\mu$ и частоты ошибок секвенирования были получены независимо для каждой особи с применением метода максимального правдоподобия, реализованного в программе mlRho. Оценки основаны на сайтах с покрытием прочтениями Illumina в рассматриваемой особи не менее чем 20×. Идентификаторы особей, принадлежащих к маленькому и большому кластеру, выделены голубым и зеленым соответственно.

4.1.24 Оценка популяционной скорости рекомбинации

Для того, чтобы рассчитать популяционную скорость рекомбинации ($4N_e c$) на основании скорости распада LD, мы оценили скорость падения r^2 с физическим расстоянием, применив нелинейную регрессию. Использованная модель основана на формуле [188], описывающей ожидаемое значение r^2 при равновесии между рекомбинацией и дрейфом генов с поправкой на низкую скорость возникновения мутаций и размер выборки (n):

$$E(r^2) = \left[\frac{(10+C)}{(2+C)(11+C)}\right] \times \left[1 + \frac{(3+C)(12+12C+C^2)}{n(2+C)(11+C)}\right] \quad (4.1)$$

где $C = 4N_e\, c_{sites}$, N_e обозначает эффективную численность популяции, а c_{sites} долю сайтов, между которыми случилась рекомбинация, среди всех сайтов на данном расстоянии. Для оценки скорости распада LD мы использовали пары полиморфных вариантов, находящихся на расстоянии не более чем 500 п.н. друг от друга, с частотой минорного варианта ≥4 (среди восьми индивидуумов большого кластера L4-L11). Мы оценили параметр $4N_e\, c$ с помощью нелинейной регрессии r^2 от расстояний между сайтами в паре с использованием

модифицированного скрипта Marroni *et al.* [189]. Оценки $4N_e c$, основанные на вариантах из набора фазированных данных 1 и набора фазированных данных 2, были равны 0.0160 (95% доверительный интервал 0.0157–0.0164) и 0.0147 (95% доверительный интервал 0.0141–0.0152) соответственно. Доверительные интервалы были получены на основании 1000 бутстреп-реплик.

Кроме того, мы получили оценки $4N_e c$ с использованием более простой модели без поправок на мутации и размер выборки [190]. В этой базовой модели [190], предполагающей равновесие между дрейфом генов и рекомбинацией, ожидаемое значение r^2 описывается уравнением:

$$E(r^2) = \frac{1}{1 + C} \qquad (4.2)$$

Оценки $4N_e c$, полученные с применением нелинейной регрессии на основании этого уравнения, составили 0.0263 (95% доверительный интервал 0.0259–0.0267) и 0.0262 (95% доверительный интервал 0.0254–0.0271) для набора фазированных данных 1 и 2 соответственно.

Дополнительно мы оценили популяционную скорость рекомбинации с применением метода моментов Вэйкли [191], имплементированного в программе LDhat [180]. Для этого мы получили оценки $4N_e c$ с использованием данного метода для всех фазированных сегментов из набора фазированных данных 1, несущих не менее чем 5 фазированных вариантов с частотой минорного варианта ≥2 среди восьми индивидуумов большого кластера L4-L11 (n = 1962). Оценки $4N_e c$ были получены для каждого сегмента и нормализованы на размер сегмента. Затем мы исключили сегменты, соответствующие статистическим выбросам по нормализованному $4N_e c$ (n = 242; определение выбросов проводилось с применением метода, основанного на расчете межквартильного диапазона). После этого были также исключены сегменты, несущие менее 20 фазированных вариантов с частотой минорного варианта ≥2, в результате чего для анализа остались доступны 245 геномных сегментов. Медианная длина сегмента из получившегося набора составила 901 п.н., а медианное число фазированных вариантов с частотой минорного варианта ≥2 составило 26. Медианное значение оценки нормализованного $4N_e c$ для этих 245 сегментов составило 0.0499, что сопоставимо с оценками, основанными на скорости распада LD.

4.1.25 Оценка гипотетической частоты мейоза в популяции *A. vaga*

Для того, чтобы ответить на вопрос о том, какая частота мейоза необходима для того, чтобы в отсутствие других процессов, сопряженных с рекомбинацией, объяснить картину распада LD у *A. vaga*, необходимо знать c, скорость рекомбинации на нуклеотид на поколение.

Оценку c в свою очередь можно получить из отношения популяционной скорости рекомбинации $4N_e c$ (где параметр N_e – эффективная численность популяции) к популяционной скорости возникновения мутаций $\theta = 4N_e \mu$, если известен параметр μ, обозначающий скорость возникновения мутаций на нуклеотид на поколение. Полученные оценки популяционной скорости рекомбинации $4N_e c$ для индивидуумов L4-L11 имеют порядок 10^{-2} (разные оценки, полученные на основании скорости распада LD, находятся в диапазоне от 0.0147 до 0.0263; см. раздел 4.1.24). Уровень генетической изменчивости указывает на то, что $4N_e \mu$ также ~10^{-2}. Таким образом, $c \sim \mu$. К сожалению, данные о скорости мутирования у *A. vaga* отсутствуют. Если эта скорость находится в диапазоне 10^{-9}–10^{-8}, как у большого числа различных многоклеточных эукариот [192,193], то c тоже имеет порядок 10^{-9}–10^{-8}.

Наконец, мы можем оценить частоту мейоза, необходимую для того, чтобы получить такие значения c. Обозначим как G размер гаплоидного генома в нуклеотидах и как n число хромосом в гаплоидном наборе. Если исходить из предположения о том, что все хромосомы одного размера и что на каждую пару хромосом за одно мейотическое событие случается один кроссинговер, вероятность мейоза на поколение можно оценить как:

$$\frac{G \times c}{n} \qquad (4.3)$$

Число нуклеотидов в гаплоидном геноме *A. vaga* ~10^8, а число хромосом в диплоидном наборе у этого вида 2n = 12 (n = 6) [194]. Таким образом, гипотетическая частота мейоза у *A. vaga* может быть оценена как $\frac{G \times c}{n} = \frac{10^8 \times c}{6}$.

Следовательно, для того, чтобы получить c в диапазоне 10^{-9}–10^{-8}, необходимо 1 событие мейоза на ~10–100 поколений.

Из этого можно заключить, что для того, чтобы в отсутствие других процессов, сопряженных с рекомбинацией, объяснить картину распада LD, мейоз у *A. vaga* должен происходить достаточно часто. Это предположение плохо согласуется с данными об отсутствии самцов среди нескольких сотен тысяч проанализированных особей бделлоидных коловраток [24]. Можно предположить, что наши оценки гипотетической частоты мейоза у *A. vaga* являются завышенными. Во-первых, возможно, что истинная скорость возникновения мутаций у *A. vaga* значительно ниже 10^{-9}–10^{-8} на нуклеотид на поколение (диапазона, из которого мы исходили в расчетах). В этом случае оценки c и частоты мейоза были бы скорректированы в сторону более низких значений. Однако среди эукариот такие низкие скорости возникновения мутаций описаны только для одноклеточных видов (e.g. *Saccharomyces cerevisiae* или *Chlamydomonas reinhardtii*) [195]. Во-вторых, реципрокная митотическая рекомбинация и генная конверсия тоже могут вносить вклад в распад LD. В таком случае наши расчеты,

основанные на предположении о том, что единственной причиной распада LD является реципрокная мейотическая рекомбинация, будут приводить к завышенным оценкам частоты мейоза. И, наконец, наши данные не позволяют исключить возможность того, что в распад LD может вносить вклад не только мейотическая рекомбинация, но и горизонтальный перенос генов между индивидуумами.

4.2 Результаты и обсуждение

4.2.1 Популяционная геномика *A. vaga*

Для того, чтобы ответить на вопрос о существовании рекомбинации и обмена генетическим материалом у бделлоидных коловраток, мы секвенировали геномы 11 особей из естественной популяции *A. vaga* (Рисунок 4.2a). Для этого мы сначала получили 11 клональных линий *A. vaga*, L1-L11. Каждая из этих линий является потомством одной особи, соответствующей по морфологическим критериям виду *A. vaga*, в связи с этим далее понятие «линия» используется взаимозаменяемо с понятием «индивидуум» или «особь». Мы подтвердили принадлежность особей L1-L11 к виду *A. vaga* с помощью построения филогении по митохондриальному маркеру *COX1* (Рисунки 4.3 и 4.4). Затем каждая из полученных клональных линий, L1-L11, была секвенирована на платформе Illumina HiSeq с покрытием ~40–100× (см. Материалы и методы; Таблица 4.1). В дополнение к этому мы секвенировали одну из одиннадцати линий (L1) на платформе MiSeq, что позволило получить *de novo* геномную сборку для этой линии (Таблицы 4.3 и 4.4; Рисунок 4.1). С точки зрения «полноты» полученная сборка содержит ~90% «универсальных» генов [196] (анализ «полноты» сборки был выполнен Е. С. Герасимовым), присутствующих почти у всех эукариот или почти всех многоклеточных животных, и в этом отношении близко напоминает ранее опубликованные геномы бделлоидных коловраток *A. vaga* [25] и *A. ricciae* [26].

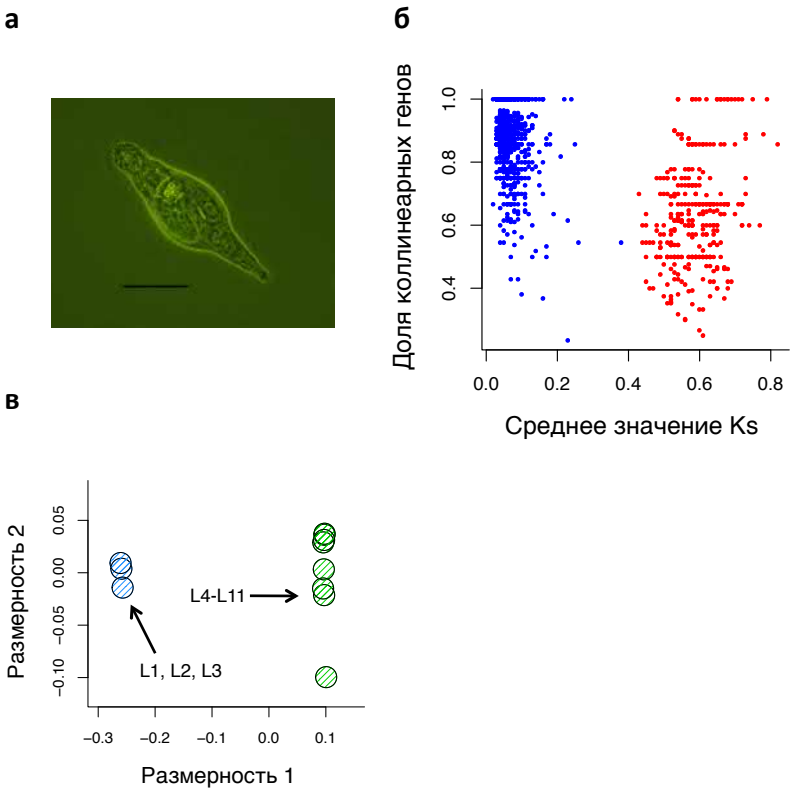

Рисунок 4.2. Полногеномное секвенирование 11 индивидуумов бделлоидной коловратки *A. vaga*. (**а**) Микрофотография особи вида *A. vaga*. Размер приведённой шкалы ~100 мкм (на основании характерного размера мастакса [156] коловраток вида *A. vaga*). Фотография приведена в целях иллюстрации. Клональные линии были получены из 11 особей, соответствующих морфологическому описанию вида *A. vaga*, и секвенированы с покрытием ~40–100×. Образцы мха, из которых были выделены коловратки, были собраны в одном из двух мест: Московской области или Костромской области, 550 км к северо-востоку. В случае клональных линий, «происходящих» из одной местности, линии были получены из коловраток, собранных с разных деревьев, находящихся на расстоянии как минимум 20 м друг от друга. (**б**) Признаки вырожденной тетраплоидности в референтном геноме *A. vaga* (линия L1). График показывает среднее число синонимических замен между коллинеарными генами на сайт (Ks) vs. доли коллинеарных генов в синтеничных блоках. Показаны только те блоки, в состав которых входит 5 или более коллинеарных генов (n = 1769). Блоки со средним значением Ks <0.4 показаны в синем цвете; блоки со средним значением Ks ≥0.4 показаны в красном цвете. (**в**) Генетическая дифференциация между 11 рассматриваемыми особями L1-L11. Многомерное шкалирование (к размерности 2) попарных IBS (сокращение от англ. термина identity-by-state) расстояний между особями L1-L11. Особи из маленького (L1-L3) и из большого (L4-L11) кластеров показаны в голубом и зелёном цветах соответственно. Кластеризация не отражает географию сбора образцов: в состав маленького кластера входят три особи (L1-L3), выделенные

из образцов мха, собранных в Московской области, в то время как большой кластер включает остальных индивидуумов из образцов мха, собранных в Московской области (L4, L6-L10), и двух индивидуумов (L5 и L11) из образцов мха, собранных в Костромской области.

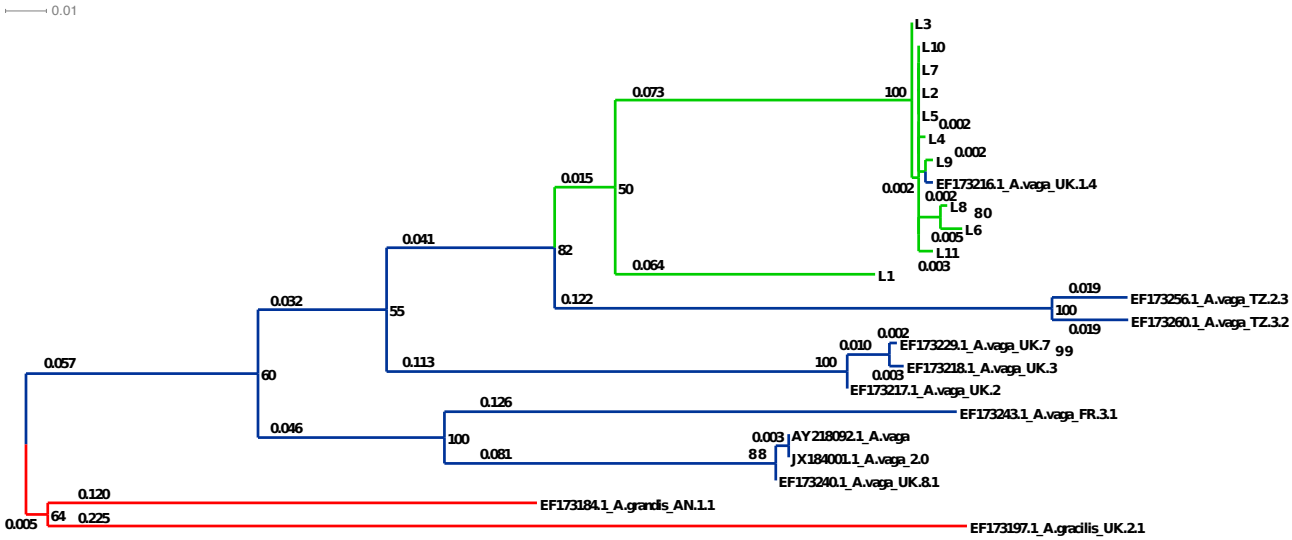

Рисунок 4.3. Филогенетическое дерево индивидуумов L1-L11 и референтных изолятов из рода *Adineta*, построенное для гена *COX1*. Филогенетическое дерево было реконструировано на основе частичных последовательностей гена *COX1* с применением метода максимального правдоподобия, реализованного в программе RAxML [187], с использованием модели GTR+G (1000 бутстреп-реплик). Ветви, ведущие к индивидуумам L1-L11 (секвенированным в данной работе), показаны зеленым цветом; ветви, ведущие к референтным изолятам *A. vaga* (секвенированным ранее в других работах), показаны синим цветом; ветви, ведущие к изолятам, относящимся к другим видам из рода *Adineta*, показаны красным цветом. Референтная линия *A. vaga*, геном которой был секвенирован для получения первой геномной сборки *A. vaga* [25], обозначена как JX184001.1_A.vaga_2.0. Значения бутстреп-поддержки не показаны на очень коротких ветвях. Данное дерево построено С. А. Науменко.

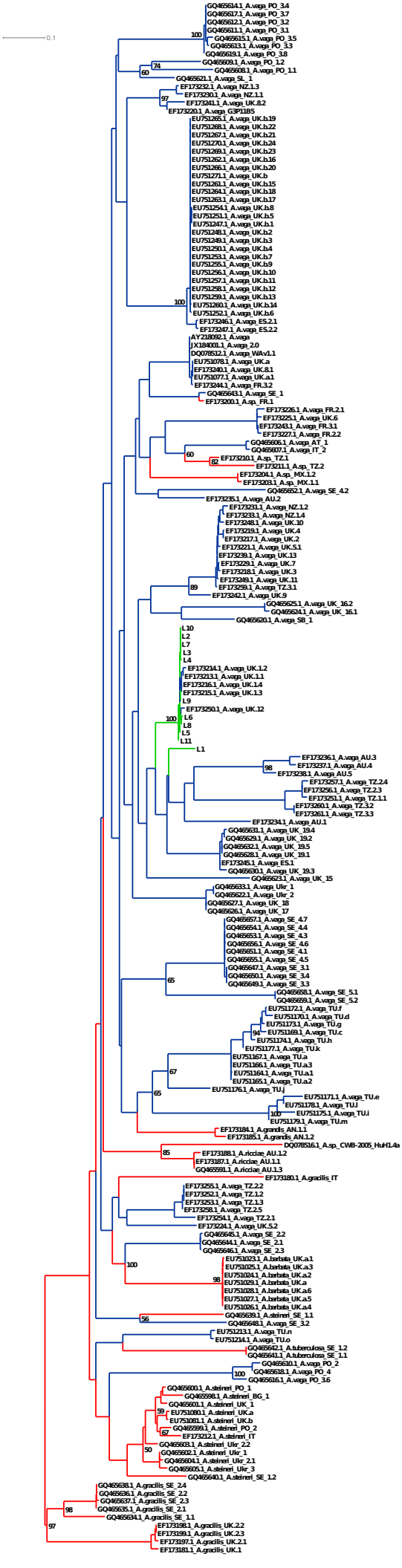
0.1
GQ465614.1_A.vaga_PO_3.4
GQ465617.1_A.vaga_PO_3.7
GQ465612.1_A.vaga_PO_3.2
GQ465611.1_A.vaga_PO_3.1
GQ465615.1_A.vaga_PO_3.5
GQ465613.1_A.vaga_PO_3.3
GQ465619.1_A.vaga_PO_3.8
GQ465609.1_A.vaga_PO_1.2
GQ465608.1_A.vaga_PO_1.1
GQ465621.1_A.vaga_SL_1
EF173232.1_A.vaga_NZ.1.3
EF173230.1_A.vaga_NZ.1.1
EF173241.1_A.vaga_UK.8.2
EF173220.1_A.vaga_G3P1185
EU751265.1_A.vaga_UK.b.19
EU751268.1_A.vaga_UK.b.22
EU751267.1_A.vaga_UK.b.21
EU751270.1_A.vaga_UK.b.24
EU751269.1_A.vaga_UK.b.23
EU751262.1_A.vaga_UK.b.16
EU751266.1_A.vaga_UK.b.20
EU751271.1_A.vaga_UK.b
EU751261.1_A.vaga_UK.b.15
EU751264.1_A.vaga_UK.b.18
EU751263.1_A.vaga_UK.b.17
EU751254.1_A.vaga_UK.b.8
EU751253.1_A.vaga_UK.b.5
EU751247.1_A.vaga_UK.b.1
EU751248.1_A.vaga_UK.b.2
EU751249.1_A.vaga_UK.b.3
EU751250.1_A.vaga_UK.b.4
EU751253.1_A.vaga_UK.b.7
EU751255.1_A.vaga_UK.b.9
EU751256.1_A.vaga_UK.b.10
EU751257.1_A.vaga_UK.b.11
EU751258.1_A.vaga_UK.b.12
EU751259.1_A.vaga_UK.b.13
EU751260.1_A.vaga_UK.b.14
EU751252.1_A.vaga_UK.b.6
EF173246.1_A.vaga_ES.2.1
EF173247.1_A.vaga_ES.2.2
AY218092.1_A.vaga
JX184001.1_A.vaga_2.0
DQ078512.1_A.vaga_WAv1.1
EU751078.1_A.vaga_UK.a
EF173240.1_A.vaga_UK.8.1
EU751077.1_A.vaga_UK.a.1
EF173244.1_A.vaga_FR.3.2
GQ465643.1_A.vaga_SE_1
EF173200.1_A.sp._FR.1
EF173226.1_A.vaga_FR.2.1
EF173225.1_A.vaga_UK.6
EF173243.1_A.vaga_FR.3.1
EF173227.1_A.vaga_FR.2.2
GQ465606.1_A.vaga_AT_1
GQ465607.1_A.vaga_IT_2
EF173210.1_A.sp._TZ.1
EF173211.1_A.sp._TZ.2
EF173204.1_A.sp._MX.1.2
EF173203.1_A.sp._MX.1.1
GQ465652.1_A.vaga_SE_4.2
EF173235.1_A.vaga_AU.2
EF173231.1_A.vaga_NZ.1.2
EF173233.1_A.vaga_NZ.1.4
EF173248.1_A.vaga_UK.10
EF173219.1_A.vaga_UK.4
EF173217.1_A.vaga_UK.2
EF173221.1_A.vaga_UK.5.1
EF173239.1_A.vaga_UK.13
EF173229.1_A.vaga_UK.7
EF173218.1_A.vaga_UK.3
EF173249.1_A.vaga_UK.11
EF173259.1_A.vaga_TZ.3.1
EF173242.1_A.vaga_UK.9
GQ465625.1_A.vaga_UK_16.2
GQ465624.1_A.vaga_UK_16.1
GQ465620.1_A.vaga_SB_1
L10
L2
L7
L3
L4
EF173214.1_A.vaga_UK.1.2
EF173213.1_A.vaga_UK.1.1
EF173216.1_A.vaga_UK.1.4
EF173215.1_A.vaga_UK.1.3
L9
EF173250.1_A.vaga_UK.12
L6
L8
L5
L11
L1
EF173236.1_A.vaga_AU.3
EF173237.1_A.vaga_AU.4
EF173238.1_A.vaga_AU.5
EF173257.1_A.vaga_TZ.2.4
EF173256.1_A.vaga_TZ.2.3
EF173251.1_A.vaga_TZ.1.1
EF173260.1_A.vaga_TZ.3.2
EF173261.1_A.vaga_TZ.3.3
EF173234.1_A.vaga_AU.1
GQ465631.1_A.vaga_UK_19.4
GQ465629.1_A.vaga_UK_19.2
GQ465632.1_A.vaga_UK_19.5
GQ465628.1_A.vaga_UK_19.1
EF173245.1_A.vaga_ES.1
GQ465630.1_A.vaga_UK_19.3
GQ465623.1_A.vaga_UK_15
GQ465633.1_A.vaga_Ukr_1
GQ465622.1_A.vaga_Ukr_2
GQ465627.1_A.vaga_UK_18
GQ465626.1_A.vaga_UK_17
GQ465657.1_A.vaga_SE_4.7
GQ465654.1_A.vaga_SE_4.4
GQ465653.1_A.vaga_SE_4.3
GQ465656.1_A.vaga_SE_4.6
GQ465651.1_A.vaga_SE_4.1
GQ465655.1_A.vaga_SE_4.5
GQ465647.1_A.vaga_SE_3.1
GQ465650.1_A.vaga_SE_3.4
GQ465649.1_A.vaga_SE_3.3
GQ465658.1_A.vaga_SE_5.1
GQ465659.1_A.vaga_SE_5.2
EU751172.1_A.vaga_TU.f
EU751170.1_A.vaga_TU.d
EU751173.1_A.vaga_TU.g
EU751169.1_A.vaga_TU.c
EU751174.1_A.vaga_TU.h
EU751177.1_A.vaga_TU.k
EU751167.1_A.vaga_TU.a
EU751166.1_A.vaga_TU.a.3
EU751164.1_A.vaga_TU.a.1
EU751165.1_A.vaga_TU.a.2
EU751176.1_A.vaga_TU.j
EU751171.1_A.vaga_TU.e
EU751178.1_A.vaga_TU.l
EU751175.1_A.vaga_TU.i
EU751179.1_A.vaga_TU.m
EF173184.1_A.grandis_AN.1.1
EF173185.1_A.grandis_AN.1.2
DQ078516.1_A.sp._CWB-2005_HuH1.4a
EF173188.1_A.ricciae_AU.1.2
EF173187.1_A.ricciae_AU.1.1
GQ465591.1_A.ricciae_AU.1.3
EF173180.1_A.gracilis_IT
EF173255.1_A.vaga_TZ.2.2
EF173252.1_A.vaga_TZ.1.2
EF173253.1_A.vaga_TZ.1.3
EF173258.1_A.vaga_TZ.2.5
EF173254.1_A.vaga_TZ.2.1
EF173224.1_A.vaga_UK.5.2
GQ465645.1_A.vaga_SE_2.2
GQ465644.1_A.vaga_SE_2.1
GQ465646.1_A.vaga_SE_2.3
EU751023.1_A.barbata_UK.a.1
EU751025.1_A.barbata_UK.a.3
EU751024.1_A.barbata_UK.a.2
EU751029.1_A.barbata_UK.a
EU751028.1_A.barbata_UK.a.6
EU751027.1_A.barbata_UK.a.5
EU751026.1_A.barbata_UK.a.4
GQ465639.1_A.steineri_SE_1.1
GQ465648.1_A.vaga_SE_3.2
EU751213.1_A.vaga_TU.n
EU751214.1_A.vaga_TU.o
GQ465642.1_A.tuberculosa_SE_1.2
GQ465641.1_A.tuberculosa_SE_1.1
GQ465610.1_A.vaga_PO_2
GQ465618.1_A.vaga_PO_4
GQ465616.1_A.vaga_PO_3.6
GQ465600.1_A.steineri_PO_1
GQ465598.1_A.steineri_BG_1
GQ465601.1_A.steineri_UK_1
EU751080.1_A.steineri_UK.a
EU751081.1_A.steineri_UK.b
GQ465599.1_A.steineri_PO_2
EF173212.1_A.steineri_IT
GQ465603.1_A.steineri_Ukr_2.2
GQ465602.1_A.steineri_Ukr_1
GQ465604.1_A.steineri_Ukr_2.1
GQ465605.1_A.steineri_Ukr_3
GQ465640.1_A.steineri_SE_1.2
GQ465638.1_A.gracilis_SE_2.4
GQ465636.1_A.gracilis_SE_2.2
GQ465637.1_A.gracilis_SE_2.3
GQ465635.1_A.gracilis_SE_2.1
GQ465634.1_A.gracilis_SE_1.1
EF173198.1_A.gracilis_UK.2.2
EF173199.1_A.gracilis_UK.2.3
EF173197.1_A.gracilis_UK.2.1
EF173181.1_A.gracilis_UK.1

Рисунок 4.4. Филогенетическое дерево индивидуумов L1-L11 и расширенного набора референтных изолятов из рода *Adineta*, построенное для гена *COX1*. Филогенетическое дерево было реконструировано на основе частичных последовательностей гена *COX1* с применением метода максимального правдоподобия, реализованного в программе RAxML [187], с использованием модели GTR+G (1000 бутстреп-реплик). Ветви, ведущие к индивидуумам L1-L11 (секвенированным в данной работе), показаны зеленым цветом; ветви, ведущие к референтным изолятам *A. vaga* (секвенированным ранее в других работах), показаны синим цветом; ветви, ведущие к изолятам, относящимся к другим видам из рода *Adineta*, показаны красным цветом. Референтная линия *A. vaga*, геном которой был секвенирован для получения первой геномной сборки *A. vaga* [25], обозначена как JX184001.1_A.vaga_2.0. Значения бутстреп-поддержки ≥50% показаны рядом с соответствующими ветвями (некоторые значения бутстреп-поддержки ≥50% не показаны из-за ограничений места). Данное дерево построено С. А. Науменко.

Анализ геномной сборки, полученной для L1 (далее именуется как «референтный геном»), выявил подписи, указывающие на древнюю «вырожденную» тетраплоидность, аналогичные тем, которые были описаны для ранее опубликованного генома *A. vaga* [25]. Так, внутри генома L1 можно выделить пары гомологичных друг другу участков, в которых гены идут в одном и том же порядке (так называемые «коллинеарные блоки»). При этом найденные гомологичные блоки достаточно четко распадаются на две группы (Рисунок 4.2б). Первая группа представлена блоками с высоким процентом коллинеарных генов и низким уровнем дивергенции между гомологичными генами в синонимических сайтах (Рисунок 4.2б) и, по всей видимости, соответствует парам гаплотипов, присутствующим в сборке [25]. Вторая группа представлена блоками со значительно более высоким уровнем синонимической дивергенции и меньшей долей коллинеарных генов, и, вероятно, представляет собой результат древней полногеномной дупликации [25].

Поскольку большая часть первоначальной сборки L1 представлена двумя гаплотипами, прежде чем перейти к популяционным анализам, мы получили гаплоидное представление генома L1 (далее именуется как «гаплоидная сборка»; см. Материалы и методы и Таблицу 4.3). Гаплоидная сборка была использована как референтный геном, на который картировались парно-концевые прочтения для всех одиннадцати образцов. Полученные картирования парно-концевых прочтений были использованы далее для определения однонуклеотидных полиморфизмов (см. Материалы и методы).

Анализ однонуклеотидных полиморфизмов, найденных среди особей L1-L11, выявил присутствие двух генетических кластеров (Рисунки 4.2в и 4.5; Таблица 4.9; см. Материалы и

методы). Среднее попарное генотипическое расстояние между коловратками из разных кластеров равно 1.22%, в то время как среднее попарное расстояние для трех коловраток из маленького кластера (L1-L3) – 0.66%, а для восьми коловраток из большого кластера (L4-L11) – 0.54% (Таблица 4.9).

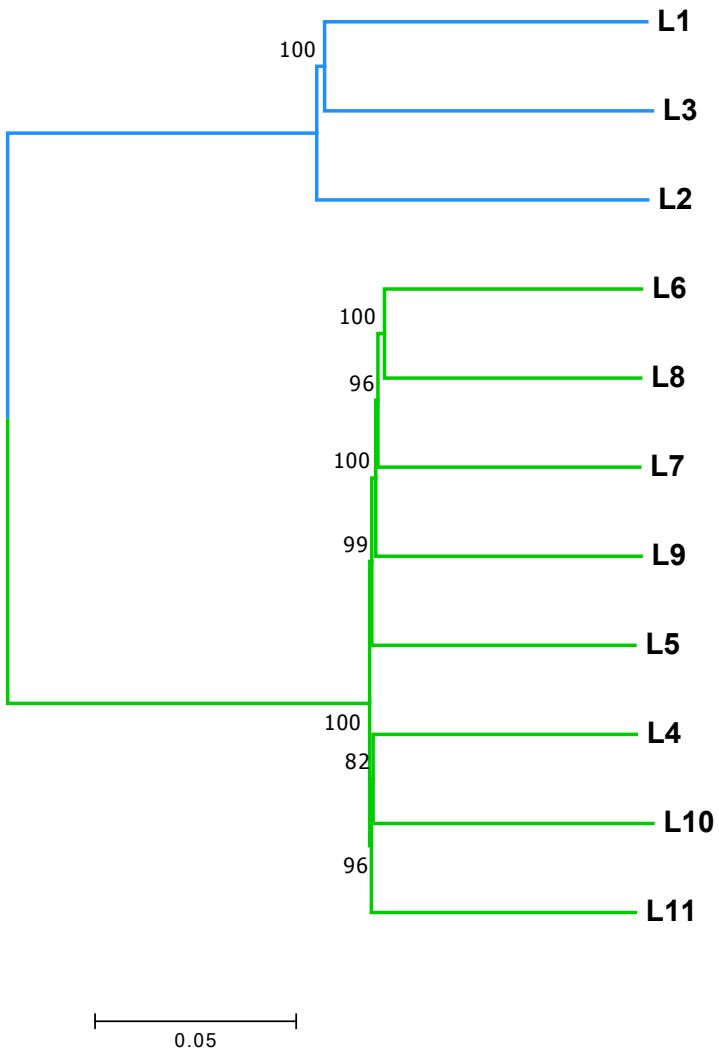

Рисунок 4.5. Дерево, построенное для особей L1-L11 методом ближайших соседей. Дерево основано на расстояниях, рассчитанных для каждой пары особей как доля аллелей различных между геномами данной пары особей. При расчете расстояний учитывались только сайты биаллельные среди L1-L11. Дерево построено с использованием «прореженного» поднабора сайтов биаллельных среди L1-L11 (n = 449,218) из набора однонуклеотидных полиморфизмов I. Неукорененное дерево было получено с использованием функции *aboot* из пакета *poppr* [197,198] для языка R и укоренено посередине в программе MEGA7 [199]. Значения бутстреп-поддержки, рассчитанные на основании 1000 бутстреп-реплик и округленные до ближайшего целого числа, показаны рядом с соответствующими узлами. В данном случае доли

отличающихся аллелей определяли для биаллельных сайтов (мономорфные сайты не учитывались), в связи с чем расстояния, лежащие в основе дерева, по построению значительно выше, чем расстояния, полученные с учетом мономорфных сайтов (Таблица 4.9). Ветви, ведущие к особям маленького и большого кластеров, показаны в синем и зеленом цвете соответственно.

	L1	**L2**	**L3**	**L4**	**L5**	**L6**	**L7**	**L8**	**L9**	**L10**	**L11**
L1	0%										
L2	0.657%	0%									
L3	0.650%	0.675%	0%								
L4	1.212%	1.217%	1.221%	0%							
L5	1.215%	1.217%	1.222%	0.529%	0%						
L6	1.214%	1.217%	1.220%	0.538%	0.528%	0%					
L7	1.213%	1.219%	1.220%	0.538%	0.530%	0.526%	0%				
L8	1.215%	1.219%	1.222%	0.538%	0.529%	0.511%	0.521%	0%			
L9	1.217%	1.222%	1.221%	0.535%	0.533%	0.530%	0.528%	0.531%	0%		
L10	1.229%	1.232%	1.238%	0.542%	0.547%	0.552%	0.553%	0.555%	0.556%	0%	
L11	1.214%	1.215%	1.220%	0.531%	0.532%	0.536%	0.537%	0.535%	0.536%	0.544%	0%

Таблица 4.9. Попарные генотипические расстояния между особями *A. vaga* L1-L11. Генотипические расстояния были рассчитаны на основе сайтов гаплоидной сборки, в которых генотипы были одновременно определены для всех особей L1-L11. Для данного анализа использовались только мономорфные и биаллельные сайты из набора однонуклеотидных полиморфизмов III (n = 58,118,767). Для пары особей *A. vaga* генотипическое расстояние рассчитывали следующим образом: расстояние в каждом рассматриваемом сайте определялось как разница в числе нереферентных вариантов (0, 1 или 2), затем полученные таким образом для всех сайтов значения были суммированы и нормализованы на 2n. Расстояния между особями из разных кластеров выделены оранжевым, расстояния между особями внутри маленького и большого кластеров выделены голубым и зеленым соответственно.

При этом особи, принадлежащие к двум кластерам, значительно отличаются по уровню внутригеномной гетерозиготности: так, средняя доля гетерозиготных сайтов для коловраток из маленького кластера (L1-L3) составляет 1.98%, в то время как соответствующая доля для коловраток из большого кластера (L4-L11) – всего 0.63% (Таблица П4.1; Рисунок 4.6). Соответствующие значения для четырехкратно вырожденных синонимических сайтов для малого и большого кластеров равны 3.75% и 1.21% соответственно (Таблица П4.1).

В представленной в данной главе работе для того, чтобы минимизировать возможный эффект популяционной структуры на результаты, мы сосредоточились в первую очередь на анализе восьми особей, составляющих большой кластер (L4-L11).

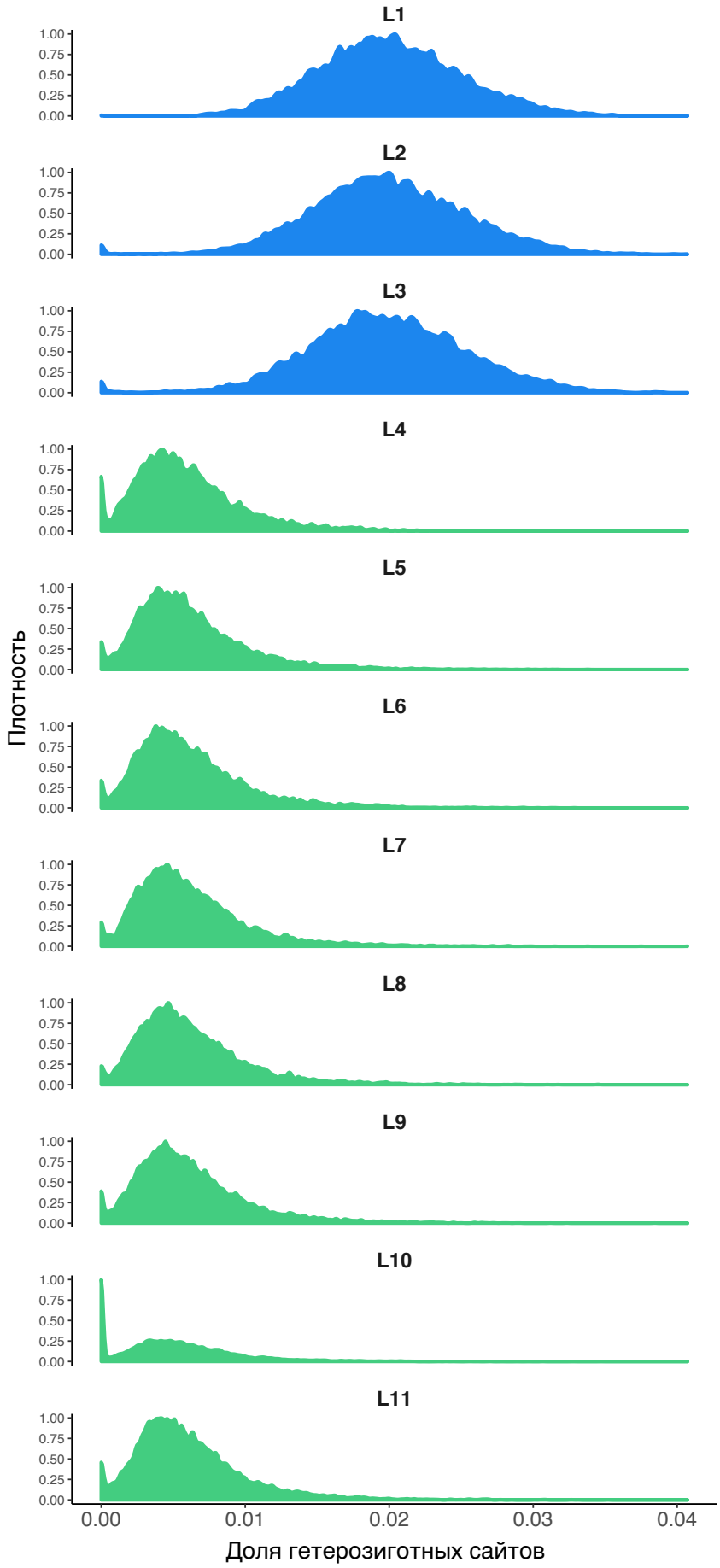
L1
L2
L3
L4
L5
L6
L7
L8
L9
L10
L11
1.00
0.75
0.50
0.25
0.00
Плотность
0.00
0.01
0.02
0.03
0.04
Доля гетерозиготных сайтов

Рисунок 4.6. Распределение плотности гетерозиготности в окнах размером 5 кб в геномах особей *A. vaga* L1-L11. Нормализованные распределения плотности доли гетерозиготных сайтов в геноме показаны для всех 11 особей из маленького (синий цвет) и большого (зеленый цвет) кластеров.

4.2.2 Анализ митохондриальной изменчивости L1-L11

Филогении, построенные для митохондриального гена *COX1*, не поддерживают то же разделение особей L1-L11 на два генетических кластера (L1-L3 и L4-L11), на которое указывают данные по изменчивости ядерных геномов. Важно отметить, что в дереве *COX1* особи L2-L11 перемешаны, в то время как ветка L1 значительно длиннее. Более того, в этом дереве L1 даже группируется не с особями L2-L11, а с референтными изолятами, вероятно, относящимися к другому криптическому виду *A. vaga* (Рисунок 4.4).

Чтобы проверить, может ли этот результат быть связан с техническими артефактами, мы провели дальнейший анализ митохондриальной изменчивости. Поскольку мы секвенировали клональные культуры, полученные из отдельных особей *A. vaga*, а не индивидуальных коловраток, в принципе, представлялось возможным, что дивергентный митохондриальный гаплотип L1 мог быть результатом контаминации. Однако в этом случае мы бы ожидали, что в прочтениях Illumina, полученных для L1, будут присутствовать два или более частых митохондриальных гаплотипа. Та же логика может быть применена для проверки на контаминацию в случае других секвенированных культур: несмотря на то, что некоторая степень митохондриальной гетерогенности может присутствовать даже в клональной культуре, все копии митохондриального генома в такой культуре должны быть очень похожи друг на друга.

Для того, чтобы провести поиск возможных признаков контаминации, мы сначала извлекли митохондриальные контиги из диплоидной сборки L1 и из высоко фрагментированных геномных сборок, полученных для остальных особей L2-L11 из имеющихся прочтений Illumina HiSeq. Для этого мы использовали простой подход, основанный на поиске (с использованием blastn) в индивидуальных геномах L1-L11 контигов похожих на митохондриальный контиг из первого опубликованного генома *A. vaga* [25].

В свою очередь, чтобы извлечь митохондриальный контиг из первой опубликованной геномной сборки *A. vaga*, мы провели поиск blastn против этой сборки с использованием референтных митохондриальных геномов двух других видов бделлоидных коловраток доступных в GenBank: *Philodina citrina* (идентификатор в GenBank FR856884.1; длина 14,003 п.н.) и *Rotaria rotatoria* (идентификатор в GenBank GQ304898.1; длина 15,319 п.н.). Лучшее выравнивание для обеих последовательностей было получено с контигом 2917 из сборки

генома *A. vaga* 2013 года (CAWI020038741.1). Выравнивание между этим контигом и референтными митохондриальными геномами содержит значительную часть референтных митохондриальных последовательностей: 81% для *Philodina* (нуклеотидное сходство 80.07%) и 74% для *Rotaria* (нуклеотидное сходство 79.49%). Двумя лучшими находками для контига 2917 в базе данных nt были соответственно митохондриальные геномы *Philodina* и *Rotaria*. Третья по «весу» находка соответствует гену *COX1 A. vaga* (JX184001.1) [200]; это выравнивание включает только 6% контига, но имеет 100% нуклеотидное сходство. Далее мы использовали контиг 2917 в качестве референтной последовательности митохондриального генома *A. vaga*.

Анализ находок blastn для этого митохондриального контига *A. vaga* в геномах L1-L11 выявил, что основная часть митохондриального генома обычно содержалась в небольшом числе контигов (от одного до трех). В частности, в диплоидной сборке L1 основная часть митохондриального генома была распределена между тремя контигами: contig8072 (длина = 7124 bp), contig11064 (длина = 3631 bp) и contig12085 (длина = 2836 bp), соответствующим трем фрагментам референтного контига *A. vaga*. Контиг L1 contig12085 содержит значительное число коротких тандемных повторов (с размером повторяющегося участка от 13 до 36 п.н.), выявленных с помощью программы Tandem repeats finder [201]. В связи с этим данный контиг не использовался для большинства последующих анализов, однако включение этого контига не оказывало существенного влияния на дальнейшие результаты.

В геномных сборках, полученных для L2, L3, L4, L7, L10 и L11, основная часть митохондриального генома содержалась в одном контиге длиной ~14,000 п.н. (диапазон длин: от 13,781 п.н. для L3 до 14,052 п.н. для L4). В геномных сборках для L5 и L9 существенная часть митохондриального генома содержалась в двух контигах, а в сборках для L6 и L8 в трех контигах (с общей длиной от 9360 п.н. для L6 до 13,967 п.н. для L5).

В соответствии с анализом филогений *COX1* митохондриальные гаплотипы особей L2-L11 очень похожи друг на друга: среднее значение нуклеотидного сходства между митохондриальным контигом L4 и лучшей находкой blastn для этого контига против митохондриальных контигов L2-L3 и L5-L11 составило 99.66% (диапазон значений: от 99.33% для L4 против L11 до 99.88% для L4 против L2). Кроме того, также в соответствии с анализом филогений *COX1*, анализ митохондриальных геномов показал, что митохондриальный гаплотип L1 отличается от гаплотипов L2-L11 значительно сильнее, чем гаплотипы L2-L11 друг от друга. Среднее значение нуклеотидного сходства между митохондриальным контигом L1 contig8072 и лучшей находкой blastn для этого контига против митохондриальных контигов L2-L11 составило всего 91.41% (диапазон значений для поиска против митохондриальных геномов разных особей: от 90.43% до 92.75%). Соответствующее значение для контига L1 contig11064 было равно 91.07% (диапазон значений: от 90.97% до 91.18%). Нуклеотидное сходство между

контигом L1 contig12085 и митохондриальными геномами L2-L11 оказалось еще ниже (среднее значение нуклеотидного сходства лучшей находки blastn в митохондриальных контигах L2-L11 для contig12085 составило 84.27%; диапазон значений: от 83.14% до 84.58%), что в том числе отражает значительное количество инсерций и делеций в выравниваниях, вероятно, связанное с присутствием тандемных повторов в этом участке митохондриального генома.

Поиск blastn с использованием митохондриальных контигов L1 против полных геномных сборок L2-L11 не выявил «альтернативных» митохондриальных гаплотипов с более высокой степенью сходства с гаплотипом L1. Аналогичным образом, поиск blastn с использованием митохондриальных контигов L2-L11 против геномной сборки L1 также не выявил присутствия другого варианта митохондриального гаплотипа среди контигов L1, который был бы более похож на гаплотипы L2-L11. Эти наблюдения подтверждают, что положение L1 в дереве *COX1*, по всей видимости, не объясняется контаминацией культуры L1, а действительно связано с тем, что L1 имеет дивергентный митохондриальный гаплотип, далекий от гаплотипов L2-L11.

4.2.3 Подписи рекомбинации в геномах *A. vaga*

В качестве одного из наиболее очевидных признаков рекомбинации и полового размножения обычно рассматривают падение неравновесия по сцеплению (далее обозначается кратко как LD от англ. термина linkage disequilibrium) с увеличением физического расстояния между парами полиморфных сайтов [202]. И, напротив, у бесполых организмов в случае отсутствия рекомбинации степень сцепления между аллелями в разных локусах не должна зависеть от расстояния между локусами (Рисунок 4.7).

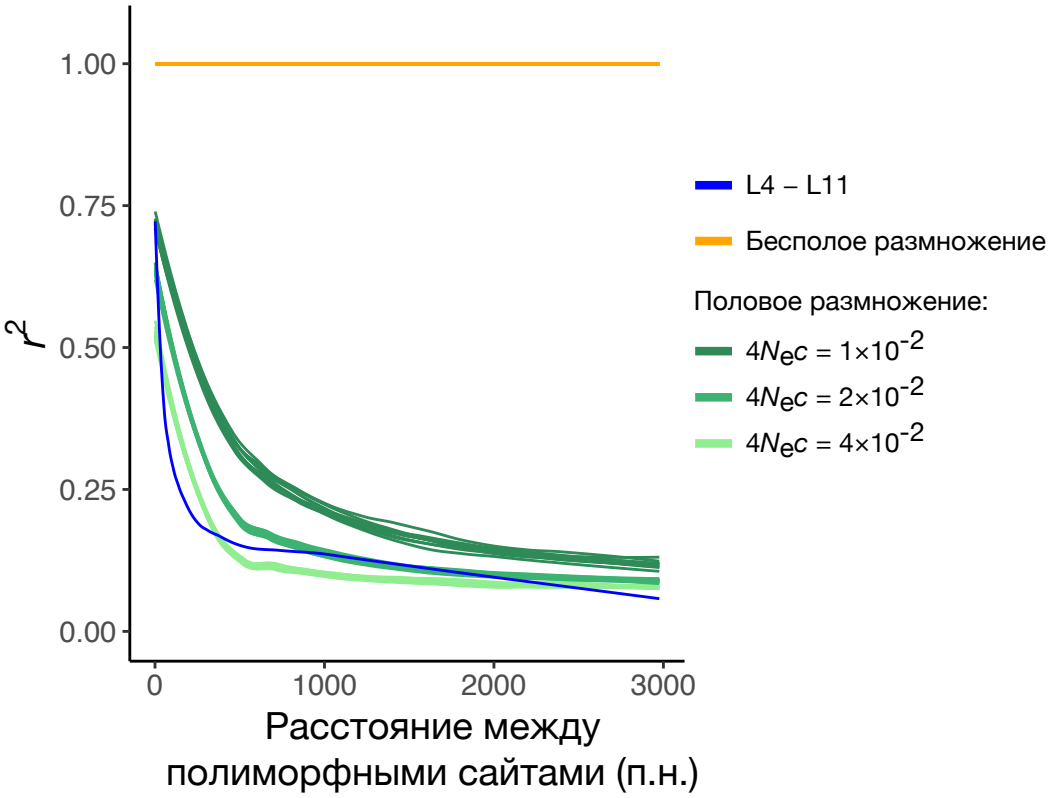

Рисунок 4.7. Зависимость LD от физического расстояния между сайтами в популяциях с бесполым и половым размножением. На рисунке представлены кривые, показывающие аппроксимацию зависимости r^2 от физического расстояния, полученную с помощью локальной регрессии (LOESS) полиномами второй степени (с коэффициентом сглаживания 0.4) для сайтов полиморфных в популяциях, симулированных в SLiM [182], или для сайтов полиморфных среди L4-L11. Параметры симуляций (симулировали популяции со строго бесполым размножением или со строго половым размножением) были подобраны так, чтобы популяционная скорость возникновения мутаций $4N_e\mu$ в симуляциях была близка к значению этого параметра, оцененного из данных для особей L4-L11 (Таблица 4.8): $N_e = 2500$, $\mu = 10^{-6}$. В симуляциях популяций со строго половым размножением значение популяционной скорости рекомбинации $4N_ec$ было равно 1×10^{-2}, 2×10^{-2} или 4×10^{-2}. Каждый вариант симуляции был проведен в 10 репликах. Через 200,000 поколений из каждой реплики каждой симуляции случайно выбирали 8 особей (что соответствует числу анализируемых особей L4-L11) и оставляли только полиморфные среди этих особей сайты с частотой минорного варианта ≥4. Для L4-L11 показаны те же данные, которые представлены на Рисунке 4.8а (биаллельные сайты с частотой минорного варианта ≥4).

Исходя из этих предсказаний, для того, чтобы определить, происходит ли у *A. vaga* рекомбинация, мы начали с анализа неравновесия по сцеплению в большом кластере (L4-L11),

используя полученный на предыдущем этапе анализа данных набор однонуклеотидных полиморфизмов (см. Материалы и методы). В качестве подготовительного этапа для данного анализа мы провели локальную реконструкцию последовательностей гаплотипов (далее коротко – «фазирование») для каждой из 11 особей *A. vaga* по отдельности с помощью программы HapCUT2 [173] (см. Материалы и методы и Таблицы 4.6 и 4.7). В результате этой процедуры для каждой коловратки был получен набор участков генома, в которых для данной особи удалось восстановить гаплотипы (далее такие участки именуются «фазированные сегменты»). Реконструированные гаплотипы были подвергнуты жесткой фильтрации, в результате чего было получено два набора профильтрованных фазированных сегментов: набор фазированных данных 1, использовавшийся для основных анализов, и вспомогательный набор фазированных данных 2, подвергнутый еще более жесткой фильтрации (см. Материалы и методы). Оба набора фазированных сегментов были профильтрованы на основании присутствия противоречий между картированными парно-концевыми прочтениями, набор фазированных данных 2 был дополнительно профильтрован с использованием оценок вероятности ошибок реконструкции гаплотипов (рассчитанных с помощью HapCUT2).

В противоположность тому, что бы ожидалось в отсутствие рекомбинации, оказалось, что в участках генома *A. vaga*, входящих в набор фазированных данных 1, в которых удалось восстановить гаплотипы для всех восьми коловраток из большого кластера (L4-L11), LD быстро убывает с увеличением расстояния между полиморфными сайтами (Рисунок 4.8а) и достигает значений, близких к среднему значению для сайтов на разных контигах, на расстоянии ~2600–2700 нуклеотидов. Такое падение LD с расстоянием сравнимо с тем, что наблюдается в видах с исключительно половым размножением: так, например, LD падает с сопоставимой скоростью у *D. melanogaster*, в то время как у человека наблюдается более медленное снижение LD [203].

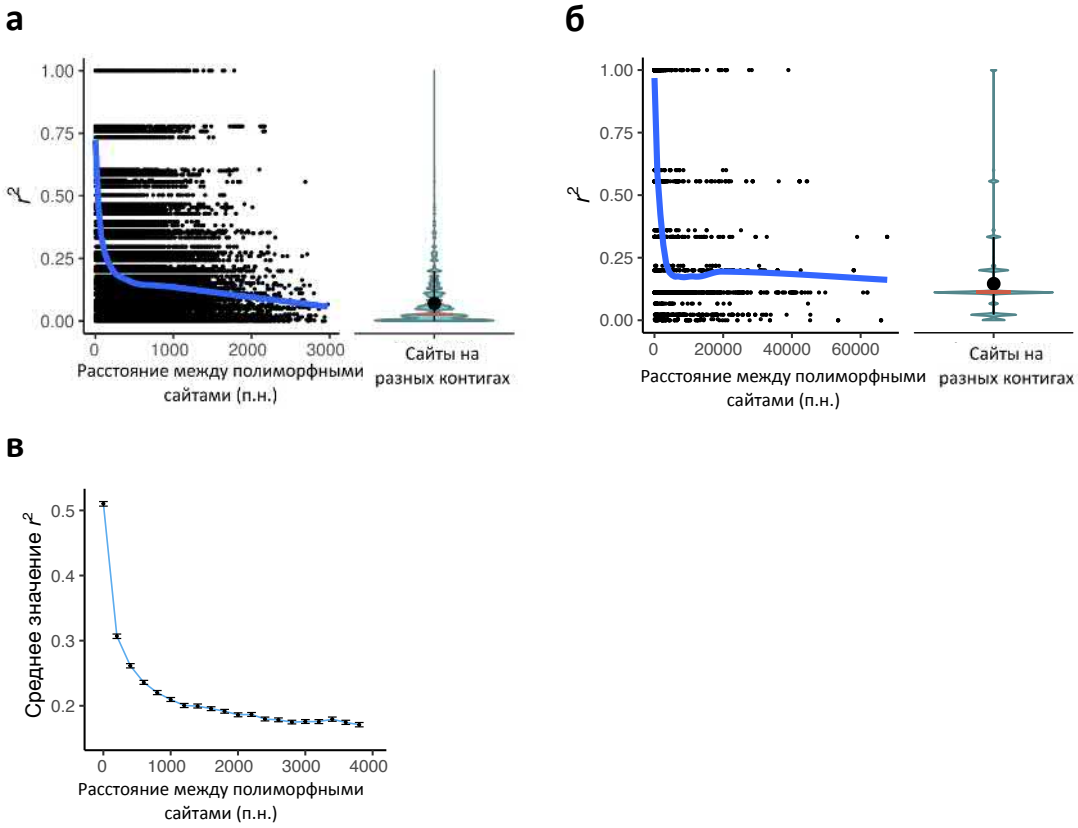

Рисунок 4.8. Распад неравновесия по сцеплению (LD) с физическим расстоянием у *A. vaga*. **а**, **б** LD выражено как r^2. Распад r^2 с физическим расстоянием оценивали с использованием фазированных данных (восстановленных гаплотипов, **а**) и нефазированных данных (генотипов, **б**). Синие кривые показывают аппроксимацию падения r^2 с физическим расстоянием с помощью локальной регрессии (LOESS) полиномами второй степени (с коэффициентом сглаживания 0.4). Скрипичные графики показывают распределение значений r^2 для пар полиморфных сайтов на разных контигах. Концы усов соответствуют 10-й и 90-й перцентилям. Среднее значение и медиана показаны с помощью черной точки и красной горизонтальной линии соответственно. (**а**) Значения r^2 рассчитывали для биаллельных сайтов, находящихся внутри сегментов референтного генома, в которых гаплотипы были восстановлены для всех индивидуумов большого кластера (L4-L11). (**б**) Оценки r^2, полученные с использованием нефазированных данных (генотипов). Значения r^2 рассчитывали для биаллельных сайтов гомозиготных в геномах всех индивидуумов большого кластера. (**в**) LD выражено как квадрат коэффициента корреляции между генотипами. Квадрат коэффициента корреляции рассчитывали между 10,000 случайно выбранными биаллельными сайтами и всеми остальными биаллельными сайтами (с частотой минорного варианта среди L4-L11 ≥4). Пары сайтов разделяли на бины по расстоянию, разделяющему сайты, с шагом 200 п.н.; на графике представлены данные для пар сайтов, находящихся на расстоянии ≤4000 п.н. Чёрные точки показывают среднее значение квадрата коэффициента корреляции для разных бинов

(координаты по оси X соответствуют нижним границам бинов). Планки погрешности соответствуют 95% доверительным интервалам, рассчитанным на основе 1000 бутстреп-реплик.

Затем мы применили тесты, использующиеся для выявления рекомбинации, к индивидуальным фазированным сегментам по отдельности. В частности, мы оценивали, есть ли зависимость между значениями r^2 и расстоянием между полиморфными сайтами [177], и использовали тест на «сумму расстояний» [178] (англ. sum of the distances), а также PHI тест [179]. Эти три теста с помощью разных метрик позволяют оценить, есть ли статистически значимая зависимость между LD и расстоянием, на котором полиморфные сайты находятся друг от друга [177–179]. Все эти три теста выявили сигнал рекомбинации в фазированных сегментах (см. Материалы и методы). В частности, среди 434 фазированных сегментов, для 362 была выявлена статистически значимая отрицательная корреляция между r^2 и физическим расстоянием, разделяющим пары сайтов (на уровне значимости 0.05). После применения поправки Бонферрони отрицательная корреляция осталась значимой для 159 фазированных сегментов. Согласно результатам теста на «сумму расстояний» [178] и PHI теста [179] сигнал рекомбинации был найден соответственно в 108 и 190 сегментах из 434 (P-значение <0.05 после поправки Бонферрони; см. Материалы и методы).

Важно отметить, что сигнал падения LD с расстоянием может в принципе возникнуть и в отсутствие рекомбинации из-за артефактов обработки данных. Во-первых, он может быть связан с ошибочным выравниванием парно-концевых прочтений с паралогичными участками генома. Однако данный сигнал сохраняется и в том случае, если для анализа использовать только поднабор полиморфных локусов, попадающих в участки, соответствующие аллельным блокам [166] (Рисунок 4.9а). Существование двух гаплотипов в геноме L1 в таких участках дополнительно подтверждено присутствием гомологичных генов, идущих в одном и том же порядке и имеющих высокое сходство последовательностей. Поскольку ожидается, что выравнивание прочтений в таких участках будет более надежным, распад LD в них с меньшей вероятностью может быть объяснен ошибками выравнивания прочтений.

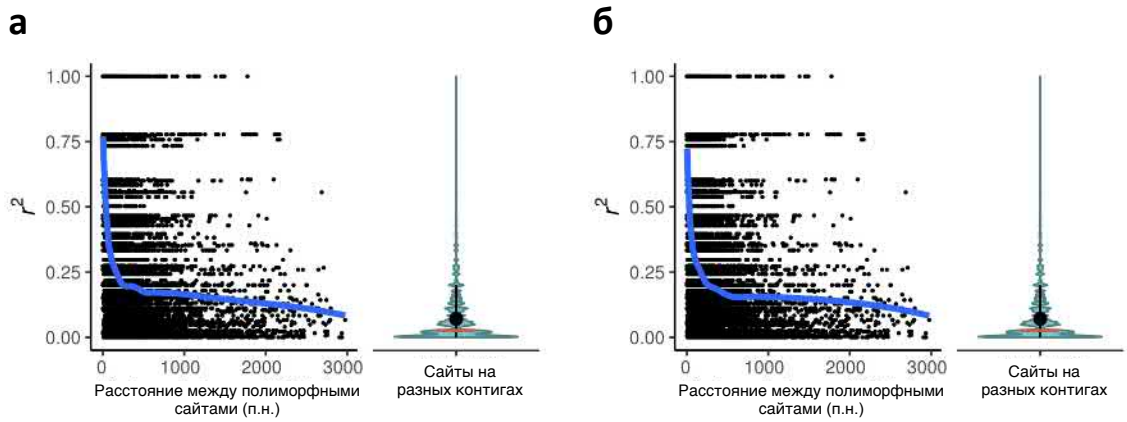

Рисунок 4.9. Падение LD с физическим расстоянием среди особей *A. vaga* не объясняется ошибочным выравниванием прочтений и ошибками фазирования. LD выражено как r^2. Распад r^2 с физическим расстоянием оценивали с использованием фазированных данных. Синие кривые показывают аппроксимацию падения r^2 с физическим расстоянием с помощью локальной регрессии (LOESS) полиномами второй степени (с коэффициентом сглаживания 0.4). Скрипичные графики показывают распределение значений r^2 для пар полиморфных сайтов на разных контигах. Концы усов соответствуют 10-й и 90-й перцентилям. Среднее значение и медиана показаны с помощью черной точки и красной горизонтальной линии соответственно. Значения r^2 рассчитывали для биаллельных сайтов, находящихся внутри сегментов референтного генома, в которых гаплотипы были восстановлены для всех индивидуумов большого кластера (L4-L11). (**а**) Значения r^2 рассчитывали с использованием сегментов из набора фазированных данных 1, принадлежащих к аллельным блокам, определенным внутри генома *A. vaga* (см. Материалы и методы). (**б**) Значения r^2 рассчитывали с использованием сегментов из набора фазированных данных 2, подвергнутого более строгой фильтрации (см. Материалы и методы).

Во-вторых, сигнал падения LD с расстоянием может быть связан с ошибками реконструкции гаплотипов, включая те, которые могут возникать из-за переключения матриц в ходе полимеразной цепной реакции [174]. Мы оценили достоверность фазирования, сравнив фазы, восстановленные для гаплотипов одного и того же индивидуума с использованием разных наборов данных секвенирования (см. раздел 4.2.5). Для этого мы сопоставили гаплотипы, реконструированные на основании прочтений Illumina HiSeq, с гаплотипами, восстановленными на основании прочтений PacBio (для индивидуума L1), и с гаплотипами, восстановленными на основании прочтений Illumina MiSeq (для индивидуумов L1, L2 и L11). Мы оценивали вероятность ошибочного фазирования как частоту случаев, в которых фазы одних и тех же полиморфных позиций у одного и того же индивидуума были определены по-

разному. С помощью этого анализа нам удалось показать, что частота ошибочного фазирования в наших данных достаточно низкая: оценки доли контигов с ошибками фазирования (среди контигов, несущих хотя бы один фазированный сегмент) составляют менее 2% и менее 1% для набора фазированных данных 1 и набора фазированных данных 2 соответственно (Таблица П4.2). В соответствии с ожиданием частота ошибок фазирования для набора фазированных данных 2, подвергнутого дополнительной фильтрации, ниже, чем для набора фазированных данных 1 (Таблица П4.2).

В наборе фазированных данных 1 вероятность ошибочного определения фазы для пары полиморфных сайтов растет с увеличением физического расстояния между ними, что при анализе гаплотипов из разных индивидуумов может быть ошибочно трактовано как распад LD. Однако важно отметить, что в наборе фазированных данных 2 подобная зависимость не наблюдается (Рисунок 4.10). Учитывая это, мы повторили анализ LD для индивидуумов L4-L11, используя гаплотипы из набора фазированных данных 2 (см. Материалы и методы), и показали, что распад LD наблюдается и в этом наборе фазированных сегментов (Рисунок 4.9б). Таким образом, вероятнее всего, распад LD среди проанализированных индивидуумов *A. vaga*, L4-L11, не может быть объяснен ошибками реконструкции гаплотипов. Более подробно анализ, проведенный для того, чтобы оценить частоту ошибок реконструкции гаплотипов, описан в разделе 4.2.5.

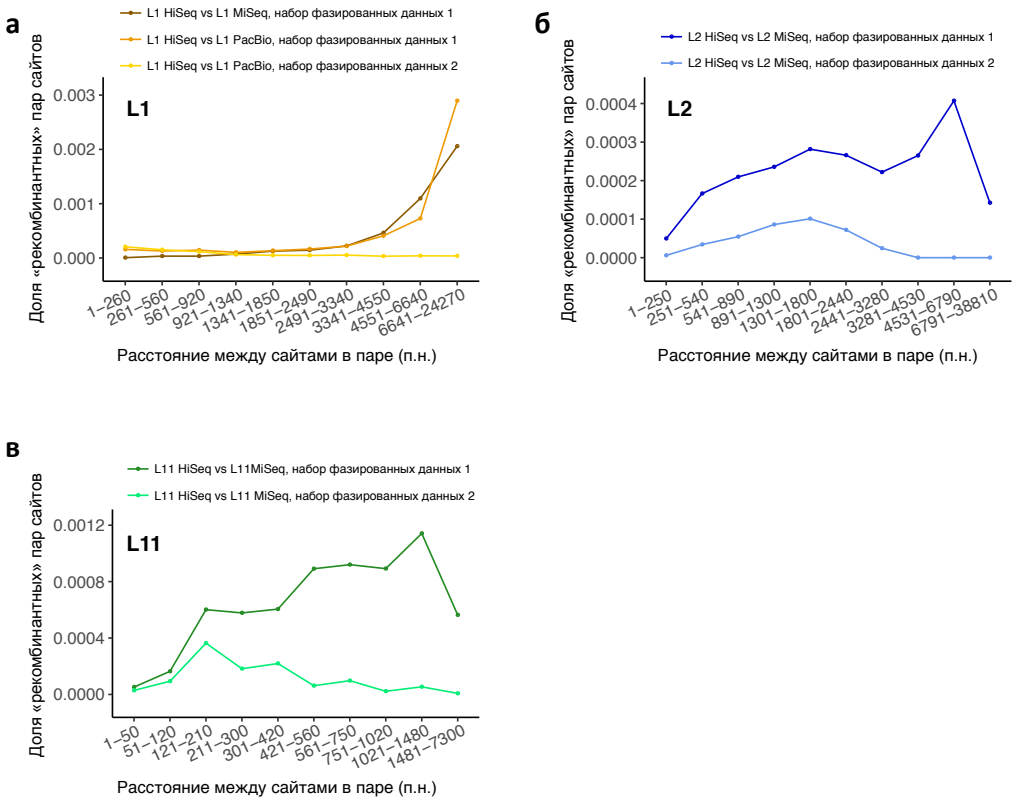

Рисунок 4.10. Частота ошибок фазирования, определенная с помощью сравнения фазированных блоков, реконструированных для одного индивидуума на основании разных наборов прочтений. На рисунке представлены данные для трех индивидуумов (L1 – **а**, L2 – **б**, L11 – **в**), для которых было доступно более одного набора прочтений. Для оценки частоты ошибок фазирования на разных расстояниях мы применили модифицированный вариант четырехгаметного теста к фазированным блокам, восстановленным для одного и того же индивидуума на основании прочтений Illumina HiSeq и Illumina MiSeq (для индивидуумов L1, L2 и L11) или на основании прочтений Illumina HiSeq и прочтений PacBio (для L1). «Рекомбинантные» пары сайтов соответствуют парам сайтов, для которых результаты фазирования, полученные для одного индивидуума на основании разных наборов прочтений, противоречат друг другу. Для каждого индивидуума бины по расстоянию между сайтами выбраны так, чтобы разные бины содержали близкие числа пар гетерозиготных сайтов для набора данных с наименьшим количеством одновременно фазированных пар гетерозиготных сайтов.

Чтобы дополнительно подтвердить, что распад LD не вызван исключительно ошибками фазирования, мы независимо оценили LD, используя нефазированные данные по генотипам L4-L1. Для этого мы применили два разных подхода: в то время как первый подход заключается в использовании вариабельных сайтов гомозиготных во всех индивидуумах L4-L11, второй

подход основан на использовании в качестве меры LD коэффициентов корреляции между генотипами. Распад LD был выявлен как с помощью первого, так и с помощью второго подхода (Рисунок 4.8б и Рисунок 4.8в). Поскольку эти анализы выполнены без использования данных фазирования, результаты этих анализов подтверждают, что сигнал распада LD с физическим расстоянием не определяется ошибками при реконструкции гаплотипов.

Кроме того, в качестве еще одного способа измерения LD, мы оценили корреляцию зиготности для пар сайтов внутри геномов отдельных особей, использовав подход, опирающийся на метод максимального правдоподобия [175,176]. Результаты этого анализа, проведенного для каждого индивидуума по отдельности, также подтверждают присутствие сигнала распада LD с расстоянием среди особей *A. vaga* (Рисунок 4.11). В совокупности результаты данных анализов исключают ошибки фазирования гаплотипов как единственную причину наблюдаемого распада LD.

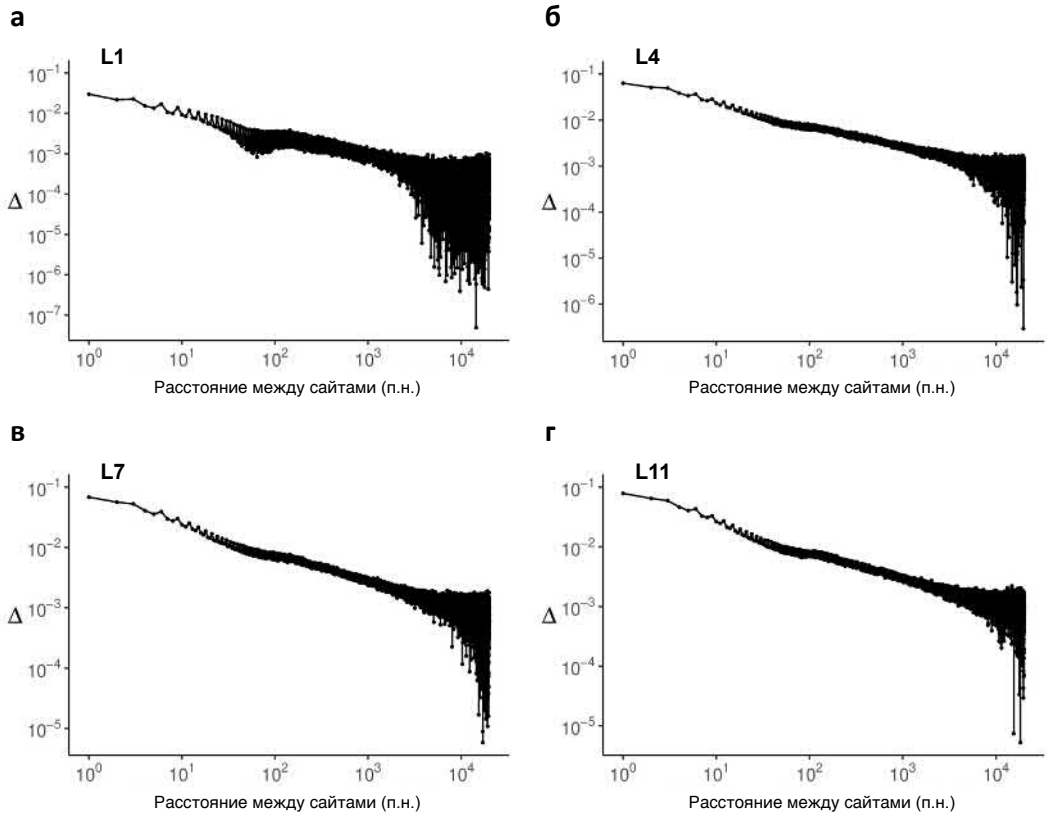

Рисунок 4.11. Падение корреляции зиготности (Δ) между парами сайтов с расстоянием в геномах индивидуумов *A. vaga*. Отдельные панели показывают зависимость Δ от расстояния в геномах четырех индивидуумов *A. vaga* (L1 – **а**, L4 – **б**, L7 – **в**, L11 – **г**). Оценки максимального правдоподобия Δ были получены с помощью программы mlRho [176] для каждой особи независимо.

4.2.4 Подписи реципрокной рекомбинации у *A. vaga*

Необходимо отметить, что даже в отсутствие ошибок реконструкции гаплотипов, сигнал распада LD может возникать без реципрокной рекомбинации в результате действия генной конверсии на аллельные участки генома. Генная конверсия является нереципрокным процессом, при котором происходит «копирование» фрагмента ДНК с одной хромосомы на гомологичную ей [74,75,204]. При этом генная конверсия может происходить как при кроссоверном, так и некроссоверном разрешении события рекомбинации [204]. Другими словами, несмотря на то, что подписи генной конверсии часто связаны с реципрокной рекомбинацией, они не обязательно указывают на нее. В работах, посвященных изучению неравновесия по сцеплению у человека, генную конверсию рассматривали как возможный фактор, повышающий скорость распада неравновесия по сцеплению между близко расположенными локусами [72,73]. Возможное действие генной конверсии обсуждали ранее и в контексте геномики бделлоидных коловраток: в первом опубликованном геноме *A. vaga* были выявлены вероятные подписи этого процесса [25].

Чтобы систематически проверить, может ли распределение полиморфных вариантов по гаплотипам в популяции *A. vaga* быть объяснено исключительно действием генной конверсии, мы применили модифицированный вариант четырехгаметного теста Хадсона [205]. В классическом четырехгаметном тесте [205] присутствие в популяции всех четырех возможных гаплотипов для пары биаллельных полиморфных локусов рассматривается как свидетельство в пользу реципрокной рекомбинации. Однако одного события мутации с последующим событием генной конверсии достаточно для того, чтобы объяснить существование всех четырех гаплотипов в популяции без предположения о реципрокном обмене ДНК между гомологичными участками генома (Рисунок 4.12а). Тем не менее действие генной конверсии может превратить гетерозиготный генотип в гомозиготный, но не наоборот. Таким образом, пара локусов, одновременно гетерозиготных в двух особях и представленных всеми четырьмя возможными гаплотипами в этих двух особях, не может появиться исключительно в результате генной конверсии (Рисунок 4.12б). В то же время такая пара локусов может очевидным образом появиться в результате реципрокной рекомбинации (или, гипотетически, в результате нереципрокной рекомбинации при трансформации). Мы используем эту логику в модифицированном варианте четырехгаметного теста Хадсона [205], сформулированного так, чтобы выявить сигнал рекомбинации, который не мог бы быть связан исключительно с действием генной конверсии: мы ищем пары полиморфных сайтов одновременно гетерозиготных в двух особях и представленных в них всеми четырьмя возможными гаплотипами. Такие пары дальше именуются как «рекомбинантные пары сайтов». Мы показали,

что среди пар сайтов одновременно гетерозиготных в двух особях доля сайтов, представленных в этих особях четырьмя возможными гаплотипами, растет с увеличением физического расстояния между сайтами (Рисунок 4.12в–г). Важно отметить, что несмотря на то, что рекуррентные мутации могут приводить к появлению «рекомбинантных пар сайтов», проходящих модифицированный вариант четырехгаметного теста, доля таких пар, возникающих из-за рекуррентных мутаций, не должна увеличиваться с расстоянием, разделяющим сайты. Таким образом, если увеличение доли «рекомбинантных пар сайтов» с расстоянием не объясняется ошибками реконструкции гаплотипов, это наблюдение несовместимо с действием генной конверсии как единственной причиной распада LD с расстоянием.

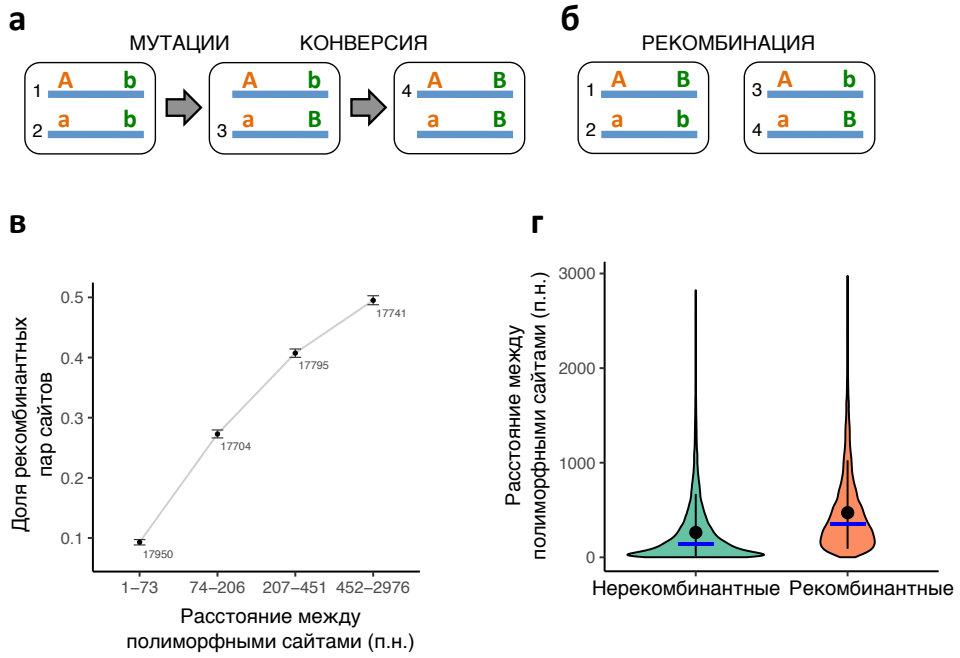

Рисунок 4.12. Результаты модифицированного четырехгаметного теста указывают на вероятное существование реципрокной рекомбинации у *A. vaga*. (**а**) Возникновение четырех гаплотипов для пары биаллельных сайтов без реципрокной рекомбинации и обмена генетическим материалом за счет мутаций и конверсии. Прямоугольники символизируют особей. (**б**) Схематическое изображение рекомбинантной пары сайтов, которая не могла возникнуть исключительно за счет генной конверсии: пара полиморфных сайтов представлена всеми четырьмя возможными гаплотипами в двух особях. Такие пары сайтов проходят модифицированный четырехгаметный тест. (**в**) Зависимость доли пар сайтов, проходящих модифицированный четырехгаметный тест, от физического расстояния между сайтами в паре. Для каждого попарного сравнения особей L4-L11 рассматривали только те пары полиморфных сайтов, которые одновременно гетерозиготны в этих двух особях. Пары сайтов,

удовлетворяющие условиям модифицированного четырехгаметного теста, были разделены на 4 бина по расстоянию между сайтами в паре с примерно равным количеством наблюдений в разных бинах. Черные точки обозначают доли рекомбинантных пар сайтов для разных бинов. Общее число проанализированных пар сайтов для каждого бина указано рядом с соответствующей точкой. Планки погрешности соответствуют 95% доверительным интервалам, рассчитанным на основе 1000 бутстреп-реплик для каждого бина. Разница в долях рекомбинантных пар сайтов значима для всех попарных сравнений между бинами (двустороннее P-значение $< 6 \times 10^{-4}$, пермутационный тест). (**г**) Скрипичные графики показывают распределение расстояний между сайтами в паре для нерекомбинантных (n = 48,657) и рекомбинантных пар сайтов (n = 22,533). Концы усов соответствуют 10-й и 90-й перцентилям. Среднее значение и медиана для каждой группы показаны с помощью черной точки и синей горизонтальной линии соответственно. Для нерекомбинантных пар сайтов среднее расстояние между сайтами в паре 261.7, медиана 146 п.н.; для рекомбинантных пар сайтов среднее расстояние между сайтами в паре 470.6, медиана 353 п.н. Разница в средних расстояниях между сайтами для нерекомбинантных и рекомбинантных пар значима (двустороннее P-значение $< 1 \times 10^{-4}$, пермутационный тест).

Для того, чтобы проверить, какой вклад в результаты этого анализа могут вносить ошибки реконструкции гаплотипов, мы применили модифицированный вариант четырехгаметного теста к двум парам индивидуумов, для которых мы располагали данными фазирования, полученными на основе независимых наборов прочтений (L2-L1 и L11-L1). Это позволило нам сравнить результаты теста, полученные для разных индивидуумов, с результатами, полученными при применении теста к наборам гаплотипов, восстановленных для одного и того же индивидуума на основании независимых данных секвенирования. В том случае, если поиск «рекомбинантных» пар сайтов проводится на основании наборов фазированных сегментов, восстановленных с использованием разных данных для одного и того же индивидуума, то выявленные в результате «рекомбинантные» пары сайтов, вероятнее всего, являются результатом ошибочного фазирования. В том же случае, если рассматриваются разные индивидуумы, помимо «рекомбинантных» пар сайтов, отражающих ошибки фазирования, мы также ожидаем найти пары сайтов, отражающие настоящие события рекомбинации (разумеется, в том случае, если события рекомбинации имели место). В соответствии с тем, что ожидается при рекомбинации, доля «рекомбинантных» пар сайтов, выявленных при сравнении гаплотипов из разных индивидуумов, на два порядка (или больше) выше соответствующей доли, определенной при сравнении наборов гаплотипов,

восстановленных для одного и того же индивидуума с использованием разных данных (порядка 10^{-3} или ниже). Более подробно данный анализ описан в разделе 4.2.5.

Таким образом, наши данные указывают на существование реципрокной рекомбинации у *A. vaga*. Однако описанные выше результаты совместимы не только с рекомбинацией, сопровождающейся обменом генетическим материалом между индивидуумами, но и с митотической рекомбинацией, происходящей в отсутствие переноса ДНК между особями [206].

4.2.5 Анализ частоты ошибок реконструкции гаплотипов

В данном разделе более подробно обсуждается анализ, проведенный для того, чтобы оценить частоту ошибок реконструкции гаплотипов, результаты которого обсуждаются в разделах 4.2.3 и 4.2.4.

Для того, чтобы оценить частоту ошибок реконструкции гаплотипов, мы сравнили результаты фазирования, полученные на основе разных наборов прочтений (HiSeq vs PacBio для L1 и HiSeq vs MiSeq для L1, L2 и L11), с использованием команды “compare”, включенной в программу WhatsHap [207]. Несоответствия между фазами гаплотипов, восстановленными для одного индивидуума на основании разных наборов данных секвенирования, мы рассматривали как ошибки фазирования.

При сравнении «сырых» фазированных блоков доля гаплоидных контигов, для которых были найдены несоответствия в фазах между гаплотипами, восстановленными на основании прочтений HiSeq и MiSeq, составила 0.023 для L1, 0.037 для L2 и 0.058 для L11 (Таблица П4.2). Однако после исключения блоков, содержащих пары «конфликтующих» сайтов, доля таких контигов с вероятными ошибками фазирования значительно снизилась и составила 0.0007 для L1, 0.0030 для L2 и 0.0154 для L11 (Таблица П4.2).

Этот результат показывает, что частота ошибок фазирования для блоков из набора фазированных данных 1, использовавшегося в большинстве анализов, действительно ниже, чем для «сырых» фазированных блоков. Дополнительная фильтрация на основании вероятностей ошибок фазирования, определенных HapCUT2 (такая фильтрация была применена к набору фазированных данных 2), привела к дальнейшему понижению доли контигов, для которых фазы гаплотипов были восстановлены по-разному (доля таких контигов снизилась до 0, 0.0006 и 0.0031 для L1, L2 и L11 соответственно).

Стоит отметить, что оценки доли контигов с вероятными ошибками при реконструкции гаплотипов скорее всего отражают как ошибки фазирования на основе прочтений HiSeq, так и ошибки фазирования на основе прочтений MiSeq (или PacBio, см. ниже).

Более высокий уровень несоответствия между гаплотипами, восстановленными с использованием HiSeq и MiSeq, для индивидуума L11 вероятно связан с более высокой

геномной дивергенций L11 (большой кластер) от L1 и L2 (маленький кластер), что должно приводить к повышению частоты ошибочного картирования прочтений. Анализы, представленные в этой главе, в первую очередь проведены на данных секвенирования для индивидуумов из большого кластера (L4-L11). Можно предположить, что более высокая частота ошибок фазирования (по сравнению с индивидуумами из маленького кластера, L1-L3), скорее всего, характерна для всех индивидуумов из большого кластера. Однако в то время как доля контигов с выявленными несоответствиями между фазами гаплотипов для фазированных блоков L11 из набора фазированных данных 1 составила 0.0154, соответствующая доля для фазированных блоков L11 из набора фазированных данных 2 значительно ниже (0.0031). Поскольку картина распада LD с расстоянием для набора фазированных данных 2 очень похожа на картину, наблюдаемую для набора фазированных данных 1 (Рисунок 4.9б и Рисунок 4.8а), распад LD у *A. vaga*, показанный в данной главе, по всей видимости, не может быть объяснен ошибками фазирования.

Интересно, что сравнение гаплотипов, восстановленных на основании прочтений HiSeq и PacBio, выявило более высокий уровень потенциальных ошибок фазирования, чем сравнение гаплотипов, восстановленных на основании прочтений HiSeq и MiSeq (Таблица П4.2). Так, сопоставление фазированных блоков из набора фазированных данных 2, восстановленных на основании прочтений HiSeq, с фазированными блоками, восстановленными на основании прочтений MiSeq, не выявило ни одного контига с несоответствиями (среди 4894 гаплоидных контигов, для которых удалось восстановить фазированные блоки на основании обоих наборов прочтений). В то же время для 37 из 5364 (0.0069) гаплоидных контигов из того же набора фазированных данных были выявлены несоответствия при сопоставлении с фазированными блоками, реконструированными с использованием прочтений PacBio. Остается непонятным, объясняется ли это картированием коротких, но аккуратных прочтений HiSeq или ошибками в длинных, но «шумных» прочтениях PacBio. Важно отметить, что мы не подвергали фильтрации гаплотипы, реконструированные с использованием прочтений PacBio, поскольку применение основного шага фильтрации (исключение блоков, включающих пары сайтов, представленных в одном индивидууме более чем двумя гаплотипами) представляется нецелесообразным в случае данных PacBio.

На следующем шаге для того, чтобы оценить до какой степени на наши результаты могли повлиять ошибки при реконструкции гаплотипов, мы использовали модифицированный вариант четырехгаметного теста. Здесь важно отметить, что модифицированный вариант четырехгаметного теста может быть применен не только к гаплотипам разных индивидуумов, но и к наборам гаплотипов, восстановленных для одного и того же индивидуума с использованием разных данных. Мы сопоставили доли рекомбинантных пар полиморфных

сайтов, найденнных при сравнении фазированных блоков из разных индивидуумов, с долями пар сайтов, которые были определены как «рекомбинантые» при сравнении фазированных блоков, восстановленных для одного и того же индивидуума на основании разных наборов прочтений.

Для этой цели мы применили модифицированный вариант четырехгаметного теста (см. раздел 4.2.4) к гаплотипам, восстановленным для L1, L2 и L11 с использованием разных данных. Для данного анализа мы сопоставляли наборы фазированных блоков попарно и искали рекомбинантные пары сайтов в двух данных наборах блоков. Специальным случаем являлось сопоставление наборов фазированных блоков, полученных для одного индивидуума на основе разных данных: в такой ситуации «рекомбинантые» пары сайтов соответствуют расхождениям в фазах гаплотипов, восстановленных для одного индивидуума с использованием разных прочтений.

Мы начали с применения модифицированного варианта четырехгаметного теста к разным наборам фазированных блоков, восстановленных для одного и того же индивидуума (L1, L2 или L11). Поскольку в такой ситуации рассматривается один индивидуум и мы не ожидаем событий рекомбинации, пары сайтов, проходящие четырехгаметный тест, вероятнее всего, отражают ошибки фазирования в одном или другом наборе блоков. Тем не менее оказалось, что в наборах фазированных блоков, прошедших только основную ступень фильтрации (исключение блоков с «конфликтующими» парами сайтов, такая фильтрация была применена к набору фазированных данных 1), наблюдалось повышение доли «рекомбинантных» пар сайтов с расстоянием (Рисунок 4.10). Увеличение частоты дискордантного фазирования с расстоянием не является удивительным, поскольку ожидается, что точность реконструкции гаплотипов будет падать с увеличением физического расстояния между сайтами. Однако доля пар сайтов с расхождениями в фазах, выявленными при сравнении разных наборов фазированных блоков, оставалась очень низкой для всех проанализированных расстояний (в большинстве случаев порядка 10^{-4}–10^{-3}, Рисунок 4.10). В наборах фазированных блоков, подвергнутых дальнейшей фильтрации на основании оцененных вероятностей ошибок фазирования (такая фильтрация была применена к набору фазированных данных 2), были выявлены еще более низкие доли пар сайтов с дискордантным фазированием на основании разных данных (Рисунок 4.10). Важно отметить, что для наборов фазированных блоков, подвергнутых такой двуступенчатой фильтрации, не было обнаружено очевидной тенденции увеличения с расстоянием доли пар сайтов с дискордантным фазированием («рекомбинантных» пар сайтов; Рисунок 4.10).

Картина распада LD, наблюдающаяся при сравнении гаплотипов из разных индивидуумов, входящих в подвергнутый двуступенчатой фильтрации набор фазированных

данных 2 (Рисунок 4.9б), схожа с той, что наблюдается для основного набора фазированных данных 1, подвергнутого только одному шагу фильтрации (Рисунок 4.8а). Поскольку согласно приведенным выше оценкам в наборе фазированных данных 2 частота ошибок фазирования очень низкая и для этого набора данных не наблюдается очевидного увеличения вероятности ошибочного фазирования с расстоянием, распад LD, описанный в данной главе для *A. vaga*, по всей видимости, не может быть объяснен ошибками при реконструкции гаплотипов.

Тем не менее мы провели еще один анализ для того, чтобы проверить, какой вклад в распад LD вносят артефакты фазирования. Для этого мы применили четырехгаметный тест к двум парам индивидуумов (L2-L1 и L11-L1), для которых были доступны фазированные данные, восстановленные с использованием разных наборов прочтений. Затем мы сравнили результаты данного теста, в котором сопоставляли гаплотипы разных индивидуумов, с результатами аналогичного теста, проведенного на разных наборах фазированных данных, полученных для одного и того же индивидуума.

При анализе гаплотипов из двух разных индивидуумов помимо «рекомбинантных» пар сайтов, являющихся результатом ошибок фазирования, мы ожидаем обнаружить настоящие рекомбинантные пары сайтов, возникшие в результате событий рекомбинации (если такие события имели место). Действительно, как ожидается при истинной рекомбинации, доля рекомбинантных пар сайтов, выявленных при сравнении разных индивидуумов, значительно (на два порядка или больше) выше, чем соответствующяя доля, определенная при сравнении гаплотипов одного и того же индивидуума, восстановленных с использованием разных данных (порядка 10^{-3} или меньше; Рисунки 4.13 и 4.14).

Например, при сравнении гаплотипов L2, реконструированных с использованием прочтений HiSeq и MiSeq, доля пар сайтов, определяемых как «рекомбинантные», среди всех пар гетерозиготных сайтов на расстоянии от 1 до 230 п.н. составила 5×10^{-5}, в то время как при сравнении гаплотипов L1 и L2, восстановленных на основании прочтений HiSeq, соответствующая доля (в таком же диапазоне расстояний, разделяющих сайты) составила 8.7×10^{-3}. Еще более сильный контраст при сравнении доли пар сайтов, определяемых как «рекомбинантные», для гаплотипов из разных индивидуумов и одного инидвидуума был выявлен при проведении анализа для пары индивидуумов из разных кластеров L11-L1: 0.39 пар гетерозиготных сайтов на расстоянии >1940 п.н. были определены как рекомбинантные при сопоставлении гаплотипов L11-L1 против 4.7×10^{-4} при сопоставлении гаплотипов L11, восстановленных из прочтений HiSeq и MiSeq.

Доля рекомбинантных пар сайтов при сопоставлении гаплотипов из разных индивидуумов увеличивается с расстоянием со схожей скоростью без очевидной зависимости от того, какие прочтения использовались для реконструкции гаплотипов, и от того, насколько

строго фильтровали фазированные блоки (Рисунки 4.13 и 4.14). Это указывает на то, что сигнал распада LD с расстоянием, показанный с помощью четырехгаметного теста, присутствует в данных фазирования, восстановленных с применением разных прочтений (HiSeq, MiSeq и PacBio), и не обусловлен недостаточно жесткой фильтрацией фазированных блоков.

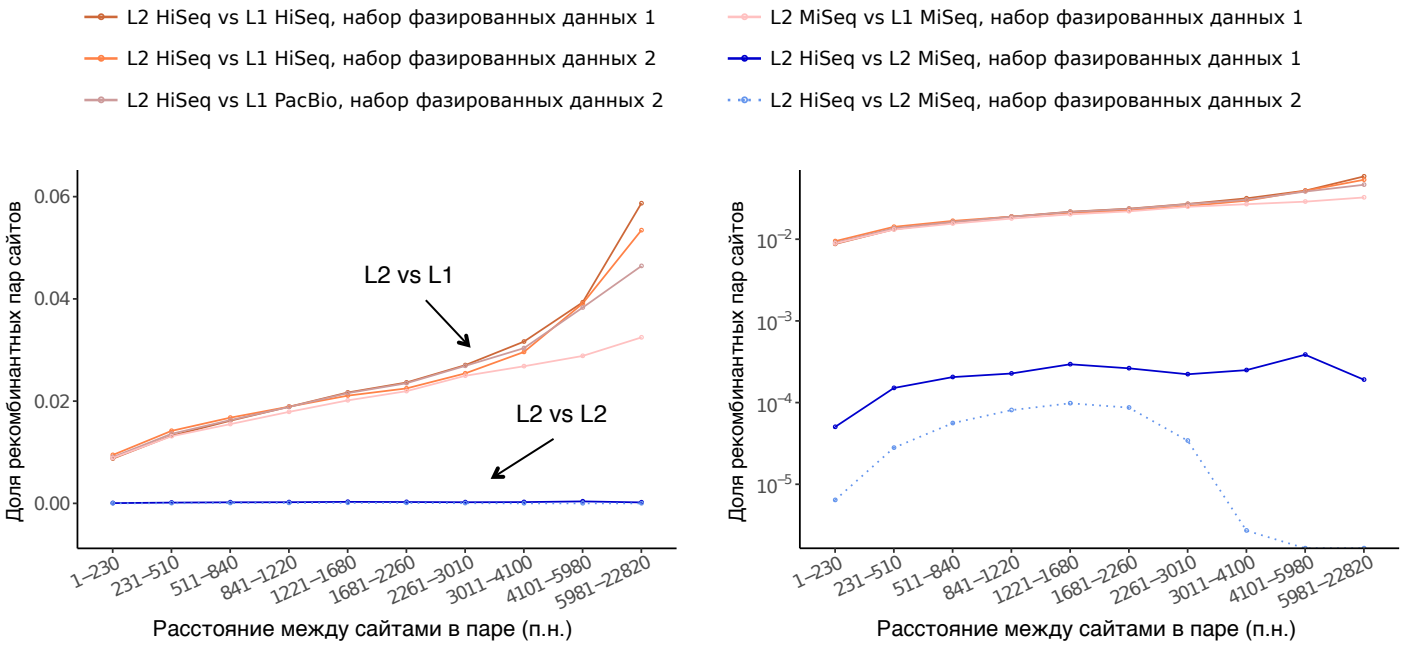

Рисунок 4.13. Результаты модифицированного четырехгаметного теста, примененного к фазированным блокам из разных индивидуумов (L2 и L1) и к фазированным блокам, восстановленным для одного и того же индивидуума (L2) на основании разных наборов прочтений. Рекомбинантные пары сайтов, выявленные при сравнении данных фазирования, полученных для одного и того же индивидуума на основании разных наборов прочтений, соответствуют ошибкам фазирования. В то же время ожидается, что рекомбинантные пары сайтов, найденные при сравнении разных индивидуумов, могли возникнуть не только из-за ошибок фазирования, но и в результате истинных событий рекомбинации. На правой панели представлены те же данные, что и на левой панели, но ось Y имеет логарифмическую шкалу.

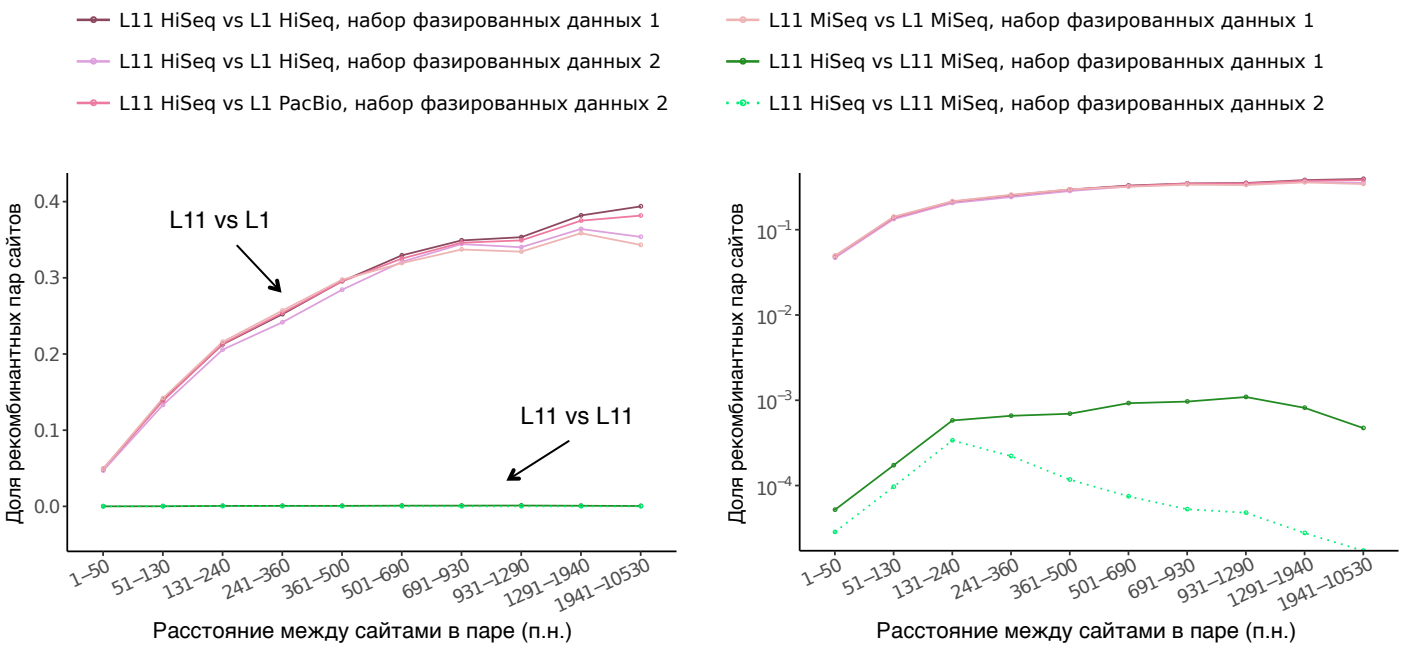

Рисунок 4.14. Результаты модифицированного четырехгаметного теста, примененного к фазированным блокам из разных индивидуумов (L11 и L1) и к фазированным блокам, восстановленным для одного и того же индивидуума (L11) на основании разных наборов прочтений. Рекомбинантные пары сайтов, выявленные при сравнении данных фазирования, полученных для одного и того же индивидуума на основании разных наборов прочтений, соответствуют ошибкам фазирования. В то же время ожидается, что рекомбинантные пары сайтов, найденные при сравнении разных индивидуумов, могли возникнуть не только из-за ошибок фазирования, но и в результате истинных событий рекомбинации. На правой панели представлены те же данные, что и на левой панели, но ось Y имеет логарифмическую шкалу.

Представляется вероятным, что некоторые области генома (например, участки богатые повторами) могут быть более подвержены ошибкам при фазировании, что может приводить к тому, что один и тот же сценарий ошибочного фазирования будет воспроизводиться в фазированных блоках, восстановленных на основании разных прочтений. В таком случае сравнение наборов фазированных блоков, реконструированных для одного индивидуума с использованием разных данных, не выявит расхождений в фазировании, поскольку неправильно собранный гаплотип будет присутствовать во всех наборах. В связи с этим вероятно, что наш подход не выявляет часть ошибок, возникающих при реконструкции гаплотипов. Тем не менее одинаковые ошибки должны быть более характерны для фазированных блоков, восстановленных на основании прочтений Illumina (HiSeq и MiSeq), чем для фазированных блоков, восстановленных на основании прочтений Illumina и PacBio.

Действительно, в соответствии с этим ожиданием сравнение гаплотипов L1, реконструированных с использованием прочтений HiSeq, с гаплотипами L1, реконструированными с использованием прочтений PacBio, выявило более высокую частоту расхождений при фазировании, чем сравнение гаплотипов L1, реконструированных с использованием прочтений HiSeq и MiSeq (Таблица П4.2). Однако несмотря на то, что оценки частоты ошибочного фазирования, полученные при сравнении фазированных блоков, реконструированных с использованием прочтений HiSeq и PacBio, выше, чем оценки, полученные с использованием прочтений HiSeq и MiSeq, оценки доли гаплоидных контигов с ошибками фазирования (среди контигов с фазированными блоками) для L1 на основании прочтений PacBio остаются ниже 1% (Таблица П4.2).

Более того, результаты четырехгаметного теста (примененного к гаплотипам разных индивидуумов) для фазированных блоков, восстановленных на основании прочтений PacBio, практически неотличимы от результатов, полученных на основании прочтений HiSeq или MiSeq (Рисунки 4.13 и 4.14). Доля пар сайтов, определенных как рекомбинантные при сравнении разных индивидуумов, выше вплоть до четырех порядков, чем соответствующая доля, выявленная при сопоставлении разных наборов фазированных гаплотипов, восстановленных для одного и того же индивидуума (Рисунки 4.13 и 4.14). В совокупности эти результаты указывают на то, что распад LD у *A. vaga* не может быть объяснен исключительно за счет ошибок в процессе реконструкции гаплотипов.

4.2.6 Подписи обмена генетическим материалом между индивидуумами

Чтобы исследовать возможность того, что выявленный сигнал рекомбинации может быть не связан с обменом ДНК между индивидуумами, мы проанализировали изменчивость среди индивидуумов большого кластера, L4-L11, на уровне отдельных биаллельных сайтов. При облигатном бесполом размножении два аллеля в одном локусе накапливают мутации независимо друг от друга. В популяции конечного размера это создает ибыток гетерозигот по сравнению с ожидаемым числом гетерозигот при равновесии Харди-Вайнберга [69,79], что соответствует отрицательным значениям коэффициента инбридинга F_{IS}. В то время как ожидаемое значение F_{IS} при равновесии Харди-Вайнберга равно 0, ожидаемое значение F_{IS} при облигатном бесполом размножении −1. Мы проанализировали распределение значений F_{IS} в сайтах биаллельных среди L4-L11 и обнаружили, что это распределение сосредоточено вокруг 0 (медиана = 0.0, среднее значение = −0.03; Рисунки 4.15a и 4.16; Таблица 4.10), что, вероятно, указывает на то, что рассматриваемая популяция *A. vaga* находится близко к равновесию Харди-Вайнберга. Схожие значения F_{IS} характерны и для участков генома, плоидность которых подтверждается присутствием коллинеарных аллельных генов (среднее значение F_{IS} = −0.03 и

−0.04 для подмножества аллельных блоков и для подмножества аллельных генов соответственно; медианное значение F_{IS} = 0.0 и в том, и в другом случае; Таблица 4.10).

Набор данных	Общее число биаллельных сайтов	Общее число биаллельных сайтов с частотой минорного варианта ≥4	F_{IS}				
			Среднее	Медиана	Стандартное отклонение	Q1	Q3
Весь геном, большой кластер, L4-L11	1,106,582	440,564	−0.03	0	0.39	−0.33	0.25
Аллельные блоки, большой кластер, L4-L11	285,043	112,236	−0.03	0	0.38	−0.33	0.25
Аллельные гены, большой кластер, L4-L11	194,384	77,543	−0.04	0	0.38	−0.33	0.25

Таблица 4.10. Статистика по значениям коэффициента инбридинга F_{IS} для особей L4-L11. Значения F_{IS} определяли для биаллельных полиморфных сайтов из профильтрованного набора полиморфизмов II с частотой минорного варианта среди особей большого кластера L4-L11 ≥4.

Для того, чтобы напрямую оценить, какие значения F_{IS} ожидаются при разной частоте бесполого размножения в выборке такого же размера, как анализируемая (8 особей), мы симулировали популяции [182], варьируя частоту бесполого размножения от 0 (строго половое размножение) до 1 (строго бесполое разможение). Затем мы получили случайные подвыборки «особей» из симулированных популяций, включающие, как и анализируемая популяция, 8 индивидуумов, и оценили распределение значений F_{IS} в таких подвыборках (см. Материалы и методы).

В соответствии с результатами ранее опубликованных работ [69] мы обнаружили, что даже относительно редких событий полового размножения достаточно для того, чтобы наблюдаемые значения F_{IS} были очень близки к ожидаемым в случае строго полового размножения [69] (Рисунок 4.15б). Важно отметить, что значения F_{IS}, наблюдаемые в L4-L11, значимо выше ожидаемых в случае облигатного бесполого размножения (частота бесполого размножения = 1.0, P-значение < 0.01) или очень редкого полового размножения (частота бесполого размножения = 0.999, P-значение = 0.03); однако они совместимы с любым из симулированных сценариев с частотой полового размножения ≥1% (Рисунок 4.15б).

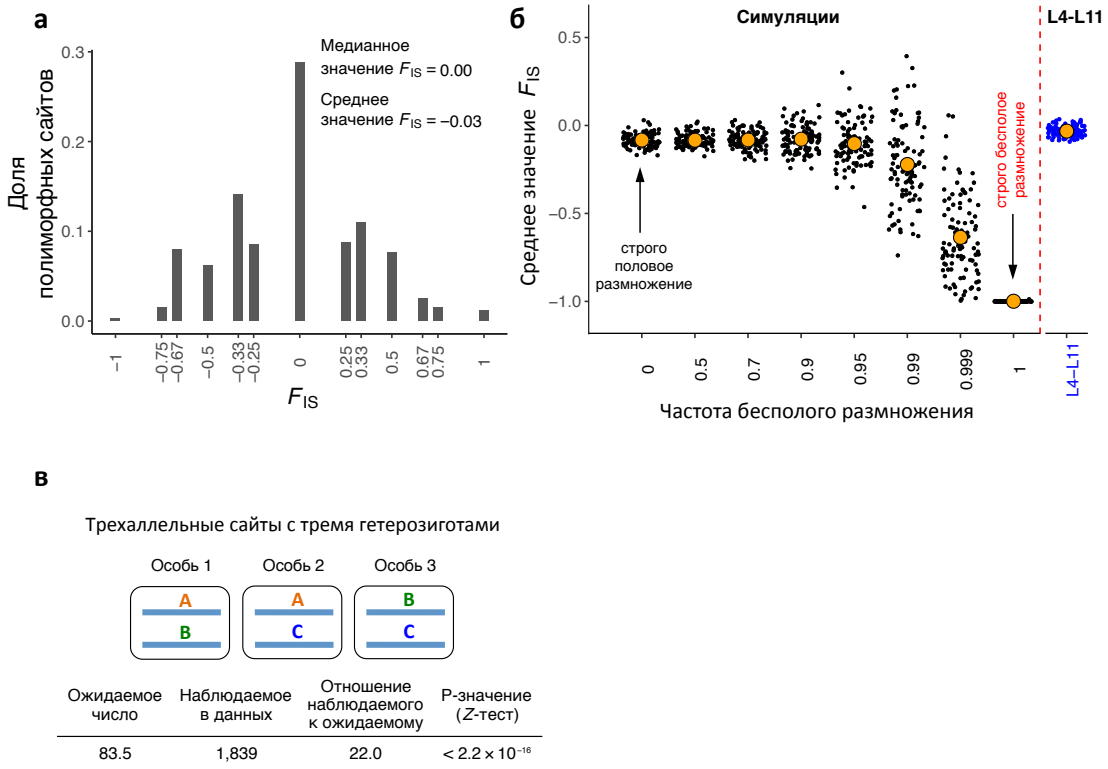

Рисунок 4.15. Анализ биаллельных и трехаллельных сайтов указывает на вероятное существование обмена генетическим материалом у *A. vaga*. (**а**) Распределение значений коэффициента инбридинга (F_{IS}) для сайтов биаллельных у особей L4-L11 (частота минорного варианта ≥4, n = 440,564). В равновесии Харди-Вайнберга F_{IS} = 0; отрицательные и положительные значения F_{IS} соответствуют избытку гетерозигот и гомозигот соответственно. (**б**) Значения F_{IS}, наблюдаемые для сайтов биаллельных у особей L4-L11, в сравнении со значениями, ожидаемыми при разной частоте бесполого размножения. График показывает средние значения F_{IS} для случайных выборок сайтов (размер каждой выборки 200 сайтов) полиморфных среди L4-L11 (синие точки) или сайтов полиморфных в популяциях, симулированных в SLiM [182] с разными частотами бесполого размножения (черные точки). Параметры симуляций были подобраны так, чтобы популяционная скорость возникновения мутаций $4N_e\mu$ в симуляциях была близка к значению этого параметра, оцененного из данных (Таблица 4.8): N_e = 2500, $\mu = 10^{-6}$. Для каждой рассмотренной частоты бесполого размножения (0, 0.5, 0.7, 0.9, 0.95, 0.99, 0.999 и 1) было проведено 100 реплик симуляции. Через 200,000 поколений из каждой реплики каждой симуляции случайно выбирали 8 особей (что соответствует числу анализируемых особей L4-L11), оставляли только полиморфные среди этих особей сайты с частотой минорного варианта ≥4 и затем случайно отбирали 200 полиморфных сайтов. В результате этой процедуры для каждой симуляции было получено 100 выборок полиморфных сайтов. Отдельные черные точки показывают среднее значение F_{IS} в каждой такой выборке. Для сравнения с данными мы аналогичным образом определили

средние значения F_{IS} для 100 случайных выборок сайтов (n = 200) биаллельных среди L4-L11 (частота минорного варианта ≥4). Средние значения F_{IS} для каждой такой выборки показаны синими точками. Оранжевые круги показывают среднее средних значений F_{IS} среди 100 выборок для каждой симуляции или среди 100 выборок, полученных из данных. (**в**) Схематическое изображение трехаллельного сайта с тремя аллелями A, B и C, представленного всеми тремя возможными гетерозиготными генотипами среди анализируемых особей *A. vaga*. В нижней части рисунка показано ожидаемое в данных число таких сайтов в случае отсутствия обмена генетическим материалом и наблюдаемое число среди L4-L11.

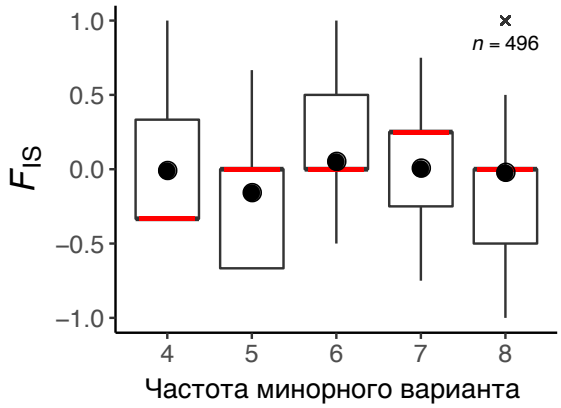

Рисунок 4.16. Диаграмма размаха, показывающая распределение значений коэффициента инбридинга (F_{IS}) для полиморфных сайтов с разной частотой минорного варианта. При проведении данного анализа рассматривались только сайты биаллельные в особях L4-L11 с частотой минорного варианта в этих особях ≥4 (n = 440,564). Количество сайтов с частотой минорного варианта 4, 5, 6, 7 и 8 составило 113,866, 100,501, 92,682, 89,419 и 44,096 соответственно. Нижняя и верхняя границы «ящика» для каждой группы соответствует первой и третьей квартили распределения, в то время как усы доходят до минимального/максимального значения в пределах 1.5 × IQR от соответствующей границы «ящика», где IQR обозначает межквартильный диапазон. Выбросы отмечены крестиком (выбросы были найдены только для частоты минорного варианта равной 8; все полиморфные варианты, соответствующие выбросам, имели одинаковое значение F_{IS} равное 1). Общее число полиморфных вариантов, являющихся выбросами по F_{IS}, обозначено под крестиком. Среднее значение и медиана для каждой группы показаны с помощью черной точки и красной горизонтальной линии соответственно.

Независимое свидетельство в пользу обмена генетическим материалом между индивидуумами *A. vaga* было получено в ходе анализа трехаллельных [208] сайтов. Трехаллельные сайты могут возникать у видов, размножающихся как бесполым, так и половым

путем, за счет множественных мутаций, затрагивающих один и тот же сайт. В гипотетическом полиморфном сайте с тремя аллелями A, B и C в принципе возможны три различных гетерозиготных генотипа: A/B, A/C и B/C (Рисунок 4.15в). Генетические обмены между индивидуумами будут приводить к тому, что в триаллельных сайтах все три возможные гетерозиготные генотипа будут часто сосуществовать. И, напротив, в бесполой популяции все три гетерозиготных генотипа могут возникнуть только за счет рекуррентных или обратных мутаций, которые происходят относительно редко.

Мы оценили, сколько трехаллельных сайтов, несущих три гетерозиготных генотипа, мы бы ожидали увидеть среди индивидуумов большого кластера, L4-L11, если бы такие сайты возникали исключительно за счет рекуррентных и обратных мутаций. Для этого, опираясь на долю трехаллельных сайтов среди полиморфных сайтов, мы сначала оценили вероятность, с которой мутации повторно затрагивали одни и те же сайты. Мы использовали следующую логику: трехаллельный сайт может возникнуть только в результате по меньшей мере одной мутации в уже существующем биаллельном сайте. Вероятность такой мутации в истории выборки генотипов может быть оценена как: $P3 = N3/(N2 + N3)$, где N2 и N3 обозначают количество сайтов с двумя и тремя аллелями соответственно. В отсутствие обмена генетическим материалом, для возникновения трехаллельного сайта, представленного тремя возможными гетерозиготами, необходима еще одна мутация. Таким образом, ожидаемое число трехаллельных сайтов с тремя гетерозиготами можно приблизительно оценить как $P3 \times N3$. Согласно полученной таким образом оценке в случае отсутствия обмена генетическим материалом ожидаемое число трехаллельных сайтов, одновременно представленных в L4-L11 тремя гетерозиготными генотипами, составляет 83.5. Однако в данных по полиморфизму L4-L11 мы идентифицировали 1839 таких сайтов (P-значение $< 2.2 \times 10^{-16}$, одновыборочный Z-тест; Рисунки 4.15в и 4.17; Таблица 4.11), что соответствует 22-кратному превышению над ожиданием. Схожее обогащение трехаллельными сайтами, представленными тремя гетерозиготами, относительно ожидания было обнаружено и для участков с достоверной плоидностью (Таблица 4.11).

Набор данных	Число биаллельных сайтов	Число трехаллельных сайтов	Доля трехаллельных сайтов среди би- и трехаллельных сайтов	Ожидаемое число трехаллельных сайтов, представленных тремя гетерозиготными генотипами	Число трехаллельных сайтов, представленных тремя гетерозиготными генотипами, в данных (L4-L11)	Отношение наблюдаемого к ожидаемому	P-значение
Весь геном	1,126,303	9,738	0.0086	**83.47**	**1,839**	**22.03**	**$<2.2\times10^{-16}$**
Аллельные блоки	290,498	2,043	0.0070	**14.27**	**399**	**27.97**	**$<2.2\times10^{-16}$**
Аллельные гены	198,151	1,448	0.0073	**10.50**	**295**	**28.08**	**$<2.2\times10^{-16}$**

Таблица 4.11. Ожидаемое и наблюдаемое в данных число трехаллельных сайтов, одновременно представленных всеми тремя возможными гетерозиготными генотипами. Анализ проводили на основании сайтов полиморфных среди L4-L11 (набор однонуклеотидных полиморфизмов II). Значимость разницы между наблюдаемыми и ожидаемыми значениями оценивали с помощью одновыборочного *Z*-теста.

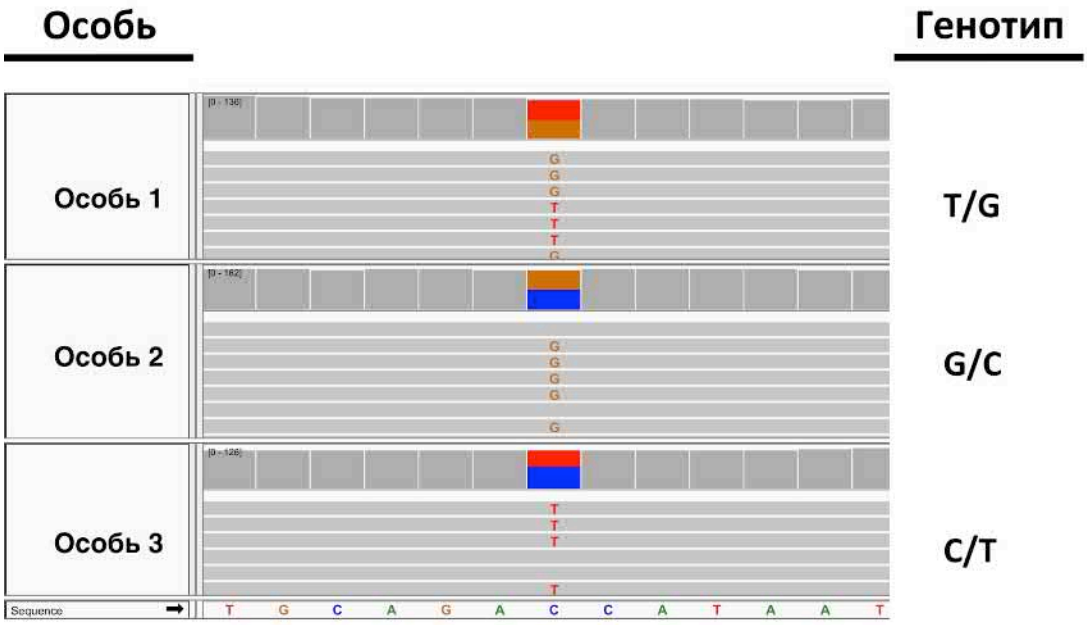

Рисунок 4.17. Пример трехаллельного сайта с аллелями C, T и G, одновременно представленного всеми тремя гетерозиготными генотипами среди анализируемых особей *A. vaga*. Изображение получено с помощью геномного браузера IGV. Показанный участок расположен на контиге contig353 в диплоидной сборке генома *A. vaga* L1. Отдельные прочтения показаны с помощью серых полосок.

Одно из возможных объяснений присутствия в данных значительного количества трехаллельных сайтов, представленных тремя гетерозиготами, – экспериментальная контаминация между клональными линиями. Для того, чтобы проверить, присутствуют ли в данных указания на контаминацию, мы отдельно рассмотрели только те сайты с тремя гетерозиготными генотипами среди L4-L11, в которых наиболее редкий из трех гетерозиготный генотип присутствовал только в одном индивидууме (n = 607). Мы проанализировали, как такие «приватные» гетерозиготные генотипы распределены среди индивидуумов. В том случае, если L4-L11 на самом деле размножались бы клонально, но некоторые линии были бы контаминированы, мы бы ожидали, что большая часть «приватных» гетерозиготных генотипов была бы найдена в этих линиях. Однако такие генотипы оказались распределены практически

равномерно между разными линиями таким образом, что числа «приватных» гетерозиготных генотипов, найденных в разных линиях, близки друг к другу. Среднее число сайтов с гетерозиготными генотипами «приватными» для данного индивидуума равно 75.9, минимальное и максимальное число таких сайтов среди индивидуумов L4-L11 – 64 и 88 соответственно (Таблица 4.12). Это наблюдение можно рассматривать в качестве аргумента против контаминации как источника сайтов, несущих все три гетерозиготных генотипа, и в качестве аргумента в пользу обмена генетическим материалом в популяции *A. vaga* как основного механизма их возникновения. Кроме того, свидетельства против контаминации между линиями *A. vaga* были получены в ходе анализа митохондриальной изменчивости, проведенного на основании последовательностей полных митогеномов (см. раздел 4.2.7).

Особь	Число сайтов с «уникальным» гетерозиготным генотипом, найденным только в данной особи, среди трехаллельных сайтов, представленных всеми тремя гетерозиготными генотипами (L4-L11)
L4	77
L5	77
L6	88
L7	76
L8	79
L9	74
L10	72
L11	64

Таблица 4.12. Статистика по числу сайтов с «уникальным» гетерозиготным генотипом, найденным только в данной особи, среди трехаллельных сайтов, представленных всеми тремя гетерозиготными генотипами. Рассматривались только те трехаллельные сайты, представленные всеми тремя гетерозиготными генотипами среди L4-L11, для которых был найден только один «уникальный» гетерозиготный генотип (n = 607).

4.2.7 Поиск потенциальных признаков контаминации между культурами *A. vaga* в данных по митохондриальной изменчивости

Мы использовали данные по митохондриальной изменчивости для того, чтобы провести поиск признаков, которые могли бы указывать на потенциальную контаминацию между секвенированными культурами *A. vaga*. Для этого мы определили однонуклеотидные митохондриальные полиморфизмы среди особей L1-L11. В качестве референтной последовательности для этого анализа мы использовали митохондриальный контиг L4 (эта особь была отнесена к большому кластеру на основании ядерной изменчивости; Рисунки 4.2в и

4.5), поскольку этот контиг имеет наибольшую длину среди митохондриальных контигов L1-L11. Дополнительно мы повторили данный анализ, используя в качестве референтной последовательности митохондриальный контиг особи, отнесенной к другому генетическому кластеру на основании ядерных полиморфизмов (L3, маленький кластер). Для каждой особи мы картировали прочтения Illumina HiSeq на референтный митохондриальный контиг (L4 или L3) с помощью Bowtie 2 (версия 2.3.2) с параметрами "--no-mixed --no-discordant" и максимальной длиной вставки 800 п.н. (для каждой пары прочтений выводилось только одно лучшее выравнивание). Фильтрация картирований прочтений к митохондриальным контигам была проведена аналогично соответствующей процедуре, использованной при анализе ядерных полиморфизмов: мы не рассматривали прочтения, для которых было найдено более одного выравнивания (с меткой XS); среди прочтений с уникальными картированиями мы выбирали только те, которые принадлежали к корректно картированным парам и имели MAPQ ≥20. Как и в случае ядерных полиморфизмов, определение однонуклеотидных митохондриальных полиморфизмов проводилось с использованием утилиты mpileup из пакета SAMtools (v.1.4.1) [168] с параметрами "-aa -u -t DP,AD,ADF,ADR", вывод которой обрабатывался дальше с применением команды "bcftools call" с опцией "-m". Для поиска гетероплазмичных сайтов и сайтов, которые потенциально могли бы свидетельствовать о контаминации, определение генотипов проводили в базовом «диплоидном» режиме. В случае присутствия единственного митохондриального гаплотипа в клональной культуре ожидается, что все генотипы для этой культуры будут определены как «гомозиготные». В то же время значительная контаминация другим митохондриальным гаплотипом должна приводить к появлению генотипов, определенных как «гетерозиготные». Кроме того, «гетерозиготные» митохондриальные сайты могут также быть результатом митохондриальной гетероплазмии (присутствия нескольких вариантов митохондриального генома в клетке) [209,210] или различий, накопленных между индивидуумами внутри клональной культуры.

Сначала мы провели определение митохондриальных генотипов совместно для всех 11 особей L1-L11. Однако в то время как в случае определения ядерных полиморфизмов низкий уровень ядерной дивергенции между разными особями (среднее значение дивергенции между особями большого и маленького кластеров по ядерному геному 1.22%) позволил картировать прочтения для всех особей на контиги L1 и одновременно определить полиморфизмы для всех одиннадцати особей, оказалось, что митохондриальный гаплотип L1 слишком сильно отличается от гаплотипов L2-L11 для одновременного определения митохондриальных полиморфизмов среди L1-L11 с помощью стандартных подходов. Прочтения Illumina для L1 картировались на митохондриальные контиги L4 (L3) очень неравномерно, что может объясняться разницей в степени консервативности между участками митохондриального

генома: участки с высоким покрытием прочтениями L1 чередовались с участками без покрытия или с очень низким покрытием. Как следствие, из 14,033 проанализированных сайтов митохондриального контига L4 у 7484 сайтов покрытие прочтениями L1 было менее 5×, а в 5526 сайтах генотип L1 не был определен. Присутствие в данных большого числа сайтов с неопределенными генотипами осложняло интерпретацию результатов, полученных с их использованием. В связи с этим мы решили сосредоточиться на анализе митохондриальных данных для L2-L11, на первом этапе исключив L1 из рассмотрения. С этой целью мы отдельно определили митохондриальные генотипы для особей L2-L11, не используя данные по L1 при проведении этой процедуры. Полученные генотипы были профильтрованы следующим образом: были исключены (1) полиморфные сайты на расстоянии менее или равном десяти нуклеотидов от инсерции или делеции, (2) инсерции и делеции, (3) сайты, в которых генотипы для части особей не были определены, и сайты с покрытием DP <50 в одной или более особях L2-L11, (4) сайты с ненадежно определенными генотипами (QUAL <15). В результате данной процедуры для одновременного анализа в особях L2-L11 осталось доступно 13,768 сайтов (или 13,644 в случае использования контига L3 в качестве референтной последовательности). Таким образом, значительная часть митохондриального генома могла быть одновременно проанализирована в особях L2-L11 как относительно референтной последовательности L4, так и относительно референтной последовательности L3.

Затем для каждой особи L2-L11 мы определили число сайтов с «гетерозиготными» и «гомозиготными» генотипами согласно результатам SAMtools/BCFtools. Практически все доступные для анализа сайты (13,765 из 13,768) были определены как «гомозиготные» во всех десяти особях. Это наблюдение согласуется с тем, что ожидается в случае, если в каждой клональной культуре присутствует только один митохондриальный гаплотип. Только три сайта были определены как «гетерозиготные», т.е. в этих трех сайтах прочтения поддерживали присутствие двух разных нуклеотидных вариантов внутри одной культуры. Три «гетерозиготных» сайта были выявлены в трех разных особях (L5, L6 и L9): в каждой из этих особей было найдено по одному такому сайту.

Чтобы получить грубую оценку того, сколько сайтов, определенных как «гетерозиготные», мы бы ожидали увидеть в случае контаминации между культурами L2-L11, мы рассчитали попарные расстояния между митохондриальными гаплотипами L2-L11, используя только те сайты, которые были определены как «гомозиготные» во всех особях. Особи L2-L11 отличались друг от друга в небольшой доле митохондриальных сайтов: только 214 из 13,764 рассмотренных сайтов (3 «гетерозиготных» и один мультиаллельный сайт были исключены) являлись вариабельными. Тем не менее для каждой особи L2-L11 существовало подмножество митохондриальных сайтов, в которых данная особь отличалась от всех

остальных особей среди L2-L11. Минимальное число таких сайтов составило 5 (для особи L2), а максимальное 57 (для особи L11) (Таблица 4.13). Минимальное расстояние между двумя особями, рассчитанное как число митохондриальных сайтов, в которых эти две особи отличались друг от друга (т.е. одна имела генотип ‘0/0’, а другая ‘1/1’), составило 11 (для пары L2-L10). Большая часть пар отличались по ≥20 сайтам (Таблица 4.14). Практически такие же результаты были получены при использовании митохондриального контига L3 вместо контига L4 в качестве референтной последовательности (Таблицы 4.15 и 4.16). Следовательно, в случае контаминации между L2-L11 мы бы ожидали получить образец, имеющий по крайней мере 11 «гетерозиготных» сайтов (в среднем 57.9 таких сайтов; Таблица 4.14). Однако среди трех образцов с выявленной митохондриальной гетерогенностью (L5, L6 и L9, см. выше) в каждом образце был найден только один гетерогенный сайт. С учетом этого представляется более вероятным, что эти три случая объясняются не контаминацией, а митохондриальной гетероплазмией [209,210], мутациями, накопленными особями внутри клональной культуры, или техническими артефактами (такими как ошибки в определении генотипов, вызванными, например, присутствием псевдогенов митохондриального происхождения в ядерном геноме [211]).

Особь	Общее число проанализированных митохондриальных сайтов	Число сайтов с вариантом, найденным только в данной особи
L2	13,764	5
L3	13,764	10
L4	13,764	8
L5	13,764	17
L6	13,764	9
L7	13,764	10
L8	13,764	12
L9	13,764	18
L10	13,764	6
L11	13,764	57

Таблица 4.13. Статистика по числу сайтов с «уникальным» митохондриальным вариантом, найденным только в данной особи (в качестве референтной последовательности использовался митохондриальный контиг L4). Для каждой особи L2-L11 в таблице представлено число митохондриальных сайтов, в которых данная особь отличается от всех остальных особей L2-L11. Представленные числа были рассчитаны для митохондриальных сайтов, в которых генотипы были одновременно определены для всех особей L2-L11 (n = 13,764) с использованием митохондриального контига L4 в качестве референтной последовательности.

	L2	L3	L4	L5	L6	L7	L8	L9	L10
L3	19								
L4	19	26							
L5	48	53	53						
L6	69	76	70	57					
L7	18	25	21	52	73				
L8	70	77	71	58	23	74			
L9	50	57	55	40	59	54	60		
L10	11	20	20	49	70	19	71	51	
L11	94	99	99	90	99	100	104	88	95

Таблица 4.14. Попарные расстояния между митохондриальными гаплотипами особей *A. vaga* L2-L11, определенные с использованием митохондриального контига L4 в качестве референтной последовательности. Для каждой пары особей L2-L11 в таблице представлено абсолютное число нуклеотидных отличий между митохондриальными гаплотипами этих двух особей. Число нуклеотидных отличий для каждой пары особей рассчитывали на основании набора митохондриальных однонуклеотидных полиморфизмов, определенных для особей L2-L11 с использованием митохондриального контига L4 в качестве референтной последовательности (длина = 14,052 п.н.). Представленные числа были рассчитаны для митохондриальных сайтов, в которых генотипы были одновременно определены для всех особей L2-L11 (n = 13,764).

Особь	Общее число проанализированных митохондриальных сайтов	Число сайтов с вариантом, найденным только в данной особи
L2	13,640	5
L3	13,640	10
L4	13,640	8
L5	13,640	18
L6	13,640	10
L7	13,640	10
L8	13,640	12
L9	13,640	17
L10	13,640	6
L11	13,640	56

Таблица 4.15. Статистика по числу сайтов с «уникальным» митохондриальным вариантом, найденным только в данной особи (в качестве референтной последовательности использовался митохондриальный контиг L3). Для каждой особи L2-L11 в таблице представлено число митохондриальных сайтов, в которых данная особь отличается от всех остальных особей L2-L11. Представленные числа были рассчитаны для митохондриальных сайтов, в которых генотипы были одновременно определены для всех особей L2-L11 (n = 13,640) с использованием митохондриального контига L3 в качестве референтной последовательности.

	L2	L3	L4	L5	L6	L7	L8	L9	L10
L3	20								
L4	19	27							
L5	48	54	53						
L6	69	77	70	59					
L7	18	26	21	52	73				
L8	69	77	70	59	24	73			
L9	48	56	53	40	59	52	59		
L10	11	21	20	49	70	19	70	49	
L11	93	97	98	91	100	99	104	87	94

Таблица 4.16. Попарные расстояния между митохондриальными гаплотипами особей *A. vaga* L2-L11, определенные с использованием митохондриального контига L3 в качестве референтной последовательности. Для каждой пары особей L2-L11 в таблице представлено абсолютное число нуклеотидных отличий между митохондриальными гаплотипами этих двух особей. Число нуклеотидных отличий для каждой пары особей рассчитывали на основании набора митохондриальных однонуклеотидных полиморфизмов, определенных для особей L2-L11 с использованием митохондриального контига L3 в качестве референтной последовательности (длина = 13,781 п.н.). Представленные числа были рассчитаны для митохондриальных сайтов, в которых генотипы были одновременно определены для всех особей L2-L11 (n = 13,640).

Аналогичным образом мы провели поиск гетерогенных митохондриальных сайтов у L1. Для этого мы картировали прочтения HiSeq для L1 на митохондриальные контиги L1 (извлеченные из геномной сборки L1, основанной на прочтениях MiSeq). Важно отметить, что прочтения HiSeq и MiSeq для L1 были получены в результате секвенирования двух независимых геномных библиотек. Визуальная инспекция выравниваний прочтений HiSeq подтвердила аккуратность сборки митохондриальных контигов L1, полученных из прочтений MiSeq. Среди 10,716 сайтов с определенными генотипами на контигах L1 contig8072 и contig11064, оставшихся после фильтрации (13,531 сайт, если в анализ включали contig12085), согласно результатам SAMtools/BCFtools ни для одного сайта не было найдено данных, поддерживающих присутствие второго варианта. Более того, во всех случаях митохондриальный вариант, восстановленный для L1 на основании прочтений HiSeq, соответствовал нуклеотиду, присутствовавшему в контиге L1.

Однако представлялось возможным, что низкий уровень митохондриальной гетерогенности, найденный в индивидуальных клональных культурах, связан с тем, что программа SAMtools/BCFtools не выявляет варианты, присутствующие в прочтениях Illumina

на средних и низких частотах. Варианты на очень низких частотах не представляли большого интереса с точки зрения выявления возможной контаминации, поскольку такие варианты вероятнее всего связаны с гетероплазмией или ошибками секвенирования. Однако значительное число вариантов на средней частоте может указывать на потенциальную контаминацию между культурами. В связи с этим для того, чтобы более тщательно исследовать возможность присутствия гетерогенных митохондриальных сайтов, мы также провели анализ митохондриальной изменчивости внутри секвенированных культур с использованием программы Mutect2, предназначенной для определения соматических вариантов. Программа Mutect2 разработана специально для поиска соматических мутаций, включая те, которые присутствуют в тканях на низких частотах [212]. Mutect2 не подходит для сравнения генотипов разных особей, но позволяет находить сайты, в которых нереферентный вариант поддерживается небольшой долей прочтений. Определение вариабельных митохондриальных сайтов с помощью Mutect2 (версия GATK 4.1.2.0) [212,213] было проведено для каждой особи в режиме, предназначенном для работы с митохондриальными данными (опция “--mitochondria-mode true”). Для особей L2-L11 в качестве референтной последовательности использовался митохондриальный контиг L4, а для особи L1 – митохондриальные контиги L1. Как и ранее, мы исключили из рассмотрения инсерции и делеции, полиморфные сайты на расстоянии менее или равном десяти нуклеотидов от инсерции или делеции и сайты с покрытием DP <50. Варианты были профильтрованы с помощью утилиты FilterMutectCalls (версия GATK 4.1.2.0) также с применением фильтров, предназначенных для митохондриальных данных (“--mitochondria-mode true” опция). Для последующего анализа были отобраны варианты с метками ‘PASS’ или ‘weak_evidence’ в поле FILTER. Мы не исключали варианты с меткой ‘weak_evidence’, поскольку соответствующая фильтрация не позволяет симметрично рассматривать референтные и альтернативные (нереферентные) аллели: обычно в результате фильтрации по данной метке исключаются сайты, в которых альтернативный вариант поддержан маленьким числом прочтений, но не исключаются сайты, в которых референтный вариант присутствует в маленьком числе прочтений.

Полученные данные были затем использованы для поиска митохондриальной гетерогенности внутри секвенированных культур. Для каждой особи L1-L11 мы определили число сайтов, в которых с помощью Mutect2 в картированных прочтениях, полученных для этой культуры, было найдено два нуклеотидных варианта таких, что более редкий вариант был поддержан не менее чем 3 прочтениями. Как и ожидалось, число гетерогенных митохондриальных сайтов, найденных Mutect2, выше, чем число таких сайтов по результатам SAMtools/BCFtools. Для каждой особи было найдено от 1 (L7 и L10) до 11 (L6) таких сайтов (Таблица 4.17). Однако в большей части случаев минорный (более редкий из двух) вариант

присутствовал только в совсем небольшой доле прочтений (Таблицы 4.17 и 4.18). Например, среди 6 гетерогенных сайтов, найденных для L2, в среднем только 2.4% прочтений (медианное значение 0.7%) поддерживали минорный вариант (Таблица 4.18). Если мы требовали, чтобы минорный вариант был поддержан ≥1% прочтений, число гетерогенных сайтов значительно уменьшалось: ни для одной особи не было найдено более трех гетерогенных сайтов, удовлетворяющих данному критерию (Таблица 4.17). В соответствии с результатами, полученными с помощью SAMtools/BCFtools, гетерогенные сайты с хорошей поддержкой прочтениями, определяемые как сайты, в которых минорный вариант присутствует в ≥10% прочтений, были найдены только для трех особей (Таблица 4.17). Это те же самые три особи, которые были определены по результатам SAMtools/BCFtools (L5, L6 и L9), с единственным отличием – в случае L9 вместо одного гетерогенного сайта с помощью Mutect2 было найдено два таких сайта.

	Митохондриальные сайты, в которых согласно результатам Mutect2 в прочтениях из одной особи было найдено два нуклеотидных варианта		
Особь	Общее число	Более редкий вариант поддержан ≥1% прочтений	Более редкий вариант поддержан ≥10% прочтений
L1	2	0	0
L2	6	2	0
L3	2	1	0
L4	8	1	0
L5	7	3	1
L6	11	2	1
L7	1	0	0
L8	5	1	0
L9	7	2	2
L10	1	1	0
L11	7	0	0

Таблица 4.17. Статистика по числу митохондриальных сайтов, в которых согласно результатам Mutect2 в прочтениях из одной особи было найдено два нуклеотидных варианта. Для особей L2-L11 в качестве референтной последовательности для поиска гетерогенных митохондриальных сайтов с помощью Mutect2 использовался митохондриальный контиг L4, для особи L1 использовались митохондриальные контиги L1. Рассматривались только те гетерогенные митохондриальные сайты, в которых более редкий вариант был поддержан ≥3 прочтениями. Для каждой особи в таблице представлена информация об общем числе таких сайтов и данные о числе гетерогенных митохондриальных сайтов, в которых более редкий вариант поддержан ≥1% и ≥10% картированных прочтений в соответствующей геномной позиции.

Особь	Митохондриальные сайты, в которых согласно результатам Mutect2 в прочтениях из одной особи было найдено два нуклеотидных варианта						
	Общее число	Среднее покрытие сайта прочтениями	Медианное покрытие сайта прочтениями	Среднее число прочтений, поддерживающих более редкий вариант	Медианное число прочтений, поддерживающих более редкий вариант	Средняя доля прочтений, поддерживающих более редкий вариант	Медианная доля прочтений, поддерживающих более редкий вариант
L1	2	634.5	634.5	4.5	4.5	0.7%	0.7%
L2	6	736.5	760.5	16.2	4.5	2.4%	0.7%
L3	2	1,727.0	1,727	21.5	21.5	1.3%	1.3%
L4	8	1,046.0	1,157	4.8	3.5	0.5%	0.3%
L5	7	1,578.9	1,409	48.1	11	5.8%	0.8%
L6	11	2,121.3	2,448	26.4	4	1.8%	0.2%
L7*	1	1,223	NA	3	NA	0.2%	NA
L8	5	1,718.8	1,559	6.0	5	0.5%	0.3%
L9	7	1,112.3	1,167	61.7	4	5.8%	0.3%
L10*	1	100	NA	7	NA	7.0%	NA
L11	7	2,314.1	2,375	5.9	3	0.2%	0.2%

Таблица 4.18. Статистика по доле и числу прочтений, поддерживающих более редкий вариант в гетерогенных митохондриальных сайтах, найденных для особей L1-L11 с помощью Mutect2. Для особей L2-L11 в качестве референтной последовательности для поиска гетерогенных митохондриальных сайтов с помощью Mutect2 использовался митохондриальный контиг L4, для особи L1 использовались митохондриальные контиги L1. Рассматривались только те гетерогенные митохондриальные сайты, в которых более редкий вариант был поддержан ≥3 прочтениями. В случае L7 и L10, для которых был найден только один такой сайт (отмечены звездочкой), показаны точное число/доля прочтений, поддерживающих более редкий вариант в этом сайте.

Таким образом, сценарий контаминации представляется маловероятным: напротив, результаты анализа митохондриальной изменчивости для особей L1-L11 хорошо согласуются с тем, что ожидается в ситуации, когда в каждой культуре присутствует один преобладающий вариант митохондриального генома, возможно, с некоторой гетерогенностью в нескольких сайтах, которая может объясняться гетероплазмией, заменами, накопленными разными особями внутри культуры, или техническими артефактами.

4.2.8 Исследование возможных сценариев обмена генетическим материалом у *A. vaga*

Результаты анализа популяционной изменчивости у *A. vaga*, представленные в этой главе, указывают на существование рекомбинации у этого вида и вероятного обмена генетическим материалом между индивидуумами. Однако полученные результаты совместимы как минимум с двумя разными механизмами генетического обмена: горизонтальным переносом генов и половым размножением, включающим формирование гамет за счет мейоза.

Гипотеза о горизонтальном переносе генов внутри популяций бделлоидных коловраток была ранее выдвинута на основе присутствия в их геномах значительного числа генов,

перенесенных из видов, не относящихся к *Metazoa* [26,67,214]. Так, в геномах бделлоидных коловраток были найдены гены бактериального и растительного происхождения, а также гены, по всей видимости, перенесенные из геномов грибов [26,67,214]. Т.к. присутствие таких «чужеродных» генов предполагает существование механизма горизонтального переноса ДНК в геномы бделлоидных коловраток из далеких видов, возможным представляется и присутствие горизонтального переноса между особями бделлоидных коловраток внутри одного вида [67].

Возможность существования у бделлоидных коловраток полового размножения *sensu stricto* тоже ранее обсуждалась в литературе [28]. При этом помимо классического мейоза также рассматривалась гипотеза об атипичном мейозе, протекающем по схеме, описанной для некоторых видов растений из рода *Oenothera* [28]. При мейозе такого типа сегрегация происходит без выравнивания гомологичных хромосом по длине относительно друг друга и преимущественно без рекомбинации (рекомбинация предположительно возможна в теломерных областях) [28]. Ранее половое размножение с атипичным мейозом по схеме *Oenothera* предложили как возможный способ обмена генетическим материалом у бделлоидных коловраток на основании паттерна наследования аллелей в нескольких геномных локусах у индивидуумов вида *Macrotrachela quadricornifera* [28].

В противоположность тому, что ожидается при строго клональном размножении, любой тип генетического обмена должен приводить к тому, что филогении для двух гаплотипов одного индивидуума будут отличаться [28] (Рисунок 4.18). Однако то, как инконгруэнтность филогений для двух гаплотипов проявляется в разных геномных локусах, может быть использовано для того, чтобы отличить атипичный мейоз по схеме *Oenothera* от двух других сценариев (классический мейоз и горизонтальный перенос генов).

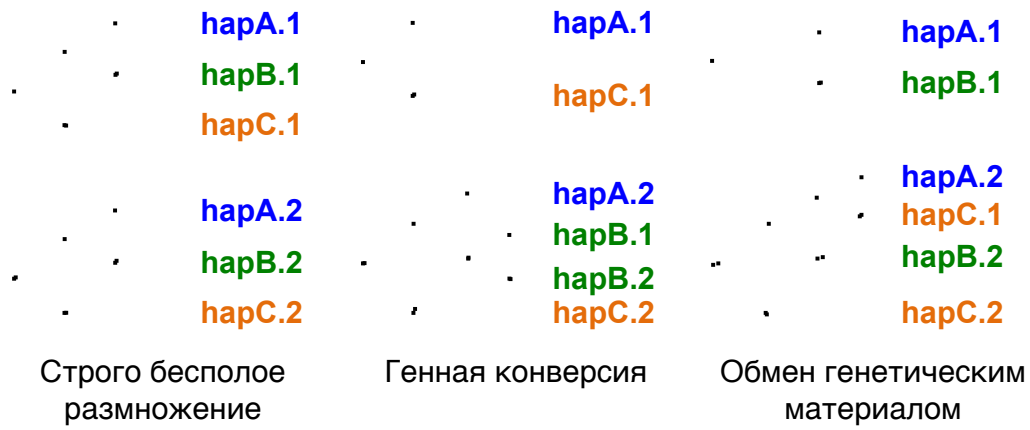

Рисунок 4.18. Филогенетические взаимоотношения между гаплотипами разных особей, ожидаемые или совместимые с разными эволюционными сценариями [28,76]. Гаплотипы трех разных особей показаны разными цветами. Индексы 1 и 2 обозначают два гаплотипа одной особи.

Ключевой особенностью атипичного мейоза, протекающего по схеме *Oenothera*, является присутствие двух комплексов гаплотипов, устроенных таким образом, что все хромосомы, принадлежащие к одному комплексу, всегда наследуются совместно [28]. Из-за этого ожидается, что при мейозе такого вида паттерн инконгруэнтности для двух гаплотипов одного и того же индивидуума будет одинаковым для всего генома [28]. И, напротив, ожидается, что как классический мейоз, так и горизонтальный перенос генов будут создавать различные паттерны инконгруэнтности для разных локусов генома [28].

Мы сравнили филогении для двух гаплотипов каждого индивидуума в разных геномных локусах (см. Материалы и методы). Этот анализ выявил значительное количество случаев, в которых два гаплотипа одного и того же индивидуума кластеризовались с гаплотипами из разных индивидуумов (Рисунок 4.19; Таблица 4.19). Такие случаи инконгруэнтности были найдены для всех индивидуумов из большого кластера (L4-L11). При этом в разных геномных локусах для двух гаплотипов одного и того же индивидуума наблюдались различные паттерны инконгруэнтности (Рисунок 4.19; Таблицы 4.19 и 4.20).

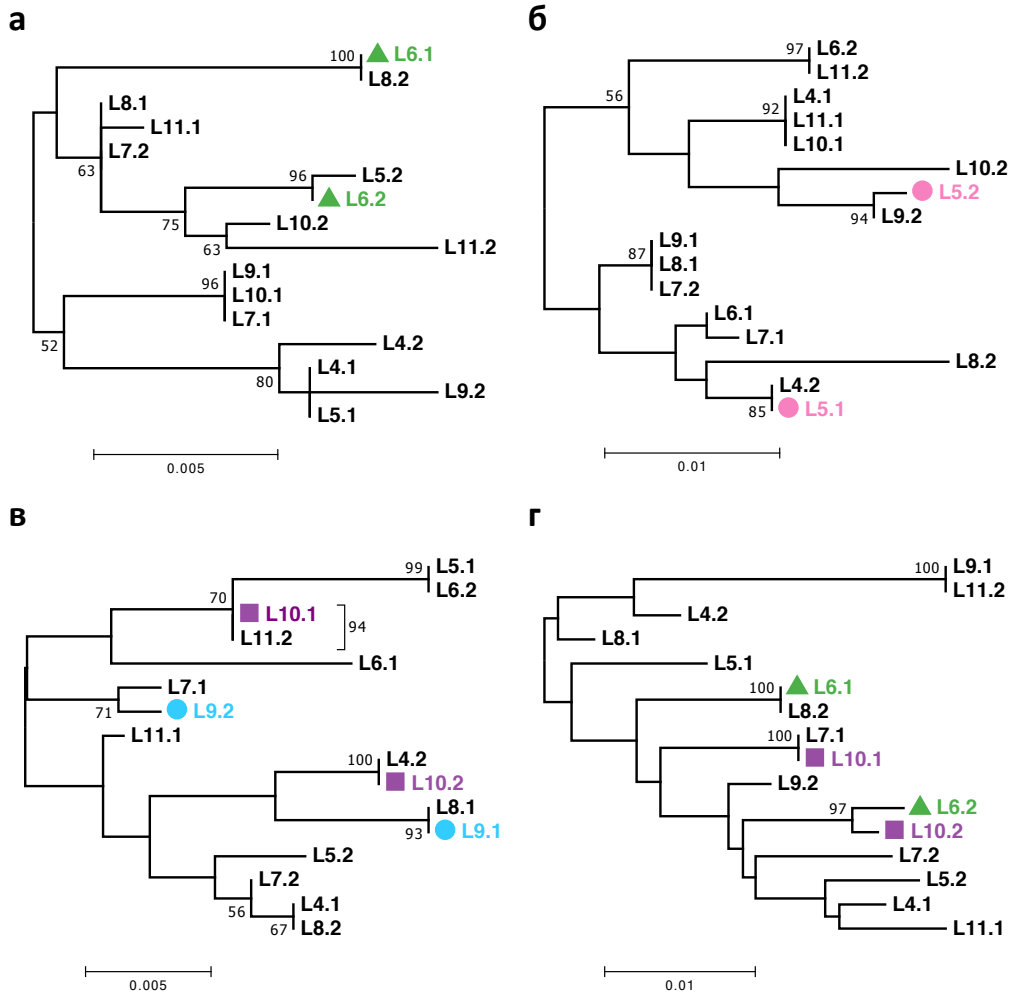

Рисунок 4.19. Филогенетический анализ гаплотипов индивидуумов L4-L11 указывает на обмен генетическим материалом у *A. vaga*. Укорененные посередине филогенетические деревья, построенные для четырех фазированных сегментов, в которых была найдена инконгруэнтность двух гаплотипов одного индивидуума. В 52 из 303 проанализированных сегментов «ближайшие соседи» двух гаплотипов одного индивидуума были найдены в двух разных индивидуумах (Таблица 4.19). Филогенетические деревья, показанные на данном рисунке, были построены для четырех фазированных сегментов, находящихся на разных контигах в диплоидной сборке L1. Длина сегментов составляет 896 (**а**), 604 (**б**), 1348 (**в**) и 1198 (**г**) п.н. Филогенетические деревья были построены с применением метода максимального правдоподобия, реализованного в программе PhyML [183], с использованием модели GTR+G (1000 бутстреп-реплик). Значения бутстреп-поддержки ≥50% показаны рядом с соответствующими ветвями. Значение бутстреп-поддержки для клады, включающей гаплотипы L10.1 и L11.2, в панели (**в**) показано рядом со скобкой. Значения бутстреп-поддержки округлены до ближайшего целого числа. Для построения филогенетических деревьев были использованы только те полиморфные сайты, которые прошли все шаги фильтрации и были одновременно фазированы у индивидуумов L4-

L11 (n = 23, 21, 33 и 57 полиморфных сайтов с частотой минорного варианта ≥2 в панелях **а**–**г** соответственно); остальные сайты рассматривались как мономорфные. Индексы 1 и 2 обозначают два гаплотипа одного индивидуума. В данных четырех сегментах два гаплотипа одного индивидуума кластеризовались с гаплотипами из разных индивидуумов с достойной бутстреп-поддержкой (≥70%) для L5 (**б**; розовый), L6 (**а**, **г**; зеленый), L9 (**в**; голубой) и L10 (**в**, **г**; фиолетовый).

Особь	Число проанализированных фазированных сегментов	Число сегментов, для которых была выявлена инконгруэнтность	Число разных паттернов инконгруэнтности	Выявленные паттерны инконгруэнтности
L4	303	9	7	L5-L9 (1), L5-L11 (1), L6-L7 (1), L6-L8 (1), L6-L11 (3), L7-L11 (1), L10-L11 (1)
L5	303	9	8	L4-L7 (1), L4-L9 (1), L4-L10 (2), L6-L9 (1), L6-L10 (1), L6-L11 (1), L7-L11 (1), L9-L10 (1)
L6	303	13	9	L4-L10 (1), L5-L8 (2), L5-L9 (1), L7-L10 (2), L7-L11 (1), L8-L9 (1), L8-L10 (1), L8-L11 (3), L9-L11 (1)
L7	303	4	2	L4-L8 (1), L5-L9 (3)
L8	303	11	8	L4-L5 (1), L4-L7 (1), L5-L7 (1), L5-L10 (1), L6-L7 (1), L6-L9 (2), L6-L11 (3), L9-L10 (1)
L9	303	5	5	L4-L8 (1), L5-L10 (1), L5-L11 (1), L6-L8 (1), L7-L8 (1)
L10	303	7	6	L4-L9 (1), L4-L11 (2), L5-L6 (1), L6-L7 (1), L7-L8 (1), L7-L11 (1)
L11	303	7	6	L4-L7 (1), L5-L6 (1), L5-L8 (2), L6-L10 (1), L7-L8 (1), L7-L9 (1)

Таблица 4.19. Случаи инконгруэнтности двух гаплотипов, найденные в особях L4-L11. Для каждой особи мы определили число фазированных сегментов таких, что реципрокные «ближайшие соседи» двух гаплотипов этой особи принадлежали разным особям. Рассматривались только случаи инконгруэнтности с бутстреп-поддержкой ≥70%. Для каждой особи L4-L11 показано общее число сегментов, для которых была выявлена инконгруэнтность, а также общее число разных паттернов инконгруэнтности. В самом правом столбце для каждой особи перечислены все найденные паттерны инконгруэнтности и число сегментов, в которых данный паттерн встретился (указано в скобках). Для каждой особи А каждая пара особей, в которых были найдены реципрокные «ближайшие соседи» двух гаплотипов особи А в одном или более локусах, соответствует отдельному паттерну инконгруэнтности. Данный анализ был проведен на сегментах генома *A. vaga*, в которых по крайней мере для 15 полиморфных сайтов была одновременно определена фаза во всех особях L4-L11 (синглтоны не рассматривались). В анализ не были включены сегменты, для которых было найдено более двух находок с высокой степенью нуклеотидного сходства (≥90%) в геноме L1, а также сегменты, для которых в геноме L1 было найдено большое количество гомологичных участков (независимо от степени сходства). В общей сложности инконгруэнтность для двух гаплотипов одной особи была выявлена в 52 из 303 проанализированных сегментов (включены только случаи с бутстреп-поддержкой ≥70%). Данные по общему числу случаев, в которых встретился каждый паттерн инконгруэнтности (среди перечисленных в самом правом столбце данной таблицы), представлены в Таблице 4.20.

	L4	L5	L6	L7	L8	L9	L10
L5	1						
L6	0	2					
L7	3	1	3				
L8	2	4	2	3			
L9	2	5	3	1	1		
L10	3	2	2	2	1	2	
L11	2	2	7	4	3	1	1

Таблица 4.20. Паттерны инконгруэнтности двух гаплотипов, выявленные для разных фазированных сегментов генома *A. vaga*. Для каждой пары особей L4-L11 мы определили число фазированных сегментов (среди 303 проанализированных сегментов из набора A), таких что была найдена третья особь, один гаплотип которой был сильнее всего похож на гаплотип первой особи из пары, а другой гаплотип был сильнее всего похож на гаплотип второй особи из пары (рассматривались только случаи с бутстреп-поддержкой ≥70%, см. Таблицу 4.19). Для некоторых сегментов такой паттерн был найден более чем для одной пары особей.

Мы воспроизвели результаты этого анализа, используя для реконструкции гаплотипов набор однонуклеотидных полиморфизмов, дополнительно профильтрованный для того, чтобы уменьшить возможное влияние таких артефактов как ошибочное демультиплексирование данных секвенирования Illumina, вызванное аберрантным отнесением индексов к прочтениям, а также ошибочное определение гетерозиготных генотипов как гомозиготных (см. Материалы и методы; Таблица 4.21). Поскольку результаты, полученные с использованием этого дополнительно профильтрованного набора данных, схожи с результатами основного анализа, можно заключить, что инкогруэнтность филогений гаплотипов не объясняется упомянутыми выше артефактами.

Особь	Число проанализированных фазированных сегментов	Число сегментов, для которых была выявлена инконгруэнтность	Число разных паттернов инконгруэнтности	Выявленные паттерны инконгруэнтности
L4	190	2	2	L5-L11 (1), L6-L11 (1)
L5	190	2	2	L4-L10 (1), L6-L11 (1)
L6	190	6	5	L5-L8 (2), L5-L9 (1), L8-L10 (1), L8-L11 (1), L9-L11 (1)
L7	190	3	3	L4-L8 (1), L5-L9 (1), L10-L11 (1)
L8	190	8	6	L4-L5 (1), L5-L7 (1), L5-L10 (1), L6-L9 (2), L6-L10 (1), L6-L11 (2)
L9	190	2	2	L5-L10 (1), L6-L8 (1)
L10	190	6	5	L4-L11 (1), L6-L7 (1), L7-L8 (1), L7-L11 (2), L8-L9 (1)
L11	190	4	4	L4-L7 (1), L6-L10 (1), L7-L8 (1), L7-L9 (1)

Таблица 4.21. Случаи инконгруэнтности двух гаплотипов, найденные в особях L4-L11 для фазированных сегментов из набора B. Данная таблица аналогична Таблице 4.19, но представленные результаты получены с использованием фазированных сегментов из набора B.

Сегменты, включенные в набор B, были дополнительно профильтрованы с целью исключения полиморфных сайтов, возможно затронутых ошибками определения полиморфизмов, в том числе ошибками, связанными с аберрантным отнесением индексов к прочтениям или ошибочным определением гетерозиготных генотипов как гомозиготных (см. Материалы и методы). В общей сложности инконгруэнтность для двух гаплотипов одной особи была выявлена в 25 из 190 проанализированных сегментов, входящих в набор B (включены только случаи с бутстреп-поддержкой ≥70%).

Существование множественных паттернов инконгруэнтности для двух гаплотипов одного индивидуума является дополнительным свидетельством в пользу генетического обмена у *A. vaga* и делает атипичный мейоз по схеме *Oenothera* маловероятным объяснением генетического обмена у бделлоидных коловраток. Тем не менее наши данные не позволяют исключить того, что наблюдаемая инконгруэнтность филогений гаплотипов может быть вызвана сочетанием атипичного мейоза и генной конверсии и/или митотической рекомбинации [92] (Рисунок 4.20). Однако подобные сценарии не представляются максимально экономным объяснением для наших наблюдений. Как горизонтальный перенос генов, так и классический мейоз являются более экономными объяснениями, т.к. требуют предположения о существовании меньшего числа различных процессов.

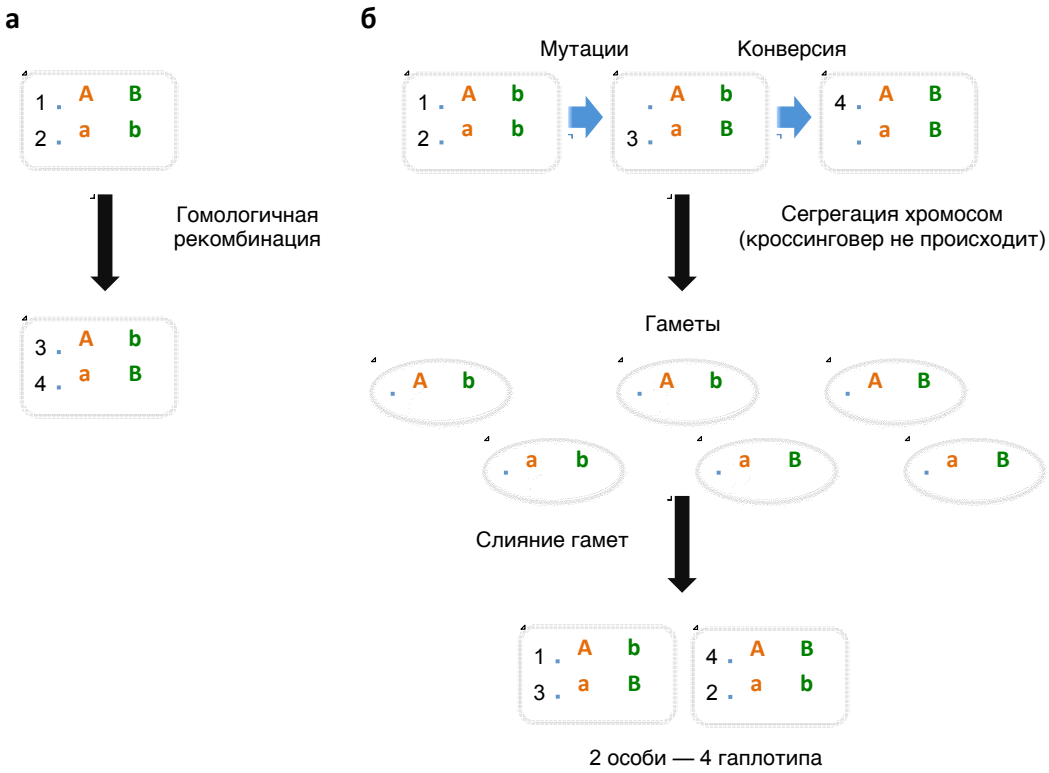

Рисунок 4.20. Возникновение четырех гаплотипов для пары гетерозиготных сайтов в двух особях в результате реципрокной рекомбинации или трансформации (**а**) или в результате генной конверсии, сопряженной с половым размножением с атипичным мейозом по схеме *Oenothera* (**б**).

Помимо генетического обмена другой возможной причиной инконгруэнтности может быть параллельное закрепление конвергентных мутаций в независимых линиях *A. vaga*. Чтобы проверить эту гипотезу, мы провели аннотацию сегментов генома, для которых была показана инконгруэнтность филогений для двух гаплотипов (далее именуются сокращенно как «инконгруэнтные сегменты»). Значительная доля инконгруэнтных сегментов имеет существенное пересечение с белок-кодирующими генами (0.77); однако это доля не отличается значимо от соответствующей доли среди всех проанализированных сегментов (P = 0.9968) или среди сегментов, оставшихся после исключения инконгруэнтных (P = 0.9962). Это наблюдение указывает на то, что независимое закрепление конвергентных мутаций у *A. vaga* под действием положительного отбора является маловероятным объяснением инконгруэнтности.

Паттерны инконгруэнтности в большом кластере (L4-L11) не позволяют сделать заключение о том, происходит ли у *A. vaga* горизонтальный перенос генов или половое размножение с классическим мейозом. Для того, чтобы получить дополнительные данные, которые могли бы помочь отличить эти два сценария, мы провели расширенный анализ филогений гаплотипов, включив в него все 11 особей (L1-L11). При проведении этого анализа

были использованы только те сегменты генома, в которых гаплотипы были восстановлены одновременно для всех 11 индивидуумов. Несмотря на небольшое число сегментов, удовлетворяющих этому условию (n = 152), мы выявили несколько случаев (n = 11) инконгруэнтности гаплотипов индивидуумов L1-L3 из маленького кластера (Таблица 4.22). Интересно, что во всех таких случаях, когда лучшие «хиты» для двух гаплотипов индивидуума (обозначим их как H1 и H2) из маленького кластера были найдены в разных индивидуумах, один из гаплотипов (H1) всегда кластеризовался с гаплотипом другого индивидуума из маленького кластера, а второй гаплотип (H2) – с гаплотипом индивидуума из большого кластера.

Особь	Число проанализированных фазированных сегментов	Число сегментов, для которых была выявлена инконгруэнтность	Число разных паттернов инконгруэнтности	Выявленные паттерны инконгруэнтности
L1	152	5	4	L2-L5 (1), L3-L4 (1), L3-L5 (1), L3-L10 (2)
L2	152	2	2	L3-L8 (1), L3-L11 (1)
L3	152	4	2	L1-L4 (2), L2-L9 (2)
L4	152	1	1	L5-L11 (1)
L5	152	0	0	
L6	152	2	2	L1-L8 (1), L5-L8 (1)
L7	152	0	0	
L8	152	2	2	L6-L9 (1), L6-L11 (1)
L9	152	1	1	L3-L6 (1)
L10	152	1	1	L3-L4 (1)
L11	152	2	2	L1-L5 (1), L7-L9 (1)

Таблица 4.22. Случаи инконгруэнтности двух гаплотипов, найденные в особях L4-L11 для фазированных сегментов из набора C. Данная таблица аналогична Таблице 4.19, но представленные результаты получены с использованием фазированных сегментов из набора C, включающих сегменты, фазированные во всех 11 особях L1-L11 (несущие ≥15 полиморфных сайтов, для которых одновременно определена фаза у L1-L11). Сегменты, включенные в набор C, были дополнительно профильтрованы с целью исключения полиморфных сайтов, возможно затронутых ошибками определения полиморфизмов, в том числе ошибками, связанными с аберрантным отнесением индексов к прочтениям или ошибочным определением гетерозиготных генотипов как гомозиготных (см. Материалы и методы). В общей сложности инконгруэнтность для двух гаплотипов одной особи была выявлена в 16 из 152 проанализированных сегментов, входящих в набор B (включены только случаи с бутстреп-поддержкой ≥70%).

Чтобы получить более подробную картину того, как гаплотипы особей из малого и большого кластеров соотносятся друг с другом, мы проанализировали филогенетические деревья для всех 152 сегментов, в которых гаплотипы были одновременно восстановлены для

11 индивидуумов L1-L11. Филогенетические деревья были построены для этой цели с применением метода максимального правдоподобия, реализованного в программе PhyML [183], и укоренены в центре (см. Материалы и методы). Дальнейший анализ полученных деревьев выявил, что в большинстве случаев L1, L2 и L3 кластеризовались по одному из двух гаплотипов, но не по второму. В совокупности из 152 сегментов в 100 три гаплотипа (по одному гаплотипу каждого из индивидуумов L1-L3) формировали кладу с хорошей поддержкой, в то время как три оставшихся гаплотипа из L1-L3 были перемешаны с гаплотипами L4-L11 (Рисунок 4.21а–г). В 36 из 100 случаев укоренение посередине разделило дерево на монофилетическую группу, содержащую три гаплотипа из L1, L2 и L3, и остальные гаплотипы (Рисунок 4.21в–г). Случаев, в которых в дереве была бы найдена монофилетическая группа, состоящая из всех шести гаплотипов L1, L2 и L3, выявлено не было. В соответствии с этими результатами в 117 из 152 сегментов дерево, построенное методом максимального правдоподобия без наложения ограничений на топологию, статистически значимо отличалось от дерева, для которого было задано ограничение на то, чтобы шесть гаплотипов, принадлежащие трем индивидуумам из маленького кластера, формировали монофилетическую группу (тест Swofford–Olsen–Waddell–Hillis [185,186], P-значение < 0.05 после поправки Бонферрони; см. Материалы и методы).

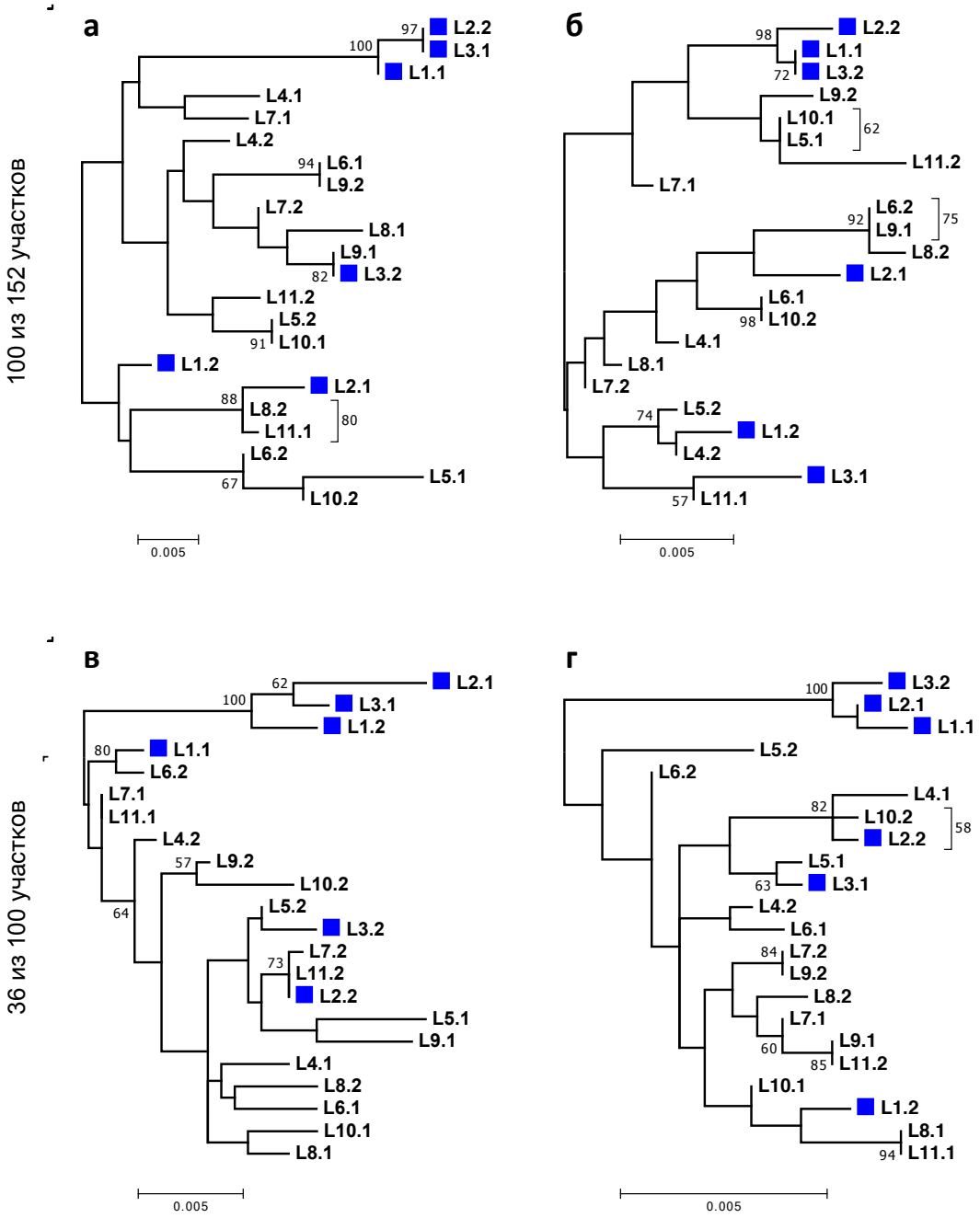

Рисунок 4.21. Филогенетический анализ гаплотипов особей L1-L11 намекает на гибридное происхождение особей из маленького кластера. Укорененные посередине филогенетические деревья, построенные для четырех сегментов, в которых фазы полиморфизмов удалось восстановить одновременно для 11 особей из обоих кластеров (L1-L11), и, предположительно, указывающие на гибридное происхождение индивидуумов из маленького кластера. Гаплотипы особей из маленького кластера показаны с помощью синих квадратов. В 100 из 152 сегментов была найдена монофилетическая группа, в состав которой входило три гаплотипа – по одному гаплотипу каждой особи L1, L2 и L3 (как в панелях **а**–**г**). Среди этих 100 сегментов в 36 корень, помещенный посередине, отделял монофилетическую группу, включающую три гаплотипа L1,

L2 и L3, от всех остальных гаплотипов (как в панелях **в** и **г**). Филогенетические деревья, показанные на данном рисунке, были построены для четырех фазированных сегментов, находящихся на разных контигах в диплоидной сборке L1. Длина сегментов составляет 1058 (**а**), 1452 (**б**), 1782 (**в**) и 1824 (**г**) п.н. Филогенетические деревья были построены с применением метода максимального правдоподобия, реализованного в программе PhyML [183], с использованием модели GTR+G (1000 бутстреп-реплик). Значения бутстреп-поддержки ≥50% показаны рядом с соответствующими ветвями или рядом со скобками. Значения бутстреп-поддержки округлены до ближайшего целого числа. Для построения филогенетических деревьев были использованы только те полиморфные сайты, которые прошли все шаги фильтрации и были одновременно фазированы у особей L1-L11 (n = 46, 32, 43 и 30 полиморфных сайтов с частотой минорного варианта ≥2 в панелях **а**–**г** соответственно); остальные сайты рассматривались как мономорфные. Индексы 1 и 2 обозначают два гаплотипа одной особи.

Эти результаты было бы сложно объяснить в рамках механизма горизонтального переноса генов. В то же время они совместимы с классическим мейозом, если мы предположим, что три индивидуума из маленького кластера имеют гибридное происхождение и являются потомством, появившимся в результате скрещивания особей из популяции, генетически близкой к популяции большого кластера, и другой популяции, находящейся на большем генетическом расстоянии от популяции большого кластера. Этот сценарий также мог бы объяснить более высокий уровень гетерозиготности индивидуумов из маленького кластера, отражающий присутствие двух относительно дивергентных гаплотипов (Рисунок 4.6).

Интересно, что анализ митохондриальной изменчивости указывает на то, что особи L1-L3 не могли быть результатом одного события полового размножения, – единственного события скрещивания между особями из двух разных популяций в данном случае было бы недостаточно (Рисунки 4.3, 4.4, 4.22 и 4.23). В то время как митохондриальные гаплотипы L2 и L3 очень близки к митохондриальным гаплотипам L4-L11, индивидуум L1 несет высоко дивергентный митохондриальный гаплотип, находящийся на значительном расстоянии от гаплотипов остальных десяти особей (Рисунки 4.3, 4.4 и 4.23). Поскольку единственного события скрещивания между особями из двух разных популяций было бы недостаточно, чтобы создать подобный паттерн наследования митохондриальных и ядерных маркеров, это наблюдение может указывать на повторные события гибридизации. Так, например, наши данные совместимы со сценарием, включающим по меньшей мере два реципрокных события гибридизации: одно событие, в результате которого был унаследован митохондриальный гаплотип из «популяции большого кластера» (L2 и L3), и другое событие, в результате которого

был унаследован митохондриальный геном из второй неизвестной популяции (L1). К сожалению, данную гипотезу невозможно проверить в отсутствие данных по митохондриальным геномам особей из второй гипотетической популяции.

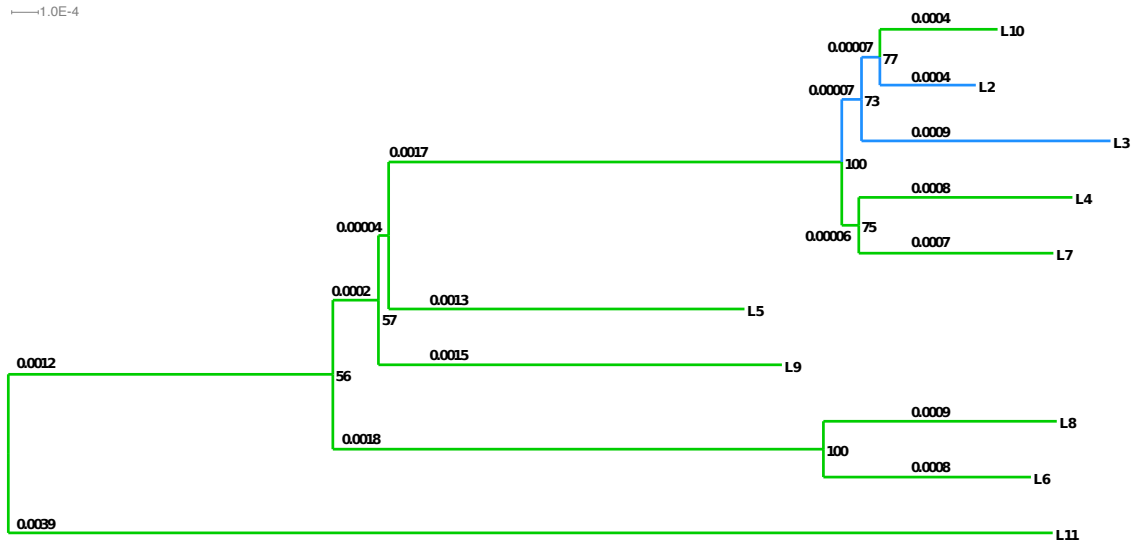

Рисунок 4.22. Митохондриальное филогенетическое дерево особей L2-L11, восстановленное методом максимального правдоподобия. Филогенетическое дерево было построено с использованием митохондриальных гаплотипов L2-L11, реконструированных на основании митохондриальных генотипов этих особей. В качестве референтной последовательности для определения митохондриальных генотипов использовали митохондриальный контиг L4 (длина 14,052 п.н.). Филогенетическое дерево было реконструировано с применением метода максимального правдоподобия, реализованного в программе RAxML [187], с использованием модели GTR+G (1000 бутстреп-реплик) и укоренено по самой длинной ветви (L11). Значения бутстреп-поддержки показаны рядом с узлами; длины ветвей указаны рядом с ветвями. Ветви, ведущие к индивидуумам маленького и большого кластеров, показаны синим и зеленым цветом соответственно. Данное дерево построено С. А. Науменко.

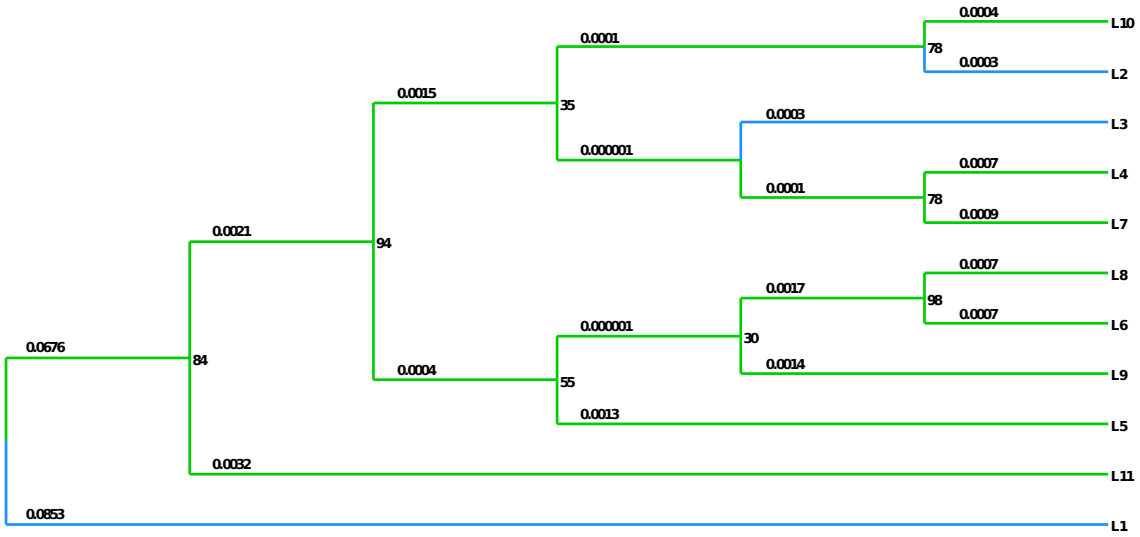

Рисунок 4.23. Митохондриальное филогенетическое дерево особей L1-L11, восстановленное методом максимального правдоподобия. Филогенетическое дерево было построено для участка митохондриального генома, соответствующего самому длинному митохондриальному контигу L1 (contig8072, длина = 7124 п.н.). Дерево построено с использованием выравнивания контига L1 contig8072 с митохондриальными гаплотипами L2-L11, реконструированными на основании митохондриальных генотипов этих особей (общая длина выравнивания 7126 п.н.). В качестве референтной последовательности для определения митохондриальных генотипов L2-L11 использовали митохондриальный контиг L4. Филогенетическое дерево было реконструировано с применением метода максимального правдоподобия, реализованного в программе RAxML [187], с использованием модели GTR+G (1000 бутстреп-реплик) и укоренено по самой длинной ветви (L1). Значения бутстреп-поддержки показаны рядом с узлами; длины ветвей указаны рядом с ветвями (для того, чтобы подчеркнуть топологию дерева, ветви изображены без сохранения пропорций длин). Ветви, ведущие к индивидуумам маленького и большого кластеров, показаны синим и зеленым цветом соответственно. Данное дерево построено С. А. Науменко.

Тем не менее данные по уровню дивергенции гаплотипов внутри особи и между особями косвенно поддерживают гипотезу о реципрокных событиях гибридизации. Если предположить, что особи L1-L3 действительно имеют гибридное происхождение, то уровень дивергенции между двумя гаплотипами L1-L3 отражает генетическое расстояние между ядерными геномами двух популяций, особи из которых участвовали в гибридизации. Степень такой дивергенции,

оцененная как доля гетерозиготных сайтов в геномах L1-L3, составляет ~2% (Таблица П4.1). В то же время, если L2-L3 имеют митохондриальный гаплотип, унаследованный от «популяции большого кластера», а митохондриальный гаплотип L1 унаследован от второй неизвестной популяции, вовлеченной в гибридизацию, то степень митохондриальной дивергенции между этими двумя популяциями составляет ~9% (эта оценка получена на основании лучших находок blastn для митохондриальных контигов L1 contig8072 и contig11064 в митохондриальных геномах L2-L11; см. раздел 4.2.2). Данные оценки гипотетической ядерной и митохондриальной дивергенции нужно сравнивать с большой осторожностью, поскольку в то время как оценки ядерной дивергенции получены с использованием генотипов, определенных на основании картировании прочтений на референтный геном, митохондриальные оценки сделаны с помощью сравнения *de novo* собранных контигов. И все же митохондриальный гаплотип L1 очевидно отличается от митохондриальных гаплотипов L2-L11 значительно сильнее, чем два ядерных гаплотипа L1-L3 друг от друга. Это наблюдение согласуется с тем, что ожидалось бы в случае реципрокных событий рекомбинации. Хорошо известно, что скорость возникновения мутаций в митохондриях и степень митохондриальной дивергенции у большинства видов по крайней мере в несколько раз выше, чем соответствующие показатели для ядерного генома [215]. Следовательно, ожидается, что в случае, если разница между ядерными геномами особей из двух популяций составляет ~2%, уровень митохондриальной дивергенции между этими двумя популяциями будет гораздо выше. Это предсказание соответствует тому, что наблюдается в наших данных. Таким образом, результаты сопоставления данных по митохондриальной и ядерной изменчивости являются дополнительным аргументом в пользу сценария с реципрокными скрещиваниями.

Интересно, что в митохондриальных филогениях две особи из маленького кластера с похожими гаплотипами (L2 и L3) не формируют монофилетическую группу: L2 группируется не с L3, а с L10 (Рисунки 4.22 и 4.23). Это позволяет предположить, что особи L2 и L3 появились в результате независимых событий полового размножения. Если это так, то для того, чтобы объяснить возникновение особей маленького кластера, необходимо предположить три события полового размножения.

4.2.9 Оценка гипотетической частоты мейоза

Далее мы задались вопросом о том, какая частота мейоза необходима, чтобы в отсутствие других механизмов объяснить наблюдаемый распад LD с расстоянием. Чтобы ответить на этот вопрос, необходимо знать c, скорость рекомбинации на нуклеотид на поколение (см. Материалы и методы). c, в свою очередь, можно оценить из отношения популяционной скорости рекомбинации $4N_{\mathrm{e}}c$ (где N_{e} – эффективная численность популяции) к

популяционной скорости возникновения мутаций $4N_e\mu$, если известен параметр μ, обозначающий скорость возникновения мутаций на нуклеотид на поколение. Полученные оценки $4N_e c$ для индивидуумов L4-L11 имеют порядок 10^{-2} (см. Материалы и методы). Уровень генетической изменчивости указывает на то, что $4N_e\mu$ также $\sim 10^{-2}$. Таким образом, $c \sim \mu$. К сожалению, данные о скорости мутирования у *A. vaga* отсутствуют. Если эта скорость находится в диапазоне 10^{-9}–10^{-8}, как у большого числа различных многоклеточных эукариот, то c тоже имеет порядок 10^{-9}–10^{-8}. Такие оценки c соответствуют одному событию мейоза на ~10–100 поколений (см. Материалы и методы).

4.2.10 Обсуждение

В этой главе мы проанализировали полные геномы 11 индивидуумов *A. vaga* и показали, что изменчивость в природной популяции этой бделлоидной коловратки значительно рандомизирована. Это проявляется в том, что внутри одного локуса генотипы находятся в пропорциях в целом близких к тем, которые ожидаются при равновесии Харди-Вайнберга, а неравновесие по сцеплению между аллелями в разных локусах быстро распадается с расстоянием. Такое отсутствие статистических ассоциаций между аллелями представляет собой свидетельство в пользу существования рекомбинации и генетического обмена [216] между индивидуумами у *A. vaga*. В частности, модифицированный вариант четырехгаметного теста исключает генную конверсию в качестве единственной причины распада LD, в то время как анализ трехаллельных сайтов и филогений гаплотипов указывает на существование обмена генетическим материалом между индивидуумами.

Однако генетический обмен между индивидуумами может происходить за счет двух очень разных механизмов. Одним механизмом является половое размножение *sensu stricto*, включающее скрещивание между индивидуумами и реципрокную мейотическую рекомбинацию. Другой механизм – горизонтальный перенос генов (возможно, в результате трансформации), сопровождающийся нереципрокной рекомбинацией, обеспечивающей инкорпорацию ДНК из внешней среды в геном индивидуума. Можем ли мы понять, какой из этих механизмов в действительности присутствует у бделлоидных коловраток?

Наиболее серьезным аргументом в пользу классического полового размножения является существование гибридов. Здесь важно подчеркнуть, что горизонтальный перенос генов не может объяснить их возникновение. Если наша интерпретация корректна, и три индивидуума, L1-L3, составляющие маленький кластер, действительно имеют гибридное происхождение, это указывает на существование полового размножения у *A. vaga*. Однако при отсутствии других процессов, сопряженных с рекомбинацией, половое размножение должно

происходить относительно часто – каждые ~10–100 поколений для того, чтобы обеспечить наблюдаемую скорость распада LD.

Необходимо отметить, что наша оценка необходимой частоты мейоза может быть завышена. У этого может быть несколько возможных причин. Во-первых, если настоящая скорость возникновения мутаций *A. vaga* значительно ниже 10^{-9}–10^{-8}, диапазона, из которого мы исходили в расчетах, оценки c и частоты мейоза должны быть скорректированы в сторону более низких значений. Впрочем, у эукариот сообщения о скоростях возникновения мутаций ниже 10^{-9} встречаются только для одноклеточных видов [195]. Во-вторых, митотическая рекомбинация и генная конверсия тоже могут вносить вклад в распад LD. Если это так, то наша оценка частоты мейоза является завышенной, поскольку она получена на основе расчетов, проведенных в предположении о том, что единственной причиной распада LD является реципрокная мейотическая рекомбинация. И, наконец, существует возможность того, что у *A. vaga* присутствуют оба механизма – половое размножение и горизонтальный перенос генов. В таком случае оба эти процесса вносили бы вклад в распад LD и наша оценка частоты мейоза опять оказалась бы завышенной.

В заключение мы должны сказать, что механизмы обмена генетическим материалом и рекомбинации у бделлоидной коловратки *A. vaga* остаются неясными и представляют большой интерес для последующих исследований. Однако анализ данных по популяционной изменчивости *A. vaga*, представленный в этой главе, содержит серьезные доказательства того, что эти процессы регулярно происходят у этого вида бделлоидных коловраток.

Заключение

В данной работе был проведен поиск сигнала отрицательного отбора и рекомбинации на различных геномных данных. Анализ ортологичных интронов в далеких парах видов выявил сигнал отрицательного отбора, продолжающего действовать на имеющие общее происхождение некодирующие участки генома после того, как их последовательности разошлись до неузнаваемости. Так, интроны, несущие сегмент значимого локального сходства или регуляторный элемент в одной паре видов, с повышенной частотой также несут сегмент локального сходства во второй паре видов, несмотря на то, что выравнивание между видами из разных пар отсутствует. Кроме того, последовательность интронов, сохранившихся в одной паре видов, более консервативна во второй паре видов, несмотря на отсутствие сквозного выравнивания между интронами из разных пар видов. Мы предполагаем, что эти результаты могут быть объяснены сохранением предковой функции разошедшихся до неузнаваемости некодирующих последовательностей. Анализ, выполненный в данной работе, проводился на интронах, поскольку общее происхождение интронов может быть установлено даже на больших филогенетических расстояниях на основании выравнивания фланкирующих экзонов. Вероятно, что явление сохранения функции без сохранения сходства последовательностей характерно не только для интронов, но и для межгенных областей генома. Изучение возможного продолжения действия отрицательного отбора на разошедшиеся до неузнаваемости ортологичные межгенные участки может стать направлением для дальнейших исследований, однако установление ортологических соответствий для межгенных участков на значительных филогенетических расстояниях требует разработки отдельного подхода.

С помощью анализа распределения мутационной нагрузки в популяциях *D. melanogaster* нами была выявлена подпись синергического эпистатического отбора, действующего на аллели, вызывающие потерю функции гена, у *D. melanogaster*. Ожидается, что в случае синергических эпистатических взаимодействий дисперсия распределения мутационной нагрузки должна быть ниже аддитивной дисперсии. Такой сигнал был выявлен для аллелей, вызывающих потерю функции гена, в двух популяциях *D. melanogaster*, а также для подмножества несинонимических аллелей, попадающих в гены, находящиеся под сильным давлением отрицательного отбора, в замбийской популяции *D. melanogaster*. Свидетельство о существовании распространенных синергических эпистатических взаимодействий между вредными мутациями является важным как с точки зрения возможного объяснения парадокса мутационного груза, так и с точки зрения поиска причин преобладания полового размножения среди эукариот. Действительно, как показано в нескольких теоретических работах, в случае

существования синергического эпистаза мутационный груз половой популяции может быть значительно ниже, чем в случае независимого влияния мутаций на приспособленность [10–12]. При этом, если скорость вредных мутаций на геном на поколение достаточно высока, в присутствии синергического эпистаза половое размножение может получать преимущество перед бесполым [11,12].

Несмотря на то, что точные причины преобладания полового размножения остаются неизвестными, по всей видимости, оно дает существенные преимущества, поскольку отказ от полового размножения обычно приводит к вымиранию. Это делает особенно интересным изучение возможных механизмов, с помощью которых немногочисленные древние группы бесполых видов противостоят долгосрочным последствиям отказа от полового размножения. В данной работе мы провели анализ популяционной изменчивости для вида *A. vaga*, принадлежащего к группе бделлоидных коловраток, которых в течение длительного времени рассматривали как наиболее яркий пример древнего бесполого таксона. Однако сравнение геномов разных особей *A. vaga*, проведенное в данной работе, выявило сигнал рекомбинации и обмена генетическим материалом. Распад неравновесия по сцеплению с увеличением расстояния между полиморфными сайтами указывает на существование рекомбинации у *A. vaga*. При этом результаты модифицированного четырехгаметного теста позволяют заключить, что распад неравновесия по сцеплению у *A. vaga* не может быть объяснен исключительно за счет генной конверсии. Результаты, полученные при анализе отдельных полиморфных сайтов, дают основания полагать, что в популяции бделлоидной коловратки *A. vaga* происходит обмен генетическим материалом. В частности, на обмен генетическим материалом у *A. vaga* указывает избыток над мутационным ожиданием трехаллельных сайтов, представленных всеми тремя гетерозиготными генотипами. Свидетельства в пользу обмена генетическим материалом у *A. vaga* были получены и при анализе филогений гаплотипов разных индивидуумов. Распределение гаплотипов в филогениях, построенных одновременно для особей из большого и маленького кластеров, согласуется с тем, что ожидалось бы в том случае, если бы три особи маленького кластера имели гибридное происхождение. Если наша интерпретация верна, то половое размножение представляется более вероятным объяснением обмена генетическим материалом в популяции *A. vaga*, чем горизонтальный перенос генов. Тем не менее определение механизма обмена генетическим материалом у бделлоидных коловраток безусловно заслуживает дальнейших исследований. Вне зависимости от способа, с помощью которого у бделлоидных коловраток происходит обмен генетическим материалом, по всей видимости, бделлоидные коловратки не относятся к древним группам бесполых видов: структура популяционной изменчивости у *A. vaga* несовместима с исключительно клональным наследованием ДНК. Свидетельства в пользу рекомбинации у бделлоидных коловраток,

которых рассматривали как наиболее удивительный пример древней группы бесполых видов, подчеркивают важность рекомбинации для долгосрочного эволюционного успеха.

Выводы

1. В ортологичных интронах из филогенетически далеких пар видов присутствует сигнал, свидетельствующий о продолжении действия отрицательного отбора на имеющие общее происхождение, но разошедшиеся до неузнаваемости, некодирующие последовательности. Возможное объяснение этого явления заключается в том, что предковая функция некодирующего участка генома, вероятно, может сохраняться дольше, чем сходство последовательностей.

2. Распределение мутационной нагрузки для аллелей *D. melanogaster*, вызывающих потерю функции гена, характеризуется понижением дисперсии относительно аддитивной дисперсии. Кроме того, понижение дисперсии распределения мутационной нагрузки относительно аддитивной дисперсии наблюдается и для подмножества несинонимических аллелей, попадающих в гены, находящиеся под сильным давлением отрицательного отбора, в замбийской популяции *D. melanogaster*. Данная картина соответствует тому, что ожидается в случае существования у *D. melanogaster* синергических эпистатических взаимодействий между вредными аллелями.

3. Данные по внутрипопуляционной изменчивости бделлоидной коловратки вида *A. vaga* указывают на существование рекомбинации и обмена генетическим материалом в популяции этого вида. Подписи рекомбинации, выявленные при анализе геномов *A. vaga*, не могут быть объяснены исключительно за счет генной конверсии и, вероятно, являются результатом реципрокной рекомбинации. Структура изменчивости у *A. vaga* как на уровне отдельных полиморфных сайтов, так и на уровне геномных сегментов несовместима с исключительно клональным способом размножения. Распределение гаплотипов особей *A. vaga* из большого и маленького кластеров в филогениях, построенных для разных геномных локусов, наиболее экономно объясняется в рамках сценария, в котором обмен генетическим материалом у *A. vaga* происходит в результате полового размножения.

Благодарности

Я безмерно признательна своим учителям Г. А. Базыкину и А. С. Кондрашову за те годы, которые мне посчастливилось работать вместе с ними, за огромную помощь и поддержку при выполнении работы, за замечательную научную среду, которую они создают вокруг себя. Я благодарна Я. Р. Галимову и Е. А. Мнацакановой, выполнившим значительную часть экспериментальной работы с клональными культурами *A. vaga*. Я признательна своим соавторам – Е. А. Мнацакановой, Я. Р. Галимову, Т. В. Неретиной, Е. С. Герасимову, С. А. Науменко, С. Г. Озеровой, А. О. Залевскому, И. А. Юшеновой, Ф. Родригезу, И. А. Архиповой, А. А. Пенину, М. Д. Логачевой, М. Сохаил, Ш. Р. Сюняеву. Я благодарю А. С. Микаеляна и И. Ю. Баклушинскую за предоставление некоторых реагентов и доступа к оборудованию. Я благодарю И. А. Архипову и Ф. Родригеза за предоставление данных PacBio и обсуждение результатов. Я признательна А. С. Касьянову за консультации и советы по фазированию гаплотипов и Я. Ю. Сафоновой за помощь с подбором параметров SPAdes. Я благодарна за помощь в установке программ А. О. Залевскому, Е. С. Герасимову и С. А. Науменко. Кроме того, я благодарна А. О. Залевскому за помощь в оформлении Рисунка 4.19. Я благодарна профессору М. Мезельсону за обсуждение результатов, представленных в Главе 4. Я благодарю своих друзей и коллег Е. Р. Набиеву, Н. М. Гомберг, С. А. Науменко, В. Е. Раменского, Е. В. Лёушкина, С. К. Гарушянц, М. А. Андрианову за поддержку и обсуждение работы. Я благодарю за поддержку свою семью, особенно Т. Л. Вахрушеву.

Выполнение работы было частично поддержано фондом РФФИ (проект 16-34-01303 мол_а «Изучение полногеномного полиморфизма в популяции бделлоидных коловраток *Adineta vaga*», 2016–2017). Кроме того, я признательна за финансовую поддержку Обществу молекулярной биологии и эволюции.

Список сокращений и условных обозначений

ДИ – доверительный интервал

кб – тысяча пар нуклеотидов

Мб – миллион пар нуклеотидов

п.н. – пар нуклеотидов

DEG – Database of Essential Genes

DGRP – Drosophila Genetic Reference Panel

DPGP3 – Drosophila Population Genomics Project

GABP – growth associated binding protein

IBS – идентичный по состоянию, от англ. identical by state

IQR – межквартильный диапазон, от англ. interquartile range

LD – неравновесие по сцеплению, от англ. linkage disequilibrium

LUCA – последний универсальный общий предок, от англ. last universal common ancestor

NA – данные отсутствуют или недоступны, от англ. not available

SNP – однонуклеотидный полиморфизм, от англ. single nucleotide polymorphism

SOWH тест – тест Swofford–Olsen–Waddell–Hillis

SRF – serum response factor

vs – versus

Список литературы

1. Cooper M.A. et al. Population genetics provides evidence for recombination in Giardia // Curr. Biol. 2007. Vol. 17, № 22. P. 1984–1988.
2. Signorovitch A.Y., Dellaporta S.L., Buss L.W. Molecular signatures for sex in the Placozoa // Proc Natl Acad Sci U S A. 2005. Vol. 102, № 43. P. 15518–15522.
3. Dermitzakis E.T. et al. Evolutionary discrimination of mammalian conserved non-genic sequences (CNGs) // Science. 2003. Vol. 302, № 5647. P. 1033–1035.
4. Lowe C.B. et al. Three periods of regulatory innovation during vertebrate evolution // Science. 2011. Vol. 333, № 6045. P. 1019–1024.
5. Bergman C.M., Kreitman M. Analysis of Conserved Noncoding DNA in Drosophila Reveals Similar Constraints in Intergenic and Intronic Sequences // Genome Res. 2001. Vol. 11, № 8. P. 1335–1345.
6. Visel A. et al. Ultraconservation identifies a small subset of extremely constrained developmental enhancers // Nature Genetics. 2008. Vol. 40, № 2. P. 158–160.
7. Xie X. et al. Systematic discovery of regulatory motifs in conserved regions of the human genome, including thousands of CTCF insulator sites // PNAS. 2007. Vol. 104, № 17. P. 7145–7150.
8. Taher L. et al. Genome-wide identification of conserved regulatory function in diverged sequences // Genome Res. 2011. Vol. 21, № 7. P. 1139–1149.
9. Fisher S. et al. Conservation of RET regulatory function from human to zebrafish without sequence similarity // Science. 2006. Vol. 312, № 5771. P. 276–279.
10. Kimura M., Maruyama T. The Mutational Load with Epistatic Gene Interactions in Fitness // Genetics. 1966. Vol. 54, № 6. P. 1337–1351.
11. Kondrashov A.S. Selection against harmful mutations in large sexual and asexual populations // Genet. Res. 1982. Vol. 40, № 3. P. 325–332.
12. Charlesworth B. Mutation-selection balance and the evolutionary advantage of sex and recombination // Genet. Res. 1990. Vol. 55, № 3. P. 199–221.
13. Kondrashov A.S. Deleterious mutations and the evolution of sexual reproduction // Nature. 1988. Vol. 336, № 6198. P. 435–440.
14. Bank C. et al. A Systematic Survey of an Intragenic Epistatic Landscape // Mol Biol Evol. 2015. Vol. 32, № 1. P. 229–238.
15. Puchta O. et al. Network of epistatic interactions within a yeast snoRNA // Science. 2016. Vol. 352, № 6287. P. 840–844.
16. Muller H.J. The relation of recombination to mutational advance // Mutat. Res. 1964. Vol. 106. P. 2–9.
17. Haigh J. The accumulation of deleterious genes in a population--Muller's Ratchet // Theor Popul Biol. 1978. Vol. 14, № 2. P. 251–267.
18. Fisher R.A. The Genetical Theory of Natural Selection. Oxford: Oxford University Press, 1930.
19. Muller H.J. Some Genetic Aspects of Sex // The American Naturalist. 1932. Vol. 66, № 703. P. 118–138.
20. Charlesworth B., Morgan M.T., Charlesworth D. The Effect of Deleterious Mutations on Neutral Molecular Variation // Genetics. 1993. Vol. 134, № 4. P. 1289–1303.
21. Rice W.R., Chippindale A.K. Sexual recombination and the power of natural selection // Science. 2001. Vol. 294, № 5542. P. 555–559.
22. Kondrashov A.S. Classification of hypotheses on the advantage of amphimixis // J. Hered. 1993. Vol. 84, № 5. P. 372–387.
23. Stearns S.C. The masterpiece of nature: The evolution and genetics of sexuality // Evolution and Human Behavior. 1984. Vol. 5, № 1. P. 73–75.
24. Birky C.W. Positively negative evidence for asexuality // J. Hered. 2010. Vol. 101 Suppl 1. P. S42-45.

25. Flot J.-F. et al. Genomic evidence for ameiotic evolution in the bdelloid rotifer Adineta vaga // Nature. 2013. Vol. 500, № 7463. P. 453–457.
26. Nowell R.W. et al. Comparative genomics of bdelloid rotifers: Insights from desiccating and nondesiccating species // PLOS Biology. 2018. Vol. 16, № 4. P. e2004830.
27. Simion P. et al. Chromosome-level genome assembly reveals homologous chromosomes and recombination in asexual rotifer Adineta vaga // Sci Adv. 2021. Vol. 7, № 41. P. eabg4216.
28. Signorovitch A. et al. Allele Sharing and Evidence for Sexuality in a Mitochondrial Clade of Bdelloid Rotifers // Genetics. 2015. Vol. 200, № 2. P. 581–590.
29. Li L., Stoeckert C.J., Roos D.S. OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes // Genome Res. 2003. Vol. 13, № 9. P. 2178–2189.
30. Clark A.G. The Search for Meaning in Noncoding DNA // Genome Res. 2001. Vol. 11, № 8. P. 1319–1320.
31. Casillas S., Barbadilla A., Bergman C.M. Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila // Mol Biol Evol. 2007. Vol. 24, № 10. P. 2222–2234.
32. Jareborg N., Birney E., Durbin R. Comparative Analysis of Noncoding Regions of 77 Orthologous Mouse and Human Gene Pairs // Genome Res. 1999. Vol. 9, № 9. P. 815–824.
33. Shabalina S.A., Kondrashov A.S. Pattern of selective constraint in C. elegans and C. briggsae genomes // Genet. Res. 1999. Vol. 74, № 1. P. 23–30.
34. Siepel A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes // Genome Res. 2005. Vol. 15, № 8. P. 1034–1050.
35. Dermitzakis E.T., Clark A.G. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover // Mol. Biol. Evol. 2002. Vol. 19, № 7. P. 1114–1121.
36. Murzin A.G., Bateman A. Distant homology recognition using structural classification of proteins // Proteins. 1997. Vol. Suppl 1. P. 105–112.
37. Schuster P. et al. From sequences to shapes and back: a case study in RNA secondary structures // Proc. Biol. Sci. 1994. Vol. 255, № 1344. P. 279–284.
38. McGaughey D.M. et al. Metrics of sequence constraint overlook regulatory sequences in an exhaustive analysis at phox2b // Genome Res. 2008. Vol. 18, № 2. P. 252–260.
39. Vavouri T. et al. Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans // Genome Biology. 2007. Vol. 8, № 2. P. R15.
40. Crow J.F. Genetic Loads and the Cost of Natural Selection // Mathematical Topics in Population Genetics / ed. Kojima K. Berlin, Heidelberg: Springer, 1970. P. 128–177.
41. Eyre-Walker A., Keightley P.D. High genomic deleterious mutation rates in hominids // Nature. 1999. Vol. 397, № 6717. P. 344–347.
42. Muller H.J. Our load of mutations // Am J Hum Genet. 1950. Vol. 2, № 2. P. 111–176.
43. Kondrashov A.S., Crow J.F. A molecular approach to estimating the human deleterious mutation rate // Hum. Mutat. 1993. Vol. 2, № 3. P. 229–234.
44. Crow J.F. Some possibilities for measuring selection intensities in man // Hum Biol. 1958. Vol. 30, № 1. P. 1–13.
45. Maynard Smith J. The evolution of sex. New York: Cambridge University Press, 1978.
46. Visser J.A.G.M. de, Elena S.F. The evolution of sex: empirical insights into the roles of epistasis and drift // Nat Rev Genet. 2007. Vol. 8, № 2. P. 139–149.
47. Felsenstein J. Sex and the evolution of recombination. // The Evolution of Sex: An Examination of Current Ideas. R.E. Michod and B.R. Levin eds. P. 74–86.
48. Bernstein H. et al. Genetic damage, mutation, and the evolution of sex // Science. 1985. Vol. 229, № 4719. P. 1277–1281.
49. Bulmer M.G. The mathematical theory of quantitative genetics. Oxford: Oxford University Press, 1980. 276 p.
50. Redfield R.J. Evolution of bacterial transformation: is sex with dead cells ever better than no sex at all? // Genetics. 1988. Vol. 119, № 1. P. 213–221.

51. Besenbacher S. et al. Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios // Nature Communications. 2015. Vol. 6. P. 5969.
52. Rands C.M. et al. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage // PLoS Genet. 2014. Vol. 10, № 7. P. e1004525.
53. Maynard Smith J. Evolution: contemplating life without sex // Nature. 1986. Vol. 324, № 6095. P. 300–301.
54. Williams G.C. Sex and evolution. Princeton, New Jersey: Princeton University Press, 1975.
55. Bell G. The Masterpiece of Nature: The Evolution and Genetics of Sexuality. Berkeley, California: University of California Press, 1982. 635 p.
56. White M.J.D. Modes of speciation. San Francisco, California: W. H. Freeman, 1978.
57. Martens K., Rossetti G., Horne D.J. How ancient are ancient asexuals? // Proc. Biol. Sci. 2003. Vol. 270, № 1516. P. 723–729.
58. Schön I., Martens K. No slave to sex // Proc. Biol. Sci. 2003. Vol. 270, № 1517. P. 827–833.
59. Maraun M. et al. Radiation in sexual and parthenogenetic oribatid mites (Oribatida, Acari) as indicated by genetic divergence of closely related species // Exp. Appl. Acarol. 2003. Vol. 29, № 3–4. P. 265–277.
60. Maraun M. et al. Molecular phylogeny of oribatid mites (Oribatida, Acari): evidence for multiple radiations of parthenogenetic lineages // Exp. Appl. Acarol. 2004. Vol. 33, № 3. P. 183–201.
61. Poinar G.O., Ricci C. Bdelloid rotifers in Dominican amber: Evidence for parthenogenetic continuity // Experientia. 1992. Vol. 48, № 4. P. 408–410.
62. Tang C.Q. et al. Sexual species are separated by larger genetic gaps than asexual species in rotifers // Evolution. 2014. Vol. 68, № 10. P. 2901–2916.
63. Mark Welch J.L., Mark Welch D.B., Meselson M. Cytogenetic evidence for asexual evolution of bdelloid rotifers // Proc. Natl. Acad. Sci. U.S.A. 2004. Vol. 101, № 6. P. 1618–1621.
64. Judson O.P., Normark B.B. Ancient asexual scandals // Trends Ecol. Evol. (Amst.). 1996. Vol. 11, № 2. P. 41–46.
65. Smith R.J., Kamiya T., Horne D.J. Living males of the 'ancient asexual' Darwinulidae (Ostracoda: Crustacea) // Proc Biol Sci. 2006. Vol. 273, № 1593. P. 1569–1578.
66. Heethoff M. et al. Parthenogenesis in Oribatid Mites (Acari, Oribatida): Evolution Without Sex // Lost Sex: The Evolutionary Biology of Parthenogenesis. 2009. P. 241–257.
67. Gladyshev E.A., Meselson M., Arkhipova I.R. Massive horizontal gene transfer in bdelloid rotifers // Science. 2008. Vol. 320, № 5880. P. 1210–1213.
68. Schurko A.M., Neiman M., Logsdon J.M. Signs of sex: what we know and how we know it // Trends in Ecology & Evolution. 2009. Vol. 24, № 4. P. 208–217.
69. Balloux F., Lehmann L., de Meeûs T. The population genetics of clonal and partially clonal diploids. // Genetics. 2003. Vol. 164, № 4. P. 1635–1644.
70. Rosendahl S., Taylor J.W. Development of multiple genetic markers for studies of genetic variation in arbuscular mycorrhizal fungi using AFLP^{TM} // Molecular Ecology. 1997. Vol. 6, № 9. P. 821–829.
71. Stukenbrock E.H., Rosendahl S. Clonal diversity and population genetic structure of arbuscular mycorrhizal fungi (Glomus spp.) studied by multilocus genotyping of single spores // Mol. Ecol. 2005. Vol. 14, № 3. P. 743–752.
72. Ardlie K. et al. Lower-Than-Expected Linkage Disequilibrium between Tightly Linked Markers in Humans Suggests a Role for Gene Conversion // The American Journal of Human Genetics. 2001. Vol. 69, № 3. P. 582–589.
73. Przeworski M., Wall J.D. Why is there so little intragenic linkage disequilibrium in humans? // Genet. Res. 2001. Vol. 77, № 2. P. 143–151.
74. Lee P.S. et al. A fine-structure map of spontaneous mitotic crossovers in the yeast Saccharomyces cerevisiae // PLoS Genet. 2009. Vol. 5, № 3. P. e1000410.

75. Yim E. et al. High-resolution mapping of two types of spontaneous mitotic gene conversion events in Saccharomyces cerevisiae // Genetics. 2014. Vol. 198, № 1. P. 181–192.
76. Mark Welch D.B., Meselson M. Evidence for the Evolution of Bdelloid Rotifers Without Sexual Reproduction or Genetic Exchange // Science. 2000. Vol. 288, № 5469. P. 1211–1215.
77. Schaefer I. et al. No evidence for the “Meselson effect” in parthenogenetic oribatid mites (Oribatida, Acari) // J. Evol. Biol. 2006. Vol. 19, № 1. P. 184–193.
78. Butlin R.K. Virgin rotifers // Trends in Ecology & Evolution. 2000. Vol. 15, № 10. P. 389–390.
79. Weir W. et al. Population genomics reveals the origin and asexual evolution of human infective trypanosomes // eLife Sciences. 2016. Vol. 5. P. e11473.
80. Schwander T., Henry L., Crespi B.J. Molecular Evidence for Ancient Asexuality in Timema Stick Insects // Current Biology. 2011. Vol. 21, № 13. P. 1129–1134.
81. Warren W.C. et al. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly // Nature Ecology & Evolution. 2018. Vol. 2, № 4. P. 669–679.
82. Martens K., Schon I. Ancient asexuals: darwinulids not exposed // Nature. 2008. Vol. 453, № 7195. P. 587–587.
83. Lost Sex: The Evolutionary Biology of Parthenogenesis / ed. Schön I., Martens K., Dijk P. van. Springer Netherlands, 2009.
84. Norton R.A., Palmer S.C. The distribution, mechanisms and evolutionary significance of parthenogenesis in oribatid mites // The Acari: Reproduction, development and life-history strategies / ed. Schuster R., Murphy P.W. Dordrecht: Springer Netherlands, 1991. P. 107–136.
85. Taberly G. Recherches sur la parthénogenèse thélytoque de deux espèces d’acariens oribatides: Trypochthonius tectorum (Berlese) et Platynothrus peltifer (Koch). IV. Observation sur les males ataviques. // Acarologia. 1988. Vol. 29. P. 95–107.
86. Hsu W.S. Oogenesis in the Bdelloidea rotifer Philodina roseola // Cellule. 1956. Vol. 57. P. 283–296.
87. Segers H. Annotated checklist of the rotifers (Phylum Rotifera), with notes on nomenclature, taxonomy and distribution // Zootaxa. 2007. Vol. 1564, № 1. P. 1–104.
88. Wesenberg-Lund C. Contributions to the biology of the rotifers, part II: the periodicity and sexual periods. Copenhagen, Denmark: A.F.Host and Son, 1930.
89. Nogrady T., Wallace R., Snell T. Rotifera: biology, ecology and systematics. Guides to the identification of the microinvertebrates of the continental waters of the world. The Hague, The Netherlands: SPB Academic Publishing, 1993.
90. Debortoli N. et al. Genetic Exchange among Bdelloid Rotifers Is More Likely Due to Horizontal Gene Transfer Than to Meiotic Sex // Curr. Biol. 2016. Vol. 26, № 6. P. 723–732.
91. Wilson C.G., Nowell R.W., Barraclough T.G. Cross-Contamination Explains “Inter and Intraspecific Horizontal Genetic Transfers” between Asexual Bdelloid Rotifers // Curr. Biol. 2018. Vol. 28, № 15. P. 2436-2444.e14.
92. Signorovitch A. et al. Evidence for meiotic sex in bdelloid rotifers // Current Biology. 2016. Vol. 26, № 16. P. R754–R755.
93. Iyer L.M. et al. Evolutionary history and higher order classification of AAA+ ATPases // J. Struct. Biol. 2004. Vol. 146, № 1–2. P. 11–31.
94. Jaillon O. et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype // Nature. 2004. Vol. 431, № 7011. P. 946–957.
95. Odronitz F., Becker S., Kollmar M. Reconstructing the phylogeny of 21 completely sequenced arthropod species based on their motor proteins // BMC Genomics. 2009. Vol. 10. P. 173.
96. Heger A., Ponting C.P. Evolutionary rate analyses of orthologs and paralogs from 12 Drosophila genomes // Genome Res. 2007. Vol. 17, № 12. P. 1837–1849.
97. Ostlund G. et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis // Nucleic Acids Res. 2010. Vol. 38, № Database issue. P. D196-203.
98. Lander E.S. et al. Initial sequencing and analysis of the human genome // Nature. 2001. Vol. 409, № 6822. P. 860–921.

99. Wheeler D.L. et al. Database resources of the National Center for Biotechnology Information // Nucleic Acids Res. 2008. Vol. 36, № Database issue. P. D13-21.
100. Mouse Genome Sequencing Consortium et al. Initial sequencing and comparative analysis of the mouse genome // Nature. 2002. Vol. 420, № 6915. P. 520–562.
101. Dehal P. et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins // Science. 2002. Vol. 298, № 5601. P. 2157–2167.
102. Aparicio S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes // Science. 2002. Vol. 297, № 5585. P. 1301–1310.
103. Nene V. et al. Genome sequence of Aedes aegypti, a major arbovirus vector // Science. 2007. Vol. 316, № 5832. P. 1718–1723.
104. Adams M.D. et al. The genome sequence of Drosophila melanogaster // Science. 2000. Vol. 287, № 5461. P. 2185–2195.
105. Drosophila 12 Genomes Consortium et al. Evolution of genes and genomes on the Drosophila phylogeny // Nature. 2007. Vol. 450, № 7167. P. 203–218.
106. Arensburger P. et al. Sequencing of Culex quinquefasciatus establishes a platform for mosquito comparative genomics // Science. 2010. Vol. 330, № 6000. P. 86–88.
107. Honeybee Genome Sequencing Consortium. Insights into social insects from the genome of the honeybee Apis mellifera // Nature. 2006. Vol. 443, № 7114. P. 931–949.
108. Ensembl Genomes: extending Ensembl across the taxonomic space. - PubMed - NCBI [Electronic resource]. URL: https://www.ncbi.nlm.nih.gov/pubmed/19884133/ (accessed: 15.03.2019).
109. Sayers E.W. et al. Database resources of the National Center for Biotechnology Information // Nucleic Acids Res. 2012. Vol. 40, № Database issue. P. D13-25.
110. Lawson D. et al. VectorBase: a data resource for invertebrate vector genomics // Nucleic Acids Res. 2009. Vol. 37, № Database issue. P. D583-587.
111. Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity // BMC Bioinformatics. 2004. Vol. 5. P. 113.
112. Haddrill P.R. et al. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content // Genome Biol. 2005. Vol. 6, № 8. P. R67.
113. Roca X., Sachidanandam R., Krainer A.R. Determinants of the inherent strength of human 5’ splice sites // RNA. 2005. Vol. 11, № 5. P. 683–698.
114. Altschul S.F. et al. Basic local alignment search tool // J. Mol. Biol. 1990. Vol. 215, № 3. P. 403–410.
115. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project // Science. 2004. Vol. 306, № 5696. P. 636–640.
116. Ernst J. et al. Mapping and analysis of chromatin state dynamics in nine human cell types // Nature. 2011. Vol. 473, № 7345. P. 43–49.
117. Kharchenko P.V. et al. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster // Nature. 2011. Vol. 471, № 7339. P. 480–485.
118. modENCODE Consortium et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE // Science. 2010. Vol. 330, № 6012. P. 1787–1797.
119. Woolfe A. et al. Highly conserved non-coding sequences are associated with vertebrate development // PLoS Biol. 2005. Vol. 3, № 1. P. e7.
120. Putnam N.H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization // Science. 2007. Vol. 317, № 5834. P. 86–94.
121. Heintzman N.D. et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression // Nature. 2009. Vol. 459, № 7243. P. 108–112.
122. Yang S. et al. Functionally conserved enhancers with divergent sequences in distant vertebrates // BMC Genomics. 2015. Vol. 16, № 1. P. 882.
123. Lesecque Y., Keightley P.D., Eyre-Walker A. A Resolution of the Mutation Load Paradox in Humans // Genetics. 2012. Vol. 191, № 4. P. 1321–1330.

124. Kondrashov A.S. Dynamics of unconditionally deleterious mutations: Gaussian approximation and soft selection // Genet. Res. 1995. Vol. 65, № 2. P. 113–121.
125. Lack J.B. et al. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population // Genetics. 2015. Vol. 199, № 4. P. 1229–1241.
126. Mackay T.F.C. et al. The Drosophila melanogaster Genetic Reference Panel // Nature. 2012. Vol. 482, № 7384. P. 173–178.
127. Huang W. et al. Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines // Genome Res. 2014. Vol. 24, № 7. P. 1193–1208.
128. Pool J.E. et al. Population Genomics of Sub-Saharan Drosophila melanogaster : African Diversity and Non-African Admixture // PLOS Genet. 2012. Vol. 8, № 12. P. e1003080.
129. Duchen P. et al. Demographic Inference Reveals African and European Admixture in the North American Drosophila melanogaster Population // Genetics. 2013. Vol. 193, № 1. P. 291–301.
130. Langley C.H. et al. Circumventing heterozygosity: sequencing the amplified genome of a single haploid Drosophila melanogaster embryo // Genetics. 2011. Vol. 188, № 2. P. 239–246.
131. Kent W.J. et al. The Human Genome Browser at UCSC // Genome Res. 2002. Vol. 12, № 6. P. 996–1006.
132. Crosby M.A. et al. FlyBase: genomes by the dozen // Nucleic Acids Res. 2007. Vol. 35, № Database issue. P. D486–D491.
133. McBride C.S. Rapid evolution of smell and taste receptor genes during host specialization in Drosophila sechellia // Proc. Natl. Acad. Sci. U.S.A. 2007. Vol. 104, № 12. P. 4996–5001.
134. Gardiner A. et al. Drosophila chemoreceptor gene evolution: selection, specialization and genome size // Mol. Ecol. 2008. Vol. 17, № 7. P. 1648–1657.
135. Lee Y.C.G., Reinhardt J.A. Widespread polymorphism in the positions of stop codons in Drosophila melanogaster // Genome Biol Evol. 2012. Vol. 4, № 4. P. 533–549.
136. McBride C.S., Arguello J.R. Five Drosophila Genomes Reveal Nonneutral Evolution and the Signature of Host Specialization in the Chemoreceptor Superfamily // Genetics. 2007. Vol. 177, № 3. P. 1395–1416.
137. Corbett-Detig R.B., Hartl D.L. Population Genomics of Inversion Polymorphisms in Drosophila melanogaster // PLOS Genetics. 2012. Vol. 8, № 12. P. e1003056.
138. Yang R.C. Zygotic associations and multilocus statistics in a nonequilibrium diploid population // Genetics. 2000. Vol. 155, № 3. P. 1449–1458.
139. Cargill M. et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes // Nat. Genet. 1999. Vol. 22, № 3. P. 231–238.
140. Halushka M.K. et al. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis // Nat. Genet. 1999. Vol. 22, № 3. P. 239–247.
141. MacArthur D.G. et al. A systematic survey of loss-of-function variants in human protein-coding genes // Science. 2012. Vol. 335, № 6070. P. 823–828.
142. Hentze M.W., Kulozik A.E. A perfect message: RNA surveillance and nonsense-mediated decay // Cell. 1999. Vol. 96, № 3. P. 307–310.
143. Baker K.E., Parker R. Nonsense-mediated mRNA decay: terminating erroneous gene expression // Curr. Opin. Cell Biol. 2004. Vol. 16, № 3. P. 293–299.
144. Anna A., Monika G. Splicing mutations in human genetic disorders: examples, detection, and confirmation // J Appl Genet. 2018. Vol. 59, № 3. P. 253–268.
145. Burset M., Seledtsov I.A., Solovyev V.V. Analysis of canonical and non-canonical splice sites in mammalian genomes // Nucleic Acids Res. 2000. Vol. 28, № 21. P. 4364–4375.
146. Thanaraj T.A., Clark F. Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions // Nucleic Acids Res. 2001. Vol. 29, № 12. P. 2581–2593.
147. Luo H. et al. DEG 10, an update of the database of essential genes that includes both protein-coding genes and noncoding genomic elements // Nucleic Acids Res. 2014. Vol. 42, № Database issue. P. D574-580.

148. Larracuente A.M. et al. Evolution of protein-coding genes in Drosophila // Trends Genet. 2008. Vol. 24, № 3. P. 114–123.
149. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood // Mol Biol Evol. 2007. Vol. 24, № 8. P. 1586–1591.
150. Eyre-Walker A., Woolfit M., Phelps T. The Distribution of Fitness Effects of New Deleterious Amino Acid Mutations in Humans // Genetics. 2006. Vol. 173, № 2. P. 891–900.
151. Kirkpatrick M., Barton N. Chromosome Inversions, Local Adaptation and Speciation // Genetics. 2006. Vol. 173, № 1. P. 419–434.
152. Sohail M. et al. Negative selection in humans and fruit flies involves synergistic epistasis // Science. 2017. Vol. 356, № 6337. P. 539–542.
153. Kiezun A. et al. Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency // PLoS Genet. 2013. Vol. 9, № 2. P. e1003301.
154. Bell G. The Masterpiece of Nature. The Evolution and Genetics of Sexuality. Berkeley, California: University of California Press, 1982.
155. Schwander T., Henry L., Crespi B.J. Molecular evidence for ancient asexuality in timema stick insects // Curr. Biol. 2011. Vol. 21, № 13. P. 1129–1134.
156. Kutikova L.A. The Bdelloid rotifers of the fauna of Russia. KMK Scientific Press Ltd, Moscow. [Electronic resource]. 2005.
157. Bolger A.M., Lohse M., Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data // Bioinformatics. 2014. Vol. 30, № 15. P. 2114–2120.
158. Fontaneto D. et al. Independently evolving species in asexual bdelloid rotifers // PLoS Biol. 2007. Vol. 5, № 4. P. e87.
159. Bankevich A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing // J. Comput. Biol. 2012. Vol. 19, № 5. P. 455–477.
160. Gurevich A. et al. QUAST: quality assessment tool for genome assemblies // Bioinformatics. 2013. Vol. 29, № 8. P. 1072–1075.
161. Kumar S. et al. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots // Front Genet. 2013. Vol. 4. P. 237.
162. Blanchette M. et al. Aligning multiple genomic sequences with the threaded blockset aligner // Genome Res. 2004. Vol. 14, № 4. P. 708–715.
163. Schwartz S. et al. Human-mouse alignments with BLASTZ // Genome Res. 2003. Vol. 13, № 1. P. 103–107.
164. Stanke M., Waack S. Gene prediction with a hidden Markov model and a new intron submodel // Bioinformatics. 2003. Vol. 19 Suppl 2. P. ii215-225.
165. Lomsadze A. et al. Gene identification in novel eukaryotic genomes by self-training algorithm // Nucleic Acids Res. 2005. Vol. 33, № 20. P. 6494–6506.
166. Wang Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity // Nucleic Acids Res. 2012. Vol. 40, № 7. P. e49.
167. Langmead B., Salzberg S.L. Fast gapped-read alignment with Bowtie 2 // Nature Methods. 2012. Vol. 9, № 4. P. 357–359.
168. Li H. et al. The Sequence Alignment/Map format and SAMtools // Bioinformatics. 2009. Vol. 25, № 16. P. 2078–2079.
169. Danecek P. et al. The variant call format and VCFtools // Bioinformatics. 2011. Vol. 27, № 15. P. 2156–2158.
170. Quinlan A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis // Curr Protoc Bioinformatics. 2014. Vol. 47. P. 11.12.1-34.
171. Cingolani P. et al. Using Drosophila melanogaster as a Model for Genotoxic Chemical Mutational Studies with a New Program, SnpSift // Front Genet. 2012. Vol. 3. P. 35.
172. Purcell S. et al. PLINK: a tool set for whole-genome association and population-based linkage analyses // Am. J. Hum. Genet. 2007. Vol. 81, № 3. P. 559–575.
173. Edge P., Bafna V., Bansal V. HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies // Genome Res. 2017. Vol. 27, № 5. P. 801–812.

174. Kebschull J.M., Zador A.M. Sources of PCR-induced distortions in high-throughput sequencing data sets // Nucleic Acids Res. 2015. Vol. 43, № 21. P. e143.
175. Lynch M. Estimation of nucleotide diversity, disequilibrium coefficients, and mutation rates from high-coverage genome-sequencing projects // Mol. Biol. Evol. 2008. Vol. 25, № 11. P. 2409–2419.
176. Haubold B., Pfaffelhuber P., Lynch M. mlRho - a program for estimating the population mutation and recombination rates from shotgun-sequenced diploid genomes // Mol. Ecol. 2010. Vol. 19 Suppl 1. P. 277–284.
177. Awadalla P., Eyre-Walker A., Smith J.M. Linkage Disequilibrium and Recombination in Hominid Mitochondrial DNA // Science. 1999. Vol. 286, № 5449. P. 2524–2525.
178. Meunier J., Eyre-Walker A. The Correlation Between Linkage Disequilibrium and Distance: Implications for Recombination in Hominid Mitochondria // Mol Biol Evol. 2001. Vol. 18, № 11. P. 2132–2135.
179. Bruen T.C., Philippe H., Bryant D. A simple and robust statistical test for detecting the presence of recombination // Genetics. 2006. Vol. 172, № 4. P. 2665–2681.
180. McVean G., Awadalla P., Fearnhead P. A coalescent-based method for detecting and estimating recombination from gene sequences // Genetics. 2002. Vol. 160, № 3. P. 1231–1241.
181. Catchen J. et al. Stacks: an analysis tool set for population genomics // Mol. Ecol. 2013. Vol. 22, № 11. P. 3124–3140.
182. Haller B.C., Messer P.W. SLiM 3: Forward Genetic Simulations Beyond the Wright-Fisher Model // Mol. Biol. Evol. 2019. Vol. 36, № 3. P. 632–637.
183. Guindon S., Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood // Syst. Biol. 2003. Vol. 52, № 5. P. 696–704.
184. Huerta-Cepas J., Serra F., Bork P. ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data // Mol. Biol. Evol. 2016. Vol. 33, № 6. P. 1635–1638.
185. Swofford D.L. et al. Phylogenetic Inference. // Molecular Systematics, 2nd Edition. Hillis, D.M., Moritz C. and Mable B.K., editors. Sinauer Associates, Sunderland (MA), 1996. P. 407–514.
186. Church S.H., Ryan J.F., Dunn C.W. Automation and Evaluation of the SOWH Test with SOWHAT // Syst Biol. 2015. Vol. 64, № 6. P. 1048–1058.
187. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies // Bioinformatics. 2014. Vol. 30, № 9. P. 1312–1313.
188. Hill W.G., Weir B.S. Variances and covariances of squared linkage disequilibria in finite populations // Theor Popul Biol. 1988. Vol. 33, № 1. P. 54–78.
189. Marroni F. et al. Nucleotide diversity and linkage disequilibrium in Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene // Tree Genetics & Genomes. 2011. Vol. 7. P. 1011–1023.
190. Sved J.A. Linkage disequilibrium and homozygosity of chromosome segments in finite populations // Theor Popul Biol. 1971. Vol. 2, № 2. P. 125–141.
191. Wakeley J. Using the variance of pairwise differences to estimate the recombination rate // Genet. Res. 1997. Vol. 69, № 1. P. 45–48.
192. Lynch M. et al. Population Genomics of Daphnia pulex // Genetics. 2017. Vol. 206, № 1. P. 315–332.
193. Shendure J., Akey J.M. The origins, determinants, and consequences of human mutations // Science. 2015. Vol. 349, № 6255. P. 1478–1483.
194. Mark Welch J.L., Meselson M. Karyotypes of bdelloid rotifers from three families // Hydrobiologia. 1998. Vol. 387, № 0. P. 403–407.
195. Lynch M. et al. Genetic drift, selection and the evolution of the mutation rate // Nature Reviews Genetics. 2016. Vol. 17, № 11. P. 704–714.
196. Simão F.A. et al. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs // Bioinformatics. 2015. Vol. 31, № 19. P. 3210–3212.
197. Kamvar Z.N., Tabima J.F., Grünwald N.J. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction // PeerJ. PeerJ Inc., 2014. Vol. 2. P. e281.

198. Kamvar Z.N., Brooks J.C., Grünwald N.J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality // Front. Genet. Frontiers, 2015. Vol. 6.
199. Kumar S., Stecher G., Tamura K. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets // Mol Biol Evol. 2016. Vol. 33, № 7. P. 1870–1874.
200. Lasek-Nesselquist E. A Mitogenomic Re-Evaluation of the Bdelloid Phylogeny and Relationships among the Syndermata // PLOS ONE. Public Library of Science, 2012. Vol. 7, № 8. P. e43554.
201. Benson G. Tandem repeats finder: a program to analyze DNA sequences // Nucleic Acids Res. 1999. Vol. 27, № 2. P. 573–580.
202. Hill W.G., Robertson A. Linkage disequilibrium in finite populations // Theor. Appl. Genet. 1968. Vol. 38, № 6. P. 226–231.
203. Shifman S. et al. Linkage disequilibrium patterns of the human genome across populations // Hum Mol Genet. 2003. Vol. 12, № 7. P. 771–776.
204. Duret L., Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes // Annu Rev Genomics Hum Genet. 2009. Vol. 10. P. 285–311.
205. Hudson R.R., Kaplan N.L. Statistical properties of the number of recombination events in the history of a sample of DNA sequences // Genetics. 1985. Vol. 111, № 1. P. 147–164.
206. Andersen S.L., Sekelsky J. Meiotic versus Mitotic Recombination: Two Different Routes for Double-Strand Break Repair // Bioessays. 2010. Vol. 32, № 12. P. 1058–1066.
207. Patterson M. et al. WhatsHap: Weighted Haplotype Assembly for Future-Generation Sequencing Reads // J. Comput. Biol. 2015. Vol. 22, № 6. P. 498–509.
208. Hodgkinson A., Eyre-Walker A. Human triallelic sites: evidence for a new mutational mechanism? // Genetics. 2010. Vol. 184, № 1. P. 233–241.
209. Wilton P.R. et al. A Population Phylogenetic View of Mitochondrial Heteroplasmy // Genetics. 2018. Vol. 208, № 3. P. 1261–1274.
210. Stewart J.B., Chinnery P.F. The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease // Nat Rev Genet. 2015. Vol. 16, № 9. P. 530–542.
211. Yuan J.D. et al. Nuclear pseudogenes of mitochondrial DNA as a variable part of the human genome // Cell Res. 1999. Vol. 9, № 4. P. 281–290.
212. Cibulskis K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples // Nat Biotechnol. 2013. Vol. 31, № 3. P. 213–219.
213. McKenna A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data // Genome Res. 2010. Vol. 20, № 9. P. 1297–1303.
214. Eyres I. et al. Horizontal gene transfer in bdelloid rotifers is ancient, ongoing and more frequent in species from desiccating habitats // BMC Biol. 2015. Vol. 13. P. 90.
215. Allio R. et al. Large Variation in the Ratio of Mitochondrial to Nuclear Mutation Rate across Animals: Implications for Genetic Diversity and the Use of Mitochondrial DNA as a Molecular Marker // Mol Biol Evol. 2017. Vol. 34, № 11. P. 2762–2772.
216. Smith J.M. et al. How clonal are bacteria? // Proc. Natl. Acad. Sci. U.S.A. 1993. Vol. 90, № 10. P. 4384–4388.

Приложения

Таблица П3.1. Список образцов *D. melanogaster*, геномы которых включены в набор данных DPGP3. **Образцы, использованные в работе, отмечены в данном столбце единицами, образцы, не включенные в анализ, отмечены в данном столбце нулями.

Образец	**Общая длина последовательностей «псевдохромосом» доступных для данного образца**	**Общее число маскированных геномных позиций (замененных на знак 'N')**	**Доля маскированных геномных позиций (замененных на знак 'N')**	**Образец использовался/ не использовался в работе****
ZI10	119,029,689	7,836,611	0.066	1
ZI103	119,029,689	8,094,905	0.068	1
ZI104	119,029,689	7,672,397	0.064	1
ZI114N	119,029,689	7,687,255	0.065	1
ZI117	119,029,689	8,386,240	0.070	1
ZI118N	119,029,689	7,961,600	0.067	1
ZI126	119,029,689	8,214,472	0.069	1
ZI129	119,029,689	8,930,261	0.075	1
ZI134N	119,029,689	7,892,469	0.066	1
ZI136	119,029,689	8,207,454	0.069	1
ZI138	119,029,689	7,949,329	0.067	1
ZI152	119,029,689	8,633,834	0.073	1
ZI157	119,029,689	7,955,970	0.067	1
ZI161	119,029,689	8,182,663	0.069	1
ZI164	119,029,689	8,073,943	0.068	1
ZI165	119,029,689	8,310,517	0.070	1
ZI167	119,029,689	8,027,815	0.067	1
ZI170	119,029,689	8,491,189	0.071	1
ZI172	119,029,689	8,110,197	0.068	1
ZI173	119,029,689	8,369,385	0.070	1
ZI176	119,029,689	8,300,963	0.070	1
ZI177	119,029,689	8,106,655	0.068	1
ZI178	119,029,689	8,041,756	0.068	1
ZI179	119,029,689	8,707,431	0.073	1
ZI181	119,029,689	8,184,156	0.069	1
ZI182	119,029,689	7,864,508	0.066	1
ZI184	119,029,689	8,342,194	0.070	1
ZI185	119,029,689	7,996,983	0.067	1
ZI188	119,029,689	8,172,961	0.069	1
ZI190	119,029,689	8,145,301	0.068	1
ZI191	119,029,689	8,665,521	0.073	1
ZI193	119,029,689	8,190,331	0.069	1
ZI194	119,029,689	8,106,916	0.068	1
ZI196	119,029,689	8,150,502	0.068	1
ZI197N	119,029,689	7,833,431	0.066	1
ZI198	119,029,689	8,900,155	0.075	1
ZI199	119,029,689	8,073,287	0.068	1
ZI200	119,029,689	9,342,205	0.078	0
ZI202	119,029,689	8,631,575	0.073	1

ZI206	119,029,689	8,129,988	0.068	1
ZI207	119,029,689	8,306,674	0.070	1
ZI210	119,029,689	8,365,705	0.070	1
ZI211	119,029,689	7,837,763	0.066	1
ZI212	119,029,689	7,987,858	0.067	1
ZI213	119,029,689	8,207,652	0.069	1
ZI214	119,029,689	8,129,299	0.068	1
ZI216N	119,029,689	8,060,419	0.068	1
ZI218	119,029,689	8,910,464	0.075	1
ZI219	119,029,689	8,198,077	0.069	1
ZI220	119,029,689	8,350,146	0.070	1
ZI221	119,029,689	8,523,623	0.072	1
ZI225	119,029,689	8,146,434	0.068	1
ZI226	119,029,689	8,103,462	0.068	1
ZI227	119,029,689	8,162,626	0.069	1
ZI228	119,029,689	8,109,626	0.068	1
ZI230	119,029,689	8,471,962	0.071	1
ZI231	119,029,689	8,090,536	0.068	1
ZI232	119,029,689	8,222,750	0.069	1
ZI233	119,029,689	7,961,094	0.067	1
ZI235	119,029,689	8,191,126	0.069	1
ZI237	119,029,689	8,366,830	0.070	1
ZI239	119,029,689	8,110,026	0.068	1
ZI240	119,029,689	7,915,701	0.067	0
ZI241	119,029,689	7,983,506	0.067	1
ZI250	119,029,689	8,318,077	0.070	1
ZI251N	119,029,689	8,449,839	0.071	1
ZI252	119,029,689	8,241,462	0.069	1
ZI253	119,029,689	8,095,580	0.068	1
ZI254N	119,029,689	7,997,196	0.067	1
ZI255	119,029,689	8,607,171	0.072	1
ZI26	119,029,689	8,195,912	0.069	1
ZI261	119,029,689	8,277,804	0.070	1
ZI263	119,029,689	8,073,705	0.068	1
ZI264	119,029,689	8,137,217	0.068	1
ZI265	119,029,689	8,097,127	0.068	1
ZI267	119,029,689	8,396,381	0.071	1
ZI268	119,029,689	8,233,611	0.069	1
ZI269	119,029,689	8,379,345	0.070	1
ZI27	119,029,689	8,182,088	0.069	1
ZI271	119,029,689	8,539,358	0.072	1
ZI273N	119,029,689	8,155,009	0.069	1
ZI276	119,029,689	8,437,580	0.071	1
ZI279	119,029,689	8,087,925	0.068	1
ZI28	119,029,689	8,123,804	0.068	1
ZI281	119,029,689	8,216,100	0.069	1
ZI284	119,029,689	7,852,835	0.066	1
ZI286	119,029,689	8,250,028	0.069	1
ZI291	119,029,689	8,015,428	0.067	1
ZI292	119,029,689	8,610,383	0.072	1
ZI293	119,029,689	9,075,834	0.076	0
ZI295	119,029,689	8,421,141	0.071	1
ZI296	119,029,689	8,229,161	0.069	1
ZI303	119,029,689	8,276,962	0.070	1

ZI311N	119,029,689	8,030,838	0.067	1
ZI313	119,029,689	7,922,717	0.067	0
ZI314	119,029,689	7,954,502	0.067	1
ZI316	119,029,689	8,470,564	0.071	1
ZI317	119,029,689	8,113,963	0.068	1
ZI319	119,029,689	8,224,154	0.069	1
ZI31N	119,029,689	8,491,254	0.071	1
ZI320	119,029,689	8,365,179	0.070	1
ZI321	119,029,689	8,873,002	0.075	1
ZI324	119,029,689	8,180,277	0.069	1
ZI329	119,029,689	8,117,749	0.068	1
ZI33	119,029,689	8,590,616	0.072	1
ZI332	119,029,689	8,195,979	0.069	1
ZI333	119,029,689	8,105,403	0.068	1
ZI335	119,029,689	7,907,266	0.066	1
ZI336	119,029,689	8,057,973	0.068	1
ZI339	119,029,689	8,524,683	0.072	1
ZI341	119,029,689	8,110,098	0.068	1
ZI342	119,029,689	7,826,843	0.066	1
ZI344	119,029,689	8,063,790	0.068	1
ZI348	119,029,689	8,111,026	0.068	1
ZI351	119,029,689	7,825,190	0.066	1
ZI352	119,029,689	8,038,988	0.068	1
ZI353	119,029,689	8,016,366	0.067	1
ZI357N	119,029,689	8,331,155	0.070	1
ZI358	119,029,689	8,041,183	0.068	1
ZI359	119,029,689	8,823,718	0.074	1
ZI362	119,029,689	8,251,285	0.069	1
ZI364	119,029,689	8,129,886	0.068	1
ZI365	119,029,689	8,663,047	0.073	1
ZI368	119,029,689	8,196,655	0.069	1
ZI370	119,029,689	8,489,440	0.071	1
ZI373	119,029,689	8,421,682	0.071	1
ZI374	119,029,689	8,038,723	0.068	1
ZI377	119,029,689	7,931,930	0.067	1
ZI378	119,029,689	8,297,946	0.070	1
ZI379	119,029,689	8,357,349	0.070	1
ZI380	119,029,689	8,048,886	0.068	1
ZI381	119,029,689	8,302,065	0.070	1
ZI382	96,606,862	6,653,953	0.069	0
ZI384	119,029,689	8,412,221	0.071	1
ZI386	119,029,689	8,555,341	0.072	1
ZI388	119,029,689	8,985,024	0.075	1
ZI392	119,029,689	8,172,188	0.069	1
ZI394N	119,029,689	8,689,386	0.073	1
ZI395	119,029,689	8,035,950	0.068	1
ZI396	119,029,689	7,994,298	0.067	1
ZI397N	119,029,689	7,923,244	0.067	1
ZI398	119,029,689	8,108,948	0.068	1
ZI400	119,029,689	8,684,034	0.073	1
ZI402	119,029,689	8,312,994	0.070	1
ZI405	119,029,689	7,957,513	0.067	1
ZI413	119,029,689	8,252,301	0.069	1
ZI418N	119,029,689	7,956,946	0.067	1

ZI420	119,029,689	7,904,290	0.066	1
ZI421	119,029,689	8,462,623	0.071	1
ZI429	119,029,689	8,001,946	0.067	1
ZI431	119,029,689	8,270,698	0.069	1
ZI433	119,029,689	8,239,035	0.069	1
ZI437	119,029,689	8,075,016	0.068	1
ZI443	119,029,689	8,614,864	0.072	1
ZI444	119,029,689	8,248,653	0.069	1
ZI445	119,029,689	8,141,440	0.068	1
ZI446	119,029,689	8,080,133	0.068	1
ZI447	119,029,689	8,125,572	0.068	1
ZI448	119,029,689	8,557,799	0.072	1
ZI453	119,029,689	8,120,225	0.068	1
ZI455N	119,029,689	8,129,321	0.068	1
ZI456	119,029,689	8,091,788	0.068	1
ZI457	119,029,689	8,314,939	0.070	1
ZI458	119,029,689	8,380,009	0.070	1
ZI460	119,029,689	8,151,764	0.068	1
ZI466	119,029,689	8,690,400	0.073	1
ZI467	119,029,689	8,103,566	0.068	1
ZI468	119,029,689	8,121,991	0.068	1
ZI471	119,029,689	7,887,056	0.066	1
ZI472	119,029,689	8,182,081	0.069	1
ZI476	119,029,689	8,564,998	0.072	1
ZI477	119,029,689	8,150,917	0.068	1
ZI486	119,029,689	7,676,441	0.064	1
ZI488	119,029,689	8,407,017	0.071	1
ZI490	119,029,689	7,941,635	0.067	1
ZI491	119,029,689	7,746,090	0.065	1
ZI504	119,029,689	8,410,081	0.071	1
ZI505	119,029,689	7,904,528	0.066	1
ZI508	119,029,689	7,923,546	0.067	1
ZI50N	119,029,689	8,082,201	0.068	1
ZI514N	119,029,689	8,475,463	0.071	1
ZI517	119,029,689	8,617,004	0.072	1
ZI523	119,029,689	7,915,367	0.066	1
ZI524	119,029,689	8,167,621	0.069	1
ZI527	119,029,689	8,211,477	0.069	1
ZI530	119,029,689	8,188,917	0.069	1
ZI56	119,029,689	7,832,605	0.066	0
ZI59	119,029,689	8,056,827	0.068	1
ZI61	119,029,689	8,165,893	0.069	1
ZI68	119,029,689	8,615,192	0.072	1
ZI76	119,029,689	8,019,319	0.067	1
ZI81	119,029,689	8,632,295	0.073	1
ZI85	119,029,689	8,191,835	0.069	1
ZI86	119,029,689	8,743,861	0.073	1
ZI90	119,029,689	8,045,205	0.068	1
ZI91	119,029,689	8,259,724	0.069	1
ZI99	119,029,689	8,381,957	0.070	1

Таблица П3.2. Список образцов *D. melanogaster*, геномы которых включены в набор данных DGRP. **Образцы, использованные в работе, отмечены в данном столбце единицами, образцы, не включенные в анализ, отмечены в данном столбце нулями.

Образец	**Общая длина последовательностей «псевдохромосом» доступных для данного образца**	**Общее число маскированных геномных позиций (замененных на знак 'N')**	**Доля маскированных геномных позиций (замененных на знак 'N')**	**Образец использовался/ не использовался в работе****
RAL-100	119,029,689	40,310,568	0.339	0
RAL-101	119,029,689	36,520,786	0.307	0
RAL-105	119,029,689	14,387,313	0.121	1
RAL-109	119,029,689	37,216,912	0.313	0
RAL-129	119,029,689	11,393,666	0.096	1
RAL-136	119,029,689	59,544,274	0.500	0
RAL-138	119,029,689	22,488,371	0.189	1
RAL-142	119,029,689	20,412,669	0.171	1
RAL-149	119,029,689	21,987,283	0.185	1
RAL-153	119,029,689	17,968,740	0.151	1
RAL-158	119,029,689	16,301,338	0.137	1
RAL-161	119,029,689	16,441,199	0.138	1
RAL-176	119,029,689	19,094,065	0.160	1
RAL-177	119,029,689	15,094,486	0.127	1
RAL-181	119,029,689	14,033,108	0.118	1
RAL-189	119,029,689	16,963,697	0.143	1
RAL-195	119,029,689	15,563,411	0.131	1
RAL-208	119,029,689	14,626,053	0.123	1
RAL-21	119,029,689	14,737,533	0.124	1
RAL-217	119,029,689	16,871,013	0.142	1
RAL-223	119,029,689	19,472,864	0.164	1
RAL-227	119,029,689	12,553,065	0.105	1
RAL-228	119,029,689	14,985,455	0.126	1
RAL-229	119,029,689	24,393,121	0.205	0
RAL-233	119,029,689	13,341,487	0.112	1
RAL-235	119,029,689	12,305,975	0.103	1
RAL-237	119,029,689	53,004,097	0.445	0
RAL-239	119,029,689	14,072,156	0.118	1
RAL-256	119,029,689	14,133,835	0.119	1
RAL-26	119,029,689	13,227,013	0.111	1
RAL-28	119,029,689	11,185,615	0.094	1
RAL-280	119,029,689	12,827,754	0.108	1
RAL-287	119,029,689	15,047,028	0.126	1
RAL-301	119,029,689	42,449,874	0.357	0
RAL-303	119,029,689	91,577,960	0.769	0
RAL-304	119,029,689	30,093,551	0.253	0
RAL-306	119,029,689	34,081,239	0.286	0
RAL-307	119,029,689	38,351,769	0.322	0
RAL-309	119,029,689	42,516,704	0.357	0
RAL-31	119,029,689	63,634,568	0.535	0
RAL-310	119,029,689	14,578,458	0.122	1
RAL-313	119,029,689	16,046,643	0.135	1
RAL-315	119,029,689	16,434,407	0.138	1
RAL-317	119,029,689	44,638,819	0.375	0
RAL-318	119,029,689	13,643,874	0.115	1

RAL-319	119,029,689	35,373,558	0.297	0
RAL-32	119,029,689	31,770,010	0.267	0
RAL-320	119,029,689	20,782,500	0.175	1
RAL-321	119,029,689	31,912,633	0.268	0
RAL-324	119,029,689	29,780,137	0.250	0
RAL-325	119,029,689	22,232,220	0.187	1
RAL-332	119,029,689	29,766,264	0.250	0
RAL-335	119,029,689	25,174,552	0.211	0
RAL-336	119,029,689	37,709,095	0.317	0
RAL-338	119,029,689	51,947,380	0.436	0
RAL-340	119,029,689	44,632,171	0.375	0
RAL-348	119,029,689	19,180,729	0.161	1
RAL-350	119,029,689	40,231,520	0.338	0
RAL-352	119,029,689	41,453,732	0.348	0
RAL-354	119,029,689	20,378,366	0.171	1
RAL-355	119,029,689	17,885,063	0.150	1
RAL-356	119,029,689	25,829,126	0.217	0
RAL-357	119,029,689	11,384,611	0.096	1
RAL-358	119,029,689	15,604,293	0.131	1
RAL-359	119,029,689	26,544,049	0.223	0
RAL-360	119,029,689	21,885,146	0.184	1
RAL-361	119,029,689	47,573,263	0.400	0
RAL-362	119,029,689	23,134,414	0.194	1
RAL-365	119,029,689	19,501,896	0.164	1
RAL-367	119,029,689	21,978,305	0.185	1
RAL-370	119,029,689	13,677,489	0.115	1
RAL-371	119,029,689	15,700,826	0.132	1
RAL-373	119,029,689	47,717,864	0.401	0
RAL-374	119,029,689	12,743,161	0.107	1
RAL-375	119,029,689	22,747,726	0.191	1
RAL-377	119,029,689	51,412,671	0.432	0
RAL-379	119,029,689	15,175,514	0.127	1
RAL-38	119,029,689	45,799,678	0.385	0
RAL-380	119,029,689	14,250,561	0.120	1
RAL-381	119,029,689	37,859,177	0.318	0
RAL-382	119,029,689	17,480,263	0.147	1
RAL-383	119,029,689	12,364,831	0.104	1
RAL-385	119,029,689	17,087,583	0.144	1
RAL-386	119,029,689	18,131,996	0.152	1
RAL-390	119,029,689	44,512,131	0.374	0
RAL-391	119,029,689	17,548,723	0.147	1
RAL-392	119,029,689	32,113,418	0.270	0
RAL-395	119,029,689	18,433,730	0.155	1
RAL-397	119,029,689	43,489,810	0.365	0
RAL-399	119,029,689	12,256,361	0.103	1
RAL-40	119,029,689	15,550,800	0.131	1
RAL-405	119,029,689	39,112,240	0.329	0
RAL-406	119,029,689	16,257,451	0.137	1
RAL-409	119,029,689	69,603,138	0.585	0
RAL-41	119,029,689	17,460,518	0.147	1
RAL-42	119,029,689	16,016,402	0.135	1
RAL-426	119,029,689	52,920,064	0.445	0
RAL-427	119,029,689	13,545,130	0.114	1
RAL-437	119,029,689	22,534,188	0.189	1

RAL-439	119,029,689	17,030,010	0.143	1
RAL-440	119,029,689	41,842,301	0.352	0
RAL-441	119,029,689	12,847,240	0.108	1
RAL-443	119,029,689	39,705,377	0.334	0
RAL-45	119,029,689	15,344,822	0.129	1
RAL-461	119,029,689	13,059,720	0.110	1
RAL-48	119,029,689	50,691,215	0.426	0
RAL-486	119,029,689	12,072,427	0.101	1
RAL-49	119,029,689	17,291,798	0.145	1
RAL-491	119,029,689	11,628,972	0.098	1
RAL-492	119,029,689	13,367,747	0.112	1
RAL-502	119,029,689	33,527,520	0.282	0
RAL-505	119,029,689	13,469,481	0.113	1
RAL-508	119,029,689	13,204,392	0.111	1
RAL-509	119,029,689	12,774,965	0.107	1
RAL-513	119,029,689	14,365,840	0.121	1
RAL-517	119,029,689	15,531,587	0.130	1
RAL-528	119,029,689	63,318,279	0.532	0
RAL-530	119,029,689	15,317,159	0.129	1
RAL-531	119,029,689	29,878,262	0.251	0
RAL-535	119,029,689	13,269,356	0.111	1
RAL-551	119,029,689	47,751,889	0.401	0
RAL-555	119,029,689	16,150,439	0.136	1
RAL-559	119,029,689	50,128,803	0.421	0
RAL-563	119,029,689	53,501,510	0.449	0
RAL-566	119,029,689	37,116,715	0.312	0
RAL-57	119,029,689	18,540,067	0.156	1
RAL-584	119,029,689	35,328,745	0.297	0
RAL-589	119,029,689	19,925,672	0.167	1
RAL-59	119,029,689	21,749,762	0.183	1
RAL-595	119,029,689	15,652,575	0.132	1
RAL-596	119,029,689	18,059,025	0.152	1
RAL-627	119,029,689	106,012,649	0.891	0
RAL-630	119,029,689	68,867,579	0.579	0
RAL-634	119,029,689	39,779,250	0.334	0
RAL-639	119,029,689	14,939,303	0.126	1
RAL-642	119,029,689	12,578,306	0.106	1
RAL-646	119,029,689	14,167,883	0.119	1
RAL-69	119,029,689	13,277,428	0.112	1
RAL-703	119,029,689	12,121,590	0.102	1
RAL-705	119,029,689	16,871,714	0.142	1
RAL-707	119,029,689	16,349,755	0.137	1
RAL-712	119,029,689	14,358,250	0.121	1
RAL-714	119,029,689	21,670,499	0.182	1
RAL-716	119,029,689	13,416,672	0.113	1
RAL-721	119,029,689	12,217,201	0.103	1
RAL-727	119,029,689	18,775,105	0.158	1
RAL-73	119,029,689	16,685,067	0.140	1
RAL-730	119,029,689	13,540,681	0.114	1
RAL-732	119,029,689	45,625,477	0.383	0
RAL-737	119,029,689	44,303,502	0.372	0
RAL-738	119,029,689	68,363,354	0.574	0
RAL-748	119,029,689	22,908,820	0.192	1
RAL-75	119,029,689	20,671,787	0.174	1

RAL-757	119,029,689	29,473,757	0.248	0
RAL-761	119,029,689	13,459,516	0.113	1
RAL-765	119,029,689	15,200,685	0.128	1
RAL-774	119,029,689	46,234,361	0.388	0
RAL-776	119,029,689	24,227,006	0.204	0
RAL-783	119,029,689	11,821,089	0.099	1
RAL-786	119,029,689	13,109,706	0.110	1
RAL-787	119,029,689	13,848,674	0.116	1
RAL-790	119,029,689	13,121,754	0.110	1
RAL-796	119,029,689	23,383,742	0.196	1
RAL-799	119,029,689	10,878,022	0.091	1
RAL-801	119,029,689	34,024,118	0.286	0
RAL-802	119,029,689	75,696,592	0.636	0
RAL-804	119,029,689	23,052,186	0.194	1
RAL-805	119,029,689	12,261,349	0.103	1
RAL-808	119,029,689	38,285,591	0.322	0
RAL-810	119,029,689	12,082,351	0.102	1
RAL-812	119,029,689	31,349,955	0.263	0
RAL-818	119,029,689	30,252,427	0.254	0
RAL-819	119,029,689	20,747,758	0.174	1
RAL-820	119,029,689	13,764,049	0.116	1
RAL-821	119,029,689	46,901,492	0.394	0
RAL-822	119,029,689	38,744,765	0.326	0
RAL-83	119,029,689	30,726,681	0.258	0
RAL-832	119,029,689	18,581,947	0.156	1
RAL-837	119,029,689	14,260,692	0.120	1
RAL-843	119,029,689	18,990,233	0.160	1
RAL-849	119,029,689	38,718,120	0.325	0
RAL-85	119,029,689	40,361,773	0.339	0
RAL-850	119,029,689	17,148,214	0.144	1
RAL-852	119,029,689	48,931,311	0.411	0
RAL-853	119,029,689	54,779,544	0.460	0
RAL-855	119,029,689	33,308,685	0.280	0
RAL-857	119,029,689	49,618,194	0.417	0
RAL-859	119,029,689	13,900,970	0.117	1
RAL-861	119,029,689	19,279,659	0.162	1
RAL-879	119,029,689	12,958,129	0.109	1
RAL-88	119,029,689	44,375,546	0.373	0
RAL-882	119,029,689	11,569,586	0.097	1
RAL-884	119,029,689	39,299,285	0.330	0
RAL-887	119,029,689	12,115,118	0.102	1
RAL-890	119,029,689	15,854,773	0.133	1
RAL-892	119,029,689	28,238,683	0.237	0
RAL-894	119,029,689	23,918,606	0.201	0
RAL-897	119,029,689	12,228,068	0.103	1
RAL-900	119,029,689	34,981,573	0.294	0
RAL-907	119,029,689	63,791,211	0.536	0
RAL-908	119,029,689	12,926,807	0.109	1
RAL-91	119,029,689	15,654,335	0.132	1
RAL-911	119,029,689	60,612,540	0.509	0
RAL-913	119,029,689	77,704,313	0.653	0
RAL-93	119,029,689	12,850,565	0.108	1

Таблица П4.1. Доля гомозиготных и гетерозиготных сайтов в геномах 11 анализируемых особей *A. vaga*. Представленные числы основаны на сайтах гаплоидной сборки, включенных в набор однонуклеотидных полиморфизмов III (n = 58,158,930; использовались только те сайты, генотипы в которых были определены для всех особей L1-L1). Отдельно приведены числа, рассчитанные для всего генома, и числа, полученные для четырехкратно вырожденных и для несинонимических сайтов.

Весь геном					
Особь	**Число проанализированных сайтов**	**Число гомозиготных сайтов**	**Доля гомозиготных сайтов**	**Число гетерозиготных сайтов**	**Доля гетерозиготных сайтов**
L1	58,158,930	56,995,588	0.9800	1,163,342	0.0200
L2	58,158,930	57,012,714	0.9803	1,146,216	0.0197
L3	58,158,930	57,015,629	0.9803	1,143,301	0.0197
L4	58,158,930	57,791,265	0.9937	367,665	0.0063
L5	58,158,930	57,779,417	0.9935	379,513	0.0065
L6	58,158,930	57,779,528	0.9935	379,402	0.0065
L7	58,158,930	57,780,928	0.9935	378,002	0.0065
L8	58,158,930	57,778,532	0.9935	380,398	0.0065
L9	58,158,930	57,779,197	0.9935	379,733	0.0065
L10	58,158,930	57,849,205	0.9947	309,725	0.0053
L11	58,158,930	57,785,526	0.9936	373,404	0.0064
Четырехкратно вырожденные сайты					
Особь	**Число проанализированных сайтов**	**Число гомозиготных сайтов**	**Доля гомозиготных сайтов**	**Число гетерозиготных сайтов**	**Доля гетерозиготных сайтов**
L1	3,612,576	3,474,302	0.9617	138,274	0.0383
L2	3,612,576	3,477,630	0.9626	134,946	0.0374
L3	3,612,576	3,478,920	0.9630	133,656	0.0370
L4	3,612,576	3,568,719	0.9879	43,857	0.0121
L5	3,612,576	3,567,218	0.9874	45,358	0.0126
L6	3,612,576	3,567,712	0.9876	44,864	0.0124
L7	3,612,576	3,568,015	0.9877	44,561	0.0123
L8	3,612,576	3,567,442	0.9875	45,134	0.0125
L9	3,612,576	3,567,537	0.9875	45,039	0.0125
L10	3,612,576	3,576,202	0.9899	36,374	0.0101
L11	3,612,576	3,568,202	0.9877	44,374	0.0123
Несинонимические сайты					
Особь	**Число проанализированных сайтов**	**Число гомозиготных сайтов**	**Доля гомозиготных сайтов**	**Число гетерозиготных сайтов**	**Доля гетерозиготных сайтов**
L1	14,099,199	13,939,008	0.9886	160,191	0.0114
L2	14,099,199	13,942,639	0.9889	156,560	0.0111
L3	14,099,199	13,943,747	0.9890	155,452	0.0110
L4	14,099,199	14,051,948	0.9966	47,251	0.0034
L5	14,099,199	14,050,690	0.9966	48,509	0.0034

L6	14,099,199	14,050,556	0.9965	48,643	0.0035
L7	14,099,199	14,050,949	0.9966	48,250	0.0034
L8	14,099,199	14,050,596	0.9966	48,603	0.0034
L9	14,099,199	14,050,979	0.9966	48,220	0.0034
L10	14,099,199	14,060,739	0.9973	38,460	0.0027
L11	14,099,199	14,051,511	0.9966	47,688	0.0034

Таблица П4.2. Оценки частоты ошибок фазирования для использованных в работе наборов фазированных данных. Оценки частоты ошибок фазирования были получены для трех особей, для которых было доступно более одного набора данных секвенирования (L1, L2 и L11). Для особи L1 были секвенированы три независимо полученные геномные библиотеки с использованием инструментов Illumina HiSeq, Illumina MiSeq и PacBio. Для особи L11 были секвенированы две независимо полученные геномные библиотеки с использованием инструментов Illumina HiSeq и Illumina MiSeq. L2 также секвенировали с использованием инструментов Illumina HiSeq и Illumina MiSeq, но и в том, и в другом случае была использована та же самая библиотека. Оценки частоты ошибок фазирования основаны на сравнении фазированных блоков, восстановленных с использованием прочтений HiSeq и MiSeq (L1, L2, L11) или прочтений HiSeq и PacBio (L1). Выявленные несоответствия в результатах фазирования, полученных для одной и той же особи на основании разных данных, могут являться результатом переключения матриц в ходе ПЦР, ошибочного выравнивания прочтений с паралогичными областями генома или других артефактов. В случае данных фазирования, восстановленных на основе прочтений HiSeq, одноступенчатая фильтрация (исключение блоков, содержащих пары «конфликтующих» сайтов) соответствует набору фазированных данных 1, а двухступенчатая фильтрация (дополнительно включающая шаг, основанный на оценках вероятности ошибок фазирования, определенных HapCUT2) соответствует набору фазированных данных 2. Оценки, полученные с применением набора фазированных данных 1, использовавшегося для большинства анализов, выделены жирным шрифтом.

Особь	**Наборы фазированных данных, использованные для сравнения**	**Фильтрация**		Число гаплоидных контигов с фазированными блоками, пересекающимися между двумя наборами фазированных данных	**Гаплоидные контиги с фазированными блоками, пересекающимися между двумя наборами фазированных данных, и ≥10 парами гетерозиготных сайтов, включенными в анализ**		
		Исключение блоков, содержащих пары «конфликтующих» сайтов	Фильтрация на основании оценок вероятности ошибок фазирования		Общее число	Число контигов с ошибками фазирования	Доля контигов с ошибками фазирования
L1	L1 HiSeq vs L1 MiSeq	–	–	6,760	6,197	142	0.0229
	L1 HiSeq vs L1 MiSeq	+	–	6,077	**5,446**	**4**	**0.0007**
	L1 HiSeq vs L1 MiSeq	+	+	5,676	4,894	0	0
	L1 HiSeq vs L1 PacBio	–	–	6,761	6,208	197	0.0317
	L1 HiSeq vs L1 PacBio	+ (только для данных фазирования на основе прочтений HiSeq)	–	6,394	**5,843**	**44**	**0.0075**
	L1 HiSeq vs L1 PacBio	+ (только для данных фазирования на основе прочтений HiSeq)	+ (только для данных фазирования на основе прочтений HiSeq)	6,051	5,364	37	0.0069
L2	L2 HiSeq vs L2 MiSeq	–	–	6,589	6,045	223	0.0369
	L2 HiSeq vs L2 MiSeq	+	–	5,611	**4,940**	**15**	**0.0030**
	L2 HiSeq vs L2 MiSeq	+	+	5,368	4,621	3	0.0006
L11	L11 HiSeq vs L11 MiSeq	–	–	6,207	4,603	268	0.0582
	L11 HiSeq vs L11 MiSeq	+	–	5,959	**4,339**	**67**	**0.0154**
	L11 HiSeq vs L11 MiSeq	+	+	5,641	3,844	12	0.0031