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

На правах рукописи

Сеплярский Владимир Борисович

ПАТТЕРНЫ И МОЛЕКУЛЯРНЫЕ МЕХАНИЗМЫ МУТАГЕНЕЗА У ЭУКАРИОТ

03.01.09 - математическая биология, биоинформатика

Диссертация на соискание учёной степени кандидата биологических наук

Научный руководитель:

кандидат биологических наук

Базыкин Георгий Александрович

Москва – 2015

Содержание

Введение

Изучение закономерностей мутагенеза имеет большое значение для медицинских и эволюционных исследований. Так, предположения о том, как распределены мутации вдоль генома, лежат в основе поиска генов, ассоциированных с различными заболеваниями; неверная оценка локальной скорости мутирования приводит к тому, что генами, ассоциированными с заболеваниями, считают гены с высокой скоростью мутирования [1]. Данные по полным геномам для различных организмов, для внутривидового полиморфизма и для троек «пара родителей-потомок» дают возможность изучать мутационный процесс для замен, передающихся по наследству. К сожалению, наиболее прямые данные по *de-novo* мутациям, полученные из сравнения геномов родителей с ребёнком, пока бедны: для человека описано только несколько десятков тысяч таких мутаций [2–5], что ограничивает возможности исследования факторов, влияющих на мутационный процесс, по таким данным. Особенности мутагенеза в разных типах рака можно исследовать, используя множество пар образцов здоровой и раковой ткани. Такие данные начинают появляться в обилии; так, крупнейшая база данных по раковым сиквенсам TCGA на данный момент включает 10247 полных экзомов [1,6].

Гетерогенность скорости мутирования вдоль генома давно исследуется с использованием межвидовых сравнений [7,8]. Более того, в моделях наибольшего правдоподобия, рассчитывающих вероятность нуклеотидных замен на ветках

филогенетического дерева, локальная неравномерность скорости мутирования уже учитывается [9]. Описанная поправка крайне важна для получения релевантных результатов о воздействии отбора, в особенности на некодирующую последовательность.

Предмет изучения в нашей работе - это локальная изменчивость скоростей точечного мутирования (Глава 1) и её влияние на соотношение различных типов мутаций (Глава 2), а также сложные мутации, одновременно меняющие несколько близлежащих нуклеотидов (Главы 3 и 4). Наше исследование сосредоточено на хорошо изученных модельных организмах – *Drosophila melanogaster* и *Homo sapiens* с использованием геномов ближайших к ним видов. Мы разработали простой подход для поиска сложных мутаций по данным о межвидовой дивергенции. Этот метод позволил исследовать, какие геномные свойства определяют распределение двойных мутаций GC→TT/AA, являющихся мутационной подписью полимеразы ζ [10–12]. Что позволило нам сделать вывод о том, что работа этой полимеразы частично объясняет изменчивость скорости мутирования вдоль по геному.

1. Гетерогенность скорости мутирования вдоль по геному

Однонуклеотидные мутации являются причиной генетических заболеваний и служат материалом для эволюции. Они – самый частый тип мутационных событий. Скорость однонуклеотидных мутаций не менее чем на порядок превосходит скорость других мутаций – сложных мутаций, инсерций, делеций,

инверсий и др. [13,14]. При изучении белковой эволюции наблюдают радикальное изменение локальной скорости эволюции [15,16]. Таким образом, крайне важно понимать локальную гетерогенность скорости мутирования, чтобы разделять вклад отборной и мутационной компоненты в неё.

Полные геномы различных видов, а также полногеномные данные по внутривидовому полиморфизму, позволяют изучать гетерогенность скорости мутирования. Для исследования мутационных процессов необходимо исключить влияние отбора, и изучение лишь нейтрально эволюционирующих участков генома является приемлемым решением. Позиции в геноме, особенно подверженные мутациям, называются горячими точками мутагенеза. Одно из объяснений для этого феномена – неканонический механизм удвоения или починки ДНК. Можно ожидать, что соотношение различных типов мутаций в них будет обладать другими свойствами.

1.1. Подходы к изучению скорости мутирования и её мелкомасштабной изменчивости

Для сбора данных об однонуклеотидных мутациях можно сравнивать: геномы разных тканей, что рассказывает о соматических мутациях - мутациях, накопленных за время жизни особи [17,18]; полные геномы родителей и их потомков [2–4]; геномы родителей и их потомков через несколько поколений в системах, где отбор не эффективен («линии накопления мутаций») [19,20]; древние датированные геномы и геномы ныне живущих особей [21]; или же

последовательности геномов разных видов [22]. Также можно использовать базы данных по генетическим заболеваниям, содержащие очень много последовательностей для некоторых локусов [13].

К сожалению, все перечисленные источники данных по однонуклеотидным мутациям имеют свои недостатки. Количество известных *de-novo* мутаций между родителями и потомками или в линиях накопления мутаций мало, а для изучения локальной гетерогенности скоростей мутирования необходимо оперировать сразу большим количеством однонуклеотидных мутаций. Базы данных заболеваний очень неравномерно покрывают различные локусы, что также будет приводить к смещению в оценке распределения скорости мутирования. Соматические мутации часто недостаточно высокого качества, и это может приводить к локальной кластеризации однонуклеотидных мутаций. Более того, на уровне соматических клеток часто работают мутационные процессы, не проявляющиеся в зародышевых линиях [1,6,18,23]. На данный момент, внутривидовой полиморфизм и межвидовая дивергенция являются приемлемым источником данных для исследования локальной гетерогенности скорости мутирования. Тем не менее, на распределение нуклеотидных отличий накопленных за много поколений влияет отбор и, работая с полиморфизмом или межвидовой дивергенцией, надо исключать возможное влияние отбора.

1.2. Мутационные контексты и мелкомасштабная гетерогенность скорости мутирования

Каждый из 4 нуклеотидов может замениться на любой из 3 оставшихся, таким образом, существует 12 типов однонуклеотидных мутаций (6 если рассматривать комплементарные мутации вместе). Частоты каждого типа мутаций могут зависеть от ближайшего нуклеотидного окружения, что было детально исследовано ранее [6,24]. Наиболее мутабильный контекст для замен в зародышевой линии у млекопитающих – это динуклеотид CpG. Цитозин в динуклеотиде CpG часто метилирован, что повышает вероятность спонтанного деаминирвания с образованием тимина [25]. Скорость мутирования CpG → TpG более чем в 10 раз превышает среднюю по геному [24]. Другая контекстно-зависимая мутация, имеющая повышенную частоту в приматах – A<u>TN</u>→ACN, однако механизм, вызывающий мутабильность этого контекста, не известен [24]. Остальные контексты повышают скорость замен гораздо слабее [24]. Мутабильность контекстов может различаться между популяциями; так, частота мутаций C→T в T<u>C</u>C контексте варьирует в 1.5 раза между популяциями человека [26], эти мутации могут быть вызваны действием ультрафиолета [27]. Для раковых мутаций известен контекст T<u>C</u>A→ TTA и T<u>C</u>A→ TGA, связанный с работой белка APOBEC [6,28–30].

Однонуклеотидные мутации в геноме распределены неравномерно, и эта неравномерность сохраняется в ходе эволюции: вероятность однонуклеотидного полиморфизма (SNP – single nucleotide polymorphism) в человеческой популяции

вдвое выше в тех сайтах, где наблюдается SNP в шимпанзе [31]. Однако на соседние нуклеотиды этот эффект почти не распространяется: скорость мутирования у человека почти не увеличена в сайтах, соседних с SNP в шимпанзе [31]. Очень сходные паттерны наблюдали при сравнении замен в парах человек-шимпанзе и орангутанг-макака [32]. Наличие одиночных сайтов в геноме с вдвое повышенной скоростью мутирования не объяснятся мутационными контекстами и является свойством «криптической» изменчивости скорости мутирования вдоль генома. Для соматических раковых мутаций известна кластеризация мутаций определённого типа на одной цепи ДНК, что показывает, что в онкогенезе участки генома могут локально подвергаться влиянию конкретного мутагена, и во многих случаях можно предсказать, какого именно [33].

1.3. Практическая важность изучения локальных мутационных процессов

Детальное знание изменчивости мутагенеза имеет прикладное значение, в том числе в области персонализированной медицины. Модели, предполагающие равномерную вероятность мутирования по геному, в ряде исследований по поиску драйверов рака выдавали неверные гены-кандидаты [1]. Так, в [34] как онкоген был выявлен очень длинный ген титин, поскольку мутации случались в нем многократно из-за его большой длины; а также запаховые рецепторы, имеющие повышенную скорость мутагенеза. Верная модель мутирования, учитывающая гетерогенность этого процесса, объяснила большую долю рекуррентных мутаций просто свойствами мутационного процесса [1]. Отрицательный отбор на

последовательность может приводить и к уменьшению частоты замен функционально важного участка генома и, как следствие, к его эволюционной консервативности. Гены млекопитающих, находящиеся под действием отрицательного отбора, содержат больше CpG динуклеотидов, чем нейтральная последовательность, и поэтому в них чаще попадают *de-novo* мутации [13,35].

Кроме того, локальное соотношение типов мутаций (спектр мутирования) может быть косвенным свидетельством различий в мутационных механизмах между разными участками генома. Например, мутации W→S (где W соответствует нуклеотиду A или T, а S – нуклеотиду C или G) гораздо чаще встречаются в сегментах генома с высоким уровнем рекомбинации из-за того, что с рекомбинацией сопряжена смещённая генная конверсия [36,37]. Смещённая генная конверсия – это процесс асимметричной репарации возникающего в ходе рекомбинации гетеродуплекса (участка двухцепочечной ДНК, где комплиментарные цепи пришли из разных гомологичных хромосом, и который поэтому может содержать не спаренные основания) в пользу варианта (аллеля) с большим содержанием G или C нуклеотидов. Ещё один пример связи скорости мутирования и молекулярных механизмов – это пониженная скорость мутирования CpG островов. Низкая скорость мутирования CpG динуклеотидов в составе CpG островов частично объясняется отсутствия метилирования таких цитозинов, что понижает вероятность спонтанного деаминирования.

2. Отношение транзиций к трансверсиям как базовая характеристика мутационного процесса

Локальные свойства мутационного процесса можно описывать не только вероятностью однонуклеотидных мутаций на сайт, но и соотношением различных типов таких мутаций. Каждый нуклеотид может замениться на 3 других нуклеотида, причём в двух из трёх случаев замены будут приводить к трансверсии – смене типа нуклеинового основания (замена пурина на пиримидин и наоборот); в третьем случае замена не приведёт к смене типа азотистого основания, такая замена называется транзицией.

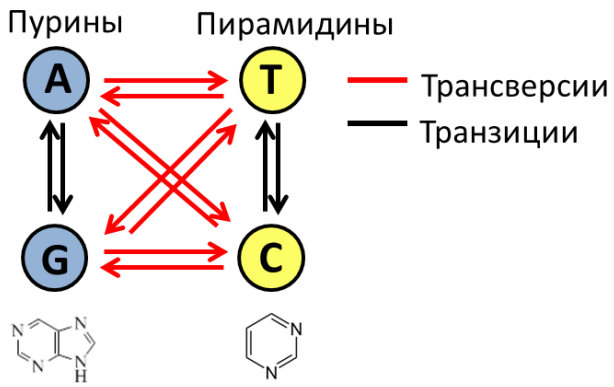

Рис. 1. Схематическое изображение транзизиций и трансверсий.

Таким образом, классов трансверсий вдвое больше, чем классов транзиций (4 против 8). Отношение числа транзиций к числу трансверсий, умноженное на 2 для того, чтобы нормировать на различное количество классов этих мутаций, обозначается **к** и является широко употребляемой однопараметрической

характеристикой мутационного спектра. Отношение транзиций к трансверсиям – это наиболее простая характеристика мутационного спектра, что важно при анализе малого количества мутаций, как например в [38,39]. Для всех известных эукариот, с измеренной **к, к** > 1, кроме *Podisma pedestris,* для которого **к**=1.13 и недостоверно отличается от 1. **к** > 1 свидетельствует о том, что повреждения ДНК, приводящие к транзициям, случаются чаще, или же, что репарация транзиций менее эффективна.

Таблица 1. Значения **к** в эукариотах. Звёздочкой отмечены оценки, полученные в данной работе. Кроме *Podisma pedestris,* все приведенные значения **к** достоверно превышают 1.

Организм	K
Homo sapiens	3.86*
Drosophila melanogaster	2.06*
Podisma pedestris	1.13 [38]
Schizophyllum commune	4.28*
Saccharomyces cerevisiae	3.2 [40]
Caenorhabditis elegans	>2 [41]
Arabidopsis thaliana	2.4 [39]

Значение **к** отличается не только между организмами, но и вдоль генома. Один из важнейших детерминантов скорости мутирования – время репликации влияет и на соотношение частот транзиций и трансверсий: в участках поздней репликации доля трансверсий выше [42,43]. Среди мутаций, вызываемых склонными к ошибкам полимеразами, например полимеразой зета (пол ζ), больше доля трансверсий, по сравнению с основными репликативными полимеразами [10,44]. Есть гипотеза, что повышенная скорость мутирования в участках поздней репликации, как и меньшие значения **к,** являются следствием работы склонных к

ошибкам полимераз [42]. Кроме того, увеличение доли трансверсий наблюдают в ближайшей окрестности инсерций или делеций (инделов) [45,46], что также связывают с работой склонных к ошибкам полимераз. Короткие повторы обогащены мутациями, инделами и имеют пониженную **к** [46]**.**

Изменчивость соотношения транзиций и трансверсий в геноме может свидетельствовать в пользу локальной смены мутационного механизма.

3. Мультинуклеотидные мутации

Существуют различные типы мутаций: однонуклеотидные – замена одной пары нуклеотидных оснований, мультинуклеотидные – одновременная замена нескольких пар оснований, инделы – выпадение или вставка последовательности ДНК, сложные события – любая комбинация инделов и однонуклеотидных или мультинуклеотидных мутаций, инверсии и различные виды геномных перестроек. Среди событий, затрагивающих лишь малый сегмент ДНК, мультинуклеотидные мутации являются самым малоизученным типом.

Понимание того, как устроены мультинуклеотидные замены, может пролить свет на определённые мутационные механизмы. Например, известно, что пол ζ склонна делать GC→TT/AA динуклеотидные мутации [10,11]. Исследование таких динуклеотидных мутаций может быть крайне информативным для понимания особенностей работы пол ζ.

В раковых опухолях также известны множественные мутации, ассоциированные с конкретными мутационнынми механизмами. Считается, что

соматические CC→TT мутации в раках вызваны, прежде всего, ультрафиолетовым (УФ) излучением, а кластеры TpC→TpT и TpC→TpG мутаций вызваны работой фермента APOBEC [6,33].

С другой стороны, знание мультинуклеотидного мутационного ландшафта необходимо для изучения эпистатических эффектов в близко расположенных сайтах. В [15] было показано, что пары замен в одном кодоне между видами происходят скоррелированно во времени из-за действия отбора. Чтобы верно разделить мутационный и отборный эффекты, необходимо понимать, как устроены мультинуклеотидные мутации. Более того, динуклеотидные мутации могут позволять «пересекать адаптивные долины» – делая возможными эволюционные события, которые не могут произойти как пара однонуклеотидных мутаций из-за их вредности поодиночке [47].

3.1. Методы изучения мультинуклеотидных мутаций

Мультинуклеотидные мутации (MHM) – более сложный для изучения объект, чем однонуклеотидные мутации или инделы, т.к. их следует отличать от нескольких последовательных однонуклеотидных замен. Оценки количества динуклиотидных мутаций можно получить по тем же данным, что и для однонуклеотидных замен: по *de-novo* мутациям [5,20,35,48]; по раковым данным [6,49]; по многократно секвенированным генам [13,50,51]; по полиморфизму [12,48]; или по межвидовым сравнениям [52]. При этом, из-за того что мультинуклеотидные мутации редки, требуется ещё больше данных чем для

изучения гетерогенности локальной скорости мутирования, и вышеописанные подходы для получения мультинуклеотидных мутаций имеют сходные недостатки, что и для получения точечных мутаций.

В большинстве работ количество динуклеотидных мутаций нормируют на количество однонуклеотидных. Рассмотрим $\alpha_d(k)$ – отношение частот динуклеотидных мутаций, затрагивающих два нуклеотида на расстоянии k друг от друга, к числу однонуклеотидных мутаций. В большинстве обсуждаемых работ рассматривался случай соседних нуклеотидов (k=1).

При высоком качестве коллирования динуклиотидных мутаций оценка $\alpha_d(1)$ по *de-novo* мутациям имеет лишь один, но существенный недостаток, а именно – очень малое количество данных даже для большого количества секвенированных троек родители – потомок. Так, в [48] обнаружили 7 мультинуклеотидных мутаций *de-novo* для $k<20$, а в [5], проанализировав геномы 250 семей, нашли только 78 мультинуклеотидных мутаций.

Измерение $\alpha_d(k)$ по полиморфизму требует сложных моделей, т.к. рекомбинация может разбить динуклеотидную мутацию, или, напротив две мутации, произошедшие не одновременно, могут выглядеть как динуклеотидная мутация. Однако, если учесть все эти ограничения, можно получить большое количество динуклеотидных событий [12,48].

Оценки частоты динуклеотидных мутаций, полученные по соматическим мутациям в раках, могут не соответствовать тому, что наблюдаеться в зародышевых линиях [1,6,33,49]. Так, самая частая динуклеотидная мутация в

полиморфизме человека и в генах, ассоциированных с заболеваниями – это GC→AA/TT (около 10%) [12,50,51], а в раках – CC/GG→ AA/TT (31%) [53].

Основная проблема оценок частоты мультинуклеотидных мутаций, полученных по базам данных заболеваний – смещение в пользу замен, приводящих к болезням.

Межвидовые сравнения – возможно, самый удачный источник данных по динуклеотидным мутациям, однако в работе [52] потребовалось дополнительное предположение о том, что $\alpha_d(k)=0$ для $k>1$, что, вероятно, занизило оценку $\alpha_d(1)$ полученную в этой статье.

Несмотря на вышеперечисленные проблемы различных методов, оценки для $\alpha_d(1)$, полученные разными методами, очень схожи [54], и их среднее составляет 0.41%.

3.2. Молекулярные причины возникновения мультинуклиотидных мутаций

Для некоторых раковых динуклиотидных мутаций известны механизмы их образования. В меланомах в динуклеотидном мутационном спектре преобладает мутация CC/GG→TT/AA, которая вызывается сшивкой пиримидиновых димеров под действием УФ и их последующим спонтанным дезаминированием [33]. CC/GG→AA/TT – наиболее частая динуклеотидная мутация в раках легких; эта мутация вызывается ацетальдегидом [55], содержащимся в табачном дыме.

Эксперименты на модельных организмах – ещё один способ изучать причины мультинуклеотидных мутаций [49]. В экспериментальной системе на

дрожжах и на клетках млекопитающих было показано, что подверженная ошибкам пол ζ особенно часто делает динуклеотидные мутации [10,11,56]. В этих системах пол ζ в первую очередь вставляет динуклеотидные мутации GC→AA/TT.

Информация о том, какие именно причины вызывают сложные мутации, позволяет решать обратную задачу, а именно – исследовать мутационные механизмы, используя мутационные подписи. В раковых данных даже однонуклеотидные мутации могут свидетельствовать о мутационном механизме. Так, избыток TCC→TTC мутаций является сильным свидетельством в пользу УФ облучения пациента [1,6,33]; кластеры TCA→TGA или TCA→TTA мутаций, найденные у пациентов с раком груди, являются убедительным свидетельством работы APOBEC в этих опухолях [57]. Есть множество других мутационных подписей в раках, и они зачастую легко различимы в мутационном спектре из-за огромной доли мутаций данного типа [1,6]. При изучении мутаций по межвидовым сравнениям, даже в десятки раз повышенную скорость мутирования участка, реплицирующегося неточной (low fidelity) полимеразой, невозможно определить, если в череде поколений склонная к ошибкам полимераза реплицирует этот участок не слишком часто. Напротив, мультинуклеотидные мутации из-за редкости их возникновения иногда можно соотнести с конкретным механизмом. Так, в статье [12] обнаружили, что GC→AA/TT - самая частая динуклеотидная мутация в человеческом полиморфизме. В экспериментальных условиях пол ζ особенно часто вызывает такие динуклеотидные мутации [10].

Кроме того, спектр динуклеотидных мутаций в человеческом полиморфизме даже для сайтов, удаленных друг от друга больше чем на один нуклеотид, похож на спектр, соответствующий пол ζ [12]. На этом основании был сделан вывод о том, что пол ζ играет важную роль в мутагенезе зародышевых линий человека. Стоит отметить, что если какой-то геномный участок будет часто реплицироваться благодаря пол ζ, можно будет изучать не только редкие динуклеотидные мутации, но и видеть другие свойственные этой полимеразе особенности, например, повышенную скорость мутирования и пониженную **к.** В [46] заметили, что сложные мутации, а именно инделы, сопровождающиеся заменой, часто происходят в повторах, и предположили, что в повторяющихся участках ДНК часто работают неточные полимеразы. Косвенно эту гипотезу подтвердили тем, что в повторах была существенно повышена скорость мутирования и доля трансверсий [46].

Материалы и методы

1. Данные для расчёта замен

Множественное выравнивание *Callithrix jacchus, Macaca mulatta, Pongo pygmaeus, Gorilla gorilla,* и *Pan troglodytes* на *H. sapiens* было загружено с UCSC [58]. Данные по внутривидовой изменчивости человека были получены по 9 диплоидным геномам, загруженным из Galaxy bioinformatics [59,60]; таким образом, вместе с референтным геномом, мы имеем 19 гаплоидных генотипов. Мы использовали следующие диплоидные геномы: KB1, ABT [60], NA18507 [61],

NA19240 [62], Craig Venter [63], NA12891, NA12892, геном китайца [64] и геном корейца [65]. Данные по аннотированным генам Known Genes были закачены из UCSC [58].

Множественное выравнивание *D. simulans, D. yakuba,* и *D. erecta* на *D. melanogaster* (dm3) было загружено с UCSC [58]. Данные по полиморфизму 37 полных геномов *D. melanogaster*, выровненные на dm3 геном *D. melanogaster* [66] и 6 геномов *D. simulans*, выровненых на dm2 геном *D. melanogaster* [67], были загружены с DPGP (http://www.dpgp.org/). dm2 были переведены в dm3 с использованием программы liftover (http://hgdownload.cse.ucsc.edu/goldenPath/dm3/liftOver/). Для аннотации генов мы использовали FlyBase genes (BDGP release 5) [68].

Мы исключили все аннотированные экзоны, а также 10-нуклеотидные участки по краям интронов, 5’ и 3’ UTR, нуклеотидные сайты, маскированные RepeatMasker и все не-A, C, G или T нуклеотиды, а также X и Y хромосомы. У *Drosophila* мы также исключали консервативные участки со значением phastCons > 0.1 [69]. Из-за влияния предкового полиморфизма на дивергенцию в *Drosophila* мы также исключали полиморфные сайты при анализе межвидовых различий [70].

Для исследования ветко-специфичных нуклеотидных замен мы использовали 2 подхода. Первый подход основан на методе наибольшей экономии. В этом случае мы сравнивали тройки видов: два сестринских вида и ещё один вид в качестве внешнего вида. В приматах мы рассматривали несколько

троек с растущим филогенетическим расстоянием между сестринскими видами: *H. sapiens* и *Pan troglodytes* (*G. gorilla* в качестве внешнего вида), *H. sapiens* и *G. gorilla* (*P. pygmaeus* в качестве внешнего вида), *H. sapiens* и *P. pygmaeus* (*M. mulatta* в качестве внешнего вида) и *H. sapiens* и *M. mulatta* (*C. jacchus* в качестве внешнего вида). В роде *Drosophila* в качестве сестринских видов мы использовали *D. melanogaster* и *D. simulans*, а в качестве внешнего вида были использованы позиции, совпадающие между *D. yakuba* и *D. erecta*; позиции различные между *D. yakuba* и *D. erecta* были исключены. Мы предполагали, что замена произошла на одной из сестринских линий в том случае, когда нуклеотид отличался в этом виде, а у другого сестринского вида и внешнего вида совпадал. Второй подход применялся в роде *Drosophila*, где восстановление предкового состояния менее надежно из-за больших филогенетических расстояний между видами, и мы дополнительно использовали метод наибольшего правдоподобия. В этом случае мы использовали программу baseml из пакета PAML [9].

В гоминидах мы исключали CpG сайты. При анализе мультинуклеотидных замен мы также исключали тандемные замены, которые могли произойти через промежуточное CpG состояние; т.е. мы исключали случаи, когда в первой позиции в предковом или производном состоянии был “C”, а во второй позиции “G”. Чтобы свести вклад ошибок секвенирования к минимуму, мы исключали синглтоны (полиморфизмы, в которых редкий вариант встречался только у одной особи).

Скорость мутирования измерялась как количество замен, делённое на количество сайтов.

Неаллельная генная конверсия может приводить к МНМ, если участок МНК заменяется на паралог отличающейся в двух соседних сайтах. Чтобы исключить влияние неаллельной генной конверсии на частоту МНМ, мы исключали участки, содержащие МНМ, если находили в геноме паралог, неаллельная конверсия с которым может привести к наблюдаемой МНМ.

Для изучения мутагенеза по полиморфизму гриба *S. commune* мы получили собственные данные полногеномного секвенирования (эти данные получены в ходе коллаборации, и легли в основу также других проектов, не выполнявшихся мной). Для этого мы собрали плодовые тела *S. commune* в 3 различных местах в США и в 3 различных местах в России. Из каждого плодового тела были получены одиночные мейоспоры, и генотипы гаплоидного мицелия, выросшего из этих спор (12 американских и 12 российских), были секвенированы. ДНК извлекали из высушенного мицелия с использованием метода CTAB. Библиотеки были подготовлены с использованием TruSeq DNA комплекта (Illumina, США), затем ДНК секвенировали с использованием Illumina HiSeq 2000, парноконцевыми ридами длиной в 101 нуклеотид.

Каждый генотип был собран *de-novo* с использованием SPAdes [71]. Мы получили N50 от 48,928 до 104,000 для различных особей, со средним 68,688 по 24 особям.

Множественное выравнивание с референтным геномом *S. commune* [72] было получено с использованием программы multiz [73]. Мы исключали регионы, несобравшиеся или невыровненные в более чем 8 особях, и сегменты ДНК на расстояниях до 1000 нуклеотидов от таких регионов. По аннотации референтного генома мы нашли 7,987 генов. Мы перевыровняли эти гены программой MACSE [74].

Для анализа мутирования у *S. commune* мы использовали 4-кратно вырожденные синонимичные сайты.

2. Анализ повторяющихся мутаций в одном сайте

Мы использовали 4 различных типа анализов для изучения неравномерности скорости мутирования и **к** вдоль генома (рис. 2).

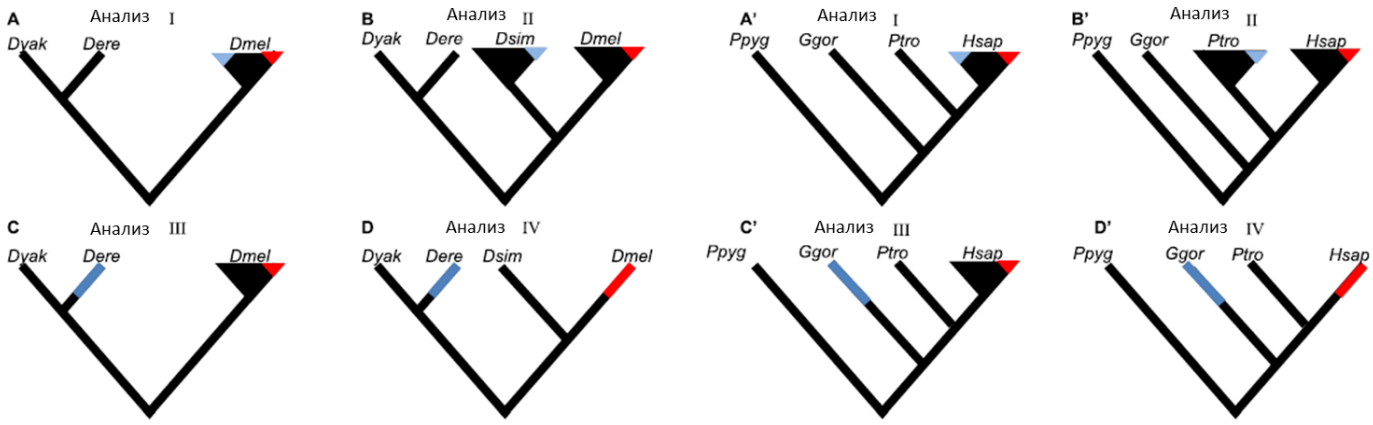

Рис. 2. Четыре типа филогенетических анализов для изучения неравномерности скорости мутирования и **к** вдоль по геному. Линии обозначают филогенетические ветви, закрашенные треугольники обозначают полиморфизм, черным обозначен предковый аллель, а синим и красным - различные производные аллели. В каждом анализе измеряли вероятность SNP или замены; на участке дерева, обозначеном красным цветом (геном, для которого мы изучаем такие события, мы назвали целевым геномом), при «условии» мутации, обозначенного синим цветом (если синее событие произошло в другом геноме (B-D), такой геном мы назвали прокси геномом). *Dyak, Dere, Dsim* и *Dmel* обозначают *D. yakuba, D. erecta, D. simulans* и *D. melanogaster; Ppyg, Ggor, Ptro,* и *Hsap* обозначают *Pongo pygmaeus, Gorilla gorilla, Pan troglodytes,* и *Homo sapiens*.

Для каждой пары нуклеотидов *x* и *y* частоту межвидовых замен или SNP с предковым нуклеотидом *x* и производным *y* мы вычисляли по формуле:

$$f_s(x, y) = N_s(x, y)/N_a(x)$$

где *x, y* ϵ {A,C,G,T}, $N_a(x)$ – количество сайтов с предковым нуклеотидом *x*, $N_s(x,y)$ – количество сайтов с межвидовой заменой или SNP $x \rightarrow y$, f_s – частота мутаций,

произошедших по «красному» сценарию (рис. 2), аналогичную величину для мутаций по «синему» сценарию мы обозначили f_{sc} (рис. 2).

Для пары производных аллелей y и z ($z \in$ {A,C,G,T}; $z \neq x$, $z \neq y$) частоту событий, включающих две замены, можно рассчитать по формуле:

$$f_m(x,y,z) = N_m(x,y,z)/N_a(x),$$

где $N_m(x,y,z)$ – число межвидовых замен или SNP $x \rightarrow y$ (красное на рисунке 2) и $x \rightarrow z$ (синее на рисунке 2). Ожидаемое количество рекуррентных мутаций – $e_m(x,y,z)$:

$$e_m(x,y,z) = f_s(x,y) * f_{sc}(x,z) * N_a(x),$$

Усреднённое отношение наблюдаемого числа рекуррентных мутаций к ожидаемому числу мутаций ($\overline{r_m}$) рассчитывается как:

$$\overline{r_m} = (\sum_x \sum_y \sum_z \frac{f_m(x,y,z)}{e_m(x,y,z)})/24$$

В каждом из анализов мы вычисляли отношение транзиций и трансверсий для пар мутаций с одинаковым предковым нуклеотидом:

$$\kappa_{tv}(x,w{:}\, x \approx w) = N_s(x,y{:}\, x \sim y)/N_s(x,z{:}\, x \approx z),$$

где $w \in$ {A,C,G,T}; $w \neq x$, $w \neq y$, $w \neq z$; символом $\approx$ мы обозначили пару замен, разделённых трансверсией, а символом $\sim$ мы обозначили пару замен, разделённых транзицией. Так как при фиксированном предковом нуклеотиде возможно лишь три типа замен, то для удобства отношение транзиций к трансверсиям для пары мутаций мы обозначали третьей мутацией из того же предкового нуклеотида – $\kappa_{tv}(x,w{:}\, x \approx w)$ (см. формулу выше).

Отношение транзиций к трансверсиям для «красных» событий при условии «синей» замены (рис. 2) мы вычисляли по формуле:

$$\kappa_{tvc}(x, w{:}\, x \approx w) = N_m(x, y, w{:}\, x{\sim}y)/N_m(x, z, w{:}\, x \approx z)$$

Изменение отношения транзиций к трансверсиям при условии дополнительной замены вычисляется как:

$$r_\kappa(x, y{:}\, x \approx y) = \kappa_{tvc}(x, y{:}\, x \approx y)/\kappa_{tv}(x, y{:}\, x \approx y)$$

Наконец, среднее значение r_κ рассчитывалось как среднее по всем 8 возможным парам транзиций и трансверсий:

$$\overline{r_\kappa} = \sum_{x,y:x\approx y} (r_\kappa(x, y))\,/8$$

3. Анализ скорости мутирования в сайтах, соседствующих с мутацией, и оценка частоты множественных замен

Частоту мутаций в сайтах, удаленных на расстояние k («красные» события на рис. 2) от случившейся мутации («синие» события на рис. 2), вычисляли как общее количество наблюдаемых событий, нормированное на число рассматриваемых сайтов.

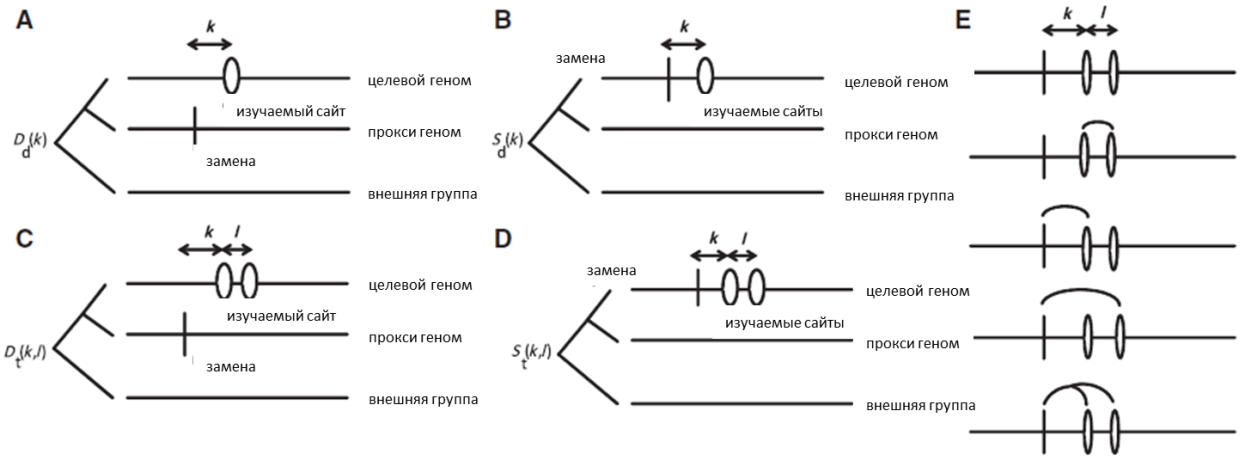

Рис. 3 Вычисление частот динуклеотидных (A,B) и тринуклеотидных (C-E) мутаций. Схематическое дерево слева изображает целевой геном и прокси геном, а также внешнюю группу; горизонтальные линии изображают множественное выравнивание. (A,B) Частоты замен в целевом геноме в сайтах (овалы), таких, что другая замена (вертикальная линия) произошла на расстоянии k от него в прокси (A) или целевом (B) геноме ($d_d(k)$ и $s_d(k)$ соответственно). (C, D) частоты пар замен в целевом геноме измерялись в парах сайтов (пары овалов на расстоянии l друг от друга), таких, что другая замена (вертикальная линия) произошла на расстоянии k от них в прокси (A) или целевом (B) геноме ($d_t(k)$ и $s_t(k)$ соответственно). E, пять сценариев, приводящих к трем соседствующим заменам в целевом геноме, сверху вниз: три независимые мутации; две мутации, одна из которых двойная (скобкой соединены сайты, вовлечённые в МНМ) и одна одиночная; тройная мутация.

Для того чтобы оценить долю α однонуклеотидных замен, происходящих в составе МНМ, мы сравнивали частоты замен в двух сестринских видах, используя

геномы *H. sapiens* и *D. melanogaster* в качестве «целевых», а сестринский геном в качестве прокси генома. Во всех анализах для расчётов скоростей мутирования мы использовали целевой геном из-за его высокого качества.

Доля однонуклеотидных замен, происходящих как динуклеотидные мутации (ДНМ), рассчитывалась следующим образом: для расстояний k ϵ (1..100) мы сравнивали две величины: 1) $d_d(k)$, частоту замен в сайтах целевого генома, удалённых на k нуклеотидов от замены в прокси геноме:

$$d_d(k) = \frac{D_d(k)}{P}$$

и 2) $s_d(k)$, частоту замен в сайтах целевого генома, удаленных на k нуклеотидов от другой замены в целевом геноме:

$$s_d(k) = \frac{S_d(k)}{F},$$

где $D_d(k)$ - число пар нуклеотидных сайтов с координатами (i, i+ k), таких, что в первом из них произошла замена в прокси геноме, а во втором – в целевом геноме; $S_d(k)$ - число пар нуклеотидных сайтов с координатами (i, i+ k), таких, что в них произошли замены в целевом геноме (рис. 3 A и B); P и F – общее число замен в прокси и целевом геноме соответственно.

Поскольку пара замен на разных филогенетических линиях может произойти только как пара независимых событий, то

$$D_d(k) = x_{pf}(k) * P,$$

где $x_{pf}(k)$ – вероятность одиночной нуклеотидной замены в целевом геноме на расстоянии k от замены в прокси геноме.

Напротив, две замены на одной филогенетической линии могут произойти не только как две независимых однонуклеотидных мутации, но и как одна ДНМ. Таким образом,

$$S_d(k) = x_{ff}(k) * F + \alpha_d(k) * F,$$

где $x_{ff}(k)$ - вероятность одиночной нуклеотидной замены в целевом геноме на расстоянии k от другой замены в целевом геноме, и $\alpha_d(k)$ – вероятность того, что однонуклеотидная замена является частью двойной мутации, вовлекающей замену на расстоянии k. Тогда

$$s_d(k) = x_{ff}(k) + \alpha_d(k)$$

Если свойства мутирования одинаковы между сестринскими геномами, то

$$x_{ff}(k)=x_{pf}(k)$$

$$\alpha_d(k) = s_d(k) - d_d(k)$$

Доля однонуклеотидных замен, происходящих как тринуклеотидные мутации (ТНМ), рассчитывалась следующим образом. Мы вычисляли частоту пар мутаций на расстоянии l друг от друга в целевом геноме, при условии третьей мутации на расстоянии k от этой пары в прокси геноме $d_t(k,l)$ или целевом геноме $s_t(k,l)$ (рис 3):

$$d_t(k,l) = \frac{D_t(k,l)}{P}$$

$$s_t(k,l) = \frac{S_t(k,l)}{F}$$

где $D_t(k,l)$ – число троек нуклеотидных сайтов с координатами (i, i+ k, i+ k+l), таких, что в первом из них произошла замена в прокси геноме, а во втором и третьем сайтах произошла замена в целевом геноме; $S_d(k)$ – число троек нуклеотидных сайтов с координатами (i, i+ k, i+ k+l), таких, что в них произошли замены в целевом геноме (рис. 3 C и D). Первый паттерн может быть следствием трёх независимых мутаций или одной одиночной мутации и одной двойной мутации:

$$D_t(k,l) = x_{pf}(k)x_{pf}(k+l)P + x_{pf}(k)\alpha_d(l)P$$

Второй паттерн может происходить по пяти различным сценариям (рис. 3E):

$$S_t(k,l) = x_{ff}(k)x_{ff}(k+l)F + x_{ff}(k)\alpha_d(l)F + x_{ff}(l+k)\alpha_d(k)F + x_{ff}(k)\alpha_d(k+l)F + \alpha_t(k,l)F$$

где $\alpha_t(k,l)$ – вероятность, что однонуклеотидная замена произошла как часть ТНМ, вовлекающей замены на расстоянии k и k+l от неё. Таким образом,

$$\alpha_t(k,l) = s_t(k,l) - d_t(k,l) - x_{ff}(l+k)\alpha_d(k) - x_{ff}(k)\alpha_d(k+l)$$

Мы не только можем оценить частоту ДНМ, но и рассчитать локальную скорость мутирования вокруг неё. Имея данные по локальной скорости мутирования вокруг тандемных мутаций, произошедших на разных ветках (для удобства $D_d(k)$ для k=1 мы также называем ТМРВ (тандемные мутации, разные ветки)) и вокруг тандемных мутаций, произошедших на одной ветке (для удобства $S_d(k)$ для k=1 мы также называем ТМОВ (тандемные мутации, одна ветка)), мы можем рассчитать локальную скорость мутирования вокруг ДНМ. Для этого из локальной скорости мутирования вокруг ТМОВ нужно вычесть вклад локальной скорости мутирования, объясняемый подмножеством тандемных мутаций, произошедших как 2 независимые мутации, а затем поделить на долю ДНМ среди ТМОВ.

$$\mu_n(\text{ДНМ}) = \frac{\mu_n(\text{ТМОВ}) - (1 - \alpha_d(1)/s_d(1)) * \mu_n(\text{ТМРВ})}{\frac{\alpha_d(1)}{s_d(1)}},$$

где $\mu_n(\text{ТМОВ})$ – средняя скорость мутирования на расстоянии n от ТМОВ, $\mu_n(\text{ТМРВ})$ – средняя скорость мутирования на расстоянии n от ТМРВ, $\mu_n(\text{ДНМ})$ – средняя скорость мутирования на расстоянии n от ДНМ.

В качестве прокси для локальной скорости мутирования вокруг независимых тандемных мутаций, произошедших на одной ветке, мы использовали локальную скорость мутирования вокруг ТМРВ.

4. Геномные свойства

Время репликации было взято из статьи [42]. Сайты гиперчувствительности к ДНКазе I (DHS) и данные о гистоновой модификации H3K9me3 из клеточной линии Gm12878 были загружены с сайта проекта ENCODE (https://www.encodeproject.org/). Данные по рекомбинации были взяты из статьи [75]. Все геномные треки были наложены на сборку hg19, которая и использовалась в анализах. Время репликации, DHS сайты, H3K9me3, скорость рекомбинации и ГЦ состав были усреднены по 50Кб непересекающимся окнам.

Для регрессионных анализов, для каждого 50Кб окна, мы строили вектор, содержащий средние величины изучаемых геномных свойств (времени репликации, DHS сайты, H3K9me3, скорость рекомбинации, ГЦ, скорость однонуклеотидных и динуклеотидных мутаций). Эти вектора подавались на вход пуассоновской регрессии. Коэффициенты регрессии и их значимость были посчитаны с использованием функции **glm()** в программе **R**.

5. Спектр аллельных частот

Для анализа спектра аллельных частот использовались сайты, в которых 30 из 37 (для *D. melanogaster*) и 16 из 18 (для *H. sapience*) генотипов содержали символы A, C, G или T. Частоты минорного аллеля вычислялись среди 30 (16) генотипов; если генотипов с рассматриваемыми символами было больше, мы случайным образом выбирали среди них 30 (16). Сайты, содержащие более двух нуклеотидных вариантов, были исключены из этих анализов.

Глава 1. Гетерогенность локальной скорости мутирования

Мы впервые исследовали локальною изменчивость скорости мутирования в роде *Drosophila*, а для семейства Hominidae расширили результаты предыдущих работ [31,32,76] .

Для этой цели мы использовали 4 типа анализов, рассматривая сайты, содержащие SNP в одном виде (анализ I, рис. 2A и A`), SNP в разных видах (анализ II, рис. 2B и B`), SNP в одном виде и однонуклеотидную замену в другом виде (анализ III, рис. 2C и C`), и замены в двух различных видах (анализ VI, рис. 2D и D`). В основе этих анализов лежит общая идея. Каждый раз мы сравнивали две величины: среднюю плотность SNP или замен по геному (отмечено красным на рис. 2); и плотность SNP или замен в сайтах, содержащих или находящихся поблизости от другого SNP или замены (синим на рис. 2). Первая величина отображает среднюю скорость мутирования; вторая отображает скорость мутирования в регионах, которые могут быть особенно подвержены мутациям. Это сравнение позволяет оценить зависимость между наблюдаемыми мутациями и оценить дисперсию скорости мутирования. У этого подхода есть два ограничения. Во-первых, мы не можем изучать возникновение одной и той же мутации дважды (например, A→T и A→T) в одном сайте (здесь и в дальнейшем гомологичные сайты в разных видах мы будем называть одним сайтом). В анализе I такая повторная мутация приведет к возникновению пары аллелей с одинаковым производным нуклеотидом (например, предковый аллель A и дважды возникший

производный аллель T), что неотличимо от ситуации, когда аллель T образовался вследствие одной мутации.

В анализе II одинаковый биаллелный полиморфизм *в D. melanogaster* и *D. simulans* может быть следствием того, что он остался в обоих видах, появившись в их общем предке [67]. Даже в человеке и шимпанзе, в которых общий полиморфизм редок [31], мутации с одинаковыми производными вариантами попадают в один сайт гораздо чаще [31,32] различных мутаций, что говорит о специфичности механизма, ответственного за появления параллельных полиморфизмов. В анализах III и IV аллели с одинаковыми производными вариантами встречаются много чаще [15,32], что снова свидетельствует об особенностях такого мутирования. Поэтому в нашей работе мы рассматриваем мутации, приводящие к различным производным состояниям.

Второе ограничение возникает при изучении влияния наблюдённого SNP на вероятность SNP в близлежащих сайтах, так как такие замены могут быть следствием мутирования нескольких нуклеотидов за одно событие [46,48,76]. Чтобы избежать примеси таких сложных мутаций мы рассматривали только те SNP, производные аллели которых наблюдают в нескольких разных генотипах.

В анализе I для одного сайта рассматриваются ситуации, при которых в одном сайте одновременно находится предковый аллель и два производных аллеля (триаллельный SNP). Частота триаллельных SNP в *D. melanogaster* в 3.5 раза превышает ожидание, основанное на частотах соответствующих биаллельных SNP (Таблица 2). В анализе I для близлежащих сайтов в *D.*

melanogaster мы наблюдали повышение скорости мутирования для 15 сайтов, соседствующих с SNP (рис. 4А).

Таблица 2. Число и частота (в скобках) одной и двух мутаций в сайте для четырёх типов анализов для *Drosophila*. Ожидание во всех случаях достоверно отличается от наблюдения ($P<10^{-100}$, χ^2).

	Тип участка	**Одна мутация в сайте**	**Две мутации в сайте, наблюдение**	**Две мутации в сайте, ожидание**	**Наблюдение /ожидание**
Триаллельный SNP (анализ I)	Интроны	392688 ($6.91*10^{-3}$)	7315 ($1.24*10^{-4}$)	2525 ($4.12*10^{-5}$)	3.48
	Межгенные интервалы	523,130 ($5.72*10^{-3}$)	9725 ($1.00*10^{-4}$)	3268 ($3.23*10^{-5}$)	3.59
Совпадающий SNP (анализ II)	Интроны	392688 ($6.91*10^{-3}$)	2743 ($4.59*10^{-5}$)	1,223 ($2.07*10^{-5}$)	2.28
	Межгенные интервалы	523130 ($5.27*10^{-3}$)	3486 ($3.59*10^{-5}$)	1,502 ($1.65*10^{-6}$)	2.35
SNP в сайтах с заменой (анализ III)	Интроны	392688 ($6.91*10^{-3}$)	39398 ($3.14*10^{-4}$)	20,291 ($1.76*10^{-4}$)	1.95
	Межгенные интервалы	523130 ($5.72*10^{-3}$)	53533 ($3.34*10^{-4}$)	25,953 ($1.63*10^{-4}$)	2.07
Замена в сайтах с другой заменой (анализ IV)	Интроны	766220 ($9.26*10^{-3}$)	80597 ($5.02*10^{-4}$)	37,319 ($2.32*10^{-4}$)	2.23
	Межгенные интервалы	914971 ($9.73*10^{-3}$)	95564 ($5.02*10^{-4}$)	47,947 ($2.51*10^{-4}$)	2.06

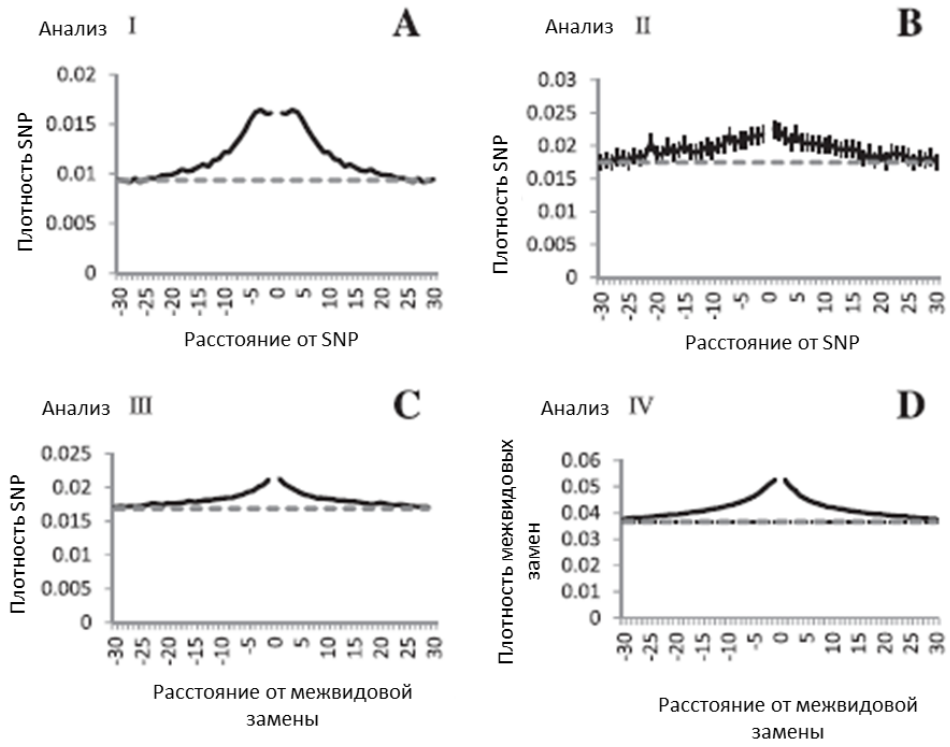

Рис. 4. Плотность SNP или нуклеотидных замен как функция расстояния от другого SNP или нуклеотидной замены для межгенных интервалов в *Drosophila*; анализы соответствуют анализам на рисунке 2. Положительные расстояния соответствуют 3` позициям рассматриваемых нуклеотидов, а отрицательные соответствуют 5` позициям рассматриваемых нуклеотидов. Черная линия обозначает плотность SNP (A-C) или межвидовых замен (D) поблизости от «условнуй» мутации; серая пунктирная линия показывает соответствующее значение, посчитанное по 100 нуклеотидному окну, центрированному на изучаемом сайте. Ошибки посчитаны как 95% биномиальные доверительные интервалы; мы не рисовали доверительные интервалы, когда они были очень малы и плохо различимы.

В анализе II для одного сайта рассматриваются ситуации, при которых в одном и том же сайте одновременно находятся биаллельные полиморфизмы в двух разных видах. SNP, наблюдаемый в *D. simulans*, повышает вероятность совпадающего полиморфизма (биаллельные SNP с различными производными состояниями и одинаковым предковым вариантом, случившиеся в одном сайте у двух различных видов) в *D. melanogaster* в 2.3 раза (Таблица 2). Для близлежащих сайтов анализ II рассчитывает изменение плотности полиморфизма в *D. melanogaster* поблизости от SNP в *D. simulans*. Мы наблюдаем слабый, но статистически значимый эффект в сайтах на расстоянии до15 нуклеотидов (рис. 4B).

В анализе III для одного сайта рассматриваются ситуации, при которых в сайте, в котором произошла замена в одном виде, наблюдается биаллельный SNP в другом. Замена между видами *D. yakuba* и *D. erecta* повышает вероятность SNP в том же сайте в *D. melanogaster* в 2.1 раза (Таблица 2). Для близлежащих сайтов анализ III рассчитывает изменение плотности полиморфизма в *D. melanogaster*, поблизости от замены между *D. yakuba* и *D. erecta*. Плотность SNP повышена в 10 нуклеотидах поблизости от замены (рис. 4C).

В анализе IV для одного сайта рассматриваются ситуации, при которых в сайте, в котором произошла замена в одном виде, наблюдается замена в другом виде. Замена между видами *D. yakuba* и *D. erecta* повышает вероятность замены в том же сайте на ветке *D. melanogaster* после отделения от *D. simulans* в 2.1 раза (Таблица 2). Это наблюдение сходится с предыдущими результатами, полученными по синонимичным сайтам [77]. Для близлежащих сайтов анализ IV рассчитывает изменение плотности замен на ветке *D. melanogaster* поблизости от замены между *D. yakuba* и *D. erecta*. Мы видим, что наблюдаемая замена между *D. yakuba* и *D. erecta* повышает плотность замен в 20 близлежащих сайтах у *D. melanogaster* (рис. 4D).

Те же 4 типа анализов можно проделать и для Homindae, несмотря на другую топологию дерева (рис. 2). Однако данных по полногеномному полиморфизму шимпанзе не было, и мы не могли проделать анализ II. Предыдущие статьи, изучающие локальную скорость мутирования, соответствуют нашим анализам I [76] и IV [32].

В анализе I для одного сайта частота триаллельных SNP в человеке в 2.3 раза превышает ожидание (Таблица 3), что подтверждает предыдущее результаты [76]. В анализе I для близлежащих сайтов плотность полиморфизмов повышена 1.6 раза, но только на расстояни 1 (рис. 5A).

Таблица 3. Число и частота (в скобках) для случаев одной и двух мутаций в сайте для трех из четырех типов анализов, для которых были данные для Hominidae. Ожидание во всех случаях достоверно отличается от наблюдения ($P<10^{-100}$, χ^2).

	Тип участка	**Одна мутация в сайте**	**Две мутации в сайте, наблюдение**	**Две мутации в сайте, ожидание**	**Наблюдение /ожидание**
Триаллельный SNP (анализ I)	Интроны	981033 ($5.82*10^{-4}$)	819 ($4.92*10^{-7}$)	442 ($2.73*10^{-7}$)	2.55
	Межгенные интервалы	1321296 ($6.93*10^{-4}$)	1113 ($5.88*10^{-7}$)	703 ($3.97*10^{-7}$)	2.21
SNP в сайтах с заменой (анализ III)	Интроны	575893 ($3.84*10^{-4}$)	2403 ($8.13*10^{-7}$)	1442 ($5.03*10^{-7}$)	1.83
	Межгенные интервалы	720188 ($4.28*10^{-4}$)	3139 ($9.66*10^{-7}$)	1,854 ($5.86*10^{-7}$)	1.85
Замена в сайтах с другой заменой (анализ IV)	Интроны	1582163 ($1.52*10^{-3}$)	8159 ($3.98*10^{-6}$)	5,591 ($2.83*10^{-6}$)	1.76
	Межгенные интервалы	1873583 ($1.67*10^{-3}$)	1880 ($4.90*10^{-6}$)	7,239 ($3.41*10^{-6}$)	1.83

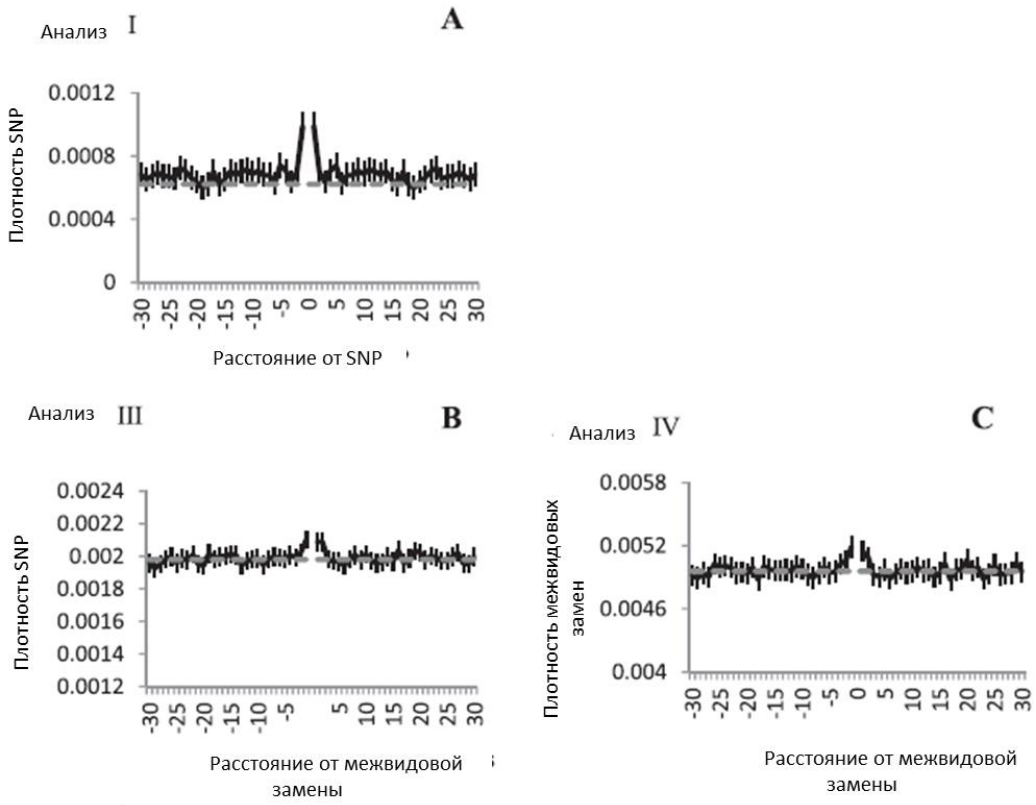

Рис. 5. Плотность SNP или нуклеотидных замен как функция расстояния от другого SNP или нуклеотидной замены для межгенных интервалов в Hominidae; анализы соответствуют анализам на рисунке 2. Положительные расстояния соответствуют 3` позициям, а отрицательные соответствуют 5` позициям. Черная линия обозначает плотность SNP (A, B) или межвидовых замен (C) поблизости от «условной» мутации; серая пунктирная линия показывает соответствующее значение, посчитанное по 100 нуклеотидным окнам, центрированным на изучаемом сайте. Ошибки посчитаны как 95% биномиальные доверительные интервалы.

В анализе III для одного сайта замена в линии гориллы повышает вероятность SNP в том же сайте в человеке в 1.9 раза (Таблица 3). Для близлежащих сайтов анализ III рассчитывает изменение плотности полиморфизма в человеке поблизости от замены на ветке гориллы. Плотность SNP слабо повышена в 2 соседних нуклеотидах (рис. 5B).

В анализе IV для одного сайта замена в линии гориллы повышает вероятность замены в том же сайте на линии человека в 1.8 раза (Таблица 3). Для близлежащих сайтов анализ IV рассчитывает изменение плотности замен в человеке поблизости от замены на ветке гориллы. Плотность замен на филогенетической ветке, ведущей к человеку, слабо повышена в двух соседних нуклеотидах от замены на ветке гориллы (рис. 5C).

Для *S. commune* мы проделали анализ, аналогичный анализу II: мы подготовили данные по SNP в российской и американской популяции грибов и вычислили избыток ситуаций, когда биаллельные SNP с разными минорными аллелями в двух популяциях попадали в один и тот же сайт. Мы наблюдали превышение наблюдения над ожиданием в 1.42 раза для таких полиморфизмов.

Глава 2. Локальная гетерогенность отношения транзиций к трансверсям

Измерение локальной скорости мутирования не учитывает изменения соотношения различных типов мутаций. Мы изучали гетерогенность **к** - отношения транзиций к трансверсиям вдоль по геному. Мы рассматривали зависимости **к** от мутаций, произошедших в тех же сайтах или сайтах в окрестности наблюдаемой мутации. Мы использовали ту же логику, что и в анализах изменчивости общей скорости мутирования: мы сравнивали **к** в сайтах или поблизости от сайтов, в которых произошла «условная» мутация, со средним по геному. Мы проделали те же 4 типа анализов, что и для общей скорости

мутирования (рис. 2) Ввиду нетривиальности анализов в одном сайте для расчета **к**, опишем их подробнее.

Рассмотрим сайт с известным предковым нуклеотидом (например, A). В этом сайте может случиться 2 различных трансверсии (A→C и A→T) и одна транзиция (A→G); как и в анализах гетерогенности скорости мутирования, мы исключаем параллельные мутации. Таким образом, для сайтов, в которых произошла трансверсия (например, A→C), мы можем вычислить вероятность того, что в них произойдет вторая трансверсия (A→T) или транзиция (A→G), и получить отношение транзиций к трансверсиям (A→G/A→T) при условии трансверсии $\boldsymbol{\kappa_{tvc}}$. Полученное значение можно сравнить с величиной $\boldsymbol{\kappa_{tv}}$ для соответствующих мутаций, но в сайтах, в которых произошла лишь одна мутация. Отличие $\boldsymbol{\kappa_{tvc}}$ от $\boldsymbol{\kappa_{tv}}$ будет свидетельствовать о том, что отношение тразиций к трансверсиям меняется в сайтах, в которых произошла независимая трансверсия. Поскольку в сайте с известным предковым нуклеотидом может произойти лишь одна транзиция, мы не можем изучать отношение тразиций к трансверсиям при условии случившейся транзиции (Таблицы 4 и 5); это ограничение пропадает при изучении сайтов, соседствующих с рассматриваемой мутацией.

Данные по $\boldsymbol{\kappa_{tvc}}$ и $\boldsymbol{\kappa_{tv}}$ для анализа I в *D. melanogaster* по всем 8 возможным парам транзиций и трансверсий представлены в таблице 4. Отношение трансверсий к транзициям ниже в сайтах, в которых наблюдается другая трансверсия, и $\overline{r_\kappa}$ – среднее по 8 возможным парам отношений $\boldsymbol{\kappa_{tvc}}$ и $\boldsymbol{\kappa_{tv}}$ получилось равным 0.76 (Таблица 4). Для 30 близлежащих сайтов анализ I показывает снижение **к** поблизости от трансверсии, но не транзиции (рис. 6A).

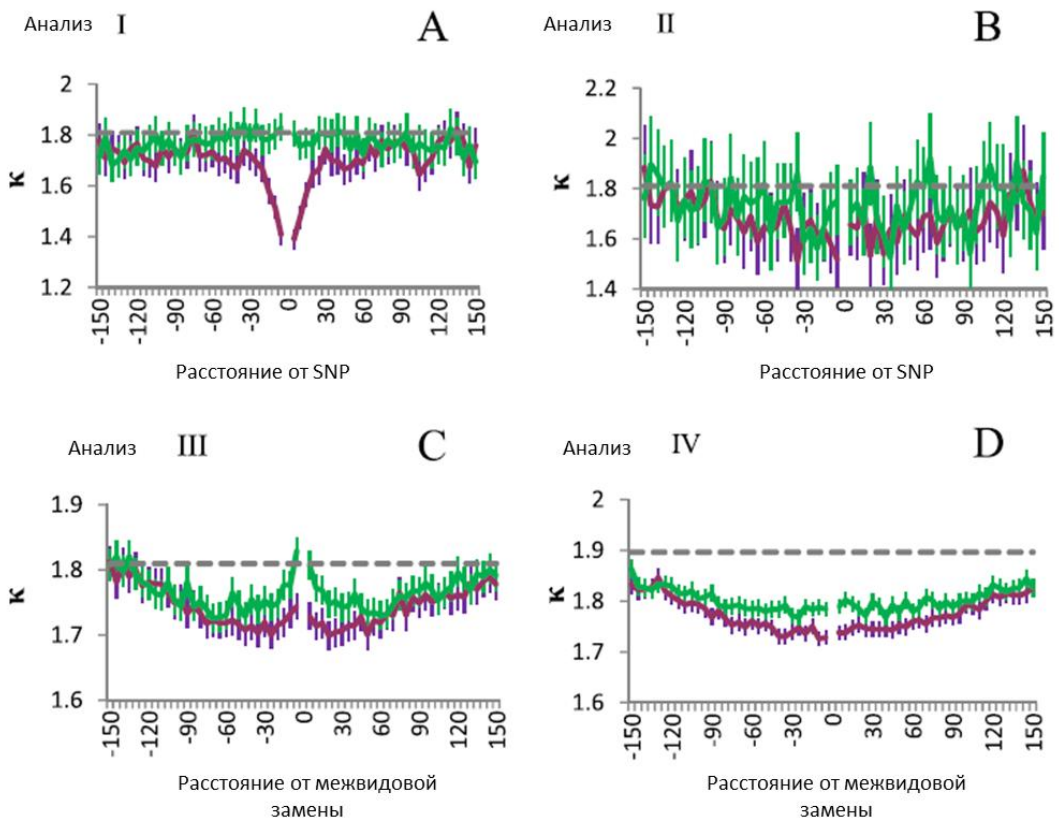

Рис. 6. Отношение транзиций к трансверсиям (**к**) как функция расстояния до тразиции (зеленый) или трансверсии (фиолетовый) для межгенных интервалов в *Drosophila*; анализы соответствуют анализам на рисунке 2. Положительные расстояния соответствуют 3` позициям, а отрицательные соответствуют 5` позициям. Серая пунктирная линия соответствует среднегеномному значению **к**. Значения **к** даны по 5 нуклеотидным окнам. Ошибки посчитаны как 95% доверительные интервалы по 1000 испытаний.

В анализе II для одного сайта «условная» трансверсия в *D. simulans* снижает **к** в полиморфизме *D. melanogaster*, и $\overline{r_\kappa}$ =0.91. При перекрестном анализе, в котором мы рассматривали влияние «условной» трансверсии в *D. melanogaster* на значение **к** в *D. simulans*, мы получили сходные результаты.

Таблица 4. Отношение частот транзиций к трансверсиям в биаллельных и триалленых SNP (анализ I) в *Drosophila*. *,**,*** обозначают P<0.05, P<0.01 и P<0.001 соответственно для отличия $\boldsymbol{\kappa_{tvc}}$ и $\boldsymbol{\kappa_{tv}}$.

Транзиция и трансверсия, использованные для расчёта κ_{tv}	«Условная» трансверсия, использованная для расчёта κ_{tvc}	Интроны			Межгенные интервалы		
		$\boldsymbol{\kappa_{tv}}$	$\boldsymbol{\kappa_{tvc}}$	$\boldsymbol{\kappa_{tvc}/\kappa_{tv}}$	$\boldsymbol{\kappa_{tv}}$	$\boldsymbol{\kappa_{tvc}}$	$\boldsymbol{\kappa_{tvc}/\kappa_{tv}}$
A→G/A→T	A→C	1.02	0.78	0.76***	1.03	0.82	0.80**
A→G/A→C	A→T	1.88	1.76	0.94	1.88	1.70	0.90*
G→A/G→T	G→C	1.84	1.45	0.79***	1.83	1.26	0.69***
G→A/G→C	G→C	3.52	2.16	0.61***	3.30	2.02	0.61***
C→T/C→G	C→A	3.50	2.17	0.62***	3.33	2.14	0.64***
C→T/C→A	C→G	1.85	1.26	0.68***	1.81	1.30	0.72***
T→C/T→G	T→A	1.85	1.50	0.81***	1.85	1.61	0.87***
T→C/T→A	T→G	1.01	0.81	0.81**	1.02	0.84	0.82**
Среднее значение		2.06	1.49	0.75	2.01	1.46	0.76

Для близлежащих сайтов из-за недостаточного количества данных мы усредняли **к** по 100 сайтам вблизи мутации и наблюдали достоверный эффект как по снижению **к** как вблизи транзиций ($P<10^{-8}$), так и вблизи трансверсий ($P<10^{-20}$); но вблизи трансверсий эффект оказался сильнее ($P=2*10^{-4}$) (рис. 6В).

В анализе III для одного сайта трансверсия на ветке *D. erecta* снижает **к** в полиморфизме *D. melanogaster*, и $\overline{r_к}$ =0.95. Для близлежащих нуклеотидов мы усредняли **к** по 100 сайтам вблизи мутации и наблюдали достоверный эффект как по снижению **к** как вблизи транзиций ($P<10^{-6}$), так и вблизи трансверсий ($P<10^{-50}$), но вблизи трансверсий эффект был сильнее ($P<10^{-18}$) (рис. 6D). Наименьшие значения **к** наблюдаются на расстоянии 20-30 нуклеотидов, а не в непосредственном соседстве с заменой, что может отражать особенности мутационного спектра плодовой мушки в 10- нуклеотидной близости от замены.

В анализе VI для одного сайта трансверсия между *D. yakuba* и *D. yakuba* снижает **к** среди замен на ветке *D. melanogaster*, и $\overline{r_к}$ =0.86. Для близлежащих сайтов мы наблюдали достоверный эффект по снижению **к** как вблизи транзиций ($P<10^{-14}$), так и вблизи трансверсий ($P<10^{-48}$); как и в анализах I-IV, вблизи трансверсий эффект сильнее ($P<10^{-6}$) (рис. 6D).

Таким образом, **к** снижена как вблизи транзиций, так и вблизи трансверсий. Этот эффект может объясняться повышенной локальной скоростью мутирования. Чтобы исследовать эту возможность, мы рассмотрели плотность мутаций вблизи от анализируемых сайтов и оценили, влияет ли количество мутаций на наблюдаемую **к** (рис 7). Локальная гипермутабильность действительно частично объясняет понижение **к** вблизи от мутации; так, в окнах, содержащих больше мутаций, мы наблюдали большую долю трансверсий. Тем не менее, даже с фиксированным количеством мутаций в окне **к** понижена поблизости от трансверсий сильнее, чем поблизости от транзиций.

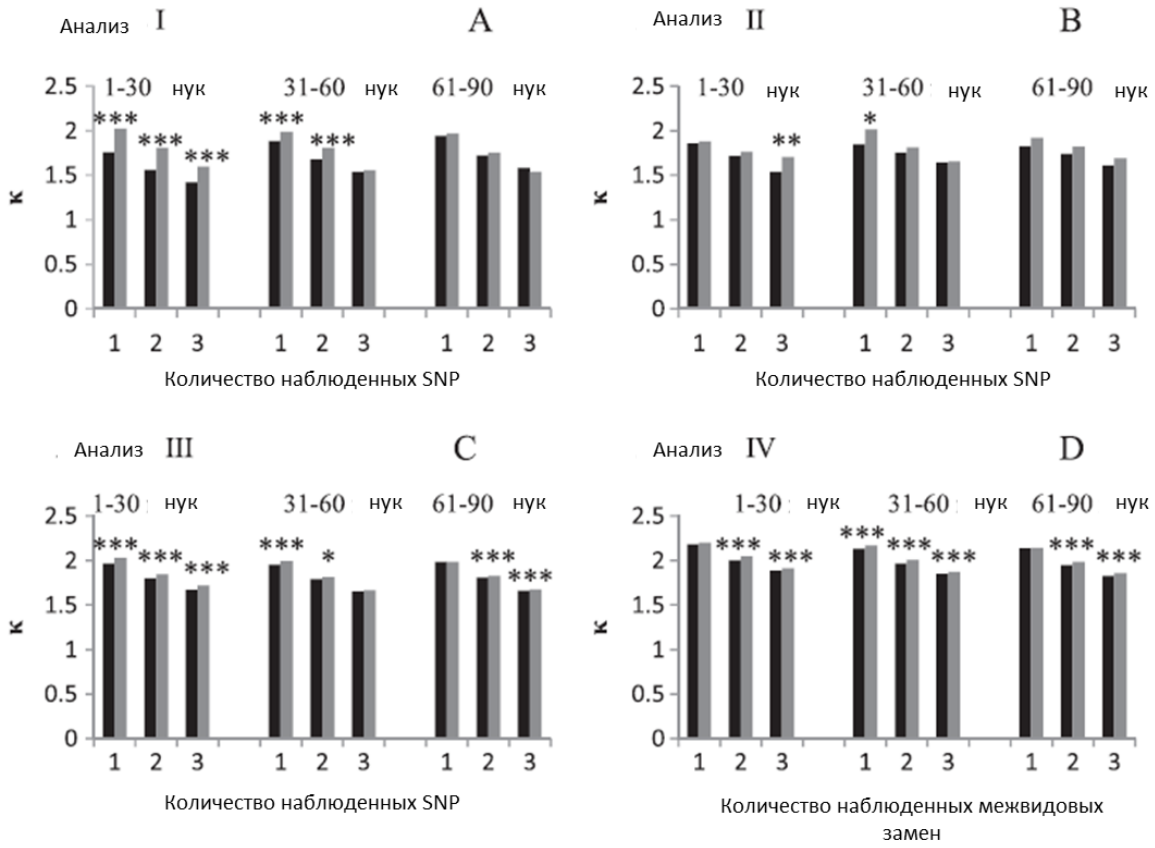

Рис. 7. Отношение транзиций к трансверсиям (**к**) как функция количества мутаций в окнах, лежащих на различных расстояниях от трансверсии (черный) или транзиции (серый) в межгенных интервалах *Drosophila*. Каждое окно длиной 60 нуклеотидов содержит сайты по обе стороны от мутации на расстояниях 1-30, 31-60 и 61-90 относительно центральной мутации. *,**,*** обозначают P<0.05, P<0.01 и P<0.001 соответственно для различий **к** вокруг сайта с трансверсей и транзицией.

Более сильный отрицательный отбор против трансверсий может быть причиной понижения **к** в регионах с высокой частотой SNP или межвидовых замен и более сильной ассоциации трансверсий с трансверсиями, нежели с транзициями, поскольку при таком режиме отбора трансверсии сильнее транзиций будут кластеризоваться в нейтральных участках. Чтобы проверить это, мы построили спектр частот минорного аллеля (СЧМА) отдельно по транзициям и трансверсиям. У *Drosophila* СЧМА трансверсий смещен в пользу редких

вариантов по сравнению с транзициями (рис. 8); значение статистики Tajima's D [78] по трансверсиям (-1.51 для интронов и -1.54 для межгенных интервалов) ниже, чем по транзициям (-0.80 для интронов и -0.84 для межгенных интервалов), что свидетельствует о слабом отрицательном отборе против трансверсий.

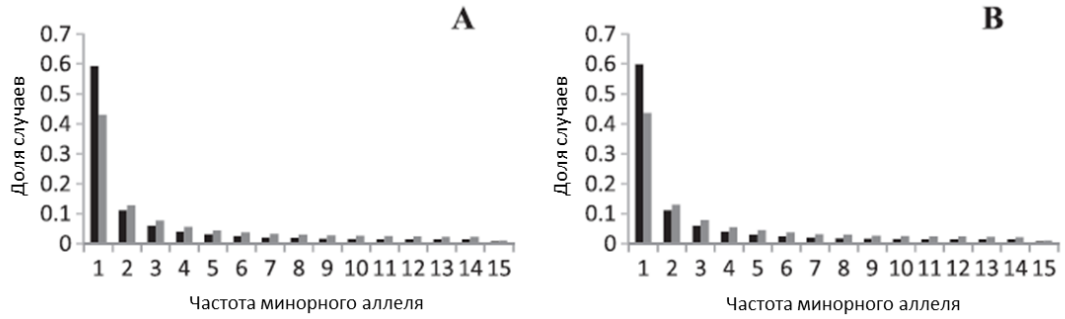

Рис. 8. Спектр частот минорного аллеля для полиморфных трансверсий (черный) и транзиций (серый) в *D.melanogaster* в интронах (A) и межгенных интервалах (B). Горизонтальная ось соответствует количеству генотипов (из 30), несущих минорный аллель.

В анализе I для одного сайта для человека в триаллельных SNP трансверсия более чем втрое снижает отношение транзиций к трансверсиям $\overline{r_\kappa}$ =0.31 (Таблица 4).

В анализе III для одного сайта трансверсия на ветке гориллы снижает **к** в полиморфизме человека, и $\overline{r_\kappa}$ =0.91.

В анализе IV для одного сайта трансверсия на ветке гориллы снижает **к** по заменам на ветке человека, и $\overline{r_\kappa}$ =0.58.

Таблица 5. Отношение частот транзиций и трансверсий в биаллельных и триаллельных SNP (анализ I) в Hominidae. $P < 10^{-3}$ для всех сравнений $\boldsymbol{\kappa_{tvc}}$ и $\boldsymbol{\kappa_{tv}}$.

Транзиция и трансверсия, использованные для расчета κ_{tv}	«Условная» трансверсия, использованная для расчёта κ_{tvc}	Интроны			Межгенные интервалы		
		$\boldsymbol{\kappa_{tv}}$	$\boldsymbol{\kappa_{tvc}}$	$\boldsymbol{\kappa_{tvc}/\kappa_{tv}}$	$\boldsymbol{\kappa_{tv}}$	$\boldsymbol{\kappa_{tvc}}$	$\boldsymbol{\kappa_{tvc}/\kappa_{tv}}$
A→G/A→T	A→C	4.76	1.74	0.37	4.35	1.51	0.35
A→G/A→C	A→T	3.91	1.56	0.40	3.84	1.30	0.34
G→A/G→T	G→C	3.48	1.18	0.34	3.20	1.34	0.42
G→A/G→C	G→C	3.01	0.57	0.19	3.31	1.17	0.35
C→T/C→G	C→A	3.01	0.93	0.31	3.36	1.04	0.31
C→T/C→A	C→G	3.58	1.34	0.37	3.22	1.09	0.34
T→C/T→G	T→A	3.82	1.08	0.28	3.89	1.12	0.29
T→C/T→A	T→G	4.87	1.00	0.21	4.34	1.38	0.32
Среднее значение		3.82	1.18	0.31	3.69	1.24	0.34

В анализе I для соседних сайтов **к** снижена для 5 нуклеотидов вблизи «условной» трансверсии, но не «условной» транзиции; мы не наблюдали эффекта на расстоянии более 5 нуклеотидов (рис. 9A). В анализах III и IV **к** меньше среднегеномной поблизости от трансверсий, но не транзиций (рис 9B, C).

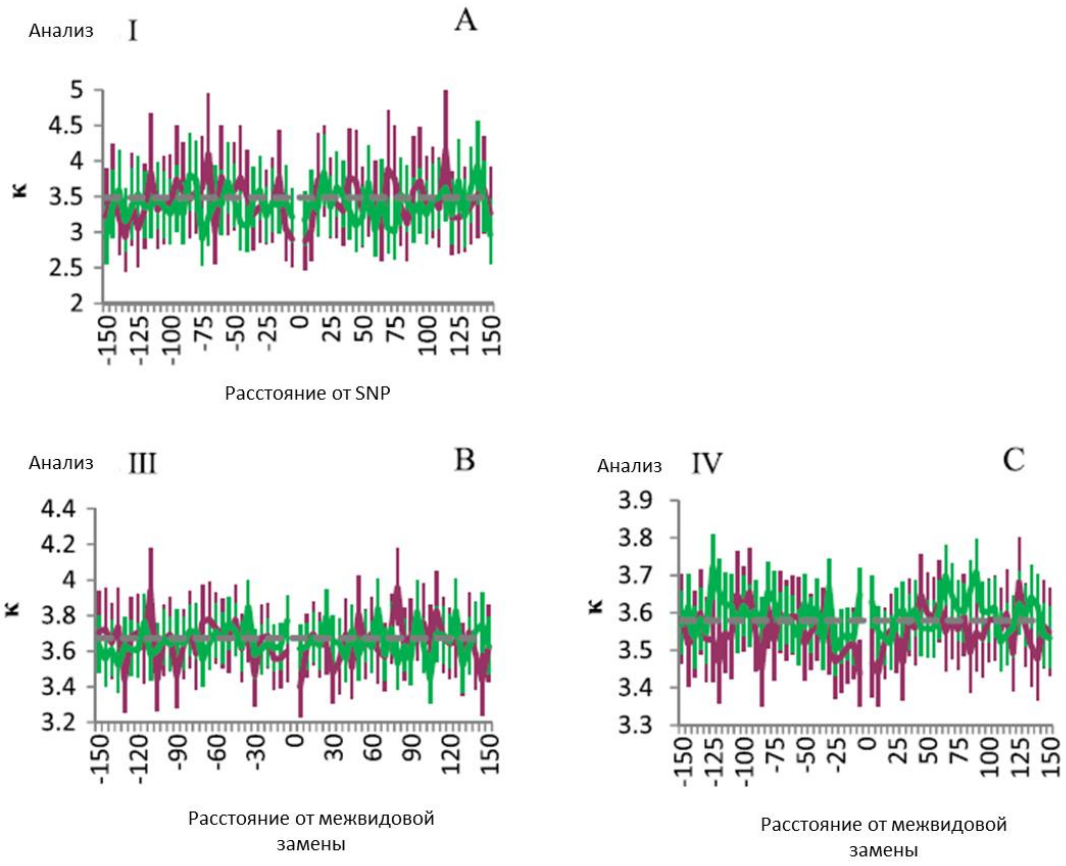

Рис. 9. Отношение транзиций к трансверсиям (**к**) как функция расстояния до транзиции (зеленый) или трансверсии (фиолетовый) для межгенных интервалов у Hominidae; анализы соответствуют анализам на рисунке 2. Положительные расстояния соответствуют 3` позициям, а отрицательные соответствуют 5` позициям. Серая пунктирная линия соответствует среднегеномному значению **к**. Значения **к** даны по 5 нуклеотидным окнам. Ошибки посчитаны как 95% доверительные интервалы по 1000 испытаний.

В людях СЧМА по транзициям и трансверсиям неотличимы (рис. 10); это свидетельствует о том, что отбор не влияет на паттерны изменчивости **к** у людей. Значения Tajima’s D по транзициям и трансверсиям также очень схожи (для

трансверсий -0.46 в интронах, -0.42 в межгенных интервалах; для транзиций -0.43 в интронах, -0.39 в межгенных интервалах).

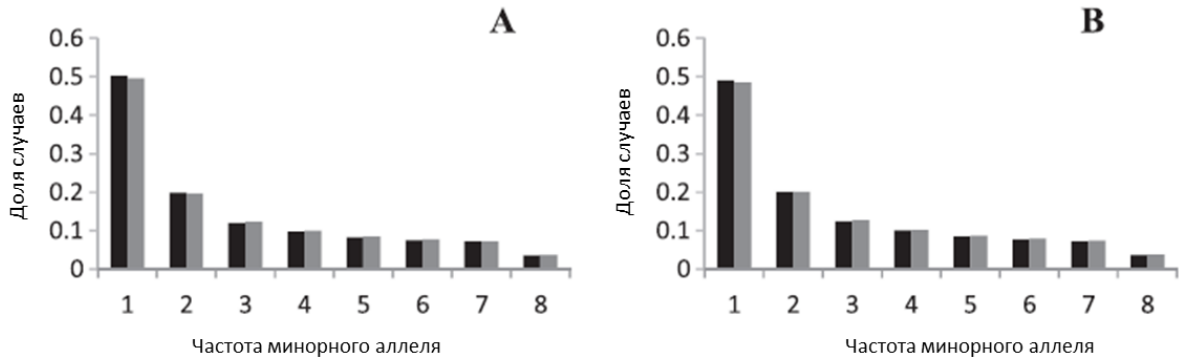

Рис. 10. Спектр частот минорного аллеля для полиморфных трансверсий (черный) и транзиций (серый) в человеке в интронах (A) и межгенных интервалах (B). Горизонтальная ось соответствует количеству генотипов (из 16), несущих минорный аллель.

В *S. commune* в вышеописанном тесте, являющимся аналогом анализа II, мы обнаружили, что избыток совпадающих полиморфизмов выше для случаев, когда оба биаллельных SNP – трансверсии (1.58), по сравнению со случаем, когда один биаллельный SNP транзиция, а другой трансверсия (1.32). Описанные паттерны качественно соответствуют результатм по плодовой мушке и человеку.

Таким образом, мы наблюдаем сходный эффект увеличения доли трансверсий в сайтах множественных мутаций в трех очень далеких группах животных: позвоночных, насекомых и грибах. Этот эффект может быть лишь частично объяснен отбором, поскольку мы наблюдаем самое сильное снижение **к** в некодирующим полиморфизме человека, где действие отбора выражено наиболее слабым образом. Подобное изменение мутационного спектра в пользу трансверсий в сайтах, подверженных мутациям, скорее всего свидетельствует о смене мутационных механизмов в этих сайтах. Например, такое может

наблюдаться если эти позиции особенно часто реплицируются неточными полимеразами, которые часто совершают ошибки, приводящие к трансверсиям [10,11]. Мы наблюдаем очень сильное снижение **к** в полиморфизме человека (более чем в три раза), и подробное изучение этого эффекта поможет продвинуть понимание молекулярных механизмов, лежащих в основе мутагенеза и приводящих к «загадочной» изменчивости скорости мутирования [31], сопутствующей изменению мутационного спектра.

Глава 3. Мультинуклеотидные замены в эволюции приматов и *Drosophila*.

В этой главе описанны результаты совместной работы коллектива авторов и мой вклад в неё – методологичкая разработка проекта и получение результатов по роду *Drosophila*.

Эволюция на всех уровнях, включая и эволюцию последовательностей ДНК, как правило, происходит благодаря большому числу небольших событий. Простые эволюционные модели не включают сложные события, предполагая, что замены в разных сайтах происходят независимо [79]. Мутация, затрагивающая несколько сайтов одновременно, является более дальним шагом для эволюционирующей последовательности и позволяет перейти из одной последовательности в другую, минуя промежуточные состояния.

Мы изучали множественные мутации в геномах с наиболее высоким качеством внутри клад Hominidae и *Drosophila:* человеке и *D. melanogaster* соответственно. Оценить частоту двойных замен (ДНМ) можно как разницу частот замен в геноме человека (*D. melanogaster*) при условии замены на той же ветке, и вычесть частоту замен в геноме человека (*D. melanogaster*) при условии замены на сестринской ветке филогенетического древа. Если две замены произошли на одной ветке, тогда первая замена может произойти как часть

двойной мутации, затрагивающей условную замену, или как независимое событие. Если две замены произошли на разных ветках, тогда первая замена может случиться только как независимое событие (см. рис. 3 и секцию 5 Материалов и Методов).

Сходный подход можно применить для расчёта числа тройных замен (ТНМ) (рис. 3E).

Описанный метод учитывает гетерогенность скорости мутирования вдоль генома, если эта гетерогенность сохраняется между прокси геномом и фокальным геномом. Он позволяет рассчитать скорости ДНМ и ТНМ для сайтов на разных расстояниях, и легко обобщаеться на случай множественных, замен включающих более трёх позиций, или же сложных мутаций, в которые также вовлечены выпадения или вставки сегментов последовательности.

Для приматов мы рассматривали 4 тройки видов. Для всех троек в качестве фокального генома мы использовали *H. sapiens*, а в качестве прокси генома и внешнего вида использовалось множество видов на различном филогенетическом расстоянии от человека. Хотя мы и не наблюдаем локального повышения скорости мутирования в приматах, за исключением сайта ближайшего к мутации в другом виде (рис. 5 и рис. 11, красные линии), частота замен поблизости от мутации в том же виде заметно повышена (рис. 11, синие линии). Таким образом, значительная часть мутаций, затрагивающих несколько сайтов в одной линии, происходят как мультинуклеотидные мутации (МНМ). Наши результаты свидетельствуют о том что ДНМ, затрагивающие сайты на расстоянии до 10 нуклеотидов друг от друга составляют $\sum_{1\leq k\leq 10}\alpha_d(k) = 0.023$ от всех однонуклеотидных замен. $\alpha_d(k)$ быстро убывает с увеличением расстояния между сайтами; так, $\alpha_d(1)$ больше, чем $\alpha_d(10)$, в 6 раз (рис 11, зеленые линии). 36% ДНМ на расстоянии до 10 нуклеотидов происходят в непосредственно соседних сайтах.

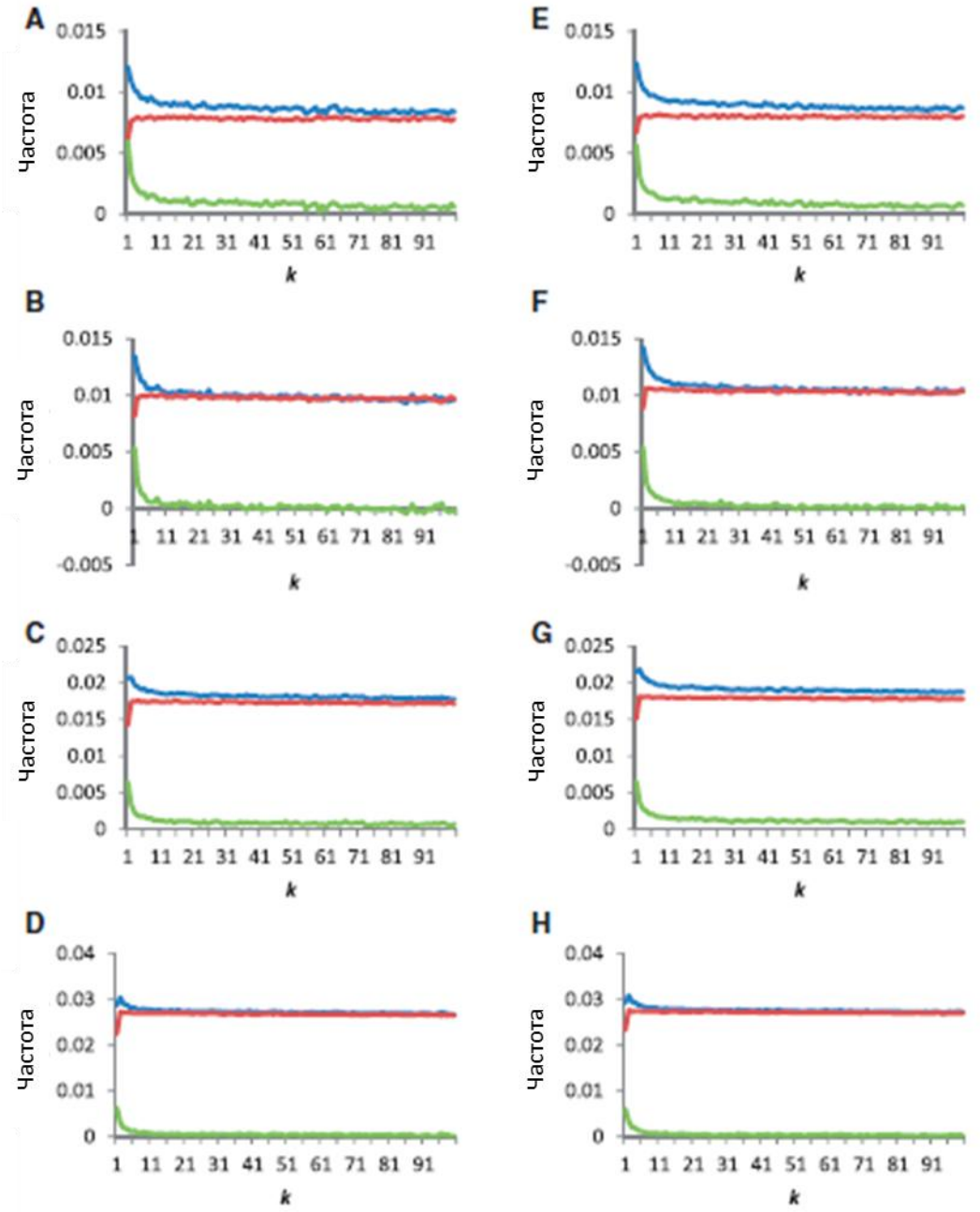

Рис. 11. Частота ДНМ в приматах для сайтов на разных расстояниях. $d_d(k)$ (красным), $s_d(k)$ (синим) и $\alpha_d(k)$ (зеленым) для расстояний между сайтами 1≤k≤100 (горизонтальная ось). Левая колонка (A-D), интроны; правая колонка (E-H), межгенные интервалы. *Homo sapiens* и *Pan troglodytes* (*Gorilla gorilla* в качестве внешнего вида), (B,F) *H. sapiens* и *G. gorilla* (*Pongo pygmaeus* в качестве внешнего вида), (C,G) *H. sapiens* и *P. pygmaeus* (*Macaca mulatta* в качестве внешнего вида), and (D,H) *H. sapiens* и *M. mulatta (Callithrix jacchus* в качестве внешнего вида). Исключение CpG динуклеотидов приводит к симметричной недооценке $d_d(1)$ и $s_d(1)$.

Значительная доля замен происходит как ТНМ. На рисунке 12 представлены результаты для случая, когда две из трёх мутаций, вовлечённых в ТНМ, случились в соседних сайтах (l=1), а на рисунке 13А показаны значения $\alpha_t(k,l)$ для 1≤k≤10 и 1≤l≤10. ТНМ для сайтов, удалённых не более чем на 10 нуклеотидов, составляют $\sum_{1\le k+l\le 10}\alpha_t(k,l) = 0.006$ от нуклеотидных замен, или 26% от ДНМ, для сайтов на тех же расстояниях. Также как и для ДНМ, скорость ТНМ быстро падает с увеличением расстояния между сайтами; так, $\alpha_t(1,1)$ больше, чем $\alpha_t(1,10)$, в 7 раз (рис 12, зеленые линии). 7.8% ТНМ на расстоянии до 10 нуклеотидов происходят в соседних сайтах (против ожидания в 2.7% при равномерном распределении по всем значениям k+l≤10, т.к. существует 36 возможных пар k и l) (рис. 13А).

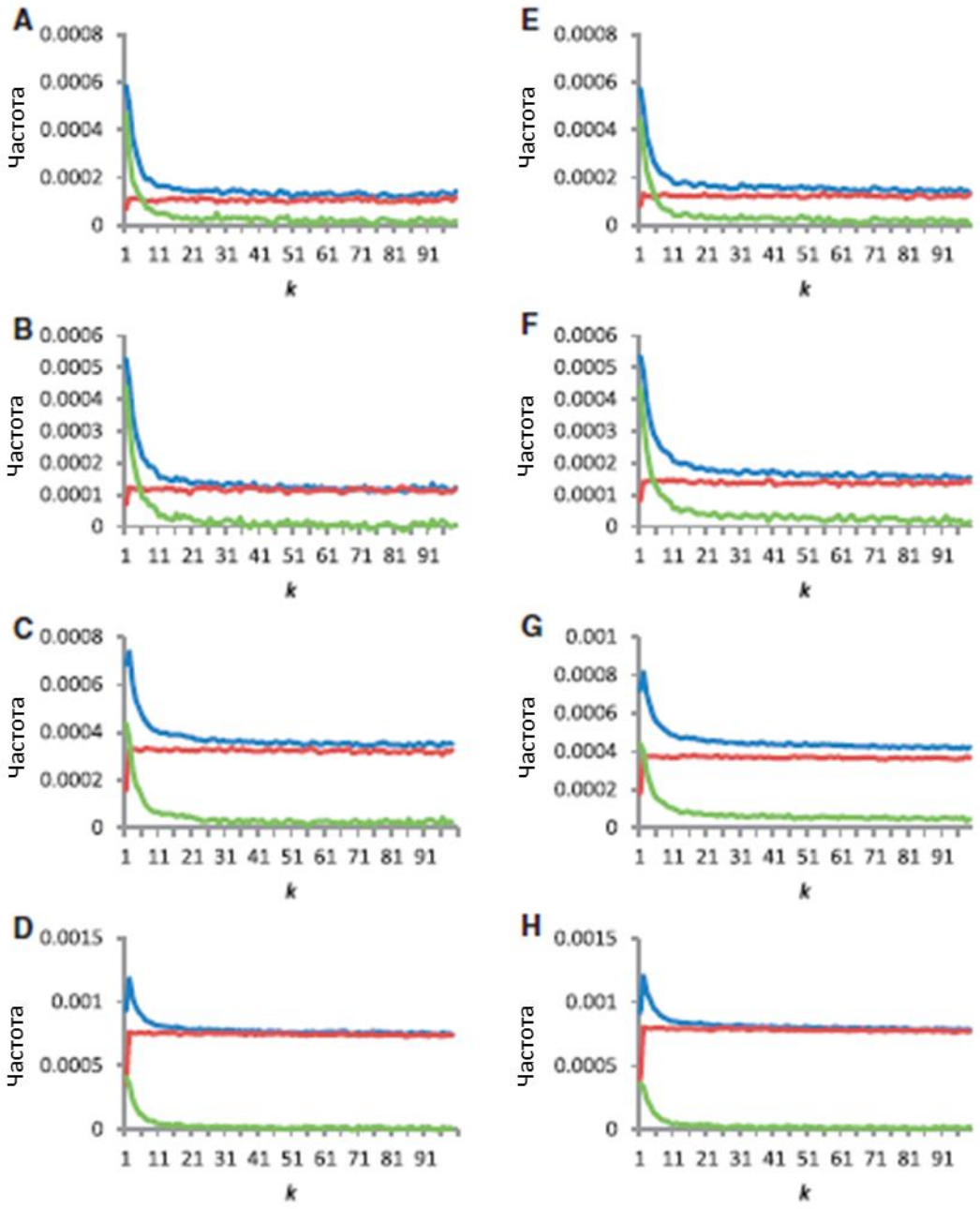

Рис. 12. Частота THM для случая, когда две из трех мутаций, вовлеченных в THM, случились в соседних сайтах (l=1), для различных расстояний до третьего сайта в приматах. $d_t(k,1)$ (красный), $s_t(k,1)$ (синий) и $\alpha_t(k,1)$ (зеленый) 1≤k≤100 (горизонтальная ось). Панели соответствуют панелям на рисунке 11. Исключение CpG динуклеотидов приводит к симметричной недооценке $d_d(1)$ и $s_d(1)$.

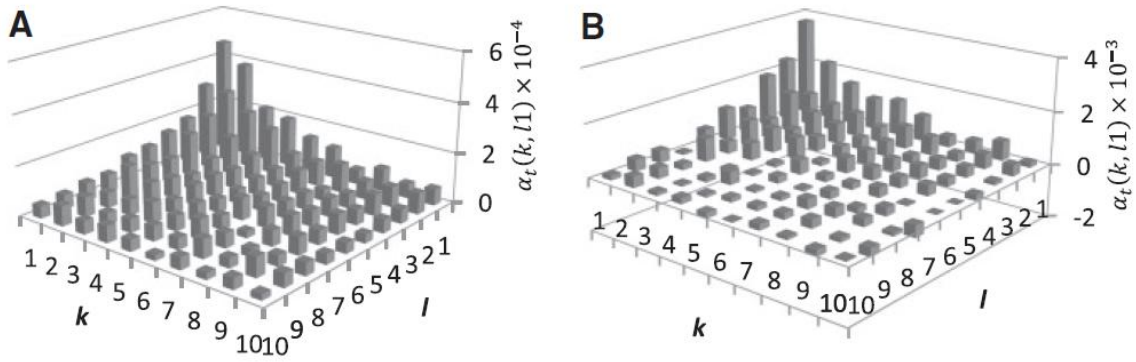

Рис. 13. Частоты ТНМ. $\alpha_t(k,l)$ для 1≤k≤10 и 1≤l≤10. (A) Рассматриваемая пара видов человек-шимпанзе, (B) пара видов *D. melanogaster–D. simulans.* Только интроны использовались в этом тесте.

Для четырёх анализируемых филогений длина ветки, ведущей к фокальному геному (*H. sapiens*), отличается до 3.8 раз. Количество замен в соседних сайтах $D_d(k)$ (которое зависит квадратично от длины ветки) отличается более, чем в 10 раз. Однако $\alpha_d(k)$ - доля ДНМ от частоты одиночных замен (как и доля любых сложных или простых событий в мутационном спектре) с длиной ветки меняться не должна, что мы и видим (рис. 14A). Это демонстрирует устойчивость наших результатов.

Хотя значение $\alpha_d(k)$ и сходно для разных филогений, их доля среди замен, произошедших в близлежащих сайтах в одной линии, падает с длиной ветки ($D_d(k)$ зависит квадратично от длины ветки, а F*$\alpha_d(k)$ - линейно). Действительно, в сравнении человек-шимпанзе 50% пар замен в соседних сайтах в линии человека случаются как ДНМ (рис. 14B), и 83% замен в трех соседних сайтах происходят как ТНМ. Эта доля меньше для более далеких пар видов; так, для пары человек-макака ДНМ и ТНМ составляют только 22% и 44% от замен, затрагивающих сразу два или три сайта в линии человека соответственно. Мы этого и ожидаем, т.к. в близких видах замены в соседних сайтах с небольшой вероятностью будут происходить как серия независимых событий, а скорее случатся как одна МНМ.

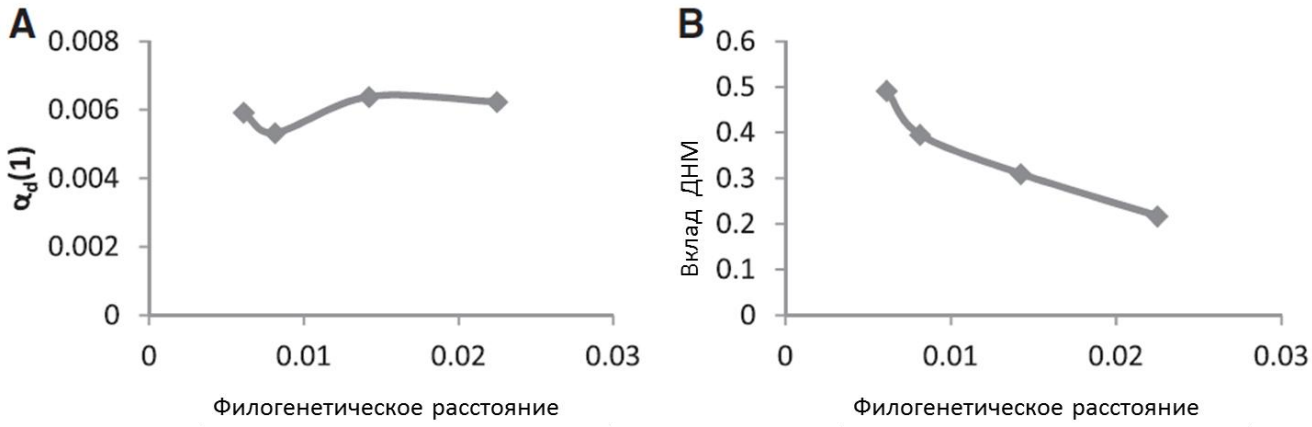

Рис. 14. Зависимость вклада ДНМ от длины филогенетической линии в интронах. (A) $\alpha_d(1)$ как функция $d_d(k)$ и (B) $\alpha_d(1)/s_d(1)$ как функция $d_d(k)$. На обеих графиках четыре точки соответствуют заменам, случившимся на линии *H. sapiens* после отхождения от общего предка с *P. troglodytes*, *G. gorilla*, *P. pygmaeus* и *M. mulatta* соответственно.

В отличие от приматов, у *Drosophila* выражена локальная изменчивость скорости мутирования, и плотность замен повышена на расстоянии более 10 нуклеотидов от замены в другом виде (рис. 4, 15, 16 красные линии).

Тем не менее, замена, произошедшая на линии *D. melanogaster*, повышает частоту других замен в *D. melanogaster* поблизости от данной значительно сильнее, чем замена в прокси геноме (рис. 15, 16), что свидетельствует о вкладе МНМ в замены, происходящие в близлежащих сайтах. ДНМ и ТНМ, происходящие на расстоянии до 10 нуклеотидов, составляют $\sum_{1\le k\le 10}\alpha_d(k) = 0.056$ и $\sum_{1\le k+l\le 10}\alpha_t(k,l) = 0.039$ от всех однонуклеотидных замен. ДНМ и ТНМ, затрагивающие два и три непосредственно соседних сайта, составляют $\alpha_d(1) = 0.02$ и $\alpha_t(1,1) = 0.004$ от всех однонуклеотидных замен. Таким образом, вклад ДНМ и ТНМ в 2 и 6 раз больше в *Drosophila*, чем в приматах. 34% пар замен в соседних сайтах в линии *D. melanogaster* после отхождения от общего предка с *D. simulans* случаются как ДНМ, и 43% замен в трёх соседних сайтах происходят как ТНМ. Как и в приматах, $\alpha_d(k)$ и $\alpha_t(k,1)$ быстро падают с увеличением k (рис. 15,

16, зелёные кривые). Результаты остаются такими же, если предковое состояние восстановить не методом наибольшей экономии, а методом наибольшего правдоподобия, и исключить сайты, в которых наблюдается более 2 различных нуклеотидов.

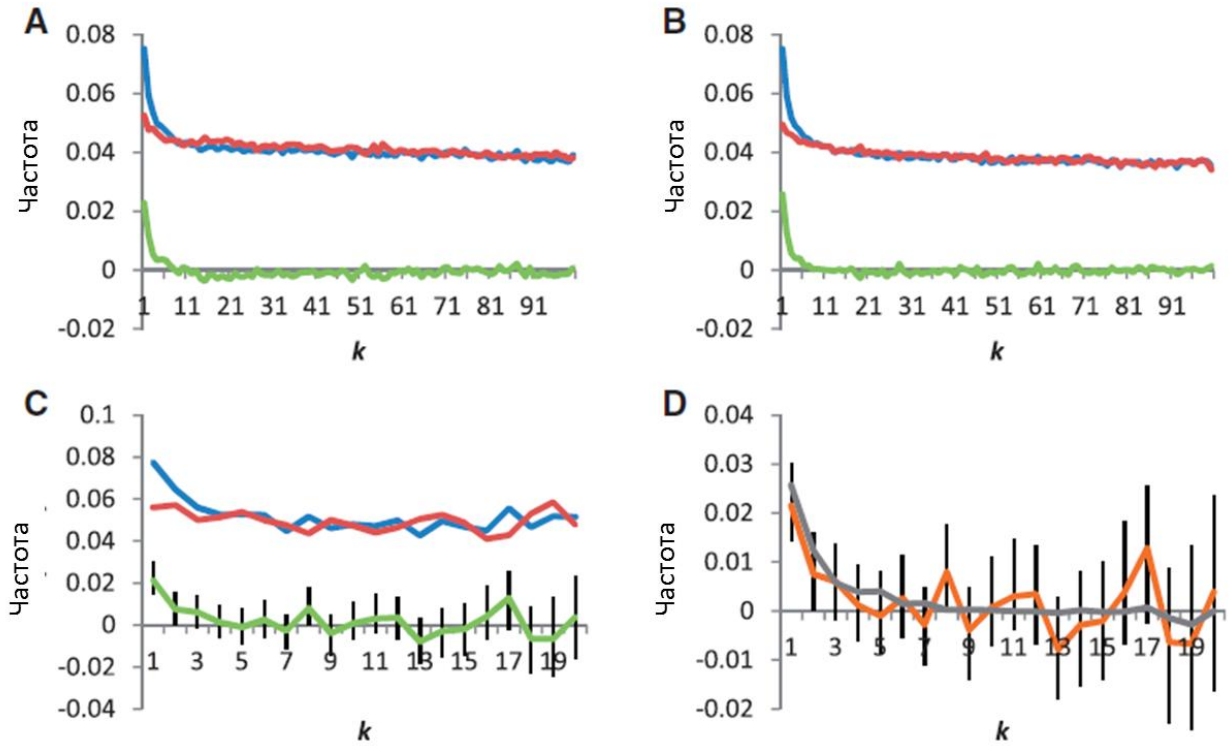

Рис. 15. Частота ДНМ у *Drosophila* для сайтов на разных расстояниях. (A-C) $d_d(k)$ (красный), $s_d(k)$ (синий) и $\alpha_d(k)$ (зеленый) для различных расстояний между сайтами k (горизонтальная ось). (A) интроны; (B) межгенные интервалы; (C) позиции 8-30 в интронах длинной до 120 нуклеотидов. (D) совмещенные кривые по $\alpha_d(k)$ с картинок B (серый) и C (оранжевый). Ошибки для $\alpha_d(k)$ посчитаны как 95% доверительные интервалы по 1000 испытаний; мы не рисовали доверительные интервалы, когда они были малы и плохо различимы.

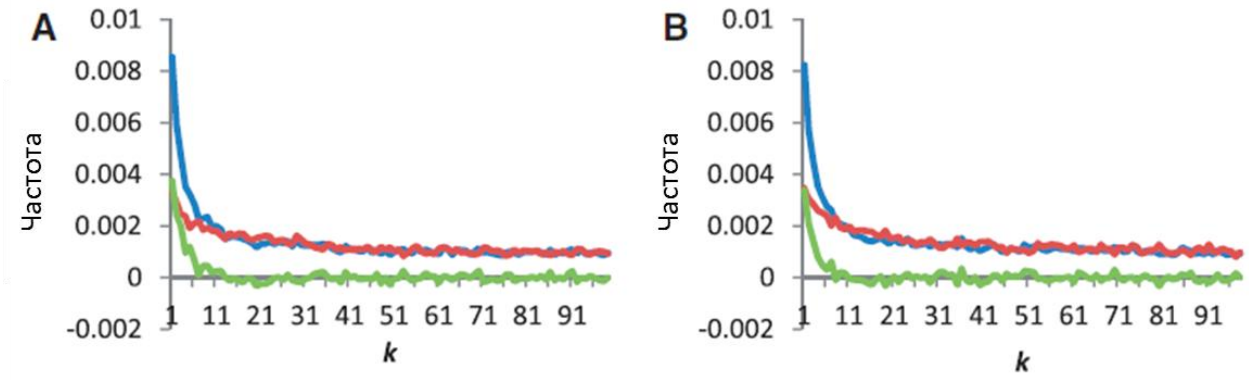

Рис. 16. Частота ТНМ для случая, когда две из трех мутаций, вовлечённых в ТНМ, случились в соседних сайтах (l=1), для различных расстояний до третьего сайта в *Drosophila*. $d_t(k,1)$ (красный), $s_t(k,1)$ (синий) и $\alpha_t(k,1)$ (зеленый) для $1 \leq k \leq 100$ (горизонтальная ось). (A) интроны; (B) межгенные интервалы.

Таким образом, мы показали, что ДНМ, затрагивающие пары сайтов на расстоянии до 10 нуклеотидов, составляют 2.3% и 5.6% от однонуклеотидных замен в приматах и *Drosophila* соответственно. Эти оценки учитывают локальную неравномерность скорости мутирования, если эта неравномерность сохраняется между близкими видами, использованными в анализах. Как мы показали в главе 1, эта неравномерность свойственна скорее для *Drosophila*, чем для приматов.

Например, если скорость мутирования автокоррелирует на расстоянии до 10 нуклеотидов, $d_d(k)$ и $s_d(k)$ для k≤10 будут одинаково повышены (т.к. эта автокорреляция приведет к кластеризации мутаций вдоль последовательности вне зависимости от того, лежат ли мутации на одной или разных филогенетических ветках), и $\alpha_d(k)$ останется неизменной.

Однако избыток замен на одной филогенетической линии может быть вызван различными причинами: неравномерностью скорости мутирования или отбора, которая не сохраняется между видами; ошибками выравнивания или сборки; мультинуклеотидными мутациями; эпистатическими взаимодействиями; а также неаллельной генной конверсией.

Локальная гетерогенность скорости мутирования может меняться между видами, что приведет к завышенной оценке МНМ, ведь в этом случае замены будут сильнее кластеризоваться внутри одной линии. Для того, чтобы проверить эту возможность, мы брали четыре разных тройки видов приматов с разным расстоянием между видами (из-за слишком больших длин веток подобный анализ невозможен для *Drosophila*). $\alpha_d(k)$ оставалась почти неизменным при увеличении филогенетического расстояния в 4 раза (рис. 14А), что говорит о том, что изменение локальной скорости мутирования или отбора для столь близких видов почти не влияет на наши результаты.

Ошибки сборки или выравнивания могут приводить к тому, что наблюдаются кластрированые замены. Ошибки секвенирования также могут к этому приводить, если распределены не равномерно [80–82]. Описанные артефакты в прокси геноме не влияют на значения $\alpha_d(k)$ или $\alpha_t(k,l)$. Лишь, ошибки в нескольких сайтах фокального генома, а именно геномов *H. sapiens* или *D. melanogaster,* будут смещать наши оценки, но эти геномы очень высокого качества, и подобные ошибки в них редки. Чтобы подтвердить, что наши оценки не подвержены описанным артефактам, мы повторили наш анализ, используя только нуклеотиды, совпадающие в человеке и шимпанзе, для анализа в тройке видов *H. sapiens*+*P. troglodytes* и *P. pygmaeus* (*M.mulatta* как внешняя группа). Сайты фокальной линии, подкреплённые ещё одним независимо собранным геномом, дают почти такие же результаты, что и неподкреплённый человеческий геном (рис. 17). Из этого следует, что вышеперечисленные артефакты практически не искажают наши оценки.

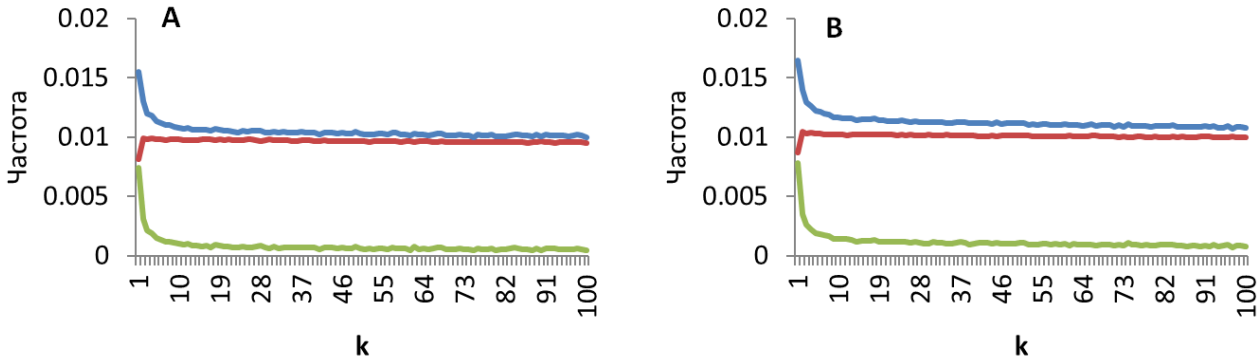

Рис. 17. Частота ДНМ на линии человека после отделения от общего предка с орангутангом для сайтов на разных расстояниях. $d_d(k)$ (красным), $s_d(k)$ (синим) и $\alpha_d(k)$ (зеленым) для расстояний между сайтами $1 \leq k \leq 100$ (горизонтальная ось). Графики (A) и (B) аналогичны графикам C и F рисунка 11, но здесь рассматривались сайты с совпадающими нуклеотидами в человеке и шимпанзе.

Положительный эпистаз приводит к увеличению приспособленности второй замены на фоне первой [15]. Этот феномен вызывает кластеризацию несинонимичных замен на одной линии [15,16,83,84]. Однако эпистаз теоретически невозможен в отсутствие отбора на последовательность [85], что подтверждено и эмпирическим путем [83]. В своей работе мы фокусировались на некодирующей последовательности и исключили все аннотированные экзоны вместе с 10 нуклеотидными участками по краям интронов, а также 5’ и 3’ UTR, что уменьшает количество нуклеотидов, на которые действует сильный отбор. Доля некодирующих позиций, находящихся под отбором, в человеке не превышает 10% [86–89]. Отбор в некодирующей последовательности *Drosophila* более распространен: 40% и 50% нуклеотидов в интронах и межгенных интервалах, соответственно, находятся под отбором [90]. Чтобы минимизировать влияние отбора, мы исключили 60% наиболее консервативных некодирующих нуклеотидов в геноме *D. melanogaster*, что превышает ожидаемое количество консервативных сайтов.

Чтобы быть еще более уверенными в том, что отбор не искажает наших оценок МНМ в мутагенезе мушек, мы использовали наиболее нейтрально эволюционирующую категорию сайтов: позиции 8-30 в коротких интронах [91] (рис. 15C), и получили результаты, статистически не отличимые от наблюдаемого паттерна для всех интронов (рис. 15D). Итак, вклад отбора в оценку $\alpha_d(k)$ минимален.

Наконец, неаллельная генная конверсия может приводить к кластеризации мутаций на одной филогенетической линии [36]. В результате неаллельной генной конверсии небольшой участок генома заменяется на другой участок, присутствующий в геноме и сходный по последовательности. Генная конверсия затрагивает участок более 1000 нуклеотидов [36,92,93], а значения $\alpha_d(k)$ становятся очень малы для расстояния между сайтами более 10 нуклеотидов (рис. 11, 15). Кроме того, мы исключили такие паралоги, генная конверсия с которыми могла бы привести к наблюдаемым ДНМ. Итак, наши наблюдения не являются последствием генной конверсии.

Таким образом, $\alpha_d(k)$ должна быть хорошей оценкой доли ДНМ в геноме. Значение $\sum_{1\leq k\leq 10}\alpha_d(k) = 0.023$ для k≤10 в человеке сходно с результатами, полученными другими группами по полиморфизму: 1.8% [12] и 2.0% [48]. В недавних статьях был также описан феномен МНМ по данным для *de novo* мутаций человека [5,35], однако в этих работах нет численных оценок, и они подтверждают наши результаты лишь качественно. Полученное нами значение $\alpha_d(1)$ по человеку наиболее точно совпадает с тем, что на данный момент принято в литературе [54]. Наша оценка $\alpha_d(k)$ для *Drosophila* превышает значения, полученные по *de novo* мутациям – 2.3% [19] и 2.8% [20]. Это можно объяснить эволюцией локальной скорости мутирования между *D. melanogaster* и *D. simulans*. К сожалению, филогенетическое расстояние между этими видами много больше, чем между человеком и шимпанзе, и мы никак не можем оценить эффект изменчивости локальной скорости мутирования между этими видами.

Глава 4. Использование динуклеотидной мутационной подписи для изучения свойств полимеразы зета.

В этой главе описаны результаты совместной работы коллектива авторов и мой вклад в неё – идея и дизайн проекта и получение данных по распределению одиночных и двойных мутаций в геноме.

В экспериментальных работах на пекарских дрожжах [10,56,94] и млекопитающих [95] было показано, что способная проходить повреждения полимераза зета (пол ζ) склонна совершать ошибки, приводящие к ДНМ. В дрожжевой линии, с активированной пол ζ, 64% от всех ДНМ составляли GC→AA/TT ДНМ и TC→AA ДНМ. Избыток этих типов ДНМ пропадал в линии дрожжей с нокаутированной пол ζ. При детектировании ДНМ использовался репортёрный ген, в котором легче детектировать TC→AA ДНМ, чем GC→AA/TT ДНМ, так как TC→AA ДНМ чаще приводят к стоп кодонам [51]. В экспериментах на клетках млекопитающих WGCW контекст особенно подвержен мутациям, которые делает пол ζ [11]. Таким образом, GC→AA/TT ДНМ может служить подписью пол ζ. Напротив, свидетельства о том, что TC→AA ДНМ является подписью пол ζ – недостаточны.

Используя GC→AA/TT ДНМ, произошедшие в линии человека после отхождения от общего предка с шимпанзе, как подпись пол ζ, мы исследовали, с какими свойствами ДНК связана работа этой полимеразы.

В согласии с результатами, полученными по полиморфизму человека [12,50] и по мутациям, вызывающим заболевания [50], мы обнаружили, что тандемные мутации на одной ветке (ТМОВ, примерно половина из которых случаются как ДНМ; см. главу 3) обогащены трансверсиями, в сравнении с однонуклеотидными мутациями (ОНМ); **к**, посчитанная по ТМОВ, на 60% меньше, чем **к,** посчитанная по тандемным мутациям на разных ветках (ТМРВ, χ^2 $P<1.1*10^{-67}$). Это отличие свидетельствует о разных механизмах, приводящих к ДНМ и ОНМ. Более слабый эффект мы наблюдали для ТМРВ, этот тип мутаций характеризуются **к** пониженной на 18% в сравнении с ОНМ (рис. 9, k=±1).

Участки генома с повышенной скоростью ОНМ скоррелированы с повышенной скоростью ДНМ (рис. 18А, $P=2*10^{-23}$, тест Корхан-Армитаж (в дальнейшем КА)). Этот результат не является артефактом метода, т.к. скорость ДНМ нормирована на локальную скорость мутирования. Таким образом, скорости ОНМ и ДНМ зависимы в 50 Кб окне, что свидетельствует о том, что, хотя механизмы, вызывающие ДНМ и ОНМ, могут быть различны, они скоррелированы вдоль последовательности генома.

Скорость ДНМ слабо зависит от времени репликации (рис. 18В, $P<0.03$, КА).

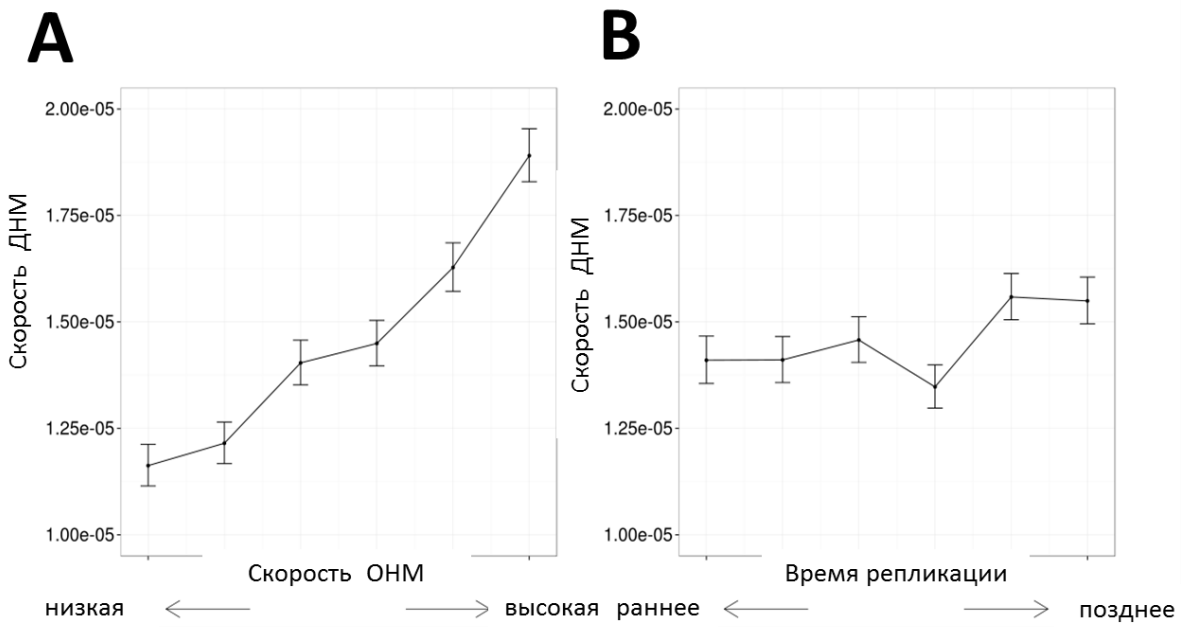

Рис. 18. Свойства ДНМ. (A) Скорость ДНМ сильно связана со скоростью ОНМ ($P<2 *10^{-23}$, КА). (B) Скорость ДНМ слабо связана со временем репликации (P=0.03, КА).

На данных по человеческому полиморфизму было показано, что GC→AA/TT ДНМ - наиболее частая ДНМ [12]. Мы подтвердили это наблюдение (рис. 19); более того, мы обнаружили, что среди GC→AA/TT ТМОВ ДНМ составляют 79%, когда среди всех ТМОВ эта доля близка к половине (рис. 14, 19).

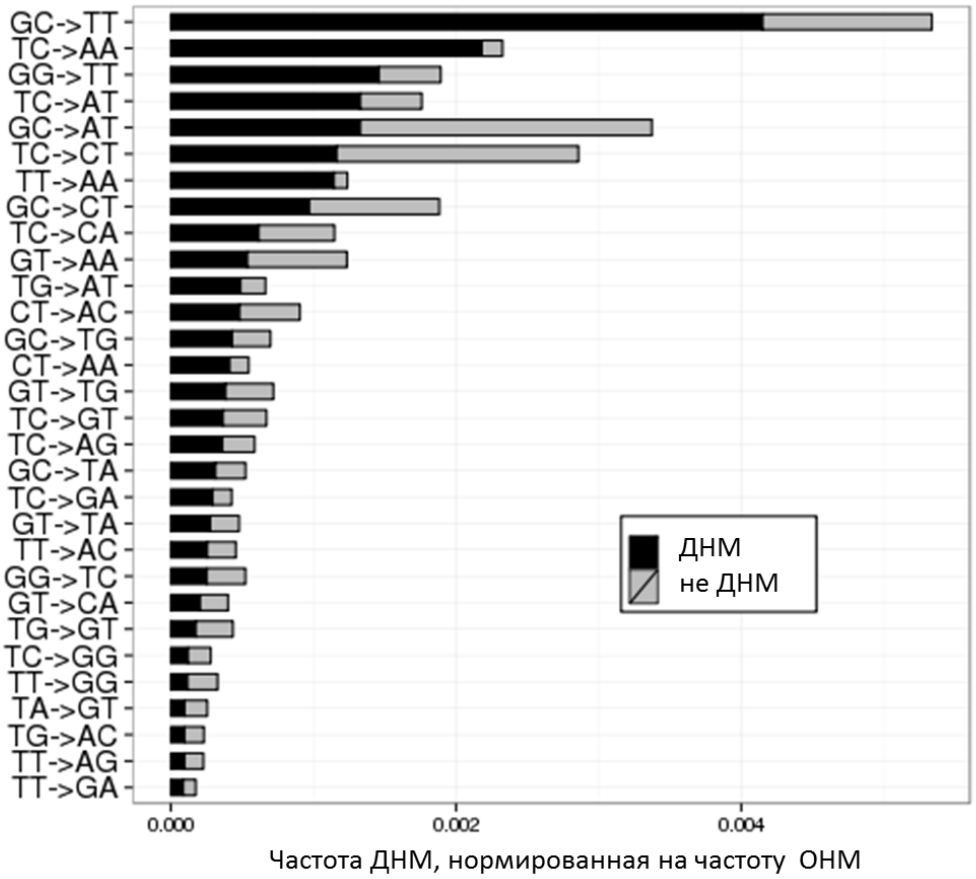

Рис. 19. 30 наиболее частых ДНМ. Мы объединяли пары комплементарных мутаций. Частота ДНМ (черным) и тандемных мутаций, произошедших как два независимых события (серым), различна для разных классов. GC→AA/TT ДНМ – наиболее частая среди всех классов ДНМ.

Время репликации и активность склонной к ошибкам пол ζ – это два фактора, влияющие на накопление мутаций. Мы использовали подпись пол ζ, чтобы изучить взаимодействие времени репликации и активности пол ζ. В противоположность слабой зависимости общей скорости ДНМ от времени репликации, скорость GC→AA/TT ДНМ на 61% выше в поздно

реплицирующихся участках (рис. 20B, P<2 *10^{-6}, КА). Эта зависимость может быть следствием других свойств генома, ассоциированных со временем репликации. Чтобы изучить зависимость между временем репликации и плотностью GC→AA/TT ДНМ и проконтролировать на возможные скрытые эффекты, мы включили сайты чувствительности к ДНКазе (DHS), скорость рекомбинации, обогащённость модификацией гистона H3K9me3 и ГЦ-состав, а также частоту ОНМ и ДНМ в качестве объясняющих переменных, а число GC→AA/TT ТМОВ – как объясняемую переменную, и количество ГЦ динуклеотидов – как ответственную переменную, в пуассоновскую регрессию. В отличие от других свойств ДНК, время репликации влияет на скорость GC→AA/TT замен на одной ветке достоверно (P < 3*10^{-5}, таблица 6). Более того, позиция, где случилась GC→AA/TT ТМОВ, расположена в более позднем времени репликации относительно GC→AA/TT ТМРВ (P<0.01, тест Манна-Уитни).

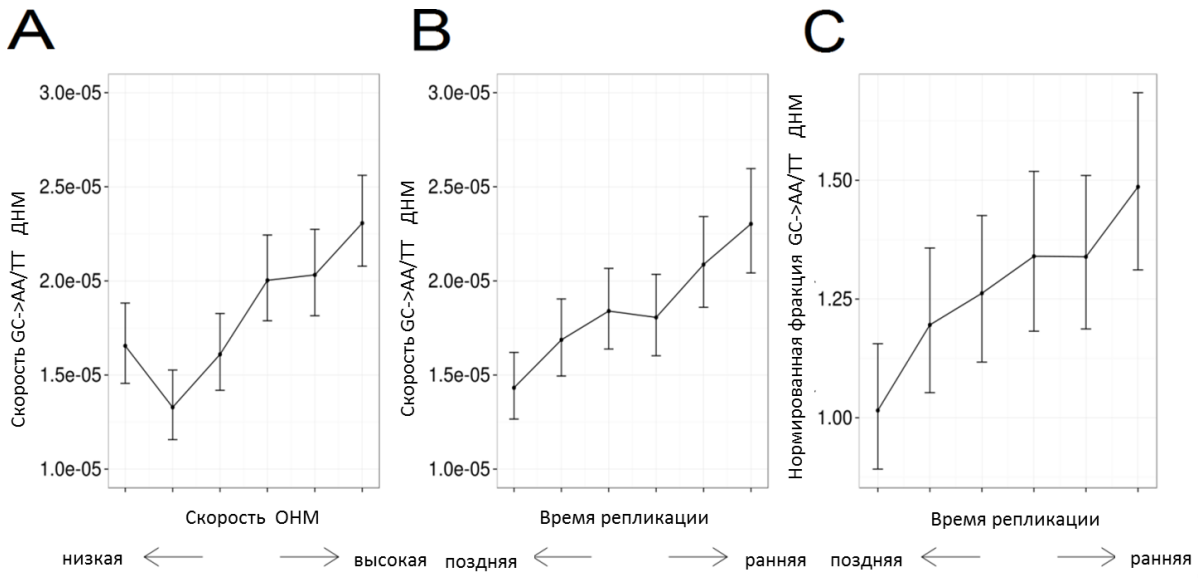

Рис. 20. Связь между скоростью GC→AA/TT ДНМ и временем репликации. (A) Скорость GC→AA/TT ДНМ растет со скорость ОНМ. (B) Скорость GC→AA/TT ДНМ растет от раннего к позднему времени репликации, как и доля GC→AA/TT ДНМ среди всех ДНМ (C).

Таблица 6. Зависимости скорости GC→AA/TT ДНМ от свойств ДНК, посчитанные пуассоновской регрессией.

Объясняющие переменные	P-value
Сокрость рекомбинации (Проект 1000 геномов)	0.37
DHS	0.50
Плотность H3K9me3	0.29
Время репликации	$2.75*10^{-5}$
ГЦ-состав	0.39
Скорость ОНМ	$1.61*10^{-13}$
Скорость ДНМ	0.0016

Чтобы разделить зависимость плотности мутационной подписи пол ζ и общей скорости ДНМ от времени репликации, мы рассматривали долю GC→AA/TT ДНМ среди всех ДНМ как функцию от времени репликации. Это отношение монотонно растёт от участков, реплицирующихся рано, к участкам, реплицирующимся поздно, и оно на 46% выше для участков поздней репликации (рис. 20C, $P<6*10^{-4}$, 1000 испытаний), что свидетельствует о разных механизмах возникновения мутаций, являющихся подписью пол ζ и остальных ДНМ.

Подпись пол ζ составляют две комплементарные мутации – GC→AA и GC→TT. Мы проверили, связана ли одна из комплементарных мутаций с цепью ДНК, являющейся лидирующей при репликации. У человека, лидирующая и отстающая цепи реплицируются полимеразами эпсилон и дельта соответственно [96]. Из-за особенностей, связанных с репликативными полимеразами или другими свойствами, различающими лидирующую и отстающую цепь,

мутационный процесс отличается между нитями ДНК [97,98]. Так как мы обнаружили связь активности пол ζ с временем репликации, мы решили проверить асимметрию её мутационной подписи между лидирующей и отстающей цепями. Мы использовали производную времени репликации для восстановления направления репликационной вилки [97,98]. Мы не наблюдали асимметрии (рис. 21), что говорит о том, что активность пол ζ не завязана на механизм, специфичный для одной из цепей.

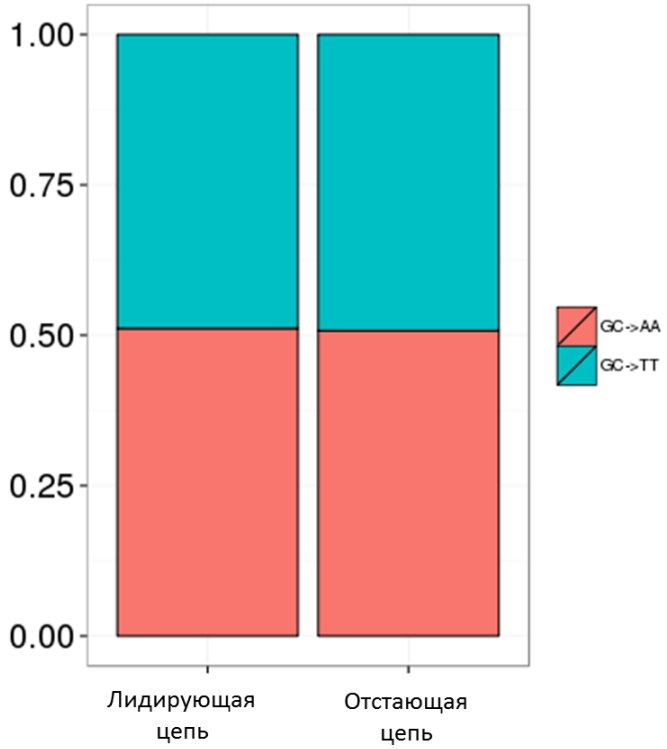

Рис. 21. Мутационная подпись пол ζ симметрична в лидирующей и отстающей цепях ДНК. Мы выбрали участки, которые реплицируются с большой вероятностью как лидирующая или отстающая цепь, и обнаружили, что частота комплементарных мутаций GC→AA и GC→TT на них одинакова.

Скорость и спектр мутаций могут зависеть от транскрипции [6,99]. Скорость мутирования в зародышевых линиях коррелирует с уровнем экспрессии генов из-за того, что ДНК расплетается во время транскрипции и находится в

одноцепочечном состоянии [100]. Нетранскрибируемая цепь больше времени проводит в одноцепочечном состоянии, чем транскрибируемая. Более того, ассоциированная с транскрипцией репарация (TC-NER) способна обнаруживать и удалять повреждённые нуклеотиды с транскрибируемой цепи. Подверженность мутациям нетранскрибируемой цепи, как и репарация, чинящая мутации специфически на транскрибируемой цепи, приводят к более высокой скорости мутирования на нетранскрибируемой цепи в сравнении с транскрибируемой цепью [101], и к асимметрии между цепями [99,102]. В дрожжах пол ζ вовлечена в мутагенез, ассоциированный с транскрипцией [103,104]. Мы решили проверить, отличается ли соотношение GC→AA и GC→TT мутаций на транскрибируемой и не транскрибируемой нитях ДНК. GC→TT ДНМ происходит на 40% чаще на нетранскрибируемой цепи в сравнении с транскрибируемой цепью или межгенными интервалами (рис. 22). Этот паттерн подтверждает, что активность пол ζ связана с транскрипцией. Более того, повышение частоты GC→TT ДНМ на не транскрибирующейся нити в сравнении с межгенными интервалами говорит о том, что асимметрия вызвана не TC-NER, активность которой не может привести к увеличению количества мутаций в транскрибируемом регионе [105]. Итак, у нас есть основания предполагать, что пол ζ во время репликации играет роль в прохождении повреждений ДНК, накопившихся при транскрипции.

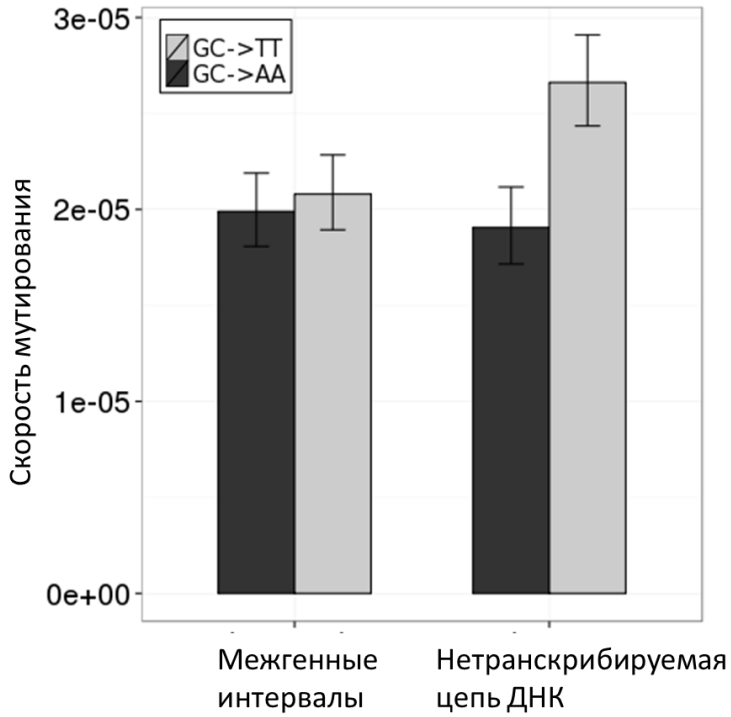

Рис. 22. Ассоциированная с транскрипцией асимметрия между GC→AA и GC→TT ДНМ. Для межгенных интервалов частоты GC→AA и GC→TT ДНМ посчитаны для референтной цепи.

Мы заметили, что GC→TT ДНМ вызывает синдром Костелло [53] и, возможно, эта мутация связана с активностью пол ζ. GC→TT ДНМ в GCG контексте происходит в первом экзоне гена HRAS и приводит к замене 12-го глицина на валин (рис. 23B). Эта мутация часто происходит в сперматозоидах; так, ее частота в 1.7 раз выше, чем частота ОНМ G35→T35 [53], хотя эта ОНМ приводит к той же аминокислотной замене. Это означает, что необычайно высокая частота GC→TT ДНМ в гене HRAS не может быть полностью объяснена

внутритканевым отбором на эту мутацию, и свидетельствует в пользу того, что эта ДНМ является следствием неканонического мутационного механизма.

пол ζ, в сравнении с основными репликативными полимеразами гораздо чаще вставляет не правильные нуклеотиды и её ошибки чаще бывают рансверсиями [10,95,106]; кроме того, пол ζ характеризуется низкой процессивностью и за один раз синтезирует участок ДНК длинной до 1000 нуклеотидов [56,107,108]. Таким образом, если пол ζ часто принимает участие в репликации экзона 1 гена HRAS, это должно приводить к увеличению скорости мутирования и уменьшению **к** в окрестности нескольких сотен нуклеотидов от сайта многократно наблюденной GC→TT ДНМ.

Мы сравнили геномы человека и мыши для того, чтобы по некодирующим межвидовым заменам оценить скорость мутирования и **к** в окрестности сайта, в котором многократно наблюдали GC→TT ДНМ. Действительно, при сравнении скорости мутирования на расстоянии до 300 нуклеотидов от GC→TT ДНМ, наблюдённой в сперматозоидах, мы обнаружили, что частота замен выше в 2 раза, а **к** ниже в 2 раза, в сравнении с сайтами на расстоянии более 1 Kb (рис. 23A). GC→TT ДНМ в гене HRAS наблюдали в сперматозоидах, но эта мутация не зафиксирована между видами, поэтому локально увеличенная скорость мутирования не является последствием сложной мутации, включающей GC→TT ДНМ. Более того, описанная замена вредна, и её закреплению противодействует отбор, таким образом, гипермутабильный GC динуклеотид не заменяется на протяжении миллионов лет эволюции. Таким образом, мы наблюдаем горячую

точку мутагенеза со спектром, смещённым в сторону трансверсий, локализованную в пределах 1000 нуклеотидов от сайта, в котором многократно наблюдали GC→TT ДНМ, все это свидетельствует в пользу того, что пол ζ ответственна за GC→TT ДНМ, вызывающую синдром Костелло. Для того, чтобы проконтролировать на другие причины локальной гипермутабильности, мы проанализировали ген KRAS – гомолог гена HRAS, в котором, вместо GC динуклеотида в гомологичном сайте, находится GT динуклеотид. Мы не обнаружили изменённой **к** или локально повышенной скорости мутирования в ближайшей окрестности GT сайта (рис. 23C).

Мутации в гене SOD1 ответственны за 20% случаев амиотрофического латерального склероза (АЛС) [109]. В гене SOD1 три раза независимо наблюдали возникновение GC→TT ДНМ в четвёртом, шестом и десятом кодонах, приводящих к АЛС [109–111], поэтому знание механизма, вызывающего эти мутации, имеет медицинское значение. Мы сделали анализы, аналогичные описанным для гена HRAS, и обнаружили локально повышенную скорость мутирования ($P<0.05$, χ^2 тест) и повышенную долю трансверсий ($P<0.1$, χ^2 тест, если сравнивать с **к** по всей хромосоме, а не для 1Kb участка, $P=0.001$) (рис. 23D). В двух случаях из трёх GC→TT ДНМ происходила в GCG контексте и все GC→TT ДНМ произошли на нематричной цепи ДНК, что соответствует паттернам, наблюдаемым в гене HRAS. То, что все описанные случаи GC→TT ДНМ происходят на нематричной цепи, согласуется с наблюдениями на уровне всего генома (рис. 22).

Вероятно, пол ζ участвует в синтезе ДНК на цепи, комплементарной AP (apurinic/apyrimidinic) сайту, возникающему при репарации ДНК после дезаминирования метилцитозина в CpG контексте [44,112,113], поэтому мы и наблюдаем мутации в CGC контексте.

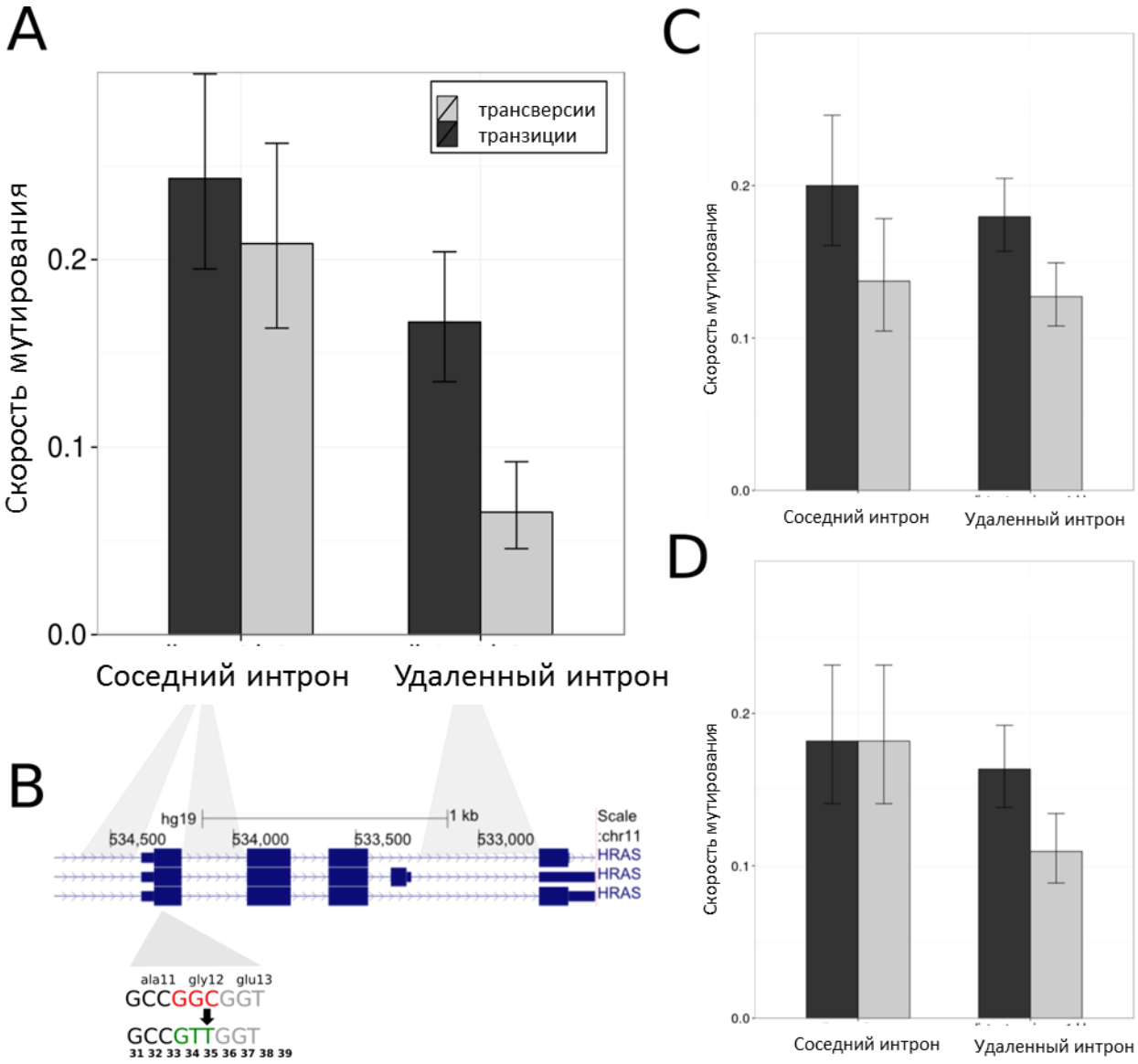

Рис. 23. GC→TT/AA ДНМ в генах HRAS and SOD1, ассоциированных с синдромом Костелло и амиотрофическим латеральным склерозом,

соответственно. (A) ген HRAS. Интроны, соседствующие с горячей точкой для GC→TT ДНМ, мутируют значительно быстрее и имеют пониженную **к,** по сравнению с чуть более удалёнными интронами. (B) Структура гена HRAS. GC→TT ДНМ происходит в двенадцатом глицине первого экзона. (C) Ген KRAS – контроль для гена HRAS. (D) ген SOD1.

Мы наблюдаем локальную гипермутабильность в окрестности GC→AA/TT ДНМ, многократно возникающей в генах HRAS и SOD1, и решили проверить эти паттерны на уровне всего генома. Для этих целей мы изучали локальную скорость ОНМ поблизости от ТМОВ и ТМРВ. Плотность ОНМ повышена на расстоянии до нескольких тысяч нуклеотидов от ТМОВ и ТМРВ (рис. 24A). Увеличение мутабильности в близи ТМОВ могжет быть вызвано связью между ДНМ и локальными горячими точками мутагенеза; однако повышенная мутабильность поблизости от ТМОВ и ТМРВ также может быть следствием того, что тандемные мутации чаще происходят в горячих точках мутагенеза [32]. Мы можем вычленить эффект ДНМ, зная скорости мутирования в окрестности ТМОВ и ТМРВ и долю ДНМ среди ТМОВ (см. методы).

Удивительно, но частота ОНМ вокруг ТМРВ меньше, чем вокруг ТМОВ, кроме как для 10 ближайших нуклеотидов (рис. 24A). Таким образом, вопреки ожиданию, две мутации в соседних сайтах, случившиеся в линии человека, хуже предсказывают горячую точку мутагенеза в сравнении с двумя мутациями, случившимися на разных линиях. Тем не менее, частота ОНМ повышена и

поблизости от ДНМ. Паттерны локальной скорости мутирования схожи вокруг GC→TT/AA ДНМ и всех типов ДНМ (рис 24B).

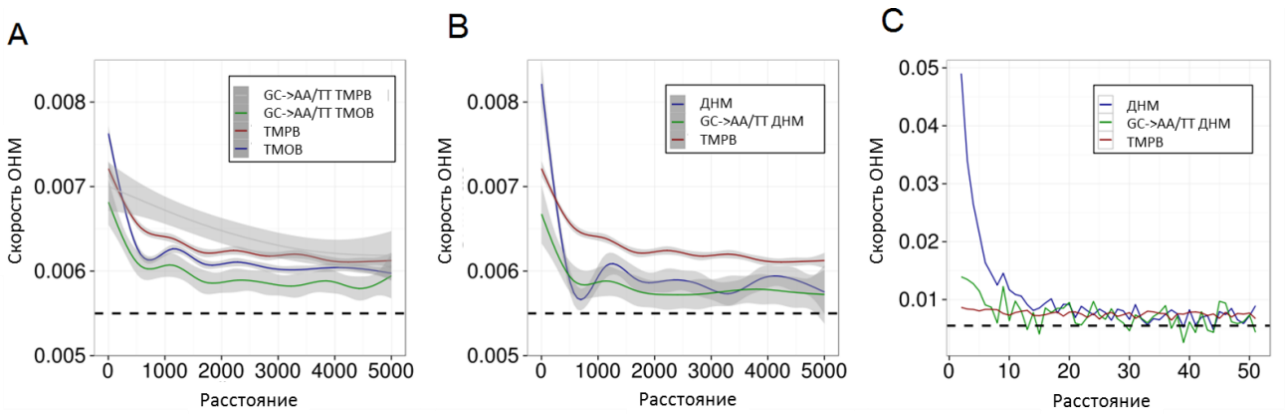

Рис. 24. ДНМ ассоциированы с горячими точками мутагенеза. Черная пунктирная линия показывает среднегеномную скорость ОНМ. (A) Средняя скорость ОНМ падает при удалении от тандемных мутаций (GC→AA/TT TMPB, GC→AA/TT TMOB, TMPB, TMOB). (B) и (C) Средняя скорость ОНМ падает при удалении от ДНМ. В (A) и (B) скорость ОНМ аппроксимировалась с использованием генерализованной линейной модели.

Паттерны для 10 ближайших нуклеотидов для TMOB и TMPB сильно отличаются (рис. 24B). Для этих расстояний частота ОНМ для тандемных мутаций на одной ветке значительно выше. Этот эффект, вероятно, является причиной высокой доли ТНМ мутаций, содержащих TMOB и ОНМ, что согласуется с результатами главы 3 (рис. 13); такие мутации случаются на

расстоянии до 10 нуклеотидов и обогащены трансверсиями [12], что мы и наблюдаем в наших данных (рис 24C, 25). Эти сложные мутации не результат генной конверсии, т.к. мы исключали паралогичные последовательности (см. методы). В противоположность всем ДНМ, в 10 соседних с GC→AA/TT ДНМ сайтах нет резкого скачка частоты ОНМ (рис 24C). Это подтверждает гипотезу, что данный тип ДНМ, как правило, является мутационной подписью пол ζ, в противоположность остальным ДНМ, которые часто происходят в составе более сложных мутаций (рис. 25B).

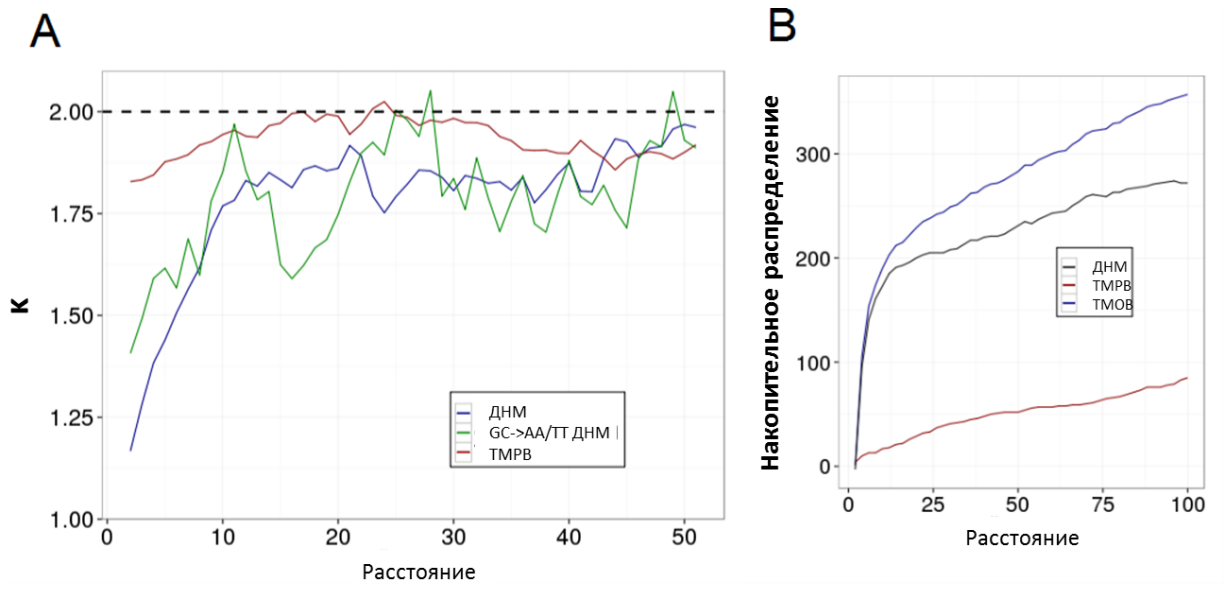

Рис. 25. Мутационные свойства ОНМ в окрестности двойных и тандемных мутаций. (A) **к** значительно понижена для 10 сайтов, соседних со всеми типами ДНМ или GC→AA/TT ДНМ, но лишь слабо понижена для 10 сайтов соседних с TMPB (также см. рис. 9C). **к** посчитана в 10-нуклеотидном скользящем окне. Чёрная пунктирная линия отображает среднегеномное значение **к.** (B) ДНМ кластеризуются на коротких расстояниях. По горизонтальной оси – расстояния

между парами тандемных или двойных мутаций (две ТМОВ, две ТМРВ, две ДНМ); по вертикальной оси – количество событий на расстоянии меньше данного. Линейная зависимость соответствует независимому накоплению мутаций, а наблюдаемая нелинейность соответствует избытку мутаций на малом расстоянии друг от друга.

Таким образом, скорость мутирования в окрестности подписи пол ζ сходна с тем, что мы наблюдаем для остальных ДНМ, и отличается от паттернов для генов HRAS и SOD1, где поблизости от подписи мы видим повышение скорости в 1.5-2 раза. Если горячая точка мутагенеза ассоциирована с GC мотивом, то при закреплении GC→AA/TT ДНМ эта горячая точка исчезает. В генах HRAS и SOD1 замена GC→TT приводит к тяжелым заболеваниям; таким образом, отбор сохраняет этот мотив вопреки его мутабильности, в результате чего мы видим повышенную скорость мутирования вблизи от этих сайтов.

Тандемные мутации могут кластеризоваться на одной филогенетической ветке [15] или в одном гаплотипе под действием эпистатического отбора [16]. Однако такая кластеризация, вызванная отбором, сильнее всего проявляется в регионах с наибольшей консервативностью [83,114]. Мы не рассматриваем UTR, границы интронов и экзоны, и таким образом избавляемся от большинства регионов генома под сильным отбором. К тому же под отбором находится лишь 8% генома человека [86], и столь малая доля не должна сильно влиять на наши результаты. Замаскировав паралоги, мы исключили влияние генной конверсии, а

исключив регионы, маскированные RepeatMasker [58], мы избежали артефактов, связанных с особенностями мутагенеза повторов.

Результаты нашей работы согласуются с результатами других исследований, в которых с использованием данных о внутривидовом полиморфизме или данных по болезням показывают, что GC→AA/TT – наиболее частая ДНМ в человеке [12,50,51,53].

Мы показали, что мутационная подпись пол ζ сильно ассоциирована с участками поздней репликации. Этот результат не является следствием влияния других факторов, определяющих изменчивость скорости мутирования вдоль человеческого генома (GC состав, чувствительность к ДНКазе, рекомбинация, модификации гистонов) [23], как и просто скорости ОНМ и ДНМ. Наблюдаемое обогащение участков поздней репликации GC→AA/TT ДНМ может быть следствием пониженной эффективности репарации неспаренных нуклеотидов [115,116], а не ассоциацией пол ζ с временем репликации. Тем не менее, если общая скорость ДНМ может зависеть от эффективности систем репарации, то нет причин ожидать, что репарация повреждений, приводящих к GC→AA/TT ДНМ и к другим типам ДНМ, будут значительно различаться. Кроме того, система репарации неспаренных нуклеотидов особенно эффективно чинит транзиции [115,116], а мы наблюдаем, что скорость GC→AA/TT ДНМ ассоциирована с пониженной **к** ($P<10^{-3}$, КА рис. 26), а также **к** снижена вблизи многократно наблюдаемых GC→ TT ДНМ в генах HRAS и SOD1. В итоге наиболее вероятное объяснение для роста доли GC→AA/TT среди всех ДНМ с временем репликации

– больший вклад пол ζ в позднем времени репликации. Это означает, что повышенная активность подверженных ошибкам полимераз является одним из факторов повышения скорости мутирования в позднем времени репликации. Наши наблюдения подтверждают эксперименты, в которых обнаружили, что активность пол ζ выше на поздних стадиях синтетической фазы клеточного цикла [117], особенно при условии нехватки нуклеотидов. В дрожжах при нокауте REV1, гена, необходимого для работы пол ζ [10,44], исчезала зависимость между скоростью мутирования и временем репликации для нескольких репортерных генов [118]. Частота ошибок, совершаемых пол ζ, на несколько порядков выше, чем для основных репликативных полимераз δ и ε [95,106,119]. Таким образом, даже малая часть генома, реплицируемая пол ζ, может влиять на среднюю скорость мутирования по геному.

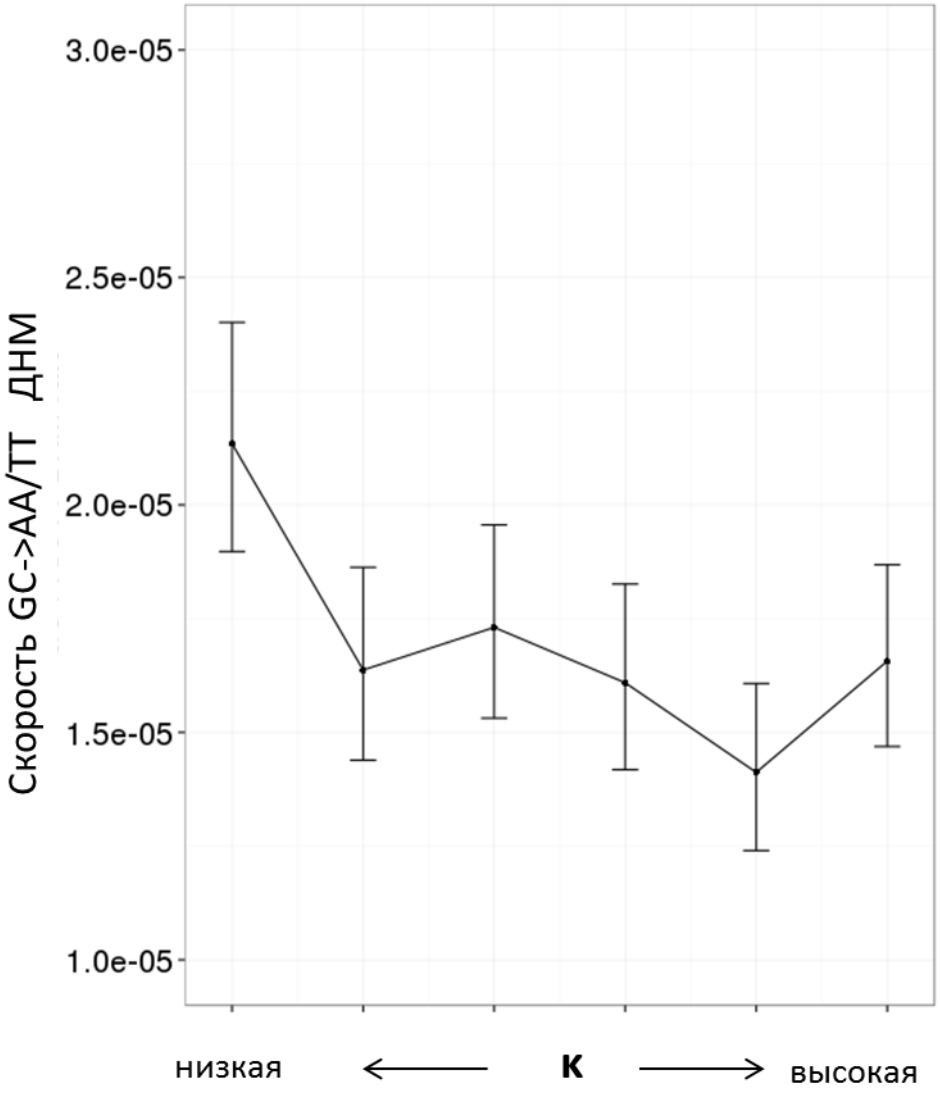

Рис. 26. Скорость GC→AA/TT ДНМ падает с увеличением **к**. Мы отсортировали 50 Kb окна по значению **к** и разбили эти окна на 6 групп равного размера.

Мы предположили, что пол ζ частично может объяснять зависимость скорости мутирования от времени репликации, и решили проверить, какую долю дисперсии скорости ОНМ вдоль генома объясняют эти факторы. Двухсторонняя ANOVA показала, что и GC→AA/TT ДНМ, и время репликации достоверно

связаны с локальной скоростью мутирования, и объясняют 0.27% и 10% изменчивости скорости мутирования соответственно. Хотя GC→AA/TT ДНМ самостоятельно объясняют лишь 0.27% дисперсии ($P < 2*10^{-16}$), низкое число (только 1770) наблюдаемых GC→AA/TT ДНМ усложняют интерпретацию этих цифр. Такое же количество ОНМ объясняет 0.51% скорости ОНМ по геному (хотя все ОНМ объяснят 100%) (рис. 27). Получается, что пол ζ вносит вклад в локальную скорость мутирования, однако численно оценить этот вклад нельзя, используя столь малое количество ДНМ, являющихся мутационной подписью пол ζ. В неданем анализе *de novo* мутаций человека обнаружили 161 МНМ со специфическим мутационным спектром, смещённым в сторону трансверсий по сравнению с некластеризованными мутациями [5]. Спектр кластерных мутаций был сильнее всего обогащен заменами C→G, что, как правило, связывают с активностью REV1 [44,120], которая необходима для рекрутирования пол ζ [44]. C→G мутации могут быть следствием синтеза ДНК полимеразой, склонной к ошибкам [5]. Недавние свидетельства о том что пол ζ способна синтезировать участки в несколько тысяч нуклеотидов, позволяют предположить, что часть из увиденных кластеров является результатом работы этой полимеразы.

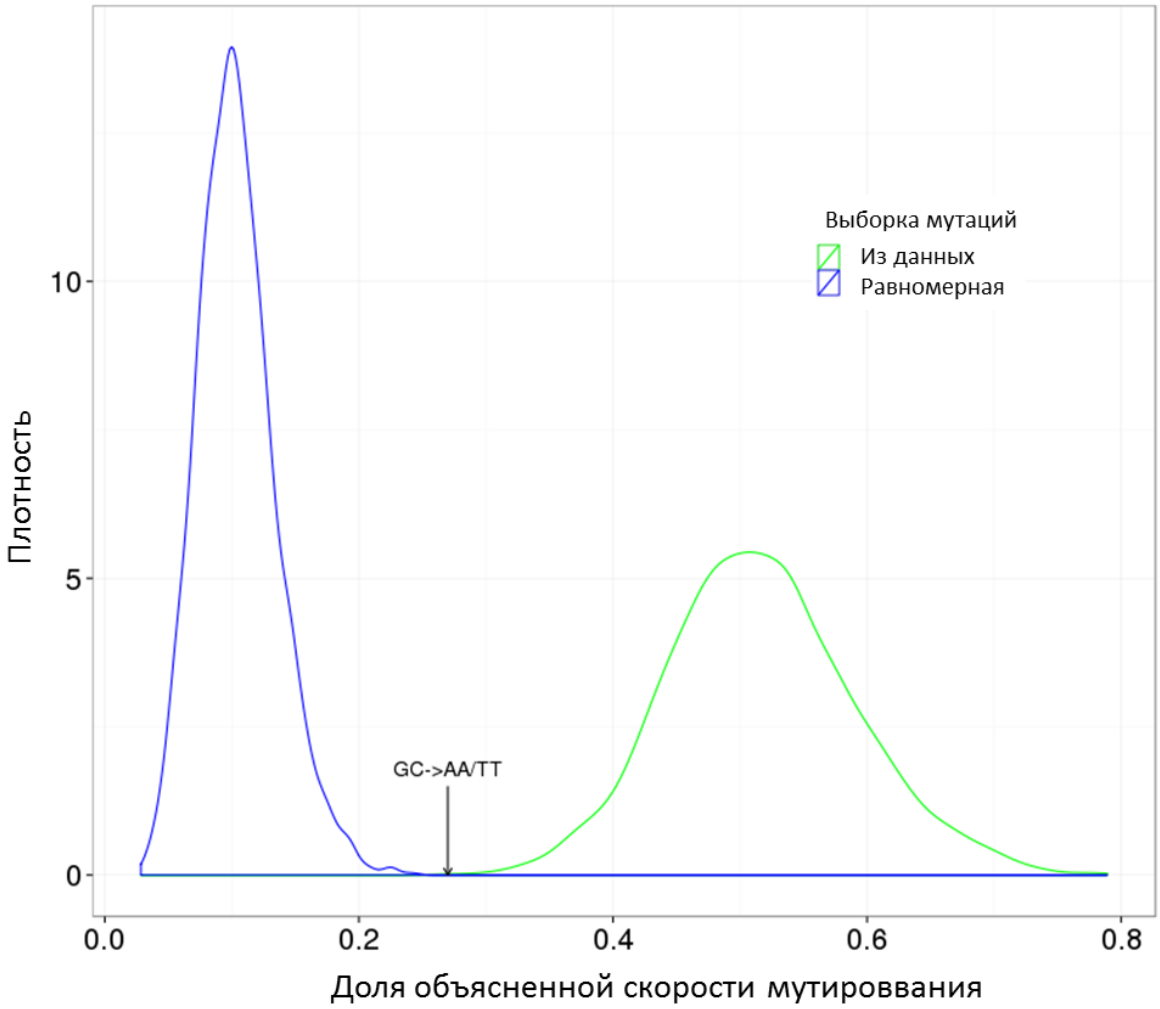

Рисунок 27. Доля объясненной скорости мутирования GC→AA/TT ДНМ и тем же количеством ОНМ. Мы выбирали мутации либо предполагая равную скорость мутирования вдоль по геному (синяя кривая), либо случайным образом из данных. Затем мы, используя функцию aov() из R, измеряли долю объясненной дисперсии; так мы проделали 2000 раз и построили приведенные распределения. Значение для GC→AA/TT ДНМ обозначено стрелкой.

Многократно наблюдаемые мутации GC→TT ДНМ в генах HRAS и SOD1 позволяют изучить горячие точки для GC→TT ДНМ, консервативные в

эволюции. Мы обнаружили, что рядом с такими сайтами скорость ОНМ и доля трансверсий повышены, что позволяет предполагать, что мутагенез, приводящий синдрому Костелло и к АЛС, осуществляется пол ζ.

GC→TT ДНМ характеризуется свойствами, отличными от свойств, характерных для других ДНМ. Во первых, они гораздо сильнее ассоциированы с временем репликации. Во вторых, несмотря на то, что GC→TT ДНМ, как и остальные ДНМ, чаще происходят в участках с высокой плотностью мутаций как на больших (рис. 18A, 20A), так и на малых шкалах (рис. 24), но на расстоянии до 10 нуклеотидов поведение GC→TT ДНМ и остальных ДНМ радикально отличается, и подпись пол ζ гораздо реже вовлечена в МНМ (рис. 24C, 25B). Эти особенность также свидетельствуют в пользу различных мутационных механизмов, приводящих к GC→TT ДНМ и большинству других ДНМ.

В дивергенции близких видов большая доля множественных замен, случившихся на одной ветке, являются последствием сложной мутации, а не серии независимых событий (рис. 14B). Использование дивергенции для изучения типов ДНМ и их геномных свойств проливает свет на механизмы мутагенеза и даже позволяет предполагать, как именно происходят мутации, приводящие к болезням.

Выводы

1) SNP попадает в сайт, содержащий другой SNP, в 3.5, 2.5 и 1.4 раза чаще среднего у *D. melanogaster*, *H. sapiens* и *S. commune*, соответственно, что говорит о сильной гетерогенности скорости мутирования на уровне однонуклеотидных позиций

2) Соотношение транзиций и трасверсий смещается в сторону трансверсий в сайтах поблизости от трансверсий и в сайтах, в которых наблюдалась трансверсия; так, в человеке доля трансверсий растёт втрое для сайтов, в которых наблюдалась другая трансверсия

3) Многонуклеотидные мутации – распространённый феномен у *Metazoa*, и доля мутаций, затрагивающих 2 сайта на расстоянии до 10 нуклеотидов**,** составляет 5.6% и 2.3% от частоты однонуклеотидных мутаций у *D. melanogaster* и *H. sapiens*

4) Около половины тандемных замен в линии человека после отделения от шимпанзе возникли как динуклеотидные мутации

5) Активность полимеразы зета связана со временем репликации, данного сегмента ДНК

6) Полимераза зета в 1.4 раза интенсивнее работает на нетранскрибируемой цепи.

7) Вероятно, полимераза зета приводит к мутациям, вызывающим синдром Костелло и боковой амиотрофический склероз

Благодарности

Кондрашову А.С., Солдатову Р.А., Терехановой Н.В., Андреяновой М.А., Леушкину Е.В., Вахрушевой О.А., Виноградовой С.В., Науменко С.А., Гарушянц С.К. за ценные обсуждения и замечания.

Список публикаций по теме диссертации

Seplyarskiy VB, Kharchenko P, Kondrashov AS, Bazykin GA. Heterogeneity of the transition/transversion ratio in Drosophila and Hominidae genomes. Mol Biol Evol. 2012 Feb 15.

Terekhanova NV, Bazykin GA, Neverov A, Kondrashov AS, Seplyarskiy VB. Prevalence of multinucleotide replacements in evolution of primates and Drosophila. Mol Biol Evol. 2013 Feb 27.

Seplyarskiy VB, Logacheva MD, Penin AA, Baranova MA, Leushkin EV, Demidenko NV, Klepikova AV, Kondrashov FA, Kondrashov AS, James TY. Crossing-over in a hypervariable species preferentially occurs in regions of high local similarity. Mol Biol Evol. 2014 Aug 18.

Baranova MA, Logacheva MD, Penin AA, Seplyarskiy VB, Safonova YY, Naumenko SA, Klepikova AV, Gerasimov ES, Bazykin GA, James TY, Kondrashov AS. Extraordinary Genetic Diversity in a Wood Decay Mushroom. Mol Biol Evol. 2015 Jul 10.

Seplyarskiy VB, Bazykin GA, Soldatov RA. Polymerase ζ activity is linked to replication timing in humans: evidence from mutational signatures. Mol Biol Evol. 2015 Sep 15.

Список литературы

1 Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* 2013; **499**:214–218.

2 Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 2012; **488**:471–475.

3 Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* 2014; **46**:818–825.

4 Besenbacher S, Liu S, Izarzugaza JMG, Grove J, Belling K, Bork-Jensen J, *et al.* Novel variation and de novo mutation rates in population-wide de novo assembled Danish trios. *Nat Commun* 2015; **6**:5969.

5 Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, *et al.* Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet* Published Online First: 18 May 2015. doi:10.1038/ng.3292

6 Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, *et al.* Signatures of mutational processes in human cancer. *Nature* 2013; **500**:415–421.

7 Lercher MJ, Williams EJ, Hurst LD. Local similarity in evolutionary rates extends over whole chromosomes in human-rodent and mouse-rat comparisons: implications for understanding the mechanistic basis of the male mutation bias. *Mol Biol Evol* 2001; **18**:2032–2039.

8 Hodgkinson A, Eyre-Walker A. Variation in the mutation rate across mammalian genomes. *Nat Rev Genet* 2011; **12**:756–766.

9 Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* 2007; **24**:1586–1591.

10 Stone JE, Lujan SA, Kunkel TA, Kunkel TA. DNA polymerase zeta generates clustered mutations during bypass of endogenous DNA lesions in Saccharomyces cerevisiae. *Environ Mol Mutagen* 2012; **53**:777–786.

11 Saribasak H, Maul RW, Cao Z, Yang WW, Schenten D, Kracker S, *et al.* DNA polymerase ζ generates tandem mutations in immunoglobulin variable regions. *J Exp Med* 2012; **209**:1075–1081.

12 Harris K, Nielsen R. Error-prone polymerase activity causes multinucleotide mutations in humans. *Genome Res* 2014; **24**:1445–1454.

13 Kondrashov AS. Direct estimates of human per nucleotide mutation rates at 20 loci causing Mendelian diseases. *Hum Mutat* 2003; **21**:12–27.

14 Britten RJ. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci U S A* 2002; **99**:13633–13635.

15 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; **429**:558–562.

16 Callahan B, Neher RA, Bachtrog D, Andolfatto P, Shraiman BI. Correlated evolution of nearby residues in Drosophilid proteins. *PLoS Genet* 2011; **7**:e1001315.

17 Behjati S, Huch M, van Boxtel R, Karthaus W, Wedge DC, Tamuri AU, *et al.* Genome sequencing of normal cells reveals developmental lineages and mutational processes. *Nature* 2014; **513**:422–425.

18 Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, *et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* 2015; **348**:880–886.

19 Keightley PD, Trivedi U, Thomson M, Oliver F, Kumar S, Blaxter ML. Analysis of the genome sequences of three Drosophila melanogaster spontaneous mutation accumulation lines. *Genome Res* 2009; **19**:1195–1201.

20 Schrider DR, Houle D, Lynch M, Hahn MW. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. *Genetics* 2013; **194**:937–954.

21 Rieux A, Eriksson A, Li M, Sobkowiak B, Weinert LA, Warmuth V, *et al.* Improved calibration of the human mitochondrial clock using ancient genomes. *Mol Biol Evol* 2014; **31**:2780–2792.

22 Nachman MW. Single nucleotide polymorphisms and recombination rate in humans. *Trends Genet TIG* 2001; **17**:481–485.

23 Schuster-Böckler B, Lehner B. Chromatin organization is a major influence on regional mutation rates in human cancer cells. *Nature* 2012; **488**:504–507.

24 Panchin AY, Mitrofanov SI, Alexeevski AV, Spirin SA, Panchin YV. New words in human mutagenesis. *BMC Bioinformatics* 2011; **12**:268.

25 Bird AP. DNA methylation and the frequency of CpG in animal DNA. *Nucleic Acids Res* 1980; **8**:1499–1504.

26 Harris K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc Natl Acad Sci U S A* 2015; **112**:3439–3444.

27 Drobetsky EA, Sage E. UV-induced G:C-->A:T transitions at the APRT locus of Chinese hamster ovary cells cluster at frequently damaged 5'-TCC-3' sequences. *Mutat Res* 1993; **289**:131–138.

28 Roberts SA, Lawrence MS, Klimczak LJ, Grimm SA, Fargo D, Stojanov P, *et al.* An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers. *Nat Genet* 2013; **45**:970–976.

29 Burns MB, Temiz NA, Harris RS. Evidence for APOBEC3B mutagenesis in multiple human cancers. *Nat Genet* 2013; **45**:977–983.

30 Chan K, Roberts SA, Klimczak LJ, Sterling JF, Saini N, Malc EP, *et al.* An APOBEC3A hypermutation signature is distinguishable from the signature of background mutagenesis by APOBEC3B in human cancers. *Nat Genet* Published Online First: 10 August 2015. doi:10.1038/ng.3378

31 Hodgkinson A, Ladoukakis E, Eyre-Walker A. Cryptic variation in the human mutation rate. *PLoS Biol* 2009; **7**:e1000027.

32 Johnson PLF, Hellmann I. Mutation rate distribution inferred from coincident SNPs and coincident substitutions. *Genome Biol Evol* 2011; **3**:842–850.

33 Roberts SA, Gordenin DA. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat Rev Cancer* 2014; **14**:786–800.

34 Cancer Genome Atlas Research Network. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* 2012; **489**:519–525.

35 Michaelson JJ, Shi Y, Gujral M, Zheng H, Malhotra D, Jin X, *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* 2012; **151**:1431–1442.

36 Duret L, Galtier N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet* 2009; **10**:285–311.

37 Lesecque Y, Mouchiroud D, Duret L. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol Biol Evol* 2013; **30**:1409–1419.

38 Keller I, Bensasson D, Nichols RA. Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLoS Genet* 2007; **3**:e22.

39 Ossowski S, Schneeberger K, Lucas-Lledó JI, Warthmann N, Clark RM, Shaw RG, *et al.* The rate and molecular spectrum of spontaneous mutations in Arabidopsis thaliana. *Science* 2010; **327**:92–94.

40 Agier N, Fischer G. The mutational profile of the yeast genome is shaped by replication. *Mol Biol Evol* 2012; **29**:905–913.

41 Solorzano E, Okamoto K, Datla P, Sung W, Bergeron RD, Thomas WK. Shifting patterns of natural variation in the nuclear genome of caenorhabditis elegans. *BMC Evol Biol* 2011; **11**:168.

42 Koren A, Polak P, Nemesh J, Michaelson JJ, Sebat J, Sunyaev SR, *et al.* Differential relationship of DNA replication timing to different forms of human mutation and variation. *Am J Hum Genet* 2012; **91**:1033–1040.

43 Chen C-L, Rappailles A, Duquenne L, Huvet M, Guilbaud G, Farinelli L, *et al.* Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. *Genome Res* 2010; **20**:447–457.

44 Chan K, Resnick MA, Gordenin DA. The choice of nucleotide inserted opposite abasic sites formed within chromosomal DNA reveals the polymerase activities participating in translesion DNA synthesis. *DNA Repair* 2013; **12**:878–889.

45 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; **455**:105–108.

46 McDonald MJ, Wang W-C, Huang H-D, Leu J-Y. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. *PLoS Biol* 2011; **9**:e1000622.

47 Averof M, Rokas A, Wolfe KH, Sharp PM. Evidence for a high frequency of simultaneous double-nucleotide substitutions. *Science* 2000; **287**:1283–1286.

48 Schrider DR, Hourmozdi JN, Hahn MW. Pervasive multinucleotide mutational events in eukaryotes. *Curr Biol CB* 2011; **21**:1051–1054.

49 Meier B, Cooke SL, Weiss J, Bailly AP, Alexandrov LB, Marshall J, *et al.* C. elegans whole-genome sequencing reveals mutational signatures related to carcinogens and DNA repair deficiency. *Genome Res* 2014; **24**:1624–1636.

50 Zhu W, Cooper DN, Zhao Q, Wang Y, Liu R, Li Q, *et al.* Concurrent Nucleotide Substitution Mutations in the Human Genome are Characterized by a Significantly Decreased Transition/Transversion Ratio. *Hum Mutat* Published Online First: 25 December 2014. doi:10.1002/humu.22749

51 Chen J-M, Férec C, Cooper DN. Patterns and mutational signatures of tandem base substitutions causing human inherited disease. *Hum Mutat* 2013; **34**:1119–1130.

52 Smith NGC, Webster MT, Ellegren H. A low rate of simultaneous double-nucleotide mutations in primates. *Mol Biol Evol* 2003; **20**:47–53.

53 Giannoulatou E, McVean G, Taylor IB, McGowan SJ, Maher GJ, Iqbal Z, *et al.* Contributions of intrinsic mutation rate and selfish selection to levels of de novo HRAS mutations in the paternal germline. *Proc Natl Acad Sci U S A* 2013; **110**:20152–20157.

54 Chen J-M, Cooper DN, Férec C. A new and more accurate estimate of the rate of concurrent tandem-base substitution mutations in the human germline: ∼0.4% of the single-nucleotide substitution mutation rate. *Hum Mutat* 2014; **35**:392–394.

55 Matsuda T, Kawanishi M, Yagi T, Matsui S, Takebe H. Specific tandem GG to TT base substitutions induced by acetaldehyde are due to intra-strand crosslinks between adjacent guanine bases. *Nucleic Acids Res* 1998; **26**:1769–1774.

56 Northam MR, Moore EA, Mertz TM, Binz SK, Stith CM, Stepchenkova EI, *et al.* DNA polymerases ζ and Rev1 mediate error-prone bypass of non-B DNA structures. *Nucleic Acids Res* 2014; **42**:290–306.

57 Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012; **149**:979–993.

58 Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, *et al.* The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* 2014; **42**:D764–770.

59 Taylor J, Schenck I, Blankenberg D, Nekrutenko A. Using galaxy to perform large-scale interactive data analyses. *Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al* 2007; **Chapter 10**:Unit 10.5.

60 Schuster SC, Miller W, Ratan A, Tomsho LP, Giardine B, Kasson LR, *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* 2010; **463**:943–947.

61 Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* 2008; **456**:53–59.

62 Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 2010; **327**:78–81.

63 Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, *et al.* The diploid genome sequence of an individual human. *PLoS Biol* 2007; **5**:e254.

64 Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, *et al.* The diploid genome sequence of an Asian individual. *Nature* 2008; **456**:60–65.

65 Ahn S-M, Kim T-H, Lee S, Kim D, Ghang H, Kim D-S, *et al.* The first Korean genome sequence and analysis: full genome sequencing for a socio-ethnic group. *Genome Res* 2009; **19**:1622–1629.

66 Jordan KW, Carbone MA, Yamamoto A, Morgan TJ, Mackay TFC. Quantitative genomics of locomotor behavior in Drosophila melanogaster. *Genome Biol* 2007; **8**:R172.

67 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; **5**:e310.

68 Tweedie S, Ashburner M, Falls K, Leyland P, McQuilton P, Marygold S, *et al.* FlyBase: enhancing Drosophila Gene Ontology annotations. *Nucleic Acids Res* 2009; **37**:D555–559.

69 Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005; **15**:1034–1050.

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

71 Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol J Comput Mol Cell Biol* 2012; **19**:455–477.

72 Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, *et al.* Genome sequence of the model mushroom Schizophyllum commune. *Nat Biotechnol* 2010; **28**:957–963.

73 Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AFA, Roskin KM, *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 2004; **14**:708–715.

74 Ranwez V, Harispe S, Delsuc F, Douzery EJP. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS One* 2011; **6**:e22594.

75 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; **467**:1061–1073.

76 Hodgkinson A, Eyre-Walker A. Human triallelic sites: evidence for a new mutational mechanism? *Genetics* 2010; **184**:233–241.

77 Bazykin GA, Kondrashov FA, Brudno M, Poliakov A, Dubchak I, Kondrashov AS. Extensive parallelism in protein evolution. *Biol Direct* 2007; **2**:20.

78 Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 1989; **123**:585–595.

79 Yang null, Bielawski null. Statistical methods for detecting molecular adaptation. *Trends Ecol Evol* 2000; **15**:496–503.

80 Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* 2008; **320**:1632–1635.

81 Mallick S, Gnerre S, Muller P, Reich D. The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res* 2009; **19**:922–933.

82 Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol Evol* 2009; **1**:114–118.

83 Bazykin GA, Kondrashov AS. Major role of positive selection in the evolution of conservative segments of Drosophila proteins. *Proc Biol Sci* 2012; **279**:3409–3417.

84 Bazykin GA, Dushoff J, Levin SA, Kondrashov AS. Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites. *Proc Natl Acad Sci U S A* 2006; **103**:19396–19401.

85 Kimura M. The role of compensatory neutral mutations in molecular evolution. *J Genet* 1985; **64**:7–19.

86 Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the Human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* 2014; **10**:e1004525.

87 Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci U S A* 2007; **104**:12410–12415.

88 Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* 2012; **337**:1675–1678.

89 Dermitzakis ET, Reymond A, Antonarakis SE. Conserved non-genic sequences - an unexpected feature of mammalian genomes. *Nat Rev Genet* 2005; **6**:151–157.

90 Andolfatto P. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 2005; **437**:1149–1152.

91 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; **27**:1226–1234.

92 Wang J, Fan HC, Behr B, Quake SR. Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 2012; **150**:402–412.

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

94 Harfe BD, Jinks-Robertson S. DNA polymerase zeta introduces multiple mutations when bypassing spontaneous DNA damage in Saccharomyces cerevisiae. *Mol Cell* 2000; **6**:1491–1499.

95 Sakamoto AN, Stone JE, Kissling GE, McCulloch SD, Pavlov YI, Kunkel TA. Mutator alleles of yeast DNA polymerase zeta. *DNA Repair* 2007; **6**:1829–1838.

96 Shinbrot E, Henninger EE, Weinhold N, Covington KR, Göksenin AY, Schultz N, *et al.* Exonuclease mutations in DNA polymerase epsilon reveal replication strand specific mutation patterns and human origins of replication. *Genome Res* 2014; **24**:1740–1750.

97 Chen C-L, Duquenne L, Audit B, Guilbaud G, Rappailles A, Baker A, *et al.* Replication-associated mutational asymmetry in the human genome. *Mol Biol Evol* 2011; **28**:2327–2337.

98 Baker A, Audit B, Chen C-L, Moindrot B, Leleu A, Guilbaud G, *et al.* Replication fork polarity gradients revealed by megabase-sized U-shaped replication timing domains in human cell lines. *PLoS Comput Biol* 2012; **8**:e1002443.

99 Polak P, Arndt PF. Transcription induces strand-specific mutations at the 5’ end of human genes. *Genome Res* 2008; **18**:1216–1223.

100 Park C, Qian W, Zhang J. Genomic evidence for elevated mutation rates in highly expressed genes. *EMBO Rep* 2012; **13**:1123–1129.

101 Hoang ML, Chen C-H, Sidorenko VS, He J, Dickman KG, Yun BH, *et al.* Mutational signature of aristolochic acid exposure as revealed by whole-exome sequencing. *Sci Transl Med* 2013; **5**:197ra102.

102 Green P, Ewing B, Miller W, Thomas PJ, NISC Comparative Sequencing Program, Green ED. Transcription-associated mutational asymmetry in mammalian evolution. *Nat Genet* 2003; **33**:514–517.

103 Datta A, Jinks-Robertson S. Association of increased spontaneous mutation rates with high levels of transcription in yeast. *Science* 1995; **268**:1616–1619.

104 Kim N, Abdulovic AL, Gealy R, Lippert MJ, Jinks-Robertson S. Transcription-associated mutagenesis in yeast is directly proportional to the level of gene expression and influenced by the direction of DNA replication. *DNA Repair* 2007; **6**:1285–1296.

105 Polak P, Lawrence MS, Haugen E, Stoletzki N, Stojanov P, Thurman RE, *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat Biotechnol* 2014; **32**:71–75.

106 Zhong X, Garg P, Stith CM, Nick McElhinny SA, Kissling GE, Burgers PMJ, *et al.* The fidelity of DNA synthesis by yeast DNA polymerase zeta alone and with accessory proteins. *Nucleic Acids Res* 2006; **34**:4731–4742.

107 Nelson JR, Lawrence CW, Hinkle DC. Thymine-thymine dimer bypass by yeast DNA polymerase zeta. *Science* 1996; **272**:1646–1649.

108 Lee Y-S, Gregory MT, Yang W. Human Pol ζ purified with accessory subunits is active in translesion DNA synthesis and complements Pol η in cisplatin bypass. *Proc Natl Acad Sci U S A* 2014; **111**:2954–2959.

109 Kim N-H, Kim H-J, Kim M, Lee K-W. A novel SOD1 gene mutation in a Korean family with amyotrophic lateral sclerosis. *J Neurol Sci* 2003; **206**:65–69.

110 Morita M, Aoki M, Abe K, Hasegawa T, Sakuma R, Onodera Y, *et al.* A novel two-base mutation in the Cu/Zn superoxide dismutase gene associated with familial amyotrophic lateral sclerosis in Japan. *Neurosci Lett* 1996; **205**:79–82.

111 Baek W, Koh S-H, Park JS, Kim YS, Kim HY, Kwon MJ, *et al.* A novel codon4 mutation (A4F) in the SOD1gene in familial amyotrophic lateral sclerosis. *J Neurol Sci* 2011; **306**:157–159.

112 Chen J, Miller BF, Furano AV. Repair of naturally occurring mismatches can induce mutations in flanking DNA. *eLife* 2014; **3**:e02001.

113 Roberts SA, Gordenin DA. Clustered and genome-wide transient mutagenesis in human cancers: Hypermutation without permanent mutators or loss of fitness. *BioEssays News Rev Mol Cell Dev Biol* Published Online First: 26 February 2014. doi:10.1002/bies.201300140

114 Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol* 2011; **28**:2651–2660.

115 Lujan SA, Clausen AR, Clark AB, MacAlpine HK, MacAlpine DM, Malc EP, *et al.* Heterogeneous polymerase fidelity and mismatch repair bias genome variation and composition. *Genome Res* 2014; **24**:1751–1764.

116 Supek F, Lehner B. Differential DNA mismatch repair underlies mutation rate variation across the human genome. *Nature* Published Online First: 23 February 2015. doi:10.1038/nature14173

117 Kotov IN, Siebring-van Olst E, Knobel PA, van der Meulen-Muileman IH, Felley-Bosco E, van Beusechem VW, *et al.* Whole genome RNAi screens reveal a critical role of REV3 in coping with replication stress. *Mol Oncol* 2014; **8**:1747–1759.

118 Lang GI, Murray AW. Mutation rates across budding yeast chromosome VI are correlated with replication timing. *Genome Biol Evol* 2011; **3**:799–811.

119 Kochenova OV, Daee DL, Mertz TM, Shcherbakova PV. DNA polymerase ζ-dependent lesion bypass in Saccharomyces cerevisiae is accompanied by error-prone copying of long stretches of adjacent DNA. *PLoS Genet* 2015; **11**:e1005110.

120 Northam MR, Robinson HA, Kochenova OV, Shcherbakova PV. Participation of DNA polymerase zeta in replication of undamaged DNA in Saccharomyces cerevisiae. *Genetics* 2010; **184**:27–42.