Факультет биоинженерии и биоинформатики Федерального государственного бюджетного образовательного учреждения высшего профессионального образования Московского государственного университета имени М.В. Ломоносова,

Сектор молекулярной эволюции Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук

На правах рукописи

Леушкин Евгений Владимирович

АНАЛИЗ ЭВОЛЮЦИИ ИНСЕРЦИЙ И ДЕЛЕЦИЙ В ПОСЛЕДОВАТЕЛЬНОСТИ ДНК, ПРОВОДИМЫЙ НА ОСНОВЕ СРАВНЕНИЯ ПОЛНЫХ ГЕНОМОВ

03.01.09 - математическая биология, биоинформатика

Диссертация на соискание ученой степени
кандидата биологических наук

Научный руководитель:
кандидат биологических наук
Базыкин Георгий Александрович

Москва – 2014

Содержание

Введение

Выявление закономерностей в молекулярной эволюции в первую очередь основывается на сравнении геномов разных видов и генотипов разных особей одного вида. Технологии секвенирования нового поколения (next-generation sequencing) вызвали в последние годы экспоненциальный рост числа секвенированных геномов, что значительно расширило масштаб сравнительно-геномных исследований. Согласно сайту Genomes OnLine Database (http://www.genomesonline.org/), на 12.09.2013 прочитано 311 эукариотических геномов, 6349 геномов бактерий и 227 геномов архей. Некоторые из этих геномов, например, *Homo sapiens* (http://www.1000genomes.org/), *Drosophila melanogaster* (https://www.hgsc.bcm.edu/projects/dgrp/, http://www.dpgp.org/), *Arabidopsis thaliana* (http://www.1001genomes.org/), были прочитаны для многих индивидуумов, что даёт возможность изучать внутривидовые различия.

Сравнение геномов разных видов позволяет исследовать процессы, действовавшие в ходе их эволюции после дивергенции от общего предка, а сравнение генотипов особей одной популяции - популяционно-генетические факторы, определяющие полиморфизм. Использование полногеномных данных по дивергенции и полиморфизму позволяет с высокой точностью измерить скорости мутагенеза для мутаций даже редких типов, определить интенсивность отрицательного и положительного отбора, действующего на мутации, оценить

эффект мутаций на геномное окружение и проследить его на разных эволюционных расстояниях, а также выяснить, как влияет на мутации генная конверсия. Основным объектом исследования были выбраны инсерции и делеции, как мутации с значительно более радикальным эффектом на приспособленность, чем однонуклеотидные замены. Основной организм, в котором проводились исследования – плодовая мушка *Drosophila melanogaster* – выбран, в первую очередь, по тем причинам, что является хорошо изученным модельным организмом, имеет высококачественные данные по секвенированию и пересеквенированию генома, а также, в отличие от человека, высокую эффективную численность и высокую популяционную изменчивость. Некоторые тесты были также выполнены на последовательностях геномов позвоночных и дрожжей.

1. Инсерции и делеции

Инсерции и делеции (инсерции и делеции в последовательности ДНК), наряду с однонуклеотидными заменами, представляют собой важнейший фактор эволюции генома. Инсерции и делеции происходят приблизительно в 10 раз реже однонуклеотидных замен. Однако общее число нуклеотидов, подвергающихся инсерции или делеции, сопоставимо с числом замен, а зачастую, как, например, в геноме человека и приматов, – даже превосходит его [1]. Также инсерция/делеция – в среднем событие более радикальное для участка ДНК, чем нуклеотидная

замена, то есть с большей вероятностью влияет на функцию, выполняемую данным участком генома.

1.1. Механизмы возникновения инсерций и делеций

Было предложено несколько моделей для механизма возникновения коротких инсерций и делеций. Большая часть из этих гипотез основывается на том факте, что подавляющее большинство инсерций происходит в участках тандемных повторов [2–5]. Например, в работе [4] было показано, что 98,4% инделов у *Blochmкannia chromaiodes* происходит в таких участках.

Предполагается, что короткие инсерции/делеции возникают в основном за счёт проскальзывания ДНК-полимеразы относительно матрицы, в результате чего образуется микропетля либо на матричной, либо на вновь синтезированной цепи ДНК. Таким образом, некоторый участок ДНК будет соответственно либо пропущен (делеция), либо реплицирован дважды (инсерция) [6,7] (Рисунок1).

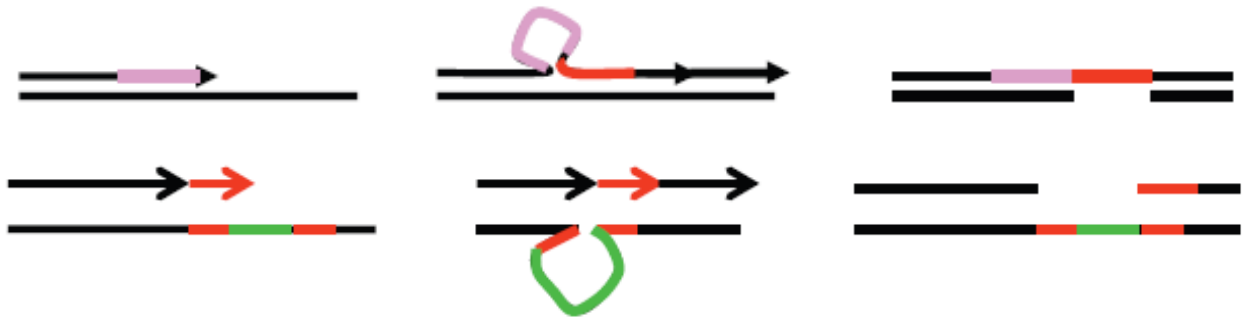

Рис. 1. Механизм возникновения инсерций и делеций за счёт эффекта проскальзывания ДНК-полимеразы в ходе репликации ДНК. Сверху показано, как деспирализация двойной цепи ДНК может вызвать образование петли (лиловый)

и последующую инсерцию (красный). В нижней части рисунка петля (зелёный и красный), образовавшаяся из-за наличия участков микрогомологии (отмечены красным), влечёт за собой делецию. Из работы [7].

Для более длинных инсерций предполагается наличие механизма, основанного на процессе негомологичного склеивания концов ДНК в местах двуцепочечных разрывов (NHEJ – nonhomologous end joining). Часто в местах таких разрывов образуются достаточно протяжённые липкие концы. При наличии участков микрогомологии может происходить ошибочное склеивание концов, за которым следует заполнение оставшихся одноцепочечных брешей ДНК. Результат такого процесса будет выглядеть как дупликация фрагмента ДНК. При этом две копии обычно будут разделены небольшим участком недуплицированной последовательности, соответствующей, согласно данному механизму, участку микрогомологии [8].

NHEJ также может вносить вклад в генерацию делеций: в процессе репарации двуцепочечных разрывов ДНК концы последовательностей часто подвергаются частичной деградации, следствием чего будет делеция этого участка. Но в некоторых случаях в ходе такой репарации может происходить инсерция экзогенного фрагмента ДНК – редкий случай инсерции, не являющейся дупликацией [8].

Ещё один механизм – неравный кроссинговер в участке тандемных повторов, который, меняя число копий участка ДНК, приводит к инсерциям и

делециям [9]. Этот механизм не объясняет образование первого повтора и, скорее всего, мало применим к "размножению" коротких повторов, поскольку маловероятно, что короткие повторы могли бы обеспечить достаточную гомологию для неравного кроссинговера; однако он может играть роль в увеличении копийности длинных повторов, в первую очередь – рибосомальных генов.

1.2. Темпы инсерционного и делеционного мутагенеза

Частоты инсерций и делеций, как и частоты других мутаций, можно оценить по наблюдаемому уровню внутривидового полиморфизма или межвидовой дивергенции. Однако в кодирующей последовательности таким оценкам мешает действие отбора. Для обхода этого ограничения можно использовать псевдогены: имея практически тот же нуклеотидный состав, что и функциональные гены, они не испытывают существенного действия отбора и накапливают мутации, в том числе инсерции и делеции, практически нейтрально. В Таблице 1 приведены темпы инсерционного и делеционного мутагенеза в расчете на число однонуклеотидных замен в псевдогенах, полученные из данных по дивергенции [10]. Во всех рассматривавшихся организмах короткие делеции встречаются значительно чаще коротких инсерций.

Скорость мутагенеза можно также измерять напрямую, секвенируя обоих родителей и потомка. Для человека несколько десятков таких троек генотипов были получены в ходе нескольких крупных проектов, в т.ч. в одной из фаз

проекта "1000 genomes". Для однонуклеотидных мутаций скорость мутагенеза составила 1,0–1.2 × 10^{-8} [11]. Однако для инделов таких прямых измерений пока не получено. В [12] оценки скоростей мутагенеза инсерций и делеций произведены на основе сравнения полиморфных инделов и однонуклеотидных замен в 62 локусах генома *H. sapiens*, ассоциированных с менделевскими заболеваниями; скорость мутагенеза для инсерций составила 0,20 × 10^{-9}, для делеций – 0,58 × 10^{-9}.

Таблица 1. Темпы инсерционного и делеционного мутагенеза, рассчитанные из анализа межвидовой дивергенции последовательностей псевдогенов (из [10]).

	Drosophila sp.	*Laupala sp.*	*Podisma sp.*	Млекопитающие (приматы и грызуны)
Размер генома (Мб)	179	1910	18150	~3000
Число делеций в расчёте на 1 нуклеотидную замену	0,13	0,07	0,06	0,05
Число инсерций в расчёте на 1 нуклеотидную замену	0,015	0,02	0,03	0,01
Средний размер делеций	35	7,0	1,6	3,2

Средний размер инсерций	2,9	6,5	1,2	2,4
Среднее число потерянных нуклеотидов в расчёте на 1 нуклеотидную замену	4,5	0,34	0,06	0,13

Следует отметить, что инсерции и делеции, как и другие мутации, не распределены по геному равномерно, а, напротив, часто образуют скопления – так называемые горячие точки мутагенеза. К примеру, в работе [13] показано, что положения полиморфных позиций (SNP – single nucleotide polymorphism) в популяциях разных видов сильно коррелированны между собой, и что рядом с такими позициями повышается частота других SNP, что можно объяснить только повышенной скоростью мутагенеза на данном участке. При этом повышенная частота инделов в сегменте последовательности коррелированна с повышенной частотой точечных нуклеотидных мутаций [14]. Частота возникновения инделов также сильно зависит от нуклеотидных контекстов. Первостепенным фактором, определяющим вероятность возникновения индела, считается количество повторов последовательности ДНК. В [15] показано, что в микросателлитах частота полиморфных инделов резко возрастает при наличии ≥10 тандемных повторов одного нуклеотида или ≥5-6 повторов двух нуклеотидов. Также существуют определенные мотивы последовательности ДНК, в которых инделы

возникают чаще [16]. Возникновению инделов сильно способствует низкий GC-состав последовательности [8,15].

В функциональных последовательностях инделы подвергаются действию отбора. В большинстве своём инсерции и делеции в белок-кодирующей последовательности – это вредные мутации. Если длина индела не кратна трём, то он приводит к сдвигу рамки считывания. Но даже если длина кратна трём, изменение количества аминокислот в белке может сильно сказываться на его пространственной структуре. Численно это выражается в том, что частота инделов, кратных 3, в кодирующей области примерно в 2 раза ниже, чем в некодирующей, а инделов, некратных 3, – в >100 раз ниже. Инделы происходят в основном в тех белках, в которых ослаблено действие отрицательного отбора, а внутри белка – в менее консервативных участках: в петлях и на границах доменов [5].

1.3. Практическая важность инделов

Поскольку инделы – это один из наиболее распространённых типов мутаций, часто оказывающий существенное влияние на функцию генов, изучение инделов в геноме человека важно с медицинской точки зрения. По данным [17], у человека изменчивость числа копий генов – т.е. внутрипопуляционный полиморфизм, создаваемый длинными инделами – затрагивает в общей сложности 360 Мб (12% генома), и определённые варианты могут вызывать заболевания. Так, делеция 1,4 Мб, затрагивающая ген HNF1B, повышает риск

заболевания аутизмом и шизофренией [18]. Для делеции генов CDY2 и HSFY показана ассоциация с мужским бесплодием [19]. Известно, что заболевания также могут быть ассоциированы с делециями некодирующих участков, затрагивающих регуляторные области генов. Например, делеция перед геном *IRGM* чаще встречается у людей с болезнью Крона [20], делеция длиной 7,4 Кб в регуляторной последовательности *FOXL2* вызывает блефорофимоз [21].

Отдельный класс заболеваний вызывается экспансией тринуклеотидных повторов в кодирующих областях генома. Хорея Хантингтона возникает при наличии >35 повторов кодона CAG в гене Хантингтина (нормальное содержание 10-29 повторов) [22]. Хантингтин с повышенным числом повторов глутамина вызывает повреждение клеток мозга, что приводит к нарушению координации и снижению когнитивных способностей человека.

Экспансия другого кодона, CGG, в гене FMR1 в Х-хромосоме приводит к метилированию участка генома, содержащего данный ген, и, как следствие, к подавлению экспрессии гена [23]. Для больных характерна умственная отсталость, нарушение речи и координации, часто развивается аутизм.

2. Естественный отбор, методы выявления

Роль естественного отбора в эволюции генома зависит от множества факторов биологии вида. Доля последовательности генома человека, находящейся под действием отбора, не превышает ~15%. Напротив, у *Drosophila melanogaster*

под отбором находится большая часть генома: среди *de novo* нуклеотидных мутаций ~90% несинонимичных мутаций в экзонах [24] и ~50% мутаций в межгенных участках и длинных интронах [25–27] находятся под отрицательным отбором, сила которого достаточна, чтобы радикально уменьшить вероятность закрепления новой мутации. Оценки доли мутаций α, закреплённых под действием положительного отбора, существенно различаются в для разных организмов и для разных методов исследования [24,26,28,29]; однако очевидно, что положительный отбор играет большую роль в закреплении мутаций во всех видах.

Ниже рассмотрены основные методы для выявления следов действия отбора на нуклеотидную последовательность.

2.1. Тест dn/ds

Самым простым тестом для определения действия отбора и его направления в кодирующей последовательности является вычисление отношения числа несинонимичных (D_n) и синонимичных (D_s) замен, приходящихся на несинонимичный (синонимичный) сайт. Так как синонимичные замены не изменяют структуру белка, то в грубом приближении можно считать, что они не вызывают изменения приспособленности. Следовательно, эволюционировать синонимичные сайты будут нейтрально, с постоянной скоростью, определяемой скоростью мутирования. Напротив, несинонимичные замены могут подвергаться действию отбора. Положительный отбор, действующий на сайты определённого

класса, увеличивает вероятность закрепления новых вариантов в этих сайтах, что увеличивает количество замен в них. Таким образом, dn/ds будет больше 1 в участках исключительного действия положительного отбора. Отрицательный отбор действует против нового варианта, уменьшая вероятность его закрепления, что уменьшает количество замен в сайте, а также увеличивает время фиксации слабовредных вариантов, т.е. вариантов, на которые действие отрицательного отбора недостаточно велико, чтобы предотвратить их фиксацию. Таким образом, dn/ds меньше 1 в участках отрицательного отбора, что способствует сохранению аминокислотной последовательности белка.

2.2 Тест Макдональда-Крейтмана

Отбор также действует на расщепление полиморфизма в популяции. Данный факт можно использовать для выявления следов отбора. В тесте Макдональда-Крейтмана [30] по аналогии с сайтами дивергенции (D_n и D_s – фиксированные замены между двумя сайтами) вводится понятие полиморфных сайтов – позиций, в которых наблюдается расщепление различных аллелей внутри популяции. По влиянию на приспособленность все мутации делятся на 3 категории: нейтральные, вредные и благоприятные. В нейтральном случае все замены имеют равный шанс закрепиться, поэтому отношение частот несинонимичных полиморфизмов P_n и синонимичных полиморфизмов P_s будет совпадать с отношением частот замен на сайт соответствующей категории. При действии отрицательного отбора D_n/D_s будет снижен по сравнению с P_n/P_s, так как

вредные мутации будут иметь меньший шанс закрепиться, чем нейтральные; напротив, при действии положительного отбора благоприятные мутации будут фиксироваться быстрее, и D_n/D_s будет выше P_n/P_s.

В тесте D_n/D_s нулевой гипотезой, говорящей о нейтральной эволюции, считалось D_n/D_s =1. Это условие оказывается слишком консервативным, так как дивергенция несинонимичных сайтов обычно оказывается ниже, чем синонимичных. Таким образом, D_n/D_s = 1 при P_n/P_s = 0,5 будет говорить о большой доле адаптивных замен. Можно показать, что число адаптивных замен рассчитывается по формуле [31]:

$$\alpha = 1 - \frac{P_n/P_s}{D_n/D_s}$$

Однако даже в тесте Макдональда-Крейтмана число адаптивных замен остаётся существенно недооценённым из-за слабовредных замен, расщепляющихся в популяции на низких частотах [32].

2.3 Тест двойных замен

Рассмотрим однонуклеотидные мутации, которые происходят после расхождения двух линий. Двойные мутации (мутации в двух соседних сайтах) в случае нейтральности будут происходить в 50% случаев в одной и той же линии, а в 50% случаев – в разных линиях. Избыток замен, произошедших в одной и той же линии, будет говорить о наличии положительного отбора [33]. Долю двойных замен δ, произошедших под действием положительного отбора, можно оценить по формуле

$$\delta = 1 - f(P_1)/(2l_1l_2),$$

где $f(P_1)$ – доля случаев, когда 2 замены происходят в одной и той же линии, а l_1 и l_2 число замен на нуклеотидный сайт, которое произошло в каждой из ветвей с момента их расхождения [34].

3. Отбор в сцепленных локусах

Сцепленное наследование генетической информации приводит к тому, что действие отбора на определённый участок генома распространяется также и на соседние участки. Например, быстрое закрепление положительной мутации часто ведёт к закреплению сцепленных с ней мутаций, которые могут быть нейтральными или даже славбовредными ("hitchhiking effect") [35,36]. И наоборот, присутствие вредного аллеля в локусе может препятствовать распространению в популяции сцепленных с ним вариантов (background selection) [37–39]. Действие отбора на сцепленные сайты наиболее выражено в участках с низкой частотой рекомбинации и быстро ослабевает с ростом частоты рекомбинации [40–42]. Как следствие, в участках с низкой частотой рекомбинации наблюдается снижение эффективности отрицательного отбора [42]; кроме того, такие участки имеют пониженный полиморфизм [43–46].

4. Генная конверсия

Одной из проблем при изучении естественного отбора является смещённая генная конверсия, так как такая конверсия, как и отбор, меняет вероятности закрепления мутаций по сравнению с нейтральным ожиданием. Генной конверсией называют направленный перенос информации с одной цепи ДНК на другую. В большинстве случаев генная конверсия в геномах высших организмов происходит в процессе репарации гетеродуплексов, возникающих в процессе мейоза, в результате чего потомки несут один и тот же вариант. Чем выше частота рекомбинации, тем больше происходит конверсионных событий. Известно, что репарация неспаренных оснований ДНК чаще происходит в пользу гуанина и цитозина [47]. В дрожжах *Saccharomyces cerevisiae* репарация гетеродуплексов S=W, возникающих в результате мейоза, происходит в сторону GC в 50,62% случаев [48]. В клетках млекопитающих проводились эксперименты с репарацией ДНК, содержащей специально сконструированные гетеродуплексы. По этим данным репарация гетеродуплексов S=W происходит в сторону GC в 78,5% случаев, но величина смещения сильно зависит от типа гетеродуплекса [49]. Несмотря на то, что эффект конверсии является исключительно мутационным, в тестах он может приводить к смещённым оценкам положительного отбора.

В наше работе мы исследовали, какой эффект оказывает генная конверсия на инсерции и делеции – пару антагонистических мутаций, аналогичную паре S→W и W→S мутаций.

5. Адаптивный ландшафт

Ландшафтом приспособленности (адаптивным ландшафтом) называют функцию, которая каждому возможному генотипу или фенотипу ставит в соответствие определённое значение приспособленности. Поскольку естественный отбор действует на различия в приспособленности между генотипами, отображение генотип-приспособленность можно рассматривать как один из ключевых факторов, определяющих ход эволюции. В самом деле, эволюцию можно представить как последовательные перемещения эволюционирующего объекта по ландшафту приспособленности [50–54]. Поскольку вредные мутации редко закрепляются между видами, закрепившиеся мутации в последовательности ДНК, как правило, либо не приводят к существенному изменению приспособленности (такие мутации называют нейтральными), либо ведут к увеличению приспособленности (адаптивные). Поэтому популяции могут "застревать" на локальных максимумах приспособленности, не добираясь до глобального пика. Одним из способов обхода этого препятствия для эволюции могут быть радикальные изменения положения – "прыжки" в пространстве приспособленности.

Рисунок 1 схематично изображает последовательность событий, связанных с крупномасштабным перемещением («прыжком») эволюционирующего объекта в пространстве генотипов [55,56], при условии, что исходно данный объект находился на одном из локальных пиков приспособленности (синяя точка на Рисунке 1). Такой «прыжок», возможно, соответствующий мутации со

значительным эффектом, может быть адаптивным (повышать приспособленность, зелёная точка на Рисунке 2), нейтральным (не менять приспособленность) или слабовредным (приводить к небольшому снижению приспособленности). В любом из вариантов после такого «прыжка» маловероятно, что новый генотип окажется в точности на пике приспособленности. Скорее всего, он окажется на склоне нового пика, причём с высокой вероятностью новый пик будет выше старого, поскольку изменения, приводящие к значительному снижению приспособленности, не закрепляются в эволюции. Таким образом, вслед за «прыжком» мы можем ожидать определённое число адаптивных изменений, приводящих, в конечном итоге, эволюционирующий объект на новый адаптивный пик (жёлтая точка).

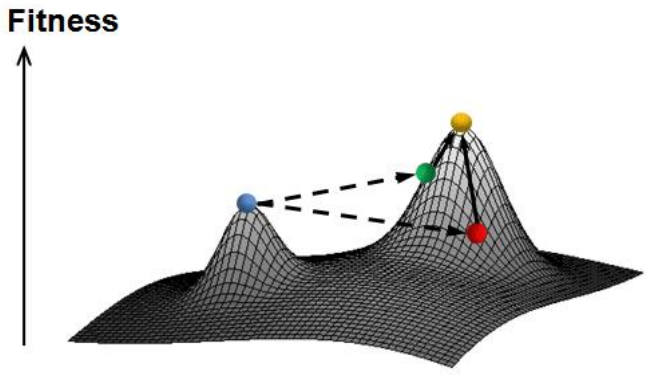

Рис. 2. Эволюционные траектории, которые сопряжены с длинным прыжком в пространстве генотипов. На рисунке показана схема адаптивного ландшафта. Исходный объект находится на локальном пике приспособленности (синий). Радикальное изменение, как, например, инсерция или делеция в белковой последовательности (пунктирные стрелки), может переместить объект на склон

нового пика (зеленый и красный). Это вызывает адаптивную «прогулку», состоящую из последовательности малых изменений, таких как аминокислотные замены (сплошные стрелки), которая в конечном итоге приводит объект на адаптивный пик (жёлтый).

Основываясь на комбинаторных соображениях, Дж. Гиллеспи показал, что последовательность аллельных замещений в эволюционирующем белке, происходящих под положительным отбором (так называемая «адаптивная прогулка» – adaptive walk), должна включать 2–5 шагов [57,58]. Эта оценка была получена в предположении, что адаптивные ландшафты являются некоррелированными, что не выполняется для реальных белков, в которых схожие последовательности имеют близкие значения приспособленности. В гладких коррелированных ландшафтах количество субоптимальных пиков меньше, и ожидается более длинная прогулка [59]. С другой стороны, многочисленные исследования [60–63] свидетельствуют о том, что ландшафты приспособленности для белков и тРНК гладкими отнюдь не являются: отбор часто является эпистатическим в смысле того, что приспособленность аллеля в определённом локусе зависит от других участков генома [64]. Таким образом, ландшафт приспособленности биологических объектов часто является коррелированным и негладким, что сильно осложняет предсказание длины адаптивных прогулок. Поэтому измерение длины и длительности таких прогулок

могло бы во многом способствовать пониманию структуры адаптивных ландшафтов.

Адаптивная прогулка, как отмечалось ранее, может быть вызвана крупномасштабным изменением [50,51]. В случае с белками, инсерции или делеции, которые представляют собой более существенные изменения в структуре белка, чем однонуклеотидные замены, могут быть связаны с «прыжками» на ландшафте приспособленности и вызывать последующие аминокислотные замещения, соответствующие адаптивным изменениям.

Материалы и методы

1. Геномные данные

Полногеномные выравнивания 11 видов *Drosophila* с *D. melanogaster* [26], выравнивания *Pan troglodytus* и *Pongo pigmaeus* с *Homo sapiens*, а также выравнивания *S. paradoxus* и *S. mikatae* с *S. cerevisiae* были загружены из базы данных UCSC http://hgdownload.cse.ucsc.edu/. Кодирующие последовательности были вырезаны из выравниваний по аннотации базы данных FlyBase для выравниваний *Drosophila* [65], по аннотации базы данных UCSC для выравниваний приматов [66] и по аннотации базы данных Ensembl для выравниваний дрожжей [67]. Данные по однонуклеотидному полиморфизму для *D. melanogaster* были получены по полным генотипам 158 инбредных линий, загруженных из http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/sequences/. Данные по частотам рекомбинации для генома *D. melanogaster* были получены из [68] (разрешение до 2Кб). Данные по частотам рекомбинации для генома *H. sapiens* получены из [69] (разрешение ~0,6Мб). В качестве показателя частоты рекомбинации в *S. cerevisiae* использовали данные связывания ДНК белком Spo11 [70].

2. Идентификация закрепившихся инсерций и делеций в участках дрозофил, приматов и дрожжей.

Для идентификации закрепившихся инделов и однонуклеотидных мутаций в геномах дрозофил, приматов и дрожжей использовались множественные выравнивания референсных геномов. *D. sechellia* и *D. erecta* использовались для определения предкового состояния у *D. melanogaster*, *Pan troglodytus* и *Pongo pigmaeus* – у *Homo sapiens*, *S. paradoxus* и *S. mikatae* – у *S. cerevisiae*. Сайты, в которых не удавалось определить предковое состояние, исключались из анализа.

3. Идентификация закрепившихся инсерций и делеций в белок-кодирующих участках последовательностей дрозофил для анализа изменений в адаптивном ландшафте на разных филогенетических расстояниях.

Для идентификации сайтов инделов были использованы сравнения референсных последовательностей 6 видов *Drosophila*, а именно *D. melanogaster*, *D. sechellia*, *D. erecta*, *D. ananassae*, *D. pseudoobscura* и *D. virilis*. Анализировались инделы с длиной, кратной 3 нуклеотидам, то есть не сдвигающие рамку (если индел попадал на границу экзона, учитывалась только экзонная часть индела). Инделы укоренялись последовательностями *D. pseudoobscura* и *D. virilis*, как показано на Рисунке 3; инделы, противоречащие представленным филогенетическим конфигурациям, исключались из анализа. Чтобы избежать участков выравнивания с низким качеством, мы требовали, чтобы ни одна из 6

анализируемых последовательностей не содержала гэпов или отличных от ATCG-нуклеотидов символов в участке ±10 нуклеотидов от сайта индела.

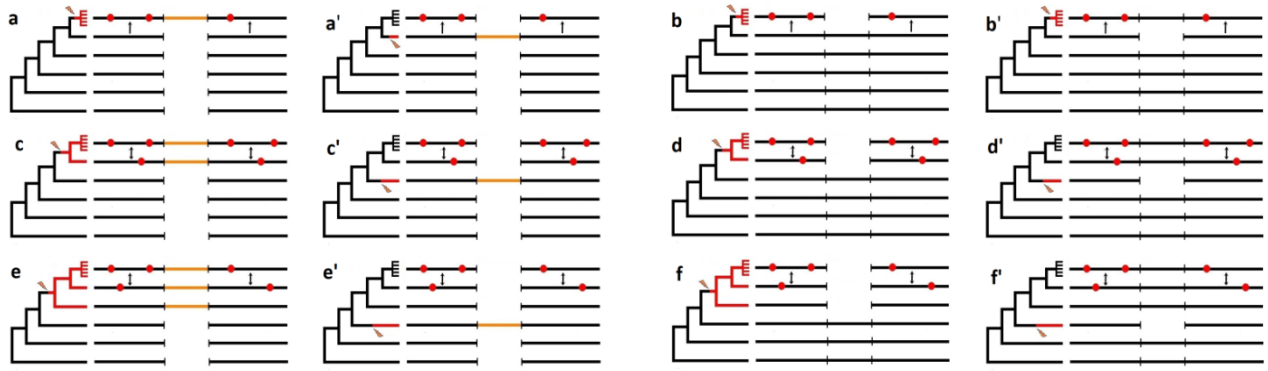

Рис. 3. Филогенетическая схема аминокислотных замен и инделов в белках дрозофил, используемых в анализе. На каждой части рисунка слева изображена филогения (((((*D. melanogaster*, *D. sechellia*), *D. erecta*), *D. ananassae*), *D.pseudoobscura*), *D.virilis*); время происхождения индела отмечено молнией, а сегмент эволюционного древа, содержащий индел, выделен красным. Для *D. melanogaster* имеются данные по полиморфизму (обозначено гребёнкой). На рисунках a, a', c, c', e, e' изображены инсерции, а на рисунках b, b', d, d', f, f' – делеции. На рисунках a – b' (нижний ряд) представлены инделы, которые произошли до ответвления *D. melanogaster* от *D. sechellia*, а на рисунках c – d' – инделы, которые произошли на участке дерева между ответвлениями *D. erecta* – (*D. melanogaster* – *D. sechellia*) и *D. melanogaster* – *D. sechellia*; наконец, рисунки e – f' представляют инделы, которые произошли на участке дерева между *D. ananassae* – ((*D. melanogaster*, *D. sechellia*), *D. erecta*) и *D. erecta* – (*D. melanogaster* – *D. sechellia*). На рисунках

a, b, c, d, e, f представлены инделы, ведущие к линии *D. melanogaster*, а на рисунках a', b', c', d', e', f' – инделы, произошедшие в боковых ветвях. Стрелка с двумя наконечниками обозначает 2 сравниваемых участка, использованных для подсчёта числа замен; стрелка с одним наконечником обозначает участок, на котором замены считались с помощью поляризации. Красными точками обозначены нуклеотидные замены.

4. Идентификация полиморфных инделов в *D. melanogaster*

Поиск полиморфных инделов осуществляли программой mpileup из пакета SAMtools [27] (v. 0.1.17, http://samtools.sourceforge.net/). Поиск был произведён на данных по полногеномному секвенированию, полученных из http://www.hgsc.bcm.tmc.edu/projects/dgrp/freeze1_July_2010/Illumina/. Идентифицированные инделы фильтровались согласно их весу Phred – брались только инделы с различием веса Phred для генотипов >10. Участки выравнивания, содержащие полиморфные в *D. melanogaster* инделы, были перевыравнены программой MUSCLE (v. 3.7) в целях улучшения качества выравнивания с последовательностями других видов. *D. sechellia* и *D. erecta* использовались для укоренения инделов; инделы исключались, если последовательности этих двух видов противоречили друг другу или если более 50% особей *D. melanogaster* не имело данных по секвенированию для данного участка генома. Количество идентифицированных инделов в различных участках генома отображено на Рисунке 4.

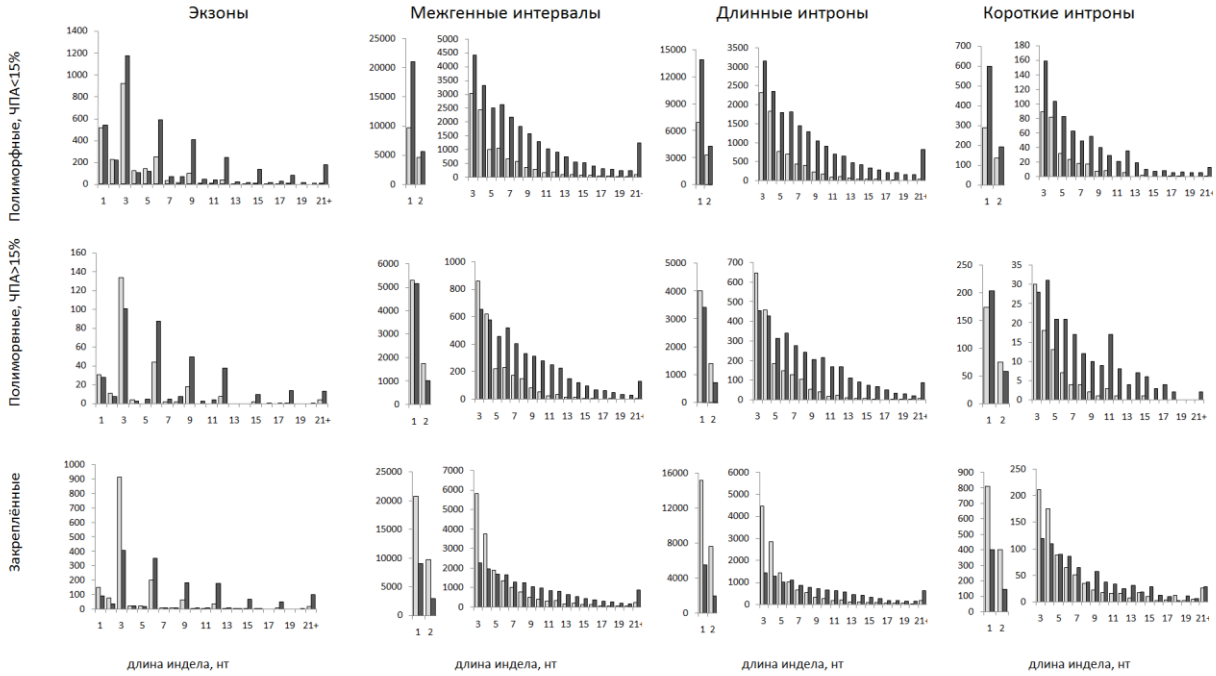

Рис. 4. Распределение длин полиморфных и закрепившихся инделов и однонуклеотидных замен в различных участках генома. Верхний ряд, полиморфизмы с частотой производного аллеля (ЧПА) < 15%; средний ряд, полиморфизмы с ЧПА > 15%; нижний ряд, закрепившиеся мутации. Светло-серым обозначены инсерции, тёмно-серым обозначены делеции.

Гэпы в выравниваниях, интерпретируемые нами как инделы, могут возникать также из-за ошибок при секвенировании или сборке. Доля таких ошибочных инделов должна быть наибольшей среди инделов, сдвигающих рамку, так как реальных инделов в этой категории должно быть мало из-за сильного отрицательного отбора против них. Чтобы оценить верхний порог на частоту ошибок, мы использовали отношение числа инделов с длинами, некратными трём нуклеотидам, в экзонах к числу таких инделов в межгенных участках. Это

отношение равно ~0,05 для инсерций и делеций с частотами ниже 15%, и лишь 0,007 – для инсерций и делеций с частотой выше 15%. Из-за того, что отбор менее эффективен на низких частотах, данные оценки для доли ошибок являются сильно завышенными.

5. Оценки относительных скоростей мутагенеза инделов

Данные по полиморфизму могут быть использованы для оценки скоростей возникновения мутаций различных типов, если выбрать такие полиморфизмы, что влияния отбора и смещённой генной конверсии на них пренебрежимо малы в сравнении с действием генетического дрейфа (который не смещает относительные частоты мутаций разного типа). Такому условию удовлетворяют полиморфизмы, которые расщепляются на очень низких частотах [71]. Мы использовали полиморфные нуклеотидные сайты, в которых производный аллель наблюдался в 1–4 особях из выборки в 158 линий *D. melanogaster*, то есть частота производного аллеля (ЧПА) <3%. Короткие интроны являются наименее отбираемыми последовательностями в геноме [27,28,42,72]. На этом основании для оценки скоростей мутагенеза мы брали только сайты из интронов с длинами 70–300 нуклеотидов.

6. Оценка интенсивности отрицательного отбора

Отрицательный отбор уменьшает количество полиморфных сайтов в сравнении с числом, ожидаемым при отсутствии отбора. Даже слабый отрицательный отбор ($-5 < N_e s < -1$, где s – коэффициент отбора, а N_e –

эффективная численность популяции) может существенно уменьшить число высокочастотных (ЧПА>15%) полиморфизмов; однако только сильный отрицательный отбор ($N_e s$ < -20) может значительно уменьшить число низкочастотных (ЧПА<15%) полиморфизмов [71]. Для того, чтобы оценить долю *de novo* мутаций, отсеянных действием отрицательного отбора в определённом участке генома, мы сравнили число низкочастотных полиморфизмов в этом участке с числом таких полиморфизмов в коротких интронах, где силы отбора минимальны. Для инделов мы проводили такое сравнение отдельно для разных длин; в частности, инделы с длинами, кратными 3 нт, в коротких интронах использовались как контроль для инделов с длинами, кратными 3 нт, в экзонах, а инделы, некратные 3 нт в коротких интронах – как контроль для инделов, некратных 3 нт, в экзонах. Чтобы получить ожидаемые количества миссенс- и нонсенс-мутаций в экзонах, мы брали случайные тройки нуклеотидов из коротких интронов с теми же частотами, что и частоты кодонов в кодирующих последовательностях.

Силу отрицательного отбора, действующего на расщепляющиеся в популяции мутации, можно оценить из спектра аллельных частот. Для расчёта отклонений спектра мутаций от нейтрального мы используем следующую статистику:

$$\xi = \frac{P_{SH}/P_{SL}}{P_{NH}/P_{NL}} \qquad (1)$$

Здесь P – число полиморфных сайтов в соответствующем классе частот; индексы S и N относятся к сайтам под отбором (**S**elected) и нейтральным сайтам (**N**eutral)

соответственно; индексы H и L относятся к высокочастотным (**H**igh-frequency) и низкочастотным (**L**ow-frequency) интервалам частот производного аллеля. Если распределение частот аллелей для исследуемого участка совпадает с таковым для нейтральных, что свидетельствует о нейтральности, ξ будет равно 1. $\xi < 1$ означает, что частоты аллелей для данного класса мутаций ниже таковых для нейтральных мутаций, что свидетельствует об отрицательном отборе. Например, $\xi = 0.3$ подразумевает, что среди мутаций, наблюдаемых в низкочастотном (L-) интервале в исследуемой выборке, 70% не достигают высокочастотного (H-) интервала из-за действия отрицательного отбора.

Мы рассчитали ξ(L,H) для инделов и однонуклеотидных замен. Для того, чтобы облегчить сравнение инделов в кодирующих и некодирующих областях, здесь были рассмотрены только инделы с длинами, кратными 3. Так как не существует общепризнанного нейтрального стандарта для инсерций и делеций, мы использовали однонуклеотидные полиморфизмы и замены в участках с 8 по 30 нт в интронах с длинами до 65 нт в качестве меры P_N как для инделов, так и для однонуклеотидных замен [27,28,42]. Мы использовали ЧПА<15% как L-интервал, а ЧПА>15% – как H-интервал. Тест Макдональда-Крейтмана может давать смещённую оценку из-за слабовредных замен, расщепляющихся на низких частотах; для избежания этого эффекта рекомендуется исключать полиморфные сайты с аллельными частотами ниже 15% [32]. Этот же порог применим и здесь, так как слабовредные мутации редко достигают таких высоких частот, и подавляющее большинство мутаций на этих частотах нейтральны.

7. Оценка интенсивности положительного отбора

Для анализа положительного отбора, действующего на участки гена, соседние с инделом использовали тест МакДональда-Крейтмана. Мы брали кодоны с 1 по 100 слева и справа от сайта индела и разбивали на подгруппы по 10 коднов. В анализе были использованы только те нуклеотидные сайты, которые в последовательностях всех шести анализируемых видов не несли гэпы (кроме связанных с гэпами исследуемого индела) или отличные от ATCG-нуклеотидов символы в участке ±10 нуклеотидов. Только четырёхкратно вырожденные (невырожденные) сайты были использованы для определения числа синонимичных (несинонимичных) замен. Синонимическая дивергенция D_s, несинонимическая дивергенция D_n, синонимический полиморфизм P_s и несинонимический полиморфизм P_n рассчитывались как доля несовпадений в соответствующих сайтах. Для инделов, произошедших в терминальном сегменте линии *D. melanogaster* (Рисунок 2 c-f'), учитывались только замены в сайтах, которые совпадали между *D. sechellia* и *D. erecta*. Для инделов, которые произошли в более ранние периоды дивергенции, и *D. melanogaster* и *D. sechellia* подвергались влиянию индела. Таким образом, мы рассчитывали D_s и D_n как долю различий между *D. melanogaster* и *D. sechellia*.

Доля несинонимичных замен, закреплённых под действием положительного отбора, оценивалась по данным, представленным в табл. 2, как $\alpha = 1 - (P_n/P_s)/(D_n/D_s)$ [31]. Как и в случае с отрицательным отбором мы исключали

полиморфные сайты с аллельными частотами ниже 15%, чтобы избежать занижения α слабовредными мутациями [32].

8. Оценка интенсивности генной конверсии, смещённой в сторону инсерций.

Отношение закрепившихся мутаций к низкочастотным (ЧПА<3%) (ОЗНЧ) использовалось в качестве показателя вероятности закрепления мутации. В участках генома с очень низкой скоростью рекомбинации < 0,01 сМ/Мб вероятность закрепления инделов была много меньше таковой для нейтральных мутаций, подразумевая отрицательный отбор; однако она практически не зависела от длины индела (Рисунок 2А), подразумевая, что отбор не влияет на распределение длин коротких инделов. Помимо этого, отношение числа закрепившихся инсерций к числу закрепившихся делеций для инделов с длинами 5–10 нт не зависит от рекомбинации (Рисунок 1В, нижний ряд), подразумевая отсутствие смещённой генной конверсии для инделов с такими длинами. Исходя из вышеперечисленного, увеличение (уменьшение) вероятности закрепления инсерций (делеций) длин 1–4 нуклеотида в сравнении с инсерциями (делециями) длин 5–10 нуклеотидов в участках с скоростью рекомбинации > 0,01 сМ/Мб мы приписали действию генной конверсии, смещённой в сторону инсерций. Общегеномное увеличение (уменьшение) числа закрепившихся инсерций (делеций) длины ℓ, вызванное действием конверсии было рассчитано как

$$\left(\frac{\text{Наблюдаемое}}{\text{Ожидаемое}}\right)_{\ell} = \frac{\sum_{i=1}^{6}(\text{ИНДЕЛ}_{\text{о}}(\ell;i))}{\sum_{i=1}^{6}(\text{ИНДЕЛ}_{\text{о}}(\ell;i)\frac{\text{ОЗНЧ}(5-10\text{нт};i)}{\text{ОЗНЧ}(\ell;i)})}$$

Здесь, $\text{ИНДЕЛ}_{\text{о}}$ (ℓ,i) – наблюдаемое число инделов, а ОЗНЧ(ℓ,i) – отношение числа закреплённых мутаций к низкочастотным (ЧПА < 3%) для инделов длины ℓ в интервале рекомбинации i.

9. Расчёт длины адаптивной прогулки

Длина адаптивной прогулки определялась как число адаптивных замен в терминальном сегменте линии *D. melanogaster* и вычислялась следующим образом. Для инделов, которые произошли в терминальном сегменте линии *D. melanogaster* (Рисунок 2a, b) или линии *D. sechellia* (Рисунок 2a', b'), мы считали число аминокислотных различий между *D. melanogaster* and *D. sechellia* в сайтах, совпадающих между *D. sechellia* and *D. erecta*, учитывая таким образом только замены, специфичные для *D. melanogaster*. Для инделов, которые произошли на более ранних этапах (Рисунок 2c-f), мы считали число аминокислотных различий между *D. melanogaster* и *D. sechellia*, и делили пополам, предполагая равные длины веток. Итоговый путь рассчитывался как разница между суммарным числом замен, которые произошли в ветке *D. melanogaster*, для случаев, когда индел также был в ветке *D. melanogaster* и когда индел произошёл в сестринской ветке.

95% доверительные интервалы для всех оценок был получены методом бутстрепа с возвращением. В каждом случае использовалось 1000 бутстреп-реплик.

10. Анализ эволюции в аминокислотных сайтах с различной консервативностью

Аминокислотный сайт считался консервативным, если кодируемая аминокислота была одинакова для всех следующих видов *Drosophila*: *D. pseudoobscura*, *D. persimilis*, *D. willistoni*, *D. virilis*, *D. mojavensis* и *D. grimshawi*. Все другие сайты считались неконсервативными.

11. Теоретическое распределение частот аллелей.

Ожидаемое распределение частот аллелей для случая нейтральности было получено из формулы $f(x) = 1/x$, задающей отношения времён, которые производный аллель проводит на каждой частоте [71,73].

Глава 1. Анализ инсерций и делеций в популяции D. melanogaster

1.1 Оценка относительных скоростей мутагенеза на данных по низкочастотным инделам

Данные по низкочастотному полиморфизму могут быть использованы для оценки относительных скоростей мутагенеза, так как на низких аллельных частотах действие отбора или генной конверсии ничтожно мало в сравнении с дрейфом [71]. Короткие интроны находятся под более слабым отбором, чем любой другой участок генома [27,72]. По этой причине мы использовали интроны с длинами до 300 нуклеотидов для оценки скоростей мутагенеза. Среди полиморфизмов с очень низкими частотами (такими, что производный аллель наблюдался в 1–4 генотипах из 158 особей *D. melanogaster*) (Рисунок 5A), частота инсерций (делеций) составляет ~0,041 (~0,091) от частоты нейтральных однонуклеотидных замен (10^{-8} на нуклеотид на поколение; [74]). Длина средней низкочастотной инсерции (делеции), обнаруженной в коротких интронах, составляет ~3,10 (~3,54) нуклеотида.

Однако данные оценки скоростей мутагенеза могут быть заниженными, так как значительное число делеций в коротких интронах попадают на экзон-интронную границу либо на сайт сплайсинга, а также из-за отбора против очень коротких интронов [72]. По этой причине был взят следующий по нейтральности участок генома – межгенные интервалы [26]; число низкочастотных однонуклеотидных полиморфизмов на сайт в межгенных интервалах (0,015)

близко к таковому для коротких интронов (0,014) (Рисунок 5B). Сравнение низкочастотных мутаций показало, что частоты инсерций и делеций составляют соответственно ~0,037 и ~0,095 от частоты нейтральных однонуклеотидных замен, а их средние длины составляют 3,46 нуклеотидов для инсерций и 5,08 нуклеотидов для делеций. Различия длин инделов в коротких интронах и межгенных интервалах согласуются с более сильным отбором против длинных делеций в коротких интронах.

Таким образом, данные по межгенным интервалам показывают, что в отсутствие отбора инсерции с длинами 1–60 нуклеотидов удлиняли бы геном на 0,13 нуклеотида в расчёте на одну нейтральную замену; в то же время делеции сокращали бы геном на 0,48 нуклеотида, что в сумме приводило бы к потере 0,35 нуклеотида на каждую нейтральную замену. Стоит отметить, что даже в самом нейтральном участке генома (коротких интронах) отбор против инсерций и, в особенности, против делеций сильнее, чем против однонуклеотидных замен (см. раздел 1.2); таким образом, данные оценки могут быть занижены.

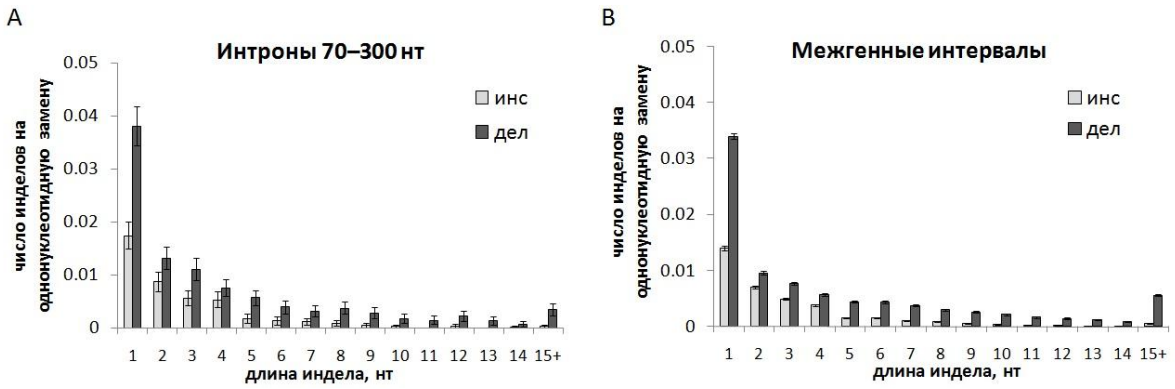

Рис. 5. Количество инсерций и делеций разных длин в расчёте на однонуклеотидную замену для мутаций с очень низкими частотами (1–4 генотипа из 158) в коротких интронах (A) и межгенных интервалах (B).

1.2 Отрицательный и положительный отбор на инделы в различных участках генома.

Отрицательный отбор уменьшает количество полиморфных сайтов, но только сильный отрицательный отбор ($N_e s$ < -20) может существенно снизить число низкочастотных (т.е. частота производного аллеля или ЧПА< 15%) вредных мутаций (fig.1 в [71]). Мы можем оценить сильный отрицательный отбор, сравнивая встречаемость низкочастотных полиморфизмов в различных участках генома (Рисунок 6). Низкочастотные инделы обладают практически одинаковой встречаемостью в интронах различных длин и в межгенных интервалах, предполагая тем самым отсутствие сильного отбора на *de novo* мутации (Рисунок 2B-D, светло- и тёмно- серые столбики). Также примерно одинаковые уровни встречаемости наблюдаются в этих участках генома и для однонуклеотидных мутаций (Рисунок 6B-D, синие столбики).

Напротив, в экзонах даже низкочастотные инделы встречаются значительно реже (Рисунок 6A). Количество инделов снижено в экзонах более чем в 2 раза по сравнению с остальными участками генома. Встречаемость несинонимичных замен также снижена и составляет ~0,35 от числа однонуклеотидных замен в

данном типе участков генома. Наиболее радикальное снижение встречаемости полиморфизмов наблюдается для нонсенс-мутаций: число сдвигающих рамку инсерций (делеций) составляет всего лишь ~0,09 (~0,06) от наблюдаемого в коротких интронах; число однонуклеотидных мутаций, ведущих к возникновению стоп-кодонов, также снижено и составляет ~0,09 от ожидания (Рисунок 7). Наблюдаемое снижение, вызванное сильным отбором, не зависит от длины индела по крайней мере для длин <20 нт (Рисунок 8).

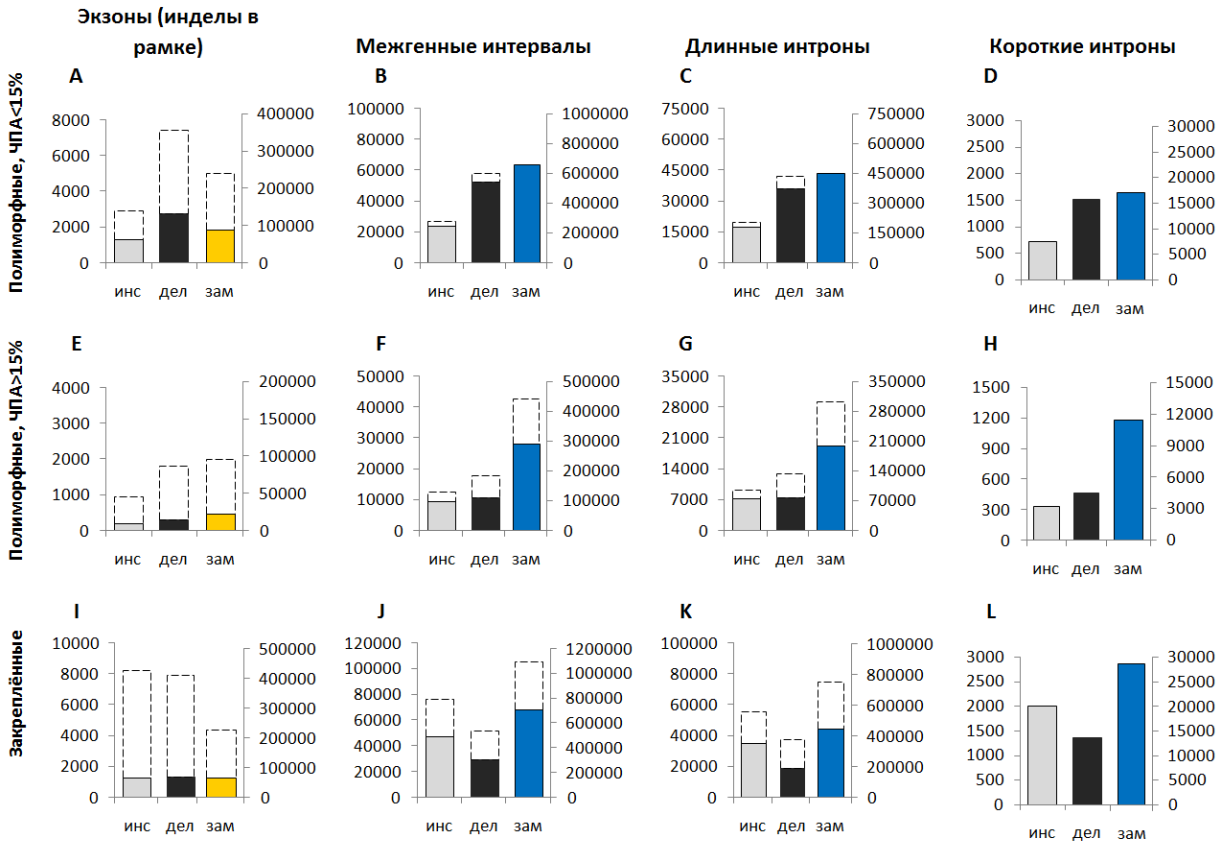

Рис. 6. Число полиморфных и закреплённых инделов и однонуклеотидных замен в различных участках генома. Верхний ряд (A–D) _ полиморфизмы с ЧПА <15%; средний ряд (E–H), полиморфизмы с ЧПА >15%; нижний ряд (I–L),

фиксированные мутации. A, E, I: кратные трём инделы и миссенс-замены в экзонах; B, F, J: мутации в межгенных интервал; C, G, K: мутации в длинных (>300 нуклеотидов) интронах; D, H, L: мутации в коротких (70–300 нуклеотидов). Светло-серый соответствует инсерциям; тёмно-серый соответствует делециям; жёлтый соответствует миссенс-заменам; синий, однонкулеотидным заменам в некодирующих участках. На каждом графике левая вертикальная ось показывает количество инделов, правая вертикальная ось показывает количество однонуклеотидных замен.

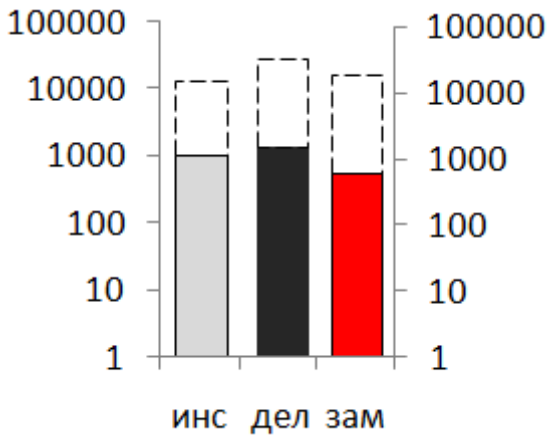

Рис. 7. Количество полиморфных инделов с длинами не кратными 3 и полиморфные однонуклеотидные замены, ведущие к образованию стоп-кодона (нонсенс-мутации), с ЧПА < 15%. Светло-серым показаны инсерции (инс), тёмно-серым показаны делеции (дел), красным показаны нонсенс-мутации (зам). Штриховкой отмечено количество полиморфизмов, ожидаемое при нейтральной эволюции (посчитано по коротким интронам).

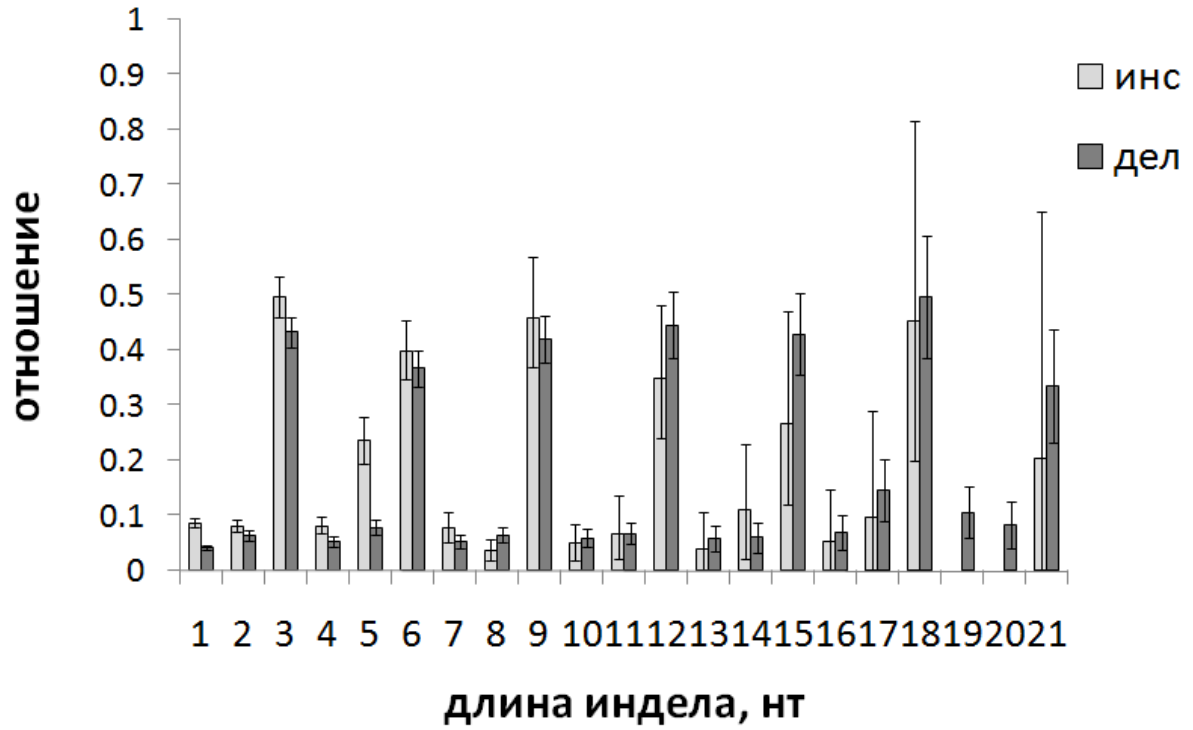

Рис. 8. Отношение числа низкочастотных (ЧПА < 15%) инсерций (инс) и делеций (дел) в экзонах к числу низкочастотных инделов в межгенных интервалах. 95% доверительные интервалы построены по 1000 бутстреп-репликам.

Слабый отрицательный отбор (-1 > $N_e s$ > -5), будучи неэффективным на низких частотах, существенно снижает встречаемость высокочастотных полиморфизмов и фиксированных мутаций. В самом деле, встречаемости высокочастотных и фиксированных инделов, а также высокочастотных однонуклеотидных замен, наиболее высоки в коротких интронах, но ниже в других участках генома (Рисунок 6*E-L*). Поскольку оценки отрицательного отбора, основанные только на дивергенции, могут быть затруднены из-за

действия положительного отбора, отрицательный отбор лучше оценивать по полиморфным мутациям.

Сравнение спектров аллельных частот не зависит от скорости мутагенеза (поскольку множественные мутации в сайте редки, и редко приводят к одновременному расщеплению нескольких производных аллелей разного происхождения). Это позволяет нам сравнивать мутации разных типов и использовать однонуклеотидный полиморфизм в позициях 8–30 в интронах с длинами до 65 нт как нейтральный стандарт для всех типов мутаций. Мы используем ξ, отношение высокочастотных полиморфизмов к низкочастотным для потенциально отбираемых мутаций, нормированное на то же самое отношение, но для нейтрального класса мутаций, как верхнюю границу нейтральных мутаций, а 1- ξ – как нижнюю границу на число слабовредных мутаций (см. разд. Материалы и методы).

Сравнение ξ для разных типов мутаций показывает, что в каждом участке генома инделы более вредны, чем однонуклеотидные замены, при этом делеции вреднее инсерций (Рисунок 9А). Доля вредных мутаций среди полиморфизмов является наибольшей в экзонах: по крайней мере ~79% инсерций и ~88% делеций являются вредными; для сравнения, доля вредных среди расщепляющихся однонуклеотидных миссенс-замен составляет ~72% . В межгенных интервалах и в длинных интронах доля вредных мутаций среди инсерций равна ~71%, что меньше, чем среди делеций (~82%), но много больше, чем среди однонуклеотидных замен (~49%) (Рисунок 9А). Отбор оказался наиболее слабым

для всех типов мутаций в коротких интронах (Рисунок 9A); однако даже в этом участке генома инделы гораздо чаще являются вредными (в ~68% случаях для инсерций и ~73% случаев для делеций), чем однонуклеотидные замены (~23%).

Очень короткие интроны (с длинами <60 нт) заслуживают отдельного рассмотрения [72]: ограничение на минимальную длину интрона ведёт здесь к дополнительному отбору на делеции. В самом деле, зависимость отбора против делеций от длины интрона не является монотонной: доля вредных делеций выше в длинных и в очень коротких интронах, чем в интронах с медианной длиной. Она является наибольшей (~81%) для делеций, которые делают интрон короче 50 нуклеотидов. Напротив, доля вредных инсерций является наименьшей в этом классе интронов (Рисунок 9B).

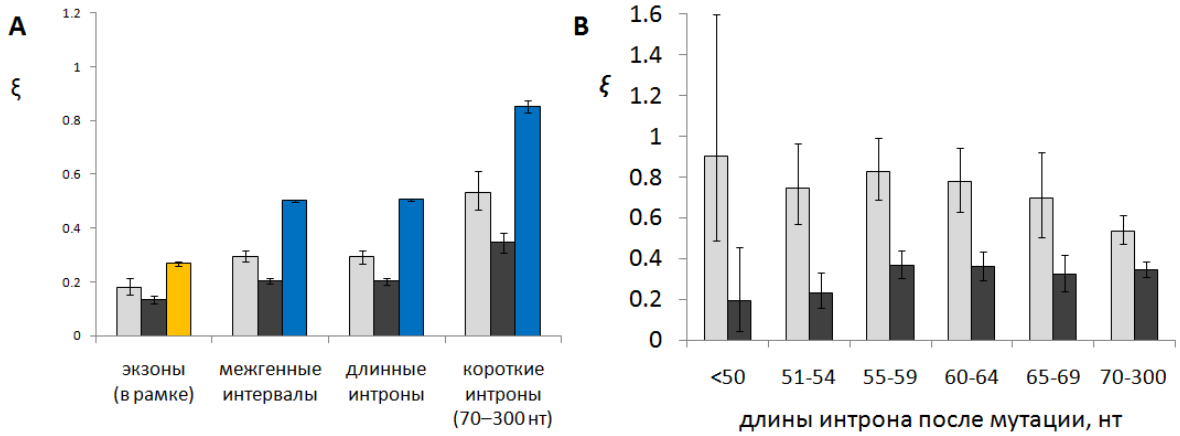

Рис. 9. ξ для инделов и однонуклеотидных замен. Низкие значения ξ соответствуют высокой доле вредных мутаций и наоборот. (A), средние значения ξ в различных участках генома: экзонах (кратные трём инделы и миссенс-замены), межгенные интервалы, длинные (>300 нт) и короткие (70–300 нт) интроны. (B), ξ

для инделов в коротких и очень коротких интронах. Светло-серым показаны инсерции; темно-серым – делеции; жёлтым – миссенс-замены; синим – однонуклеотидные замены в некодирующих участках генома. Доверительные интервалы построены по 1000 бутстреп-репликам.

Часть мутаций находится под действием положительного отбора. Рисунок 10 показывает, что положительный отбор на инделы распространён повсеместно по геному *D. melanogaster*. Доля α инсерций, фиксированных под действием положительного отбора, очень велика во всех участках генома: она достигает ~61% в экзонах и ~67% в межгенных интервалах. Для делеций α составляет ~48% в экзонах, но ниже в других участках генома: 32–36%. Оценки α для однонуклеотидных замен значительно ниже: ~15% для миссенс-замен, и около 0% для однонуклеотидных замен в некодирующих участках.

Наши оценки α для однонуклеотидных замен значительно ниже полученных в других исследованиях [28,29]. Это объясняется двумя факторами. Во-первых, при вычислении дивергенции мы использовали только сайты, мономорфные в *D. melanogaster,* тогда как в большинстве других работ дивергенция считается как среднее попарное различие между индивидами двух видов [28,75], что ведёт ко включению, помимо закреплённых различий, множества несинонимичных полиморфизмов и к завышению α [76]. Во-вторых, в качестве нейтрального стандарта мы использовали нуклеотиды с 8 по 30 в коротких интронах, а не синонимичные сайты, как, например, в [29]. Нуклеотиды с 8 по 30 в коротких

интронах наименее подвержены действию отбора, тогда как мутации в синонимичных сайтах, особенно мутации S→W, находятся под действием отрицательного отбора. Хотя этот отбор является слабым ($N_e s$ > -0.5), он уменьшает дивергенцию в синонимичных сайтах в ~1,3 раза (Рисунок 11). В нашем анализе использование нуклеотидов с 8 по 30 в коротких интронах как нейтрального стандарта вместо синонимичных сайтов приводит к уменьшению α с 34% до 14%.

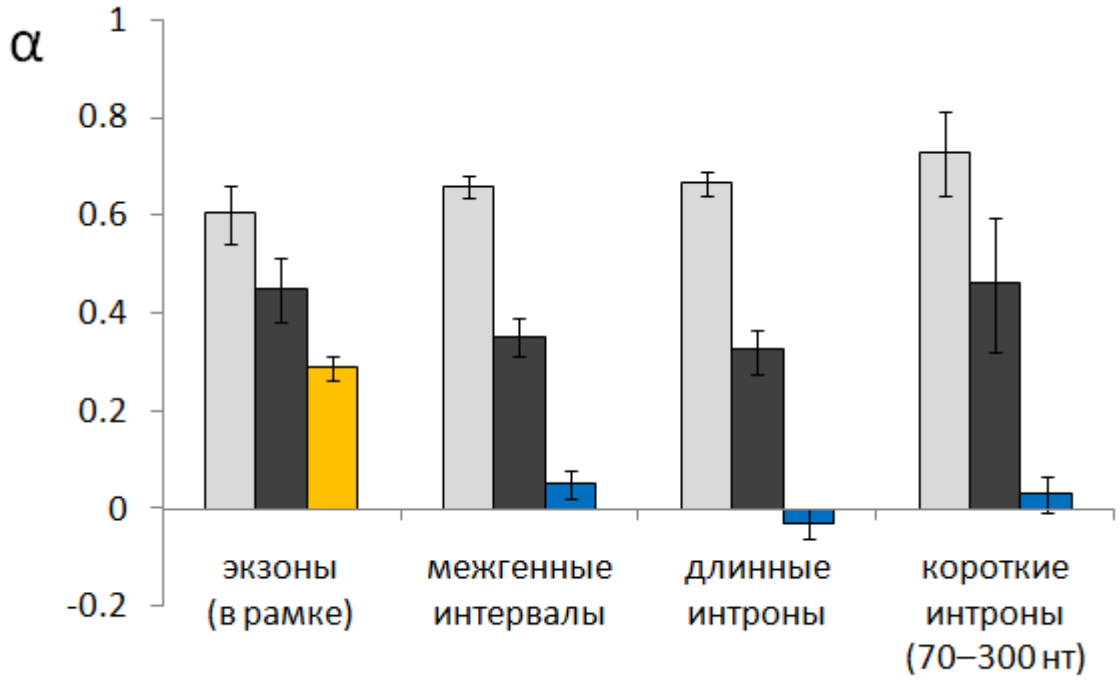

Рис. 10. α для инделов и однонуклеотидных замен в различных участках генома: экзонах (кратные трём инделы и миссенс-замены), межгенные интервалы, длинные (>300 нт) и короткие (70–300 нт) интроны. (B), ξ для инделов в коротких и очень коротких интронах. Светло-серым показаны инсерции; темно-серым – делеции; жёлтым – миссенс-замены; синим – однонуклеотидные замены в

некодирующих участках генома. Доверительные интервалы построены по 1000 бутстреп-репликам.

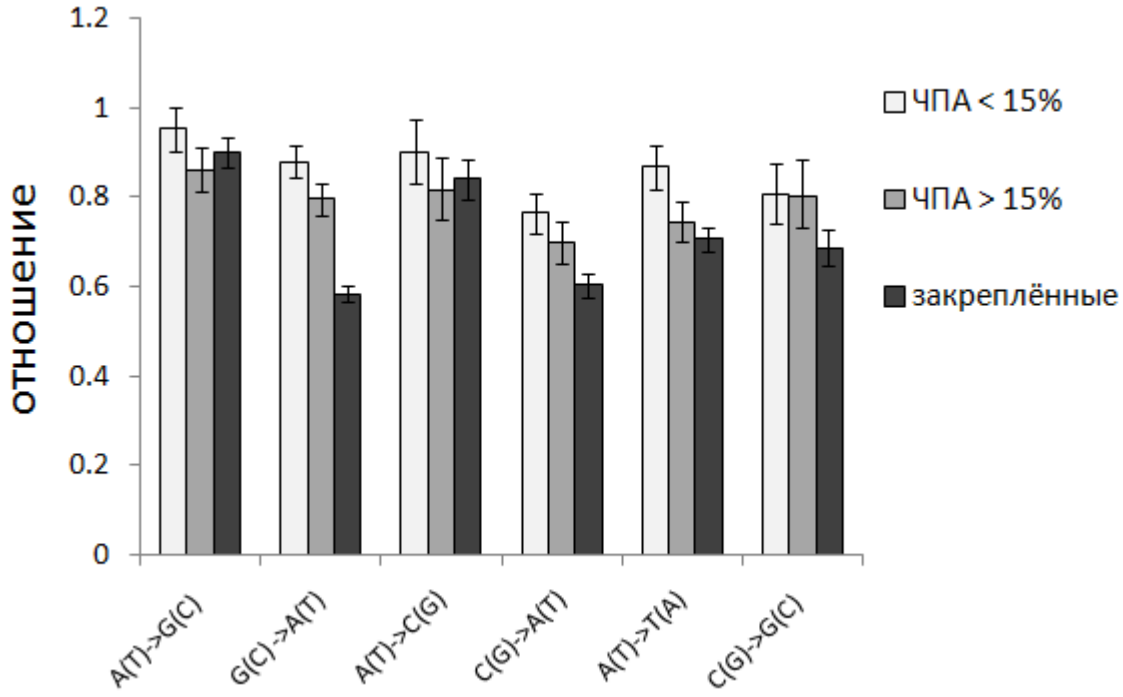

Рис. 11. Отрицательный отбор в синонимичных сайтах. Каждый столбик показывает отношение встречаемости мутаций в четырёхкратно-вырожденных синонимичных сайтах к мутациям в позициях с 8 по 30 в коротких интронах. Доверительные интервалы построены по 1000 бутстреп-репликам.

Глава 2. Влияние генной конверсии на закрепление мутаций

Чтобы изучить эффект рекомбинации на мутации различных типов, мы проанализировали эволюцию в линии *D. melanogaster*, сравнивая участки с разными скоростями рекомбинации. Сперва мы рассмотрели однонуклеотидные

замены. Скорость их фиксации самая низкая в участках с самой низкой рекомбинацией, и в ~2 раза выше в участках с высокой рекомбинацией (Рисунок 12A, левая часть). Эта зависимость была описана ранее, и, вероятно, возникает из-за фиксации слабовредных мутаций в участках с низкой скоростью рекомбинации [42,77]. Для позиций 8–30 в интронах длин <65нт, которые считаются участками с самым слабым отбором в геноме, данного эффекта не наблюдается ([42]; Рисунок 13) .

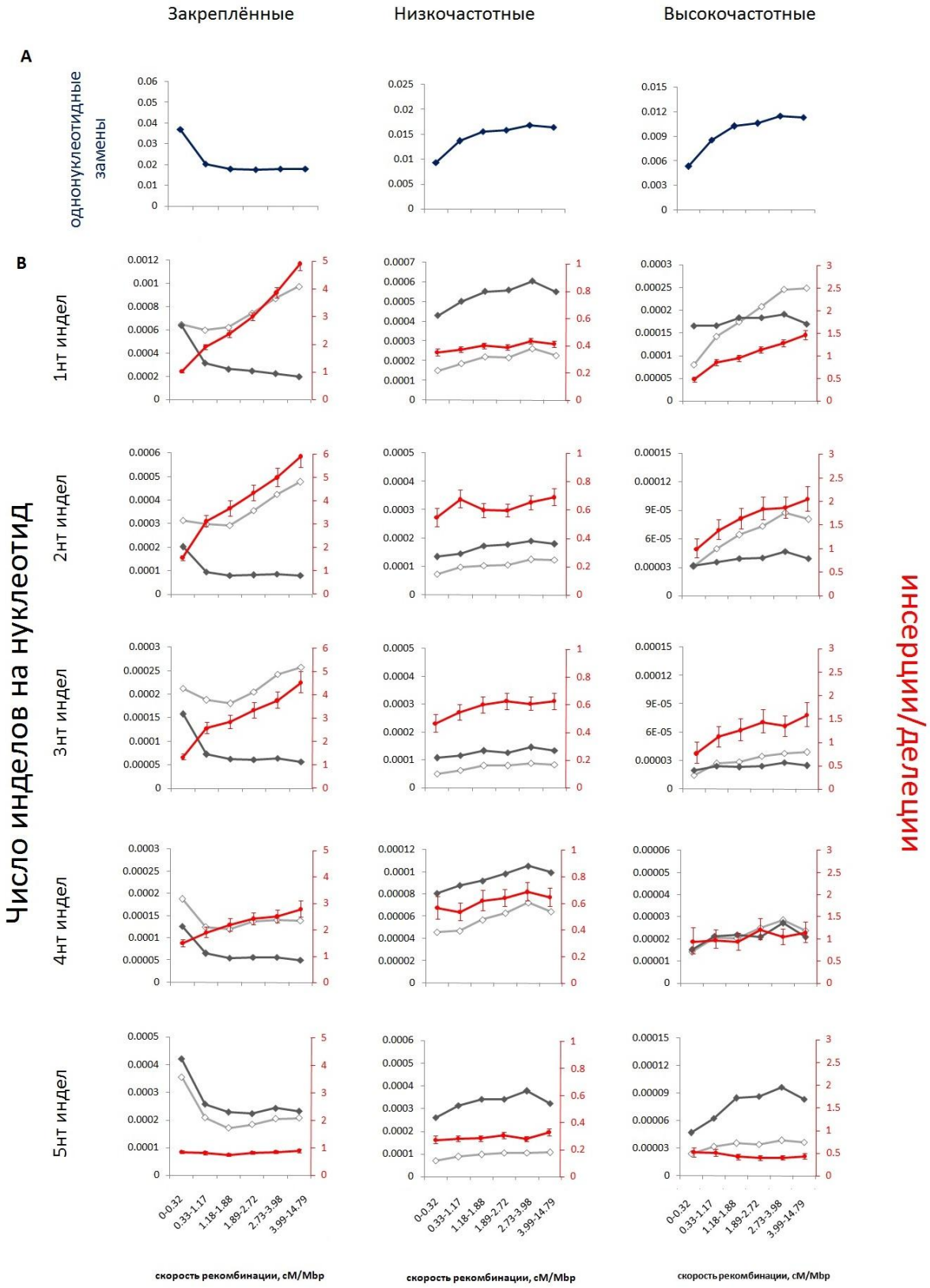
Закреплённые
Низкочастотные
Высокочастотные
A
однонуклеотидные замены
B
1нт индел
2нт индел
3нт индел
4нт индел
5нт индел
Число инделов на нуклеотид
инсерции/делеции
0-0.32
0.33-1.17
1.18-1.88
1.89-2.72
2.73-3.98
3.99-14.79
скорость рекомбинации, cM/Mbp
скорость рекомбинации, cM/Mbp
скорость рекомбинации, cM/Mbp

Рис. 12. Закрепившиеся и полиморфные мутации в некодирующих участках *D. melanogaster* в различными скоростями рекомбинации. (A) Однонуклеотидные мутации. (B) Инсерции и делеции. Серым показаны инсерции, чёрным – делеции, красным – отношение инсерций к делециям, синим – однонуклеотидные замены. Левая часть – закрепившиеся мутации, средняя часть – низкочастотные мутации (<3%), правая часть – мутации с частотой выше 15%.

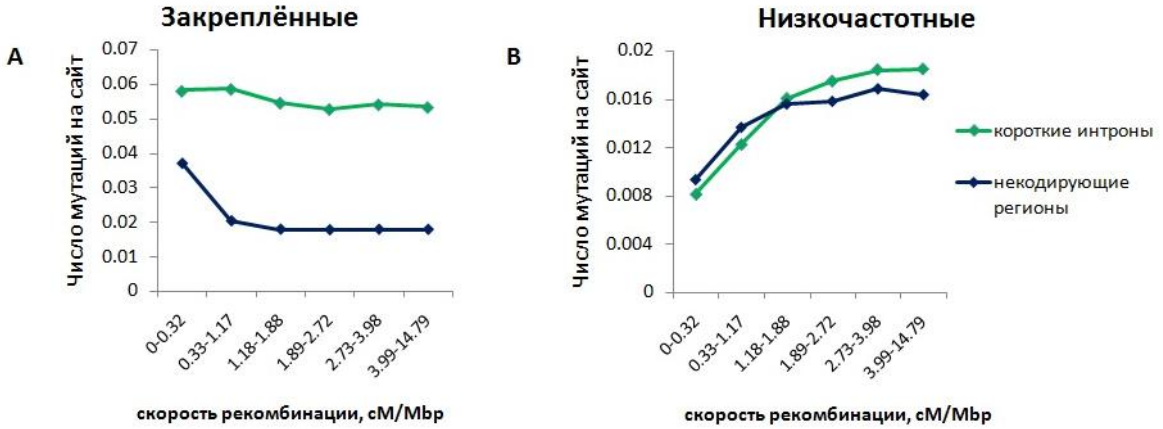

Рис. 13. Нейтральные и отбираемые однонуклеотидные замены в некодирующих участках *D. melanogaster* с различными скоростями рекомбинации. (A) закреплённые мутации, (B) низкочастотные мутации (ЧПА < 3%). На каждом графике зелёная линия соответствует мутациям в позициях 8–30 в коротких (<65 нт) интронах (нейтральный стандарт), а синяя линия соответствует мутациям во всех некодирующих участках.

Сходная картина наблюдается для инделов длин 5-10нт, закрепившихся в популяции: как инсерции, таки и делеции легче закрепляются в участках с более низкой рекомбинацией. Зависимость от рекомбинации одинакова для инсерций и делеций, в результате чего отношение инсерций к делециями от скорости рекомбинации не зависит. Совершенно по-другому зависит от рекомбинации отношение инсерций к делециям для коротких (1-4 нт) инделов. Вероятность закрепления в популяции делеций длин 1-4 нт, как и для более длинных делеций, отрицательно коррелирована со скоростью рекомбинации, причём эта корреляция сильнее, чем для более длинных делеций. Для инсерций напротив, отрицательная корреляция со скоростью рекомбинации сильнее для длинных инсерций, чем для коротких, а для инсерций с длинами 1–3 нт и вовсе наблюдается положительная корреляция со скоростью рекомбинации. В результате мы наблюдаем, что отношение инсерций к делециям значительно возрастает с ростом скорости рекомбинации (Рисунок 12B, левая часть). Самый большой контраст (в 5 раз) между участками высокой и низкой рекомбинации наблюдается для однонуклеотидных инделов; он снижается с увеличением длины и практически отсутствует для инделов длиннее 5 нт (Рисунок 12B, левая часть).

Корреляция между отношением инсерций к делециям и рекомбинацией могла возникнуть из-за неоднородности возникновения мутаций вдоль генома [78,79], если бы эта неоднородность по-разному влияла на инсерции и делеции. В таком случае мы должны были бы увидеть те же самые эффекты на уровне внутривидового полиморфизма. Чтобы это проверить, мы проанализировали

частоты полиморфных однонуклеотидных замен, а также инсерций и делеций, которые расщепляются в популяции на низких частотах (<3%). Встречаемость полиморфных однонуклеотидных замен, как известно [29,43,80], положительно скоррелированна с рекомбинацией, вероятно из-за пониженного влияния отбора в сцепленных сайтах на участки с высокой скоростью рекомбинации [38,39]. В наших данных эта корреляция наблюдается как для однонуклеотидных полиморфизмов (Рисунок 12A, средняя часть), так и для инделов всех длин (Рисунок 12B, средняя часть). Однако, в отличие от закрепившихся инделов, встречаемость полиморфных инделов зависит от скорости рекомбинации одинаково, и отношение инсерций к делециям практически не зависит от скорости рекомбинации (Рис 12B, средняя часть).

Таким образом, различие в скоростях мутагенеза не может объяснить зависимость отношения инсерций к делециям от рекомбинации, наблюдаемую на данных по дивергенции. Эта зависимость возникает в процессе закрепления мутаций. Для мутаций, расщепляющихся на более высоких частотах (>15%), мы видим паттерн, промежуточный между тем, что наблюдали для *de novo* мутаций и для закрепившихся мутаций (Рисунок 12B, правая часть).

В процессе закрепления две направленные силы могут влиять на частоты аллеля: естественный отбор и смещённая генная конверсия. Отбор действует на аллели, различающиеся по приспособленности, так, что положительный отбор увеличивает, а отрицательный отбор уменьшает вероятность закрепления нового аллеля. Для всех типов мутаций отрицательный отбор гораздо распространённее

положительного. В отличие от отбора, смещённая генная конверсия действует невзирая на приспособленности аллелей. Действуя на пару антагонистических мутаций, таких как A→G против G→A, или в нашем случае инсерции и делеции, она пропорционально увеличивает частоту одних и уменьшает частоту других [47,81].

Чтобы отличить отбор от смещённой генной конверсии, мы проанализировали процесс фиксации более детально, используя отношение числа закрепившихся мутаций к числу низкочастотных (ОЗНЧ) как оценку вероятности фиксации мутации. Как нейтральный стандарт мы использовали полиморфизмы в позициях 8–30 в интронах <65нт; этот класс считается наиболее нейтральным [28]. ОЗНЧ для нейтральных сайтов было выше в участках низкой рекомбинации (Рисунок 14) в соответствии с действием сцепленного отбора, уменьшающего полиморфизм в этих участках [38,39].

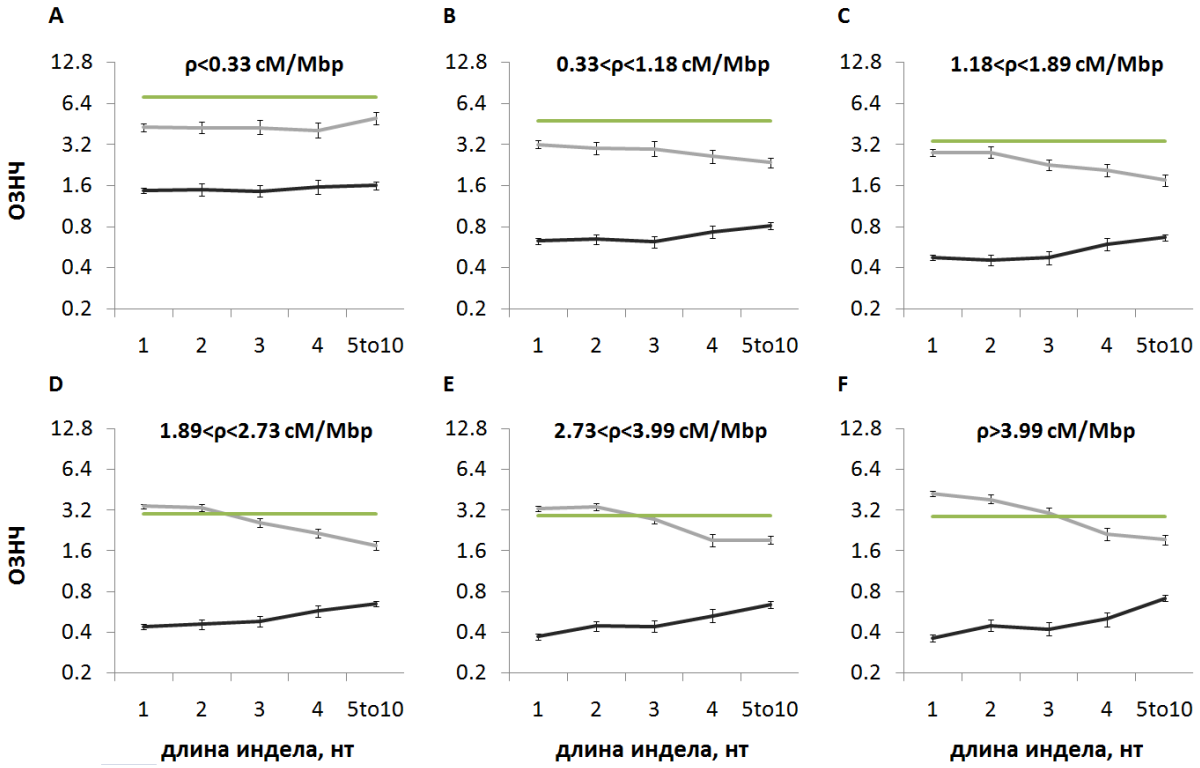

Рис. 14. Отношение числа закреплённых мутаций к числу низкочастотных (ОЗНЧ) для участков с разными скоростями рекомбинации (ρ). Зелёным показан нейтральный стандарт, серым инсерции, чёрным делеции.

В целом как для инсерций, так и для делеций ОЗНЧ ниже, чем для нейтрального стандарта, что соответствует действию отрицательного отбора против закрепления обоих типов мутаций. Однако естественный отбор не может объяснить наблюдаемые зависимости отношения инсерций к делециям от рекомбинации. Во-первых, отношение инсерций к делециям не зависит от длины индела в участках с самой низкой рекомбинацией (ρ<0,32 сМ/Мб), где действие конверсии предполагается минимальным, в то время как вероятности закрепления как инсерций, так и делеций значительно ниже, чем для нейтральных мутаций,

что говорит о том, что отбор здесь сохраняется (Рисунок 14A). Другими словами, в отсутствие рекомбинации отбор не меняет распределение длин инделов. В участках же высокой рекомбинации (Рисунок 14B-F) наблюдается корреляция вероятности закрепления с длиной индела, отрицательная для инсерций и положительная для делеций. В этих участках длинные делеции имеют более высокую вероятность закрепления в сравнении с короткими делециями, что несовместимо с гипотезой об отборе.

Также несовместима с отбором зависимость отношения инсерций к делециям для закрепившихся мутаций от скорости мутагенеза. Это отношение равномерно растёт со скоростью рекомбинации, согласуясь с линейно растущим числом событий конверсии (Рисунок 12), тогда как отбор предсказывает сильно-нелинейную зависимость: даже небольшого уровня рекомбинации достаточно, чтобы устранить эффект сцепленного отбора [40–42].

Есть также два дополнительных свидетельства того, что, наблюдаемое смещение не было вызвано отбором. Во-первых, отбор очень слаб в коротких интронах [28], тем не менее, если рассматривать отдельно интроны с длиной <70 нт, наблюдаются те же самые значения положительной корреляции, что и для межгенных интервалов (Рисунок 15). Во-вторых, отбор должен быть сильнее в более консервативных участках. Однако отношение числа инсерций к числу делеций в участках с различной консервативностью (по значению phastCons [82]) различалось лишь незначительно (Рисунок 16). В то же время в каждом интервале консервативности наблюдалось сильное различие между участками с низкой и

высокой скоростью рекомбинации: в 5 раз для инделов длины 1 нт, в 2–4 раза для инделов длины 2–4 нт (Рисунок 16). Таким образом, рекомбинация, но не степень консервативности, является главным фактором, влияющим на отношение числа инсерций к числу делеций.

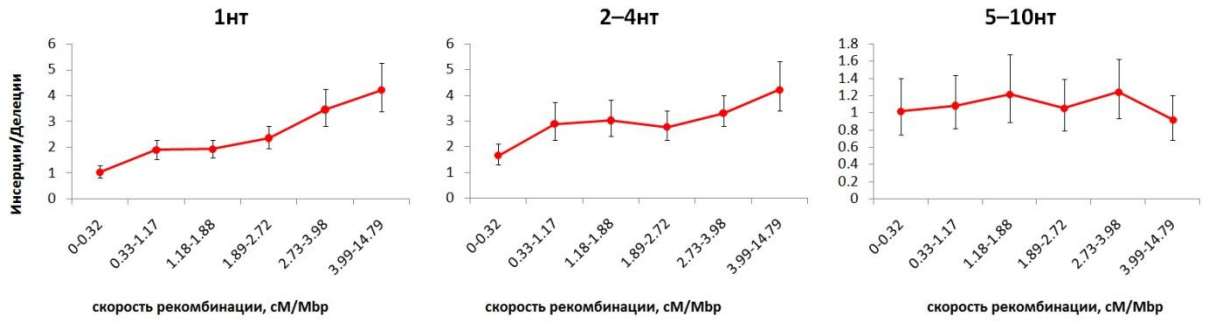

Рис. 15. Отношение инсерции/делеции для закреплённых инделов в коротких (<75 нуклеотидов) интронах *D. melanogaster*. 95% доверительные интервалы построены по 1000 бутстреп-репликам.

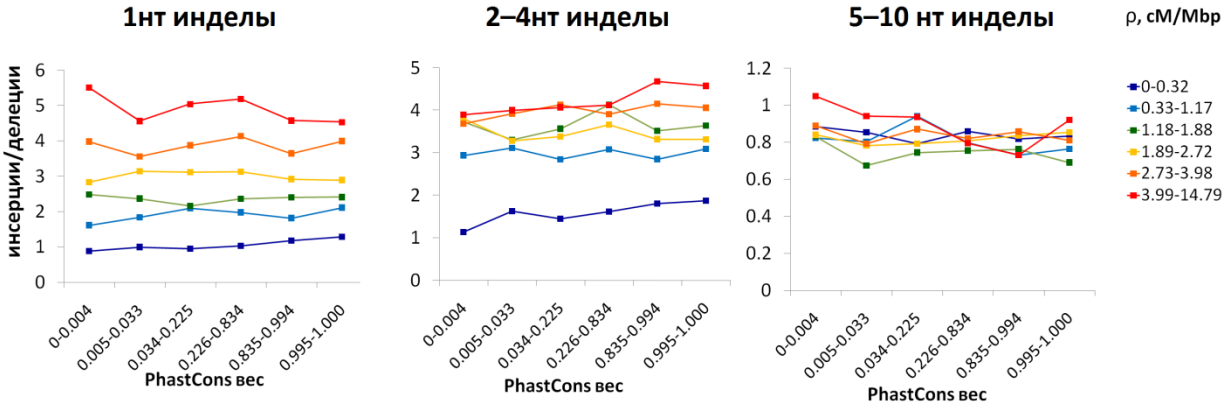

Рис. 16. Отношение числа закреплённых инсерций к числу закреплённых делеций для некодирующих последовательностей *D. melanogaster* с различной степенью консервативности (phastCons вес) и разными скоростями рекомбинации (ρ).

Таким образом, единственным правдоподобным объяснением наблюдаемого паттерна является генная конверсия, смещённая в сторону инсерций. Как увеличение вероятности закрепления инсерций длины 1–4 нт, так и снижение вероятности закрепления делеций длины 1–4 нт объясняются конверсией, благоприятствующей закреплению инсерций и препятствующей закреплению делеций.

Чтобы оценить силу смещённой в сторону инсерций генной конверсии, мы сравнили вероятности закрепления коротких инделов (1–4нт) и инделов с длинами >4 нт, подразумевая, что конверсионное смещение незначимо для последних (см. раздел 8 в Материалах и методах). В среднем по геному для инсерций за счёт генной конверсии наблюдается увеличение вероятности закрепления в 1,60, 1,55, 1,31 и 1,05 раз, а для делеций уменьшение вероятности закрепления в 1,43, 1,34, 1,35 и 1,16 раз для инделов с длинами 1, 2, 3 и 4 нт соответственно. В участках с наибольшей скоростью рекомбинации эффект смещённой в сторону инсерций генной конверсии был наибольшим (Рисунок 14F), достигая увеличения в ~2 раза для однонуклеотидных инсерций и уменьшения в ~2 раза для однонуклеотидных делеций. Смещение вероятности

закрепления соответствует усреднённому по геному конверсионному преимуществу для инсерций $N_e\omega$ = ~0,3–0,4. Здесь N_e – эффективный размер популяции, а $\omega=2\kappa(\beta-0,5)$, где $(\beta-0,5)$ – величина смещения конверсии, а κ – вероятность, что нуклеотидный сайт подвергнется рекомбинации [81]. Отсутствие признаков смещённой в сторону инсерций генной конверсии в участках с низкой скоростью рекомбинации (Рисунок 14А) говорит о том, что рекомбинация является основным источником конверсии. Таким образом, κ можно рассчитать как $\kappa = \rho lx$, где ρ – скорость рекомбинации, l – длина конверсионного тракта, x – частота конверсии относительно частоты рекомбинации. В *D. melanogaster* ρ находится в пределах от 10^{-10} (низкая скорость рекомбинации) до 10^{-7} (высокая скорость рекомбинации) на нуклеотид на поколение [68], а $l \approx 350$ нт [83]. Предполагая, что конверсия происходит с той же частотой, что и рекомбинация (x = 1), в участках с высокой скоростью рекомбинации $\kappa \approx 10^{-5}$. Наконец, $N_e = 10^{-5}$ для *D. melanogaster*, что означает, что смещение генной конверсии в сторону инсерций β = 0,7.

С чем связано смещение генной конверсии, наблюдаемое для инделов длин 1–4 нт, в сторону инсерций? Возможно, инделы могут вызывать репарацию шпильки в гетеродуплексе ДНК, такую, что репликация участка петли более вероятна, чем вырезание петли. Ослабление смещения с увеличением длины индела может быть связано с ослаблением интенсивности репарации из-за образования более стабильных структур ДНК для более длинных инделов [84].

Сходная, хотя и более слабая зависимость отношения числа инсерций к делециям для коротких инделов от частоты рекомбинации наблюдается для мутаций, закрепившихся в линии человека после ответвления шимпанзе, а также для мутаций, закрепившихся в линии *S. cerevisiae* после ответвления *S. paradoxus* (Рисунок 17A). Наблюдение такого паттерна в трёх далёких друг от друга филогенетических группах даёт основания полагать, что генная конверсия, смещённая в сторону инсерций – универсальный фактор эволюции животных.

Отношение числа инсерций к числу делеций у *H. sapiens* и *S. cerevisiae* зависит от рекомбинации так же, как и отношение W→S/S→W однонуклеотидных замен (Рисунок 17B), что означает, что в этих видах генная конверсия действует на инделы и на нуклеотидные замены с равной силой. Также у *H. sapiens* и *S. cerevisiae* с увеличением скорости рекомбинации растёт процентное содержание GC. Напротив, у *D. melanogaster* ни отношение W→S/S→W, ни частота GC не зависят от рекомбинации (Рисунок 18); по-видимому, GC-смещённая генная конверсия, даже если и присутствует, не оказывает значимого воздействия на эволюцию генома в этом организме. Возможно, что отсутствие GC-смещённой генной конверсии связано с отсутствием метилтрансферазы в роде *Drosophila*, так как наибольшее смещение конверсии сторону GC наблюдается в CpG динуклеотидах, подверженных метилированию [85].

GC-смещённая генная конверсия повышает GC-состав в участках генома с высокой скоростью рекомбинации (Рисунок 17C), по всей видимости, способствуя формированию и сохранению изохор [47,86]. Сходным образом, генная конверсия, смещённая в сторону инсерций, может влиять на длины сегментов генома. Этот эффект должен быть наибольшим для небольших нефункциональных фрагментов генома, разделяющих консервативные блоки, так как влияние на них естественного отбора и длинных инделов мало. В самом деле, у *D. melanogaster* очень короткие интроны (короче 60 нуклеотидов) чаще встречаются в участках низкой рекомбинации, чем более длинные интроны (60–75), подтверждая влияние генной конверсии (тест Манна-Уитни для участков с $\rho<0.01$ и $\rho>3.66$, $P = 10^{-15}$; Рисунок 19). Отношение инсерций к делециям в интронах с длинами до 65 нт растёт с рекомбинацией только для инделов короче 4 нуклеотидов; таким образом, наблюдаемое различие в распределении длин интронов объясняется смещённой генной конверсией, а не отбором. Ситуация, однако, меняется на противоположную для длинных интронов (>70 нт) (тест Манна-Уитни для участков с $\rho<0,01$ и $\rho>3,66$, $P = 10^{-7}$; Рисунок 20), для которых, как предполагается, характерно действие отбора, способствующего сокращению длины [87–89].

Смещённую генную конверсию нужно иметь в виду при изучении адаптивной эволюции, так как она увеличивает вероятность закрепления инсерций длин 1–4 нт по сравнению с более длинными инсерциями. Для инсерций с длинами 1–2 нуклеотида в участках с высокой скоростью

рекомбинации вероятность закрепления даже превышает вероятность закрепления нейтральных замен (Рисунок 14F), что обычно принимается как свидетельство положительного отбора.

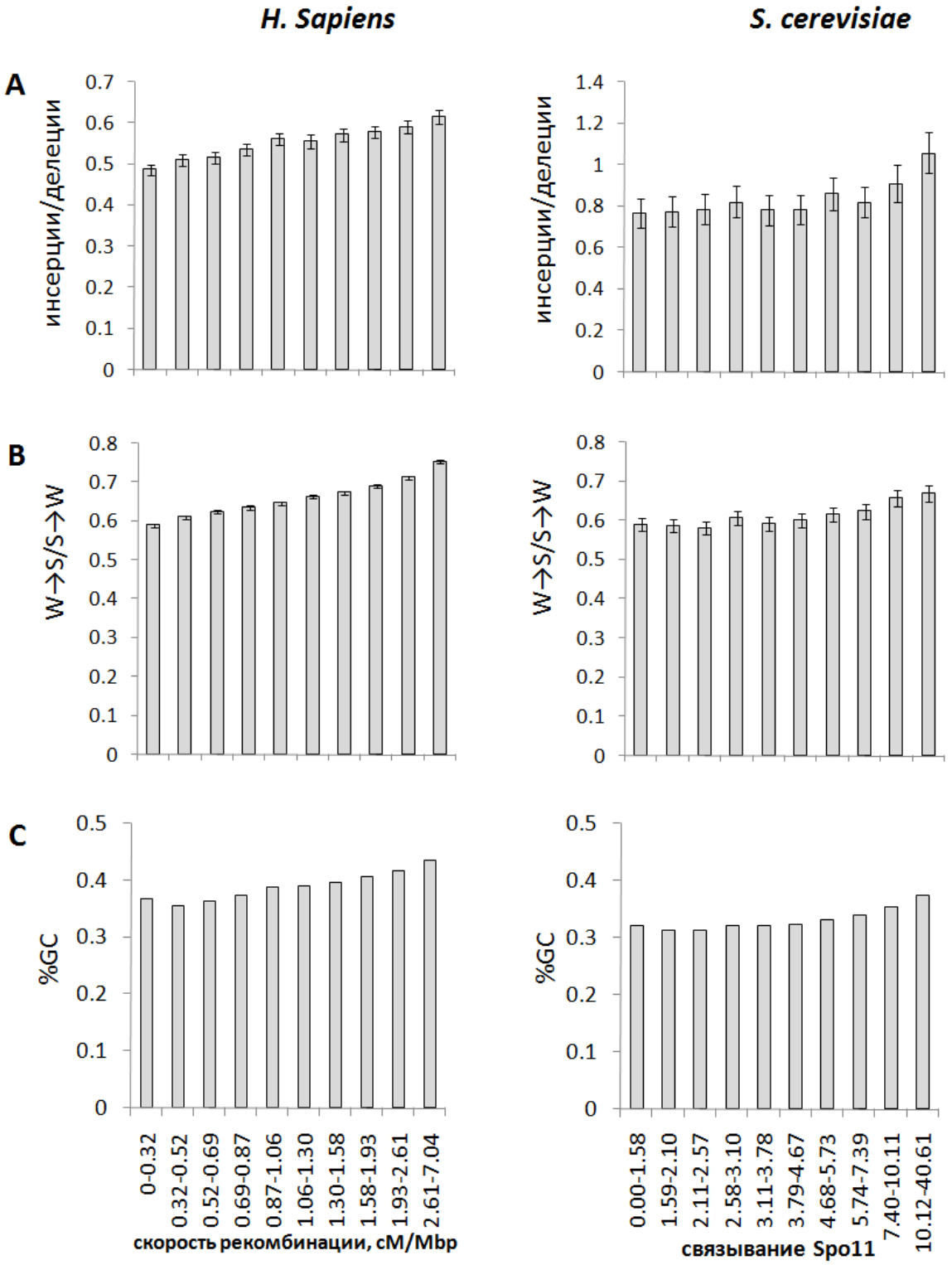

Рис. 17 Генная конверсия в некодирующих участках *H. sapiens* и *S. cerevisiae*. (A) Отношение числа инсерций к числу делеций для инделов с длинами 1–5 нуклеотидов; (B) отношение W→S/S→W для однонуклеотидных замен,

произошедших в линии *H. sapiens (S. cerevisiae)* после расхождения с *P. troglodytes (S. paradoxus);* и (C) GC-состав в зависимости от скорости рекомбинации. Отношение W→S/S→W было посчитано как число замен W→S в A(T)-сайтах, делённое на число замен S→W в G(C)-сайтах. Для *S. cerevisiae* в качестве оценки скорости рекомбинации в определённом участке использовалась степень связывания рекомбинационного белка Spo11 [70]. 95% доверительные интервалы построены по 1000 бутстреп-репликам.

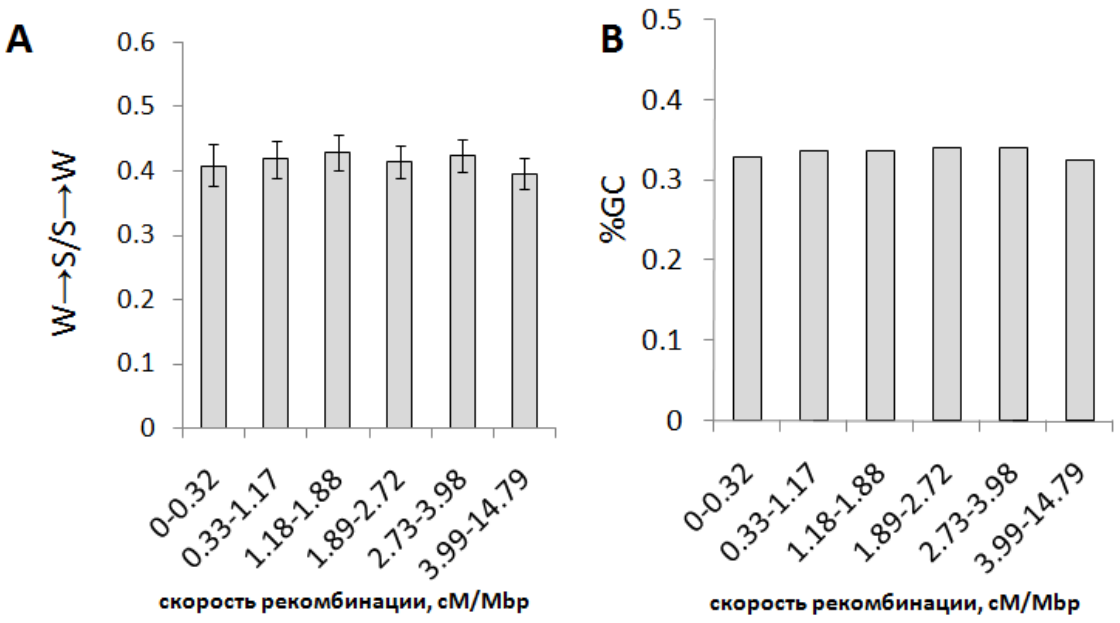

Рис. 18. Паттерн однонуклеотидных мутаций и GC-состав в зависимости от скорости рекомбинации в участках с 8-го по 30-й нуклеотид в коротких интронах (<65 нуклеотидов) *D. melanogaster*. (A) отношение W→S/S→W для однонуклеотидных замен, произошедших в линии *D. melanogaster* после расхождения с *D. sechellia*, и (B) GC-состав в зависимости от скорости

рекомбинации. 95% доверительные интервалы построены по 1000 бутстреп-репликам.

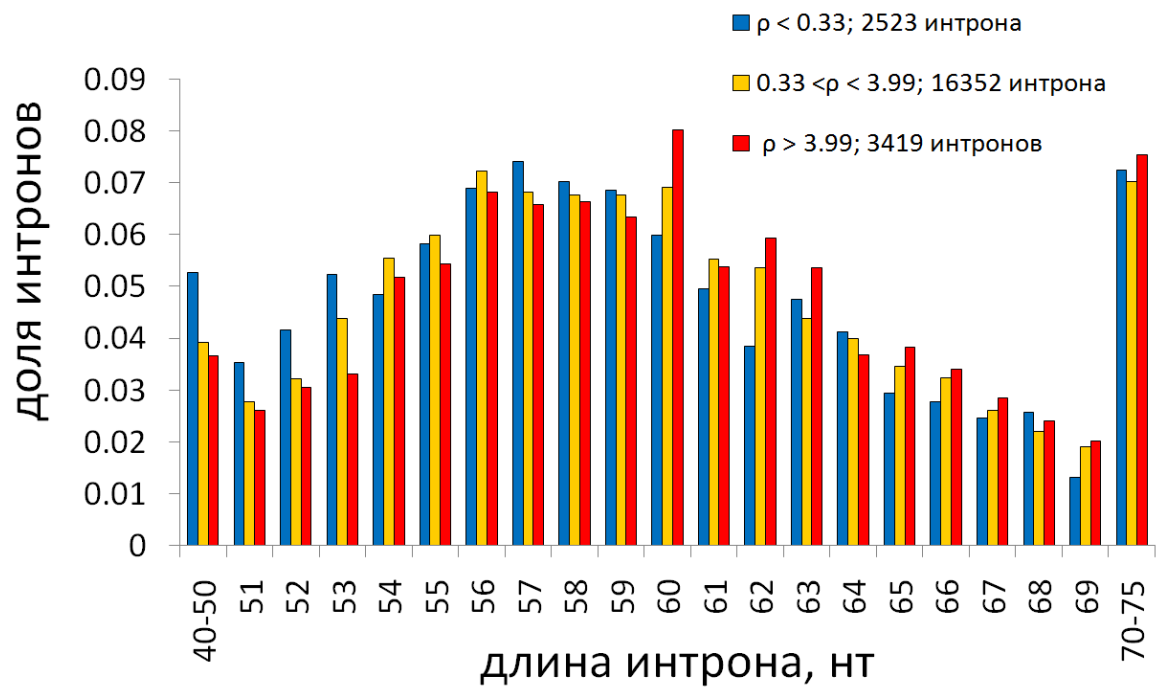

Рис. 19. Распределение длин коротких интронов в участках генома *D. melanogaster* с различными скоростями рекомбинации (ρ). Распределение длин интронов показано для коротких (40–75 нуклеотидов) интронов в участках с низкой (ρ < 0,01), промежуточной (0,01 < ρ < 3,66) и высокой (ρ > 3,66) скоростью рекомбинации.

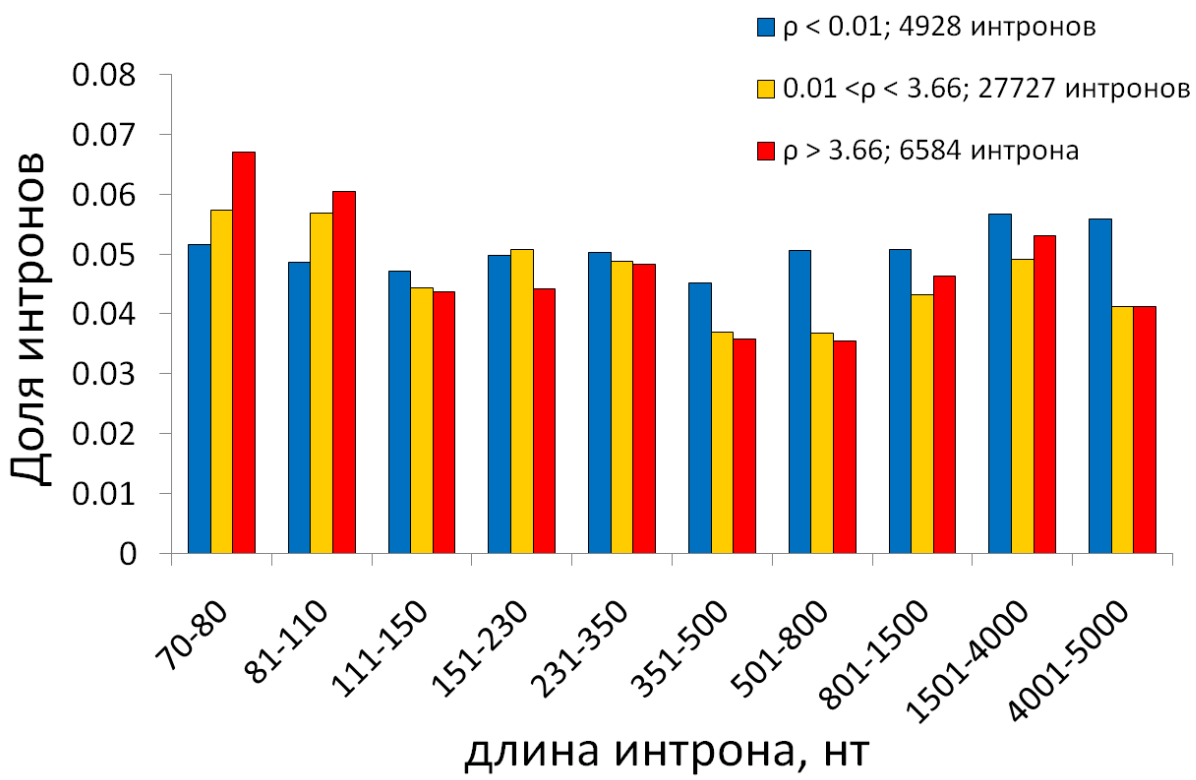

Рис. 20. Распределние длин интронов для интронов длиннее 70 нуклеотидов в участках генома *D. melanogaster* с низкой (ρ < 0,01), промежуточной (0,01 < ρ < 3,66) и высокой (ρ > 3,66) рекомбинацией.

Глава 3. Изменение адаптивного ландшафта при возникновении инсерций и делеций.

Логично ожидать, что такое радикальное событие в последовательности ДНК, как инсерция или делеция, может существенно повлиять на эволюцию окружающей последовательности – например, привести к изменению скорости мутагенеза или/и интенсивности отбора в данном участке. Так, можно ожидать ослабления отрицательного отбора вследствие нарушения структуры и функции участка, содержащего индел, и, как результат, увеличения числа нейтральных замен. Или же, напротив, после индела могут происходить замены, интегрирующие его в структуру белка; в таком случае мы ожидаем увидеть увеличение числа адаптивных замен.

Для определения интенсивности положительного отбора мы применили тест МакДональнда-Крейтмана (МК-тест) [30,31] для аминокислотных замен, которые случилось на расстоянии до 100 аминокислот от сайта индела в линии *D. melanogaster* после ответвления от линии *D. sechellia*, соотнеся данные по дивергенции между *D. melanogaster* и *D. sechellia* с данными по полиморфизму для *D. melanogaster* в синонимичных и несинонимичных сайтах (Рисунок 21-24). Мы сравнили результаты теста для случаев, когда индел произошёл в ветке *D. melanogaster* (опыт), с соответствующими им случаями, когда индел произошёл в

сестринской ветке (контроль; Рисунок 3 a против a', b против b' и т.д.). Очевидно, что только в опыте, а не в контроле, аминокислотные замены могли быть вызваны инделом.

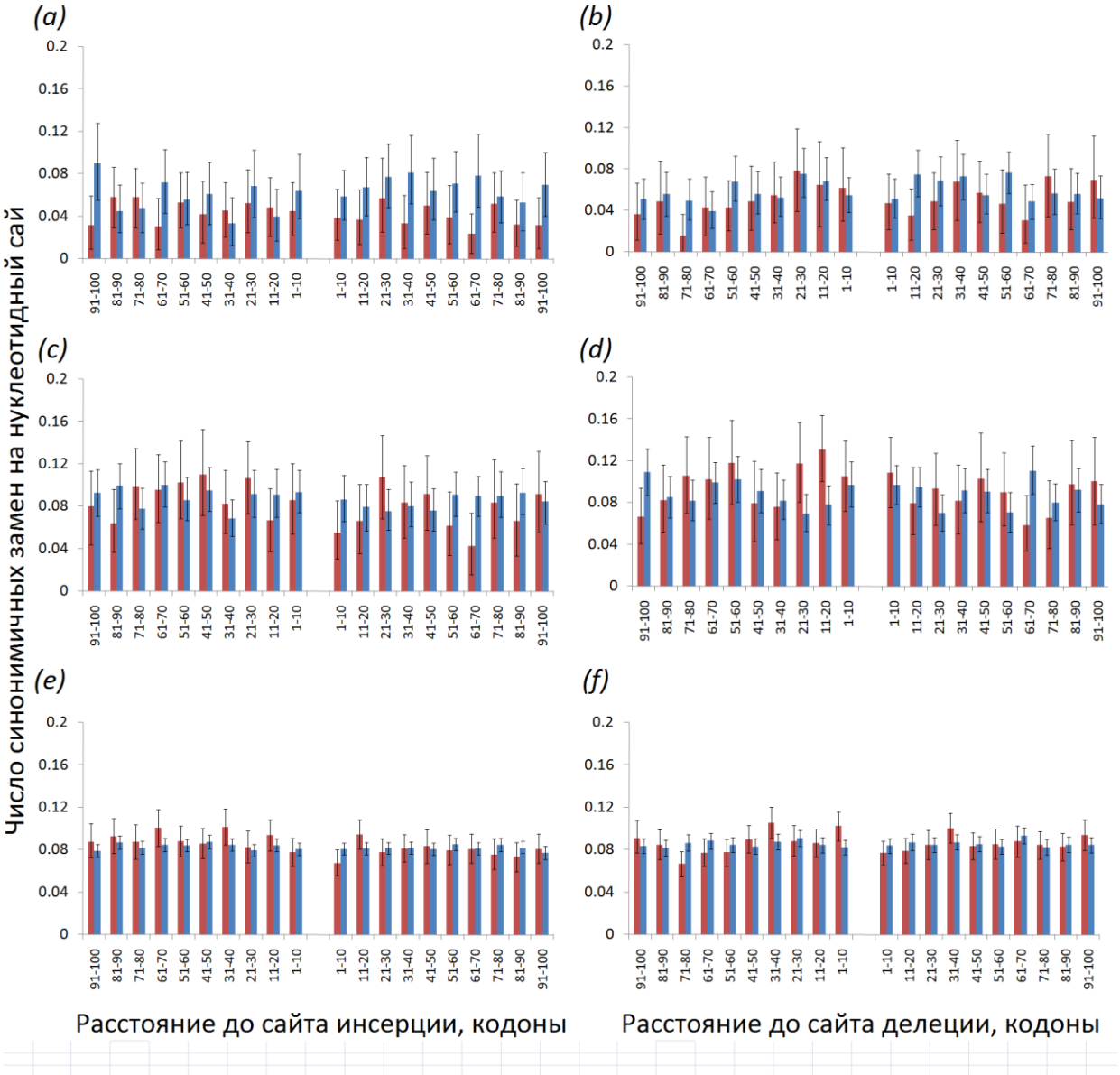

Рис. 21. Скорость синонимичной эволюции D_s в линии *D. melanogaster* (a,b) или в линиях *D. melanogaster* и *D. sechellia* (c–f) в окружении сайтов инсерций и делеций. Графики a, c, e соответствуют инсерциям; графики b, d, f соответствуют

делециям. Графики *a* и *b* (верхний ряд) соответствуют инделам, которые случились после расщепления *D. melanogaster – D. sechellia*; графики *c* и *d* (средний ряд) соответствуют инделам, которые случились между расщеплениями *D. erecta – (D. melanogaster – D. sechellia*) и *D. melanogaster – D. sechellia*; графики e и f (нижний ряд) соответствуют инделам, которые случились между расщеплениями *D. ananassae* – ((*D. melanogaster*, *D. sechellia*), *D. erecta*) and *D. erecta – (D. melanogaster – D. sechellia*). На каждом графике горизонтальная ось – расстояние от сайта индела; красные столбцы соответствуют инделам, которые случились в кладах, включающих *D. melanogaster* (Рисунок 2 a, b, c, d, e, f); синие столбцы соответствуют инделам, которые случились в кладах, не включающих *D. melanogaster* (Рисунок 2 a', b', c', d', e', f'). 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаниями. Согласно точному тесту Фишера, значимых различий между исследуемым случаем (красный) и контролем (синий) не было.

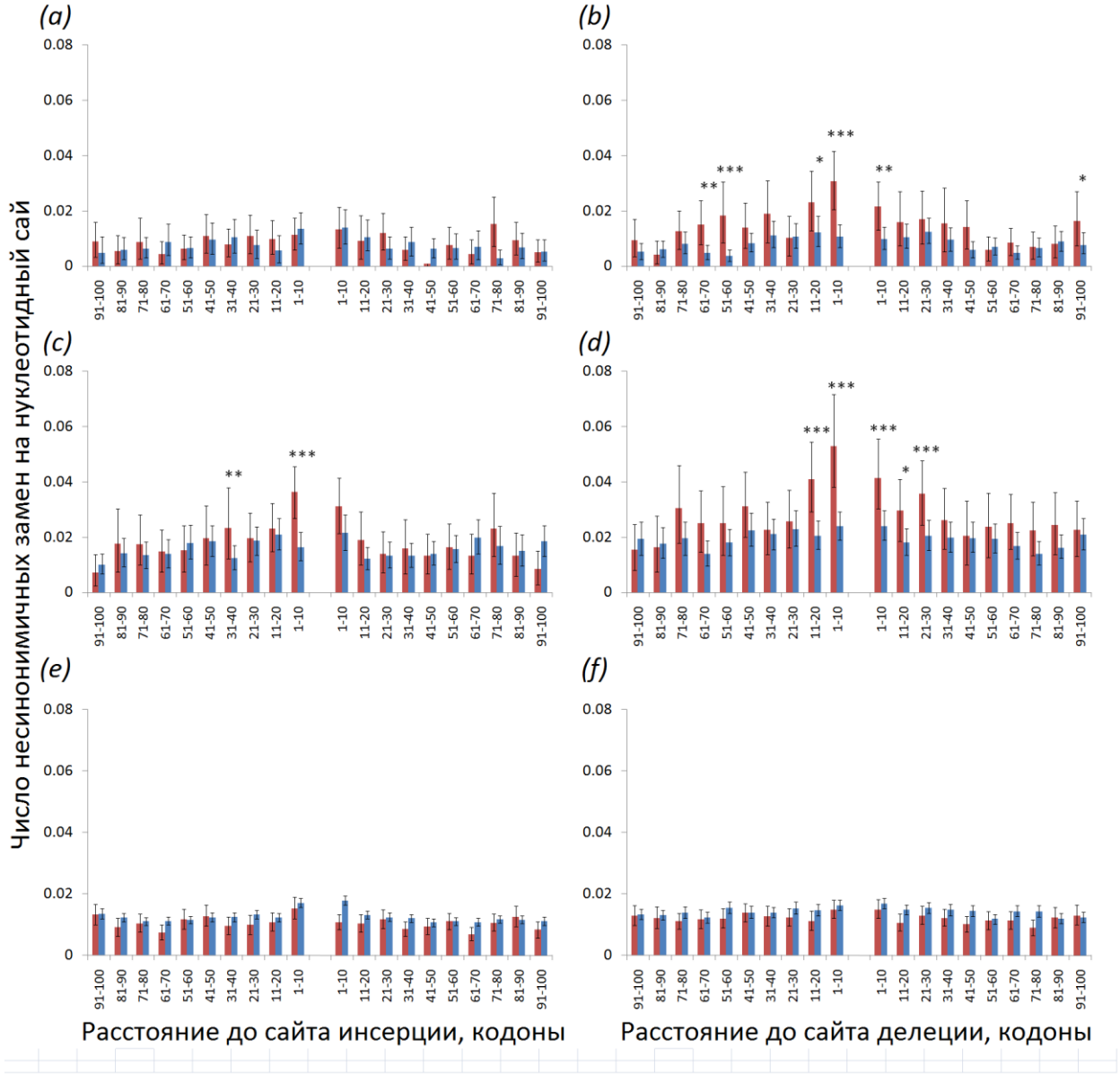

Рис. 22. Скорость несинонимчной эволюции D_n в окружении сайтов инсерций и делеций. Графики и обозначения соответствуют рис. 21. 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаний.

Значимость различий исследуемого случая (красный) и контроля (синий) проверена точным тестом Фишера; *P < 0,05, **P < 0,01,***P < 0,005.

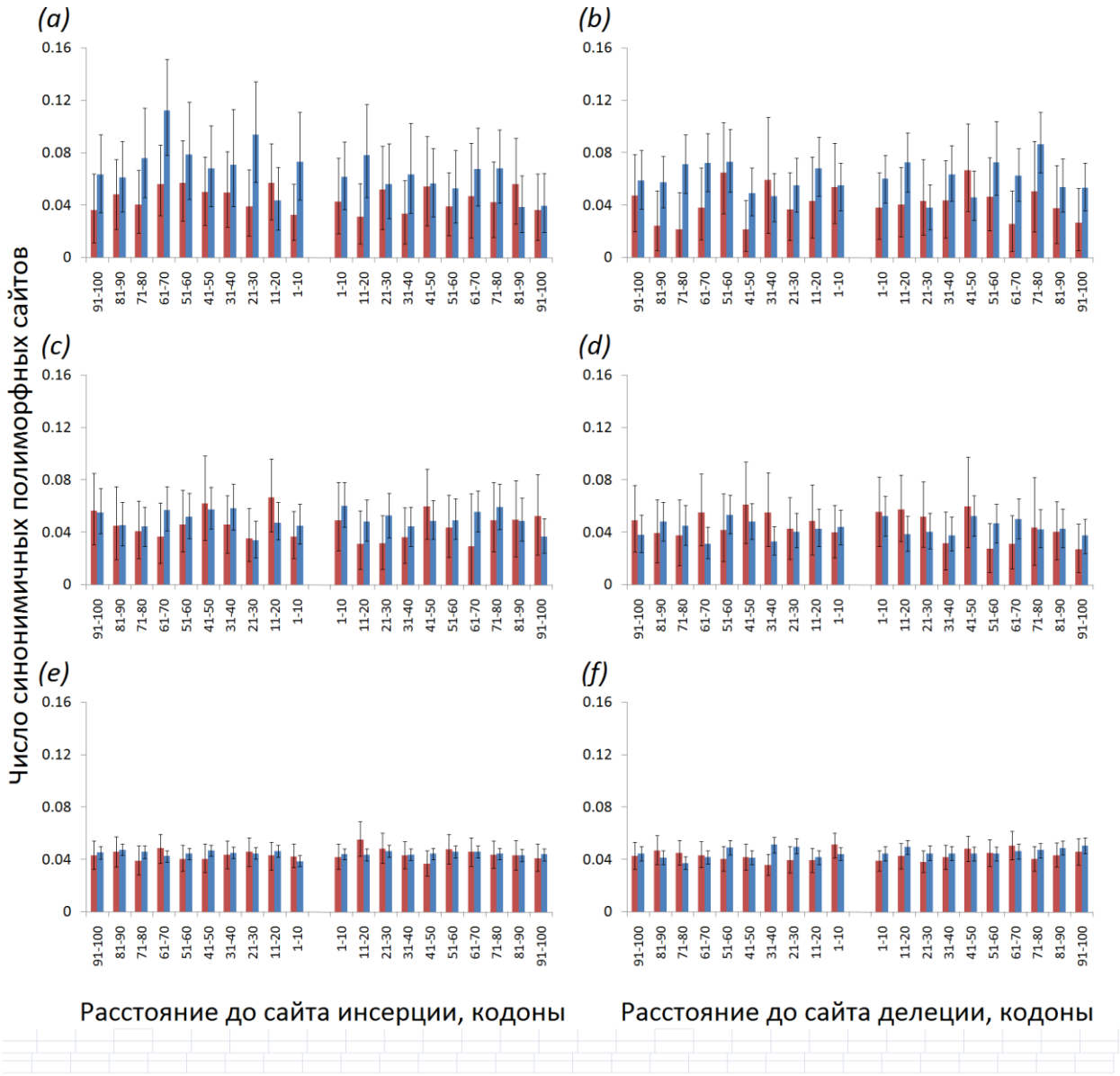

Рис. 23. Частоты синонимичного полиморфизма P_s в окружении сайтов инсерций и делеций. Графики и обозначения соответствуют рис. 21. 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаний.

Согласно точному тесту Фишера, значимых различий между исследуемым случаем (красный) и контролем (синий) не было.

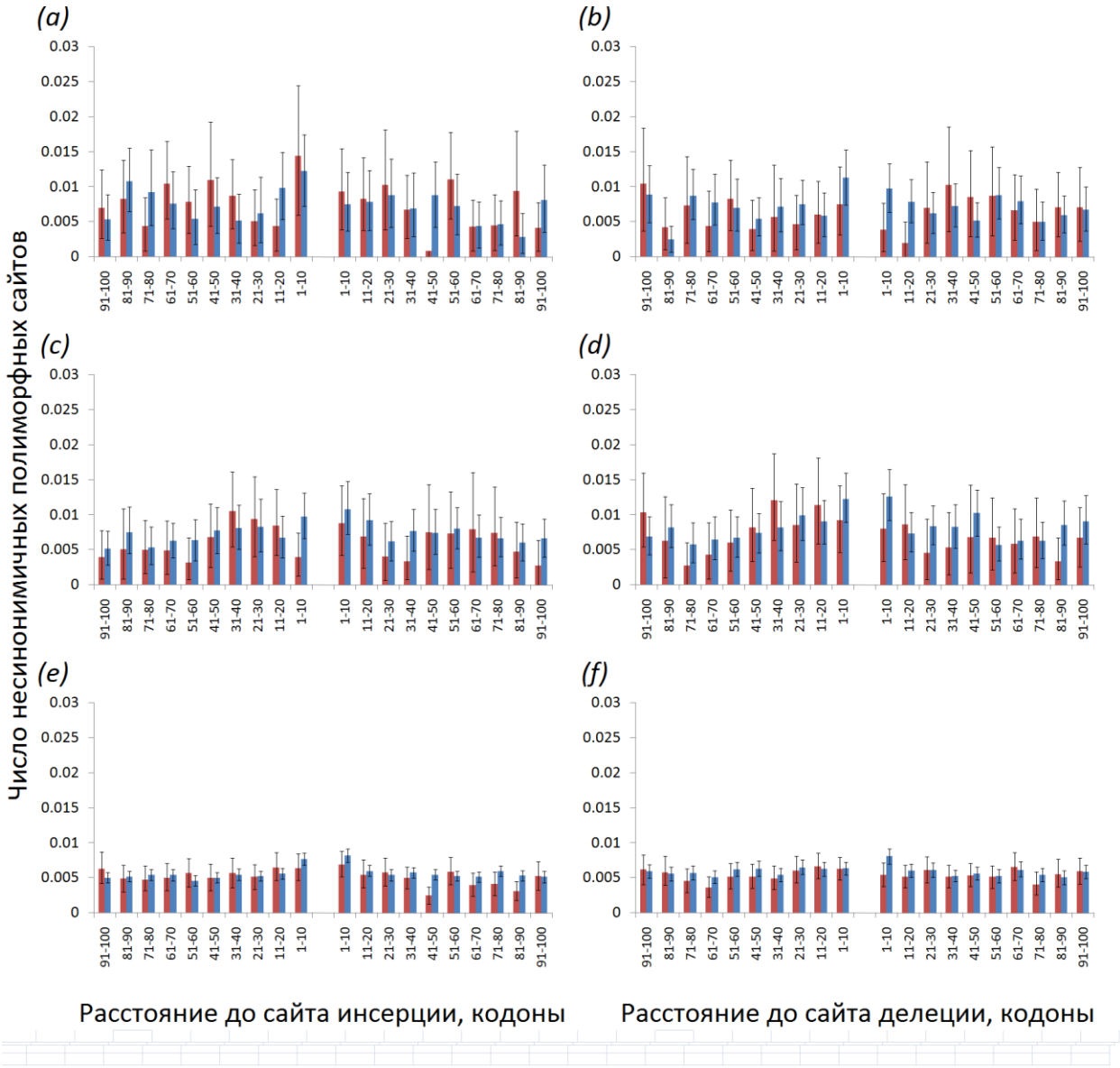

Рис. 24. Частоты несинонимичного полиморфизма P_n в окружении сайтов инсерций и делеций. Графики и обозначения соответствуют рис. 21. 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаний.

Согласно точному тесту Фишера, значимых различий между исследуемым случаем (красный) и контролем (синий) не было.

Рисунки 25 и 26 представляют данные по эволюции и положительному отбору в терминальном (т.е. после отщепления от линии *D. sechellia*) сегменте линии *D. melanogaster* для сайтов, соседствующих с инделом. В соответствии с предыдущими наблюдениями [30,31], мы видим, что инделы чаще происходят в быстроэволюционирующих участках белков, как видно из того, что в аминокислотных сайтах, наиболее близких к сайту индела, число замен выше, по сравнению с более удалёнными сайтами. Это верно как для исследуемого случая, так и для контроля (Рисунок 3, верхний ряд). Однако анализ нуклеотидной последовательности показывает, что такое ускорение не связано с повышенным мутагенезом вблизи сайта индела, как предполагалось в [14]. На это указывает тот факт, что число синонимичных замен остаётся постоянным, при том что число несинонимичных замен увеличивается (Рисунок25, средний и нижний ряды, Рисунок 21, 22).

Ускоренная эволюция вокруг индела также не является следствием повышения мутагенеза, который, как предполагалось в [14], может иметь место в гетеродуплексах по инделу, возникающих при рекомбинации или репарации. На это указывает отсутствие различий в дивергенции и полиморфизме синонимичных сайтов для исследуемого случая и контроля (Рисунок 21, 23). Также ускорение эволюции не есть следствие ослабления отрицательного отбора,

как предполагалось в [90], так как частоты несинонимичного полиморфизма не повышены в исследуемом случае. И только число несинонимичных замен было существенно повышено в соседствующих с инделом сайтах для исследуемых случаев при сравнении с контролем (Рисунок 22). Увеличение скорости эволюции было значимым для филогенетических конфигураций, представленных на Рисунке 2c (инсерции) и Рисунках 2b,d (делеции). В целом, в среднем 1,03±0,75 дополнительных аминокислотных замен происходит после события инсерции и 4,77±1,03дополнительных замен происходит после события делеции в 100 аминокислотных сайтах слева и справа от участка индела.

Тот факт, что различие в скоростях аминокислотной эволюции обусловлено только несинонимичными заменами, означает, что все дополнительные аминокислотные замены (Рисунок 25, Рисунок 27) произошли благодаря положительному отбору. Формально доля замен, случившихся под действием положительного отбора, может быть рассчитана с помощью МК-теста [30,31]. Как следует из Рисунка 26, подавляющее большинство замен, вызванных событием индела, закрепляется под действием положительного отбора. Некоторое увеличение числа аминокислотных замен рядом с сайтом индела в контрольных случаях (Рисунок 27) было связано также и с увеличением числа несинонимических полиморфизмов (Рисунок 24), вследствие чего избытка адаптивных замен вблизи сайта индела не было (Рисунки 26 a', b' и т.д.). В то же время дополнительные замены, которые случились в ветке *D. melanogaster*, были адаптивными, так как они могут быть объяснены положительным отбором

(Рисунок 4 b, c, d). Стоит отметить, что большинство замен, вызванных событием индела, произошли к 5'-концу от сайта индела (Рисунок26 и Рисунок 27).

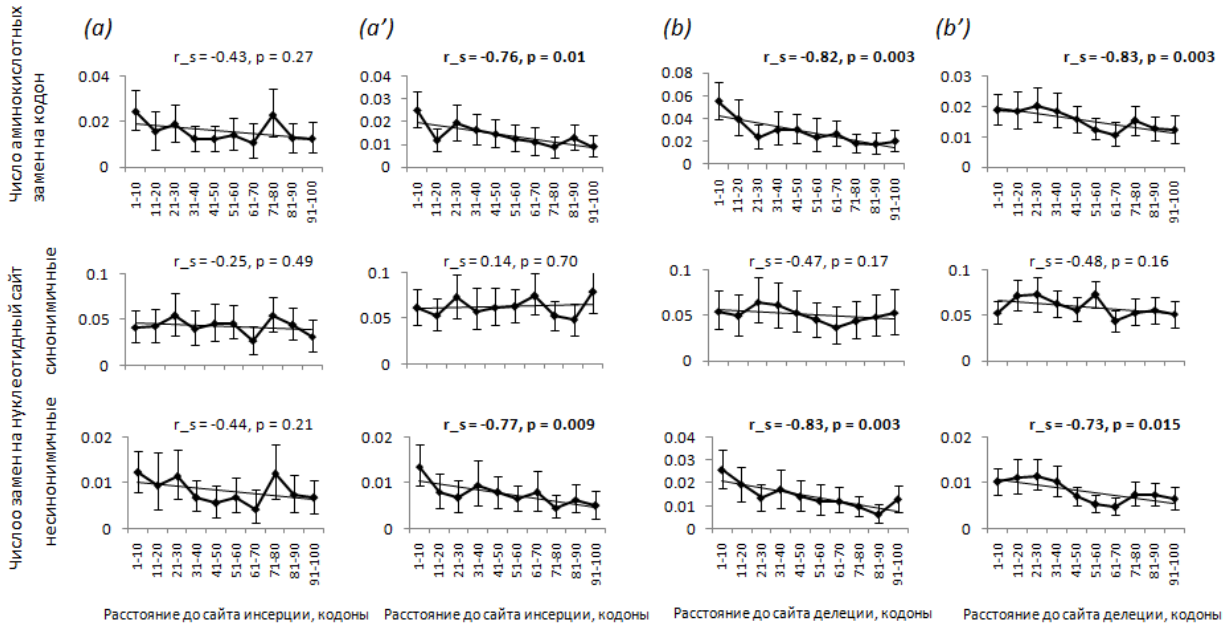

Рис. 25. Ускоренная эволюция в несинонимичных, но не в синонимичных сайтах вблизи индела. Графики a-b' на данном рисунке соответствуют графикам a-b' на рис. 3. Верхний ряд показывает число аминокислотных замещений на аминокислотный сайт, средний ряд показывает число синонимичных замен на синонимичный сайт в ДНК, нижний ряд показывает число несинонимичных замен на несинонимичный сайт в ДНК. 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаний. Корреляция числа замен с расстоянием от индела проанализирована тестом Спирмана; жирным выделены случаи, когда корреляция была значимой (P<0,05).

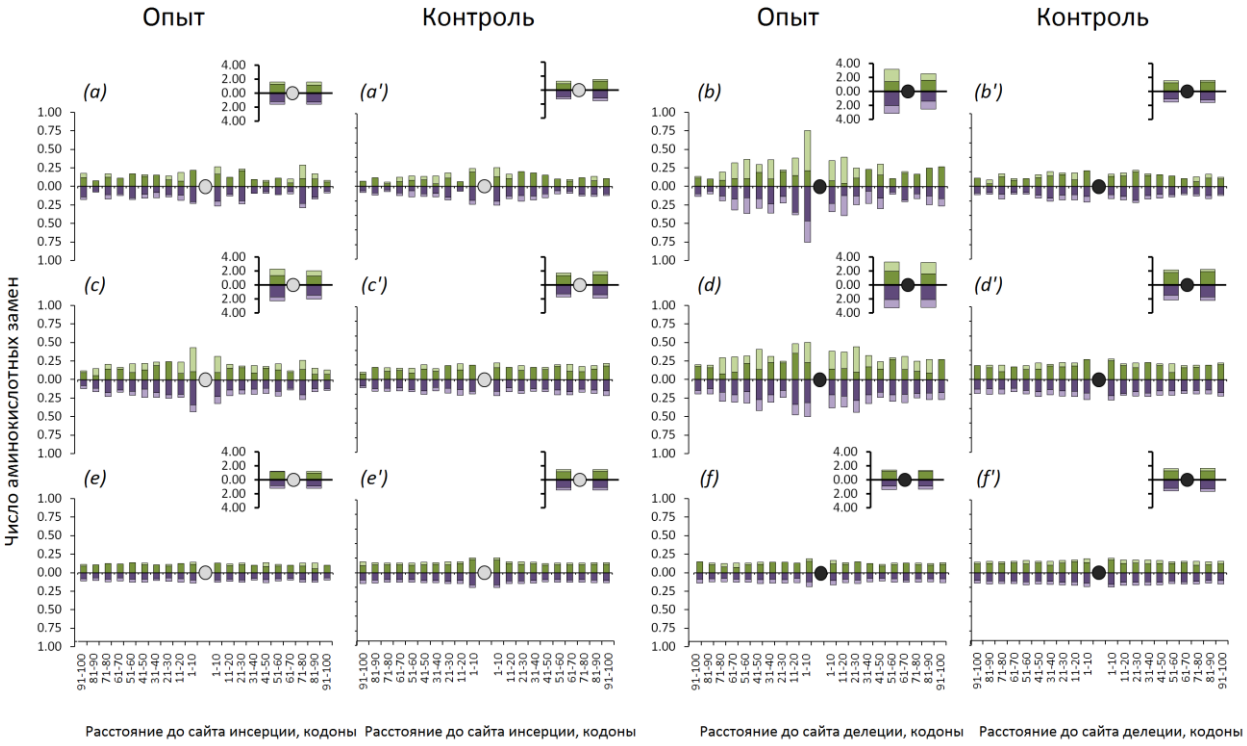

Рис. 26 Ускорение адаптивной эволюции в аминокислотных сайтах под действием делеций и инсерций. Графики расположены также, как филогенетические схемы на рисунке 3. Серым кружком обозначен сайт инсерции, чёрным кружком – сайт делеции. Слева от сайта индела – N-конец белка, справа – C-конец. Высота столбца показывает общее количество аминокислотных замен на данном участке, произошедших на терминальном сегменте линии *D. melanogaster*. Над горизонтальной осью светло-зелёным обозначена доля замен, закреплённых под действием положительного отбора, тёмно-зелёным – доля остальных замен. Под осью светло-лиловым обозначена доля замен, которые произошли в абсолютно консервативных сайтах. На каждой рисунке в правом верхнем углу показана суммарная информация по 100 кодонам слева и справа от сайта индела.

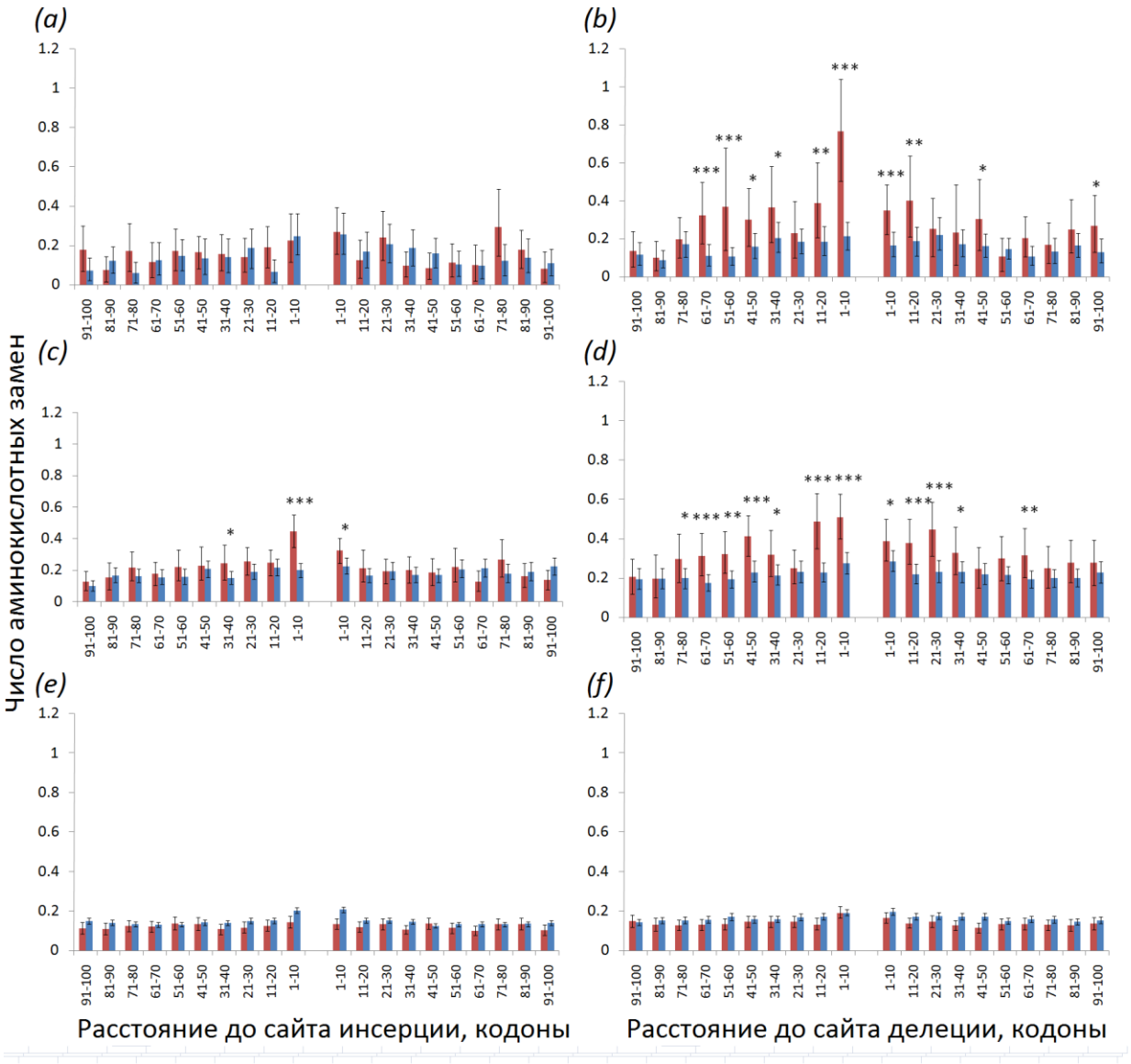

Рис. 27. Число аминокислотных замен в окружении сайтов инсерций и делеций. Графики и обозначения соответствуют рис. 21. 95% доверительные интервалы посчитаны бутстреп-анализом с 1000 испытаний. Значимость различий исследуемого случая (красный) и контроля (синий) проверена точным тестом Фишера; $*P < 0{,}05$, $**P < 0{,}01$,$***P < 0{,}005$.

Наблюдаемое ускорение адаптивной эволюции зависит от времени прошедшего с момента возникновения мутации. Наш анализ позволяет сравнивать эволюцию для инделов, произошедших сравнительно недавно, в терминальном сегменте линии *D. melanogaster* или до ответвления от *D. erecta,* с эволюцией инделов, которые произошли до ответвления линии *D. melanogaster* от линий *D. ananassae* и *D. erecta* (Рисунок 3 e–f'). Такие древние инделы уже не вызывают ускорения эволюции в терминальном сегменте (Рисунок 26 e–f'), что означает, что адаптивная прогулка продолжается в течение не очень длительного времени.

Значительный контраст наблюдается при сравнении инсерций с делециями. По сравнению с инсерцией, делеция вызывает замены в более широком участке белка (до 100 аминокислотных остатков, по сравнению с ~40 аминокислотными остатками для инсерций); большая часть замен на всём участке закрепляется под действием положительного отбора (Рисунок 26 a против b, c против d). Для делеций (Рисунок 26 b-b'), но не для инсерций (Рисунок 26 a-a'), значимое увеличение числа аминокислотных замен наблюдалось даже когда индел случился в терминальном сегменте линии *D. melanogaster*, где более слабый эффект ожидался по причине того, что в такой конфигурации невозможно различить замены, которые произошли после индела, от тех, что произошли до.

Такой контраст может объясняться тем, что делеция сильнее влияет на структуру и функцию белка, а значит, вновь возникшая делеция имеет более высокую вероятность быть для белка вредной, чем инсерция. Данные по частотам

инсерций и делеций, расщепляющихся в популяции *D. melanogaster* (Рисунок 28), показывают, что инсерции расщепляются на более высоких частотах по сравнению с делециями, что соответствует выдвинутой гипотезе. Значение статистики Tajima's D [21] для делеций (-1,64) ниже, чем для инсерций (-1,19), что означает, что делеция в гене в среднем вреднее инсерции.

Среди замен, которые произошли после события индела, доля тех, что произошли в консервативных сайтах, была выше в исследуемом случае по сравнению с контролем (Рисунок 26). Это различие также было более выражено для делеций, чем для инсерций (Рисунок 26 b против b', d против d', f против f'). Таким образом, индел изменяет отбор на близлежащие аминокислоты, что соответствует длинному «прыжку» на ландшафте приспособленности (Рисунок2).

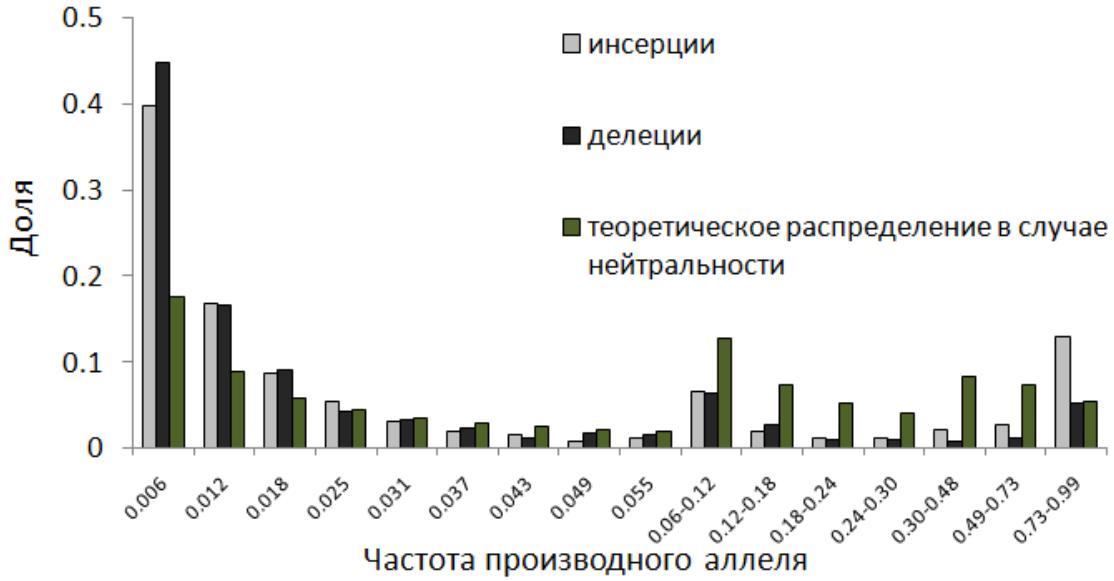

Рис. 28 Спектр аллельных частот для делеций и инсерций одной аминокислоты в белках *D. melanogaster*. Средняя частота производного аллеля

составляет 0,097 для инсерций (серый) и 0,057 для делеций (чёрный) (значимость различия в тесте Вилкоксона, *P=0,0006*). Оба значения значительно меньше 0,175 (средняя частоты производного аллеля, предсказнная в модели бесконеченого числа сайтов для нейтрального случая).

Выводы

1) На 1 нейтральную однонуклеотидную замену у *D. melanogaster* в среднем происходит 0,036–0,039 инсерций (средняя длина 3,23 нт) и 0,085–0,092 делеций (средняя длина 4,78 нт). Таким образом, на уровне мутагенеза наблюдается сильное смещение в сторону делеций; при отсутствии отбора происходило бы сокращение генома на 0.3нт на каждую однонуклеотидную замену.

2) Действие отрицательного и положительного отбора препятствуют сокращению генома. Среди новых мутаций доля отсеиваемых отрицательным отбором составляет 71% для инсерций и 82% для делеций. Среди мутаций, закрепляющихся в межвидовой эволюции, доля закреплённых под действием положительного отбора составляет 67% для инсерций и 36% для делеций.

3) Инсерции и делеции подвергаются действию генной конверсии, смещённой в сторону инсерций. Смещение вероятности закрепления соответствует усреднённому по геному конверсионному преимуществу для инсерций $N_e\omega$ = ~0,3–0,4.

4) Скорость эволюции в синонимических сайтах не меняется после возникновения индела. Из этого следует, что инделы не обладают мутагенным эффектом на окружающую последовательность.

5) Скорость эволюции в несинонимических сайтах значительно возрастает после возникновения индела, причём этот эффект сильнее для делеций.

Несинонимичные замены, вызванные инделами, носят исключительно адаптивный характер.

6) Событие инсерции вызывает приблизительно 1 дополнительную аминокислотную замену, а событие делеции – приблизительно 5 дополнительных аминокислотных замен. Различие между инсерциями и делециями, вероятно, связано с тем, что делеции снижают приспособленность сильнее инсерций.

7) Анализ инделов разных возрастов показывает, что адаптивная прогулка происходит за время, которое соответствует 0,1–0,7 синонимическим заменам.

8) Среди замен, которые произошли после события индела, доля тех, что произошла в консервативных сайтах, была выше, чем для замен, не связанных с инделами. Для делеций это различие выражено сильнее, чем для инсерций.

Благодарности

Кондрашову А.С., Вахрушевой О.А., Сеплярскому В.Б., Хайруллину А.Ф., Виноградовой С.В., Науменко С.А. за ценные обсуждения и замечания, а также Заике А.В и Виноградову Д.В за помощь в работе на вычислительном кластере.

Список публикаций по теме диссертации

Статьи в изданиях, рекомендованных перечнем ВАК

Leushkin EV, Bazykin GA, Kondrashov AS. Insertions and deletions trigger adaptive walks in *Drosophila* proteins // *Proc Biol Sci.*- 2012. - Vol. 279. - P. 3075-3082

Leushkin EV, Bazykin GA, Kondrashov AS. Strong mutational bias towards deletions in the *Drosophila melanogaster* genome is compensated by selection // *Genome Biol Evol*. -2013. - Vol. 5. - P. 514-524

Leushkin EV, Bazykin GA. Short indels are subject to insertion-biased gene conversion // *Evolution*. - 2013. - Vol. 67. - P. 2604-4613

Тезисы конференций

Leushkin EV, Bazykin GA, Kondrashov AS. Adaptive amino acid replacements triggered by indels in *Drosophila* proteins // MCCMB 2011: Proceedings of the International Moscow Conference on Computational Molecular Biology. - Moscow, Russia, 2011. - P. 198

Leushkin E.V., Bazykin G.A., Kondrashov A.S. Insertions and deletions trigger adaptive walks in *Drosophila* proteins. // Otto Warburg International Summer School and Research Symposium 2011 on Evolutionary Genomics. - Berlin, Germany. 2011.

Leushkin EV, Kondrashov AS, Bazykin GA. Insertion-biased gene conversion for short indels // SMBE 2012: Annual Meeting of the Society for Molecular Biology and Evolution. - Dublin, Ireland, 2012. - P. 1082

Leushkin E.V. Bazyking G.A.. Insertion-biased gene conversion for short indels. // ITaS 2012: Information Technology and Systems – 2012. - Petrozavodsk, Russia, 2012. - P. 319

Список литературы

1 Britten RJ. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels // *Proc Natl Acad Sci U S A*. - 2002. - Vol. 99. - P. 13633–13635.

2 Ananda G, Walsh E, Jacob KD, Krasilnikova M, Eckert KA, Chiaromonte F, *et al.* Distinct mutational behaviors differentiate short tandem repeats from microsatellites in the human genome // *Genome Biol Evol*. - 2013. - Vol. 5. - P. 606–620.

3 Kofler R, Schlötterer C, Luschützky E, Lelley T. Survey of microsatellite clustering in eight fully sequenced species sheds light on the origin of compound microsatellites // *BMC Genomics*. - 2008. - Vol. 9. - P. 612.

4 Williams LE, Wernegreen JJ. Sequence context of indel mutations and their effect on protein evolution in a bacterial endosymbiont // *Genome Biol Evol*. - 2013. - Vol. 5. - P. 599–605.

5 Messer PW, Arndt PF. The majority of recent short DNA insertions in the human genome are tandem duplications // *Mol Biol Evol*. - 2007. - Vol. 24. - P. 1190–1197.

6 Viguera E, Canceill D, Ehrlich SD. Replication slippage involves DNA polymerase pausing and dissociation // *EMBO J*. - 2001. - Vol. 20. - P. 2587–2595.

7 Tanay A, Siggia ED. Sequence context affects the rate of short insertions and deletions in flies and primates // *Genome Biol*. - 2008. - Vol. 9. - P. R37.

8 Chaux N de la, Messer PW, Arndt PF. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage // *BMC Evol Biol*. - 2007. - Vol. 7. - P. 191.

9 Zhang J. Evolution by gene duplication: an update // *Trends Ecol Evol*. - 2003. - Vol. 18. - P. 292–298.

10 Petrov DA. DNA loss and evolution of genome size in Drosophila // *Genetica*. - 2002. - Vol. 115. - P. 81–91.

11 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A, Brooks LD, Durbin RM, *et al.* A map of human genome variation from population-scale sequencing // *Nature*. - 2010. - Vol. 467. - P. 1061–1073.

12 Lynch M. Rate, molecular spectrum, and consequences of human mutation // *Proc Natl Acad Sci*. - 2010. - Vol. 107. - P. 961–968.

13 Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate // *PLoS Biol*. - 2009. - Vol. 7. - P. e1000027.

14 Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, *et al.* Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes // *Nature*. - 2008. - Vol. 455. - P. 105–108.

15 Kelkar YD, Strubczewski N, Hile SE, Chiaromonte F, Eckert KA, Makova KD. What Is a Microsatellite: A Computational and Experimental Definition Based upon Repeat Mutational Behavior at A/T and GT/AC Repeats // *Genome Biol Evol*. - 2010. - Vol. 2. - P. 620–635.

16 Kondrashov AS, Rogozin IB. Context of deletions and insertions in human coding sequences // *Hum Mutat*. - 2004. - Vol. 23. - P. 177–185.

17 Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, *et al.* Global variation in copy number in the human genome // *Nature*. - 2006. - Vol. 444. - P. 444–454.

18 Moreno-De-Luca D, Mulle JG, Kaminsky EB, Sanders SJ, Myers SM, Adam MP, *et al.* Deletion 17q12 Is a Recurrent Copy Number Variant that Confers High Risk of Autism and Schizophrenia // *Am J Hum Genet*. - 2010. - Vol. 87. - P. 618–630.

19 Stahl PJ, Mielnik AN, Barbieri CE, Schlegel PN, Paduch DA. Deletion or underexpression of the Y-chromosome genes CDY2 and HSFY is associated with maturation arrest in American men with nonobstructive azoospermia // *Asian J Androl*. - 2012. - Vol. 14. - P. 676–682.

20 McCarroll SA, Huett A, Kuballa P, Chilewski SD, Landry A, Goyette P, *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease // *Nat Genet*. - 2008. - Vol. 40. - P. 1107–1112.

21 D'haene B, Attanasio C, Beysen D, Dostie J, Lemire E, Bouchard P, *et al.* Disease-Causing 7.4 kb Cis-Regulatory Deletion Disrupting Conserved Non-Coding Sequences and Their Interaction with the FOXL2 Promotor: Implications for Mutation Screening // *PLoS Genet*. - 2009. - Vol. 5. doi:10.1371/journal.pgen.1000522

22 A novel gene containing a trinucleotide repeat that is expanded and unstable on Huntington's disease chromosomes. The Huntington's Disease Collaborative Research Group // *Cell*. - 1993. - Vol. 72. - P. 971–983.

23 Loomis EW, Eid JS, Peluso P, Yin J, Hickey L, Rank D, *et al.* Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile X gene // *Genome Res*. - 2012. - . - P. gr.141705.112.

24 Eyre-Walker A, Keightley PD. Estimating the Rate of Adaptive Molecular Evolution in the Presence of Slightly Deleterious Mutations and Population Size Change // *Mol Biol Evol*. - 2009. - Vol. 26. - P. 2097 – 2108.

25 Casillas S, Barbadilla A, Bergman CM. Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila // *Mol Biol Evol*. - 2007. - Vol. 24. - P. 2222 –2234.

26 Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila // *Nature*. - 2005. - Vol. 437. - P. 1149–1152.

27 Halligan DL, Keightley PD. Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison // *Genome Res*. - 2006. - Vol. 16. - P. 875–884.

28 Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila // *Mol Biol Evol*. - 2010. - Vol. 27. - P. 1226–1234.

29 Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, *et al.* The Drosophila melanogaster Genetic Reference Panel // *Nature*. - 2012. - Vol. 482. - P. 173–178.

30 McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila // *Nature*. - 1991. - Vol. 351. - P. 652–654.

31 Smith NGC, Eyre-Walker A. Adaptive protein evolution in Drosophila // *Nature*. - 2002. - Vol. 415. - P. 1022–1024.

32 Charlesworth J, Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations // *Mol Biol Evol*. - 2008. - Vol. 25. - P. 1007–1015.

33 Bazykin GA, Kondrashov FA, Ogurtsov AY, Sunyaev S, Kondrashov AS. Positive selection at sites of multiple amino acid replacements since rat-mouse divergence // *Nature*. - 2004. - Vol. 429. - P. 558–562.

34 Bazykin GA, Kondrashov AS. Major role of positive selection in the evolution of conservative segments of Drosophila proteins // *Proc Biol Sci*. - 2012. - Vol. 279. - P. 3409–3417.

35 Smith JM, Haigh J. The hitch-hiking effect of a favourable gene // *Genet Res*. - 1974. - Vol. 23. - P. 23–35.

36 Kaplan NL, Hudson RR, Langley CH. The “hitchhiking effect” revisited. // *Genetics*. - 1989. - Vol. 123. - P. 887–899.

37 Charlesworth B, Morgan MT, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. // *Genetics*. - 1993. - Vol. 134. - P. 1289–1303.

38 Hudson RR. How can the low levels of DNA sequence variation in regions of the drosophila genome with low recombination rates be explained? // *Proc Natl Acad Sci*. - 1994. - Vol. 91. - P. 6815–6818.

39 Charlesworth B. The Effects of Deleterious Mutations on Evolution at Linked Sites // *Genetics*. - 2012. - Vol. 190. - P. 5 –22.

40 Birky CW, Walsh JB. Effects of linkage on rates of molecular evolution // *Proc Natl Acad Sci*. - 1988. - Vol. 85. - P. 6414–6418.

41 McVean GA, Charlesworth B. The effects of Hill-Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. // *Genetics*. - 2000. - Vol. 155. - P. 929–944.

42 Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over // *Genome Biol*. - 2007. - Vol. 8. - P. R18.

43 Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster // *Nature*. - 1992. - Vol. 356. - P. 519–520.

44 Nachman MW. Single nucleotide polymorphisms and recombination rate in humans // *Trends Genet TIG*. - 2001. - Vol. 17. - P. 481–485.

45 Cutter AD, Payseur BA. Selection at Linked Sites in the Partial Selfer Caenorhabditis elegans // *Mol Biol Evol*. - 2003. - Vol. 20. - P. 665–673.

46 Nordborg M, Innan H. Molecular population genetics // *Curr Opin Plant Biol*. - 2002. - Vol. 5. - P. 69–73.

47 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.

48 Mancera E, Bourgon R, Brozzi A, Huber W, Steinmetz LM. High-resolution mapping of meiotic crossovers and non-crossovers in yeast // *Nature*. - 2008. - Vol. 454. - P. 479–485.

49 Buard J, de Massy B. Playing hide and seek with mammalian meiotic crossover hotspots // *Trends Genet*. - 2007. - Vol. 23. - P. 301–309.

50 Kauffman S, Levin S. Towards a general theory of adaptive walks on rugged landscapes // *J Theor Biol*. - 1987. - Vol. 128. - P. 11–45.

51 Gillespie JH. Molecular Evolution Over the Mutational Landscape // *Evolution*. - 1984. - Vol. 38. - P. 1116–1129.

52 Orr HA. The genetic theory of adaptation: a brief history // *Nat Rev Genet*. - 2005. - Vol. 6. - P. 119–127.

53 Orr HA. Fitness and its role in evolutionary genetics // *Nat Rev Genet*. - 2009. - Vol. 10. - P. 531–539.

54 Kryazhimskiy S, Tkacik G, Plotkin JB. The dynamics of adaptation on correlated fitness landscapes // *Proc Natl Acad Sci U S A*. - 2009. - Vol. 106. - P. 18638–18643.

55 Smith JM. Natural selection and the concept of a protein space // *Nature*. - 1970. - Vol. 225. - P. 563–564.

56 Romero PA, Arnold FH. Exploring protein fitness landscapes by directed evolution // *Nat Rev Mol Cell Biol*. - 2009. - Vol. 10. - P. 866–876.

57 Gillespie JH. *The Causes of Molecular Evolution* Oxford University Press, USA; 1994.

58 Orr HA. A minimum on the mean number of steps taken in adaptive walks // *J Theor Biol*. - 2003. - Vol. 220. - P. 241–247.

59 Orr HA. The population genetics of adaptation on correlated fitness landscapes: the block model // *Evol Int J Org Evol*. - 2006. - Vol. 60. - P. 1113–1124.

60 Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness // *Nature*. - 2010. - Vol. 464. - P. 279–282.

61 Chou H-H, Chiu H-C, Delaney NF, Segrè D, Marx CJ. Diminishing returns epistasis among beneficial mutations decelerates adaptation // *Science*. - 2011. - Vol. 332. - P. 1190–1192.

62 Khan AI, Dinh DM, Schneider D, Lenski RE, Cooper TF. Negative epistasis between beneficial mutations in an evolving bacterial population // *Science*. - 2011. - Vol. 332. - P. 1193–1196.

63 Kvitek DJ, Sherlock G. Reciprocal sign epistasis between frequently experimentally evolved adaptive mutations causes a rugged fitness landscape // *PLoS Genet*. - 2011. - Vol. 7. - P. e1002056.

64 Wolf JB, III EDB, Wade MJ. *Epistasis and the Evolutionary Process* 1st ed. Oxford University Press, USA; 2000.

65 Crosby MA, Goodman JL, Strelets VB, Zhang P, Gelbart WM, The FlyBase Consortium. FlyBase: genomes by the dozen // *Nucleic Acids Res*. - 2007. - Vol. 35. - P. D486–D491.

66 Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D. The UCSC Known Genes // *Bioinformatics*. - 2006. - Vol. 22. - P. 1036 –1046.

67 Hubbard T, Barker D, Birney E, Cameron G, Chen Y, Clark L, *et al.* The Ensembl genome database project // *Nucleic Acids Res*. - 2002. - Vol. 30. - P. 38 –41.

68 Comeron JM, Ratnappan R, Bailin S. The Many Landscapes of Recombination in Drosophila melanogaster // *PLoS Genet*. - 2012. - Vol. 8. doi:10.1371/journal.pgen.1002905

69 Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, *et al.* A high-resolution recombination map of the human genome // *Nat Genet*. - 2002. - Vol. 31. - P. 241–247.

70 Buhler C, Borde V, Lichten M. Mapping meiotic single-strand DNA reveals a new landscape of DNA double-strand breaks in Saccharomyces cerevisiae // *PLoS Biol*. - 2007. - Vol. 5. - P. e324.

71 Messer PW. Measuring the rates of spontaneous mutation from deep and large-scale polymorphism data // *Genetics*. - 2009. - Vol. 182. - P. 1219–1232.

72 Parsch J. Selective constraints on intron evolution in Drosophila // *Genetics*. - 2003. - Vol. 165. - P. 1843–1851.

73 Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence // *Genetics*. - 1992. - Vol. 132. - P. 1161–1176.

74 Haag-Liautard C, Dorris M, Maside X, Macaskill S, Halligan DL, Charlesworth B, *et al.* Direct estimation of per nucleotide and genomic deleterious mutation rates in Drosophila // *Nature*. - 2007. - Vol. 445. - P. 82–85.

75 Begun DJ, Holloway AK, Stevens K, Hillier LW, Poh Y-P, Hahn MW, *et al.* Population genomics: whole-genome analysis of polymorphism and divergence in Drosophila simulans // *PLoS Biol*. - 2007. - Vol. 5. - P. e310.

76 Keightley PD, Eyre-Walker A. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small // *J Mol Evol*. - 2012. - Vol. 74. - P. 61–68.

77 Campos JL, Charlesworth B, Haddrill PR. Molecular Evolution in Nonrecombining Regions of the Drosophila melanogaster Genome // *Genome Biol Evol*. - 2012. - Vol. 4. - P. 278–288.

78 Yang Y, Sterling J, Storici F, Resnick MA, Gordenin DA. Hypermutability of Damaged Single-Strand DNA Formed at Double-Strand Breaks and Uncapped Telomeres in Yeast Saccharomyces cerevisiae // *PLoS Genet*. - 2008. - Vol. 4. - P. e1000264.

79 Hicks WM, Kim M, Haber JE. Increased Mutagenesis and Unique Mutation Signature Associated with Mitotic Gene Conversion // *Science*. - 2010. - Vol. 329. - P. 82–85.

80 Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? // *PLoS Genet*. - 2009. - Vol. 5. - P. e1000495.

81 Lynch M. *The Origins of Genome Architecture* 1st ed. Sinauer Associates; 2007.

82 Siepel A, Haussler D. Combining Phylogenetic and Hidden Markov Models in Biosequence Analysis // *J Comput Biol*. - 2004. - Vol. 11. - P. 413–428.

83 Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. Meiotic Gene Conversion Tract Length Distribution within the Rosy Locus of Drosophila Melanogaster // *Genetics*. - 1994. - Vol. 137. - P. 1019–1026.

84 Garcia-Diaz M, Kunkel TA. Mechanism of a genetic glissando*: structural biology of indel mutations // *Trends Biochem Sci*. - 2006. - Vol. 31. - P. 206–214.

85 Bill CA, Duran WA, Miselis NR, Nickoloff JA. Efficient Repair of All Types of Single-Base Mismatches in Recombination Intermediates in Chinese Hamster Ovary Cells: Competition Between Long-Patch and G-T Glycosylase-Mediated Repair of G-T Mismatches // *Genetics*. - 1998. - Vol. 149. - P. 1935–1943.

86 Dreszer TR, Wall GD, Haussler D, Pollard KS. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion // *Genome Res*. - 2007. - Vol. 17. - P. 1420–1430.

87 Carvalho AB, Clark AG. Intron size and natural selection // *Nature*. - 1999. - Vol. 401. - P. 344.

88 Comeron JM, Kreitman M. The correlation between intron length and recombination in drosophila. Dynamic equilibrium between mutational and selective forces // *Genetics*. - 2000. - Vol. 156. - P. 1175–1190.

89 Barton NH, Charlesworth B. Why Sex and Recombination? // *Science*. - 1998. - Vol. 281. - P. 1986–1990.

90 Zhang Z, Huang J, Wang Z, Wang L, Gao P. Impact of indels on the flanking regions in structural domains // *Mol Biol Evol*. - 2011. - Vol. 28. - P. 291–301.