МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

БИОЛОГИЧЕСКИЙ ФАКУЛЬТЕТ
Кафедра биологической эволюции

Дмитрий Андреевич Биба

Скомпенсированные сдвиги рамки считывания и их роль в эволюции белковых последовательностей

Выпускная квалификационная работа бакалавра

Научные руководители:

Галина Викторовна Клинк
Георгий Александрович Базыкин
Сергей Николаевич Лысенков

Москва – 2020 г.

Оглавление

Оглавление

Введение

Механизмы появления новых аминокислотных последовательностей в эволюции - одна из нерешённых проблем эволюционной биологии. Было предложено немалое количество способов появления новых белков и их функций (см. обзор литературы), многие из которых подтверждаются наблюдениями. Тем не менее, новые механизмы, такие как появление целых генов из некодирующих участков, всё еще открываются.

О роли мутаций сдвига рамки считывания (инсерций и делеций длины, не кратной трём) в эволюции белковых последовательностей кое-что известно. Очевидно, потенциально такие мутации могут генерировать совершенно новые аминокислотные последовательности большой длины. Однако считается, что такие мутации обычно очень вредны, поскольку вызывают образование преждевременных стоп-кодонов, тем самым значительно нарушая структуру белка, и, следовательно, редко фиксируются и играют незначительную роль в эволюции (Ohno, 1970). Однако примеры, когда мутации сдвига рамки фиксируются, известны, и хотя иногда это ведет к потере функции, в некоторых случаях белок продолжает работать (Hahn, Lee, 2005). Более того, есть примеры, когда сдвиг рамки порождает новую функцию белка (Ohno, 1984; Vandenbussche et al., 2003).

Между тем, несложно представить сценарии, в котором негативный эффект от мутации сдвига рамки будет минимизирован. Во-первых, если мутация произойдет вблизи 3′-конца, преждевременный стоп-кодон, который она может вызвать, будет не таким уж преждевременным. Поэтому, собственно, инделы (**ин**серции или **дел**еции), сдвигающие рамку, чаще находят близко к концу белка (Hu, Ng, 2012). Во-вторых, и это сценарий, о котором в дальнейшем и пойдет речь, вблизи уже случившейся мутации сдвига рамки может произойти еще одна, компенсирующая её, то есть такая, что их суммарная длина будет кратна трём. В этом случае преждевременный стоп-кодон может и не образоваться (вероятность его образования уменьшается с уменьшением расстояния между двумя мутациями), а аминокислотная последовательность, тем не менее, может значительно измениться. Свидетельства того, что это иногда происходит, присутствуют (Hu, Ng, 2012). Также есть основания полагать, что при подобных событиях структура белка нарушается несильно, поскольку генетический код устроен таким образом, что при сдвиге рамки получаются похожие по свойствам аминокислоты (Wang et al., 2016; Bartonek et al., 2020). Конечно, остается открытым вопрос о том, что происходит с организмом, когда он уже несет одну из мутаций, но еще не получил вторую. Однако есть данные о том, что подобное состояние может стабилизироваться тем, что рибосома может пропускать преждевременные стоп-кодоны, образовавшиеся в результате сдвига рамки (Rockah-Shmuel et al., 2013). Также ген, в котором произошла только одна мутация сдвига рамки, может быть не критически важен для выживания организма, благодаря чему может “дождаться” компенсирующей мутации.

Обзор литературы.
Механизмы образования новых аминокислотных последовательностей в процессе эволюции.

Дупликации

Дупликации считаются главным источником возникновения новых генов (Kaessmann, 2010). О возможности появления новых генов путём дупликации старых и последующего накопления мутация писал ещё Мёллер (Muller, 1935). С тех пор было обнаружено, что дупликации генов – обычное явление и у эукариот (Lynch, 2007), и у прокариот (Romero and Palacios, 1997), а также придумано множество вариантов развития событий после дупликации, ведущих к закреплению и сохранению в популяции двух копий гена.

Рисунок 1. Модели эволюции гена после дупликации (по Иннану и Кондрашову, 2010). Красным цветом отмечен ген в предковом состоянии. Яркие цвета – мутировавший ген с другой функцией. Белый цвет – “поломанный ген”, без функции. Знак “+” отмечает действие положительного отбора на особь с дупликацией.

Систематический подход к классификации моделей эволюционных событий после дупликации был предложен Иннаном и Кондрашовым (Innan and Kondrashov, 2010). Их классификация основывается на динамике закрепления дупликаций (под действием дрейфа или отбора) и молекулярной эволюции во время и после их закрепления. Я приведу некоторые из этих моделей, основываясь на их классификации (Рисунок 1).

Неофункционализация

Модель неофункционализации была предложена Оно в 1970 году в книге “Evolution by Gene Duplication” (Ohno, 1970). Это была первая публикация, в которой постулировалась главнейшая роль дупликации генов в создании материала для эволюционных преобразований. Оно писал, что пока ген, имеющий функцию в организме, находится под действием отрицательного отбора, никаких мутаций в нем закрепляться не должно. Лишь с ослаблением отрицательного отбора вследствие дупликации в гене могут накапливаться изменения, потенциально ведущие к новым функциям. Вера в реальность модели Оно быстро распространилась в научном сообществе. Однако настоящее подтверждение она получила только в 1990-е, когда были получены данные по секвенированию многих геномов (Zhang, 2003).

Один из примеров был найден в геноме обезьян рода *Pygatrix*, семейство Мартышковые (Cercopithecidae) (Zhang, Zhang, Rosenberg, 2002). Пигатрикс знаменит тем, что питается в основном листьями, а не фруктами или насекомыми, как большинство других приматов. Листья перевариваются симбиотическими бактериями в кишечнике обезьяны, а сам пигатрикс переваривает этих бактерий. Один из ферментов, с помощью которых он это делает – панкреатическая рибонуклеаза (RNASE1). Этот ген претерпел дупликацию в геноме пигатрикса (получились две копии – RNASE1 и RNASE1B). Было показано, что одна из его копий (RNASE1) почти не претерпела изменений, а другая (RNASE1B) эволюционировала под действием положительного отбора таким образом, чтобы увеличить эффективность расщепления РНК бактерий в условиях кишечника (оптимальный pH для работы RNASE1B – 6.3, в то время как для RNASE1 – 7.4). Таким образом, одна из копий гена специализировалась для выполнения функции переваривания, в то время как другая продолжила выполнять физиологические функции вне кишечника (например, расщепление двухцепочечных РНК; другие функции RNASE1, впрочем, пока не ясны).

Модель DDC (Duplication-Degeneration-Complementation)

Модель DDC, предложенная в конце 90-х (Force et al., 1999) по сравнению с неофункционализацией учитывает два факта: (а) мутации после дупликации могут накапливаться в обеих копиях и (б) они будут в основном снижать способность этих копий к выполнению анцестральной функции. Однако, хотя каждая копия отдельно будет выполнять функцию менее эффективно, пока вместе обе копии будут справляться с ней так же хорошо, как предковый ген, отбор не будет влиять на закрепление или потерю дупликации. Если в результате дрейфа такая дупликация закрепится, в популяции будет два гена, вместе выполняющих одну функцию, вместо одного. Похожим образом, предковый ген мог экспрессироваться в различных органах, в то время как копии могут потерять способность к экспрессии в некоторых из них (главное, чтобы хотя бы одна копия экспрессировалась в каждом органе, в котором экспрессировался предковый ген).

Пример дупликации, развивающейся по сценарию DDC - гены SYN2A и SYN2B, найденные у рыбы фугу (гены, кодирующие синапсины – белки, регулирующие выброс нейромедиаторов в синапсах). Предковый ген, встречающийся, например, у человека – SYN2 – имеет две сплайсоформы (SYN2a и SYN2b), в то время как SYN2A потерял сплайсоформу SYN2b, а SYN2B – сплайсоформу SYN2a (Yu, Brenner and Venkatesh, 2003).

Специализация

Эта модель, предложенная Хьюзом (Hughes, 1994), предполагает наличие у предкового гена двух конкурирующих функций – то есть таких, которые он не может одновременно выполнять с максимальной эффективностью. Тогда после дупликации отбор будет поощрять мутации в копиях, увеличивающие эффективность выполнения одной функции в ущерб другим – поскольку другие функции может выполнять другая копия. Таким образом, получится два гена, лучше оптимизированных под выполнение функций предкового гена, чем он сам. Эта модель также называется “уходом от адаптивного конфликта”. Хьюз сам пишет, что этот механизм сейчас вряд ли имеет большое значение, поскольку гены, выполняющие конфликтующие функции, должны встречаться нечасто (как минимум потому, что большинство из них должны были уйти от адаптивного конфликта за миллиарды лет эволюции).

Адаптивное увеличение числа копий (positive dosage effect)

Суть модели проста и вытекает из названия: если приспособленность организма с дупликацией выше приспособленности организма без нее, отбор будет способствовать закреплению дупликации. Это может произойти из-за повышенной приспособленности при увеличенном уровне экспрессии гена, поскольку при увеличении числа копий гена обычно увеличивается уровень его экспрессии (Anderson, Roth, 1977).

Пример адаптивного увеличения числа копий можно найти у *Drosophila melanogaster* (Cardoso-Moreira et al., 2016). В исследовании было найдено несколько дупликаций, находящихся под положительным отбором. Например, ген CG9186, который влияет на процесс запасания жиров, в результате дупликации стал экспрессироваться более интенсивно, что вкупе с положительным отбором на дупликацию позволяет предположить адаптивность увеличения числа копий.

Защита от вредных мутаций

Этот вариант был предложен ещё Холдейном в 1933 году в контексте эволюционных эффектов рекуррентных мутаций (мутаций, имеющих повышенную вероятность возникновения, и поэтому происходящих чаще других) (Haldane, 1933). Он говорил о том, что если мутация часто происходит в каком-то гене, преимущество будут получать организмы, у которых есть несколько копий этого гена – при возникновении мутации в одной копии остальные продолжат работу. Поэтому

дупликации (и вообще амплификации) генов могут поддерживаться отбором. Впрочем, селективное преимущество организмов с дупликациями при таком сценарии, скорее всего, будет небольшим из-за редкости мутаций. Оно может быть значительным разве что у организмов с высокой частотой мутаций или в участках повышенного мутагенеза. Время жизни копии гена после дупликации при таком сценарии должно быть не очень большим – ведь единственное “назначение” такой копии заключается в том, чтобы она разрушилась под действием мутаций (Innan and Kondrashov, 2010).

Приобретение новых функций вследствие дупликации

Дупликация может произойти таким образом, что копия гена сразу будет находиться не в одинаковом положении с предковым геном. Например, ген может быть копирован не полностью, без регуляторных последовательностей, что может привести к экспрессии в других органах или в ответ на другие факторы среды (Lynch and Katju, 2004). Примерами могут служить Hox-гены ftz, zen и bcd у дрозофилы (Averof, Dawes, Ferrier, 1996)

Переход к постоянному гетерозиготному состоянию

Эта модель предполагает наличие балансирующего отбора по типу преимущества гетерозигот в рассматриваемом локусе. После дупликации гена, находящегося под балансирующим отбором, может произойти рекомбинация между новой копией и другим аллелем этого гена. Тогда на одной хромосоме будет оба аллеля, и отбор превратится из балансирующего в положительный, поддерживающий эту “постоянную гетерозиготу” (Spofford, 1969).

Баланс количества генов

В этой модели предполагается существование генов, оптимальное количество которых в геноме тесно связано с количеством других генов. Например, гена А должно быть столько же копий, сколько гена Б. Тогда в случае крупных дупликаций (например, полногеномных), обоих генов станет по два, и ни одна из копий не сможет потеряться, поскольку это нарушит баланс (Birchler et al., 2005).

Модель была вдохновлена наблюдениями за геномами растений (например, *Arabidopsis thaliana*) – было обнаружено, что после полногеномных дупликаций регуляторные гены сохраняются чаще, а именно для них характерна важность баланса количества копий (ibid).

Из этого неполного списка моделей (другие можно найти в обзоре Иннана и Кондрашова (Innan and Kondrashov, 2010); мы не включили их в этот обзор, потому что посчитали их отличия от описанных моделей незначительными) видно, что причин, по которым дупликации могут закрепляться и оставаться в геноме – великое множество. Вопрос о распространённости каждой из них в природе остаётся открытым, но

реальность многих из них подтверждается различными геномными данными (см. выше).

Дупликация – событие макромасштаба. Она поставляет материал для эволюции, но этот материал не нов. Новые же аминокислотные последовательности после дупликации могут получаться различными способами – точечными мутациями, инсерциями, делециями. Многие механизмы, о которых пойдёт речь дальше, могут вступать в действие уже после дупликации. Таким образом, сама дупликация поставляет только материал для создания новых последовательностей, но не является механизмом их образования.

Перемешивание доменов (Domain shuffling/exon shuffling)

В большинстве белков можно распознать доменную структуру – некоторые части белка приспособлены к выполнению одних функций, другие – под выполнение других. Было показано, что эти функциональные домены примерно соответствуют экзонам (Souza et al., 1996). Согласно модели перемешивания доменов, новые белки могут появляться путем перемещения экзонов в другие участки генома и, следовательно, добавления старым белкам новых функциональных доменов (Kaessmann et al., 2002). Есть два основных механизма, по которым это может происходить (Long et al., 2003). Первый – эктопическая рекомбинация (рекомбинация между негомологичными участками). Второй – ретротранспозиция (ретротранспозон может захватить с собой экзон функционального белка). Оба этих механизма были обнаружены в природе (Moran, 1999).

Слияние генов

Еще один способ образования новых генов - путём слияния (полного или частичного) старых. Появившиеся таким образом гены называют химерными. Некоторые механизмы, лежащие в основе образования химерных генов известны. Например, это сквозное прочтение двух последовательных генов. Другой возможный механизм — обратная транскрипция (ретропозиция) иРНК одного гена внутрь другого гена (что также можно отнести к перемешиванию доменов). По данным Вана и коллег (Wang et al., 2006) образование химерных генов по механизму ретропозиции встречается очень часто в геномах растений. У прокариот также встречается процесс, обратный слиянию — разделение одного гена на два (Long et al., 2003).

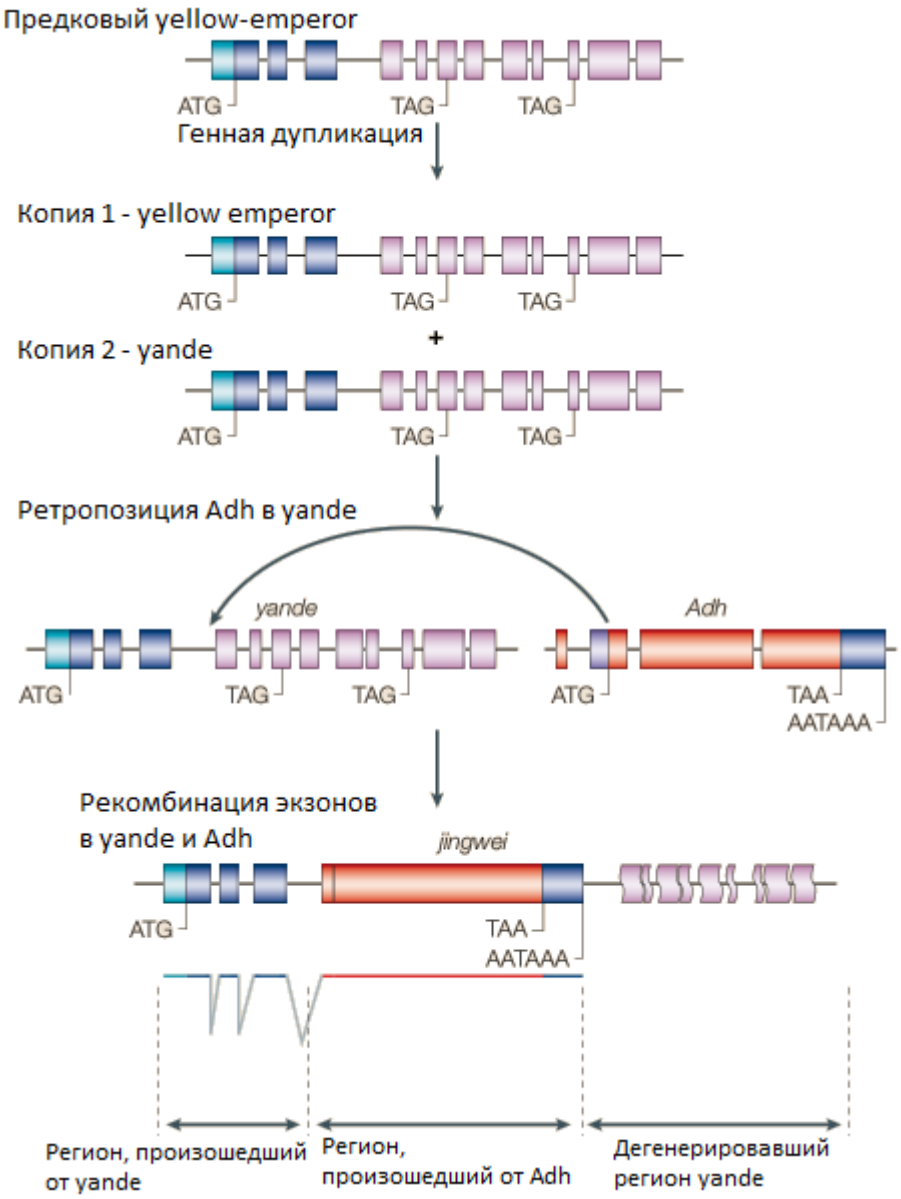

Рисунок 2. Образование химерного гена на примере гена jingwei (см. в тексте). По Long et al., 2003

Есть как минимум две причины, по которым отбор может сохранять химерные гены (Akiva et al., 2006). Во-первых, такие гены могут нести две функции, унаследованные от двух предковых генов, выполнение которых по каким-то причинам более удобно вместе, чем отдельно (см., например, Pradet-Balade, 2002). Во-вторых, может случиться так, что химерный ген сохранит функцию одного предкового гена и регуляцию другого. Это происходит и при ретропозиции (новый генетический контекст ведет к новой регуляции) и при сквозном прочтении (регуляторные участки такой химерный ген наследует от предкового гена, находившегося ближе к 5'-концу).

Насколько процесс слияния генов широко распространён, до конца непонятно. По некоторым оценкам (Akiva et al., 2006), около 5% генов человека подвергаются сквозному прочтению.

Пример химерного гена, образованного в результате считывания двух последовательных генов - человеческий ген Kua-UEV (Thomson, 2000). Ген UEV-1 кодирует один из коэффекторов убиквитинирования, локализованный в ядре. Функции гена Kua не известны. Образовшийся в результате их слияния Kua-UEV получил от Kua домен, обуславливающий его локализацию в цитоплазме, в результате чего он предположительно может участвовать в убиквитинировании цитоплазматических мишеней.

Пример гена, образовавшегося в результате ретропозиции - ген jingwei у некоторых видов дрозофил (Long and Langley, 1993). Было показано, что у этих видов ген алкоголь-дегидрогеназы вставился в другой участок генома, и получившийся химерный ген содержит помимо алкоголь-дегидрогеназы 77 аминокислот из этого участка (Рисунок 2). Опять же, функции этого гена не известны. Их наличие доказывается тем, что на гене jingwei видны следы положительного отбора.

Увеличение числа повторов

Повторы — участки ДНК с повторяющейся несколько раз одинаковой последовательностью составляют значительную часть генома многих эукариот. Например, у некоторых млекопитающих около 95% генома представляют собой повторы различной длины (Gibbs et al., 2004). Большая часть этих повторов составляет белок-некодирующие участки ДНК, такие как интроны, транспозоны или теломеры, но в белок-кодирующих участках они также встречаются (Marcotte et al., 1999).

В череде поколений длина повтора может увеличиваться или уменьшаться (Рисунок 3), например за счет рекомбинации или проскальзывания ДНК-полимеразы при репликации (Hancock and Simon, 2005). В последнем случае, процесс отличается от обычных инсерций или делеций повышенной интенсивностью - риск ошибки ДНК полимеразы в области повтора во много раз выше, чем в других участках (Schlötterer, 2000). В случае, если повтор находится в белок-кодирующем участке, изменения в его длине приведут к изменению размеров белка. Такой процесс наблюдается в генах, кодирующих коллаген: длина повторов варьирует от 100 до 500 аминокислот (Marcotte et al., 1999). Потенциально увеличение длины белка за счет экспансии повторов может вести к приобретению им новых функций либо просто за счет изменения размера либо за счет накопления мутаций во вновь образовавшихся участках.

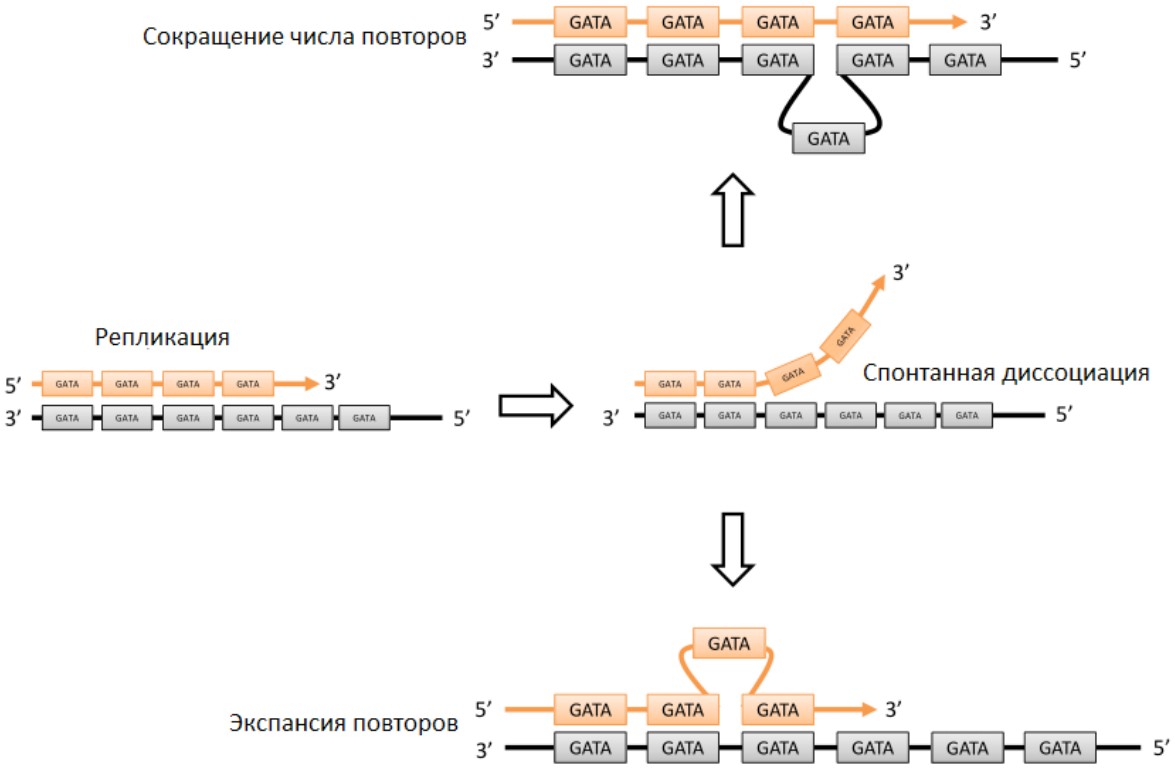

Рисунок 3. Изменение числа повторов в результате проскальзывания ДНК-полимеразы. По Hansson, 2018.

Смена поколений у эукариот более медленная по сравнению с прокариотами. Это ведет к более низкой скорости эволюции. По мнению Маркота с коллегами (Marcotte et al., 1999), высокая частота повторов в геноме эукариот может быть призвана скомпенсировать эти потери в скорости эволюции, поскольку области с повторами эволюционируют гораздо быстрее других. Другая версия, подтверждающаяся их данными, состоит в том, что повторы хороши для выполнения специфических для белков эукариот функций. Например, повторы содержат белки цитоскелета, мышц, синапсов и клеточной адгезии. Все эти белки выполняют функции, нехарактерные для прокариот.

Двойное кодирование в одном участке

Белок-кодирующий участок генома не всегда кодирует один белок. Не говоря об альтернативном сплайсинге, встречаются случаи двойного кодирования – когда один и тот же участок может быть прочитан РНК-полимеразой двумя способами, что приведет к синтезу двух разных иРНК. Механизмы двойного кодирования – это кодирование в альтернативных рамках считывания (со сдвигом на 1 нуклеотид вперед или назад) и кодирование в другой цепи ДНК. В обоих случаях синонимичная мутация в одной аминокислотной последовательности будет несинонимичной в другой: в

первом – из-за использования одних и тех же нуклеотидов, во втором – из-за однозначного соответствия нуклеотидов на комплементарных цепях ДНК.

Почему отбор может сохранять двойное кодирование на некоторых участках генома? Есть несколько возможных причин.

Прежде всего, очевидно, двойное кодирование позволяет сделать геном меньше. Отбор на малый объем генома идет у многих прокариот, но прежде всего у вирусов (особенно ретровирусов). Поэтому неудивительно, что случаи двойного кодирования встречаются у прокариот (Kovacs et al., 2010) и у вирусов (Belshaw, Pybus, Rambaut, 2007), однако они также были найдены и у эукариот, у которых отбор на малый объём генома если и есть, то в значительной степени ослаблен (Kovacs et al., 2010).

Другая возможная причина сохранения двойного кодирования состоит в том, что в среднем замена аминокислоты влияет очень мало на функцию белка (Eyre-Walker, Keightley, 2007). Это, вероятно, обусловлено обилием неупорядоченных участков во многих белках. Поскольку в них нет жесткой структуры, требующей конкретных аминокислот, изменение последовательности не сильно сказывается на приспособленности. В этом случае двойное кодирование накладывает лишь небольшие ограничения и получается ничем не хуже обыкновенного одинарного (Kovacs et al., 2010).

Также кодирование двух белков одним участком может быть полезно, если оба эти белка вовлечены в разные пути одного процесса. Тогда путем модификации иРНК можно быстро переключать эти пути, синтезируя разные белки (Szklarczyk et al., 2007).

Примером кодирования на комплементарной цепи служит ген TRA - рецептор к трийодтиронину (Keese, Gibbs, 1992). Это ядерный рецептор, состоящий из двух частей – ДНК-связывающего и лиганд-связывающего доменов. У человека и крысы ген этого рецептора имеет две сплайсоформы: TRA1 и TRA2. 8 экзонов (370 аминокислот) общие для этих сплайсоформ и кодируют ДНК-связывающий домен. 9-й экзон (40 аминокислот) используется в TRA1 и кодирует лиганд-связывающий домен, а 10-й (120 аминокислот) – в TRA2, его аминокислотная последовательность не связывает трийодтиронин. TRA2 является отрицательным регулятором экспрессии TRA1. Последние 263 нуклеотида 10-ого экзона перекрываются последним экзоном гена ear-1, кодируемого на комплементарной цепи. Этот участок ear-1 тоже связывает трийодтиронин и кодирующая его последовательность похожа на 9-й экзон TRA1. Похоже, TRA1 и ear-1 – паралоги, а 10-й экзон TRA возник de novo на альтернативной цепи (Рисунок 4).

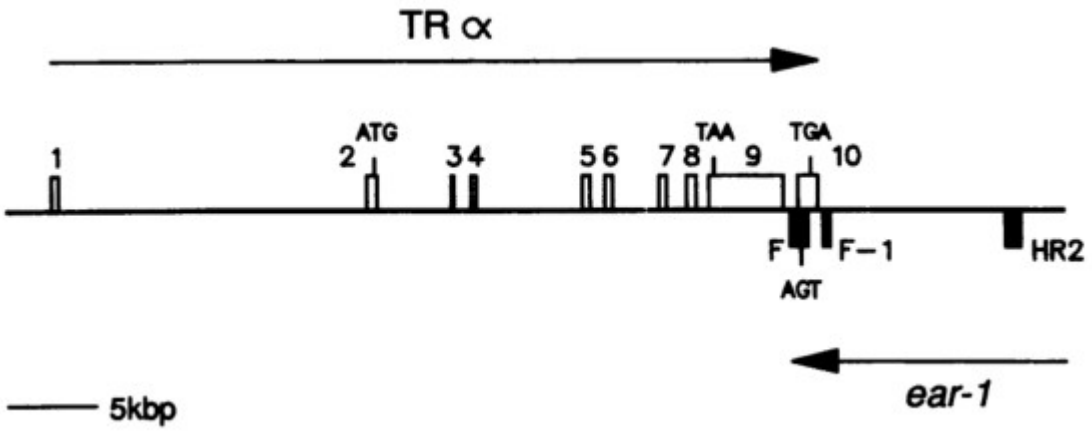

Рисунок 4. Генетическая карта гена рецептора трийодтиронина TRA и его паралога ear-1. Отмечены экзоны, старт- и стоп-кодоны. Стрелки указывают направление транскрипции. По Keese&Gibbs, 1992.

Пример генов, кодируемых перекрывающимися альтернативными рамками считывания – человеческие гены INK4a и ARF, кодирующие белки p16 и p14 соответственно. Оба этих белка участвуют в подавлении раковых заболеваний. Оба транскрипта начинаются разными экзонами (1A и 1B), следующий (второй) экзон считывается в двух альтернативных рамках (Рисунок 5). С учетом того, что белки p16 и p14 вовлечены в разные метаболические пути, эволюционное преимущество от их кодирования в одном участке генома не ясно (Szklarczyk et al., 2007).

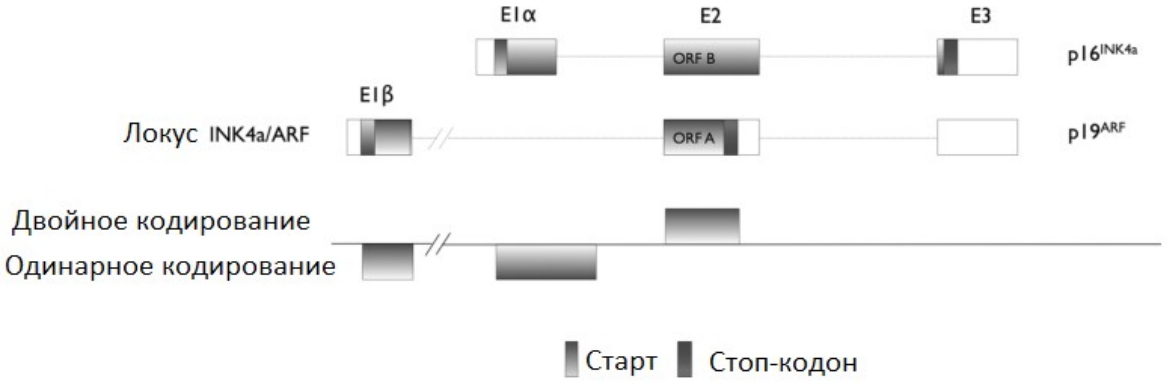

Рисунок 5. Строение участка, кодирующего INK4a и ARF. Сверху показана структура экзонов обоих транскриптов. Снизу отмечено, какие экзоны кодируют одну аминокислотную последовательность, а какие – две. По Szklarczyk et al., 2007.

“Из мусора”

Долгое время считалось, что новые гены могут появляться лишь в результате модификации старых (путем дупликации, слияния и т.д.). Возникновение гена из некодирующей последовательности рассматривалось как совершенно невероятное.

Франсуа Жакоб в 1977 году писал: “The probability that a functional protein would appear de novo by random association of amino acids is practically zero” (Jacob, 1977). Тем не менее, последние 10 лет накапливаются свидетельства того, что гены из некодирующих последовательностей всё-таки возникают. Появление гена таким способом (“de novo”) сложно отследить – даже если ортолог среди близкородственных видов найти не получается, всегда остается вероятность того, что ген быстро эволюционировал и стал совершенно не похож на свои ортологи (Domazet-Loso, Tautz, 2003). Но чем более близкие виды мы рассматриваем, тем больше можем быть уверены в происхождении гена. И более-менее достоверные случаи появления генов из некодирующих участков были найдены у некоторых растений, малярийного плазмодия, дрожжей, дрозофилы, мышей и человека (Neme,Tautz, 2013).

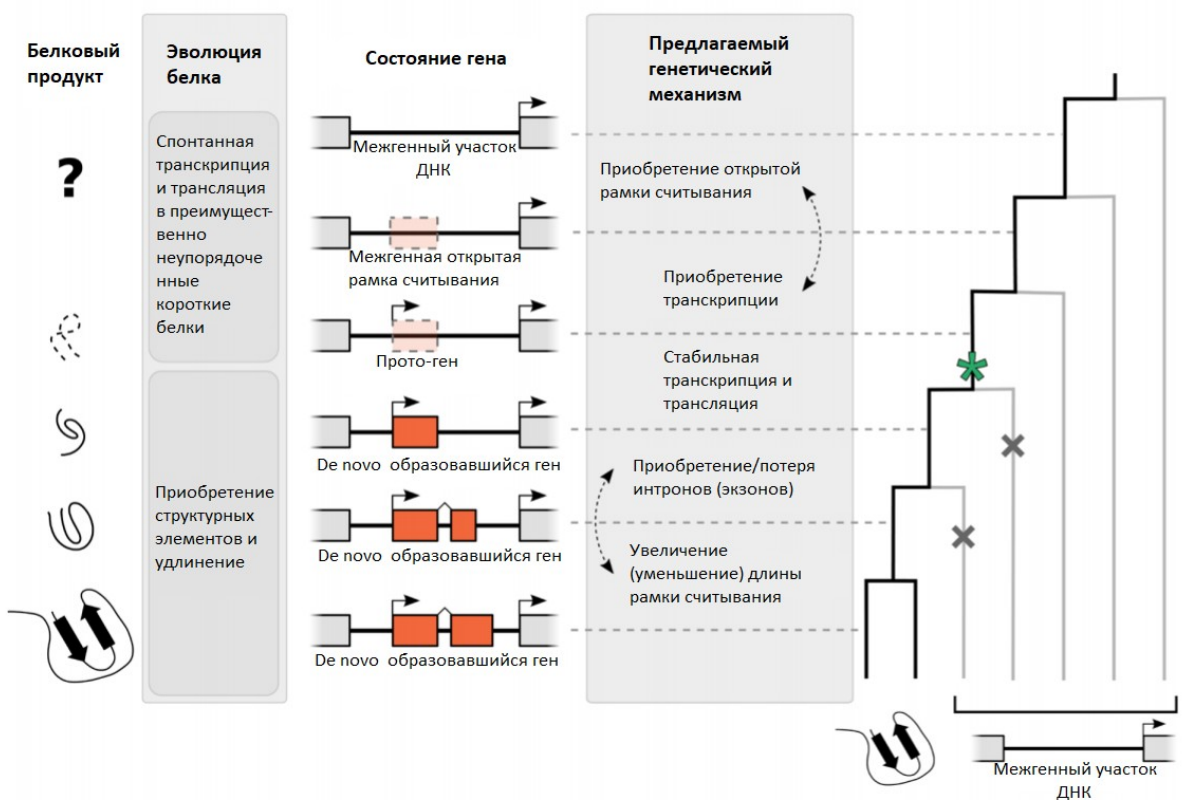

Рисунок 6. Гипотетическая последовательность событий при de novo возникновении белок-кодирующего гена. По Шмитсу и Борнерг-Бауэру, 2017.

Есть два возможных сценария de novo возникновения белок-кодирующего гена (Рисунок 6). В первом из них сначала в результате мутаций в межгенном участке появляется открытая рамка считывания (т.е. старт- и стоп-кодоны), которая затем начинает транскрибироваться и транслироваться. Такие транскрибируемые единицы называют прото-генами. Второй сценарий предполагает, что сначала межгенный участок начинает транскрибироваться в некодирующую РНК, а уже затем в нем в результате мутаций появляются старт- и стоп-кодоны (Schmitz, Bornberg-Bauer, 2017).

Пример гена, появившегося de novo у человека – FLJ33706 (Li et al., 2010). Этот ген кодирует последовательность из 194 аминокислот и несет на себе следы

отрицательного отбора, что предполагает наличие функции. Он экспрессируется в нейронах мозжечка, среднего мозга и коры больших полушарий. Точная функция FLJ33706 не известна, но было показано, что он участвует в развитии болезни Альцгеймера (ibid).

Похоже, образование новых генов из некодирующих последовательностей ДНК играет немалую роль в эволюции. Исследование 2008 года (Zhou et al., 2008) показало, что 11.9% генов новых генов у *Drosophila melanogaster* произошли именно таким образом.

Сдвиги рамки считывания

Мутации сдвига рамки считывания до некоторых пор считались неважными для эволюции, поскольку последовательность с такой мутацией скорее всего будет содержать ранний стоп-кодон и последовательность случайных аминокислот перед ним, что вряд ли может оказаться функционально значимым (Ohno, 1970). Однако существует несколько случаев, когда эти мутации могут быть не такими уж вредными. Во-первых, если мутация случится недалеко от С-конца белка, будет изменено немного аминокислот, и длина и аминокислотный состав белка изменятся мало. Во-вторых, если одна мутация будет скомпенсирована другой (так, что в сумме их длина будет кратна трем), аминокислоты изменятся только между ними, что тоже может привести к небольшой потере приспособленности (или не привести вообще). И, наконец, если на ген был в силу тех или иных причин ослаблен отбор (из-за дупликации или из-за потери актуальности функции гена), радикальные изменения в его длине и составе аминокислот могут не сильно влиять на приспособленность (Raes, van de Peer, 2005).

Закрепленные сдвиги рамки на конце белков были обнаружены по большей части в трансмембранных белках и в транскрипционных факторах (ibid). В первых – потому что С-конец трансмембранных белков как раз и несет функциональное значение – связывание лиганда или передача сигнала. Поэтому изменение С-конца может привести к новым функциям. Во вторых – потому что С-концы транскрипционных факторов обычно вовлечены в белок-белковые взаимодействия и трансактивацию и репрессию генов, и, опять же, мутации в них могут привести к новым функциям.

Скомпенсированные сдвиги рамки считывания также были обнаружены у млекопитающих (Hu, Ng, 2012) (Рисунок 7) и у растений (Long et al., 2013). Их функциональное значение пока не ясно, но факт их существования говорит о том, что скомпенсированные сдвиги рамки могут делать вклад в образование аминокислотных последовательностей в эволюции, несмотря на относительную радикальность таких изменений.

Рисунок 7. Пример скомпенсированного сдвига рамки считывания у собаки. 1 и 2 не кратны трём, в то время как 1+2=3 кратно трём, поэтому в последовательности после второй делеции рамка не сдвинута. По Hu, Ng, 2013.

Мы поставили своей целью изучить роль скомпенсированных сдвигов рамки считывания в эволюции белок-кодирующих генов животных. Мы использовали данные о двух группах: позвоночных и двукрылых. Для выполнения этой цели мы поставили следующие задачи:

1) Получить данные, подходящие для наших задач.
2) Разработать и имплементировать алгоритм поиска скомпенсированных сдвигов рамки считывания в этих данных.
3) Придумать и реализовать способы верификации найденных случаев.
4) Изучить действие отбора на белки со скомпенсированными сдвигами рамки считывания и на участки, в которых они произошли.
5) Изучить эффект от сдвига рамки и последующих однонуклеотидных замен на структуру белка.

Материалы и методы

Данные и их предобработка

В качестве первичных данных мы использовали выравнивания экзонов 100 позвоночных и 124 двукрылых (ссылки на списки животных можно найти в приложении), выполненные с помощью MULTIZ, находящиеся в открытом доступе в геномном браузере UCSC (Rosenbloom et al., 2015). Однако из-за некоторых неудобных особенностей этих данных нам пришлось их перевыравнивать. Первая из таких особенностей — это, похоже, ошибка в алгоритме выравнивания MULTIZ, поскольку некоторые делеции, присутствующие в этом выравнивании, отсутствуют в геномах соответствующих животных, использовавшихся для построения этого выравнивания. Вторая особенность состоит в том, что авторы выравнивания использовали геном человека (или *Drosophila melanogaster* в случае двукрылых) в качестве референсного (это значит, все последовательности в множественном выравнивании выравнивались на него), и все позиции, в которых у человека была делеция, были вырезаны из выравнивания. Хотя такая обработка выравнивания вполне годится для многих задач, для наших она совершенно не подходит. Поэтому мы сделали собственное выравнивание. Мы использовали экзонную аннотацию из соответствующих выравниваний MULTIZ, чтобы идентифицировать экзоны в геномах (мы использовали те же геномы, что и авторы выравнивания MULTIZ). Полученные экзоны мы соединили и выровняли с помощью mafft (Katoh et al., 2002) (параметры: --maxiterate 1000 --globalpair --preservecase). В выравнивании двукрылых мы полностью убрали геномы D_pseudoobscura_1 и A_gambiae_1, поскольку не смогли их найти (геномы нужны были, чтобы достать из них гены по аннотации). Впрочем, это не должно влиять на результаты, поскольку в изначальное выравнивание эти геномы были добавлены лишь для поддержки droPse3 (*Drosophila pseudoobscura*) и anoGam3 (*Anopheles gambiae*) — других геномов тех же самых животных. Мы также убирали из выравнивания куски генов в начале и в конце, если они не выравнивались на ген референсного животного, потому что генная аннотация была сделана по геному референсного животного, и, значит, гены остальных животных должны начинаться и заканчиваться в тех же самых местах. По не вполне понятным причинам (скорее всего, из-за большой длины некоторых генов) выровнять получилось только 21208 генов из 21521 у позвоночных и 27341 генов из 30482 у двукрылых, и только их мы использовали в дальнейшем анализе.

Алгоритм поиска скомпенсированных сдвигов рамки считывания

Чтобы найти скомпенсированные сдвиги рамки считывания, мы придумали алгоритм и имплементировали его в скрипте на языке python (https://github.com/Captain-Blackstone/Compensatory-frameshifts). Он получает на вход файл в формате phylips, содержащий множественное выравнивание одного гена, и

выдаёт на выходе названия (инсерции, делеции), длины, позиции инделов и названия животных, в которых эти инделы были найдены.

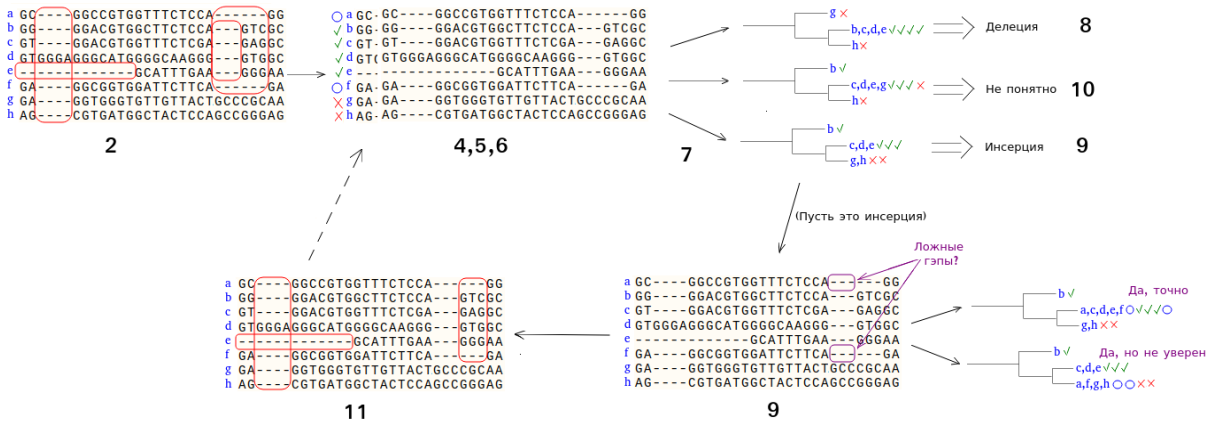

Рисунок 8. Схема работы алгоритма нахождения инсерций и делеций в выравнивании. Буквами a-h обозначены кусочки генов разных животных. Зелеными галочками обозначены животные, имеющие рассматриваемую в данный момент дыру (см. описание алгоритма). Красными крестиками - животные, у которых она отсутствует. Синими кружочками - остальные (для них нет названия, у них она и не присутствует и не отсутствует). Числа соответствуют пунктам в описании алгоритма.

Алгоритм работает так (удобнее всего будет сверяться с рисунком 8 при чтении пунктов 2-11):

1. Убрать все гены, состоящие только из гэпов (вот таких символов: «-»)
2. Определить “дыры” в выравнивании. Дырой я называю череду гэпов, идущих подряд, которая в дальнейшем может быть классифицирована как делеция (у животных с ней) или инсерция (у животных без неё). Число гэпов в этой череде я называю длиной дыры. Позицию первого из них я называю позицией дыры.
3. Пока список дыр не пустой, повторять пункты 3-11.
4. Выбрать одну из самых коротких дыр в списке.
5. Найти всех животных, у которых есть эта дыра (дыра такой же длины на такой же позиции). Обратите внимание, что если у животного на этой позиции есть дыра большего размера, то я не считаю, что у него есть эта дыра.
6. Найти всех животных, у которых на месте этой дыры (на всех позициях, входящих в эту дыру) стоят только нуклеотиды.
7. Ответить на вопрос: если рассматривать только две эти группы животных, формирует ли кладу одна из этих групп? Или обе?
8. Если только первая группа (с дырой) формирует кладу, классифицировать дыру как делецию.
9. Если только вторая группу (без дыры) формирует кладу, классифицировать дыру как инсерцию. Здесь появляются некоторые сложности. Если дыра — это результат инсерции у животных, не имеющих этой дыры, все гэпы находящиеся на позициях дыры могут быть также результатом этой инсерции, даже если они не относятся к этой дыре (а, например, относятся к более крупной, «проходящей» через эти позиции дыре). Например, пусть мы нашли инсерцию длины 3 на позиции 23 в одном животном, а у другого животного есть дыра длиной 6 на этой же позиции 23. Тогда мы, возможно,

должны рассматривать 3 гэпа в этой дыре как «ложные», появившиеся в результате инсерции у другого животного. Значит, дыра длины 6 превратится в дыру длины 3. Возможно, однако, и не должны, если эта дыра представляет собой делецию, длины 6, произошедшую после инсерции длины 3. Такие случаи двусмысленны, и, насколько мы понимаем, принципиально не могут быть разрешены с использованием одной лишь информации из выравнивания. Поэтому такие гэпы мы всегда считали ложными, но помечали их как «случаи, в которых мы не уверены».
10. Если обе группы формируют кладу или обе не формируют, оставить дыру неклассифицированной и игнорировать её. Впрочем, она помечается как «странный случай».
11. Убрать дыру из списка найденных дыр.
12. После того, как все инсерции и делеции классифицированы, отфильтровать гены. Убрать гены
- длиной не кратной трём или
- не имеющие начала или конца или
- имеющие стоп-кодон не на конце.
13. Для каждого оставшегося животного выполнить пункты 14-18.
14. Выбросить инделы длиной кратной трём. Они точно не скомпенсированные.
15. Также выбросить длинные инделы (>20 нуклеотидов), общие для данного животного и всех потомков родительского узла этого животного. Этот шаг нужен для того, чтобы длинные невыровнявшиеся участки у базальных животных не считались инсерциями у других животных (пусть, например, есть дыра длиной 100 в паре базальных животных — скорее всего, не выровнявшийся экзон. Алгоритм определит последовательности в остальных животных как инсерции, хотя это неразумно. Эта процедура проводится как раз для того, чтобы избавиться от таких «неправильных» инсерций). Такая ситуация представлена на рисунке 9.

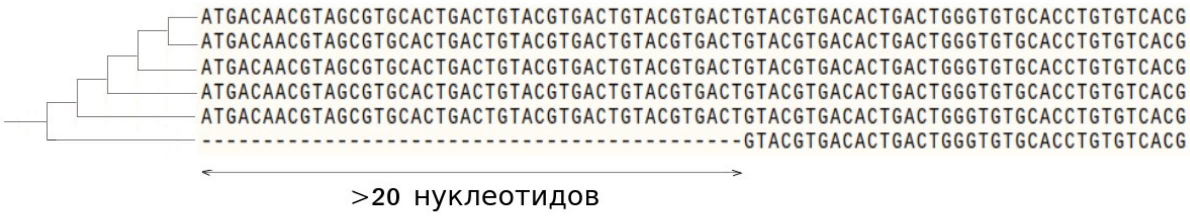

Рисунок 9. Алгоритм классифицирует показанную ситуацию как инсерцию у 5 верхних животных. Нам кажется, что в таких случаях более правдоподобно предположение о плохо выравнявшемся куске, поэтому такие инсерции мы отбрасывали.
16. Если после у животного осталось более 2 инделов, выбросить это животное (не выполнять пункты 17-18). Скорее всего, этот ген плохо отсеквенирован или плохо выравнен. Впрочем, обработка генов с множественными инделами была имплементирована, но мы не использовали её, чтобы не иметь дела с потенциально подозрительными случаями.
17. Если инделы имеют одинаковое название (инсерция-инсерция или делеция-делеция), считать их скомпенсированными, если их суммарная длина кратна трём (их длины по отдельности не кратны трём, иначе мы бы их уже отфильтровали). Если

инделы имеют разные названия (инсерция-делеция или делеция-инсерция), считать их скомпенсированными, если разность их длин кратна трём.
18. Если скомпенсированные инделы были найдены, добавить выброшенных ранее (в пункте 13) животных с такими же инделами в рассмотрение. Если у них такие же инделы, как у хороших, «доверенных» животных, они, вероятно, настоящие.
19. Также, если скомпенсированные инделы были найдены, проверить, случились они одновременно или нет. Для этого сравнить последнего общего предка животных с первым инделом с последним общим предком животных со вторым инделом. Если они одинаковые, инделы произошли одновременно, если нет — нет. Запомнить: животное, в котором были найдены инделы, названия инделов, длины, позиции, расстояние между инделами в нуклеотидах, одновременность происхождения инделов.
20. Выдать всё запомненное в 19 пункте для каждого животного.

Таблица 1. Пример выдачи алгоритма. Из подобных строк создаются два файлы, отдельно для одновременных и неодновременных инделов.

Код животного	Название файла	Первый индел: Тип_позиция_длина	Второй индел: Тип_позиция_длина	Расстояние между инделами
calJac3	uc010utv.1	Del_137_2	Del_312_1	130
equCab2	uc010utv.1	In_1379_1	Del_1392_1	12

Фильтрация найденных скомпенсированных инделов

После нахождения всех скомпенсированных инделов, мы решили рассматривать только короткие — которые имеют суммарную длину не более 4 (1 и 2 для инделов с одинаковыми названиями, 1 и 1 или 2 и 2 для инделов с разными), потому что вероятность встретить индел падает с увеличением его длины (Leushkin et al., 2013). Мы просмотрели вручную все такие инделы, чтобы исключить ошибки выравнивания. Затем мы сконцентрировались на случаях, где скомпенсированные инделы нашлись в более чем одном животном, поскольку такие инделы наиболее надёжны.

Поскольку именно такие сдвиги мы считаем наиболее заслуживающими доверия, мы приложили некоторые усилия к тому, чтобы увеличить их число в позвоночных. Для каждого гена с одним животным со скомпенсированным сдвигом мы попробовали найти ортологичный ген в животном, отсутствующем в нашем выравнивании, но присутствующем в базе данных ENSEMBL, которое было бы ближе по филогенетическому дереву к животному со сдвигом, чем любое присутствующее в выравнивании. Если у такого животного из ENSEMBL тоже находился

скомпенсированный сдвиг, случай тоже считался надежным и добавлялся в список генов со скомпенсированным сдвигом у нескольких животных.

Для дальнейшего подтверждения существования наблюдаемых сдвигов мы искали в базе данных NCBI те же самые гены у всех животных со сдвигом и у некоторых без сдвига с помощью BLAST. Если было подтверждено наличие инсерций и делеций у животных со сдвигом и их отсутствие у животных без сдвига, случай считался надёжным.

Результаты

Найденные инделы

Мы нашли 536 генов, в которых хотя бы у одного животного был скомпенсированный сдвиг рамки считывания, не похожий на ошибку выравнивания, в позвоночных и 182 таких гена в двукрылых. Однако только в 15 из этих генов в позвоночных и 12 в двукрылых один и тот же скомпенсированный сдвиг был у более чем одного животного. Мы нашли 55 генов, для которых была возможна проверка в ENSEMBL (см. раздел “Фильтрация найденных скомпенсированных инделов” в Материалах и методах). Только в одном из них животное со скомпенсированным сдвигом имело близкого родственника в ENSEMBL с таким же сдвигом. В 36 из этих генов скомпенсированные инсерции и делеции отсутствовали в ENSEMBL даже в животном, в котором они присутствовали в наших данных. Если доверять ENSEMBL больше и считать такие случаи ложноположительными, можно оценить FDR = 0.65 ± 0.13 (95% доверительный интервал по методу Вальда), и настоящее число найденных генов со скомпенсированными сдвигами рамки можно оценить как 185 ± 70 для позвоночных и 63 ± 24 для двукрылых. Поскольку мы хотели сконцентрироваться на самых надёжных случаях, мы в дальнейшем работали только с генами, в которых скомпенсированные сдвиги встречались у нескольких животных. После проверки с помощью NCBI BLAST таких генов осталось всего 6 у позвоночных и 5 у двукрылых (Таблица 2).

Таблица 2. Гены с несколькими животными с одним и тем же скомпенсированным сдвигом, прошедшие проверку с помощью BLAST. * - ген, для которого был найден близкий родственник со скомпенсированным сдвигом в ENSEMBL.

	Название гена	Число животных со скомпенированным сдвигом	Числло нуклеотидов между скомпенированными инделами	Типы инделов
Позвоночные	RAB36	11	99	Делеция-делеция
	ARHGAP6	9	7	Делеция-делеция
	INSL6	9	14	Инсерция-инсерция
	NCR3LG1	4	32	Делеция-инсерция
	RASSF4	3	26	Делеция-делеция
	SPATA24*	1	14	Делеция-инсерция
Двукрылые	Diap1	11	16	Делеция-делеция
	CG31530-RA	2	10	Делеция-делеция
	wds	4	19	Делеция-делеция
	Cyp6a23	4	5	Делеция-делеция
	Osi10	4	15	Делеция-делеция

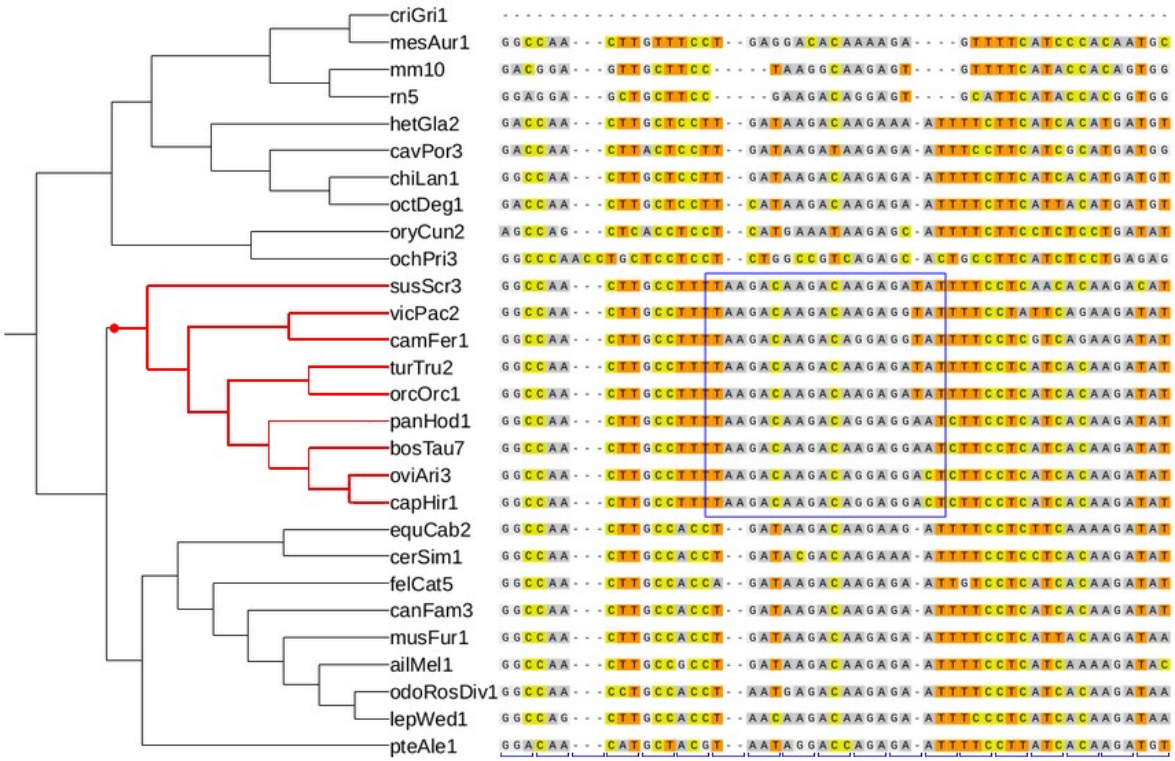

Рисунок 10. Пример скомпенсированного сдвига рамки считывания в гене INSL6. Красным отмечена клада, в которой произошёл скомпенсированный сдвиг рамки. Снизу показана предковая рамка считывания, нарушенная при сдвиге на участке, выделенном синим прямоугольником.

Действие отбора на гены со скомпенсирванными сдвигами рамки считывания

Мы проверили гипотезу о том, что скомпенсированные сдвиги рамки происходят с большей вероятностью в генах с ослабленным отрицательным отбором. Для этого мы использовали значение ω - параметр, отвечающий за силу отбора в упрощенной модели замен Goldman and Yang (1994), описанной в Yang et al., 1998. Этот параметр - изменение вероятности происхождения замены благодаря тому, что эта замена несинонимичная. Другими словами, если синонимичные замены происходят со скоростью q (например, на нуклеотид на поколение), несинонимичные будут происходить со скоростью ωq. ω=1 говорит о нейтральности, ω<1 - об отрицательном отборе, ω>1 - о положительном. Мы посчитали ω для всего дерева для каждого гена в нашем распоряжении, с помощью PAML, чтобы получить референсное распределение, и сравнили ω генов со сдвигами с этим распределением (Рисунок 11). Гены под ослабленным отрицательным или под положительным отбором оказываются в правой части графика. Как видно из рисунка, хотя отметку 0.05 (условную границу значимости) проходят только INSL6 и Diap6, большая часть генов (все, кроме RAB36 и wds) находится в правой части распределения.

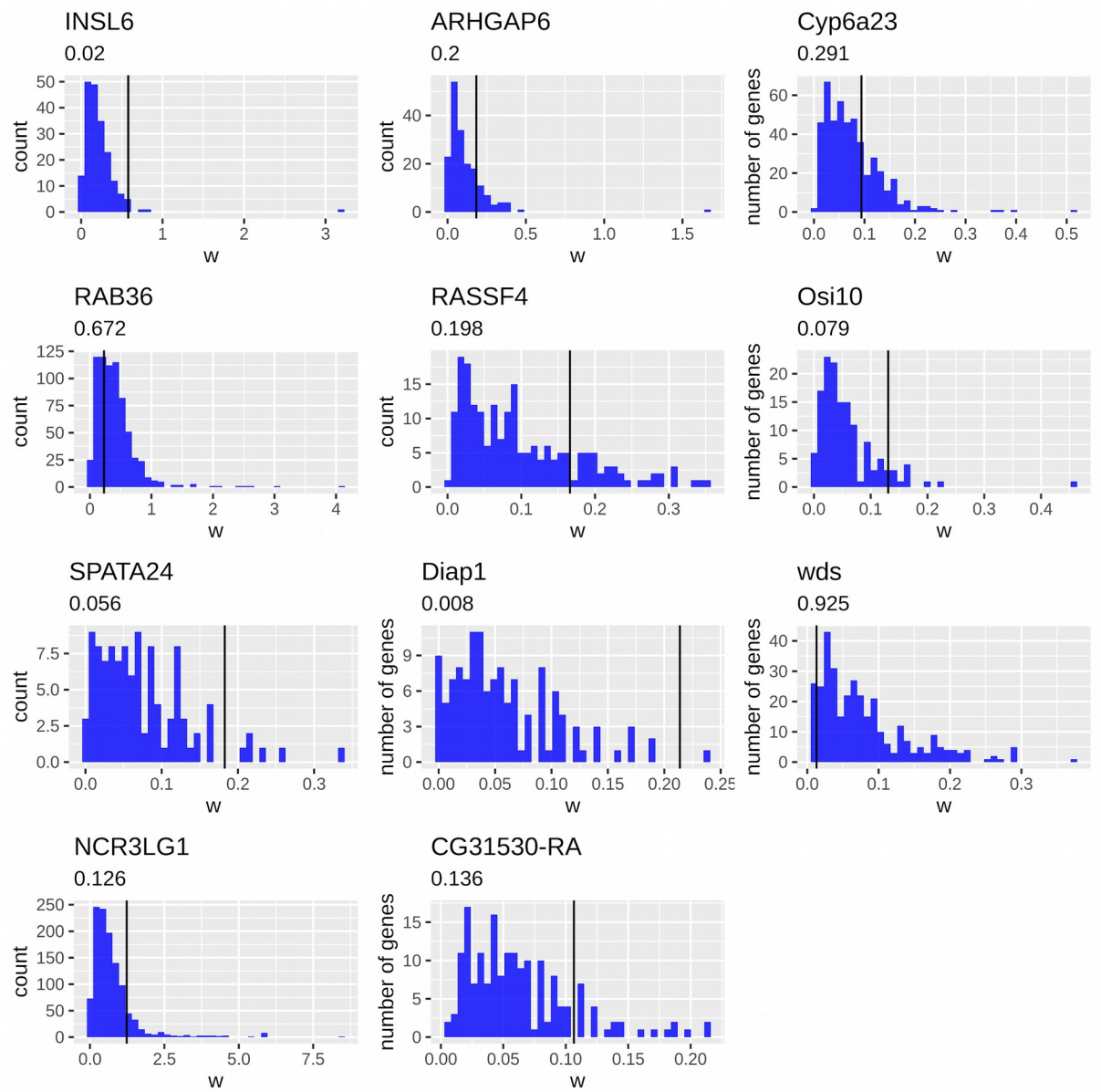

Рисунок 11. ω, вычисленная для каждого гена для всего дерева позвоночных или двукрылых. Каждому гену со скомпенсированным сдвигом отведена отдельная картинка. Синие гистограммы отображают количество генов из референсного распределения с соответствующей ω. Черная линия на графике соответствует ω гена со компенсированным сдвигом рамки. Доля генов справа от черной линии выписана под названием гена. Поскольку ω были определены с использованием только животных, проходящих по фильтрам (число нуклеотидов в гене кратно трем и остальные, описанные в методах), число животных варьировало между генами, что может зачительно влиять на ω, определенную для всего дерева. Для каждого гена со сдвигом рамки в референсном распределении учитывались только гены с тем же числом животных.

Также мы проверили гипотезу о том, что режим отбора меняется в генах, в которых происходят скомпенсированные сдвиги рамки считывания. Для этого мы запустили PAML с параметром model=2 (с возможностью подбора одной ω для ветки со скомпенсированными сдвигами, а второй — для остального дерева). Мы сравнили эти результаты с моделью с одной ω с помощью отношения правдоподобия (Таблица 2).

Только для одного гена, CG31530-RA у двукрылых, тест показал значимое увеличение ω на ветке со скомпенсированными инделами.

Таблица 2. Результаты LRT для позвоночных. ω – омега для всего дерева в предыдущей модели, ω1 – омега для дерева, кроме ветки животных со скомпенсированными сдвигами рамки, ω2 – омега для ветки животных со скомпенсированными сдвигами рамки. Колонка значимости рассчитана с учётом поправки Бонферрони.

Белок	ω	ω1	ω2	LRT	p-value	Значимо?
RASSF4	0.166	0.163	0.254	3.78	0.05	Нет
NCR3LG1	1.23	1.71	0.867	1.76	0.183	Нет
RAB36	0.229	0.24	0.211	0.22	0.636	Нет
SPATA24	0.183	0.181	0.284	0.734	0.391	Нет
INSL6	0.578	0.605	0.438	2.04	0.153	Нет
ARHGAP6	0.184	0.184	0.187	0.01	0.912	Нет
Diap1	0.214	0.213	0.235	0.207	0.649	Нет
CG31530-RA	0.107	0.101	0.155	11.013	0.001	Да!
wds	0.013	0.013	0.015	0.128	0.721	Нет
Cyp6a23	0.094	0.094	2.129	0	1	Нет
Osi10	0.131	0.128	0.177	3.45	0.063	Нет

Консервативность белков со скомпенсированными сдвигами рамки считывания

Мы проверили ещё два предположения:

- о том, что сдвиг рамки скорее произойдёт в консервативных белках, чем в неконсервативных.
- и о том, что сдвиг рамки скорее произойдет в менее консервативных участках белков, чем в более консервативных;

Консервативностью белка мы считали среднюю консервативность всех его аминокислотных позиций. Консервативность аминокислотной позиции определялась как доля наиболее часто встречающейся аминокислоты в данной позиции (так, например, если у всех животных в данной позиции стояла одна и та же аминокислота, консервативность равнялась 1, а если у всех разная — 1/n, где n — число животных в выравнивании).

Для проверки первой гипотезы мы определили консервативность для каждого белка в нашем распоряжении, имеющегося у данного таксона, и сравнили консервативность белков со сдвигами с этим распределением (Рисунок 12). Видно, что, хотя белки со скомпенсированным сдвигом рамки находятся не вблизи экстремальных значений консервативности, они все же чаще менее консервативны, чем в среднем.

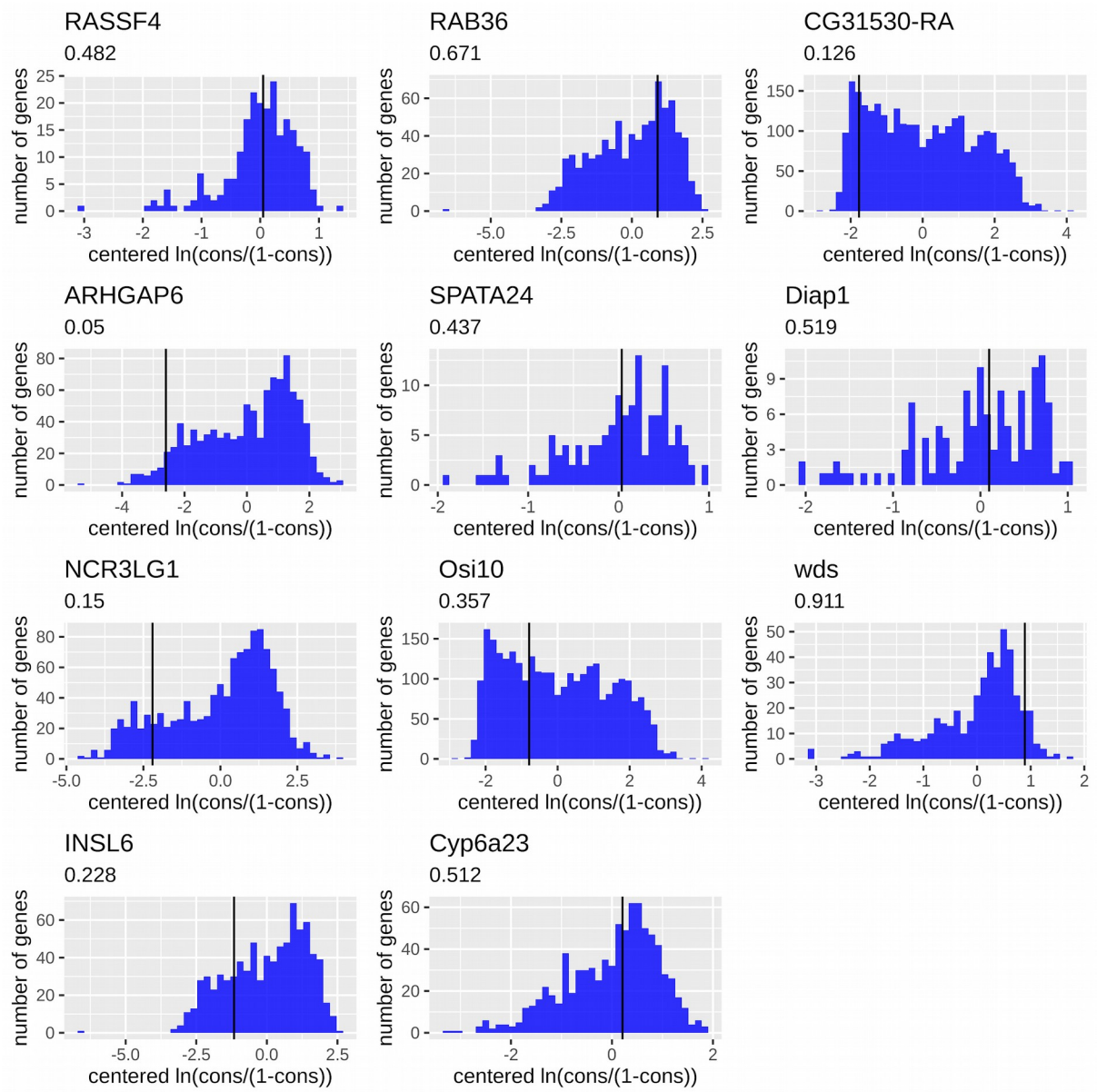

Рисунок 12. Консервативность белков со скомпенсированными инделами по сравнению с остальными белками соответствующего таксона. Каждому гену со скомпенсированным сдвигом отведена отдельная картинка. Синие гистограммы отображают количество генов из референсного распределения с соответствующей консервативностью. Черная линия на графике соответствуют консервативности гена со компенсированным сдвигом рамки. Доля генов справа от черной линии выписана под названием гена. Как в случае с омегами, для каждого гена со сдвигом рамки в референсном распределении учитывались только гены с тем же числом животных в выравнивании (после фильтрации). Заметьте, что по оси x на этот раз отложена, опять же, не сама консервативность, а центрированный на 0 логарифм отношения шансов для консервативности. Благодаря этому преобразованию распределение становится похожим на нормальное, слева от нуля оказываются менее консервативные, чем в среднем, гены, а справа - более консервативные.

Для проверки второй гипотезы мы сравнили среднюю консервативность участка между скомпенсированными инделами с распределением консервативностей позиций в том же самом белке (Рисунок 13). На графиках видно, что средняя консервативность

участка между инделами всегда ниже (черная линия правее красной), чем средняя консервативность белка и обычно лежит в 25% самых неконсервативных позиций.

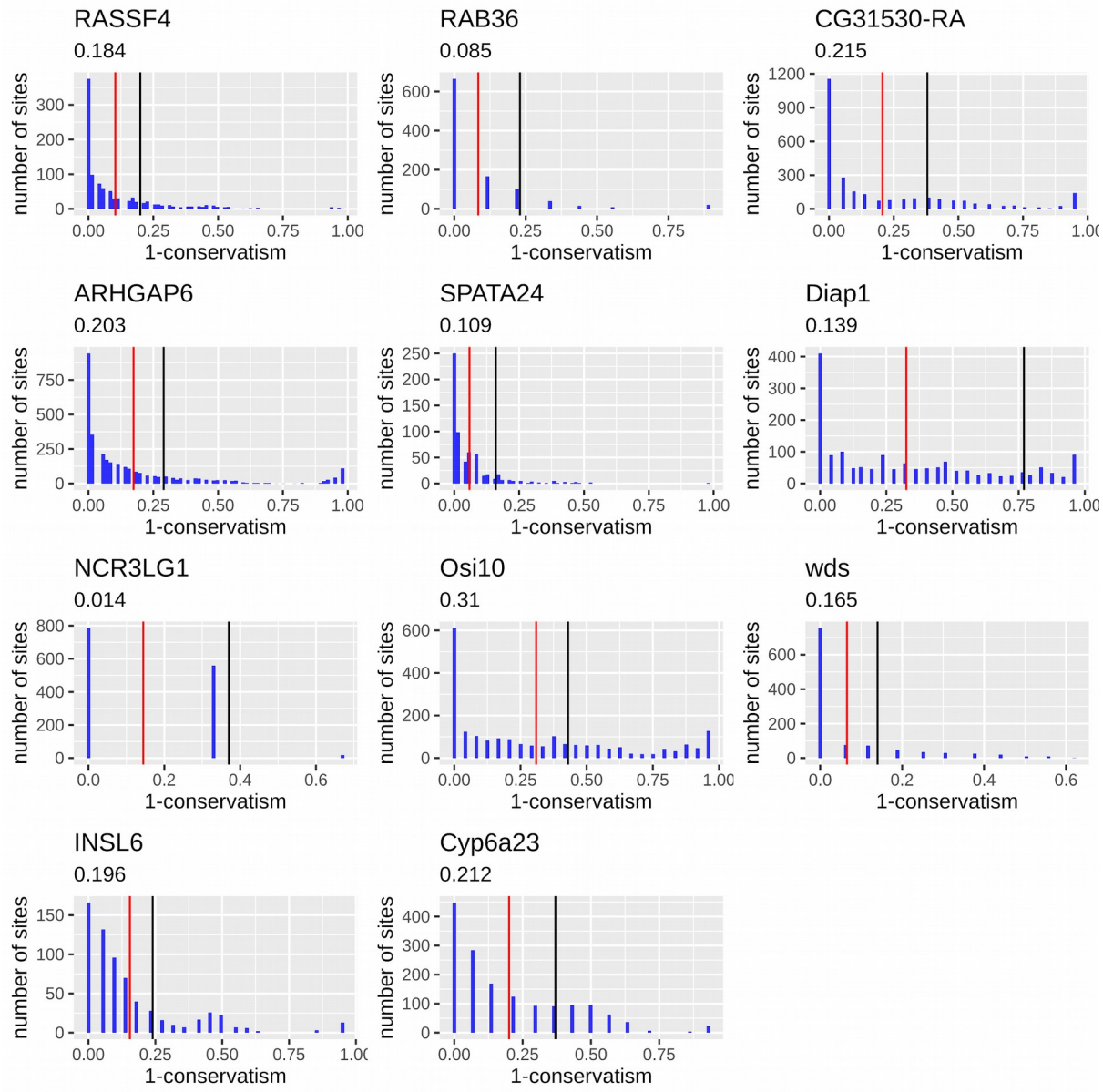

Рисунок 13. Консервативность участка между скомпенсированными инделами. Синие гистограммы — распределение консервативностей во всех позициях белка. Красная линия — средняя консервативность белка. Чёрная линия - средняя консервативность участка между скомпенсированными инделами. Число под названием гена означает долю распределения, находящуюся справа от чёрной линии Заметьте, что на самом деле по оси x отложено значение 1-консервативность, то есть самые консервативные белки позиции оказываются слева, а неконсервативные - справа.

Эффект скомпенсированного сдвига рамки считывания

Следующая гипотеза, которую мы проверили - о том, что скомпенсированные сдвиги рамки считывания фиксируются в таких местах, в которых они имеют малый эффект и вследствие этого нейтральны. Чтобы проверить эту гипотезу, мы сравнили аминокислотную последовательность между скомпенсированными инделами до того, как они произошли, и после (Рисунки 15 и 16).

Поскольку мы работаем с генами, в которых обнаружены сдвиги рамки у нескольких животных, в качестве последовательности, в которой происходили эти сдвиги, мы использовали восстановленное с помощью метода максимального правдоподобия в программе MEGAX предковое состояние. Какое именно предковое состояние использовать, зависит от некоторых предположений, поскольку мы не знаем, что произошло раньше на ветке общего предка - точечные мутации (однонуклеотидные замены) или сдвиги рамки. Если предположить,что сначала происходили мутации, можно взять реконструированного общего предка для всех животных со сдвигом рамки (зеленая линия на рисунках 15 и 16). Если же предположить,что сначала произошли инделы, а затем мутации, следует использовать предковое состояние животных со сдвигом и их ближайшего родственника (красная линия). Для лучшего понимания разницы между этими случаями обратитесь к рисунку 14.

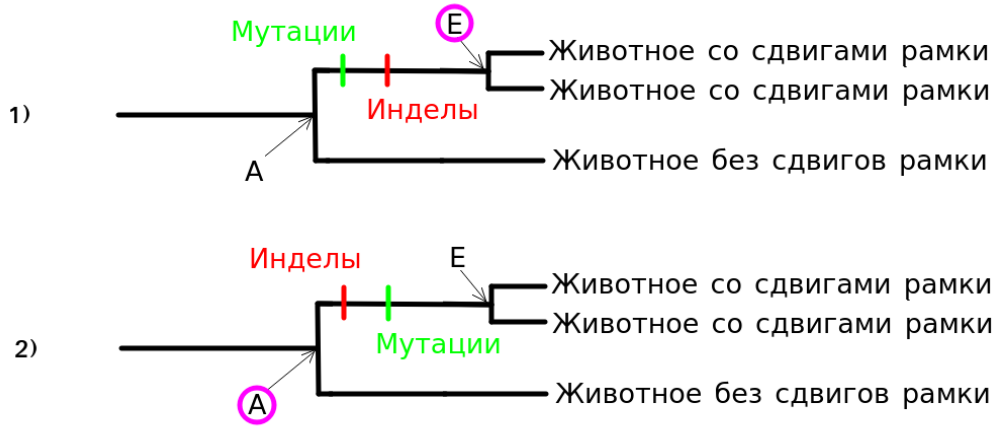

Рисунок 14. Два предположения (первыми произошли мутации (1) или инделы (2)) ведут выбору разных последовательностей в качестве тех, в которых происходили инделы (обведены сиреневым цветом).

Мы использовали две метрики близости двух последовательностей. Первая - разница в среднем индексе гидрофобности последовательности (Kyte, Doolittle, 1982). Для последовательностей с очень похожей средней гидрофобностью он близок к 0, для последовательностей с максимально непохожей - близок к 9. Вторая метрика - среднее расстояние по матрице Мияты (Miyata et al., 1979) между аминокислотами, стоящими друг напротив друга в выравнивании двух последовательностей. Если в выравнивании напротив аминокислоты стоит гэп, такой паре присваивалось максимально возможное расстояние. Эта метрика также близка к 0 для очень похожих последовательностей, но для максимально далеких она близка 5.13 (максимальное расстояние между

аминокислотами, глицином и триптофаном). Эта метрика учитывает различие в полярности и размере аминокислот.

Таким образом, если до сдвига рамки и после последовательности статистически значимо похожи друг на друга, мы считаем гипотезу подтвержденной. Для проверки значимости мы построили референсное распределение соответствующих метрик по всем нашим данным. А именно, мы вносили инделы в разные части разных белков с сохранением типов (инсерция/делеция), длин и расстояний между ними, и сравнивали аминокислотную последовательность до внесения инделов и после. Так мы получили “ожидаемое” распределение метрик.

Мы уверены, что данные согласуются с гипотезой, только если оба сравнения (в предположении первенства мутаций и в предположении первенства инделов) лежат в 5% самых наименее изменившихся после сдвига рамки последовательностей. Это так для среднего расстояния Мияты для гена Osi10 (хотя совершенно не так для индекса гидрофобности этого гена). Также в предположении того, что первыми происходили инделы, ген RAB36 по обеим метрикам и ген CG31530-RA по расстоянию Миаты попадают в эту область.

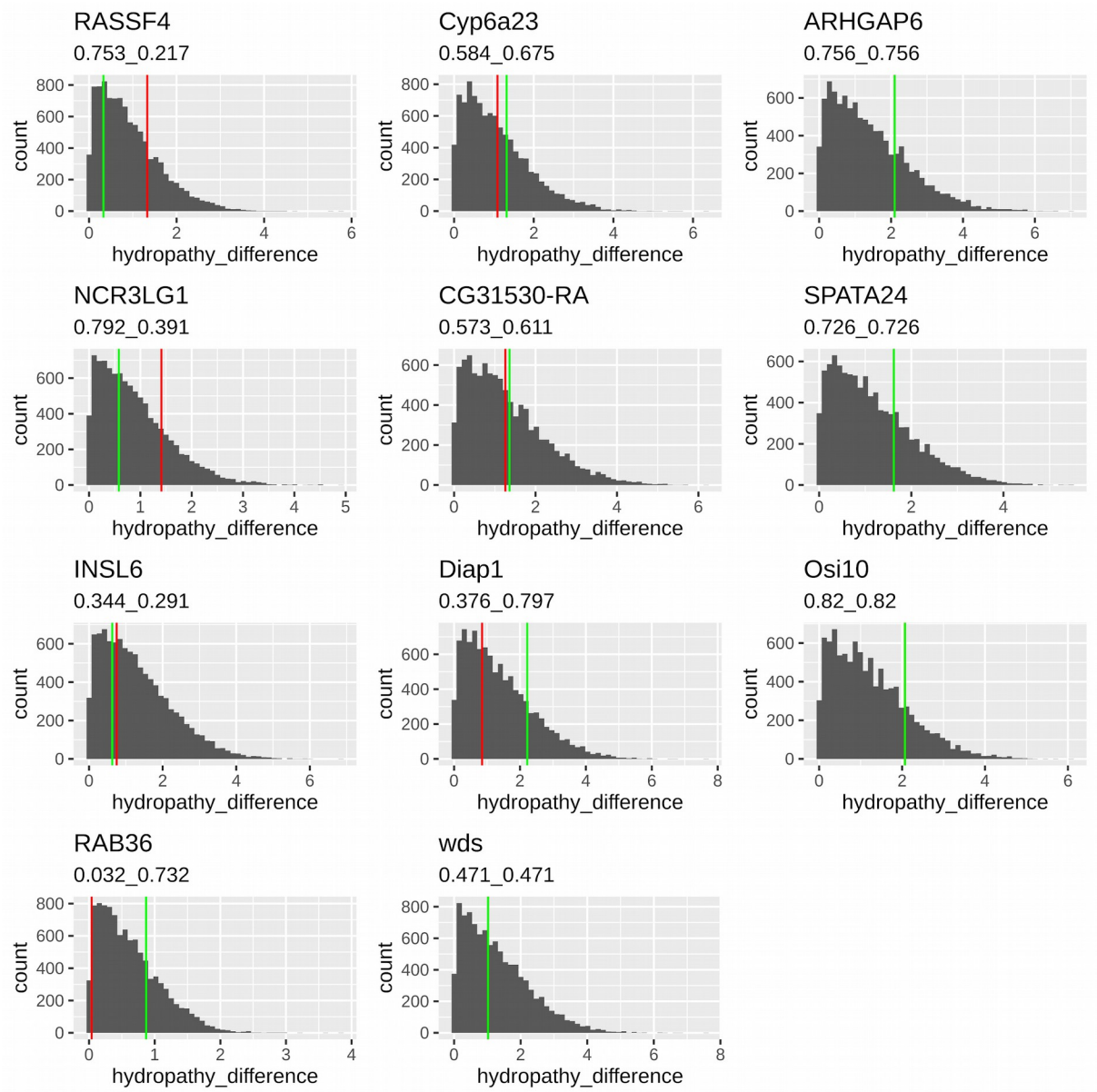

Рисунок 15. Эффект скомпенсированного сдвига рамки считывания на индекс гидрофобности последовательности. Красная линия соответствует предположению о том, что первыми происходили инделы, зеленая - мутации. Если нарисована только зеленая линия, мутаций на ветке общего предка животных с инделами не происходило. Первое число под названием гена - доля референсного распределения, лежащая слева от красной линии, второе число - доля, лежащая слева от зеленой.

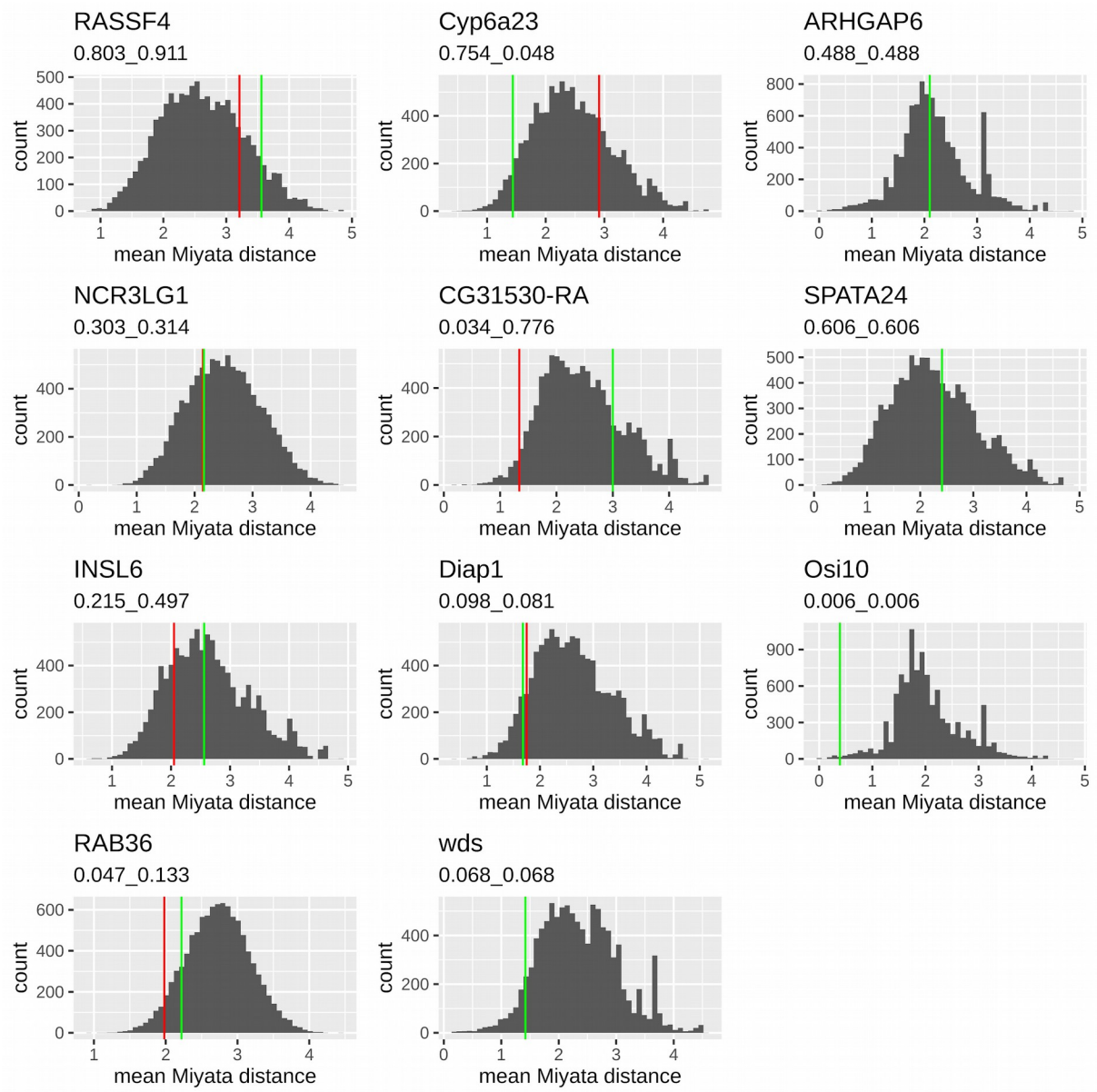

Рисунок 16. Среднее расстояние Мияты между последовательностями до сдвига и после. Обозначения такие же, как на рисунке 15.

Взаимодействие скомпенсированных инделов с однонуклеотидными заменами

О взаимодействии однонуклеотидных замен (в дальнейшем просто “мутаций”) со скомпенсированными сдвигами рамки считывания также можно составить некоторое количество гипотез. Так, если мутации на ветке общего предка произошли раньше инделов, они могли быть или не быть пермиссивными (облегчающими фиксацию инделов), если позже - быть или не быть компенсаторными (нейтрализующими эффект инделов). Мутации же на последующих ветках могли только быть либо не быть компенсаторными, поскольку мы наверняка знаем, что они произошли позже инделов.

Если взять за основу предположение о том, что предковое состояние (до инделов) оптимально, а скачок от него (который происходит при фиксации

скомпенсированных инделов) приводит к нежелательным последствиям, можно выразить эти гипотезы численно. Так, в случае мутаций на ветке общего предка, следует рассмотреть 2 ситуации.

1) Мутации произошли раньше инделов.

В этом случае можно измерить два расстояния:

- Расстояние между предковым состоянием до мутаций и инделов (A) и его версией со сдвинутой рамкой (Afr, от “A frameshifted”);
- Расстояние между предковым состоянием с мутациями но еще без инделов (Amut, от “A mutated”) и предковым состоянием с мутациями и инделами (E).

Если второе расстояние меньше, чем первое, то мутации можно считать пермиссивными - они “облегчают” фиксацию инделов, уменьшая их эффект. Если первое расстояние меньше, мутации пермиссивными не были. Можно заметить, что двум этим сравнениям как раз и соответствуют красная и зеленая линии на рисунках 15 и 16 в предыдущем разделе, но если при регистрации эффекта инделов нас интересовало положение этих линий относительно референсного распределения, сейчас нас интересует их относительное положение.

2) Мутации произошли позже инделов.

В этом случае также можно измерить два расстояния:

- Расстояние между предковым состоянием до мутаций и инделов (A) и его версией со сдвинутой рамкой (Afr); точно такое же как первое расстояние в первом случае;
- Расстояние между предковым состоянием до мутаций и инделов (A) и предковым состоянием после мутаций и инделов (E).

Если второе расстояние меньше, чем первое, мутации можно считать компенсаторными, уменьшающими эффект инделов.

Надо заметить, что поскольку никакой статистической обработки эти рассуждения не предполагают, читатель в любой момент вполне справедливо может задаться вопросом, достаточна ли наблюдаемая разница в расстояниях для того, чтобы делать какие-то выводы? Я могу лишь ответить, что никаких выводов, собственно, и не делается, а происходит описание наблюдаемых явлений. Если в одном контексте инделы имеют меньший эффект, чем в другом, и это видно по соответствующим измерениям, то утверждается всего лишь это, и ничего больше.

Результаты соответствующих измерений расстояний изображены на рисунках 17 (расстояния по индексу гидрофобности) и 18 (расстояния по Мияте).

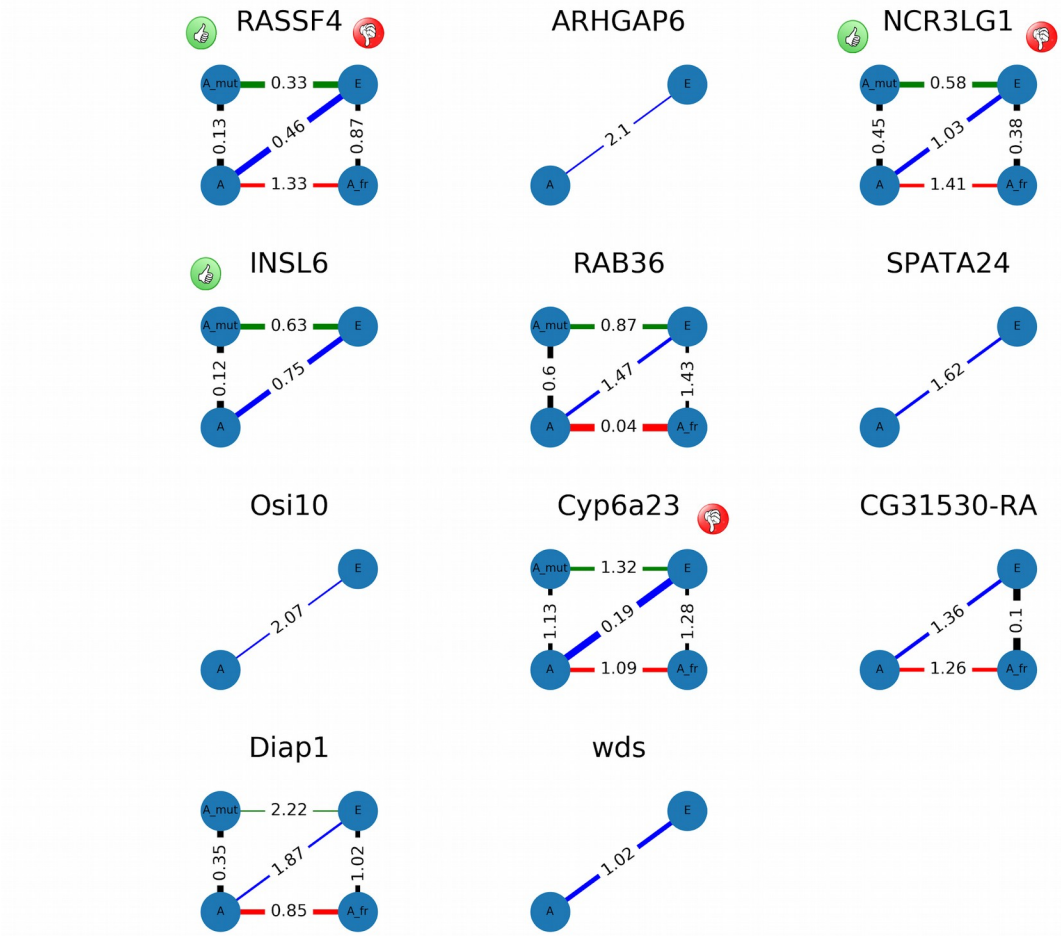

Рисунок 17. Измеренные расстояния между различными предковыми состояниями по индексу гидрофобности. Толщина линии обратно пропорциональна расстоянию (далекие точки соединены тонкой линией). Переход из узла ниже в узел выше соответствует появлению мутаций. Переход из узла левее в узел правее - появлению инделов. В предположении первенства мутаций следует сравнивать красную линию с зеленой (“красное” расстояние > “зеленое” расстояние ⇒ мутации пермиссивны, отмечено на рисунке большим пальцем вверх на зелёном фоне). В предположении первенства инделов, следует сравнивать красную линию с синей (“красное” расстояние > “зеленое” расстояние ⇒ мутации компенсаторны, отмечено на рисунке большим пальцем вниз на красном фоне). В случае отсутствия зелёной или красной линии ее следует заменить синей в первом сравнении (в отсутствии зеленой линии расстояние от Amut до E - то же самое, что от A до E, т.е. синяя линия; в отсутствии красной линии расстояние от A до Afr - это то же самое, что от A до E, т. е., опять же, синяя линия). Черные линии не участвуют в сравнениях и изображены для полноты картины. Наличие только двух вершин в графе говорит об отсутствии несинонимичных мутаций на ветке предка животных с инделами. Наличие только трёх - о синонимичности мутаций в соответствующем контексте (вследствие чего получается, что их как бы нет).

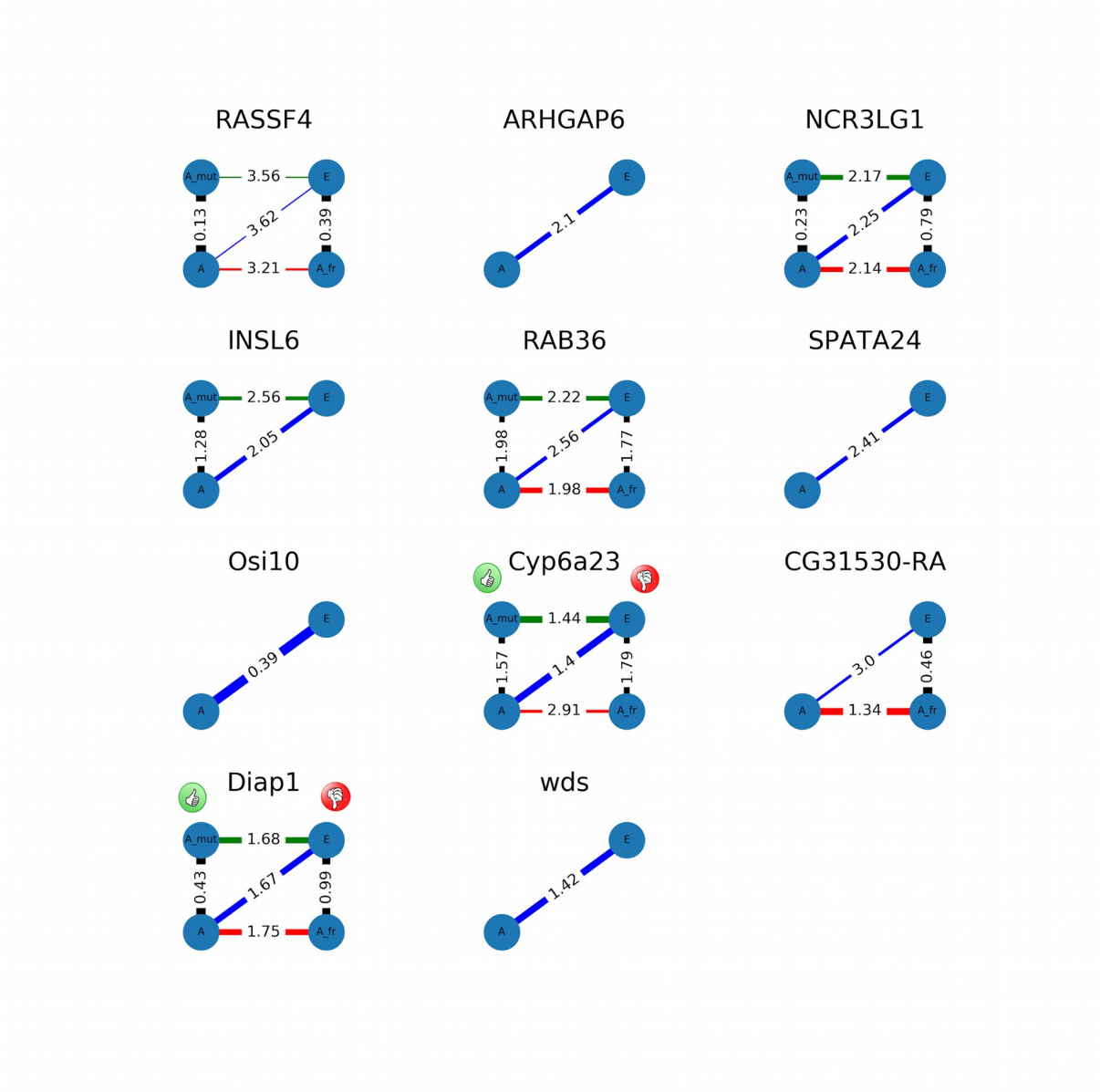

Рисунок 18. Измеренные расстояния между различными предковыми состояниями по расстоянию Мияты. Обозначения как на рисунке 17.

Результаты сравнений показывают, что в предположении первенства мутаций, они получаются пермиссивными в RASSF4, NCR3LG1 и INSL6 по индексу гидрофобности и пермиссивными в Cyp6a23 и Diap1 по расстоянию Мияты. В предположении первенства инделов, мутации получаются компенсаторными в RASSF4, NCR3LG1 и Cyp6a23 по индексу гидрофобности и в Cyp6a23 и Diap1 по расстоянию Мияты.

Для чтобы понять, были ли компенсаторными мутации, происходившие на ветках после инделов, следует сравнить другие два расстояния:

- Расстояние между предковым состоянием до мутаций и инделов (A) и предковым состоянием после мутаций и инделов (E); такое же, как второе расстояние в предположении первенства инделов, обозначено синей линией на рисунках 19 и 20.

- Расстояние между предковым состоянием до мутаций и инделов (A) и состоянием на конце ветки, с инделами и всеми мутациями.
  Если второе расстояние меньше,чем первое, мутации, происходившие на ветках после инделов, были компенсаторными.

Такое сравнение можно проделать для каждого животного со скомпенсированными инделами, что мы и сделали (Рисунки 19 и 20 для индекса гидрофобности и расстояния Мияты соответственно).

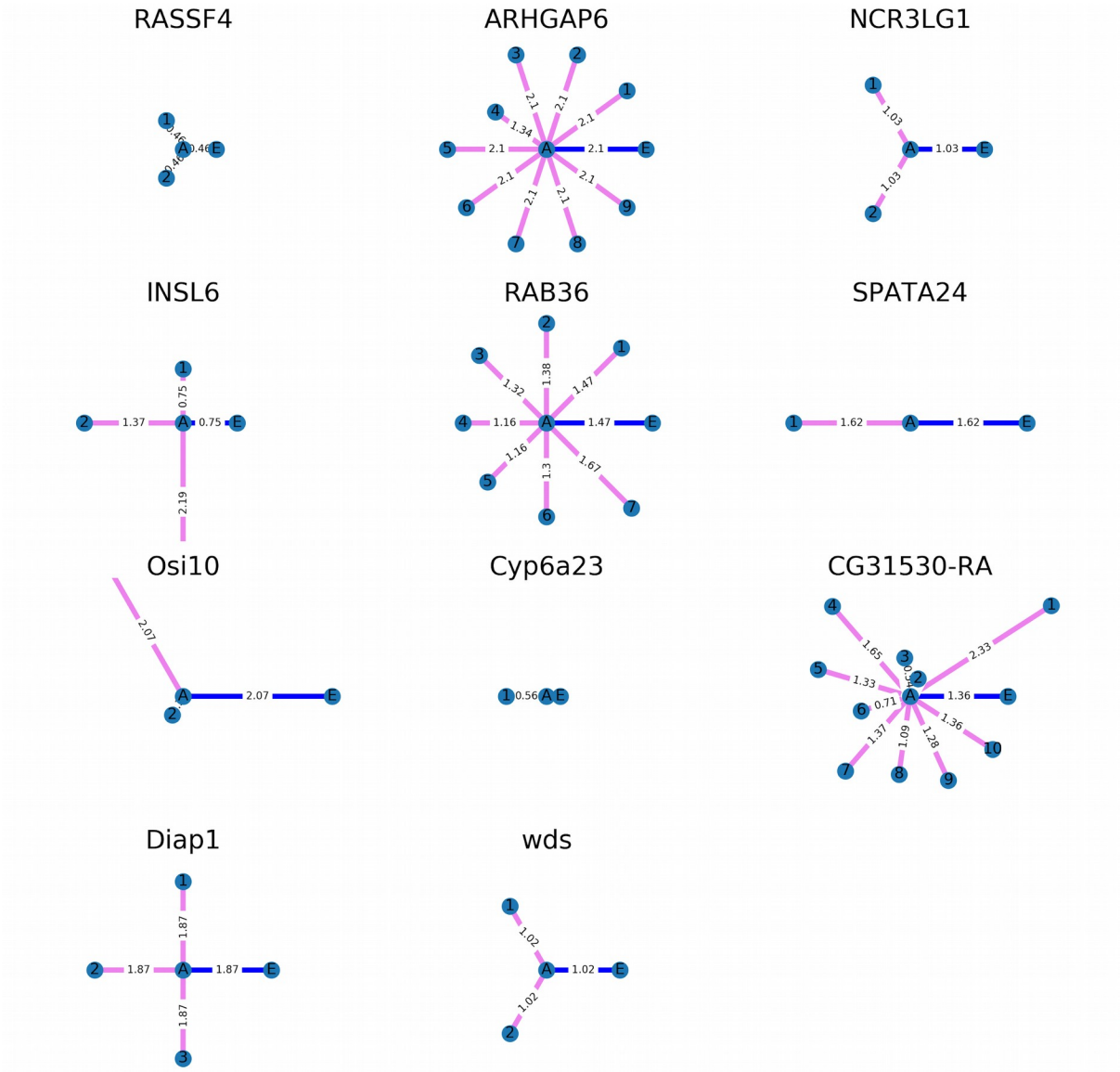

Рисунок 19. Сравнение расстояний по индексу гидрофобности для определения эффектов мутаций на ветках после инделов. Синие линии соответствуют синим линиям на предыдущих картинках - расстоянию между A и E. Сиреневые линии соответствуют расстояниям между предковым расстоянием A и нынешними состояниями последовательностей у различных животных. Длины линий пропорциональны расстояниям. Мутации являются компенсаторными, если сиреневые линии короче синей.

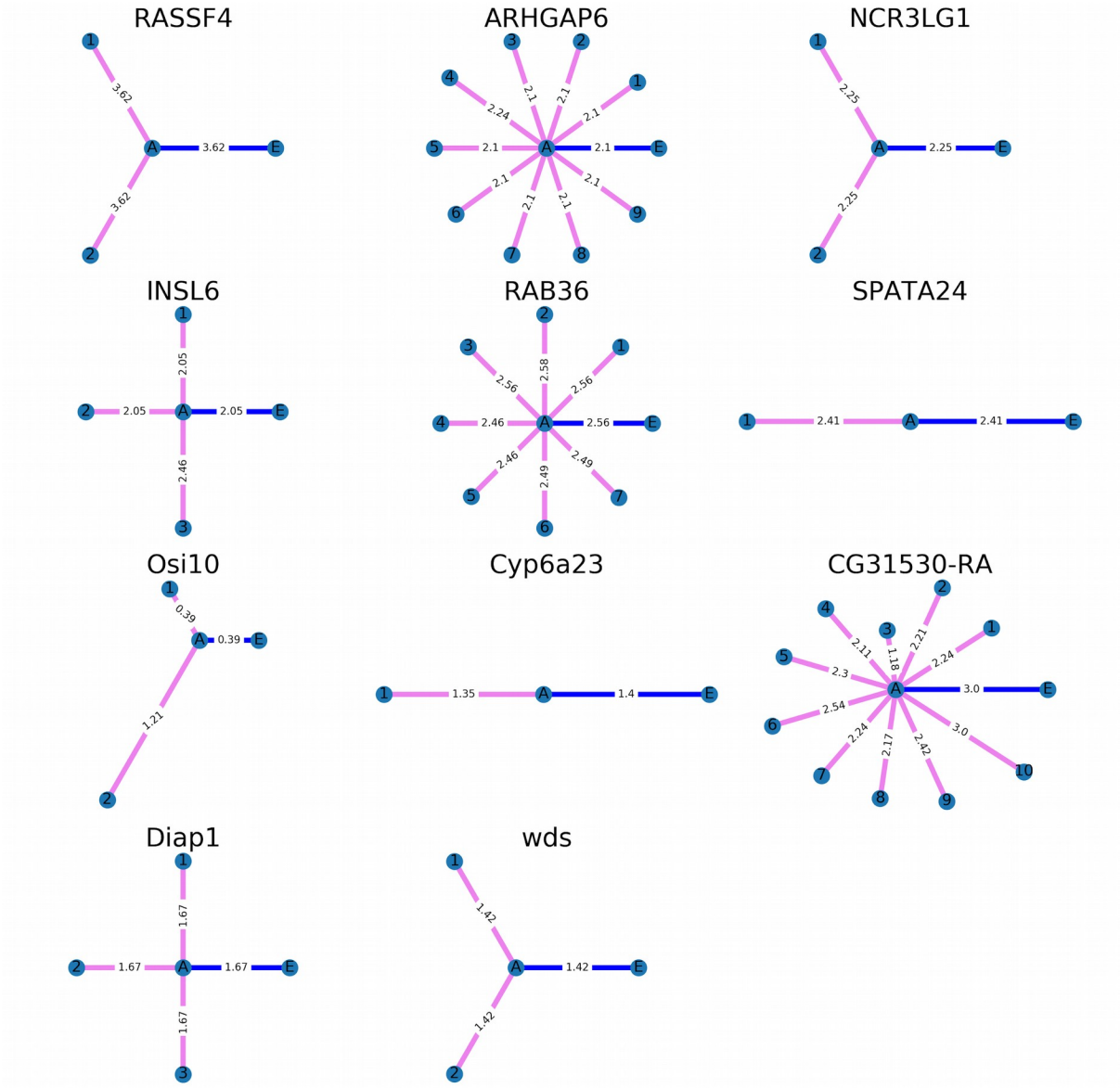

Рисунок 20. Сравнение расстояний по среднему расстоянию Мияты для определения эффектов мутаций на ветках после инделов.

Эти сравнения показывают, что по индексу гидрофобности мутации получаются компенсаторными в некоторых животных по генам ARHGAP6, RAB36,CG31530-RA и Osi10. По среднему расстоянию Мияты мутации получаются компенсаторными в некоторых животных по генам RAB36, CG31530-RA и Cyp6a23.

По индексу гидрофобности всего на 13 ветках мутации уменьшили расстояние до предкового состояния (на рисунке 19 сиреневая линия короче синей), и 7 увеличили (сиреневая линия длиннее синей). По расстоянию Мияты - 14 уменьшили и 4 увеличили (то же на рисунке 20).

Обсуждение

Оценка ожидаемого числа инделов

Пять скомпенсированных сдвигов рамки у насекомых и шесть у позвоночных выглядят очень маленькими числами. С другой стороны, сколько мы ожидаем их увидеть в таких данных? Может, примерно столько и ожидаем?

Прикидочную оценку можно получить следующим образом. Пусть L - средняя длина гена в наших данных; p - вероятность произойти однонуклеотидной или двухнуклеотидной инсерции или делеции на нуклеотид на поколение (примем для простоты одинаковой вероятность вне зависимости от длины); s - коэффициент отбора против одиночной мутации сдвига рамки считывания. Тогда в популяции будет находиться нескомпенсированный сдвиг рамки считывания на частоте m1/s, где m1 - вероятность происхождения индела в расчете на ген (Crow, Kimura, 1970). m1 несложно посчитать из p:

$m1 = ¿,$

где ¿ - вероятность того, что конкретный индел не произойдет в данном нуклеотиде, ¿- вероятность того, что он не произойдет в гене длиной L, $1 - ¿$- вероятность того, что произойдет (строго говоря, хотя бы один раз, но, думаю, вероятностью множественного происхождения инделов можно пренебречь). Умножение на 4 нужно, чтобы учесть все комбинации типов (инсерция, делеция) и длин (1, 2) инделов.

Итак, m1/s - частота, на которой держится первый индел в популяции. Если рассматривать одно поколение, ожидаемая частота генов, содержащих два индела, вследствие того, что второй индел произошёл в гене, где уже был первый, равна $m2 * m1/s$, где

$m2 = ¿$

Поскольку для каждого из 4 вариантов инделов есть только 2 компенсирующих (инсерцию длины 1 можно компенсировать только инсерцией длины 2 или делецией длины 1, но не инсерцией длины 1 или делецией длины 2), m2 отличается от m1 тем, что множитель при p равен 2, а не 4. Также p возводится в сотую степень, а не в степень L, поскольку далее мы хотим использовать предположение нейтральности, для которого звучит разумно, чтобы второй индел происходил неподалеку от первого (скажем, в пределах 50 нуклеотидов). Далее, если предположить, что генотипы со скомпенсированными инделами нейтральны, вероятность фиксации такого мутанта будет равна его частоте, то есть тоже $m1\,m2/s$. Итак, в озвученных предположениях должно фиксироваться примерно по одному скомпенсированному инделу на ген каждые $s/m1\,m2$ поколений. Осталось оценить число поколений, охваченных в нашем анализе, и ожидаемое число скомпенсированных инделов составит $N/(s/m1\,m2) = Nm1\,m2/s$ на ген, где N - количество поколений, охваченных нашим анализом.

Для оценки этих параметров возьмем $p = 10^{-10}$ (Schrider et al,. 2013; Uchimura et al., 2015), $s = 0.0015$ (Langley et al., 1981), $N = 176\,536\,337$ и $L = 1897$для позвоночных и $N = 403\,272\,889$, $L = 2403$ для двукрылых. s взято как среднее значение коэффициента

отбора для мутации потери функции в генах, потеря которых не летальна; N оценено как средняя длина деревьев, построенных для каждого гена только по животным, прошедшим фильтры (и без учета терминальных ветвей, поскольку если индел произошел на терминальной ветви, мы отбросим его как ненадежный). Рассчитанное значение $Nm1m2/s=0.004$ для позвоночных и 0.005 для двукрылых, другими словами, мы ожидаем увидеть 4 или 5 скомпенсированных индела на 1000 генов. Поскольку в позвоночных мы рассмотрели 21208 генов, а в двукрылых - 27341, ожидаемое число скомпенсированных инделов, в них составляет 87±18 и 141±23 соответственно (95% доверительный интервал, посчитан в приближении нормальным распределением биномиального распределения с параметрами n=21208, p=0.004 и n=27341, p=0.005), что как видно, намного больше, чем найденные 6.

Столь большое отклонение можно объяснить, например, тем, что скомпенсированные сдвиги рамки считывания, по-видимому, в среднем вовсе не являются нейтральными, как мы предположили при расчётах. Возможно, отрицательный отбор против них столь силён, что фиксируется значительно меньшее их число, чем ожидается в нейтральном случае. Также следует отметить, что рассчитанное значение является ожидаемым числом пар инделов, действительно присутствующих в данных, а не числом, которое мы ожидаем найти. Последнее число рассчитать довольно сложно, поскольку не понятно, как часто мы встречаемся с ошибками сборки, выравнивания, аннотации (напомню, что гены большинстве животных аннотированы по похожести на соответствующие гены референсного животного) и прочими, препятствующими детекции скомпенсированных сдвигов рамки считывания.

Мы также хотим отметить, что в статье коллег (Hu, Ng, 2012) упоминалось, что они наблюдали среди прочего случаи скомпенсированных сдвигов рамки считывания и, конечно, мы ожидали увидеть те же самые случаи. В качестве примера Hu и Ng показывают выравнивание белка FLJ43860 на рисунке, который также вставлен в обзор литературы данной работы. Мы захотели проверить, есть ли этот белок среди найденных нами, и когда обнаружили, что его там нет, открыли соответствующий файл выравнивания. Оказалось, что этот белок не имеет ничего общего с представленным на рисунке, причём последовательность с рисунка не находится в NCBI BLAST, в то время как наша последовательность, согласно тому же источнику, соответствует белку FLJ43860. Похоже, это указывает на какую-то ошибку в данных Hu и Ng, и говорит о том, что они, возможно, не видели скомпенсированных сдвигов рамки считывания. Мы тем не менее, благодарны этим авторам, поскольку их статья была одной из тех, которые воодушевили нас заняться этой проблемой.

Отбор и консервативность

Анализ действия отбора на гены со скомпенсированными инделами показал, что отрицательный отбор на большую их часть, как и ожидалось, ослаблен, что, видимо, облегчает фиксацию инделов (уменьшает s в формуле $m1m2/s$). Также один ген (CG31530-RA), похоже, претерпел изменение в режиме отбора (а именно, ослабление отрицательного отбора) примерно в момент фиксации инделов. Его версия с инделами

есть у *Drosophila melanogaster* и еще 10 видов дрозофил. Согласно UniProt (UniProt Consortium, 2008)**;** https://www.uniprot.org/uniprot/D2NUL5), это трансмембранный белок, вовлеченный в трансмембранный транспорт, и никаких подробностей его функций, тем более особенностей этих функций в *Drosophila melanogaster* не известно.

В то же время, анализ консервативности позиций белков показывает, что инделы происходят в наименее консервативных участках, что опять же согласуется с ожиданиями. Высокая консервативность означает пристальное внимание отбора к данному участку, а низкая консервативность, наоборот, говорит об ослабленном отрицательном отборе, и, соответственно, инделы чаще фиксируются именно в таких участках.

Результаты по сравнению консервативности белков, в которых произошли сдвиги рамки, с консервативностью остальных, показывают, что белки со скомпенсированными инделами чаще бывают менее консервативными, чем в среднем, хотя эффект не очень выраженный. Это также согласуется с ожиданиями, поскольку консервативность - во многом отражение действия отрицательного отбора на последовательность, и соответственно, последовательности под отбором будут более консервативны и мутации сдвига рамки будут с меньшей вероятностью фиксироваться в них.

Эффект инделов

Ещё один результат, который мы получили, состоит в том, что скомпенсированные сдвиги рамки считывания происходят не обязательно там, где их эффект минимален. Можно было бы думать, что они не имеют особенного эффекта и фиксируются за счет дрейфа, но, поскольку видно, что эффект они в большинстве своём имеют, остается предположить,что либо регионы, в которых они происходят, не важны (что подтверждается анализом консервативности), либо белки, в которых они происходят, не важны (что не подтверждается анализом консервативности), либо же, наконец, они фиксируются за счет отбора, будучи полезными. В последнем случае, конечно, было бы интересно понять, какое преимущество скомпенсированный сдвиг рамки даёт своему хозяину, но для этого нужны более подробные данные о различиях в функционировании белков в разных позвоночных и двукрылых.

Можно заметить, что две использованные нами метрики дают совершенно разные результаты, и если эффект от сдвига рамки получается малым по одной из этих метрик, он вовсе не обязательно будет малым по другой. Между тем, обе должны быть чувствительны к изменению гидрофобности - это с очевидностью верно для индекса гидрофобности, и также верно для расстояния Мияты, поскольку при его расчёте используются данные о полярности аминокислот, очевидно, скоррелированные с гидрофобностью. По-видимому, в тех случаях, когда расстояние Мияты увеличивается, а по индексу гидрофобности - нет, вклад в расстояние даёт именно размер аминокислот, а в обратных ситуациях, вклад от гидрофобности и от размера может быть обратным (расстояние по гидрофобности увеличивается, а по размеру - уменьшается).

Эффект мутаций

Однонуклеотидные замены, происходящие на ветке со скомпенсированными инделами, могут быть как-то с ними связаны. Замены, происходящие до скомпенсированных инделов, могут быть пермиссивными - то есть такими, в контексте которых инделы будут иметь меньший эффект. Замены, происходящие после них, могут быть компенсаторными - тоже уменьшающими эффект от инделов. Мы действительно обнаружили ряд таких замен, но, как уже было замечено, не можем делать никаких утверждений о том, почему они произошли. Они могли произойти под действием положительного отбора, будучи полезными именно потому, что “чинят” последовательность, частично нарушенную после пусть скомпенсированных, но инделов. С другой стороны, фиксирующаяся мутация будет неизбежно либо увеличивать, либо уменьшать эффект от инделов, в том числе, если она фиксируется по независящим от инделов причинам (или даже под действием дрейфа). Поскольку нет повода ожидать, что замены будут непременно увеличивать эффект инделов, нет ничего удивительного в том, что мы обнаружили такие, которые его уменьшают. Во всяком случае, если взять наивную нулевую гипотезу о том, что мутации с одинаковой вероятностью могут как уменьшать расстояние до предкового состояния, так и увеличивать его, наблюдаемые различия в числе случаев, когда расстояние уменьшается и случаев, когда оно увеличивается, не позволяет отклонить эту нулевую гипотезу (13 против 8, критерий знаков, p-value = 0.38) для индекса гидрофобности, и формально говоря, позволяет для расстояния Мияты (14 против 4, p=0.03), но вряд ли этому можно доверять, учитывая зависимость большинства случаев, где расстояние уменьшилось (разные животные в одной кладе могут иметь одни и те же мутации).

Выводы

1. Мы разработали алгоритм поиска скомпенсированных сдвигов рамки считывания по данным множественного выравнивания белок-кодирующих генов и филогении. Используя его мы нашли 11 белок-кодирующих генов со скомпенсированными инделами в данных UCSC по 100 позвоночным и 122 двукрылым.
2. Скомпенсированные сдвиги рамки считывания в большинстве своём происходят в генах с ослабленным отрицательным отбором, а также в участках низкой консервативности. В одном случае из 11 удалось зарегистрировать ослабление отрицательного отбора примерно одновременно с фиксацией скомпенсированных инделов.
3. Скомпенсированные сдвиги рамки считывания происходят не только и не столько в местах, где их эффект минимален.
4. Некоторые мутации, происходящие на ветках со скомпенсированными инделами, имеют пермиссивный или компенсаторный характер, но некоторые лишь увеличивают эффект от сдвига рамки; систематического паттерна в относительном эффекте сдвига рамки и мутаций не обнаружено.
5. Различия в количестве найденных скомпенсированных сдвигах рамки по сравнению с ожидаемым наталкивает на предположение о в среднем большом коэффициенте отбора против даже скомпенсированных инделов и, следовательно, их низкой роли в образовании новых аминокислотных последовательностей белок-кодирующих генов.

Благодарности

Мы благодарим Алексея Кондрашова за помощь в выводе формулы для оценки ожидаемого числа скомпенсированных инделов, лабораторию эволюционной геномики (http://evolgenomics.fbb.msu.ru/) за предоставление ресурсов на кластере, и её сотрудников за помощь в обсуждении проекта и подборе литературных материалов.

Литература

Akiva P., Toporik A., Edelheit S., Peretz Y., Diber A., Shemesh R., Novik A., Sorek R. Transcription-mediated gene fusion in the human genome. // Genome Res. 2006. V. 16(1): P. 30–36.

Anderson R.P., Roth J.R. Tandem genetic duplications in phage and bacteria. // Annual review of microbiology 1977. V. 31: P. 473–505.

Averof M., Dawes R., Ferrier D. Diversification of arthropod Hox genes as a paradigm for the evolution of gene functions. // Seminars in Cell & Developmental Biology 1996. V. 7(4): P. 539–551.

Bartonek L., Braun D., Zagrovic B. Frameshifting preserves key physicochemical properties of proteins. // Proc. Natl. Acad. Sci. USA 2020. V. 117: P. 5907–5912.

Belshaw R., Pybus O.G., Rambaut A. The evolution of genome compression and genomic novelty in RNA viruses. // 2007. V. Genome research 17(10): P. 1496–1504.

Birchler J.A., Riddle N.C., Auger D.L., Veitia R.A. Dosage balance in gene regulation: biological implications. // Trends in genetics 2005. V. 21(4): P. 219–226.

Cardoso-Moreira M., Arguello J.R., Gottipati S., Harshman L.G., Grenier J.K., Clark A.G. Evidence for the fixation of gene duplications by positive selection in Drosophila. // Genome research 2016. V. 26(6): P. 787–798.

Crow J.F., Kimura M. An introduction to population genetics theory. // An introduction to population genetics theory. P. 1970.

Domazet-Loso T., Tautz D. An evolutionary analysis of orphan genes in Drosophila. // Genome research 2003. V. 13(10): P. 2213–2219.

Eyre-Walker A., Keightley P.D. The distribution of fitness effects of new mutations. // Nature reviews. Genetics 2007. V. 8(8): P. 610–618.

Force A., Lynch M., Pickett F.B., Amores A., Yan Y.-l., Postlethwait J. Preservation of Duplicate Genes by Complementary, Degenerative Mutations. // Genetics 1999. V. 151(4): P. 1531–1545.

Gibbs R.A., Weinstock G.M., Metzker M.L., Muzny D.M., Sodergren E.J., Scherer S., Scott G., Steffen D., Worley K.C., Burch P.E., Okwuonu G., Hines S., Lewis L., DeRamo C., Delgado O., Dugan-Rocha S., Miner G., Morgan M., Hawes A., Gill R., Celera, Holt R.A., Adams M.D., Amanatides P.G., Baden-Tillson H., Barnstead M., Chin, S. et al., Genome sequence of the Brown Norway rat yields insights into mammalian evolution. // Nature 2004. V. 428(6982): P. 493–521.

Hahn Y., Lee B.. Identification of nine human-specific frameshift mutations by comparative analysis of the human and the chimpanzee genome sequences. // Bioinformatics 2005. V. 21 Suppl 1: P. 186-94.

Haldane J.B.S. The Part Played by Recurrent Mutation in Evolution. // The American Naturalist 1933. V. 67(708): P. 5–19.

Hancock J.M., Simon, M. Simple sequence repeats in proteins and their significance for network evolution. // Gene 2005. V. 345(1): P. 113–118.

Hu J., Ng P.C. Predicting the effects of frameshifting indels. // Genome biology 2012. V. 13(2): R9.

Hughes A. L. The evolution of functionally novel proteins after gene duplication. // Proceedings of the Royal Society of London 1994: P. 119–124.

Innan H., Kondrashov F. The evolution of gene duplications: classifying and distinguishing between models. // Nature reviews. Genetics 2010. V. 11(2): P. 97–108.

Jacob F. Evolution and tinkering. // Science 1977. V. 196(4295): P. 1161–1166.

Kaessmann H. Origins, evolution, and phenotypic impact of new genes. // Genome research 2010. V. 20(10): P. 1313–1326.

Kaessmann H., Zöllner S., Nekrutenko A., Li W.-H. Signatures of domain shuffling in the human genome. // Genome research 2002. V. 12(11): P. 1642–1650.

Katoh K., Misawa K., Kuma K., Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. // Nucleic Acids Res. 2002. V. 30: P. 3059–3066.

Keese P.K., Gibbs A. Origins of genes: “big bang” or continuous creation? // Proceedings of the National Academy of Sciences of the United States of America 1992. V. 89(20): P. 9489–9493.

Kovacs E., Tompa P., Liliom K., Kalmar L. Dual coding in alternative reading frames correlates with intrinsic protein disorder. // Proceedings of the National Academy of Sciences of the United States of America 2010. V. 107(12): P. 5429–5434.

Kyte J., Doolittle R.F.. A simple method for displaying the hydropathic character of a protein. // Journal of molecular biology 1982. V. 157: P. 105–132.

Langley C.H., Voelker R.A., Brown A.J., Ohnishi S., Dickson B., Montgomery E. Null allele frequencies at allozyme loci in natural populations of Drosophila melanogaster. // Genetics 1981. V. 99: P. 151–156.

Leushkin E.V., Bazykin G.A., Kondrashov A.S. Strong mutational bias toward deletions in the Drosophila melanogaster genome is compensated by selection. // Genome Biol. Evol. 2013. V. 5: P. 514–524.

Li C.-Y., Zhang Y., Wang Z., Zhang Y., Cao C., Zhang P.-W., Lu S.-J., Li X.-M., Yu Q., Zheng X. et al., A human-specific de novo protein-coding gene associated with human brain functions. // PLoS computational biology 2010. V. 6(3): e1000734.

Long M., Betrán E., Thornton K., Wang, W. The origin of new genes: glimpses from the young and old. // Nature reviews 2003. V. Genetics 4(11): P. 865–875.

Long M., Langley C. Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. // Science 1993. V. 260(5104): P. 91–95.

Long Q., Rabanal F.A., Meng D., Huber C.D., Farlow A., Platzer A., Zhang Q., Vilhjálmsson B.J., Korte A., Nizhynska V. et al., Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. // Nature genetics 2013. V. 45(8): P. 884–890.

Lynch M., Katju V. The altered evolutionary trajectories of gene duplicates. // Trends in genetics 2004. V. 20(11): P. 544–549.

Lynch M.The origins of genome architecture. // Sinauer Associates, Sunderland, MA. 2007.

Marcotte E.M., Pellegrini M., Yeates T.O., Eisenberg D. A census of protein repeats. // Journal of molecular biology 1999. V. 293(1): P. 151–160.

Miyata T., Miyazawa S., Yasunaga T. Two types of amino acid substitutions in protein evolution. // Journal of molecular evolution. 1979. V. 12: P. 219–236.

Moran J. V. Exon Shuffling by L1 Retrotransposition. // Science 1999. V. 283(5407): P. 1530–1534.

Muller H.J. The origination of chromatin deficiencies as minute deletions subject to insertion elsewhere. // Genetica 1935. V. 17(3-4): P. 237–252.

Neme R., Tautz D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. // BMC genomics 2013. V. 14: P. 117.

Ohno S. Evolution by Gene Duplication. // Springer-Verlag 1970. V., New York: 160 p.

Ohno S. Birth of a unique enzyme from an alternative reading frame of the preexisted, internally repetitious coding sequence. // Proc Natl. Acad. Sci. USA 1984. V. 81: P. 2421–2425.

Pradet-Balade B. An endogenous hybrid mRNA encodes TWE-PRIL, a functional cell surface TWEAK-APRIL fusion protein. // The EMBO Journal 2002. V. 21(21): P. 5711–5720.

Raes J., van de Peer Y. Functional divergence of proteins through frameshift mutations. // Trends in genetics 2005. V. 21(8): P. 428–431.

Rockah-Shmuel L., Tóth-Petróczy Á., Sela A., Wurtzel O., Sorek R., Tawfik D.S. Correlated occurrence and bypass of frame-shifting insertion-deletions (InDels) to give functional proteins. // PLoS Genet. 2013. V. 9: e1003882.

Romero D., Palacios R. Gene amplification and genomic plasticity in prokaryotes. // Annual review of genetics 1997. V. 31: P. 91–111.

Rosenbloom K.R., Armstrong J., Barber G.P., Casper J., Clawson H., Diekhans M., Dreszer T.R., Fujita P.A., Guruvadoo L., Haeussler M., et al. The UCSC Genome Browser database: 2015 update. // Nucleic Acids Res. 2015. V. 43: P. 670-81.

Schlötterer C. Evolutionary dynamics of microsatellite DNA. // Chromosoma 2000. V. 109(6): P. 365–371.

Schmitz J.F., Bornberg-Bauer E. Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. // F1000Research 2017. V. 6: P. 57.

Schrider D.R., Houle D., Lynch M., Hahn M.W. Rates and genomic consequences of spontaneous mutational events in Drosophila melanogaster. // Genetics 2013. V. 194: P. 937–954.

Souza S.J. de, Long M., Schoenbach L., Roy S.W., Gilbert W. Intron positions correlate with module boundaries in ancient proteins. // Proceedings of the National Academy of Sciences 1996. V. 93(25): P. 14632–14636.

Spofford J.B. Heterosis and the Evolution of Duplications. // The American Naturalist 1969. V. 103(932): P. 407–432.

Szklarczyk R., Heringa J., Pond,S.K., Nekrutenko A. Rapid asymmetric evolution of a dual-coding tumor suppressor INK4a/ARF locus contradicts its function. // Proceedings of the National Academy of Sciences of the United States of America 2007. V. 104(31): P. 12807–12812.

Thomson T.M. Fusion of the Human Gene for the Polyubiquitination Coeffector UEV1 with Kua, a Newly Identified Gene. // Genome Res. 2000. V. 10(11): P. 1743–1756.

Uchimura A., Higuchi M., Minakuchi Y., Ohno M., Toyoda A., Fujiyama A., Miura I., Wakana S., Nishino J., Yagi T. Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. // Genome Res. 2015. V. 25: P. 1125–1134.

UniProt Consortium. The universal protein resource (UniProt). // Nucleic Acids Res. 2008. V. 36: P. 190-5.

Vandenbussche M,, Theissen G,, Van de Peer Y,, Gerats T. Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. // Nucleic Acids Res. 2003. V. 31: P. 4401–4409.

Wang X., Dong Q., Chen G., Zhang J., Liu Y., Zhao J, Peng H., Wang Y., Cai Y., Wang X., et al. Why are frameshift homologs widespread within and across species? // 2016. BioRxiv.

Wang W., Zheng H., Fan C., Li J., Shi J., Cai Z., Zhang G., Liu D., Zhang J., Vang S. et al., High rate of chimeric gene origination by retroposition in plant genomes. // The Plant cell 2006. V. 18(8): P. 1791–1802.

Yu W.-P., Brenner S., Venkatesh B. Duplication, degeneration and subfunctionalization of the nested synapsin–Timp genes in Fugu. // Trends in Genetics 2003. V. 19(4): P. 180–183.

Zhang J. Evolution by gene duplication: an update. // Trends in Ecology & Evolution 2003. V. 18(6): P. 292–298.

Zhang J., Zhang Y.-p., Rosenberg H.F. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. // Nature genetics 2002. V. 30(4): P. 411–415.

Zhou Q., Zhang G., Zhang Y., Xu S., Zhao R., Zhan Z., Li X., Ding Y., Yang S., Wang W. On the origin of new genes in Drosophila. // Genome research 2008. V. 18(9): P. 1446–1455.

Приложение

Списки животных, использованных в выравниваниях с расшифровками их кодов можно найти в геномном браузере UCSC
https://genome.ucsc.edu/cgi-bin/hgc?hgsid=797731595_9NYu3l7qnNix8QWbdCSMUjNACl7a&c=chr21&l=33024806&r=33045960&o=33024806&t=33045960&g=phyloP100wayAll&i=phyloP100wayAll - для позвоночных
https://genome.ucsc.edu/cgi-bin/hgTrackUi?db=dm6&g=cons124way - для двукрылых