Автономная некоммерческая образовательная организация высшего образования “Сколковский Институт Науки и Технологии”

На правах рукописи

Иванов Тимофей Михайлович

Альтернативный сплайсинг тандемно дуплицированных экзонов

Специальность 1.5.8 —
«математическая биология, биоинформатика»

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
кандидат физико-математических наук
Первушин Дмитрий Давидович

Москва — 2022

Оглавление

Стр.

Стр.

Стр.

Введение

Основной движущей силой молекулярной эволюции являются мутационные процессы, вносящие передаваемые из поколения в поколение изменения в генетический материал. Наиболее частым и наиболее хорошо изученным типом мутаций являются однонуклеотидные полиморфизмы, т.е. мутации, затрагивающие отдельные нуклеотиды, однако не менее важны и другие виды изменений в последовательности ДНК, такие как, например, дупликации. Как следует из названия, при дупликации определенный участок ДНК оказывается удвоенным. Дупликации возникают благодаря нескольким молекулярным механизмам: негомологичной рекомбинации, ошибкам при репликации ДНК, связанным с диссоциацией и реассоциацией ДНК полимеразы в неправильном положении на ДНК, и ретротранспозиции чужеродного генетического материала в ДНК хозяина. Эти механизмы ответственны также и за другие типы геномных изменений, включая инсерции, делеции, инверсии и транслокации.

Дупликации могут различаться по размеру и охватывать разные масштабы от сотен нуклеотидов до целых геномов. Например, дупликация, затрагивающая всю хромосому (анеуплоидия), возникает из-за нерасхождения этой хромосомы, что приводит к аномальному числу хромосом. Дупликации меньшего масштаба — это так называемые сегментные дупликации, которые представлены длинными последовательностями ДНК (обычно более 1 т.п.н. в длину), которые имеют высокий уровень идентичности последовательностей (более 90%) и представлены в геноме в нескольких копиях. Сегментные дупликации могут быть тандемными, т. е. непосредственно примыкающими друг к другу, или разнесенными в пространстве.

У эукариот тандемные дупликации могут затрагивать целые гены, как белоккодирующие, так и некодирующие, или только части генов. В последнем случае дупликация приводит к удвоению только части последовательности гена, что влияет на экзон-интронную структуру. Процесс, при котором один и тот же экзон некоторого гена дублируется два или более раз или экзоны из разных генов эктопически сближаются, называется перемешиванием экзонов. Молекулярные механизмы перемешивания экзонов в целом такие же, как и в других типах дупликаций, включая ретротранспозицию, кроссинговер во время половой рекомбинации родительских геномов и проскальзывание репликации.

Однако, в то время как ретротранспозиция может приводить как к тандемным, так и нетандемным дупликациям, негомологичная рекомбинация и ошибки репликации с большой вероятностью должны приводить именно к тандемным дупликациям экзонов, т. е. повторению нуклеотидной последовательности экзона несколько раз в одном и том же месте генома. Предметом изучения данной диссертационной работы являются тандемные дупликации экзонов.

В настоящее время в литературе появляется все больше сообщений о том, что альтернативные изоформы транскриптов, образующиеся в результате тандемных дупликаций экзонов, широко распространены в геноме человека и имеют важное значение для заболеваний [1]. Также из литературы известно, что транскрипты, содержащие тандемно дуплицированные экзоны, часто оказываются подвержены особому виду альтернативного сплайсинга пре-мРНК, при котором один и только один из нескольких экзонов в кластере включается в зрелую мРНК [2—4]. Такие экзоны называются взаимоисключающими (взаимоисключающие экзоны, ВИЭ). С другой стороны, в литературе имеется много примеров ВИЭ, сплайсинг которых регулируется с помощью так называемых конкурирующих структур РНК — групп регуляторных элементов в пре-мРНК, которые конкурируют друг с другом за комплементарное спаривание с одним и тем же общим элементом [5; 6]. А именно, такие транскрипты содержат несколько последовательностей, называемых селекторными сайтами, каждый из которых комплементарен одному и тому же регуляторному элементу, называемому докерным сайтом. Считается, что одновременно может образоваться только одна из конкурирующих структур РНК между селекторным и докерным сайтами, что открывает только один экзон из кластера для распознавания сплайсосомой, однако детали молекулярного механизма взаимоисключающего сплайсинга пока во многом остаются неясными [7].

Тот факт, что тандемные дупликации часто приводят к образованию ВИЭ, сплайсинг которых часто регулируется конкурирующими структурами РНК, наводит на вопрос о существовании общего молекулярного механизма, связанного с природой геномных дупликаций, который мог бы объяснить образование конкурирующих структур РНК. Поиску ответа на этот вопрос и посвящена настоящая диссертация.

Тандемные дупликации, приводящие к взаимоисключающему сплайсингу экзонов, являются важным механизмом расширения разнообразия протеомов, в связи с чем они активно обсуждаются в научной литературе [8]. По со-

временным оценкам геном *D. melanogaster* содержит как минимум 60 генов, обладающих 261 аннотированными ВИЭ, и еще в 744 генах предсказано существование 3583 ВИЭ [9]. Также существует база данных взаимоисключающих экзомов эукариот [10]. В геноме человека с высокой степенью надежности аннотировано 855 ВИЭ, многие из которых играют важную роль в развитии функции сердечной мышцы, а из данных высокопроизводительного секвенирования РНК и других источников предсказано еще 6541 ВИЭ, многие из которых обогащены патогенными мутациями, а их пространственно-временная экспрессия связана с заболеваниями [11]. В большинстве известных случаев, взаимоисключающий сплайсинг экзонов управляется конкурирующими структурами РНК, причем в отдельных генах было показано существование нескольких групп сложно организованных многодоменных структур [8]. Таким образом, в литературе накоплен большой объем сведений о ВИЭ и регуляции их сплайсинга конкурирующими структурами РНК.

Целью данной работы является выявление эволюционных механизмов, объясняющих общее происхождение взаимоисключающих экзонов и конкурирующих структур РНК при тандемных дупликациях.

Для достижения поставленной цели необходимо было решить следующие **задачи**:

1. Исследовать представленность тандемных дупликаций экзонов в генах человека, *D. melanogaster* и *C. elegans*.
2. Исследовать свойства нуклеотидных последовательностей в кластерах взаимоисключающих экзонов человека и *D. melanogaster*, в частности их способность образовывать конкурирующие структуры РНК.
3. Предсказать конкурирующие структуры РНК, регулирующие взаимоисключающий сплайсинг экзонов в гене *Ate1* человека.

Научная новизна:

1. Впервые показано, что тандемные дупликации экзонов широко распространены не только в кодирующих, но также и в нетранслируемых областях генов животных.
2. Получено описание неизвестных ранее тандемных дупликаций экзонов в геноме человека с указанием степени консервативности их последовательностей и уровня экспрессии в тканях.

3. Впервые показано, что интроны в кластерах взаимоисключающих экзонов склонны образовывать конкурирующие структуры РНК, состоящие из докерного и множества селекторных сайтов.
4. Впервые выдвинута гипотеза о том, что тандемные дупликации, затрагивающие экзоны и части фланкирующих интронов, неизбежно приводят к образованию конкурирующих структур РНК и, как следствие, к взаимоисключающему типу сплайсинга.
5. Впервые предсказаны конкурирующие структуры РНК, ответственные за взаимоисключающий сплайсинг экзонов 7a и 7b гена *Ate1* человека.

Практическая значимость. Высказанное в данной работе предположение о закономерном возникновении конкурирующих структур РНК в результате тандемных дупликаций приводит к общему знаменателю и обобщает все известные на данный момент случаи регуляции взаимоисключающего сплайсинга такими структурами. С практической точки зрения это обобщение направляет исследования механизмов взаимоисключающего сплайсинга, в том числе в генах, связанных с болезнями человека, на путь изучения цис-регуляторных элементов вторичной структуры РНК. Исследования в этом направлении также имеют смысл и для тандемных дупликаций, не связанных с аннотированными взаимоисключающими экзонами, поскольку, как здесь было показано, в геномах животных имеется большое число неаннотированных случаев дупликаций экзонов, среди которых часто встречаются взаимоисключающие. Результаты работы представлены через геном браузер, что позволяет исследователям легко визуализировать интересующие их участки генома и находить тандемные дупликации экзонов в них. Конкурирующие структуры РНК, управляющие взаимоисключающим сплайсингом экзонов 7a и 7b гена *Ate1*, можно рассматривать как возможные терапевтические мишени в опухолях, в которых соотношение сплайс-изоформ отличается от физиологического. Таким образом, полученные в данной диссертации результаты имеют фундаментальную научную значимость и широкую практическую применимость.

Методология и методы исследования. Для решения поставленных задач использовались методы биоинформатики, включающие в себя методы выравнивания нуклеотидных последовательностей, методы предсказания вторичной структуры РНК и оценки ее свободной энергии, методы сравнительной геномики и методы анализа данных высокопроизводительного секвенирования РНК нового поколения. Для выявления тандемных дуплика-

ций использовались специализированные методы выравнивания нуклеотидных последовательностей с учетом экзон-интронной структуры. Для предсказания вторичной структуры РНК использовались методы, моделирующие термодинамику взаимодействий РНК-РНК с учетом открытия сайта связывания, и методы, использующие профили доступности с приблизительной энергетической моделью. Для оценки уровней экспрессии в тканях использовались данные высокопроизводительного секвенирования РНК из консорциумов Экспрессия Генотипа Ткани (Genotype-Tissue Expression, GTEx) [12] и Атласа Ракового Генома (The Cancer Genome Atlas, TCGA) [13] с применением стандартных методов обработки и картирования чтений и вычисления покрытия. Для визуализации полученных результатов и представления их в публичном доступе использовались трек-хабы на базе геном браузера UCSC [14].

Основные положения, выносимые на защиту:

1. Интроны и нетранслируемые области генов человека, *D. melanogaster* и *C. elegans* содержат участки, имеющие высокую степень идентичности с аннотированными экзонами и, предположительно, представляющие из себя неаннотированные тандемные дупликации экзонов. Функциональность этих участков подтверждается эволюционной консервативностью и данными по экспрессии в больших панелях транскриптомных экспериментов.
2. Интроны в кластерах взаимоисключающих экзонов человека и *D. melanogaster* обладают повышенным процентом идентичности внутри кластера, который коррелирует с процентом идентичности фланкирущих экзонов, а также повышенной склонностью к образованию комплементарных спариваний, совместимых с моделью докерных и селекторных сайтов, по сравнению с интронами в группах экзонов, подверженных другим типам сплайсинга.
3. Свойства интронов в кластерах взаимоисключающих экзонов указывают на общее происхождение конкурирующих структур РНК и взаимоисключающего сплайсинга через тандемные дупликации, затрагивающие экзоны и части структуры РНК в интронах.
4. Интроны в кластере экзонов 7a и 7b гена *Ate1* человека содержат конкурирующие структуры РНК, предположительно определяющие взаимоисключающий характер сплайсинга этих экзонов и образовавшиеся в результате тандемной дупликации.

Достоверность результатов о роли тандемных дупликаций в образовании конкурирующих структур РНК подтверждается тем, что они, с одной стороны, хорошо согласуются со всеми известными из литературы примерами регуляции взаимоисключающего сплайсинга, и, с другой стороны, дополняют и обобщают их. Предсказанная роль конкурирующих структур РНК в регуляции взаимоисключающего сплайсинга экзонов 7a и 7b гена *Ate1* подтверждается экспериментами с антисенс-олигонуклеотидами и мутагенезом, не вошедшими в данную диссертацию [15]. Предсказания существования неаннотированных случаев дупликаций экзонов подтверждаются данными выскопроизводительного секвенирования, в том числе с помощью покрытия и сплит-чтений. Таким образом, результаты находятся в соответствии с публичными транскриптомными данными и с результатами, полученными другими авторами.

Апробация работы. Основные результаты работы докладывались на 26-м ежегодном собрании Общества РНК в 2021 году (RNA Society Annual meeting, RNA 2021 — онлайн), на Московской международной конференции по вычислительной молекулярной биологии в 2019 и 2021 годах (Moscow Conference on Computational Molecular Biology, MCCMB 2019 и MCCMB 2021 — Москва, Россия), а также на конференции “Информационные технологии и системы” в 2018 году (ИТиС 2018 — Казань, Россия). По материалам диссертации опубликовано три статьи в рецензируемых научных журналах.

Личный вклад. Автор принимал активное участие в разработке плана исследования, получении доступа к данным высокопроизводительного секвенирования РНК и их обработке, выполнении исследования, а также в планировании экспериментальных работ по материалам данной диссертационной работы. Представленные результаты получены автором лично.

Публикации. Основные результаты по теме диссертации изложены в 7 печатных изданиях, 3 из которых изданы в журналах, рекомендованных ВАК, 3 — в периодических научных журналах, индексируемых Web of Science и Scopus, 4 — в тезисах докладов.

Объем и структура работы. Диссертация состоит из введения, 4 глав, заключения и 1 приложения. Полный объём диссертации составляет 104 страницы, включая 39 рисунков и 4 таблицы. Список литературы содержит 157 наименований.

Диссертационная работа была выполнена при поддержке гранта Российского фонда фундаментальных исследований №19-34-90174 “Эволюция взаи-

моисключающих экзонов и регуляция альтернативного сплайсинга вторичной структурой РНК" (руководитель Первушин Д. Д.).

Глава 1. Обзор литературы

1.1 Дупликации генов и дупликации экзонов

Дупликации являются одним из основных механизмов, с помощью которых в молекулярной эволюции генерируются новые молекулярные функции. Дупликации могут охватывать различные геномные масштабы, от генов или их частей до целых геномов. В этом разделе кратко описываются дупликации генов и дупликации экзонов.

1.1.1 Дупликации генов

Дупликация гена создает генетическую избыточность, при которой ген-копия часто свободен от селективного давления, а мутации в нем не оказывают вредного воздействия на организм, поскольку исходный ген все еще выполняет свою функцию. Следовательно, дуплицированные гены обычно приобретают мутации быстрее, чем гены с одной копией, и одна из двух копий может развить новую функцию. Этот процесс называется неофункционализацией. Другой возможностью является субфункционализация, при которой каждая из копий накапливает вредные мутации до тех пор, пока дефекты дополняются другой копией [16; 17].

Классическим примером дупликации генов является человеческий α-like глобиновый генный кластер [18]. Составляющие его гены *hba1* и *hba2* — это паралоги, которые произошли в результате дупликации гена-предшественника с последующим расхождением функций. Примечательно, что оба названных гена имеют почти идентичные последовательности ДНК из-за гомогенизации последовательностей путем генной конверсии [19]. Еще один пример дупликации — это семейство генов *map4k*, кодирующих протеинкиназы, участвующие в каскаде передачи сигнала MAP-киназы [20]. В группе obscura дрозофилы произошла дупликация гена *caf1-55*, в результате чего образовался ген *caf1-55dup*. При этом одна из копий — *caf1-55* находилась под постоянным отрицатель-

ным отбором, а производная копия *caf1-55dup*, которая возникла в результате дупликации ∼18 миллионов лет назад, подверглась неофункционализации [21].

Дупликации могут происходить по нескольким сценариям: полногеномная дупликация (whole genome duplication, WGD), хромосомная дупликация и сегментная дупликация (segment duplication, SD). Считается, что полногеномные дупликации оказывают наибольшее влияние на эволюцию растений [22], однако исследования одноклеточной эукариоты *Paramecium tetraurelia* показали, что её геном, состоящий примерно из 40000 генов, претерпел не менее трех полногеномных дупликаций [23]. Дивергенции рыб и четвероногих предшествовали два различных цикла полногеномных дупликаций у позвоночных [24]. Также считается, что около 25% всех генов позвоночных возникли в результате событий дупликации [25].

Данные типы дупликаций следуют так называемой модели Охно, которая описывает сценарий “сначала дупликация, затем неофункционализация”, но неофункционализация гена после этого события происходит редко, потому что делеция или дрейф новой копии гораздо более вероятны, чем приобретение новой функции под давлением отбора, заключающимся в наличии двух копий региона [26]. Поэтому в настоящее время считается, что большая часть неофункционализированных генов возникает благодаря так называемой модели “инновация-амплификация-дивергенция” (рис. 1.1). Эта модель предполагает, что сначала ген приобретает побочную функцию, тем самым подвергая себя селективному давлению, чтобы сохранить множество своих копий в геноме, а затем полезные мутации в одной копии наследуются, что приводит к образованию паралогов.

1.1.2 Дупликации экзонов

Тандемные дупликации также могут происходить на уровне экзонов. Ранее сообщалось, что для человека, мухи и червя примерно 10% их генов содержат дуплицированные экзоны [28]. Такое обилие дуплицированных экзонов можно объяснить расширением разнообразия белков без необходимости хранить несколько копий целого гена. Большинство примеров дупликаций тандемных экзонов связаны с ВИЭ — типом альтернативного сплайсинга, который

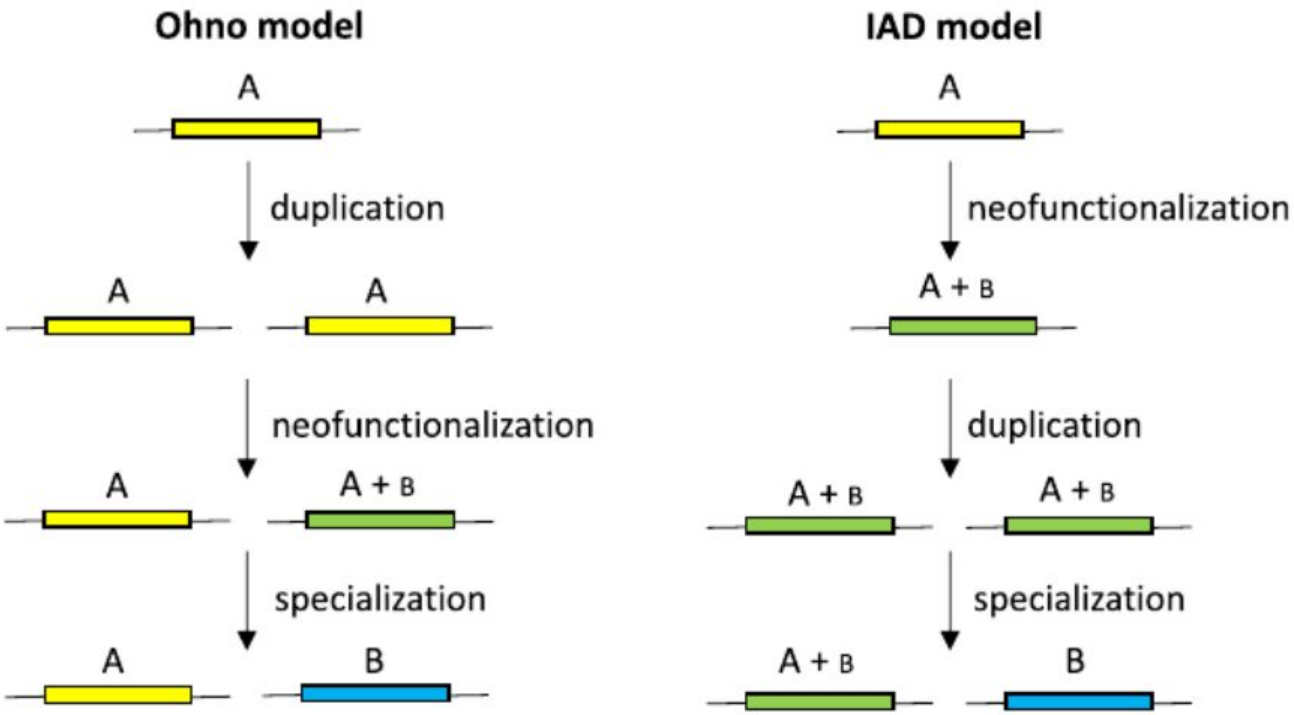

Рисунок 1.1 — Модели Охно и Инновация-Амплификация-Дивергенция. Изображение заимствовано из [27].

подробно обсуждается в следующем разделе. Следует отдельно отметить то, что альтернативные изоформы транскриптов, содержащих тандемно дуплицированные экзоны, имеют большое клиническое значение [11].

Первое и на сегодняшний день единственное систематическое исследование тандемных дупликаций экзонов было посвящено изучению гомологии между экзонами в генах млекопитающих и поиску гомологии между экзонами и соседними интронами [28]. В этом исследовании был выявлен 12291 случай тандемных дупликаций экзонов в геномах человека, мухи и червя. Также сообщалось о 4660 областях в последовательностях интронов, которые имеют высокое сходство с последовательностями экзонов. Некоторые из этих участков (66) были на 100% идентичны и относились к артефактам сборки генома. Для оценки функциональности дуплицированные области, которые не были аннотированы как экзоны, сравнивали с экспрессируемыми последовательностями (expressed sequence tags, EST) и кодирующими ДНК (кДНК). При этом поддержка экспрессии экзонов в EST была обнаружена в 35% неаннотированных и 41% аннотированных областей. Неаннотированные дуплицированные участки также обнаруживались в последовательностях из баз данных RefSeq [29], однако их частота была намного ниже (21% против

63,3% для аннотированных дуплицированных участков). При этом 1578 из 4660 найденных неаннотированных участков содержали стоп-кодоны. Оценка избирательного давления проводилась методом Янга и Нильсена [30], причем соотношение K_a/K_s количества несинонимичных замен на несинонимичный сайт к количеству синонимичных замен на синонимичный сайт показало, что значимой разницы между аннотированными экзонами и неаннотированными дуплицированными областями, т.е. экзонами-кандидатами, не наблюдается за исключением тех случаев, когда область содержит стоп-кодон. Затем было выбрано подмножество неаннотированных областей для проверки возможности их сплайсинга взаимоисключающим образом. 963 из этих регионов были достаточно похожи друг на друга для дальнейшего изучения. Из них 62,3% имели длину, не кратную трем. Кроме того, поиск в базах данных EST и кДНК показал, что дуплицированные экзоны подвержены взаимоисключающему сплайсингу и только 0,6% из них экспрессируются одновременно.

Однако данное исследование основано на версиях баз данных двадцатилетней давности и изучает только гомологию между аннотированными экзонами и соседними с ними интронами [28]. В данной диссертационной работе заново рассматривается проблема обнаружения гомологии между экзонами и интронами, которая распространяется также и на некодирующие области, а также на межгенные области за пределами аннотированных границ генов.

1.1.3 Механизмы геномных дупликаций

Молекулярные механизмы дупликаций представлены тремя основными классами: механизмы, зависящие от репликации, дупликация с помощью мобильных элементов и неравный кроссинговер.

Механизмы, зависящие от репликации

При репликации ДНК могут происходить процессы, в результате которых ДНК-полимераза может перемещаться с одной ДНК-матрицы на другую.

Согласно модели FoSTeS (Fork Stalling and Template Switching) вилка репликации ДНК останавливается, отстающая цепь отделяется от исходной матрицы и переключается на другую вилку репликации на основе гомологии между измененным сайтом матрицы и исходной вилкой. Такое переключение с одной ДНК-матрицы на другую может привести либо к удалению, либо к дублированию, в зависимости от того, расположена ли новая вилка ниже или выше в направлении 5'-конца ведущей цепи [31]. Ошибки, связанные с переключением ДНК-полимеразы с одной ДНК-матрицы на другую играют важную роль в патологических процессах и способствуют развитию таких заболеваний человека, как рак. Например, линии клеток рака молочной железы человека HCC1187 и HCC1806 содержат множественные дупликации, охватывающие 46 тыс. п.о. и 200 тыс. п.о., соответственно, а также делеции, охватывающие 29 тыс. п.о. и 31 млн. п.о., соответственно, которые, как предполагается, происходят из репликационного пузыря, состоящего из двух репликационных вилок [32].

Дупликация с помощью мобильных элементов

Второй механизм, приводящий к дупликациям генетического материала, связан с действием транспозонов. Транспозоны, или мобильные генетические элементы представляют из себя участки ДНК, способные к передвижению (транспозиции) и размножению в пределах генома. Транспозиция может происходить двумя путями: с участием РНК в качестве медиатора и без неё. В первом случае исходный участок ДНК остается нетронутым, а промежуточный РНК-транскрипт обратно транскрибируется в ДНК, которая позже встраивается в геном. Полученные при этом новые последовательности ДНК называются ретротранспозонами. Транспозиция без РНК в качестве посредника вырезает исходный участок ДНК и помещает его в другое место генома, причем дупликации как таковой не происходит. Дуплицированные таким образом гены обычно лишены промоторов и часто становятся псевдогенами, однако они могут приобретать новые функции при благоприятных обстоятельствах, таких как слияние с генами в месте интеграции [33]. Примеры дупликаций экзонов с помощью мобильных элементов известны в человеческих генах *ptpn1* и *gzma*. Их

кодирующие части содержат LINE-подобный повтор L3, который в гене *ptpn1* охватывает пять экзонов [34].

Неравный кроссинговер

Основным источником дупликаций является неравный кроссинговер [35]. Кроссинговер (или рекомбинация) — это обмен эквивалентными участками ДНК между гомологичными хроматидами в процессе мейоза. Неравный кроссинговер происходит тогда, когда гомологичные хроматиды обмениваются неэквивалентными сегментами ДНК. В этом случае участок одной хроматиды присоединяется к неэквивалентному ему сайту другой, в результате чего образуются две хроматиды разной длины: одна с удаленным сегментом, а другая с соседним дуплицированным сегментом. Этому процессу способствует наличие повторяющихся последовательностей, расположенных вблизи сайта кроссинговера. Если область с повторяющимися последовательностями дуплицирована один раз, то она может быть дуплицирована снова и снова из-за повторяющихся элементов, которые повышают вероятность неравного кроссинговера. Этот процесс приводит к образованию тандемно дуплицированных регионов (рис. 1.2). Например, заболевание, связанное с MYH9 (MYH9-related disease, MYH9RD), представляет собой редкое аутосомно-доминантное заболевание, вызванное мутациями в гене *myh9*, кодирующем тяжелую цепь немышечного миозина IIA. Считается, что патология MYH9RD происходит из-за мутации в окрестности экзона 24 благодаря наличию повтора из 16 нуклеотидов, который предположительно отвечает за неравный кроссинговер и вызывает дупликацию [36].

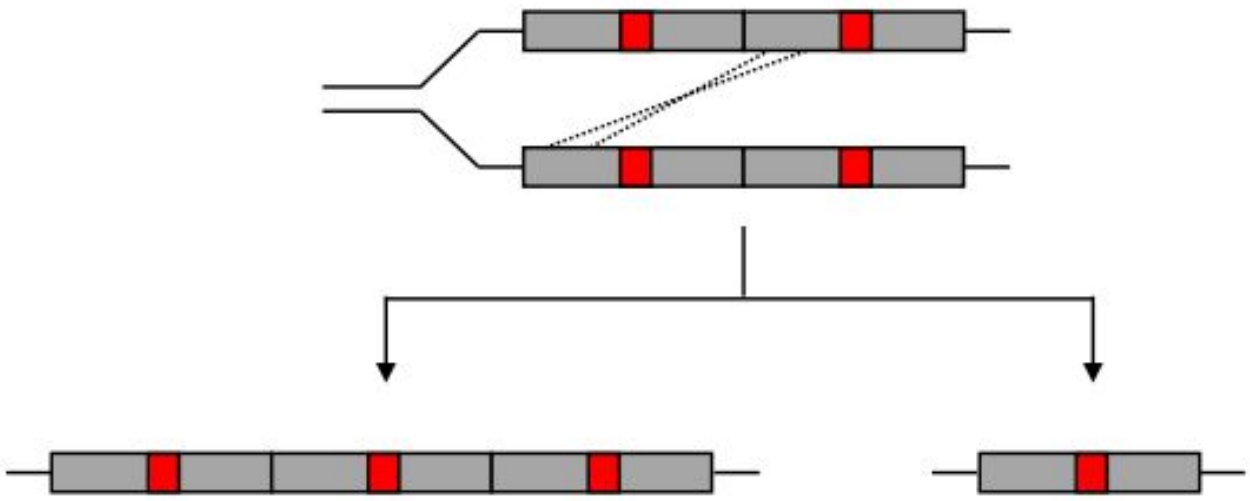

Рисунок 1.2 — Неравный кроссинговер между гомологичными областями дуплицированных сегментов во время репликации приводит к возникновению дочерних клеток либо с большим, либо с меньшим количеством копий дуплицированного участка. Изображение заимствовано из [27].

1.2 Взаимоисключающие экзоны (ВИЭ)

1.2.1 Сплайсинг пре-мРНК

Одновременно с синтезом пре-мРНК на ДНК-матрице в процессе транскрипции она подвергается сплайсингу — процессу, при котором из пре-мРНК вырезаются интроны, а оставшиеся экзоны сшиваются. Как правило интроны представляют собой некодирующую часть пре-мРНК, в то время как экзоны как правило являются кодирующими. Однако некодирующие экзоны также существуют и располагаются в нетранслируемых областях, а также кодирующие экзоны могут пропускаться и становиться интронами в результате альтернативного сплайсинга (см. разд. 1.2.2).

Сплайсинг осуществляется макромолекулярным комплексом, называемым сплайсосомой, который состоит из пяти малых ядерных рибонуклеопротеинов (мяРНП), которые в свою очередь состоят из малых ядерных РНК (мяРНК) U1, U2, U4, U5 и U6 и более ста белков [37; 38]. Этот комплекс собирается на каждом интроне котранскрипционно, т.е. в ходе процесса тран-

скрипции [39; 40]. В простейшем представлении, процесс вырезания интрона состоит из двух основных стадий, каждая из которых включает реакции трансэтерификации. Сначала расщепляется фосфодиэфирная связь между интроном и экзоном, расположенным ближе к 5'-концу пре-мРНК, образуя интронную петлю (лариат, lariat) с так называемой точкой ветвления (branch point, BP). Затем освободившийся экзон соединяется с экзоном, расположенным ближе к 3'-концу пре-мРНК, одновременно высвобождая интронный лариат. Сборка сплайсосомы зависит от консервативных консенсусных последовательностей на границах экзона, называемых сайтами сплайсинга, последовательности точки ветвления и так называемого полипиримидинового тракта — богатого пиримидинами участка, расположенного в 3'-конце интрона. Большинство интронов содержит консервативные консенсусные последовательности сайтов сплайсинга GT/AG (GT для донорного сайта и AG для акцепторного сайта), однако также существует и минорная сплайсосома, распознающая консенсус AT/AC [41].

В ходе протекающих в сплайсосоме реакций, сначала U1 мяРНК распознает 5'-сайт сплайсинга, образуя так называемый ранний комплекс E (рис. 1.3). Затем 3'-сплайс сайт распознается U2 мяРНК и привлекаются факторы SF1 и U2AF, тем самым завершая формирование E-комплекса. В ходе АТФ-зависимого процесса, катализируемого геликазами DExD/H (Prp5 и Sub2), мяРНК U2 распознает консенсусную последовательность точки ветвления и взаимодействует с мяРНП U1, образуя так называемую пре-сплайсосому (A-комплекс). У большинства многоклеточных организмов, в частности у млекопитающих, длина интронов во много раз превышает длину экзонов, составляющих несколько сотен и даже тысяч нуклеотидов. Процесс определения экзонов происходит посредством взаимодействия мяРНП U1 и U2, образующих комплекс распознавания экзонов, а затем мяРНП U1 и U2 образуют взаимодействия, охватывающие всю длину интрона, называемые комплексом распознавания интрона. После сборки комплекса A привлекаются мяРНП U4/U6 и U5 и вместе образуют комплекс B в результате реакции, катализируемой геликазой DExD/H Prp28. Образовавшийся комплекс претерпевает конформационные изменения, образуя каталитически активный комплекс B*. Для этого требуется ряд геликаз РНК (Brr2, Snu114 и Prp2), что приводит к образованию связи между мяРНП U2 и U6 и диссоциации мяРНП U1 и U4 от сплайсосомы, что приводит к высвобождению 5'-конца мяРНК U6. Затем комплекс B* выполняет первую стадию каталитического сплайсинга, образуя комплекс C, содержащий свободный

первый экзон, промежуточный продукт из интрона и второй экзон. Образовавшийся комплекс претерпевает дополнительную АТФ-зависимую перестройку и осуществляет вторую стадию каталитического сплайсинга с участием Prp8, Prp16 и Slu7, что приводит к образованию постсплайсосомного комплекса, содержащего интроны, а также претерпевшую сплайсинг мРНК. Наконец, мяРНП U2, U5 и U6 высвобождают мРНК, катализируемую геликазой DExD/H Prp22, и диссоциируют друг от друга. Процесс сплайсинга подробно описан в [37; 38; 42], и описывать его более детально здесь не потребуется.

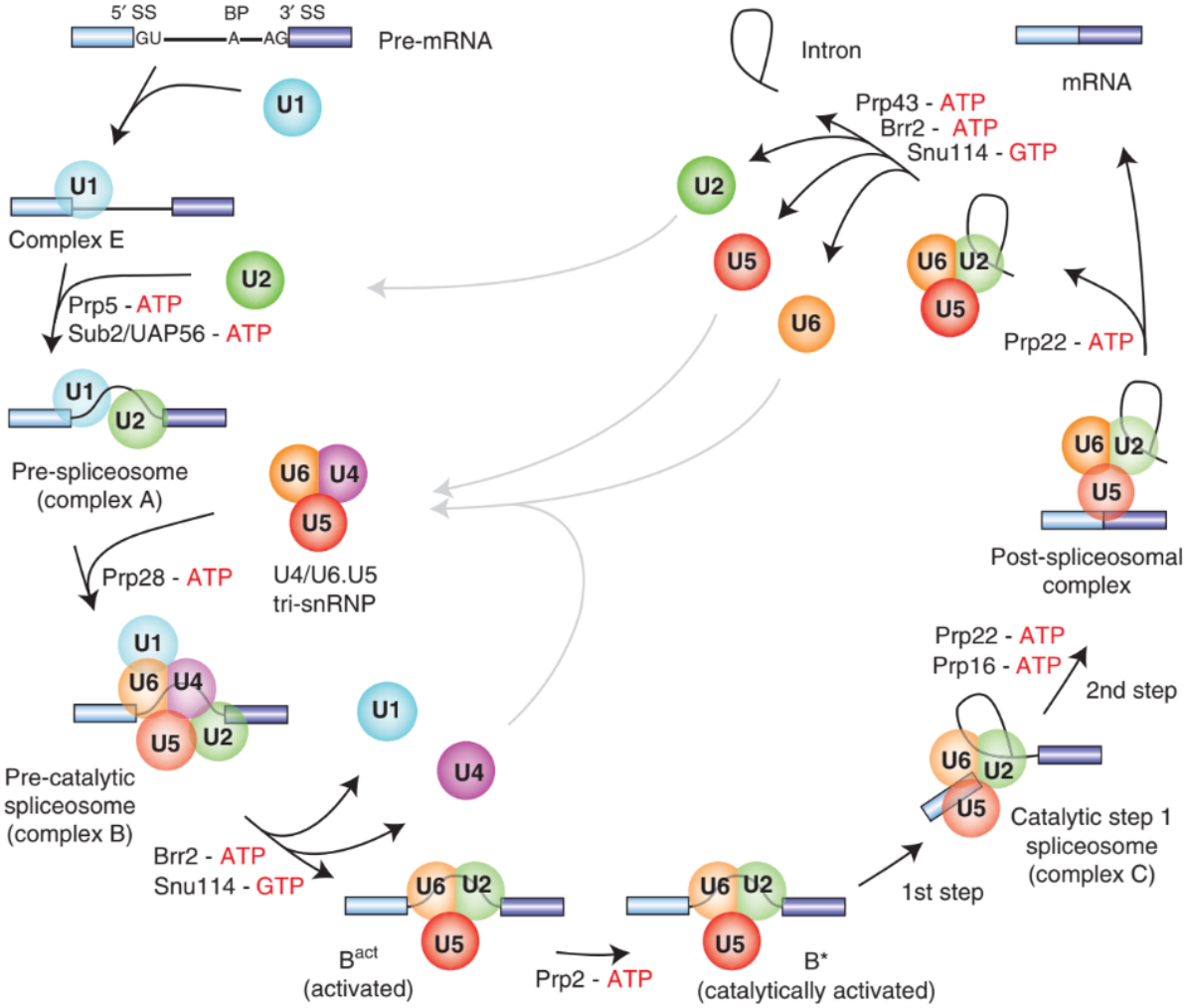

Рисунок 1.3 — Сплайсинг пре-мРНК сплайсосомой U2-типа. Для простоты показаны только взаимодействия между мяРНП (обозначены кружками), но не между белками, не являющиеся мяРНП. Сплайсосомные комплексы названы в соответствии с номенклатурой многоклеточных животных. Последовательности экзонов и интронов обозначены прямоугольниками и линиями, соответственно. Изображение заимствовано из [38].

1.2.2 Альтернативный сплайсинг

Экзоны не всегда сплайсируются одинаково в транскриптах одного и того же гена. У высших эукариот около 95% генов, состоящих из нескольких экзонов, подвергаются альтернативному сплайсингу (АС), при котором зрелые транскрипты одного и того же гена могут состоять из разного набора экзонов [43; 44]. При этом альтернативно сплайсированные мРНК могут транслироваться, что позволяет генерировать несколько изоформ белка из одного гена. Наличие геномных последовательностей и данных секвенирования РНК позволило идентифицировать множество альтернативных вариантов транскриптов, функции большинства из которых неизвестны [45].

Одним из примеров транскриптов, которые подвергаются альтернативному сплайсингу и транслируются в разные белковые изоформы, является транскрипт гена *FGFR2*. Различия в механизме сплайсинга в разных типах клеток приводят к образованию различных тканеспецифических изоформ зрелых транскриптов — FGF-R2IIIb и FGF-R2IIIc [46]. Пример гена опухоли Вильмса *WT1* показывает, что изменение в уровне экспрессии альтернативно сплайсируемых изоформ может вызывать заболевания, в частности синдром Фрейзера. Транскрипт *WT1* кодирует домен цинкового пальца, который способен связываться с ДНК, а его экзоны 5 и 9 сплайсируются альтернативно, что приводит к образованию четырех различных белковых изоформ. Точечная мутация в интроне *WT1* вызывает нарушение нормального альтернативного сплайсинга в донорном сайте сплайсинга экзона 9, что приводит к нарушению синтеза одной из изоформ, что в конечном итоге приводит к развитию синдрома Фрейзера [47].

Биоинформатические способы изучения альтернативного сплайсинга базируются на аннотации транскриптома, т.е., на некотором списке транскриптов и их составных частей, для которых сообщаются их координаты в референсном геноме. В настоящее время имеется две больших базы данных для аннотации транскриптома человека: база данных GENCODE [48] и база данных RefSeq [29]. GENCODE, в свою очередь, состоит из нескольких частей. В частности, исчерпывающая аннотация GENCODE содержит больше информации об альтернативном сплайсинге и охватывает большие части генома по сравнению с RefSeq, в то время как базовая аннотация GENCODE почти не отличается

от RefSeq. Тем не менее, между этими базами данных имеются существенные различия [49].

Наиболее универсальным способом математического описания событий альтернативного сплайсинга на основании аннотации транскриптома является построение графа сплайсинга [50; 51]. Граф сплайсинга — это ориентированный ациклический граф, вершины которого соответствуют позициям или интервалам в нуклеотидной последовательности, а ребра — событиям сплайсинга. Такой граф можно определить несколькими различными способами, например, положив, что вершины соответствуют сайтам сплайсинга, а ребра соответствуют экзонам и интронам. В последнем случае граф естественно снабжается двудольной структурой, т.е., его вершины делятся на два класса (донорные и акцепторный сайты), а ребра соединяют вершины из разных классов. Независимо от способа определения графа сплайсинга, наибольший интерес в нем представляют из себя подграфы — мотивы, которые можно назвать элементарными событиями альтернативного сплайсинга, из которых собираются более сложные комбинации.

Наиболее распространенные среди элементарных событий альтернативного сплайсинга приведены на рис. 1.4. Среди них кассетные, или пропущенные экзоны (cassette or skipped exons) являются наиболее часто встречающимся подтипом, в котором сам экзон и фланкирующие его интроны или полностью вырезаются из транскриптов, или экзон включается в зрелый транскрипт. Включение экзона обычно увеличивает длину транскрипта и соответствующего белка, однако оно может также привести к включению преждевременного стоп-кодона (premature termination codon, PTC) или сдвигу рамки считывания. По современным оценкам около 38% событий альтернативного сплайсинга у человека являются кассетными экзонами [52].

Использование альтернативных сайтов сплайсинга составляет около 28% всех событий альтернативного сплайсинга у человека [52]. В результате использования альтернативных сайтов сплайсинга один и тот же донорный, т.е. 5'-сайт сплайсинга может соединяться с несколькими различными акцепторными, т.е. 3'-сайтами сплайсинга (альтернативный акцепторный сайт), или один и тот же акцепторный, т.е. 3'-сайт сплайсинга может соединяться с несколькими различными донорными, т.е. 5'-сайтами сплайсинга (альтернативный донорный сайт). Альтернативные акцепторные и альтернативные донорные сайты составляют около 18% и 10% всех событий альтернативного сплайсинга человека,

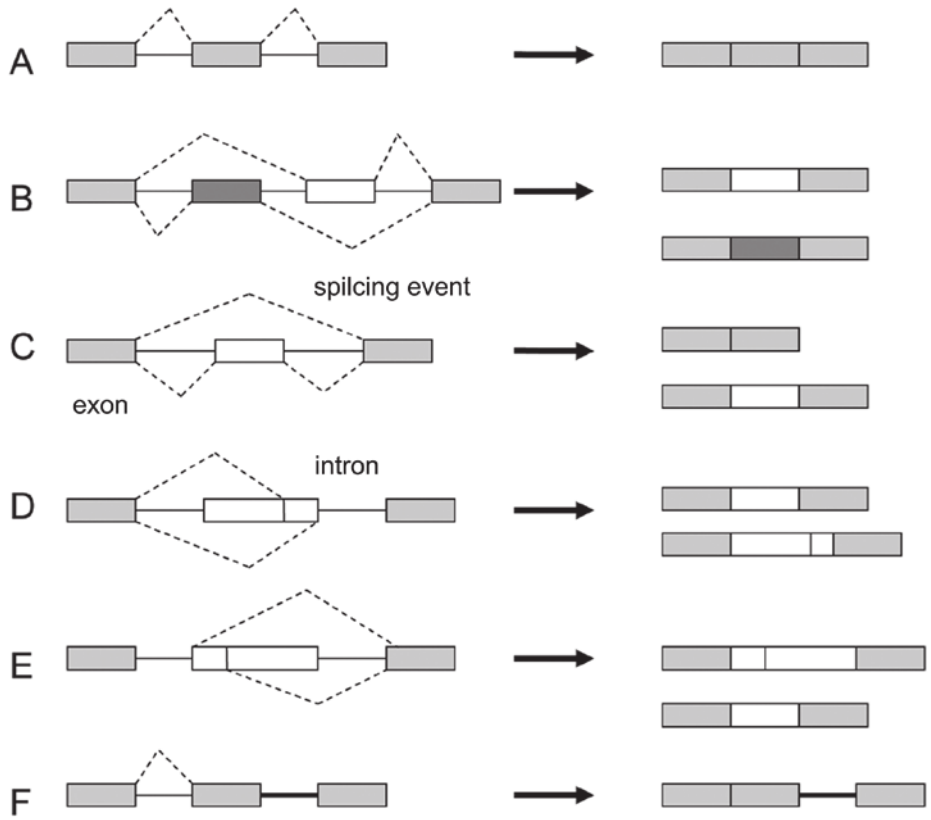

Рисунок 1.4 — Пять основных типов событий сплайсинга. (A) конститутивный сплайсинг; (B) взаимоисключающие экзоны; (C) кассетный альтернативный экзон; (D) альтернативный 3'-сайт сплайсинга; (E) альтернативный 5'-сайт сплайсинга; и (F) удержание интрона. Изображение заимствовано из [52].

соответственно. События удержания интрона, при которых интронная последовательность оказывается включена в зрелую мРНК, составляют около 3% событий альтернативного сплайсинга. В норме события удержания интронов у млекопитающих встречаются редко, однако они часто наблюдаются в опухолях при инактивации опухолевых супрессоров [53].

Наконец, взаимоисключающие экзоны составляют менее 2% событий альтернативного сплайсинга [52] и позволяют включить в зрелую мРНК один и только один из нескольких последовательных экзонов, которые называются взаимоисключающими. Подробному описанию этого типа альтернативного сплайсинга полностью посвящен разд. 1.2.3.

1.2.3 Взаимоисключающие экзоны (ВИЭ)

Взаимоисключающий выбор экзонов — вид альтернативного сплайсинга, при котором один и только один из тандемно расположенных экзонов включается в мРНК, присущ многим генам эукариот [8]. Несмотря на то, что многие примеры ВИЭ были изначально описаны в модельных организмах [54; 55], недавние исследования, основанные на анализе экспериментов секвенирования РНК (RNA sequencing, RNA-seq), подтверждают, что ВИЭ также часто встречаются у высших позвоночных, нематод и растений [9; 10; 56] и содержат мутации, связанные с патогенными состояниями [11]. Этот тип сплайсинга играет важную роль в диверсификации белков, создавая изоформы, которые отличаются только областями, кодируемыми последовательностями ВИЭ. Хотя большинство кластеров ВИЭ состоят всего из двух экзонов, также известны случаи, когда кластеры ВИЭ содержат десятки экзонов.

У множества генов функции взаимоисключающих экзонов хорошо изучены. Например, классические кадгерины представляют собой гомофильные молекулы клеточной адгезии с характерными мотивами внеклеточной последовательности (кадгериновые повторы) и консервативным цитоплазматическим доменом, который обеспечивает взаимодействие с внутриклеточными лигандами (катенины). В нервной системе кадгерины локализуются в синапсах с ранних стадий синаптогенеза [57; 58]. Эти свойства (ранняя синаптическая локализация, молекулярное разнообразие, селективная адгезия и региональная экспрессия) подтверждают предположение о том, что опосредованное кадгерином межклеточное распознавание важно для синаптической специфичности [59]. Ген Cadherin-N (*cadn*) *D. melanogaster*'s содержит три кластера ВИЭ и в целом может продуцировать до 12 различных изоформ транскриптом. Они используются для синтеза различных белков N-кадгерина, которые необходимы для соединения нейронов фоторецепторов R7 [60; 61].

Ген *mhc D. melanogaster* кодирует тяжелую цепь миозина — белок, участвующий в сокращении мышц посредством гидролиза АТФ. Животные, такие как млекопитающие, куры и нематоды, содержат в своем геноме несколько копий этого гена, возникших в результате дупликации, но у *D. melanogaster* ген *mhc* только один. Изменчивость изоформ белков достигается за счет АС. Один ген дрозофилы содержит 5 кластеров ВИЭ с 2, 4, 3, 5 и 2 экзонами [62], соот-

ветственно, и может продуцировать до 480 различных изоформ белка, которые по-разному экспрессируются в личиночных и взрослых мышцах. Изоформы, содержащие экзон 9a, экспрессируются в непрямых летательных мышцах, экзоны 9a и 9b экспрессируются в прыгательных мышцах, экзоны 9b и 9c экспрессируются в других личиночных и взрослых мышцах [63].

Ген *14-3-3ζ D. melangaster* кодирует три различные изоформы белка с помощью одного кластера ВИЭ. Функция этого белка связана с сигнальными путями, в частности с каскадом Ras/MAPK. Он участвует в таких процессах, как эмбриогенез, полярность эпителиальных клеток, активен в нейронах, и необходим для правильного функционирования механизмов памяти и обучения. Альтернативный сплайсинг этого гена в *D. melangaster* достигается за счет кластера ВИЭ с тремя гомологичными экзонами, причем различные изоформы по-разному экспрессируются в таких структурах, как кровеносная система, эмбриональная/личиночная пищеварительная система, и нервная система. Как и в предыдущем примере, другие животные генерируют разные изоформы этого белка с помощью изменения копийности этого гена (7 различных генов *14-3-3ζ* у млекопитающих), в то время как у *D. melangaster's* многообразие изоформ достигается за счет взаимоисключающего сплайсинга [64].

Ген *srp* кодирует транскрипционный фактор, играющий важную роль в развитии гемоцитов, лимфатических желез и других производных мезодермы. У *D. melanogaster* он содержит кластер из двух ВИЭ, который отвечает за две различные изоформы белка. Эти изоформы обладают различными активационными свойствами in vivo [65]. Два типа встречающихся в природе изоформ, кодируемых srp (с N-пальцевым доменом или без него), обладают разными свойствами связывания с ДНК, и только изоформы, включающие N-пальцевой домен, могут взаимодействовать с FOG homologue U-shaped [65].

Ген *mrp D. melanogaster*, ассоциированный с множественной лекарственной устойчивостью, содержит два кластера ВИЭ с двумя и восемью экзонами, соответственно, что приводит к образованию до 16 альтернативных изоформ белков. MRP представляет собой трансмембранный белок, участвующий в транспорте анионов и обеспечивающий перекрестную устойчивость клеток к широкому спектру лекарственных средств [66]. Предполагается, что различные изоформы отвечают за распознавание или афинность к различным субстратам [3]. Было показано, что ген *mrp* сплайсируется альтернативно с большей частотой в опухолях яичников, чем в контрольных нормальных тканях. Неко-

торые из изоформ дают устойчивость к доксорубицину. Экспрессия факторов сплайсинга *ptb* и *srp20* тесно связана с альтернативным сплайсингом гена *mrp* [67].

Семейство факторов транскрипции млекопитающих forkhead box играет важную роль в клеточно- и тканеспецифической регуляции транскрипции. Один из членов этого семейства — ген *foxp1*, который регулирует дифференцировку клеток, содержит кластер ВИЭ с двумя экзонами, причем одна из образующихся изоформ стимулирует гены плюрипотентности и подавляет гены, связанные с дифференцировкой в эмбриональных стволовых клетках, а другая делает противоположное в дифференцированных стволовых клетках [68].

Другим примером альтернативного сплайсинга факторов транскрипции является ген *mef2d*. Он содержит кластер ВИЭ с двумя экзонами. Изоформы белка, синтезированные с этого гена, являются тканеспецифичными и играют существенную роль в дифференцировке мышц [69]. Изоформы *mef2d* действуют антагонистически, модулируя дифференцировку клеток. Анализ секвенирования методом иммунопреципитации хроматина (ChIP) показывает, что изоформы *mef2d* связывают перекрывающийся набор генов, однако только одни из его изоформ (*Mef2D*α*2*) активирует позднюю мышечную транскрипцию [69].

Семейство генов тропомиозина млекопитающих кодирует белки, которые действуют как агенты, связывающие актин, и участвуют в сокращении мышц. В немышечных клетках они действуют как цитоскелет. Один из членов этого семейства — ген *tpm1-3* сплайсируется взаимоисключающим образом, а полученные альтернативные изоформы специфичны для мышц [70]. Соотношение изоформ этого белка важно для правильного развития и функционирования сердца, в частоности, увеличение частоты одной из изоформ было обнаружено в сердце пациентов с дилатационной кардиомиопатией [71].

Наконец, ген *ate1*, кодирующий фермент аргинилтрансферазу у млекопитающих, который участвует в посттрансляционной конъюгации аргинина с остатками глутамата или N-концевым аспартатом, содержит кластер из двух ВИЭ. Этот ген обладает функцией супрессора опухолей, которая зависит от изоформы белка, содержащей один или второй ВИЭ [15]. Данный ген является предметом изучения в данной диссертационной работе и подробно обсуждается в гл. 4.

1.2.4 Идентификация взаимоисключающих экзонов

Экзоны внутри кластеров ВИЭ часто демонстрируют высокую степень идентичности нуклеотидных последовательностей, что свидетельствует о том, что они образовались в результате тандемных дупликаций [2]. В недавнем обзоре тандемно дуплицированные ВИЭ у эукариот были предсказаны на основании изучения длин экзонов и степени идентичности их последовательностей [72]. Исследования отдельных примеров, таких как филогенетический анализ тандемных экзонов гена *MRP* у многоклеточных животных, выявил множественные независимые дупликации экзонов в разных типах и позволил предположить конвергентную эволюцию взаимоисключающего альтернативного сплайсинга [3]. С другой стороны, также существуют негомологичные ВИЭ, что говорит о том, что не все случаи взаимоисключающего сплайсинга происходят из дупликаций экзонов [8].

В другом исследовании, посвященном характеризации ВИЭ, описана следующая процедура их идентификации [9]. Для того чтобы группу экзонов можно было рассматривать как ВИЭ, они должны располагаться рядом друг с другом, т. е. образовывать кластер. Если рассматривать экзоны с одинаковыми сигналами сплайсинга, т.е. случаи, в которых каждый из интронов кластера фланкирован консенсусными последовательностями GT/AG, GC/AG или AT/AC, а также ограничиться случаями, в которых включение экзонов не нарушает рамку считывания, экзоны имеют одинаковую длину и высокую степень идентичности, то по этим критериям можно предсказать 539 кандидатных кластеров ВИЭ у *D. melanogaster*, 389 у *C. elegans* и 1399 у человека. При этом большинство ВИЭ кластеров (1116) у человека состоят из двух экзонов, а максимальное количество экзонов в кластере равно десяти. С использованием данных секвенирования РНК был подтвержден взаимоисключающий характер сплайсинга у 855 из 1399 найденных ВИЭ [9; 11] (табл. 1).

Организм	Количество ВИЭ	Метод	Источник
D. melanogaster	Около 7,1% всех генов содержат тандемно дуплицированные экзоны; В общей сложности 251 из 23 859 (1,1%) событий сплайсинга; 539 кандидатов ВИЭ	Предсказание на основе гомологии; РНК-сек; Предсказания пайплайна in silico	[28]; [73]; [9]
C. elegans	Около 7,5% всех генов содержат тандемно дуплицированные экзоны; В общей сложности 55 из 4049 (1,4%) событий сплайсинга относятся к ВИЭ; 389 ВИЭ в 138 генах	Предсказание на основе гомологии; РНК-сек; Предсказания пайплайна in silico	[28]; [74]; [10]
H. sapiens	Около 10,7% всех генов содержат тандемно дуплицированные экзоны; 336 ВИЭ, 118 кластеров ВИЭ в 114 генах; 1399 ВИЭ в 629 кластерах	Предсказание на основе гомологии; Предсказания пайплайна in silico;	[28]; [75]; [11]
Arabidopsis thaliana	Не аннотированы; 166 ВИЭ в 66 генах	РНК-сек; Предсказания пайплайна in silico	[76]; [10]

Таблица 1 — Аннотация и предсказания тандемных дупликаций экзонов.

1.2.5 Механизмы регуляции взаимоисключающего сплайсинга

Взаимоисключающий тип сплайсинга регулируется более сложным образом по сравнению с другими видами альтернативного сплайсинга поскольку в состав зрелой мРНК должен входить только один экзон из кластера, а активация включения этого экзона должна происходить одновременно с инактивацией включения остальных экзонов кластера. Для объяснения этого явления было предложено несколько механизмов [8].

Стерическая интерференция

Для того, чтобы сплайсосома млекопитающих и дрозофилы могла сплайсировать интрон, он должен иметь длину более 50-60 нт, т.к. в противном случае возникают стерические затруднения [77; 78]. Несмотря на то, что это требование выполняется в гене альфа-тропомиозина млекопитающих и в кластере экзона 17 в гене *Dscam1* дрозофилы, у них расстояние между точкой ветвления и 5'-сайтом сплайсинга короче, чем ∼50 нт, что также приводит к затруднениям при сборке сплайсосомы [79; 80], вследствие чего из двух ВИЭ в зрелый транскрипт может включиться только один. Стерический механизм может объяснить взаимоисключающий сплайсинг кластеров только с двумя экзонами (рис. 1.5).

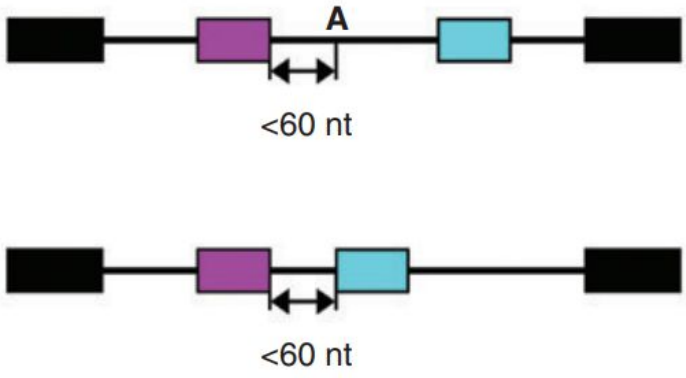

Рисунок 1.5 — Механизм стерической интерференции сплайсинга взаимоисключающих экзонов. Изображение заимствовано из [8].

Сплайсосомная несовместимость

Взаимоисключительность экзонов может достигаться за счет использования разных консенсусных последовательностей на границах сплайсируемых экзонов. Сайты сплайсинга, используемые мажорной (U1/U2) и минорной (U11/U12) сплайсосомой, имеют разные консенсусные последовательности и не совместимы друг с другом. Так, например, интрон с 5'-сайтом сплайсинга U1 и 3'-сайтом сплайсинга U12 не может сплайсироваться ни одной из двух сплайсосом, как это, например, имеет место в человеческом гене *jnk1*, кодирующем активируемую стрессом протеинкиназу [7], а также в других генах [11]. Однако сайты сплайсинга типа U12 встречаются редко [9; 56], что означает ограниченную распространенность данного механизма для регуляции сплайсинга взаимоисключающих экзонов (рис. 1.6).

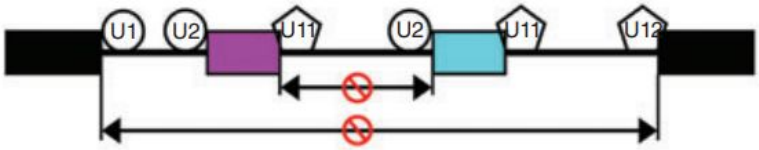

Рисунок 1.6 — Механизм сплайсосомной несовместимости, приводящий к взаимоисключающему сплайсингу. Изображение заимствовано из [8].

Нонсенс-опосредованный распад

Нонсенс-опосредованный распад (nonsense mediated decay, NMD) представляет из себя механизм защиты эукариотических клеток от вредоносных транскриптов, содержащих преждевременные стоп-кодоны [7]. Если длины каждого из двух экзонов не кратны трем, а положение рамки считывания дает полноценный транскрипт только при включении одного из них, то включение обоих экзонов или пропуск обоих экзонов вызывают сдвиг рамки считывания (рис. 1.7) и приводят к образованию транскриптов, содержащих преждевременные стоп-кодоны [81]. Механизм взаимоисключения экзонов, опосредованный

системой NMD, реализуется приблизительно для 60% кластеров ВИЭ, состоящих из пар экзонов [9; 28; 56]. Однако этот механизм не применим к кластерам, состоящим из трех и большего числа экзонов, так как из трех чисел, не кратных трем всегда можно составить комбинацию, кратную трем. Также известно, что многие экзоны в кластерах ВИЭ имеют длину, кратную трем, но при этом сплайсируются взаимоисключающим образом. Поэтому данный механизм не может полностью объяснить все случаи взаимоисключающего сплайсинга.

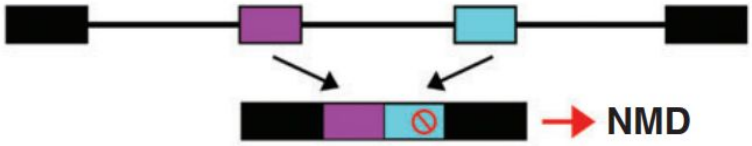

Рисунок 1.7 — Взаимоисключение экзонов по механизму нонсенс-опосредованного распада. Изображение заимствовано из [8].

Однако наиболее распространенным механизмом, за счет которого обеспечивается взаимоисключающий сплайсинг экзонов, являются конкурирующие комплементарные спаривания, образующие взаимоисключающие структуры РНК. Этот механизм подробно обсуждается в разд. 1.3.

1.3 Конкурирующие структуры РНК

Конкурирующие структуры РНК, т.е, группы комплементарных спариваний, которые имеют общие участки и поэтому не могут существовать одновременно, были впервые обнаружены у прокариот, например в системе терминатор-антитерминатор [82]. Для эукариот известно, что взаимоисключающий сплайсинг в ряде случаев также зависит от вторичной структуры пре-мРНК, причем эта вторичная структура образована конкурирующими комплементарными спариваниями.

1.3.1 Конкурирующие структуры РНК в гене *Dscam1*

Самым выдающимся (и исторически первым описанным) примером конкурирующих вторичных структур в эукариотических генах является вторичная структура РНК в кластере экзонов 6 гена *Dscam1* — молекулы клеточной адгезии синдрома Дауна 1 у *D. melanogaster*, которая необходима для регуляции альтернативного сплайсинга [55; 83]. Этот ген содержит 4 кластера ВИЭ, состоящих из 2, 12, 33, 48 экзонов, соответственно (рис. 1.8). Интроны одного из кластеров ВИЭ (кластер экзонов 6) содержат два типа консервативных элементов: так называемый докерный сайт, расположенный в интроне между конститутивным экзоном 5 и первым экзоном из кластера экзонов 6, и множество селекторных сайтов, расположенных по одному в каждом из интронов перед 48 альтернативными вариантами экзона 6. Эти классы элементов способны образовывать комплементарные спаривания таким образом, что каждый селекторный сайт комплементарен к части докерного сайта, а комплементарное спаривание сопоставляет один и только один альтернативный экзон с конститутивным экзоном 5. В связи с этим было высказано предположение, что взаимоисключающая природа взаимодействий докерного и селекторного сайтов является центральным компонентом механизма, гарантирующего, что только один вариант экзона 6 включается в зрелую мРНК.

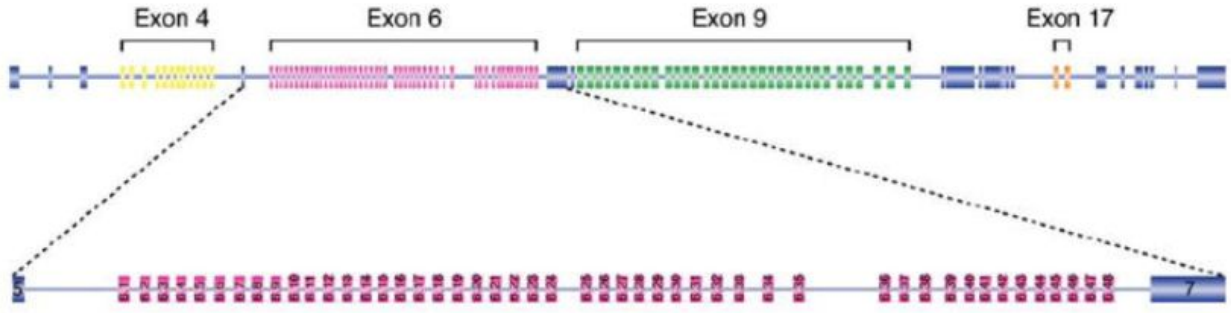

Рисунок 1.8 — Организация гена *Dscam1* у *D. melanogaster* Всего он содержит 115 экзонов, 95 из которых образуют 4 кластера ВИЭ. Большинство экзонов кодируют вариабельные домены иммуноглобулинов. Кластер экзонов 17 содержит только два экзона, которые кодируют альтернативные варианты трансмембранного домена. Изображение заимствовано из [83].

Примечательно то, что конкурирующие вторичные структуры в гене *Dscam1* были впервые обнаружены с помощью методов сравнительной геномики [83]. Множественное выравнивание последовательностей интронов в кластере экзонов 6 позволило обнаружить консервативные интронные участки, позднее названные докерным (рис. 1.9) и селекторными сайтами (рис. 1.10). Эти участки имеют высокую степень консервативности среди членистоногих, что указывает на их важность для регуляции сплайсинга пре-мРНК *Dscam1*.

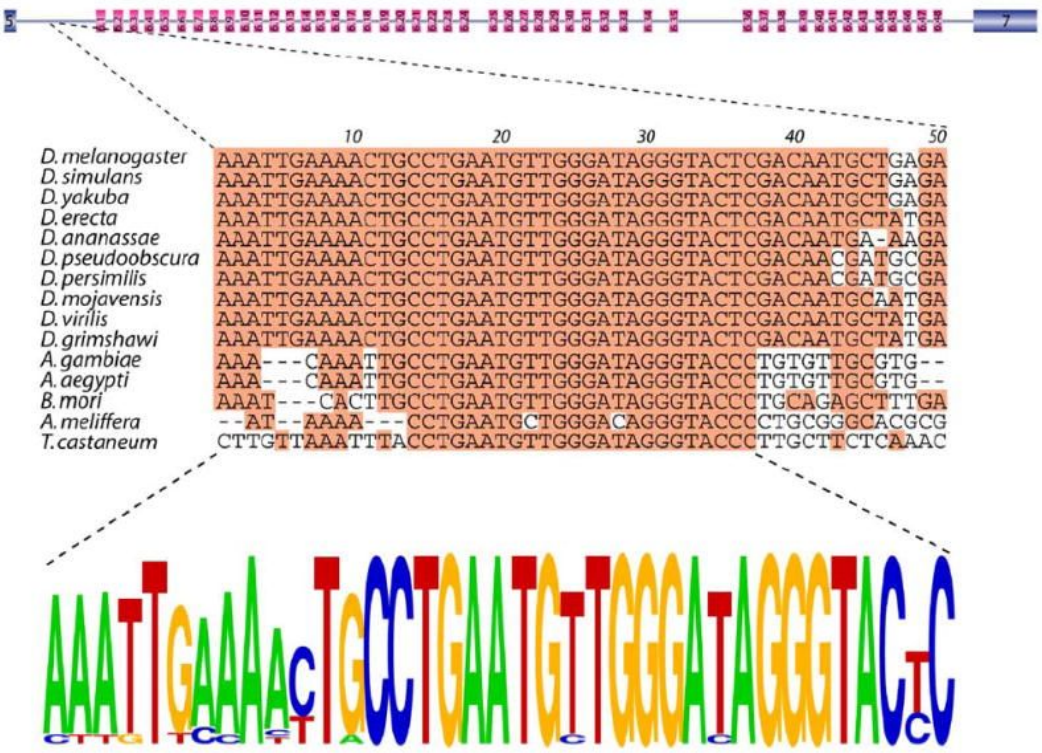

Рисунок 1.9 — Консенсусные последовательности докерного сайта кластера экзонов 6 *Dscam1*. Изображение заимствовано из [7].

Позднее аналогичные регуляторные элементы были обнаружены и для других кластеров ВИЭ этого гена (кластер экзонов 4 и кластер экзонов 9) [5]. Кроме конкурирующих вторичных структур, ген *Dscam1* содержит сложные тандемные мульти-субъединичные структуры РНК — так называемый локус контроля, который обеспечивает выбор только одного варианта экзона из кластера в сочетании с взаимодействиями докерных и селекторных последовательностей [84]. Сравнительно недавно был обнаружен набор скрытых вторичных структур РНК, которые уравновешивают стохастический выбор ВИЭ, находящихся на разных расстояниях от фланкирующих конститутивных экзонов [84]. Мутационный анализ *in vivo* выявил двойную функцию этих балансирующих взаимодействий в управлении стохастическим выбором ВИЭ за

счет усиления включения дистальных экзонов и одновременного подавления включения проксимальных экзонов.

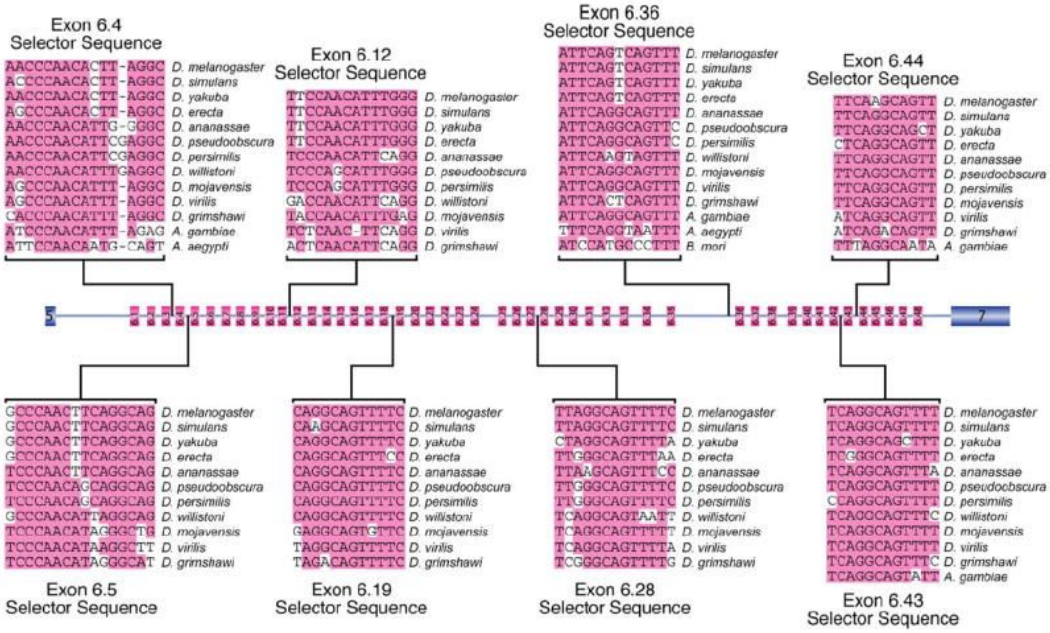

Рисунок 1.10 — Консенсусные последовательности селекторного сайта кластера экзонов 6 *Dscam1*. Изображение заимствовано из [7].

Таким образом обеспечивается взаимоисключающий сплайсинг в каждом из кластеров, а благодаря всевозможным сочетаниям в общей сложности могут образоваться до 38016 различных изоформ белка. Считается, что такое огромное разнообразие необходимо для правильного развития нервной системы *D. melanogaster*, и в частности для роста и избегания соединений между аксонами. В каждом нейроне вырабатывается уникальная комбинация изоформ белков, которые определяют нормальное формирование связей между дендритами и аксонами разных нервных клеток, одновременно предотвращая самосцепление аксонов [85—93]. В частности, у мух со смещенным соотношением взаимоисключающих изоформ экзона 4 и 9 наблюдаются заметные дефекты грибовидного тела [94].

1.3.2 Другие примеры конкурирующих структур РНК

После того, как в 2005 году проф. Брентоном Гревели была выдвинута гипотеза о регуляции взаимоисключающего сплайсинга в гене *Dscam1* конку-

рирующими структурами РНК, в литературе начали появляться сведения о существовании таких структур и в других генах [8]. На рис. 1.11 кроме гена *Dscam1* также показана архитектура кластеров ВИЭ в генах *Mhc* [95] и *MRP1* [66] у *D. melanogaster*, гене *Spn4* у *M. sexta* [96], гене *MRP* у *B. floridae* [3] и гене *CD55* человека [9].

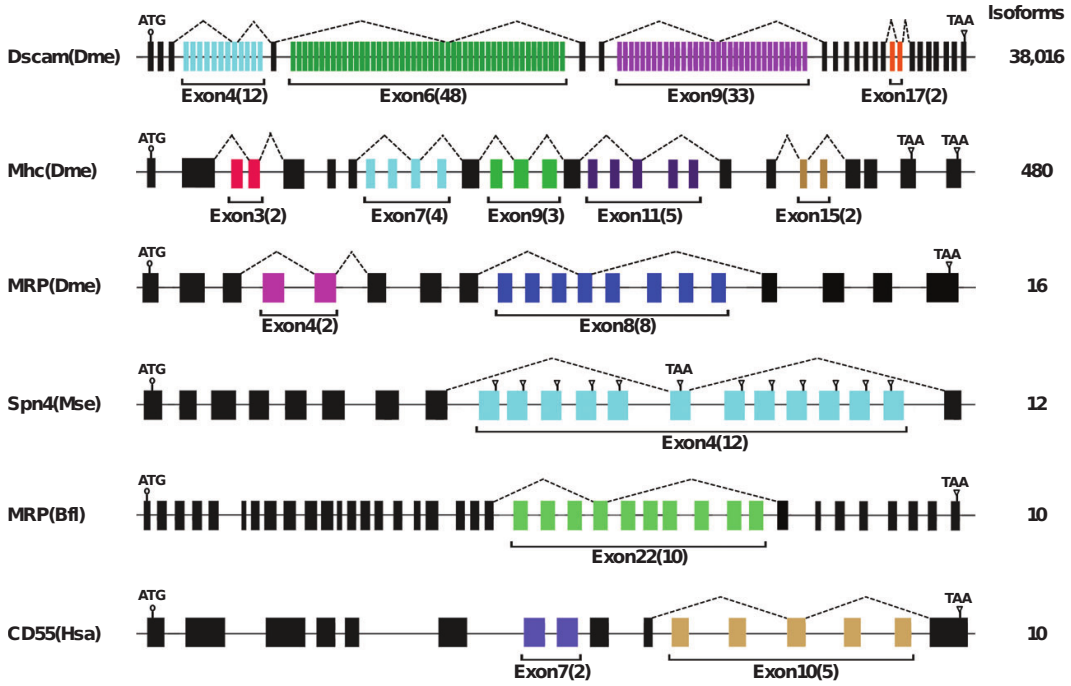

Рисунок 1.11 — Примеры генов, содержащих кластеры ВИЭ. Показаны конститутивные экзоны (черные прямоугольники), альтернативные экзоны (цветные прямоугольники) и интроны (линии). Пунктирные линии изображают события альтернативного сплайсинга. Общее число потенциальных изоформ указано справа. Изображение заимствовано из [7].

Примечательно то, что в гене *14-3-3ζ* у *D. melanogaster* конкурирующие структуры РНК были обнаружены в кластере экзонов 5, претерпевшем сравнительно недавнюю дупликацию [5] (кластеры у *D. melanogaster* и у *D. yakuba* состоят из разного числа экзонов). Сравнительный анализ нуклеотидных последовательностей 33 видов семейства Drosophilidae, разошедшихся приблизительно 80 миллионов лет назад, обнаружил консервативные элементы IE1 и IE2 и обратно комплементарную им последовательность IEa в интронах этого кластера. Было показано, что взаимоисключающий выбор экзона в этом кластере нарушается после внесения замен в IE1, IE2, и IEa, и восстанавливается после внесения компенсаторных замен. Сравнительный анализ геномов 73 видов членистоногих также обнаружил похожие элементы в гене *14-3-3ζ*,

специфичные для вида или клады и эволюционно консервативные на уровне вторичной структуры.

Существенно более сложная организация конкурирующих вторичных структур РНК наблюдается при выборе одного из двух ВИЭ в кластере экзонов 4 гена *srp* у *D. melanogaster* [97]. Взаимоисключительность экзонов здесь достигается за счет существования в кластере двух независимых наборов комплементарных областей, образованных консервативными элементами и образующими псевдоузел. Несмотря на то, что псевдоузлы в молекулах РНК могут существовать, авторы работы предположили, что из-за относительного расположения одна из комплементарных областей выпетливается, когда образуется второе спаривание, что предотвращает образование вторичной структуры с её участием [97].

1.3.3 Механизмы сплайсинга и взаимоисключительность структур РНК

Механистическая модель регуляции взаимоисключающего сплайсинга экзонов конкурирующими структурами РНК состоит в том, что группа экзонов выпетливается и вырезается тогда, когда происходит спаривание комплементарных оснований между докерным и одним из селекторных сайтов. На первый взгляд такое объяснение является логичным для взаимоисключения, однако молекулярные детали отдельных событий сплайсинга остаются во многом неясными.

Ключевым компонентом модели Гревели [83], основаной на дополнительных данных РНК-интерференционного скрининга регуляторов сплайсинга [98], является некоторый репрессор сплайсинга, который предотвращает сплайсинг инактивированных вариантов экзона 6 [83]. Гревели предположил, что вследствие какого-то неизвестного механизма взаимодействие докерного и селекторного сайтов инактивирует репрессор сплайсинга на ближайшем в направлении 3'-конца экзоне и, следовательно, активирует его включение, а включение оставшихся вариантов экзона продолжает подавляться.

В дополнение к этому, другие авторы [8] выделяют различные типы расположения конкурирующих структур: с докерным сайтом в 5'-интроне,

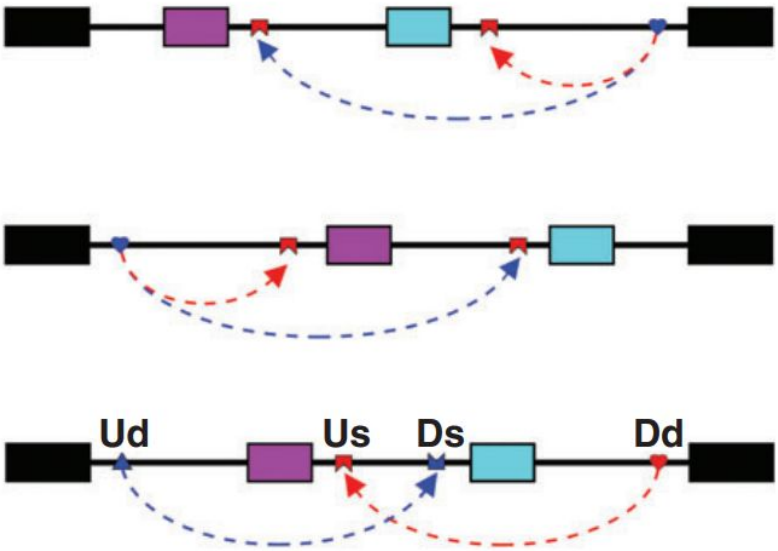

Рисунок 1.12 — Варианты расположения (сверху вниз) докерного и селекторных сайтов: с докерным сайтом в 5'-интроне, 3'-интроне и двустороннее расположение, при котором образуется псевдоузел. Изображение заимствовано из [8].

докерным сайтом в 3'-интроне, и комбинацию этих двух случаев (рис. 1.12). Репрессия альтернативных экзонов требует множества сайленсерных элементов, транс-факторов и слабых сайтов сплайсинга, которые вместе инактивируют альтернативные экзоны, в то время как специфическая активация подразумевает пространственное сближение сайтов сплайсинга. Таким образом, конкурирующие структуры РНК создают пространственную конфигурацию, направляющую сплайсосому по нужному пути для специфической активации нужного экзона, причем в этом процессе могу участвовать как репрессоры, так и активаторы. Таким образом, специфическая активация экзона регулируется на многих уровнях, а степень участия в ней различных факторов может различаться в разных кластерах экзонов.

Таким образом, с одной стороны ВИЭ часто образуются в результате тандемных дупликаций, а с другой стороны их сплайсинг регулируется конкурирующими структурами РНК. В связи с этим возникает вопрос, не связаны ли друг с другом тандемные дупликации экзонов, взаимоисключающий сплайсинг и конкурирующие структуры РНК, и нет ли общего молекулярного механизма, связанного с природой геномных дупликаций, который может быть ответственен за связь между всеми тремя явлениями. Получению ответа на этот вопрос и посвящена настоящая диссертационная работа.

Глава 2. Представленность тандемных дупликаций экзонов в генах человека, *D. melanogaster* и *C. elegans*

Как уже отмечалось, тандемные дупликации тесно связаны с образованием взаимоисключающих экзонов, т.е. типом альтернативного сплайсинга, при котором один и только один экзон из группы тандемно расположенных экзонов включается в зрелый транскрипт. 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 ВИЭ путем идентификации гомологии с соседними экзонами или с близлежащими участками ДНК [28]. Однако тандемные дупликации экзонов могут также охватывать интронные и нетранслируемые области (НТО, untranslated region, UTR), которые не примыкают непосредственно к аннотированным экзонам, а также базы данных аннотаций генома значительно расширились с тех пор.

В этой главе с использованием биоинформатических методов заново рассматривается проблема идентификации тандемных дупликаций экзонов в эукариотических генах и показывается, что тандемно дуплицированные экзоны широко распространены не только в кодирующих частях генов, но и в нетранслируемых областях. В ней представлена динамическая картина экзонных дупликаций в зависимости от гомологии нуклеотидных последовательностей, идентифицированы неаннотированные тандемно дуплицированные экзоны, и приведены статистические свидетельства их экспрессии с использованием больших панелей транскриптомных данных. Результаты, представленные в этой главе, опубликованы автором в работе [99].

2.1 Методы

2.1.1 Процедура поиска тандемных дупликаций

Геномы и аннотации

Геномная последовательность человека (сборка hg19, GRCh37.p13) была получена из Genome Reference Consortium [100]. Исчерпывающая аннотация транскриптов генов человека (GENCODE comprehensive gene annotation) версии 19 была получена из базы данных GENCODE [101]. Геномы *D. melanogaster* (BDGP Release 6, dm6) и *C. elegans* (WBcel235, ce11) были получены из базы данных UCSC Genome Browser [14]. Аннотации транскриптов *D. melanogaster* были получены из базы данных FlyBase релиз dmel_r6.32 [102]. Аннотации транскриптов *C. elegans* релиз 104 были получены из базы данных Wormbase [103]. Аннотации транскриптов RefSeq были получены из базы данных NCBI RefSeq database [29]. Рассматривались только экзоны и транскрипты белок-кодирующих генов. Число уникальных экзонов для человека, *D. melanogaster* и *C. elegans* составляло 329983, 83276 и 172984, соответственно.

Поиск гомологичных экзонов

Для идентификации тандемных дупликаций экзонов была использована программа `exonerate` [104]. Для этого нуклеотидная последовательность каждого экзона была выровнена с нуклеотидной последовательностью содержащего его гена, которая удлинялась в обоих направлениях на 15% длины гена. Программа запускалась в исчерпывающем (exhaustive) режиме для получения точного выравнивания. Минимальный процент отсечения идентичности был установлен на 20%. Последовательности выравниваний были извлечены с помощью программы `getfasta` из пакета bedtools [105]. Выравнивания последовательностей были организованы в таблицу в формате 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 в таблице содержалось 116320, 5244 и 5605 пар запрос-мишень для генов человека, *D. melanogaster* и *C. elegans*, соответственно.

Процедура фильтрации для пар запрос-мишень

Для того, чтобы идентифицировать неаннотированные тандемные дупликации экзонов, таблица пар запрос-мишень была отфильтрована с помощью программы `bedtools intersect` следующим образом. Пары запрос-мишень, в которых последовательность мишени пересекает по крайней мере один аннотированный экзон более чем по 5% ее длины, были удалены. Кроме того, 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 [14], также были удалены.

2.1.2 Данные транскриптомных экспериментов

Были использованы данные РНК-секвенирования из 6625 образцов консорциума Genotype-Tissue Expression Project (GTEx) версии v7 [12]. Короткие чтения (риды) были выравнены на геном человека с помощью картировщика STAR v2.4.2a [106]. Чтения с разрывами (сплит-чтения, split reads), поддерживающие экзон-экзонные соединения, были извлечены с помощью пакета программ IPSA с настройками по умолчанию [107] (порог энтропии Шеннона 1.5 бит). Учитывались только сплит-чтения с каноническими динуклеотидами GT/AG. Уникально картированные чтения были выбраны на основе наличия тега NH:1 из файлов в формате BAM. Среднее покрытие чтениями и показатели консервативности PhastCons были рассчитаны с использованием программного пакета Deeptools [108].

2.2 Результаты

2.2.1 Коэффициент дупликации

Для обнаружения экзонных дупликаций были использованы самые большие на сегодняшний день наборы данных по аннотации экзонов, включая базы данных GENCODE [48] и RefSeq [29]. Поиск гомологии последовательностей для каждого экзона в расширенной на 15% длины гена нуклеотидной последовательности содержащего его гена был выполнен с помощью программы `exonerate` [104]. В дальнейшем аннотированные экзоны будут называться запросными последовательностями или просто запросами, а их соответствующие гомологи, которые были обнаружены с помощью exonerate — мишенями (рис. 2.1). Каждая пара запрос-мишень характеризуется ковариатами, относящимися к запросу (например, местоположение в пределах кодирующего участка (КУ, coding sequence, CDS) или НТО, ковариатами, относящимися к мишени (например, доля длины мишени, перекрывающаяся с аннотированными экзонами), и величиной процента идентичности между запросом и мишенью. Поскольку многие экзоны подвергаются альтернативному сплайсингу и, таким образом, вносят вклад в качестве перекрывающихся областей в наборы аннотаций экзонов, был введен показатель коэффициента дупликации (КД), который определяется как отношение суммарного числа нуклеотидов, покрытых мишенями, к суммарному числу нуклеотидов, покрытых запросами, с заданным или более высоким процентом идентичности нуклеотидных последовательностей. По построению, КД всегда больше 1, поскольку каждый запрос служит своей собственной мишенью со 100% идентичностью последовательности. КД можно вычислить для всех экзонов, а также только для экзонов в кодирующих или только для экзонов в нетранслируемых областях. Таблицы, перечисляющие пары запрос-мишень, доступны в онлайн репозитории https://zenodo.org/record/5474863.

Как и ожидалось, значения КД уменьшаются с увеличением порога на процент идентичности последовательностей (рис. 2.2). Около 2% экзонных нуклеотидов человека в белок-кодирующих областях подвергаются дупликациям при пороге отсечения по проценту идентичности последовательностей в 80%

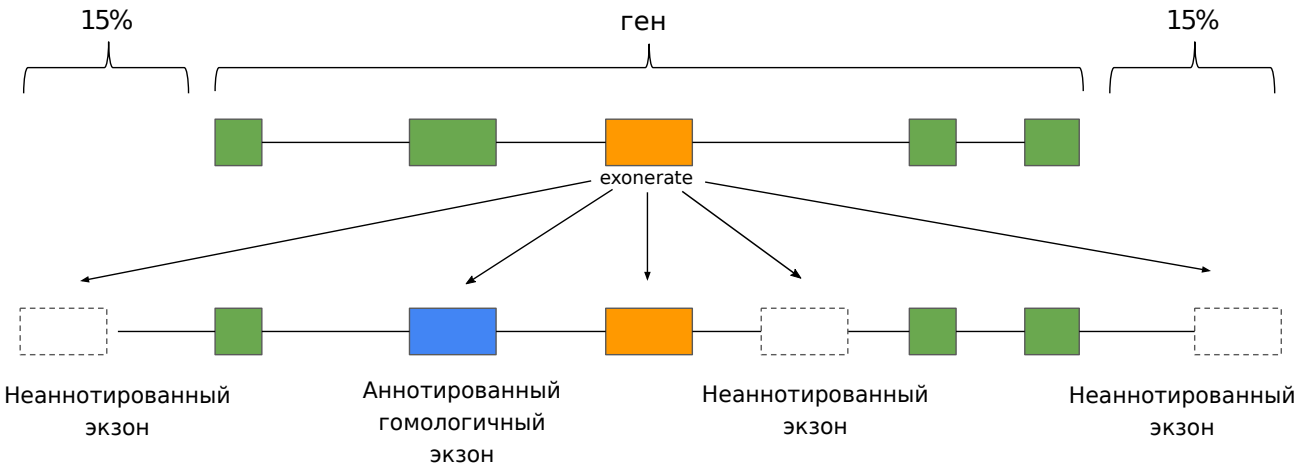

Рисунок 2.1 — Схема поиска тандемных дупликаций экзонов. Нуклеотидная последовательность каждого экзона выравнивается на нуклеотидную последовательность содержащего его гена, расширенную в обе стороны на 15% длины.

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 НТО, почти 15% экзонных нуклеотидов человека подвергаются дупликациям при пороге отсечения по проценту идентичности последовательностей в 80% или более (рис. 2.2), 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 НТО.

Как было показано выше, экзоны, расположеные в НТО, чаще подвергаются дупликациям, чем экзоны из кодирующих областей, однако не ясно какая доля нуклеотидов в НТО является результатом дупликаций. Для ответа на этот вопрос был зафиксирован порог отсечения по проценту идентичности последовательностей в 80% и вычислена доля нуклеотидов, принадлежащих мишеням, в кодирующих участках, НТО, и части межгенных участков, прилегающих к генам, в которых производился поиск. Оказалось, что доля нуклеотидов, принадлежащих мишеням, в НТО действительно выше, чем в кодирующих и межгенных участках (рис. 2.3).

Таким образом, из полученных результатов можно сделать вывод о том, что экзоны, расположеные в НТО, чаще подвергаются дупликациям, чем

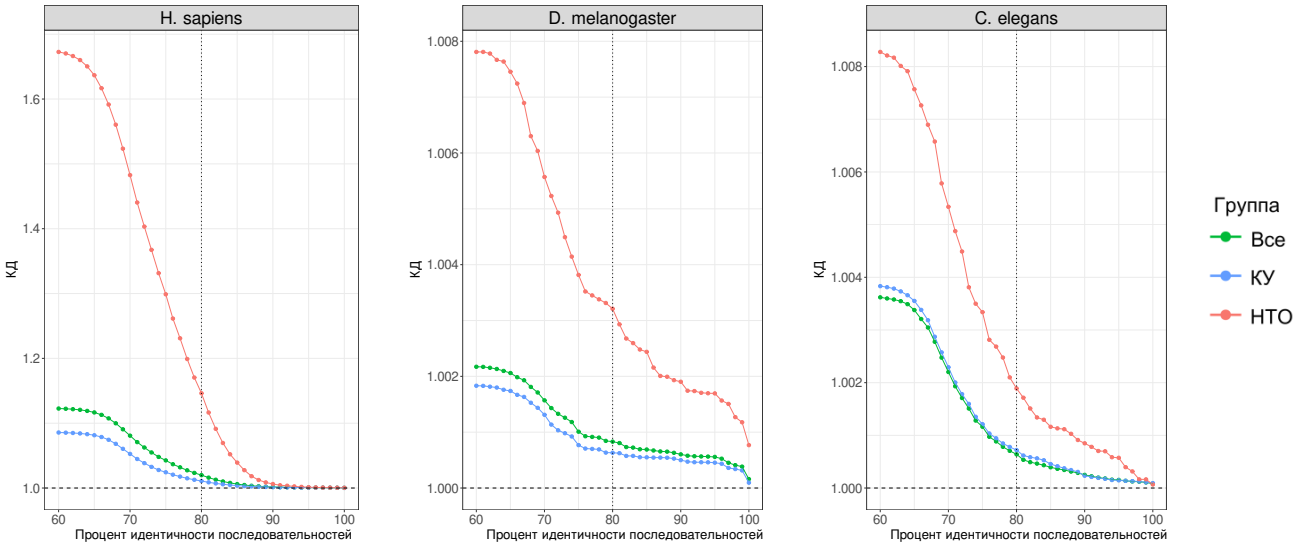

Рисунок 2.2 — Коэффициенты дупликации (КД) в геномах человека, *D. melanogaster* и *C. elegans* как функция процента идентичности нуклеотидных последовательностей запрос-мишень для всех экзонов, экзонов в кодирующих участках (КУ) и экзонов в НТО.

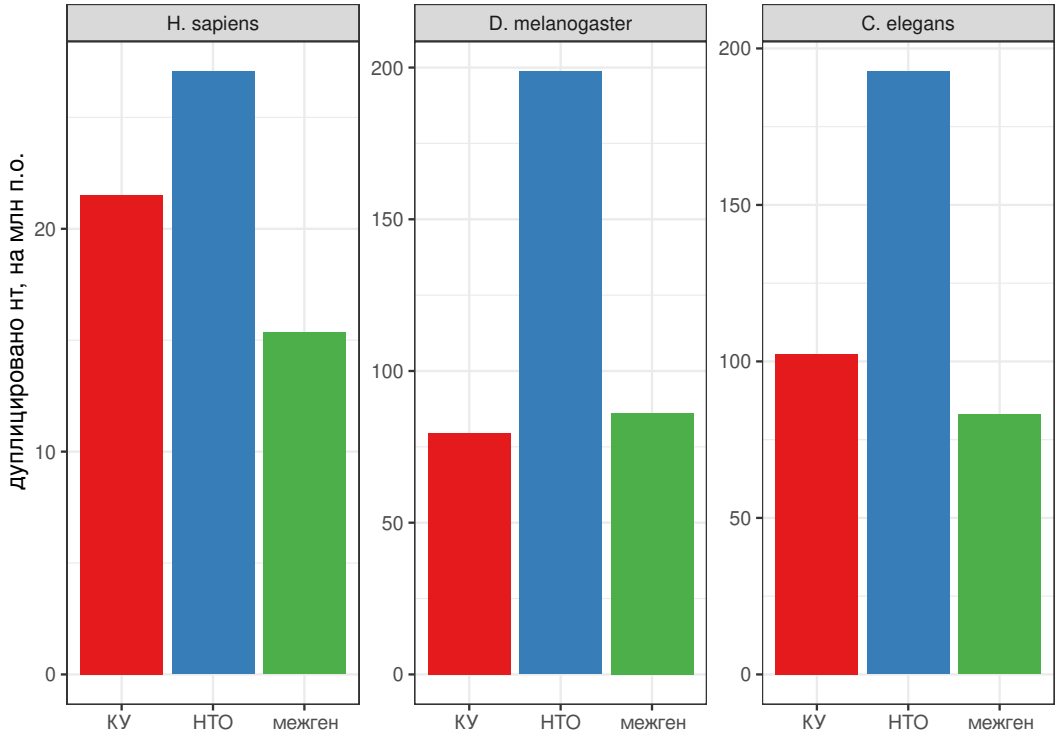

Рисунок 2.3 — Число нуклеотидов на млн. п.о., принадлежащих мишеням с порогом отсечения по проценту идентичности последовательностей в 80%, в кодирующих участках, НТО, и межгенных участках, прилегающих к генам.

экзоны из кодирующих областей, а сами НТО содержат бóльшую долю дуплицированных нуклеотидов.

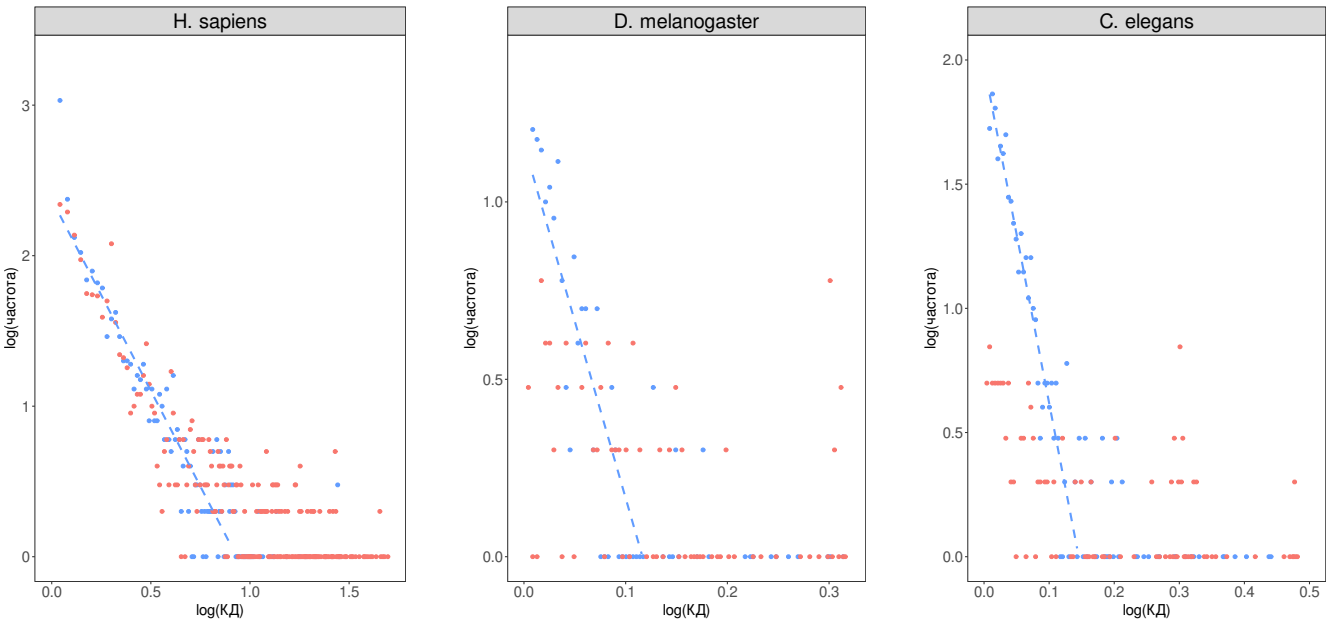

Рисунок 2.4 — Распределение частот значений КД для генов человека, *D. melanogaster* и *C. elegans* при пороге процента идентичности нуклеотидных последовательностей запрос-мишень 80% в $\log_{10} - \log_{10}$ координатах. Цвета как на (рис. 2.3).

Для того, чтобы ответить на вопрос о том, не являются ли одни гены более склонными к тандемным дупликациям экзонов, чем другие, значения КД были вычислены для каждого аннотированного гена отдельно, сгруппированы по интервалам, а затем построены их частотные распределения (рис. 2.4). Частоты значений КД подчиняются степенному закону распределения, о чем свидетельствует близкая к линейной зависимость логарифма частоты от логарифма значения КД со значительным отклонением в сторону более высоких частот для больших значений КД в некоторых генах (рис. 2.4). Гены человека с необычно высокими значениями КД для кодирующих экзонов включают в себя *CAMK1D* (кальций/кальмодулин-зависимой протеинкиназы), *CLYBL* (цитрамалил-КоА лиаза) и *NBPF20* (neuroblastoma breakpoint family), однако некоторые гены человека также имеют заметно высокие значения КД и для нетранслируемых областей, например *OBSCN* (обскурин) и *NEB* (небулин). У *D. melanogaster* заметными отклонениями по числу тандемных дупликаций были гены *dpy*, *hydra* и *heph*.

Для дальнейшего исследования структуры экзонных дупликаций в этих генах был создан трек-хаб для геном браузера в качестве инструмента визуализации для всех пар запрос-мишень. Метод успешно идентифицировал кластеры

тандемно дуплицированных экзонов в генах, в которых такие кластеры были известны из литературы [62; 64—66]. Чтобы обнаружить новые неаннотированные тандемные дупликации экзонов, пары запрос-мишень, которые перекрывают любой аннотированный экзон, были исключены из рассмотрения, а мишени, которые пересекают аннотированные повторы или участки ДНК низкой сложности были отфильтрованы, поскольку последние могут содержать экзоны, возникшие по другому механизму, например, через экзонизацию транспозонов [109]. Затем были исследованы статистические свидетельства экспрессии вновь обнаруженных экзонов, используя данные секвенирования РНК из проекта Genotype Tissue Expression Project [12], а именно их покрытие короткими чтениями и поддержка экзон-экзонных соединений.

2.2.2 Примеры неаннотированных тандемных дупликаций

Обскурин (OBSCN)

Одним из примеров генов человека, подверженных тандемным дупликациям экзонов, является обскурин (*OBSCN*). Он охватывает более 150000 пар оснований и содержит более 80 экзонов [110]. Белок, кодируемый этим геном, принадлежит к семейству гигантских сакромерных сигнальных белков, в которое также входят титин и небулин [111]. *OBSCN* высоко экспрессируется в сердце (RPKM 8.6), простате (RPKM 2.9) и других тканях [12] (см. также разд. 2.2.3).

Подавляющее большинство экзонов обскурина гомологичны друг другу и имеют одинаковую длину, что позволяет предположить, что они образовались в результате тандемной дупликации (рис. 2.5). Наличие повторов в промежуточных интронах свидетельствует о том, что они возникли в результате нескольких раундов геномных дупликаций, вероятно, в результате негомологичной рекомбинации. Примечательно, что один из промежуточных интронов содержит область, которая гомологична другим экзонам, но не аннотирована как экзон (рис. 2.5, показана голубым цветом). Функциональность этой области подтверждается высокой степенью консервативности (phastCons) и наличием

сплит-чтений, поддерживающих экзон-экзонные соединения. Интересно, что тот же самый промежуточный интрон содержит и другую область с высоким уровнем консервативности, которая также содержит сплит-чтения, поддерживающие экзон-экзонные соединения. Однако эта область обладает меньшей степенью гомологии с другими экзонами (процент идентичности последовательностей 62,4% против 78,9% для других областей).

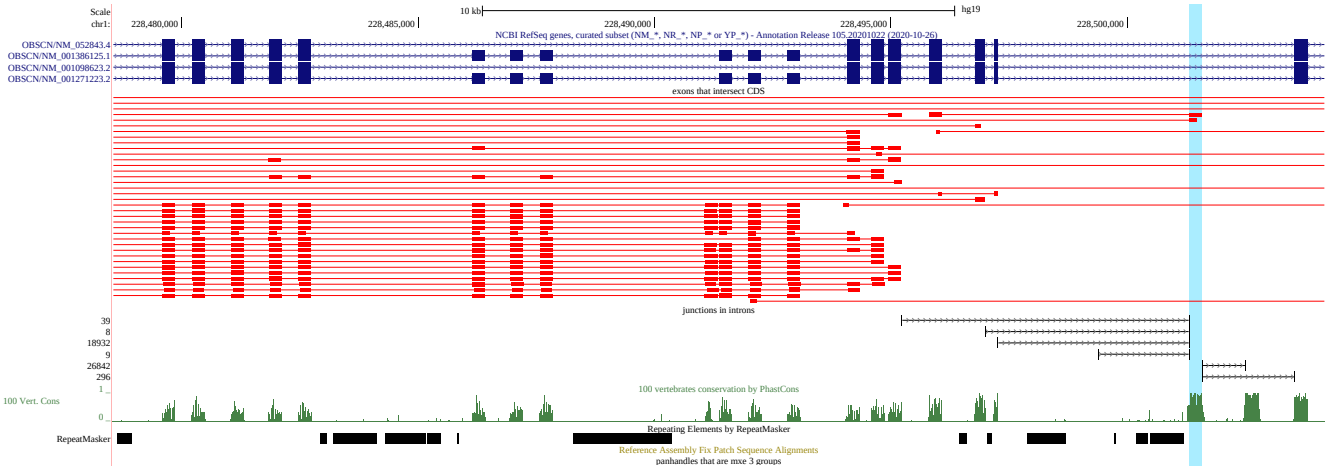

Рисунок 2.5 — Диаграмма тандемных экзонных дупликаций в обскурине *OBSCN*. Темно-синим цветом показаны аннотированные транскрипты (GENCODE и RefSeq). Красным цветом показаны пары запрос-мишень; запросы показаны толстыми прямоугольниками, а мишени – тонкими. Следующий трек показывает поддержку экзонных границ сплит-чтениями. Значения степени консервативности (PhastCons) показаны зеленым цветом.

UDP-глюкуронозилтрансфераза (UGT1A)

Ген *UGT1A* человека кодирует UDP-глюкуронозилтрансферазу и содержит тринадцать уникальных альтернативных начальных экзонов, за которыми следуют четыре конститутивных экзона. Этот ген ассоциирован с такими заболеваниями, как синдром Гилберта [112] и синдром Криглера-Наджара [113]. Каждый начальный экзон регулируется собственным промотором и кодирует сайт связывания субстрата, в результате чего образуются белки с разными N-концами и идентичными C-концами. Проведенный анализ показывает, что вариабельные начальные экзоны этих генов гомологичны друг другу (рис. 2.6),

что дает основания предположить, что они произошли в результате серии тандемных дупликаций. В 5'-НТО этого гена есть консервативная область, которая гомологична начальным экзонам, но не аннотирована как экзон (рис. 2.6). Следует отметить, что все начальные экзоны этого гена включаются в зрелый транскрипт взаимоисключающим образом.

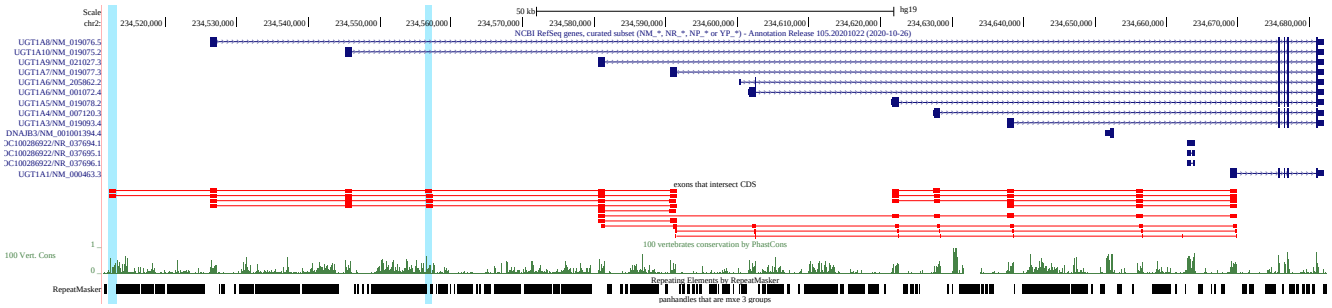

Рисунок 2.6 — Диаграмма тандемных экзонных дупликаций в гене UDP-глюкуронозилтрансферазы *UGT1A*; обозначения как и на рис. 2.5.

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

Два интересных примера тандемных дупликаций экзонов в нетранслируемых областях генов *D. melanogaster* представлены генами *hydra* и *pip* (рис. 2.7 и рис. 2.8). В гене *hydra* имеется девять гомологичных начальных экзонов, которые сплайсируются взаимоисключающим образом, тогда как в гене *pip* имеются восемь тандемно повторяющихся гомологичных кластеров взаимоисключающих терминальных экзонов. Было показано, что начальный экзон гена *hydra* подвергся рекуррентным дупликациям, и семь из этих альтернативных начальных экзонов фланкированы на своей 3'-стороне транспозоном DINE-1 [114]. По крайней мере четыре из девяти дуплицированных начальных экзонов могут функционировать как альтернативные сайты начала транскрипции [114]. Однако 3'-нетранслируемая область гена *pip*, который кодируют сульфотрансферазу и вносит вклад в формирование и полярность дорсально-вентральной оси эмбриона, изучена гораздо менее полно. Недавно было показано, что взаимоисключающее использование экзонов в 3'-НТО этого гена зависит от конкурирующих вторичных структур РНК [115].

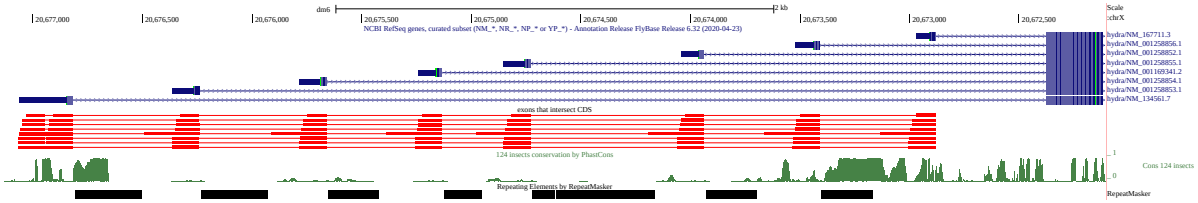

Рисунок 2.7 — Диаграмма тандемных экзонных дупликаций в гене *hydra* у *D. melanogaster*.

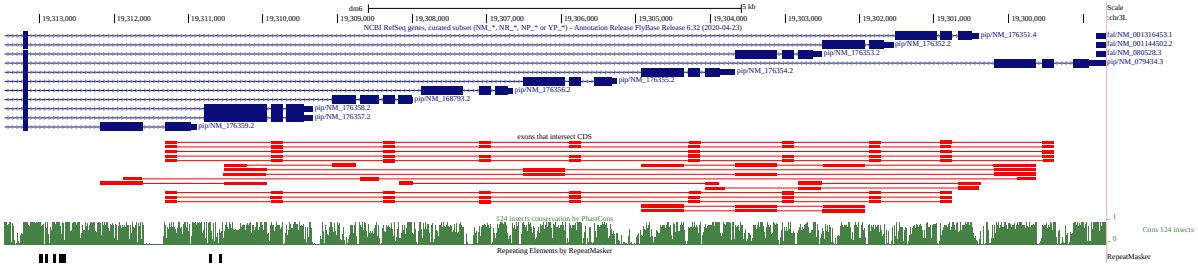

Рисунок 2.8 — Диаграмма тандемных экзонных дупликаций в гене *pip* у *D. melanogaster*.

2.2.3 Экспрессия тандемно дуплицированных экзонов

Для того, чтобы оценить экспрессию тандемных дупликаций экзонов с использованием транскриптомных данных, были рассмотрены пары запрос-мишень в генах человека, в которых мишень не пересекается ни с аннотированными экзонами, ни с повторами, и объединили оставшиеся мишени с помощью програмы bedtools merge. Эта процедура дала 4027 интронных мишеней. Каждой из этих мишеней случайным образом сопоставляли контрольную область такой же длины, которая располагалась на 30 нт 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 все короткие чтения, которые картировались более чем на одну позиции в геноме были исключены из анализа, затем было вычислено среднее покрытие чтениями для каждой мишени и для соответствующей контрольной области в каждом из транскриптомов 53 тканей из проекта Genotype Tissue Expression (GTEx) [12],

используя только однозначные картирования. Затем был вычислен показатель $\text{logFC}_i = \log_{10}(1 + TC_i) - \log_{10}(1 + CC_i)$, где TC_i — среднее покрытие мишени в ткани i, а CC_i — среднее покрытие контрольного региона в ткани i. Ткани с недостаточным количеством значений logFC_i (мочевой пузырь, эндоцервикс и эктоцервикс шейки матки) были исключены из дальнейшего анализа. В группе мишеней, которые обладали по меньшей мере 80% идентичностью нуклеотидной последовательности с запросом, наблюдалось значительное положительное отклонение метрики logFC_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, мышце, мозге, а также в некоторых трансформированных клетках (рис. 2.9). Аналогично определялась величина $\text{logFC}_i = \log_{10}(1+TS_i)-\log_{10}(1+CS_i)$, где TS_i — суммарное число сплит-чтений, поддерживающих границу мишени в ткани i, а CS_i — суммарное число сплит-чтений, поддерживающих границу запроса в ткани i. Наблюдалось увеличение количества сплит-чтений, поддерживающих экзон-экзонные соединения для тандемно дуплицированных экзонов с более высокой идентичностью нуклеотидных последовательностей (рис. 2.10). Эти результаты демонстрируют, что по крайней мере некоторые из неаннотированных тандемно дуплицированных экзонов действительно могут экспрессироваться, причем тканеспецифичным образом.

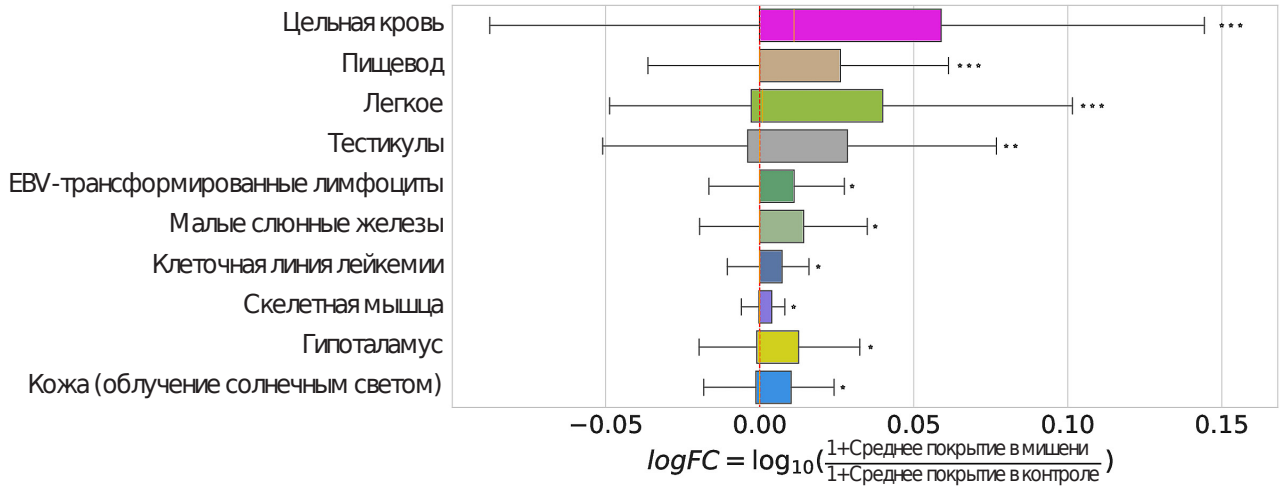

Рисунок 2.9 — Распределение значений метрики $\text{logFC}_i = \log_{10}(1 + TC_i) - \log_{10}(1+CC_i)$, где TC_i (CC_i, соответственно) равно среднему покрытию мишени (контрольного региона, соответственно) в ткани i, для мишеней с по крайней мере 80% идентичностью нуклеотидной последовательности.

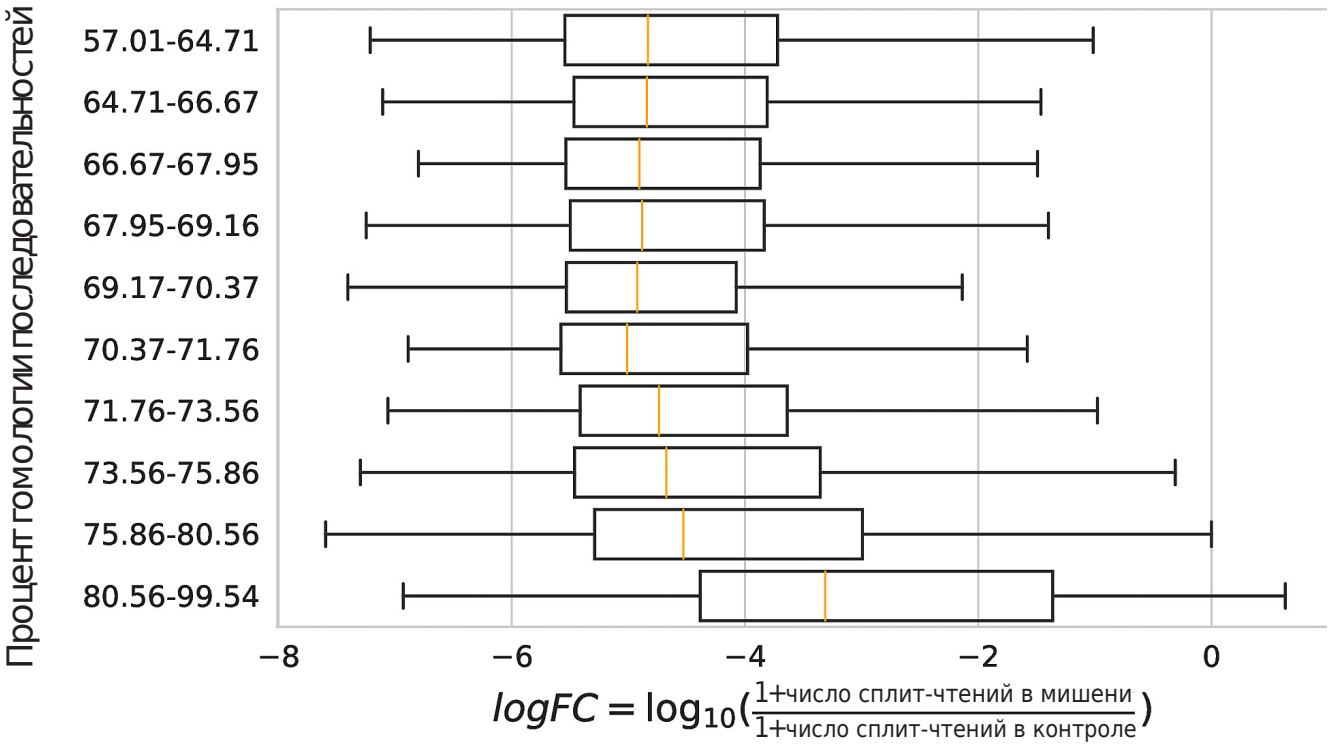

Рисунок 2.10 — Распределение значений метрики $\mathrm{logFC}_i = \log_{10}(1 + TS_i) - \log_{10}(1 + CS_i)$, где TS_i (CS_i, соответственно) равно суммарному числу уникально картированных сплит-чтений поддерживающих экзонные границы мишени (запроса, соответственно) в ткани i.

В заключение была вычислена разница между средними показателями степени консервативности [116], полученными из множественного выравнивания геномов 100 видов позвоночных, для каждой мишени и соответствующего ей контрольного региона. Мишени оказались в среднем более эволюционно консервативными, чем контрольные области (знаковый ранговый критерий Вилкоксона, P-значение = 0.009), что также указывает на их возможную функциональность.

В прил. А приводятся таблицы наиболее экспрессируемых областей в геноме человека, предположительно являющихся тандемными дупликациями экзонов из нетранслируемых (табл. 3) и кодирующих областей (табл. 4) с указанием их координат, ткани, в которой они экспрессируется, среднего уровня покрытия чтениями, среднего уровня консервативности и числа поддерживающих сплит-чтений. В данном случае наибольшее число неаннотированных экспрессирущихся мишеней обнаруживаются в ткани легкого, а не в крови (см рис. 2.9), так как сообщаются абсолютные уровни экспрессии.

2.3 Обсуждение и выводы

Расширение существующих результатов о представленности тандемных дупликаций экзонов, полученное в данной работе и находящееся в открытом онлайн репозитории https://zenodo.org/record/5474863, является полезным ресурсом для ученых, изучающих сплайсинг эукариотических генов. В частности, наблюдения, сделанные для обскурина и других генов, показывают, что существующая аннотация тандемно дуплицированных экзонов не полна, и что найденные мишени, поддерживаемые высокой степенью эволюционной консервативности и сплит-чтениями из данных секвенирования РНК, соответствуют неизвестным ранее экзонам. Также следует отметить, что не все тандемные дупликации экзонов могут быть обнаружены с помощью метода `exonerate` из-за низкого процента идентичности последовательностей и особенностей реализации этого метода.

Интересное наблюдение, сделанное в этой работе, заключается в том, что тандемные дупликации экзонов преобладают не только в кодирующих областях, но также и в нетранслируемых областях эукариотических генов и, более того, они, по-видимому, связаны с взаимоисключающим выбором тандемно дуплицированных начальных и конечных экзонов. Таким образом, частое образование ВИЭ в результате тандемных дупликаций относится не только ко внутренним, но и к терминальным экзонам в нетранслируемых частях генов, что очередной раз заставляет задуматься о том, не могут ли тандемные дупликации экзонов в нетранслируемых областях контролироваться конкурирующими структурами РНК некоторым общим образом.

Действительно, 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 [115] (рис. 2.11). Эти структуры РНК совместно регулируют отбор 3'-НТО посредством активации проксимального 3'-сайта сплайсинга и одновременного подавления интрон-проксимального 5'-сайта сплайсинга вместе со стерической конкуренцией за спаривание РНК [115]. Сходная регуляторная программа также действует в 3'-вариабельных областях генов *D. melanogaster CG42235* и *pip*.

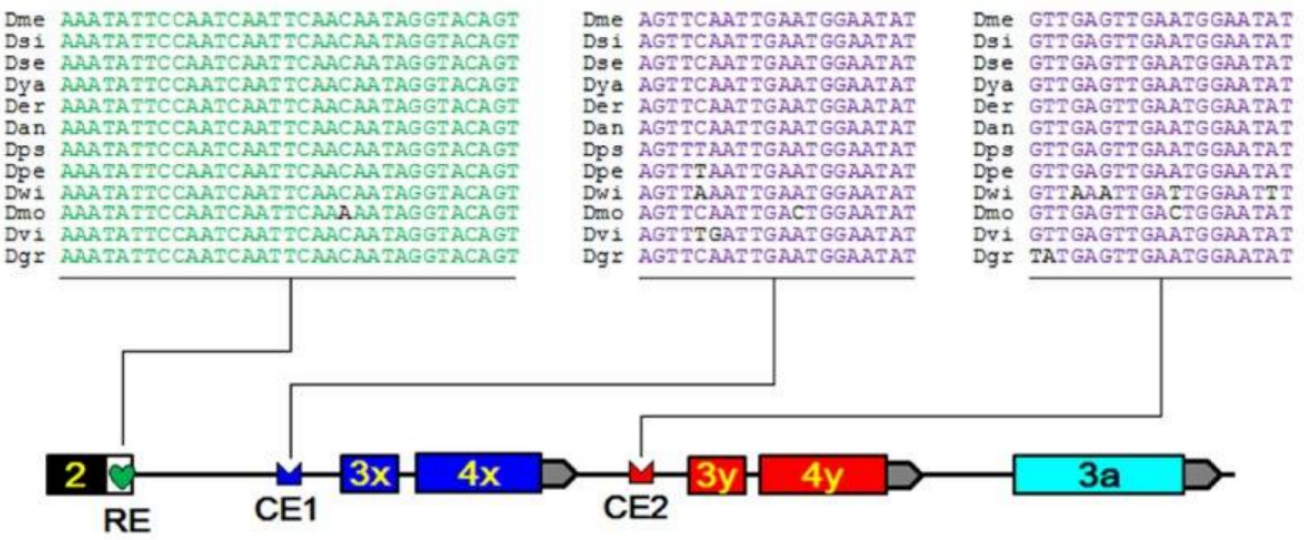

Рисунок 2.11 — Расположение цис-регуляторных элементов в вариабельной 3'-НТО гена *PGRP-LC* у *D. melanogaster*. Включение изоформ PGRP-LCx и LCy обусловлено конкурирующими парами РНК. Изображение заимстовано из [115].

Таким образом, тандемные дупликации экзонов широко представлены не только в кодирующих частях, но и в нетранслируемых областях эукариотических генов. Являются ли конкурирующие РНК-структуры вовлеченными в регуляцию взаимоисключающего сплайсинга этих экзонов и могут ли они образовываться как побочный продукт тандемных геномных дупликаций, пока остается открытым вопросом.

Глава 3. Конкурирующие структуры РНК и ВИЭ могут возникать в результате тандемных дупликаций

В этой главе будет получен предположительный ответ на вопрос о том, каким именно образом тандемные дупликации приводят к образованию ВИЭ и какую роль в их образовании могут играть конкурирующие структуры РНК. Результаты, представленные в этой главе, опубликованы автором в работе [117].

На протяжении всей этой главы используются обозначения, показанные на рис. 3.1. Каждый кластер ВИЭ расположен между двумя фланкирующими конститутивными экзонами (экзон 1 и экзон 3) и состоит из n альтернативных экзонов, обозначенных $2.1, \ldots, 2.n$, и $n+1$ и промежуточных интронов, пронумерованных от 0 до n. Интроны, фланкирующие конститутивные экзоны, т. е. интроны 0 и n, также будут называться “левый” и “правый”, соответственно. Система координат независима от цепи, т. е. левый интрон находится ближе к 5’-концу пре-мРНК, чем правый.

3.1 Методы

Геномы и аннотации

Геномные последовательности и аннотации транскриптомов человека и *Drosophila melanogaster* были получены так же, как и в разд. 2.1.1. Как и ранее, из файлов аннотации генома были извлечены координаты интронов и экзонов,

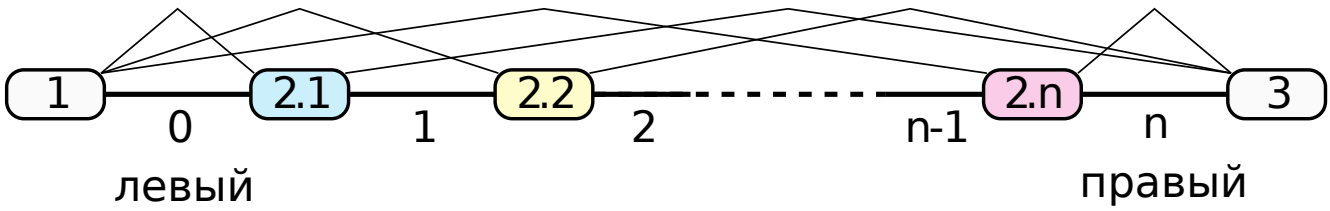

Рисунок 3.1 — Экзоны 1 и 3 конститутивны; экзоны $2.1, 2.2, \ldots, 2.n$ являются взаимоисключающими. Интроны пронумерованы от 0 до n. Концевые фланкирующие интроны (0 и n) называются “левым” и “правым”, соответственно.

а нуклеотидные последовательности интронов и экзонов выделяли с помощью инструмента `bedtools getfasta` [105].

3.1.1 Граф сплайсинга

Граф сплайсинга определялся как двудольный граф, узлами которого являются аннотированные сайты сплайсинга, а ребрами — экзоны или интроны аннотированных транскриптов. Для нахождения кластеров ВИЭ была написана программа, которая сопоставляет каждой паре вершин a и b множество путей из a в b в графе сплайсинга и отбирает вершинно-независимые пути, т.е. пути, не имеющие общих вершин кроме начала и конца. Затем в графе сплайсинга были выбраны пары вершин с двумя или более вершинно-независимыми путями, которые содержат ровно один экзон и два интрона. Такие пары вершин соответствуют кластерам ВИЭ, а количество вершинно-независимых путей между ними равно количеству ВИЭ в кластере.

Большинство аннотированных кластеров ВИЭ состоят только из двух экзонов. У *D. melanogaster* насчитывается 126 кластеров ВИЭ; из них 102 состоят из двух экзонов, остальные имеют кластеры из трех и более ВИЭ. У человека аннотация транскриптома содержит 526 двухэкзонных кластеров ВИЭ, а все остальные аннотированные кластеры содержат три или более ВИЭ.

3.1.2 Консервативность и анализ гомологии

В качестве меры эволюционной консервативности отрезка нуклеотидной последовательности генома была использована доля нуклеотидов, идентифицированных как эволюционно консервативные с помощью филогенетической скрытой марковской модели (philo-HMM) [116]. Для этого списки консервативных участков `phastConsElements`, вычисленные предварительно авторами метода по множественным выравниваням последовательностей 100 позвоночных и 15 насекомых, соответственно, были загружены с веб-сайта геном браузера UCSC [14].

Для оценки степени идентичности между двумя нуклеотидными последовательностями были использованы локальные выравнивания Смита-Уотермана, реализованные в библиотеке pairwise2 в модуле Biopython [118]. Была использована функция `pairwise2.align.localms` с параметрами match = 1, mismatch = −0.2, gap opening = −1, gap extension = −0.5. Последовательности длиной более 10 000 нуклеотидов не рассматривались. Нормализованная мера сходства σ между двумя последовательностями x и y вычислялась как

$$\sigma(x,y) = \frac{s(x,y)}{\max\{s(x,x), s(y,y)\}}, \tag{3.1}$$

где $s(x,y)$ — значение целевой функции локального выравнивания последовательностей x и y. Эта мера принимает значения от 0 до 1 и достигает максимального значения, когда x и y последовательности идентичны.

3.1.3 Перемешивание последовательностей и контроль

В качестве контроля для каждой нуклеотидной последовательности была сгенерирована перемешанная последовательность, которая содержит точно такие же частоты динуклеотидов, как и исходная, но отличается от нее частотами тринуклеотидов. Для этого была использована следующая процедура [119]. В исходной последовательности случайно выбираются три идущих последовательно нуклеотида x,y,z, находится вхождение динуклеотида x,z, а затем перестанавливаются так, чтобы x,y,z стало x,z и наоборот. По построению эта процедура не влияет на частоты динуклеотидов. Перестановки повторяются l раз, где l — длина последовательности.

Другая контрольная процедура заключалась в замене каждого интрона (экзона) случайно выбранным интроном (соответственно, экзоном), совпадающим по длине. Для этого нуклеотидные последовательности в каждом классе интервалов (интрон, экзон) сортировались по возрастанию длины, а затем каждому интервалу случайно сопоставлялся другой, отстоящий от него в отсортированном списке не более чем на 10 позиций. При этом длины интервалов в среднем не меняются, а нуклеотидные последовательности случайным образом заменяются на другие.

3.1.4 Свободная энергия гибридизации

Вторичные структуры РНК, связанные со взаимоисключающим сплайсингом, принадлежит к классу структур дальнего действия, поэтому их следует рассматривать как межмолекулярные [120; 121]. Для оценки свободной энергии гибридизации двух молекул РНК были использованы следующие программы из пакета ViennaRNA [122]: `RNAup`, которая моделирует термодинамику взаимодействий РНК-РНК, дополнительно оценивая энергию, необходимую для открытия сайта связывания, и энергию, полученную в результате гибридизации [123], и `RNAplex` — программа, которая позволяет быстро находить возможные сайты гибридизации для двух РНК, используя предварительно вычисленные профили доступности с приблизительной энергетической моделью [124]. Обе программы запускались с ключом –noLP для избегания однонуклеотидных спариваний. Программы запускались на парах исходных последовательностей, а затем снова на парах контрольных последовательностей, либо перемешанных с сохранением частот динуклеотидов, либо выбранных случайным образом из одного и того же класса (интрон, экзон) схожих по длине (разд. 3.1.3). В результате этих действий были получены оценки минимальной свободной энергии гибридизации.

Разность минимальных свободных энергий (minimum free energy, MFE) гибридизации, $\Delta\Delta G = \Delta G - \Delta G_0$, где ΔG — минимальная свободная энергия гибридизации для истинной пары, а ΔG_0 — минимальная свободная энергия гибридизации для контрольной пары, использовалась как скорректированная мера способности двух нуклеотидных последовательностей к комплементарному спариванию. Значения $\Delta\Delta G$, превышающие 30 ккал/моль по абсолютной величине, отбрасывались. В случаях, когда требовалось ограничить анализ на консервативную часть последовательности, последняя пересекалась с интервалами `phastConsElements` (разд. 3.1.2) с помощью программы `intersectBed`, а полученные части конкатенировались. Найденные комплементарные регионы были визуализированы с помощью UCSC геном браузера [14].

3.1.5 Статистические методы

Данные были проанализированы и визуализированы с использованием программного обеспечения `R Statistics` версии 3.4.1 и пакета `ggplot2`. Статистические тесты проводились без поправки на конечный размер генеральной совокупности на выборках геномных интервалов (экзоны или интроны), рассматриваемых как простые случайные выборки, взятые из множества всех таких геномных интервалов. Приводятся односторонние P-значения. Значимые различия на 5%, 1%, и 0.1% уровне значимости на всех рисунках обозначены *, **, ***, соответственно.

3.2 Результаты

3.2.1 Эволюционная консервативность фланкирующих ВИЭ интронов

Для того, чтобы выяснить, отличается ли степень эволюционной консервативности интронов, фланкирующих ВИЭ, по сравнению с интронами, фланкирующими другие классы экзонов, была использована доля интронных нуклеотидов, принадлежащих `phastConsElements` (разд. 3.1.2), как показатель средней степени консервативности по интронам [14]. Эта доля вычислялась для для каждого интрона, фланкирующего ВИЭ, и сравнивалась с соответствующей долей для случайно выбранного интрона схожей длины (рис. 3.2).

Оказалось, что у *D. melanogaster* (соответственно, *H. sapiens*) в среднем 29,0% (соответственно 10.5%) нуклеотидов эволюционно консервативны в интронах, фланкирующих ВИЭ, в то время как в случайно выбранных интронах соответствующие доли составляют 18,3% (соответственно, 6%), и в обоих случаях разница статистически значима (критерий Манна-Уитни, P-значение $< 10^{-6}$). Таким образом, фланкирующие ВИЭ интроны обладают повышенной консервативностью, что указывает на то, что они могут содержать функцио-

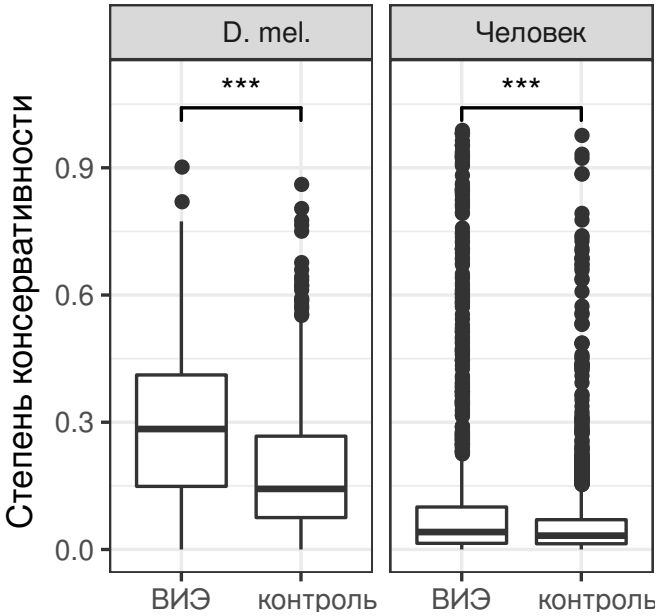

Рисунок 3.2 — Сравнение степени консервативности в интронах, фланкирующих ВИЭ, и степени консервативности в случайной выборке интронов, фланкирующих другие экзоны (контроль) у *D. melanogaster* и человека.

нальные регуляторные элементы, которые обеспечивают взаимоисключающий сплайсинг.

3.2.2 Степень идентичности фланкирующих интронов внутри кластера ВИЭ

Предположение о том, что некоторые кластеры ВИЭ возникли в результате тандемных геномных дупликаций основано на том наблюдении, что экзоны внутри кластеров ВИЭ часто гомологичны друг другу и имеют одинаковую длину [2; 11]. В то время как сходство экзонов сохраняется из-за эволюционных ограничений на кодирующую последовательность, гораздо более слабый отбор действует на интронные последовательности. Однако не известно, остались ли следы сходства между интронами в кластерах ВИЭ.

Для того, чтобы ответить на этот вопрос, для каждой пары соседних интронов, фланкирующих ВИЭ, была вычислена мера сходства $\sigma(x_i,x_{i+1})$ их последовательностей, которая затем была нормализована до максимального значения, которое она могла бы иметь, если бы последовательности были идентичны (разд. 3.1.2). Эта мера улавливает короткие похожие мотивы в соседних

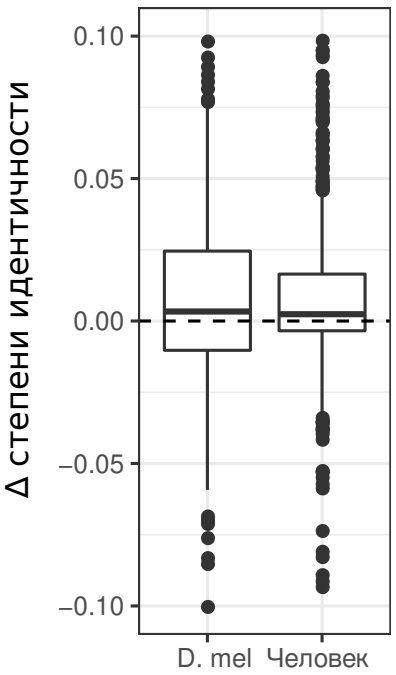

Рисунок 3.3 — Разность степени идентичности Δстепени идентичности = $\sigma - \sigma_0$, где σ — это степень идентичности соседних интронов, i и $i+1$, фланкирующих ВИЭ, а σ_0 — степень идентичности соседних интронов в контрольной выборке сопоставимых по длине интронов, у *D. melanogaster* и человека.

интронах, причем чем длиннее общий мотив, тем больше показатель сходства. В качестве контроля были вычислены соответствующие показатели для контрольной выборки интронов, фланкирующих экзоны, которые не являются ВИЭ. Ожидалось, что похожие мотивы, если они существуют, будут систематически смещать $\sigma(x_i,x_{i+1})$ в сторону больших значений в интронах, фланкирующих ВИЭ, по сравнению с другими интронами.

Действительно, выборка соседних интронов, фланкирующих ВИЭ, обладает значимо большим средним значением $\sigma(x_i,x_{i+1})$, чем контрольная выборка (критерий Манна-Уитни, P-значение $< 10^{-6}$). Поскольку более длинные интроны с большей вероятностью будут содержать похожие мотивы по случайным причинам, для ответа на этот вопрос необходимо получить контрольную выборку из интронов той же длины и изучить соответствующую выборку разностей (рис. 3.3). Как и ожидалось, медианная разность $\sigma(x_i,x_{i+1})$ была значимо больше нуля (критерий Уилкоксона, P-значение $< 10^{-5}$ у *D. melanogaster* и P-значение $< 10^{-10}$ у человека, соответственно), хотя и мала по абсолютной величине (медиана разности 2% и 1%, соответственно). Положительный сдвиг указывает на то, что фланкирующие ВИЭ соседние интроны могут содержать больше гомологичных регуляторных элементов, таких как селекторные после-

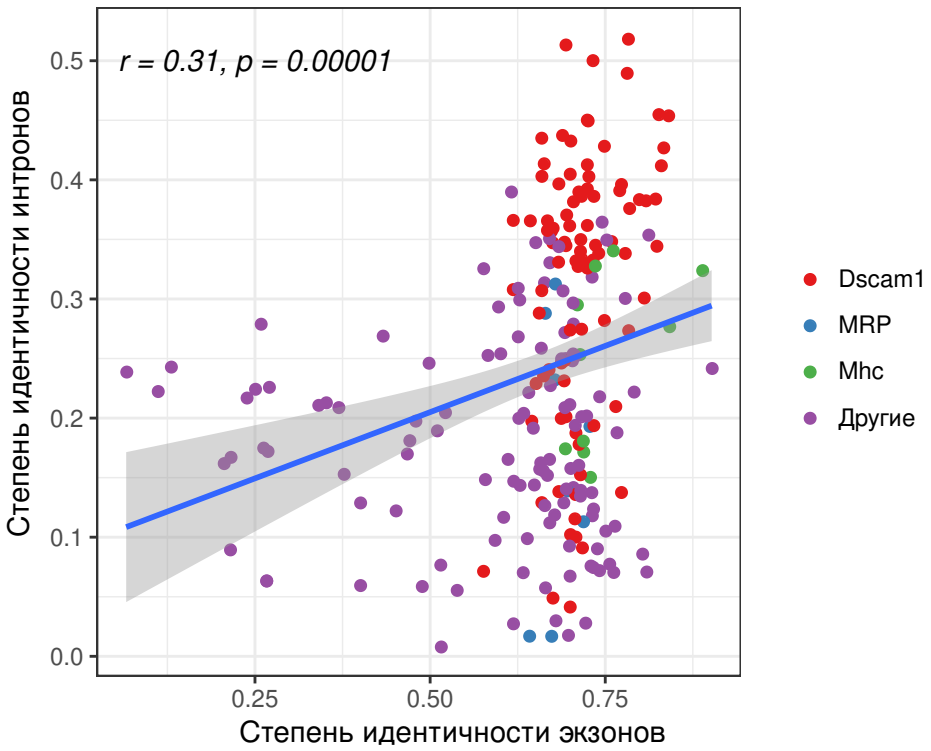

Рисунок 3.4 — Степень идентичности соседних интронов, i и $i+1$, (ось ординат) по сравнению со степенью идентичности соседних экзонов, i и $i+1$, (ось абсцисс) у *D. melanogaster*.

довательности, чем другие интроны. При этом небольшое абсолютное значение этого сдвига не является неожиданным, так как в известных случаях оно составляет порядка 1-2% для консервативной структуры из 14 нуклеотидов в 600-нуклеотидном интроне [125].

Также было обнаружено, что степень идентичности двух соседних интронов в кластерах ВИЭ положительно коррелирует со степенью идентичности фланкируемых ими экзонов ($r = 0.31$, рис. 3.4), причем у *D. melanogaster* эта корреляция остается значимой ($r = 0.18$, $n = 139$, t-критерий, P-значение $= 0.017$) даже после удаления кластера ВИЭ из гена *Dscam1*, гомология экзонов в котором хорошо известна, т.е., чем больше похожи соседние ВИЭ, тем больше похожи их фланкирующие интроны. Поэтому для ВИЭ, которые возникли в результате тандемных дупликаций, это наблюдение указывает на то, что дупликация затронула не только экзон, но и часть соседнего интрона.

Таким образом, соседние интроны, фланкирующие ВИЭ, в среднем более гомологичны друг другу, чем интроны, фланкирующие другие классы экзонов, а степень их идентичности коррелирует с степенью идентичности соответствующих экзонов. Эти наблюдения указывают на то, что часть интронного пространства была дуплицирована вместе с экзоном и оставалась под давлени-

ем отбора, которое, как было показано в разд. 3.2.1, выше во фланкирующих ВИЭ интронах.

3.2.3 Комплементарные спаривания во фланкирующих ВИЭ интронах

Поскольку фланкирующие ВИЭ интроны подвергаются более сильному отбору, чем другие интроны, то разумно спросить, не может ли это быть связано с комплементарными парами оснований. Для ответа на этот вопрос программа `RNAup` применялась ко всем попарным комбинациям левых и правых интронов по сравнению с внутренними интронами в трех больших кластерах ВИЭ (экзон 4, экзон 6 и экзон 9) гена *Dscam1* у *D. melanogaster*. Минимальная свободная энергия гибридизации предсказывалась для каждой комбинации, а затем ту же процедуру применяли к последовательностям, перемешанным с сохранением динуклеотидного состава. Разность между свободной энергией для исходных и перемешанных пар, $\Delta\Delta G$, отражает склонность двух последовательностей к комплементарному спариванию, причем более отрицательные значения $\Delta\Delta G$ указывают на то, что настоящие последовательности складываются в более стабильные РНК структуры, чем перемешанные.

В соответствии с тем, что известно о конкурирующих структурах РНК в *Dscam1*, наблюдалась тенденция к формированию структуры РНК во всех трех ее рассмотренных кластерах (рис. 3.5). Однако в то время как кластер экзона 4 (12 экзонов) и кластер экзона 9 (33 экзона) содержат докерный сайт в правом интроне, что полностью соответствует литературным сведениям [5], кластер экзона 6 (48 экзонов), как оказалось, имеет склонность к образованию комплементарных спариваний за счет докерных сайтов как в левом, так и в правом интроне.

Затем изучались предсказания `RNAup` для внутримолекулярных спариваний в кластере экзонов 6 (рис. 3.6). В левом интроне программа `RNAup` правильно идентифицировала уникальную последовательность известного докерного сайта, в то время как спаривания с правым интроном предсказали несколько кандидатных докерных сайтов. Примечательно то, что правые докерные последовательности располоожены не случайно, а образуют несколько

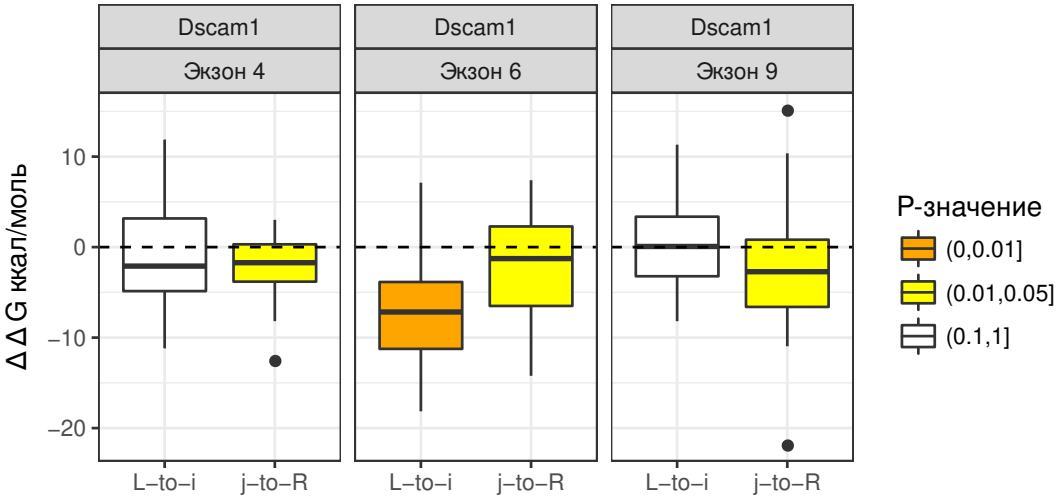

Рисунок 3.5 — Изменение минимальной свободной энергии ($\Delta\Delta G$, ккал/моль) по сравнению с перемешанными последовательностями для гибридизации левого (L-to-i) или правого (j-to-R) интронов со внутренними интронами для трех кластеров ВИЭ гена *Dscam1* у *D. melanogaster*.

групп, которые перекрывают консервативные области. Это наблюдение согласуется с моделью двунаправленного контроля, которая была предложена для кластеров экзонов 4 и экзонов 9 *Dscam*, хотя и у других видов [126].

Склонны ли левый и правый интроны образовывать комплементарные спаривания с внутренними интронами в других кластерах ВИЭ? Поскольку программа `RNAup` не эффективна для длинных последовательностей, то для ответа на этот вопрос была использована программа `RNAplex`, которая не учитывает внутримолекулярную структуру и поэтому может быть использована в тех же условиях для обнаружения комплементарности в интронах с перемешанными последовательностями (рис. 3.7). При применении `RNAplex` к консервативным частям интронных последовательностей статистически значимое изменение свободной энергии гибридизации с внутренними интронами наблюдается для левого и правого интронов (критерий Уилкоксона, P-значение < 0.005), но не для гибридизации внутренних интронов друг с другом (P-значение $\simeq 0.45$).

Таким образом, для левых и правых интронов в аннотированных кластерах ВИЭ наблюдается тенденция образовывать комплементарные спаривания с внутренними интронами, а для спаривания внутренних интронов друг с другом такой тенденции не наблюдается. Это говорит о том, что механизм взаимоисключающего выбора экзонов, основанный на спаривании докерного и селекторных сайтов, который известен для кластеров ВИЭ гена *Dscam1* у *D. melanogaster*, может быть также присущ и многим другим кластерам ВИЭ.

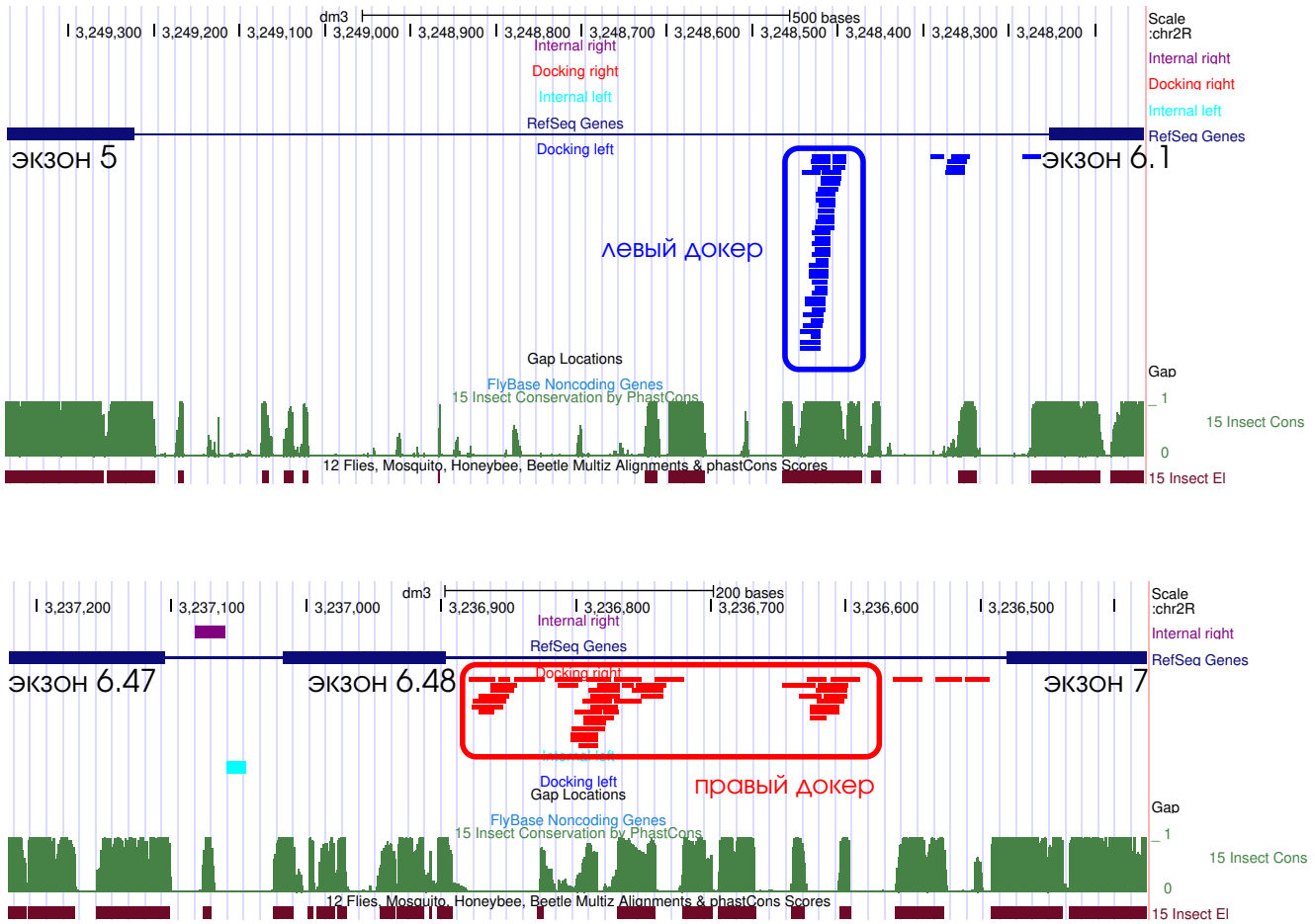

Рисунок 3.6 — Предсказанные докерные сайты в левом интроне (синий) и правом интроне (красный) кластера экзонов 6 гена *Dscam1* у *D. melanogaster*.

Это, а также следы гомологии между сосдениими интронами, наводят на мысль о том, что, возможно, существует эволюционный механизм, ответственный за конвергентную эволюцию сплайсинга ВИЭ, который связан с их происхождением в результате тандемных геномных дупликаций. В разд. 3.3 мы предлагаем один из таких возможных механизмов.

3.3 Обсуждение и выводы

3.3.1 Общее происхождение ВИЭ и конкурирующих структур РНК

С эволюционной точки зрения дупликацию экзонов можно рассматривать как способ создания функционального разнообразия белков, наряду с дуплика-

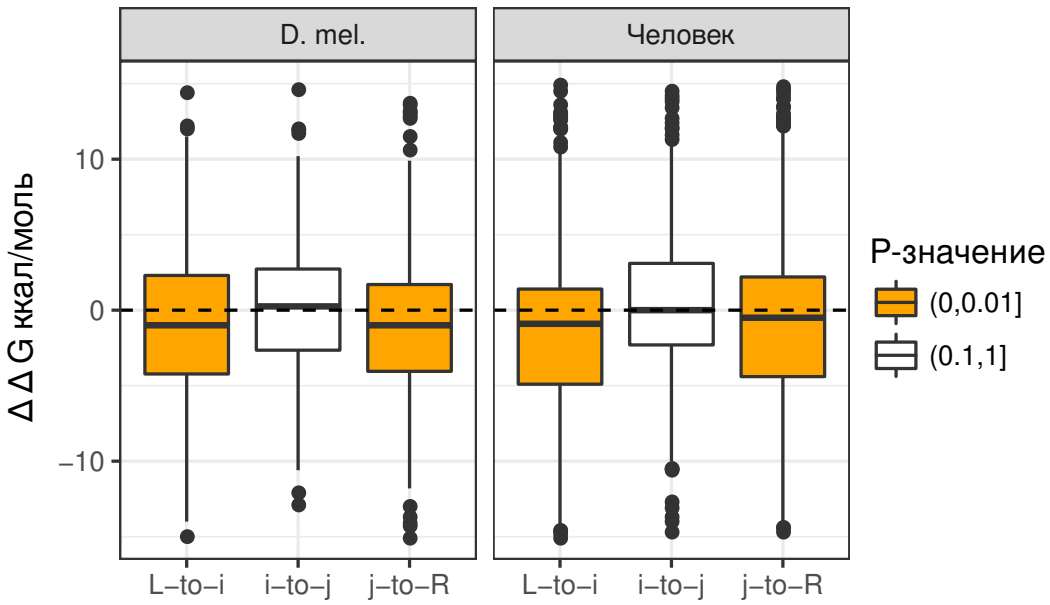

Рисунок 3.7 — Изменение минимальной свободной энергии ($\Delta\Delta G$, ккал/моль) по сравнению с перемешанными последовательностями для гибридизации левого и внутренних интронов (L-to-i), гибридизации правого и внутренних интронов (j-to-R) и гибридизации между внутренними интронами (i-to-j) у *D. melanogaster* и человека. Цветовые коды показывают значимые отличия от нуля по знаковому критерию Уилкоксона.

цией генов [2]. Дупликации экзонов могут происходить с помощью различных механизмов, включая события, опосредованные транспозонами, и негомологичную рекомбинацию [28]. В результате часть генной последовательности тандемно дуплицируется вместе с регуляторными сигналами, определяющими экзон-интронную структуру.

В этой главе было продемонстрировано, что фланкирующие ВИЭ интроны более консервативны, чем другие интроны, что они более похожи друг на друга внутри кластеров ВИЭ по сравнению со случайно выбранными интронами сопоставимой длины, и что степень этого сходства растет по мере увеличения сходства между соседними экзонами. Эти наблюдения указывают на то, что тандемные дупликации могут затрагивать как экзонную, так и интронную части гена. Поскольку, как было показано, первый и последний интроны в кластерах ВИЭ обладают высокой склонностью к гибридизации с внутренними интронами, вместе все эти наблюдения позволяют предположить следующую модель.

Рассмотрим дупликацию, затрагивающую область генома, содержащую экзон (экзон 2) и часть левого фланкирующего интрона (рис. 3.8). Предположим дополнительно, что интрон, находящийся между экзонами 1 и 2, содержит

пару комплементарных последовательностей, a и a', которые способны образовывать шпилечную структуру, и что только одна из двух комплементарных частей, a', подверглась дупликации. Такие структуры распространены в эукариотических интронах и их часто связывают с конститутивным и альтернативным сплайсингом [127]. На самом деле они могут охватывать большие расстояния и функционировать как элементы, сближающие удаленные сайты сплайсинга [128]. В результате такой дупликации две копии экзона 2 располагаются тандемно, а a' и его копия a'' комплементарны a. Это создает пару конкурирующих структур РНК, в которых a спаривается либо с a', либо с a''; в первом случае включается экзон 2.1, а во втором случае он выпетливается и пропускается. Этот сценарий дает правдоподобное объяснение тому, как последовательности докерных и селекторных сайтов могут самопроизвольно возникать в результате тандемных дупликаций. Аналогичным образом дупликации могут создавать конкурирующие структуры РНК с правым докерным сайтом (рис. 3.9).

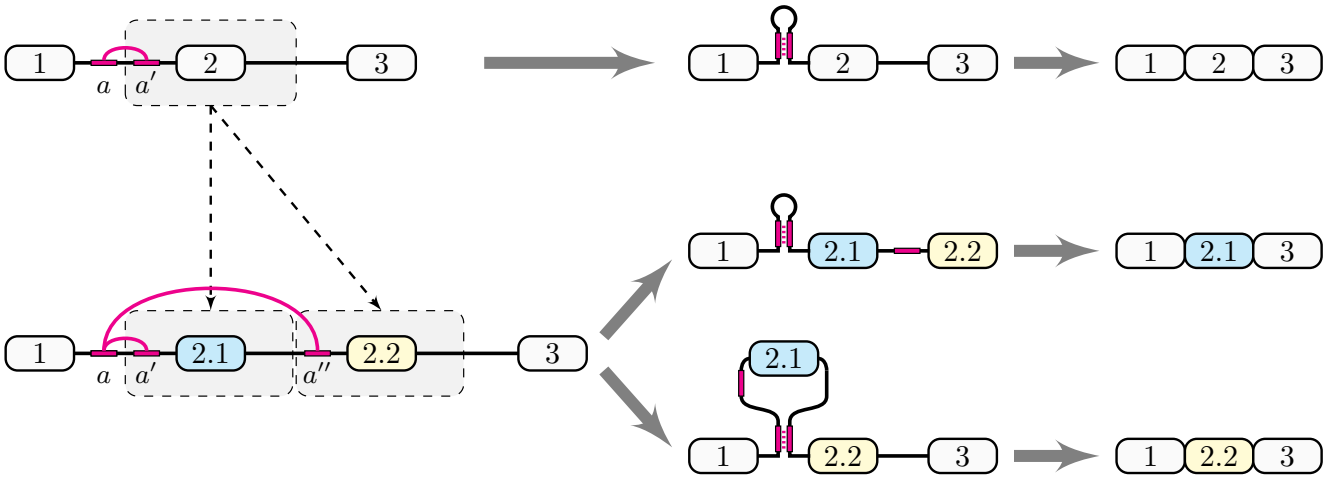

Рисунок 3.8 — Механизм образования конкурирующих структур РНК с левым докерным сайтом посредством дупликации. Если дупликация затрагивает экзон и одно плечо шпилечной структуры, расположенной в левом фланкирующем интроне, то образуется пара селекторных сайтов, которые могут конкурировать за один докерный сайт.

Однако молекулярный механизм включения ВИЭ в модели с докерными и селекторными сайтами сложнее, чем образование конкурирующих структур в пре-мРНК. Согласно существующим моделям он регулируется взаимодействием между активаторами и репрессорами сплайсинга, их соответствующими цис-регуляторными элементами и силой сайтов сплайсинга [129].

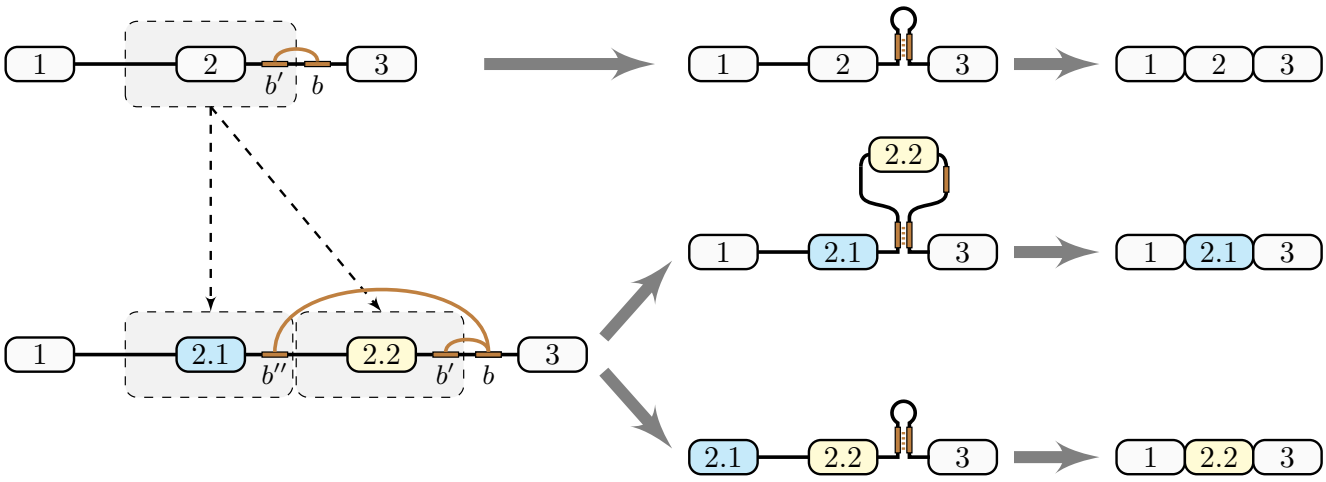

Рисунок 3.9 — Механизм образования конкурирующих структур РНК с правым докерным сайтом посредством дупликации (аналогично рис. 3.9).

Комбинация конкурирующих структур РНК, глобально действующего репрессора сплайсинга *hrp36* и более слабых сайтов сплайсинга ВИЭ совместно удерживают большинство альтернативных вариантов экзонов инактивированными. Взаимодействие докерного и селекторного сайтов селективно активирует экзон-мишень, способствуя распознаванию его сайтов сплайсинга, вероятно, в результате их пространственного сближения. Эта модель, однако, даёт механистическое объяснение сплайсинга только одного из двух альтернативных интронов. Считается, что сплайсинг второго интрона, не связанного с взаимодействием докерного и селекторного сайтов, может определяться силой сайта сплайсинга [129].

Недавно была предложена так называемая двунаправленная модель регуляции взаимоисключающего сплайсинга, основанная на структуре РНК внутри кластера экзонов 4 гена *srp* [126]. В этом гене две пары консервативных комплементарных интронных элементов окружают два альтернативных экзона. Каждая из двух пар способствует активации соответствующего экзона-мишени и инактивации второго экзона. Комплементарные спаривания в этих двух структурах не являются взаимоисключающими, но они образуют псевдоузел. Аналогичная двунаправленная модель применима к другим генам *D. melanogaster*, таким как *RIC-3* и *MRP1*, кластеру экзонов 4 перепончатокрылых и кластеру экзонов 9 чешуекрылых генов семейства *Dscam* [3; 126]. Следовательно, регуляция, основанная на двунаправленных взаимодействиях нескольких групп донорных и селекторных сайтов, может представлять собой общий механизм регуляции взаимоисключающего сплайсинга.

Небольшая модификация сценария, показанного на рис. 3.8, может объяснить эволюционный механизм, который способен генерировать также и двунаправленные конкурирующие структуры РНК. Рассмотрим геномную дупликацию, затрагивающую экзон, а также два его фланкирующих интрона, которые содержат две пары комплементарных последовательностей, a и a', а также b' и b (рис. 3.10). В результате такой дупликации две копии экзона 2 снова будут расположены тандемно, а также образуются две конкурирующие структуры РНК, в которых $a - a'$ конкурирует с $a - a''$, а $b' - b$ конкурирует с $b''-b$. Примечательно, что они будут расположены так, что b'' будет располагаться в направлении 5'-конца гена относительно a'', что совпадает с наблюдаемым расположением в гене *srp*. Несмотря на то, что каждая пара конкурирующих структур может образовываться независимо от другой пары, не все четыре комбинации равновероятны из-за псевдоузла. Если a спаривается с a', а b'' спаривается с b, то экзон 2.2 выпетливается и пропускается. Наоборот, если a спаривается с a'', а b' спаривается с b, то экзон 2.1 выпетливается и пропускается. По прежнему возможно, что a соединяется с a', а b' соединяется с b так, что ни один экзон не выпетливается, и это приводит к одновременному включению обоих экзонов. В отличие от однонаправленной модели (рис. 3.8), двунаправленная модель механически объясняет подавление обоих вариантов ВИЭ.

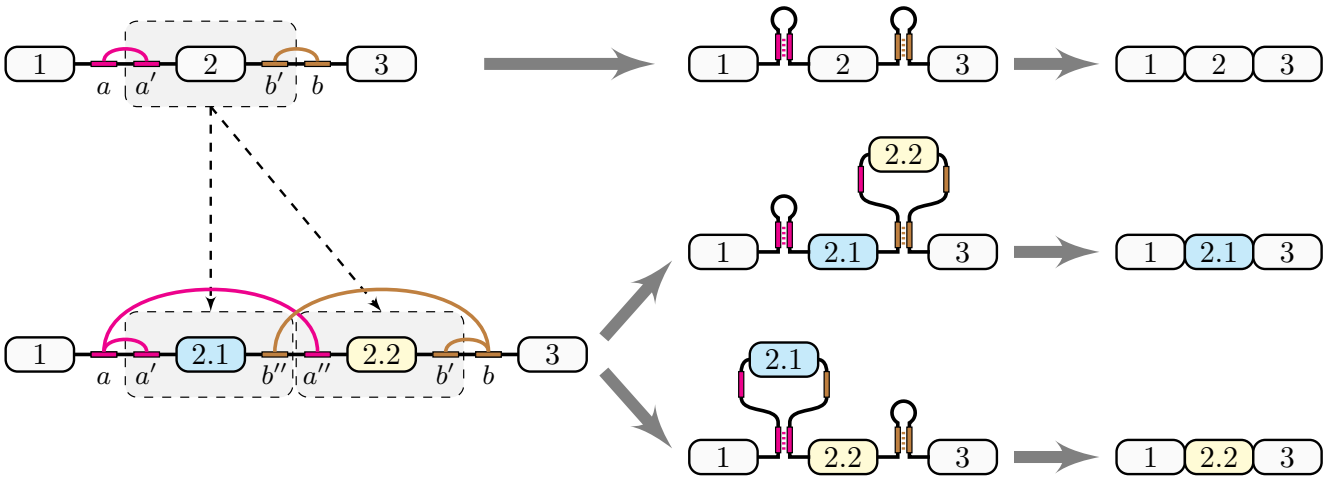

Рисунок 3.10 — Если тандемная дупликация затрагивает экзон и части двух его фланкирующих интронов, каждая из которых содержит шпилечную структуру, она может создать две пары конкурирующих комплементарных последовательностей, каждая из которых будет выпетливать один из экзонов 2.1 и 2.2. Расположение комплементарных частей в точности соответствует двунаправленной модели в гене *srp*.

3.3.2 Регулируемость ВИЭ конкурирующими структурами РНК

Расположение докерного и селекторных сайтов в модели регуляции взаимоисключающего сплайсинга допускает два варианта. В первом варианте докерный сайт расположен перед кластером ВИЭ в левом интроне, как это происходит в кластере экзонов 6 гена *Dscam1*. В другом варианте докерный сайт располагается после всех ВИЭ в правом интроне, как это происходит в генах *14-3-3ζ*, *Mhc*, *MRP1* и других кластерах ВИЭ [3; 5]. Двунаправленная модель допускает оба этих варианта одновременно. Повышение уровня консервативности и склонности к образованию конкурирующих РНК структур не одинаково для левого и правого интронов, с несколько более высокими показателями для последнего. Естественно возникает вопрос: почему некоторые гены предпочитают левые докерные сайты, в то время как другие используют правые для регуляции сплайсинга ВИЭ?

Хотя на этот вопрос нет очевидного ответа, фундаментальное различие между левыми и правыми докерными сайтами может заключаться в регулируемости их спаривания с селекторными последовательностями. Поскольку структура РНК формируется котранскрипционно [130], то спаривание $a' - a$ получает кинетическое преимущество по сравнению со спариванием $a - a''$, если докерный сайт расположен в левом интроне, в противоположность тому, что b'' и b' транскрибируются последовательно и получают равные шансы на спаривание с b, который появляется последним, если докер-сайт расположен в правом интроне. Хотя это может объяснить преобладание правых докерных сайтов в генах, в которых были обнаружены докерные и селекторные сайты (табл. 2), многие другие факторы также могут влиять на кинетику котранскрипционного фолдинга и сплайсинга пре-мРНК [131].

Другим важным фактором, который может влиять на формирование взаимоисключающих конформаций, является термодинамика структуры РНК. В то время как a' и его копия a'' (рис. 3.8 и рис. 3.10) могут образовывать одинаково стабильные дуплексы с a, другие структурные элементы, такие как петли, спирали $b - b'$ и псевдоузлы, также вносят свой вклад в свободную энергию структуры. Вопрос о том, может ли свободная энергия структуры РНК влиять на степень включения экзона, остается спорным. С одной стороны, соотношение альтернативных изоформ в пре-мРНК *Gug* меняется с увеличением свободной

ген	кластер ВИЭ	длина структуры	охват	докерный сайт
Dscam1	Экзон 4	13	4500	правый
Dscam1	Экзон 6	16	11000	левый
Dscam1	Экзон 9	16	14000	правый
Mhc	Экзон 7	14	2500	правый
Mhc	Экзон 9	14	1600	правый
Mhc	Экзон 11	15	2600	правый
srp	Экзон 4	21	450	оба
14-3-3-ζ	Экзон 5	22	1200	правый

Таблица 2 — Правые докерные сайты преобладают над левыми. Приводятся длина комплементарной части структуры (нт), охват, т.е., расстояние между комплементарными участками (нт), и позиция докерного сайта.

энергии дуплекса [125]. С другой стороны, Мэй и др. связали свободную энергию структуры РНК с изменением степени включения экзонов в минигенах *Dscam1* и обнаружили, что корреляция незначительна [55]. Взаимосвязь между свободной энергией и включением экзона *in vivo* должна быть еще сложнее, потому что докерные и селекторные сайты продолжали эволюционировать со времени дупликации для адаптации новых биологических функций. Анализ естественных переключателей между взаимоисключающими конформациями пре-мРНК является предметом будущих исследований.

Глава 4. Конкурирующие структуры РНК в кластере ВИЭ гена *Ate1* человека

В этой главе рассматривается кластер ВИЭ в гене *Ate1* человека и предсказывается существование в нем конкурирующих структур РНК. Результаты, представленные в этой главе, опубликованы автором в работе [15].

Аргинилирование — широко распространенная посттрансляционная модификация белка, при которой остаток L-аргинила переносится с Arg-тРНК на полипептидную цепь [132]. Аргинилирование выполняется аргинилтрансферазой *Ate1* [133], которая необходима в большинстве эукариотических систем и участвует в регуляции физиологических путей, включая протеолиз [134; 135], реакцию на стресс и тепловой шок [136—138], эмбриогенез [139—141], регенеративные процессы [142—144] и старение [145; 146]. Ген *Ate1* недавно был идентифицирован как основной регулятор, влияющий на некоторые процессы, связанные с заболеваниями [147—149], а его нокаут приводит к эмбриональной летальности и серьезным дефектам развития у мышей [140; 141; 150].

Как и большинство эукариотическиех генов, *Ate1* генерирует несколько изоформ мРНК посредством альтернативного сплайсинга [151]. У млекопитающих они отличаются взаимоисключающим выбором двух соседних гомологичных экзонов длиной 129 п.н. (экзоны 7a и 7b) и альтернативным выбором начального экзона (экзоны 1a или 1b) [152]. Двумя основными изоформами мРНК *Ate1* являются *Ate1-1* (1b7a) и *Ate1-2* (1b7b), в то время как изоформы, содержащие оба экзона 7a и 7b, подавляются через взаимоисключающий сплайсинг [149]. У мышей *Ate1-1* и *Ate1-2* стабильно экспрессируются во всех тканях, но их соотношение варьирует от 0,1 в скелетных мышцах до 10 в семенниках [151; 153; 154]. В то время как *Ate1-2* является почти полностью цитозольным, *Ate1-1* локализуется как в цитозоле, так и в ядре [151] и может специфически взаимодействовать с *Liat1* — молекулой, специфичной для семенников [155]. Кроме того, *Ate1*-нокаутные клетки могут образовывать опухоли в подкожном мышином ксенотрансплантате, где рост опухоли может быть частично подавлен повторным введением стабильно экспрессируемого *Ate1-1*, но не *Ate1-2* [147]. Изоформы *Ate1-3* (1a7a) и *Ate1-4* (1a7b) кодируют вариант аргинилтрансферазы, который специфичен для N-концевого цистеина с тканеспецифической экспрессией, клеточной локализацией и канцерогенным

потенциалом, подобным к таковым из *Ate1-1* и *Ate1-2* соответственно [152]. Соотношение изоформ *Ate1*, содержащих экзоны 7a и 7b, существенно меняется во время мейоза у самцов мышей, что указывает на роль в переходе от митотического к мейотическому циклу зародышевых клеток [154]. Все эти наблюдения позволяют предположить, что последовательности аминокислот, кодируемые экзонами 7a и 7b, приводят к функционально различным аргинилтрансферазам.

4.1 Методы

Данные секвенирования полиаденилированной фракции РНК были загружены в BAM формате с веб-порталов консорциумов Экспрессия Генотипа Ткани (Genotype-Tissue Expression, GTEx) и Атласа Ракового Генома (The Cancer Genome Atlas, TCGA) [12; 13]. Было получено 8,555 образцов из консорциума GTEx и 731 пара образцов опухоль–здоровая ткань из консорциума TCGA. С помощью программы IPSA из BAM файлов были получены частоты сплит-чтений, которые были отфильтрованы по энтропии (порог энтропии Шеннона 1.5 бит) и каноническим динуклеотидам GT/AG в сайтах сплайсинга [107]. Степень включения экзонов (ψ, PSI или Percent-Spliced-In) рассчитывалась в соответствии с уравнением

$$\psi = \frac{inc}{inc + 2{\cdot}exc},$$

где inc — количество ридов, поддерживающих включение экзона, а exc — количество ридов, поддерживающих его исключение. Значения ψ со знаменателем меньше 20 считались ненадежными и отбрасывались. Статистическую значимость логарифмического изменения соотношения изоформ экзонов 7a и 7b в образцах TCGA оценивали с помощью знакового рангового критерия Уилкоксона с поправкой Бонферони-Холма для множественного тестирования (групповая вероятность ошибки первого рода, familywise error rate, $FWER < 0.05$), примененного к отношению суммарного числа сплит-чтений, поддерживающих включение экзона 7a, к суммарному числу сплит-чтений, поддерживающих включение экзона 7b.

Для проведения сравнительного анализа последовательностей, множественное выравнивание последовательностей 45 геномов позвоночных с геномом человека (GRCh37) было получено из UCSC геном браузера в MAF формате [156]. Фрагмент гена *Ate1* был получен с помощью утилиты `maftools`.

4.2 Результаты

4.2.1 Экспрессия и распространенность изоформ

Распределение степени включения экзонов согласно данным проектов TCGA и GTEx говорит о том, что эти экзоны действительно сплайсируются взаимоисключающим образом в здоровых тканях (рис. 4.1) и опухолях (рис. 4.2). Экзоны 7a и 7b имеют широкий диапазон уровней включения в тканях (медианы 33% и 67% соответственно) с наиболее заметным отклонением в семенниках (медианы 60% и 39% соответственно).

Также наблюдается значительное увеличение числа изоформ, содержащих экзон 7a, по сравнению с экзоном 7b в образцах аденокарциномы простаты по сравнению с соответствующими значениями в нормальных тканях (FWER<0.05), а также в других эпителиальных опухолях, включая аденокарциному желудка, прямой кишки, толстой кишки и плоскоклеточный рак легкого (рис. 4.3).

Таким образом, изучение молекулярного механизма взаимоисключающего сплайсинга экзонов 7a и 7b гена *Ate1* человека представляется интересным для понимания как тканеспецифического, так и опухолеспецифического сплайсинга.

4.2.2 Консервативные конкурирующие структуры РНК

Для того, чтобы исследовать механизм, отвечающий за взаимоисключающий сплайсинг экзонов 7a и 7b, мы использовали сравнительный анализ

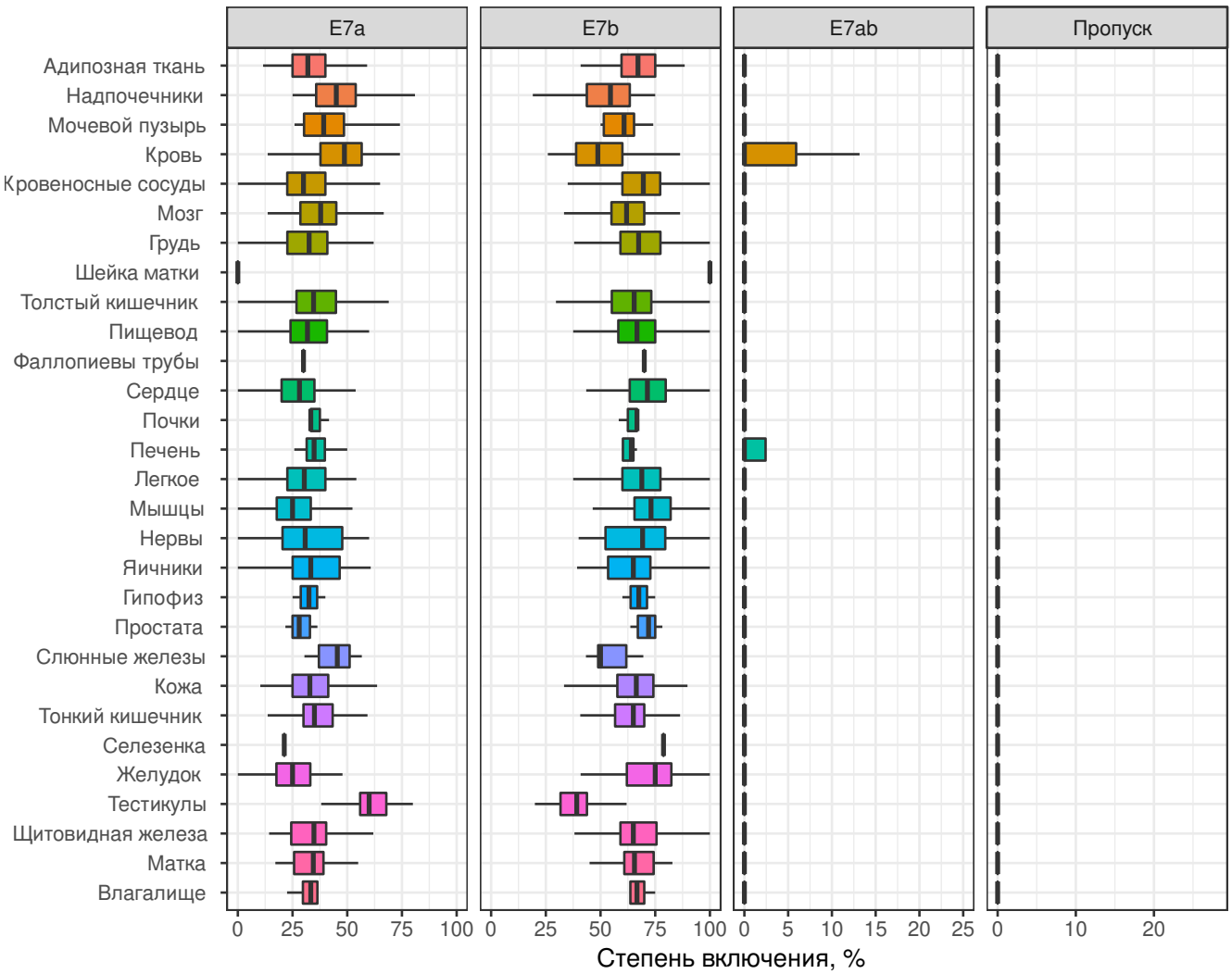

Рисунок 4.1 — Степень включения экзонов 7a, 7b, а также двойных экзонов (7a7b) и одновременного пропуска обоих экзонов (Пропуск) в тканях человека из проекта GTEx.

последовательностей гена *Ate1* у позвоночных для поиска потенциальных регуляторных последовательностей во фланкирующих интронах (рис. 4.4) и каталог консервативных комплементарных областей, опубликованный в недавней работе [157]. Последовательности, расположенные непосредственно перед экзонами 7a и 7b, обозначенные как R1 и R4, обладают высокой степенью идентичности и консервативны у разных видов позвоночных. Промежуток между экзонами 7a и 7b содержит две консервативные интронные последовательности, обозначенные как R2 и R3, причем R3 комплементарна как R1, так и R4, а R2 комплементарна другой высококонсервативной области R5, расположенной в интроне между экзонами 7b и 8 на расстоянии приблизительно 30 т.п.о. в направлении 3'-конца. Комплементарное спаривание оснований между R1 и R3 было предсказано в более ранних работах [120; 157]. Таким образом, мы приходим к выводу, что

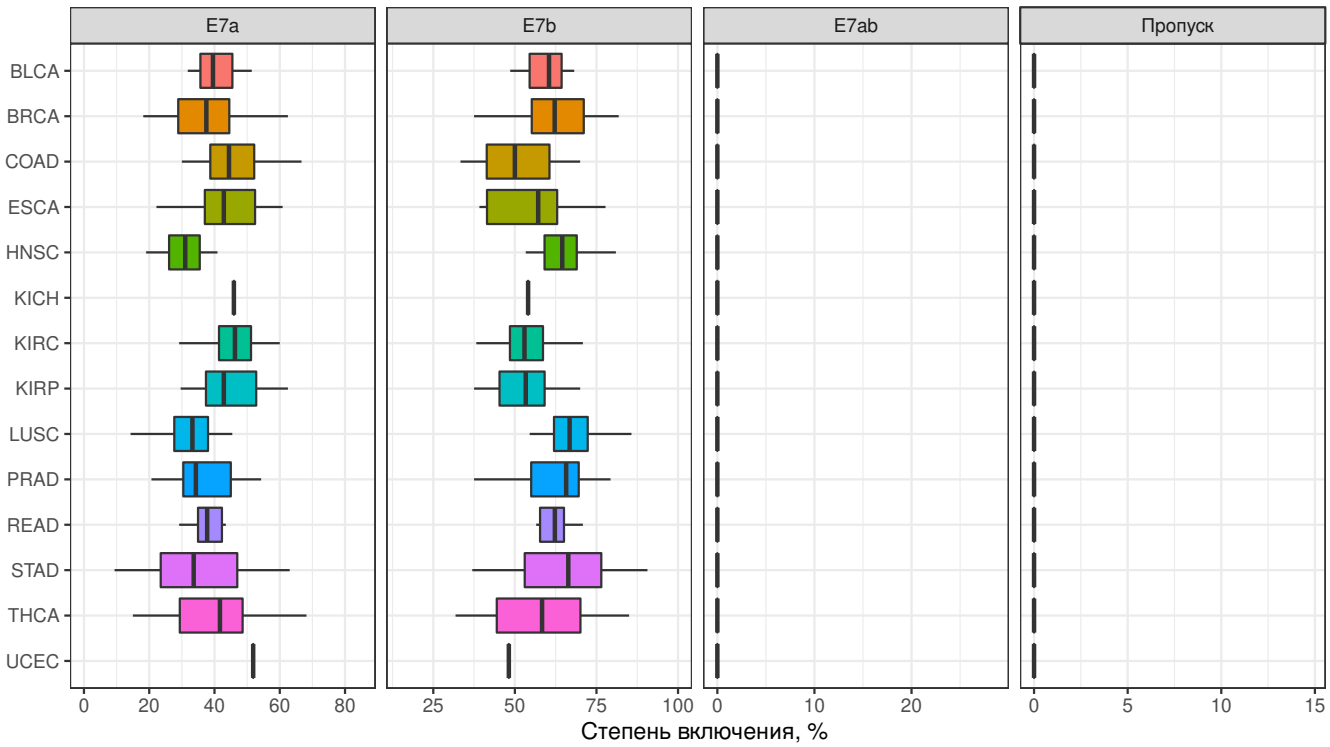

Рисунок 4.2 — Частота включения экзонов 7a, 7b, а также двойных экзонов (7a7b) и одновременного пропуска обоих экзонов (Skip) в опухолях из проекта TCGA. Обозначения опухолей: BLCA — уротелиальная карцинома мочевого пузыря; COAD — рак толстой кишки; BRCA — рак молочной железы; ESCA — рак пищевода; HNSC — плоскоклеточный рак головы и шеи; KICH — хромофобный почечно-клеточный рак; KIRC — карцинома почки; KIRP — почечная папиллярно-клеточная карцинома; LUSC — плоскоклеточный рак легкого; PRAD — аденокарцинома предстательной железы; READ — аденокарцинома прямой кишки; STAD — аденокарцинома желудка; THCA — рак щитовидной железы; UCEC — рак матки.

R1 и R4 могут конкурировать друг с другом за гибридизацию с R3, а вместе со спариванием R2 с R5 они образуют псевдоузел (рис. 4.5).

В соответствии с высказанным нами предположением, конкурирующие структуры РНК в гене *Ate1*, отвечающие за взаимоисключающий сплайсинг экзонов 7a и 7b, могли образоваться в результате тандемной дупликации, охватывающей часть последовательности интрона. Для того, чтобы выяснить когда такая дупликация могла произойти, необходимо найти ген-предшественник, в котором между экзонами 6 и 8 располагался бы только один вариант экзона 7. Подобный гомолог в настоящее время сохранился только у *C. intestinalis*, однако степень идентичности последовательностей в интронах этого гена крайне

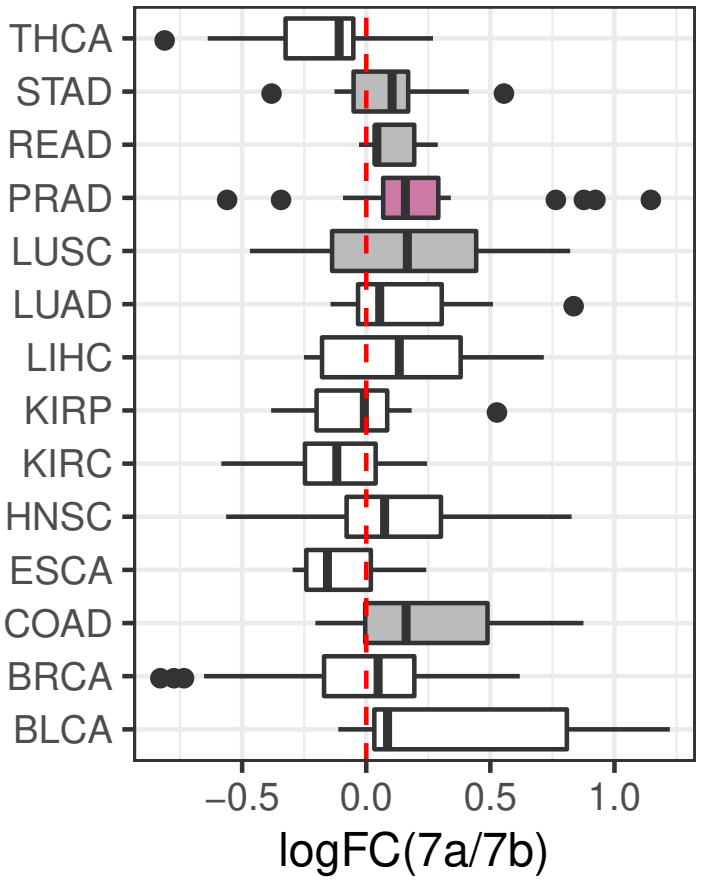

Рисунок 4.3 — Относительная экспрессия изоформ экзона 7a/7b (разница $\log_{10} \frac{7a}{7b}$ между опухолью и нормальной тканью из одного и того же донора) по данным консорциума TCGA. Относительная экспрессия значительно повышена в раке желудка (STAD), раке прямой кишки (READ), колоректальном раке (COAD), аденокарциноме предстательной железы (PRAD) и в плоскоклеточном раке легкого (LUSC). Критерий Уилкоксона FWER < 0.05 (фиолетовый), $P < 0.05$ (серый).

мала, и комплементарных участков, гомологичных R1, R3 и R4, в них обнаружить не удается.

Тем не менее, представляется интересным выяснить как изменялись длины интронов в кластере ВИЭ гена *Ate1* в ходе эволюции. Для ответа на этот вопрос были проанализированы длины интронов в гомологах *Ate1*, определяемые как расстояние от ВИЭ до ближайших конститутивных экзонов по геномной последовательности. Оказалось, что длина фланкирующего интрона в сторону 3' конца гена от экзона 7b уменьшается по мере увеличения эволюционного расстояния от человека (рис. 4.6).

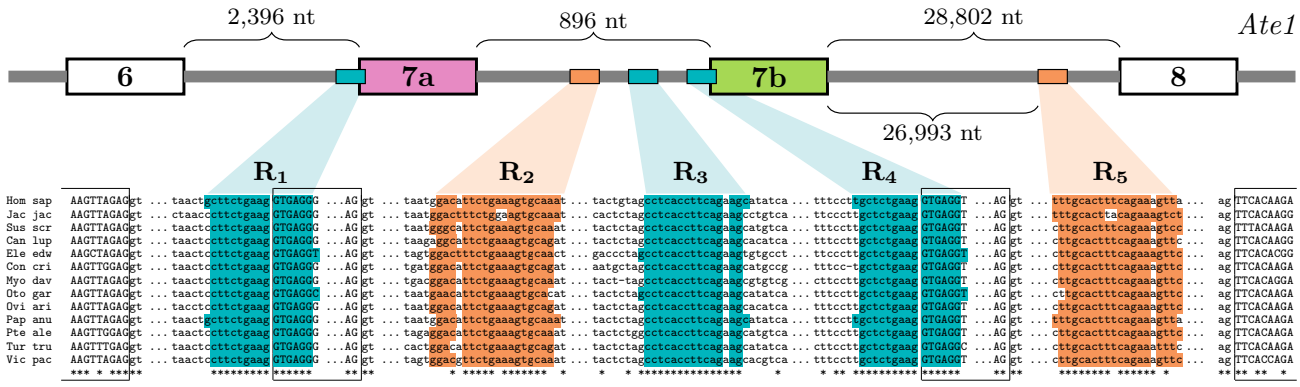

Рисунок 4.4 — Выравнивание кластера ВИЭ в гене *Ate1*. Консервативные интронные элементы обозначены R1–R5.

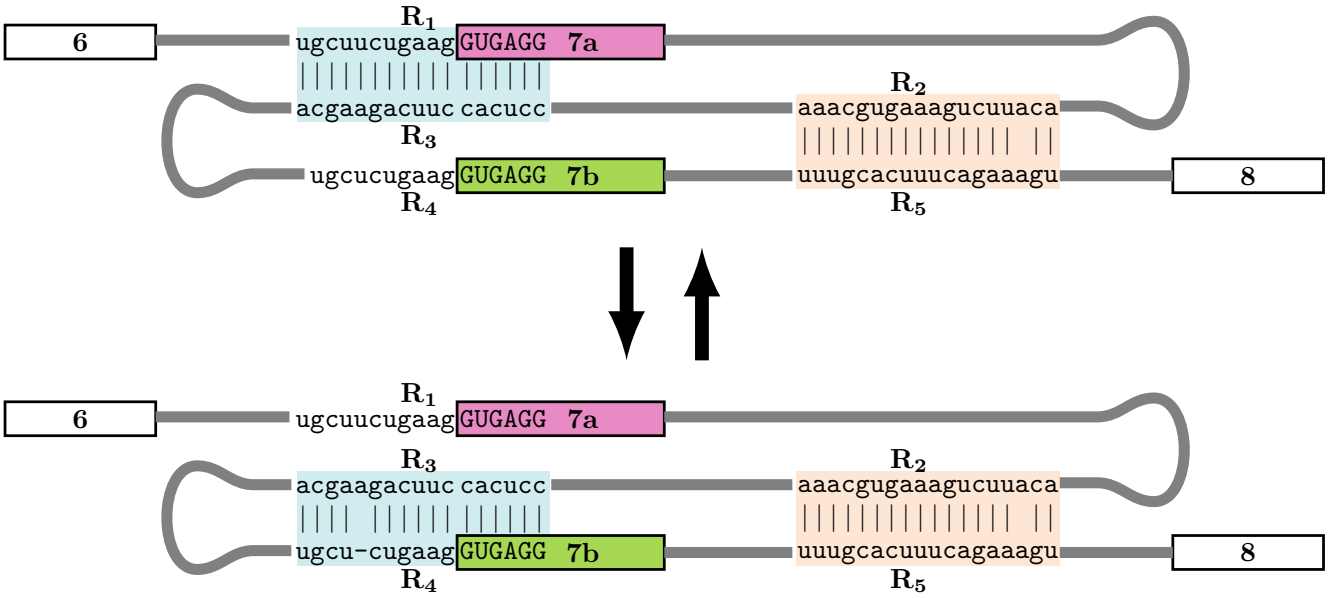

Рисунок 4.5 — Конкурирующие структуры РНК в гене *Ate1* состоят из последовательностей R1 и R4, которые конкурируют друг с другом за гибридизацию с R3, а вместе со спариванием R2 с R5 образуют псевдоузел.

4.3 Обсуждение и выводы

Взаимоисключающий сплайсинг экзонов 7a и 7b в гене *Ate1* человека приводит к образованию функционально различных изоформ аргинилтрансферазы, причем изоформа с преобладанием экзона 7a экспрессируется в тестикулах, а также суперэкспрессирована при аденокарциноме предстательной железы. Механизм, объясняющий взаимоисключающий сплайсинг этих экзонов, до настоящего времени не был известен.

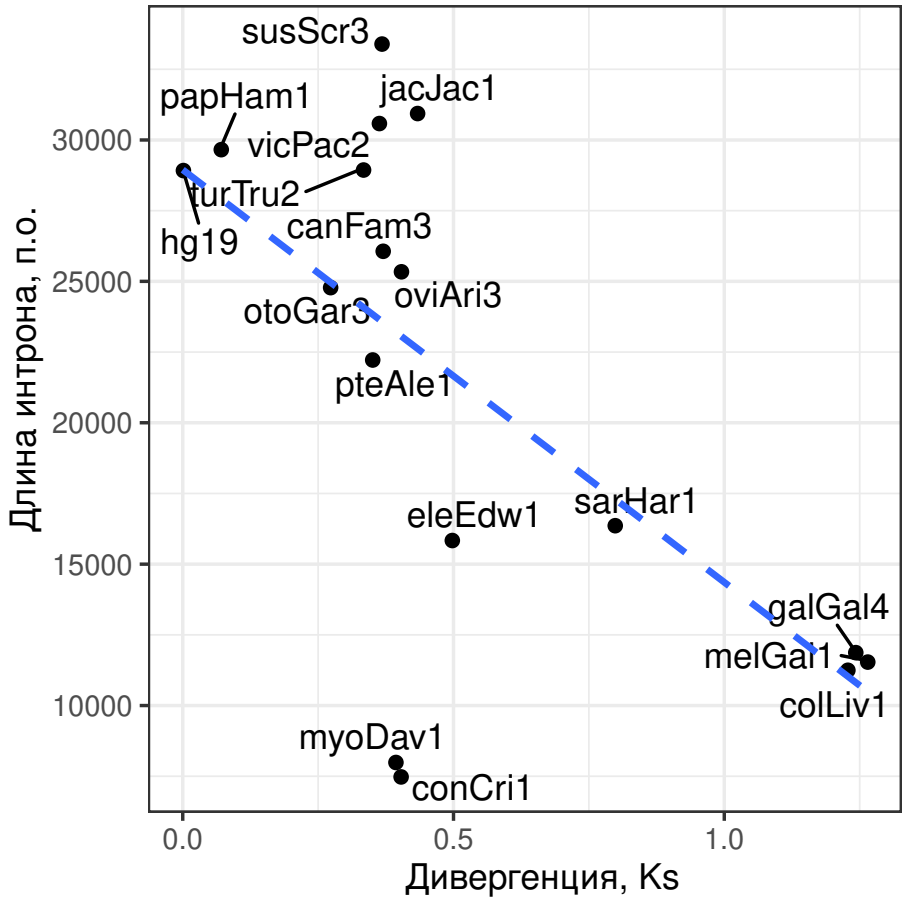

Рисунок 4.6 — Длина интрона между экзоном 7b и экзоном 8 уменьшается с увеличением эволюционного расстояния, измеряемого как среднее количество замен на синонимичный сайт (Ks). Последний был получен из филогенетического дерева в ходе множественного выравнивания multiz[14].

Каталог консервативных комплементарных областей, с помощью которого были обнаружены комплементарные области R1–R5, на сегодняшний день дает наиболее полную характеризацию консервативных вторичных структур РНК в белоккодирующих генах человека [157]. Консервативные комплементарные участки часто возникают внутри интронов, препятствуют включению выпетливаемых экзонов и подавляют криптические и неактивные сайты сплайсинга. Их двухцепочечная структура подтверждается сниженной доступностью нуклеотидов при модификации, большим количеством сайтов редактирования РНК и частым появлением раздвоенных пиков eCLIP. В случае гена *Ate1* наблюдается подавление акцепторных сайтов экзонов 7a и 7b, которые перекрываются с учатсками R1 и R4, в результате образования конкурирующих спариваний с R3. Справедливость этого механизма была показана нами экспериментально, однако этот результат выходит за рамки данной диссертационной работы [15].

В соответствии с высказанной в разд. 3.3 гипотезой, можно предположить, что предшественник гена *Ate1*, обладающий только одним вариантом экзона 7, также мог содержать шпилечную структуру, которая перекрывалась с акцепторным сайтом экзона. После дупликации, которая затронула экзон 7 и часть последовательности интрона, содержащую только одну из комплементарных частей шпильки, образовались конкурирующие структуры R1–R3 и R3–R4. Однако гомологи *Ate1*, обладающие только одним вариантом экзона 7, удается обнаружить только у *C. intestinalis*, в то время гомологи у позвоночных содержат как экзон 7a, так и 7b. Последовательнось интрона, фланкирующего экзон 7 в направлении 5' конца гена не содержит последовательности, гомологичной R3. Это, однако, не противоречит тому, что шпилечная структура в интроне существовала у общего предка *C. intestinalis* и позвоночных и была утеряна *C. intestinalis* в ходе эволюции.

Таким образом, в случае экзонов 7a и 7b у *Ate1* дупликация, вероятно, произошла после радиации хордовых, поскольку гомолог экзона 7b отсутствует у беспозвоночных, но отследить происхождение R1, R3 и R4 не представляется возможным, поскольку конкурирующие вторичные структуры РНК обычно не сохраняются на больших эволюционных расстояниях [11]. Интересно, что длина интрона в сторону 3' конца гена от экзона 7b уменьшается по мере увеличения эволюционного расстояния от человека, что позволяет предположить, что дополнительное комплементарное спаривание между элементами R2 и R5 могло образоваться после дупликации, чтобы противодействовать экспансии интрона в сторону 3' конца гена от экзона 7b.

Заключение

Основные результаты работы заключаются в следующем.

1. Показано, что тандемные дупликации экзонов широко распространены не только в кодирующих, но и в нетранслируемых областях генов животных, а также получено описание неизвестных ранее тандемных дупликаций экзонов в геноме человека с указанием степени их консервативности и уровня экспрессии в тканях.
2. Показано, что интроны в кластерах взаимоисключающих экзонов склонны образовывать конкурирующие структуры РНК, состоящие из докерного и множества селекторных сайтов.
3. Высказано предположение о том, что тандемные дупликации, затрагивающие экзоны и части фланкирующих интронов, неизбежно приводят к образованию конкурирующих структур РНК и, как следствие, ко взаимоисключающему типу сплайсинга. А именно, если в интроне, прилегающем к предковому экзону содержалась шпилечная структура, а дупликация затронула только одну из двух ее комплементарных частей, то после дупликации образуется пара селекторных сайтов, конкурирующих за один и тот же докерный сайт. Данное предположение согласуются со всеми известными из литературы примерами регуляции взаимоисключающего сплайсинга конкурирующими структурами РНК, а также дополняет и обобщает их.
4. Предсказаны конкурирующие структуры РНК, отвечающие за взаимоисключающий сплайсинг экзонов 7a и 7b гена *Ate1* человека.

В заключение автор выражает благодарность и большую признательность научному руководителю Д. Д. Первушину за поддержку, помощь, и бесконечное терпение. Также автор благодарит Сколковский Институт Науки и Технологии и научные коллективы лабораторий Д. Д. Первушина и О. А. Донцовой, в особенности Светлану Калмыкову, Марину Калинину и Дмитрия Скворцова за помощь в работе над совместным проектом по конкурирующим структурам РНК, а также авторов шаблона Russian-Phd-LaTeX-Dissertation-Template, без которого оформление данной диссертационной работы было бы намного более трудоемким.

Список сокращений

АС : Альтернативный сплайсинг
ВИЭ : Взаимоисключающие экзоны
КД : Коэффициент дупликации
кДНК : Кодирующая ДНК
КУ : Кодирующий участок, coding sequence, CDS
мяРНК : малые ядерные РНК
мяРНП : малые ядерные рибонуклеопротеины
нт : нуклеотид
НТО : Нетранслируемая область, untranslated region, UTR
п.о. : пара оснований
BP : точка ветвления, branch point
ChIP : Иммунопрецепитация хроматина, chromatin immunoprecepitation
EST : Экспрессируемые последовательности, expressed sequence tags
FoSTeS : Модель Fork Stalling and Template Switching
FWER : Групповая вероятность ошибки I рода, familywise error rate
GTEx : Консорциум Genotype Tissue Expression project
MFE : Минимальная свободная энергия, minimum free energy
NMD : Нонсенс-опосредованный распад, nonsense mediated decay
PTC : Преждевременный стоп-кодон, premature termination codon
RNA-seq : Секвенирование РНК, RNA sequencing
RPRM : экспрессия, reads per kilobase per million of mapped reads
SD : сегментная дупликация, segment duplication
TCGA : Консорциум Атлас Ракового Генома, The Cancer Genome Atlas
WGD : Полногеномная дупликация, whole genome duplication

Список литературы

1. The clinical importance of tandem exon duplication-derived substitutions [Текст] / L. Martinez Gomez [и др.] // Nucleic Acids Res. — 2021. — Авг. — Т. 49, № 14. — С. 8232—8246.
2. *Kondrashov, F. A.* Origin of alternative splicing by tandem exon duplication [Текст] / F. A. Kondrashov, E. V. Koonin // Hum Mol Genet. — 2001. — Нояб. — Т. 10, № 23. — С. 2661—2669.
3. Role and convergent evolution of competing RNA secondary structures in mutually exclusive splicing [Текст] / Y. Yue [и др.] // RNA Biol. — 2017. — Окт. — Т. 14, № 10. — С. 1399—1410.
4. Noncardiac chest pain: is the esophagus really a frequent source? [Текст] / A. J. Limburg [и др.] // Scand J Gastroenterol. — 1990. — Авг. — Т. 25, № 8. — С. 793—798.
5. RNA secondary structure in mutually exclusive splicing [Текст] / Y. Yang [и др.] // Nat Struct Mol Biol. — 2011. — Февр. — Т. 18, № 2. — С. 159—168.
6. Role of RNA secondary structures in regulating Dscam alternative splicing [Текст] / B. Xu [и др.] // Biochim Biophys Acta Gene Regul Mech. — 2019. — Т. 1862, № 11/12. — С. 194381.
7. *Smith, C. W.* Alternative splicing–when two's a crowd [Текст] / C. W. Smith // Cell. — 2005. — Окт. — Т. 123, № 1. — С. 1—3.
8. Mutually exclusive alternative splicing of pre-mRNAs [Текст] / Y. Jin [и др.] // Wiley Interdiscip Rev RNA. — 2018. — Май. — Т. 9, № 3. — e1468.
9. *Hatje, K.* Expansion of the mutually exclusive spliced exome in Drosophila [Текст] / K. Hatje, M. Kollmar // Nat Commun. — 2013. — Т. 4. — С. 2460.
10. *Hatje, K.* Kassiopeia: a database and web application for the analysis of mutually exclusive exomes of eukaryotes [Текст] / K. Hatje, M. Kollmar // BMC Genomics. — 2014. — Февр. — Т. 15. — С. 115.
11. The landscape of human mutually exclusive splicing [Текст] / K. Hatje [и др.] // Mol Syst Biol. — 2017. — Дек. — Т. 13, № 12. — С. 959.

12. Human genomics. The human transcriptome across tissues and individuals [Текст] / M. Melé [и др.] // Science. — 2015. — Май. — Т. 348, № 6235. — С. 660—665.

13. The Cancer Genome Atlas Pan-Cancer analysis project [Текст] / J. N. Weinstein [и др.] // Nat Genet. — 2013. — Окт. — Т. 45, № 10. — С. 1113—1120.

14. The human genome browser at UCSC [Текст] / W. J. Kent [и др.] // Genome Res. — 2002. — Июнь. — Т. 12, № 6. — С. 996—1006.

15. Multiple competing RNA structures dynamically control alternative splicing in the human ATE1 gene [Текст] / M. Kalinina [и др.] // Nucleic Acids Res. — 2021. — Янв. — Т. 49, № 1. — С. 479—490.

16. *Rastogi, S.* Subfunctionalization of duplicated genes as a transition state to neofunctionalization [Текст] / S. Rastogi, D. A. Liberles // BMC Evol Biol. — 2005. — Апр. — Т. 5. — С. 28.

17. *Sandve, S. R.* Subfunctionalization versus neofunctionalization after whole-genome duplication [Текст] / S. R. Sandve, R. V. Rohlfs, T. R. Hvidsten // Nat Genet. — 2018. — Июль. — Т. 50, № 7. — С. 908—909.

18. Mutations in the paralogous human alpha-globin genes yielding identical hemoglobin variants [Текст] / K. Moradkhani [и др.] // Ann Hematol. — 2009. — Июнь. — Т. 88, № 6. — С. 535—543.

19. *Michelson, A. M.* Boundaries of gene conversion within the duplicated human alpha-globin genes. Concerted evolution by segmental recombination [Текст] / A. M. Michelson, S. H. Orkin // J Biol Chem. — 1983. — Дек. — Т. 258, № 24. — С. 15245—15254.

20. *Keshet, Y.* The MAP kinase signaling cascades: a system of hundreds of components regulates a diverse array of physiological functions [Текст] / Y. Keshet, R. Seger // Methods Mol Biol. — 2010. — Т. 661. — С. 3—38.

21. *Calvo-Martín, J. M.* Evidence of neofunctionalization after the duplication of the highly conserved Polycomb group gene Caf1-55 in the obscura group of Drosophila [Текст] / J. M. Calvo-Martín, M. Papaceit, C. Segarra // Sci Rep. — 2017. — Янв. — Т. 7. — С. 40536.

22. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants [Текст] / X. Qiao [и др.] // Genome Biol. — 2019. — Февр. — Т. 20, № 1. — С. 38.

23. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia [Текст] / J. M. Aury [и др.] // Nature. — 2006. — Нояб. — Т. 444, № 7116. — С. 171—178.

24. *Dehal, P.* Two rounds of whole genome duplication in the ancestral vertebrate [Текст] / P. Dehal, J. L. Boore // PLoS Biol. — 2005. — Окт. — Т. 3, № 10. — e314.

25. *Singh, P. P.* OHNOLOGS v2: a comprehensive resource for the genes retained from whole genome duplication in vertebrates [Текст] / P. P. Singh, H. Isambert // Nucleic Acids Res. — 2020. — Янв. — Т. 48, № D1. — С. D724—D730.

26. *Bergthorsson, U.* Ohno's dilemma: evolution of new genes under continuous selection [Текст] / U. Bergthorsson, D. I. Andersson, J. R. Roth // Proc Natl Acad Sci U S A. — 2007. — Окт. — Т. 104, № 43. — С. 17004—17009.

27. *Copley, S. D.* Evolution of new enzymes by gene duplication and divergence [Текст] / S. D. Copley // FEBS J. — 2020. — Апр. — Т. 287, № 7. — С. 1262—1283.

28. *Letunic, I.* Common exon duplication in animals and its role in alternative splicing [Текст] / I. Letunic, R. R. Copley, P. Bork // Hum Mol Genet. — 2002. — Июнь. — Т. 11, № 13. — С. 1561—1567.

29. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation [Текст] / N. A. O'Leary [и др.] // Nucleic Acids Res. — 2016. — Янв. — Т. 44, № D1. — С. D733—745.

30. *Yang, Z.* Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models [Текст] / Z. Yang, R. Nielsen // Mol Biol Evol. — 2000. — Янв. — Т. 17, № 1. — С. 32—43.

31. Copy number variation in human health, disease, and evolution [Текст] / F. Zhang [и др.] // Annu Rev Genomics Hum Genet. — 2009. — Т. 10. — С. 451—481.

32. Large duplications at reciprocal translocation breakpoints that might be the counterpart of large deletions and could arise from stalled replication bubbles [Текст] / K. D. Howarth [и др.] // Genome Res. — 2011. — Апр. — Т. 21, № 4. — С. 525—534.

33. *Kaessmann, H.* RNA-based gene duplication: mechanistic and evolutionary insights [Текст] / H. Kaessmann, N. Vinckenbosch, M. Long // Nat Rev Genet. — 2009. — Янв. — Т. 10, № 1. — С. 19—31.

34. *Gotea, V.* Do transposable elements really contribute to proteomes? [Текст] / V. Gotea, W. Makałowski // Trends Genet. — 2006. — Май. — Т. 22, № 5. — С. 260—267.

35. *Achaz, G.* Study of intrachromosomal duplications among the eukaryote genomes [Текст] / G. Achaz, P. Netter, E. Coissac // Mol Biol Evol. — 2001. — Дек. — Т. 18, № 12. — С. 2280—2288.

36. Identification of the first duplication in MYH9-related disease: a hot spot for unequal crossing-over within exon 24 of the MYH9 gene [Текст] / D. De Rocco [и др.] // Eur J Med Genet. — 2009. — Т. 52, № 4. — С. 191—194.

37. *Matera, A. G.* A day in the life of the spliceosome [Текст] / A. G. Matera, Z. Wang // Nat Rev Mol Cell Biol. — 2014. — Февр. — Т. 15, № 2. — С. 108—121.

38. *Will, C. L.* Spliceosome structure and function [Текст] / C. L. Will, R. Lührmann // Cold Spring Harb Perspect Biol. — 2011. — Июль. — Т. 3, № 7.

39. Splicing and transcription touch base: co-transcriptional spliceosome assembly and function [Текст] / L. Herzel [и др.] // Nat Rev Mol Cell Biol. — 2017. — Окт. — Т. 18, № 10. — С. 637—650.

40. Regulatory roles and mechanisms of alternative RNA splicing in adipogenesis and human metabolic health [Текст] / Y. Chao [и др.] // Cell Biosci. — 2021. — Апр. — Т. 11, № 1. — С. 66.

41. Minor spliceosome and disease [Текст] / B. Verma [и др.] // Semin Cell Dev Biol. — 2018. — Июль. — Т. 79. — С. 103—112.

42. *Wilkinson, M. E.* RNA Splicing by the Spliceosome [Текст] / M. E. Wilkinson, C. Charenton, K. Nagai // Annu Rev Biochem. — 2020. — Июнь. — Т. 89. — С. 359—388.

43. Apoptosis and abundance of Bcl-2 family and transforming growth factor β1 signaling proteins in canine myxomatous mitral valves [Текст] / S. Surachetpong [и др.] // J Vet Cardiol. — 2013. — Сент. — Т. 15, № 3. — С. 171—180.

44. *Blencowe, B. J.* Alternative splicing: new insights from global analyses [Текст] / B. J. Blencowe // Cell. — 2006. — Июль. — Т. 126, № 1. — С. 37—47.

45. Identification of novel alternative splicing isoform biomarkers and their association with overall survival in colorectal cancer [Текст] / H. Lian [и др.] // BMC Gastroenterol. — 2020. — Июнь. — Т. 20, № 1. — С. 171.

46. *Goldstrohm, A. C.* Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing [Текст] / A. C. Goldstrohm, A. L. Greenleaf, M. A. Garcia-Blanco // Gene. — 2001. — Окт. — Т. 277, № 1/2. — С. 31—47.

47. Frasier syndrome is caused by defective alternative splicing of WT1 leading to an altered ratio of WT1 +/-KTS splice isoforms [Текст] / B. Klamt [и др.] // Hum Mol Genet. — 1998. — Апр. — Т. 7, № 4. — С. 709—714.

48. GENCODE reference annotation for the human and mouse genomes [Текст] / A. Frankish [и др.] // Nucleic Acids Res. — 2019. — Янв. — Т. 47, № D1. — С. D766—D773.

49. Comparison of GENCODE and RefSeq gene annotation and the impact of reference geneset on variant effect prediction [Текст] / A. Frankish [и др.] // BMC Genomics. — 2015. — Т. 16 Suppl 8. — S2.

50. *Ma, C.* Exact transcript quantification over splice graphs [Текст] / C. Ma, H. Zheng, C. Kingsford // Algorithms Mol Biol. — 2021. — Май. — Т. 16, № 1. — С. 5.

51. *Chacko, E.* Comprehensive splicing graph analysis of alternative splicing patterns in chicken, compared to human and mouse [Текст] / E. Chacko, S. Ranganathan // BMC Genomics. — 2009. — Июль. — Т. 10 Suppl 1. — S5.

52. Mechanism of alternative splicing and its regulation [Текст] / Y. Wang [и др.] // Biomed Rep. — 2015. — Март. — Т. 3, № 2. — С. 152—158.

53. *Dvinge, H.* Widespread intron retention diversifies most cancer transcriptomes [Текст] / H. Dvinge, R. K. Bradley // Genome Med. — 2015. — Т. 7, № 1. — С. 45.

54. *Kofuji, P.* Mutually exclusive and cassette exons underlie alternatively spliced isoforms of the Na/Ca exchanger [Текст] / P. Kofuji, W. J. Lederer, D. H. Schulze // J Biol Chem. — 1994. — Февр. — Т. 269, № 7. — С. 5145—5149.

55. Competing RNA secondary structures are required for mutually exclusive splicing of the Dscam exon 6 cluster [Текст] / G. E. May [и др.] // RNA. — 2011. — Февр. — Т. 17, № 2. — С. 222—229.

56. *Kuroyanagi, H.* Comprehensive analysis of mutually exclusive alternative splicing in C. elegans [Текст] / H. Kuroyanagi, S. Takei, Y. Suzuki // Worm. — 2014. — Т. 3. — e28459.

57. *Benson, D. L.* N-cadherin redistribution during synaptogenesis in hippocampal neurons [Текст] / D. L. Benson, H. Tanaka // J Neurosci. — 1998. — Сент. — Т. 18, № 17. — С. 6892—6904.

58. The catenin/cadherin adhesion system is localized in synaptic junctions bordering transmitter release zones [Текст] / N. Uchida [и др.] // J Cell Biol. — 1996. — Нояб. — Т. 135, № 3. — С. 767—779.

59. *Shapiro, L.* The diversity of cadherins and implications for a synaptic adhesive code in the CNS [Текст] / L. Shapiro, D. R. Colman // Neuron. — 1999. — Июль. — Т. 23, № 3. — С. 427—430.

60. An isoform-specific allele of Drosophila N-cadherin disrupts a late step of R7 targeting [Текст] / A. Nern [и др.] // Proc Natl Acad Sci U S A. — 2005. — Сент. — Т. 102, № 36. — С. 12944—12949.

61. Drosophila N-cadherin functions in the first stage of the two-stage layer-selection process of R7 photoreceptor afferents [Текст] / C. Y. Ting [и др.] // Development. — 2005. — Март. — Т. 132, № 5. — С. 953—963.

62. *George, E. L.* Functional domains of the Drosophila melanogaster muscle myosin heavy-chain gene are encoded by alternatively spliced exons [Текст] / E. L. George, M. B. Ober, C. P. Emerson // Mol Cell Biol. — 1989. — Июль. — Т. 9, № 7. — С. 2957—2974.

63. Alternative myosin hinge regions are utilized in a tissue-specific fashion that correlates with muscle contraction speed [Текст] / V. L. Collier [и др.] // Genes Dev. — 1990. — Июнь. — Т. 4, № 6. — С. 885—895.

64. A third functional isoform enriched in mushroom body neurons is encoded by the Drosophila 14-3-3zeta gene [Текст] / G. Messaritou [и др.] // FEBS Lett. — 2009. — Сент. — Т. 583, № 17. — С. 2934—2938.

65. Two isoforms of Serpent containing either one or two GATA zinc fingers have different roles in Drosophila haematopoiesis [Текст] / L. Waltzer [и др.] // EMBO J. — 2002. — Окт. — Т. 21, № 20. — С. 5477—5486.

66. *Grailles, M.* The Drosophila melanogaster multidrug-resistance protein 1 (MRP1) homolog has a novel gene structure containing two variable internal exons [Текст] / M. Grailles, P. T. Brey, C. W. Roth // Gene. — 2003. — Март. — Т. 307. — С. 41—50.

67. Alternative splicing of the multidrug resistance protein 1/ATP binding cassette transporter subfamily gene in ovarian cancer creates functional splice variants and is associated with increased expression of the splicing factors PTB and SRp20 [Текст] / X. He [и др.] // Clin Cancer Res. — 2004. — Июль. — Т. 10, № 14. — С. 4652—4660.

68. An alternative splicing switch regulates embryonic stem cell pluripotency and reprogramming [Текст] / M. Gabut [и др.] // Cell. — 2011. — Сент. — Т. 147, № 1. — С. 132—146.

69. Tissue-specific splicing of a ubiquitously expressed transcription factor is essential for muscle differentiation [Текст] / S. Sebastian [и др.] // Genes Dev. — 2013. — Июнь. — Т. 27, № 11. — С. 1247—1259.

70. *Gooding, C.* Tropomyosin exons as models for alternative splicing [Текст] / C. Gooding, C. W. Smith // Adv Exp Med Biol. — 2008. — Т. 644. — С. 27—42.

71. Molecular and functional characterization of a novel cardiac-specific human tropomyosin isoform [Текст] / S. Rajan [и др.] // Circulation. — 2010. — Янв. — Т. 121, № 3. — С. 410—418.

72. Predicting mutually exclusive spliced exons based on exon length, splice site and reading frame conservation, and exon sequence homology [Текст] / H. Pillmann [и др.] // BMC Bioinformatics. — 2011. — Июнь. — Т. 12. — С. 270.

73. The developmental transcriptome of Drosophila melanogaster [Текст] / B. R. Graveley [и др.] // Nature. — 2011. — Март. — Т. 471, № 7339. — С. 473—479.

74. Genome-wide analysis of alternative splicing in Caenorhabditis elegans [Текст] / A. K. Ramani [и др.] // Genome Res. — 2011. — Февр. — Т. 21, № 2. — С. 342—348.

75. *Suyama, M.* Mechanistic insights into mutually exclusive splicing in dynamin 1 [Текст] / M. Suyama // Bioinformatics. — 2013. — Сент. — Т. 29, № 17. — С. 2084—2087.

76. Transcriptome survey reveals increased complexity of the alternative splicing landscape in Arabidopsis [Текст] / Y. Marquez [и др.] // Genome Res. — 2012. — Июнь. — Т. 22, № 6. — С. 1184—1195.

77. *Kennedy, C. F.* Pyrimidine tracts between the 5' splice site and branch point facilitate splicing and recognition of a small Drosophila intron [Текст] / C. F. Kennedy, S. M. Berget // Mol Cell Biol. — 1997. — Май. — Т. 17, № 5. — С. 2774—2780.

78. *Ruskin, B.* Cryptic branch point activation allows accurate in vitro splicing of human beta-globin intron mutants [Текст] / B. Ruskin, J. M. Greene, M. R. Green // Cell. — 1985. — Июль. — Т. 41, № 3. — С. 833—844.

79. *Smith, C. W.* Mutually exclusive splicing of alpha-tropomyosin exons enforced by an unusual lariat branch point location: implications for constitutive splicing [Текст] / C. W. Smith, B. Nadal-Ginard // Cell. — 1989. — Март. — Т. 56, № 5. — С. 749—758.

80. Regulation of Dscam exon 17 alternative splicing by steric hindrance in combination with RNA secondary structures [Текст] / Y. Yue [и др.] // RNA Biol. — 2013. — Дек. — Т. 10, № 12. — С. 1822—1833.

81. The nonsense-mediated decay pathway and mutually exclusive expression of alternatively spliced FGFR2IIIb and -IIIc mRNAs [Текст] / R. B. Jones [и др.] // J Biol Chem. — 2001. — Февр. — Т. 276, № 6. — С. 4158—4167.

82. *Oxender, D. L.* Attenuation in the Escherichia coli tryptophan operon: role of RNA secondary structure involving the tryptophan codon region [Текст] / D. L. Oxender, G. Zurawski, C. Yanofsky // Proc Natl Acad Sci U S A. — 1979. — Нояб. — Т. 76, № 11. — С. 5524—5528.

83. *Graveley, B. R.* Mutually exclusive splicing of the insect Dscam pre-mRNA directed by competing intronic RNA secondary structures [Текст] / B. R. Graveley // Cell. — 2005. — Окт. — Т. 123, № 1. — С. 65—73.

84. An RNA architectural locus control region involved in Dscam mutually exclusive splicing [Текст] / X. Wang [и др.] // Nat Commun. — 2012. — Т. 3. — С. 1255.

85. The molecular diversity of Dscam is functionally required for neuronal wiring specificity in Drosophila [Текст] / B. E. Chen [и др.] // Cell. — 2006. — Май. — Т. 125, № 3. — С. 607—620.

86. Cell-intrinsic requirement of Dscam1 isoform diversity for axon collateral formation [Текст] / H. He [и др.] // Science. — 2014. — Июнь. — Т. 344, № 6188. — С. 1182—1186.

87. Homophilic Dscam interactions control complex dendrite morphogenesis [Текст] / M. E. Hughes [и др.] // Neuron. — 2007. — Май. — Т. 54, № 3. — С. 417—427.

88. Axonal targeting of olfactory receptor neurons in Drosophila is controlled by Dscam [Текст] / T. Hummel [и др.] // Neuron. — 2003. — Янв. — Т. 37, № 2. — С. 221—231.

89. Dendrite self-avoidance is controlled by Dscam [Текст] / B. J. Matthews [и др.] // Cell. — 2007. — Май. — Т. 129, № 3. — С. 593—604.

90. Drosophila sensory neurons require Dscam for dendritic self-avoidance and proper dendritic field organization [Текст] / P. Soba [и др.] // Neuron. — 2007. — Май. — Т. 54, № 3. — С. 403—416.

91. Drosophila Dscam is required for divergent segregation of sister branches and suppresses ectopic bifurcation of axons [Текст] / J. Wang [и др.] // Neuron. — 2002. — Февр. — Т. 33, № 4. — С. 559—571.

92. Analysis of Dscam diversity in regulating axon guidance in Drosophila mushroom bodies [Текст] / X. L. Zhan [и др.] // Neuron. — 2004. — Сент. — Т. 43, № 5. — С. 673—686.

93. Dendritic patterning by Dscam and synaptic partner matching in the Drosophila antennal lobe [Текст] / H. Zhu [и др.] // Nat Neurosci. — 2006. — Март. — Т. 9, № 3. — С. 349—355.

94. Intron-targeted mutagenesis reveals roles for Dscam1 RNA pairing architecture-driven splicing bias in neuronal wiring [Текст] / W. Hong [и др.] // Cell Rep. — 2021. — Июль. — Т. 36, № 2. — С. 109373.

95. Drosophila muscle myosin heavy chain encoded by a single gene in a cluster of muscle mutations [Текст] / S. I. Bernstein [и др.] // Nature. — 1983. — Т. 302, № 5907. — С. 393—397.

96. Organization of serpin gene-1 from Manduca sexta. Evolution of a family of alternate exons encoding the reactive site loop [Текст] / H. Jiang [и др.] // J Biol Chem. — 1996. — Нояб. — Т. 271, № 45. — С. 28017—28023.

97. *Nusbaum, M. J.* The content of calcium, magnesium, copper, iron, sodium, and potassium in amniotic fluid from eleven to nineteen weeks' gestation [Текст] / M. J. Nusbaum, A. Zettner // Am J Obstet Gynecol. — 1973. — Янв. — Т. 115, № 2. — С. 219—226.

98. Identification of alternative splicing regulators by RNA interference in Drosophila [Текст] / J. W. Park [и др.] // Proc Natl Acad Sci U S A. — 2004. — Нояб. — Т. 101, № 45. — С. 15974—15979.

99. *Ivanov, T. M.* Tandem Exon Duplications Expanding the Alternative Splicing Repertoire [Текст] / T. M. Ivanov, D. D. Pervouchine // Acta Naturae. — 2022. — Т. 14, № 1. — С. 73—81.

100. Modernizing reference genome assemblies [Текст] / D. M. Church [и др.] // PLoS Biol. — 2011. — Июль. — Т. 9, № 7. — e1001091.

101. GENCODE: the reference human genome annotation for The ENCODE Project [Текст] / J. Harrow [и др.] // Genome Res. — 2012. — Сент. — Т. 22, № 9. — С. 1760—1774.

102. *Marygold, S. J.* Using FlyBase, a Database of Drosophila Genes and Genomes [Текст] / S. J. Marygold, M. A. Crosby, J. L. Goodman // Methods Mol Biol. — 2016. — Т. 1478. — С. 1—31.

103. WormBase: a modern Model Organism Information Resource [Текст] / T. W. Harris [и др.] // Nucleic Acids Res. — 2020. — Янв. — Т. 48, № D1. — С. D762—D767.

104. *Slater, G. S.* Automated generation of heuristics for biological sequence comparison [Текст] / G. S. Slater, E. Birney // BMC Bioinformatics. — 2005. — Февр. — Т. 6. — С. 31.

105. *Quinlan, A. R.* BEDTools: a flexible suite of utilities for comparing genomic features [Текст] / A. R. Quinlan, I. M. Hall // Bioinformatics. — 2010. — Март. — Т. 26, № 6. — С. 841—842.

106. STAR: ultrafast universal RNA-seq aligner [Текст] / A. Dobin [и др.] // Bioinformatics. — 2013. — Янв. — Т. 29, № 1. — С. 15—21.

107. *Pervouchine, D. D.* Intron-centric estimation of alternative splicing from RNA-seq data [Текст] / D. D. Pervouchine, D. G. Knowles, R. Guigó // Bioinformatics. — 2013. — Янв. — Т. 29, № 2. — С. 273—274.

108. deepTools: a flexible platform for exploring deep-sequencing data [Текст] / F. Ramírez [и др.] // Nucleic Acids Res. — 2014. — Июль. — Т. 42, Web Server issue. — W187—191.

109. *Schmitz, J.* Exonization of transposed elements: A challenge and opportunity for evolution [Текст] / J. Schmitz, J. Brosius // Biochimie. — 2011. — Нояб. — Т. 93, № 11. — С. 1928—1934.

110. *Fukuzawa, A.* Complete human gene structure of obscurin: implications for isoform generation by differential splicing [Текст] / A. Fukuzawa, S. Idowu, M. Gautel // J Muscle Res Cell Motil. — 2005. — Т. 26, № 6—8. — С. 427—434.

111. *Kontrogianni-Konstantopoulos, A.* Obscurin: a multitasking muscle giant [Текст] / A. Kontrogianni-Konstantopoulos, R. J. Bloch // J Muscle Res Cell Motil. — 2005. — Т. 26, № 6—8. — С. 419—426.

112. A Gilbert syndrome-associated haplotype protects against fatty liver disease in humanized transgenic mice [Текст] / S. Landerer [и др.] // Sci Rep. — 2020. — Май. — Т. 10, № 1. — С. 8689.

113. *Strassburg, C. P.* Variability and function of family 1 uridine-5'-diphosphate glucuronosyltransferases (UGT1A) [Текст] / C. P. Strassburg, S. Kalthoff, U. Ehmer // Crit Rev Clin Lab Sci. — 2008. — Т. 45, № 6. — С. 485—530.

114. Evolution of hydra, a recently evolved testis-expressed gene with nine alternative first exons in Drosophila melanogaster [Текст] / S. T. Chen [и др.] // PLoS Genet. — 2007. — Июль. — Т. 3, № 7. — e107.

115. Competing RNA pairings in complex alternative splicing of a 3' variable region [Текст] / H. Pan [и др.] // RNA. — 2018. — Нояб. — Т. 24, № 11. — С. 1466—1480.

116. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes [Текст] / A. Siepel [и др.] // Genome Res. — 2005. — Авг. — Т. 15, № 8. — С. 1034—1050.

117. *Ivanov, T. M.* An Evolutionary Mechanism for the Generation of Competing RNA Structures Associated with Mutually Exclusive Exons [Текст] / T. M. Ivanov, D. D. Pervouchine // Genes (Basel). — 2018. — Июль. — Т. 9, № 7.

118. Biopython: freely available Python tools for computational molecular biology and bioinformatics [Текст] / P. J. Cock [и др.] // Bioinformatics. — 2009. — Июнь. — Т. 25, № 11. — С. 1422—1423.

119. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts [Текст] / M. Jiang [и др.] // BMC Bioinformatics. — 2008. — Апр. — Т. 9. — С. 192.

120. *Pervouchine, D. D.* IRBIS: a systematic search for conserved complementarity [Текст] / D. D. Pervouchine // RNA. — 2014. — Окт. — Т. 20, № 10. — С. 1519—1531.

121. *Pervouchine, D. D.* Towards Long-Range RNA Structure Prediction in Eukaryotic Genes [Текст] / D. D. Pervouchine // Genes (Basel). — 2018. — Июнь. — Т. 9, № 6.

122. ViennaRNA Package 2.0 [Текст] / R. Lorenz [и др.] // Algorithms Mol Biol. — 2011. — Нояб. — Т. 6. — С. 26.

123. Thermodynamics of RNA-RNA binding [Текст] / U. Mückstein [и др.] // Bioinformatics. — 2006. — Май. — Т. 22, № 10. — С. 1177—1182.

124. Fast accessibility-based prediction of RNA-RNA interactions [Текст] / H. Tafer [и др.] // Bioinformatics. — 2011. — Июль. — Т. 27, № 14. — С. 1934—1940.

125. Modulation of alternative splicing by long-range RNA structures in Drosophila [Текст] / V. A. Raker [и др.] // Nucleic Acids Res. — 2009. — Авг. — Т. 37, № 14. — С. 4533—4544.

126. Long-range RNA pairings contribute to mutually exclusive splicing [Текст] / Y. Yue [и др.] // RNA. — 2016. — Янв. — Т. 22, № 1. — С. 96—110.

127. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures [Текст] / D. D. Pervouchine [и др.] // RNA. — 2012. — Янв. — Т. 18, № 1. — С. 1—15.

128. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges [Текст] / M. T. Lovci [и др.] // Nat Struct Mol Biol. — 2013. — Дек. — Т. 20, № 12. — С. 1434—1442.

129. A regulator of Dscam mutually exclusive splicing fidelity [Текст] / S. Olson [и др.] // Nat Struct Mol Biol. — 2007. — Дек. — Т. 14, № 12. — С. 1134—1140.

130. Coupling of RNA Polymerase II Transcription Elongation with Pre-mRNA Splicing [Текст] / T. Saldi [и др.] // J Mol Biol. — 2016. — Июнь. — Т. 428, № 12. — С. 2623—2635.

131. *Herschlag, D.* RNA chaperones and the RNA folding problem [Текст] / D. Herschlag // J Biol Chem. — 1995. — Сент. — Т. 270, № 36. — С. 20871—20874.

132. *Kashina, A. S.* Protein Arginylation: Over 50 Years of Discovery [Текст] / A. S. Kashina // Methods Mol Biol. — 2015. — Т. 1337. — С. 1—11.

133. Cloning and functional analysis of the arginyl-tRNA-protein transferase gene ATE1 of Saccharomyces cerevisiae [Текст] / E. Balzi [и др.] // J Biol Chem. — 1990. — Май. — Т. 265, № 13. — С. 7464—7471.

134. N-terminal arginylation generates a bimodal degron that modulates autophagic proteolysis [Текст] / Y. D. Yoo [и др.] // Proc Natl Acad Sci U S A. — 2018. — Март. — Т. 115, № 12. — E2716—E2724.

135. The N-end rule pathway catalyzes a major fraction of the protein degradation in skeletal muscle [Текст] / V. Solomon [и др.] // J Biol Chem. — 1998. — Сент. — Т. 273, № 39. — С. 25216—25222.

136. Protein arginylation regulates cellular stress response by stabilizing HSP70 and HSP40 transcripts [Текст] / K. Deka [и др.] // Cell Death Discov. — 2016. — Т. 2. — С. 16074.

137. *Lamon, K. D.* Stress-induced increases in rat brain arginyl-tRNA transferase activity [Текст] / K. D. Lamon, W. H. Vogel, H. Kaji // Brain Res. — 1980. — Май. — Т. 190, № 1. — С. 285—287.

138. Posttranslational arginylation of soluble rat brain proteins after whole body hyperthermia [Текст] / G. Bongiovanni [и др.] // J Neurosci Res. — 1999. — Апр. — Т. 56, № 1. — С. 85—92.

139. An essential role of N-terminal arginylation in cardiovascular development [Текст] / Y. T. Kwon [и др.] // Science. — 2002. — Июль. — Т. 297, № 5578. — С. 96—99.

140. Arginylation-dependent neural crest cell migration is essential for mouse development [Текст] / S. Kurosaka [и др.] // PLoS Genet. — 2010. — Март. — Т. 6, № 3. — e1000878.

141. Arginyltransferase regulates alpha cardiac actin function, myofibril formation and contractility during heart development [Текст] / R. Rai [и др.] // Development. — 2008. — Дек. — Т. 135, № 23. — С. 3881—3889.

142. *Tanaka, Y.* Incorporation of arginine by soluble extracts of ascites tumor cells and regenerating rat liver [Текст] / Y. Tanaka, H. Kaji // Cancer Res. — 1974. — Сент. — Т. 34, № 9. — С. 2204—2208.

143. *Chakraborty, G.* N-terminal arginylation and ubiquitin-mediated proteolysis in nerve regeneration [Текст] / G. Chakraborty, N. A. Ingoglia // Brain Res Bull. — 1993. — Т. 30, № 3/4. — С. 439—445.

144. *Wang, Y. M.* N-terminal arginylation of sciatic nerve and brain proteins following injury [Текст] / Y. M. Wang, N. A. Ingoglia // Neurochem Res. — 1997. — Дек. — Т. 22, № 12. — С. 1453—1459.

145. *Kaji, H.* Correlated Measurement of Endogenous ATE1 Activity on Native Acceptor Proteins in Tissues and Cultured Cells to Detect Cellular Aging [Текст] / H. Kaji, A. Kaji // Methods Mol Biol. — 2015. — Т. 1337. — С. 39—48.

146. *Lamon, K. D.* Arginyl-tRNA transferase activity as a marker of cellular aging in peripheral rat tissues [Текст] / K. D. Lamon, H. Kaji // Exp Gerontol. — 1980. — Т. 15, № 1. — С. 53—64.

147. Arginyltransferase suppresses cell tumorigenic potential and inversely correlates with metastases in human cancers [Текст] / R. Rai [и др.] // Oncogene. — 2016. — Авг. — Т. 35, № 31. — С. 4058—4068.

148. Reduced Arginyltransferase 1 is a driver and a potential prognostic indicator of prostate cancer metastasis [Текст] / M. D. Birnbaum [и др.] // Oncogene. — 2019. — Февр. — Т. 38, № 6. — С. 838—851.

149. *Galiano, M. R.* Post-translational protein arginylation in the normal nervous system and in neurodegeneration [Текст] / M. R. Galiano, V. E. Goitea, M. E. Hallak // J Neurochem. — 2016. — Авг. — Т. 138, № 4. — С. 506—517.

150. *Leu, N. A.* Conditional Tek promoter-driven deletion of arginyltransferase in the germ line causes defects in gametogenesis and early embryonic lethality in mice [Текст] / N. A. Leu, S. Kurosaka, A. Kashina // PLoS One. — 2009. — Нояб. — Т. 4, № 11. — e7734.

151. *Kwon, Y. T.* Alternative splicing results in differential expression, activity, and localization of the two forms of arginyl-tRNA-protein transferase, a component of the N-end rule pathway [Текст] / Y. T. Kwon, A. S. Kashina, A. Varshavsky // Mol Cell Biol. — 1999. — Янв. — Т. 19, № 1. — С. 182—193.

152. *Rai, R.* Identification of mammalian arginyltransferases that modify a specific subset of protein substrates [Текст] / R. Rai, A. Kashina // Proc Natl Acad Sci U S A. — 2005. — Июль. — Т. 102, № 29. — С. 10123—10128.

153. Arginyltransferase, its specificity, putative substrates, bidirectional promoter, and splicing-derived isoforms [Текст] / R. G. Hu [и др.] // J Biol Chem. — 2006. — Окт. — Т. 281, № 43. — С. 32559—32573.

154. The splicing landscape is globally reprogrammed during male meiosis [Текст] / R. Schmid [и др.] // Nucleic Acids Res. — 2013. — Дек. — Т. 41, № 22. — С. 10170—10184.

155. Liat1, an arginyltransferase-binding protein whose evolution among primates involved changes in the numbers of its 10-residue repeats [Текст] / C. S. Brower [и др.] // Proc Natl Acad Sci U S A. — 2014. — Нояб. — Т. 111, № 46. — E4936—4945.

156. The UCSC Genome Browser database: 2019 update [Текст] / M. Haeussler [и др.] // Nucleic Acids Res. — 2019. — Янв. — Т. 47, № D1. — С. D853—D858.

157. Conserved long-range base pairings are associated with pre-mRNA processing of human genes [Текст] / S. Kalmykova [и др.] // Nat Commun. — 2021. — Апр. — Т. 12, № 1. — С. 2300.

Список рисунков

Список таблиц

Приложение А

Неаннотированные тандемные дупликации экзонов в геноме человека

В данном разделе приводятся списки наиболее экспрессируемых по данным GTEx геномных участков, не аннотированных как экзоны, но обладающих высокой гомологией с аннотированными экзонами. Для визуализации таблиц в UCSC геном браузере следует указать ссылку `https://raw.githubusercontent.com/tim-ivanov/tracks_bb/master/hub.txt` в окне геном браузера по адресу `http://genome.ucsc.edu/cgi-bin/hgHubConnect`.

А.1 Неаннотированные тандемные дупликации в НТО

Список 100 наиболее экспрессируемых по данным GTEx геномных участков, гомологичных экзонам из НТО и не аннотированных как экзоны. В таблице перечисляются координаты участка (сборка генома человека GRCh37); ткань, в которой он экспрессируется; c, средний уровень покрытия чтениями (число чтений на нуклеотид); e, средний уровень консервативности (на нуклеотид) по phastCons; j, число поддерживающих сплит-чтений, и название гена.

Таблица 3 — Неаннотированные тандемные дупликации экзонов в НТО.

№	Координаты	Ткань	c	e	j	Ген
1	chr17:27407097-27407173	Скелетная мышца	4006.91	0.53	80428	MYO18A
2	chr10:38737638-38737717	Семенники	1043.37	0.27	89991	LINC00999
3	chr10:38737288-38737368	Семенники	801.10	0.47	89991	LINC00999
4	chr10:38737156-38737236	Семенники	707.45	0.21	89991	LINC00999
5	chr11:65415132-65415215	Подвздошная кишка	253.13	0.17	177048	SIPA1
6	chr19:54411881-54412020	Фронтальная кора мозга	196.81	0.61	3	PRKCG
7	chr10:38714746-38714811	Семенники	193.75	0.10	168139	LINC00999
8	chr13:76287258-76287317	Фибробласты	166.39	0.44	1139	LMO7
9	chr16:30934666-30934750	Семенники	150.00	0.36	1042	FBXL19
10	chr9:131870921-131871021	Семенники	113.16	0.17	38	CRAT
11	chr11:117702922-117702997	Легкое	99.75	0.51	431	FXYD6-FXYD2
12	chr6:34205390-34205498	Фибробласты	94.43	0.22	818531	HMGA1
13	chr9:131799043-131799130	Мозжечок	94.01	0.10	813	FAM73B
14	chr3:105148508-105148582	Легкое	92.49	0.26	597877	ALCAM
15	chr2:144960152-144960235	Легкое	91.53	0.25	138495	GTDC1
16	chr5:63461492-63461615	Легкое	87.16	0.13	33223	RNF180
17	chr7:80142113-80142209	Легкое	86.17	0.21	511	CD36
18	chr7:90390351-90390443	Легкое	86.12	0.92	965	CDK14

19	chr1:217064543-217064641	Легкое	85.86	0.93	922	ESRRG
20	chr4:123221928-123222012	Легкое	85.00	0.11	37179	KIAA1109
21	chr6:90999696-90999780	Легкое	84.64	0.64	930	BACH2
22	chr5:41178409-41178518	Легкое	84.36	0.12	499	C6
23	chr4:62088233-62088299	Легкое	83.77	0.62	24368	LPHN3
24	chr3:168739651-168739709	Легкое	83.00	0.62	19	MECOM
25	chr3:196014255-196014342	Легкое	82.59	0.33	43	PCYT1A
26	chr18:53005236-53005341	Легкое	81.35	0.82	7002	TCF4
27	chr2:145095908-145095994	Легкое	80.58	0.10	138495	GTDC1
28	chr3:114100393-114100462	Легкое	80.46	0.36	90	ZBTB20
29	chr7:14971198-14971298	Легкое	79.70	0.12	4419	DGKB
30	chr1:198209395-198209520	Легкое	79.68	0.11	10117	NEK7
31	chr6:35227626-35227732	Щитовидная железа	79.47	0.78	152409	ZNF76
32	chr5:75380660-75380754	Легкое	79.00	0.30	1373	SV2C
33	chr10:93724755-93724843	Легкое	78.62	0.14	80542	BTAF1
34	chr8:102890285-102890389	Легкое	78.48	0.87	743	NCALD
35	chr1:82497554-82497666	Легкое	77.07	0.63	4639	LPHN2
36	chr15:100192066-100192142	Легкое	76.97	0.11	332274	MEF2A
37	chr1:161601788-161601890	Легкое	76.57	0.15	175	FCGR2B
38	chr11:12746145-12746243	Легкое	75.70	0.45	1219	TEAD1
39	chr5:175499520-175499715	Легкое	75.31	0.11	8771	FAM153B
40	chr2:12858928-12859011	Легкое	75.13	0.23	23939	TRIB2
41	chr1:81873084-81873188	Легкое	74.13	0.88	4639	LPHN2
42	chr17:45501538-45501632	Полушарие мозжечка	73.98	0.11	7038	EFCAB13
43	chr8:140715529-140715652	Полушарие мозжечка	72.98	0.89	72628	TRAPPC9
44	chr16:57668083-57668177	Легкое	72.97	0.18	15916	GPR56
45	chr7:90416365-90416455	Легкое	72.56	0.37	965	CDK14
46	chr17:27482674-27482755	Легкое	71.95	0.30	80428	MYO18A
47	chr3:188122593-188122688	Легкое	71.45	0.55	6514	LPP
48	chr1:216673666-216673754	Легкое	71.02	0.21	922	ESRRG
49	chr1:33721467-33721555	Легкое	69.25	0.26	23429	ZNF362
50	chr4:185017809-185017917	Легкое	69.18	0.10	63	ENPP6
51	chr3:38690842-38690938	Легкое	68.94	0.22	23371	SCN5A
52	chr4:169609904-169610002	Легкое	68.86	0.16	22449	PALLD
53	chr3:149688291-149688394	Легкое	68.63	0.17	1986	PFN2
54	chr14:56047514-56047628	Легкое	68.51	0.33	157894	KTN1
55	chr15:75488318-75488405	Легкое	68.10	0.47	167053	C15orf39
56	chr1:94050734-94050817	Легкое	67.98	0.14	3055	BCAR3
57	chr21:30890750-30890864	Легкое	67.77	0.10	2185	BACH1
58	chr2:155389831-155389896	Легкое	66.46	0.39	28247	GALNT13
59	chr4:115531970-115532060	Легкое	66.44	0.11	3331	UGT8
60	chr18:59000686-59000802	Легкое	66.29	0.21	3	CDH20
61	chr11:1889688-1889784	Легкое	66.23	0.14	3264	LSP1
62	chr7:63773715-63773933	Легкое	66.11	0.16	56	ZNF736
63	chr14:67716224-67716305	Легкое	65.36	0.26	92922	MPP5
64	chr7:23148225-23148366	Легкое	64.46	0.31	9795	KLHL7
65	chr2:176866682-176866800	Легкое	64.37	0.11	41760	KIAA1715
66	chr11:46151042-46151111	Легкое	63.43	0.25	984	PHF21A
67	chr16:49867094-49867211	Легкое	63.26	0.26	4	ZNF423
68	chr14:57857446-57857560	Легкое	62.68	0.79	21182	NAA30
69	chr17:66508763-66508864	Легкое	62.47	0.57	747787	PRKAR1A
70	chr11:47279320-47279407	Легкое	62.30	0.34	31408	NR1H3
71	chr5:175499744-175499939	Легкое	61.63	0.35	8771	FAM153B
72	chr1:198126442-198126541	Легкое	61.32	0.14	44	NEK7
73	chr17:1987871-1987953	Легкое	61.01	0.96	3453	SMG6
74	chr11:17229212-17229336	Легкое	60.74	0.17	13455	PIK3C2A
75	chr7:5323082-5323199	Легкое	60.68	0.16	12895	SLC29A4
76	chr5:176046257-176046344	Легкое	60.46	0.13	465547	SNCB
77	chr16:3364306-3364398	Легкое	60.39	0.52	218256	ZNF75A
78	chr16:57481537-57481634	Легкое	59.23	0.28	337	COQ9
79	chr1:16085691-16085788	Аорта	59.05	0.18	393078	FBLIM1

80	chr4:102267555-102267636	Легкое	56.44	0.79	4	PPP3CA
81	chr4:7477234-7477340	Легкое	55.36	0.29	3	SORCS2
82	chr2:74774731-74774815	Легкое	55.29	0.12	30082	LOXL3
83	chr1:33358586-33358670	Легкое	51.69	0.11	241625	HPCA
84	chr5:175501285-175501457	Легкое	50.35	0.58	8771	FAM153B
85	chr10:93922163-93922242	Легкое	49.78	0.16	2535	CPEB3
86	chr15:100106347-100106472	Легкое	49.53	0.12	1557	MEF2A
87	chr4:170947223-170947336	Легкое	49.42	0.13	9015	MFAP3L
88	chr19:37187889-37187990	Легкое	49.16	0.14	31708	ZNF567
89	chr3:71446334-71446421	Легкое	48.79	0.28	117291	FOXP1
90	chr14:24476276-24476670	Легкое	48.19	0.73	5934	DHRS4L2
91	chr14:65882775-65882874	Легкое	45.24	0.11	168620	FUT8
92	chr4:95679518-95679620	Легкое	42.96	0.25	6327	BMPR1B
93	chr6:52226797-52226917	Легкое	42.08	0.33	59061	PAQR8
94	chr5:177445366-177445561	Легкое	41.14	0.31	11152	FAM153C
95	chr16:639877-639933	Легкое	39.59	0.18	9996	RAB40C
96	chr5:177445142-177445337	Легкое	38.48	0.44	11152	FAM153C
97	chrX:149533193-149533305	Легкое	36.23	0.20	4975	MAMLD1
98	chrX:79693095-79693222	Легкое	35.46	0.21	549	FAM46D
99	chr5:177190692-177190864	Легкое	32.10	0.82	3890	FAM153A
100	chr5:177192210-177192405	Легкое	31.77	0.65	3890	FAM153A

A.2 Неаннотированные тандемные дупликации в кодирующих областях

Список 100 наиболее экспрессируемых по данным GTEx геномных участков, гомологичных кодирующим экзонам и не аннотированных как экзоны. Названия столбцов как и в табл. 4.

Таблица 4 — Неаннотированные тандемные дупликации экзонов в кодирующих областях.

№	Координаты	Ткань	c	e	j	Ген
1	chr16:20740551-20740679	Яичники	696.97	0.69	212616	ACSM3
2	chr9:133254768-133254854	Большеберцовая артерия	625.95	0.65	53921	HMCN2
3	chr11:35215971-35216088	Кожа обл.	514.68	0.62	158223	CD44
4	chr9:133249881-133249965	Большеберцовая артерия	508.75	0.51	53921	HMCN2
5	chr16:20744228-20744271	Яичники	464.12	0.76	136748	ACSM3
6	chr16:20740775-20740875	Яичники	436.75	0.76	208102	ACSM3
7	chr17:37562641-37562734	Кожа обл.	358.15	0.98	149977	MED1
8	chr16:20739709-20739811	Яичники	357.71	0.52	182471	ACSM3
9	chr16:20739484-20739564	Яичники	345.43	0.64	184024	ACSM3
10	chr16:20742941-20743062	Яичники	324.96	0.58	238589	ACSM3
11	chr1:228501314-228501587	Скелетная мышца	313.79	0.88	438820	OBSCN
12	chr2:152379250-152379358	Скелетная мышца	301.85	0.90	3684384	NEB
13	chr16:20738664-20738788	Яичники	271.21	0.75	187294	ACSM3
14	chr16:20729669-20729815	Яичники	269.32	0.71	176680	ACSM3
15	chr12:2851670-2851774	Пищевод	257.17	0.99	546659	CACNA1C
16	chr16:20736683-20736763	Яичники	257.11	0.60	187187	ACSM3
17	chr12:2101206-2101321	Матка	191.07	0.52	300203	CACNA1C
18	chr11:13736520-13736693	Семенники	166.42	0.94	398879	FAR1

19	chr4:120517777-120517865	Аорта	163.09	0.56	313490	PDE5A
20	chr2:16741137-16741255	Легкое	155.88	0.75	492649	FAM49A
21	chr10:73484903-73484990	Пищевод	150.60	0.50	51510	CDH23
22	chr12:52660329-52660497	Пищевод	137.47	0.63	341518	KRT86
23	chr12:2851239-2851346	Полушарие мозжечка	127.57	0.95	539674	CACNA1C
24	chr5:102519739-102519846	Кожа обл.	113.38	0.53	90681	PPIP5K2
25	chr4:20523635-20523787	Пищевод	111.11	0.57	142552	SLIT2
26	chr3:9857379-9857498	Пищевод	109.05	0.81	400252	TTLL3
27	chr4:20523897-20524063	Пищевод	106.43	0.56	168582	SLIT2
28	chr15:49056452-49056568	Легкое	98.53	0.67	57965	CEP152
29	chr8:79585873-79585946	Легкое	96.15	0.56	246145	ZC2HC1A
30	chr5:39388602-39388665	Легкое	91.43	0.51	961937	DAB2
31	chr7:117524041-117524104	Легкое	87.90	0.52	167333	CTTNBP2
32	chr10:112744616-112744756	Легкое	87.39	0.92	530586	SHOC2
33	chr2:63265685-63265794		86.56	0.78	1112676	EHBP1
34	chr10:78868043-78868135	Легкое	86.35	1.00	381981	KCNMA1
35	chr2:152628119-152628221	Легкое	85.09	0.74	98223	NEB
36	chr21:22596158-22596269	Легкое	84.57	0.51	105476	NCAM2
37	chr15:57089592-57089691	Легкое	84.09	0.59	189785	ZNF280D
38	chr2:166238994-166239100	Легкое	82.92	0.66	88924	SCN2A
39	chr5:66149547-66149655	Легкое	82.90	0.81	65164	MAST4
40	chr20:9224962-9225081	Легкое	82.86	0.74	157710	PLCB4
41	chr4:128858648-128858704	Пищевод	82.43	0.88	174557	MFSD8
42	chr2:157454149-157454253	Легкое	82.07	0.96	233922	GPD2
43	chr1:164564998-164565074	Легкое	82.05	0.62	404944	PBX1
44	chr2:225799489-225799569	Легкое	81.20	0.79	129281	DOCK10
45	chr12:28636615-28636714	Легкое	80.34	0.95	807021	CCDC91
46	chr6:117850537-117850608	Легкое	80.30	0.63	111905	DCBLD1,GOPC
47	chr6:144721035-144721139	Легкое	79.56	0.73	160563	UTRN
48	chr1:12508909-12509006	Легкое	79.25	0.68	859447	VPS13D
49	chr10:22627941-22628030	Легкое	78.52	0.96	81520	SPAG6
50	chr18:47794273-47794420	Полушарие мозжечка	78.00	0.85	88844	MBD1
51	chr5:159825734-159825823	Полушарие мозжечка	77.22	0.74	840092	C5orf54
52	chr15:93532974-93533146	Легкое	76.97	0.64	481935	CHD2
53	chr2:157365839-157365942	Легкое	76.63	0.84	278542	GPD2
54	chr2:58431769-58431870	Легкое	76.31	0.90	301275	FANCL
55	chr7:25172961-25173049	Легкое	75.89	0.60	154894	C7orf31
56	chr12:15992445-15992524	Легкое	75.80	0.63	1052245	EPS8
57	chr5:141029222-141029341	Легкое	75.60	0.77	573143	FCHSD1
58	chr5:142574294-142574395	Легкое	74.95	0.67	127406	ARHGAP26
59	chr9:72842756-72842850	Легкое	74.23	0.54	998632	SMC5-AS1
60	chr1:161563215-161563272	Легкое	74.16	0.61	94022	FCGR2B
61	chr4:140696560-140696650	Легкое	74.11	0.61	84073	MAML3
62	chr4:54786931-54787059	Легкое	73.55	0.56	659812	FIP1L1
63	chr12:2848654-2848789	Легкое	73.22	0.80	484053	CACNA1C
64	chr7:120975470-120975541	Легкое	72.42	0.83	325191	WNT16
65	chr8:70378289-70378352	Легкое	72.33	0.87	963013	SULF1
66	chr6:114282943-114283068	Легкое	71.69	0.59	485527	HDAC2
67	chr3:141217945-141218019	Легкое	71.23	0.53	66400	RASA2
68	chr4:148972400-148972522	Легкое	70.79	0.82	114349	ARHGAP10
69	chr9:128582857-128582948	Легкое	70.78	1.00	625141	PBX3
70	chr15:64002301-64002432	Легкое	69.44	0.52	113251	HERC1
71	chr4:62929498-62929592	Легкое	69.26	0.69	110762	LPHN3
72	chr17:59202107-59202177	Легкое	68.70	1.00	274512	BCAS3
73	chr5:102147489-102147579	Легкое	68.40	0.59	2011985	PAM
74	chr16:20730548-20730710	Яичники	68.36	0.57	209975	ACSM3
75	chr2:209050729-209050811	Легкое	67.56	0.94	124569	C2orf80
76	chr2:206159432-206159554	Легкое	67.38	0.80	189470	PARD3B
77	chr1:237758663-237758759	Легкое	67.30	0.71	555415	RYR2
78	chr12:52661128-52661325	Легкое	66.68	0.76	381544	KRT86
79	chr4:151858247-151858342	Легкое	66.63	0.94	83223	LRBA

80	chr4:99226957-99227069	Легкое	66.51	0.70	336869	RAP1GDS1
81	chr11:126236018-126236086	Легкое	65.35	0.55	753243	ST3GAL4
82	chr6:86225496-86225611	Легкое	65.32	0.59	1269064	SNX14
83	chr16:3364745-3364872		65.15	0.83	139589	ZNF75A
84	chr5:71627856-71627975	Легкое	64.25	0.51	123296	PTCD2
85	chr4:110015457-110015525	Легкое	64.16	0.76	64198	COL25A1
86	chr12:52671294-52671512	Легкое	63.68	0.60	381544	KRT86
87	chr12:46208533-46208750	Легкое	63.32	0.87	80128	ARID2
88	chr11:8892982-8893060	Легкое	61.26	1.00	217176	ST5
89	chr10:21288743-21288841	Легкое	60.60	0.72	282041	NEBL
90	chr1:82207301-82207397	Легкое	60.10	0.62	719775	LPHN2
91	chr3:121514889-121514929	Поджелудочная железа	59.42	0.72	210630	IQCB1
92	chr19:45737763-45737926	Щитовидная железа	57.77	0.75	81182	MARK4
93	chr7:33103545-33103641	Легкое	56.80	0.57	192266	AVL9
94	chr12:21404647-21404712	Легкое	56.66	0.63	61054	SLCO1B1
95	chr2:55529788-55529897	Легкое	56.40	0.73	92956	CCDC88A
96	chr21:38523728-38523795	Легкое	56.03	0.63	908120	TTC3
97	chr3:57258962-57259072	Легкое	54.76	0.72	56450	HESX1
98	chr15:43952137-43952234	Мозжечок	53.68	0.99	50759	CATSPER2
99	chr2:55537899-55538092	Легкое	53.57	0.80	209341	CCDC88A
100	chr16:20722794-20723002	Легкое	53.49	0.63	143849	ACSM3