*Математическая биология и биоинформатика*
*2017. Т. 12. № 2. С. 302–316. doi: 10.17537/2017.12.302*

========== **БИОИНФОРМАТИКА** ==========

УДК: 577.21

# Слайдинг и вариабельность длины интронов для генов, обогащенных длинными интронами фазы 1

**Поверенная И.В.*[1,2], Горев Д.Д.[3], Астахова Т.В.**[4], Цитович И.И.[3,6], Яковлев В.В.[4,5], Ройтберг М.А.[3,4,5]**

[1]*Факультет биоинженерии и биоинформатики Московский государственный университет имени М.В. Ломоносова, Москва*
[2]*Институт общей генетики им. Н.И. Вавилова РАН, Москва*
[3]*Московский физико-технический институт, Москва*
[4]*Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша РАН, Пущино*
[5]*Национальный исследовательский университет «Высшая школа экономики», Москва*
[6]*Институт проблем передачи информации им. А.А. Харкевича РАН, Москва*

***Аннотация.*** В связи с высоким уровнем мутагенеза в последовательности интрона, эволюцию интронов обычно рассматривают только в рамках эволюции экзон-интронной структуры гена в целом. Смещение границ интрона на небольшое расстояние (редкое эволюционное событие, называемое слайдингом) может привести к смене фазы интрона. Проведен анализ экзон-интронной структуры эукариотических генов в целях выявить предпочтительный выбор фазы интрона во время слайдинга и изучить зависимость между длинами ортологичных интронов. Для определения ортологичных интронов мы построили соответствующие выравнивания экзон-интронных структур. Анализ полученных выравниваний выявил несколько событий слайдинга со смещением фазы, однако результаты не подтвердили нашу первоначальную гипотезу, что в процессе слайдинга интроны предпочитают менять свою фазу на наиболее часто встречающуюся фазу 0. Для окончательных выводов, тем не менее, необходимо провести поиск случаев изменения фазы на большей выборке генов. Несмотря на общеизвестную высокую вариабельность длины интрона, в некоторых таксономических группах значения длины схожи. Более того, можно увидеть некоторую консервативность, если вместо длины интрона $L$ рассматривать нормализованную длину $N = (L–A)/A$, где $A$ – это средняя длина по группе ортологичных интронов. Например, для гена *ptprd* в случае птиц (28 видов) нормализованная длина находится в интервале (–0.15; 0.15) для 85.2 % интронов, что существенно выше значения для случайной выборки длин в соответствии с распределением длин интронов. Такая консервативность длины ведет нас к вопросу о длине древних интронов.

***Ключевые слова:*** *интрон, фаза, выравнивание, слайдинг, консервативность.*

## ВВЕДЕНИЕ

Экзон-интронная структура гена (ЭИС), свойственная эукариотическим организмам, уже долгое время представляет большой интерес с различных точек

*ipoverennaya@gmail.com
**astakhova@lpm.org.ru

зрения, например, изучения альтернативного сплайсинга. Интронам, долгое время считавшимся лишь еще одной группой «мусорной» ДНК, стало уделяться больше внимания лишь в последние десятилетия, когда появилось больше сведений об их функциональной значимости, в частности, об их регуляторной роли в сплайсинге и экспрессии генов [1]. Например, почти во всех эукариотических организмах гены, кодирующие рибосомальные белки, содержат на 5’-конце интроны с регуляторными элементами.

Важной характеристикой интрона является его фаза, отражающая положение интрона в последовательности относительно рамки считывания. Интрон, расположенный между кодонами, имеет фазу 0; интроны, расположенные между 1 и 2 или между 2 и 3 нуклеотидами кодона, имеют фазы 1 и 2, соответственно. В ходе целого ряда исследований, посвященных распределению фаз интронов, выяснилось, что во всех изученных организмах приблизительно 50 % интронов имеют фазу 0, 30 % − фазу 1, а 20 % − фазу 2 [2, 3]. Такая закономерность получила название "золотое правило 50/30/20" или "соотношение 5:3:2". Среди причин такого устойчивого распределения фаз интронов назывались процесс "перетасовки" экзонов [4], связь интронов фазы 1 и 2 с консервативными участками последовательности [5], и даже особая структура белка [6]. Тем не менее, ни одно из представленных предположений не стало общепринятым.

Эволюцию интронов обычно рассматривают только с точки зрения эволюции экзон-интронной организации гена: когда произошло встраивание или вырезание интронов. Споры вокруг теорий о раннем или позднем происхождении интронов привели к созданию синтетической теории: интроны заселили древнего предка эукариотических организмов во время митохондриального эндосимбиоза, и с тех пор происходит как постепенная потеря интронов, так и их появление. Согласно этой теории, альфа-протеобактериальный предок митохондрии содержал большое количество интронов группы II (вид самосплайсирующихся интронов), которые мигрировали в ДНК хозяина и стали обычными сплайсосомальными интронами [7].

Количество интронов в разных эукариотах сильно варьируется: от нескольких интронов на геном до десятков интронов на ген. Из-за такого разброса значений количества интронов в геноме эукариотические организмы условно стали разделять на интрон-бедные (большинство одноклеточных эукариот) и интрон-богатые (животные, растения, некоторые грибы и некоторые одноклеточные, такие как хламидомонада) геномы. Интрон-бедные и интрон-богатые организмы различаются не только разной плотностью интронов, но и распределением интронов в геноме: для интрон-богатых геномов характерно относительно равномерное распределение интронов по геному, в то время как в интрон-бедных геномах интроны перепредставлены в 5’-областях генов [8].

Для интрон-богатых организмов (в частности позвоночных) свойственно наличие многочисленных длинных интронов. Например, как оказалось, примерно 90 % и 40 % генов в геномах приматов имеют хотя бы один интрон длиной более 1000 п.о. и 10000 п.о., соответственно. Наличие таких длинных интронов можно объяснить с точки зрения регуляции экспрессии гена, например, они могут содержать в себе регуляторные элементы или, как можно предположить, сами по себе фактически являются регуляторами за счет длины – чем длиннее интрон, тем дольше он будет транскрибироваться. Также длинные интроны могут являться энхансерами мейотического кроссинговера между белок-кодирующими последовательностями, поскольку вероятность кроссинговера между экзонами, разделенными длинными интронами, гораздо больше, чем между такими же последовательностями, но без интронов [9].

В работе [10] обнаружена положительная зависимость между уровнем эволюционной консервации гена и его интронной нагрузкой. В тоже время имеет место

отрицательная зависимость между уровнем экспрессии гена и числом интронов, а также общим размером интронной области. Одним из возможных объяснений этого является идея «баланса затрат и выгод». Эволюционно консервативные (функционально важные) гены могут "позволить себе" негативные последствия поддержания нескольких интронов. Известно, что транскрипция интронов связана со значительными расходами энергии. Логично предположить, что только функционально важные гены могут поддерживать нагрузку большого количества интронов (или большой общий размер интронов). Интересным фактом является то, что в генах с большим содержанием длинных интронов, по-видимому, нарушено «золотое правило» распределения фаз, так как недавно было показано, что длинные интроны предпочитают иметь фазу 1, а не обычно перепредставленную фазу 0 [11]. Одним из объяснений перепредставленности фазы 0 является создание интронами фазы 0 большей гибкости для процесса кодирования белка, так как они не влияют на концевые нуклеотиды экзонов, где обычно расположены сайты сплайсинга [12]. Поэтому изучение консервативности фаз в генах-ортологах с длинными интронами представляет особый интерес: есть ли случаи изменения фазы, насколько часто они происходят, и предпочитает ли фаза 1 переходить в фазу 0. Другой задачей данной работы является анализ распределения длин ортологичных интронов, которые принято считать низко-консервативными в связи с высоким уровнем возникновения мутации.

## СЕМЕЙСТВА ГЕНОВ

В продолжение работы [11], для анализа экзон-интронной структуры гена были выбраны такие семейства генов, в которых преобладают длинные интроны фазы 1. Для этого были созданы выборки из 100 генов с подобными характеристиками для геномов *Homo sapiens, Macaca mulatta, Mus musculus, Anolis carolinensis* и *Danio rerio*. Четыре гена, представленные во всех выборках, – *ptprd, mdga2, dscam* и *tenm1* – были использованы в данной работе. Интересно, что помимо нарушенного соотношения фаз интронов, эти гены также имеют схожие биологические функции: кодируемые ими белки локализованы на поверхности клетки, где участвуют в процессах клеточной адгезии. Они играют важную роль в развитии и правильном функционировании нервной системы, в клетках которой происходит их основная экспрессия. Три гена (*ptprd, mdga2* и *dscam*) содержат иммуноглобулиновые домены, что свидетельствует об их принадлежности к суперсемейству иммуноглобулинов.

### *Ген ptprd*

Белок PTPRD (тирозинфосфатаза дельта), кодируемый геном *ptprd*, является членом семейства белковых тирозиновых фосфатаз (PTP), связанного с процессами клеточного роста и дифференцировки, митоза и онкогенной трансформации. PTPRD – белок рецепторного типа, в его структуре есть внеклеточный участок, один трансмембранный участок и два тандемных цитоплазматических домена. Внеклеточный участок содержит три Ig-подобных домена и 8 доменов, подобных фибронектину III типа. Для гена *ptprd* человека известно наличие многих альтернативных изоформ.

У животных белок PTPRD экспрессируется в некоторых участках головного мозга, например, в областях CA2 и CA3 гиппокампа, а также в B-лимфоцитах и в мозговом веществе тимуса. PTPRD является важным регулятором синаптической пластичности и играет важную роль в процессе обучения и памяти [13].

Значительное увеличение мутаций в гене *ptprd* было показано у больных раком молочной железы [14], учащение мутаций/делеций гена также связывают с легочной аденокарциномой [15]. В 2012 *ptprd* стал объектом изучения в качестве опухолевого супрессора в нейробластоме, но его роль оказалась незначительной [16].

### *Ген mdga2*

Ген *mdga2* (изначальное название – *mamdc1*) впервые был выделен из мозга крыс в 2004 году [17]. Кодируемый им белок длиной 956 аминокислотных остатков содержит несколько Ig-подобных доменов и один высоко-консервативный MAM домен, фактически являющийся соединением трех доменов – меприна, A5 белка, а также тирозин-фосфатазы mu рецепторного типа. В отличие от белка MDGA1 в белке MDGA2 нет домена, подобного фибронектину III типа. На С-конце белка есть сайт связывания с гликозилфосфатидилинозитольным якорем – гликолипидом, который навешивается на белок в процессе посттрансляционных модификаций. Ген *mdga2*, в основном, экспрессируется в центральной и периферической нервной системе в различных субпопуляциях нейронов, например, в коре головного мозга и в базилярной борозде варолиева моста.

Однонуклеотидные полиморфизмы в гене *mdga2* предположительно связаны с проявлением симптомов невротизма [18], а также с появлением сыпи и повышенной светочувствительности при системной красной волчанке [19].

### *Ген dscam*

Ген *dscam* человека является членом иммуноглобулинового суперсемейства молекул клеточной адгезии (Ig-CAMs). Его основные функции связаны с развитием центральной и периферической нервной системы. Таким образом, он активно экспрессируется в эмбриональной нервной системе, с наибольшим уровнем экспрессии в головном мозге эмбриона. Гиперэкспрессия гена *dscam* ведет к синдрому Дауна.

Белок DSCAM содержит 10 Ig-подобных доменов и 6 доменов фибронектина III типа. Структурно это трансмембранный белок с N-концевым внеклеточным и С-концевым цитоплазматическим доменами.

У насекомых ген *dscam* также играет крайне важную роль для развития нервной системы и для правильного функционирования иммунной системы. При этом, в отличие от млекопитающих, ген *dscam* некоторых насекомых очень вариативен: например, потенциально возможное количество продуктов гена *dscam* дрозофилы равно 38016. Такое количество связано с тем, что иммуноглобулин-подобные домены в гене *dscam* дрозофилы закодированы не конститутивными экзонами, а экзонными кластерами – взаимоисключающимися альтернативно-сплайсируемыми экзонами.

### *Ген tenm1*

Продукт гена *tenm1* (синонимичное название: *odz1*) принадлежит к подсемейству трансмембранных белков тенейринов из семейства тенаскинов (гликопротеины из внеклеточного матрикса). Свое название тенейрины образуют из слов «тенаскин» и «нейрон», который является основным местом экспрессии генов тенейринов, в частности гена *tenm1*. Белок TENM1 высококонсервативен как у позвоночных, так и беспозвоночных животных. Он содержит три домена: N-концевой внутриклеточный домен, содержащий 3 EF-руки и 2 пролин-богатых повтора, трансмембранный домен и С-концевой внеклеточный домен с 8 EGF-подобными повторами, EF-рукой и множественными YD повторами [20].

TENM1 участвует во внутриклеточных сигнальных каскадах [21], а также в синаптической передаче сигнала между дендритами проекционных нейронов и аксонами обонятельных нейронов. Так, потеря тенейринов вызывает нарушения во взаимодействии некоторых обонятельных и проекционных нейронов [22].

## МЕТОДИКА АНАЛИЗА ЭИС СЕМЕЙСТВА ГЕНОВ-ОРТОЛОГОВ

Список генов-ортологов для каждого анализируемого гена получен из раздела «Orthologs from Annotation Pipeline» в БД Gene на сайте NCBI [23]. Данный раздел представляет ссылку на таблицу ортологов, созданную по результатам аннотации геномных сборок организмов позвоночных. Информация о генах в виде файлов в GenBank-формате была скачана из БД GenBank [24].

В случае наличия альтернативных изоформ в гене в качестве канонической последовательности, то есть изоформы, представляющей данный ген в дальнейшем анализе, выбиралась изоформа гена с максимальной длиной белкового продукта либо с наибольшим числом экзонов (при равной длине белков). Во избежание возможных ошибок была проведена верификация данных об экзон-интронной структуре гена (выбранной изоформы) через проверку целостности рамки считывания и сравнения транслированной последовательности продукта гена, полученной по координатам ЭИС, с последовательностью белка данного гена из курируемой БД RefSeq [25].

Для определения ортологичных интронов в генах-ортологах были построены выравнивания экзон-интронных структур, представляющие собой две схематичные таблицы, отдельно для экзонов и отдельно для интронов, в колонках которых представлены порядковые номера выровненных между собой гомологичных экзонов (интронов). Построение подобного выравнивания экзонов было основано на сравнении длины экзонов и множественном выравнивании белковых последовательностей. Выравнивание интронов было построено с помощью выравнивания экзонов по следующем принципу: интрон $i$ гена $a$ выравнивается с интроном $j$ гена $b$, если $i$-ый экзон гена $a$ гомологичен $j$-ому экзону гена $b$, а $(i + 1)$-ый экзон гена $a$ – $(j + 1)$-ому экзону гена $b$.

## РЕЗУЛЬТАТЫ

В построенных выравниваниях ЭИС мы выделили экзонные группы, содержащие гомологичные экзоны, и интронные группы, состоящие из выровненных гомологичных интронов. Строго говоря, такие интроны являются потенциально гомологичными, поскольку известны случаи параллельного встраивания интронов в одно и тоже место. Однако подобные события обычно характерны только для очень далеких организмов (например, таких как растения и животные) [26], а в данной работе мы рассматриваем только позвоночных животных. Таким образом, для гена *ptprd* было найдено 37 и 36 выровненных экзонных и интронных групп; для гена mdga2 – 18 и 17; для *tenm1* – 32 и 31; и, наконец, для гена *dscam* – 33 и 32 экзонных и интронных групп, соответственно (подробнее см. табл. 1). Ввиду потери интронов у отдельных организмов, количество интронов в интронных группах варьировалось: например, от 56 до 132 интронов у гена *dscam*, от 31 до 87 интронов у гена *ptrpd*.

**Таблица 1.** Данные о построенных выравниваниях экзон-интронных структур генов *ptprd, mdga2, dscam, tenm1*

| | ***ptprd*** | ***mdga2*** | ***dscam*** | ***tenm1*** |
|---|---|---|---|---|
| Число генов-ортологов | 90 | 134 | 137 | 145 |
| Число экзонных групп | 37 | 18 | 33 | 32 |
| Число интронных групп | 36 | 17 | 32 | 31 |
| Максимальное число ортологичных экзонов в группе | 88 | 128 | 133 | 142 |
| Минимальное число ортологичных экзонов в группе | 32 | 69 | 57 | 70 |



Математическая биология и биоинформатика. 2017. Т. 12. № 2. doi: 10.17537/2017.12.302

## Анализ фазы интрона

Сравнение фаз ортологичных интронов показало их ожидаемую, с точки зрения альтернативного сплайсинга, высокую консервативность, но, тем не менее, для всех генов были найдены случаи изменения фазы или – говоря об эволюционных процессах – случаи слайдинга, то есть смещения границ интрона на небольшие расстояния в процессе эволюции. Необходимо отметить, что слайдинг интронов необязательно приводит к смене фазы, поскольку интрон может сместиться на длину, кратную трем, и тем самым сохранить свою изначальную фазу. Для анализируемых генов *dscam, tenm1, mdga2* и *ptprd* удалось найти всего два случая слайдинга без изменения фазы, но слабовыровненные концы экзонов в месте слайдинга ставят под сомнение надежность данных находок, поэтому они не были учтены в дальнейшем анализе.

Сравнение фаз интронов в интронных группах 4 генов обнаружило 10 случаев потенциального слайдинга с изменением фазы. Достоверность 7 случаев стоит под вопросом из-за, по-видимому, ошибочного определения экзон-интронных границ, связанного со слабым выравниванием на границах экзонов и/или неизвестной последовательностью интрона (см. Дополнительные материалы к статье). Оставшиеся три случая представлены в таблице 2 и на рисунках 1 и 2.

**Таблица 2.** Найденные случаи слайдинга

| | ***dscam*** | ***mdga2*** | ***tenm1*** |
|---|---|---|---|
| Организм *A*, в гене которого произошел слайдинг интрона *x* | *Camelus bactrianus* | *Melopsittacus undulatus* | *Callorhinchusmilii* |
| Изменение фазы | 1→2 | 1→2 | 1→0 |
| Порядковый номер интрона *x* в гене из организма *A* | 22 | 6 | 8 |
| Номер интронной группы | 24 | 6 | 6 |
| Длины экзонов слева и справа в гене из организма *A* | 100–188 | 274–329 | 131–201 |
| Длины экзонов слева и справа в других организмах | 99–189 | 270–330 | 153–200 |

Как видно из рисунка 1, выравнивание на границах экзонов, где произошел слайдинг, является высококонсервативным для всех трех генов *dscam, mdga2* и *tenm1*. Граница экзонов в генах-гомологах *dscam* происходит в последовательности кодонов IRG||YIL (она указана двойной вертикальной чертой), при этом в гомологе из генома двугорбого верблюда (*Camelus bactrianus*), помимо смещения интрона на один нуклеотид с соответственным изменением фазы с 1-ой на 2-ую, произошла замена в экзонной последовательности: с тирозина Y на серин S (соответственно, IRG||SIL). Тирозин – ароматическая аминокислота, в то время как серин – полярная незаряженная алифатическая, однако они обе содержат гидроксильную группу в радикальном остатке. Если рассмотреть отдельно нуклеотидную последовательность конца предыдущего и начала следующего экзонов, а также начало и конец интрона, то видно, что граница экзонов проходит не по кодону тирозина (серина), а по кодону предыдущей аминокислоты глицина. Поэтому в данном случае просто произошла одиночная мутация в кодирующем тирозин триплете TAC, приведшая к образованию кодона TCC.

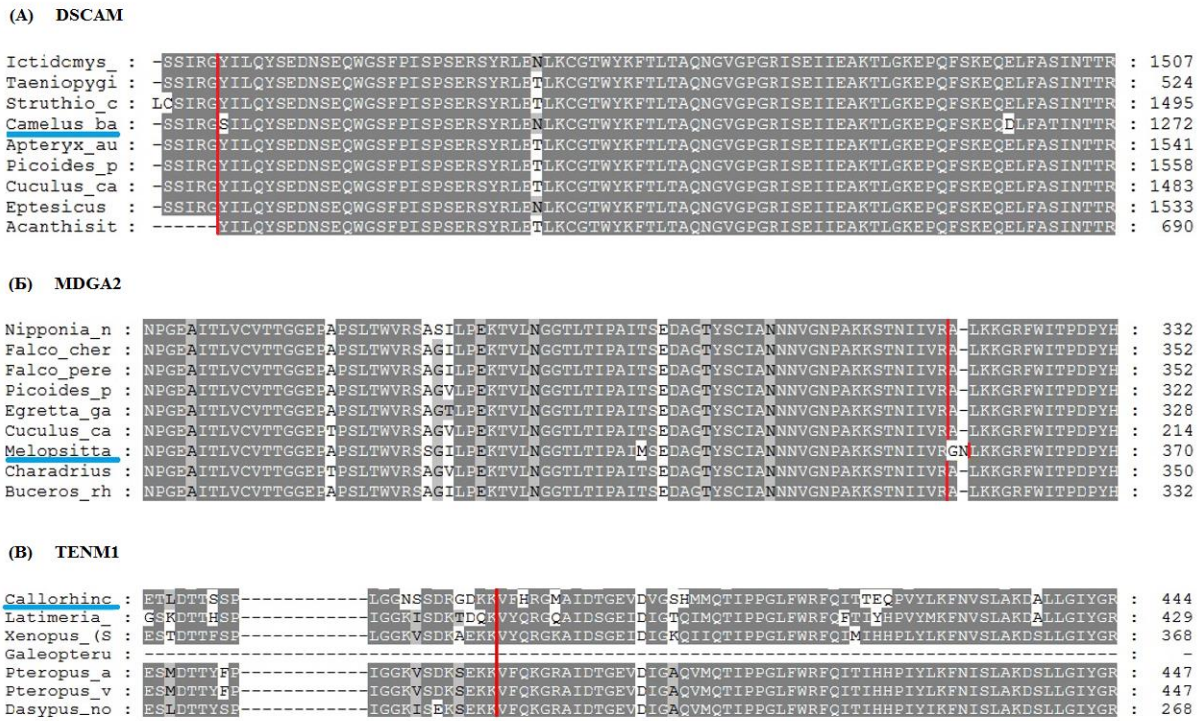


**Рис. 1.** Фрагменты множественного белкового выравнивания ортологов генов *dscam* (А), *mdga2* (Б) и *tenm1* (В). Синей линией подчеркнут организм, в гене которого предположительно произошел слайдинг, красной линией – граница экзонов. Обозначения организмов: Ictidomys_ – *Ictidomys tridecemlineatus*; Eptesicus_ – *Eptesicus fuscus*; Acanthisit – *Acanthisitta chloris*; Taeniopygi – *Taeniopygia guttata*; Struthio_c – *Struthio camelus*; Camelus_ba – *Camelus bactrianus*; Apteryx_au – *Apteryx australis*; Picoides_p – *Picoides pubescens*; Cuculus_ca – *Cuculus canorus*; Nipponia_n – *Nipponia nippon*; Falco_cher – *Falco cherrug*; Falco_pere – *Falco peregrinus*; Egretta_ga – *Egretta garzetta*; Melopsitta – *Melopsittacus undulatus*; Charadrius – *Charadrius vociferus*; Buceros_rh – *Buceros rhinoceros*; Callorhinc – *Callorhinchus milii*; Latimeria – *Latimeria chalumnae*; Xenopus_(S – *Xenopus (Silurana) tropicalis*; Galeopteru – Galeopterus variegatus; Pteropus_a – Pteropus alecto; Pteropus_v – *Pteropus vampyrus*; Dasypus_no – *Dasypus novemcinctus*.

Гомолог гена *mdga2* в организме волнистого попугайчика (*Melopsittacus undulates*) на самой границе экзонов, интрон между которыми сместился на одну фазу (с 1-ой на 2-ую), содержит вставку в виде аспарагина N (рис. 1,Б). Хотя отсутствие аспарагина в данной позиции в других гомологах может свидетельствовать об ошибке аннотации, более детальное рассмотрение ближайшего нуклеотидного окружения подтверждает правильность определения экзон-интронной структуры. Начало последующего экзона из генома малой белой цапли (*Egretta garzetta*) содержит непрерывную последовательность из 7 аденинов, а аналогичный ей экзон из *M. undulates* имеет последовательность только из 6 аденинов (рис. 2,Б). Тем не менее, вырожденность кодонов позволила сохранить аминокислотную последовательность LKK неизменной. Кроме того, начало и конец интересующих интронов, несмотря на общеизвестное сильное расхождение интронных последовательностей, почти идентичны.

В отличие от случаев слайдинга, произошедших в генах *dscam* и *mdga2*, где интрон изменил фазу с первой на вторую, смещение восьмого интрона в гене *tenm1* представителя хрящевых рыб австралийского каллоринха (*Callorhinchus milii*) привело к замене фазы интрона 1 на фазу 0. Интрон сместился на один нуклеотид назад, на уровне аминокислотного выравнивания граница экзонов не изменилась, более того сдвиг интрона не повлек за собой образования мутации (замены аминокислоты) в данном месте.

Количество вышеперечисленных событий слайдинга не позволяет сделать выводы о предпочтительности «старой» и «новой» фазы при смещении интрона, хотя уже можно

заметить, что во всех случаях смещались интроны фазы 1. Это может быть объяснено количественным превосходством интронов данной фазы в анализируемых генах. Вопреки ожиданию, интроны чаще сменяли фазу на более редкую фазу 2, чем на самую распространенную фазу 0.

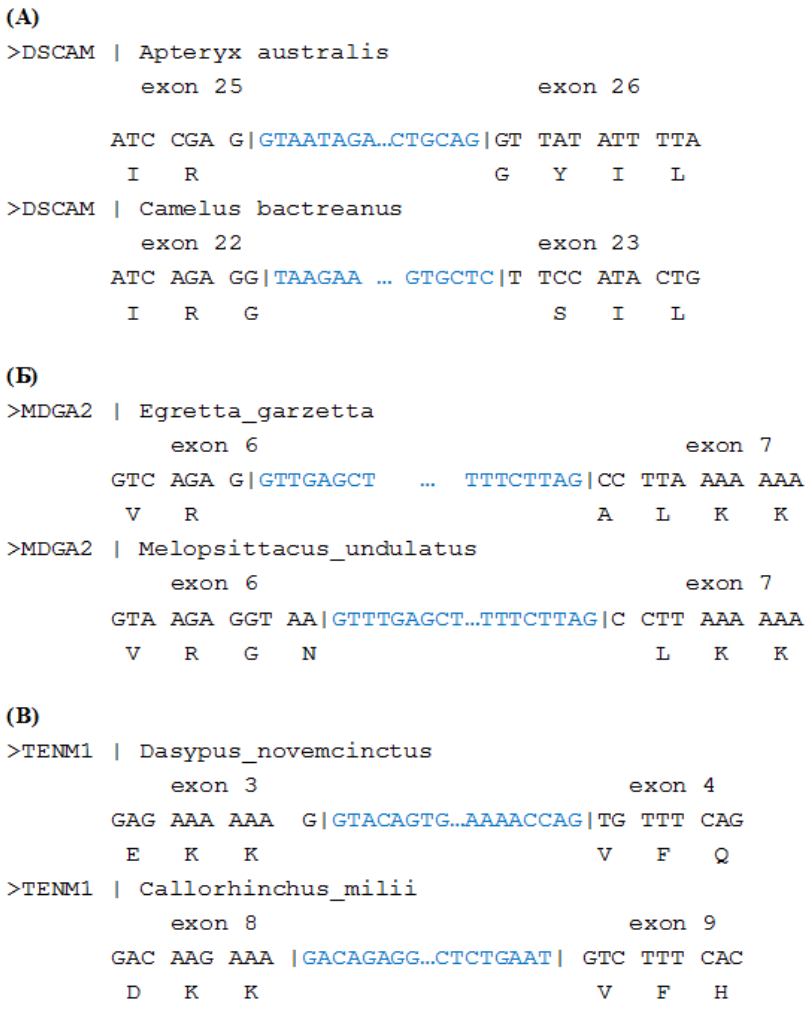


**Рис. 2.** Случаи слайдинга в экзон-интронной структуре генов *dscam* (А), *mdga2* (Б) и *tenm1* (В).

Сместившиеся интроны генов *dscam* и *tenm1* имеют неканоничные донорные и акцепторные сайты сплайсинга (рис. 2): |TA и TC| и |GA и AT| вместо каноничных |GT и AG|. Наличие неканоничных сайтов характерно только для ~1.5 % интронов млекопитающих [27]; в гомологах генов *dscam, tenm1, mdga2* и *ptprd* неканоничные сайты встретились у 11 из 4121 интронов, 9 из 4183, 4 из 1952 и 5 из 2880 представленных интронов, соответственно, причем в разных позициях. Из-за столь низкой частоты встречаемости неканоничных сайтов возникает вопрос о неслучайности их наличия в тех интронах, в которых предположительно произошел слайдинг, и, поскольку основной гипотезой механизма слайдинга считается двойной альтернативный сплайсинг [28], дальнейшее изучение слайдинга может пролить свет на механизм перемещения интронов.

### Анализ длины интрона

Анализ длин интронов показал, что длины гораздо менее консервативны, чем фазы, они могут значительно изменяться внутри одной и той же группы ортологичных

интронов (особенно длинных), что ранее уже подтверждалось данными литературы [29].

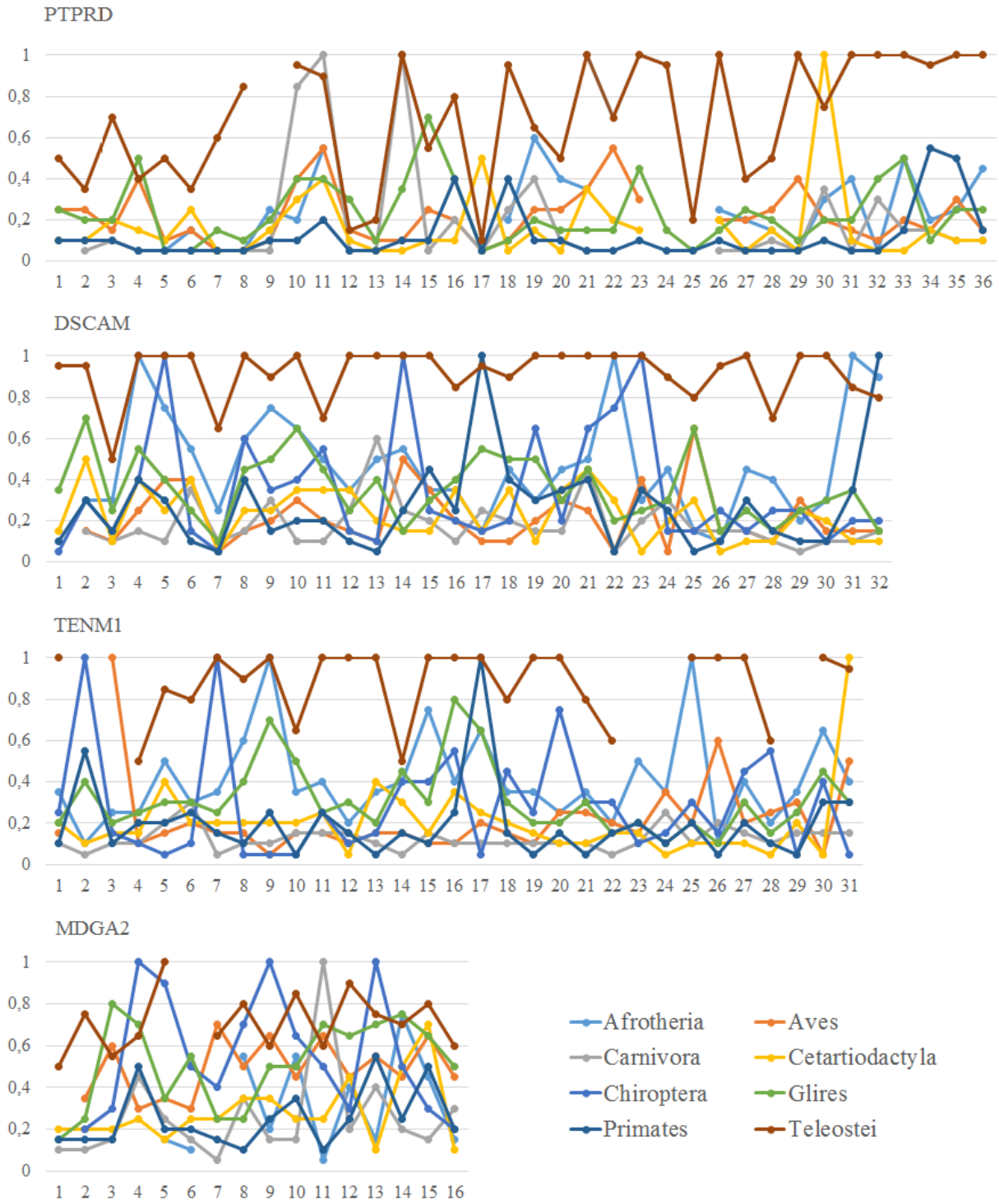


**Рис. 3.** Интервал расхождений от средней длины интрона в группах ортологичных интронов для разных таксономических групп. Каждому интрону приписывается нормализованная длина $N = (L–A)/A$, где $L$ – это длина данного интрона, а $A$ – это средняя длина по группе ортологичных интронов. По оси абсцисс – номер интронной группы, по оси ординат – значение порога x такое, что в интервал нормализованных длин $(–x, x)$ попадают более 90 % представителей данной интронной группы. Интронные группы, в которых было менее 3-х интронов, не учитывались.

Тем не менее, на уровне таких таксономических групп, как классы и отряды, можно наблюдать некоторую консервативность, если вместо длины интрона $L$ мы будем

рассматривать нормализованную длину $N = (L–A)/A$, где $A$ – это средняя длина по группе ортологичных интронов (рис. 3). Например, для гена *ptprd* в случае птиц (28 видов) нормализованная длина в среднем выпадает из интервала (–0.15; 0.15) только для 14.8 % интронов. А при расширении интервала до (–0.5; 0.5) доля таких интронов составляет всего 3.2 %. На рисунке 3 для каждого гена приведены графики границ интервала, таких, что в данный интервал попадает 90 % представителей из группы ортологичных интронов. Как видно, подобные границы для большинства анализируемых интронов из геномов птиц (*Aves*), приматов (*Primates*), хищников (*Carnivora*) и парнокопытных животных (*Cetartiodactyla*) варьируют в пределах 0.05–0.4. Однако, такую консервативность можно наблюдать не для всех интронных групп анализируемых генов, например, интроны костистых рыб (*Teleostei*) имеют очень широкий разброс значений длин (см. рис. 3). При этом, костистые рыбы имеют самые короткие интроны (рис. 4).

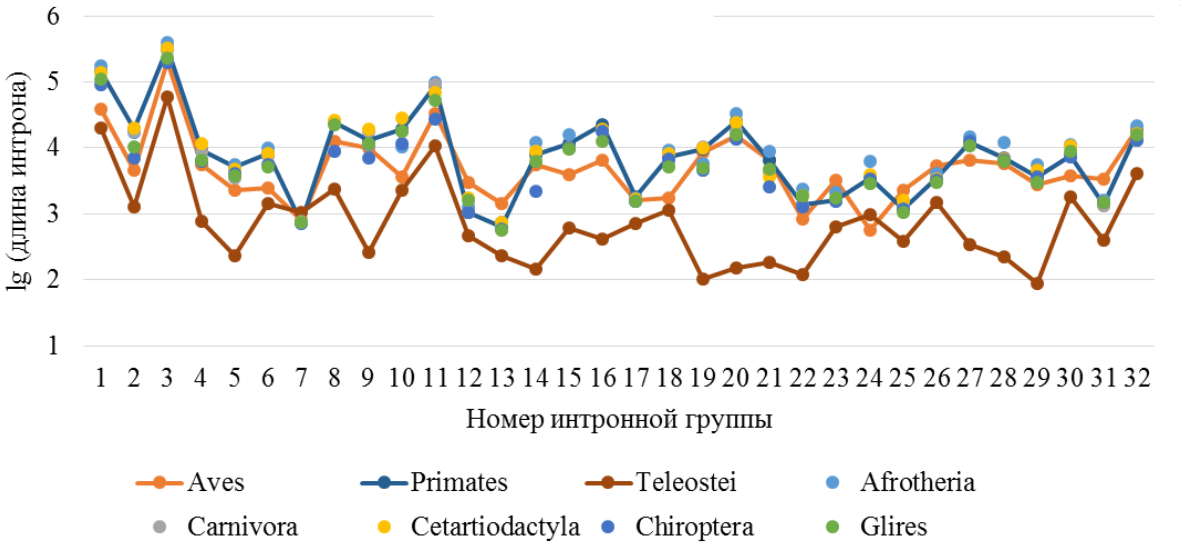


**Рис. 4.** Медиана распределения длин интронов в интронных группах в логарифмическом масштабе для разных таксономических групп в гене *dscam*. (Таксономические группы, содержащие менее 3-х членов, не представлены).

Среди групп интронов гена *dscam* птиц выделяются группы 7, 13, 22 и 24, интроны в которых имеют маленький разброс длин (90 % интронов из групп 7, 22 и 24 с нормализованной длиной попадают в интервал (–0.05; 0.05), для группы 13 такой интервал равен (–0.1; 0.1)), а также относительно небольшую длину: примерно 800, 1400, 850, и 600 п.о. Множественные выравнивания последовательностей из данных групп демонстрируют довольно высокое сходство последовательностей. По итогам расширения выравниваний путем включения интронных последовательностей из остальных организмов выяснилось, что некоторые участки сохраняют свою консервативность и в других таксономических группах; например, первые 27 п.о. на 5’-концах интронов из группы 7 и последние 50 п.о. на 3’-концах интронов из группы 22 (см. Приложение). Поскольку интронные последовательности известны высокой степенью мутагенеза и, в соответствии с этим, большой вариативностью, сохранение данных участков под действием естественного отбора, по-видимому, указывает на биологическую важность этих подпоследовательностей. Мы предполагаем, что эти консервативные участки несут в себе регуляторные последовательности, хотя поиск известных мотивов в них результатов не принес.

Наиболее вероятное объяснение значительного увеличения длины интронов связано с накоплением в интронах большого количества мобильных элементов и повторов. Для подтверждения данного предположения с помощью программы RepeatMasker

(версия 4.0.6) [29] мы провели поиск всевозможных повторов внутри последовательности каждого интрона. Полученные результаты показали значительную положительную корреляцию между длиной интрона (*LIntr*) и общей длиной найденных в нем повторов и мобильных элементов (далее для краткости «повторов», *LRep*). Эта зависимость наблюдается как при рассмотрении наборов интронов, относящихся к отдельным организмам, так и наборов интронов, относящихся к отдельным генам и интронным группам. Отметим, что значимая зависимость между длиной интрона и долей длины, которую в интроне занимают обнаруженные повторы (*PRep*), отсутствует.

**Таблица 3.** Коэффициенты корреляции Пирсона между длиной интрона *LIntr* и суммарной длиной повторов *LRep*, а также между длиной интрона и долей длины повторов в интроне *PRep* для всех интронов из генов *dscam, mdga2, ptprd* и *tenm1*

| Ген | ***LIntr* vs. *LRep*** | ***LIntr* vs. *PRep*** |
|---|---|---|
| *dscam* | 0.899 | 0.215 |
| *mdga2* | 0.948 | 0.277 |
| *ptprd* | 0.920 | 0.314 |
| *tenm1* | 0.933 | 0.431 |
| Все анализируемые гены | 0.922 | 0.293 |

**Таблица 4.** Коэффициенты корреляции Пирсона между длиной интрона *LIntr* и суммарной длиной повторов *LRep*, а также между длиной интрона и долей длины повторов в интроне *PRep* для интронов из генов человека *dscam, mdga2, ptprd* и *tenm1*

| Набор данных | ***LIntr* vs. *LRep*** | ***LIntr* vs. *PRep*** |
|---|---|---|
| *dscam, H. sapiens* | 0.994 | 0.274 |
| *mdga2, H. sapiens* | 0.987 | 0.589 |
| *ptprd, H. sapiens* | 0.981 | 0.308 |
| *tenm1, H. sapiens* | 0.993 | 0.562 |
| Все анализируемые гены, *H. sapiens* | 0.979 | 0.373 |

**Таблица 5а.** Данные для интронных групп из гена *mdga2*. Коэффициенты корреляции Пирсона рассчитаны между длиной интрона *LIntr* и суммарной длиной повторов *LRep*, а также между длиной интрона и долей длины повторов в интроне *PRep*

| группа | медиана длины интронов | ***LIntr* vs. *LRep*** | ***LIntr* vs. *PRep*** | группа | медиана длины интрона | ***LIntr* vs. *LRep*** | ***LIntr* vs. *PRep*** |
|---|---|---|---|---|---|---|---|
| 1 | 332418 | 0.908 | 0.765 | 9 | 30246.5 | 0.764 | 0.369 |
| 2 | 77263 | 0.918 | 0.796 | 10 | 36278 | 0.830 | 0.654 |
| 3 | 62275 | 0.876 | 0.726 | 11 | 4726 | 0.752 | 0.030 |
| 4 | 12142 | 0.871 | 0.544 | 12 | 4910.5 | 0.839 | 0.518 |
| 5 | 23599.5 | 0.817 | 0.666 | 13 | 491 | 0.576 | 0.174 |
| 6 | 35249.5 | 0.813 | 0.532 | 14 | 13674 | 0.847 | 0.548 |
| 7 | 24199 | 0.755 | 0.309 | 15 | 9948 | 0.862 | 0.696 |
| 8 | 62463 | 0.865 | 0.756 | 16 | 2556 | 0.635 | 0.481 |

В таблицах 3, 4 и 5 показаны значения коэффициентов корреляции Пирсона между величинами (а) *LIntr* и *LRep* (б) *LIntr* и *PRep*, рассчитанные для различных групп интронов. В таблице 3 рассмотрены интроны всех организмов, относящиеся к определенному гену; в таблице 4 рассмотрены только гены человека.

**Таблица 5б.** Данные для интронных групп из гена *tenm1*. Коэффициенты корреляции Пирсона рассчитаны между длиной интрона *LIntr* и суммарной длиной повторов *LRep*, а также между длиной интрона и долей длины повторов в интроне *PRrep*

| группа | медиана длины интрона | *LIntr* vs. *LRep* | *LIntr* vs. *PRep* | группа | медиана длины интрона | *LIntr* vs. *LRep* | *LIntr* vs. *PRep* |
|---|---|---|---|---|---|---|---|
| 1 | 65002 | 0.855 | 0.672 | 17 | 2598 | 0.598 | 0.359 |
| 2 | 1656 | 0.649 | 0.079 | 18 | 12802 | 0.917 | 0.802 |
| 3 | 145387.5 | 0.746 | 0.576 | 19 | 6298.5 | 0.773 | 0.607 |
| 4 | 26376.5 | 0.899 | 0.766 | 20 | 7571 | 0.890 | 0.852 |
| 5 | 30440.5 | 0.897 | 0.763 | 21 | 4902 | 0.624 | 0.474 |
| 6 | 17873.5 | 0.878 | 0.782 | 22 | 25669 | 0.849 | 0.771 |
| 7 | 1422 | 0.928 | 0.782 | 23 | 26359 | 0.835 | 0.625 |
| 8 | 4799.5 | 0.673 | 0.457 | 24 | 1687 | 0.042 | –0.055 |
| 9 | 1364.5 | 0.672 | 0.313 | 25 | 10771 | 0.869 | 0.704 |
| 10 | 3241 | 0.343 | 0.167 | 26 | 1178 | 0.692 | 0.219 |
| 11 | 58967.5 | 0.916 | 0.766 | 27 | 10721 | 0.790 | 0.656 |
| 12 | 1565 | 0.845 | 0.627 | 28 | 5597.5 | 0.878 | 0.623 |
| 13 | 1959 | 0.575 | 0.472 | 29 | 823 | 0.806 | 0.512 |
| 14 | 11202.5 | 0.715 | 0.523 | 30 | 583 | 0.707 | 0.422 |
| 15 | 16486.5 | 0.870 | 0.715 | 31 | 1450 | 0.694 | 0.409 |
| 16 | 5447.5 | 0.910 | 0.768 | | | | |

В таблицах 5а и 5б приведены данные по отдельным интронным группам для генов *mdga2* (табл. 5а) и *tenm1* (табл. 5б). Данные для других генов аналогичны и приведены в Дополнительных материалах. Низкие значения коэффициента корреляции *LIntr* vs. *LRep* могут быть объяснены малой длиной интрона.

**Таблица 6.** Коэффициент корреляции Пирсона между длиной интрона и суммарной длиной найденных повторов и мобильных элементов для интронов из гена *mdga2*

| ген *mdga2* | Интронная группа | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| Все интроны | 0.91 | 0.92 | 0.88 | 0.87 | 0.82 | 0.81 | 0.76 | 0.86 | 0.76 | 0.83 | 0.75 | 0.84 | 0.58 | 0.85 | 0.86 | 0.63 |
| Интроны содержащие менее 5 % символов N | 0.89 | 0.94 | 0.93 | 0.89 | 0.86 | 0.89 | 0.79 | 0.90 | 0.82 | 0.91 | 0.13 | 0.83 | 0.59 | 0.94 | 0.86 | 0.64 |
| Интроны содержащие менее 1 % символов N | 0.93 | 0.97 | 0.93 | 0.94 | 0.86 | 0.89 | 0.80 | 0.91 | 0.76 | 0.91 | 0.20 | 0.82 | 0.59 | 0.93 | 0.86 | 0.64 |

Тем не менее, были найдены и обратные случаи, когда в длинных интронах повторы были не обнаружены вовсе или их доля очень мала (например, интрон 1-й группы из гомолога гена *mdga2* в *Xenopus tropicalis* имеет длину почти 150 тыс. п.о. и ни одного повтора). Это может быть вызвано как минимум двумя причинами. Во-первых, последовательность интрона не всегда аннотирована полностью, в таких случаях неизвестные участки отображаются символами N; поиск каких-либо функциональных элементов на таких участках просто невозможен. Так, при рассмотрении только тех последовательностей, где содержание символа N не превышает 5 %, можно наблюдать небольшое увеличение коэффициента корреляции во многих интронных группах (таблица 6). Однако, интересно отметить, что в группе 11 гена *mdga2* такой отбор привел к резкому уменьшению коэффициента корреляции.

Во-вторых, уровень аннотации мобильных элементов в разных организмах сильно варьируется, поэтому вероятно, что программа RepeatMasker, которая изначально была направлена на поиск мобильных элементов человека, могла пропустить часть элементов в недостаточно изученных организмах.

## ВЫВОДЫ

Мы построили выравнивания экзон-интронных структур около 100 организмов различных таксонов для ортологов четырех генов, в структуре которых превалируют длинные интроны фазы 1: *ptprd, dscam, tenm1* и *mdga2*. Полученные выравнивания позволили выделить группы ортологичных интронов, необходимые для анализа изменения длины и случаев перемещения границ интрона в ходе эволюции (слайдинга). Мы нашли три события слайдинга с изменением фазы и десять случаев без изменения фазы. Выявленные случаи слайдинга с изменением фазы не подтвердили нашу первоначальную гипотезу о том, что в процессе смещения интроны предпочитают менять свою фазу на более часто встречающуюся нулевую. Однако в связи с небольшим числом найденных случаев для окончательных выводов необходимо провести поиск случаев изменения фазы на большей выборке генов. Полученные результаты поднимают вопрос о связи слайдинга интронов с присутствием неканонических сайтов сплайсинга.

Несмотря на высокую вариативность длин ортологичных интронов, использование нормализованных длин на уровне таких таксономических групп как класс или порядок позволило обнаружить некоторую консервативность длины. Найденные в ортологичных группах с относительно короткими интронами консервативные последовательности наводят на мысль, что сохранение длины в течение эволюции может являться сигналом о наличии в интронах регуляторных элементов. Несмотря на недостаточность в настоящий момент данных о последовательностях интронов и мобильных элементах в немодельных организмах, наши результаты частично подтверждают предположение, что значительные увеличения длин интронов связаны с накоплением мобильных элементов внутри интронной последовательности.

Работа выполнена при поддержке Российского фонда фундаментальных исследований (грант № 16-04-01640).

## СПИСОК ЛИТЕРАТУРЫ

1. Patel A.A., McCarthy M., Steitz J.A. The splicing of U12-type introns can be a rate-limiting step in gene expression. *EMBO J.* 2002. V. 21. № 14. P. 3804–3815.
2. de Souza S.J., Long M., Klein R.J., Roy S., Lin S., Gilbert W. Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci. U.S.A.* 1998. V. 95. № 9. P. 5094–5099.

3. Long M., de Souza S.J., Rosenberg C., Gilbert W. Relationship between ‘proto-splice sites’ and intron phases: evidence from dicodon analysis. *Proc. Natl. Acad. Sci. U.S.A.* 1998. V. 95. № 1. P. 219–223.
4. Fedorov A., Suboch G., Bujakov M., Fedorova L. Analysis of nonuniformity in intron phase distribution. *Nucleic Acids Res.* 1992. V. 20. № 10. P. 2553–2557.
5. Endo T., Fedorov A., de Souza S.J., Gilbert W. Do introns favor or avoid regions of amino acid conservation? *Mol. Biol. Evol.* 2002. V. 19. № 4. P. 521–525.
6. Gilbert W., de Souza W., Long M. Origin of genes. *Proc. Natl. Acad. Sci. U.S.A.* 1997. V. 94. № 15. P. 7698–7703.
7. Rogozin I.B., Carmel L., Csuros M., Koonin E.V. Origin and evolution of spliceosomal introns. *Biol. Direct.* 2012. V. 7. № 1. P. 11.
8. Sakurai A., Fujimori S., Kochiwa H., Kitamura-Abe S., Washio T., Saito R., Carninci P., Hayashizaki Y., Tomita M. On biased distribution of introns in various eukaryotes. *Gene.* 2002. V. 300. № 1–2. P. 89–95.
9. Fedorova L., Fedorov A. Introns in gene evolution. *Genetica*. 2003. V. 118. № 2–3. P. 123–131.
10. Gorlova O., Fedorov A., Logothetis Ch., Amos Ch., Gorlov I. Genes with a large intronic burden show greaterevolutionary conservation on the protein level. *BMC Evolutionary Biology*. 2014. V. 14. № 1. P. 50. doi: 10.1186/1471-2148-14-50.
11. Астахова Т.В., Ройтберг М.А., Цитович И.И., Яковлев В.В. Закономерности, связанные с распределением длин интронов. *Математическая биология и биоинформатика.* 2014. V. 9. № 2. P. 482–490. doi: 10.17537/2014.9.482.
12. Ruvinsky A., Ward W. Intron Framing Exonic Nucleotides: A Compromise Between Protein Coding and Splicing Constraints. *Open EV. J.* 2008. V. 2. P. 7–12.
13. Uetani N., Kato K., Ogura H., Mizuno K., Kawano K., Mikoshiba K., Yakura H., Asano M., Iwakura Y. Impaired learning with enhanced hippocampal long-term potentiation in PTPdelta-deficient mice. *EMBO J.* 2000. V. 19. № 12. P. 2775–2785.
14. Koboldt D.C., Fulton R.S., McLellan M.D., Schmidt H., Kalicki-Veizer J., McMichael J.F., Fulton L.L., Dooling D.J., Ding L., Mardis E.R. at al. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012. V. 490. № 7418. P. 61–70.
15. Ding L., Getz G., Wheeler D.A., Mardis E.R., McLellan M.D., Cibulskis K., Sougnez C., Greulich H., Muzny D.M., Morgan M.B. at al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*. 2008. V. 455. № 7216. P. 1069–1075.
16. Clark O., Schmidt F., Coles C.H., Tchetchelnitski V., Stoker A.W. Functional analysis of the putative tumor suppressor PTPRD in neuroblastoma cells. *Cancer Invest.* 2012. V. 30. № 5. P. 422–432.
17. Litwack E.D., Babey R., Buser R., Gesemann M., O’Leary D.D.M. Identification and characterization of two novel brain-derived immunoglobulin superfamily members with a unique structural organization. *Mol. Cell. Neurosci.* 2004. V. 25. № 2. P. 263–274.
18. Van den Oord E.J.C.G., Kuo P.-H., Hartmann A.M., Webb B.T., Möller H.-J., Hettema J.M., Giegling I., Bukszár J., Rujescu D. Genomewide association analysis followed by a replication study implicates a novel candidate gene for neuroticism. *Arch. Gen. Psychiatry*. 2008. V. 65. № 9. P. 1062–1071.
19. Wu Q., Yu B., Chen Y., Shao Y., Zhang J., Zhong Q., Peng X., Yang H., Hu X., Chen B., Guan M., Wan J., Zhang W. Single-nucleotide polymorphisms of MAMDC1 are associated with rash and photosensitivity, but not disease risk, of systemic lupus erythematosus in Chinese mainland population. *Clin. Rheumatol.* 2011. V. 30. № 10. P. 1373–1378.
20. Minet A.D., Rubin B.P., Tucker R.P., Baumgartner S., Chiquet-Ehrismann R. Teneurin-1, a vertebrate homologue of the Drosophila pair-rule gene ten-m, is a neuronal protein with a novel type of heparin-binding domain. *J. Cell Sci.* 1999. V. 112. P. 2019–2032.

21. Nunes S.M., Ferralli J., Choi K., Brown-Luedi M., Minet A.D., Chiquet-Ehrismann R. The intracellular domain of teneurin-1 interacts with MBD1 and CAP/ponsin resulting in subcellular codistribution and translocation to the nuclear matrix. *Exp. Cell Res.* 2005. V. 305. № 1. P. 122–132.
22. Mosca T.J., Hong W., Dani V.S., Favaloro V., Luo L. Trans-synaptic Teneurin signalling in neuromuscular synapse organization and target choice. *Nature.* 2012. V. 484. № 7393. P. 237–241.
23. *NCBI Gene Database*. URL: http://www.ncbi.nlm.nih.gov/gene/ (дата обращения: 10.09.2017).
24. *NCBI GenBank Database*. URL: http://www.ncbi.nlm.nih.gov/genbank/ (дата обращения: 10.09.2017).
25. *NCBI RefSeq Database*. URL: http://www.ncbi.nlm.nih.gov/refseq/ (дата обращения: 10.09.2017).
26. Rogozin I.B., Wolf Y.I., Sorokin A.V., Mirkin B.G., Koonin E.V. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* 2003. V. 13. № 17. P. 1512–1517.
27. Burset M., Seledtsov I.A, Solovyev V.V. Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Res.* 2000. V. 28. № 21. P. 4364–4375.
28. Marais G., Nouvellet P., Keightley P.D., Charlesworth B. Intron size and exon evolution in Drosophila. *Genetics.* 2005. V. 170. № 1. P. 481–485.
29. *RepeatMasker Home page*. URL: http://repeatmasker.org (дата обращения: 10.09.2017).

Рукопись поступила в редакцию 07.08.2017.
Дата опубликования 19.09.2017.