Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук

На правах рукописи

Клинк Галина Викторовна

Расположение аминокислотных замен на эволюционном дереве как показатель изменчивости однопозиционного адаптивного ландшафта

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание ученой степени кандидата биологических наук

научный руководитель:
д.б.н., профессор
Базыкин Г.А.

Москва – 2020

Оглавление

Введение

Актуальность

Эволюция последовательностей ДНК всех организмов происходит под давлением естественного отбора. Его действие заключается в том, что более приспособленные к условиям среды особи дают больше потомства. Число потомков, оставляемых особью – ее приспособленность – может быть измерена экспериментально, однако такие измерения очень трудоёмки. Отбор действует на фенотип, но эволюция происходит за счёт изменения генотипа. Функция, связывающая генотип и приспособленность его обладателя, называется адаптивным ландшафтом, или ландшафтом приспособленности. Полный адаптивный ландшафт отражает приспособленность организма при всех сочетаниях нуклеотидов во всех позициях генома. Зная эту функцию, можно не только понять законы эволюции в прошлом и настоящем, но и предугадывать её возможные пути в будущем. Понимание устройства адаптивных ландшафтов может быть очень важно для определения эффекта мутаций у человека, изучения динамики эволюции патогенов, оценки состояния популяций, поиска функциональных взаимодействий между сайтами белков. Однако ландшафт приспособленности многомерен: его размерность равна числу позиций в рассматриваемой последовательности. Вместо полного адаптивного ландшафта можно изучать его срезы. При этом часть информации теряется, но снижается число возможных сочетаний вариантов. Но даже ландшафт для аминокислотных позиций одного белка длиной в 100 аминокислот имеет размерность 100 и содержит приспособленности для 20^{100} возможных вариантов последовательности. Получение и изучение таких ландшафтов – непосильная задача для существующих методов и вычислительных мощностей.

Минимальный адаптивный ландшафт, который имеет смысл рассматривать в отдельности – это однопозиционный адаптивный ландшафт – ОПАЛ [5]. ОПАЛ каждого сайта можно представить в виде вектора приспособленности длиной 20, где для каждой аминокислоты указано, какой эффект она окажет на приспособленность, если будет стоять в этом сайте. Будем называть этот эффект приспособленностью аминокислоты в данном сайте. ОПАЛ сайта может меняться во времени из-за изменений в его геномном контексте: приспособленность одной и той же аминокислоты в сайте может зависеть от того, какие аминокислоты стоят в других позициях того же или другого белка. Тогда говорят, что аминокислотные позиции эпистатически взаимодействуют. Также ОПАЛ сайта, как и адаптивный ландшафт любой размерности, зависит от факторов среды, в которой функционирует белок [5].

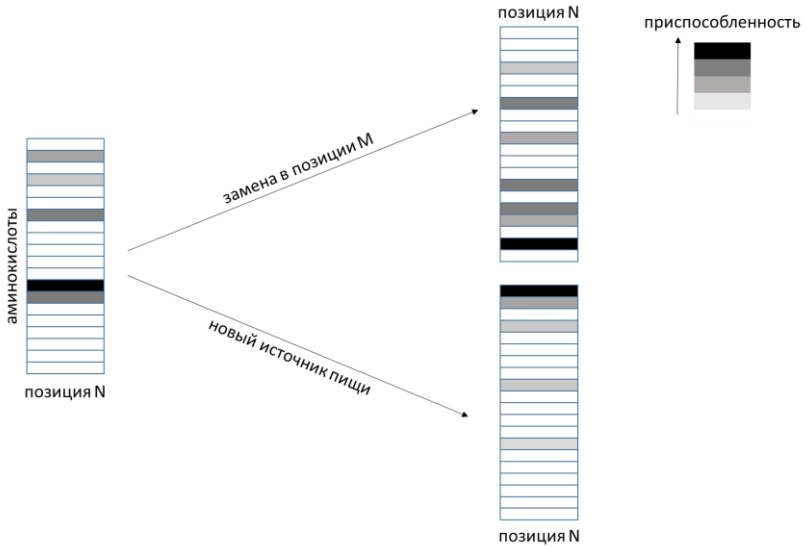


Рисунок 1. *Однопозиционный адаптивный ландшафт (ОПАЛ) может меняться во времени из-за изменения геномного контекста или условий среды.*

Насколько сходны ОПАЛы одних и тех же сайтов у разных видов? Насколько точно можно предсказать ОПАЛ сайта для одного вида живых существ по ОПАЛу для другого вида? Изучение изменчивости ОПАЛов в молекулярной эволюции – ключ к пониманию поведения сложных, многомерных адаптивных ландшафтов.

Для некоторых белков бактерий, вирусов и дрожжей методами насыщающего мутагенеза показано влияние каждой аминокислотной замены в каждом сайте на приспособленность организма [16, 25, 73]. Сравнение полученных на основании таких экспериментов матриц предпочитаемых аминокислот для гомологичных белков в разных видах позволяет сказать, насколько разнятся их однопозиционные адаптивные ландшафты, и определить сайты, в которых они изменяются [12].

Однако изучение изменения приспособленности аминокислот экспериментальными методами возможно не для всех организмов. Кроме того, отбор, проводящийся в условиях лаборатории, в некоторых аспектах отличается от отбора в природе [25]. Разработка методов изучения ОПАЛов, основанных на анализе данных секвенирования, необходима для ответа на фундаментальные вопросы об эволюции белков, а также для изучения ландшафтов приспособленности белков человека и его патогенов.

Вирусы – самые быстро эволюционирующие патогены. Аминокислотные сайты вирусных белков эволюционируют с разной скоростью. Ранее было показано, что скорость эволюции в пределах сайта также может быть непостоянной [4]. Большой проблемой для лечения вирусных заболеваний является приобретение этими патогенами устойчивости к противовирусным препаратам [69]. Появление в вирусной популяции отбора на устойчивость к лекарствам меняет ОПАЛы сайтов вирусных белков [7, 19]. Поиск и изучение аминокислотных сайтов с переменным ландшафтом приспособленности в вирусных белках позволит глубже понять механизмы приобретения вирусами устойчивости к иммунитету и лекарствам.

Данная работа посвящена изучению непостоянства однопозиционных адаптивных ландшафтов филогенетическими методами.

Цели и задачи

Цель исследования – оценить роль изменчивости однопозиционных адаптивных ландшафтов в эволюции белков. Для достижения этой цели поставлены задачи:

1. Оценить изменчивость ОПАЛов митохондриальных белков с помощью анализа расположения аминокислотных замен на филогенетических деревьях.

2. Изучить, как меняется приспособленность референтных, нейтральных и патогенных аллелей митохондриальных белков человека с увеличением филогенетического расстояния от него.

3. Разработать метод поиска аминокислот, приспособленность которых неодинакова у разных видов. С его помощью найти сайты с переменным адаптивным ландшафтом в вирусных белках.

Новизна и практическая значимость

Появление одной и той же аминокислоты в конкретном сайте белка может неодинаково отразиться на его функциональности в разных организмах. В таких случаях говорят, что ОПАЛ сайта изменчив [5]. Эволюционная роль этого явления подтверждается с помощью многих методов. Например, показано, что замены на определённую аминокислоту (гомоплазии) распределены на филогенетическом дереве неравномерно. Действительно, частое возникновение определённой аминокислоты в сайте белка в пределах одной клады и отсутствие замен на неё в этом же сайте в другой кладе может говорить о том, что в первой кладе приспособленность этой аминокислоты выше. Кластеризация на филогении показана для разных типов гомоплазий: реверсий, конвергентных и параллельных замен [22, 63, 68, 71, 98]. В отличие от существующих методик анализа взаиморасположения аминокислотных замен, разработанный нами метод

поиска изменения ОПАЛов [37] устойчив к кластеризации гомоплазий в близких видах, не связанной с изменением адаптивного ландшафта.

Ранее было показано, что вероятность наблюдать патогенную для человека аминокислоту как аллель дикого типа в другом биологическом виде либо не зависит от эволюционного расстояния до линии человека [40, 79], либо возрастает с этим расстоянием [32]. Мы впервые показали, что в митохондриальных белках замены на известные патогенные для человека варианты могут кластеризоваться в видах, эволюционно близких к человеку [38]. Это наблюдение может помочь меньше ошибаться в интерпретации филогенетической информации при предсказании эффектов мутаций для человека и других организмов.

В данной работе мы не только изучили степень непостоянства однопозиционных адаптивных ландшафтов статистически, но и разработали метод поиска отдельных аминокислот, меняющих приспособленность между видами, основанный на анализе данных секвенирования. С помощью этого метода можно изучать и сравнивать ОПАЛы отдельных сайтов. Также метод может помочь в определении эффекта мутаций у видов, для которых невозможно изучать этот вопрос экспериментально.

С помощью нового метода мы нашли для поверхностного белка вируса иммунодефицита человека 1 (ВИЧ1) аминокислоты, меняющие приспособленность между вирусными подтипами A, B и C. Результаты выложены в открытый доступ в виде веб-сервиса (http://makarich.fbb.msu.ru/galkaklink/hiv_landscape).

Таким образом, в диссертации решается важная для эволюционной биологии задача – оценка роли непостоянства однопозиционных адаптивных ландшафтов в эволюции белков.

Основные результаты и положения, выносимые на защиту

1. В митохондриальных белках видов *Opisthokonta* филогенетическое расстояние между параллельными заменами в среднем на 20% меньше, чем филогенетическое расстояние между дивергентными заменами. Это характеризует степень непостоянства однопозиционных адаптивных ландшафтов в эволюции митохондриальных белков.

2. Виды, в которых происходят замены на аминокислоты, представляющие референтные, минорные нейтральные и патогенные аллели в митохондриальных белках человека, более близко родственны ему, чем виды, в которых происходят замены на аминокислоты, не встречающиеся у человека.

3. Информация о взаимном расположении аминокислотных замен на филогенетическом дереве позволяет находить аминокислоты, меняющие приспособленность между кладами. Это может быть полезным в поиске причин различий в особенностях заражения, течения инфекции и ответа на лечение между разными штаммами одного вируса.

Объекты и методы

В работе изучены ОПАЛы кодируемых и работающих в митохондриях белков *Opisthokonta*, а также ОПАЛы поверхностных белков gp160 вируса иммунодефицита человека (ВИЧ-1) и гемагглютинина вируса гриппа человека А. Эти объекты выбраны, так как последовательности генов митохондриальных и вирусных белков прочитаны для тысяч видов/штаммов, что необходимо для применения наших методик изучения ОПАЛов. Кроме того, для указанных вирусных белков опубликованы экспериментально измеренные аминокислотные предпочтения в некоторых штаммах, что позволило нам сравнить наши результаты с экспериментальными.

В работе использованы как стандартные методы работы с данными секвенирования (для выравнивания последовательностей, реконструкции филогенетических деревьев и поиска положительного отбора), так и оригинальные методы анализа данных для поиска непостоянства ОПАЛов.

Апробация работы и публикации по теме диссертации

Результаты и основные положения работы доложены на международных конференциях SMBE-2014, SMBE-2015, SMBE-2017, SMBE-2018, российской конференции ИтиС-2015, а также на семинаре группы эволюционной геномики австрийского Института науки и технологии (г. Клостернойбург, Австрия, Австрия, 2018).

По теме диссертации опубликовано 2 работы в рецензируемых международных научных журналах, индексируемых в базах данных Scopus и Web of Science.

Работа по теме диссертации была поддержана грантом 18-34-00358 мол_а (руководитель Г.В.Клинк).

Публикации в изданиях из перечня ВАК

1) Klink, G. V., Bazykin, G. A. Parallel evolution of metazoan mitochondrial proteins / G. V. Klink, G. A. Bazykin // Genome biology and evolution. – 2017. – Vol. 9. – Is. 5. – P. 1341–1350

2) Klink, G. V., Golovin, A. V., Bazykin, G. A. Substitutions into amino acids that are pathogenic in human mitochondrial proteins are more frequent in lineages closely related to human than in distant lineages/ G. V. Klink, A. V. Golovin, G. A. Bazykin // PeerJ. – 2017. – Vol. 12. – e4143.

Тезисы конференций

Klink G.V., Bazykin G.A. Inference of prevalence of epistasis from huge phylogenies / G. V. Klink, G. A. Bazykin // Society for Molecular Biology & Evolution (SMBE 2014), Сан-Хуан, Пуэрто-Рико, 2014.

Klink G.V., Bazykin G.A. Analysis of prevalence of epistasis on the basis of huge phylogenies / G. V. Klink, G. A. Bazykin // Society for Molecular Biology & Evolution (SMBE 2015), Вена, Австрия, 2015.

Klink G.V., Bazykin G.A. Analysis of prevalence of epistasis on the basis of huge phylogenies / G. V. Klink, G. A. Bazykin // Информационные технологии и системы (ИТИС 2015), Сочи, Россия, 2015.

Klink G.V., Bazykin G.A. Inference of changes of fitness landscape from sequence data with single-position resolution / G. V. Klink, G. A. Bazykin // Society for Molecular Biology & Evolution (SMBE 2017), Остин, США, 2017.

Klink G.V., Bazykin G.A. Inference of changes of HIV-1 gp160 protein fitness landscape from sequence data with single-position resolution / G. V. Klink, G. A. Bazykin // Society for Molecular Biology & Evolution (SMBE 2018), Йокогама, Япония, 2018.

Структура и объём работы

Диссертация состоит из введения, 4 глав, выводов и списка литературы, состоящего из 99 источников. Содержательная часть работы изложена на 115 страницах текста, включает 25 рисунков, 2 таблицы и 14 приложений.

Глава 1. Обзор литературы

1.1. Свидетельства изменчивости однопозиционных адаптивных ландшафтов в эволюции белков

Анализ последовательностей геномов разных организмов с использованием филогенетических деревьев позволил выявить важную роль непостоянства однопозиционных ландшафтов приспособленности в эволюции белков. Ниже представлены свойства процесса эволюции белков, которые интерпретируют как результат вариабельности ОПАЛов.

1.1.1. Продолжительный положительный отбор на дереве жизни

Поддержание положительного отбора в сайте в течение долгого времени говорит о динамичности его ОПАЛа. Когда ОПАЛ неизменен, на больших временах увеличивающие и снижающие приспособленность замены в сайте происходят с одинаковой скоростью и уравновешивают друг друга: за слабовредной заменой, которая произошла в результате генетического дрейфа, будет следовать замена, восстанавливающая приспособленность. Но если ОПАЛ постоянно меняется, оптимальные аминокислоты становятся неоптимальными, и процесс адаптации преобладает над генетическим дрейфом [61].

Доля аминокислотных замен, происходящих под положительным отбором на линии *Drosophila melanogaster* после дивергенции с *Drosophila virilis* 60 миллионов лет назад, оставалась неизменной и оценивается в 20 – 50% [6]. Свойства положительного отбора у *D.melanogaster* хорошо описываются моделью, включающей непрерывное изменение ландшафта приспособленности [61]. Однако продолжительный положительный отбор характерен не для всех филогенетических групп. Например, на линии *Homo sapiens* доля произошедших под положительным отбором аминокислотных

замен упала с 50% до 0% после разделения подсемейств Понгины и Гоминины [6].

1.1.2. Продолжительное расхождение последовательностей гомологичных белков

Если векторы приспособленности аминокислот в каждом сайте белка неизменны во времени, то аминокислотные различия гомологичных белковых последовательностей должны достигать асимптотического уровня со скоростью порядка скорости расхождения нейтрально эволюционирующих последовательностей [39]. Однако даже самые древние белки продолжают накапливать различия в своих последовательностях в разных видах организмов. Это наблюдение можно объяснить, приняв условие, что набор доступных для сайта аминокислот меняется с течением времени и неодинаков в разных видах [68]. В пользу этого говорит и наблюдение, что в одних и тех же белковых сайтах скорость аминокислотных замен между близкими видами ниже, чем ожидается из числа аминокислот, встречающихся в этих сайтах при захвате более далёких эволюционных расстояний [9].

1.1.3. Скоррелированность замен в разных сайтах

Если замена аминокислоты в сайте белка меняет аминокислотные предпочтения в другом сайте того же или другого белка, то говорят, что эти сайты связаны эпистатическими взаимодействиями. Таким образом, замена в одном сайте может вызвать изменение ОПАЛов и последующие замены в эпистатически связанных с ним сайтах. На геномах мух рода *Drosophila* было показано, что вероятность аминокислотной замены в сайте белка значительно возрастает, если недавно произошла замена аминокислоты в пределах десяти сайтов от него. Поскольку близкие по последовательности сайты, как правило, находятся рядом и в трёхмерной структуре белка, их

ОПАЛы могут быть взаимозависимыми. Авторы статьи показали, что замены, произошедшие в одной линии на расстоянии в пределах 10 кодонов, часто имеют противоположный эффект на заряд белка. Тем самым, вторая замена компенсирует возникшее в результате первой изменение заряда [10]. Скоррелированными по времени могут быть и замены в сайтах разных белков, что показано для поверхностных белков вируса гриппа А (H1N1 и H3N2) – гемагглютинина и нейраминидазы [64].

1.1.4. Неодинаковый эффект мутации у разных организмов

Непостоянство ОПАЛов может выражаться и в том, что одна и та же аминокислота в сайте белка в одних видах вызывает болезни, а в других представляет собой дикий тип. Поскольку предполагается, что патогенная в одних видах аминокислота становится предпочтительной в других видах за счёт разрешающих (предшествующих) или компенсирующих (последующих) замен в геноме, такие случаи называют компенсированными патогенными отклонениями [32, 40, 79]. Какова скорость такой компенсации и как она изменяется с филогенетическим расстоянием, пока не полностью понятно. По некоторым данным, вероятность замены на патогенный вариант человека неизменна с филогенетическим расстоянием от человеческой ветви и равна 10% [40]. Другое исследование, основанное на распределении эволюционных расстояний от ветви человека до ближайшего вида, в котором патогенная для человека аминокислота является диким типом, говорит, что вероятность замены на патогенную для вида аминокислоту возрастает с филогенетическим расстоянием от него. При этом для большинства патогенных аминокислот нужна всего одна компенсирующая замена, чтобы стать разрешёнными [32].

1.2. Взаиморасположение замен в сайте белка на филогенетическом дереве как отражение адаптивного ландшафта

Вероятность замены одной аминокислоты на другую монотонно зависит от выигрыша в приспособленности, который даёт новая аминокислота по сравнению с текущей, поэтому частоты разных аминокислотных замен, происходивших в сайте в процессе эволюции, отражают относительное положение аминокислот в ОПАЛе. Следовательно, отличия в паттернах аминокислотных замен в сайте между разными участками филогенетического дерева могут свидетельствовать о непостоянстве ОПАЛа между этими участками.

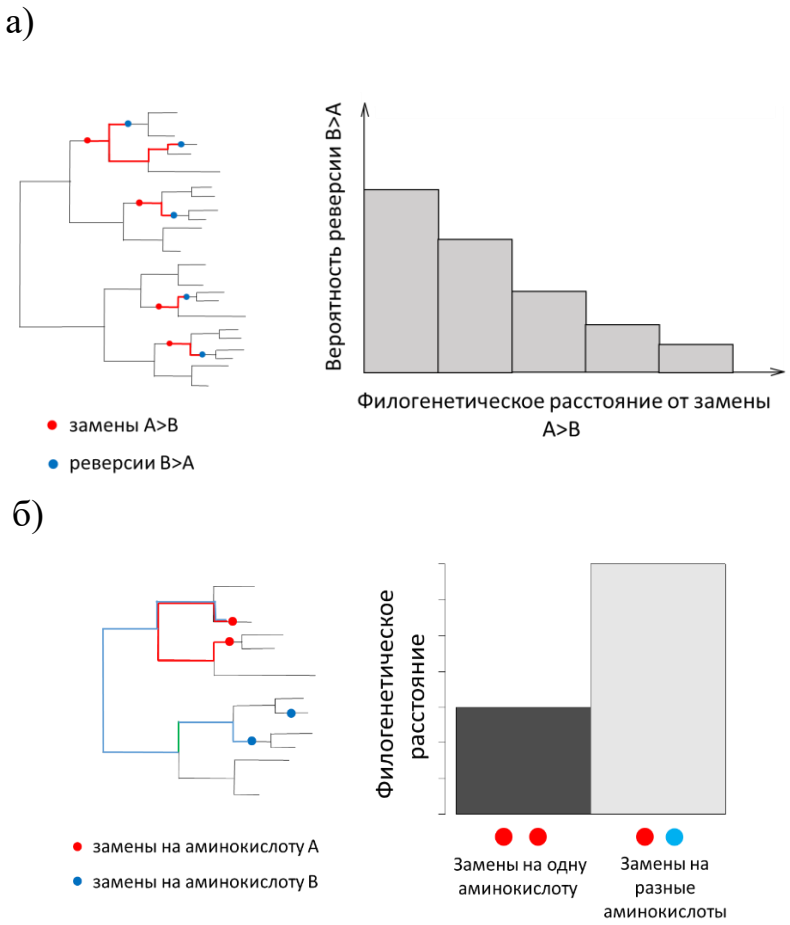


Рисунок 2. *Проявление вариабельности ОПАЛов в распределении замен на филогенетическом дереве.*

а) – С филогенетическим расстоянием от замены A⟶B вероятность реверсии (замены, возвращающей первоначальную аминокислоту) B⟶A снижается
б) – Филогенетическое расстояние между заменами на одну и ту же аминокислоту меньше, чем расстояние между заменами на разные аминокислоты

После аминокислотной замены вероятность реверсии (обратной замены) снижается со временем, как и вероятность встретить предковую аминокислоту в качестве полиморфизма (Рисунок 2, а). Это происходит из-за параллельного снижения приспособленности предкового варианта и повышения приспособленности аминокислоты, на которую произошла замена [63].

Замены на одну и ту же аминокислоту у разных видов в одном сайте белка называются гомоплазиями. Если такие замены происходят из разных предковых аминокислот, они называются конвергентными, если из одинаковой – параллельными. Замены, не являющиеся гомоплазиями, называют дивергентными. Несколько научных групп с помощью разных методов показали, что гомоплазии чаще происходят в эволюционно более близких видах (Рисунок 2, б). Рогозин с коллегами (2008) увидели, что параллельные замены между двумя кладами на филогенетическом дереве эукариот чаще происходят на более близких глубоких предковых ветвях, чем на более далёких терминальных [71]. Гольдштейн и коллеги (2015) показали, что в митохондриальных белках позвоночных отношение частот конвергентных и дивергентных замен снижается с эволюционным расстоянием [22] (Рисунок 3). Зу и Жанг (2015), анализируя белки млекопитающих и насекомых, выяснили, что отношение реального числа гомоплазий к ожидаемому при неизменном ландшафте приспособленности снижается с филогенетическим расстоянием между видами [98].

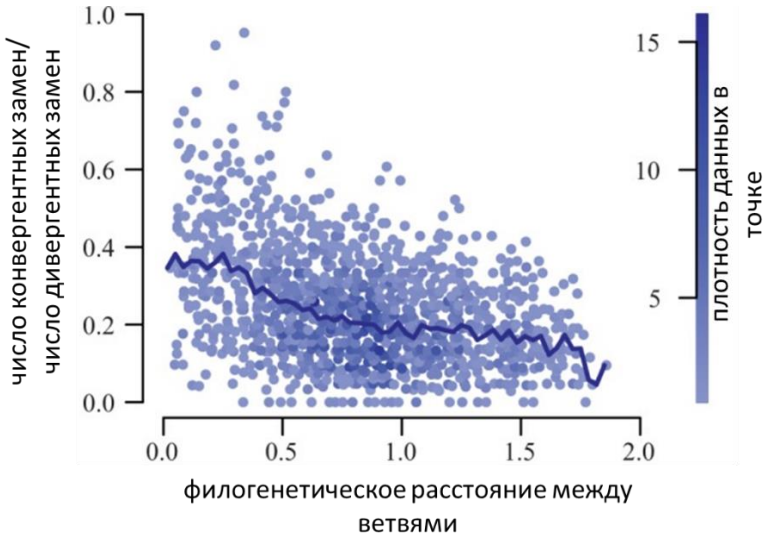


Рисунок 3. *Отношение числа конвергентных и дивергентных замен на двух ветвях филогенетического дерева падает с эволюционным расстоянием между ветвями. Перекрывающиеся точки слиты, цвет определяет число объединённых точек в одну в соответствии со шкалой справа. Филогенетическое расстояние измерено как среднее число аминокислотных замен на сайт. Синяя линия показывает среднее отношение для окон длиной 0.03* [22].

Для поиска изменения скоростей замен на разные аминокислоты в сайте белка необходимо иметь гомологичные белковые последовательности из большого числа видов на разных эволюционных расстояниях. В связи с развитием методов высокопроизводительного секвенирования, такие данные сейчас можно найти для некоторых белков, например, белков клеточных органелл [9] и вирусов [77]. Имеющиеся сейчас методы множественного выравнивания последовательностей и реконструкции филогенетических деревьев позволяют работать с наборами белков из тысяч видов организмов.

1.3. Исследование адаптивных ландшафтов и поиск сайтов с изменчивыми ОПАЛами

Выше описано, как разным научным группам удалось увидеть, что непостоянство ОПАЛов играет важную роль в эволюции белков. Однако эти методы, основанные на суммарных статистиках, не подходят для систематического поиска конкретных сайтов с непостоянными ОПАЛами.

Определение таких сайтов важно не только для понимания динамики изменения ландшафтов приспособленности, но и для более точного предсказания эффектов мутаций у человека и других организмов. Существует несколько аналитических и экспериментальных методов, позволяющих определять сайты с непостоянными ОПАЛами.

1.3.1. Экспериментальные методы

Экспериментальное определение адаптивного ландшафта основано на мутагенезе с последующим измерением приспособленности мутантных последовательностей. Поскольку невозможно рассмотреть пространство вариантов последовательности всего белка, исследователи могут делать упор либо на «ширину» экспериментального ландшафта, стараясь определить ОПАЛы всех позиций белка при неизменном контексте других сайтов, либо на «глубину» ландшафта, исследуя все возможные сочетания аминокислот в небольшом числе сайтов.

Одним из экспериментальных способов определения адаптивных ландшафтов белков является глубокое мутационное сканирование (deep mutational scanning) [20, 21]. Этот метод заключается в том, что полученные случайным мутагенезом или химическим синтезом варианты изучаемой последовательности подвергают отбору в условиях конкуренции (Рисунок 4). Далее с помощью глубокого секвенирования определяются частоты каждой мутантной последовательности до и после отбора. Приспособленность последовательности считается пропорциональной разнице этих частот [25]. Таким образом, в одном эксперименте определяется относительная приспособленность всех мутантных вариантов. С помощью этого метода изучены эффекты одиночных мутаций в аминокислотных позициях некоторых белков [28, 54, 55, 21], в том числе поверхностных белков вирусов ВИЧ [24, 25] и гриппа А [12, 44]. Зная

относительные приспособленности аминокислот в сайтах белка для двух организмов, можно оценить, насколько разнятся ОПАЛы каждой позиции белка между этими организмами.

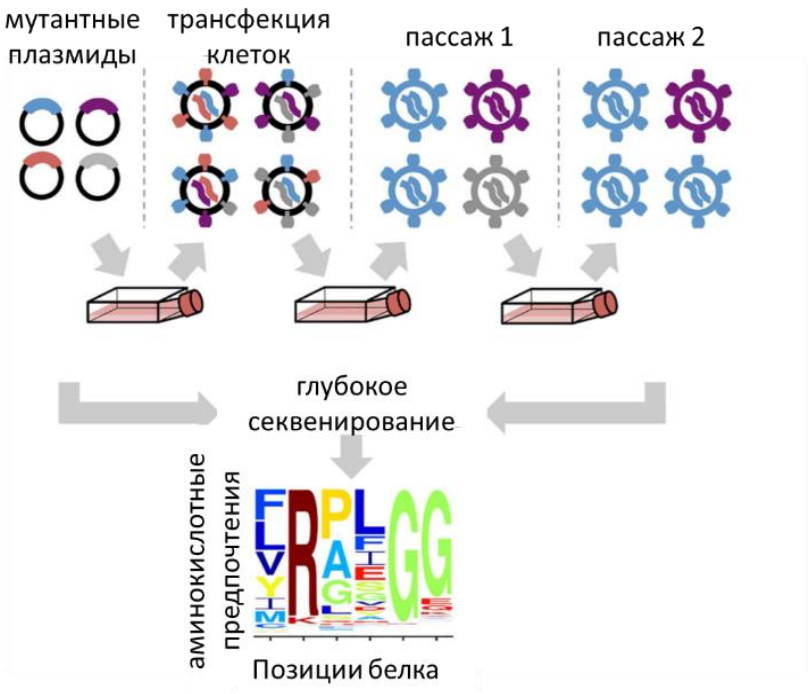


Рисунок 4. *Глубокое мутационное сканирование белка оболочки ВИЧ1 – схема эксперимента. Библиотеки провирусных плазмид создают путём случайных мутаций гена поверхностного белка. Плазмиды трансфецируют в компетентные клетки. Далее отбирают функциональные варианты путём двух пассажей образующихся вирусных частиц в новые клетки. С помощью секвенирования определяют относительное содержание каждого варианта в вирусной популяции до и после отбора. По изменению распространённости аминокислотных вариантов вычисляют их относительные приспособленности в каждом сайте. Размер символов для аминокислот в каждой позиции белка в матрице аминокислотных предпочтений отражает их экспериментально измеренную относительную приспособленность в этой позиции* [25].

Недавно такое сравнение было проведено для белка оболочки ВИЧ (gp160) двух относительно близких штаммов, принадлежащих одному вирусному подтипу. Позиции gp160 двух штаммов идентичны на 86%. Исследователи нашли 30 сайтов (4.5% от всех рассмотренных), имеющих достоверно отличные (ожидаемая доля ложных отклонений < 0.1) аминокислотные предпочтения в двух штаммах. В основном это сайты, где только одна аминокислота имеет высокую приспособленность в штамме с менее стабильным gp160, а в штамме с более стабильным белком одинаково

высокой приспособленностью обладают несколько аминокислот. Найденные сайты оказались близко расположенными в трёхмерной структуре белка [24].

Та же научная группа сравнила аминокислотные предпочтения в сайтах гемагглютинина из двух эволюционно далёких штаммов вируса гриппа А, принадлежащих разным подтипам (H1N1 и H3N2). Только 42% исследуемых аминокислотных позиций оказались идентичными в двух штаммах, и для многих сайтов экспериментальные аминокислотные предпочтения сильно различались. Экспериментально измеренные относительные приспособленности аминокислот для H1 коррелировали с встречаемостью аминокислот в изолятах H3N2 хуже, чем экспериментально измеренные приспособленности для H3 (коэффициент корреляции Спирмена r=0,17, p=0.002 для H1 и r=0,24, p<0.0001 для H3). Для некоторых сайтов изменения ОПАЛов оказались связаны с различиями в структуре белка между штаммами [44].

Глубокое мутационное сканирование используют и при анализе двойных мутантов, чтобы найти эпистатически взаимодействующие пары сайтов [55, 65, 75]. Недавняя работа показала, что полученная в таких экспериментах информация о парных эпистатических взаимодействиях позволяет точно определять пространственные структуры белков [72].

Саркисян и коллеги (2016) построили адаптивный ландшафт зелёного флуоресцентного белка (green fluorescent protein, GFP) из медузы *Aequorea 21erevisi*, используя в качестве меры приспособленности силу флуоресценции. С помощью случайного мутагенеза исследователи получили 51715 аминокислотных последовательностей, содержащих в среднем 3.7 мутаций, отличающих их от последовательности дикого типа. 75% одиночных мутаций снижало свечение белка, при этом 9.4% мутаций имели сильный эффект и снижали флуоресценцию более, чем в 5 раз. Среди аминокислот с сильным отрицательным эффектом 10% зафиксированы в

других видах, что говорит о непостоянстве ОПАЛов позиций, в которых возникли эти мутации [76].

Экспериментальное исследование всевозможных сочетаний аминокислот во всех сайтах белка невозможно. Можно сузить круг рассматриваемых позиций и мутаций при изучении адаптивных ландшафтов, исследуя всевозможные сочетания аминокислот в небольшом функциональном участке белка. Так была показана важная роль эпистатических взаимодействий четырёх позиций протеинкиназы PhoQ бактерии *Escherichia coli* в её связывании с лигандом [66]. Другой способ ограничения числа изучаемых сайтов – вносить в сайты белков только мутации на аминокислоты, стоящие в этих сайтах у группы близких видов [23, 41, 67]. Этот подход является биологически оправданным при изучении роли вариабельности адаптивных ландшафтов в естественной эволюции белков. Так был исследован адаптивный ландшафт белка His3 дрожжей: в белок в разных сочетаниях ставили аминокислотные варианты, встречающиеся среди дрожжей двадцати одного вида, чей общий предок жил 4 миллиона лет назад. Мутантные гены вводились в штамм *Saccharomyces 22erevisiae* с выключенным геном для His3, и измерялась скорость роста таких дрожжей, которая служила мерой приспособленности. Приспособленность 85% аминокислот, встречающихся у рассмотренных видов, зависела от аминокислот в других сайтах. Причём в 67% сайтов белка одна и та же мутация могла иметь положительный или отрицательный эффект на приспособленность в зависимости от контекста других сайтов (это явление называется знаковым эпистазом) [67]. Таким образом, большинство сайтов белка His3 имеют разные ОПАЛы у близких видов дрожжей.

С помощью экспериментальных подходов была показана важная роль непостоянства ОПАЛов в достижении устойчивости к антибиотикам [91] и противовирусным препаратам [52].

Экспериментальные системы позволяют измерять приспособленность напрямую, что может давать более явный результат и позволяет получить более полные ОПАЛы, чем анализ возникших в процессе естественной эволюции последовательностей, где достоверность оценки приспособленности сильно зависит от количества и информативности данных. Однако экспериментальные методы применимы не для всех организмов и очень трудоёмки. Кроме того, условия отбора могут отличаться в природе и лаборатории, поэтому экспериментально измеренные и существующие в природе ОПАЛы могут не совпадать даже при одинаковом геномном контексте[8, 25, 27].

1.3.2. Построение адаптивных ландшафтов белков на основе встречаемости аминокислот в множественном выравнивании

При достаточном количестве данных о гомологичных белковых последовательностях в разных организмах, можно частично сконструировать адаптивный ландшафт белка по встречаемости разных аминокислот в его сайтах. Получающийся в результате анализа множественного выравнивания аминокислотных последовательностей ландшафт – «ландшафт распространения» («prevalence landscape») – отражает вероятность любой заданной последовательности быть встреченной в множественном выравнивании природных образцов [51]. В наиболее простой модели аминокислоты из разных сайтов вносят независимые вклады в вероятность наблюдать последовательность. Но есть методы, позволяющие учитывать при построении такого ландшафта совместную встречаемость аминокислот в парах сайтов, отражающую попарные эпистатические взаимодействия [14, 15, 30, 48, 51, 53, 89]. С помощью этих методов можно искать пары сайтов с взаимозависимыми ОПАЛами.

Фиглиуцци с коллегами (2016) разработали способ оценки эффекта конкретных мутаций на функцию белка, основанный на одном из таких методов – анализе прямых взаимодействий. Разница вклада дикого и мутантного аллелей в активность белка при сохранении остальной его последовательности неизменной определяется разницей значений специальных статистик исходной и мутантной последовательностей. В неэпистатической модели значение статистики для последовательности пропорционально суммарной встречаемости её аминокислот в множественном выравнивании. Для учета вклада попарных эпистатических взаимодействий в фенотип в модель включают совместную встречаемость каждой пары аминокислот в выравнивании. Метод был применён к мутациям бета-лактамазы *Escherichia coli*, и эпистатическая модель показала более высокую корреляцию с экспериментально измеренными фенотипическими эффектами, чем модель с независимыми сайтами [14, 51, 53]. Теоретически таким же образом можно определить вклад одной и той же аминокислоты в фенотип при разных аминокислотах в других позициях белка.

Ландшафты распространённости аминокислот (prevalence landscapes), учитывающие попарные взаимодействия сайтов, построены для некоторых белков ВИЧ [14, 51, 53], в том числе для поверхностного белка ВИЧ – gp160. Этот белок является самым длинным и вариабельным белком вируса, что осложняет вычисления параметров модели. Относительные приспособленности, предсказанные с помощью ландшафтов предпочтения для вирусных штаммов из подтипа B, хорошо коррелировали с экспериментально измеренными. Также авторы статьи определили для пар сайтов скоррелированность встречаемости в них разных аминокислот. Оказалось, что большинство наиболее связанных по этому критерию пар сайтов находятся в прямом контакте в пространственной структуре белка [51].

Как правило, включение в эволюционную модель попарных эпистатических взаимодействий сайтов улучшает предсказание эффектов мутаций, что является ещё одним подтверждением изменчивости ОПАЛов [15, 30, 51]. Однако, расширение существующих моделей до взаимодействий более высоких порядков невозможно из-за огромного числа параметров. В 2018 г. разработан метод предсказания эффектов мутаций с помощью нелинейной генеративной модели со скрытой переменной, позволяющий найти аминокислоты с разным эффектом в разных контекстах. Скрытая переменная (то есть, переменная, которую нельзя измерить напрямую) учитывает эпистатические взаимодействия высоких порядков, относительная приспособленность мутантной последовательности находится с помощью нейронной сети [70].

Поиск ОПАЛов, основанный на встречаемости аминокислот в выравнивании, требует применения явных моделей, которые не могут учитывать все возможные причины вариабельности однопозиционных адаптивных ландшафтов. ОПАЛ сайта может изменяться не только за счёт замен в сайтах того же белка, но и за счёт изменений в других участках генома. Однако определение ландшафта распространённости с учётом всех позиций затруднительно из-за большого объема вычислений. Кроме того, в разных условиях аминокислота может иметь разную приспособленность вне зависимости от контекста, что также не учитывается при поиске ОПАЛов по множественным выравниваниям.

1.3.3. Поиск сайтов с неодинаковым ОПАЛом в двух кладах филогенетическим методом

Используя филогенетический подход, Тамури и коллеги (2009) нашли в белках вируса гриппа А позиции с разными ОПАЛами между штаммами, заражающими птиц и человека. Для каждого аминокислотного сайта авторы сравнивали правдоподобие двух моделей: гомогенной с

независимым от типа хозяина вектором приспособленности аминокислот и негомогенной, в которой вирусы из двух клад имели разные сайт-специфические векторы приспособленности аминокислот. Большинство сайтов, для которых изменение ландшафта приспособленности показано экспериментально, удалось найти и этим методом [84].

Основным ограничением этого метода является необходимость указать на филогении точку, разделяющую клады с разными ОПАЛами, что возможно далеко не всегда. Также сравнение двух моделей является время затратным.

Заключение

На данный момент многие научные группы с помощью экспериментальных и аналитических методов нашли свидетельства вариабельности однопозиционных ландшафтов приспособленности. По-видимому, это явление играет важную роль в эволюции белков, но многие его аспекты остаются не полностью понятными. В частности: в каких именно сайтах и с какой скоростью меняются приспособленности аминокислот, как между белками варьирует изменчивость однопозиционных ландшафтов?

Поиск изменения ОПАЛов конкретных сайтов в разных геномных и средовых контекстах можно проводить как с помощью мутационных экспериментов, так и на основании множественного выравнивания последовательностей или филогении с помощью моделей, учитывающих эти контексты. Экспериментальное определение эффектов мутаций на приспособленность возможно для ограниченного числа систем. Модели, учитывающие контекст при анализе последовательностей, требуют определения значений большого числа параметров и сложных вычислений. Разработка метода поиска сайтов с переменными ОПАЛами на основании

анализа белковых последовательностей без явной модели, не требующего определения параметров, является важной и пока не решённой задачей. В виду растущего объема данных геномного секвенирования такой метод может стать широко применимым в решении как фундаментальных, так и практических задач.

Глава 2. Исследование степени вариабельности однолокусных адаптивных ландшафтов митохондриальных белков *Metazoa* и *Opisthokonta* с помощью анализа взаиморасположения аминокислотных замен на филогенетическом дереве

В этой главе описан наш филогенетический метод оценки роли изменения ОПАЛов в эволюции белков и его применение к митохондриальным белкам *Matazoa* и *Opisthokonta*.

2.1. Материалы и методы

2.1.1. Выравнивание последовательностей и построение филогении

Мы использовали множественные аминокислотные выравнивания 12 митохондриально-кодируемых белков *Metazoa*, опубликованные в статье [9], взяв в анализ 77% сайтов – все, где аминокислота определена более, чем в 99% видов. Поскольку для имеющейся подвыборки видов нет общепринятого филогенетического дерева, мы использовали гибридный подход для его реконструкции, взяв топологию дерева из базы данных «ITOL» [interactive tree of life; 46] и используя в дальнейшем анализе только виды, которые были на тот момент в базе данных. После фильтрации мы располагали последовательностями каждого белка более, чем для 900 видов (Таблица 1). Для разрешения имевшихся мультифуркаций и оценки длин ветвей мы использовали программу «RAxML 8.0.0» [81] с моделью «GTRGAMMA». Предковые состояния восстанавливали в программе «codeml» пакета «PAML» [95] с параметрами, оцененными в «RAxML».

Таблица 1. Аминокислотные замены в митохондриальных генах *Metazoa*

ген	число видов	аминокислотные сайты	аминокислоты на сайт	замены на сайт	Аминокислоты на сайт в симуляции	замены на сайт в симуляции
ATP6	2931	186	9.5	152.1	12.1	154.6
COX1	4366	404	6.1	63.8	10.6	117.3
COX2	4131	165	8.9	137.2	12.4	206.8
COX3	2152	198	9.2	131.6	13	167.4
CYTB	5995	327	9.4	174.2	13.3	252.0
ND1	2013	253	8.7	92.9	12.7	124.4
ND2	5765	299	10.2	259.6	13.6	313.9
ND3	2766	94	9.7	182.3	12.8	194.4
ND4	2007	392	9.1	127.1	13.1	165.9
ND4L	1759	82	11.3	139.3	13.6	175.4
ND5	926	516	7.9	57.6	11.3	75.7
ND6	996	119	10.5	76.6	13.2	103.4

Независимо с использованием тех же методов и параметров, мы построили филогенетическое дерево, включающее 3586 видов хордовых, 586 – нехордовых и 178 видов грибов (всего 4350 видов), на основе сконкатенированных аминокислотных последовательностей пяти митохондриальных генов, последовательности каждого из которых выравнивали в программе «MUSCLE» [13]. Длина сконкатенированной последовательности составила 1524 аминокислоты.

Мы разделяли позиции на трансмембранные и немембранные согласно базе данных UniProt database [86].

2.1.2. Кластеризация замен на филогенетическом дереве

По восстановленным аминокислотным состояниям сайтов во всех узлах филогенетических деревьев мы определили, на каких ветвях происходили параллельные и дивергентные аминокислотные замены (Рисунок 5а). Мы считали, что все замены происходят на серединах филогенетических ветвей. Мы рассматривали только пары замен, которые произошли на независимых ветвях филогенетического дерева, то есть,

таких, что одна ветвь не является предковой для другой. Для нашего метода информативными являются сайты, где можно определить хотя бы одну пару параллельных и одну пару дивергентных замен из одной и той же предковой аминокислоты. Далее описан анализ только информативных сайтов. Мерой длины филогенетической ветви (филогенетического расстояния) служило число аминокислотных замен, произошедших на этой ветви по результатам моделирования в программе «RAxML», в пересчёте на один аминокислотный сайт. Мы определяли филогенетическое расстояние между двумя заменами как сумму расстояний от центров ветвей, где произошли замены, до узла – общего предка этих ветвей (Рисунок 5а).

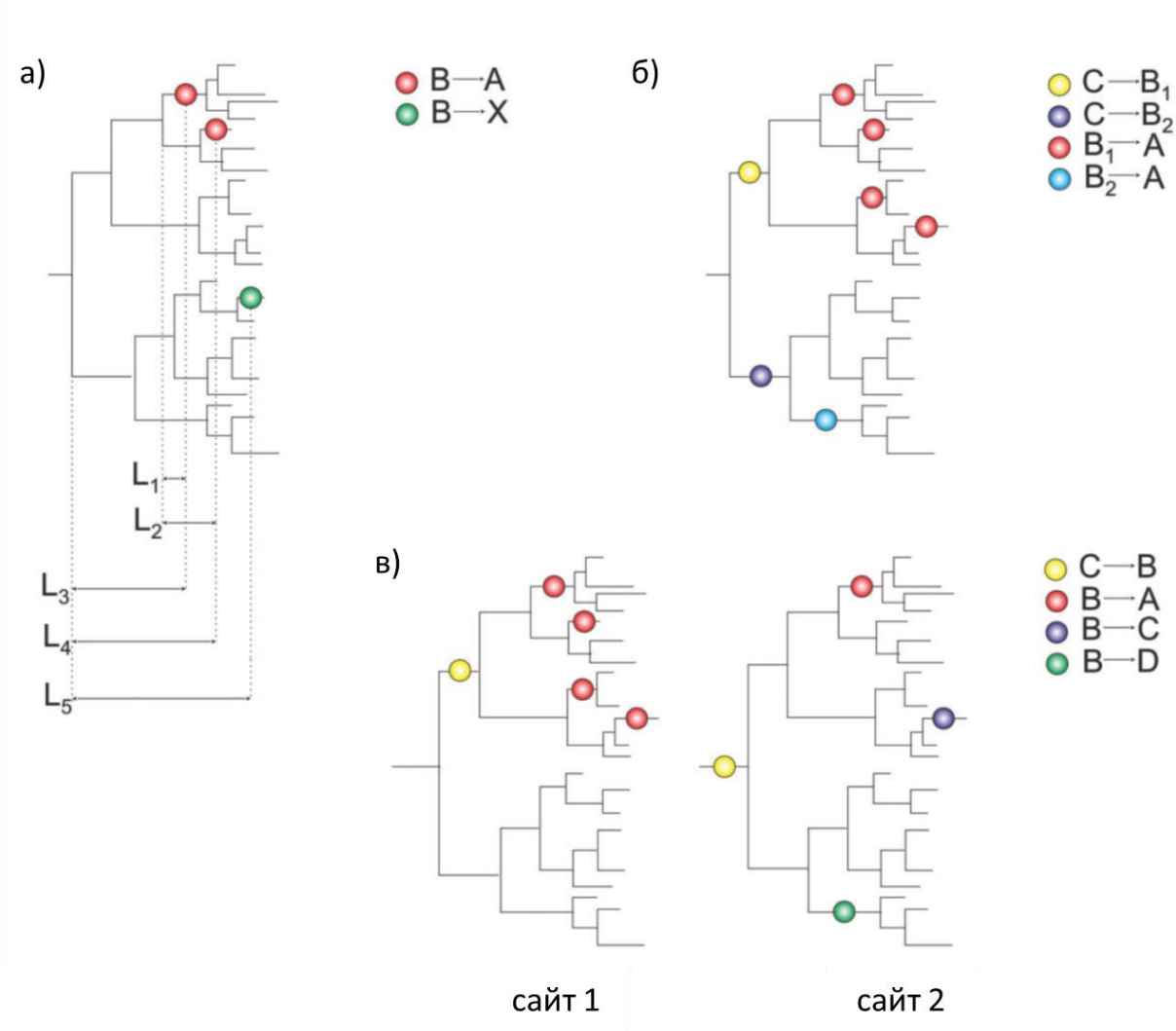


Рисунок 5. *Определение филогенетических расстояний между параллельными и дивергентными заменами. Кружками отмечены аминокислотные замены, произошедшие на соответствующих ветвях филогенетического дерева.*

а) – для каждой пары аминокислот (B, A) в сайте белка мы считаем филогенетические расстояния между всеми параллельными заменами $B \longrightarrow A$ (L1+L2) и расстояния между всеми дивергентными заменами $B \longrightarrow A$ и $B \longrightarrow X$ (L3+L5, L4+L5), где X – любая аминокислота, кроме A и B.
б) – замена $B_1 \longrightarrow A$ более частая, чем $B_2 \longrightarrow A$, что при отсутствии контроля на предковую аминокислоту даст избыток гомоплазий на коротких филогенетических расстояниях.
в) – объединение информации из сайтов с разными свойствами также может приводить к избытку гомоплазий на коротких филогенетических расстояниях.

Мы сравнивали расстояния между параллельными и дивергентными заменами и считали суммарную статистику (подробнее о статистике далее) для всех аминокислот и всех сайтов белка. При этом, чтобы предотвратить возможные влияния объединения сайтов и аминокислот на результат (Рисунок 5 б, в), для каждой предковой аминокислоты в информативном сайте мы выбирали максимально возможное одинаковое число пар параллельных и дивергентных замен, повторяя эту процедуру для всех предковых аминокислот во всех сайтах и получая две подвыборки параллельных и дивергентных замен одинакового размера. В обеих подвыборках для каждой пары замен мы измеряли филогенетические расстояния между заменами. Полученные расстояния разделяли на окна, для каждого окна считая отношение числа параллельных и дивергентных замен, попавших в это окно – П/Д (parallel to divergent ratio, P/D), Это отношение напоминает O-кольцевую статистику (O-ring statistics), широко применяемую в пространственной экологии в качестве меры аггрегации видов в сообществе [90].

Чтобы удостовериться, что разные сайты вносят согласованный вклад в значение статистики, мы проводили бутстрэппинг (взятие подвыборки с возвращением) сайтов в 1000 повторностях, каждый раз повторяя всю процедуру расчёта статистики. На основании полученных данных мы считали средние значения и квартили распределения нашей статистики.

2.1.3. Борьба с ненадёжностью восстановления топологии филогенетического дерева и предковых состояний

Мы выполнили отдельный анализ только для пар замен, которые определили, как «надёжные». Для выявления таких замен, для каждой ветви филогенетического дерева мы определили бутстрэп-поддержку на основании ста повторностей процедуры бутстрэппинга сайтов с помощью программы «RAxML». Пару замен считали «надёжной», если по крайней мере одна ветвь между ними имела 100% поддержку бутстрэпа, и для каждой замены из пары предковая и производная аминокислоты имели оценку максимального правдоподобия 1 по расчётам в программе «codeml». Это обеспечивало надёжность топологии и реконструкции предковых состояний для пары замен.

2.1.4. Симуляция эволюции

Мы симулировали эволюцию каждого гена при постоянных ОПАЛах с помощью программы «evolver» пакета «PAML» [95], используя модель «empirical F» с дискретным гамма-распределением скоростей эволюции между сайтами. Филогенетическое дерево, матрица замен и параметры распределения скоростей между сайтами были взяты из результатов программы «RAxML», посчитанных на основании имеющегося выравнивания. По симулированным аминокислотам в конечных узлах дерева (ныне живущих видах), мы восстановили предковые состояния в программе «codeml» с теми же параметрами.

Чтобы проверить влияние на результат отличий, которые могут возникнуть в матрицах замен между кладами филогенетического дерева из-за специфичных для отдельных клад особенностей использования аминокислот, в программе «RAxML» мы построили индивидуальные матрицы замен для каждой из трёх основных групп видов, входящих в совместную филогению для пяти генов: хордовых, нехордовых и грибов.

Затем мы использовали эти матрицы для симуляции эволюции в соответствующих группах видов дерева из 4350 видов и провели анализ для этих данных.

2.2. Результаты

2.2.1. Кластеризация замен на филогенетическом дереве – свидетельство изменчивости ОПАЛов

Разработанный нами подход для анализа филогенетической сближенности параллельных замен в сайте сходен с ранее предложенными [22, 68, 98], но более устойчив к не связанной с изменением ОПАЛов кластеризации замен и позволяет контролировать на потенциальное искажение результатов из-за совместного использования информации от разных аминокислот из разных сайтов. Объединение информации из разных сайтов применяется для повышения статистической мощности при поиске вариабельности ОПАЛов в ходе эволюции. Однако поскольку различные свойства (например, скорость эволюции) сайтов неодинаковы, такое действие может привести к ошибочным результатам, которые будут приняты за свидетельства непостоянства ОПАЛов.

Во-первых, невозможность различать параллельные и конвергентные замены, ведущие к одной и той же аминокислоте, из-за структуры генетического кода может привести к неверным выводам о изменении ОПАЛов. Например, на Рисунке 5б аминокислота A возникает из предковых аминокислот B_1 и B_2. Если аминокислота B_1 является предковой в одной кладе, а B_2 – в другой, и мутация $B_1 \longrightarrow A$ более частая, чем мутация $B_2 \longrightarrow A$, то замены на A будут превалировать в первой кладе. Но здесь кластеризация замен не будет являться свидетельством изменения ОПАЛа сайта, поскольку произойдёт за счёт причин, не связанных с отбором. Для того, чтобы контролировать на это, мы не берём в анализ конвергентные

замены и рассматриваем параллельные и дивергентные замены для каждого предкового варианта B отдельно.

Во-вторых, даже независимое рассмотрение разных предковых вариантов не предотвращает возможности ложной кластеризации гомоплазий, когда смешивается информация из сайтов и аминокислот с разными свойствами. Представим, что мы анализируем филогенетические расстояния между параллельными и дивергентными заменами в смешанной выборке сайтов. Рассмотрим ситуацию, показанную на Рисунке 5в. В сайте 1 аминокислота B появляется только в относительно небольшой кладе. Поэтому и замены B⟶A, и замены B⟶X автоматически будут филогенетически близкими друг к другу. Напротив, в сайте 2 B – долгоживущая аминокислота, поэтому расстояния между заменами могут быть больше. Если аминокислотные предпочтения в сайтах 1 и 2 отличаются, это может привести к видимости непостоянства ОПАЛов. Например, если в сайтах, где B – короткоживущая аминокислота (как в сайте 1), аминокислота A имеет высокую относительную приспособленность, а в сайтах, где аминокислота B распространена широко, одинаковую приспособленность имеют многие аминокислоты, то объединение таких сайтов может привести к избытку гомоплазий на коротких филогенетических расстояниях, поскольку параллельные замены B⟶A будут приходить в основном из сайтов первого типа, а дивергентные замены – в основном из сайтов второго типа.

Мы обходим эту проблему, в каждом информативном сайте (см. раздел 2.1.2) рассматривая все предковые аминокислоты отдельно. Для каждой предковой аминокислоты B, мы берём подвыборку пар замен так, что для каждой пары параллельных замен (B⟶A, B⟶A) случайно выбираем пару дивергентных замен (B⟶A и B⟶X) из того же сайта, объединяем подвыборки из разных сайтов и анализируем в полученной выборке расстояния между параллельными и дивергентными заменами.

Поскольку каждой паре параллельных замен в объединённой выборке соответствует одна пара дивергентных замен из той же предковой аминокислоты в том же сайте, такой подход исключает возможность получения ошибочных результатов из-за того, что предковые аминокислоты по-разному распределены на филогенетическом дереве. Для объединённых данных мы вычисляем отношение числа параллельных и дивергентных пар замен (П/Д) в каждом промежутке филогенетических расстояний.

2.2.2. Параллельные замены сближены филогенетически в белках митохондрий

В большинстве сайтов из 12 митохондриальных белков более чем для 900 видов *Metazoa*, мы обнаружили множество аминокислотных вариантов, что согласуется с результатами из [9]. После восстановления предковых состояний и аминокислотных замен в каждом сайте оказалось, что большинство аминокислот возникали больше одного раза, что дало нам возможность изучать распределение гомоплазий (Таблица 1 и Приложение 1).

Мы обнаружили избыток параллельных замен на близких филогенетических расстояниях (Рисунки 6 и 7). Это согласуется с исследованием митохондриальных белков позвоночных по меньшему объему данных [22]. По нашим результатам, отношение П/Д достигало 1.7-2.5 на филогенетических расстояниях меньше 0.1, быстро падая до 1.0 с увеличением расстояния (Рисунок 8, Приложение 2). В симулированных данных наблюдается незначительное снижение значения П/Д с расстоянием, что может быть следствием несбалансированности филогении.

Мы сравнили средние значения П/Д для сайтов, эволюционирующих с разной скоростью, и не нашли систематических отличий (Приложение 3).

Также мы не нашли значимых отличий между средними значениями П/Д для трансмембранных и немембранных аминокислотных остатков, хотя в большинстве белков эффект был немного ниже –то есть, отношение П/Д было ближе к 1 - для трансмембранных участков (Приложение 4).

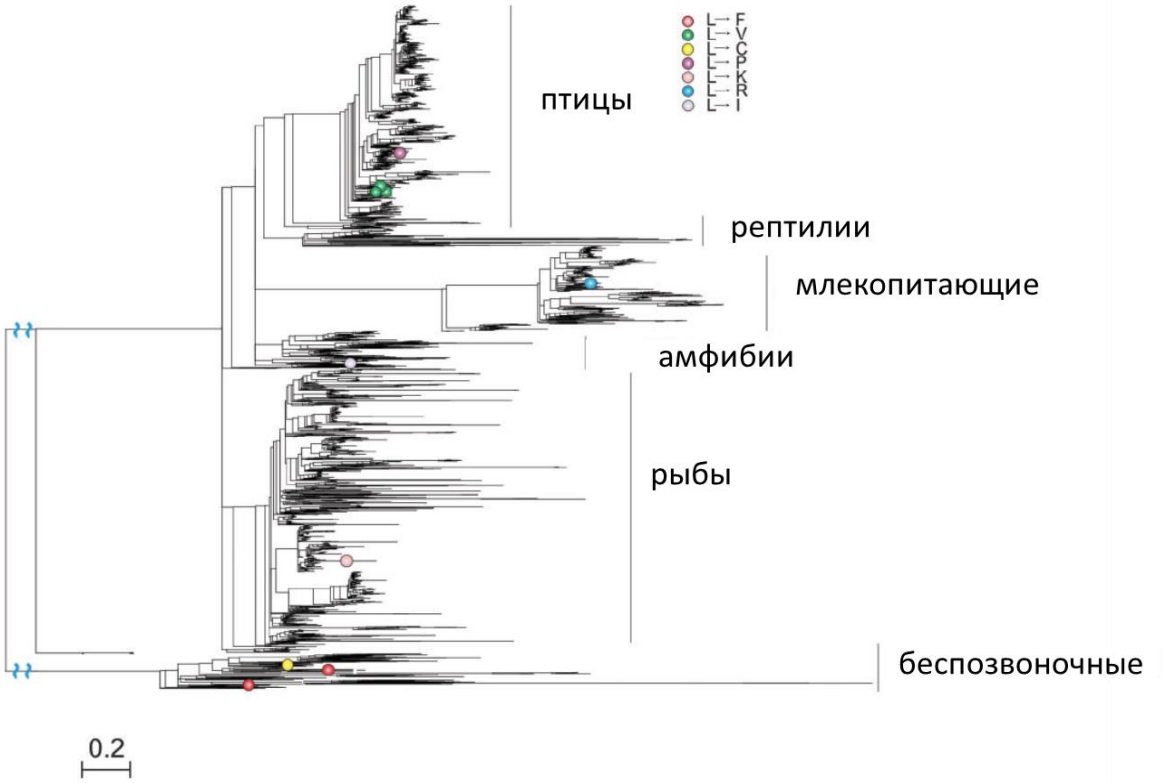


Рисунок 6. *Параллельные и дивергентные замены в сайте 202 белка ATP6 (по нумерации сайтов референтной последовательности человека в базе данных NCBI). Аминокислотные замены из предкового варианта (L) происходят по всему филогенетическому дереву. При этом две параллельные замены L⟶E произошли в эволюционно близких видах. То же верно и для трёх параллельных замен L⟶V. Филогенетические расстояния измерены как среднее число аминокислотных замен на сайт. Ветви, отмеченные синими волнистыми линиями, на рисунке укорочены на 1.2 единицы расстояния.*

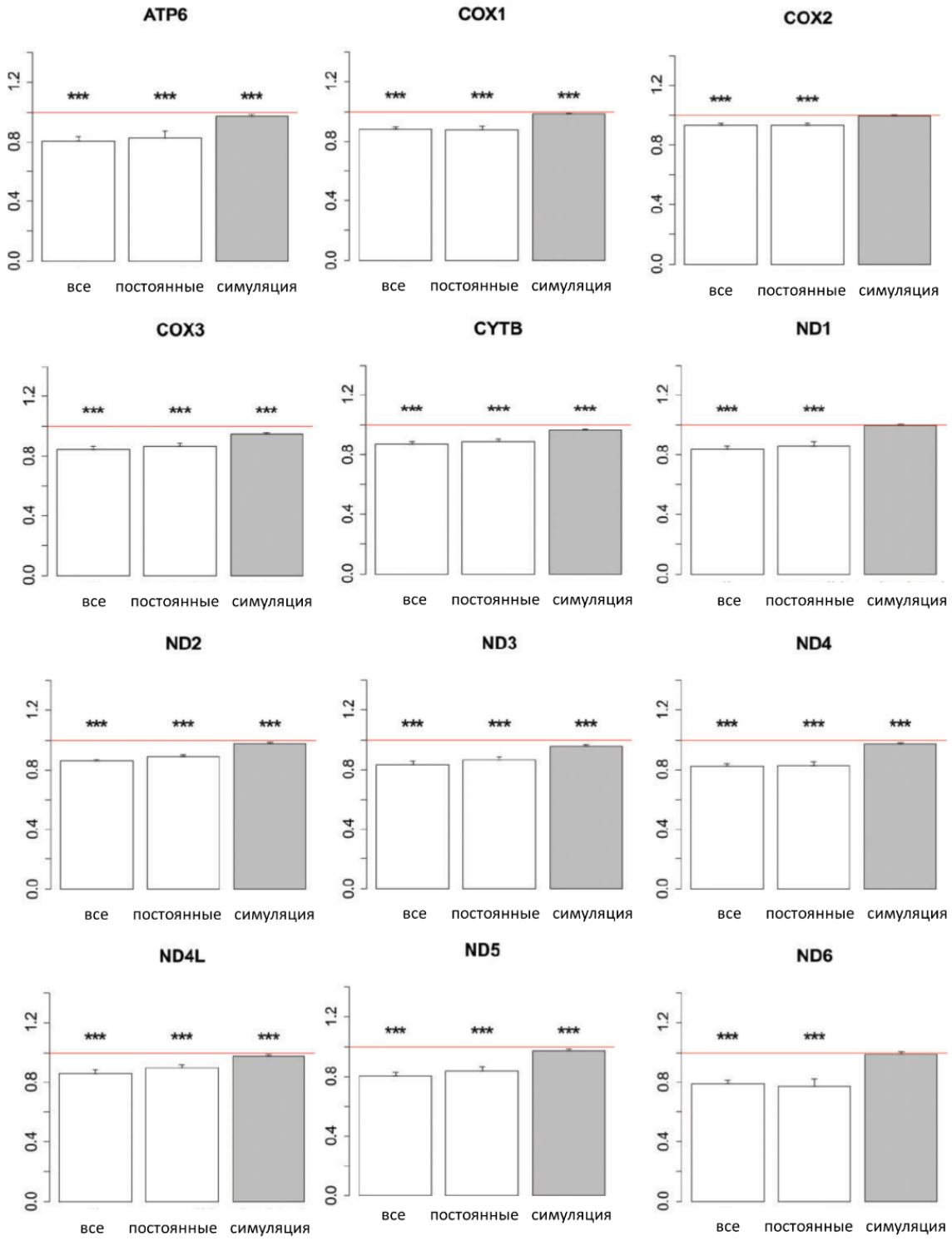


Рисунок *7. Отношение филогенетических расстояний между параллельными и дивергентными заменами на филогении Metazoa. Значения ниже 1 означают, что параллельные замены расположены ближе друг к другу на филогении, чем дивергентные. Высота столбца и усы обозначают соответственно медиану и 95% доверительный интервал для 1000 бустрэпов (выборок с возвращением) сайтов. Звёздочки показывают значимость отличия отношения от 1/1 (красная линия): *** - p < 0.001; нет звёздочки - p > 0.05). «**все**» - настоящие данные для всех замен;*

*«**постоянные**» - настоящие данные только для замен между «постоянными» парами аминокислот (см. текст); «**симуляция**» - симулированные данные.*

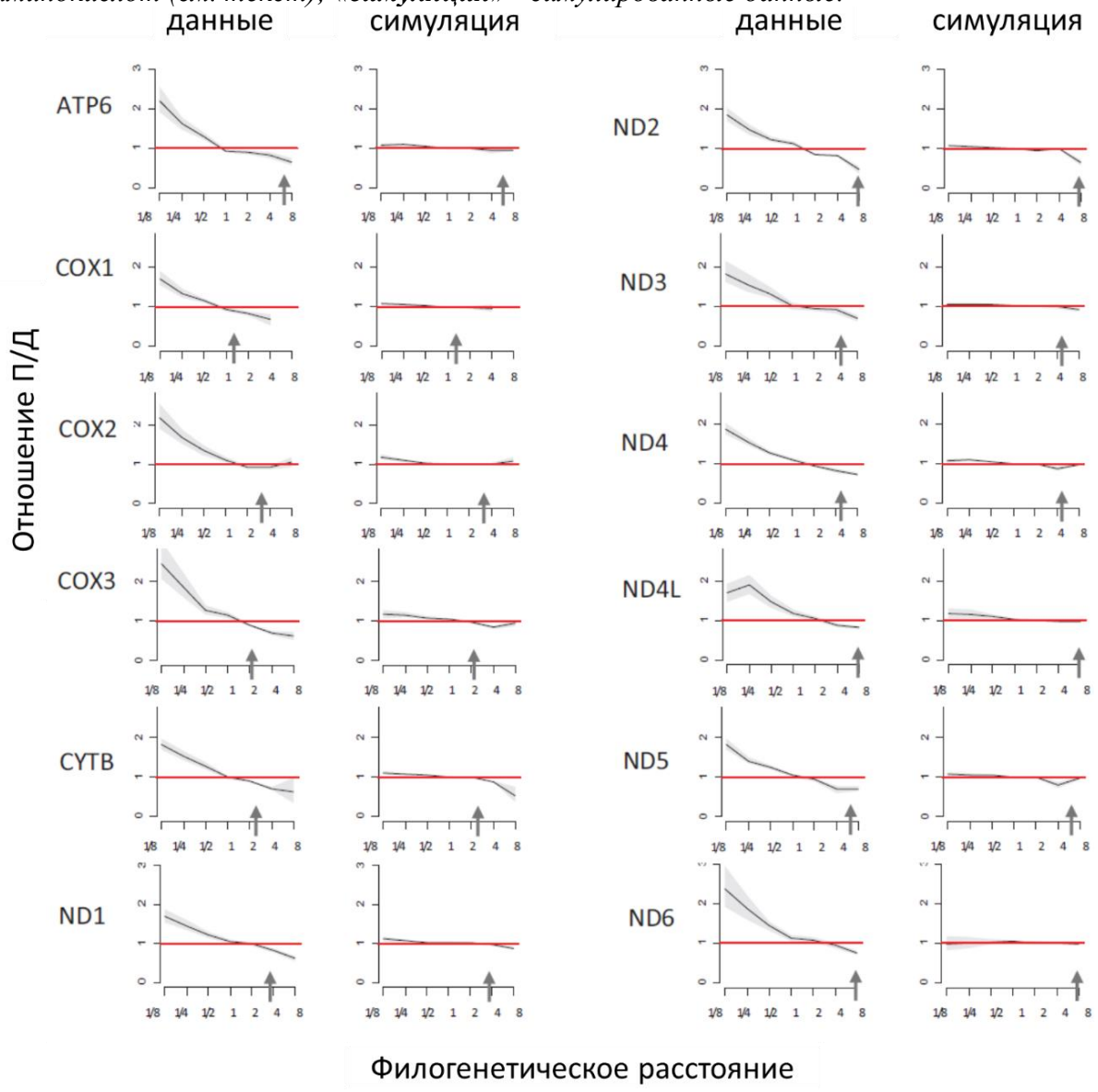


***Рисунок 8**. Доля параллельных замен выше между эволюционно близкими видами. Горизонтальная ось - филогенетическое расстояние между ветвями филогенетического дерева, несущими замены. Расстояния измерены в числе аминокислотных замен на сайт и объединены в группы так, что в одну группу попадают расстояния с одинаковым округлённым до целого числа $\log_2$. Вертикальная ось - отношение П/Д на соответствующем расстоянии. Чёрная линия - среднее; серая область - 95% доверительный интервал для 1000 бутстрэп-выборок сайтов; красная линия – ожидаемое отношение (единица); стрелка – расстояние между филогенетическими ветвями человека и дрозофилы.*

2.2.3. Избыток параллельных замен на коротких филогенетических расстояниях не является артефактом

Теоретически, снижение отношения П/Д с расстоянием может быть артефактом восстановления филогении, ведь при ошибочном разделении сестринских клад одна замена может быть принята за две параллельные, что приведёт к избытку гомоплазий в близких видах [56]. Мы проверили, вносят ли подобные артефакты существенный вклад в наши результаты, проведя анализ только для пар замен, которые мы обозначили как «надёжные» (см. подробнее в разделе 2.1.3). Подвыборка таких сайтов оказалась очень мала для имеющихся данных (Breen et al., 2012), поэтому мы построили филогенетическое дерево для 4350 видов по 5 белкам. Анализ, проведённый для «надёжных» замен на этом дереве (Рисунок 9), показал монотонное снижение отношения П/Д с расстоянием между заменами. Это значит, что наши результаты не являются следствием ошибок в восстановлении филогении.

Изменение П/Д с увеличением филогенетического расстояния отражает изменение скорости замен B⟶A относительно других замен. Эта скорость зависит от скоростей мутирования и фиксации и может меняться, если значения этих факторов разнятся между кладами.

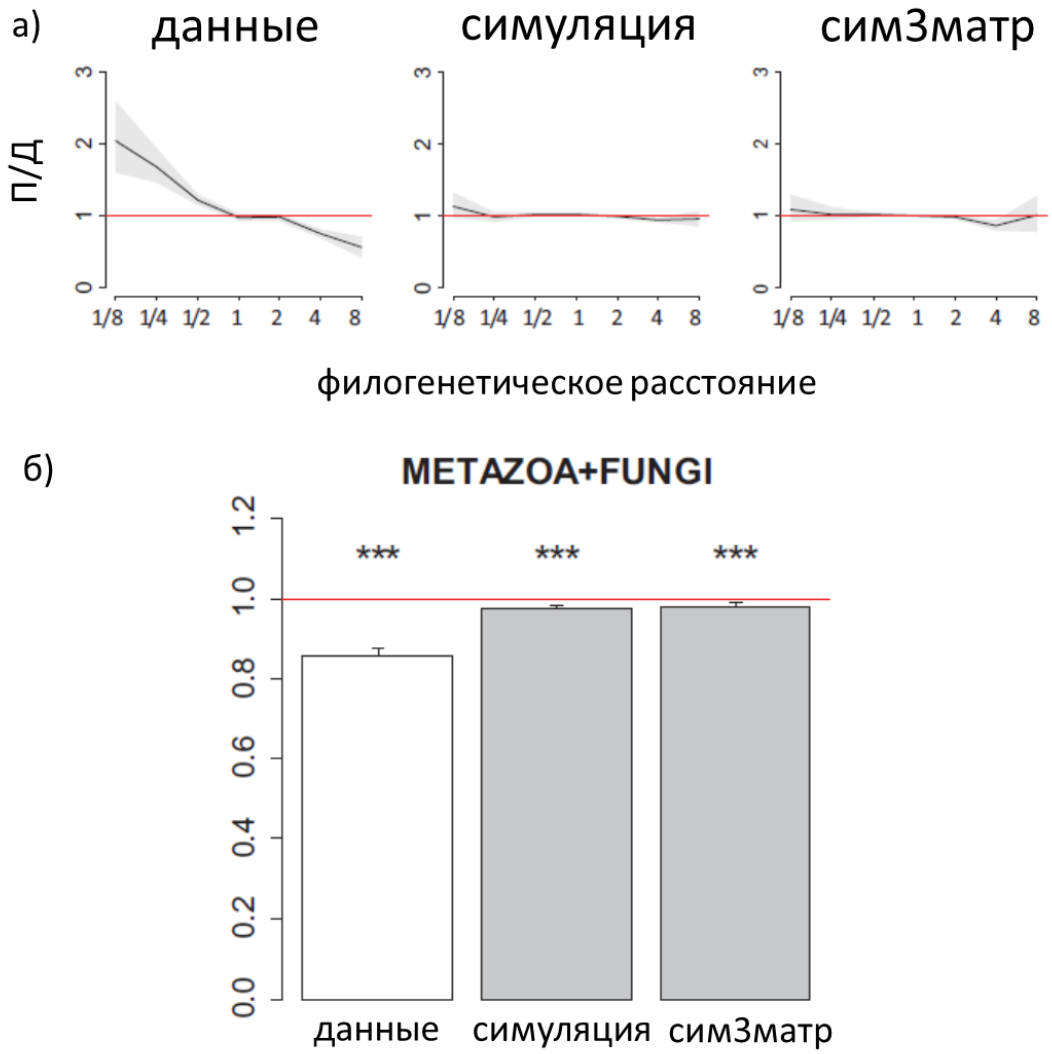


Рисунок 9. *Отношение числа параллельных и дивергентных замен на разных филогенетических расстояниях (а) и отношение средних филогенетических расстояний между параллельными и дивергентными заменами (б) в филогении для 4350 видов для «надёжных» пар замен.* ***сим3матр*** *– симуляция с независимыми матрицами замен в каждой кладе.*

Могут ли изменения скоростей замен с филогенетическим расстоянием наблюдаться при постоянных приспособленностях аминокислот? Это возможно при двух сценариях. Во-первых, мутационные спектры митохондрий могут разниться между кладами, что потенциально может привести к различным вероятностям одной и той же мутации в разных кладах. Если скорость определённой мутации в какой-то кладе намного выше, чем в других частях филогенетического дерева, то в этой кладе может возникнуть избыток гомоплазий. Для того, чтобы выяснить

влияние этого механизма на наш результат, мы симулировали эволюцию трёх основных клад филогенетического дерева для 4350 видов, используя для каждой клады отдельную матрицу замен (детали см. в разделе Методы настоящей главы). Мы провели свой анализ для полученных таким образом последовательностей и получили результаты, неотличимые от результатов для симуляции, проведённой с использованием одной матрицы замен для всего дерева (Рисунок 9). Следовательно, полученные нами результаты не могут объясняться описанным механизмом. Кроме того, большинство изменений в величине П/Д происходит на небольших филогенетических расстояниях (Рисунок 8), где мутационные матрицы не должны сильно отличаться.

Во-вторых, даже не будучи вызванными различиями мутационных матриц, изменения в значении П/Д могут быть следствием неодинакового использования кодонов в разных кладах. Если разные кодоны для аминокислоты B превалируют в разных кладах, то частота замен B⟶A будет выше в кладе, где более вероятна мутация из кодона аминокислоты B в кодон аминокислоты A. Чтобы проверить влияние этого фактора на наш результат, мы определили «постоянные» пары аминокислот (B, A), такие, что A может быть получена из любого кодона B за единственную однонуклеотидную мутацию, и провели анализ только для таких замен. И снова мы обнаружили избыток параллельных замен на коротких филогенетических расстояниях (Рисунок 7), а это означает, что наш результат не вызван структурой генетического кода.

Наконец, чтобы проверить, не вызван ли наблюдаемый нами эффект параллельными заменами, происходящими только на определённых парах филогенетических ветвей, мы построили распределение пар филогенетических ветвей по числу параллельных замен, которые произошли на них (Приложение 5). Мы не увидели систематических различий между распределениями для настоящих и симулированных

данных. Это значит, что наш результат формируется за счёт замен на многих парах филогенетических ветвей. Кроме того, если пара ветвей несёт параллельные замены во многих сайтах, эти ветви будут длинными, тогда как мы наблюдаем избыток параллельных замен на близких филогенетических расстояниях (Рисунок 8).

2.3. Обсуждение

Вероятность аминокислотной замены – монотонная функция от приспособленности производного варианта по отношению к предковому, поэтому изменения скоростей замен в процессе эволюции указывают на непостоянство относительных приспособленностей аминокислот, а, следовательно, на изменчивость ОПАЛов. Изменения скоростей замен в белке могут быть найдены с помощью обобщённых статистик, каковой является изменение степени параллелизма со временем: если аминокислотная замена становится более вредной, её вероятность падает, и она реже встречается на филогении.

Мы увидели, что параллельные замены в эволюции митохондриальных белков *Metazoa* тяготеют друг к другу на филогении. То есть, такие замены более вероятны в эволюционно близких видах, чем в далёких. В результате, расстояние между двумя параллельными заменами на филогенетическом дереве в среднем на 20% меньше, чем расстояние между двумя дивергентными заменами или чем ожидается при постоянных скоростях одних и тех же замен (Рисунок 7). Мы показали, что этот результат не является следствием несовершенной реконструкции филогенетического дерева или обобщения информации для сайтов и аминокислот с разными свойствами. Наши результаты нельзя объяснить только моделью ковариона, где сайт может переключаться между состояниями нейтрального и важного [17, 18], потому что наблюдаемые

нами изменения не связаны с изменениями скорости эволюции в сайте. По той же причине, наши результаты не объясняются и более широким классом моделей гетеротахии, где скорость эволюции сайта меняется со временем [49, 60, 96]. Для объяснения наших наблюдений необходима гетеропециллия [74, 84] – непостоянство скоростей индивидуальных замен во времени. Мы продемонстрировали, что наблюдаемые нами паттерны не вызваны систематическими различиями в матрицах замен между кладами, которые могут возникать из-за различий в скоростях мутирования, разного использования кодонов или генной конверсии.

Мы считаем, что наши наблюдения отражают изменения ОПАЛов [5, 61], происходящие в процессе эволюции. Действительно, сайт-специфическое изменение частоты замен на определённую аминокислоту означает, что приспособленность этой аминокислоты относительно других меняется в этом сайте со временем. Снижение частоты параллельных замен с филогенетическим расстоянием может происходить из-за снижения приспособленности данного варианта или повышения приспособленности других вариантов. Различить эти причины сложно, но, скорее всего, оба фактора играют роль [63]. Также мы не можем отличить изменения ОПАЛов из-за эволюции эпистатически связанных сайтов того же или другого белка, изменений среды или комбинации этих факторов.

Сравнение числа замен на одну и ту же или разные аминокислоты, то есть, числа конвергентных и дивергентных замен на разных филогенетических расстояниях, уже использовали при изучении эволюции. В древних белках степень конвергенции монотонно снижается с филогенетическим расстоянием, и на расстоянии в 10% аминокислотных различий половина реверсий (обратных замен) запрещена отбором [68]. Отношение скоростей конвергентных и дивергентных замен в позвоночных падает с филогенетическим расстоянием более, чем вдвое [22]. Аналогично, частота конвергентных замен падает с филогенетическим расстоянием в

млекопитающих и плодовых мушках [98]. Мы провели анализ для большего разброса филогенетических расстояний, чем [22]; однако также наблюдали локальный эффект (Рисунок 8). С помощью модели, в которой аминокислота в позиции белка может быть в «разрешённом» либо «запрещённом» состоянии, для данных из разных источников показано, что в одной позиции белка переключения аминокислот между состояниями происходят в 5 раз чаще, чем аминокислотные замены [87]. По нашим данным, скорость изменения ОПАЛов сильно разнится между белками: филогенетическое расстояние, за которое отношение П/Д достигает единицы, варьирует от 0.6 для ATP6 до 2.1 для ND6. Это расстояние зависит также от размера и формы филогенетического дерева. В целом, скорость изменения ОПАЛов в наших данных сопоставима со скоростью аминокислотных замен (Рисунок 8), что согласуется с [61], где показано, что флуктуации ОПАЛов в белках дрозофилы происходят со скоростями порядка скорости нейтральной эволюции.

Заключение

Непостоянство однопозиционных ландшафтов приспособленности играет заметную роль в эволюции митохондриальных белков животных. Скорость изменения ОПАЛов сопоставима со скоростью аминокислотных замен. В какой степени непостоянство ОПАЛов является следствием изменения геномного контекста, а в какой – изменения среды, остаётся темой будущих исследований.

Глава 3. Вариабельность приспособленности аллелей в митохондриальных белках человека

В этой главе проанализирована встречаемость различных аллелей митохондриальных белков человека в других видах *Opisthokonta*. Мы показали, что замены на такие аминокислоты чаще происходят в более родственных человеку линиях. Это говорит о том, что человеческий вариант может с большей вероятностью быть вредным в филогенетически далёких от человека видах. Мы наблюдали эту тенденцию и для известных патогенных вариантов человека, но не для аминокислот, не встречающихся в людях. Таким образом, аминокислоты, имеющие высокую приспособленность в близких видах, чаще можно встретить в людях как нейтральные или патогенные, чем аминокислоты, встречающиеся только у филогенетически далёких видов.

3.1. Материалы и методы

3.1.1. Подготовка данных

Для двенадцати митохондриальных белков более чем 900 видов *Metazoa* и пяти митохондриальных белков 4350 видов *Opisthokonta* в качестве референтных, нереферентных нейтральных и патогенных вариантов мы брали аминокислоты, которые наблюдаются в исследуемых белках человека. В качестве референтной последовательности человека мы использовали Кэмбриджскую Референтную Последовательность (revised Cambridge Reference Sequence, rCRS) митохондриальной ДНК человека [3], в качестве нейтральных полиморфизмов - аминокислотные аллели из раздела «варианты последовательности кодирующих участков митохондриальной ДНК и РНК» («mtDNA Coding Region & RNA Sequence Variants») базы данных MITOMAP [50]. Мы считали патогенными варианты из раздела «Мутации в кодирующих и регуляторных участках

митохондриальной ДНК, для которых была показана связь с заболеваниями» («Reported Mitochondrial DNA Base Substitution Diseases: Coding and Control Region Point Mutations») базы данных MITOMAP со статусом «упоминавшиеся» («reported»; рассматривались как патогенные хотя бы в одной публикации) или «подтверждённые» («confirmed»; патогенность подтверждена как минимум двумя независимыми лабораториями).

3.1.2. Определение избытка замен на аминокислоты человека вблизи его ветви на филогении

Для каждой позиции белка мы рассматривали аминокислотные варианты, которые (1) являются референтным аллелем человека, (2) существуют в популяции человека как нейтральные полиморфизмы или (3) указаны как патогенные для человека. Для дальнейшего анализа мы оставили только аминокислоты, для которых на филогении произошло хотя бы по одной замене на неё (гомоплазии) и на другую аминокислоту (дивергентной замене) из одной и той же предковой аминокислоты, без учета замен, случившихся между корнем филогенетического дерева и ветвью *Homo sapiens*. Предковая аминокислота для гомоплазий и дивергентных замен не обязательно должна была повторять предковый вариант для человека.

Для каждого такого аллеля, мы сравнили филогенетические расстояния между ветвью человека и заменами на эту аминокислоту с расстояниями между ветвью человека и дивергентными заменами, используя описанный в разделе 2.1.2 метод, который учитывает возможные различия в ОПАЛах разных сайтов и в вероятностях разных мутаций. Для каждой предковой аминокислоты в каждой позиции белка мы брали одинаковое число гомоплазий на человеческую аминокислоту (референтную, нереферентную нейтральную или патогенную – Nhomo) и

дивергентных замен (Ndiverg) на другие аминокислоты. Размер подвыборки составлял min(Nhomo, Ndiverg). Мы повторили эту процедуру для всех предковых аминокислот в каждом сайте, получив таким образом подвыборки гомоплазий и дивергентных замен одинакового размера, и в каждой подвыборке измерили филогенетические расстояния от замен до ветви человека. Затем разделили все подсчитанные расстояния по равным отрезкам и для каждого отрезка посчитали отношение Г/Д числа попавших в него гомоплазий (Г) к числу попавших в него дивергентных замен (Д). Для получения среднего значения и 95% доверительного интервала для отношения Г/Д, мы провели бутстрэппинг сайтов в 1000-кратной повторности, каждый раз заново беря подвыборку замен для каждой предковой аминокислоты. Для контроля, мы провели такой же анализ, беря вместо аминокислоты человека случайную аминокислоту из тех, что встречались в сайте, а также проанализировали данные, полученные симуляцией эволюции в каждом сайте программой «evolver» пакета «PAML» (подробности в разделе 2.1.4).

3.2. Результаты

3.2.1. Замены на референтные аминокислоты человека чаще встречаются в филогенетически близких человеку видах

Референтный аллель человека наблюдается в среднем более чем в половине видов, составляющих наши данные (71.2% видов *Opisthokonta* в данных по пяти митохондриальным белкам, 60% видов *Metazoa* в данных по 12 митохондриальным белкам). В 7.1% (*Opisthokonta*) и 10% (*Metazoa*) таких видов, эти аминокислоты не делят происхождение с аллелем человека, а возникают независимо в результате в среднем 30 (*Opisthokonta*) и 21 (*Metazoa*) гомоплазий на сайт.

Мы проверили предположение, что филогенетические позиции замен на аминокислоты, представляющие референтный аллель человека,

смещены к ветви человека по сравнению с заменами на другие аминокислоты. Устройство нашего метода исключает ошибки, связанные с совместным анализом разных сайтов и аминокислот (см. раздел 3.1.2). Метод не позволяет включать в анализ самые консервативные сайты, но большинство сайтов в рассмотренных данных были достаточно вариабельными (Приложения 6 – 8).

Для большинства белков средние филогенетические расстояния от ветви человека до независимых гомоплазий на аминокислоту – референтный аллель человека на 10% короче, чем расстояния до замен на другие аминокислоты (Рисунок 10а; Приложение 9). Ни в одном из контролей (случайная встречающаяся в сайте аминокислота вместо аллеля человека или симулированные данные) такой эффект не наблюдался (Рисунок 10а; Приложение 9). Отношение числа замен на референтную аминокислоту человека к числу замен на аминокислоты, не замеченные в популяции *H. sapiens* (отношение Г/Д), монотонно снижается с эволюционным расстоянием от человека (Рисунок 10б; Приложение 10). Для большинства генов, относительное число замен на человеческую референтную аминокислоту падает в 1.5 – 3 раза с филогенетическим расстоянием от ветви человека (Рисунок 10б; Приложение 10), отражая склонность таких замен происходить в более родственных человеку видах.

Наш результат не обусловлен различиями в использовании кодонов в разных частях филогении, поскольку эффект по-прежнему наблюдается, если взять в анализ только пары аминокислот, где любой кодон предковой аминокислоты может дать кодон производной аминокислоты за одну нуклеотидную замену («постоянные» пары аминокислот; Приложение 9).

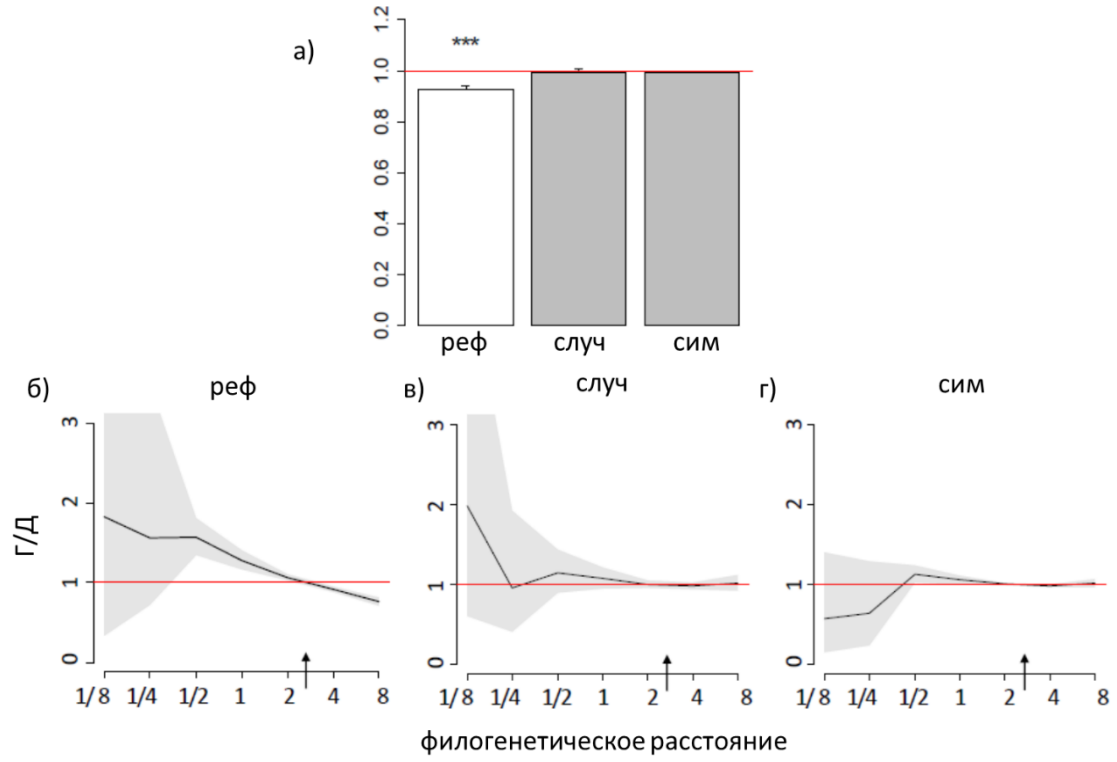


Рисунок 10. *Замены на референтный аллель человека чаще происходят в эволюционно близких к человеку видах:*
Сниженное отношение средних филогенетических расстояний от ветки H. sapiens до замен на рассматриваемую аминокислоту и средних филогенетических расстояний от ветви H. sapiens до замен на другие аминокислоты в том же сайте (а) и повышенное отношение числа замен на рассматриваемую аминокислоту к числу замен на другие аминокислоты (Г/Д) на близких филогенетических расстояниях от ветки человека (б) наблюдаются для референтного аллеля человека, но не для случайной аминокислоты (а, в) и не в симулированных данных (а, г), в филогении 4350 видов Opisthokonta.
*а) - Г/Д <1 означает, что замены на рассматриваемую аминокислоту происходят на меньших филогенетических расстояниях от человека, чем замены на другие аминокислоты. Высота столбца и усы обозначают соответственно медиану и 95% доверительный интервал для 1000 бустрэпов (выборок с возвращением) сайтов. Звёздочки показывают значимость отличия отношения от 1/1 (красная линия): *** - p < 0.001; **, p < 0.01, *, p < 0.05.* ***реф*** *- референтный аллель человека;* ***случ*** *- случайная аминокислота среди тех, что встречаются в сайте у других видов, но не у человека;* ***сим*** *- аллель человека в симулированных данных.*
б), в), г) - горизонтальная ось, филогенетические расстояния между ветвями, на которых случились замены, и ветвью H. sapiens. *Расстояния измерены в числе аминокислотных замен на сайт и соединены в группы так, что в одну группу попали расстояния с одинаковым* $\log_2$*, округлённым до целого числа. Вертикальная ось, отношение Г/Д на соответствующем расстоянии. Чёрная линия - среднее; серая область - 95% доверительный интервал для 1000 бутстрэп-выборок сайтов. Красная линия - ожидаемое отношение 1. Стрелки показывают расстояние между ветками человека и дрозофилы.*

3.2.2. Замены на аминокислоты, представляющие собой нейтральные полиморфизмы человека, чаще происходят в эволюционно близких к человеку видах

Мы провели анализ для аминокислотных полиморфизмов в митохондриальных белках из базы данных «MITOMAP». Эти полиморфизмы представляют собой варианты из многих популяций человека, но, как и в любой подобной базе данных, наиболее представлены в ней частые аллели. Мы проанализировали филогенетическое распределение замен на аминокислоты, наблюдающиеся у человека как нереферентные (как правило, минорные) аллели (Приложение 7).

Как и в случае референтного аллеля человека, замены на нереферентные аллели имели тенденцию происходить ближе к ветви человека на филогенетическом дереве по сравнению с заменами на аминокислоты, не наблюдаемые в популяции человека (Рисунок 11; Приложения 11, 12). И снова среднее филогенетическое расстояние от ветви человека до замен на аминокислоту, которая встречается в популяции человека как нереферентный аллель, было на 10% меньше, чем до других замен (Рисунок 11а; Приложение 11), и плотность замен на такие аминокислоты снижалась с отдалением от ветви человека (Рисунок 11б; Приложение 12). Как и в разделе 3.3.1, при рассмотрении только «постоянных» пар аминокислот эффект кластеризации гомоплазий сохранялся для человеческих аллелей, но не наблюдался для случайных аминокислот (Приложение 11).

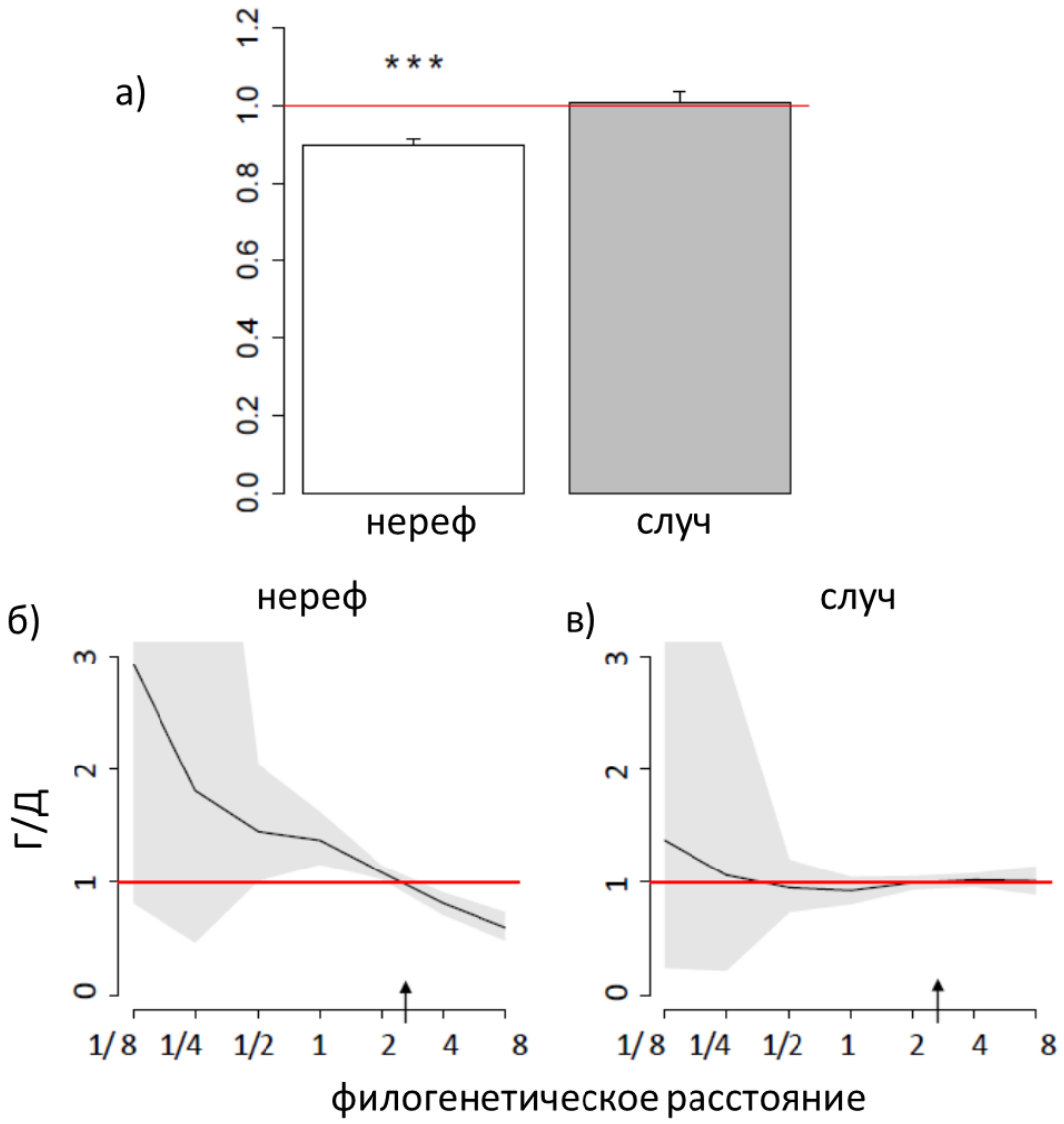


Рисунок 11. *Замены на нереферентный аллель человека чаще происходят в эволюционно близких к человеку видах.*
Сниженное отношение средних филогенетических расстояний от ветви H. sapiens до замен на рассматриваемую аминокислоту и до замен на другие аминокислоты в том же сайте (а) и повышенное отношение числа замен на рассматриваемую аминокислоту к числу замен на другие аминокислоты (гомоплазии/дивергентные, Г/Д) на близких филогенетических расстояниях от ветви человека (б) наблюдаются для нереферентного аллеля человека, но не для случайной аминокислоты (в), в филогении 4350 видов Opisthokonta. Условные обозначения см. на рисунке 11.

3.2.3. Замены на аминокислоты, представляющие собой патогенные варианты человека, чаще происходят в эволюционно близких к человеку видах

Наконец, мы рассмотрели аллели человека, выделенные в базе данных «MITOMAP» как вызывающие заболевания. Поскольку для каждого гена таких мутаций в базе данных оказалось немного (Приложение 8), дисперсия отношения Г/Д, как и ожидалось, велика. Тем не менее, в данных по *Opisthokonta* (Рисунок 12), а также в 5 из 12 белков из данных по *Metazoa* (Приложения 13, 14), патогенный для человека вариант чаще возникает в результате независимых замен в филогенетически более близких к человеку линиях. Противоположная ситуация, когда патогенный для человека вариант встречается предпочтительно в филогенетически далёких от человека видах, не наблюдалась ни для одного гена. Эффект был ещё более выражен для шести мутаций, патогенность которых для человека подтверждена двумя или более независимыми исследованиями (Рисунок 12а). Отметим, что описанный эффект наблюдался для гомоплазмичных (зафиксированных в индивиде) мутаций, но не для мутаций, которые были описаны только как гетероплазмичные (то есть, не зафиксированные, а полиморфные в индивиде; Рисунок 12б). Как и прежде, результат для патогенных аллелей не вызван предпочтительным использованием кодонов, для которых велика вероятность мутации на человеческий вариант, в филогенетически близких к человеку видах (Приложение 13).

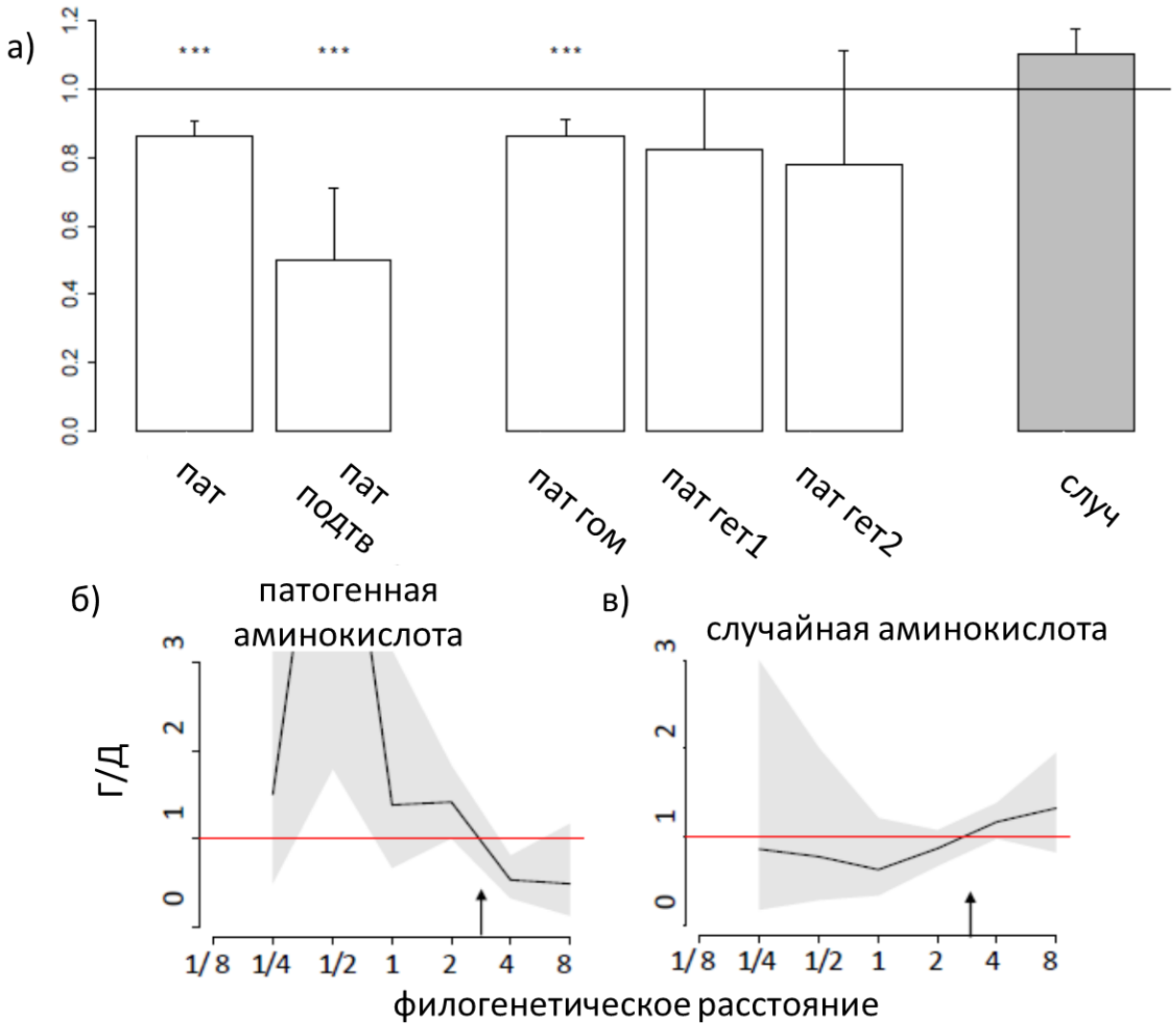


Рисунок 12. *Замены на известные патогенные для человека аминокислоты чаще происходят в эволюционно близких к человеку видах.*

*а) - в филогении 4350 видов Opisthokonta филогенетические расстояния между ветвью человека и заменами на рассматриваемую аминокислоту в среднем короче расстояний между ветвью человека и заменами на не встречающиеся у человека аминокислоты в том же сайте для всех (**пат**), подтверждённых (**пат подтв**) и гомоплазмичных (**гом**) патогенных аминокислотных вариантов человека, но не для случайных аминокислот (**случ**) и не для патогенных аминокислот, которые наблюдались только в состоянии гетероплазмии (**пат гет1,2**); **пат гет1** – патогенные варианты, наблюдавшиеся только как гетероплазмичные; **пат гет2** - патогенные варианты, наблюдавшиеся только как гетероплазмичные, с двумя и более литературными ссылками в «MITOMAP». Остальные условные обозначения повторяют рисунки 11 и 12.*

б), в) - повышенное отношение Г/Д на близких филогенетических расстояниях от ветви человека наблюдается для патогенных аллелей человека, но не для случайной аминокислоты, в филогении 4350 видов Opisthokonta. Условные обозначения те же, что на рисунках 11 и 12.

3.2.4. Патогенные аллели человека биохимически сходны с нормальными аминокислотными вариантами

Чтобы понять, отчего и нормальные, и патогенные аллели человека имеют тенденцию встречаться в филогенетически близких к человеку видах, мы проанализировали сходство этих вариантов. Оказалось, что и нормальные, и патогенные нереферентные аллели человека ближе к референтному аллелю по своим биохимическим свойствам согласно матрице Мияты [57], чем аминокислоты, встречающиеся на филогении, но не наблюдавшиеся в популяции человека (Рисунок 13). В свою очередь, такие аминокислоты больше походили на референтный аллель человека, чем аминокислоты, не встречающиеся на филогении в данном сайте.

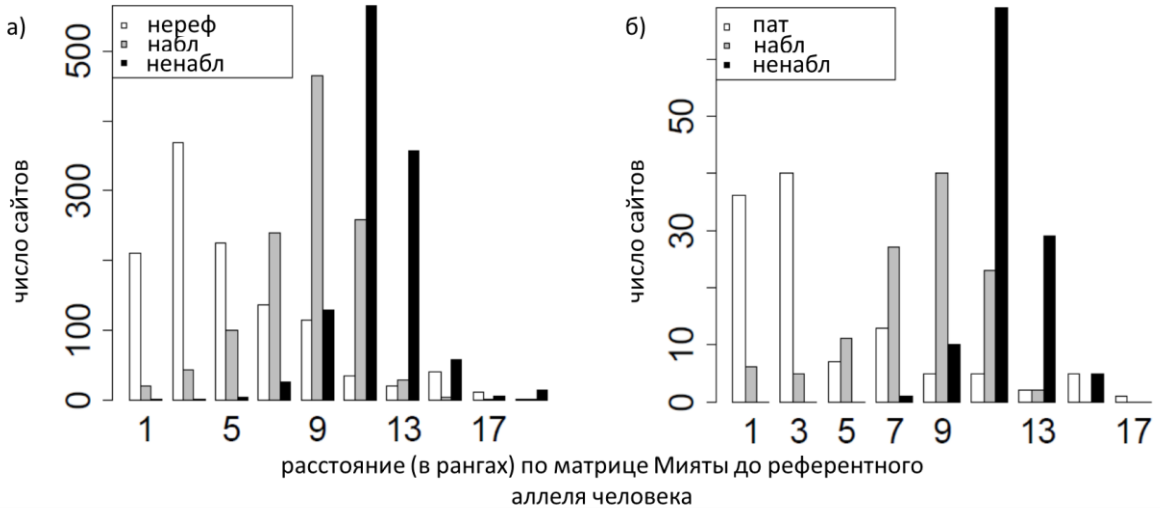


Рисунок 13. *Ранг расстояния по матрице Мияты между референтным аллелем человека и его нереферентными аллелями (белый цвет), другими аминокислотными вариантами, которые наблюдались (серый) или не наблюдались (чёрный) в том же сайте:*
а) нереферентный аллель – нейтральный полиморфизм,
б) нереферентный аллель – патогенный вариант.
Для каждого сайта с известными нейтральными или патогенными вариантами, мы проранжировали (от минимума к максимуму) все аминокислоты по расстоянию Мияты от референтного аллеля человека и определили ранг для патогенного (или нейтрального) варианта человека, средний ранг для аминокислот, встречающихся в сайте в других видах и средний ранг для остальных аминокислот.

3.2.5. Индивидуальные мутации

Чтобы проиллюстрировать наблюдавшуюся филогенетическую сгруппированность замен на аллели человека, мы схематично изобразили на филогенетическом дереве *Opisthokonta* распределение замен в шести подтвержденных аминокислотных сайтах из базы данных «MITOMAP». На этих схемах видно, что замены на патогенные аллели человека происходят ближе к ветви человека, чем другие замены из таких же предковых аминокислот (Рисунок 14).

Для данных по *Metazoa* мы также изобразили три индивидуальных аминокислотных сайта с известными патогенными мутациями. Повышенная плотность замен на аминокислоту вблизи ветви человека в этих сайтах может означать, что её относительная приспособленность выше в видах, близко родственных человеку.

По существующим данным, мутация V113A белка ND1 вызывает биполярное расстройство. В эксперименте она снижала потенциал митохондриальной мембраны и активность ND1 [59]. Согласно базе данных «mtDB» [31], частота аллеля A113 в популяции человека составляет 0.5%. Однако, мы видим, что аланин независимо возникает в результате замен V113A в трёх кладах позвоночных: обезьяны Старого света (*Cercopithecidae*), летающие лемуры (*Cynocephalidae*) и черепахи (*Geoemydidae*), тогда как большинство других замен из валина в этом сайте происходят в беспозвоночных (Рисунок 15). В результате, среднее филогенетическое расстояние между веткой человека и параллельными заменами V113A равно 2.35 (медиана 0.75), тогда как для замен из валина на другие аминокислоты это расстояние составило 4.12 (медиана 4.6).

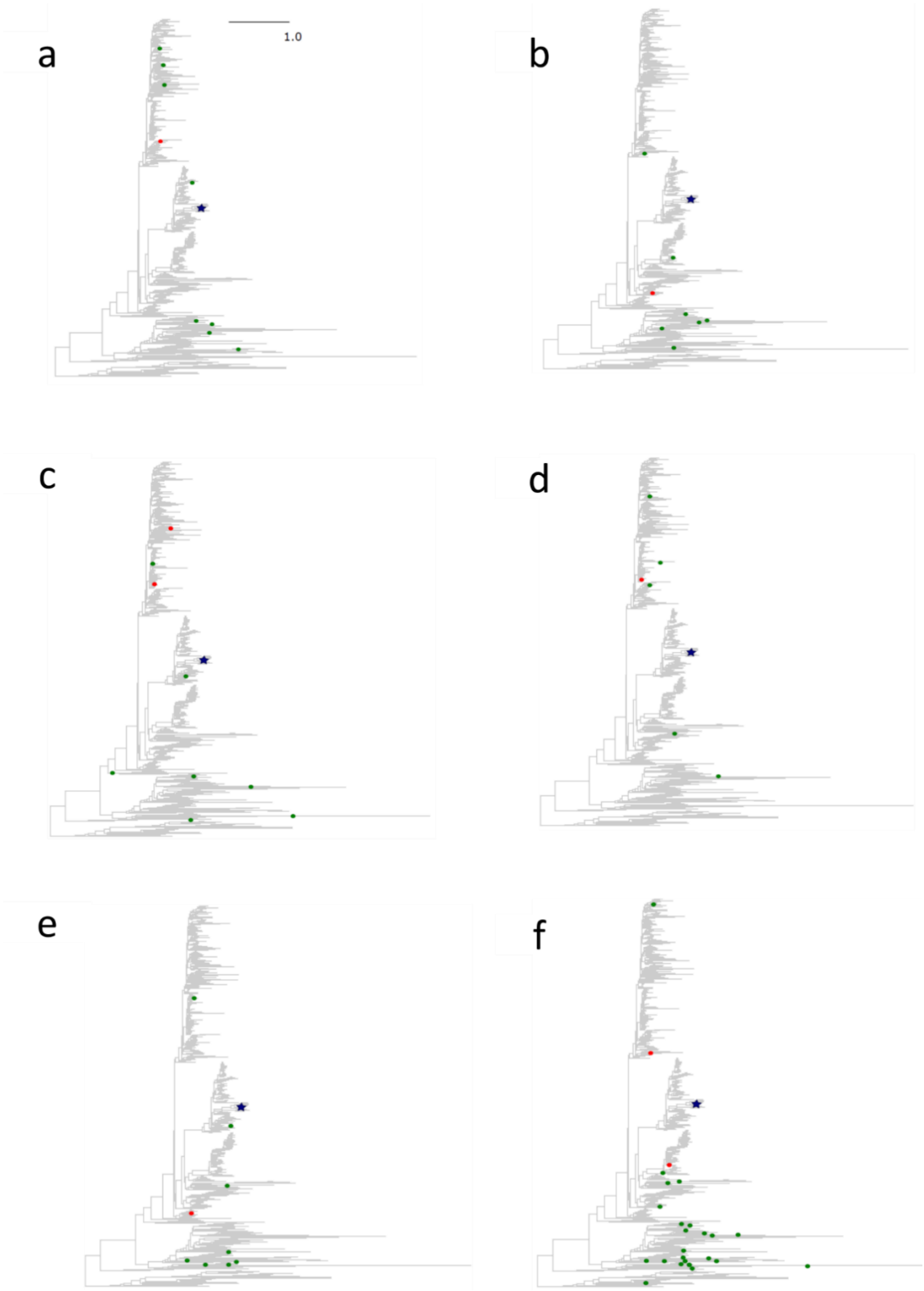


Рисунок 14. *Замены на патогенный аллель человека (красные кружки) и на другие аминокислоты (зелёные кружки) на филогении 4350 видов Opisthokonta. a – сайт 156 белка ATP6; b – сайт 220 белка ATP6; c – сайт 220 белка ATP6; d – сайт 278 белка CYTB; e – сайт 35 белка CYTB; f – сайт 40 белка CYTB. Синяя звёздочка – H. sapiens. Филогенетические расстояния измерены в числе аминокислотных замен на сайт.*

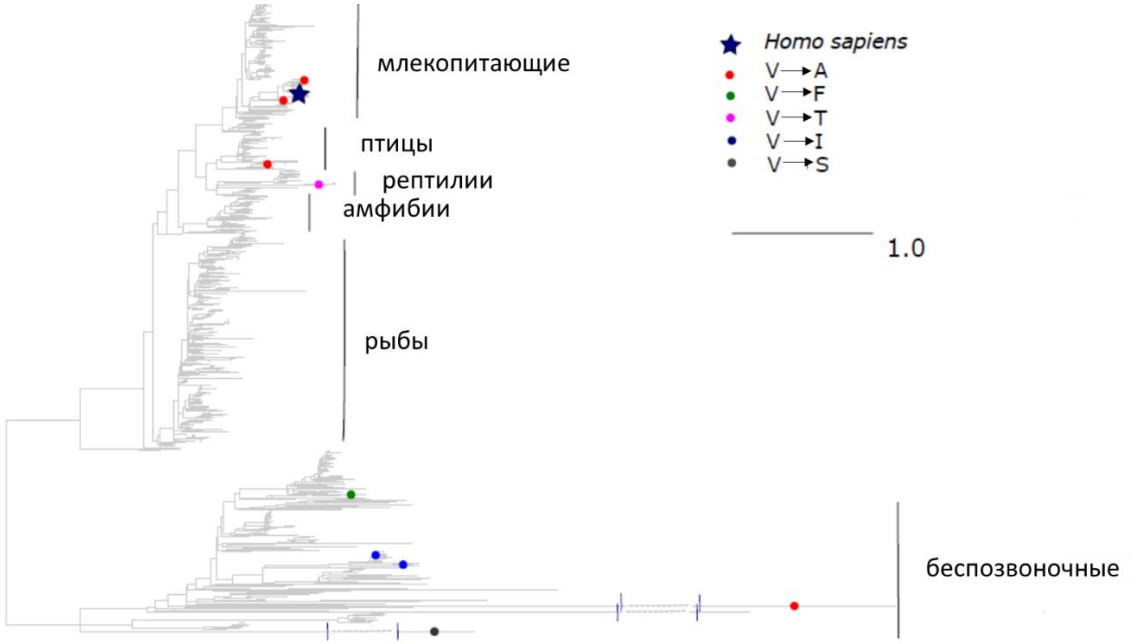


Рисунок 15. *Замены в сайте 113 белка ND1. Синяя звёздочка - H. sapiens; красные кружки – замены с валина (V) на аланин (A), который является патогенным аллелем человека; кружки других цветов – замены валина на другие аминокислоты. Филогенетические расстояния измерены в числе аминокислотных замен на сайт. Ветви, отмеченные синими линиями, укорочены на 2 единицы длины.*

Мутация V91A белка COX3 может приводить к синдрому Лейха [58]. В наших данных аланин независимо возникал в этом сайте семь раз, из них шесть раз из валина и один раз из изолейцина. Все эти замены, кроме одной, произошли у млекопитающих, тогда как десятки других замен из валина и изолейцина происходили по всей филогении (Рисунок 16). В результате, среднее филогенетическое расстояние от ветви человека до замен V91A равно 0.6 (медиана 0.7), а до других замен из V – 0.8 (медиана 2.1). Значения соответствующих расстояний для предковой аминокислоты I составили 0.6 (медиана 0.6) и 2.7 (медиана 2.2) соответственно.

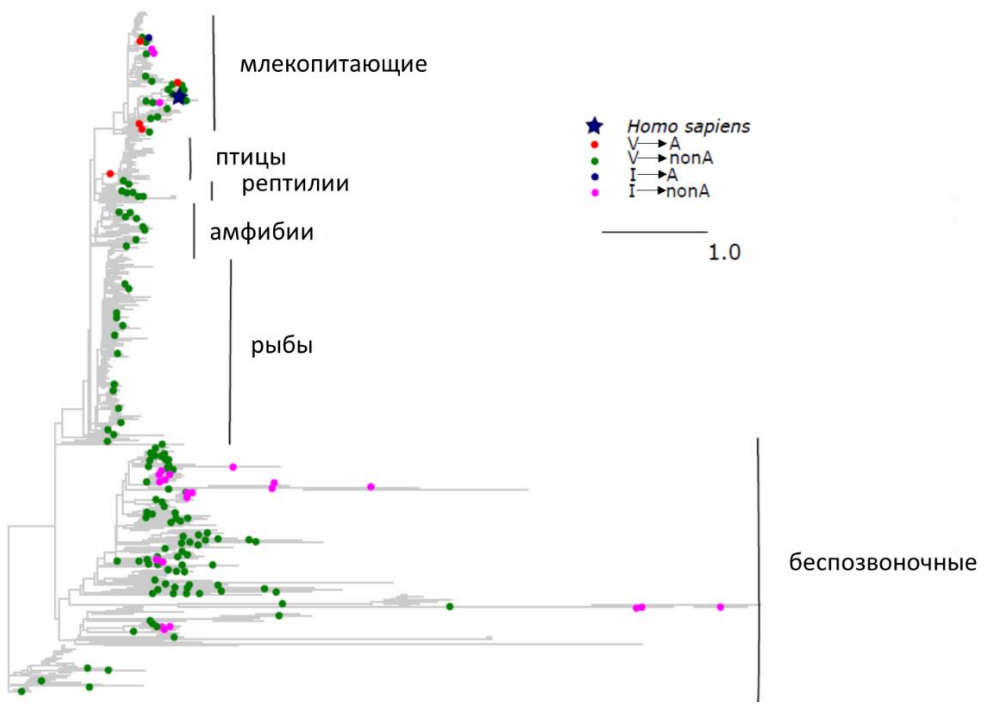


Рисунок 16. *Замены в сайте 91 белка COX3. Синяя звёздочка - H.sapiens; красные кружки - замены валина (V) на аланин (A), который является патогенным аллелем человека; тёмно-синий кружок - замена изолейцина (I) на аланин; остальные кружки - замены на другие аминокислоты. Филогенетические расстояния измерены в числе аминокислотных замен на сайт.*

По некоторым данным, вариант I33V в ND6 вызывает диабет второго типа [85]. Согласно базе данных «mtDB», этот аллель в человеческой популяции имеет частоту 0.1%. Однако, независимые замены I33V неоднократно происходили у млекопитающих и амфибий, тогда как у беспозвоночных были часты другие замены из I (Рисунок 17). Среднее филогенетическое расстояние от ветви человека до параллельных замен I33V составило 2.4 (медиана 1.5), а до других замен из I – 12.9 (медиана 14.8).

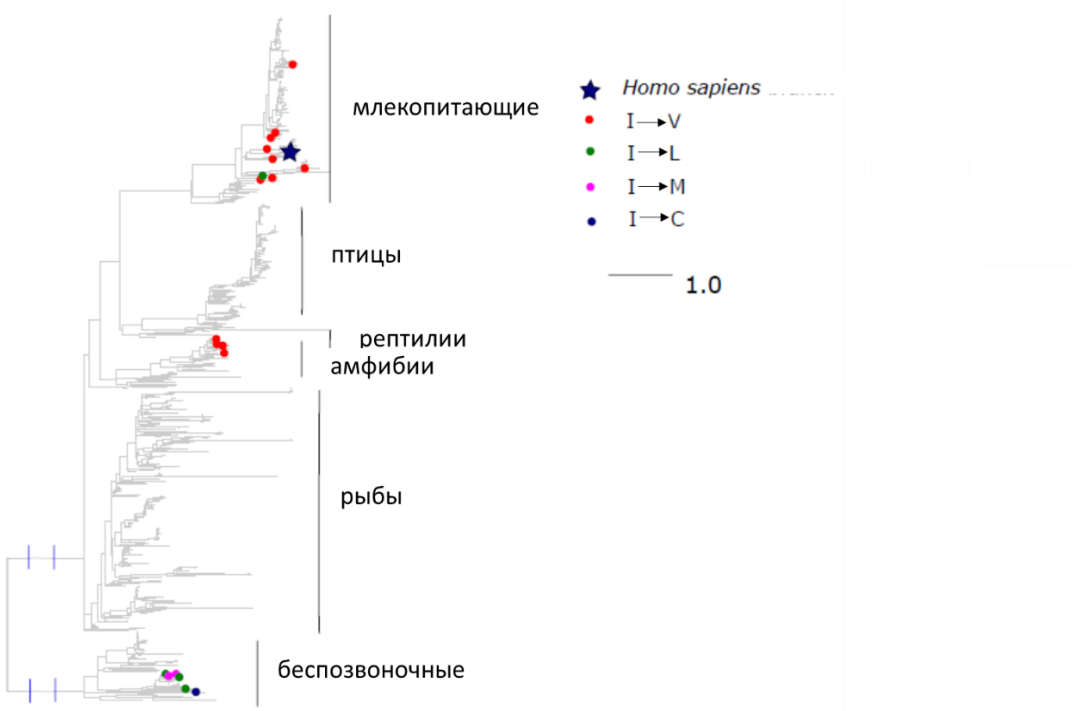


Рисунок 17. *Замены в сайте 33 белка ND6. Синяя звёздочка - H.sapiens; красные кружки, замены изолейцина (I) на валин (V), который является патогенным вариантом для человека; кружки других цветов - замены на другие аминокислоты. Филогенетические расстояния измерены в числе аминокислотных замен на сайт. Ветви, отмеченные синими линиями, укорочены на 3 единицы длины.*

3.3. Обсуждение

Патогенный для человека аминокислотный вариант может быть зафиксирован в других видах. Это можно объяснить разрешающими или компенсаторными заменами в остальном геноме [32, 34, 40]. Мы исследовали этот феномен с другой стороны: вариант, зафиксированный или являющийся полиморфизмом в популяции человека, может быть непредпочтительным в других видах. Мы обнаружили, что замены на аминокислоты, присутствующие в людях, происходят в эволюционно более близких к человеку видах, чем замены на другие аминокислоты. Несмотря на то, что избыток параллельных замен в филогенетически близких видах может возникнуть случайно, когда филогения для гена не соответствует видовой филогении [56], мы считаем, что эта причина маловероятна для наших данных. Мы показали ранее, что филогенетическая скластеризованность параллельных замен в митохондриальных белках не

является следствием ошибок в построении филогенетического дерева (раздел 2.2.3). Теоретически, ложные замены на встречающиеся у человека варианты могут появляться в филогенетически близком к нему виде (например, шимпанзе), если эти варианты на самом деле были полиморфизмами у общего предка человека и близкого вида, а затем зафиксировались в этом виде, но до сих пор остаются полиморфизмами у человека. Однако, есть сведения, что последний общий предок человеческих митохондриальных гаплотипов жил не ранее, чем 148 тысяч лет назад – гораздо позже разделения линий человека и шимпанзе [6-8 млн.л.н.; 2, 43], что исключает возможность общего митохондриального полиморфизма у этих видов. Таким образом, мы наблюдаем действительную кластеризацию замен. Вероятность аминокислотной замены зависит от связанного с ней коэффициента отбора [35]. Таким образом, наблюдаемый эффект может быть следствием уменьшения относительной приспособленности человеческого аллеля по мере увеличения филогенетического расстояния от человека.

На первый взгляд, для патогенных аминокислотных вариантов человека ситуация должна быть противоположной, ведь такие аминокислоты с большой вероятностью могут быть патогенными и в филогенетически близких к человеку видах, но разрешёнными в более далёких за счёт изменения геномного контекста или условий среды. Однако, в нашем исследовании замены на аминокислоты, патогенные для человека, сравнительно чаще происходят в эволюционно близких к человеку видах, чем в далёких от него.

Вклад в сходство филогенетического распределения замен на нейтральные и патогенные варианты могут вносить нейтральные мутации, неправильно классифицированные как патогенные [45]. Однако, ни одна из шести патогенных аминокислот со статусом «подтверждённые» в базе данных «Mitomap» не присутствует в базе данных митохондриальных

полиморфизмов «mtDB» [31] среди 2704 последовательностей из разных популяций человека, что говорит о их вероятной вредности. В то же время, мутации на эти аминокислоты значимо кластеризуются на филогении вблизи ветви человека (Рисунки 12а, 14).

Таким образом, замены на патогенные аллели человека действительно с большей вероятностью происходят в более родственных человеку видах. Этот результат отличается от результатов предыдущих исследований, где либо не было найдено зависимости частоты встречаемости патогенной аминокислоты от филогенетического расстояния [40, 79], либо было показано повышение вероятности замен на такие аминокислоты у более далёких от человека видов [26, 32]. Причин такого несоответствия может быть несколько. Во-первых, наши данные отличались от данных в предыдущих работах: мы брали митохондриальные, а не ядерные, белки, и рассматривали больший разброс филогенетических расстояний. Во-вторых, поскольку наш анализ возможен только при условии нескольких независимых замен на патогенный аллель, наша выборка сайтов может быть смещена в сторону меньшей консервативности. Однако, в своей предыдущей работе (раздел 2.2.2) мы не обнаружили зависимости степени филогенетической неравномерности гомоплазий от консервативности сайта (Приложение 3). В-третьих, проанализированные нами патогенные варианты могут снижать приспособленность не так сильно, как рассмотренные в предыдущих работах мутации, многие из которых приводили к потере функции гена. В каждой клетке имеется множество копий митохондриальной ДНК, и фенотипическое проявление патогенных мутаций в митохондриально-кодируемых генах зависит от доли мутантной ДНК [88]. Скорее всего, находясь в состоянии гомоплазмии (то есть, во всех копиях митохондриальной ДНК), приводящие к потере функции варианты будут летальными. Поэтому мы ожидаем, что самые вредные мутации будут

наблюдаться только в состоянии гетероплазмии, а патогенные мутации, не приводящие к полной потере функции, напротив, могут быть гомоплазмичными. Это подтверждается нашим наблюдением, что в базе данных «Mitomap» все мутации, приводящие к появлению стоп-кодона (нонсенс-мутации), отмечены как гетероплазмичные. Мы наблюдали филогенетическую кластеризацию вблизи ветки человека для патогенных аминокислот, которые встречаются в состоянии гомоплазмии, но не для аминокислот, которые наблюдались у человека только в виде гетероплазмических вариантов (Рисунок 12а). Таким образом, частота замен на аминокислоты, приводящие у человека к потере функции белка, может не меняться с филогенетическим расстоянием от ветви человека.

Почему патогенные для человека аминокислоты имеют наибольшую приспособленность в его ближайших родственниках? Наш анализ направлен на изучение относительной приспособленности, которая отражается в частоте замен на аминокислоту относительно замен на другие аминокислоты в сайте. Сравнение биохимических свойств аминокислотных вариантов помогает понять, почему замена на патогенный для человека вариант имеет больший шанс произойти в эволюционно более близких к человеку видах. Мы обнаружили, что патогенные аллели человека из нашей выборки в среднем более близки по биохимическим свойствам к референтному аллелю человека, чем аминоксислоты, которые наблюдались в сайте у других видов, но не встречаются у человека. Таким образом, в контексте человеческого генома, известный патогенный вариант может меньше нарушать структуру белка, и, следовательно, снижать приспособленность не так сильно, как вариант, не наблюдающийся в популяции человека. Поэтому многие не встречающиеся у людей варианты могут быть более вредны для человека и близкородственных ему видов, чем известные патогенные аллели, будучи при этом оптимальными в геномном контексте видов, где они закреплены.

Встречаемость аминокислоты в других видах часто используется как критерий вредности мутаций для человека [1, 42]. Многочисленные свидетельства вариабельности однопозиционных ландшафтов приспособленности говорят о том, что важно учитывать степень родства этих видов и человека. Наше наблюдение, что замены на патогенные аллели человека могут происходить чаще в эволюционно более близких к нему видах, свидетельствует о том, что связь патогенности аминокислоты и расположения замен на неё на филогении относительно человека может быть неоднозначной.

Заключение

Изучив филогенетическую кластеризацию замен на аминокислоты, встречающиеся у человека как референтные, полиморфные или патогенные варианты митохондриальных белков, мы имеем основания предполагать, что приспособленность таких вариантов часто различается между видами и выше в видах, более родственных человеку.

Глава 4. Однопозиционный адаптивный ландшафт белка оболочки ВИЧ1 и гемагглютинина вируса гриппа

Предыдущие разделы диссертации были посвящены свидетельствам вариабельности ОПАЛов, которые мы получили с применением совокупных статистик, использующих информацию из множества аминокислотных сайтов белков. Этот раздел описывает разработанный нами подход, позволяющий находить аминокислоты с разной приспособленностью в группах видов при наличии аминокислотных последовательностей и восстановленной филогении. Метод основан на анализе взаимного расположения гомоплазий (конвергентных и параллельных замен) и дивергентных замен на филогении: более высокая скорость замен на аминокислоту отражает её более высокую приспособленность в кладе. Мы применили свой метод к поверхностному белку gp160 ВИЧ1 из двух вирусных подтипов A и B и к гемагглютинину вируса гриппа A из подтипов H1 и H3, и показали, что полученные нами результаты согласуются с результатами экспериментов по глубокому мутационному сканированию. Также мы показали, что изменение ОПАЛов не всегда влечёт за собой положительный отбор.

4.1. Материалы и методы

4.1.1. Поиск аминокислот с вариабельной приспособленностью

Обозначим за d филогенетическое расстояние между фокальным видом и заменой на аминокислоту A, которое вычисляется как сумма длин ветвей между соответствующими точками на филогенетическом дереве (мы считаем, что все замены происходят на серединах филогенетических ветвей). Обозначим за $\bar{d}$ среднее d для всех замен на A. Если замены на A предпочтительно происходят на близких к фокальному виду ветвях, то $\bar{d}$ будет понижено по сравнению с нулевым ожиданием. Напротив, если

замены на A недопредставлены вблизи фокального вида на филогении, то $\bar{d}$ будет повышено. Мы получали нулевое распределение $\bar{d}$ для аминокислоты A с использованием замен на другие аминокислоты (nonA) в том же сайте. $\bar{d}$ зависит от филогенетического распределения предковых для A аминокислот и от числа замена на A [37]. Для контроля на эти факторы, мы брали подвыборки замен из предковых для A аминокислот следующим образом. Для каждой предковой аминокислоты, которая хотя бы раз заменялась на A в наших данных, из всех её замен на любую аминокислоту в сайте, мы брали число замен, равное числу замен этой аминокислоты на A. Для всех выбранных замен считали среднее филогенетическое расстояние $\tilde{d}$ до фокального вида. Мы повторяли взятие подвыборки 10000 раз для получения распределения $\tilde{d}$ - нулевого распределения для $\bar{d}$.

Значимость (p-значение) смещения замен к фокальному виду определяли как перцентиль $\bar{d}$ в этом распределении, а значимость смещения замен от фокального вида – как (1 – p).

Для заданной границы значимости α аминокислота называлась проксимальной для вида, если p < α, и дистальной, если p > (1–α). Аминокислоту считали «потенциально значимой» (проксимальной или дистальной), если минимальная достижимая значимость (то есть, доля нулевых подвыборок, попавших в крайнюю точку нулевого распределения) была меньше α. Мы определяли ожидаемую долю ложноположительных результатов (false discovery rate, FDR) как argmin ((α×PotVar)/VarFit, 1), где α – уровень значимости, PotVar – число потенциально значимых аминокислот для данного α, VarFit – число значимых аминокислот для данного α. Для gp160 мы использовали α=0.01. Процент ложноположительных результатов для этого уровня значимости не превышает 11%. Поскольку частота аминокислотных замен для

гемагглютинина гриппа ниже, при анализе этого белка мы использовали уровень значимости α = 0.05, для которого процент ложноположительных результатов <39%.

Наконец, для пары видов 1 и 2, мы рассматривали аминокислоту как «аминокислоту с вариабельной приспособленностью» (АВП), если она одновременно оказывалась проксимальной для одного вида и дистальной для другого (Рисунок 18). Все аминокислоты, которые были потенциально проксимальными для одного вида и потенциально дистальными для другого, мы относили к «аминокислотам с потенциально вариабельной приспособленностью» (потАВП).

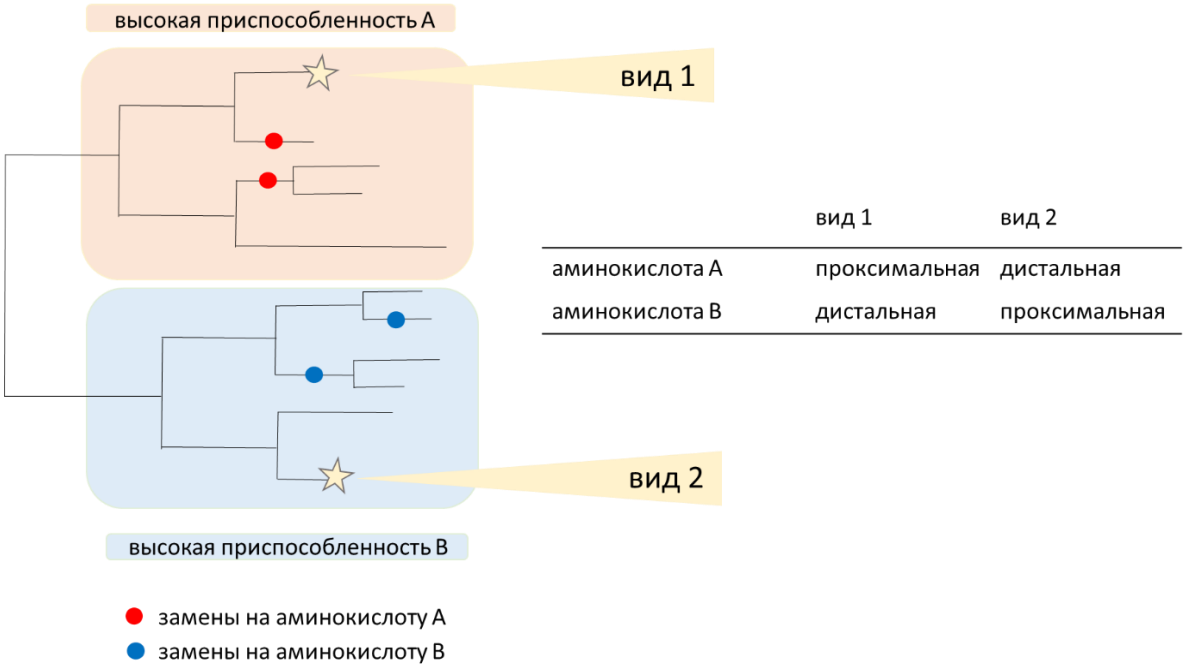


Рисунок 18. *Схематичное представление подхода для поиска меняющих приспособленность аминокислот. Здесь обе аминокислоты A и B определяются как АВП, поскольку каждая из них проксимальна для одного вида и дистальна для другого (изображение схематично. Для того, чтобы аминокислота считалась «проксимальной» или «дистальной», замены на неё должны происходить значимо ближе или дальше от фокального вида, чем ожидается).*

4.1.2. Данные по gp-160

Мы скачали из базы данных «Los Alamos HIV sequence database» [99] 4241 нуклеотидную последовательность и соответствующие

аминокислотные последовательности белка gp160 из девяти подтипов типа M ВИЧ1, используя следующие параметры: alignment type=filtered web, organism=HIV-1/SIVcpz, region=Env, subtype=M group without recombinants, year: 2016. Подтипы вируса были неравномерно представлены в базе данных, и скачанные последовательности в основном принадлежали подтипам B (>2000 изолятов) и C (>1000 изолятов). Мы отфильтровали последовательности с внутренними стоп-кодонами или длиной, не кратной трём, после чего у нас осталось 3789 последовательностей. Мы добавили к ним экспериментально изученные [24, 25] последовательности из двух штаммов, LAI (подтип B) и BF520 (подтип A), и последовательность из типа O (изолят ANT70) в качестве аутгрупа. Таким образом, объём наших данных составил 3792 последовательности. Аминокислотное и кодонное нуклеотидное выравнивания выполнили в программах «Pal2Nal» [83] и «Mafft» [33]. Используя нуклеотидное выравнивание, мы построили филогенетическое дерево методом максимального правдоподобия в программе «RAxML» [81] с моделью «GTRGAMMA». Затем в той же программе оптимизировали длины ветвей на основании аминокислотного выравнивания моделью «PROTGAMMAGTR» и реконструировали аминокислоты во внутренних узлах в программе «codeml» из пакета «PAML» [95].

Наш анализ был проведён с использованием аминокислотного выравнивания, фокальными видами служили LAI и BF520. Чтобы удостовериться в устойчивости наших результатов к ошибкам филогенетической реконструкции, поиск аминокислот с непостоянной приспособленностью параллельно выполнялся для десяти филогенетических деревьев, независимо построенных на основании одного и того же выравнивания. В дальнейшем анализировались только аминокислоты с вариабельной/постоянной приспособленностью, имевшие значимый или незначимый результат для всех десяти деревьев.

Координаты функциональных участков белка взяли из UniProt [86; entry P04578]. Информацию о доступности для растворителя аминокислотных остатков штамма LAI брали из [25]. Аминокислотные остатки с доступностью для растворителя ниже 0.25 считались погружёнными, выше – поверхностными.

4.1.3. Данные по гемагглютинину (ГА)

Мы скачали из базы данных «GISAID» [77] 2238 и 50109 нуклеотидных последовательностей ГА, принадлежащих соответственно подтипам H1N1 (сезонному) и H3N2 вируса гриппа А и отфильтровали последовательности короче 1680 нуклеотидов, не кратные трём или содержащие внутренние стоп-кодоны. Фильтрацию прошли 2142 последовательности H1 и 49543 последовательности H3. Затем в программе «cd-hit» [47] мы удалили из данных повторяющиеся последовательности, после чего осталось 1557 последовательностей из H1 и 20191 последовательность из H3. Наконец, мы случайно отобрали 1557 последовательностей из H3, чтобы выровнять представленность двух подтипов в данных. Белковые последовательности были получены из нуклеотидных с помощью скриптов «BioPerl» [80]. Все последующие действия с данными по ГА проводили так же, как для gp160 (см. раздел 4.1.2). Экспериментально измеренные аминокислотные предпочтения для штаммов A/WSN/1933 (H1N1) и A/Perth/16/2009 (H3N2) взяли из [44].

4.1.4. Симуляция эволюции

Мы провели симуляцию эволюции аминокислотной последовательности вдоль филогении для двух подтипов ВИЧ1 B и C. Филогенетическое дерево строили описанными выше методами, но только для последовательностей из подтипов B и C. Также мы сконструировали

маленькое дерево, убрав половину случайных терминальных ветвей, и большое дерево, добавив к каждой терминальной ветви по два потомка.

Симуляции проводили с помощью программы «SELVa» [Simulator of Evolution with Landscape Variation, 62], пользователь которой может указать ветви филогенетического дерева, где происходит смена ОПАЛа, и векторы приспособленности аминокислот до и после этой смены. «SELVa» симулирует аминокислотные замены согласно модели, описанной в [97]. При этом все скорости мутирования принимаются равными единице, поэтому различия в частоте замен определяются только различиями в приспособленности аминокислот.

Мы проводили симуляцию эволюции для 500 аминокислотных сайтов вдоль филогенетического дерева с вектором приспособленности аминокислот {X,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1} для значений X ϵ [3; 10] с шагом 1. При этом в подтипах B и C разные аминокислоты имели приспособленность X. Смена ОПАЛа происходила в предке подтипа C.

Для каждого X мы вычисляли два значения: 1. доля истинно положительных классификаций (True Positive Rate, TPR) – доля действительных аминокислот с переменной приспособленностью, которую обнаружил метод; 2. доля ложноположительных классификаций (False Positive Rate, FPR) – доля аминокислот с постоянной приспособленностью, которые метод ошибочно счёл аминокислотами с переменной приспособленностью.

4.1.5. Сравнение с данными глубокого мутационного сканирования

Чтобы проверить, совпадают ли наши результаты с полученными в результате экспериментов по глубокому мутационному сканированию (Deep mutational scanning, DMS), для каждой аминокислоты мы считали отношение (r) между её экспериментально измеренными приспособленностями в штамме, к которому она филогенетически более

близка (π_x), и в оставшемся штамме (π_y). Аминокислоту считали более близкой (но не «проксимальной») к штамму, в котором её p-значение меньше (независимо от значимости), поэтому отношение r могло быть посчитано для всех аминокислот. Затем мы сравнивали среднее значение r для АВП и остальных аминокислот с помощью непараметрического теста Уилкоксона.

4.1.6. Поиск сайтов под положительным отбором

Сайты gp160 тестировали на действие положительного отбора в программе «codeml» с помощью сайт-специфической модели. Для проверки наличия положительного отбора сравнивали модели M1a и M2a с помощью теста отношения правдоподобия. Вероятность нахождения каждого сайта под положительным отбором вычисляли эмпирическим методом Байеса (BEB, Bayes Empirical Bayes) в программе «codeml» независимо для филогенетических деревьев подтипа A, подтипа B и обоих подтипов вместе. Деревья для этого анализа были построены, как описано выше, но с использованием только последовательностей из подтипов A и/или B.

4.2. Результаты

4.2.1. Валидация метода

Для анализа возможностей нашего метода находить аминокислоты, меняющие свою приспособленность в сайте белка, мы использовали программу «SELVa» [62], которая позволяет симулировать эволюцию аминокислотных сайтов с переменным ландшафтом приспособленности. Скорости замен между аминокислотами определяются разницей в их приспособленности, которая может отличаться на разных участках филогенетического дерева. Мы предположили, что в каждом сайте a и в каждой точке филогении только одна аминокислота из двадцати, AA_k,

имеет популяционную приспособленность $F_a^{\mathrm{AA}_k} = 2N_e f_a^{\mathrm{AA}_k} = X > 1$, а остальные 19 аминокислот имеют приспособленности $F_a^{\mathrm{AA}_{i\neq k}} = 1$. Таким образом, замены $\mathrm{AA}_k \longrightarrow \mathrm{AA}_{i\neq k}$ происходят с скоростью $\frac{1-X}{1-e^{X-1}}$, а замены $\mathrm{AA}_{i\neq k} \longrightarrow \mathrm{AA}_k$ – со скоростью $\frac{X-1}{1-e^{1-X}}$ [97]. Мы моделировали вариабельность приспособленности аминокислот, предполагая, что смена адаптивного ландшафта происходит в середине одной заданной филогенетической ветки *b*; в этой точке случайно выбиралась новая аминокислота AA_l с приспособленностью X, а остальные 19 аминокислот получали приспособленность 1. В такой модели менялись приспособленности аминокислот AA_k и AA_l, а приспособленности остальных 18 аминокислот были постоянными.

Мы проверяли, может ли наш подход найти AA_k и AA_l в симулированных белковых последовательностях. Для симуляции мы брали филогенетическое дерево белка gp160, кодируемого геном *env* вируса ВИЧ1. Дерево состояло из двух клад, соответствующих вирусным подтипам B и C (Рисунок 19). Мы предполагали, что изменение значений относительной приспособленности аминокислот происходило между подтипами. Для того, чтобы изучить влияние количества данных на мощность метода, на основании имеющегося филогенетического дерева мы сконструировали дерево, где половина ветвей отсутствовала («разреженное дерево»), и дерево, где к каждой терминальной ветви были добавлены по 2 потомка («плотное дерево»). Вдоль каждого дерева для каждого целочисленного значения X ϵ [3;10] мы симулировали эволюцию пятисот сайтов. Результаты симуляций показаны на рисунке 20.

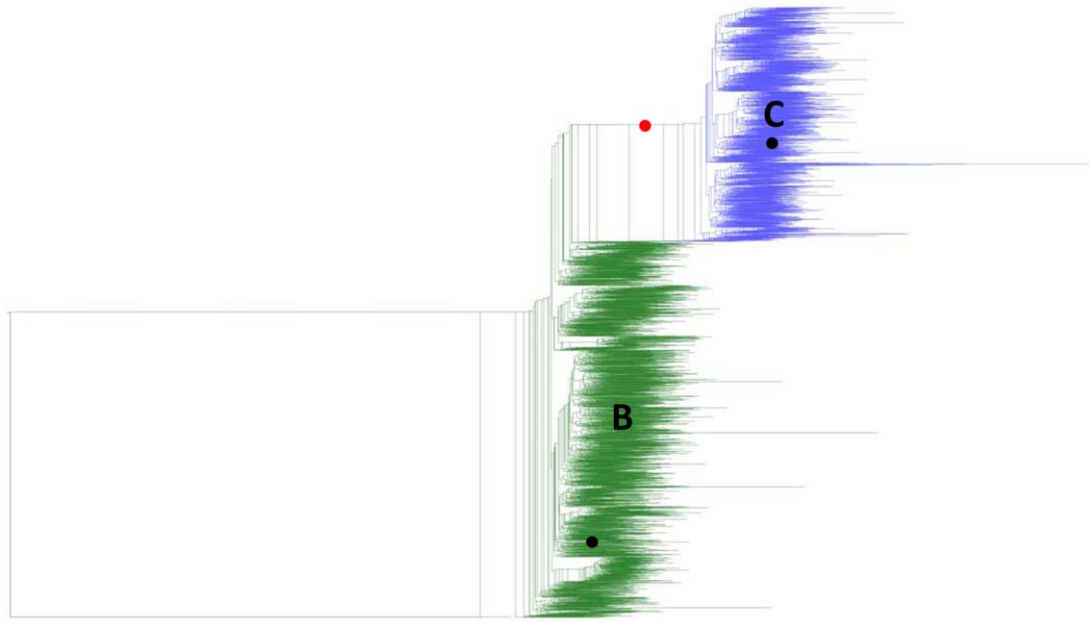


Рисунок 19. Филогенетическое дерево подтипов B и C ВИЧ1, которое мы использовали для симуляционного анализа. *Зелёная клада – подтип B; синяя– подтип C; чёрными точками обозначены два штамма, которые мы использовали как фокальные в симуляционном анализе; красная точка – ветка, на которой изменялся ОПАЛ в симуляции*

При средних коэффициентах отбора ($4 \leq X \leq 9$) наш метод правильно находит ~20%, ~40% и ~80% аминокислот с переменной приспособленностью для разреженного, исходного и плотного деревьев соответственно. При более сильном отборе ($X \approx 10$), многие сайты мономорфны, и чувствительность нашего метода падает. Однако, она по-прежнему зависит от количества данных, и для большого дерева находится 30% аминокислот с вариабельной приспособленностью. При более слабом отборе ($X<4$) чувствительность метода также падает, возможно, из-за того, что изменения таких адаптивных ландшафтов маскируются шумом, связанным с генетическим дрейфом и топологией дерева. Процент аминокислот с постоянной приспособленностью, ошибочно отнесённых к группе аминокислот с вариабельной приспособленностью, невысок для трёх деревьев и не превышает 0.6%.

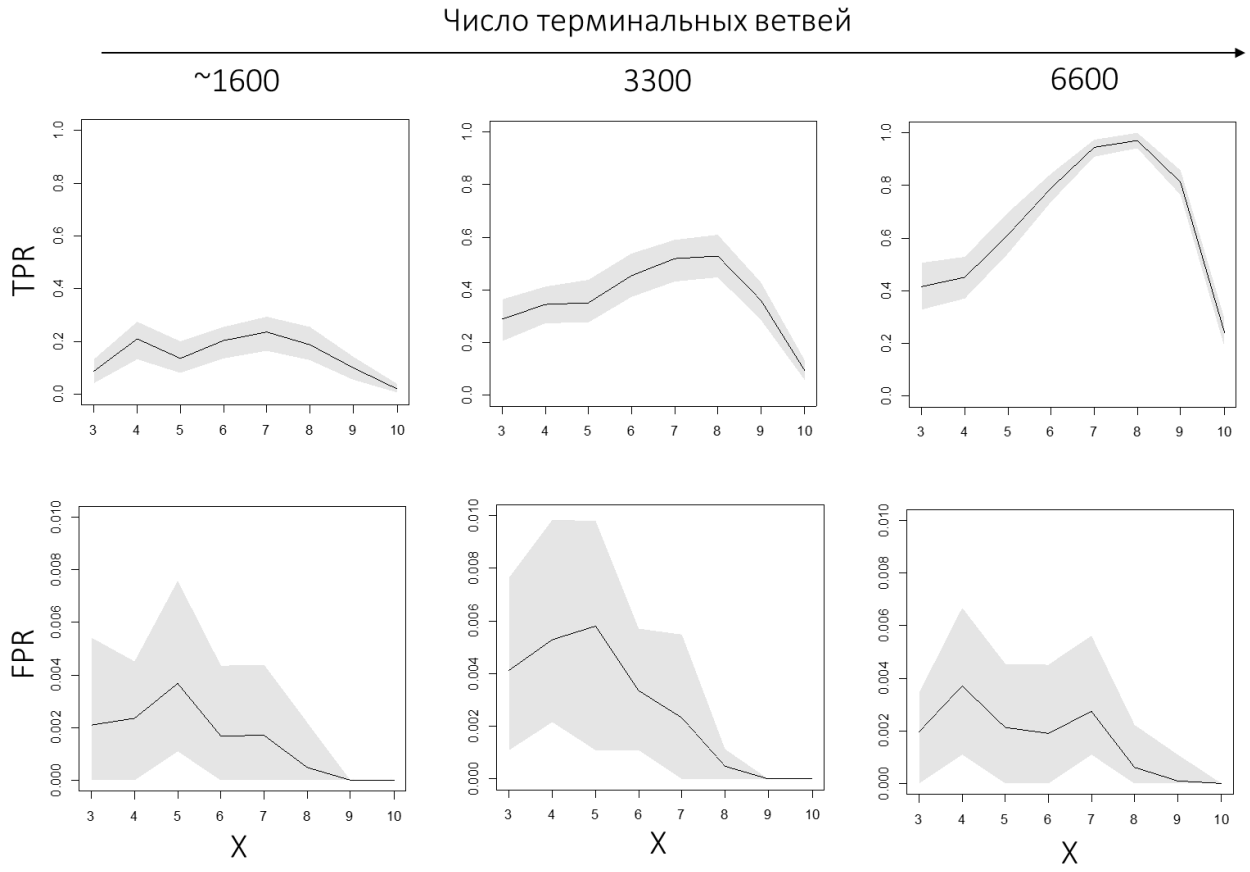


Рисунок 20. Результаты нашего подхода в симуляциях при разных значениях популяционной приспособленности (X = Ne×s) предпочтительного варианта; *верхний ряд - доля истинно положительных классификаций (True Positive Rate, TPR); нижний ряд - доля ложноположительных классификаций (False Positive Rate, TPR); чёрная линия – среднее, серая область – 90% доверительный интервал для ста случайных выборок по 1000 аминокислот из результатов симуляции*

4.2.2. ОПАЛ белка оболочки ВИЧ1

Чтобы проверить работу нашего метода на настоящих данных, мы применили его для изучения эволюции белка gp160 ВИЧ1. Этот белок эволюционирует под сильным положительным отбором [4, 94], что может отражать изменения сайт-специфических аминокислотных предпочтений. В качестве фокальных мы брали два штамма, для которых имеются данные DMS: BF520 из подтипа A и LAI из подтипа B.

Среди 3491 аминокислоты с достаточно информативной для нашего анализа историей замен (потАВП) мы нашли 59 АВП (при FDR = 11%, см.

раздел 4.1.1), распределённых по 46 из 724 выровненных аминокислотных позиций (Рисунок 21).

Замены на АВП располагались на филогении до 1.75 раз (среднее = 1.17, стандартное отклонение = 0.16) ближе к одному из фокальных штаммов, и до 1.90 (среднее = 1.19, стандартное отклонение = 0.19) раз дальше от другого, чем при случайном ожидании. Среди 59 АВП 36 аминокислот в 34 сайтах были проксимальными для штамма BF520 и дистальными для LAI, и 23 аминокислоты в 23 сайтах – проксимальными для LAI и дистальными для BF520.

Сайты с АВП находились во всех функциональных доменах и структурных участках белка, и их распределение не было сдвинуто в сторону какого-либо домена (точный тест Фишера, среднее двустороннее p-значение для функциональных доменов > 0.6).

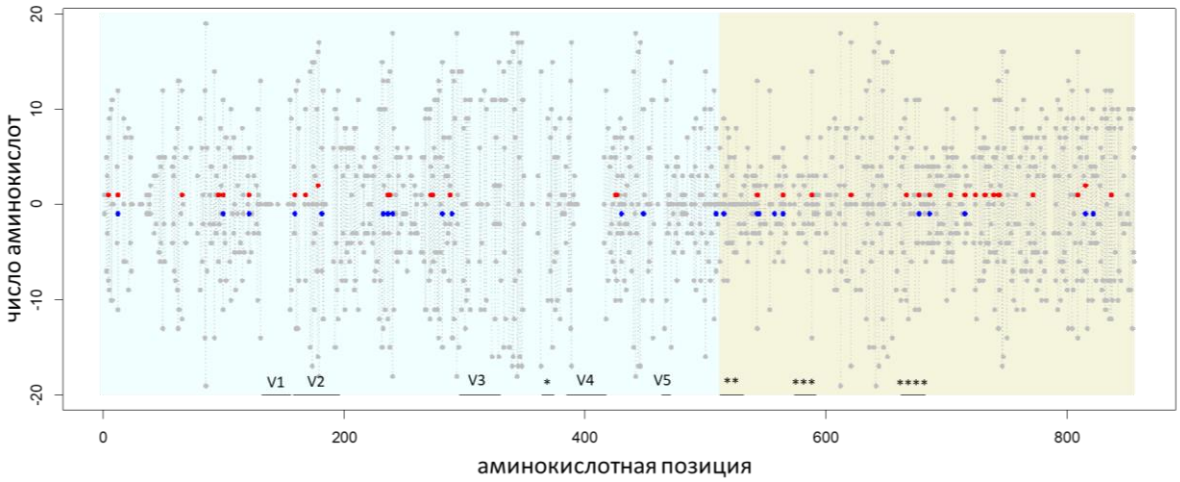


Рисунок 21. Аминокислотные сайты белка gp160, несущие АВП. *Для каждой аминокислотной позиции в белке (горизонтальная ось): красная точка - число аминокислот, проксимальных для BF520 и дистальных для LAI; синяя точка - число аминокислот, проксимальных для LAI и дистальных для BF520, умноженное на (-1). Серые точки обозначают число аминокислот в сайте, которые потенциально могут значимо менять приспособленность (потАВП, см. раздел 4.1.1). Позиции без точек – позиции, отфильтрованные из-за низкого качества выравнивания. Вертикальные серые линии соединяют точки для одного и того же сайта. Голубой фон - gp120; жёлтый фон - gp41. Вариабельные петли V1-V5 и функциональные домены по UniProt отмечены горизонтальными чёрными линиями: * - петля связывания с рецептором CD4; ** - пептид слияния; *** - иммуносупрессорный участок; **** - мембранный проксимальный внешний участок (membrane proximal external region, MPER). Нумерация белковых сайтов основана на штамме hxb2.*

4.2.3. Найденные с помощью филогении изменения приспособленности согласуются с данными DMS

Мы сравнили обнаруженные нашим методом изменения приспособленностей аминокислот между двумя штаммами с изменениями приспособленностей, определёнными в экспериментах DMS [24, 25]. Чтобы оценить изменения экспериментально измеренных приспособленностей, для каждой аминокислоты мы посчитали отношение $r = \frac{\pi_x}{\pi_y}$, где π_x – экспериментально измеренная приспособленность аминокислоты в штамме, к которому замены на неё происходят ближе, а π_y – экспериментально измеренная приспособленность той же аминокислоты в штамме, от которого замены на неё происходят дальше. В случае АВП, π_x – приспособленность, измеренная в штамме, для которого аминокислота является проксимальной, π_y – в штамме, для которого аминокислота является дистальной. В качестве контроля мы проанализировали распределение r в подвыборке аминокислот, для которых наш тест был незначимым («аминокислоты с постоянной приспособленностью», неАВП).

Значения r для АВП значительно превышали таковые в контроле (односторонний тест Уилкоксона, p = 0.005; Рисунок 22), что говорит о согласовании филогенетически и экспериментально определённых изменений аминокислотных приспособленностей. Это справедливо как для аминокислот с более высокой приспособленностью в штамме BF520, так и для аминокислот с более высокой приспособленностью в штамме LAI (Рисунок 22). Основной вклад в наблюдаемый эффект дают аминокислоты с невысокими экспериментальными приспособленностями ($\pi < 0.2$; Рисунок 22б), не являющиеся оптимальными в экспериментах DMS (Рисунок 22в).

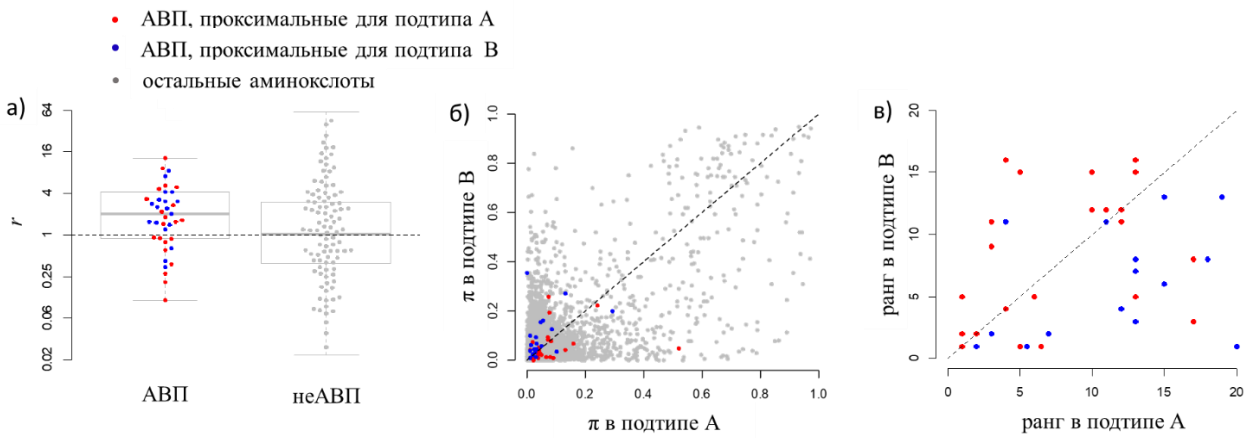


Рисунок 22. *а) диаграмма размаха для отношения экспериментально измеренных приспособленностей аминокислоты в штамме, к которому замены на неё происходят ближе, и в штамме, от которого замены на неё происходят дальше. Показаны аминокислоты со значимым (слева, $p<0.01$ для обоих штаммов) или незначимым (справа, $p>0.1$ для обоих штаммов) изменением приспособленности между штаммами; сплошная линия – медиана, границы прямоугольников обозначают межквартильный размах (IQR) между первым (Q1) и третьим (Q3) квартилями, усы обозначают интервалы Q1 – 1.5×IQR и Q3+1.5×IQR, выбросы не показаны; диаграммы построены по всем данным; каждая АВП, проксимальная к подтипу A или B, и подвыборка из 100 незначимых аминокислот показаны как красные, синие и серые точки соответственно. Пунктирная линия обозначает значение r при постоянной экспериментальной приспособленности (r =1).*
в, г) экспериментально измеренная приспособленность (б) или ранг приспособленности (в, ранг 1 соответствует наибольшей приспособленности) для каждой аминокислоты в двух штаммах ВИЧ1. Серые точки - аминокислоты, которые встречаются хотя бы в одной последовательности на нашем филогенетическом дереве; ***красные точки*** *- АВП, проксимальные к подтипу A;* ***синие точки*** *- АВП, проксимальные к подтипу B; пунктирная линия - биссектриса.*

4.2.4. Большинство сайтов с АВП не имеют признаков действия положительного отбора

Изменения приспособленности аллеля могут привести к эпизодам действия положительного отбора. Для исследования связи между этими процессами, мы нашли сайты, в которых отношение скоростей несинонимичной и синонимичной эволюции (dN/dS) превышает нейтральное значение по оценке программы «codeml» (ω >1) в подтипе A, подтипе B или обоих подтипах. Среди 46 сайтов, несущих АВП, только 20 (43%) испытывали положительный отбор в подтипе A и/или B, несмотря на то, что мощность теста dN/dS и мощность нашего теста возрастают при одном и том же условии увеличения числа замен. Ещё меньшее

перекрывание с АВП обнаружилось для сайтов под положительным отбором, описанных в предыдущих исследованиях ([94]: 17%; [92]: 9%). Несмотря на избыток сайтов под положительным отбором среди сайтов с АВП (точный тест Фишера, двусторонний P <0.001, Таблица 2), отсутствие свидетельств положительного отбора более, чем в половине сайтов с АВП, говорит о том, что эти явления часто могут быть несвязанными. Кроме того, r не был выше для АВП в сайтах под положительным отбором, чем для АВП в остальных сайтах (односторонний тест Уилкоксона, p = 0.09).

Таблица 2 *Сайты с АВП и сайты под положительным отбором*

подвыборка	**сайты с АВП**	**сайты без АВП**	**p-значение***	**источник**
сайты под + отбором	20	71	<0.001	наш анализ
остальные сайты	26	422		
сайты под + отбором	4	21	0.26	[92]
другие сайты	42	472		
сайты под + отбором	8	37	0.04	[94]
другие сайты	38	456		

Включены только сайты с потАВП; * - двусторонний точный тест Фишера

Более детальное рассмотрение конкретных случаев позволяет понять возможные взаимодействия между положительным отбором и изменением аминокислотных приспособленностей. Например, согласно результатам DMS, Y159 – наиболее предпочтительная аминокислота в подтипе A, тогда как в подтипе B наибольшую приспособленность имеет аминокислота F159. По результатам нашего теста, аминокислота Y159 является АВП, проксимальной для подтипа A. При этом в сайте 159 в программе «codeml» обнаружен положительный отбор в подтипе A (ω = 3.47), но не в подтипе B (ω = 0.10). Внутри подтипа A, замены F159Y составляют 81% от наблюдаемых в этом сайте. Таким образом, сигнал положительного отбора, который уловила программа «codeml», вызван заменами, которые представляют собой параллельную адаптацию (Рисунок 23а).

Другая ситуация наблюдается в сайте 273. Несмотря на то, что наш метод определил аминокислоту S273 как АВП, программа «codeml» не обнаружила в этом сайте положительного отбора. Замены R273S составляют только 5.7% всех замен в этом сайте в подтипе B, тогда как в остальной части дерева (включающей подтип A) они составляют 46.5%. По результатам DMS, приспособленность S273 также выше в подтипе A, чем в подтипе B. Но даже в подтипе A S273 – один из нескольких вариантов со сходными приспособленностями (Рисунок 23б). Таким образом, частота замен R273S в подтипе A повышена вследствие ослабления отрицательного отбора против 273S, а не положительного отбора в пользу этого варианта.

4.2.5. Различия адаптивных ландшафтов между тремя подтипами ВИЧ1

Наш подход может быть расширен для сравнения приспособленностей одной аминокислоты в нескольких подтипах вируса. Чтобы это продемонстрировать, мы применили разработанный метод для изучения аминокислот, меняющих приспособленность между тремя основными подтипами ВИЧ1, которые являются самыми распространёнными подтипами в Азии (подтип A), Северной Америке и Европе (B), а также Африке (C). Помимо 59 аминокислот, найденных нами при анализе для двух подтипов, мы обнаружили ещё 268 аминокислот, приспособленности которых отличаются между подтипом C и по крайней мере одним из подтипов A и B. Полученные данные доступны на странице http://makarich.fbb.msu.ru/galkaklink/hiv_landscape/, где пользователь может получить список аминокислот, приспособленности которых отличаются между выбранными подтипами.

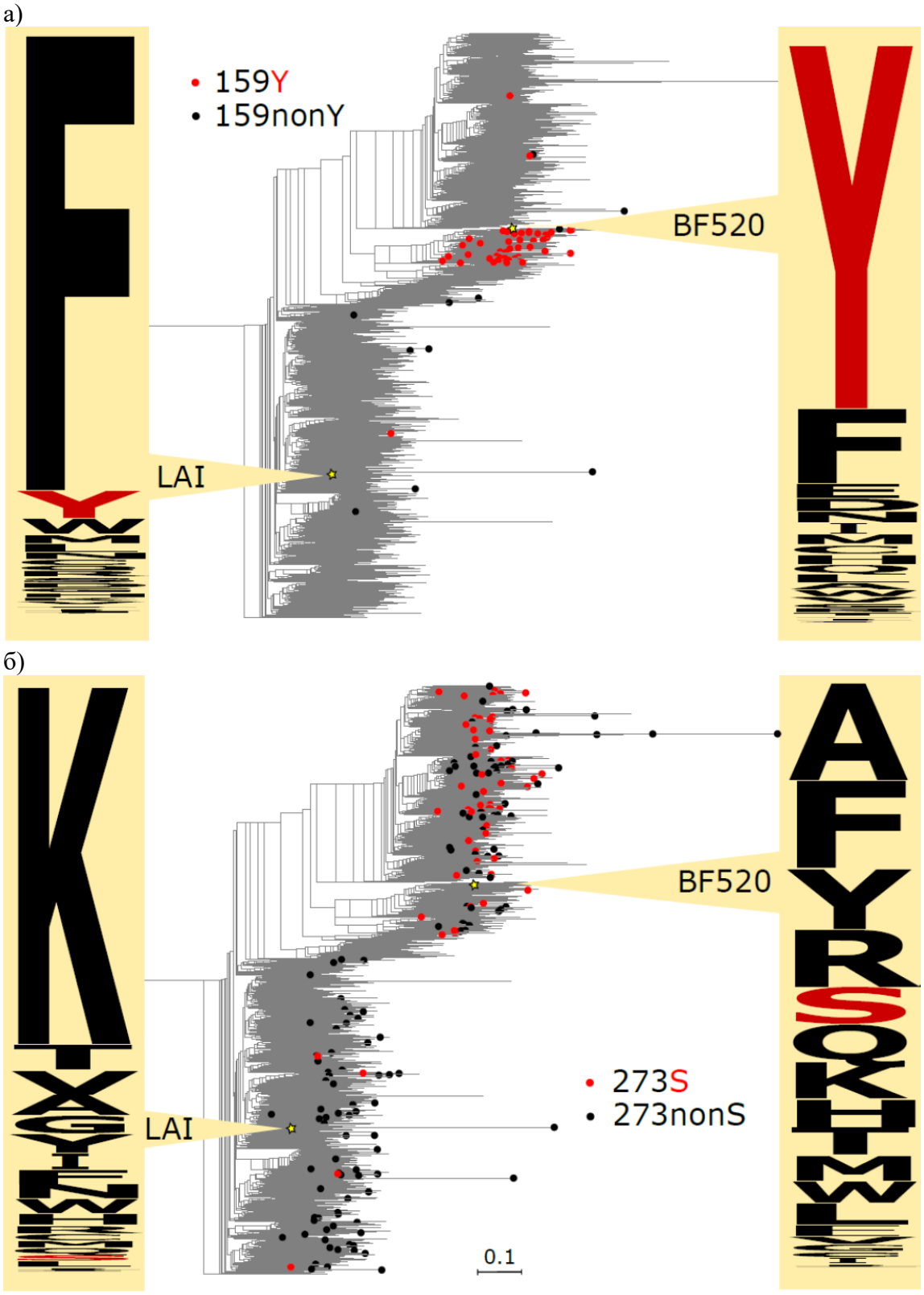


Рисунок 23. *Замены на АВП (красные кружки) и на другие аминокислоты (чёрные коужки) в сайтах 159 (а) и 273 (б) белка gp-160 и аминокислотные предпочтения в экспериментах DMS в штаммах BF520 и LAI. Филогенетические расстояния измерены в числе аминокислотных замен на сайт. Нумерация позиций белка основана на штамме*

hxb2. Картинки лого, отражающие экспериментально измеренные аминокислотные предпочтения, построены с помощью он-лайн сервиса Weblogo [11].

4.2.6. АВП гемагглютинина вируса гриппа А

Мы применили разработанный метод к гемагглютинину (ГА) подтипов H1 и H3 вируса гриппа А и нашли 9 АВП в 7 сайтах (FDR = 39%; рисунок 24).

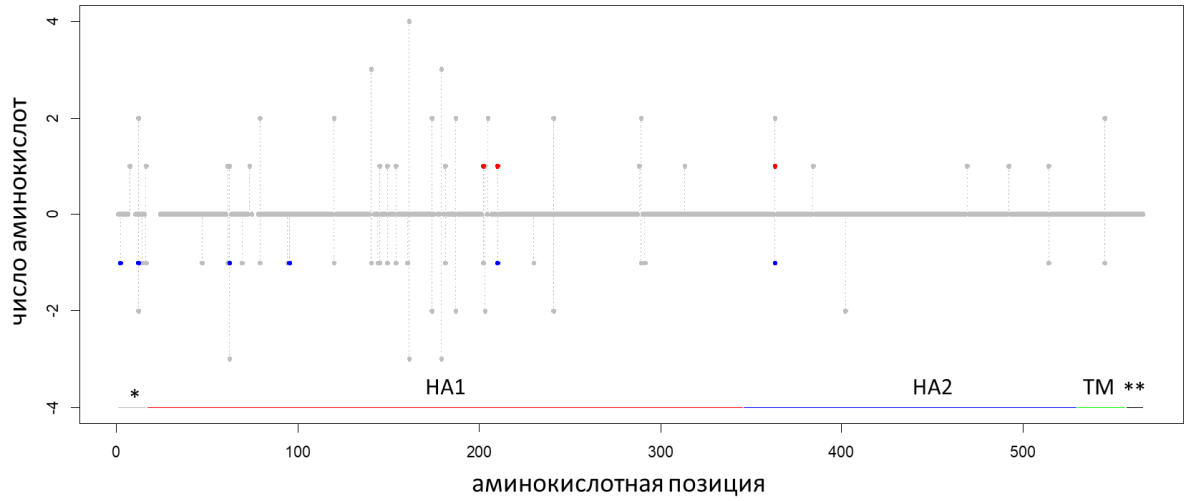


Рисунок 24 *Аминокислотные сайты ГА вируса гриппа А, несущие АВП. Для каждой позиции белка (горизонтальная ось) красная точка показывает число АВП, проксимальных для штамма из подтипа H1, а синяя точка – число АВП, проксимальных для штамма из подтипа H3, умноженное на (-1). Серые точки – число потАВП в сайте. Позиции, не имеющие точек – сайты с недостаточным качеством выравнивания. Вертикальные серые линии соединяют точки для одного сайта. Домены белка отмечены горизонтальными линиями: * - сигнальный пептид; HA1 - эктодомен HA1; HA2 - эктодомен HA2; TM - трансмембранный домен; ** - цитоплазматический хвост; нумерация позиций основана на штамме A/Perth/16/2009.*

Как и в случае с gp160, результаты нашего теста для ГА соответствовали данным эксперимента DMS: аминокислоты, замены на которые значимо кластеризовались на филогении ближе к одному из штаммов, имели в этом штамме более высокую экспериментально измеренную приспособленность по сравнению со вторым штаммом (односторонний тест Уилкоксона, P = 0.02; Рисунок 25).

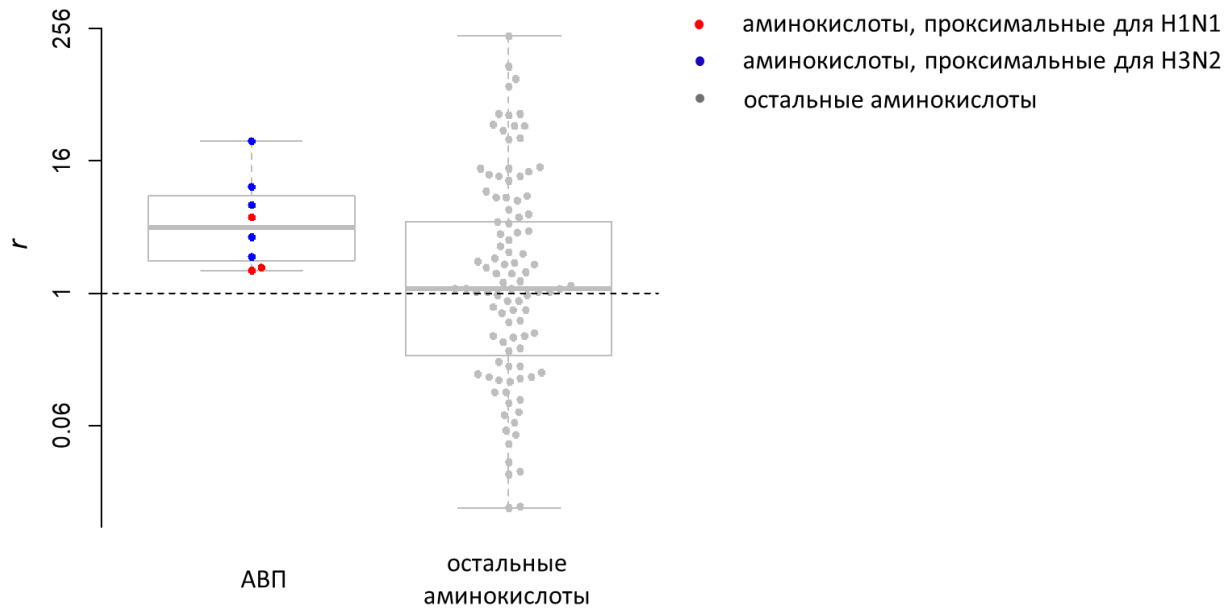


Рисунок 25. *Диаграмма размаха для отношения экспериментально измеренных приспособленностей аминокислоты ГА вируса гриппа А в штамме, к которому замены на неё происходят филогенетически ближе, и в штамме, от которого замены на неё происходят филогенетически дальше. Показаны аминокислоты со значимым (слева, $p<0.05$ для обоих штаммов) и незначимым (справа, $p>0.1$ для обоих штаммов) изменением приспособленности; обозначения – как на Рисунке 22.*

Для некоторых сайтов, изменения приспособленности АВП могут быть объясненены с функциональной точки зрения. Например, сайт 202 (в нумерации по A/Perth/16/2009) – один из двух сайтов с АВП, которые ранее были найдены как антигенные в подтипе H3N2, но не H1N1 [82]. S202 является предковой для клад H1 и H3. И наш анализ, и эксперимент DMS показали, что приспособленность варианта P202 выше в подтипе H1N1, чем в H3N2. Четыре замены S202P произошли в кладе H1, и 90% изолятов из подтипа H1N1 несут P202. Напротив, ни одной замены на P не произошло в кладе H3, и 95% изолятов из подтипа H3N2 несут G202, на который в кладе H3 сучилось три параллельные замены. Наблюдаемые изменения в ОПАЛе сайта могут отражать различия в функции сайта 202 между двумя вирусными подтипами.

4.3. Обсуждение

Паттерны замен, происходящих в процессе эволюции белковой последовательности, формируются под давлением действующего на белок отбора, и изменения этого давления могут приводить к неравномерности скоростей замен между эволюционирующими линиями. Многие подходы для поиска вариабельности отбора во времени, в основе которых лежат модели нестационарности скорости эволюции сайта в целом (гетеротахия, [49, 60, 96]) или скорости отдельных типов аминокислотных замен (гетеропециллия, [37, 38, 74, 84]), требовали объединения информации от многих сайтов для получения статистического сигнала. Таким образом, они не могли уловить изменения аминокислотных предпочтений в отдельных сайтах.

При наличии большого числа последовательностей с установленными филогенетическими связями можно повысить разрешение таких методов до отдельных аминокислотных сайтов. Этот раздел диссертации был посвящен разработанному нами подходу для поиска в индивидуальных сайтах белка аминокислот, которые имеют неодинаковую приспособленность в двух видах.

Наш метод оценивает изменение частот конвергентных (или параллельных) и дивергентных замен между эволюционирующими линиями. В то время как высокая частота замен на одну и ту же аминокислоту может говорить о её высокой приспособленности, использование дивергентных замен обеспечивает внутренний контроль, делающий метод устойчивым к другим типам нестационарности. В частности, он нечувствителен к различиям в общей скорости замен между эволюционными линиями. Кроме того, наш метод, рассматривающий отдельно каждый сайт, нечувствителен к различиям в скоростях и паттернах замен между сайтами. А благодаря контролю на предковые

аминокислоты, он нечувствителен и к различиям в скоростях разных типов замен (например, транзиций и трансверсий) в сайте. По той же причине он нечувствителен к изменениям в скорости замен на аминокислоту из-за изменения в распространённости разных предковых аминокислот [37].

Наш метод имеет и некоторые ограничения [37]. Во-первых, поскольку он основан на анализе филогенетического распределения аминокислотных замен, его мощность лимитирована числом наблюдаемых замен, что, в свою очередь, зависит от формы филогенетического дерева и режима отбора в сайте. В частности, он не применим для поиска изменений приспособленности аминокислот в близких видах или медленно эволюционирующих сайтах. Мы изучили возможности нашего метода при разной силе отбора с помощью симуляций эволюции. Согласно симуляции, метод лучше всего работает при отборе средней силы. Это легко объяснимо. При слишком слабом отборе замены и на полезные, и на вредные мутации происходят со сходной частотой в кладах с разными ОПАЛами, не позволяя нашему методу определить изменения в приспособленности. С другой стороны, при очень сильном отборе падает число сайтов, к которым метод применим: если аминокислота зафиксирована почти на всей кладе, где она предпочтительна, у нас не будет замен для анализа. В частности, наш метод не обнаруживает отбор в пользу высоко консервативного предкового варианта, поскольку замены с него, а значит, и на него редки. Таким образом, мы можем искать АВП только в достаточно вариабельных сайтах, среди которых мощность нашего метода достаточно высока.

Во-вторых, наш тест определяет любые изменения в относительной частоте замен независимо от причины этих изменений. Такие изменения могут быть вызваны не только непостоянством ОПАЛа (причина на аминокислотном уровне), но и причинами на нуклеотидном уровне: изменения мутационных предпочтений или отбор на нуклеотидный состав.

Теоретически, для контроля на такие причины можно использовать синонимические сайты.

Чтобы проиллюстрировать и проверить наш подход, мы использовали данные двух экспериментов DMS для поверхностных вирусных белков. В каждом эксперименте измеряли аминокислотные предпочтения белка в двух штаммах. В целом, наши результаты согласуются с экспериментальными: если замены на аминокислоту склонны происходить в филогенетическом окружении одного из штаммов, её экспериментально оцененная приспособленность в этом штамме выше, чем в другом.

Обнаруженные нами АВП, в основном, не обладают намного большей по сравнению с другими аминокислотами экспериментально измеренной приспособленностью π даже в штамме, для которого они являются проксимальными. Причин этого может быть несколько. Во-первых, АВП, обнаруженные нашим тестом, не обязаны быть самыми приспособленными на каком-то участке филогении. В частности, наш метод может обнаружить изменение в приспособленности только для производной, но не предковой аминокислоты. Например, экспериментальная приспособленность аминокислоты 159F белка gp160 выше в штамме LAI по сравнению со штаммом BF520. Однако поскольку этот вариант является предковым, и замены на него редки в обоих подтипах, наш филогенетический метод его не обнаруживает (Рисунок 23а). Аминокислота, приспособленность которой в кладе намного выше, чем у других вариантов, скорее всего будет преобладать в ней, что отразится в маленьком числе замен. Поэтому наш метод, требующий наличия достаточного числа замен, будет лучше работать для аминокислоты, приспособленность которой в кладе сравнима с другим вариантом (или несколькими).

Во-вторых, многие изменения ОПАЛов между подтипами А и В ВИЧ-1 не затрагивают аминокислоты с наибольшей приспособленностью. Среди 42 АВП, для которых имеются результаты DMS, только 6 (14%) имели в штамме из подтипа, к которому они проксимальны, наивысший ранг в векторе экспериментальных аминокислотных предпочтений в сайте (Рисунок 22, в). 34 аминокислоты субоптимальны в обоих штаммах. Таким образом, наш метод позволяет обнаружить изменения приспособленности субоптимальных аминокислот, что важно для понимания роли непостоянства адаптивных ландшафтов в эволюции.

Эволюционная модель, учитывающая результаты DMS экспериментов [ExpCM; 29], превосходит по точности другие модели аминокислотных замен. Это достигается оценкой силы отрицательного отбора из-за функциональных ограничений индивидуально для каждого сайта [24, 44]. Изменения наблюдаемых нами частот замен могут быть вызваны непостоянством силы отрицательного отбора между штаммами. Кроме того, наблюдаемые изменения аминокислотных предпочтений могут вести за собой эпизоды положительного отбора, если новый предпочтительный вариант был редким в предковой популяции. По нашим результатам, стандартный dN/dS тест способен детектировать положительный отбор менее, чем в половине сайтов с АВП. Таким образом, изменения ОПАЛа не всегда связаны с сильным положительным отбором на новые оптимальные варианты.

Наш метод может быть применён для двух и более фокальных видов при поиске аминокислот, относительная приспособленность которых сходна в одних эволюционных линиях и отлична в других. Мы иллюстрируем это на примере трёх подтипов ВИЧ-1. Поскольку число штаммов с доступными последовательностями значительно отличается между этими тремя кладами (223, 2035 и 1265 для подтипов A, B и C соответственно), мощность нашего теста для них различна. Но наши

результаты показывают, что при наличии достаточного объема данных можно сравнивать приспособленности одной аминокислоты между несколькими кладами. При этом, в отличие от других методов определения аминокислот с вариабельной приспособленностью, наш тест не требует знания точки изменения ОПАЛа [84] и не использует модели эпистатических взаимодействий или изменения приспособленности [51, 84].

Предсказание влияния мутаций на приспособленность часто основано на консервативности аминокислотной позиции [1, 36, 78], но функционально важные позиции не обязательно консервативны [93]. Наш метод может позволить отличить сайты с ослабленным отбором от сайтов с непостоянным ОПАЛом. Например, в gp160 в наших данных есть 10 сайтов, в каждом из которых на филогении встречается 18 разных аминокислот. Несмотря на то, что это может быть свидетельством нейтральной эволюции, по результатам нашего теста половина этих сайтов несёт АВП.

Несмотря на недавнее развитие экспериментов DMS, позволяющих проводить измерение приспособленности для большинства одиночных мутаций белка, проведение мутационного сканирования возможно только для ограниченного числа модельных систем, а из-за ресурсоёмкости эксперименты DMS нельзя провести для большого числа штаммов. Кроме того, отбор, действующий на организм в условиях лаборатории, может отличаться от отбора, под давлением которого организм эволюционирует в природе [25]. Напротив, сравнительная геномика позволяет находить изменения естественного отбора, повлиявшие на эволюцию многих немодельных видов.

Заключение

Мы разработали метод поиска аминокислот, меняющих приспособленность между несколькими участками филогенетического дерева, и с помощью симуляций эволюции определили возможности нашего метода при разной силе отбора. С помощью разработанного метода мы нашли для поверхностных белков вирусов ВИЧ1 и гриппа А аминокислоты, меняющие приспособленность между подтипами. Наши результаты согласуются с экспериментальными. Таким образом, разработанный метод может быть применён и к другим системам для поиска аминокислот с вариабельной приспособленностью при средней силе отбора, действующего на сайт.

Выводы

1. В митохондриальных белках *Opisthokonta* филогенетическое расстояние между параллельными заменами в среднем на 20% меньше расстояния между дивергентными заменами. Это является свидетельством вариабельности адаптивных ландшафтов митохондриальных белков.

2. В среднем приспособленность референтных, нейтральных и патогенных (но не летальных) аллелей митохондриальных белков человека снижается с филогенетическим расстоянием от ветви *Homo sapiens.* Таким образом, приспособленность встречающихся в популяциях человека аминокислот выше в видах, более эволюционно близких к человеку. Обнаруженный в популяциях человека патогенный вариант может требовать меньше изменений в других сайтах белка, чтобы стать разрешённым, чем вариант, не наблюдаемый в человеке.

3. Разработан и проверен на экспериментальных данных метод поиска индивидуальных аминокислот, приспособленность которых меняется между двумя участками филогенетического дерева. Найдены аминокислоты поверхностных белков ВИЧ-1 и вируса гриппа А, меняющие приспособленность между вирусными подтипами. Список таких аминокислот для поверхностного белка из подтипов A, B, C ВИЧ-1 представлен в виде веб-сервера. Эта информация может помочь в изучении отличий между вирусными подтипами, приводящих к разному течению болезни и ответу на лечение. Показано, что изменение однопозиционного адаптивного ландшафта не всегда влечёт за собой сильный положительный отбор.

Благодарности

Автор выражает самую искреннюю благодарность своему научному руководителю за терпение и веру, коллегам за поддержку и семье за закрытие путей к отступлению. Отдельное спасибо автор выражает Ивану Николаевичу Семенкову, который добровольно прочёл сей труд не один раз.

Список литературы

1. Adzhubei, I., Jordan, D.M., Sunyaev, S.R. Chapter 7. Predicting functional effect of human missense mutations using PolyPhen-2 / I. Adzhubei, D.M. Jordan, S.R. Sunyaev // *Current Protocols in Human Genetics*. 2013. – Unit7.20. – doi:10.1002/0471142905.hg0720s76.

2. Amster, G., Sella, G. Life history effects on the molecular clock of autosomes and sex chromosomes / G. Amster, G. Sella // *Proceedings of the National Academy of Sciences of the United States of America*. 2016. – Vol. 113. – P. 1588–1593. – doi:10.1073/pnas.1515798113.

3. Andrews, R.M. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA / R.M. Andrews, I. Kubacka, P.F. Chinnery et al. // *Nature Genetics*. 1999. – Vol. 23. – P. 147. – doi:10.1038/13779.

4. Bazykin, G.A. Bursts of nonsynonymous substitutions in HIV-1 evolution reveal instances of positive selection at conservative protein sites / G.A. Bazykin, J. Dushoff, S.A. Levin, A.S. Kondrashov // *Proceedings of the National Academy of Sciences of the United States of America*. – 2006. – Vol. 103. – P. 19396–19401. – doi:10.1073/pnas.0609484103.

5. Bazykin, G.A. Changing preferences: deformation of single position amino acid fitness landscapes and evolution of proteins / G.A. Bazykin // *Biology Letters*. – 2015. – Vol. 11. – doi:10.1098/rsbl.2015.0315.

6. Bazykin, G.A., Kondrashov, A.S. Detecting past positive selection through ongoing negative selection / G.A. Bazykin, A.S. Kondrashov // *Genome Biology and Evolution*. – 2011. – Vol. 3. – P. 1006–1013. – doi:10.1093/gbe/evr086.

7. Bloom, J.D., Gong, L.I., Baltimore, D. Permissive secondary mutations enable the evolution of influenza oseltamivir resistance / J.D. Bloom, L.I. Gong,

D. Baltimore // *Science (New York, N.Y.)*. – 2010. – Vol. 328. – P. 1272–1275. – doi:10.1126/science.1187816.

8. Boucher, J.I., Bolon, D.N.A., Tawfik, D.S. Quantifying and understanding the fitness effects of protein mutations: Laboratory versus nature / J.I. Boucher, D.N.A. Bolon, D.S. Tawfik // *Protein Science: A Publication of the Protein Society*. – 2016. – Vol. 25. – P. 1219–1226. – doi:10.1002/pro.2928.

9. Breen, M.S. Epistasis as the primary factor in molecular evolution / M.S. Breen, C. Kemena, P.K. Vlasov et al. // *Nature*. – 2012. – Vol. 490. – P. 535–538. – doi:10.1038/nature11510.

10. Callahan, B. Correlated evolution of nearby residues in Drosophilid proteins / B. Callahan, R.A. Neher, D. Bachtrog, et al. // *PLoS genetics*. – 2011. – Vol. 7. – e1001315. – doi:10.1371/journal.pgen.1001315.

11. Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E. WebLogo: a sequence logo generator / G.E. Crooks, G. Hon, J.-M. Chandonia, S.E. Brenner // *Genome Research*. – 2004. – Vol. 14. – P. 1188–1190. – doi:10.1101/gr.849004.

12. Doud, M.B., Ashenberg, O., Bloom, J.D. Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs / M.B. Doud, O. Ashenberg, J.D. Bloom // *Molecular Biology and Evolution*. – 2015.– Vol. 32. – P. 2944–2960. – doi:10.1093/molbev/msv167.

13. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput/ R.C. Edgar // *Nucleic Acids Research*. – 2004. – Vol. 32. – P. 1792–1797. – doi:10.1093/nar/gkh340.

14. Ferguson, A.L. Translating HIV sequences into quantitative fitness landscapes predicts viral vulnerabilities for rational immunogen design / A.L. Ferguson, J.K. Mann, S. Omarjeeet al. // *Immunity*. – 2013. – Vol. 38. – P. 606–617. – doi:10.1016/j.immuni.2012.11.022.

15. Figliuzzi, M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1 / M. Figliuzzi, H. Jacquier,

A. Schug, et al. // *Molecular Biology and Evolution*. – 2016. – Vol. 33. – P. 268–280. – doi:10.1093/molbev/msv211.

16. Firnberg, E. A Comprehensive, High-Resolution Map of a Gene's Fitness Landscape / E. Firnberg, J.W. Labonte, J.J. Gray, M. Ostermeier // *Molecular Biology and Evolution*. – 2016. – Vol. 33. – P. 1378. – doi:10.1093/molbev/msw021.

17. Fitch, W.M. Rate of change of concomitantly variable codons / W.M. Fitch // *Journal of Molecular Evolution*. – 1971.– Vol. 1. – P. 84–96.

18. Fitch, W.M., Markowitz, E. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution / W.M. Fitch, E. Markowitz // *Biochemical Genetics*. – 1970. – Vol. 4. – P. 579–593.

19. Flynn, W.F. Inference of Epistatic Effects Leading to Entrenchment and Drug Resistance in HIV-1 Protease / W.F. Flynn, A. Haldane, B.E. Torbett, R.M. Levy // *Molecular Biology and Evolution*. – 2017. – Vol. 34. – P. 1291–1306. – doi:10.1093/molbev/msx095.

20. Fowler, D.M. High-resolution mapping of protein sequence-function relationships / D.M. Fowler, C.L. Araya, S.J. Fleishman et al. // *Nature Methods*. – 2010. – Vol. 7. – P. 741–746. – doi:10.1038/nmeth.1492.

21. Fowler, D.M., Fields, S. Deep mutational scanning: a new style of protein science / D.M. Fowler, S. Fields // *Nature Methods*. – 2014. – Vol. 11. – P. 801–807. – doi:10.1038/nmeth.3027.

22. Goldstein, R.A. Nonadaptive Amino Acid Convergence Rates Decrease over Time / R.A. Goldstein, S.T. Pollard, S.D. Shah, D.D. Pollock // *Molecular Biology and Evolution*. – 2015. – Vol. 32. – P. 1373–1381. – doi:10.1093/molbev/msv041.

23. Gong, L.I., Suchard, M.A., Bloom, J.D. Stability-mediated epistasis constrains the evolution of an influenza protein / L.I. Gong, M.A. Suchard, J.D. Bloom, // *eLife*. – 2013. – Vol. 2. – doi:10.7554/eLife.00631.

24. Haddox, H.K. Mapping mutational effects along the evolutionary landscape of HIV envelope / H.K. Haddox, A.S. Dingens, S.K. Hilton et al. // *eLife*. – 2018. – Vol. 7. – doi:10.7554/eLife.34420.

25. Haddox, H.K., Dingens, A.S., Bloom, J.D., Experimental Estimation of the Effects of All Amino-Acid Mutations to HIV's Envelope Protein on Viral Replication in Cell Culture / H.K. Haddox, A.S. Dingens, J.D. Bloom, // *PLoS pathogens*. – 2016. – Vol. 12. – e1006114. – doi:10.1371/journal.ppat.1006114.

26. Harpak, A., Bhaskar, A., Pritchard, J.K. Mutation Rate Variation is a Primary Determinant of the Distribution of Allele Frequencies in Humans / A. Harpak, A. Bhaskar, J.K. Pritchard // *PLoS genetics*. – 2016. – Vol. 12. – e1006489. – doi:10.1371/journal.pgen.1006489.

27. Hartman, E.C., Tullman-Ercek, D. Learning from protein fitness landscapes: a review of mutability, epistasis, and evolution / E.C. Hartman, D. Tullman-Ercek, // *Current Opinion in Systems Biology*. – 2019. – Vol. 14. – P. 25–31. – doi:10.1016/j.coisb.2019.02.006.

28. Hietpas, R.T., Jensen, J.D., Bolon, D.N.A. Experimental illumination of a fitness landscape / R.T. Hietpas, J.D. Jensen, D.N.A. Bolon // *Proceedings of the National Academy of Sciences of the United States of America*. – 2011. – Vol. 108. – P. 7896–7901. – doi:10.1073/pnas.1016024108.

29. Hilton, S.K., Doud, M.B., Bloom, J.D. phydms: software for phylogenetic analyses informed by deep mutational scanning / S.K. Hilton, M.B. Doud, J.D. Bloom // *PeerJ*. – 2017. – Vol. 5. – e3657. – doi:10.7717/peerj.3657

30. Hopf, T.A. Mutation effects predicted from sequence co-variation / T.A. Hopf, J.B. Ingraham, F.J. Poelwijk et al. // *Nature Biotechnology*. – 2017. – Vol. 35. – P. 128–135. – doi:10.1038/nbt.3769.

31. Ingman, M., Gyllensten, U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences / M. Ingman, U. Gyllensten // *Nucleic Acids Research*. – 2006.– Vol. 34. – D749-751. – doi:10.1093/nar/gkj010.

32. Jordan, D.M. Identification of cis-suppression of human disease mutations by comparative genomics / D.M. Jordan, S.G. Frangakis, C. Golzio // *Nature*. – 2015. – Vol. 524. – P. 225–229. – doi:10.1038/nature14497.

33. Katoh, K., Misawa, K., Kuma, K., Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform / K. Katoh, K. Misawa, K. Kuma, T. Miyata // *Nucleic Acids Research*. – 2002. – Vol. 30. – P. 3059–3066. – doi:10.1093/nar/gkf436.

34. Kern, A.D., Kondrashov, F.A. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial tRNAs / A.D. Kern, F.A. Kondrashov // *Nature Genetics*. – 2004. – Vol. 36. – P. 1207–1212. – doi:10.1038/ng1451.

35. Kimura, M. *The neutral theory of molecular evolution* / Kimura, M., Cambridge: Cambridge University Press, 1983. 367 p. – doi: https://doi.org/10.1017/S0016672300025957.

36. Kircher, M. A general framework for estimating the relative pathogenicity of human genetic variants / M. Kircher, D.M. Witten, P. Jain // *Nature Genetics*. – 2014. – Vol. 46. – P. 310–315. – doi:10.1038/ng.2892.

37. Klink, G. V., Bazykin, G.A. Parallel Evolution of Metazoan Mitochondrial Proteins / G. V. Klink, G.A. Bazykin // *Genome Biology and Evolution*. – 2017. – Vol. 9. – P. 1341–1350. – doi:10.1093/gbe/evx025.

38. Klink, G. V., Golovin, A. V., Bazykin, G.A., Substitutions into amino acids that are pathogenic in human mitochondrial proteins are more frequent in lineages closely related to human than in distant lineages / G. V. Klink, A. V. Golovin, G.A. Bazykin // *PeerJ*. – 2017. – Vol. 5. – e4143. – doi:10.7717/peerj.4143.

39. Kondrashov, A.S. Rate of sequence divergence under constant selection / A.S. Kondrashov, I.S. Povolotskaya, D.N. Ivankov, F.A. Kondrashov // *Biology Direct*. – 2010. – Vol. 5. – P. 5. doi:10.1186/1745-6150-5-5.

40. Kondrashov, A.S., Sunyaev, S., Kondrashov, F.A. Dobzhansky-Muller incompatibilities in protein evolution / A.S. Kondrashov, S. Sunyaev, F.A. Kondrashov, // *Proceedings of the National Academy of Sciences of the United States of America*. – 2002. – Vol. 99. – P. 14878–14883. – doi:10.1073/pnas.232565499.

41. Kumar, A. Stability-Mediated Epistasis Restricts Accessible Mutational Pathways in the Functional Evolution of Avian Hemoglobin / A. Kumar, C. Natarajan, H.Moriyama // *Molecular Biology and Evolution*. – 2017. – Vol. 34. – P. 1240–1251. – doi:10.1093/molbev/msx085.

42. Kumar, S., Dudley, J.T., Filipski, A., Liu, L. Phylomedicine: an evolutionary telescope to explore and diagnose the universe of disease mutations / S. Kumar, J.T. Dudley, A. Filipski, L. Liu // *Trends in genetics*: TIG. – 2011. – Vol. 27. – P. 377–386. – doi:10.1016/j.tig.2011.06.004.

43. Langergraber, K.E. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution / K.E. Langergraber, K. Prüfer, C. Rowney et al. // *Proceedings of the National Academy of Sciences of the United States of America*. – 2012. – Vol. 109. – P. 15716–15721. – doi:10.1073/pnas.1211740109.

44. Lee, J.M., Huddleston, J., Doud, M.B. Deep mutational scanning of hemagglutinin helps predict evolutionary fates of human H3N2 influenza variants / J.M. Lee, J. Huddleston, M.B. Doud // *Proceedings of the National Academy of Sciences of the United States of America*. – 2018. – Vol. 115. – E8276–E8285. – doi:10.1073/pnas.1806133115.

45. Lek, M., Karczewski, K.J., Minikel, E. V. Analysis of protein-coding genetic variation in 60,706 humans / M. Lek, K.J. Karczewski, E. V. Minikel et al// Nature. – 2016. – Vol. 536. – P. 285–291. – doi:10.1038/nature19057.

46. Letunic, I., Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation / I. Letunic, P. Bork // *Bioinformatics*

(Oxford, England). – 2007. – Vol. 23. – P. 127–128. – doi:10.1093/bioinformatics/btl529.

47. Li, W., Godzik, A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences / W. Li, A. Godzik, // *Bioinformatics (Oxford, England)*. – 2006. – Vol. 22. – P. 1658–1659. – doi:10.1093/bioinformatics/btl158.

48. Lockless, S.W., Ranganathan, R. Evolutionarily conserved pathways of energetic connectivity in protein families / S.W. Lockless, R. Ranganathan // *Science (New York, N.Y.)*. – 1999. – Vol. 286. – P. 295–299. – doi:10.1126/science.286.5438.295.

49. Lopez, P., Casane, D., Philippe, H. Heterotachy, an important process of protein evolution / P. Lopez, D. Casane, H. Philippe // *Molecular Biology and Evolution*. – 2002. – Vol. 19. – P. 1–7. – doi:10.1093/oxfordjournals.molbev.a003973.

50. Lott, M.T. mtDNA Variation and Analysis Using Mitomap and Mitomaster / M.T. Lott, J.N. Leipzig, O.Derbeneva, et al. // *Current Protocols in Bioinformatics*. – 2013. – Vol. 44. – 1.23.1-26. – doi:10.1002/0471250953.bi0123s44.

51. Louie, R.H.Y. Fitness landscape of the human immunodeficiency virus envelope protein that is targeted by antibodies / R.H.Y. Louie, K.J. Kaczorowski, J.P. Barton et al. // *Proceedings of the National Academy of Sciences of the United States of America*. – 2018. – Vol. 115. – E564–E573. – doi:10.1073/pnas.1717765115.

52. Lyons, D.M., Lauring, A.S. Mutation and Epistasis in Influenza Virus Evolution / D.M. Lyons, A.S.Lauring // *Viruses*. – 2018. – Vol. 10. – doi:10.3390/v10080407.

53. Mann, J.K. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing / J.K. Mann,

J.P. Barton, A.L. Ferguson et al. // *PLoS computational biology*. – 2014. – Vol. 10. – e1003776. – doi:10.1371/journal.pcbi.1003776.

54. McLaughlin, R.N. The spatial architecture of protein function and adaptation / R.N. McLaughlin, F.J. Poelwijk, A. Raman et al. // *Nature*. – 2012. – Vol. 491. – P. 138–142. – doi:10.1038/nature11500.

55. Melamed, D. Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein / D. Melamed, D.L. Young, C.E. Gamble et al.// *RNA (New York, N.Y.).* – 2013. – Vol. 19. – P. 1537–1551. – doi:10.1261/rna.040709.113.

56. Mendes, F.K., Hahn, Y., Hahn, M.W. Gene Tree Discordance Can Generate Patterns of Diminishing Convergence over Time / F.K. Mendes, Y. Hahn, M.W.Hahn // *Molecular Biology and Evolution*. – 2016.– Vol. 33. – P. 3299–3307. – doi:10.1093/molbev/msw197.

57. Miyata, T., Miyazawa, S., Yasunaga, T. Two types of amino acid substitutions in protein evolution / T. Miyata, S. Miyazawa, T.Yasunaga // *Journal of Molecular Evolution*. – 1979. – Vol. 12. – P. 219–236.

58. Mkaouar-Rebai, E., Ellouze, E., Chamkha, I. Molecular-clinical correlation in a family with a novel heteroplasmic Leigh syndrome missense mutation in the mitochondrial cytochrome c oxidase III gene / E. Mkaouar-Rebai, E. Ellouze, I. Chamkha // *Journal of Child Neurology*. – 2011. – Vol. 26. – P. 12–20. – doi:10.1177/0883073810371227.

59. Munakata, K. Mitochondrial DNA 3644T-->C mutation associated with bipolar disorder / K. Munakata, M. Tanaka, K. Mori et al. // *Genomics*. – 2004. – Vol. 84. – P. 1041–1050. – doi:10.1016/j.ygeno.2004.08.015.

60. Murrell, B. Detecting individual sites subject to episodic diversifying selection / B. Murrell, J.O. Wertheim, S. Moola et al. // *PLoS genetics*. – 2012. – Vol. 8. – e1002764. – doi:10.1371/journal.pgen.1002764.

61. Mustonen, V., Lässig, M. Adaptations to fluctuating selection in Drosophila / V. Mustonen, M. Lässig // *Proceedings of the National Academy of*

Sciences of the United States of America. – 2007. – Vol. 104. – P. 2277–2282. – doi:10.1073/pnas.0607105104.

62. Nabieva, E., Bazykin, G.A. SELVa: Simulator of Evolution with Landscape Variation / E. Nabieva, G.A. Bazykin // *bioRxiv*. – 2019. – doi:10.1101/647834.

63. Naumenko, S.A., Kondrashov, A.S., Bazykin, G.A. Fitness conferred by replaced amino acids declines with time / S.A. Naumenko, A.S. Kondrashov, G.A. Bazykin // *Biology Letters*. – 2012. – Vol. 8. – P. 825–828. – doi:10.1098/rsbl.2012.0356.

64. Neverov, A.D. Coordinated Evolution of Influenza A Surface Proteins / A.D. Neverov, S. Kryazhimskiy, J.B. Plotkin, G.A. Bazykin // *PLoS genetics*. – 2015. – Vol. 11. – e1005404. – doi:10.1371/journal.pgen.1005404.

65. Olson, C.A., Wu, N.C., Sun, R. A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain / C.A. Olson, N.C. Wu, R. Sun // *Current biology: CB*. – 2014. – Vol. 24. – P. 2643–2651. – doi:10.1016/j.cub.2014.09.072.

66. Podgornaia, A.I., Laub, M.T. Protein evolution. Pervasive degeneracy and epistasis in a protein-protein interface / A.I. Podgornaia, M.T. Laub // *Science (New York, N.Y.)*. – 2015. – Vol. 347. – P. 673–677. – doi:10.1126/science.1257360.

67. Pokusaeva, V.O. An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape / V.O. Pokusaeva, D.R. Usmanova, E. V. Putintseva // *PLoS Genetics*. – 2019. – doi:10.1371/journal.pgen.1008079.

68. Povolotskaya, I.S., Kondrashov, F.A. Sequence space and the ongoing expansion of the protein universe / I.S. Povolotskaya, F.A. Kondrashov // *Nature*. – 2010. – Vol. 465. – P. 922–926. – doi:10.1038/nature09105.

69. Rhee, S.-Y. Geographic and temporal trends in the molecular epidemiology and genetic mechanisms of transmitted HIV-1 drug resistance: an

individual-patient- and sequence-level meta-analysis / S.-Y. Rhee, J.L. Blanco, M.R. Jordan, et al. // *PLoS medicine*. – 2015. – Vol. 12. – e1001810. – doi:10.1371/journal.pmed.1001810.

70. Riesselman, A.J., Ingraham, J.B., Marks, D.S. Deep generative models of genetic variation capture the effects of mutations / A.J. Riesselman, J.B. Ingraham, D.S. Marks // *Nature Methods.* – 2018. – Vol. 15. – P. 816–822. – doi:10.1038/s41592-018-0138-4.

71. Rogozin, I.B. Homoplasy in genome-wide analysis of rare amino acid replacements: the molecular-evolutionary basis for Vavilov's law of homologous series / I.B. Rogozin, K. Thomson, M. Csürös // *Biology Direct*. – 2008. – Vol. 3. – P. 7. – doi:10.1186/1745-6150-3-7.

72. Rollins, N.J. Inferring protein 3D structure from deep mutation scans / N.J. Rollins, K.P. Brock, F.J. Poelwijk et al. // *Nature Genetics*. – 2019. – Vol. 51. – P. 1170–1176. – doi:10.1038/s41588-019-0432-9.

73. Roscoe, B.P. Analyses of the effects of all ubiquitin point mutants on yeast growth rate / B.P. Roscoe, K.M. Thayer, K.B. Zeldovich et al. // *Journal of Molecular Biology*. – 2013. – Vol. 425. – P. 1363–1377. – doi:10.1016/j.jmb.2013.01.032.

74. Roure, B., Philippe, H. Site-specific time heterogeneity of the substitution process and its impact on phylogenetic inference / B. Roure, H. Philippe // *BMC evolutionary biology*. – 2011.– Vol. 11. – P. 17. – doi:10.1186/1471-2148-11-17.

75. Salinas, V.H., Ranganathan, R. Coevolution-based inference of amino acid interactions underlying protein function / V.H. Salinas, R. Ranganathan // *eLife*. – 2018. – Vol. 7. – doi:10.7554/eLife.34300.

76. Sarkisyan, K.S. Local fitness landscape of the green fluorescent protein / K.S. Sarkisyan, D.A. Bolotin, M. V. Meer // *Nature*. – 2016. – Vol. 533. – P. 397–401. – doi:10.1038/nature17995.

77. Shu, Y., McCauley, J. GISAID: Global initiative on sharing all influenza data - from vision to reality / Y. Shu, J. McCauley // *Euro Surveillance: Bulletin Europeen Sur Les Maladies Transmissibles = European Communicable Disease Bulletin*. – 2017. – Vol. 22. – doi:10.2807/1560-7917.ES.2017.22.13.30494.

78. Sim, N.-L. SIFT web server: predicting effects of amino acid substitutions on proteins / N.-L. Sim, P. Kumar, J. Hu et al. // *Nucleic Acids Research*. – 2012. – Vol. 40, W452-457. – doi:10.1093/nar/gks539.

79. Soylemez, O., Kondrashov, F.A. Estimating the rate of irreversibility in protein evolution / O. Soylemez, F.A. Kondrashov // *Genome Biology and Evolution*. – 2012. – Vol. 4. – P. 1213–1222. – doi:10.1093/gbe/evs096.

80. Stajich, J.E. The Bioperl toolkit: Perl modules for the life sciences / J.E. Stajich, D. Block, K. Boulez // *Genome Research*. – 2002. – Vol. 12. – P. 1611–1618. – doi:10.1101/gr.361602.

81. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies / A. Stamatakis // *Bioinformatics (Oxford, England)*. – 2014. – Vol. 30. – P. 1312–1313. – doi:10.1093/bioinformatics/btu033.

82. Stray, S.J., Pittman, L.B. Subtype- and antigenic site-specific differences in biophysical influences on evolution of influenza virus hemagglutinin / S.J. Stray, L.B. Pittman // *Virology Journal*. – 2012. – Vol. 9. – P. 91. – doi:10.1186/1743-422X-9-91.

83. Suyama, M., Torrents, D., Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments / M. Suyama, D. Torrents, P. Bork, // *Nucleic Acids Research*. – 2006. – Vol. 34. – W609-612. – doi:10.1093/nar/gkl315.

84. Tamuri, A.U. Identifying changes in selective constraints: host shifts in influenza / A.U. Tamuri, M. D. Reis, A.J. Hay, R.A. Goldstein // *PLoS*

computational biology. – 2009. – Vol. 5. – e1000564. – doi:10.1371/journal.pcbi.1000564.

85. Tawata, M. A new mitochondrial DNA mutation at 14577 T/C is probably a major pathogenic mutation for maternally inherited type 2 diabetes / M. Tawata, J.I. Hayashi, K. Isobe et al. // *Diabetes*. – 2000. – Vol. 49. – P. 1269–1272. – doi:10.2337/diabetes.49.7.1269.

86. UniProt Consortium UniProt: a worldwide hub of protein knowledge // *Nucleic Acids Research*. 2019. – Vol. D1 (47). – P. D506–D515. – doi:10.1093/nar/gky1049.

87. Usmanova, D.R. A model of substitution trajectories in sequence space and long-term protein evolution / D.R. Usmanova, L. Ferretti, I.S. Povolotskaya et al. // *Molecular Biology and Evolution*. – 2015. – Vol. 32. – P. 542–554. – doi:10.1093/molbev/msu318.

88. Wallace, D.C. Diseases of the Mitochondrial DNA / D.C. Wallace // *Annual Review of Biochemistry*. – 1992. – Vol. 61. – P. 1175–1212. – doi:10.1146/annurev.bi.61.070192.005523.

89. Weigt, M. Identification of direct residue contacts in protein-protein interaction by message passing / M. Weigt, R.A. White, H. Szurmant, et al. // *Proceedings of the National Academy of Sciences of the United States of America*. – 2009. – Vol. 106. – P. 67–72. – doi:10.1073/pnas.0805923106.

90. Wiegand, T., A. Moloney, K. Rings, circles, and null-models for point pattern analysis in ecology / T. A. Wiegand, K. Moloney // *Oikos*. – 2004. – Vol. 104. – P. 209–229. – doi:10.1111/j.0030-1299.2004.12497.x.

91. Wong, A. Epistasis and the Evolution of Antimicrobial Resistance / A. Wong // *Frontiers in Microbiology*. – 2017. – Vol. 8. – doi:10.3389/fmicb.2017.00246.

92. Wood, N. HIV evolution in early infection: selection pressures, patterns of insertion and deletion, and the impact of APOBEC / N. Wood, T.

Bhattacharya, B.F. Keele et al. // *PLoS pathogens*. – 2009. – Vol. 5. – e1000414. – doi:10.1371/journal.ppat.1000414.

93. Wu, N.C. Functional Constraint Profiling of a Viral Protein Reveals Discordance of Evolutionary Conservation and Functionality / N.C. Wu, C.A. Olson, Y. Du et al. // *PLoS genetics*. – 2015. – Vol. 11. – e1005310. – doi:10.1371/journal.pgen.1005310.

94. Yang, W., Bielawski, J.P., Yang, Z. Widespread adaptive evolution in the human immunodeficiency virus type 1 genome / W. Yang, J.P. Bielawski, Z. Yang // *Journal of Molecular Evolution*. – 2003. – Vol. 57, 212–221. – doi:10.1007/s00239-003-2467-9.

95. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood / Z. Yang // *Molecular Biology and Evolution*. – 2007. – Vol. 24. – P. 1586–1591. – doi:10.1093/molbev/msm088.

96. Yang, Z., Nielsen, R. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages / Z. Yang, R. Nielsen // *Molecular Biology and Evolution*. – 2002. – Vol. 19. – P. 908–917. – doi:10.1093/oxfordjournals.molbev.a004148.

97. Yang, Z., Nielsen, R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage / Z. Yang, R. Nielsen // *Molecular Biology and Evolution*. – 2008. – Vol. 25. – P. 568–579. – doi:10.1093/molbev/msm284.

98. Zou, Z., Zhang, J. Are Convergent and Parallel Amino Acid Substitutions in Protein Evolution More Prevalent Than Neutral Expectations? / Z. Zou, J. Zhang // *Molecular Biology and Evolution*. – 2015. – Vol. 32. – P. 2085–2096. – doi:10.1093/molbev/msv091.

99. Los Alamos HIV sequence database [Электронный ресурс]. – Режим доступа: http://www.hiv.lanl.gov/.

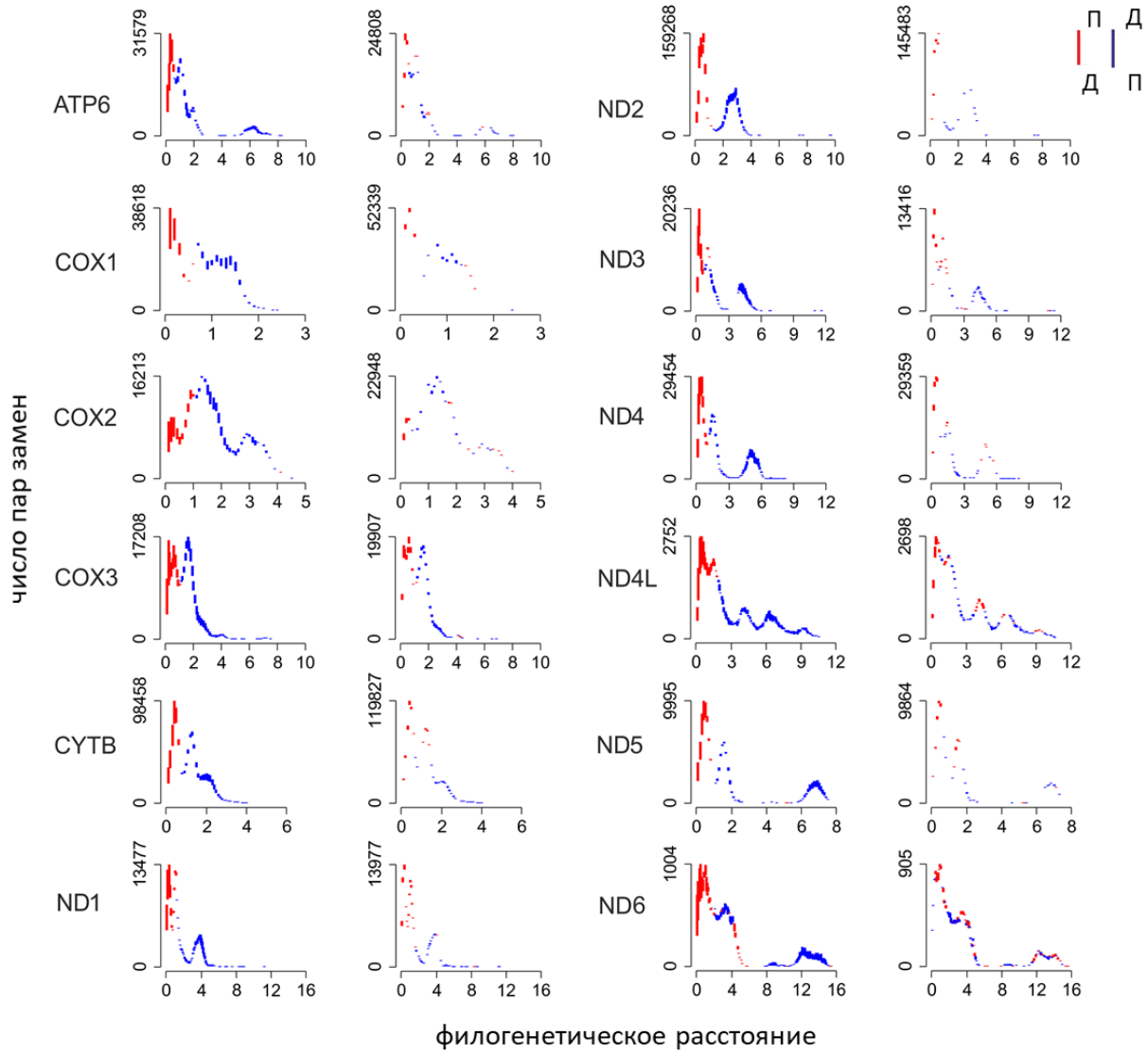
Приложения

Приложение 1. Данные для белков *Metazoa*.

ген	число сайтов	число информативных сайтов*
ATP6	186	172
COX1	404	317
COX2	165	154
COX3	198	173
CYTB	327	321
ND1	253	227
ND2	299	288
ND3	94	87
ND4	392	348
ND4L	82	79
ND5	516	404
ND6	119	111

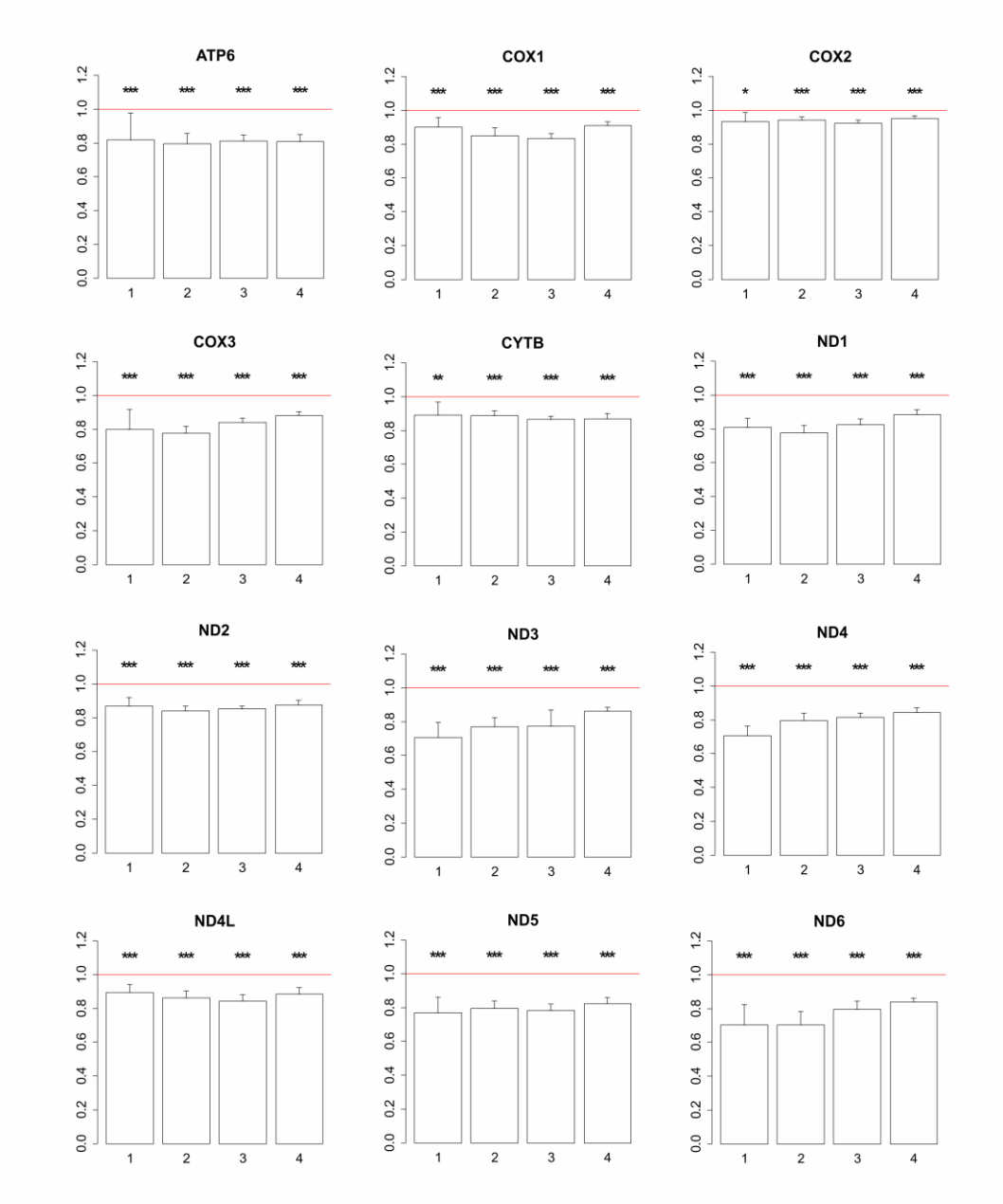
* Информативные сайты – сайты, где в пределах имеющейся филогении для одной аминокислоты находится хотя бы одна пара параллельных замен на неё и хотя бы одна пара дивергентных замен на неё и другую аминокислоту.

Приложение 2. Число пар параллельных (П) и дивергентных (Д) замен



Для каждого окна расстояний величиной 0.1 аминокислотных замен на сайт, два конца вертикальной линии показывают средние среди 1000 повторностей бутстрэпа значения П и Д так, что что высота линии соответствует абсолютному значению их разности. Красные линии нарисованы для окон, где П>Д (то есть, когда верхняя точка линии соответствует П, а нижняя – Д), а синие линии – для окон с П<Д (то есть, когда верхняя точка линии соответствует Д, а нижняя – П).

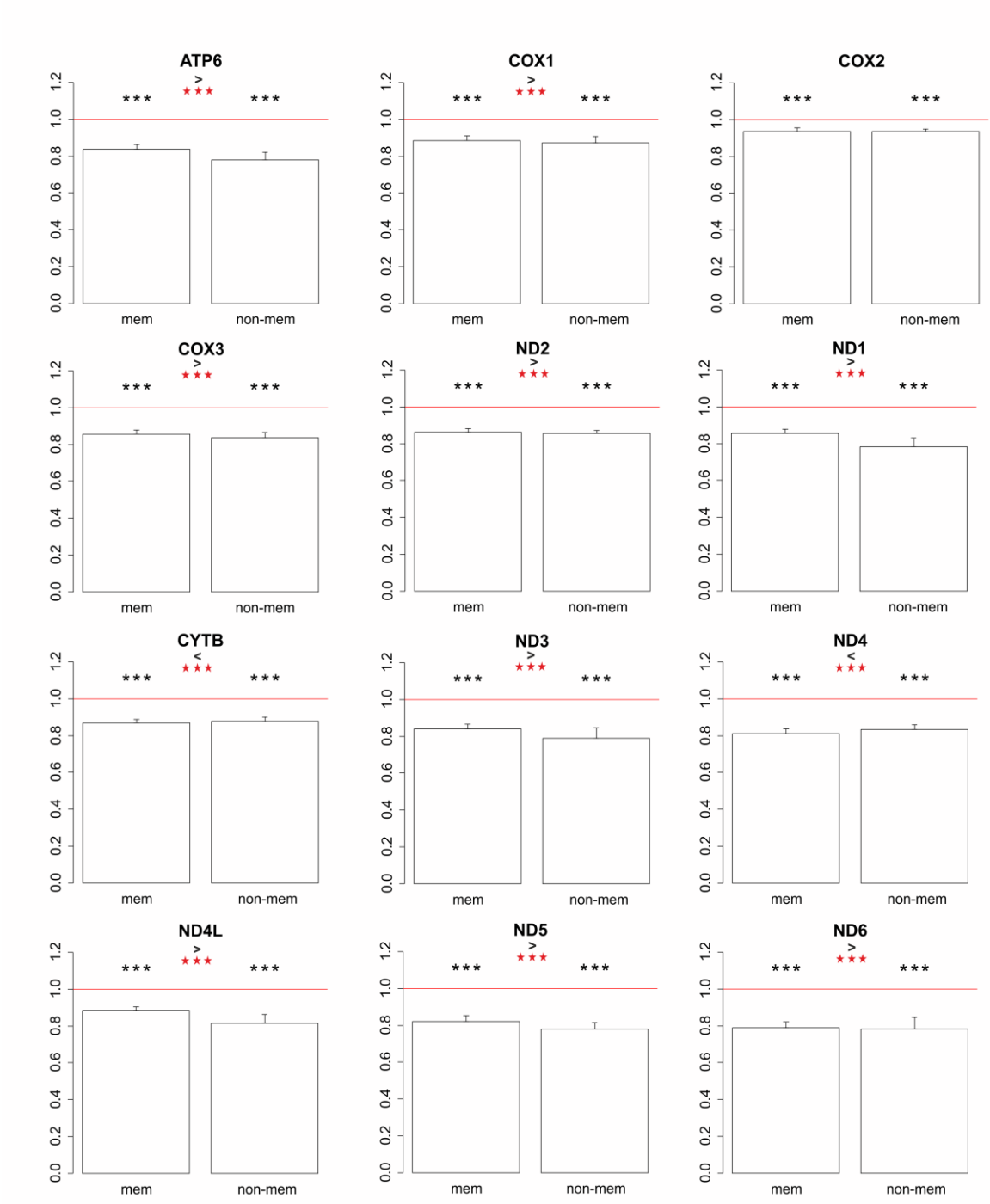
Приложение 3. Отношения филогенетических расстояний между параллельными и дивергентными заменами из разных категорий скорости эволюции сайтов по codeml на филогении *Metazoa*



1, 2, 3, 4 – группы с разными скоростями эволюции, на которые программа «codeml» разбила сайты. Скорость эволюции растёт от группы 1 к группе 4.

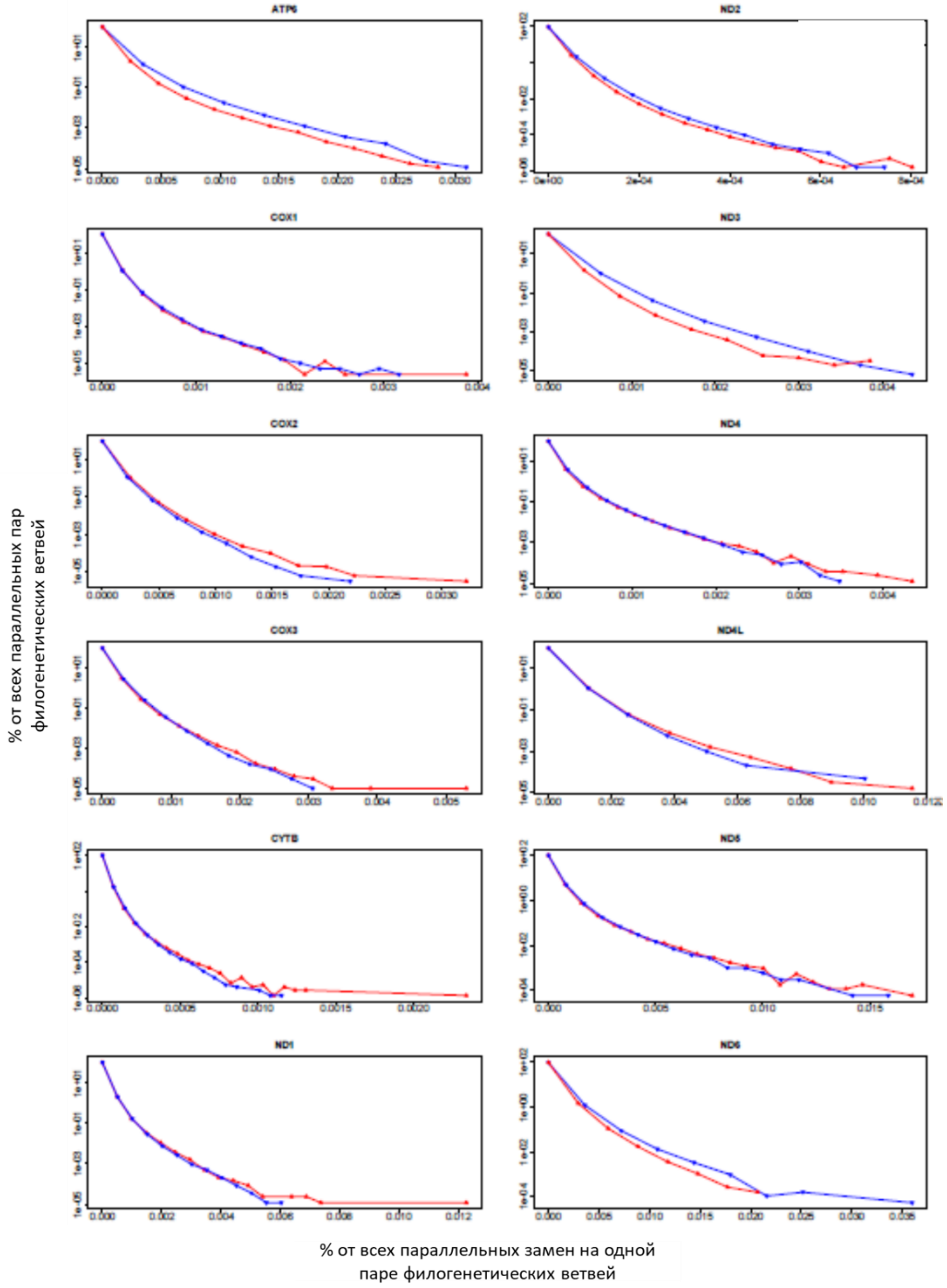
*Отношения ниже 1 означают, что параллельные замены расположены ближе друг к другу на филогении, чем дивергентные замены. Высота столбца и усы обозначают соответственно медиану и 95% доверительный интервал для 1000 бустрэпов (выборок с возвращением) сайтов. Звёздочки показывают значимость отличия отношения от 1/1 (красная линия; *** - p <0.001; ** - p < 0.01). «**все**» - настоящие данные для всех замен; «**постоянные**» - настоящие данные только для замен между «постоянными» парами аминокислот (см. текст); «**симуляция**» - симулированные данные.*

Приложение 4. Отношение филогенетических расстояний между параллельными и дивергентными заменами на филогении *Metazoa*, для трансмембранных (mem) и немембранных (non-mem) сайтов.



*Красные звёздочки показывают P-значение двустороннего теста Вилкоксона с нулевой гипотезой об отсутствии различий между двумя типами сайтов (***, P<0.001), со знаками «<» и « >», показывающими направление различия. Остальные обозначения – как в приложении 3.*

Приложение 5. Распределение пар филогенетических ветвей по доле произошедших на них параллельных замен



Красный цвет – данные; синий – симуляция.

Приложение 6. Замены на референтный аллель человека в митохондриальных белках *Metazoa* и *Opisthokonta*

ген	виды	число сайтов	число сайтов в анализе	число замен на аминокислоту человека на сайт	число замен на аминокислоту человека на сайт в симуляции
ATP6	Metazoa	186	131	25.3	16.7
COX1	Metazoa	404	146	12.1	13.9
COX2	Metazoa	165	110	22.5	22.3
COX3	Metazoa	198	128	22.4	16.2
CYTB	Metazoa	327	183	32.2	29.7
ND1	Metazoa	253	160	15.5	12.1
ND2	Metazoa	299	220	36.3	34.0
ND3	Metazoa	94	67	26.6	20.0
ND4	Metazoa	392	263	19.9	17.3
ND4L	Metazoa	82	70	18.4	19.6
ND5	Metazoa	516	302	8.7	7.3
ND6	Metazoa	119	91	11.4	10.1
ATP6+COX1+COX2+COX3+CYTB	Opisthokonta	1524	964	30.2	24.4

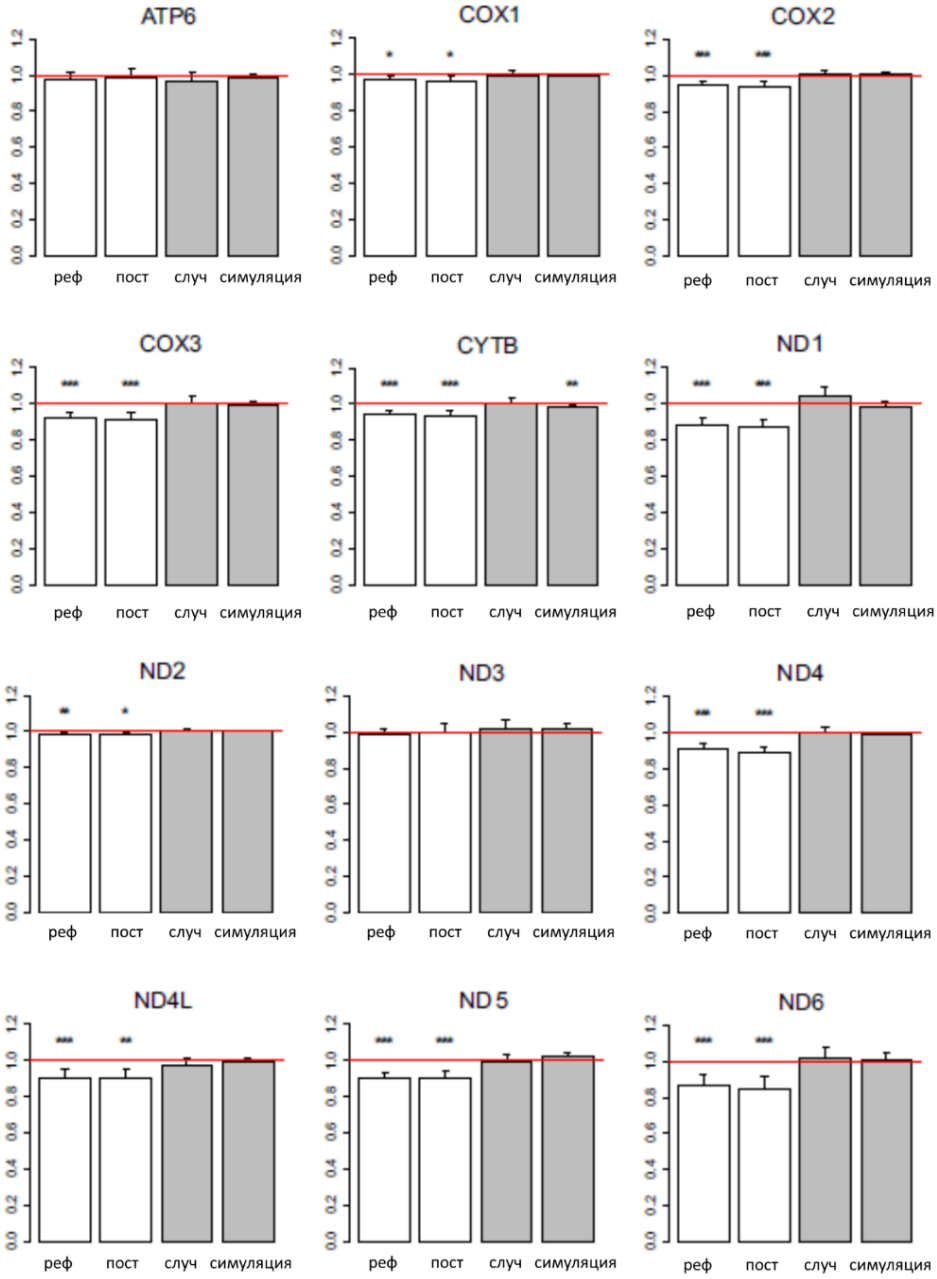
Приложение 7. Замены на нереферентный нейтральный аллель человека в митохондриальных белках *Metazoa* и *Opisthokonta*

ген	виды	число сайтов	число сайтов в анализе	число замен на аминокислоту человека на сайт	среднее число нереферентных нейтральных аллелей в полиморфном сайте
ATP6	Metazoa	142	128	21.1	1.9
COX1	Metazoa	131	104	21.0	1.4
COX2	Metazoa	81	71	28.3	1.4
COX3	Metazoa	108	92	24.8	1.6
CYTB	Metazoa	199	187	36.0	1.7
ND1	Metazoa	101	87	16.8	1.5
ND2	Metazoa	143	129	49.1	1.6
ND3	Metazoa	34	32	30.3	1.6
ND4	Metazoa	136	109	20.9	1.2
ND4L	Metazoa	29	25	17.3	1.5
ND5	Metazoa	223	171	9.9	1.6
ND6	Metazoa	55	45	11.1	1.5
ATP6+COX1+COX2+COX3+CYTB	Opisthokonta	775	516	32.2	1.6

Приложение 8. Замены на патогенный аллель человека в митохондриальных белках *Metazoa* и *Opisthokonta*

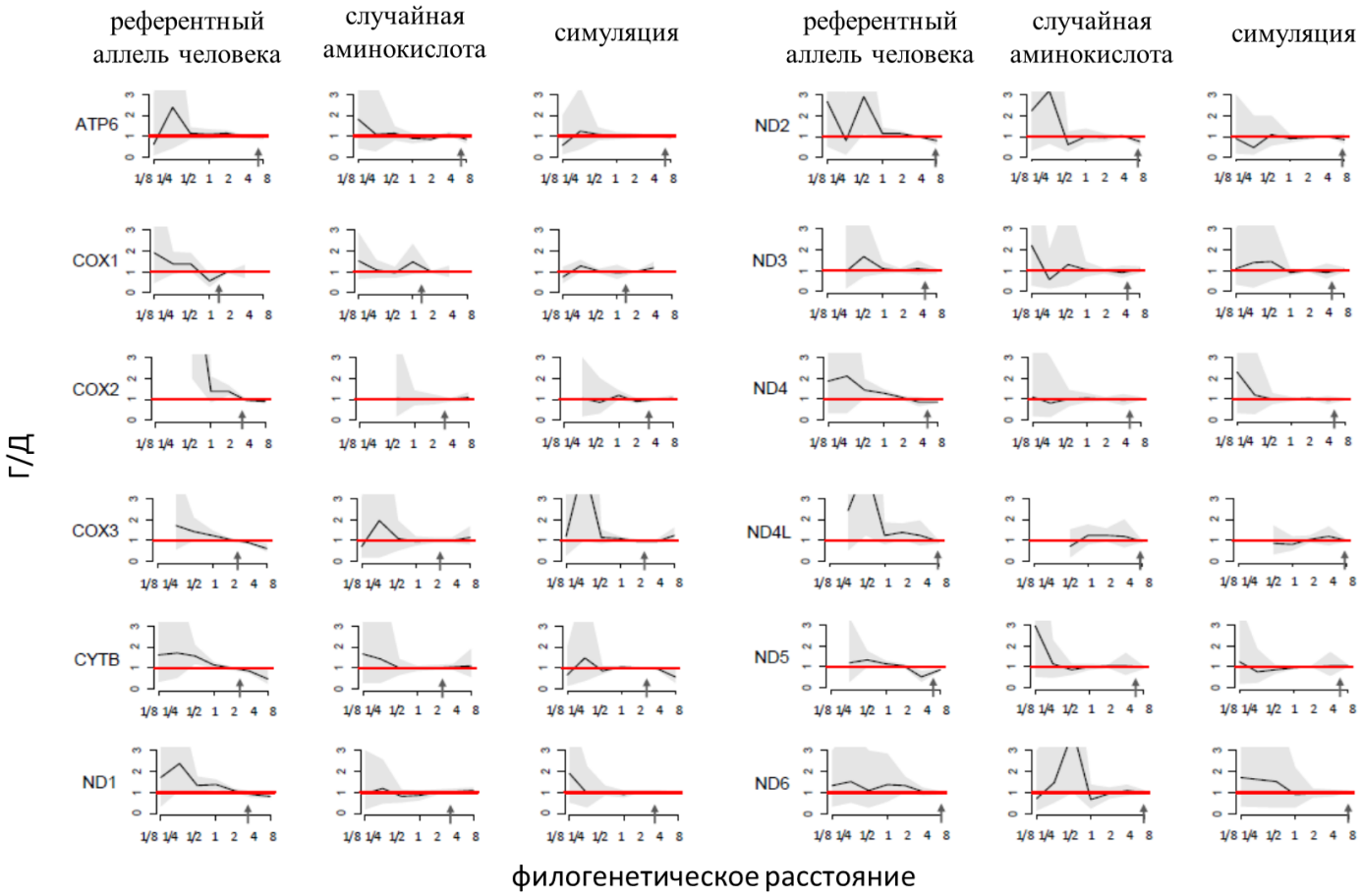
ген	**виды**	**число сайтов**	**число сайтов в анализе**	**число замен на аминокислоту человека на сайт**	**среднее число патогенных аминокислот в сайте с патогенными аллелями**
ATP6	Metazoa	15	11	9.7	1.1
COX1	Metazoa	20	15	41.0	1.0
COX2	Metazoa	10	9	44.8	1.0
COX3	Metazoa	9	7	14.8	1.0
CYTB	Metazoa	20	16	15.5	1.1
ND1	Metazoa	25	17	6.3	1.1
ND2	Metazoa	10	8	29.5	1.0
ND3	Metazoa	6	5	3.5	1.0
ND4	Metazoa	7	4	15.1	1.2
ND4L	Metazoa	3	3	6.0	1.0
ND5	Metazoa	24	8	5.1	1.0
ND6	Metazoa	13	10	8.0	1.1
ATP6+COX1+COX2+ COX3+CYTB	Opisthokonta	89	72	25.0	1.0

Приложение 9. Отношения филогенетических расстояний между веткой человека и заменами на его референтный аллель к филогенетическим расстояниям между веткой человека и заменами на аминокислоты, не встречающиеся в людях; данные по *Metazoa*



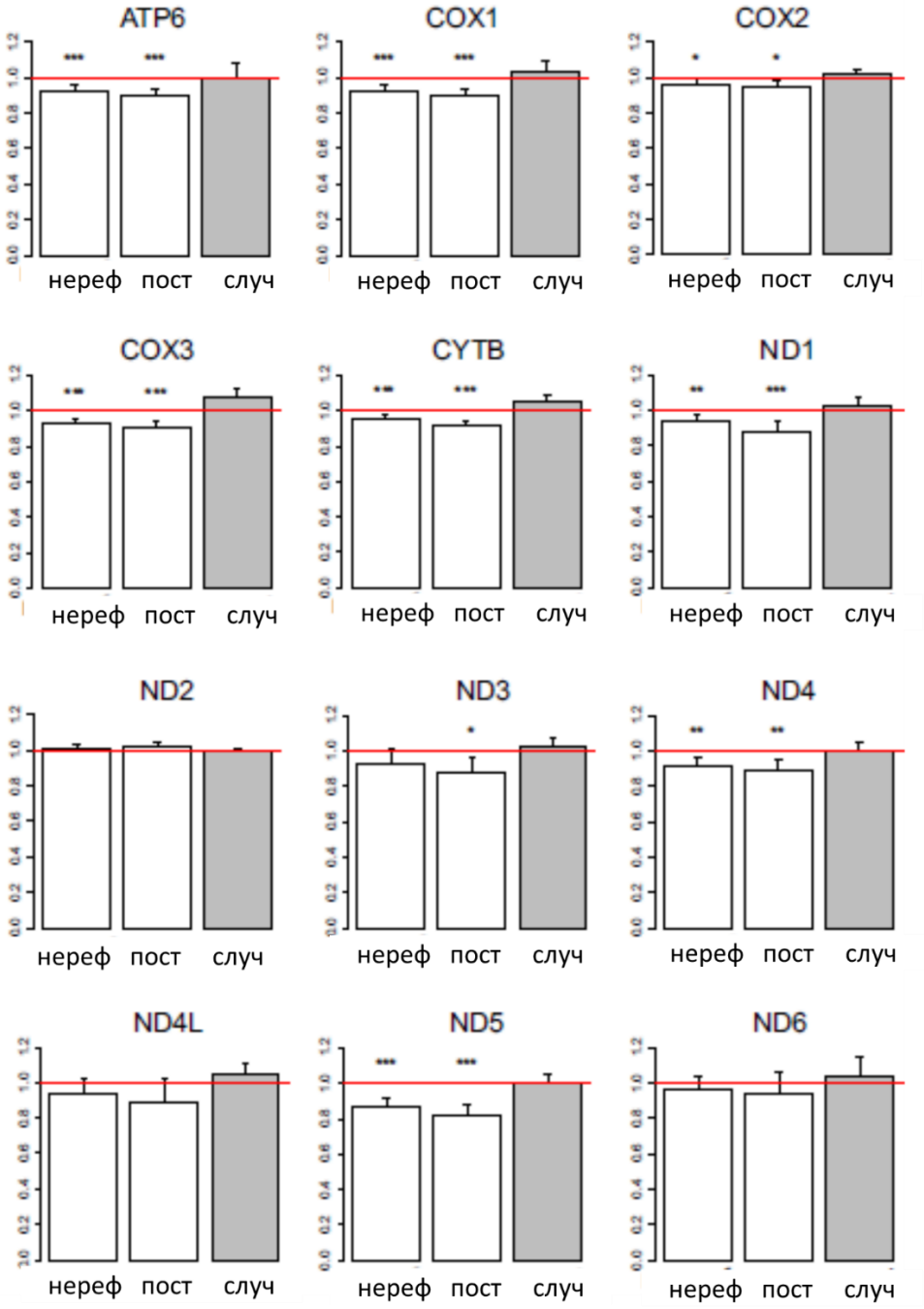
Отношения <1 означают, что в среднем независимые замены на человеческий аллель происходят филогенетически ближе к человеку, чем замены на другие аминокислоты. Высота столбца и усы представляют собой медиану и 95% доверительный интервал, полученные по 1000 повторностям бутстрэпа. Звёздочки показывают значимость отличий от отношения 1:1 (- P < 0.05; ** - P < 0.01; *** - P < 0.001).* ***реф.*** *– референтный аллель человека;* ***пост.*** *– референтный аллель человека из «постоянных» пар аминокислот (см. текст);* ***случ.*** *–случайная аминокислота из числа аминокислот, в наших данных встречающихся в сайте у других видов, но не у человека;* ***симуляция*** *– аллель человека в симуляции эволюции.*

Приложение 10. Доля гомоплазий на референтный аллель человека по сравнению с независимыми дивергентными заменами на случайные аминокислоты (Г/Д), на разных эволюционных расстояниях от ветви *H. sapiens*



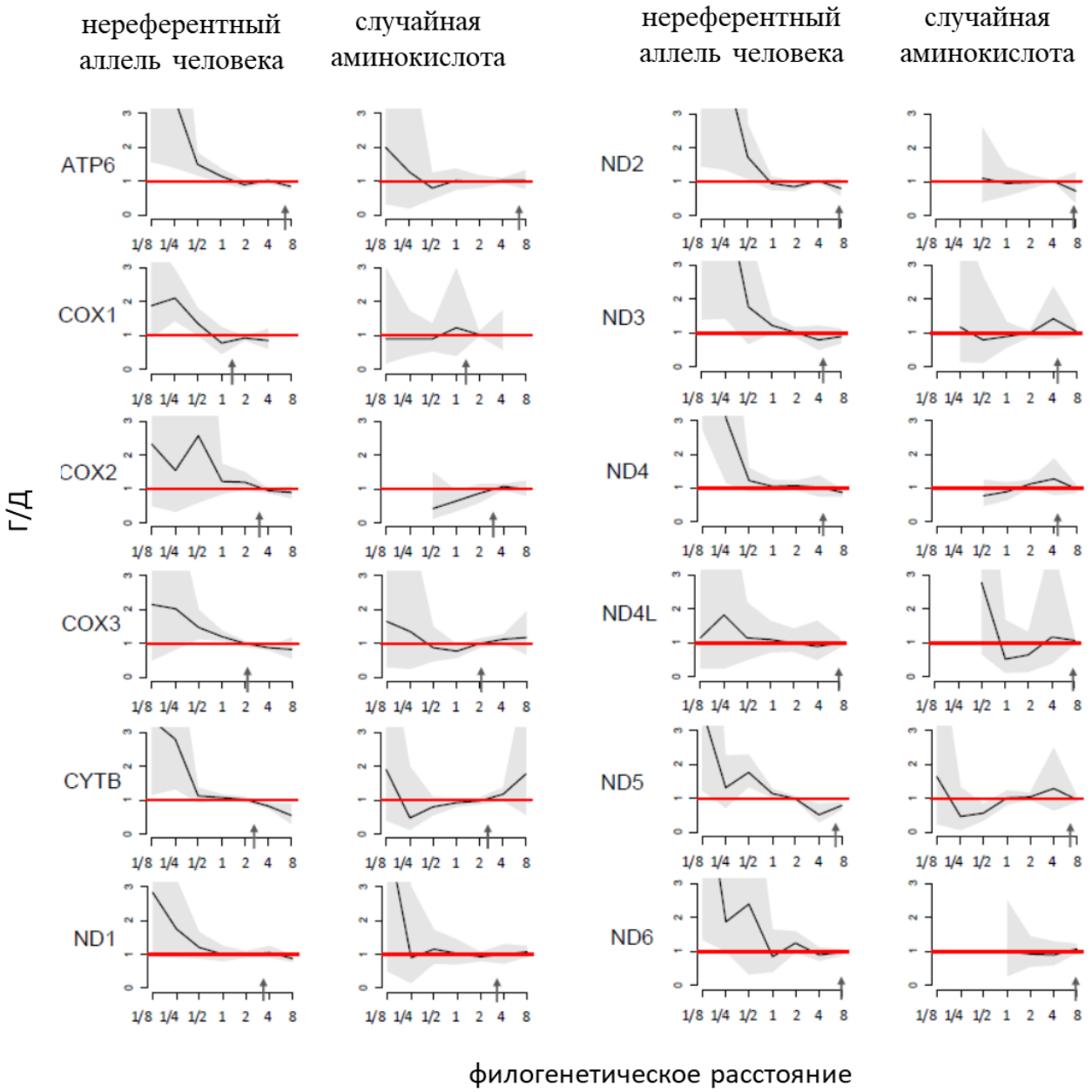
Горизонтальная ось – расстояния между ветвями, несущими замены, и веткой человека, измеренные в числе аминокислотных замен на сайт и разделённые на окна по log2(расстояние). Вертикальная ось – отношения Г/Д для замен на таком расстоянии. Чёрная линия и серая область – среднее и 95% доверительный интервал, полученные по 1000 повторностям бутстрэпа. Красная линия – ожидаемое отношение Г/Д=1. Стрелки –филогенетическое расстояние между человеком и насекомыми.

Приложение 11. Отношения филогенетических расстояний между веткой человека и заменами на его нереферентный аллель к филогенетическим расстояниям между веткой человека и заменами на аминокислоты, не встречающиеся в популяции человека; данные по *Metazoa*



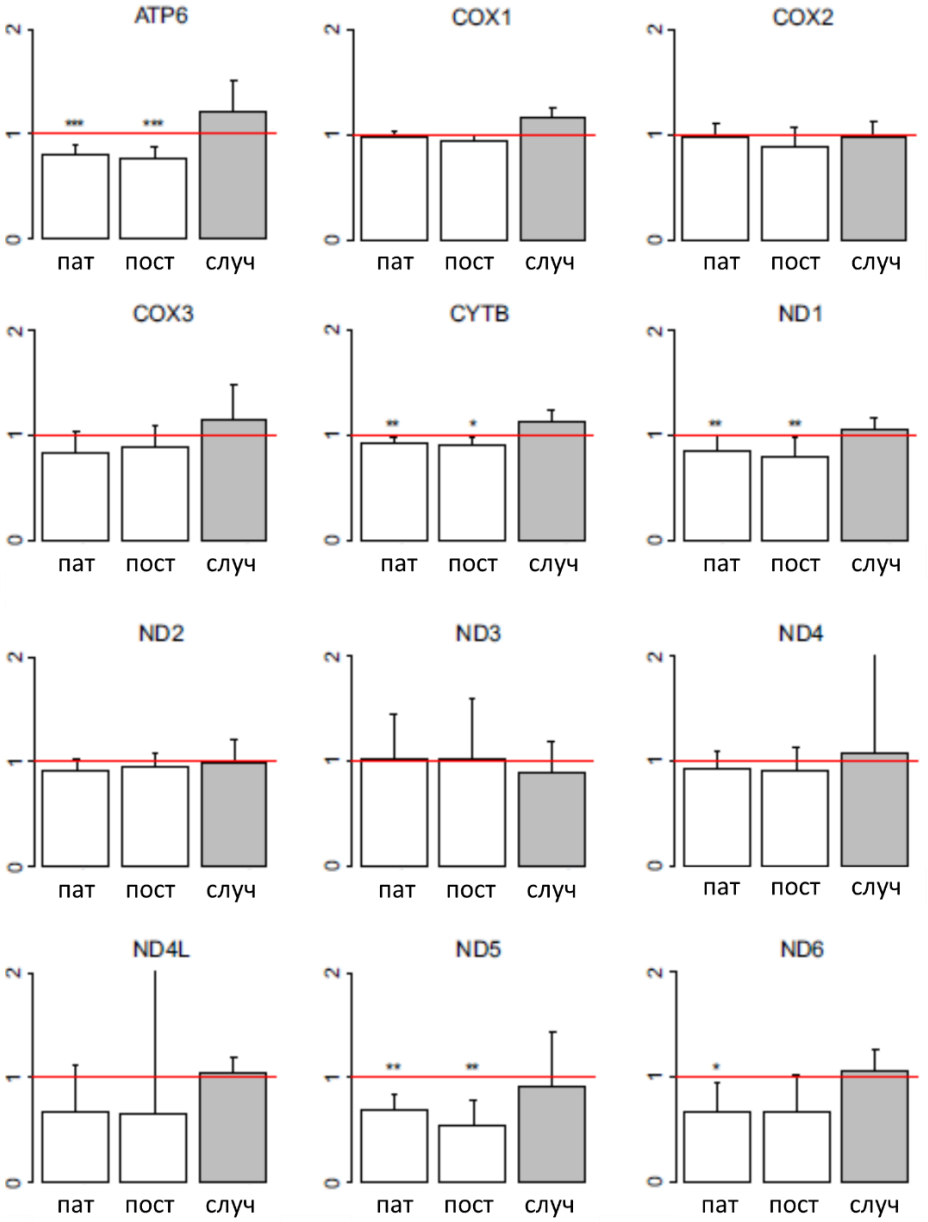
Нереф. *– нереферентный аллель человека;* ***пост.*** *– нереферентный аллель человека из «постоянных» пар аминокислот (см. текст);* ***случ.*** *– случайная аминокислота, встречающаяся в этом же сайте у других видов, но не у человека. Остальные обозначения – как в Приложении 9.*

Приложение 12. Доля гомоплазий на нереферентный аллель человека по сравнению с независимыми заменами на случайные аминокислоты (Г/Д), на разных эволюционных расстояниях от ветви *H. sapiens*



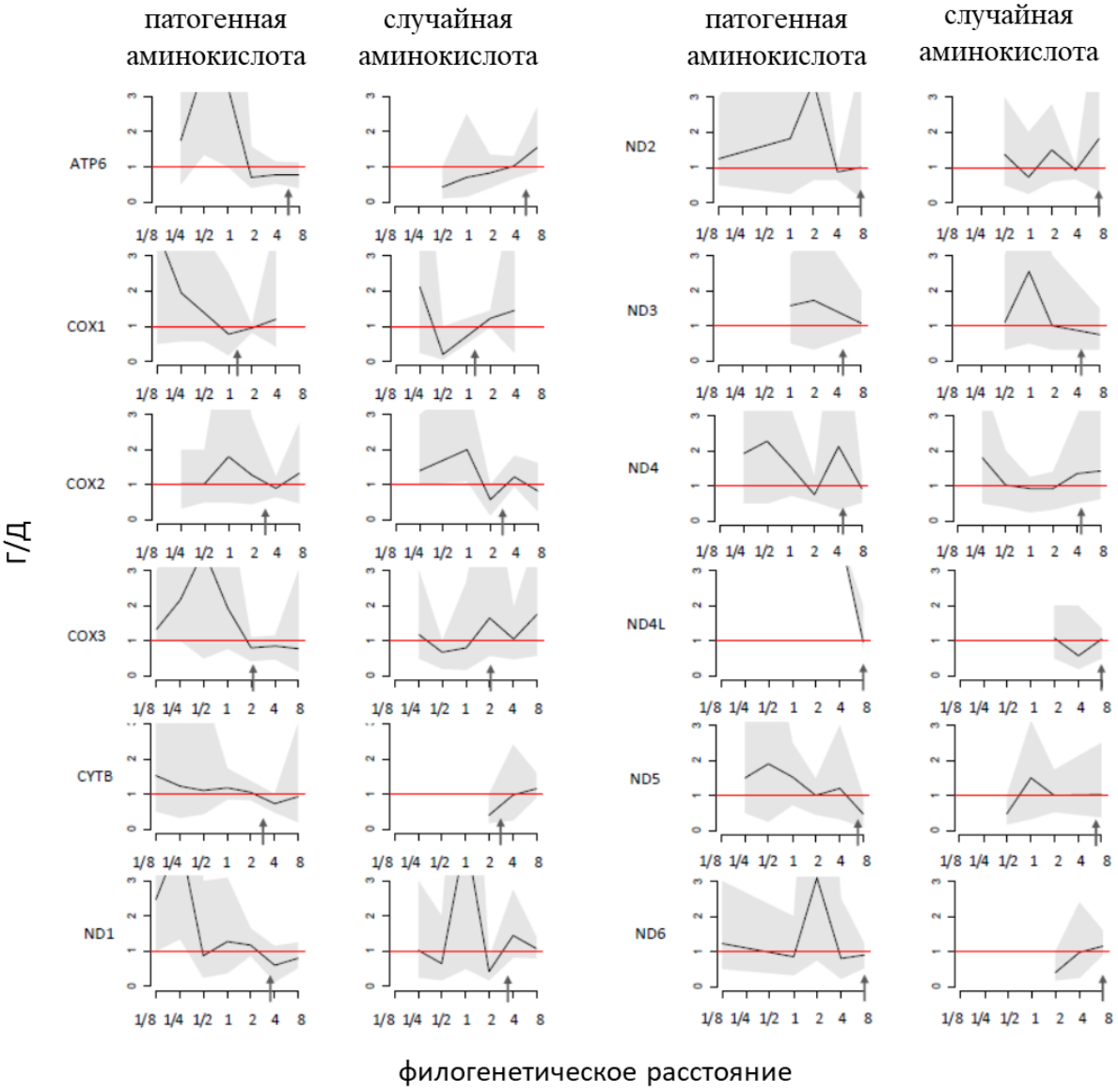
Обозначения – как в приложении 9 и 10.

Приложение 13. Отношения филогенетических расстояний между веткой человека и заменами на патогенную для него аминокислоту к филогенетическим расстояниям между веткой человека и заменами на аминокислоты, не встречающиеся в его популяции; данные по *Metazoa*



Пат. *– патогенная для человека аминокислота;* ***пост.*** *– патогенная аминокислота из «постоянных» пар аминокислот (см. текст);* ***случ.*** *–случайная аминокислота из числа аминокислот, в наших данных встречающихся в сайте у других видов, но не у человека. Остальные обозначения – как в Приложении 9.*

Приложение 14. Доля гомоплазий на патогенный вариант человека по сравнению с независимыми заменами на случайные аминокислоты (Г/Д), на разных эволюционных расстояниях от ветви *H. sapiens*; данные по *Metazoa*



Обозначения – как в Приложении 10.