УДК 577.21

Тандемные дупликации экзонов расширяют репертуар альтернативного сплайсинга

Т. М. Иванов, Д. Д. Первушин*
Центр наук о жизни, Сколковский институт науки и технологий, Москва, 121205 Россия
*E-mail: d.pervouchine@skoltech.ru
Поступила в редакцию 06.09.2021
Принята к печати 17.01.2022
DOI: 10.32607/actanaturae.11583

РЕФЕРАТ **Тандемные дупликации экзонов, обеспечивающие механизм адаптивной регуляции функций белков, играют важную роль в эволюции эукариотических генов. В недавних исследованиях установлена связь тандемных дупликаций с образованием взаимоисключающих экзонов, т.е. с типом альтернативного сплайсинга, при котором в зрелый транскрипт включается один и только один экзон из группы тандемно расположенных экзонов. С использованием биоинформатических методов нами заново рассмотрена проблема идентификации тандемных дупликаций экзонов в эукариотических генах и показано, что тандемно дуплицированные экзоны широко распространены не только в кодирующих частях генов, но и в нетранслируемых областях. Приведен ряд примеров тандемных дупликаций экзонов, идентифицированы неаннотированные тандемно дуплицированные экзоны, представлены статистические свидетельства их экспрессии с использованием больших панелей экспериментов секвенирования РНК.**

КЛЮЧЕВЫЕ СЛОВА **альтернативный сплайсинг, структура РНК, тандемные дупликации экзонов, РНК-секвенирование.**

СПИСОК СОКРАЩЕНИЙ **CDS – кодирующая последовательность ДНК; UTR – нетранслируемая область.**

ВВЕДЕНИЕ

Основной движущей силой молекулярной эволюции считаются мутации, вносящие изменения в передаваемые из поколения в поколение геномные последовательности. Наиболее частый тип мутаций – однонуклеотидные полиморфизмы, которые затрагивают отдельные нуклеотиды, но не менее важным типом изменений, происходящих с ДНК, являются дупликации. Тандемные геномные дупликации представлены протяженными участками ДНК, обычно длиной более 1000 п.н., которые непосредственно примыкают друг к другу и имеют высокий уровень гомологии [1, 2].

Тандемные геномные дупликации могут затрагивать целые гены, кодирующие или не кодирующие белки, или только части отдельных генов. В последнем случае дупликация приводит к копированию только части нуклеотидной последовательности, влияя таким образом на экзон-интронную структуру [3]. Процесс, при котором дуплицируется один экзон гена или эктопически объединяются два или более экзона из разных генов, получил название перемешивания экзонов (exon shuffling) [4, 5]. Во многих случаях перемешивание экзонов посредством тандемных дупликаций связано с взаимоисключающим выбором экзонов, т.е. с регулируемым типом альтернативного сплайсинга, при котором только один экзон из группы экзонов включается в зрелый транскрипт [6, 7].

Взаимоисключающие экзоны (mutually exclusive exons, MXE) обнаружены во многих генах, например, в генах кадгерина-N (*CadN*) [8, 9], тяжелой цепи миозина (*MHC*) [10], *14-3-3ξ* [11], *srp* [12] и гене множественной лекарственной устойчивости (*MRP*) [13] *Drosophila melanogaster*, в гене фактора транскрипции FOX млекопитающих [14], а также в генах семейства тропомиозина [15]. Возможно, наиболее яркий пример взаимоисключающего сплайсинга, возникшего в результате тандемных дупликаций, представляет ген *DSCAM1 D. melanogaster*, содержащий четыре группы кластеров MXE, которые вместе могут привести к образованию до 38016 различных изоформ белка DSCAM1 [16–21].

Проведенное в 2002 году систематическое исследование дупликаций экзонов и их роли в альтернативном сплайсинге показало, что около 10% генов животных содержат тандемно дуплицированные экзоны, и обнаружило более 2000 неаннотированных MXE-кандидатов путем идентификации гомологии с соседними экзонами или с близлежащими участками ДНК [22]. Однако тандемные дупликации

экзонов могут охватывать также интронные и нетранслируемые области (untranslated regions, UTR), которые не примыкают непосредственно к аннотированным экзонам, а базы данных аннотаций генома значительно расширились с тех пор. В данной работе мы вернулись к поиску гомологичных экзонов в полных последовательностях генов, а также в их геномных окрестностях и обнаружили, что тандемные дупликации экзонов простираются далеко за пределы белоккодирующей части гена и довольно часто встречаются в нетранслируемых областях. Нами получена динамическая картина представленности экзонных дупликаций в зависимости от гомологии нуклеотидных последовательностей и приведен ряд характерных примеров таких дупликаций.

ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

Геномы и аннотации

Геномная последовательность человека (сборка hg19, GRCh37.p13) получена из Genome Reference Consortium [23]. Исчерпывающая аннотация транскриптов генов человека (GENCODE comprehensive gene annotation) версии 19 получена из базы данных GENCODE [24]. Геномы *D. melanogaster* (BDGP Release 6, dm6) и *Caenorhabditis elegans* (WBcel235, ce11) получены из базы данных UCSC Genome Browser [25]. Аннотации транскриптов *D. melanogaster* получены из базы данных FlyBase релиз dmel_r6.32 [26]. Аннотации транскриптов *C. elegans* релиз 104 получены из базы данных Wormbase [27]. Аннотации транскриптов RefSeq получены из базы данных NCBI RefSeq database [28]. Рассматривали только экзоны и транскрипты белоккодирующих генов. Число экзонов, уникальных для человека, *D. melanogaster* и *C. elegans,* составляло 329983, 83276 и 172984 соответственно.

Поиск гомологии экзонов

Для идентификации тандемных дупликаций экзонов использовали программу exonerate EMBL-EBI [29]. Нуклеотидную последовательность каждого экзона выравнивали с последовательностью его гена, которая удлинялась в обоих направлениях на 15% длины гена. Поскольку гены человека в среднем намного длиннее генов дрозофилы, для поиска мы использовали процент длины гена, а не окно фиксированной длины. Порог отсечения 15% выбран потому, что расстояние от гена до ближайшего соседа не превосходит 15% длины гена для половины генов дрозофилы. Программу запускали в исчерпывающем (exhaustive) режиме для получения точного выравнивания. Минимальный процент отсечения идентичности установлен равным 50%, однако exonerate не обнаружил гомологию последовательностей ниже 57%. Последовательности выравниваний были извлечены с помощью программы getfasta из пакета bedtools [30]. Выравнивания последовательностей были организованы в таблицу bed12, в которой каждая строка соответствует одному выравниванию, в дальнейшем называемому парой запрос-мишень, включая выравнивания экзона на себя. После удаления выравниваний экзона на самого себя таблица содержала 116320, 5244 и 5605 пар запрос-мишень генов человека, *D. melanogaster* и *C. elegans* соответственно.

Процедура фильтрации для пар запрос-мишень

Для идентификации неаннотированных тандемных дупликаций экзонов мы отфильтровали таблицу пар запрос-мишень с помощью программы bedtools intersect следующим образом. Мы удалили пары запрос-мишень, в которых последовательность мишени пересекает по крайней мере один аннотированный экзон более чем по 5% ее длины. Кроме того, мы удалили пары запрос-мишень, в которых последовательность мишени пересекает по крайней мере один аннотированный геномный повтор или последовательность низкой сложности более чем по 10% ее длины в соответствии с треками аннотированных повторов из браузера генома UCSC [25].

Данные РНК-секвенирования

Мы использовали данные РНК-секвенирования 6625 образцов консорциума Genotype-Tissue Expression Project (GTEx) версии v7 [31]. Короткие прочтения (риды) были выравнены на геном человека с помощью картировщика STAR v2.4.2a [32]. Прочтения с разрывами (split reads, сплит-чтения), поддерживающие экзон-экзонные соединения, были извлечены с помощью пакета программ IPSA с настройками по умолчанию [33] (порог энтропии Шеннона 1.5 бит). Учитывали только сплит-чтения с каноническими динуклеотидами GT/AG. Уникально картированные прочтения выбраны на основе наличия метки (тега) NH:1 из файлов в формате BAM. Среднее покрытие ридами и показатели консервативности PhastCons рассчитывали с использованием программного пакета Deeptools [34].

РЕЗУЛЬТАТЫ

Коэффициент дупликации

Для обнаружения экзонных дупликаций мы использовали самые большие на сегодняшний день наборы данных по аннотации экзонов, включая базы GENCODE [35] и RefSeq [28]. Проведен поиск гомологии последовательностей для каждого экзона

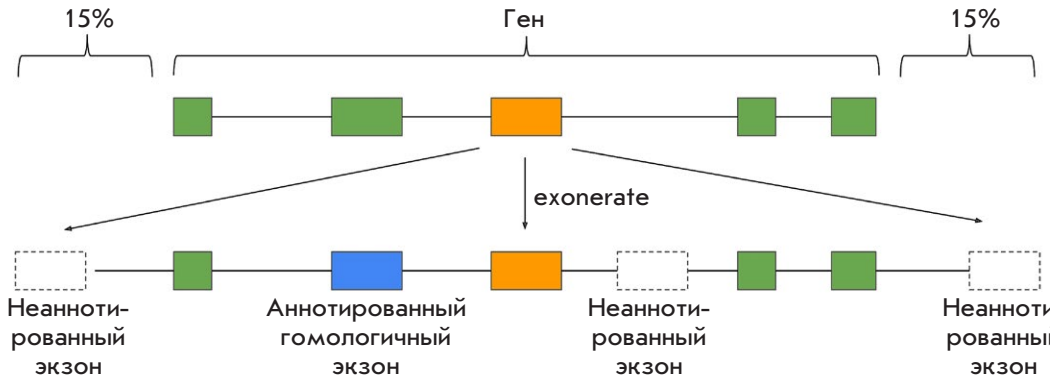

Рис. 1. Схема поиска тандемных дупликаций экзонов. Нуклеотидная последовательность каждого экзона выравнивается на нуклеотидную последовательность гена, расширенную в обе стороны на 15% длины

в расширенной нуклеотидной последовательности его гена с помощью программы exonerate [29]. В дальнейшем мы будем называть аннотированные экзоны запросными последовательностями, или просто запросами, а их гомологи, обнаруженные с помощью exonerate, – мишенями (*рис. 1*). Каждая пара запрос-мишень характеризуется ковариатами, относящимися к запросу (например, местоположение в пределах CDS или UTR), ковариатами, относящимися к мишени (например, доля длины мишени, перекрывающаяся с аннотированными экзонами), и процентом гомологии между запросом и мишенью. Поскольку многие экзоны подвергаются альтернативному сплайсингу и вносят таким образом вклад в качестве перекрывающихся областей в наборы аннотаций экзонов, мы ввели показатель коэффициента дупликации (nucleotide increase ratio, NIR), который определяется как отношение суммарного числа нуклеотидов, покрытых мишенями, к суммарному числу нуклеотидов, покрытых запросами с заданным или более высоким процентом гомологии нуклеотидных последовательностей. По построению NIR всегда больше 1, поскольку каждый запрос служит своей собственной мишенью со 100% идентичностью последовательности. NIR можно вычислить для всех экзонов, а также только для экзонов в кодирующих или только в нетранслируемых областях. Таблицы, перечисляющие пары запрос-мишень, доступны в онлайн репозитории https://zenodo.org/record/5474863.

Как и ожидалось, значения NIR уменьшаются с увеличением порога на гомологию последовательностей (*рис. 2А*). Несмотря на использование 50% порога на гомологию последовательностей, exonerate не обнаружил гомологичных пар запрос-мишень со степенью гомологии ниже 57%. Выбрав порог отсечения по гомологии 80%, что соответствует середине интервала 60–100%, в котором расположены степени гомологии всех мишеней, мы обнаружили, что около 2% экзонных нуклеотидов человека в белоккодирующих областях подвергаются дупликациям при пороге отсечения по гомологии последовательностей в 80% или более, в то время как только 0.08% экзонных нуклеотидов *D. melanogaster* и 0.06% экзонных нуклеотидов *C. elegans* подвергаются дупликациям. Очевидно, это связано с тем, что неаннотированные мишени экзонных нуклеотидов принадлежат интронным областям, а интроны человека намного длиннее, чем интроны *D. melanogaster* и *C. elegans*. Примечательно, что при рассмотрении только экзонов, которые расположены в UTR, почти 15% экзонных нуклеотидов человека подвергаются дупликациям при пороге отсечения по гомологии последовательностей в 80% или более (*рис. 2А*), а соответствующие пропорции для *D. melanogaster* и *C. elegans* составляют 0.3 и 0.2%, что указывает на значительно более высокую частоту дупликаций экзонов в UTR.

Затем мы выяснили, не являются ли одни гены более склонными к тандемным дупликациям экзонов, чем другие. Чтобы ответить на этот вопрос, мы вычислили значения NIR для каждого аннотированного гена отдельно и построили частотные распределения NIR (*рис. 2Б*). Частоты значений NIR подчиняются степенному закону распределения, о чем свидетельствует близкая к линейной зависимость логарифма частоты от логарифма значения NIR со значительным отклонением в сторону более высоких частот для больших значений NIR в некоторых генах. Гены человека с отклоняющимися значениями NIR для экзонов CDS включают в себя *CAMK1D* (кальций/кальмодулин-зависимая протеинкиназа), *CLYBL* (цитрамалил-КоА-лиаза) и *NBPF20* (neuroblastoma breakpoint family), однако некоторые гены человека также имеют отклоняющиеся значения NIR и для нетранслируемых областей, например *OBSCN* (обскурин) и *NEB* (небулин). Генами с заметными отклонениями по числу тандемных дупликаций у *D. melanogaster* были *dpy*, *hydra* и *heph*.

Различия в склонности к тандемным дупликациям между генами с высоким значением NIR по сравнению с остальными генами могут быть обусловлены различиями в длинах экзонов. Чтобы выяснить так ли это, мы сравнили значения NIR в группах экзонов разной длины, используя 10 интервалов одина-

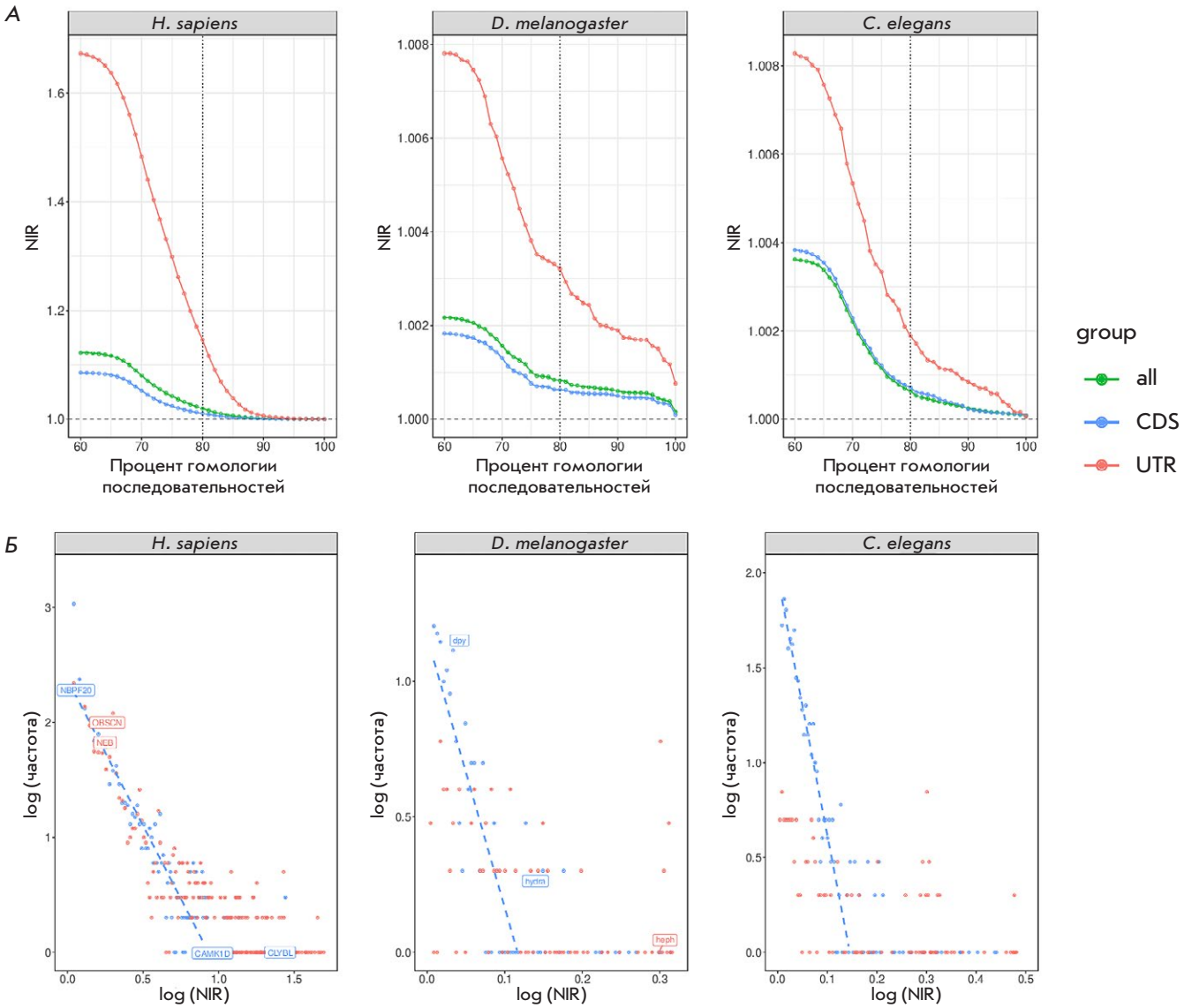

Рис. 2. Коэффициенты дупликации (NIR) в геномах человека, *D. melanogaster* и *C. elegans* как функция процента гомологии нуклеотидных последовательностей запрос-мишень (*А*) и распределение частот значений NIR в генах человека, *D. melanogaster* и *C. elegans* при пороге гомологии нуклеотидных последовательностей запрос-мишень 80% (*Б*). Приведены имена генов, существенно отклоняющихся от степенной зависимости

ковой ширины. Оказалось, что значения NIR уменьшаются примерно в 4 раза при увеличении длины экзона с 20 до 220 нуклеотидов, что указывает на то, что более длинные экзоны не вносят больший вклад в значение NIR. Действительно, чем длиннее экзон, тем меньше вероятность найти его гомолог при условии сохранения порога отсечения 80% на гомологию последовательностей. Кроме того, средняя длина экзона у 200 генов с наивысшими значениями NIR статистически значимо не отличается от средней длины экзона в генеральной совокупности всех экзонов (критерий Вилкоксона, *P*-значение = 0.2). Таким образом, длины экзонов существенно не влияют на склонность к тандемным дупликациям. Анализ 200 генов с наиболее высоким значением NIR обнаружил преобладание онтологических категорий, относящихся к клеточной адгезии и развитию нервной системы (биологическая функция), связыванию ионов и активности рецепторов (молекулярная функция), и мембранной локализации (клеточные компартменты).

Для дальнейшего исследования структуры экзонных дупликаций в этих генах создан трек-хаб для геномного браузера UCSC в качестве инструмента визуализации всех пар запрос-мишень. В качестве положительного контроля мы подтвердили, что наша процедура успешно идентифицировала кластеры тандемно дуплицированных экзонов в генах, в которых такие кластеры были известны ранее

[10–13] (данные не приведены). Чтобы обнаружить новые неаннотированные тандемные дупликации экзонов, мы исключили из рассмотрения пары запрос-мишень, которые перекрывают любой аннотированный экзон, а также отфильтровали мишени, которые пересекают аннотированные повторы или участки ДНК низкой сложности, поскольку последние могут содержать экзоны, возникшие по другому механизму, например, через экзонизацию транспозонов [36]. Мы также исследовали статистические свидетельства экспрессии вновь обнаруженных экзонов, используя данные секвенирования РНК из проекта Genotype Tissue Expression Project [31], а именно, их покрытие короткими прочтениями и поддержку экзон-экзонных соединений.

ПРИМЕРЫ

Обскурин (*OBSCN*)

Один из генов человека, подверженных тандемным дупликациям экзонов, – обскурин (*OBSCN*). Этот ген состоит из более 150000 п.н. и содержит более 80 экзонов [37]. Белок, кодируемый геном *OBSCN*, принадлежит к семейству гигантских сакромерных сигнальных белков, в которое входят также титин и небулин [38]. *OBSCN* высоко экспрессируется в сердце (RPKM 8.6), простате (RPKM 2.9) и других тканях [31].

Наш анализ показывает, что подавляющее большинство экзонов обскурина гомологичны друг другу и имеют одинаковую длину, что указывает на их образование в результате тандемной дупликации (*рис. 3*). Наличие повторов в промежуточных интронах свидетельствует об их возникновении в ходе нескольких раундов геномных дупликаций, вероятно, в результате негомологичной рекомбинации. Примечательно, что один из промежуточных интронов содержит область, гомологичную другим экзонам, но не аннотированную как экзон (*рис. 3*, показана голубым цветом). Функциональность этой области подтверждается высокой степенью консервативности (PhastCons) и наличием сплит-ридов, поддерживающих экзон-экзонные соединения. Интересно, что тот же самый промежуточный интрон содержит и другую область с высоким уровнем консервативности (PhastCons), которая также включает сплит-риды, поддерживающие экзон-экзонные соединения. Однако эта область обладает меньшей степенью гомологии с другими экзонами (процент идентичности последовательностей 62.4% против 78.9% у других областей).

UDP-глюкуронозилтрансфераза (*UGT1A*)

Ген *UGT1A* человека кодирует UDP-глюкуронозилтрансферазу и содержит 13 уникальных альтернативных начальных экзонов, за которыми следуют четыре конститутивных экзона. Этот ген ассоциирован с такими заболеваниями, как синдром Гилберта [39] и синдром Криглера–Найяра [40]. Каждый начальный экзон регулируется собственным промотором и кодирует сайт связывания субстрата, в результате чего образуются белки с разными N-концами и идентичными C-концами. Наш анализ показывает, что вариабельные начальные экзоны этих генов гомологичны друг другу (*рис. 4*), что дает основания предположить возможность их происхождения в результате серии тандемных дупликаций. В 5'-UTR этого гена содержится консервативная область, которая гомологична начальным экзонам, но не аннотирована как экзон (*рис. 4*). Следует отметить, что все начальные экзоны этого гена включаются в зрелый транскрипт взаимоисключающим образом.

Примеры тандемных дупликаций в нетранслируемых областях генов *D. melanogaster*

В качестве двух интересных примеров тандемных дупликаций экзонов в нетранслируемых областях генов *D. melanogaster* можно привести гены *hydra* (*рис. 5А*) и *pip* (*рис. 5Б*). Ген *hydra* содержит девять гомологичных начальных экзонов, которые сплайсируются взаимоисключающим образом, тогда как ген *pip* имеет восемь тандемно повторяющихся гомологичных кластеров взаимоисключающих терминальных экзонов. Показано, что начальный экзон гена *hydra* подвергся рекуррентным дупликациям, и семь из этих альтернативных начальных экзонов фланкированы на своей 3'-стороне транспозоном DINE-1 [41]. По крайней мере четыре из девяти дуплицированных начальных экзонов могут функционировать как альтернативные сайты начала транскрипции [41]. Однако 3'-нетранслируемая область гена *pip*, который кодирует сульфотрансферазу и вносит вклад в формирование и полярность дорсально-вентральной оси эмбриона, изучена гораздо менее полно. Недавно показали, что взаимоисключающее использование экзонов в 3'-UTR этого гена зависит от конкурирующих вторичных структур РНК [42].

Свидетельства экспрессии тандемно дуплицированных экзонов по данным секвенирования РНК

Экспрессию тандемных дупликаций экзонов оценивали с использованием данных секвенирования РНК, рассматривали пары запрос-мишень в генах человека, в которых мишень не пересекается ни с аннотированными экзонами, ни с повторами, оставшиеся мишени объединяли с помощью программы bedtools merge.

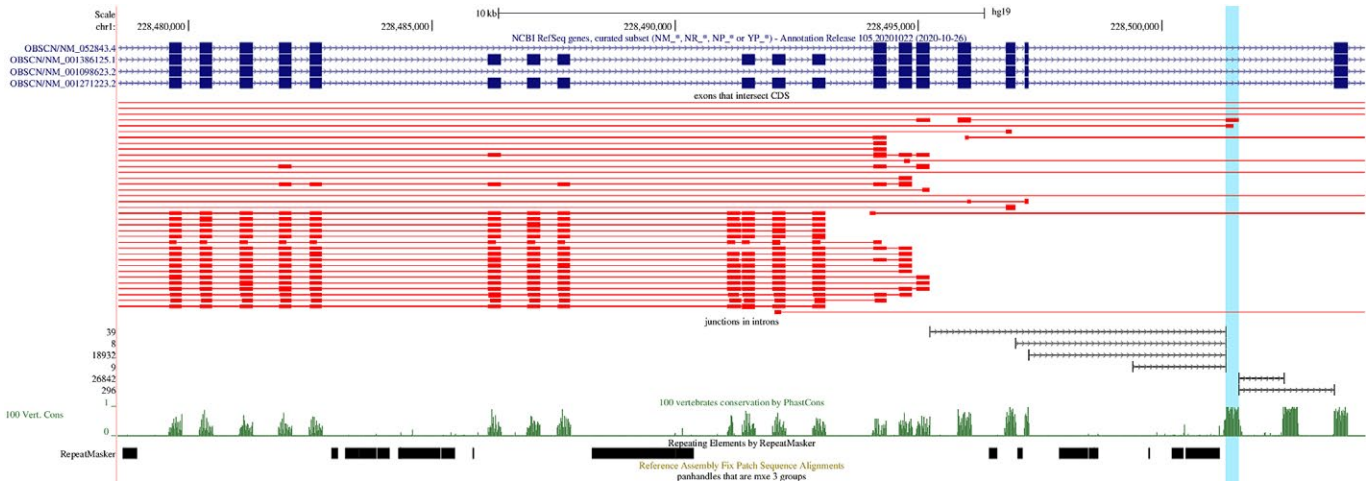

Рис. 3. Диаграмма тандемных экзонных дупликаций в гене обскурина (*OBSCN*). Темно-синим цветом показаны аннотированные транскрипты (GENCODE и RefSeq). Красным показаны пары запрос-мишень; запросы показаны толстыми прямоугольниками, а мишени – тонкими. Следующий трек показывает поддержку экзонных границ сплит-ридами. Значения степени консервативности (PhastCons) показаны зеленым цветом

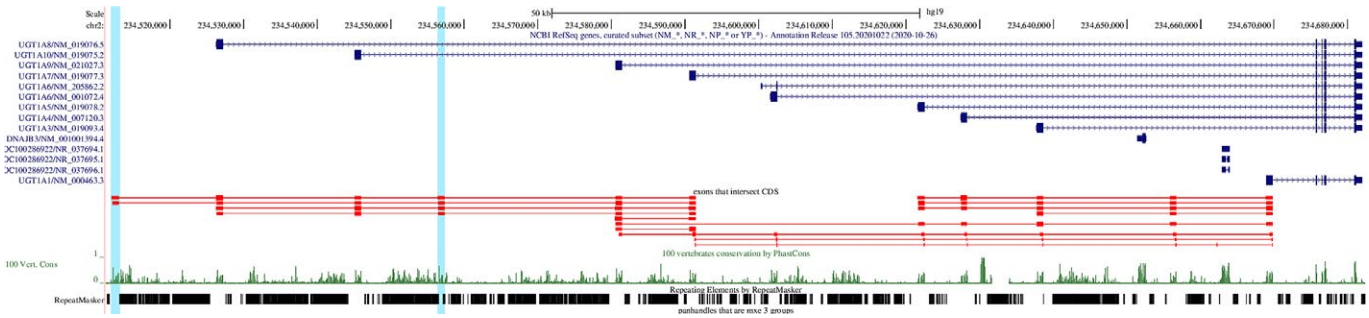

Рис. 4. Диаграмма тандемных экзонных дупликаций в гене UDP-глюкуронозилтрансферазы *UGT1A*; обозначения, как на *рис. 3*

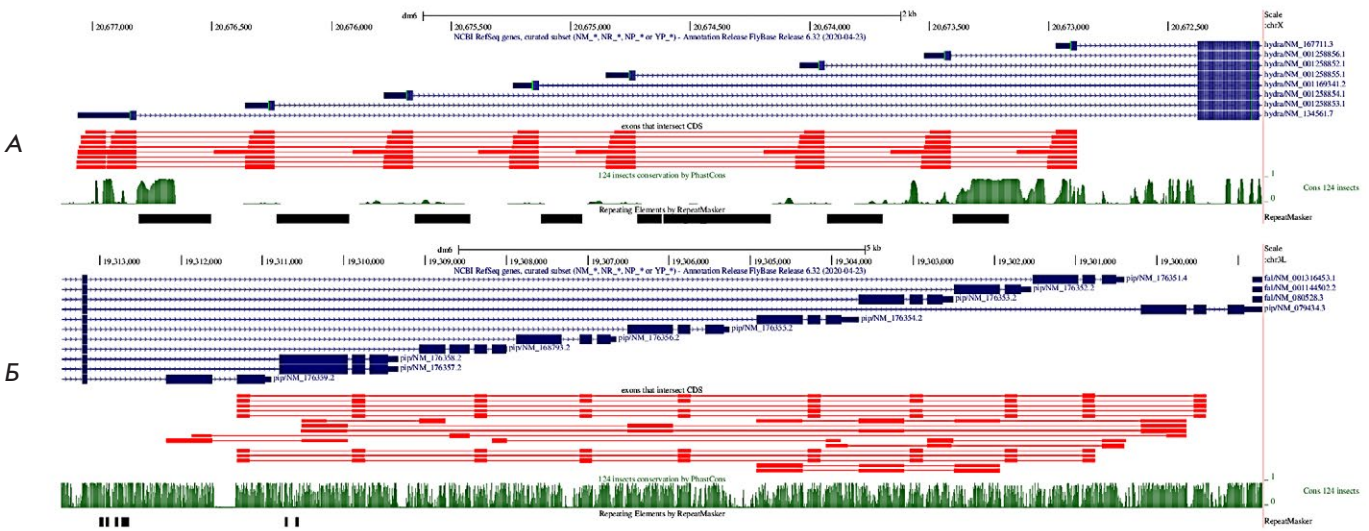

Рис. 5. Диаграмма тандемных экзонных дупликаций в генах *hydra* (*А*) и *pip* (*Б*); обозначения, как на *рис. 3*

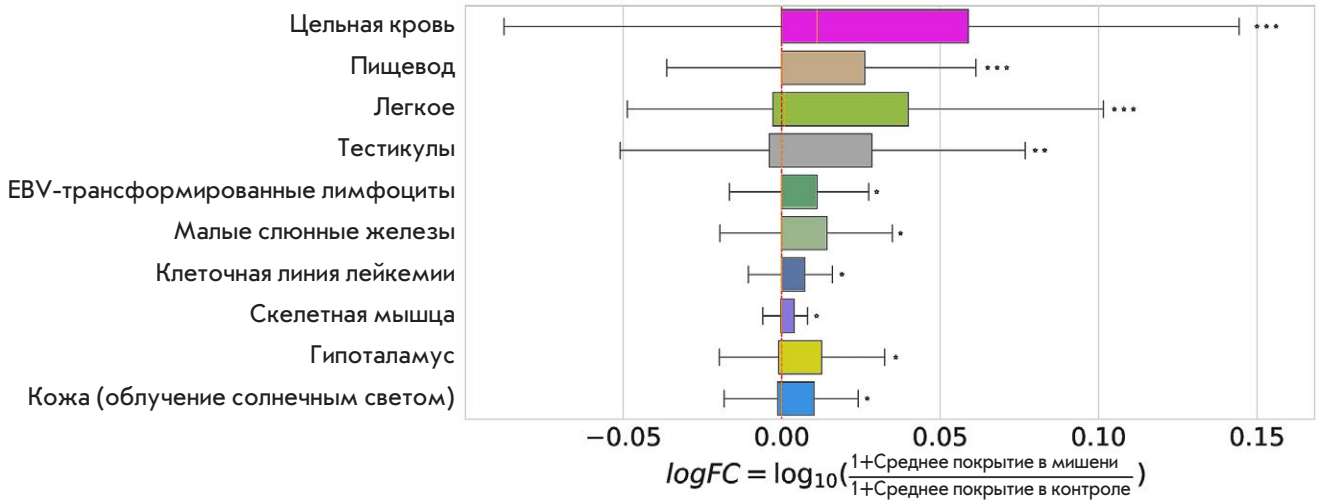

Рис. 6. Распределение значений метрики lg FC_i в тканях GTEx для мишеней с по крайней мере 80% гомологией нуклеотидной последовательности. Цвета тканей соответствуют стандартной цветовой палитре консорциума GTEx [31]. Показаны только те ткани, которые имеют значимое отклонение метрики lg FC_i от нуля (в порядке уменьшения статистической значимости). Уровни значимости вычислены с помощью знакового рангового критерия Вилкоксона после поправки Бенджамини–Хохберга на множественное тестирование

Эта процедура дала 4027 интронных мишеней, каждую из которых случайным образом сопоставляли с контрольной областью такой же длины, расположенной на 30 н. в сторону 5'- или 3'-конца гена.

Одна из основных проблем оценки экспрессии тандемных дупликаций экзонов с использованием данных секвенирования РНК заключается в том, что в случае высокой идентичности нуклеотидной последовательности запроса и мишени короткие чтения одинаково хорошо выравниваются как с последовательностью запроса, так и с последовательностью мишени. Поэтому мы исключили из анализа все короткие чтения, которые картировались более чем на одну позицию в геноме, и вычислили среднее покрытие ридами каждой мишени и соответствующей контрольной области в каждом из 53 транскриптомов тканей из проекта Genotype-Tissue Expression (GTEx) [31], используя только однозначные картирования. Затем мы вычислили показатель lg FC_i = lg(1 + $target_i$) - lg(1 + $control_i$), где $target_i$ – среднее покрытие мишени в ткани i, а $control_i$ – среднее покрытие контрольного региона в ткани i. Ткани с недостаточным количеством значений lg FC_i (мочевой пузырь, эндоцервикс и эктоцервикс шейки матки) были исключены из дальнейшего анализа. В группе мишеней, которые обладали по меньшей мере 80% гомологией нуклеотидной последовательности с запросом, мы наблюдали значительное положительное отклонение метрики lg FC_i от нуля (знаковый ранговый критерий Вилкоксона), которое в некоторых тканях оставалось значимым после коррекции Бенджамини–Хохберга на множественное тестирование, например в крови, пищеводе, легких, тестикулах, мышцах, мозге, а также в некоторых трансформированных клетках (*рис. 6*). Следует отметить, что тест Вилкоксона обнаружил статистически значимые отклонения от нуля даже в тех случаях, когда медиана выборки близка к нулю, что указывает на преобладание больших положительных значений в выборке разностей. Мы также наблюдали увеличение количества сплит-ридов, поддерживающих экзон-экзонные соединения в тандемно дуплицированных экзонах с более высокой идентичностью нуклеотидных последовательностей (*рис. 7*). Эти результаты показывают, что по крайней мере некоторые из неаннотированных тандемно дуплицированных экзонов действительно могут экспрессироваться, причем тканеспецифичным образом.

В заключение мы вычислили разницу между средними показателями степени консервативности PhastCons [43], полученной из множественного выравнивания геномов 100 видов позвоночных между мишенями и соответствующими им контрольными регионами. Мишени оказались в среднем более эволюционно консервативными, чем контрольные области (знаковый ранговый критерий Вилкоксона, $P = 0.009$), что также указывает на их возможную функциональность.

ОБСУЖДЕНИЕ

Интересное наблюдение, сделанное в этой работе, заключается в том, что тандемные дупликации экзонов преобладают не только в кодирующих, но также и в нетранслируемых областях эукарио-

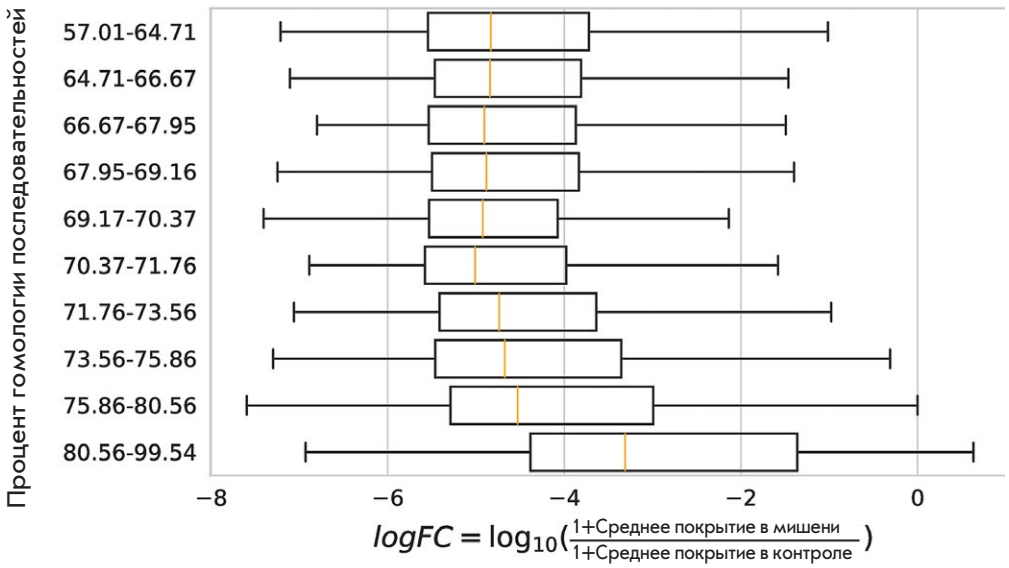

Рис. 7. Распределение значений метрики lg FC_i для сплит-ридов. Метрика lg FC_i вычислена как lg $FC_i = lg(1 + target_i) - lg(1 + control_i)$, где $target_i$ ($control_i$ соответственно) равно суммарному числу уникально картированных сплит-ридов, поддерживающих экзонные границы мишени (контроля соответственно) в ткани i

тических генов. Более того, они, по-видимому, связаны с взаимоисключающим выбором тандемно дуплицированных начальных и конечных экзонов. Недавнее исследование показало, что регуляторный механизм, лежащий в основе взаимоисключающего выбора 3'-вариабельных областей в пре-мРНК гена *PGRP-LC* у *D. melanogaster*, задействует конкурирующие структуры РНК [42]. Эти структуры РНК совместно регулируют отбор 3'-UTR посредством активации проксимального 3'-сайта сплайсинга и одновременного подавления интрон-проксимального 5'-сайта сплайсинга вместе со стерической конкуренцией за спаривание РНК [42]. Сходная регуляторная программа действует и в 3'-вариабельных областях генов *D. melanogaster CG42235* и *pip*. Это наблюдение заставляет задуматься о том, не могут ли тандемные дупликации экзонов в нетранслируемых областях контролироваться конкурирующими структурами РНК некоторым общим образом.

В недавней работе мы предложили эволюционный механизм образования конкурирующих структур РНК, связанных с взаимоисключающим сплайсингом через геномные дупликации, которые затрагивают не только экзоны, но и соседние с ними интроны со шпилечными структурами РНК [44]. Согласно этой гипотезе, дупликация одной из двух цепей интронной шпилечной структуры автоматически приводит к образованию двух последовательностей, которые конкурируют за спаривание оснований с третьей последовательностью. Это соответствует часто наблюдаемому механизму регуляции сплайсинга MXE через конкурирующие структуры РНК [13–15, 21]. В частности, из этой модели вытекает, что взаимоисключающий сплайсинг, опосредованный конкурирующими структурами РНК, является неизбежным следствием тандемных дупликаций. Принимая во внимание большое количество консервативных комплементарных областей в нетранслируемых областях генов человека [45], вероятным представляется то, что тандемные дупликации экзонов внутри UTR также могут автоматически генерировать конкурирующие структуры РНК, приводящие к взаимоисключающему включению экзонов.

ЗАКЛЮЧЕНИЕ

Тандемные дупликации экзонов широко представлены не только в кодирующих частях, но и в нетранслируемых областях эукариотических генов. Вовлечены ли конкурирующие РНК-структуры в регуляцию взаимоисключающего сплайсинга этих экзонов и могут ли они образовываться как побочный продукт тандемных геномных дупликаций? Ответ на этот вопрос остается открытым. ●

Авторы выражают благодарность М.А. Калининой, Д.А. Скворцову и О.А. Донцовой за содержательные обсуждения.

Работа поддержана грантами РФФИ № 19-34-90174 и 18-29-13020-МК. Работа по анализу данных РНК-секвенирования поддержана грантом 21-64-00006 Российского научного фонда.

Авторы заявляют, что у них нет конкурирующих интересов.

Д.П. разработал план и руководил исследованием; Т.И. провел анализ данных. Оба автора участвовали в написании текста статьи, прочитали и одобрили его.

СПИСОК ЛИТЕРАТУРЫ

1. Emanuel B.S., Shaikh T.H. // Nat. Rev. Genet. 2001. V. 2. № 10. P. 791–800.
2. Mehan M.R., Freimer N.B., Ophoff R.A. // Hum. Genomics. 2004. V. 1. № 5. P. 335–344.
3. Ma M.Y., Lan X.R., Niu D.K. // Peer J. 2016. V. 4. P. e2272.
4. Kolkman J.A., Stemmer W.P. // Nat. Biotechnol. 2001. V. 19. № 5. P. 423–428.
5. Patthy L. // Gene. 1999. V. 238. № 1. P. 103–114.
6. Suyama M. // Bioinformatics. 2013. V. 29. № 17. P. 2084–2087.
7. Kondrashov F.A., Koonin E.V. // Hum. Mol. Genet. 2001. V. 10. № 23. P. 2661–2669.
8. Nern A., Nguyen L.V., Herman T., Prakash S., Clandinin T.R., Zipursky S.L. // Proc. Natl. Acad. Sci. USA. 2005. V. 102. № 36. P. 12944–12949.
9. Ting C.Y., Yonekura S., Chung P., Hsu S.N., Robertson H.M., Chiba A., Lee C.H. // Development. 2005. V. 132. № 5. P. 953–963.
10. George E.L., Ober M.B., Emerson C.P. // Mol. Cell. Biol. 1989. V. 9. № 7. P. 2957–2974.
11.Messaritou G., Leptourgidou F., Franco M., Skoulakis E.M. // FEBS Lett. 2009. V. 583. № 17. P. 2934–2938.
12. Waltzer L., Bataillé L., Peyrefitte S., Haenlin M. // EMBO J. 2002. V. 21. № 20. P. 5477–5486.
13. Grailles M., Brey P.T., Roth C.W. // Gene. 2003. V. 307. P. 41–50.
14. Gabut M., Samavarchi-Tehrani P., Wang X., Slobodeniuc V., O'Hanlon D., Sung H.K., Alvarez M., Talukder S., Pan Q., Mazzoni E.O., et al. // Cell. 2011. V. 147. № 1. P. 132–146.
15. Gooding C., Smith C.W. // Adv. Exp. Med. Biol. 2008. V. 644. P. 27–42.
16. Chen B.E., Kondo M., Garnier A., Watson F.L., Püettmann-Holgado R., Lamar D.R., Schmucker D. // Cell. 2006. V. 125. № 3. P. 607–620.
17. He H., Kise Y., Izadifar A., Urwyler O., Ayaz D., Parthasarthy A., Yan B., Erfurth M.L., Dascenco D., Schmucker D. // Science. 2014. V. 344. № 6188. P. 1182–1186.
18. Hughes M.E., Bortnick R., Tsubouchi A., Bäumer P., Kondo M., Uemura T., Schmucker D. // Neuron. 2007. V. 54. № 3. P. 417–427.
19. Hummel T., Vasconcelos M.L., Clemens J.C., Fishilevich Y., Vosshall L.B., Zipursky S.L. // Neuron. 2003. V. 37. № 2. P. 221–231.
20. Matthews B.J., Kim M.E., Flanagan J.J., Hattori D., Clemens J.C., Zipursky S.L., Grueber W.B. // Cell. 2007. V. 129. № 3. P. 593–604.
21. Soba P., Zhu S., Emoto K., Younger S., Yang S.J., Yu H.H., Lee T., Jan L.Y., Jan Y.N. // Neuron. 2007. V. 54. № 3. P. 403–416.
22.Letunic I., Copley R.R., Bork P. // Hum. Mol. Genet. 2002. V. 11. № 13. P. 1561–1567.
23. Church D.M., Schneider V.A., Graves T., Auger K., Cunningham F., Bouk N., Chen H.C., Agarwala R., McLaren W.M., Ritchie G.R., et al. // PLoS Biol. 2011. V. 9. № 7. P. e1001091.
24. Harrow J., Frankish A., Gonzalez J.M., Tapanari E., Diekhans M., Kokocinski F., Aken B.L., Barrell D., Zadissa A., Searle S., et al. // Genome Res. 2012. V. 22. № 9. P. 1760–1774.
25. Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M., Haussler D. // Genome Res. 2002. V. 12. № 6. P. 996–1006.
26. Marygold S.J., Crosby M.A., Goodman J.L. // Meth. Mol. Biol. 2016. V. 1478. P. 1–31.
27. Harris T.W., Arnaboldi V., Cain S., Chan J., Chen W.J., Cho J., Davis P., Gao S., Grove C.A., Kishore R., et al. // Nucl. Acids Res. 2020. V. 48. № D1. P. D762–D767.
28. O'Leary N.A., Wright M.W., Brister J.R., Ciufo S., Haddad D., McVeigh R., Rajput B., Robbertse B., Smith-White B., Ako-Adjei D., et al. // Nucl. Acids Res. 2016. V. 44. № D1. P. D733– D745.
29. Slater G.S., Birney E. // BMC Bioinformatics. 2005. V. 6. P. 31.
30. Quinlan A.R., Hall I.M. // Bioinformatics. 2010. V. 26. № 6. P. 841–842.
31. Mele M., Ferreira P.G., Reverter F., DeLuca D.S., Monlong J., Sammeth M., Young T.R., Goldmann J.M., Pervouchine D.D., Sullivan T.J., et al. // Science. 2015. V. 348. № 6235. P. 660–665.
32. Dobin A., Davis C.A., Schlesinger F., Drenkow J., Zaleski C., Jha S., Batut P., Chaisson M., Gingeras T.R. // Bioinformatics. 2013. V. 29. № 1. P. 15–21.
33.Pervouchine D.D., Knowles D.G.., Guigó R. // Bioinformatics. 2013. V. 29. № 2. P. 273–274.
34.Ramírez F., Dündar F., Diehl S., Grüning B.A., Manke T. // Nucl. Acids Res. 2014. V. 42 (Web Server issue). P. W187–191.
35. Frankish A., Diekhans M., Ferreira A.M., Johnson R., Jungreis I., Loveland J., Mudge J.M., Sisu C., Wright J., Armstrong J., et al. // Nucl. Acids Res. 2019. V. 47. № D1. P. D766–D773.
36. Schmitz J., Brosius J. // Biochimie. 2011. V. 93. № 11. P. 1928–1934.
37. Fukuzawa A., Idowu S., Gautel M. // J. Muscle Res. Cell. Motil. 2005. V. 26. № 6–8. P. 427–434.
38. Kontrogianni-Konstantopoulos A., Bloch R.J. // J. Muscle Res. Cell. Motil. 2005. V. 26. № 6–8. P. 419–426.
39. Landerer S., Kalthoff S., Paulusch S., Strassburg C.P. // Sci. Rep. 2020. V. 10. № 1. P. 8689.
40.Strassburg C.P., Kalthoff S., Ehmer U. // Crit. Rev. Clin. Lab. Sci. 2008. V. 45. № 6. P. 485–530.
41. Chen S.T., Cheng H.C., Barbash D.A., Yang H.P. // PLoS Genet. 2007. V. 3. № 7. P. e107.
42. Pan H., Shi Y., Chen S., Yang Y., Yue Y., Zhan L., Dai L., Dong H., Hong W., Shi F., et al. // RNA. 2018. V. 24. № 11. P. 1466–1480.
43. Siepel A., Bejerano G., Pedersen J.S., Hinrichs A.S., Hou M., Rosenbloom K., Clawson H., Spieth J., Hillier L.W., Richards S., et al. // Genome Res. 2005. V. 15. № 8. P. 1034–1050.
44. Ivanov T.M., Pervouchine D.D. // Genes (Basel). 2018. V. 9. № 7. P. 356.
45. Kalmykova S., Kalinina M., Denisov S., Mironov A., Skvortsov D., Guigo R., Pervouchine D. // Nat. Commun. 2021. V. 12. № 1. P. 2300.