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

На правах рукописи

Первушин Дмитрий Давидович

Альтернативный сплайсинг и дальние взаимодействия в структуре эукариотических РНК

1.5.3. Молекулярная биология

Диссертация на соискание учёной степени
доктора химических наук

Москва — 2024

Оглавление

Стр.

Стр.

Стр.

Введение

Способность нуклеиновых кислот образовывать двухцепочечные структуры лежит в основе всех известных биологических процессов. В отличие от ДНК, которая, как правило, находится в двунитевом состоянии, бóльшая часть молекул РНК в клетке являются одноцепочечными, но их отдельные участки могут принимать конформации, содержащие двойные спирали, из которых формируется структура. Изменения в структуре эукариотических РНК лежат в основе механизмов регуляции многих клеточных процессов, включая сплайсинг.

Сплайсинг является одним из главных этапов созревания эукариотических пре-мРНК, при котором из них вырезаются интроны, а оставшиеся экзоны соединяются, образуя зрелые мРНК. Сплайсинг может протекать альтернативно, в результате чего из транскриптов одного и того же гена образуется множество различных сплайс-изоформ. Регуляция альтернативного сплайсинга осуществляется за счет скоординированного действия большого числа факторов, включающих в себя РНК-связывающие белки и структуру РНК. Структура РНК способна блокировать цис-регуляторные элементы сплайсинга, приближать или отдалять их друг от друга, создавая конформации, которые нужны для получения необходимых клетке сплайс-изоформ. Неправильное сворачивание пре-мРНК может вызвать нарушения в работе сплайсинга, которые являются причиной тяжелых наследственных, нейродегенеративных и онкологических заболеваний.

Принято различать четыре уровня организации структуры биополимеров: первичную, вторичную, третичную и четвертичную. Первичная структура РНК — это линейная последовательность ее нуклеотидов, соединенных ковалентными фосфодиэфирными связями. Способность нуклеотидов образовывать пары, включая канонические Уотсон-Криковские, неканонические (например, гуанин-урациловые, или G:U пары), а также имидазольные (хугстеновские) и некоторые другие пары, приводит к сворачиванию первичной структуры во вторичную, состоящую из характерных элементов: шпилек, стеблей, внутренних и множественных петель, а также псевдоузлов. Вторичные элементы впоследствии собираются в трехмерные третичные структуры, которые стабилизируются коаксиальным стекингом стеблей и взаимодействиями петель. Наконец, взаимодействия с другими макромолекулами, включая РНК-белковые

взаимодействия, приводят к образованию четвертичных структур. В данной работе, посвященной вторичному уровню организации РНК, рассматриваются только самые распространенные спаривания нуклеотидов — Уотсон-Криковские и G:U пары.

Комплементарные спаривания оснований, из которых состоит структура РНК, можно отнести к локальным и дальним взаимодействиям. Простейшим типом локальной структуры РНК является шпилька. Поскольку сворачивание пре-мРНК происходит ко-транскрипционно, основная часть ее структуры образуется за счет локальных взаимодействий. В отличие от локальных, дальние взаимодействия образуются между комплементарными сайтами, разделенными протяженными участками последовательности (более 100 нт). Дальние взаимодействия обладают некоторыми чертами третичной структуры, но, как и локальные, относятся ко вторичному уровню организации, т.е., определяют укладку полинуклеотидной цепи вследствие спаривания между основаниями. В данной диссертации термин «дальние взаимодействия» используется по отношению к комплементарным взаимодействиям внутри одной и той же молекулы РНК, а не к межмолекулярным взаимодействиям, времени формирования структуры, ее топологии или трехмерной организации.

Развитие технологий высокопроизводительного секвенирования привело к появлению ряда экспериментальных методов для одновременного определения структур в больших ансамблях молекул РНК. Однако структурная гетерогенность молекул РНК, динамическая природа сворачивания и разреженность получаемой информации значительно затрудняют интерпретацию результатов этих экспериментов. Поэтому наряду с экспериментальными методами важное практическое значение имеют вычислительные методы предсказания структуры РНК. Их можно подразделить на термодинамические и филогенетические. Термодинамические методы находят оптимальный набор спариваний оснований, при котором свободная энергия молекулы РНК минимальна. Филогенетические методы предсказывают комплементарность оснований в родственных последовательностях, используя происходящие в них компенсаторные замены. Одновременная минимизация свободной энергии и построение множественного выравнивания представляют из себя знаменитую задачу Санкова, которая не имеет эффективного вычислительного решения. Основной темой данной диссертационной работы является разработка методов, сочетающих в себе термодинамический и филогенетический подходы, для предсказания дальних

взаимодействий в структуре РНК и экспериментальная валидация их результатов.

Актуальность темы исследования. В настоящее время изучение РНК и ее структуры переживает бурный расцвет, однако подавляющее большинство вычислительных исследований моделирует структуру РНК без псевдоузлов. В действительности отсутствие псевдоузлов является техническим ограничением метода динамического программирования, широко используемого для предсказания вторичной структуры РНК. Это делает его неприменимым к исследованию дальних взаимодействий, поскольку алгоритм оптимизации предпочитает опустить высокоэнергетические дальние взаимодействия, которые вследствие запрета на псевдоузлы оказываются несовместимыми с большим числом низкоэнергетических, но суммарно более «выгодных» локальных спариваний. Поэтому разработка новых методов предсказания структуры РНК, учитывающих дальние взаимодействия, является актуальной задачей, имеющей важное фундаментальное значение.

Альтернативный сплайсинг играет определяющую роль в клеточной дифференцировке и развитии организмов, а его нарушения приводят к возникновению болезней. Сплайс-изоформы, специфически экспрессируемые в опухолевых клетках, все чаще используются для диагностики, прогноза и таргетной терапии многих типов рака. Несмотря на значительный прогресс, достигнутый в исследовании альтернативного сплайсинга, роль большинства сплайс-изоформ в физиологических и патологических процессах остается неизвестной, как полностью не изучены и управляющие регуляцией этого процесса факторы, в число которых входит структура пре-мРНК. В последние годы значительно увеличилось число экспериментально подтвержденных примеров функциональных структур РНК, влияющих на альтернативный сплайсинг, а также предпринимаются попытки идентифицировать структуру РНК высокопроизводительными методами. Несмотря на это, уровень структурированности пре-мРНК и степень распространенности дальних взаимодействий остаются во многом неизученными. Одной из задач данной диссертационной работы является составление полногеномного каталога предсказанных структур РНК в генах человека и его сопоставление с экспериментальными сведениями.

Поскольку правильное сворачивание РНК необходимо для ее нормального функционирования, вполне естественно, что неправильное сворачивание приводит к нарушению регуляции клеточных процессов. Мутации в сайтах,

которые важны для образования структуры РНК и распознавания регуляторными факторами, часто вызывают изменение сплайсинга. Так, мутации, влияющие на дальние взаимодействия в структуре РНК, изменяют частоту использования одного из важных экзонов гена *SMN2*, связанного со спинальной мышечной атрофией [1; 2]. Для для лечения этого тяжелого заболевания в 2016 году Управление по санитарному надзору за качеством пищевых продуктов и медикаментов США одобрило препарат «Спинраза» — антисмысловую олигонуклеотидную терапию, мишенью которой является структура РНК. Таким образом, исследование влияния структуры РНК на альтернативный сплайсинг, а также способов его коррекции с помощью антисмысловых нуклеотидов имеет важное практическое применение.

Функция альтернативного сплайсинга состоит не только в генерации мРНК, кодирующих различные белковые продукты, но и в посттранскрипционной регуляции экспрессии генов. В частности, в процессе так называемого непродуктивного сплайсинга из-за сдвига рамки считывания или включения ядовитых экзонов в мРНК могут вставляться преждевременные стоп кодоны, в результате чего транскрипты деградируют по механизму нонсенс-опосредованного распада. Непродуктивный сплайсинг регулирует уровни экспрессии большого числа генов, а сбои в его работе приводят к развитию патологий. Представляется особенно актуальным изучить роль структуры РНК в регуляции именно непродуктивного сплайсинга, где функции сплайс-изоформ легко прослеживаются, в отличие от альтернативного сплайсинга в целом, где функции белоккодирующих транскриптов далеко не всегда известны. Нахождению ответов на этот и другие актуальные вопросы посвящена настоящая диссертационная работа.

Степень разработанности темы. О том, что молекулы РНК имеют естественную склонность образовывать высокостабильные вторичные структуры, а изменения в этих структурах представляют собой механизм регуляции клеточных процессов было известно еще на заре молекулярной биологии. Согласно классической концепции, эукариотические РНК сразу после транскрипции покрываются РНК-связывающими белками, что препятствует их сворачиванию [3]. Поэтому долгое время считалось, что после транскрипции они могут сворачиваться лишь в течение очень ограниченного периода времени и образуют в основном локальную структуру.

Однако постепенно становилось понятно, что дальние взаимодействия в структуре РНК играют важную роль в регуляции сплайсинга. Ярким примером является открытый в 2005 году механизм взаимоисключающего сплайсинга в гене клеточной адгезии синдрома Дауна (*Dscam1*) дрозофилы, пре-мРНК которого содержит конкурирующие комплементарные спаривания оснований, взаимодействующие на расстоянии до 10000 п.о. [4]. Было показано, что конкурирующие структуры РНК определяют включение только одного из 48 вариабельных экзонов в кластере экзонов 6, а затем аналогичный механизм был обнаружен и в других экзонах этого гена. Позднее дальние взаимодействия, регулирующие альтернативный сплайсинг, были обнаружены в десятках других эукариотических генов, а также в геномах вирусов, включая SARS-CoV-2 [5; 6]. Была показана определяющая роль РНК-структур с дальними взаимодействиями в регуляции многих биологических процессов, связанных с развитием и нейрогенезом. Отдельные сообщения о структурах РНК, приближающих сайты связывания РНК-связывающих белков к сайтам сплайсинга, комплементарных областях, способствующих образованию кольцевых РНК, а также роли дальних взаимодействий в транс-сплайсинге появлялись в литературе [7; 8], однако все они касались генов, интерес к которым был обусловлен исследованием конкретных биологических систем.

Предсказание вторичной структуры РНК является второй по древности задачей биоинформатики после задачи выравнивания гомологичных последовательностей, причем в основе решения обеих лежит метод динамического программирования [9]. На сегодняшний день наиболее популярным алгоритмом предсказания структуры РНК по последовательности является метод минимизации свободной энергии, который реализован во многих программных пакетах и использует экспериментально определенные термодинамические параметры [10; 11]. Его основными ограничениями являются возрастающая неточность, увеличивающаяся сложность вычислений и неспособность учитывать дальние взаимодействия. Филогенетические методы, оценивающие частоты компенсаторных замен нуклеотидов в выравниваниях гомологичных последовательностей, еще более вычислительно затратны и, в сущности, тоже направлены на предсказание локальных структур РНК, хотя и не содержат явного запрета на псевдоузлы [12; 13].

В последние годы все большую популярность приобретают методы предсказания структуры РНК, основанные на машинном обучении. К ним относятся

как традиционные статистические методы, так и методы, основанные на использовании нейронных сетей [14; 15]. Однако в отличие от методов, в которых параметры оцениваются на основе экспериментов или эволюционных моделей, методы машинного обучения вычисляют параметры, исходя из небольшого набора известных структур, что неизбежно приводит к смещению в сторону уже исследованных структурных типов и переобучению модели. В настоящее время из литературы известны лишь десятки примеров функциональных дальних взаимодействий в структуре РНК, что не позволяет полноценно обучить сложные статистические модели.

В 2008 году первое систематическое исследование влияния структуры РНК на альтернативный сплайсинг, использующее структурную и функциональную консервативность, позволило охарактеризовать элементы структуры РНК в геноме человека, которые связаны с выбором альтернативных сплайс-сайтов [16]. Затем было показано, что тысячи гомологичных геномных областей человека и мыши, не совпадающих по нуклеотидной последовательности, тем не менее, содержат общие структуры, что позволило предсказать структуру некоторых некодирующих РНК [17]. Также была реализована задача «перевыравнивания» уже построенных полногеномных выравниваний с учетом структуры РНК для нахождения разошедшихся по нуклеотидной последовательности, но все еще консервативных на уровне структуры РНК участков [18]. Однако эти исследования так не дали ответа на вопрос о распространенности дальних взаимодействий в структуре РНК.

В настоящее время начали появляться методы, которые используют данные высокопроизводительного секвенирования для моделирования структуры РНК. Некоторые из них преобразуют показатели структурной реактивности нуклеотидов в псевдоэнергии и применяют их в моделях, использующих штрафы для спаренных оснований, однако область их применения ограничена локальной структурой [19; 20]. Ответ на вопрос о дальних взаимодействиях дают эксперименты, основанные на лигировании пространственно близких молекул, например метод конформационного секвенирования РНК, который позволил создать карты связности для различных РНК, подобные картам хроматиновых контактов [21]. Несколько лет назад был разработан метод, преобразующий данные псораленового анализа структуры РНК в вероятности образования пар между нуклеотидами, которые могут быть использованы для нахождения репрезентативных ансамблей структур [22]. Таким образом, изуче-

ние структуры РНК и различных аспектов, связанных с ее регуляторной ролью в биологических системах, является интенсивно развивающейся, современной областью исследований.

Целью данной диссертационной работы является разработка методов предсказания дальних взаимодействий в структуре РНК, объединяющих термодинамический и филогенетический подходы, сопоставление результатов их предсказаний с экспериментальными данными, а также применение полученных методов к исследованию влияния структуры РНК на альтернативный сплайсинг и изучение функциональных последствий этого влияния.

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

1. Разработать методы предсказания дальних взаимодействий в структуре РНК, реализующие принципы «сначала выравнивание, потом фолдинг» и «сначала фолдинг, потом выравнивание»;
2. Описать положение предсказанных РНК-структур относительно цис-регуляторных элементов в пре-мРНК и исследовать отклик транскриптома на замедление элонгации транскрипции в зависимости от структуры РНК;
3. Сопоставить предсказания РНК-структур с данными конформационного секвенирования РНК *in situ*;
4. Разработать основанный на данных конформационного секвенирования РНК *in situ* метод предсказания структуры за пределами консервативных областей;
5. Экспериментально валидировать влияние предсказанных РНК-структур на основные типы событий альтернативного сплайсинга;
6. Разработать методы предсказания ауто- и кросс-регуляторного непродуктивного сплайсинга по транскриптомным данным;
7. Исследовать и экспериментально валидировать роль вторичной структуры РНК в регуляции непродуктивного сплайсинга.

Научная новизна работы заключается в следующем:

1. Разработаны новые методы предсказания дальних взаимодействий в структуре РНК, применимые в масштабах эукариотических геномов.
2. Впервые описана взаимосвязь между расположением элементов вторичной структуры РНК и экзон-интронной разметкой генов, позициями

сплайс-сайтов, сайтов редактирования РНК и сайтов связывания РНК-связывающих белков.

3. Впервые показано изменение сплайсинга в зависимости от структурированности РНК при замедлении элонгации транскрипции.
4. Выдвинута гипотеза о роли вторичной структуры РНК и котранскрипционного сплайсинга в предотвращении интронного полиаденилирования и преждевременной терминации транскрипции.
5. Впервые изучены характеристики РНК-структур, поддерживаемых данными конформационного секвенирования РНК *in situ*, и предложен новый метод предсказания структуры РНК за пределами консервативных областей, основанный на этих данных.
6. Впервые экспериментально подтверждено влияние структуры РНК на альтернативный сплайсинг в генах *CG33298*, *Gug* и *Nmnat* дрозофилы, а также в генах *PHF20L1*, *CASK*, *ATE1*, *SF1* и *MARK2* человека.
7. Разработаны антисмысловые олигонуклеотиды для модуляции сплайсинга через структуру РНК в вышеперечисленных генах.
8. Показано существование нескольких функционально различных структурных модулей в пре-мРНК гена *ATE1* человека.
9. Впервые экспериментально подтверждено влияние скорости элонгации транскрипции на альтернативный сплайсинг через дальние взаимодействия в структуре пре-мРНК.
10. Предсказаны новые механизмы ауто- и кросс-регуляции непродуктивного сплайсинга.
11. Впервые предсказан и экспериментально подтвержден механизм тканеспецифической регуляции непродуктивного сплайсинга в генах *DCLK2* и *IQGAP1*.
12. В генах *BRD2* и *BRD3* впервые предсказана и экспериментально подтверждена регуляция непродуктивного сплайсинга РНК-структурами и показано их независимое приобретение в процессе конвергентной эволюции.

Теоретическая и практическая значимость. Теоретическая значимость исследования заключается во всестороннем освещении проблемы поиска дальних взаимодействий в структуре РНК и разработке методов их предсказания, применимых в масштабах эукариотических геномов. В диссертации показывается, что такие методы неизбежно имеют высокую долю

ложноположительных предсказаний, в частности, из-за консервативных мотивов, которые встречаются на противоположных цепях ДНК, а уровень вариабельности нуклеотидных последовательностей в консервативных участках недостаточен для оценки значимости РНК-структур по компенсаторным заменам. Таким образом, работа дает представление о факторах, которые ограничивают предсказательные возможности всех методов, основанных на сравнительной геномике.

Установление взаимосвязи между расположением элементов структуры РНК и цис-регуляторными элементами сплайсинга, в том числе подтверждаемое данными конформационного секвенирования РНК, дает ответ на фундаментальный вопрос молекулярной биологии о так называемом «коде сплайсинга», т.е. объясняет то, как в эукариотических РНК распознаются и вырезаются интроны. Экспериментальное изучение влияния скорости элонгации транскрипции на сплайсинг показывает, что именно структура РНК является медиатором взаимодействия между временными и пространственными компонентами в его регуляции. Эти результаты, а также представленные свидетельства того, что структура РНК и котранскрипционный сплайсинг способствуют предотвращению интронного полиаденилирования и преждевременной терминации транскрипции, имеют важное теоретическое значение.

С точки зрения практической значимости понимание структуры РНК важно для биомедицинских задач. Среди экспериментально подтвержденных структур РНК, влияющих на альтернативный сплайсинг, следует отметить регуляторные структуры в генах человека, связанных с заболеваниями. Большое прикладное значение имеют предсказания регуляторных сетей, значительно расширяющие существующие знания о непродуктивном сплайсинге. Подтверждение роли структуры РНК в регуляции непродуктивного сплайсинга важно для понимания механизмов патогенеза связанных с ним заболеваний. В диссертации продемонстрирована модуляция альтернативного сплайсинга через блокировку структуры РНК антисмысловыми олигонуклеотидами, что открывает возможности для его коррекции, основанные не только на подавлении, но и на активации включения экзонов. Разработка таких антисмысловых олигонуклеотидов может помочь получить индивидуальные лекарственные средства с независимыми правами интеллектуальной собственности.

Кроме того, в данной диссертационной работе получены несколько полногеномных каталогов РНК-структур, которые могут быть использованы

широким кругом исследователей через доступные интерфейсы визуализации[1] для изучения структур РНК в конкретных генах. Поэтому работа также имеет энциклопедическую ценность.

Объекты и методы исследования. Объектами исследования являются нуклеотидные последовательности геномов позвоночных и насекомых, их транскрипты, экзоны, интроны и содержащиеся в них комплементарные участки. Для выполнения работы применялся комплексный подход, включающий в себя вычислительные и экспериментальные методы, такие как термодинамическое моделирование структуры РНК, выравнивание последовательностей, построение хеш-таблиц, анализ компенсаторных мутаций, высокопроизводительное секвенирование РНК, конструирование минигенов, сайт-направленный мутагенез, обратная транскрипция и полимеразная цепная реакция (ОТ-ПЦР), ОТ-ПЦР в реальном времени (ОТ-ПЦР-РВ), суперэкспрессия и подавление экспрессии генов с помощью микроРНК, блокировка структуры РНК антисмысловыми олигонуклеотидами и др.

Основные положения, выносимые на защиту:

1. Комплементарные участки предпочтительно располагаются в интронах, подавляют использование криптических сплайс-сайтов и выпетливаемых экзонов, обогащены сайтами редактирования РНК и сайтами связывания РНК-связывающих белков, и поддерживаются данными конформационного секвенирования РНК *in situ*.
2. Изменение степени включения экзона при замедлении элонгации транскрипции зависит от структурированности предшествующего интрона.
3. Дальние взаимодействия в структуре РНК могут регулировать все основные типы событий альтернативного сплайсинга и альтернативное полиаденилирование, как показывают примеры в генах *CG33298*, *Gug*, *Nmnat*, *PHF20L1*, *CASK*, *ATE1*, *SF1* и *MARK2*.
4. Ген *ATE1* содержит два функционально различных структурных модуля, один из которых обеспечивает взаимоисключающий сплайсинг экзонов, а другой благодаря дальним взаимодействиям на расстоянии 30000 п.о. контролирует соотношение сплайс-изоформ через котранскрипционное сворачивание пре-мРНК.

[1] Представлены в приложениях к [23—25].

5. Непродуктивный сплайсинг может регулироваться РНК-связывающими белками и дальними взаимодействиями в структуре РНК, как показывают примеры в генах *DCLK2*, *IQGAP1*, *BRD2* и *BRD3*.

Достоверность результатов, в частности, предсказаний дальних взаимодействий в структуре РНК обеспечивается их экспериментальной валидацией в рамках данной диссертационной работы, а также сравнением с экспериментальными данными, полученными другими авторами. Все полученные результаты обосновываются оценками статистической значимости и построением доверительных интервалов. Результаты работы полностью согласуются с результатами, известными из литературных источников. Достоверность полученных результатов подтверждается публикациями в ведущих рецензируемых научных журналах.

Апробация работы. Основные результаты работы были доложены автором на следующих конференциях и конгрессах: Московская конференция по вычислительной молекулярной биологии (MCCMB), Москва, РФ (2015, 2017, 2019, 2021, 2023 гг.); VI съезд биохимиков России, Дагомыс, РФ (2019 г.); конференция «Информационные технологии и системы» (ИТиС), Казань, РФ (2018 г.); международная конференция «Вычислительные подходы к структуре и функциям РНК», Бенаске, Испания (2009, 2012, 2015, 2018 и 2022 гг.); международная конференция по интеллектуальным системам молекулярной биологии (ISMB), Берлин, ФРГ (2013 г.), Прага, Чехия (2017 г.); международная конференция по исследованиям в области вычислительной молекулярной биологии (RECOMB), Барселона, Испания (2012 г.); ежегодный конгресс консорциума «Энциклопедия элементов ДНК» (ENCODE), Сан Диего, США (2014 и 2016 гг.); международная конференция «Биология Геномов», Нью Йорк, США (2014, 2015, 2016 гг.); международный конгресс по высокопроизводительному секвенированию РНК, Барселона, Испания (2017, 2018, 2022 гг.); международный симпозиум «Регуляторные сети РНК», Лиссабон, Португалия (2019 г.); открытый семинар кафедры биомедицинской информатики, Гарвардский университет, Бостон, США (2018 г.).

Личный вклад. Биоинформатическая часть работы была выполнена автором лично либо в соавторстве при непосредственном руководстве на всех этапах проведения исследования. Имена соавторов по научным коллективам указаны в соответствующих публикациях. Вклад автора во всех опубликованных работах, за исключением публикаций в составе консорциумов [26—29],

является определяющим. Экспериментальные результаты, изложенные в гл. 4, были получены в соавторстве с проф. Юаньчао Сюэ и проф. Чанчан Цао (Китайская Академия Наук, КНР), а также проф. Юнфэн Джин (Чжэцзянский университет, КНР). Экспериментальная валидация в гл. 4 и гл. 5 проводилась в сотрудничестве с проф. Хуаном Валкарселем (Центр Геномной Регуляции, г. Барселона), проф. П.М. Рубцовым (Институт Молекулярной Биологии им. Энгельгардта РАН) и проф. О.А. Донцовой (МГУ им. М.В. Ломоносова). Эксперименты по высокопроизводительному секвенированию РНК проводились при поддержке Центра Коллективного Пользования «ГЕНОМИКА» Сколковского института науки и технологий. Под руководством автора диссертации в рамках темы данной работы подготовлены и защищены четыре кандидатские диссертации и более 20 выпускных квалификационных работ специалистов и магистров.

Диссертационная работа была выполнена при поддержке гранта Российского фонда фундаментальных исследований №10-04-00783 «Полногеномное изучение альтернативного сплайсинга и его взаимосвязи со вторичной структурой пре-мРНК», гранта Российского фонда фундаментальных исследований №19-34-90174 «Эволюция взаимоисключающих экзонов и регуляция альтернативного сплайсинга вторичной структурой РНК», гранта Российского фонда фундаментальных исследований №18-29-13020 «Идентификация и функциональная валидация опухолеспецифических изменений сплайсинга, вызванных соматическими мутациями в структурных элементах пре-мРНК», исследовательского гранта №RF-0000000653 Сколковского института науки и технологии, гранта Министерства науки и высшего образования Российской Федерации №075-10-2021-116 «Вторичная структура РНК как регулятор альтернативного сплайсинга и лекарственная мишень» и гранта Российского научного фонда №22-14-00330 «Изучение регуляторных сетей непродуктивного сплайсинга в норме и патологии», в которых автор диссертации являлся руководителем, а также при поддержке гранта Российского научного фонда №21-64-00006 «Генетические технологии создания моделей заболеваний, обусловленных нарушениями функционирования РНК», в котором автор диссертации являлся исполнителем (руководитель проф. О.А. Донцова).

Публикации. Основные результаты по теме диссертации изложены в 40 публикациях и одном патенте РФ, приравненном к публикации. Из них 33 статьи опубликованы в периодических научных журналах, индексируемых Web

of Science и Scopus, рекомендованных для защиты в диссертационном совете МГУ.014.2.

Объем и структура работы. Диссертация состоит из введения, 6 глав, и заключения. Полный объём диссертации составляет 248 страниц, включая 57 рисунков и 8 таблиц. Список литературы содержит 575 наименований.

Глава 1. Обзор литературы

1.1 Альтернативный сплайсинг (АС)

Большинство эукариотических транскриптов в процессе созревания подвергаются сплайсингу — процессу, при котором участки, называемые интронами, удаляются, а оставшиеся экзоны соединяются, образуя зрелые мРНК [30]. В подавляющем числе случаев сплайсинг катализируется сложным макромолекулярным комплексом, называемым сплайсосомой, который состоит из малых ядерных рибонуклеопротеинов, (мяРНП), в свою очередь состоящих из малых ядерных РНК (мяРНК) и связанных с ними белков [31—33].

Сплайсосома распознает цис-регуляторные элементы в пре-мРНК, среди которых следует выделить четыре основных: 5' сайт сплайсинга (5'ss), 3' сайт сплайсинга (3'ss), полипиримидиновый тракт (polypyrimidine tract, PPT) и сайт ветвления (branch point sequence, BPS) [34]. Однако сплайсинг одинаковых транскриптов может происходить по-разному из-за распознавания на них различных сплайс-сайтов, а также в результате их комбинирования в различных сочетаниях. Таким образом, вследствие альтернативного сплайсинга (АС) в клетке можно обнаружить множество различных изоформ зрелой мРНК, получаемых в результате сплайсинга транскриптов одного и того же гена.

Из многообразия событий АС можно выделить несколько основных типов, например пропуск (кассетного) экзона, использование альтернативного 5'ss или 3'ss, удержание интрона, а также выбор одного из нескольких взаимоисключающих экзонов, однако существуют и более сложные типы событий АС [35; 36]. По современным оценкам не менее 95% генов человека, состоящих из более чем одного экзона, подвергаются альтернативному сплайсингу [37; 38], а скоординированные изменения сплайсинга множества пре-мРНК являются неотъемлемой частью регуляции ряда клеточных процессов [39—41].

АС регулируется комбинацией РНК-белковых, РНК-РНК и белок-белковых взаимодействий, которые возникают между цис-регуляторными элементами и транс-действующими факторами [42; 43]. Помимо описанных выше ключевых элементов (5'ss, 3'ss, PPT, BPS) на АС оказывают влияние дополнительные цис-регуляторные элементы, которые могут располагаться как в

экзонах, так и в интронах. Они называются экзонными и интронными энхансерами и сайленсерами сплайсинга. Их взаимодействие с транс-действующими факторами стимулирует или подавляет выбор сайта сплайсинга, соответственно [44]. Результат сплайсинга зависит от согласованного действия множества энхансеров и сайленсеров [45].

1.1.1 Регуляция АС РНК-связывающими белками (РСБ)

В регуляции АС принимают участие более полутора тысяч РНК-связывающих белков (РСБ) [46]. Их можно разделить на несколько классов: гетерогенные ядерные рибонуклеопротеиды (heterogeneous nuclear ribonucleoproteins, hnRNP), серин/аргинин-богатые белки (serine/arginine-rich proteins, SR), и остальные, например тканеспецифические РНК-связывающие белки, такие как NOVA, нейрональные PTB/hnRNP I, семейство RBFOX и др. [35]. Здесь коротко перечисляются примеры, имеющие отношение к структуре РНК, а более подробные сведения о регуляции АС различными классами РСБ можно найти в литературе [35; 47—49].

Повсеместно экспрессируемые белки из семейств SR и hnRNP являются наиболее изученными медиаторами распознавания сайтов сплайсинга [3; 50—53]. SR белки участвуют как в конститутивном, так и в альтернативном сплайсинге, что делает это семейство РНК-связывающих белков уникальным по сравнению с другими РНК-связывающими белками [51]. SR белки обычно рассматриваются как положительные регуляторы сплайсинга. Они способствуют включению экзона, помогая рекрутировать U1 мяРНП в 5'ss и вспомогательный фактор U2 (U2AF) в 3'ss посредством белок-белковых взаимодействий на ранних стадиях сборки сплайсосомы [50; 54].

Белки семейства hnRNP и SR-белки считаются антагонистами. Природа этого антагонизма не совсем ясна, так как высокоаффинные сайты-связывания hnRNP нечасто перекрываются с сайтами связывания SR-белков в экзонах. Потенциальный механизм предполагает совместное связывание олигомеров hnRNP, которое распространяется вдоль транскрипта, чтобы предотвратить связывание SR-белков с РНК [53]. Наиболее охарактеризованными среди hnRNP, участвующих в регуляции сплайсинга, являются негативные регуля-

торы hnRNP A/B и белок PTB, связывающий PPT, также известный как hnRNP I. Фактор hnRNP A2/B1 в основном является ингибитором сплайсинга, который препятствует распознаванию 5'ss и 3'ss, что чаще приводит к исключению альтернативного экзона (подробно функции hnRNP A/B изложены в [55]). PTB связывается с полипиримидиновыми участками, так же как и U2AF65, который способствует связыванию U2 мяРНП с 3'ss. Это подразумевает, что PTB может мешать функциональному распознаванию 3'ss [56]. Механизм и направление действия белков семейства hnRNP зависит от расположения их сайтов связывания: при связывании перед или внутри кассетного экзона они как правило действуют как репрессоры, при связывании после — как активаторы АС [48; 57; 58].

Помимо SR и hnRNP белков охарактеризовано несколько тканеспецифических РНК-связывающих регуляторов сплайсинга. К ним относятся специфические для нейронов факторы NOVA [59], PTBP2 [60] и SRRM4 [61], а также такие тканеспецифические факторы, как белки семейства RBFOX [7], MBNL [62; 63], CELF [64], QKI [65] и TIA [66; 67]. Их действие может быть обусловлено как тканеспецифической экспрессией, так и связыванием с мотивами пре-мРНК, которыми обогащены гены, экспрессирующиеся в определенном типе клеток или ткани. Тканеспецифические регуляторы АС чаще всего исследуются в связи с различными патологиями, например, нейродегенеративными заболеваниями или мышечной дистрофией [68—70].

Для привлечения и правильного распределения факторов сплайсинга на их сайты связывания необходимо присутствие РНК-полимеразы II (RNAPII). Соответственно, транскрипция и сплайсинг взаимно влияют друг на друга за счет пространственных и кинетических механизмов [71]. РНК-полимераза II имеет С-концевой домен гептадных повторов, который используется в качестве «посадочной площадки» для доступных факторов, что позволяет увеличить их концентрацию рядом с сайтами сплайсинга [72—75]. Скорость элонгации транскрипции влияет на протекание АС, определяя то, насколько быстро сайты сплайсинга становятся доступными для конкуренции за связывание с транс-действующими факторами, в том числе за счет образования вторичной структуры пре-мРНК [76—80].

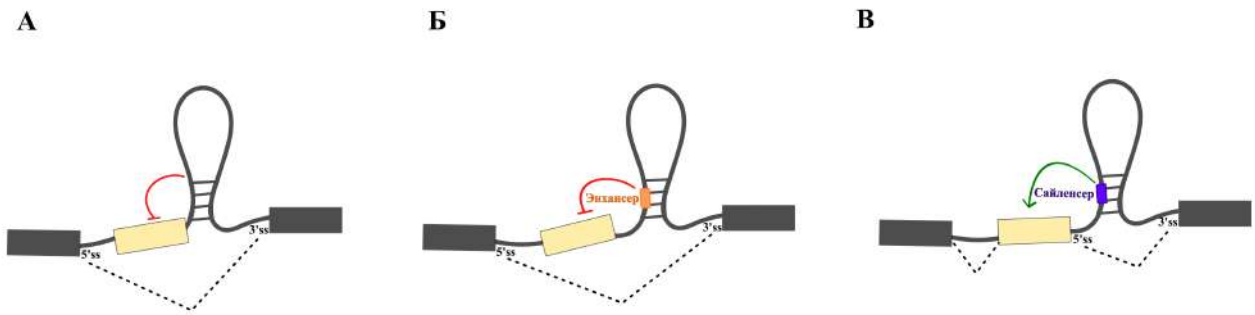


Рисунок 1.1 — Блокировка цис-регуляторных элементов сплайсинга структурой РНК. (А) Блокировка сайта сплайсинга; (Б) Блокировка интронного энхансера; (В) Блокировка интронного сайленсера сплайсинга. Красными и зелеными линиями обозначено активирующее и ингибирующее действие на сплайсинг, соответственно.

1.1.2 Регуляция АС структурой пре-мРНК

Существует множество экспериментально подтвержденных данных о регуляции АС локальной структурой пре-мРНК, например, путем предотвращения распознавания сплайсосомой 5'ss, 3'ss или элементов последовательности BPS [81]. Простейшим механизмом регуляции АС локальной структурой является блокирование сплайс-сайтов (рис. 1.1А) [82]. Например, в пре-мРНК гена *TAU* человека локальная структура маскирует 5'ss экзона 10, что не позволяет ему включаться в зрелый транскрипт [83]. Образование шпильки вблизи 5'ss сплайсинга может мешать взаимодействию пре-мРНК со сплайсосомой, как в случае экзона 7 гена *SMN2*, где такая шпилька мешает связыванию 5'-ss с U1 мяРНП, что приводит к снижению уровня включения экзона [1].

Пре-мРНК гена фибронектина (*FN1*) является ярким примером влияния структуры шпильки на функцию энхансера сплайсинга (рис. 1.1Б). Один из экзонов гена *FN1*, называемый экзоном EDA, сильно структурирован и образует 7 шпилек. Энхансер локализован в терминальной петле шпильки V и распознается транс-действующими факторами, например, SRSF1. Изменение локализации энхансера с петли на стебель приводит к снижению его регулирующей способности [84]. Сходный механизм регуляции АС с участием интронного сайленсера сплайсинга наблюдается в пре-мРНК вируса иммунодефицита человека (рис. 1.1В) [85].

Неканоническим типом структуры, влияющей на протекание АС, является G-квадруплекс (GQ). В G-квадруплексе четыре гуанозина взаимодействуют

друг с другом через имидазольные связи, а их стеки образуют четырехцепочечную спираль [86]. GQ действуют как цис-элементы в регуляции АС, обычно располагаются в интронных областях и способствуют включению экзонов. Так, например, нарушение способности образовывать GQ существенно уменьшает включение экзона 8 в гене *CD44* [87]. Некоторые регуляторы сплайсинга, например SRSF1, SRSF9, hnRNP H, hnRNP F, hnRNP U и U2AF65, могут взаимодействовать с GQ [88—90]. Формирование GQ в пре-мРНК гена *TP53* в интроне 3 регулирует сплайсинг интрона 2, что приводит к изменению соотношения активных и неактивных изоформ [91], причем удержание интрона приводит к появлению неактивной формы белка [92].

Локальные структуры в пре-мРНК также могут являться мишенями малых молекул. Например, в результате АС транскрипта гена обратной транскриптазы теломеразы человека (*hTERT*) образуются 22 изоформы, из которых только полноразмерная мРНК транслируется в активный белок с обратной транскриптазной активностью [88]. Использование стабилизатора GQ приводит к снижению уровня активной теломеразы за счет исключения экзонов 7 и 8. Это приводит к синтезу укороченного неактивного белка, называемого hTERT-β. Важным классом локальных структур РНК, которые являются мишенями малых молекул у эукариот и влияют на АС, являются рибопереключатели [93].

Дальние взаимодействия в пре-мРНК наиболее хорошо изучены у вирусов, таких как вирус табачной мозаики [94], вирус иммунодефицита человека [95—97], вирусы гепатита B и C [98; 99], вирус денге [100] и др. В настоящее время появляется все больше и больше данных о наличии дальних взаимодействий в пре-мРНК человека и их влиянии на АС [21; 23; 101—104], а в некоторых случаях и на трансляцию [105; 106].

Дальние взаимодействия могут регулировать АС с помощью различных механизмов. Во-первых, как и локальные РНК-структуры, они могут блокировать цис-регуляторные элементы [107]. Во-вторых, дальние взаимодействия могут действовать как «РНК-мосты», сближающие цис-регуляторные элементы [7]. В-третьих, дальние взаимодействия могут также отдалять цис-регуляторные элементы друг от друга. Так, дальние взаимодействия между соседними интронами могут приводить к «выпетливанию» промежуточного экзона или группы экзонов и способствовать их пропуску. Пример дальних взаимодействий в генах *CG33298* и *Gug* дрозофилы, которые функционируют как

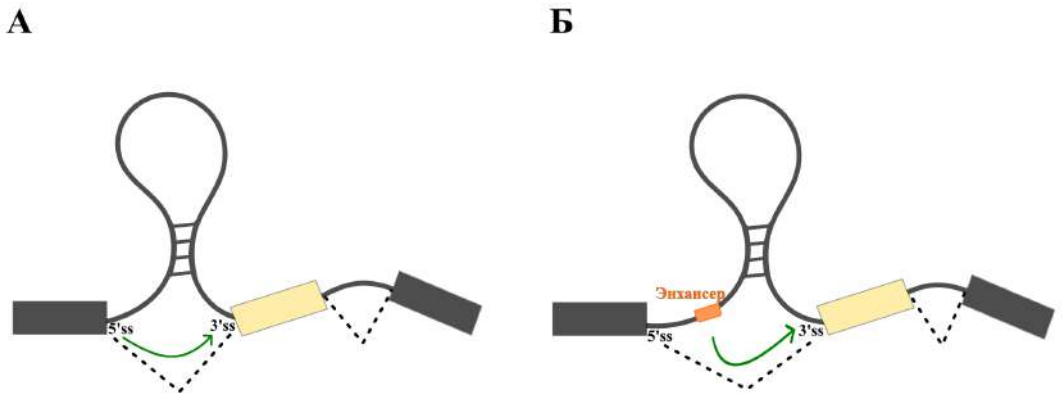


Рисунок 1.2 — Сближение цис-регуляторных элементов сплайсинга структурой РНК (РНК-мосты). (А) Сближение сайтов сплайсинга; (Б) Приближение энхансера сплайсинга к сайту сплайсинга.

РНК-мосты и одновременно блокируют сайты сплайсинга [107], показывает, что эти три механизма не исключают друг друга.

РНК-мосты могут сближать в пространстве цис-регуляторные элементы без участия вспомогательных белков (рис. 1.2А). Например, дальние взаимодействия в пре-мРНК гена *SF1* млекопитающих сближают сильный 5'ss экзона 9 и слабый 3'ss экзона 10, а разрушение образуемой ими вторичной структуры приводит к активации более сильного 3'ss, расположенного на расстоянии 21 нт в направлении 3' конца гена [108]. РНК-мосты могут также приближать интронные цис-регуляторные элементы к сайтам сплайсинга (рис. 1.2Б). Для успешной сборки сплайсосомы и протекания сплайсинга пре-мРНК гена *ENAH* необходимо, чтобы сайт связывания фактора RBFOX2 был сближен в пространстве с альтернативным экзоном, что достигается за счет взаимодействия удаленных участков пре-мРНК, образующих РНК-мост, длина которого превышает 10000 п.о. [7]. В настоящее время описано множество случаев, когда цис-регуляторные элементы находятся на значительном расстоянии от регулируемого экзона, как например у гена *14-3-3ζ* дрозофилы [109], а также генов *ENAH* и *KIF21A* человека [7]. Полногеномные карты РНК-белковых взаимодействий показывают, что бóльшая часть сайтов связывания РСБ расположена намного дальше от потенциальных экзонов-мишеней, чем 1000 нт [110].

Выпетливание части пре-мРНК вторичной структурой, с одной стороны, сближает окружающие цис-регуляторные элементы, а с другой помещает внутреннюю ее часть в петлю, что, как считается, способствует исключению выпетливаемого участка (рис. 1.3А) [111]. Например, при взаимодействиях между

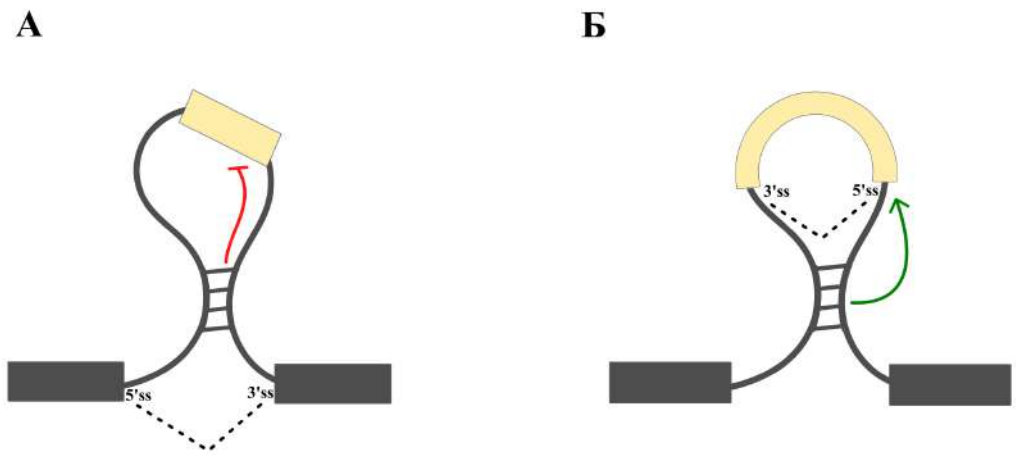


Рисунок 1.3 — Отдаление цис-регуляторных элементов сплайсинга структурой РНК (выпетливания). (А) Выпетливание участка, содержащего один или несколько экзонов и интронов. (Б) Обратный сплайсинг в интроне, приводящий к образованию кольцевой РНК. Красными и зелеными линиями обозначено активирующее и ингибирующее действие на сплайсинг, соответственно.

комплементарными основаниями в интронах, фланкирующих альтернативный экзон, увеличивается частота пропуска такого экзона [112]. Вторичная структура в гене *Nmnat* дрозофилы, выпетливает примерно 350 нт и приводит к исключению экзона 5 и сигнала полиаденилирования (поли(А)) из пре-мРНК. В этом случае структура приближает дистальный акцепторный сайт сплайсинга к донорному сайту и тем самым способствует вырезанию внутреннего терминального экзона [107]. Выпетливания экзонов характерны и для дальних взаимодействий в других генах млекопитающих, например в *CASK* и *PHF20L1* [24], гене дистонина (*DST*), в котором комплементарные участки предположительно выпетливают кластер из шести экзонов [113], а также гене теломеразы человека (*hTERT*), в котором дальние взаимодействия между тандемными повторами приводят к исключению двух экзонов [114]. Спинальная мышечная атрофия — наследственное заболевание, приводящее к ранней смерти в младенческом возрасте, вызывается пропуском экзона 7 гена моторного нейрона выживания 2 (*SMN2*), возникающем в результате разрушения интронной структуры РНК, образованной дальними взаимодействиями [115—117]. Пример вторичной структуры в пре-мРНК ответственного за Х-сцепленную лейкодистрофию Пелизеуса–Мерцбахера протеолипидного белка 1 (*PLP1*), две альтернативные сплайс-изоформы которого различаются выбором альтернативного 5’ss в интроне между экзонами 3 и 4, показывает, что выпетливания не

только экзонов, но и отдельных сплайс-сайтов, оказывают значительное влияние на сплайсинг [118].

Однако самым известным примером влияния дальних взаимодействий на АС является ген *Dscam1* дрозофилы, в транскриптах которого комплементарные спаривания могут происходить на расстоянии до 12000 п.о. Особенностью механизма сплайсинга *Dscam1* является то, что комплементарные участки образуют комплекс конкурирующих структур РНК, которые управляют взаимоисключающим выбором экзонов [4; 119]. Расположенный перед кластером экзонов 6 докерный сайт может спариваться только с одним из нескольких селекторных сайтов, находящихся перед каждым из альтернативных экзонов, тем самым не только сближая удаленные друг от друга 5'ss и 3'ss, но и выпетливая промежуточные экзоны. Взаимоисключающий механизм АС дополнительно контролируется фактором hrp36, который подавляет эктопическое включение альтернативных экзонов под действием SR белков [120]. Аналогичный механизм был обнаружен во многих других генах, содержащих кластеры взаимоисключающих экзонов (см обзор в [121]), например, *14-3-3ζ* [122], *Mhc* [109], *srp*, *RIC-3*, *MRP1* [5], *DNM1* [123], *TCF3*, *CD55* [124] и *ATE1* [79]. Было высказано предположение о том, что тандемные дупликации, в результате которых образуются кластеры взаимоисключающих экзонов, неизбежно приводят к образованию конкурирующих структур РНК и, вследствие этого, ко взаимоисключающему типу АС [125].

Выпетливание части пре-мРНК само по себе не предотвращает ее связывания с компонентами сплайсосомы, а, наоборот, может способствовать протеканию сплайсинга. Как показывает пример кольцевых РНК, комплементарные взаимодействия в интронах, в частности с участием Alu-повторов, могут способствовать протеканию так называемого обратного сплайсинга (back-splicing), ковалентно связывающего 5'- и 3'-концы РНК с образованием кольцевых транскриптов (рис. 1.3Б) [126; 127]. Из сказанного можно заключить, что блокировка, сближение и отдаление цис-регуляторных элементов являются частными случаями общего молекулярного механизма, в котором направление сплайсинга регулируется конформацией транскрипта, зависящей, в свою очередь, от дальних взаимодействий в его структуре.

1.1.3 Совместная регуляция АС структурой пре-мРНК и РСБ

Пре-мРНК образует локальную структуру котранскрипционно, одновременно со сворачиванием вступая во взаимодействие с РСБ [128]. РСБ содержат четко определенные РНК-связывающие домены (РСД), такие как РНК-распознающий домен (RRM), hnRNP K гомологичный домен (KH), цинковые пальцы (ZF) и др., которые взаимодействуют с определенными последовательностями и/или структурами в РНК [129]. Большинство РСД распознают очень короткие (3–7 нт) и вырожденные мотивы, которые часто организованы в кластеры, что позволяет увеличить специфичность связывания РСБ, имеющих несколько РСД, а также позволяет нескольким РСБ кооперировать между собой [46]. Например, высокоаффинное связывание нейрон-специфического фактора сплайсинга NOVA определяется мотивом YCAY (Y = C/U), который обычно находится в кластерах из нескольких тетрамеров [130]. Некоторые РСБ распознают разнесенные в пространстве двудольные мотивы, имеющие определенный структурный контекст [131]. Тем не менее, РСБ, узнающие схожие мотивы, могут иметь различные профили связывания, и даже высокоаффинные взаимодействия могут оказаться нефункциональными [132].

Множество данных указывает на то, что важнейшим фактором, влияющим на связывание РСБ, является структура РНК [133]. Сайты связывания РСБ могут входить в состав различных структурных элементов пре-мРНК [134]. То, что более двадцати ZF-домен-содержащих белков избирательно связывают высокоструктурированные двухцепочечные прекурсоры микроРНК, указывает связывание домена ZF с РНК-дуплексами [129]. РСБ с доменами KH, как правило, предпочитают большие петли шпилек. Учитывая, что большинство таких РСБ содержат несколько РСД, большие петли РНК-шпильки позволяют связывать сразу несколько доменов KH, как это происходит в случае с NOVA1 и PCBP2 [130; 135—137]. Можно предположить, что результат АС должен зависеть от равновесия между РНК-РНК и РНК-белковыми взаимодействиями, причем конкуренция между ними зависит от репертуара РСБ, которые экспрессируются в данном типе клеток [132]. Кроме того, сами РСБ часто функционируют комбинаторно, связываясь с сайтами и структурными элементами на общих мишенях мРНК [138].

Изменения в структуре РНК и вызванные ими изменения АС могут возникать за счет взаимодействия с другими нуклеиновыми кислотами, например с микроРНК [139], а также в результате посттранскрипционных модификаций последовательности пре-мРНК [140]. Так, например, A→I редактирование с помощью белков ADAR регулирует протекание АС за счет изменения последовательности основных цис-элементов [141—143]. Кроме того, ADAR2 может связываться с двухцепочечной РНК, образованной GA-богатой последовательностью и полипиримидиновым трактом, тем самым предотвращая рекрутирование U2AF65 [144]. Метилированный N6-аденозин (m^6A) и связанные с ним белки также могут регулировать АС [140; 145]. Например, модификация m^6A может способствовать связыванию hnRNP C за счет изменения структуры РНК-мишени и обнажения одноцепочечного сайта сплайсинга. Такой же механизм характерен для hnRNP G [146].

Структура РНК может затруднять распознавание цис-регуляторных элементов сплайсинга и сайтов связывания РСБ, однако это не единственный способ, которым она может оказывать влияние на АС. Так, для сплайсинга экзона 5 гена человеческого сердечного тропонина T (*cTNT*) требуется связывание белка MBNL1 на 3'-конце предшествующего интрона. MBNL1 связывает часть интрона в форме шпильки, тогда как фактор сплайсинга U2AF65 связывает ту же область в одноцепочечном состоянии. Стабилизация локальной структуры в форме шпильки блокирует связывание U2AF65, что не позволяет рекрутировать U2 мяРНП, и экзон пропускается [147]. Еще одним ярким примером является связывание hnRNP F с пре-мРНК, содержащей G-квадруплексы, которое стимулирует включение кассетного экзона в гене *CD44*. Интересно отметить, что другой регулятор, ESRP1, также стимулирует включение альтернативного экзона *CD44* независимо от hnRNP F, связываясь с GU-богатым мотивом, частично перекрывающимся с GQ. Это позволяет предположить, что пре-мРНК *CD44* находится в равновесии линейной формы и формы GQ, что помогает поддерживать правильное соотношение сплайс-изоформ [87].

Регуляция АС может происходить за счет РСБ-зависимой стабилизации или ослабления вторичной структуры РНК [148]. Например, было показано, что белки ZFR и ILF3 образуют гетеродимерные дуплексы с ILF2. Получившиеся комплексы неспецифически связываются с двухцепочечными участками в пре-мРНК, влияя на доступность сайтов сплайсинга и связывание транс-действующих факторов. Взаимодействие ILF3 и ZFR со структурой РНК влияет

на взаимоисключающий выбор экзонов гена *ATE1*. Было высказано предположение о том, что ZFR и ILF3 участвуют в стабилизации дуплексов РНК во время взаимоисключающего сплайсинга, хотя точный механизм их действия пока остается неизвестным.

Некоторые РСБ регулируют АС путем изменения третичной и четвертичной структуры пре-мРНК. В отличии от РНК-мостов, в этом случае именно белок-белковые, а не комплементарные взаимодействия обеспечивают необходимую для АС конформацию пре-мРНК. Например, гомодимеры белка hnRNP A1, взаимодействуя с расположенными в соседних интронах сайтами, сближают их и выпетливают экзон, приводя к его пропуску [111]. Подобный механизм также характерен для белков hnRNP F/H [149]. Также было показано, что hnRNP A1 и hnRNP H могут взаимодействовать друг с другом и с другими белками семейства hnRNP [150]. Сближением далеких участков пре-мРНК объясняют и влияние белка NOVA на сплайсинг, поскольку его сайты связывания часто располагаются в начале интрона и вблизи BPS. Это позволяет предположить, что NOVA связывается с двумя сайтами на концах интрона и образует петлю, сближая 5'ss и BPS [151].

Таким образом, гомотипические и гетеротипические взаимодействия между РСБ, которые сближают удаленные друг от друга участки пре-мРНК, являются широко распространенным механизмом регуляции АС.

1.1.4 Непродуктивный сплайсинг

Физиологические уровни экспрессии генов эукариот контролируется большим числом факторов, которые поддерживают баланс между синтезом и деградацией мРНК [152; 153]. Появление нонсенс-мутаций и сдвигающих рамку считывания ошибок сплайсинга приводит к возникновению изоформ мРНК, содержащих преждевременные стоп-кодоны (premature termination codons, PTC). У эукариот существует система селективной деградации таких транскриптов, называемая нонсенс-опосредованным распадом (nonsense-mediated decay, NMD) [154].

То, каким образом система NMD распознает PTC и отличает их от нормальных, долгое время оставалось неизвестным [155]. Современная модель

предполагает, что распознавание PTC происходит в цитоплазме с участием связанных с экзон-экзонными соединениями (ЭЭС) белков, которые депонируются на пре-мРНК в процессе сплайсинга [156; 157]. Во время первого раунда трансляции белки ЭЭС, находящиеся внутри рамки считывания, вытесняются с пре-мРНК рибосомой (рис. 1.4A) [158—160]. Поскольку нормальный сайт терминации трансляции обычно находится в последнем экзоне [161], оставшиеся связанными с пре-мРНК белки ЭЭС, находящиеся за пределами рамки считывания, служат сигналом о том, что появился PTC (рис. 1.4B). Наличие ЭЭС в 50–55 или более нуклеотидах в направлении 3'-конца от стоп кодона запускает каскад деградации транскрипта, центральную роль в котором играет белок UPF1, фосфорилированная форма которого привлекает эндонуклеазу SMG6 и другие факторы, вызывающие деаденилирование и удаление 5'-кэпа у пре-мРНК, что, в свою очередь, приводит к деградации транскрипта клеточными экзонуклеазами [160; 162—164]. Есть и другие модели, в которых преждевременность стоп-кодона определяется расстоянием от него до поли(A)-хвоста, а также модели, в которых PTC вызывает деградацию мРНК независимо от белков ЭЭС [165—167]. Существование ЭЭС-независимого механизма NMD объясняет наличие большого количества мишеней NMD в дрожжах, несмотря на почти полное отсутствие у них сплайсинга [168; 169].

Ранее считалось, что основная функция NMD состоит в предотвращении трансляции усеченных, и поэтому вредоносных белков [170]. Однако в последнее время появляется все больше свидетельств тому, что NMD повсеместно используется для регуляции уровня экспрессии генов [171; 172]. Например, многие РСБ используют NMD для подавления собственной экспрессии через петлю отрицательной обратной связи, при которой белковый продукт гена связывается с кодирующей его мРНК и индуцирует в ней АС, приводящий к появлению PTC [173; 174]. Аналогичным образом может происходить кросс-регуляция, причем в большинстве известных случаев факторы сплайсинга регулируют таким способом экспрессию друг друга [175; 176]. Механизм, при котором альтернативный сплайсинг и NMD пост-транскрипционно регулируют уровни экспрессии генов встречается у всех известных эукариот и часто является эволюционно консервативным [175; 177]. В литературе он называется регулируемым непродуктивным сплайсингом (regulated unproductive splicing and translation, RUST) или просто непродуктивным сплайсингом [171; 178].

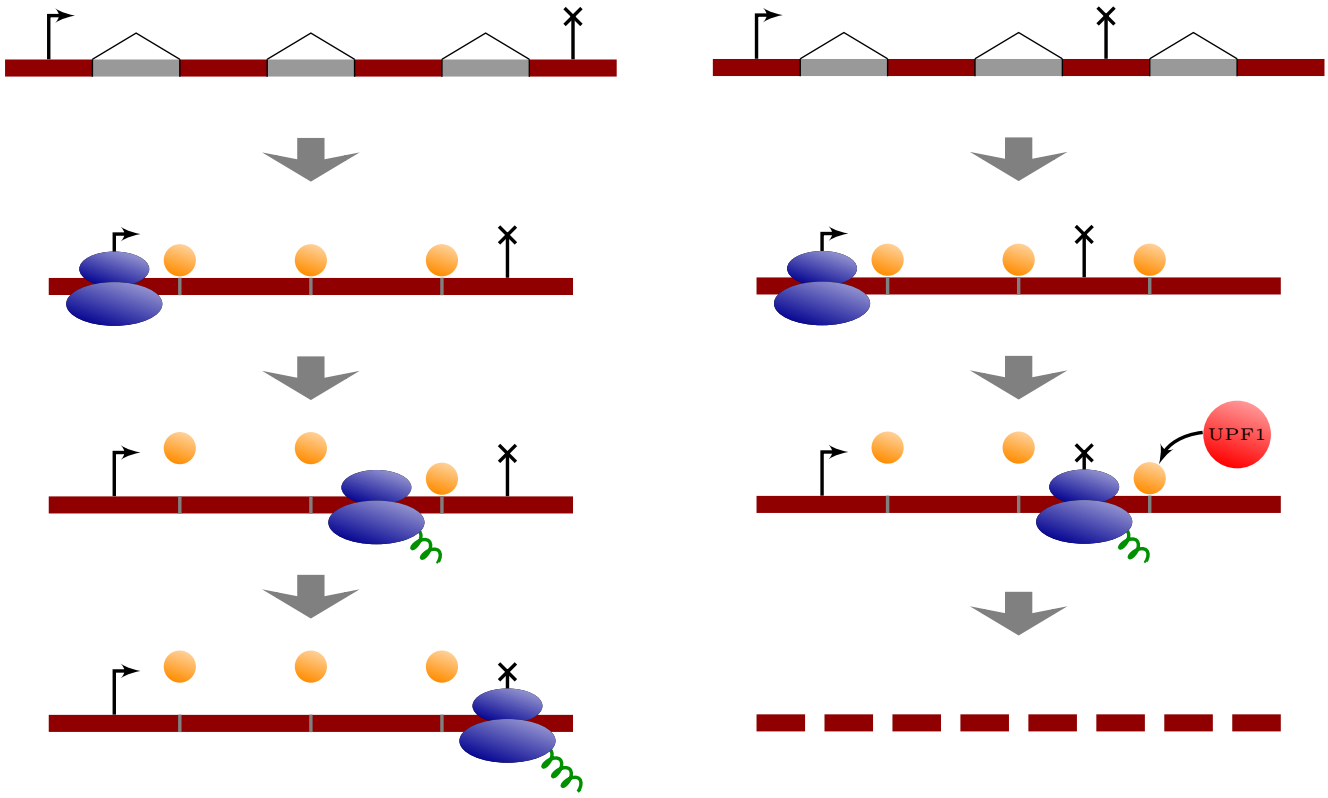


Рисунок 1.4 — ЭЭС-зависимый механизм нонсенс-опосредованного распада. **(А)** Комплексы ЭЭС (оранжевые круги) вытесняются с мРНК рибосомой во время первого раунда трансляции. **(Б)** Оставшиеся связанными с мРНК комплексы ЭЭС за пределами рамки считывания служат сигналом о том, что появился РТС.

1.1.5 Виды непродуктивного сплайсинга

АС может приводить к появлению РТС в транскрипте несколькими способами. Наиболее изучены так называемые ядовитые (poison) экзоны, которые в кодирующей изоформе пропускаются, а при включении в транскрипт приводят к образованию РТС (рис. 1.5А) [178—180]. Ядовитые экзоны могут содержать стоп-кодон как в составе самого экзона, так и индуцировать РТС в экзонах, расположенных за ними в направлении 3'-конца, за счет сдвига рамки считывания (рис. 1.5В). Обратным является случай так называемого необходимого (essential) экзона, который в кодирующей изоформе включается и вызывает появление РТС при пропуске (рис. 1.5С) [173]. Следует отметить, что необходимые экзоны обычно имеют длину, не кратную трем, и вызывают сдвиг рамки считывания, приводящий к образованию РТС в следующих за ними экзонах.

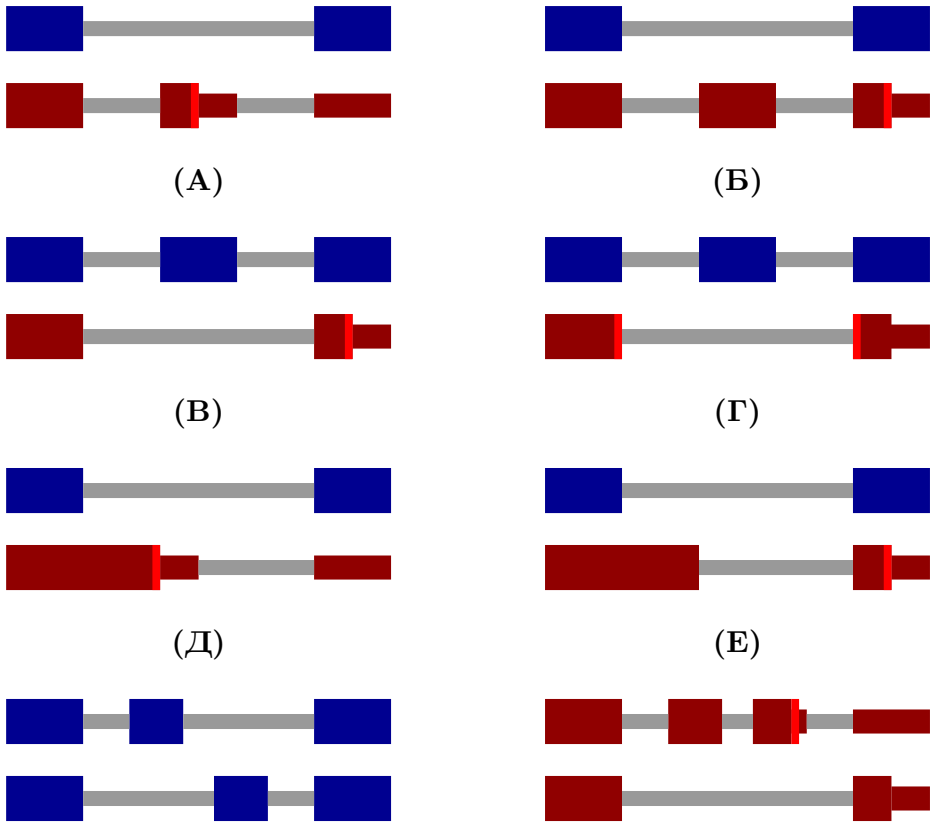


Рисунок 1.5 — Классификация событий непродуктивного сплайсинга. Белок-кодирующие изоформы обозначены синим цветом. Непродуктивные изоформы обозначены красным цветом. РТС обозначены вертикальными красными линиями. **(А)** Ядовитый экзон с РТС. **(Б)** Ядовитый экзон, вызывающий РТС посредством сдвига рамки. **(В)** Необходимый экзон, вызывающий РТС посредством сдвига рамки. **(Г)** Необходимый экзон, вызывающий РТС на ЭЭС. **(Д)** Альтернативный 5'ss, вызывающий РТС за счет удержания интрона. **(Е)** Альтернативный 5'ss, вызывающий РТС посредством сдвига рамки. **(Ж)** Пара взаимоисключающих экзонов.

Однако длина некоторых необходимых экзонов кратна трем, а РТС образуется на ЭЭС, возникающем на месте их пропуска (рис. 1.5D). Активация альтернативных 5'ss и 3'ss также может приводить образованию РТС как за счет сдвига рамки считывания, так и за счет образования новых ЭЭС (рис. 1.5E,F). Пары взаимоисключающих экзонов могут приводить к сдвигу рамки считывания если оба экзона одновременно включаются, или оба одновременно пропускаются (рис. 1.5G). Таким образом, РТС может возникать в результате включения стоп-кодона в транскрипт как на месте самого события АС, так и в следующих за ним в направлении 3'-конца экзонах.

Отдельный интерес представляют события сплайсинга в 3'-нетранслируемых областях (3'-НТО) генов. Стоп кодон, предшествующий 3'-НТО не является преждевременным, но если на расстоянии 50 нт или более от

него в направлении 3'-конца имеется интрон, то вырезание такого интрона автоматически создает NMD-мишень. Например, экспрессия AU-богатого РНК-связывающего фактора AUF1 регулируется консервативными альтернативно сплайсируемыми элементами в 3'-НТО [181]. Известно, что 3'-НТО транскриптов, экспрессия которых повышается при инактивации системы NMD, имеют в среднем бóльшую медианную длину и обогащены интронами [182]. При этом большинство мРНК, кодирующих факторы NMD, сами имеют длинные 3'-НТО и являются мишенями NMD, что указывает на ауторегуляцию их экспрессии [182; 183]. Активность сплайсинга в 3'-НТО широко распространена в онкогенах, значительно повышена в опухолях и коррелирует с плохим прогнозом [184; 185].

Следует отметить, что курируемые вручную или полученные в результате автоматической аннотации базы данных транскриптов содержат далеко не все мРНК, подверженный действию NMD [186; 187]. Неполнота существующей аннотации NMD-транскриптов объясняется тем, что уровень их экспрессии очень мал, и поэтому они не попадают в базы данных. Существует экспериментальный подход для идентификации низко экспрессируемых NMD-транскриптов, который основан на секвенировании обогащенной комплексами ЭЭС фракции РНК, которая содержит частично сплайсированную, но еще не транслированную РНК [188]. С помощью этого метода было обнаружено большое число ранее неизвестных консервативных ЭЭС, причем 70% экзонов, которые поддерживаются этими данными, не кратны трем, а среди оставшихся многие содержат стоп-кодоны [188].

1.1.6 Ауто- и кросс-регуляторный непродуктивный сплайсинг

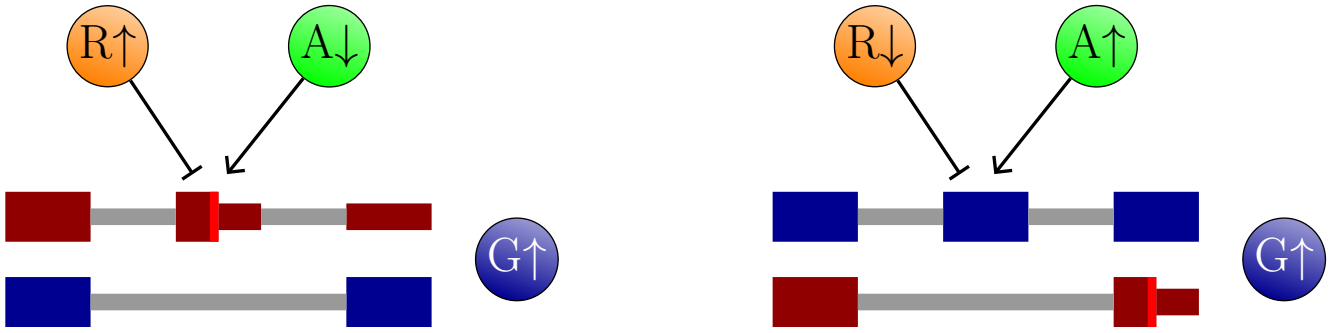
Стимулом к запуску ауторегуляторного непродуктивного сплайсинга обычно является накопление белкового продукта гена. Например, белок RBM10 связывается с собственной пре-мРНК и индуцирует пропуск двух необходимых экзонов, что приводит к смещению баланса сплайс-изоформ в сторону образования мишеней NMD, и уровень экспрессии RBM10 снижается [189]. По этому принципу регулируется экспрессия многих генов, задействованных в сплайсин-

ге, например SR белков [190—193], генов CLK [194; 195], TIAL1 [196], PTB [197; 198], hnRNPD [199], а также некоторых рибосомальных белков [200; 201].

При кросс-регуляторном непродуктивном сплайсинге один белок связывается с пре-мРНК другого и способствует образованию или подавлению NMD изоформ. Такая форма регуляции также распространена среди РСБ из семейства SR [202]. Например, белок SRSF3 наряду с ауторегуляторным включением ядовитых экзонов в собственную пре-мРНК вызывает включение ядовитых экзонов в транскрипты своих паралогов SRSF2, SRSF5 и SRSF7 [203]. Помимо SR-белков таким же образом регулируются и другие пары паралогов, такие как PTBP1/PTBP2 [204], RBM10/RBM5 [189], RBFOX2/RBFOX3 [205], hnRNPD/hnRNPDL [199] и hnRNPL/hnRNPLL [206]. Вообще, кросс-регуляция между паралогами — это весьма частое явление для многих РСБ, которое характеризуется быстрой эволюционной динамикой, в частности быстрым возникновением и исчезновением ядовитых экзонов [175].

Кросс-регуляторный непродуктивный сплайсинг имеет важное значение не только для РСБ. Например, он обуславливает тканеспецифическую экспрессию гена *MID1*, кодирующего ассоциированную с микротрубочками убиквитин-лигазу, дисфункция которой приводит к патологиям эмбрионального развития [176; 207]. Непродуктивный сплайсинг важен для многих физиологических процессов, таких как эмбриональное развитие [208], клеточная дифференцировка [209], ответ на стресс [210—212], патогенез нейродегенеративных заболеваний [213; 214] и др.

В регуляции непродуктивного сплайсинга могут участвовать как активаторы, так и репрессоры сплайсинга. Увеличение концентрации репрессора или понижение концентрации активатора включения ядовитого экзона приводят к его пропуску, вследствие чего уровень экспрессии гена-мишени увеличивается (рис. 1.6А). Аналогично, уменьшение концентрации репрессора или увеличение концентрации активатора включения необходимого экзона подавляют его пропуск, что также приводит к увеличению уровня экспрессии гена-мишени (рис. 1.6В). Следует отметить, что некоторые РСБ могут быть как активаторами, так и репрессорами сплайсинга, а выбор между активацией и репрессией зависит от положения их сайта связывания на мРНК [215]. Как будет показано в этой диссертационной работе, PTBP1 стимулирует включение ядовитого экзона в гене *DCLK2*, что приводит к повышению уровня его экспрессии в нейрональных тканях, в которых экспрессия PTBP1 понижена [216]. В то же



(А) **(Б)**

Рисунок 1.6 — Механизмы регуляции непродуктивного сплайсинга. «R» обозначает репрессор сплайсинга. «A» обозначает активатор сплайсинга. «G» обозначает ген-мишень. Цвета экзонов такие же, как на рис. 1.5. **(А)** Увеличение (↑) концентрации R или уменьшение (↓) концентрации A приводит к пропуску ядовитого экзона, и экспрессия G увеличивается. **(Б)** Уменьшение концентрации R или увеличение концентрации A подавляет пропуск необходимого экзона, и экспрессия G также увеличивается.

самое время PTBP1 подавляет включение ядовитого экзона в гене *IQGAP1*, вследствие чего уровень его экспрессии в мозге снижается.

Многие мишени непродуктивного сплайсинга сами являются РСБ и регулируют уровни включения экзонов в других РСБ, что создает множественные регуляторные петли как с положительными, так и с отрицательными обратными связями. Отрицательные обратные связи обеспечивают механизмы ауторегуляции для поддержания гомеостаза, а положительные обратные связи могут создавать бистабильные системы для включения экспрессии [217]. Например, ген *Sxl* дрозофилы использует оба этих механизма, что приводит к его аутоиндукции при малых концентрациях и, одновременно, предотвращает вредоносное перепроизводство белка [218]. Для достижения такой регуляции РСБ могут действовать одновременно как активаторы и репрессоры сплайсинга, связываясь сразу с несколькими сайтами на пре-мРНК, чем, вероятно, и объясняется высокая эволюционная консервативность нуклеотидных последовательностей вокруг событий непродуктивного сплайсинга [219].

1.1.7 Биологические функции АС

Долгое время было принято считать, что основной функцией АС является расширение спектра экспрессируемых белков [220]. Различные сплайс-изоформы потенциально кодируют различные изоформы белка, которые могут терять или приобретать определенные домены и, следовательно, различаться по своим функциональным свойствам. По оценкам на данный момент в человеческом геноме имеется около 20000 белок-кодирующих генов, но благодаря АС из них получается более 100000 различных белков [38; 221]. Согласно современным транскриптомным исследованиям, репертуар производимых белок-кодирующими генами транскриптов намного шире, чем это считалось ранее, и что до 95% человеческих генов производят более одной зрелой изоформы РНК [36; 38]. Однако исследования протеома с помощью методов масс-спектрометрии показали, что белковое разнообразие намного меньше того, которое ожидалось бы при трансляции всех продуктов АС. Это несоответствие вызвало продолжающиеся и в настоящее время дебаты о том, является ли наблюдаемое разнообразие альтернативного сплайсинга биологически функциональным, или же его причина кроется в стохастической природе сплайсинга [222; 223]. Ясного ответа на этот вопрос пока нет [224].

Однако хорошо известно то, что АС играет важнейшую роль в дифференцировке тканей [37]. Более двух третей событий АС имеют тканеспецифичные различия, причем от 47% до 65% альтернативных событий являются тканеспецифичными в зависимости от типа события сплайсинга, а мажорная сплайс-изоформа варьируется в зависимости от условий более чем в 60% кодирующих генов [225]. При этом далеко не все тканеспецифичные события АС эволюционно консервативны, что наводит на предположение о том, что различное использование экзонов лежит в основе тканеспецифической «перестройки» сетей межбелковых взаимодействий, которая необходима для возникновения морфологических различий между видами [226—228]. У приматов АС устроен очень сложно, особенно в тканях мозга, а в целом события АС эволюционировали таким образом, что они большей мере определяют принадлежность к виду, чем к органу [226; 229]. Другими словами, профиль АС печени человека больше похож на профили АС других органов человека, чем на профили АС печени шимпанзе.

Ген *CALCA* является интересным примером того, как кардинально АС может изменить последовательность кодируемого белка [230; 231]. Этот ген, экспрессирующийся в нейронах гипоталамуса и С-клетках щитовидной железы, кодирует два белка – кальцитонин и пептид, родственный гену кальцитонина (calcitonin gene-related peptide 1, CGRP1). Сочетание АС и альтернативного полиаденилирования приводит к образованию мРНК CGRP1 в нейронах и мРНК кальцитонина в С-клетках щитовидной железы, причем кодируемые ими полипептиды подвергаются клеточно-специфичным протеолитическим процессам, в результате чего образуются два зрелых белка, состоящие из совершенно разных аминокислот. Еще одним интересным примером является АС в гене *FOXP1* (forkhead box P1) из большого FOX-семейства транскрипционных факторов, которые распознают специфические последовательности ДНК. Сплайсинг взаимоисключающих экзонов E18 и E18b в этом гене определяет то, будут ли клетки вести себя как плюрипотентные эмбриональные стволовые клетки или запускать программы клеточной дифференцировки [232; 233]. Сплайс-изоформа с экзоном E18 кодирует белок, который преимущественно связывается с консенсусной последовательностью GTAAACA в ДНК и активирует транскрипцию генов дифференцировки, а сплайс-изоформа с экзоном E18b, называемая FOXP1-ES, кодирует белок, связывающийся с совершенно другими сайтами и стимулирующий экспрессию генов плюрипотентности, таких как *OCT4*, *NANOG*, *NR5A2* и *GDF3*, одновременно подавляя транскрипцию генов дифференцировки.

Не углубляясь в описание изменяющих биологические функции событий АС, которые детально обсуждаются во многих обзорах [121; 234; 235], перейдем к обсуждению роли АС в заболеваниях.

1.1.8 Роль АС в заболеваниях человека

Около 15% наследственных и онкологических заболеваний человека связаны с альтернативным сплайсингом [236—239], включая кистозный фиброз [240], спинальную мышечную атрофию [241], фронто-темпоральную деменцию [242], болезнь Хантингтона [243], расстройства аутического спектра [244; 245], и рак [246]. Многие их этих изменений определяют опухолевые фенотипы, та-

кие как увеличение ангиогенеза [247], или избегание апоптоза [248]. Сплайсинг играет важную роль в приобретении терапевтической устойчивости. Так, например у нон-респондеров к BRAF-таргетной терапии рака молочной железы, экспрессируется форма *BRAF*, лишенная экзонов 4–8, которые кодируют RAS связывающий домен [249].

На молекулярном уровне изменения альтернативного сплайсинга могут происходить из-за соматических мутаций, которые разрушают регуляторные цис-элементы в экзонах и интронах [250; 251]. Например, делеция 150 нт, затрагивающая альтернативный 5'ss экзона 11 гена ламина A (*LMNA*), а также мутация в 3'ss экзона 4, добавляющая три аминокислоты в его белковый продукт, приводит к трансляции нефункционального белка, вследствие чего развиваются синдром прогерии Хатчинсона-Гилфорда и дилатационная кардиомиопатия [252; 253]. Лобно-височная деменция и паркинсонизм, связанный с хромосомой 17 (FTDP-17), являются аутосомно-доминантным нейродегенеративным заболеванием, которое вызывается мутацией в экзонном энхансере, из-за которой увеличивается степень включения экзона 10 гена *MAPT* [254]. Мутация, создающая энхансер сплайсинга в гене *ATP6AP2* приводит к увеличению уровня включения экзона 4, что вызывает Х-сцепленный паркинсонизм со спастичностью [255]. Мутации в факторах сплайсинга часто приводят к глобальной дерегуляции всей программы сплайсинга онкогенов [256—259].

Однако наибольшее число заболеваний, связанных со сплайсингом, вызывается сбоями в работе системы NMD и нарушениями непродуктивного сплайсинга. Так, нонсенс-мутации в генах *CFTR* и *hERG* вызывают муковисцидоз и синдром удлиненного QT-интервала, соответственно, в результате деградации их транскриптов системой NMD [260; 261]. Делеции, вызывающие сдвиг рамки считывания, также вызывают дефицит важных белков. Известным примером является мышечная дистрофия Дюшенна, причиной которой являются нарушающие рамку считывания делеции в гене *DMD* [262—264].

Мутации в сплайс-сайтах могут вызывать переключение альтернативного сплайсинга на непродуктивную изоформу. Так происходит в гене *SYNGAP1*, непродуктивный сплайсинг которого регулируется PTBP1/2, что обеспечивает его тканеспецифичную экспрессию. В результате активации альтернативного 3'ss возникает NMD-изоформа и уровень экспрессии гена падает, что приводит к развитию аутизма и умственной отсталости [265; 266].

Однако не только мутации в кодирующей области и сплайс-сайтах способны создавать мишени NMD. Например, мутации в интроне 20 гена *SCN1A* увеличивают степень включения ядовитого экзона, что является причиной синдрома Драве [266; 267]. Мутации в ядовитом экзоне гена SNRPB вызывают церебро-косто-мандибулярный синдром [268]. Предположительно, они создают или разрушают сайт связывания РСБ, активирующего или подавляющего включение ядовитого экзона. Мутация в криптическом ядовитом экзоне гена *PCCA*, вызывающая пропионовую ацидемию, является редким случаем, когда механизм дерегуляции непродуктивного сплайсинга известен [269]. Эта мутация находится в сайте связывания фактора hnRNP A, который в норме подавляет включение ядовитого экзона, но мутация разрушает этот сайт и одновременно создает энхансер сплайсинга, в результате чего экспрессия *PCCA* снижается [269].

Наряду с мутациями вблизи событий непродуктивного сплайсинга к заболеваниям также может приводить неправильная работа РСБ. Точечная мутация в факторе сплайсинга SRSF2, наблюдаемая с высокой частотой у пациентов, страдающих острым миелоидным лейкозом [270; 271], вызывает включение ядовитого экзона в транскрипты метилтрансферазы гистонов *EZH2*, что ведет к снижению уровня ее экспрессии и, как следствие, развитию подавляемых ею миелоидных новообразований [270]. Мутации в факторе сплайсинга *SF3B1*, часто наблюдаемые в миелодиспластических синдромах [272], увеличивают уровень включения ядовитого экзона в транскрипты гена *BRD9*, в результате чего его экспрессия падает, что приводит к ускоренному росту и метастазированию меланом [273]. Метилирование транскриптов *SRSF3*, *SRSF6* и *SRSF11* в результате повышенной экспрессии метилтрансферазы METTL3, часто наблюдаемой в глиобластомах, приводит к пропуску ядовитых экзонов и увеличивает уровни экспрессии этих генов [274]. Примечательно, что подавление экспрессии METTL3 в глиобластомных клеточных линиях приводит к снижению пролиферации и миграции клеток, отчасти за счет изменения сплайсинга мишеней SR-белков, таких как BCL-X и NCOR2 [274].

1.1.9 Модуляция АС антисмысловыми олигонуклеотидами и малыми молекулами

Модуляция АС — многообещающая терапевтическая стратегия для лечения многих заболеваний. Изменять сплайсинг позволяют переключающие сплайсинг антисмысловые олигонуклеотиды (АОН) [275]. Комплементарно связываясь с последовательностью пре-мРНК, АОН блокируют сайты сплайсинга и/или сайты связывания РСБ, тем самым способствуя выбору нужного события альтернативного сплайсинга [275]. В конце 2016 года Управление по санитарному надзору за качеством пищевых продуктов и медикаментов США одобрило препарат Спинраза — антисмысловую нуклеотидную терапию для спинальной мышечной атрофии, разработанную совместно компаниями Biogen и Ionis Pharmaceuticals. Регистрация препарата Спинраза (Нусинерсен) ознаменовала собой начало эры клинического использования антисмысловых нуклеотидов для модуляции сплайсинга РНК и было отмечено Нобелевской премией в области медицины и физиологии [276; 277].

Поскольку основной патологией сплайсинга, которая поддается коррекции с помощью АОН, является переключение на NMD-изоформу, наиболее перспективным представляется применение АОН именно к непродуктивным событиям. АОН для изменения непродуктивного сплайсинга можно разделить на три группы: АОН для увеличения экспрессии полноразмерного белка (индукция пропуска ядовитых экзонов), АОН для сохранения экспрессии укороченного белка, когда экспрессия полноразмерного белка невозможна (индукция пропуска части экзонов или удержания интронов) и АОН для снижения экспрессии (индукция включения ядовитых экзонов).

АОН первой группы могут быть использованы для лечения заболеваний, вызванных дефицитом функционального белка, например из-за мутаций в генах *SYNGAP1*, *SCN1A*, *PCCA* и *SNRNPB* [265—269]. АОН второй группы необходимы, если в результате нонсенс-мутации или делеции со сдвигом рамки считывания возникает РТС. Они позволяют избежать деградации транскрипта и сохранить экспрессию укороченного белкового продукта. Технически АОН второй группы могут индуцировать пропуск экзона с нонсенс-мутацией (как в гене *PCCA*) или удержание интрона в направлении 3'-конца от РТС (как в гене *hERG*). Пропуск кодирующих экзонов может быть полезен в случае делеции

со сдвигом рамки считывания для того, чтобы восстановить рамку (как в гене *DMD*). Ряд действующих таким образом препаратов для лечения мышечной дистрофии Дюшенна уже одобрены [264]. АОН третьей группы можно использовать, если необходимо подавить накопление белка. Например, мутации в гене FUS, разрушающие сигнал его ядерной локализации и вызывающие экспорт в цитоплазму, ассоциированы с боковым амиотрофическим склерозом [278; 279]. Поскольку для подавления экспрессии гена *FUS* через непродуктивный сплайсинг его белковый продукт должен находиться в ядре, экспорт мутантного белка из ядра разрушает петлю ауторегуляции, что усугубляет его накопление в цитоплазме и способствует образованию агрегатов, обладающих цитотоксическим эффектом [280—283].

Несмотря на все положительные стороны АОН, имеется множество трудностей с их доставкой к целевым органам и тканям. Необходимость разработки системы доставки и использование повышенных доз препарата для того, чтобы достичь необходимой концентрации, приводит к росту цены и увеличению риска побочных эффектов [284]. Альтернативой являются малые молекулы — модуляторы сплайсинга, обладающие большей биодоступностью, чем АОН. На сегодняшний день найдено несколько молекул, специфично связывающихся с конкретными РНК [284—288]. Среди них наиболее изучен рисдиплам, который модулирует сплайсинг гена *SMN2* и может быть использован для лечения спинальной мышечной атрофии [287]. Бранаплам, сходный по структуре и механизму действия с рисдипламом, способствует включению криптического ядовитого экзона в гене *HTT*, что снижает его экспрессию и замедляет прогрессирование болезни Хантингтона [288]. Аналогичным эффектом обладает молекула PTC518, которая сейчас находится во второй фазе клинических исследований [286].

1.2 Методы определения и предсказания структуры РНК

1.2.1 Экспериментальные методы определения структуры РНК

Экспериментальное изучение структур РНК имеет долгую историю. Первым методом для определения структуры РНК была рентгеновская кристаллография, однако для нее необходимы кристаллы, которые сложно получить из-за структурной гетерогенности РНК. На практике этот крайне трудоемкий метод позволяет определить структуру только очень коротких молекул. Тем не менее, кристаллографические исследования позволили определить структуру целой рибосомы с атомным разрешением, за что их авторам Венки Рамакришнану, Тому Стейцу и Аде Йонат в 2009 году была присуждена Нобелевская премия по химии [289]. Метод ядерного магнитного резонанса (ЯМР) также подходит в основном для определения структуры небольших РНК (обычно менее 100 нуклеотидов). Прогресс в технологии криоэлектронной микроскопии (КриоЭМ) значительно улучшил разрешение и способность определять структуры макромолекул, включая РНК [290—292], но, несмотря на все эти усилия, в настоящее время в структурных базах данных насчитывается всего около шести тысяч РНК-содержащих структур, что составляет менее 3% от общего числа известных структур [293; 294].

Если не задаваться целью определить структуру РНК с атомным разрешением, а ограничиться только комплементарными взаимодействиями, то в настоящее время существует широкий спектр методов исследования структуры РНК, которые по-разному сочетают ферментативные и химические зонды с глубоким секвенированием для одновременного исследования структуры больших ансамблей молекул РНК («структурома» РНК). В широком смысле эти методы можно разделить на две основные группы в зависимости от типа получаемой структурной информации: методы, основанные на пробинге, и методы, основанные на лигировании пространственно близких молекул.

Общий принцип, лежащий в основе методов первой группы, заключается в использовании зондов для модификации РНК специфичным для структуры РНК способом [295—298]. Эти зонды оставляют «следы» на РНК в виде модифицированного основания, позицию которого можно определить с помощью

обратной транскрипции (ОТ) и последующего секвенирования. Пробинг не дает информации о спаривании оснований, а вместо этого измеряет интенсивность реакции зонда с каждым нуклеотидом и рассчитывает показатель реактивности, отражающий вероятность комплементарного спаривания. Например, локальную доступность неспаренных оснований экспериментально определяет метод SHAPE-seq [299; 300]. Однако методы, основанные на разнице в реакционной способности одноцепочечных и двухцепочечных остатков, не подходят для исследования дальних взаимодействий в структуре пре-мРНК, поскольку с их помощью можно обнаружить только спарен ли нуклеотид, но невозможно понять с каким именно другим нуклеотидом. Если локальные спаривания оснований с высокой степенью достоверности можно предсказать по близлежащим последовательностям, то на дальних расстояниях возникает неопределенность из-за слишком большого числа вариантов спаривания [299].

В отличие от пробинга, методы, основанные на лигировании пространственно близких молекул, дают информацию о том, какие основания спарены друг с другом внутри РНК (внутримолекулярная структура РНК) или между двумя молекулами РНК (межмолекулярные РНК-РНК взаимодействия) [101—104; 301—303]. Как правило, эти методы фотохимически сшивают взаимодействующие РНК, например при помощи производных псоралена, после чего РНК фрагментируются, а затем взаимодействующие пары РНК лигируются с образованием химерных молекул, которые могут быть идентифицированы в ходе секвенирования. Они широко применяются к исследованию не только вторичной структуры одиночных молекул, но также и РНК-РНК взаимодействий, которым в последние годы уделяется все больше внимания в свете растущего интереса к пониманию функций длинных некодирующих РНК [304; 305]. РНК-РНК взаимодействия управляют регуляторными программами процессинга РНК, такими, как подавление трансляции и сайленсинг генов [306; 307]. Они играют фундаментальную роль в функционировании сплайсосомы, где малые ядерные РНК (мяРНК) взаимодействуют друг с другом и с пре-мРНК, образуя гетеродуплексы [31]. Также появляется все больше свидетельств тому, что РНК-РНК взаимодействия участвуют в регуляции транскрипции [308; 309].

1.2.2 Вычислительные методы предсказания структуры РНК

Из экспериментальных данных о молекулах РНК с известной структурой можно извлечь принципы, по которым она формируется, и построить по ним модели для сворачивания молекул с неизвестной структурой. Существующие методы предсказания структуры РНК можно подразделить, с одной стороны, на термодинамические, основанные на минимизации свободной энергии, и филогенетические, основанные на поиске ковариаций. С другой стороны, их также можно подразделить на методы предсказания внутримолекулярных и межмолекулярных РНК-структур.

Методы, основанные на минимизации свободной энергии одиночной молекулы РНК, находят термодинамически наиболее стабильную вторичную структуру путем минимизации свободной энергии с использованием алгоритмов динамического программирования. Первым в этой группе методов был так называемый алгоритм Нуссинов, в котором наиболее стабильной предполагается структура РНК, содержащая максимальное число спаренных оснований [9]. В настоящее время расчет свободной энергии основан на экспериментально определенных параметрах так называемой аддитивной термодинамической модели, в которой участвуют энергии стэкинга, выпячиваний, внутренних и мультипетель, а также других элементов вторичной структуры [11; 310]. Спектр вычислительных методов для предсказания структуры РНК с помощью минимизации свободной энергии охватывает около 40 программ, таких как Mfold [311] и RNAfold [312], которые различаются своей производительностью и предположениями о типах предсказываемых структур. Их обычно применяют для сворачивания небольших молекул или фрагментов РНК (не более 200 нуклеотидов).

Основными ограничениями этих методов являются их возрастающая неточность и сложность вычислений по мере увеличения длины анализируемой РНК, а также неспособность учитывать ключевые детерминанты структуры РНК в живых клетках, такие как котранскрипционная природа сворачивания, связывание белков и модификации РНК. Действительно, РНК в клетке сворачиваются котранскрипционно с помощью белков-шаперонов, а также принимает промежуточные структурные конформации, которые помогают избежать энергетических ловушек [313; 314]. На сегодняшний день эти

динамические взаимодействия плохо изучены, а методы анализа РНК-белковых взаимодействий и кинетики транскрипции только начинают появляться [315; 316]. РНК-связывающие белки и кинетика элонгации вносят большую неопределенность в параметры моделей, которые используются для сворачивания РНК, и представляют собой основной источник расхождений между вычислительными моделями длинных РНК [317; 318].

Однако главной проблемой перечисленных методов является лежащее в их основе динамическое программирование, которое за приемлемое вычислительное время может предсказать только структуры без псевдоузлов, что делает их неприменимыми к предсказанию дальних взаимодействий, которые «шунтируются» локальными структурами [113]. Иными словами, алгоритм минимизации энергии предпочитает «не заметить» высокоэнергетические дальние взаимодействия, которые вследствие запрета на псевдоузлы оказываются несовместимыми с большим числом низкоэнергетических, но суммарно более «выгодных» локальных спариваний.

Таким образом, наша способность вычислительно сворачивать эукариотические РНК без учета псевдоузлов оказывается систематически смещена в сторону локальных структур. В то же время задача предсказания структуры РНК с произвольными псевдоузлами является NP-полной [319], а для предсказания даже наиболее простых типов псевдоузлов динамическому программированию требуется время $\mathcal{O}(n^6)$, где n — длина последовательности [320], что делает его неприменимым к эукариотическим транскриптам. Помимо этого технического ограничения, существует и более фундаментальная проблема, заключающаяся в том, что аддитивная термодинамическая модель недостаточна для описания энтропийного вклада петель в молекулах с псевдоузлами поскольку необходимо принимать во внимание важные стерические и топологические ограничения [321].

Несмотря на то, что дальние взаимодействия в структуре РНК являются внутримолекулярными, с вычислительной точки зрения их также можно рассматривать как межмолекулярные. Задача предсказания РНК-РНК взаимодействий похожа на задачу предсказания вторичной структуры одиночной РНК, но отличие состоит в том, что допускаются пары оснований как внутри одной молекулы, так и между молекулами РНК. Несмотря на то, что внутри- и межмолекулярные взаимодействия управляются одними и теми же молекулярными силами, это отличие имеет решающее значение для алгоритмов,

поскольку динамическое программирование применимо только к незаузленным структурам РНК, в то время как одновременное предсказание внутри- и межмолекулярных пар оснований эквивалентно сворачиванию РНК с псевдоузлами.

Как и в случае одной молекулы, для предсказания РНК-РНК взаимодействий можно ограничиться только простыми видами псевдоузлов. Примечательно, что самой первой работой, в которой задача поиска межмолекулярных РНК-взаимодействий была сформулирована в упрощенном виде и решена на основе динамического программирования, является работа автора этой диссертации [321]. Разработанные на ее базе современные методы предсказания межмолекулярных РНК-структур, как правило, еще более вычислительно затратны, чем внутримолекулярные методы, поскольку они моделируют РНК-РНК взаимодействия как разборку локальной структуры с последующей межмолекулярной гибридизацией [322; 323]. Некоторые их разновидности избегают рассмотрения внутримолекулярных взаимодействий, чтобы быть эффективными в вычислительном отношении [324; 325].

Наилучший на данный момент компромиссный подход использует предварительно рассчитанные профили доступности в дополнение к оценке свободной энергии открытых сайтов связывания [326]. Другие методы получают дополнительную скорость за счет упрощения модели свободной энергии, что делает их применимыми, например, к поиску мишеней микроРНК, но устранение внутренней структуры РНК приводит к резкому увеличению доли ложноположительных предсказаний [327—329]. Вклад псевдоузлов можно оценить, рассматривая отдельные спирали вместо пар оснований, но этот подход необходимо сочетать с филогенетическими методами [330]. Таким образом, в масштабе эукариотических геномов предсказание дальних взаимодействий в структуре РНК оказывается весьма сложной задачей как с точки зрения эффективности, так и с точки зрения специфичности, поскольку количество случайной комплементарности растет с увеличением длины последовательности.

Когда предсказание структуры РНК по последовательности не дает результатов, мощной альтернативой становятся методы сравнительной геномики. Во-первых, они ограничивают пространство поиска до эволюционно консервативных областей, что частично снижает сложность вычислений и повышает специфичность предсказаний [331]. Во-вторых, они приобретают статистическую силу за счет наблюдения за компенсаторными заменами, т.е. ковариациями нуклеотидов в спаренных позициях [332—336]. Их идея основана

на том, что структурно и функционально значимые пары оснований во вторичной структуре РНК должны эволюционировать одновременно для того, чтобы поддерживать комплементарность.

Для исследования компенсаторных замен существуют методы, такие как Dynalign [337] и R-scape [12], которые предсказывают структуру РНК, ограниченную результатами ковариационного анализа выравнивания гомологичных последовательностей РНК. Несмотря на то, что ковариационные модели оказались чрезвычайно успешными в открытии рибопереключателей [338; 339], недостаточный уровень вариабельности нуклеотидных последовательностей далеко не всегда позволяет их использовать [13]. Среди примеров, представленных в данной диссертации, многие функциональные структуры РНК были фактически обнаружены в множественных выравниваниях последовательностей без каких-либо компенсаторных замен.

В применении к эукариотам предсказание вторичной структуры РНК и РНК-РНК взаимодействий методами сравнительной геномики сталкиваются с множеством трудностей из-за большой длины и сложной организации эукариотических генов. Степени консервативности экзонов и интронов значительно различаются, причем интроны не всегда могут быть выровнены или их выравнивание может не быть уникальным. Некоторые методы используют модели филогенетических замен для сворачивания последовательностей, кодирующих белки [335; 336], но о применении сравнительной геномики к сворачиванию интронов и нетранслируемых областей известно гораздо меньше. При этом, основным ограничением остается то, что множественное выравнивание не всегда существует, а даже тогда, когда оно существует и анализируется как профиль с помощью алгоритма для одиночной последовательности [340; 341] или с помощью вероятностной модели [342—344], результаты предсказаний кардинально зависят от качества входного выравнивания.

Логичным решением представляется объединить задачи множественного выравнивания и предсказания структуры РНК, что представляет собой знаменитую проблему одновременного сворачивания и выравнивания, впервые сформулированную в 1984 году Давидом Санковым [345]. Алгоритм Санкова требует огромных вычислительных затрат, а его строгая реализация для двух последовательностей имеет сложность по времени и памяти $\mathcal{O}(n^6)$ и $\mathcal{O}(n^4)$ соответственно, где n — длина последовательности [346]. Его можно использовать для оптимизации существующего множественного выравнивания с учетом

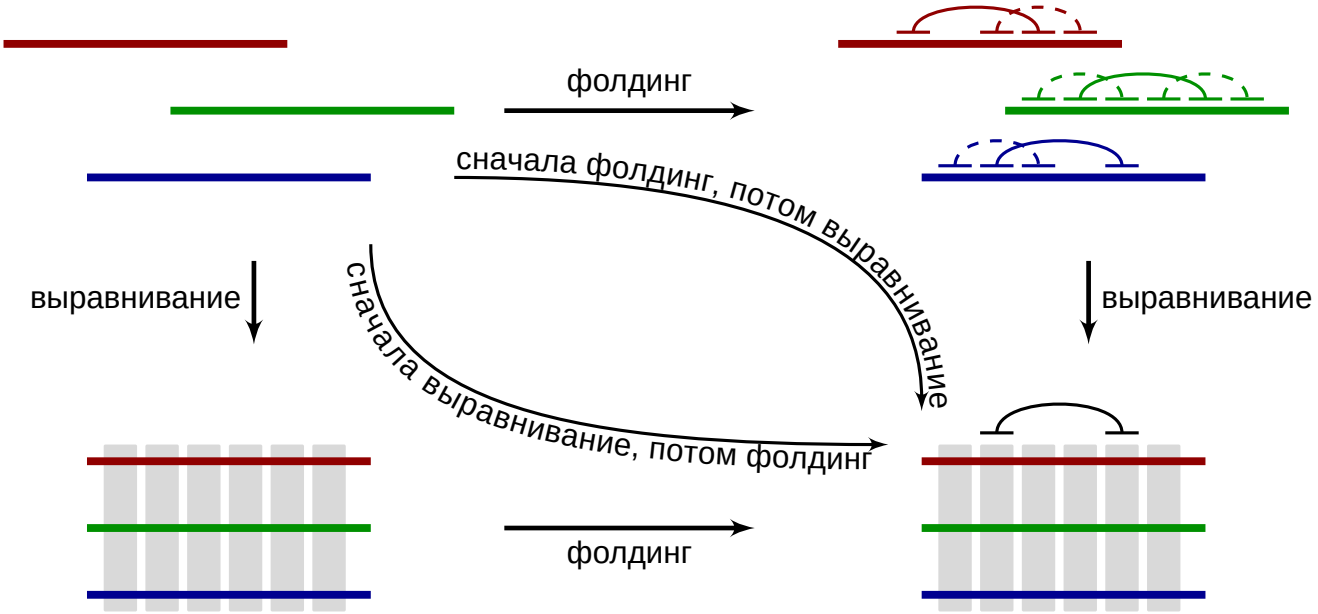


Рисунок 1.7 — Диаграмма, описывающая одновременное выравнивание последовательностей и предсказание структуры. Вверху слева: невыровненные последовательности РНК. Внизу слева: их выравнивание без учета структуры; серым цветом показаны консервативные участки. Вверху справа: предсказанные элементы вторичной структуры для каждой последовательности показаны в виде дуг. Внизу справа: структуры сопоставляются либо в результате согласованного выравнивания, либо идентифицируются непосредственно в множественном выравнивании последовательностей.

структуры, но возможности этой оптимизации весьма ограничены [18]. Применительно к дальним взаимодействиям метод Санкова выходит далеко за рамки современных вычислительных возможностей и, принимая во внимание число неизвестных факторов, влияющих на сворачивание РНК в живых клетках, работа над его усовершенствованием вряд ли имеет смысл.

У алгоритма Санкова есть две предельные реализации, которые можно назвать «сначала выравнивание, потом фолдинг» и «сначала фолдинг, потом выравнивание», которые можно условно представить в виде диаграммы на рис. 1.7. В первом случае, «сначала выравнивание, потом фолдинг», набор ортологичных последовательностей сначала выравнивается, а затем по полученному выравниванию предсказывается структура РНК. Этот подход реализован во всех существующих филогенетических методах [322], однако его чувствительность ограничена качеством входного выравнивания, например, неопределенностью выравнивания взаимоисключающих экзонов, возникающих в результате геномных дупликаций, или ошибками в выравнивании консервативных структурных элементов, которые слишком коротки и теряются в протяженном неконсервативном фоне [119; 347]. Он также неприменим к случаям, когда последовательности разошлись до неузнаваемости, но их структура

осталась неизменной. Тем не менее, «сначала выравнивание, потом фолдинг» представляет собой наиболее простой, быстрый и мощный подход, который используется во многих современных сравнительных методах, в том числе и для предсказания РНК-РНК взаимодействий [332; 348].

Вторая предельная реализация алгоритма Санкова, «сначала фолдинг, потом выравнивание», в которой для каждой последовательности находятся все возможные структуры, по которым затем строится согласованное множественное выравнивание, исследована в значительно меньшей степени. На первый взгляд она представляется нерациональной или даже невозможной, поскольку число структур для одной последовательности, возведенное в степень числа их комбинаций при множественном выравнивании, слишком велико. Первой работой в этом направлении был поиск локальной структуры у энтеровирусов с использованием филогенетического сравнения потенциальных двухцепочечных участков и последующим анализом согласованности графа вторичных структур [349]. В данной диссертационной работе эти идеи успешно обобщаются на дальние комплементарные взаимодействия благодаря существенному сокращению числа структур за счет рассмотрения только очень длинных и почти идеально комплементарных спариваний [113].

В заключение остается добавить, что в связи с быстрым развитием искусственного интеллекта методы, основанные на машинном обучении приобретают все большую популярность в предсказании вторичной структуры РНК [294; 350]. Они основаны на статистических моделях с большим числом параметров и предполагают возможность обучения этой модели на основе экспериментально определенных данных о структуре РНК. Однако параметры этих моделей трудно интерпретировать в физико-химических терминах, а качество и количество экспериментальных данных о дальних взаимодействиях, как будет показано в этой диссертационной работе, в настоящее время недостаточно для получения надежных предсказаний структуры РНК.

Глава 2. Материалы и методы

В этом разделе кратко перечисляются общие для всех глав материалы и методы. Методы, разработанные в диссертации, а также подробное описание специфических для каждого раздела методов и данных можно найти в соответствующих главах, а также в цитируемых публикациях.

2.1 Экспериментальные методы

2.1.1 Конструирование минигенов

В *D. melanogaster* мини-гены, содержащие интересующие экзоны и интроны, амплифицировали из геномной ДНК с использованием полимеразы Taq Precision Plus (Stratagene) и встраивали в плазмиду pRMHA5 под индуцируемым металлотионеиновым промотором. Клетки Schneider S2-L4 трансфицировали с использованием реагента для трансфекции Effectene (Qiagen). Промотор индуцировали через 24 часа после трансфекции добавлением в среду 10 мкМ солей меди, а еще через 24 часа клетки собирали. РНК очищали с использованием набора RNeasy Mini Kit (Qiagen). Обратная транскрипция (ОТ) осуществлялась с 1 мкг РНК с обратным праймером олиго-dT. Полуколичественная ПЦР выполнялась с прямым праймером и обратным праймером, специфичным либо для вектора, либо для гена (как указано на рисунках) с использованием 1/40 реакции ОТ. Полуколичественную ОТ-ПЦР для эндогенных мРНК проводили с использованием олиго-dT-праймера для ОТ, используя 1 мкг общей РНК клеток S2, и праймеры, которые были расположены внутри экзонов, граничащих с исследуемым событием сплайсинга. Контроль проводили без добавления фермента ОТ, чтобы различать амплификацию РНК и ДНК. Сплайсинг визуализировали на агарозных гелях, а полосы определяли количественно с помощью программы NIH ImageJ. Фрагменты кДНК продуктов сплайсинга клонировали в вектор pGEM-T Easy (Promega) и идентифицировали путем секвенирования. Мутагенез проводили с использованием метода QuikChange (Stratagene)

согласно рекомендациям производителя, а полученные мутанты проверяли секвенированием.

Для генов человека, последовательности для вставки в миниген амплифицировали из геномной ДНК клеточной линии A549 с использованием ДНК-полимеразы Q5 High-Fidelity (New England Biolabs). Фрагменты, кодирующие экзоны 18–20 гена *CASK*, экзоны 5–7 гена *PHF20L1* и экзоны 6–8 гена *ATE1*, были клонированы в экспрессионный вектор pRK5 под промотором CMV. Миниген с фрагментом гена *PHF20L1* был собран с использованием безрестрикционного клонирования. Фрагмент гена *CASK* клонировали с использованием протокола клонирования с тупым концом. Миниген *ATE1* был собран из трех фрагментов: первый был вставлен в вектор pRK5 с использованием безрестрикционного клонирования, а следующие два фрагмента были вставлены в полученную плазмиду с использованием набора для клонирования сборки ДНК NEBuilder HiFi (New England Biolabs). Миниген, кодирующий экзоны 3 и 4 гена *BRD2*, был создан по методу клонирования с тупым концом. Миниген, кодирующий экзоны 5 и 6 гена *BRD3*, был получен с использованием набора для клонирования NEBuilder HiFi DNA Assembly Master Mix. Фрагменты амплифицировали с помощью высокоточной ДНК-полимеразы Q5 (New England Biolab). Праймеры для клонирования и мутагенеза, а также условия и праймеры для экспериментов ПЦР в реальном времени можно найти в приложениях к [24; 79; 216; 351]. Все конструкции были подтверждены секвенированием.

2.1.2 Трансфекция плазмидами и АОН

Клетки S2 дрозофилы культивировали при 28°C в среде Шнайдера для дрозофилы и трансфицировали плазмидами дикого типа или несущими мутации минигенными плазмидами с использованием реагента для трансфекции Effectene (Qiagen) как рекомендовано производителем. Клетки аденокарциномы легкого человека A549 культивировали в модифицированной среде Дульбекко Игла и питательной смеси F-12 (1:1) с добавлением 10% фетальной бычьей сывороткой, 1% GlutaMAX, 50 ед./мл пенициллина и 0.05 мг/мл стрептомицина (все продукты от Thermo Fisher Scientific) при 37°C в среде 5% CO_2. Плазмиды дикого типа или несущие мутации минигенные плазмиды трансфицировали

в клетки A549 с использованием Lipofectamine 3000 (Invitrogen) по протоколу обратной трансфекции в течение 24 часов.

Антисенс олигонуклеотиды (АОН) были изготовлены на основе LNA с заменой оснований ДНК в каждом втором нуклеотиде [352]. Синтез LNA/ДНК миксмеров был осуществлен АО «Синтол». Последовательности АОН перечислены в приложениях к [24; 79; 351]. Липофекцию АОН проводили с использованием липофектамина RNAiMAX (Invitrogen) на 50–70% конфлюэнтных клеток. Клетки собирали через 48 часов после обработки. α-аманитин (Sigma) добавляли в концентрации 1 мкг/мл или 2 мкг/мл к 50–70% конфлюэнтным клеткам. Клетки собирали через 24 часа после обработки. В экспериментах, когда клетки трансфицировали одновременно минигенами и АОН, плазмиды и АОН смешивали перед трансфекцией, затем эти смеси трансфицировали с помощью Lipofectamine 3000, а через 24 часа после обработки клетки собирали. Эксперименты с обработкой α-аманитином и АОН/минигенами проводили следующим образом. Сначала клетки трансфицировали АОН/минигенами с использованием обратной трансфекции, через 12–14 ч после трансфекции среду меняли и добавляли α-аманитин, а через 24 часа клетки собирали. В экспериментах по инактивации экспрессии фактора PTBP1 клетки трансфицировали 100 нМ контрольной миРНК против гена люциферазы светлячка, или 100 нМ миРНК против фактора PTBP1. В экспериментах по инактивации системы NMD за три часа до сбора к клеткам добавляли циклогексимид, получая конечную концентрацию 300 мкг/мл.

2.1.3 Вестерн блоттинг

В экспериментах по замедлению элонгации транскрипции для вестерн-блоттинга плазмиду pCMV3-NELFE (HG15217-UT, Sino Biological) трансфицировали в клетки A549 с использованием Lipofectamin 3000 (Invitrogen). Клетки лизировали буфером RIPA через 24 часа. Клетки без трансфицированной плазмиды использовали в качестве отрицательного контроля. Клеточные лизаты (3 мкг общего белка) разделяли гель-электрофорезом на 10% полиакриламидном геле с додецилсульфатом натрия (ДДС-ПААГ) в денатурирующих условиях и переносили на нитроцеллюлозную мембрану. Сначала проводили обработку

при 4°C в течение 12 часов с использованием антител против NELFE [353] (1:500) и GAPDH (Thermo Fisher Scientific, 39-8600, 1:3000). В экспериментах по инактивации экспрессии фактора PTBP1 клеточные лизаты также разделяли гель-электрофорезом на 10% ДДС-ПААГ, белки переносили на нитроцеллюлозные мембраны. Мембраны блокировали в трис-буферном физиологическом растворе (TBS, pH=7.4), содержащем 5% бычьего сывороточного альбумина (BSA), трижды промывали TBS и инкубировали с первичным антителом против PTBP1 (Cloud-Clone Corp., PAC737Hu01) или поликлональными мышиными антителами против глицеральдегидфосфатдегидрогеназы (GAPDH) [6C5] (Abcam, ab8245) в течение ночи при 4°C (разведение 1:2000). Затем на 1 час при комнатной температуре добавляли козьи анти-кроличьи IgG, конъюгированные с пероксидазой хрена (Invitrogen, G21234, 1:2500), или вторичные антитела против мышиного IgG (H+L) (Thermo Fisher Scientific, 62-6520), с последующим обнаружением с использованием реагента Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences) и системы визуализации Bio-Rad ChemiDoc XRS.

2.1.4 Замедление элонгации транскрипции

Клеточная линия A549 поддерживалась в модифицированной среде Дульбекко Игла и питательной смеси F-12 (1:1), содержащей 10% фетальной бычьей сыворотки, 50 ед/мл пенициллина и 0.05 мг/мл стрептомицина (Thermo Fisher Scientific) при 37°C в среде 5% CO_2. Для обработки α-аманитином (Sigma) к клеткам добавляли 1 или 2 мкг/мл α-аманитина при конфлюэнтности 50–70%. После 24 часов обработки клетки собирали, РНК выделяли с использованием набора PureLink RNA Mini Kit (Thermo Fisher Scientific), а затем выделяли поли(A) фракцию мРНК с использованием Dynabeads Oligo(dT) 25 (Thermo Fisher Scientific). Библиотеки кДНК конструировали с использованием набора NEBNext Ultra II for Illumina (New England BioLabs) для подготовки библиотеки направленных РНК со временем фрагментации 10 минут. Полученные библиотеки секвенировали в двух биорепликах с использованием прибора NextSeq500 (Illumina). Для каждого образца с длиной чтения 75 п.о. было получено 33–41

миллионов чтений. Полученные чтения анализировались программой IPSA для получения числа с разрывами и вычисления степени включения экзонов [354].

2.2 Биологические источники РНК для высокопроизводительного секвенирования

В диссертации использовались данные высокопроизводительного секвенирования РНК из клеточных линий и тканей человека, полученные в международных консорциумах ENCODE (Encyclopedia of DNA Elements) и GTex (Genotype Tissue Expression project) при участии автора диссертации [26; 27; 355—359], а также данные, полученные другими авторами.

Каталоги используемых консорциумом ENCODE клеточных линий, условия их культивации, протоколы выделения РНК, получение ядерных и цитоплазматических фракций, фракционирование по длине и внутриклеточной локализации, фракционирование РНК на поли(A)$^+$ и поли(A)$^-$, удаление рибосомальных РНК, подготовка библиотек для секвенирования, а также другие экспериментальные протоколы подробно описаны в [360]. Методика забора биоматериалов и выделения РНК в консорциуме GTEx описана в [361]. Полный каталог исследуемых тканей и трансформированных клеток приводится в таблице S1 в [361]. Были использованы образцы РНК, соответствующие критерию целостности (RIN score) 6.0 и выше. Для секвенирования использовали по меньшей мере 1 мкг тотальной РНК, из которой выделяли поли(A)$^+$ фракцию. Доноры тканей выбирались из соображений увеличения статистической мощности для анализа локусов количественных признаков и получения широкого спектра тканей у некоторых отдельных доноров.

Дополнительно были использованы данные высокопроизводительного секвенирования РНК из экспериментов по замедлению элонгации транскрипции в мутантах RNAPII, данные из экспериментов по ответу транскриптома на инактивацию РСБ и их футпринтинга по методу eCLIP, а также данные из экспериментов по инактивации системы NMD. Число образцов в эксперименте, число биореплик, число образцов в контроле, число биореплик в контроле, а также коды доступа в репозитории GEO приведены в табл. 1

Таблица 1 — Источники данных высокопроизводительного секвенирования. Эксп. — число образцов x число реплик. Контр. — число образцов x число реплик.

Образец	Метод	Фракция	Эксп.	Контр.	Источник
HepG2	РНК-сек	поли(A)$^+$	1x2	1x2	ENCODE
K562	РСБ-KD + РНК-сек	поли(A)$^+$	234x2	25x2	ENCODE
HepG2	РСБ-KD + РНК-сек	поли(A)$^+$	230x2	24x2	ENCODE
HEK293	Rpb1 C4/R749H	поли(A)$^+$	1x4	1x4	GSE63375
K562	eCLIP		91x2	25x2	ENCODE
HepG2	eCLIP		75x2	24x2	ENCODE
HEK293	XRN1/UPF1-KD + РНК-сек	поли(A)$^+$	1x1	1x1	GSE57433
7 линий	RIC-seq		7x2	7x2	GSE190214
50 тканей	РНК-сек	поли(A)$^+$	8551x1	0x0	GTEx
18 опухолей	РНК-сек	поли(A)$^+$	698x1	698x1	TCGA

2.3 Обработка данных высокопроизводительного секвенирования

В разд. 3.3 использовались данные секвенирования поли(A)$^+$ РНК из клеточной линии HepG2 (номера доступа ENCFF670LIE и ENCFF074BOV), которые были загружены в BAM-формате с веб-сайта консорциума ENCODE [362]. Данные по отклику транскриптома клеточных линий HepG2 и K562 на инактивацию экспрессии 250 РСБ были загружены в BAM-формате из репозитория ENCODE [362] (коды доступа в [23]). Данные секвенирования поли(A)$^+$ РНК для дикого типа и мутантных клеток HEK293 Rpb1 C4/R749H были загружены из репозитория Gene Expression Omnibus (GEO) под номером доступа GSE63375 [363]. Данные секвенирования РНК поли(A)$^+$ РНК из клеточной линии A549, обработанной α-аманитином, были получены как описано ниже (разд. 2.1.4) и картированы на сборку GRCh37 генома человека с использованием программы STAR v2.3.1z с настройками по умолчанию.

В разд. 4.1 данные, полученные в экспериментах RIC-seq на семи клеточных линиях человека, включая GM12878, H1, HeLa, HepG2, IMR90, K562 и hNPC (по две биореплике каждый), были загружены из репозитория GEO под номерами доступа GSE127188 и GSE190214 в формате FASTQ. Контрольные эксперименты по секвенированию РНК были загружены из консорциума ENCODE под номерами доступа, указанными в таблице S4 в [24]. Все данные RIC-seq и данные по секвенированию РНК обрабатывались пакетом RNAcontacts с настройками по умолчанию [364].

В разд. 5.2.3 данные секвенирования поли(A)$^+$ РНК были загружены в BAM-формате с веб-сайтов проектов GTEx и TCGA [365; 366]. Данные секвенирования РНК в эксперименте по замедлению элонгации транскрипции в мутантах RNAPII были загружены из репозитория GEO под номером доступа GSE63375 [363].

В разд. 6.1 использовались данные по отклику транскриптома клеточной линии HEK293 Flp-In T-Rex на совместную инактивацию факторов XRN1 и UPF1 (репозиторий GEO под номером доступа GSE57433) [367], которые были картированы на сборку GRCh37 генома человека с использованием картировщика STAR-2.5.3a. Данные по отклику транскриптома клеточных линий HepG2 и K562 на инактивацию экспрессии РСБ были загружены в BAM-формате в виде готовых картирований на сборку GRCh37 генома человека из репозитория ENCODE [362]. Также были использованы данные футпринтинга (enhanced cross-linking and immunoprecipitation, eCLIP) для 115 РСБ, представленные в [315].

В разд. 3.3 и разд. 6.2 выравнивания коротких чтений, для 8551 образца из Консорциума GTEx v7 [365], были загружены из базы данных dbGaP в виде готовых картирований на сборку GRCh37 генома человека в BAM-формате (код проекта phs000424/GRU). Результаты экспериментов по пертурбации уровней экспрессии РСБ с последующим секвенированием РНК, включая эксперименты по частичной и полной инактивации и суперэкспрессии, перечисленные в таблице S3 [216], были загружены с портала ENCODE и архива SRA в форматах BAM и FASTQ, соответственно, и картированы на сборку GRCh37 генома человека с использованием картировщика STAR-2.7.7a [110]. Для анализа дифференциального сплайсинга использовался пакет rMATS v.4.1.1 [368].

В разд. 6.3 данные из Консорциума GTEx были загружены из базы данных dbGaP в формате FASTQ. Эти, а также данные, полученные из консорциума TCGA, были картированных на сборку GRCh38 генома человека с помощью картировщика STAR-2.7.3a.

В разд. 6.2 данные уровня MS1 об экспрессии белков в тканях человека были загружены с портала Proteomics DB [369]. Ткани из Proteomics DB сопоставлялись с тканями GTEx, затем вычислялись медианные значения Ψ (см. разд. 2.4.4), а затем ткани были отсортированы по возрастанию Ψ. Для проверки отрицательной взаимосвязи между уровнем экспрессии белка и непродуктивным сплайсингом, т.е. того, следует ли уровень экспрессии бел-

ка нисходящему тренду при упорядочивании по Ψ, применялся односторонний критерий Джонкхира-Терпстры.

2.4 Биоинформатические методы

2.4.1 Гены и геномы

В разд. 3.2 используются аннотации и сборки геномов млекопитающих и дрозофилид, перечисленные в таблице S1 в [113]. По умолчанию использовалась сегментация, индуцированная границами аннотированных в эталонном виде экзонов (*H. sapiens* для млекопитающих и *D. melanogaster* для дрозофилид). Генный сегмент классифицировался как экзонный если он принадлежал хотя бы одному аннотированному экзону, и как интронный в противном случае. Ортологичные экзоны были получены из парных полногеномных выравниваний BLASTz, загруженных в chain формате с веб-сайта Геномного Браузера UCSC [370]. Начиная с раздела разд. 3.3 использовалась сборка генома человека GRCh37 и аннотация GENCODE версии v35lift37, за исключением с разд. 6.3, в котором использовалась сборка генома человека GRCh38 и аннотация GENCODE версии v42 [371].

2.4.2 Оценка значимости компенсаторных замен

Для оценки количества и статистической значимости компенсаторных замен в комплементарных участках использовались глобальные множественные выравнивания геномов 99 позвоночных с геномом человека [372]. Последовательности, которые содержали вставки и делеции по сравнению с эталонным видом, удалялись. Два блока выравнивания объединялись через искусственный спейсер, содержащий десять адениновых нуклеотидов, и передавались на вход программе R-scape v1.2.340 вместе с филогенетическим деревом в формате STOCKHOLM и предсказанной вторичной структурой РНК [12]. Результиру-

ющее E-значение определялось как произведение E-значений, предсказанных R-scape, по всем парам оснований в РНК-структуре. На всех рисунках приводятся сокращения в названиях видов как в [373].

2.4.3 Статистические методы

Для статистической обработки данных во всех разделах использовалось программное обеспечение R версии 3.4.1 и выше, а также пакет для визуализации данных ggplot2. Если не оговорено противное, для парных сравнений средних (медиан) в независимых выборках использовался критерий суммы рангов Манна-Уитни, а в зависимых — знаковый критерий Вилкоксона. Тестирование выполнялось при помощи встроенных R-функций с использованием поправки на непрерывность при приближении нормальным распределением. Для сравнения пропорций использовался двухвыборочный z-тест. Отношение шансов (OR) вычислялось как отношение доли исходов в пользу воздействия в опыте к доле исходов в пользу воздействия в контроле. Стандартная ошибка и доверительные интервалы OR вычислялись с использованием нормальной аппроксимации распределения логарифма отношения шансов. В тексте диссертации в скобках приводятся двусторонние P-значения (P), если не указано иное. Для поправки на множественное тестирование использовалась поправка Бонферони-Холма для множественного тестирования (family-wise error rate $FWER < 0.05$). Ящичковые диаграммы на всех рисунках представлены медианой, верхним и нижним квартилем, а также максимальным и минимальным значением по выборке, попадающим в верхнюю и нижнюю границу усика; выбросы не показаны. Числа после знака $\pm$, а также интервалы ошибок на всех рисунках обозначают 95% доверительные интервалы. Статистически значимые различия на стандартных уровнях значимости (5%, 1% и 0.1%) помечены звездочками.

2.4.4 Уровни экспрессии генов и уровни включения экзонов

Уровни экспрессии генов рассчитывались на основании суммарного числа картированных чтений в экспериментах по секвенированию РНК. Значения уровней экспрессии были нормализованы в соответствии с методологией DESeq2 [374]. А именно, элементы строк матрицы экспрессии, столбцы которой соответствуют образцам, а строки соответствуют генам, делились на медиану по строке, а затем нормировочный фактор sf_k для образца k рассчитывался как медиана по столбцу. Значения экспрессии делили на sf_k, логарифмировали, и центрировали путем вычитания медианы по всем образцам. Для оценки дифференциальной экспрессии генов в экспериментах по пертурбациям экспрессии РСБ число чтений извлекали из соответствующих BAM-файлов с помощью программы RNA-SeQC [375] и анализировали с помощью пакета DESeq2 [374] с использованием коррекции сжатия apeglm [376].

Разрывные чтения, а также непрерывные чтения, поддерживающие удержание интронов, извлекались из выравниваний с использованием пакета IPSA с настройками по умолчанию [354]. Чтения фильтровались по значению энтропии Шеннона распределения отступов (расстояние от начала чтения до разрыва), статусу аннотации и наличию канонических динуклеотидов (GT/AG) в сайтах сплайсинга [354]. Во всех разделах, за исключением разд. 6.2, степень включения кассетного экзона Ψ (percent-spliced-in или PSI) рассчитывалась по формуле

$$\Psi = \frac{I}{I + 2 * S},$$

где I — число чтений, поддерживающих включение экзона, а S — число чтений, поддерживающих исключение экзона. Значения Ψ со знаменателем менее 20 отбрасывались. В разд. 6.2 степень включения вычислялась не только для кассетных экзонов, но и для других событий AC, например альтернативных 5'ss и 3'ss, в случаях, когда чтения позволяли различать альтернативно сплайсированные изоформы транскриптов. При этом метрика Ψ определялась по отношению к NMD-изоформе, т.е. $\Psi = 1$ для ядовитого экзона означало, что он всегда включается, а $\Psi = 1$ для необходимого экзона означало, что он всегда пропускается.

Изменение степени включения экзонов при инактивации экспрессии (или в сравнении любых двух условий) оценивали с помощью метрики $\Delta\Psi = \Psi_{KD} -$

Ψ_C, где Ψ_{KD} и Ψ_C — степени включения экзона в эксперименте по инактивации экспрессии и контрольном эксперименте, соответственно. Чтения из биореплик объединялись. Для того, чтобы учесть изменение АС в транскриптах, уровни экспрессии которых изменяются при инактивации, была использована регрессионная модель для зависимости от среднего значения знаменателей Ψ_{KD} и Ψ_C. А именно, была построена линейная модель $\Delta\Psi = \beta_0 + \beta_1 \log_{10}(FC)$ для экзонов с $\Delta\Psi \neq 0$, где FC — отношение знаменателей Ψ_{KD} и Ψ_C, после чего каждое значение $\Delta\Psi$ заменялось остатком в этой линейной модели. Для оценки статистической значимости $\Delta\Psi$ использовалась похожая методика, в которой в каждом из заранее выбранных интервалов значений $\log_{10}(SJ)$, где SJ — суммарное число разрывных чтений, поддерживающих включение и исключение экзона в эксперименте и контроле, вычислялось среднее и стандартное отклонение значений $\Delta\Psi$, на основании которых $\Delta\Psi$ преобразовывалось в z-значение и с использованием нормального распределения вычислялось P-значение. Ко всем P-значениями применялась поправка на множественное тестирование, а также они преобразовывались в q-значения [377].

2.4.5 Кластеры сайтов полиаденилирования

Для нахождения сайтов полиаденилирования в данных GTEx рассматривали чтения, содержащие область мягкого отсечения (soft clip region) размером не менее 6 нт. Требовалось, чтобы сообщаемая нуклеотидная последовательность невыровненной области содержала не менее 80% T, если мягкое отсечение находился в начале чтения, и 80% A, если оно находилось в конце. Поли(A)-чтения объединялись по геномному положению первого невыровненного нуклеотида, соответствующего положению сайта полиаденилирования, в результате чего для каждого значения отступа были получены частоты, по которым затем вычислялась энтропия Шеннона (H). Использовался порог $H \geqslant 2$ в сочетании с минимальной длиной отступа 6 нт. Сайты, которые располагались в пределах 10 нт друг от друга, были объединены в кластеры сайтов полиаденилирования (КСПА) с помощью односвязной кластеризации.

Глава 3. Предсказание дальних взаимодействий в структуре РНК

3.1 Постановка задачи предсказания дальних взаимодействий

Известные из литературы дальние взаимодействия в структуре эукариотических РНК обладают рядом характерных свойств. Во-первых, они эволюционируют в условиях отрицательного отбора и поэтому высококонсервативны, хотя степени консервативности отдельных структурных элементов зависят от того, когда они были приобретены. Во-вторых, ядро структуры часто представляет из себя собой длинное, почти идеальное комплементарное спаривание, что, вероятно, связано с ограничениями на свободную энергию для поддержания комплементарности на больших расстояниях. Наконец, почти все структурные элементы расположены в синтенных областях, например, в интронах, разделяющих ортологичные экзоны, что может быть связано с их ролью в регуляции АС.

Первое свойство обосновывает применимость подхода «сначала выравнивание, потом фолдинг» (рис. 1.7). Противоположный путь, «сначала фолдинг, потом выравнивание», ранее систематически не исследовался, поскольку предсказания оптимальной структуры для одной последовательности недостаточно точны для построения согласованного выравнивания [331; 378; 379]. Этот путь является намного более затратным с вычислительной точки зрения из-за того, что нуклеотидные последовательности могут иметь много сходных по энергетическим характеристикам комплементарных спариваний, и поэтому для построения структурного выравнивания для каждой последовательности необходимо знать множество всех ее вторичных структур.

В данной диссертационной работе развиваются оба этих подхода. В разд. 3.2 обсуждается стратегия «сначала фолдинг, потом выравнивание», которая заимствует некоторые технические идеи из метода GUUGle для поиска комплементарности с учетом G:U пар [380]. Она состоит в преобразовании исходной последовательности в хэш-таблицу, в которой хранится местоположение каждого k-мера, и ее последующем пересечении с хэш-таблицой обратных дополнений для нахождения консервативности и с хэш-таблицами ортологов для обнаружения консервативности. Преимущество этого подхода заключается

в том, что неконсервативные участки не приходится выравнивать. Наоборот, отсутствие консервативности становится полезным свойством при присвоении статистической значимости сохранившимся «островкам», погруженным в неконсервативный фон [107]. По построению не накладывается никаких ограничений ни на расстояние между парами оснований, ни на псевдоузлы.

Затем в разд. 3.3 обсуждается стратегия «сначала выравнивание, потом фолдинг», в которой разреженный алгоритм динамического программирования применяется к предсказанию дальних взаимодействий между консервативными интронными участками, рассматриваемыми как межмолекулярные РНК-РНК взаимодействия. Эта стратегия позволяет охватить большее число структур и обойти технические и комбинаторные ограничения первого метода. Приводится исчерпывающая статистическая характеризация взаимосвязи между консервативными комплементарными участками и сигналами процессинга РНК, такими как сплайс-сайты, сайты полиаденилирования и редактирования РНК, а также выдвигается гипотеза о роли структуры РНК и сплайсинга в предотвращении преждевременной терминации транскрипции.

3.2 Сначала фолдинг, потом выравнивание

В этом разделе представлены результаты, полученные мной в серии работ с 2009 по 2014 год [107; 108; 113]. Все они относятся к предсказанию интронных структур РНК, не ограниченных локальными взаимодействиями, в группах ортологичных последовательностей без использования множественного выравнивания. Парные выравнивания последовательностей генов из заранее выбранной группы организмов используются только для нахождения границ ортологичных экзонов, по которым из соображений синтении определяются ортологичные интроны. Технические детали нахождения ортологичных интронов подробно описаны в публикациях, поэтому здесь для краткости они будут опущены. Используемые в перечисленных работах методы различаются между собой, однако все они основаны на анализе k-меров, т.е., коротких подпоследовательностей длины k.

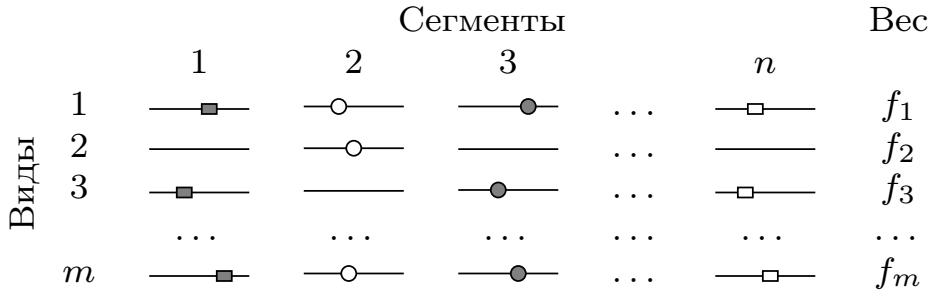


Рисунок 3.1 — Ортологичные последовательности s_{ij} индексируются идентификаторами сегментов $j = 1...n$ в каждом из видов $i = 1...m$. Серые прямоугольники (круги) комплементарны белым прямоугольникам (соответственно кругам). Все прямоугольники и круги встречаются в ортологичных сегментах у трех видов, но прямоугольники встречаются одновременно у трех видов, а круги встречаются одновременно только у двух.

3.2.1 Краткое описание метода IRBIS

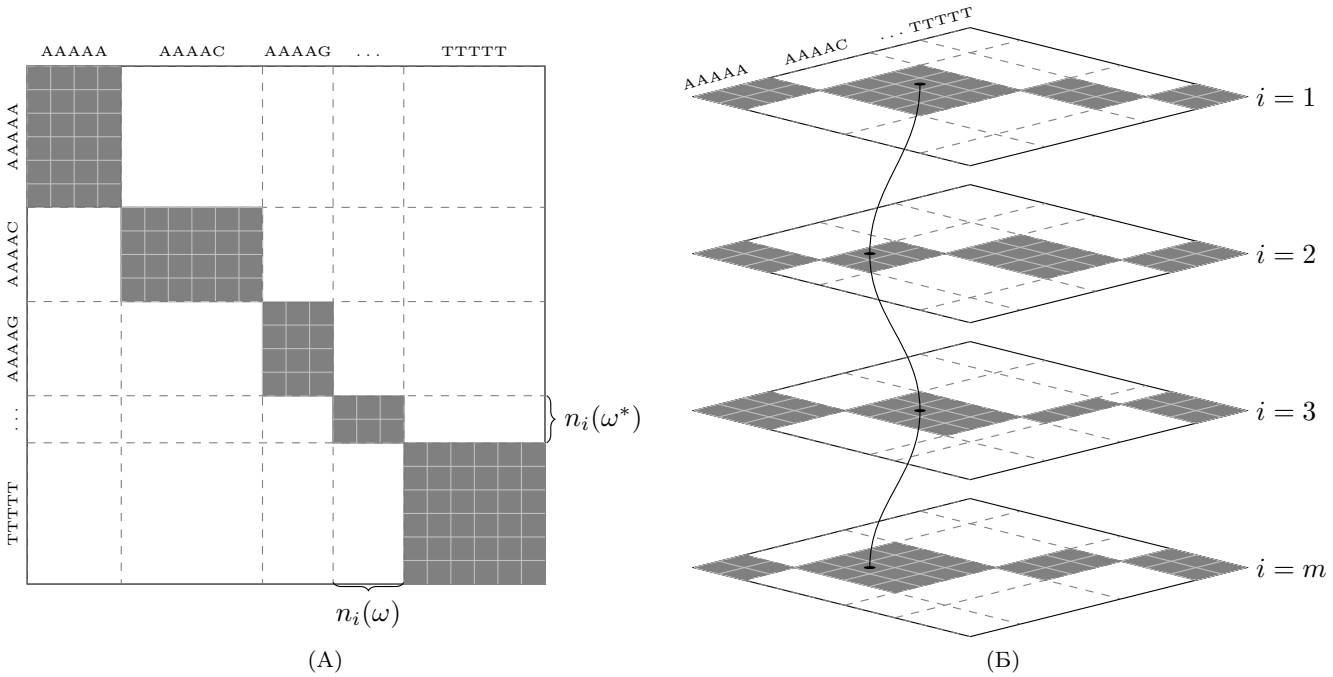
Входные данные состоят из набора невыровненных ортологичных сегментов нуклеотидных последовательностей s_{ij}, где $i = 1,...,m$ индексирует виды, а $j = 1,...,n$ — сегменты, таким образом, что ортологичные сегменты из разных видов i получают одинаковый идентификатор сегмента j (рис. 3.1). Видам присвоены весовые коэффициенты f_i, где $i = 1,...,m$, которые в сумме дают единицу. Разыскиваются короткие комплементарные слова длины k (k-меры, по умолчанию $k = 8$), которые консервативны, то есть встречаются во «многих» s_{ij} для одного и того же i. Один из способов определить «многие» — это потребовать, чтобы сумма весов, соответствующих s_{ij}, была больше некоторого порога t. Также будем требовать, чтобы комплементарные пары содержали как минимум h спариваний G:C и не более g спариваний G:U. Комплементарные k-меры могут перекрываться, образуя более длинные структуры. Будем кластеризовать перекрывающиеся k-меры и требовать как минимум L комплементарных нуклеотидов в каждом кластере (полный список параметров приводится в табл. 1 в [113]).

Для разных приложений имеет смысл искать комплементарность между разными наборами сегментов. Это формализуется путем ограничения поиска комплементарности на сегменты s_{ia} и s_{ib}, где $a \in A$, $b \in B$, а A и B — некоторые (не обязательно непересекающиеся) подмножества множества $\{1,...,n\}$. Например, при поиске мишеней малых ядрышковых РНК (мякРНК) в интронах множество A представляет собой набор сегментов мякРНК, а множество B — набор

интронных сегментов белоккодирующих генов. Число комбинаций «все против всех» для A и B может быть очень велико. Поэтому поиск можно ограничить подмножеством комбинаций, определяемых некоторым отношением $\mathcal{R}$ между элементами множеств A и B. Например, для поиска интронных структур внутри генов оба множества A и B представляют собой набор интронных сегментов всех генов, но в $\mathcal{R}$ допускаются только сегменты из одного и того же гена.

Суть метода IRBIS сводится к следующему. Для каждого вида i создается хеш-таблица, которая каждому k-меру ω сопоставляет массив $H_i(\omega)$ таких упорядоченных пар (j,p), что $(j,p)\in H_i(\omega)$ тогда и только тогда, когда ω встречается в позиции p последовательности s_{ij}. Массив $H_i(\omega)$ будет автоматически отсортирован в лексикографическом порядке, т.е. $(j,p) \leqslant (j',p')$, если $j < j'$ или $j = j$ и $p \leqslant p'$, если последовательностиs_{ij}, изначально отсортированные по возрастанию j, читать слева направо. Нас интересуют k-меры, встречающиеся во многих s_{ij} для одного и того же j. Для их поиска вводится «забывающее позицию» отношение, при котором $(j_1,p_1) \simeq (j_2,p_2)$ всякий раз, когда $j_1 = j_2$. Это — отношение эквивалентности между элементами $H_i(\omega)$ для разных i, причем наиболее консервативные k-меры соответствуют самым большим классам эквивалентности. Поскольку консервативные k-меры определяются независимо от их положения в s_{ij}, можно определить более сильную форму отношения $\simeq$, в которой $(j_1,p_1) \simeq (j_2,p_2)$ всякий раз, когда $j_1 = j_2$ и $|p_2 - p_1| < \Delta$. Однако следует понимать, что невыровненные последовательности нельзя сравнивать по позициям, и поэтому порог Δ должен быть очень большим для того, чтобы отсекать только k-меры, расположенные в заведомо разных частях последовательностей s_{ij}.

Отношение $\simeq$ слишком строгое в том смысле, что оно предполагает точное совпадение консервативных k-меров. Вместо обычных, непрерывных k-меров аналогично можно рассматривать k-меры с пробелами (gapped k-mers), что позволяет допускать небольшое число мутаций в консервативных областях, а также моделировать короткие внутренние петли в структурах РНК. Известно, что k-меры с асимметричными пробелами лучше подходят для выравнивания последовательностей по методу фильтрации с потерями [381], однако в данной задаче преимущества асимметричных пробелов не могут быть полностью использованы, поскольку перекрывающиеся k-меры в дальнейшем кластеризуются и расширяются. Для краткости описания я не буду здесь на этом останавливаться [113].



(А) (Б)

Рисунок 3.2 — **(А)** Пространство комплементарных спариваний длины k. Горизонтальная и вертикальная оси соответствуют хеш-таблицам $H_i(\omega)$ и $H_i(\omega^*)$ соответственно; ширина и высота серых прямоугольников — это $n_i(\omega)$ и $n_i(\omega^*)$ соответственно. Серая область представляет собой все возможные комбинации комплементарных k-меров. **(Б)** Для нахождения консервативных комплементарных k-меров необходимо сравнить $H_i(\omega)\times H_i(\omega^*)$ для разных i. Эквивалентные относительно отношения $\simeq$ структуры соединены путем.

Для того, чтобы найти все идеально комплементарные Уотсон-Криковские спирали длины k, необходимо для каждого k-мера $\boldsymbol{\omega}$ найти декартово произведение $H_i(\boldsymbol{\omega})\times H_i(\boldsymbol{\omega}^*)$, где $\boldsymbol{\omega}^*$ — последовательность, комплементарная к $\boldsymbol{\omega}$. На рис. 3.2А схематично изображены все $\boldsymbol{\omega}$ и $\boldsymbol{\omega}^*$ с учетом числа позиций, в которых они встречаются. По построению каждый квадрат в серой области соответствует идеально комплементарной спирали длины k, т.е., если $\boldsymbol{\omega}$ встречается $n_i(\boldsymbol{\omega})$ раз, а $\boldsymbol{\omega}^*$ встречается $n_i(\boldsymbol{\omega}^*)$ раз, то существует $n_i(\boldsymbol{\omega})\cdot n_i(\boldsymbol{\omega}^*)$ таких спиралей. Площадь серой области будет большой тогда, когда $n_i(\boldsymbol{\omega})$ велико для некоторого $\boldsymbol{\omega}$, например, когда s_{ij} состоит из полинуклеотидов (например, поли-А и поли-Т). И наоборот, указанная область будет наименьшей когда серые прямоугольники представляют собой квадраты примерно одинакового размера, но даже в этом случае объем памяти для хранения всех комбинаций оказывается слишком велик. Поэтому следует сразу исключить из рассмотрения участки низкой сложности, поскольку они значительно увеличивают число парных комбинаций k-меров, т.е, мощность множества $H_i(\boldsymbol{\omega})\times H_i(\boldsymbol{\omega}^*)$.

Для того, чтобы сравнивать $H_i(\boldsymbol{\omega})\times H_i(\boldsymbol{\omega}^*)$ для разных i (рис. 3.2В), необходим этап предварительной фильтрации. Этот предварительный этап,

называемый триммингом, использует тот факт, что $H_i(\omega)$ является упорядоченным массивом, и элементы $H_i(\omega)$ можно за линейное время сравнить для всех i для того, чтобы заранее исключить из рассмотрения неконсервативные k-меры. Тримминг принимает на вход набор хеш-таблиц $H_i(\omega)$ и возвращает набор разреженных хеш-таблиц $\widehat{H}_i(\omega)$, в которых $(j,p)\in\widehat{H}_i(\omega)$ тогда и только тогда, когда сумма весов видов, для которых $(j,p)\in H_i(\omega)$, больше заданного порога.

После тримминга, таблицы $\widehat{H}_i(\omega)$ и $\widehat{H}_i(\omega^*)$ содержат только консервативные k-меры, однако сообщить все попарные комбинации ω и ω^* было бы недостаточно, поскольку, например, если ω встречается у четных i, а ω^* — у нечетных, то вместе они не будут присутствовать ни у одного вида (рис. 3.1). Для того, чтобы это учесть, можно применить процедуру тримминга к хеш-таблице $P_i(\omega) = \widehat{H}_i(\omega)\times\widehat{H}_i(\omega^*)$ с каноническим лексикографическим порядком, возникающем на декартовом произведении. Следует отметить, что процедура тримминга наиболее эффективна тогда, когда последовательности s_{ij} содержат мало консервативных k-меров. Поэтому данный метод не подходит для нахождения структур РНК в экзонах, поскольку экзоны эволюционируют под отбором на последовательность белка, и ограничение на консервативность, отсеивающее бóльшую часть k-меров, не приведет к существенному сжатию хеш-таблиц.

В таблице $P_i(\omega)$ перечислены все попарные комбинации вхождений k-мера ω и идеально комплементарного ему k-мера ω^*. Для того, чтобы разрешить небольшое число G:U пар, можно вместо таблицы $H_i(\omega^*)$ использовать таблицу $H_i^*(\omega)$, которая является объединением таблиц $H_i(\omega')$ по всем ω', которые образуют с ω не более чем g G:U пар (обычно $g = 2$). Если рассматривается заранее определенное отношение $\mathcal{R}$ на множествах сегментов A и B, то необходимо построить отдельные хеш-таблицы $H_{A,i}$ и $H_{B,i}$, и при тримминге таблицы $P_i(\omega) = \widehat{H}_{A,i}(\omega)\times\widehat{H}_{B,i}(\omega^*)$ рассматривать только пары (j,p) и (j',p'), где $(j,j')\in\mathcal{R}$.

Для того, чтобы кластеризовать перекрывающиеся k-меры, достаточно отсортировать элементы (j,p,j',p') таблицы $P_i(\omega)$ не лексикографически, а по j, j', p, и p' (в этом порядке). Тогда перекрывающиеся k-меры будут встречаться в отсортированном списке последовательно, а самые длинные кластеры можно идентифицировать с помощью динамического программирования. После этого на стадии постпроцессинга для найденных комплементарных участков можно построить отдельное множественное выравнивание, например с помо-

щью программы MUSCLE [382], а также отдельно выровнять оставшиеся части сегментов и соединить полученные множественные выравнивания с помощью конкатенации.

Данная формулировка позволяет применить описанный метод к широкому кругу задач, связанных с поиском консервативных комплементарных участков (ККУ), например к поиску интронных структур РНК в окрестностях сплайс-сайтов или целых интронах, поиску мишеней микроРНК в 3'-НТО и т. д. Следует также отметить, что он получает на вход наборы ортологичных сегментов нуклеотидных последовательностей, нахождение которых представляет из себя отдельную нетривиальную задачу. Под сегментами могут пониматься экзоны, интроны или их части. Поэтому IRBIS содержит этап препроцессинга данных. В заранее выбранной кладе выбирается один «эталонный» вид (например, человек у плацентарных млекопитающих) и в нем производится разбиение (сегментация) геномной последовательности на интервалы, определяемые границами аннотированных экзонов, из которых затем отбираются интронные сегменты. Полногеномные выравнивания, полученные при помощи программы BLASTz [370], позволяют найти (не всегда однозначные) проекции границ экзонов эталонного вида на другие виды, из которых с помощью стандартной процедуры можно извлечь уникальные и взаимно однозначные проекции, по которым из соображений синтении определить ортологичные интроны как имеющие соответствующие границы. Ввиду технического характера этих подготовительных шагов, в данной диссертации они не рассматриваются. Детальное описание всех процедур, включая препроцессинг данных, приводится в [113].

3.2.2 Оценка чувствительности и доли ложных предсказаний

По построению описанный метод должен находить комплементарные k-меры даже в последовательностях, которые невозможно выровнять. Для того, чтобы это проверить, были сгенерированы $n = 200$ случайных последовательностей длины 1000 в $m = 16$ видах, в которые затем через одну были добавлены комплементарные k-меры ($k = 8$) на независимых равномерно распределенных позициях (в четные — ω, а в нечетные — ω^*). Комплементарность между ними была обнаружена в 100% случаев. Чувствительность можно также оценить

иначе, генерируя случайные выравнивания последовательностей. Так, с одинаковыми настройками метод RNAplex [383] идентифицировал почти в три раза меньше РНК-РНК взаимодействий по сравнению с описанным методом, причем во всех пропущенных им случаях комплементарные k-меры не были выровнены друг с другом [113].

В применении к интронам белок-кодирующих генов плацентарных млекопитающих и дрозофилид ($n = 350000$ и 60000, соответственно) с порогами на консервативность $t = 0.8$ и $t = 0.75$, соответственно, без рассмотрения геномных повторов и участков низкой сложности, метод IRBIS находит 832 и 632 пары ККУ длины 12 нт или более, если для каждой пары сегментов сообщать только одну наилучшую структуру.

Оценка специфичности подразумевает знание полного каталога «настоящих» РНК-структур для того, чтобы вычислить вероятность ошибки первого рода, что невозможно. Вместо специфичности можно оценить долю ложных положительных предсказаний, используя рандомизацию. Для этого была разработана процедура, называемая «пересоединением» (rewiring) и заключающаяся в рассмотрении химерных транскриптов, состоящих из сегментов, принадлежащих разным генам [108]. Реализовать такую процедуру очень просто, поскольку отношение $\mathcal{R}$, которое при поиске внутримолекулярных структур содержит индексы сегментов, принадлежащих одному и тому же гену, можно случайно «перемешать», выбирая сегменты из различных генов и контролируя динуклеотидный состав и уровень консервативности последовательностей так, чтобы они совпадали с характеристиками настоящих транскриптов. В предположении нулевой гипотезы о том, что транскрипты кодирующих белки генов не образуют комплементарных взаимодействий, число предсказаний в таких химерных транскриптах отражает долю ложных положительных предсказаний. Эта доля колеблется в пределах от 15% до 30% в зависимости от порога на длину и степени консервативности структуры [113]. Как мы увидим далее в разд. 3.2.5, эта оценка является крайне пессимистичной, поскольку предположение о том, что различные мРНК не взаимодействуют друг с другом не исключает того, что они могут содержать консервативные регуляторные элементы, которые встречаются на противоположных цепях ДНК.

3.2.3 Характеристики интронных структур РНК

В данном разделе кратко описываются характеристики интронных структур РНК у насекомых и млекопитающих, полученные в результате применения разных вариантов описанного метода к различным множествам интронов. В первой работе, описывающей дальние взаимодействия в структурах РНК у дрозофил, требовалось, чтобы комплементарные области располагались в аннотированных интронах на расстоянии не более 150 нт от экзонов, но при этом ограничений на расстояние между комплементарными областями не накладывалось [107]. В следующей работе, посвященной структурам РНК у дрозофил и человека, также использовалось интронное окно длины 150 нт, но расположение структурных элементов по отношению к сплайс-сайтам могло быть любым и не ограничивалось аннотированными интронами [108]. Наконец, в последней работе ограничений на расстояние от консервативных участков до границы интрона не накладывалось, но рассматривались все комбинации интронных сегментов с высоким порогом на консервативность и сообщалась только одна (наилучшая) структура для каждой пары сегментов [113]. Поэтому сравнивать общее количество найденных этими методами структур не имеет смысла.

Общая тенденция в расположении найденных во всех работах комплементарных участков заключается в том, что они расположены неслучайно по отношению к сплайс-сайтам и событиям АС, причем для некоторых типов событий АС они более характерны, чем для других. У дрозофил (*D. melanogaster*) часто наблюдаются структуры, выпетливающие кассетные экзоны, а внутри подгруппы альтернативно сплайсируемых интронов наблюдается обогащение интронами, которые содержат альтернативные акцепторные сайты [107]. Также наблюдается обогащение интронами, которые содержат сайты полиаденилирования (рис. 3.3А,Б). Внутри области поиска (интронное окно 150 нт) комплементарные участки также расположены не случайно, а предпочитают находиться на расстоянии 60 нт в сторону 3'-конца от донорного сайта и на расстоянии 80 нт в сторону 5'-конца от акцепторного сайта, избегая пересечений с полипиримидиновым трактом и сайтом ветвления (рис. 3.3В). Интроны дрозофилы имеют среднюю длину около 60 п.н., однако среди интронов со структурами РНК наблюдается существенное обогащение длинными, и особенно длинными альтернативными интронами (рис. 3.3Г,Д). Примечательно, что

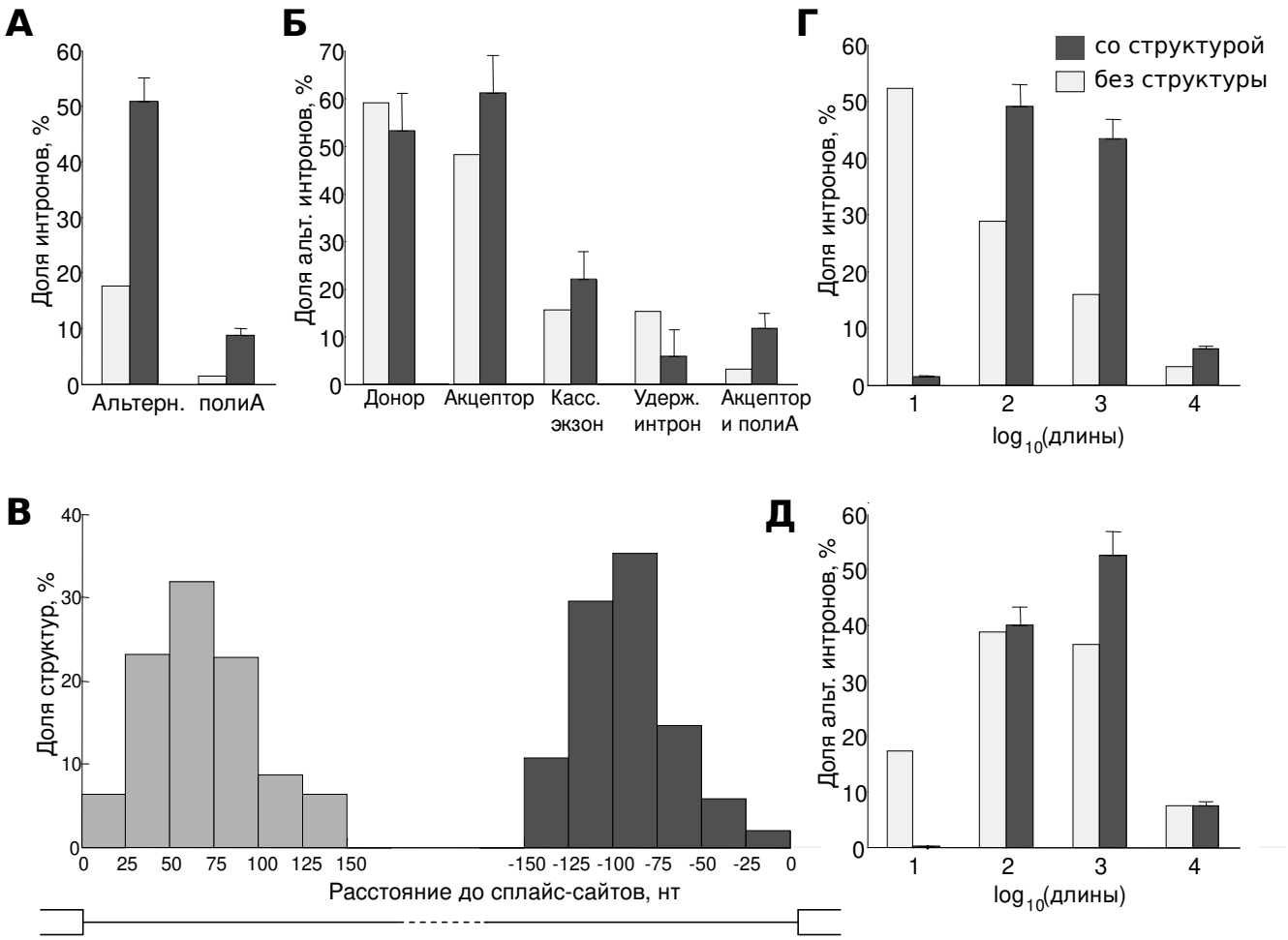


Рисунок 3.3 — Характеристики интронных структур РНК у насекомых. **(А)** Доля альтернативных интронов и интронов, содержащих события полиаденилирования (полиА) среди интронов с предсказанными структурами по сравнению со всеми интронами. Усики обозначают стандартную ошибку. **(Б)** Доля альтернативных интронов со структурой и без структуры в категориях: с альтернативным 5'ss, с альтернативным 3'ss, содержащие кассетные экзоны, удержанные интроны, с альтернативным 3'ss и внутренним сигналом поли(А). **(В)** Распределение позиций комплементарных участков относительно сайтов сплайсинга. **(Г)** Длины интронов со структурами по сравнению со всеми интронами. **(Д)** Длины интронов со структурами по сравнению со всеми альтернативными интронами.

существенной разницы в свободной энергии между структурами, расположенными в альтернативных и конститутивных интронах, не наблюдалось ($P = 0.2$).

Тенденции в расположении интронных структур РНК у млекопитающих (*H. sapiens*) в целом сходны с таковыми у дрозофил [108]. Если рассматривать пары комплементарных участков, расположенные в аннотированных интронах на расстоянии не более 150 нт в сторону 3'-конца от донорного сайта и не более 150 нт в сторону 5'-конца от акцепторного сайта, то в 33% случаев они попадают в альтернативно сплайсируемые интроны, в то время как в случайной выборке эта доля составляет лишь 10% ($P \simeq 0$). По отношению ко всем событиям сплайсинга почти все подтипы событий AC в интронах с предсказанными

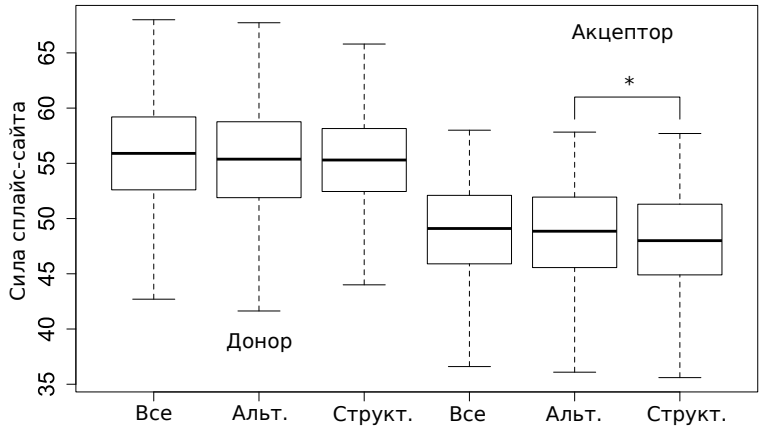


Рисунок 3.4 — Распределение сил сайтов сплайсинга (донорных и акцепторных сайтов) в интронах человека с предсказанными структурами РНК по сравнению с соответствующими распределениями сил для всех и альтернативных сайтов сплайсинга. Символ * обозначает статистически значимые различия на уровне значимости 5%.

структурами РНК наблюдались с более высокой частотой, чем в простой случайной выборке интронов того же объема, однако по сравнению с событиями AC только альтернативные акцепторные сайты являлись значимо перепредставленной категорией.

Во многих исследованиях сообщалось о влиянии структур РНК на выбор слабых сайтов сплайсинга [2]. Для того, чтобы ответить на вопрос о том, связаны ли предсказанные структуры РНК с сайтами сплайсинга, которые отличаются от других сайтов сплайсинга по своей силе, были вычислены позиционно-весовые матрицы, которые измеряют степень сходства между последовательностью сайта сплайсинга и консенсусом [108]. Оказалось, что акцепторные сайты, соседствующие с РНК-структурой, в среднем оказались даже слабее, чем альтернативные акцепторные сайты в целом (рис. 3.4). Таким образом, комплементарные интронные последовательности, влияющие на сплайсинг, преимущественно ассоциированы со слабыми альтернативными акцепторными сайтами сплайсинга.

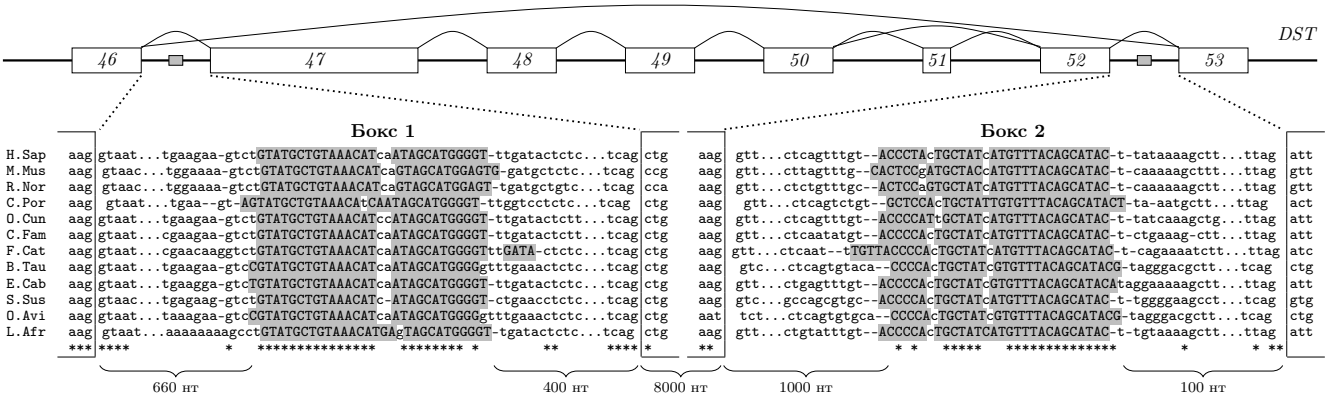


Рисунок 3.5 — Предсказанная структура РНК, регулирующая АС в гене *DST*. Верхняя панель: фрагмент человеческого гена *DST* (дистонин) длиной 9567 нт. Экзоны 47–52 либо одновременно включаются, либо одновременно исключаются. Две комплементарные последовательности (бокс 1 и бокс 2, серые прямоугольники) расположены в интронах между экзонами 46–47 и 52–53. Нижняя панель: фрагмент множественного выравнивания последовательностей интронов, содержащих боксы 1 и 2. Сокращения в названиях видов как в [373].

3.2.4 Пример внутримолекулярной структуры РНК в гене *DST*

Ген дистонина млекопитающих (*DST*, *BPAG1*, буллезный пемфигоидный антиген) принадлежит к семейству плакинов и участвует в развитии буллезного пемфигоида Левера — хронического воспалительного аутоиммунного заболевания кожи [384]. Известны несколько сплайс-изоформ этого гена, включая изоформы, кодирующие повторы плектинового типа (PTR) и повторы спектринового типа (STR) [385]. В частности, экзоны 47–52, кодирующие группу PTR, либо включаются в транскрипт, либо пропускаются, за исключением экзона 51, который может пропускаться независимо (рис. 3.5). Моделирование трехмерной структуры показало, что белок с гибридными типами повторов, который образуется в результате пропуска экзонов 47–52 может существенно отличаться по своим свойствам от белка, который образуется в случае, если эти экзоны включаются [386]. Кроме того, сплайсинг кластера экзонов 47–52 происходит по-разному в разных тканях и на разных стадиях развития. Представляется вероятным, что молекулярный механизм, управляющий включением этого кластера экзонов, с бóльшей вероятностью регулирует вырезание одного длинного интрона, чем вырезание семи более коротких последовательных интронов.

Согласно предсказаниям, интрон между экзонами 46–47 и интрон между экзонами 52–53 содержат пару ККУ, бокс 1 и бокс 2, которые обнаружива-

ются у большинства плацентарных млекопитающих (рис. 3.5). Учитывая, что комплементарные последовательности консервативны почти на 100%, а остальная часть интронной последовательности — нет, то логично предположить, что пропуск кластера опосредован структурой РНК, которая выпетливает экзоны 47–52, когда комплементарные последовательности спарены. Примечательно то, что эволюционная консервативность резко обрывается на границах комплементарных участков, что указывает на положительный отбор, действующий на эти последовательности, несмотря на разделяющее их расстояние около 10000 нт. Однако следует отметить, что отсутствие компенсаторных замен не доказывает и не опровергает наличия положительного отбора на поддержание комплементарности между ними.

3.2.5 Пример ложноположительного предсказания

Описанный метод можно применить не только для поиска внутримолекулярных комплементарных участков, но также и предсказания межмолекулярных РНК-РНК взаимодействий, например мишеней мякРНК. Хотя основная функция мякРНК заключается в проведении химических модификаций других РНК, некоторые из них или их фрагменты могут регулировать сплайсинг или трансляцию [387—389]. Так, мякРНК HBII-52 содержит последовательность, которая комплементарна мРНК гена серотонинового рецептора (HT2CR) и влияет на альтернативный сплайсинг в этом гене, ассоциированном с синдромом Прадера-Вилли [390]. Для поиска потенциальных мишеней мякРНК необходимо определить отношение $\mathcal{R}$ как $A \times B$, где A — набор сегментов мякРНК, а B — набор интронных сегментов пре-мРНК. Аналогично, для поиска потенциальных мишеней длинных некодирующих РНК в качестве множества A можно использовать их сегменты.

Действительно, число пар ККУ между сегментами мякРНК и мРНК в среднем на 20% выше, чем в соответствующем контрольном множестве, что позволяет предположить, что мякРНК способны комплементарно взаимодействовать с интронами белоккодирующих генов [113]. Аналогично, сегменты длинных некодирующих РНК имеют на 30% больше комплементарных мишеней в мРНК на кодирующей цепи по сравнению с противоположной цепью. Даже

после удаления аннотированных мякРНК из набора длинных некодирующих РНК, на кодирующей цепи мРНК оставалось на 21% больше комплементарных мишеней, чем на противоположной. При этом если выбирать непересекающиеся множества A и B из интронных сегментов кодирующих генов, то на кодирующей цепи находится только на 3–7% больше комплементарных мишеней, чем на некодирующей. Таким образом, можно предположить, что мякРНК и длинные некодирующие РНК в целом способны комплементарно взаимодействовать с мРНК и оказывать влияние на АС.

Однако следующий пример показывает, что это может быть не совсем так. А именно, длинная некодирующая РНК *RP11-439A17.4* содержит консервативную последовательность, комплементарную консервативным последовательностям на 3'-концах более чем 20 генов гистонов млекопитающих (рис. 3.6). При этом последовательности боксов 1 и 2 содержат сайт связывания транскрипционного фактора MEF-2A (миоцитарно-специфический энхансерный фактор 2A), а длинная некодирующая РНК *RP11-439A17.4* находится в антисмысловой ориентации по отношению к соседнему гену *HIST2H2BA*. Поэтому последовательность бокса 1 в *RP11-439A17.4*, вероятно, является регуляторным элементом транскрипции, который встречается на противоположных цепях ДНК, т.е., консервативная комплементарность в этом случае не отражает РНК-РНК взаимодействий.

Данное наблюдение показывает, что часть ложноположительных предсказаний внутримолекулярной структуры РНК может быть отнесена к мотивам, которые встречаются на противоположных цепях ДНК и находятся под эволюционным отбором по причинам, не связанным с комплементарностью. Более того, в них могут встречаться кажущиеся компенсаторные замены, если, например, сайт связывания транскрипционного фактора эволюционирует вместе с самим фактором, а на противоположных цепях это выглядит как поддержание вторичной структуры РНК. Такие консервативные элементы в последовательности ДНК в принципе неотличимы от консервативных комплементарных участков в РНК, что принципиально ограничивает предсказательные возможности всех основанных на сравнительной геномике методов.

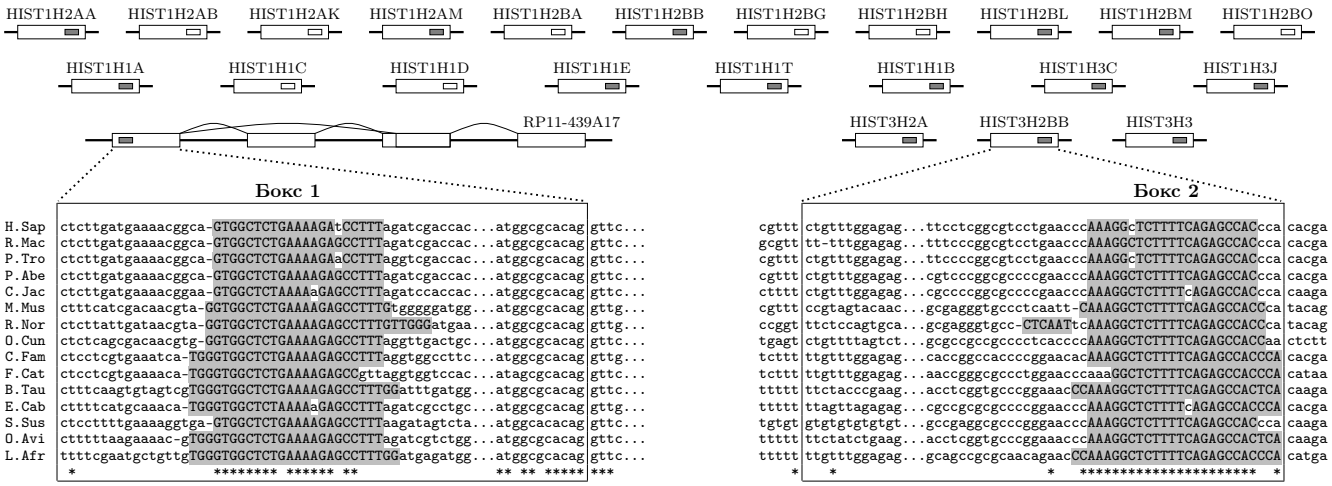


Рисунок 3.6 — Ложное предсказание комплементарного взаимодействия между длинной некодирующей РНК *RP11-439A17.4* и генами гистонов млекопитающих. Первый экзон *RP11-439A17.4* (слева внизу) содержит консервативную последовательность (бокс 1), которая комплементарна последовательности в гене *HIST3H2BB* (бокс 2), а также другим последовательностям в 3'-области как минимум 22 генов гистонов млекопитающих. Некоторые из этих последовательностей, а также бокс 1 содержат сайт связывания фактора MEF-2A, что указывает на случайную комплементарность транскрипционных регуляторных элементов, расположенных на противоположных цепях ДНК. Аннотированные (предсказанные) сайты связывания MEF-2A обозначены маленькими серыми (белыми) прямоугольниками.

3.3 Сначала выравнивание, потом фолдинг

Из результатов применения подхода «сначала фолдинг, потом выравнивание» (разд. 3.2) становится ясно, что комплементарные последовательности часто располагаются в консервативных интронных участках, которые могут быть заранее найдены в множественном выравнивании, и поэтому решать одновременно задачу множественного выравнивания и предсказания комплементарности не приходится [391]. Для реализации подхода «сначала выравнивание, потом фолдинг» был разработан метод PREPH (PREdiction of PanHandles) для предсказания консервативных комплементарных участков в геномах млекопитающих [23].

3.3.1 Описание и оценка производительности метода PREPH

Суть метода PREPH состоит в том, чтобы предсказывать гибридизацию между консервативными интронными участками белок-кодирующих генов, в которых нет отбора на аминокислотную последовательность, рассматривая все возможные комбинации таких участков, находящихся друг от друга на расстоянии не более заданного. Ограничений на псевдоузлы в нем нет, поскольку внутримолекулярная структура рассматривается как РНК-РНК взаимодействия между консервативными участками. Имея пару консервативных последовательностей и заранее сосчитав таблицу энергий гибридизации всех возможных пар k-меров (по умолчанию $k = 5$), можно эффективно предсказать комплементарные последовательности с небольшими внутренними петлями и небольшим числом G:U пар, не имеющие «рыхлых» коротких спиралей. Таким образом, метод заимствует некоторые идеи у IRBIS и может рассматриваться как разреженная реализация алгоритма Смита-Уотермана для локального выравнивания последовательностей [392] со специальными улучшениями, направленными на поиск комплементарности k-меров.

В PREPH целевая функция свободной энергии состоит из аддитивных энергетических вкладов лишь небольшого числа структурных элементов, а именно стекинга комплементарных пар оснований, коротких (не более 2 нт) выпячиваний и внутренних петель. На предварительном этапе вычисляются энергии гибридизации всех возможных пар идеально комплементарных k-меров с использованием таблиц энергий стекинга [312]. В полученную матрицу размера $4^k \times 4^k$, элементами которой являются энергии гибридизации пары k-меров, при помощи дополнительного вычисления можно добавить энергии гибридизации для структур с небольшим числом G:U пар, а значения энергий гибридизации остальных пар k-меров положить равными $+\infty$. Вычислительная эффективность метода достигается именно за счет этого шага, который выполняется только один раз и может быть использован для многих пар входных последовательностей с одним и тем же k.

На следующем этапе, получая на вход две последовательности длины n и m, соответственно, алгоритм заполняет разреженную матрицу размера $(n - k + 1) \times (m - k + 1)$, которая содержит энергии взаимодействия k-меров, расположенных в соответствующих местах последовательности, с помощью

найденной на предварительном этапе матрицы энергий. Затем реализуется алгоритм динамического программирования, в котором структура для пары позиций либо дополняется комплементарным спариванием, либо рассматриваются выпетливания не более двух нуклеотидов и внутренние петли размера не более чем 2×2 в каждой из последовательностей, или же структура начинается заново со спаривания двух комплементарных k-меров (всего 10 различных вариантов). На последнем, четвертом этапе выполняется обратный проход для идентификации непересекающихся структур. Две структуры называются непересекающимися, если комплементарные участки обеих последовательностей одной структуры не имеют общих нуклеотидов с комплементарными участками обеих последовательностей второй структуры, но при этом допускается наличие пересекающихся участков только в одной из двух последовательностей. Оптимальная структура соответствует минимальному значению в матрице динамического программирования, а также ей сопоставлен прямоугольник (путь) в матрице. Для получения субоптимальных структур алгоритм исключает из рассмотрения прямоугольник, соответствующий уже найденному оптимальному выравниванию, ищет новую минимальную энергию и проверяет, пересекает ли соответствующий ей прямоугольник оптимальное выравнивание. Если это не так, то структура сообщается и процесс повторяется до тех пор, пока (отрицательная) оптимальная энергия не станет выше заданного порога.

Разработанный метод был валидирован на тестовом наборе последовательностей, содержащих комплементарные участки в 50% длины, против других известных методов, таких как IntaRNA [393], RNAplex [383], и RIsearch [325]. Оценка доли p совпадающих пар оснований показала, что PREPH в целом верно определяет гибридизацию по сравнению с результатами IntaRNA ($p = 0.33 \pm 0.13$) и RNAplex (p=0.29 ± 0.12), но не с RIsearch ($p = 0$). Оказалось, что предсказания RIsearch, несмотря на высокую вычислительную эффективность этого метода, практически не пересекаются ни с результатами IntaRNA, ни с предсказаниями RNAplex, вследствие чего он были исключен из дальнейшего рассмотрения. То, что PREPH находит не все спаривания объясняется заложенной в нем моделью, а именно тем, что предсказываются лишь длинные, почти идеально комплементарные спаривания. С другой стороны, время работы PREPH на тестовом наборе последовательностей составляло 2.3 сек против 76 сек для RNAplex и свыше 90 сек для RNAcofold и IntaRNA. Таким образом, PREPH превосходит существующие методы по вычислительной эффективно-

сти и предсказывает оптимальную структуру РНК в последовательностях, содержащих длинные комплементарные участки, в целом не противоречащую предсказаниям более сложных, но медленных программ.

3.3.2 Консервативные комплементарные участки (ККУ)

В отличие от предыдущего метода, PREPH получает на вход не парные, а построенные заранее множественные полногеномные выравнивания, из которых вырезаются блоки, соответствующие заранее известному множеству консервативных фрагментов. Чтобы получить это множество, в геноме человека рассматривались участки нуклеотидных последовательностей белок-кодирующих генов за исключением конститутивных и альтернативных экзонов, повторов, некодирующих генов, находящихся в интронах, и других областей, находящихся под не связанным со структурой РНК отбором (рис. 3.7А). Полученные интронные фрагменты расширялись на 10 нт во фланкирующие экзоны для того, чтобы находить комплементарные участки, перекрывающиеся с сайтами сплайсинга. Полученный набор интронных фрагментов был пересечен с треком `phastConsElements` геномного браузера Калифорнийского университета в Санта-Крус (UCSC) [394], который содержит координаты высококонсервативных участков в полногеномных выравниваниях 100 позвоночных (основные виды *P. troglodytes*, *M. musculus*, *S. scrofa*, *G. Gallus*, *X. tropicalis*, *D. rerio*; самые короткие и самые длинные филогенетические расстояния 0.01 и 2.40, соответственно). В результате был получен набор из 1931116 коротких фрагментов со средней длиной 17 нт, которые далее будут называться консервативными интронными фрагментами (КИФ).

Пары консервативных комплементарных участков (ККУ) были идентифицированы во всех парных комбинациях КИФ, расположенных в не более чем L нт друг от друга и принадлежащих одному и тому же гену, с использованием описанного метода PREPH (рис. 3.7Б). При поиске последовательностей длиной не менее 10 нт со свободной энергией гибридизации $\Delta G \leqslant -15$ ккал/моль, минимальной длиной спирали $k = 5$ и предельным расстоянием $L = 10000$ между КИФ были найдены 916360 пар ККУ, в среднем 75 пар ККУ на ген и не более 295 пар ККУ в 95% генов. Средняя свободная энергия гибридизации

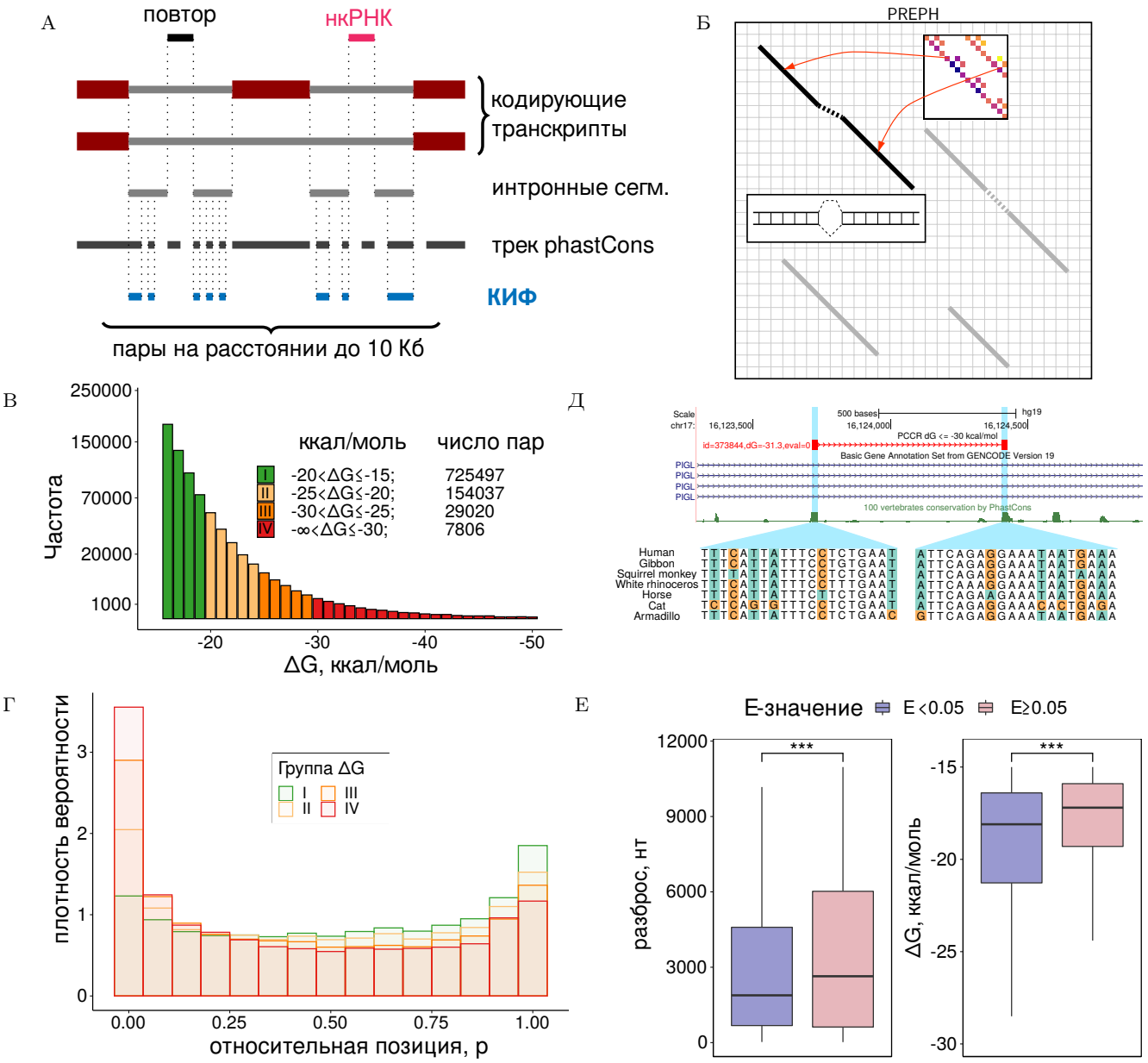


Рисунок 3.7 — Характеристики консервативных комплементарных участков (ККУ). **(А)** Пары ККУ разыскиваются в консервативных интронных фрагментах (КИФ), которые находятся на расстоянии не более 10000 нт друг от друга. **(Б)** Принцип работы метода PREPH, использующего предварительно вычисленные энергии спаривания для всех k-меров (вставка). **(В)** Распределение энергий пар ККУ состоит из четырех энергетических групп (I–IV). **(Г)** Распределение относительного положения (p) пар ККУ в гене. **(Д)** Независимые компенсаторные замены поддерживают дальние взаимодействия в структуре РНК у гена *PIGL*. **(Е)** Пары ККУ со значимыми компенсаторными заменами ($E < 0.05$, $n = 3204$) имеют в среднем меньший разброс и более стабильны, чем пары ККУ с незначимыми нуклеотидными ковариациями ($E \geqslant 0.05$, $n = 905942$); символ *** обозначает статистически значимые различия на уровне значимости 0.1%.

(ΔG) и средняя длина ККУ составляли -17.2 ккал/моль и 13 нт соответственно, причем распределение частот ожидаемо убывало при увеличении длины и стабильности структуры. Более длинные структуры имели более высокие абсолютные значения ΔG; однако энергия и длина были не прямо пропорциональны, а соотношение между ними зависело от GC состава. В дальнейшем пары ККУ подразделяются на четыре группы по энергии: группа I от -15 до -20 ккал/моль, группа II от -20 до -25 ккал/моль, группа III от -25 до -30 ккал/моль и группа IV ниже -30 ккал/моль, которые представлены единой цветовой схемой[1] (рис. 3.7В). Примечательно, что регулирующие сплайсинг дальние взаимодействия в интронной структуре РНК гена *PLP1/DM20*, а также РНК-мост в гене *ENAH* были отнесены к группе I ($\Delta G = -15.8$ и -19.4 ккал/моль, соответственно), что указывает на то, что менее стабильные структуры не менее функциональны или менее интересны, чем другие.

Расстояние между двумя ККУ в паре, называемое разбросом, имеет частотное распределение, которое достигает максимума на коротких расстояниях и убывает до некоторого предельного значения, совпадающего с предельным значением для попарных расстояний между КИФ. Для оценки относительного положения пар ККУ внутри гена использовалась метрика p, которая изменяется от 0% для ККУ, расположенных в 5'-конце гена, до 100% для ККУ, расположенных в 3'-конце гена. Распределение p также неравномерно и имеет две выраженные моды на 5'- и 3'-конце (рис. 3.7Г), что, очевидно, объясняется большей плотностью КИФ в начале и в конце гена.

3.3.3 Эволюционные подписи в ККУ

Для исследования компенсаторных мутаций к фрагментам множественных выравниваний, которые вырезаются ККУ из множественного выравнивания геномов ста позвоночных, применялась программа R-scape, которая оценивает независимое возникновение комплементарных замен на разных ветвях филогенетического дерева [12]. Отклонение от нулевой гипотезы о том, что парные ковариации в паре ККУ не обусловлены отбором на структуру

[1]Цветовую схему можно условно назвать шкалой зрелости яблок: от зеленого (совсем незрелые) до красного (уже созрели).

РНК, оценивалось как произведение E-значений, сообщаемых R-scape для всех спаренных оснований в структуре с использованием поправки Бенджамини-Хохберга. Из 909146 пар ККУ, для которых вычисление E-значений было возможно, только 3204 имели E-значение менее 5%, причем они оказались в среднем более стабильными и имеющими меньший разброс, чем остальные (рис. 3.7Е). В некоторых случаях структурное выравнивание поддерживалось ковариациями, как, например, выскоэнергетическое ($\Delta G = -31$ ккал/моль) взаимодействие, охватывающее 700 нт в первом интроне гена *PIGL* (рис. 3.7Д). Однако E-значения многих известных из литературы структур были близки к единице, что еще раз подтверждает то, что вариабельность нуклеотидных последовательностей ККУ в большинстве случаев недостаточна для оценки статистической значимости по компенсаторным заменам.

3.3.4 Оценка доли ложноположительных предсказаний

Нуклеотиды, принадлежащие ККУ, теоретически должны находиться в спаренном состоянии. Склонность отдельных нуклеотидов к спариванию можно оценить путем измерения показателя реактивности с помощью метода icSHAPE [395]. Оказалось, что средняя реактивность нуклеотидов в составе ККУ значительно ниже по сравнению с реактивностью нуклеотидов в контрольном наборе расположенных рядом с ККУ участков ($P < 10^{-60}$), а разность реактивностей ожидаемо возрастает по абсолютной величине с увеличением свободной энергии структуры (рис. 3.8А).

Поскольку показатели реактивности icSHAPE можно было вычислить только для 4551 пар ККУ, что составляет 0.5% от полного набора ККУ, представляется логичным оценить поддержку ККУ в других наборах экспериментальных данных. А именно, были исследованы взаимодействующие пары оснований в экспериментальных данных псораленового анализа взаимодействий и структур РНК (PARIS) [102] ($n = 15036$), лигирования взаимодействующих РНК с последующим высокопроизводительным секвенированием (LIGR-seq) [103] ($n = 551926$) и конформационного секвенирования

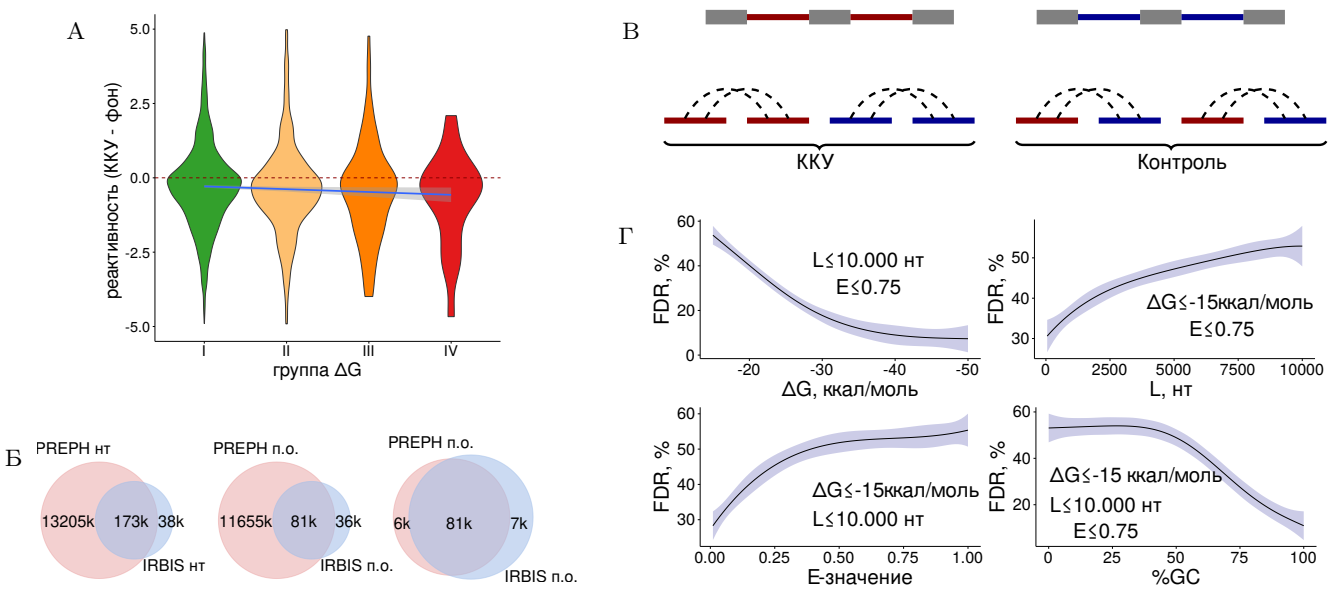


Рисунок 3.8 — Поддержка ККУ данными высокопроизводительного секвенирования. **(А)** Разница между реактивностью нуклеотидов (метод icSHAPE) внутри ККУ и средней реактивностью нуклеотидов в участках, соседних с ККУ, в энергетических группах I–IV (цвета, как на рис. 3.7Б). Линейная регрессия $\Delta\text{реактивность} = \beta_0 + \beta_1 \Delta G$ представлена наклонной линией; $\widehat{\beta}_1 = -0.03 \pm 0.01$. **(Б)** Диаграмма Венна для количества общих нуклеотидов (слева), количества общих пар оснований (в центре) и количества общих пар оснований среди общих нуклеотидов (справа) для предсказаний PREPH и IRBIS. **(В)** Оценка частоты ложноположительных результатов (FDR) путем пересоединения, то есть создания контрольного набора, состоящего из химерных КИФ, выбранных из разных генов. **(Г)** FDR как функция от энергии ΔG (вверху слева), максимального расстояния между КИФ (вверху справа), E-значения (внизу слева) и GC состава (внизу справа). Сплошные линии изображают среднее значение по $n = 16$ рандомизациям; заштрихованные области обозначают 95% доверительные интервалы, полученные с помощью локально оцененной регрессии сглаживания диаграмм рассеяния (LOESS).

РНК *in situ*[2] (RIC-seq) [21] ($n = 501144$). Если рассматривать внутримолекулярные взаимодействия внутри КИФ на расстоянии не более 10000 нт друг от друга, а также ограничиться множеством пар ККУ, которые пересекаются с экспериментальными данными, то наилучшие значения точности (P, precision) и полноты (R, recall), а также наилучшее значение условной вероятности π правильного предсказания взаимодействующего партнера ККУ при условии, что второй ККУ предсказан правильно, достигалось по сравнению с методом RIC-seq (табл. 2). Также было обнаружено, что свободная энергия пар ККУ, поддерживаемых RIC-seq и PARIS, в среднем на 1.2 ккал/моль ниже, чем у пар ККУ без экспериментальной поддержки ($P < 10^{-19}$).

[2]В 2020 году эти данные были любезно предоставлены проф. Юаньчао Сюэ и проф. Чанчан Цао в виде списка кластеров химерных чтений для внутримолекулярных взаимодействий. К более детальному анализу этих данных мы вернемся в разд. 4.1.

	RIC-seq, $n = 1{,}804$				LIGR-seq, $n = 586$				PARIS, $n = 907$			
ΔG	PREPH	$P, \%$	$R, \%$	$\pi, \%$	PREPH	$P, \%$	$R, \%$	$\pi, \%$	PREPH	$P, \%$	$R, \%$	$\pi, \%$
-15	1611	42	53	43	362	54	49	56	5901	50	60	51
-20	364	66	26	67	88	57	17	58	1725	60	33	60
-25	84	86	9	86	21	48	3	48	545	65	14	65
-30	23	91	3	91	5	40	1	40	160	72	6	73

Таблица 2 — Точность (P) и полнота (R) по сравнению с RIC-seq, LIGR-seq, и PARIS при различных ограничениях на ΔG. π — условная вероятность правильного предсказания взаимодействующего партнера ККУ при условии, что другой ККУ в паре предсказан правильно. В столбце «PREPH» показано количество пар ККУ, удовлетворяющих критериям сравнения; n обозначает количество структур в каждом эксперименте.

Затем набор предсказанных пар ККУ был сопоставлен со списком ККУ, полученным ранее при помощи метода IRBIS (разд. 3.2). Пересечение предсказаний двух методов было значительно бóльшим, чем предсказания только одного IRBIS как с точки зрения числа нуклеотидов, так и числа спаренных оснований (рис. 3.8Б). Иными словами, PREPH значительно расширяет предсказательные возможности по сравнению с IRBIS. С другой стороны, свободная энергия пар ККУ, предсказанных IRBIS, была в среднем на 2.7 ккал/моль ниже, чем у пар ККУ, предсказанных только PREPH ($P < 10^{-30}$), что, вероятно связано с тем, что PREPH допускает короткие внутренние петли и боковые выпетливания. Небольшая часть предсказаний IRBIS отсутствует в списке предсказаний PREPH из-за ограничения на консервативность, но без этого ограничения он был бы нереалистичным с вычислительной точки зрения.

Как было показано в разд. 3.2.2, долю ложноположительных предсказаний можно оценить используя логику «пересоединения» и применяя тот же метод, который использовался для предсказания пар ККУ, к контрольному набору последовательностей, которые не образуют комплементарных спариваний. Для этого PREPH был применен к химерным последовательностям КИФ, которые были выбраны случайным образом из разных генов, но при этом имели такую же длину и динуклеотидный состав, как и исходные КИФ (рис. 3.8В). Доля ложноположительных предсказаний, определяемая как число предсказаний для контрольного множества по отношению к числу настоящих предсказаний, зависит от энергии структуры, разброса, GC состава и E-значения, а также варьируется от 10% в самых строгих до более 50% в наиболее расслабленных условиях (рис. 3.8Г). Как и ожидалось, она падает с увеличением энергии, GC состава, а также с уменьшением разброса и E-значения.

3.3.5 Полиморфизмы и компенсаторные мутации в ККУ

Мутации обычно снижают стабильность вторичной структуры РНК, а некоторые из них находятся под эволюционным отбором из-за их влияния на термодинамическую стабильность пре-мРНК [396]. Оценка влияния полиморфизмов в человеческой популяции из проекта 1000 Геномов [397] на ККУ показывает, что полиморфизмы зародышевой линии значимо недопредставлены в ККУ ($P < 0.01$). Затем для каждого полиморфизма, возникающего в ККУ, было рассчитано вызванное им изменение свободной энергии в сравнении с изменением, которое наблюдалось бы, если бы та же самая мутация произошла в другой позиции того же ККУ. Оказалось, что полиморфизмы дестабилизируют структуру РНК меньше, чем можно было бы ожидать по случайным причинам ($P < 0.001$), что позволяет заключить, что они имеют тенденцию минимизировать свое дестабилизирующее воздействие на структуру РНК.

Компенсаторные мутации могут возникать в человеческой популяции, но их частота очень мала. Для оценки частоты компенсаторных замен были отобраны ККУ, содержащие полиморфизмы с аллельной частотой 1% и выше. Из 64074 пар оснований в ККУ, в которые попали полиморфизмы, только в 12 были обнаружены компенсаторные замены, но в контрольном наборе химерных ККУ в среднем наблюдалось только 0.06 ± 0.02 случаев. Плотность компенсаторных замен, нормированная на количество пар оснований с мутациями, составила 0.019% для предсказанных спариваний и $(9.0 \pm 7.3) \cdot 10^{-5}$% для рандомизированных наборов, из чего можно сделать вывод о том, что компенсаторные мутации в ККУ крайне редки в человеческой популяции, но тем не менее демонстрируют некоторое обогащение в ККУ.

3.3.6 Взаимосвязь между ККУ и цис-регуляторными элементами АС

Как отмечалось ранее, взаимодействующие на больших расстояниях комплементарные участки расположены неслучайно относительно сигналов сплайсинга. Для повторного исследования этого вопроса применительно к

парам ККУ, предсказанным PREPH, можно использовать следующую классификацию (рис. 3.9А). Если пара ККУ пересекает интрон, то она может быть расположена либо полностью внутри интрона («внутри»), либо интрон может быть полностью расположен внутри ККУ («снаружи»), либо два интервала могут пересекаться («узел»). Предпочтение этих категорий измеряется метрикой обогащения (FC), определяемой как количество пар ККУ в данной категории по отношению к числу некоторых контрольных интервалов ($FC = 1$ в отсутствие перепредставленности и недопредставленности).

Выбирать контрольные интервалы можно несколькими способами. В первом способе, называемом «случайным сдвигом», каждая пара ККУ случайным образом смещается внутри гена. Полученный таким образом контрольный интервал имеет такой же разброс, что и исходный, принадлежит тому же гену, но расположен в другой его части. Во втором способе, называемом «случайным геном», для каждой пары ККУ случайно выбирается ген той же длины, что и исходный, и в нем выбирается интервал с таким же разбросом и относительным расположением относительно границ гена. Полученный контрольный интервал имеет такой же разброс, что и исходный, расположен в том же месте, что и исходная пара ККУ, но в другом гене.

По сравнению со случайными сдвигами, пары предсказанных ККУ демонстрируют перепредставленность в категории «внутри» и недопредставленность в категориях «снаружи» и «узел» (рис. 3.9Б). Такие же закономерности обнаруживаются при выборе случайного гена, что исключает возможность того, что преимущественное расположение пар ККУ внутри интронов происходит из-за неравномерного распределения более длинных интронов вдоль гена (рис. 3.7Г). Аналогично, сравнение с набором аннотированных экзонов выявило недопредставленность пар ККУ, выпетливающих экзоны, которая усиливается по мере увеличения стабильности структуры (рис. 3.9В). Средний уровень включения выпетливаемых экзонов в клеточной линии HepG2 ниже, чем у экзонов, не окруженных ККУ, причем это различие также усиливается по мере увеличения энергии структуры (рис. 3.9Г). Эти наблюдения еще раз подтверждают то, что большинство включаемых в мРНК экзонов избегает выпетливания структурой РНК. Представленность структур РНК снижается с увеличением расстояния до границы интрона во всех четырех энергетических группах, однако менее стабильные структуры чаще встречаются в окнах, прилегающих к сайтам сплай-

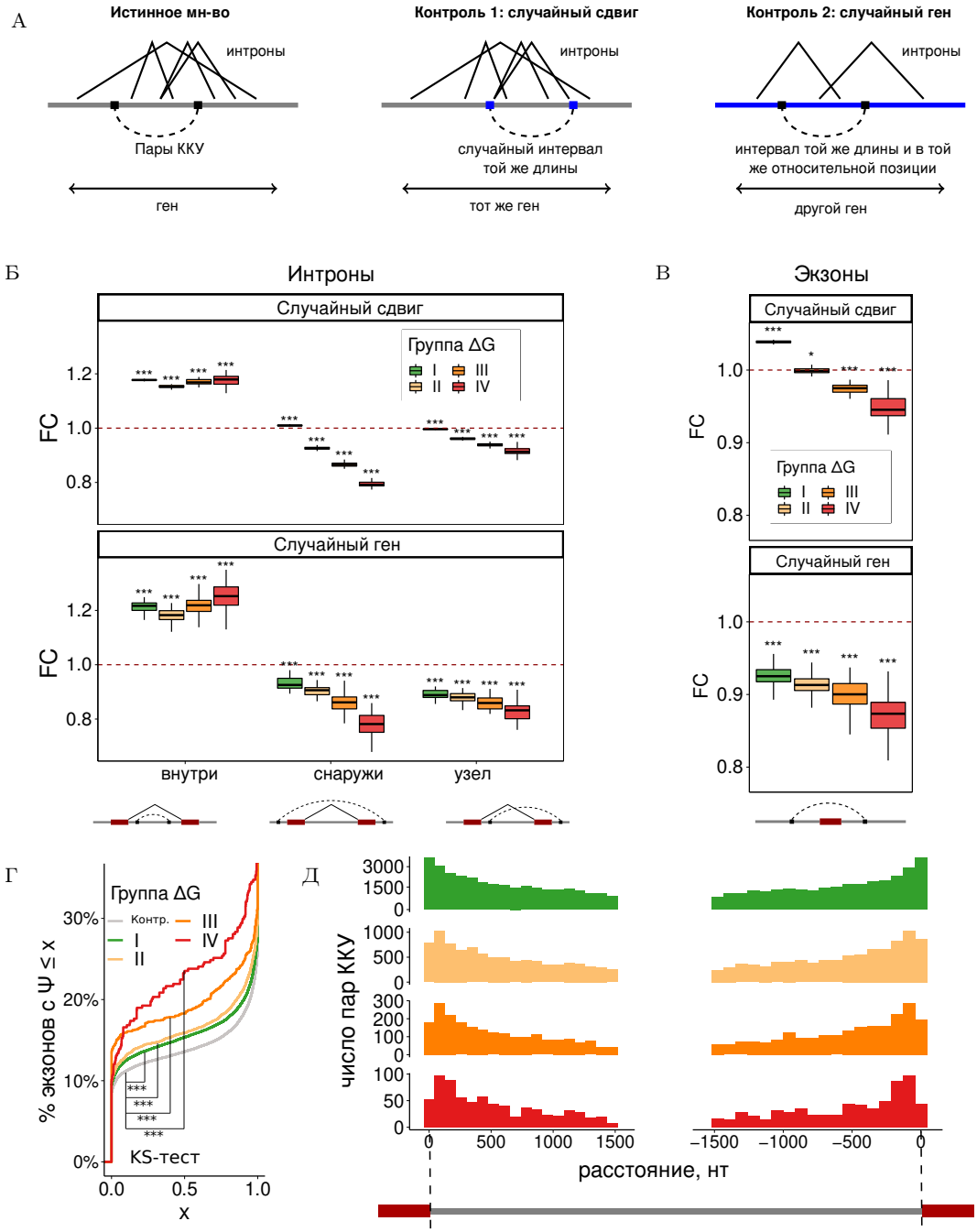


Рисунок 3.9 — ККУ и сплайсинг. **(А)** Процедуры контроля. «Случайный сдвиг» – пара ККУ случайно сдвигается внутри гена. «Случайный ген» – в другом гене на той же относительной позиции выбирается интервал такой же как и пара ККУ. **(Б)** Пары ККУ перепредставлены в категории «внутри» (см. идеограмму внизу) и недопредставлены в категориях «снаружи» и «узел». **(В)** Пары ККУ, выпетливающие экзоны, недопредставлены. **(Г)** Кумулятивное распределение среднего уровня включения экзона (Ψ) для выпетливаемых экзонов по сравнению с остальными экзонами (Контр.). **(Д)** Распределение расстояний от интронных ККУ до границ интронов. На всех панелях ящичковые диаграммы соответствуют $n = 40$ рандомизациям; символы * и *** обозначают статистически значимые различия на уровне значимости 5% и 0.1%, соответственно ($H_0 : FC = 1$).

синга, тогда как более стабильные структуры предпочитают занимать более удаленные позиции (рис. 3.9Д).

Для того, чтобы выяснить, как пары ККУ расположены по отношению к сплайс-сайтам, были найдены интронные мотивы, похожие на консенсусные последовательности 5'ss и 3'ss (криптические сайты сплайсинга), а также идентифицированы сайты сплайсинга в экспрессируемых транскриптах по данным секвенирования РНК (РНК-сек) из проекта Genotype Tissue Expression Project (GTEx) [356]. Оказалось, что ККУ, пересекающиеся с активными (т.е., принадлежащими высокоэкспрессирущимся транскриптам) сайтами сплайсинга, недопредставлены, а пересекающиеся с неактивными сайтами сплайсинга — перепредставлены (рис. 3.10А). С другой стороны, ККУ, пересекающиеся с криптическими сайтами сплайсинга, также перепредставлены во всех энергетических группах за исключением высокостабильных структур, которые, вероятно, не содержат сплайс-сайтов из-за высокого GC состава. Вторичная структура РНК также задействована в так называемом обратном сплайсинге (back-splicing), который приводит к образованию кольцевых РНК [398]. Сравнение с базой данных тканеспецифичных кольцевых РНК [399] выявило картину, противоположную таковой для обычных интронов (рис. 3.10Б), а именно перепредставленность в категории «снаружи» и недопредставленность в категориях «внутри», что подтверждает гипотезу о том, что образованию кольцевых РНК способствуют выпетливания, образуемые стабильными двухцепочечными участками [398].

Широко распространенной формой посттранскрипционной модификации РНК является ферментативное превращение аденозиновых нуклеозидов в инозин, процесс, называемый $A{\to}I$ редактированием, который катализируется аденозиндезаминазами из семейства белков ADAR [400; 401]. Поскольку субстратами ADAR являются двухцепочечные РНК, представляется интересным оценить, не обогащены ли ККУ сайтами редактирования РНК (рис. 3.10В). По мере увеличения свободной энергии структуры наблюдается увеличивающееся обогащение сайтами редактирования РНК, аннотированными в базе данных RADAR [402], в ККУ по сравнению с КИФ, не участвующими в дальних комплементарных взаимодействиях (отношение шансов $OR = 2.1 \pm 0.2$). Следует отметить, что для пар ККУ с $\Delta G \leqslant -30$ ккал/моль отношение шансов значительно увеличивается до 2.1 ± 0.2, что подтверждает предыдущие наблюдения о том, что мишенями ADAR являются именно длинные двухцепочечные РНК.

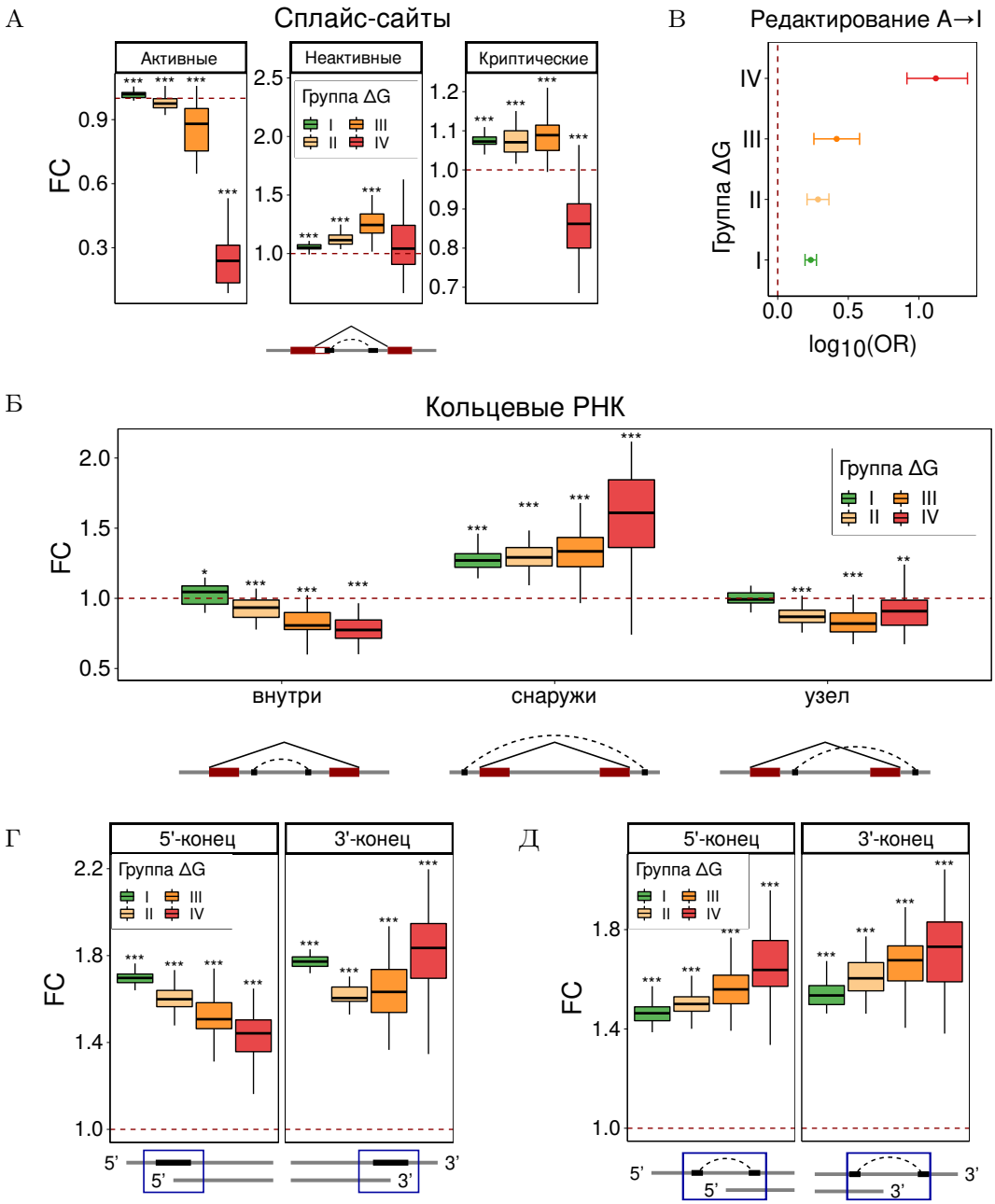


Рисунок 3.10 — ККУ, сплайсинг, редактирование РНК и концевой процессинг. **(А)** ККУ недопредставлены в активно экспрессируемых сайтах сплайсинга и перепредставлены в неактивных и криптических сайтах сплайсинга. **(Б)** Пары ККУ предпочитают находиться в положении «снаружи» по отношению обратно сплайсируемым интронам и избегают положения «внутри» и «узел». **(В)** ККУ обогащены сайтами редактирования РНК. OR — отношение шансов. Усиками показаны 95% доверительные интервалы. **(Г)** ККУ обогащены 5'- и 3'-концами аннотированных транскриптов, включая все аберрантные и неполные транскрипты, т.е., концы транскриптов часто встречаются в двухцепочечных частях ККУ. **(Д)** Пары ККУ также обогащены 5'- и 3'-концами транскриптов, т.е., аннотированные концы транскриптов часто встречаются в петле между ККУ. На всех панелях ящичковые диаграммы соответствуют $n = 40$ рандомизациям; символы *, ** и *** обозначают статистически значимые различия на уровне значимости 5%, 1% и 01% соответственно ($H_0 : FC = 1$).

Эти результаты согласуются друг с другом для сайтов из баз данных RADAR и REDIportal [402; 403] и могут рассматриваться как дополнительное подтверждение двухцепочечности предсказанных ККУ.

3.3.7 Взаимосвязь между ККУ и концевым процессингом РНК

Структура РНК играет важную роль в процессинге 5'- и 3'-концов мРНК человека [404], а конкурирующие структуры РНК могут участвовать в регуляции АС и альтернативного полиаденилирования в 3'-НТО [405]. В этой связи интересно охарактеризовать взаимосвязь между парами ККУ и 5'- и 3'-концами мРНК, используя аннотации транскриптов, сайтов полиаденилирования и каталог кэпированных мРНК (CAGE) [406; 407].

Во-первых, число аннотированных транскриптов, которые начинаются или заканчиваются внутри ККУ, включая аберрантные изоформы из базы данных GENCODE, значительно превышает ожидаемое (рис. 3.10Г), что позволяет предположить, что двухцепочечные структуры могут участвовать в подавлении экспрессии аберрантных транскриптов. Во-вторых, аннотированные 5'- и 3'-концы транскриптов значительно перепредставлены в промежутках между ККУ, причем этот эффект не обусловлен неравномерным распределением пар ККУ по длине гена (рис. 3.10Д).

Далее, зададимся вопросом, встречаются ли аннотированные концы транскриптов, метки CAGE и кластеры поли(А)-сайтов чаще в интронах, содержащих пары ККУ, чем в других интронах. Для этого все интроны были разделены на две группы, интроны с ККУ и интроны без ККУ, и из каждой группы были сделаны выборки интронов, сопоставимых по длине (38119 интронов в каждой). Около 33.5% интронов с ККУ содержали по крайней мере один 3'-конец транскрипта, тогда как среди интронов без ККУ эта доля составляла 25.1% ($OR = 1.50 \pm 0.05$). Аналогично, 31.6% интронов с ККУ содержали по крайней мере один 5'-конец транскрипта по сравнению с 23.6% для интронов без ККУ ($OR = 1.50 \pm 0.05$). Полученные результаты свидетельствуют о том, что дальние взаимодействия в структуре РНК и сплайсинг нетривиальным образом связаны с концевым процессингом эукариотических транскриптов. Более

развернутый ответ на вопросы, связанные с 3'-процессингом и сплайсингом дает следующий раздел.

3.3.8 Интронное полиаденилирование и сплайсинг

Отвлекаясь от структуры РНК, обратимся к данным высокопроизводительного секвенирования из консорциума GTEx для исследования связи между интронным полиаденилированием и сплайсингом. Огромное количество данных, полученных в рамках этого проекта, позволяет идентифицировать сайты интронного полиаденилирования при помощи чтений, содержащих фрагменты поли(А)-хвоста мРНК, и сопоставить тканеспецифическое интронное полиаденилирование с тканеспецифическим сплайсингом [408]. С помощью таких чтений, называемых поли(А)-чтениями, был получен каталог, состоящий из 164497 кластеров сайтов полиаденилирования (КСПА), и исследована их принадлежность к различным частям белоккодирующих генов, а именно к 5'-НТО, 3'-НТО и кодирующей области. Количественная оценка КСПА может проводиться не только по абсолютному числу, но и по плотности, т.е. числу КСПА на нуклеотид, а также может принимать во внимание уровень поддержки поли(А)-чтениями.

Как и ожидалось, наибольшая плотность КСПА наблюдалась в 3'-НТО, причем обогащение было более заметным, если принять во внимание количество поддерживающих поли(А)-чтений (рис. 3.11А). Плотность КСПА в интронах меньше, чем в других участках, однако по абсолютному количеству они все же встречались довольно часто (рис. 3.11Б). Однако следует учесть то, что интронные сигналы недопредставлены в данных секвенирования РНК из-за того, что интроны деградируют после сплайсинга. Для того, чтобы оценить частоту интронного полиаденилирования, принимая во внимание эту систематическую ошибку, число поли(А)-чтений было нормализовано на среднее покрытие чтениями в экзонах и интронах. С учетом этого плотность поли(А)-чтений в интронах оказалась существенно выше, чем в экзонах (рис. 3.11В). Кроме того, если сопоставить интроны, конститутивные и альтернативные экзоны по уровню покрытия и выбрать подмножество интервалов каждого типа, имеющих приблизительно одинаковое покрытие, то окажется, что поли(А)-чтения

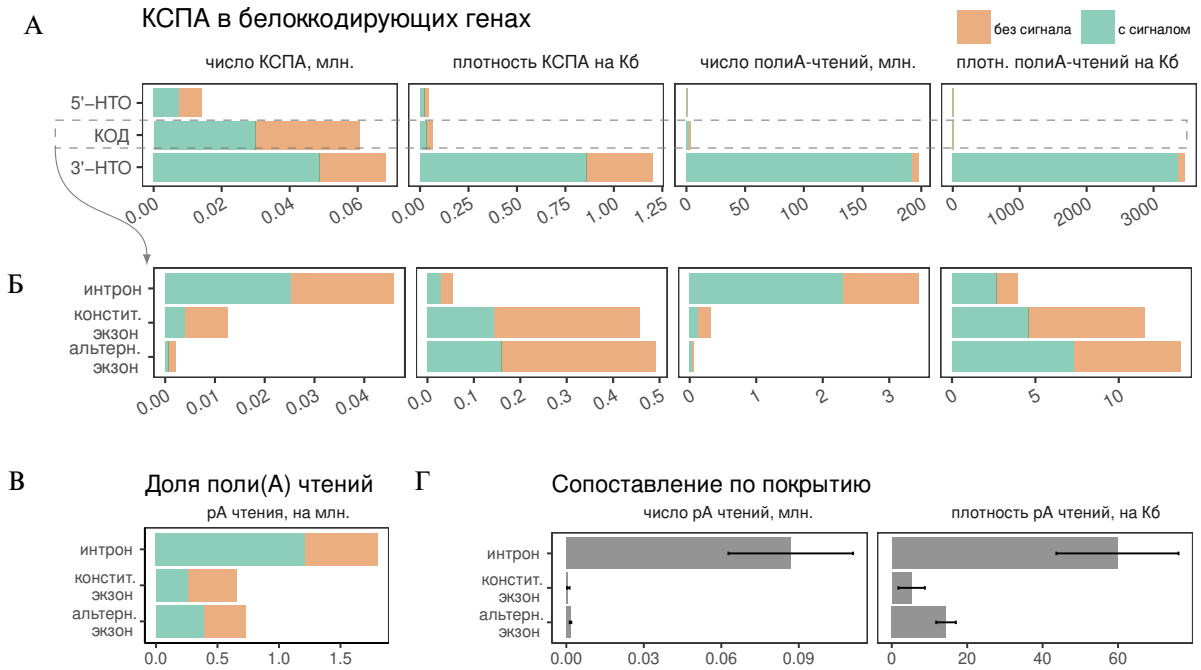


Рисунок 3.11 — Кластеры сайтов полиаденилирования (КСПА) в белоккодирующих генах. **(А)** Распределение числа КСПА в 5'-НТО, кодирующей области (КОД) и 3'-НТО. Показаны общее количество КСПА, плотность КСПА на килобазу (Кб), число поли(А)-чтений и плотность поли(А)-чтений на Кб. Цвета обозначают КСПА с консенсусным сигналом полиаденилирования (AAAUAA) и без него. **(Б)** Распределение числа и плотности КСПА из КОД в интронах, конститутивных экзонах и альтернативных экзонах. **(В)** Число поли(А)-чтений, нормализованное на среднее покрытие чтениями в соответствующей области (число поли(А)-чтений на миллион выровненных чтений). **(Г)** Число поли(А)-чтений в участках, имеющих одинаковое покрытие чтениями.

также обогащены интронах (рис. 3.11Г). В целом эти наблюдения указывает на то, что если бы интроны и экзоны были одинаково представлены в данных РНК-секвенирования, то наблюдаемая частота событий полиаденилирования в интронах оказалась бы в несколько раз большей, чем в экзонах. Данное наблюдение представляется удивительным, поскольку такие события потенциально должны приводить к преждевременной терминации транскрипции почти в каждом гене, но в действительности ее не происходит.

Альтернативные терминальные экзоны можно разделить на два класса: пропускаемые терминальные экзоны (skipped terminal exons, STE), которые могут использоваться в качестве терминальных экзонов или исключаться, и составные терминальные экзоны (composite terminal exons, CTE), которые возникают в результате удержания интрона (рис. 3.12А) [409]. Чтобы различать эти две возможности, рассматривался перепад покрытия чтениями в двух окнах, we_1 и we_2, на границе экзона, а также перепад покрытия чтениями в окнах wi_1 и wi_2 в окрестности самого КСПА (рис. 3.12А). STE должны характеризо-

ваться большим значением we_1/we_2, тогда как CTE должны характеризоваться небольшим соотношением we_1/we_2. Для количественной характеризации сплайсинга, была использована модифицированная метрика $\Psi = a/(a+b+c)$, где a — число разрывных чтений, начинающихся в 5'ss и заканчивающихся перед КСПА, b — число разрывных чтений, начинающихся в 5'ss и заканчивающихся в 3'ss, а c — число непрерывных чтений, выравнивающихся на экзон-интронную границу (рис. 3.12А). Значение $\Psi = 1$ указывает на преобладание канонического сплайсинга, соединяющего аннотированные сплайс-сайты, а $\Psi = 0$ указывает на наличие события AC перед КСПА. STE, так и CTE должны иметь $\Psi = 0$ из-за отсутствия у них канонического сплайсинга, с преобладанием b в случае STE и преобладанием c в случае CTE.

Как и ожидалось, между Ψ и числом поддерживающих КСПА поли(А)-чтений имеется отрицательная взаимосвязь (рис. 3.12Б). Она также проявляется в отрицательном сдвиге в распределении коэффициентов корреляции Пирсона между Ψ и поддержкой поли(А)-чтениями, вычисленных по тканям, по сравнению с фоновым распределением, в котором метки тканей были случайно перемешаны (рис. 3.12В, слева). В некоторых случаях перепад покрытия чтениями изменялся на два порядка, когда значение Ψ увеличилось с 25% до 100% (рис. 3.12В, справа). Эти наблюдения еще раз подтверждают, что сплайсинг и полиаденилирование естественным образом противодействуют друг другу. При рассмотрении случаев со значительным перепадом покрытия в КСПА ($wi_1/wi_2 > 10$) и с высоким покрытием чтениями в начале интрона ($wi_1 > 0.1we_1$) двумерные распределения $\log(we_1)$ и $\log(we_2)$ для 1136 аннотированных CTE и 1948 аннотированных STE разделялись прямой $we_2 = 0.25we_1$, причем первые ожидаемо группировались выше, вторые — ниже прямой, а другие КСПА образовали смесь двух распределений (рис. 3.12Г). В то время, как распределения Ψ для CTE и STE характеризовались одной модой при $\Psi \simeq 0$, что указывает на отсутствие у них канонического сплайсинга, распределения Ψ у остальных КСПА имели выраженную вторую моду при $\Psi \simeq 1$ (рис. 3.12Д). Эта вторая мода несовместима с моделями CTE и STE, поскольку в ней интронное полиаденилирование сосуществует с каноническим сплайсингом.

Интроны, которые имеют высокую поддержку разрывными чтениями ($a+b+c \geqslant 30$) и содержат КСПА с $\Psi > 0.9$, далее будут называться сплайсированными полиаденилированными интронами (spliced polyadenylated introns, SPI). Сравнение распределений значений we_2/we_1 и we_2/wi_1 среди STE, CTE

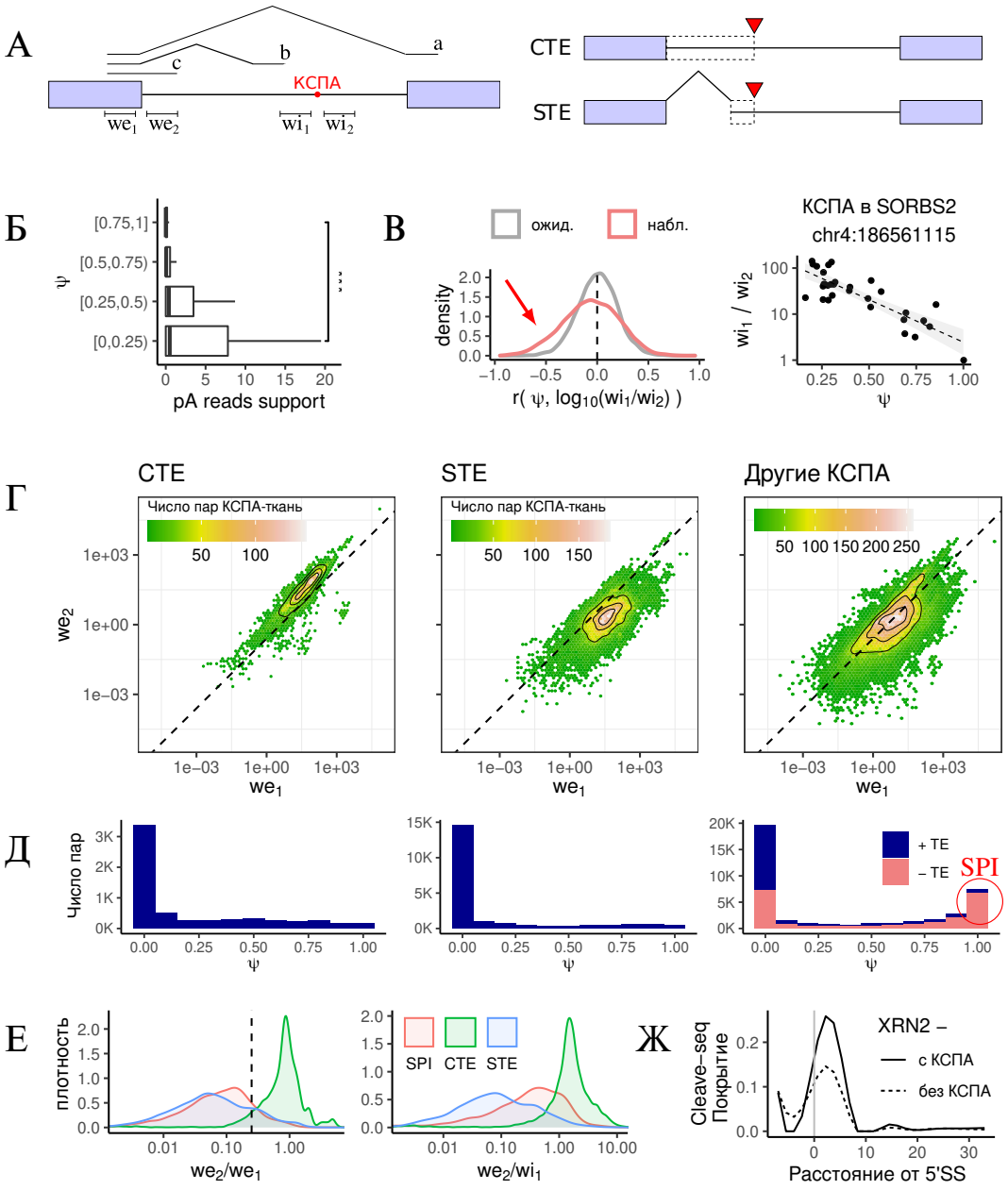


Рисунок 3.12 — Взаимосвязь между интронным полиаденилированием и сплайсингом. **(А)** Экзонные (we_1 и we_2) и интронные (wi_1 и wi_2) окна длиной 150 нт. **(Б)** Число поли(А)-чтений в четырех квартилях распределения Ψ. Символ *** обозначает статистически значимые различия на уровне значимости 0.1%. **(В)** Распределение коэффициентов корреляции Пирсона между Ψ и $\log_{10}(wi_1/wi_2)$ для $n = 12261$ КСПА по сравнению со случайным контролем (слева). Отрицательная ассоциация между Ψ и $log_{10}(wi_1/wi_2)$ в гене *SORBS2* (справа). **(Г)** Совместное распределение we_1 и we_2 в парах КСПА-ткань для CTE, STE и других КСПА. Пунктирная линия соответствует $we_2/we_1 = 0.25$. **(Д)** Распределение Ψ для CTE, STE и других КСПА; +TE (-TE) обозначают КСПА в пределах (не в пределах) 100 нт от аннотированного конца транскрипта. **(Е)** Распределение we_2/we_1 (слева) и we_2/we_1 (справа) для CTE, STE и SPI. Вертикальная пунктирная линия обозначает $we_2/we_1 = 0.25$. **(Ж)** Нормализованное покрытие Cleave-seq в интронах с КСПА и без КСПА.

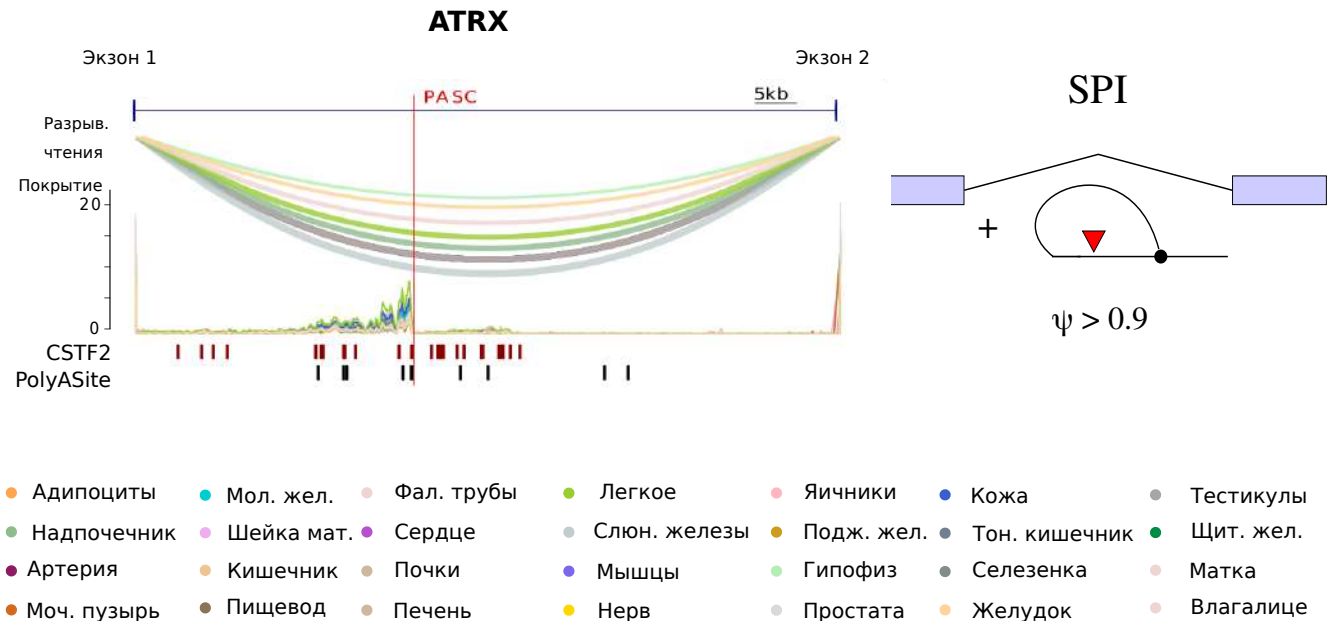


Рисунок 3.13 — Интронное полиаденилирование и сплайсинг в гене *ATRX*. Показан фрагмент гена между экзонами 1 и 2. Цвета разрывных чтений и плотности покрытия соответствуют тканям, в которых они наблюдаются. Вертикальная красная линия обозначает сайт интронного полиаденилирования, который подтверждается сигналами связывания фактора CSTF2 и данными 3'-секвенирования из базы данных PolySite 2.0.

и SPI (рис. 3.12E) показывает, что, как и STE, SPI характеризуются низким покрытием 5'-конца интрона по отношению к экзону (we_2/we_1), но при этом достаточно высоким покрытием перед КСПА относительно 5'-конца интрона (we_2/wi_1). Если SPI действительно представляют собой преждевременно полиаденилированные и сплайсированные интроны, то следует ожидать, что они должны иметь монофосфат на 5'-конце (5'-p), образующийся в результате расщепления точки ветвления ферментом DBR1 [410; 411]. Тогда линеаризованный продукт полиаденилирования должен состоять из двух отдельных молекул: одна соответствует фрагменту интрона до КСПА как с 5'-p, так и с полиА-хвостом, а другая — оставшейся части интрона. В соответствии с этим, покрытие 5'-концов РНК, идентифицированных с помощью метода Cleave-seq для исследования 3'-полиаденилированных РНК, содержащих 5'-p [412], было существенно больше в интронах с КСПА, чем в интронах без КСПА (рис. 3.12Ж). В совокупности эти результаты показывают, что SPI подвергаются как сплайсингу, так и полиаденилированию и не являются 3'-концами отдельных транскриптов, инициируемых и терминирующих в одном и том же интроне.

На рис. 3.13 приведен пример интронного сайта полиаденилирования в гене *ATRX*. Этот ген участвует в ремоделировании хроматина и связан с ря-

дом заболеваний [413]. Наблюдается увеличение покрытия в окне wi_1 перед сайтом полиаденилирования, низкое покрытие в начале интрона, которое могло бы поддерживать CTE, и отсутствие разрывных чтений, которые могли бы поддерживать STE. Единственным возможным объяснением наблюдаемой картины может быть то, что канонический сплайсинг и интронное полиаденилирование в этом гене сосуществуют и действуют одновременно, что приводит к образованию SPI. Также обращает на себя внимание характерное нарастание покрытия перед сайтом полиаденилирования, которое может объясняться постепенной деградацией интрона с 5'-конца.

3.3.9 ККУ и сайты связывания РСБ

Для того, чтобы охарактеризовать взаимосвязь между ККУ и сайтами связывания РСБ, использовались данные футпринтинга по протоколу eCLIP (enhanced crosslinking and immunoprecipitation). Сравнение частот связывания РСБ рядом с ККУ с фоновыми частотами связывания тех же РСБ в КИФ, расположенных рядом с ККУ, показало наличие предпочтений у некоторых факторов связываться с двухцепочечными участками (рис. 3.14А). Среди факторов, которые показали наиболее значимое обогащение, были РСБ, которые выполняют регуляторные функции в сочетании со структурой РНК, такие как RBFOX2 [7], а также факторы, предпочитающие структурированные мотивы, например SRSF9 и SFPQ [131; 414]. Сигналы eCLIP для факторов, связывающих одноцепочечные РНК, таких как hnRNPA1, были значимо недопредставлены [415].

Одной из особенностей протокола eCLIP является то, что в процессе сшивания РСБ может оказаться связан с любой из цепей РНК, примыкающей к двухцепочечной области. Следовательно, вероятность наблюдать сигнал eCLIP рядом с ККУ при условии, что рядом с комплементарным ему ККУ сигнал eCLIP наблюдается, должна быть выше, чем в случаях, когда рядом с комплементарным ККУ такого сигнала нет. Эта ситуация будет называться «раздвоенным сигналом eCLIP». Оценка отношения шансов показывает, что подавляющее большинство РСБ (64 из 74, $P < 10^{-12}$) действительно имеют значительно более высокую вероятность связывания рядом с ККУ при усло-

вии связывания с комплементарным ККУ (рис. 3.14Б). Этот результат можно рассматривать как независимое доказательство двухцепочечности ККУ. Интересно, что наибольшее отношение шансов наблюдалось для фактора TAF15, который сам по себе не является белком, связывающим двухцепочечную РНК, но взаимодействует с белком FUS, который способен связывать двухцепочечные участки [416]. Это наблюдение указывает на то, что связывание РСБ зависит от окружающей структуры РНК, а также на то, что сигналы eCLIP могут неправильно отражать фактические положения сайта связывания, поскольку факторы взаимодействуют не только с РНК, но и друг с другом.

3.3.10 Структура РНК и замедление элонгации транскрипции

Модель котранскрипционного сплайсинга предполагает, что замедление РНК полимеразы II (RNAPII) расширяет «окно возможностей» для распознавания слабых сайтов сплайсинга, тем самым увеличивая степень включения пропускаемых в нормальных условиях экзонов [363; 417]. Помимо влияния на распознавание сайтов сплайсинга, замедление элонгации транскрипции может также влиять на образование структуры РНК, что является еще одним важным фактором, определяющим то, как транскрипт будет процессироваться сплайсосомой [72]. Для того, чтобы изучить роль дальних взаимодействий в котранскрипционном сплайсинге, были проанализированы литературные данные по изменению сплайсинга в в клеточной линии, экспрессирующей медленный мутант РНК-полимеразы R749H [363], а также были выполнены дополнительные эксперименты по секвенированию РНК, в которых для замедления скорости элонгации RNAPII был использован α-аманитин (разд. 2.1.4).

Как и ожидалось, при замедлении элонгации RNAPII степень включения экзонов, следующих за короткими интронами, увеличивается, а степень включения экзонов, следующих за длинными интронами, уменьшается, как под действием α-аманитина, так и в клеточной линии, экспрессирующей медленный мутант R749H (рис. 3.14В). Для того, чтобы выяснить, влияет ли замедление RNAPII по-разному на интроны с ККУ и без них, каждому экзону, следующему за интроном, содержащим ККУ, был сопоставлен случайно выбранный экзон, который следует за интроном той же длины, но без ККУ. Оказалось, что

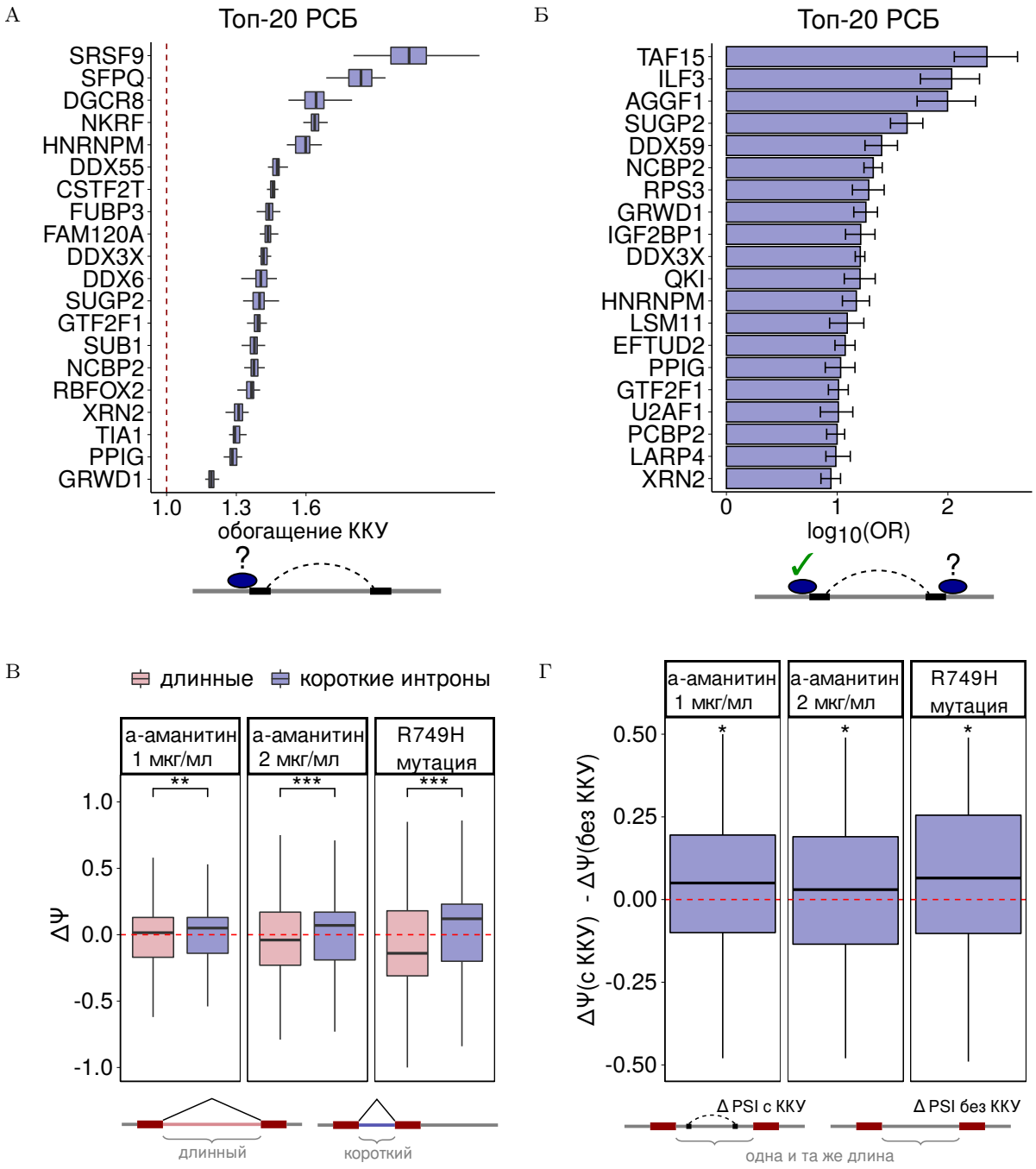


Рисунок 3.14 — ККУ и сайты связывания РСБ. **(А)** Согласно профилям eCLIP, ККУ обогащены сайтами связывания некоторых РСБ (показаны топ-20 РСБ). **(Б)** Отношения шансов (OR) связывания РСБ рядом ККУ при условии, что рядом с комплементарным ККУ он также связывается. **(В)** Изменение степени включения ($\Delta\Psi$) экзонов, следующих за короткими интронами, по сравнению с экзонами после длинных интронов в ответ на замедление RNAPII с помощью α-аманитина и в медленном мутанте R749H71. **(Г)** Разность между изменением степени включения ($\Delta\Psi$с ККУ) экзонов, следующих за интронами с ККУ, и изменением степени включения ($\Delta\Psi$без ККУ) экзонов, следующих за интронами без ККУ, в ответ на замедление RNAPII. Символы *, ** и *** обозначают статистически различимые различия на уровне значимости 5%, 1% и 0.1%, соответственно (H_0 : $\Delta\Psi$с ККУ $-$ $\Delta\Psi$без ККУ $= 0$). Усики соответствуют 95% доверительным интервалам при $n = 40$ случайных сдвигах.

при обеих концентрациях α-аманитина и в мутанте R749H степень включения у экзонов, которые следуют за интронами с ККУ, увеличивается больше, чем у экзонов, следующие за интронами без ККУ (рис. 3.14Г).

Этот результат показывает, что помимо влияния на распознавание сайтов сплайсинга замедление RNAPII может влиять на уровни включения экзонов через котранскрипционное сворачивание РНК. А именно, медленная элонгация транскрипции не только способствует распознаванию сайтов сплайсинга, но также дает достаточно времени для сворачивания структуры РНК в интронах, способствуя включению следующих за ними экзонов. Пример такого динамического действия структуры РНК на сплайсинг будет представлен в разд. 5.2.3.

3.3.11 Примеры структур РНК в консервативных областях

В этом разделе приводятся предсказания дальних взаимодействий в структуре РНК для двух конкретных регуляторных механизмов, связанных со сплайсингом, а именно РНК-мостов и выпетливаний экзонов. Для того, чтобы обнаружить РНК-мосты, т.е. приближающие сайт связывания РСБ к регулируемому экзону структуры, разыскивались пары ККУ, содержащие сигналы eCLIP в пределах 50 нт от одного из комплементарных участков и экзон в пределах 50 нт от другого. В качестве подтверждения регуляции требовалось, чтобы уровень включения экзона значимо откликался на инактивацию экспрессии РСБ.

Эта процедура позволила получить набор из 296 кандидатных РНК-мостов, включая известную из литературы РНК-структуру [7] с энергией гибридизации -19.8 ккал/моль, которая контролирует включение экзона 12 гена *ENAH* (рис. 3.15А). Эта структура окружена раздвоенными пиками eCLIP фактора RBFOX2, что отражает перекрестное сшивание рядом с двухцепочечной областью, причем уровень включение экзона 12 снижается на 43% при инактивации RBFOX2. Также обнаруживается вложенная пара ККУ с энергией гибридизации -20.4 ккал/моль, что позволяет предположить, что структура РНК простирается гораздо дальше, чем об этом сообщалось ранее [7]. В качестве ранее неизвестного примера описывается РНК-мост в 3'-конце гена *RALGAPA1*. В этом гене группа вложенных ККУ приближает сайты связывания RBFOX2 и QKI к предпоследнему экзону (рис. 3.15Б). Инактивация

каждого из этих факторов способствует пропуску экзона, что указывает на то, что АС зависит от связывания RBFOX2 и QKI через РНК-мост.

Аналогичным образом были найдены РНК-структуры, выпетливающие экзоны. Разыскивались пары ККУ, которые окружают экзон и содержат сайт связывания РСБ внутри одного из комплементарных участков. Дополнительно требовалось, чтобы экзон реагировал на инактивацию экспрессии РСБ. Эта процедура дала 1135 потенциальных структур, среди которых были две вложенные пары ККУ, выпетливающие экзон 24 гена *GPR126*, в которых один из ККУ перекрывает сайт связывания RBFOX2, а экзон отвечает на инактивацию RBFOX2 (рис. 3.15В). Другим примером является альтернативный 3'-концевой экзон в гене *FGFR1OP2*, который пропускается при инактивации QKI и выпетливается парой ККУ, перекрывающей сайт связывания QKI (рис. 3.15Г).

3.4 Обсуждение результатов и выводы

3.4.1 Сначала фолдинг, потом выравнивание, или наоборот?

В этой главе были рассмотрены две предельные реализации алгоритма Санкова применительно к дальним взаимодействиям в структуре РНК. Первая из них, использующая подход «сначала фолдинг, потом выравнивание», возможна только для протяженных комплементарных спариваний, образующих стабильный остов вторичной структуры, поскольку в общем случае число комбинаций всевозможных РНК-структур слишком велико и неизбежно привело бы нас обратно к общему случаю алгоритма Санкова, который технически не реализуем. Метод IRBIS решает проблему сжатия комбинаторного пространства структур при помощи процедуры триммингa, которая заранее отсеивает специфичные для отдельных видов k-меры и оставляет только такие, которые хотя бы потенциально могут быть выровнены друг с другом независимо от их расположения в гомологичных последовательностях. Метод имеет высокую чувствительность при нахождении «спрятанных» в любом месте длинных комплементарных спариваний. Однако ценой, которую приходится за это заплатить, является высокая доля ложноположительных предсказаний. Поэтому

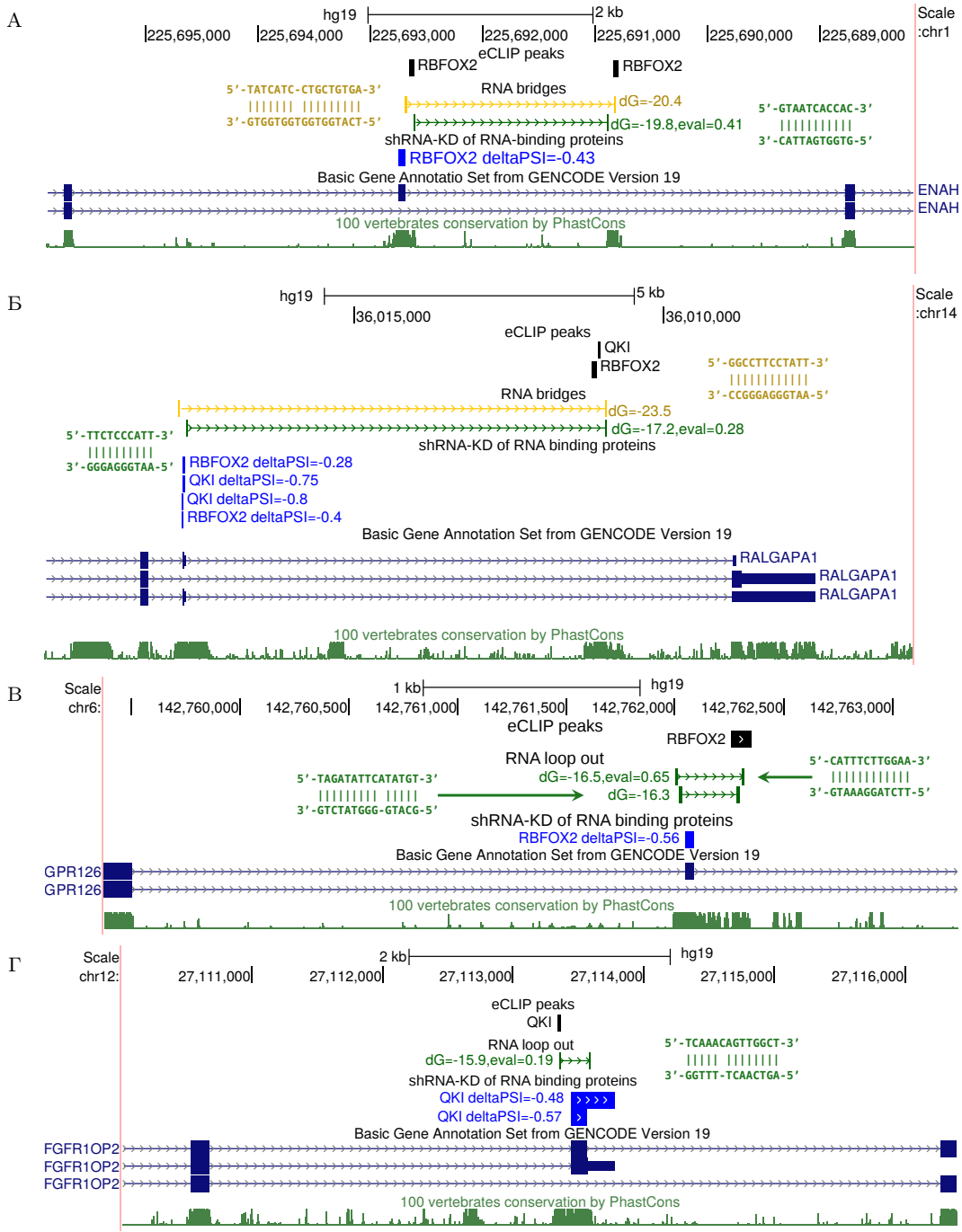


Рисунок 3.15 — Примеры РНК-структур в консервативных областях. Показаны снимки из Геномного браузера UCSC. **(А)** РНК-мост в гене *ENAH* приближает удаленный сайт связывания RBFOX2 к кассетному экзону, уровень включения которого снижается при деплеции RBFOX2 ($\Delta\Psi = -0.43$). **(Б)** РНК-мост в *RALGAPA1* приближает удаленные сайты связывания RBFOX2 и QKI к экзону, степень включения которого снижается при деплеции этих факторов ($\Delta\Psi = -0.28$ и $\Delta\Psi = -0.75$ соответственно). **(В)** Кассетный экзон в *GPR126*, реагирующий на инактивацию RBFOX2 с $\Delta\Psi = -0.56$, выпетливается парой ККУ, один из которых содержит сайт связывания RBFOX2. **(Г)** Альтернативный терминальный экзон в *FGFR1OP2* реагирует на деплецию QKI ($\Delta\Psi = -0.48$) и выпетливается парой ККУ, содержащих сайт связывания QKI. На всех панелях изменения степени включения экзонов статистически значимы ($q < 0.01$).

IRBIS применим только при очень жестких ограничениях на консервативность и длину спариваемой области. Фактически он дает доказательство осуществимости концепции «сначала фолдинг, потом выравнивание» и определяет условия, при которых она технически реализуема.

Однако, как показывают примеры, многие консервативные комплементарные последовательности можно обнаружить непосредственно в множественных выравниваниях геномов. Следует сразу отметить, что полногеномные выравнивания по построению разрывны, т.е., интронные участки, выровненные с интроном эталонного вида, совершенно не обязательно синтенны по отношению к фланкирующим экзонам. В ряде случаев это может приводить к недоразумениям, как, например, это происходит в гене *14-3-3ζ* дрозофилы, в котором пара экзонов, образовавшаяся в результате недавней дупликации, выравнивается неверно, в результате чего известные из экспериментов регуляторные структуры РНК оказываются разнесены по разным интронам [109; 113]. Тем не менее, такие ошибки являются скорее исключением, чем правилом, и применение подхода «сначала выравнивание, потом фолдинг» к полногеномным выравниваниям совершенно оправдано. Метод PREPH находит комплементарные спаривания между частями интронов, находящимися на определенном расстоянии друг от друга, исходя из заранее определенного множества консервативных участков. Поскольку консервативность определяется относительно группы видов, а также пороги, определяющие положение и длину консервативных участков можно выбирать различными способами, применимость PREPH ограничена качеством входного выравнивания и процедурой выбора консервативных участков. Как мы увидим в разд. 6.3, некоторые функциональные структуры РНК не могут быть обнаружены PREPH, поскольку они лишь частично перекрываются с консервативными участками. У метода IRBIS такой проблемы нет.

Таким образом, мы приходим к тому, что оба подхода, «сначала фолдинг, потом выравнивание» и «сначала выравнивание, потом фолдинг» имеют свои преимущества и недостатки, но ни один из них не может в полной мере компенсировать ошибки другого. Сравнивать предсказания, полученные IRBIS и PREPH, по числу найденных структур вряд ли имеет смысл. Однако, как было показано в разд. 3.3.4, предсказания PREPH более многочисленны, хотя и не покрывают полностью предсказания IRBIS. Поэтому в следующих главах будут обсуждаться предсказания PREPH. В дальнейших исследованиях

структуры эукариотических РНК представляется разумным отойти от готовых полногеномных выравниваний и усовершенствовать методы выравнивания интронов для выявления в них консервативных участков, комплементарность между которыми может находить PREPH. Оставшиеся промежуточные последовательности можно исследовать с помощью IRBIS для построения новых, обусловленных структурой РНК выравниваний, итеративно повторяя этот процесс для сокращения области поиска. Однако такой подход трудно реализовать в автоматическом режиме, и для многих генов он требует «ручной» работы.

3.4.2 Приоритизация РНК-структур

Преимуществом филогенетических методов является возможность ранжировать предсказания РНК-структур по количеству компенсаторных замен. Однако для того, чтобы наблюдать компенсаторные изменения, необходимо, чтобы замены в нуклеотидных последовательностях хоть и редко, но происходили. Условия поиска структур в IRBIS слишком строгие для того, чтобы ожидать достаточной вариабельности от консервативных k-меров. В PREPH комплементарные взаимодействия предсказываются между консервативными частями интронов, что не исключает небольшой вариабельности в отдельных позициях. Однако, как было показано, лишь малая часть предсказанных комплементарных последовательностей обладает достаточным уровнем вариабельности для анализа компенсаторных замен. Следует также отметить, что компенсаторные замены могут возникать не только из-за спаренности оснований во вторичной структуре РНК, но и для поддержания сайтов связывания РСБ, расположенных на противоположных цепях. Ярким примером является длинная некодирующая РНК RP11-439A17.4, которая находится в антисмысловой ориентации по отношению к гену HIST2H2BA и содержит сайт связывания транскрипционного фактора (разд. 3.2.5). Этот сайт также встречается почти во всех генах гистонов в смысловой ориентации, а происходящие в нем мутации приводят к кажущимся компенсаторным заменам в РНК [113]. Поэтому рассчитывать на эволюционные подписи для усиления мощности предсказаний структуры РНК, к сожалению, не приходится.

В отсутствие вариабельности нуклеотидных последовательностях важным индикатором функциональности структуры может быть соответствие между комплементарностью и консервативностью. А именно, во многих известных из литературы случаях наблюдается резкое падение уровня консервативности на границах комплементарных областей, что можно использовать для приоритизации РНК-структур. Эта идея, впервые высказанная в [107], получила развитие в последующих работах [108; 113]. Однако метрики, оценивающие соответствие между комплементарностью и консервативностью, лишь косвенно подтверждают двухцепочечность комплементарных участков и не доказывают существования отбора на спаривание оснований.

Исследование структуры РНК с помощью icSHAPE и анализ дальних взаимодействий между участками РНК с помощью фотоиндуцируемого сшивания являются наиболее современными методами глобального анализа структуры РНК *in vivo*. Однако эти данные отражают закономерности экспрессии генов, специфичные для клеточных линий, в которых они были получены, и имеют низкое покрытие чтениями в интронных областях, которые подвергаются сплайсингу и деградации. Снижение интронного сигнала также является распространенной проблемой в экспериментах eCLIP и, как было показано в разд. 3.3.8, искажает истинную частоту полиаденилирования в интронах. Поэтому валидация предсказанных РНК-структур с помощью этих данных была возможна лишь для небольшого числа пар ККУ, хотя и показала согласованный результат. Следующая глава будет целиком посвящена тому, как находить структуры РНК с использованием метода конформационного секвенирования РНК *in situ*.

3.4.3 Структура РНК, сплайсинг и полиаденилирование

Отдельные предположения о существовании взаимосвязи между дальними взаимодействиями в структуре РНК и сплайсингом высказывались в предыдущих исследованиях [418], однако в полной мере статистическая характеризация этой взаимосвязи дается в данной диссертационной работе. Наблюдаемые для дрозофил (разд. 3.2.3) закономерности, в частности, предпочтение ККУ располагаться внутри интронов рядом с сайтами сплайсинга

и выпетливание альтернативных экзонов, справедливы и для дальних взаимодействий в структуре РНК у млекопитающих. В частности, наблюдается более низкая частота включения выпетливаемых экзонов, повышенная встречаемость ККУ вокруг кольцевых РНК, избегание интронных точек ветвления и, как правило, как правило, общий ингибирующий эффект двухцепочечной структуры на цис-регуляторные элементы сплайсинга. Также обращают на себя внимание перепредставленность в ККУ сайтов редактирования РНК и высокая частота раздвоенных сигналов eCLIP, которые являются следами специфичных для двухцепочечных областей взаимодействий РНК с белковыми факторами. С одной стороны эти наблюдения подтверждают хорошо известный механизм ADAR-опосредованного редактирования РНК [419] и показывают важность структуры РНК для сборки РНК-белковых комплексов с предпочтением некоторых РСБ и избеганием других [110; 131]. В то же время их можно рассматривать как независимое подтверждение двухцепочечности предсказанных РНК-структур в дополнение к данным экспериментального профилирования структуры РНК и компенсаторным заменам.

Как уже отмечалось, компоненты аппарата процессинга РНК работают в строгой координации не только в пространстве, но и во времени. Кинетический профиль элонгации RNAPII оказывает существенное влияние на альтернативный сплайсинг. Медленная элонгация не только открывает окно возможностей для распознавания слабых сайтов сплайсинга, что приводит к увеличению частоты включения альтернативных экзонов, но также может влиять на выбор сайтов полиаденилирования за счет усиления распознавания субоптимальных сигналов [420]. В полном согласии с этим, при замедлении RNAPII с помощью α-аманитина наблюдается повышенная частота включения экзонов, которым предшествуют более короткие интроны. Однако также имеется и другая тенденция, в которой структурированные и неструктурированные РНК по-разному реагируют на замедление RNAPII. Это наблюдение указывает на то, дальние взаимодействия в структуре РНК могут быть тем самым элементом, который координирует взаимодействие между пространственной и временной компонентами в регуляции сплайсинга. Как будет показано в разд. 5.2.3, подобные механизмы действительно имеют место в эукариотических генах.

Структура РНК может участвовать в распознавании поли(А)-сайтов фактором специфичности расщепления и полиаденилирования (CPSF), сближая сайт связывания и сайт разрезания [404; 421]. Структура РНК в 5'-НТО также

может участвовать в регуляции трансляции [422]. Поскольку эти механизмы задействуют в основном локальную структуру РНК, особенно интересным представляется наблюдение взаимосвязи между дальними взаимодействиями в структуре РНК и процессингом 3'-конца пре-мРНК, которая проявляется в обогащении концов транскриптов внутри, т.е. между комплементарными последовательностями ККУ. Человеческие гены содержат тысячи «спящих» интронных поли(А)-сайтов, которые подавляются, по крайней мере частично, мяРНП U1 в процессе, называемом телескриптингом [423]. Может ли структура РНК подавлять преждевременное интронное полиаденилирование?

На рис. 3.16 представлен гипотетический механизм подавления интронного полиаденилирования путем котранскрипционного сплайсинга, который объясняет обогащение концами транскриптов в промежутках между ККУ. А именно, полиаденилирование структурированной пре-мРНК можно предотвратить путем котранскрипционного вырезания интрона, в то время как структура РНК стабилизирует молекулу посредством внутримолекулярных комплементарных спариваний несмотря на разрезание основной цепи (рис. 3.16А). Этого не должно происходить в неструктурированных РНК в случае, если сплайсинг происходит с задержкой по отношению к полиаденилированию (рис. 3.16Б). В разд. 5.1.3 будет приведен пример дальних взаимодействий в структуре РНК гена *Nmnat* дрозофилы, которые управляют не только сплайсингом, но и полиаденилированием.

Предположение о связи интронного полиаденилирования, структуры РНК, и котранскрипционного сплайсинга косвенно подтверждается сделанными в разд. 3.3.8 предположениями о существовании сплайсированных полиаденилированных интронов (SPI), которые представляют из себя промежуточные продукты сплайсинга и полиаденилирования, образующиеся одновременно с транскрипцией. Как и в механизме, представленном на рис. 3.16, если полиаденилирование опережает сплайсинг, то образуется усеченный транскрипт с композитным терминальным экзоном. Если же сплайсинг успевает вырезать интрон, в котором происходит или уже произошло полиаденилирование, то образуется полноразмерный транскрипт, а интрон, содержащий поли(А)-хвост, но все еще удерживаемый внутримолекулярной структурой, будет вырезаться и деградировать. В некоторых случаях удается наблюдать SPI — интермедиаты, простирающиеся от 5'ss до сайта полиаденилирования, содержащие как 5'-р из-за разветвления лариата, так и поли(А)-хвост. SPI разрушаются с 5'-кон-

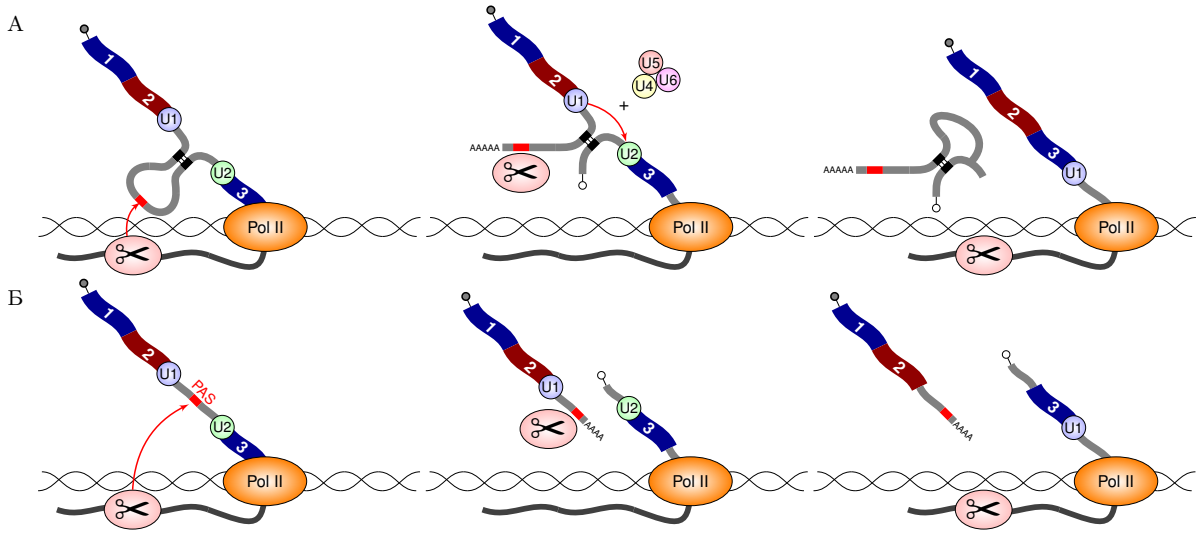


Рисунок 3.16 — Гипотеза о котранскрипционном подавлении преждевременного полиаденилирования структурой РНК. **(А)** Если сплайсинг происходит вскоре после или одновременно с полиаденилированием, то структурированный интрон будет котранскрипционно вырезан несмотря на разрезание основной цепи благодаря внутримолекулярным комплементарным спариваниям. **(Б)** Если сплайсинг происходит со значительной задержкой по отношению к полиаденилированию, то «спасения» не происходит. Переключение между (А) и (В) зависит от скорости сплайсинга, сворачивания РНК и элонгации транскрипции.

ца клеточными экзонуклеазами, о чем во многих случаях свидетельствует характерное нарастание покрытия чтениями от 5'ss до сайта полиаденилирования [408].

Существование SPI, а также различия в ответе сплайсинга на замедление элонгации транскрипции в зависимости от структуры РНК позволяют предположить, что котранскрипционный сплайсинг может выполнять побочную функцию по предотвращению преждевременной терминации транскрипции у эукариот. Это предположение оспаривает тезис о том, что после использования интронного сайта полиаденилирования интрон больше не может подвергаться сплайсингу. Временные и пространственные взаимодействия сплайсинга и полиаденилирования регулируются множеством РСБ, которые распознают одни и те же сигналы на растущей пре-мРНК и связываются с ними в одно и то же время [424—426]. Представляется возможным, что эволюция позволяет создавать такие «разменные монеты» — сайты полиаденилирования в интронах, которые благодаря структуре РНК котранскрипционно вырезаются сплайсингом и не вызывают преждевременной терминации транскрипции.

Глава 4. Структура и конформационное секвенирование РНК

В этой главе будет рассказано о том, как предсказания дальних взаимодействий в структуре РНК соотносятся с данными конформационного секвенирования РНК *in situ* (RIC-seq). Основные результаты охватывают исследования, проведенные с 2019 по 2023 год [24; 25; 364], а также вспомогательные результаты, изложенные в других работах. В разд. 4.1 рассказывается о вычислительном конвейере «RNAcontacts» для картирования данных RIC-seq. В разд. 4.2 производится развернутый анализ полученных в разд. 3.3 предсказаний и их сравнение с результатами экспериментов RIC-seq, а в разд. 4.3 найденные закономерности используются для разработки метода предсказания структуры РНК вне консервативных областей, основанного на данных RIC-seq.

4.1 Конформационное секвенирование РНК *in situ*

В методе конформационного секвенирования РНК *in situ* (RIC-seq) цепи РНК сшиваются через РСБ [21; 427], что позволяет не только устанавливать вторичную и третичную структуру РНК, но и помогает воссоздать трехмерные карты взаимодействий РНК с РСБ. Протокол RIC-seq состоит из семи основных шагов (рис. 4.1). Лигированные *in situ* комплексы РНК с белками выделяются из клеток и подвергаются обработке микрококковой нуклеазой (MNase) и термочувствительной щелочной фосфатазой (FastAP) с образованием свободных гидроксильных групп на 5'- и 3'-концах РНК. Затем происходит присоединение биотинилированного цитозина, обработка полинуклеотидкиназой T4 с образованием свободных гидроксильных групп на 3'-концах и фосфатной группы на 5'-концах. Затем происходит лигирование пространственно близких участков РНК, выделение РНК, очистка от белков и ДНК и осаждение фрагментов, содержащих биотин, с помощью стрептовидина. К полученным фрагментам добавляются адаптеры и провыполняется высокопроизводительное парно-концевое секвенирование.

При картировании результатов секвенирования на референсный геном точки лигирования пространственно близких участков РНК проявляются в

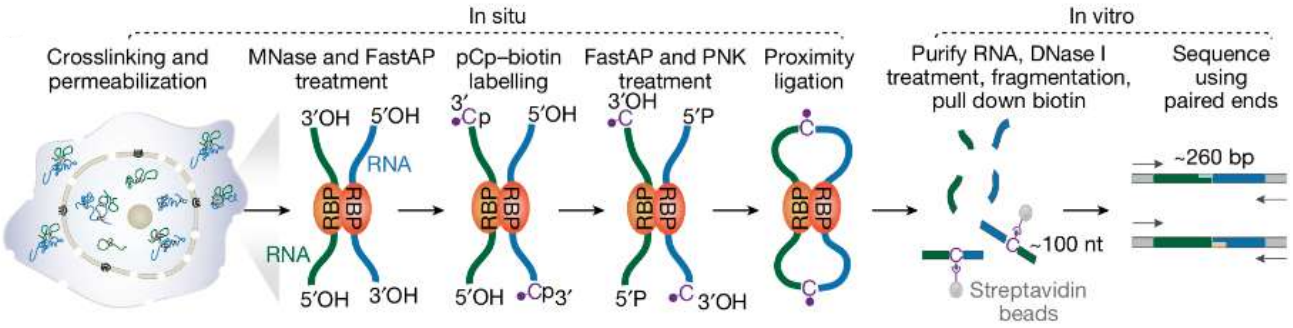


Рисунок 4.1 — Принцип конформационного секвенирования РНК *in situ* RIC-seq (RIC-seq). Изображение заимствовано из [21].

виде чтений с разрывами. Однако в отличие от ДНК-ДНК взаимодействий, проявляющихся в виде разрывных чтений, которые картируются только на два пространственно близких геномных локуса, взаимодействия РНК с РНК дают чтения, которые могут картироваться более сложным образом, поскольку пре-мРНК подвергаются сплайсингу. В частности, химерные фрагменты могут содержать как сплайсосомные интроны, так и продукты сшивки, в результате чего образуются как чтения с каноническими GT/AG разрывами, возникающими при сплайсинге, так и чтения с другими разрывами, возникающими при лигировании (рис. 4.2А). Для картирования таких чтений требуются специализированные алгоритмы, поскольку большинство программ-картировщиков может работать только с одним типом разрывов. Для этой цели в рамках данной диссертационной работы был разработан вычислительный конвейер «RNAcontacts», позволяющий картировать короткие чтения с двумя различными типами разрывов [364].

«RNAcontacts» обходит проблему нескольких типов разрывов путем выравнивания коротких чтений в двухпроходном режиме (рис. 4.2Б). На первом проходе картировщик STAR [365] выравнивает контрольный набор данных секвенирования РНК в парноконцевом режиме, используя строгий штраф для не-GT/AG-разрывов, чтобы определить интроны, которые экспрессируются в данном биологическом образце и отсутствуют в аннотации. На втором проходе чтения из эксперимента RIC-seq выравниваются с использованием ослабленного штрафа за не-GT/AG-разрывы, при этом на вход подается список интронов, идентифицированных на первом проходе, чтобы картировщик преимущественно делал разрывы в чтениях по координатам из предоставленного списка. Поскольку данные RIC-seq могут содержать химерные чтения на произвольном геномном расстоянии или *in trans*, выравнивание на втором проходе выполня-

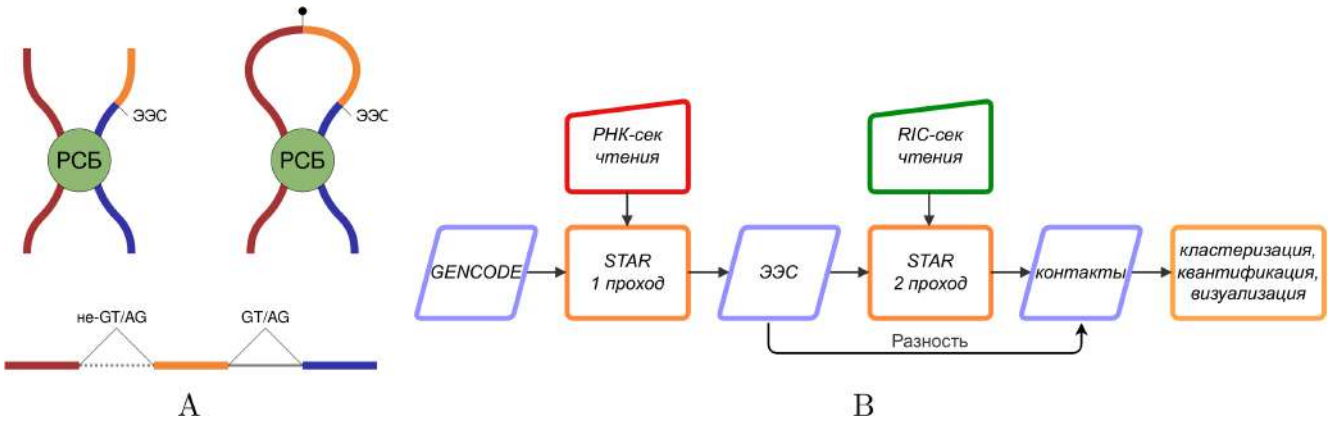


Рисунок 4.2 — Картирование чтений с двумя типами разрывов. **(А)** В протоколе RIC-seq [21] несплайсированный участок РНК может быть лигирован через РНК-связывающий белок (РСБ) с другим участком, содержащим экзон-экзонные соединения (ЭЭС). Последовательность, образованная в результате лигирования, выравнивается на геном с двумя разрывами: не-GT/AG разрывом, соответствующим точке лигирования, и каноническим GT/AG разрывом, соответствующим ЭЭС. **(Б)** Схема «RNAcontacts». На первом проходе чтения из контрольного эксперимента РНК-сек выравниваются на референсный геном для выявления экспрессируемых ЭЭС. Они используются на втором проходе как известные интроны при выравнивании данных КС для обнаружения разрывов, в которых закодированы РНК-контакты.

ется в одноконцевом режиме. Все выравнивания с разрывами, полученные на втором проходе, анализируются для извлечения РНК-контактов и исключения интронов, полученных на первом проходе.

При использовании протокола RIC-seq небольшие различия в координатах разрывов могут возникать даже при сопоставлении чтений, которые соответствуют одной и той же точке лигирования, поскольку картирование на разные цепи может приводить к смещению координат из-за отсутствия консенсусных последовательностей у разрыва. Кроме того, различные копии одной и той же РНК разрезаются и повторно лигируются стохастически, что приводит к еще большей вариабельности. Учитывая эту техническую и биологическую вариативность, следует ожидать появление кластеров точек лигирования, а не четко определенных сайтов, как это имеет место в случае сайтов сплайсинга. Поэтому соответствующие РНК-контактам точки разрыва кластеризуются с использованием односвязной кластеризации с порогом по расстоянию 10 нт. После этого РНК-контакты определяются как пары кластеров, соединяемые хотя бы одним разрывом, а число чтений, поддерживающих РНК-контакт, определяется как суммарное число чтений с разрывами внутри кластера.

4.2 Сравнение экспериментов RIC-seq и предсказаний PREPH

4.2.1 Согласованность РНК-контактов и предсказаний PREPH

В этом разделе приводится исследование согласованности между предсказаниями PREPH и РНК-контактами из экспериментов RIC-seq, проведенных на семи клеточных линиях человека, включая GM12878, H1, HeLa, HepG2, IMR90, K562 и hNPC. РНК-контакт характеризуется парой геномных координат, соответствующих точкам лигирования пространственно близких фрагментов РНК после разрезания, а также числом поддерживающих их разрывных чтений. В общей сложности в семи клеточных линиях было получено около 25 миллионов РНК-контактов и 46 миллионов поддерживающих их разрывных чтений.

Дальнейшие рассуждения основываются на предположении о том, что после лигирования рядом с двухцепочечными участками РНК должны возникать РНК-контакты, изображенные на рис. 4.3А, т.е., РНК-контакты, которые поддерживают пару ККУ с внешней и/или с внутренней стороны, что соответствует коллинеарным (2–3) или химерным (1–4) чтениям с разрывами. Из 916360 пар ККУ были отобраны пары с расстоянием не менее 200 нт между комплементарными участками. Вокруг центра каждого ККУ было выбрано окно радиуса 100 нт (рис. 4.3Б). Такое центрирование окон позволяет единообразно оценивать РНК-контакты вблизи ККУ и учитывать РНК-контакты, происходящие внутри самих комплементарных областей. В результате набор пар ККУ разделился на четыре взаимоисключающие группы: поддерживаемые хотя бы одним контактом внутри, но не снаружи (inside, I), поддерживаемые хотя бы одним контактом снаружи, но не внутри (outside, O), поддерживаемые хотя бы одним контактом как внутри, так и снаружи (inside and outside, IO), и не поддерживаемые контактами ни снаружи, ни внутри (none, N).

Из-за большой разреженности информации, получаемой в экспериментах RIC-seq, в каждой клеточной линии поддержка пар ККУ оценивалась по принципу «все или ничего», т.е., пара ККУ распределялась в категории I, O, IO и N при наличии хотя бы одного соответствующего этим группам чтения. Такая классификация также применялась по отношению к объединенному набору всех экспериментов RIC-seq во всех клеточных линиях, причем было обнаружено,

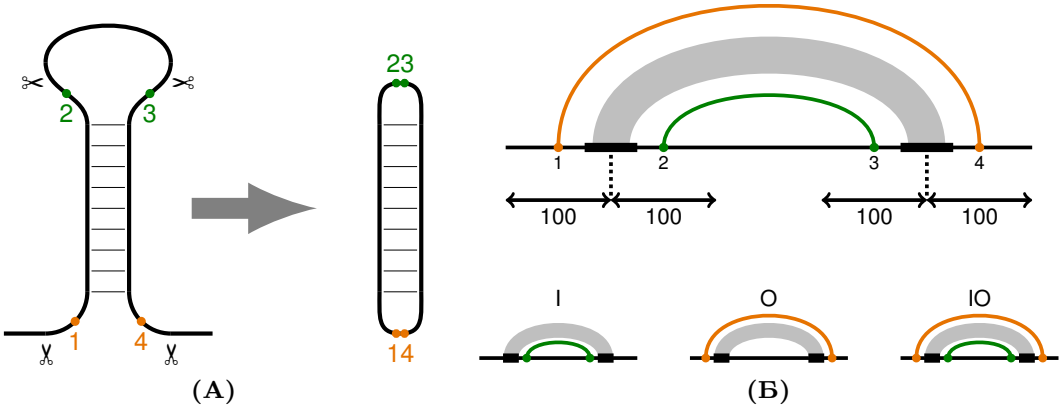


Рисунок 4.3 — Взаимосвязь между РНК-контактами и ККУ. **(А)** Разрезание и повторное лигирование цепей РНК, примыкающих к ККУ, приводит к РНК-контактам, поддерживающим структуру с внутренней (2-3) и/или внешней (1-4) стороны. **(Б)** Сверху: расположение окон для поиска РНК-контактов. Внизу: внутренние (I) и внешние (O) контакты соответствуют внутренней и внешней дугам по отношению к структуре РНК. В категории IO структура поддерживается как снаружи, так и внутри.

что многие пары ККУ поддерживаются РНК-контактами внутри в одной клеточной линии, а снаружи — в другой. Это указывает на то, что объединенные данные из всех экспериментов лучше отражают расположение РНК-контактов, чем разреженные данные, ограниченные конкретным биологическим состоянием. Затем рассматривалось пересечение между группами I, O и IO в двух биорепликах каждого эксперимента и между клеточными линиями. Как и ожидалось, группы были больше всего похожи между биорепликами, а степень сходства при сравнении клеточных линий друг с другом была ниже. Уровень согласованности между биорепликами составлял 9–15%, что еще раз показывает, что каждый отдельный эксперимент RIC-seq дает очень разреженные результаты, и что объединение результатов нескольких независимых экспериментов является наилучшей стратегией для дальнейшего анализа. Более подробно эти технические результаты изложены в [24].

4.2.2 Свойства РНК-структур с поддержкой РНК-контактами

Как было показано в разд. 3.3.6, пары ККУ характеризуются рядом параметров, а именно свободной энергией взаимодействия (ΔG), E-значением, наличием сайтов редактирования РНК и появлением раздвоенных сигналов eCLIP, которые свидетельствуют о перекрестном связывании РСБ вблизи ком-

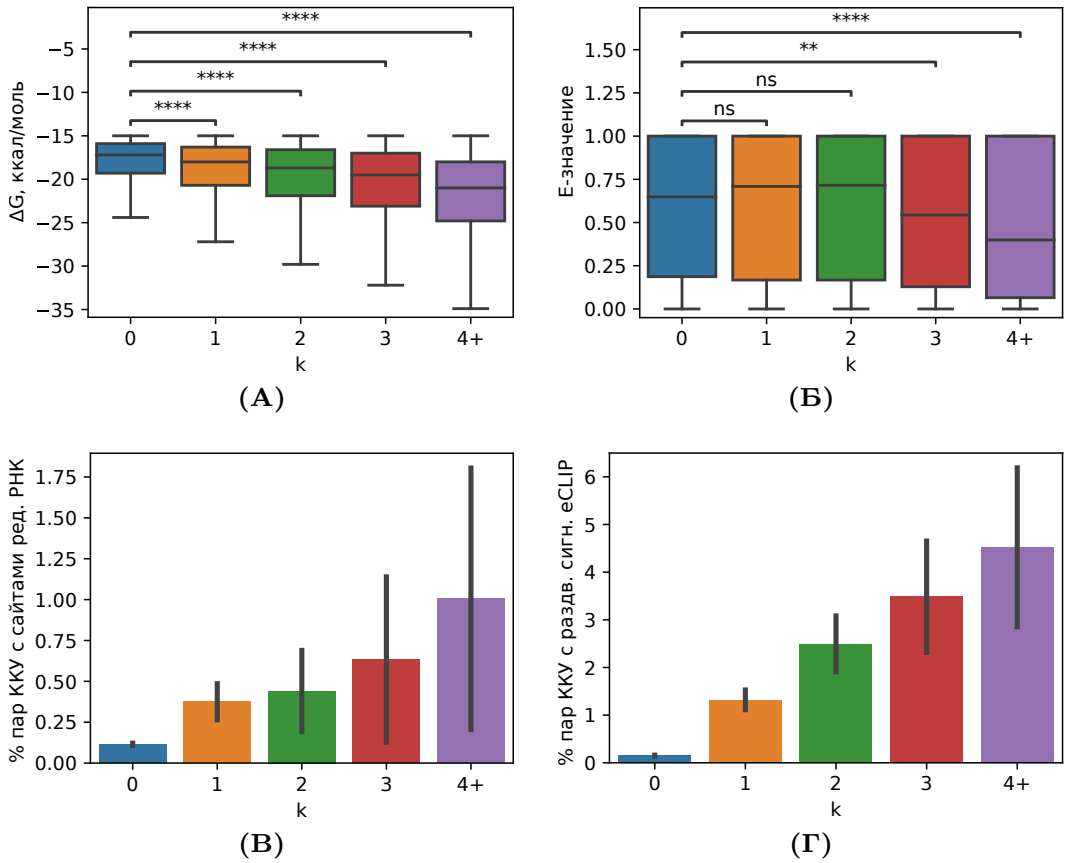


Рисунок 4.4 — Свойства ККУ, поддерживаемых РНК-контактами изнутри и снаружи (категория IO) в по крайней мере k клеточных линиях. **(А)** Свободная энергия гибридизации ΔG. **(Б)** E-значение из R-scape. **(В)** Частота сайтов редактирования РНК. **(А)** Частота раздвоенных сигналов eCLIP вблизи ККУ. Символы *, **, ***, и **** обозначают статистически значимые различия на уровне значимости 5%, 1%, 0.1% и 0.01%, соответственно. Отрезками показаны 95% доверительные интервалы для пропорций.

плементарных цепей РНК. Для того, чтобы оценить взаимосвязь между этими параметрами и наблюдаемыми РНК-контактами, ККУ, которые поддерживаются хотя бы одним РНК-контактом в группах IO, I и O по крайней мере в $k \geqslant 1$ клеточных линиях, сравнивались с ККУ, которые не поддерживаются РНК-контактами (т.е. $k = 0$). Абсолютное значение ΔG, доля ККУ с сайтами редактирования РНК и доля ККУ с раздвоенными пиками eCLIP ожидаемо увеличиваются с увеличением k, в то время как поддержка компенсаторными заменами (E-значение) значимо изменяется только при больших значениях k (рис. 4.4).

Для того, чтобы дополнительно охарактеризовать ККУ с различными уровнями поддержки, было проведено двумерное разбиение пар ККУ по количеству клеточных линий, в которых они поддерживаются I и O контактами, и построены 95% доверительные интервалы для свободной энергии взаимодействия в каждом классе (табл. 3 и табл. 4). Обнаруживается последовательная

тенденция к увеличению ΔG (по абсолютному значению) с увеличением поддержки чтениями с обеих сторон, при этом достаточно много ККУ имеют хорошо выраженную поддержку. Например, 534 ККУ поддерживаются по крайней мере в пяти клеточных линиях как внутренними, так и внешними контактами.

Таблица 3 — Количество ККУ, поддерживаемых внутренними контактами в k клеточных линиях (строки) и внешними контактами в l клеточных линиях (столбцы).

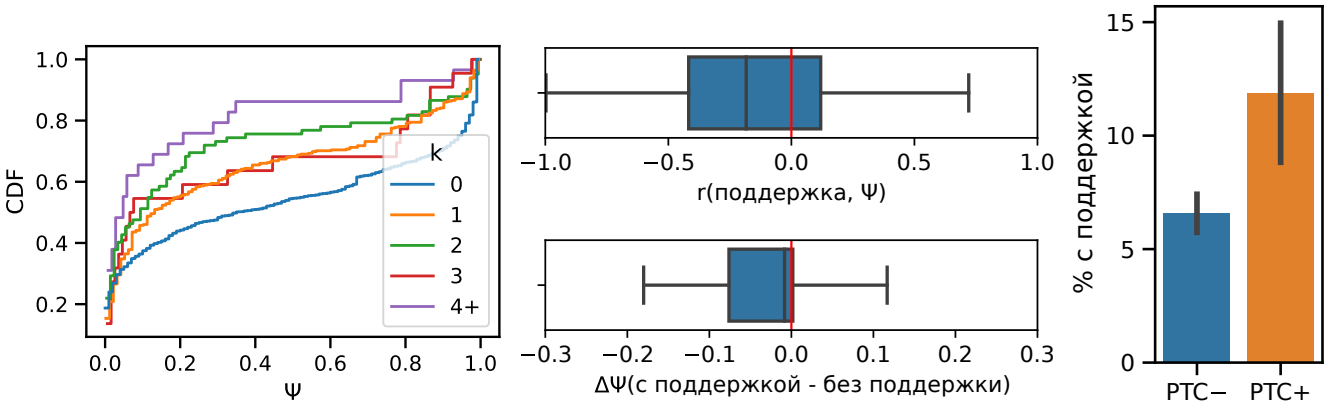
I\O	0	1	2	3	4	5+
0	0	32,320	5,288	1,582	651	466
1	36,599	9,281	2,801	991	432	285
2	7,036	3,517	1,550	713	352	253
3	2,184	1,449	878	481	252	246
4	809	733	497	310	197	221
5+	634	558	430	361	270	534

Таблица 4 — Доверительные интервалы (95% уровень доверия) для свободной энергии гибридизации ККУ в ккал/моль, поддерживаемых внутренними контактами в k клеточных линиях (строки) и внешними контактами в l клеточных линиях (столбцы).

I\O	0	1	2	3	4	5+
0	нет	(-18.5, -18.4)	(-18.7, -18.6)	(-19.2, -18.8)	(-19.5, -18.9)	(-20.5, -19.6)
1	(-18.6, -18.5)	(-19.0, -18.8)	(-19.1, -18.9)	(-19.6, -19.1)	(-20.0, -19.2)	(-20.9, -19.7)
2	(-19.2, -19.0)	(-19.4, -19.1)	(-19.8, -19.3)	(-19.9, -19.3)	(-20.5, -19.6)	(-21.3, -20.0)
3	(-19.8, -19.5)	(-20.1, -19.6)	(-20.1, -19.5)	(-20.6, -19.7)	(-22.0, -20.6)	(-21.4, -20.1)
4	(-20.0, -19.4)	(-20.4, -19.7)	(-20.9, -20.0)	(-20.9, -19.9)	(-21.7, -20.0)	(-21.5, -20.1)
5+	(-20.2, -19.6)	(-20.8, -19.9)	(-21.2, -20.3)	(-21.7, -20.6)	(-21.8, -20.6)	(-23.5, -22.5)

Затем оценивалась средняя степень включения (Ψ) экзонов, расположенных между комплементарными частями ККУ, в экспериментах РНК-сек, проведенных в тех же клеточных линиях, что и эксперименты RIC-seq. Как и ожидалось, распределение Ψ смещается в сторону меньших значений с увеличением k (рис. 4.5А) в полном согласии с предыдущими наблюдениями о том, что степень включения экзона уменьшается с увеличением стабильности окружающих ККУ.

Для анализа дифференциального АС были отобраны ККУ, которые поддерживались по крайней мере четырьмя чтениями по крайней мере в трех клеточных линиях. Сравнение уровней включения экзонов, выпетливаемых ККУ, в экспериментах с поддержкой РНК-контактами и в экспериментах



(А) **(Б)** **(В)**

Рисунок 4.5 — Свойства экзонов, выпетливаемых ККУ с поддержкой РНК-контактами. **(А)** Распределение средней степени включения выпетливаемых ККУ экзонов (Ψ), которые поддерживаются РНК-контактами как внутри, так и снаружи (категория IO) в не менее чем k клеточных линиях. CDF — кумулятивная функция распределения. **(Б)** Сверху: распределение коэффициента корреляции Пирсона между уровнем поддержки структуры (поддержка) и степенью включения выпетливаемого ей экзона (Ψ). Снизу: Распределение $\Delta\Psi = \Psi_h - \Psi_l$, где Ψ_h и Ψ_l — средние значения Ψ в клеточных линиях с поддержкой RIC-seq и без неё, соответственно. **(В)** Доля выпетливаемых экзонов с поддержкой RIC-seq среди кодирующих кассетных экзонов (PTC-) и ядовитых экзонов (PTC+).

без поддержки показало, что разность уровней включения экзонов в клеточных линиях с поддержкой и без нее ($\Delta\Psi$) значительно смещена в сторону отрицательных значений ($P < 10^{-9}$). Также наблюдалась значимая отрицательная корреляция между уровнем поддержки и степенью включения экзона (рис. 4.5Б), что позволяет предположить, что АС может регулироваться сборкой и разборкой структуры между ККУ в различных клеточных линиях. Значительно большая доля выпетливаемых экзонов, содержащих PTC, по сравнению с выпетливаемыми кассетными экзонами, не содержащими PTC, поддерживалась РНК-контактами (рис. 4.5В). Кроме того, в группе с поддержкой РНК-контактами равновесные свободные энергии ККУ, образующие выпетливания, были значительно больше по абсолютной величине для экзонов, содержащих PTC, по сравнению с экзонами без PTC ($P < 10^{-4}$), в то время как в группе без поддержки не было обнаружено существенных различий ($P = 0.2$). Эти результаты указывают на то, что структура РНК активно участвует в контроле непродуктивного сплайсинга, возможно, обеспечивая возможность регулируемого пропуска ядовитых экзонов (см. также разд. 6.3).

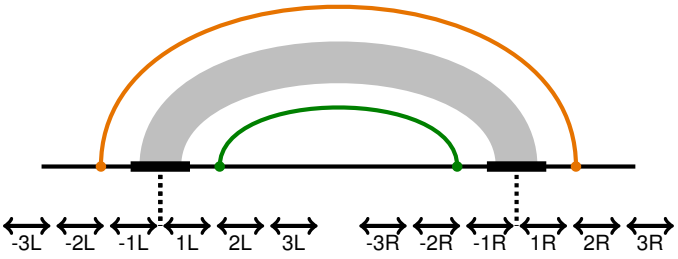
Для дополнительной характеризации взаимосвязи между РНК-контактами и ККУ, был создан классификатор на основе модели случайного леса, который предсказывает появление раздвоенных сигналов eCLIP. С этой целью оба комплементарных участка, левый (L) и правый (R), были окружены тремя

окнами по 50 нт в направлении 5'-конца РНК (-3, -2, -1) и в направлении 3'-конца (1, 2, 3) (рис. 4.6A). Поскольку плотность РНК-контактов выше для более коротких интервалов, в модель в качестве переменной также включался разброс, т.е., расстояние между комплементарными участками [364]. В качестве метрики, оценивающей качество модели, использовалась площадь под рабочей операционной кривой (AUC). Вариант модели, учитывающий только разброс, способен предсказать сигналы eCLIP вблизи ККУ с $AUC = 0.66$, в то время как добавление числа РНК-контактов в качестве независимой переменной повышает качество прогнозов до $AUC = 0.74$ (рис. 4.6B). Наличие внутренних и внешних контактов в непосредственной близости от ККУ, т. е. контактов между окнами $1L$ и $-1R$ и контактов между окнами $-1L$ и $1R$ соответственно, оказалось наиболее важным признаком для классификатора (рис. 4.6C). Этот эффект не связан с более высокой плотностью РНК-контактов на более коротких расстояниях, поскольку разброс был включен в эту модель в качестве независимой переменной.

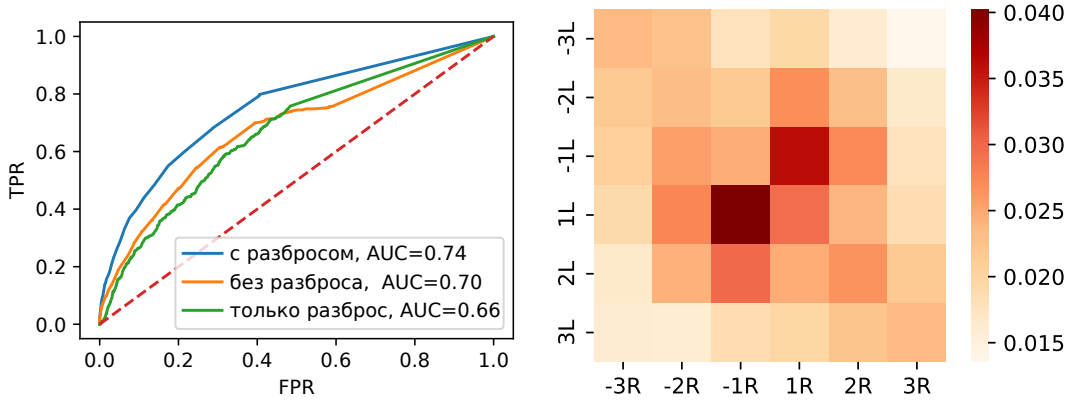
4.2.3 Примеры РНК-структур с поддержкой РНК-контактами

В разд. 3.3.11 были разобраны несколько примеров предсказанных структур РНК с дальними взаимодействиями в контексте регуляторных механизмов, связанных со сплайсингом, а именно РНК-мостов и выпетливаний экзонов. В этом разделе описываются структуры РНК в двух генах человека, *PHF20L1* и *CASK*, которые в дальнейшем были отобраны для экспериментальной валидации (разд. 5.2).

При отборе мишеней для проверки влияния предсказанных РНК-структур на АС в первую очередь оценивались уровни экспрессии генов в доступных для исследования клеточных линиях. Если выбранный ген экспрессировался на достаточном для обнаружения уровне, то при дальнейшем отборе мишеней отдавалось предпочтение генам, у которых экспрессировались ожидаемые сплайс-изоформы. Например, если структура РНК выпетливает экзон, то степень его включения в клеточной линии должна быть низкой для того, чтобы при разрушении РНК-структуры можно было обнаружить его включение. Среди оставшихся кандидатов отбирали гены, имеющие отношение к заболе-

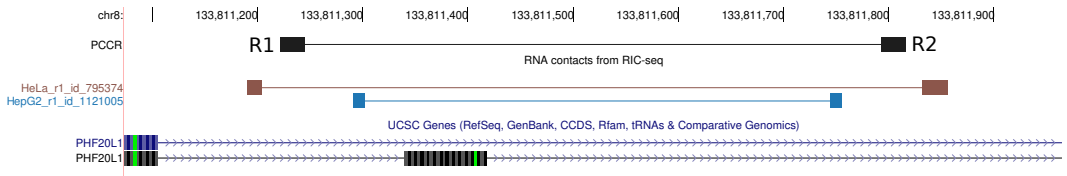


(А)

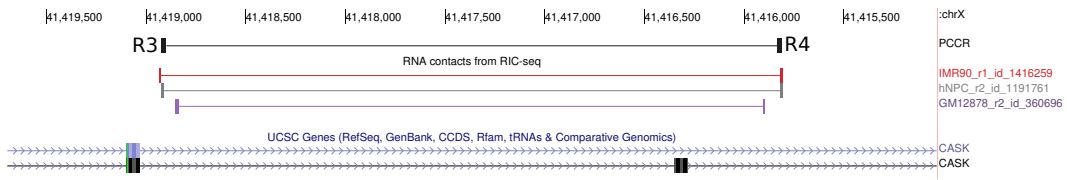


(Б) **(В)**

Рисунок 4.6 — Классификатор для предсказания раздвоенных сигналов eCLIP. **(А)** Расположение окон вокруг пары ККУ (шесть окон по 50 нт с центром в середине области спаривания). **(Б)** Рабочая операционная кривая (ROC-кривая) классификатора, предсказывающего наличие раздвоенных сигналов eCLIP в зависимости от (1) разброса ККУ, (2) числа чтений и (3) числа чтений вместе с разбросом ККУ. TPR — доля истинных положительных результатов, т.е. чувствительность. FPR — доля ложных положительных результатов, т.е. единица минус специфичность. **(С)** Важность признаков в окнах (см. панель А). Двумя наиболее важными признаками являются РНК-контакты между ячейками $1L$ и $-1R$ и между ячейками $-1L$ и $1R$, которые соответствуют внутренним и внешним контактам, непосредственно примыкающим к ККУ.



(А)



(Б)

Рисунок 4.7 — ККУ в генах *CASK* и *PHF20L1* подтверждаются данными RIC-seq. Показаны снимки из Геномного браузера UCSC. **(А)** РНК-контакты, поддерживающие ККУ в гене *PHF20L1*. **(Б)** РНК-контакты, поддерживающие ККУ в гене *CASK*. Приведены названия клеточных линий и расположение контактов.

ваниям. Среди отобранных мишеней были выбраны две: кассетный экзон 6 в гене *PHF20L1*, выпетливание которого структурой РНК подтверждалось в клеточных линиях HeLa и HepG2, а также кассетный экзон 19 в гене *CASK*, выпетливание которого структурой РНК подтверждалось в клеточных линиях IMR90, hNPC, и GM12878 (рис. 4.7). Более подробно эти мишени обсуждаются в разд. 5.2.

4.3 Структура РНК вне консервативных областей

В консервативных областях генома закодирована лишь небольшая часть всех РНК-структур, а содержащаяся в данных RIC-seq информация о РНК-контактах потенциально позволяет расширить предсказания структуры РНК за их пределы. Кроме того, из рассмотрения ранее выпадали кодирующие участки, поскольку предсказание структуры РНК в них филогенетическими методами затруднительно из-за подавления сигнала от компенсаторных замен на фоне эволюционного отбора на аминокислотную последовательность белка. В этом разделе будет описан метод PHRIC[1], который по данным RIC-seq находит пары вложенных кластеров контактов (ВКК) на всей протяженности генов, включая экзоны и целые интроны, а затем выполняет термодинамическое сворачивание последовательностей, заключенных между контактами, для нахождения комплементарных участков.

4.3.1 Вложенные кластеры РНК-контактов (ВКК)

Принцип работы метода PHRIC состоит в следующем. Исходя из предположения, что дальние взаимодействия в структуре РНК окружены внутренними и внешними РНК-контактами, которые мы наблюдали ранее в разд. 4.2, определим пары вложенных кластеров контактов (ВКК) как показано на рис. 4.8. А именно, ВКК представляет из себя пару кластеров контактов, один внутри другого, с заданным диапазоном расстояний между левыми и правыми

[1]От «Panhandle» — сковородка, т.е. пара ККУ, и «RIC-seq».

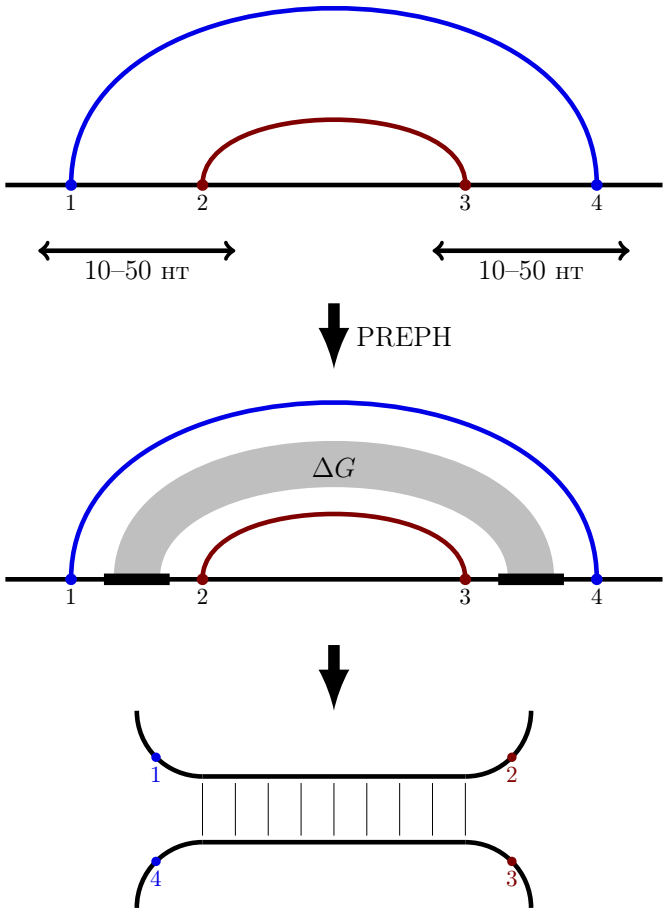


Рисунок 4.8 — Принцип работы метода PHRIC. После нахождения вложенных кластеров контактов (ВКК), нуклеотидные последовательности, расположенные между 5'-концами (1 и 2) и 3'-концами (3 и 4) внутреннего и внешнего контакта, извлекаются и передаются в программу PREPH для предсказания длинных, почти идеально комплементарных участков. Структура РНК, образованная этими участками, поддерживается РНК-контактами (1–4) и (2–3) из экспериментов конформационного секвенирования РНК.

концами. Комплементарность нуклеотидных последовательностей, заключенных между контактами, можно быстро обнаружить с помощью ядра метода PREPH (разд. 3.3). Для быстрого поиска пар ВКК можно определить два списка пар кластеров контактов A и B, таких, что левый сегмент B лежит в некотором окне от левого сегмента A, а правый сегмент B лежит в некотором окне от правого сегмента A. Тогда пересечение этих списков [428] за линейное время дает требуемый список пар ВКК, из которого после фильтрации можно извлечь потенциально гибридизующиеся нуклеотидные последовательности и передать их на вход программе PREPH. Технические детали этой процедуры можно найти в [25].

В общей сложности из экспериментов RIC-seq в семи клеточных линиях человека, содержащих 55–170 миллионов чтений на реплику, в каждой клеточной линии было получено 2–10 миллионов РНК-контактов, поддерживаемых

15–40 миллионами отдельных чтений с разрывами. Из них было получено около 35 миллионов кластеров РНК-контактов. В связи с разреженностью данных разрывные чтения из всех экспериментов объединялись. Каждый кластер РНК-контактов характеризовался набором экспериментов RIC-seq, в которых он поддерживался, и суммарным числом поддерживающих чтений. Затем были найдены пары ВКК, в которых внутренний и внешний контакт разделялись не более чем 100 нт.

Полученный набор пар ВКК был дополнительно отфильтрован для ограничения расстояния между внешним и внутренним контактами в пределах от 10 до 50 нт, при этом требовалось, чтобы они поддерживались не менее чем тремя чтениями, и исключались кластеры, пересекающие аннотированные геномные повторы. В результате было получено около 29000 пар ВКК, которые затем были переданы на вход программе PREPH. В конечном итоге было предсказано 11998 РНК-структур с отсечением по равновесной свободной энергии $\Delta G < -15$ ккал/моль.

4.3.2 Свойства поддерживаемых ВКК структур РНК

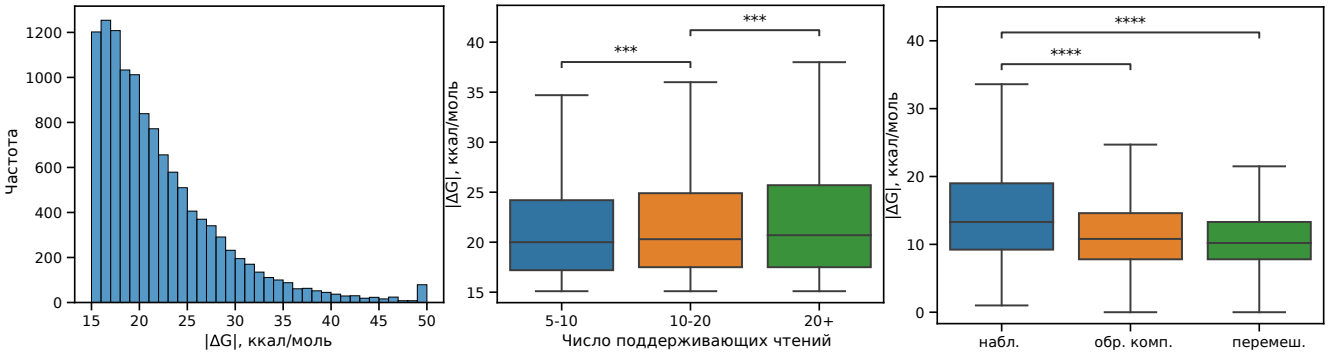
Свойства предсказанных РНК-структур без требования консервативности во многом повторяют свойства ККУ (разд. 3.3). Например, распределение свободной энергии ΔG структуры ожидаемо является убывающим (рис. 4.9А). Предсказанные структуры РНК можно разделить на три категории: интронные, в которых обе последовательности полностью располагаются в интронах, экзонные, в которых обе последовательности полностью располагаются в экзонах, и смешанные, в которые входят случаи, когда одна из последовательностей перекрывает сайт сплайсинга или экзонная последовательность контактирует с интронной. Во всех трех группах примерно 40% предсказаний поддерживались 5–10 чтениями, примерно 40% предсказаний были поддерживались 10–20 чтениями, и 20% предсказаний были поддерживались более чем 20 чтениями (табл. 5). Пороги, определяющие эти группы, являются естественными границами, разделяющими набор предсказаний на подмножества примерно одинакового размера с увеличением уровня поддержки чтениями, далее обозначаемого через r. Примечательно, что более 70% предсказанных структур располагаются

Таблица 5 — Количество РНК-структур в экзонных, интронных и смешанных областях в группах по поддержке чтениями. Проценты приведены по отношению к общему количеству в строке.

Класс	Всего	5–10	10–20	>20
Экзонные	3676	1317 (36%)	1634 (44%)	725 (20%)
Интронные	7132	2425 (34%)	2943 (41%)	1764 (25%)
Смешанные	1190	489 (41%)	499 (42%)	202 (17%)
Всего	11998	4231 (35%)	5076 (42%)	2691 (22%)

за пределами консервативных геномных элементов из выравнивания 100 позвоночных.

Естественно ожидать, что пары ВКК с более высокой поддержкой чтениями содержат более стабильные структуры РНК. Действительно, медианное абсолютное значение ΔG увеличивалось с ростом числа поддерживающих чтений (рис. 4.9Б), но величина этого увеличения была очень мала (в среднем 0.03 ккал/моль с каждым дополнительным поддерживающим чтением). Для ответа на вопрос, является ли наблюдаемое распределение свободной энергии неслучайным, использовался контроль по принципу «пересоединения», в котором соответствие между последовательностям в парах ВКК было случайным образом перемешано, а также отдельно оценивалась свободная энергия взаимодействия в парах, в которых одна из последовательностей была заменена на обратно комплементарную. Второй способ был необходим для поддержания частот динуклеотидов, которые значительно влияют на величину ΔG. Как в перемешанных парах, так и в контроле с обратной комплементарностью медианные значения ΔG оказались существенно ниже, чем исходные (рис. 4.9В). Таким образом, представленный метод предсказывает более стабильные структуры, чем можно было бы ожидать по случайным причинам, и их стабильность коррелирует с поддержкой чтениями.



(А) **(Б)** **(В)**

Рисунок 4.9 — Свойства структур РНК, предсказанных по методу PHRIC. **(А)** Распределение равновесных свободных энергий (ΔG). **(Б)** Равновесная свободная энергия (ΔG) слабо возрастает с увеличением степени поддержки чтениями. **(В)** Наблюдаемые РНК-структуры имеют в среднем бóльшую по абсолютной величине равновесную свободную энергию (ΔG), чем РНК-структуры в перемешанном наборе или РНК-структуры, образованные между интервалом 1-2 и обратно комплементарным к интервалу 3-4. Символы *** и **** обозначают статистически значимые различия на уровне значимости 0.1% и 0.01%, соответственно.

4.3.3 Структуры РНК в экзонах и интронах

Предыдущие исследования вторичной структуры РНК и ее взаимодействия РСБ показали, что интроны, как правило, более структурированы, чем экзоны [429—431]. Однако в этих исследованиях оценивалась склонность оснований РНК к образованию локальной структуры, а дальние взаимодействия не рассматривались. Предсказанные PHRIC структуры РНК можно подразделить, с одной стороны, на четыре энергетические группы и, с другой стороны, на группы с высокой ($r \geqslant 12$) и низкой ($r < 12$) поддержкой чтениями. Отметим, что порог $r = 12$ равен медиане распределения поддержки чтениями.

Таблица 6 — Количество РНК-структур в экзонных, интронных и смешанных областях по группам свободной энергии (15–20 ккал/моль, 20–25 ккал/моль, 25–30 ккал/моль и >30 ккал/моль по абсолютной величине). Проценты приведены по отношению к общему количеству в строке.

Класс	15–20	20–25	25–30	>30
Экзонные	2055 (56%)	989 (27%)	420 (11%)	212 (6%)
Интронные	3189 (45%)	1949 (27%)	1042 (15%)	952 (13%)
Смешанные	570 (48%)	349 (29%)	165 (14%)	106 (9%)

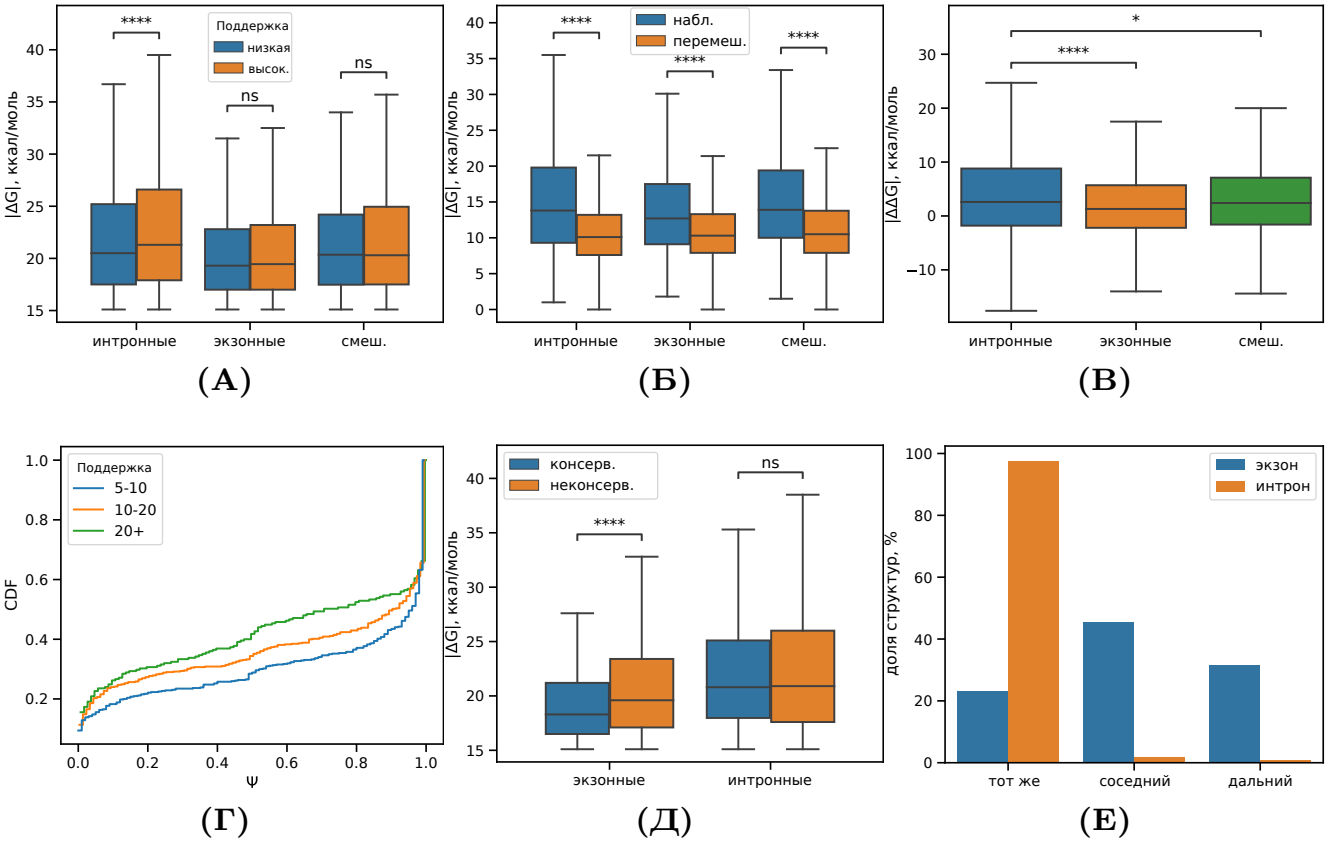


Рисунок 4.10 — Структура РНК в экзонах и интронах. **(А)** Распределение свободных энергий (ΔG) структур РНК, окруженных контактами с высокой ($r \geqslant 12$) и с низкой поддержкой чтениями ($r < 12$). **(Б)** Распределение свободных энергии наблюдаемых интронных, экзонных и смешанных структур РНК в сравнении с пересоединенным контролем. **(В)** Распределение значений $\Delta\Delta G = \Delta G_{obs} - \Delta G_{RC}$, где ΔG_{obs} - свободная энергия наблюдаемой структуры РНК, а ΔG_{RC} — свободная энергия структуры, в которой одна из последовательностей была заменена на обратно комплементарную. **(Г)** Изменение средней степени включения (Ψ) экзонов, выпетливаемых предсказанными РНК-структурами, при увеличении поддержки чтениями. **(Д)** Свободные энергии экзонных и интронных структур РНК, расположенных в консервативных и неконсервативных областях. **(Е)** Доли интронных/экзонных РНК-структур в пределах одного интрона/экзона (тот же), соседних, т.е. следующих друг за другом интронов/экзонов (соседний), и удаленных, т.е. не следующих друг за другом интронов/экзонов (дальний). Символы *, **** и ns обозначают статистически значимые различия на уровне значимости 5%, 0.01% и не значимые различия, соответственно.

Предсказанные интронные структуры РНК характеризуются большей долей двухцепочечных участков со свободной энергией, превышающей по абсолютной величине 25 ккал/моль (табл. 6). Кроме того, интронные структуры РНК с высокой поддержкой чтениями имеют значительно бóльшие значения ΔG, чем структуры с низкой поддержкой чтениями ($P < 0.1\%$), в то время как в экзонных и смешанных группах разница между группами с высокой и низкой поддержкой чтения не значима (рис. 4.10А). Сравнение значений ΔG в наблюдаемых структурах и в контрольном наборе показывает, что свободные энергии во всех трех группах (экзонные, интронные и смешанные) достоверно

больше по сравнению с контрольными ($P < 0.1\%$), причем для интронной группы разница больше, чем для экзонной (рис. 4.10Б). Второй контроль позволяет сравнить наблюдаемые значения свободной энергии (ΔG_{obs}) со значениями в наборе, в котором одна из последовательностей была заменена на обратно комплементарную (ΔG_{RC}), с помощью метрики $\Delta\Delta G = \Delta G_{obs} - \Delta G_{RC}$. И в этом случае интронные структуры оказались более стабильными по отношению к обратно комплементарному контролю по сравнению с экзонными и смешанными структурами (рис. 4.10В), что еще раз подтверждает бóльшую склонность интронов к образованию РНК-структур между собой.

Как и ранее, экзоны, которые выпетливаются РНК-структурами, имеют в среднем меньшую степень включения в клеточных линиях, чем экзоны, не окруженные структурами, причем степень включения уменьшается с увеличением поддержки чтениями (рис. 4.10Г). Неконсервативные интронные РНК-структуры так же стабильны, как и консервативные, а неконсервативные экзонные РНК-структуры в среднем даже более стабильны, чем консервативные (рис. 4.10Д). Это говорит о том, что экзонные последовательности, не имеющие ограничений на поддержание аминокислотной последовательности, могут эволюционировать в более стабильные РНК-структуры. Наконец, если разделить интронные и экзонные РНК-структуры на три класса, соответствующие взаимодействию внутри одного интрона или экзона, соседних (т.е. следующих друг за другом) интронов и экзонов и удаленных (т.е. не следующих друг за другом) интронов и экзонов, то экзонные РНК-структуры распределяются между этими группами с почти равными вероятностями, в то время как интронные РНК-структуры обычно расположены в одном и том же интроне (рис. 4.10Е). С одной стороны, наблюдаемая закономерность может быть связана с тенденцией к образованию РНК-структур в пределах одного интрона (разд. 3.3.6). С другой стороны, она также может быть вызвана тем, что РНК-контакты образуются на расстояниях, не превышающих длины интрона, и в определенной мере отражают конформацию РНК в процессе сплайсинга.

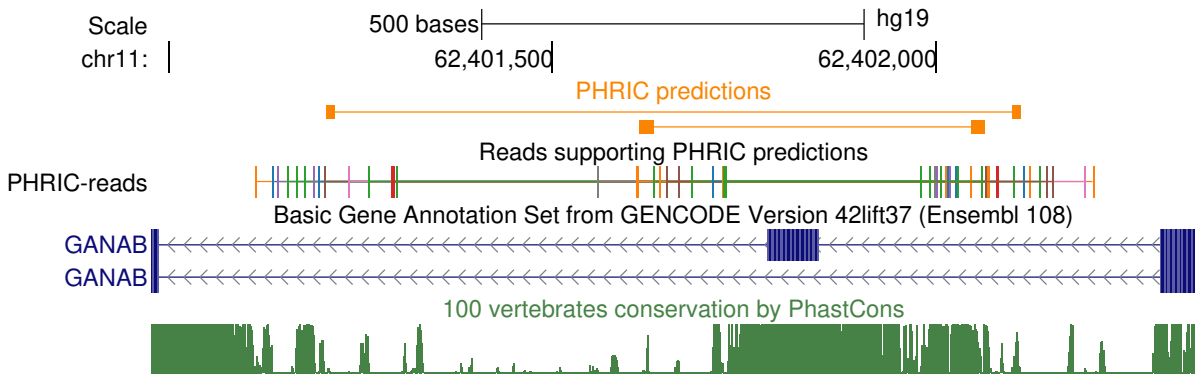
4.3.4 Примеры структур РНК в неконсервативных областях

Для визуализации результатов предсказаний PHRIC можно использовать специализированный интерфейс для Геномного браузера UCSC, называемый track hub [25]. На рис. 4.11 представлены несколько примеров структур РНК из списка предсказаний PHRIC, визуализированные с помощью этого инструмента.

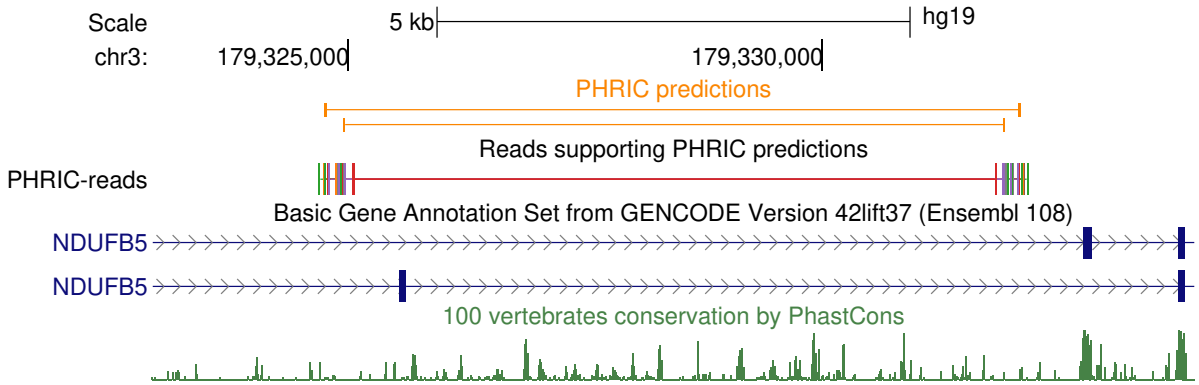
Ген *GANAB* кодирует субъединицу глюкозидазы IIα и связан с аутосомно-доминантным поликистозом почек и печени [432; 433]. Один из его внутренних экзонов, экзон 6, сплайсируется альтернативно. PHRIC обнаружил две поддерживаемые большим количеством чтений пары ВКК вокруг этого экзона и предсказал две пары комплементарных областей, образующих РНК-структуры со свободными энергиями -26.3 ккал/моль и -22.1 ккал/моль, соответственно (рис. 4.11А). Другим примером является структура РНК в гене *NDUFB5*, который кодирует субъединицу NADH-убихиноновой оксидоредуктазы. Для этого гена обнаружено три транскрипта, кодирующие различные изоформы, две из которых отличаются альтернативным включением экзона 7. PHRIC предсказал две пары комплементарных интронных последовательностей внутри ВКК со свободными энергиями -28.8 ккал/моль и -22.1 ккал/моль, соответственно (рис. 4.11Б). Наконец, в гене *ZNF655*, который участвует в регуляции транскрипции и связан с прогрессией рака поджелудочной железы [434], PHRIC обнаружил пару ВКК, способных образовывать дуплекс со свободной энергией -25.9 ккал/моль и выпетливать кассетный экзон 4 (рис. 4.11В). Во всех перечисленных примерах комплементарные области располагаются вне консервативных областей ста позвоночных.

4.3.5 Эволюционные подписи структур РНК вне консервативных областей

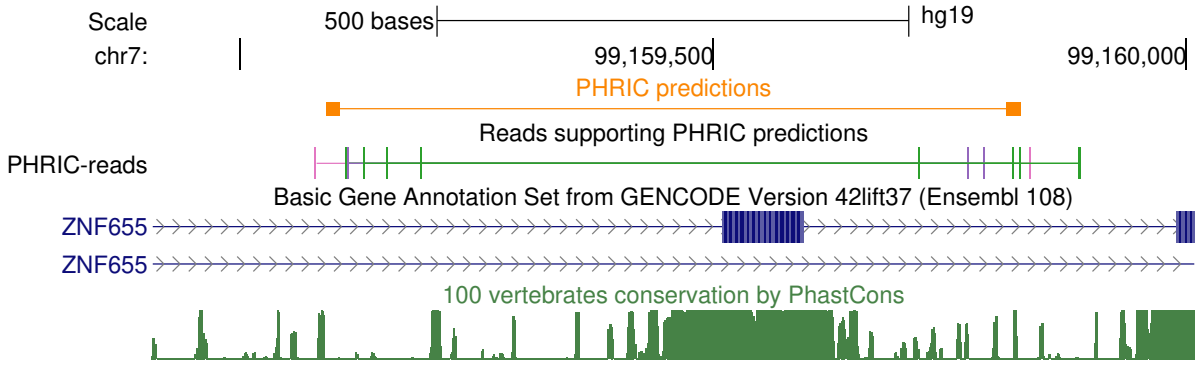
Построенный с помощью метода PREPH каталог пар ККУ ограничен так называемыми консервативными элементами РНК, которые были получены в результате выравнивания последовательностей ста геномов позвоночных по ме-



(А)



(Б)



(В)

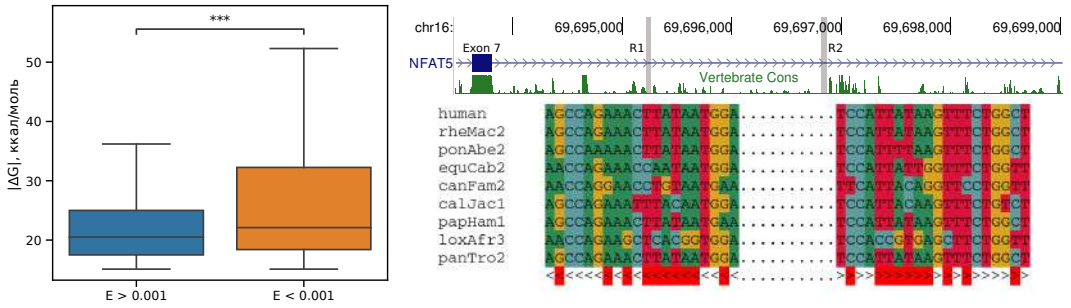
Рисунок 4.11 — Примеры интронных структур РНК в генах человека за пределами консервативных областей. Приводятся снимки из геномного браузера UCSC. Трек эволюционной консервативности по 100 позвоночным показан зеленым цветом. Предсказания PHRIC, подтвержденные в экспериментах RIC-seq, показаны оранжевым цветом. Внутренние и внешние контакты из RIC-seq показаны в треке PHRIC-reads. **(А)** Ген *GANAB*, субъединица глюкозидазы IIα. **(Б)** Ген *NDUFB5*, субъединица NADH-убихиноновой оксидоредуктазы. **(В)** Ген *ZNF655*, транскрипционный регулятор с доменом типа «цинковый палец».

тоду phylo-HMM [372; 435]. Генеративная вероятностная модель этого метода содержит состояния для консервативных сайтов и состояния для неконсервативных сайтов, а переходы между ними определяют границы консервативных элементов РНК. При этом границы и сам набор консервативных элементов РНК различаются в зависимости от числа и набора анализируемых геномов.

Если задаться вопросом, можно ли извлечь эволюционные подписи структур РНК непосредственно из множественных выравниваний геномных последовательностей, ограничившись меньшим количеством геномов, то в блоках выравнивания, соответствующим предсказаниям метода PHRIC, можно найти компенсаторные замены. С этой целью были проанализированы пары блоков выравнивания, которые вырезаются предсказанными комплементарными участками из множественного выравнивания геномов 46 млекопитающих. Столбцы, содержащие более 80% пропусков, и строки, содержащие более 10% пропусков, были удалены. Как и ранее, полученные пары были поданы на вход программе R-scape v1.2.340 [23] после соединения через спейсер вместе с ограниченным филогенетическим деревом. Из 11998 структур, первоначально предсказанных PHRIC, 11224 имели по крайней мере одну пару оснований с $E < 1$, и только 308 пар имели значимые компенсаторные замены ($E < 0.05$) после поправки Бенджамини-Хохберга на множественное тестирование.

Как и ранее, структуры РНК со значимыми компенсаторными заменами имели значимо бóльшую по абсолютной величине равновесную свободную энергию (ΔG), чем структуры РНК без значимых ковариаций ($P < 0.001$) (рис. 4.12А). Медианные длины областей спаренных оснований существенно не различались между этими двумя наборами ($P = 0.15$), что исключает возможность того, что более длинные структуры РНК вносят вклад одновременно в оба значения ΔG и E.

В гене *NFAT5*, члене семейства транскрипционных факторов активированных Т-клеток, обнаружена структура РНК с множеством независимых компенсаторных замен ($E = 1.6 \cdot 10^{-14}$). Этот ген играет центральную роль в индуцируемой транскрипции генов во время иммунного ответа. Интрон, расположенный между экзонами 7 и 8 *NFAT5*, содержит пару комплементарных последовательностей, R1 и R2, которые поддерживаются чтениями RIC-seq, но выходят за рамки консервативных элементов РНК у позвоночных (рис. 4.12Б). В множественном выравнивании последовательностей гомологов у млекопитающих участки, соответствующие R1 и R2, отсутствуют у грызунов,



(А) **(Б)**

Рисунок 4.12 — Эволюционные подписи вне консервативных областей позвоночных. **(А)** Равновесные свободные энергии (ΔG) существенно больше по абсолютной величине для РНК-структур со значимыми компенсаторными заменами ($E < 0.001$), чем для остальных РНК-структур ($E \geqslant 0.001$). **(Б)** Фрагмент гена *NFAT5* между экзонами 7 и 8; R1 и R2 — комплементарные последовательности, предсказанные PHRIC, которые выходят за пределы консервативных элементов РНК у позвоночных (вверху). Множественное выравнивание последовательностей R1 и R2 и консенсусная структура РНК (внизу). Компенсаторные замены отмечены красным цветом.

но присутствуют у приматов, канид и африканских слонов с множественными независимыми компенсаторными заменами. Примечательно, что структура, образованная R1 и R2, вложена в более крупную консервативную структуру с $\Delta G = -40.1$ ккал/моль, которая была предсказана для этого гена ранее [23].

4.4 Обсуждение результатов и выводы

4.4.1 РНК-контакты и предсказания ККУ

Лежащий в основе технологии RIC-seq принцип, основанный на разрезании и стохастическом лигировании РНК с последующим высокопроизводительным секвенированием химерных молекул, имеет много общего с принципом работы метода Hi-C для идентификации пространственных контактов между ДНК в трехмерной структуре хроматина [436—438]. Лигирование пространственно близких молекул дает информацию о структуре и функциональных структурных элементах РНК. Например, визуализация данных PARIS и данных RIC-seq позволила создать Hi-C-подобные карты связности для различных РНК, которые были названы структурными или топологическими доменами [21; 439], а Ли и др. реализовали итеративный алгоритм, основанный на

данных PARIS, для поиска оптимального иерархического разделения больших РНК и с его помощью успешно разделили РНК вируса Зика на десятки структурных доменов [439].

Однако по сравнению с Hi-C у RIC-seq есть два существенных отличия. Во-первых, поскольку пре-мРНК подвергаются сплайсингу, химерные молекулы РНК могут содержать не только продукты лигирования, но и ЭЭС, что вызывает определенные трудности при картировании получаемых чтений на геном. Эта проблема решается с помощью вычислительного конвейера RNAcontacts, разработанного в рамках данной диссертационной работы, который картирует чтения с двумя типами разрывов в двухпроходном режиме (разд. 4.1). Во-вторых, при анализе данных Hi-C обычной практикой является усреднение хроматиновых контактов в масштабе килобаз, в то время как для исследования РНК-контактов с помощью RIC-seq требуется однонуклеотидное разрешение. При этом уровень согласованности между биорепликами составляет всего 9–15%, поддержка чтениями RIC-seq для известных из литературы РНК-структур крайне мала, а подавляющее большинство контактов поддерживаются лишь одним чтением. Это естественно приводит к вопросу об оценке статистической значимости контактирующих кластеров, на который в настоящее время нет ответа.

Вместе с тем, сравнение дальних взаимодействий в структуре РНК, предсказанных методом PREPH, и РНК-контактов, полученных по объединению экспериментов RIC-seq в различных клеточных линиях, выявило ожидаемые общие тенденции, заключающиеся в увеличении стабильности структуры, значимости компенсаторных замен, встречаемости сайтов редактирования РНК и раздвоенных сигналов eCLIP по мере увеличения степени поддержки РНК-контактов чтениями. Для выпетливаний экзонов некоторые различия между биологическими условиями, обусловленные РНК-контактами, все же удалось зарегистрировать (рис. 4.5Б). Однако следует помнить, что эти контакты не обязательно соответствуют структуре РНК, а скорее подтверждают, что две цепи расположены близко друг от друга, возможно, из-за внутри- и межмолекулярных взаимодействий РНК-РНК или, возможно, из-за взаимодействий, опосредованных РСБ. Например, факторы сплайсинга hnRNPA1 и PTB, связанные с пре-мРНК, предпочитают образовывать димеры, создавая таким образом контакты, которые не вызываются комплементарными спариваниями [440; 441]. При исследовании контактов в классах структурированных РНК, таких как предшественники микроРНК, было обнаружено, что большинство структур под-

держиваются только внутренними контактами, соответствующими апикальной петле шпильки, но не внешними контактами. Тем не менее, наложение жестких ограничений на поддержку чтениями позволило предсказать пары функциональных ККУ в генах *PHF20L1* и *CASK*, которые в значительной степени поддерживаются как внутренними, так и внешними контактами.

Ранее было высказано предположение о том, что дальние взаимодействия в структуре РНК сближают 5'ss и 3'ss, тем самым облегчая распознавание интрона [442]. Здесь показывается, что это не совсем так: экзоны, которые выпетливаются парами ККУ с поддержкой RIC-seq, фланкируются более короткими интронами, в которых ККУ расположены ближе к пропускаемому экзону, а их сайты сплайсинга напоминают сайты интронов, которые сплайсируются посттранскрипционно. Это наблюдение, а также снижение частоты включения экзонов с увеличением поддержки чтениями RIC-seq вместе позволяют предположить, что экзоны, выпетливаемые ККУ, подвергаются посттранскрипционному сплайсингу, иначе бы их фланкирующие интроны были бы сплайсированы по принципу «первый пришел, первым обслужен» еще до того, как структура РНК успела собраться [443]. Еще одно примечательное наблюдение заключается в том, что ядовитые экзоны имеют тенденцию выпетливаться более стабильными РНК-структурами, что позволяет предположить, что регуляторные механизмы, контролирующие экспрессию генов посредством непродуктивного сплайсинга, могут в значительной степени зависеть от структуры РНК. В разд. 6.3 будет показано, что это действительно так.

4.4.2 О предсказании глобальной структуры РНК

Закономерности в поддержке пар ККУ контактами изнутри и снаружи, которые наблюдались при сопоставлении предсказаний PREPH с данными RIC-seq, можно интерпретировать не только как свойства, но и как требования, которые следует предъявить к потенциальным структурам РНК, постулируя двустороннюю поддержку контактами как критерий поиска. Основанный на этой идее метод PHRIC сначала определяет пары ВКК, а затем выясняет сворачиваются ли заключенные внутри них последовательности во вторичную структуру. Его важной особенностью является то, что он не ограничен тре-

бованием консервативности, что расширяет пространство поиска за пределы эволюционно консервативных участков и позволяет рассматривать не только интроны, но и экзоны. При этом наблюдаются те же закономерности, что и для PREPH, а именно зависимость стабильности структуры, наличия сайтов редактирования РНК, раздвоенных сигналов eCLIP, а также частоты включения выпетливаемых экзонов от поддержки чтениями. Интересно, что эти результаты применимы не только к альтернативным, но и к конститутивным событиям сплайсинга. Лишь небольшая часть структур РНК (менее 3%) охватывает соседние интроны, что примерно соответствует доле событий AC, которые фактически экспрессируются в клеточных линиях человека.

Признаком наличия биологической функции у цис-регуляторных элементов является эволюционная консервативность последовательности, которая в случае структуры РНК коррелирует со свободной энергией гибридизации. Можно было бы ожидать, что консервативные структуры РНК будут более термодинамически стабильными, чем неконсервативные. Однако, оказывается, что это не так: интроны генов человека содержат почти в 20 раз больше структур РНК в неконсервативных областях по сравнению с консервативными, но комплементарные спаривания в неконсервативных областях не менее стабильны, чем в консервативных. При этом структуры, не консервативные среди позвоночных, тем не менее могут иметь значимые компенсаторные замены, если рассматривается меньшее множество видов.

Разреженность данных RIC-seq не позволяет сделать точные выводы о поддержке отдельных структурных элементов в транскриптах. Однако колоссальный объем информации, содержащийся в данных RIC-seq, заставляют переформулировать задачу предсказания дальних взаимодействий в контексте глобальной структуры. Центральный вопрос этой задачи заключается в том, как объединить филогенетическую информацию, предсказания комплементарности и экспериментальные данные об РНК-контактах и связывании РСБ в единую модель с учетом динамической изменчивости структуры и протекающих одновременно со сворачиванием процессов созревания пре-мРНК. По мере накопления данных конформационного секвенирования РНК *in situ*, необходимость в переходе к этой глобальной формулировке будет возрастать.

Глава 5. Экспериментальная валидация влияния структуры РНК на АС

В этой главе приводятся результаты, полученные мной в период с 2009 по 2021 год в сотрудничестве с лабораториями под руководством проф. Хуана Валкарселя (Центр Геномной Регуляции, г. Барселона), проф. П.М. Рубцова (Институт Молекулярной Биологии им. Энгельгардта РАН) и проф. О.А. Донцовой (Московский Государственный Университет им. М.В. Ломоносова) [24; 79; 107; 108].

Для экспериментальной валидации влияния предсказанных структур РНК на АС использовался классический метод двойных мутантов. А именно, фрагменты генов встраивали в минигены с индуцируемым промотором, а также с помощью сайт-направленного мутагенеза создавали два варианта минигенов, в каждом из которых комплементарность в структуре РНК нарушалась, а при одновременном внесении обеих мутаций — восстанавливалась. Использовались различные стратегии мутагенеза, от дестабилизирующих структуру точечных мутаций до полной замены последовательности. Для скрининга предполагаемых структур использовались блокирующие комплементарность антисмысловые олигонуклеотиды, изготовленные на основе LNA (locked nucleic acid, замкнутая нуклеиновая кислота, или «недоступная РНК») с заменой оснований ДНК в каждом втором нуклеотиде. LNA содержит модифицированные нуклеотиды РНК, в которых фрагмент рибозы модифицирован дополнительным мостиком, соединяющим 2'-кислород с 4'-углеродом. Эта структура обеспечивает повышенную устойчивость к ферментативному расщеплению и увеличивает прочность дуплекса [444—446]. Изменения сплайсинга регистрировали качественно с помощью ОТ-ПЦР и гель-электрофореза, а также количественно с помощью ОТ-ПЦР-РВ (см. разд. 2.1.2).

В общей сложности были валидированы влияющие на АС структуры РНК в десяти генах — в трех генах дрозофилы и в семи генах человека. Список этих структур со ссылками на соответствующие разделы диссертации и публикации, а также указание методов, при помощи которых структура была предсказана или подтверждена, приводится в табл. 7. Структуры РНК в генах *BRD2* и *BRD3* имеют отношение к непродуктивному сплайсингу и поэтому обсуждаются в следующей главе.

Таблица 7 — Подтверждение регуляции АС структурой РНК. DM – *D. melanogaster*; HS – *H. sapiens*; знаки + и − обозначают, что структура была предсказана или подтверждена соответствующим методом.

Вид	Ген	Событие АС	IRBIS	PREPH	RIC-seq	Раздел	Публ.
DM	CG33298	Альт. 5'ss	+	+		5.1.1	[107]
DM	Gug	Альт. 3'ss	+	+		5.1.2	[107]
DM	Nmnat	Альт. термин. экзон	+	+		5.1.3	[107]
HS	SF1	Удержание интрона	+	+		5.3	[108]
HS	PHF20L1	Кассетный экзон	+	+	+	5.2.1	[24]
HS	CASK	Кассетный экзон	+	+	+	5.2.2	[24]
HS	ATE1	Взаимоискл. экзоны	+	+	+/−	5.2.3	[79]
HS	MARK2	Кассетный экзон	+	+	+	5.3	[447]
HS	BRD2	Ядовитый экзон	+	+	+	6.3	[351]
HS	BRD3	Ядовитый экзон	+	−	−	6.3	[351]

5.1 Структуры РНК в генах насекомых

В этом разделе приводятся результаты экспериментальной валидации влияния структуры РНК на АС в трех генах насекомых, а именно на выбор альтернативного донорного сайта в гене *CG33298*, выбор альтернативного акцепторного сайта в гене *Gug*, и пропуск терминального экзона в гене *Nmnat*. Для всех этих случаев были сконструированы минигены с разрушающими структуру одиночными и двойными компенсаторными мутациями (разд. 2.1.1); АОН не использовались.

5.1.1 Выбор альтернативного донорного сайта в гене *CG33298*

Ген *CG33298* предположительно кодирует фосфолипид-флиппазу, которая поддерживает асимметрию мембран в аппарате Гольджи и их холестериновый состав [448]. Экзон 13 этого гена имеет два альтернативных донорных сайта, переключение между которыми приводит к изменению С-концевого домена кодируемого белка. Предсказанные ККУ длиной 19 нт, называемые «бокс 1» и «бокс 2», расположены так, что бокс 1 перекрывается с проксимальным донорным сайтом и отстоит от бокса 2 на 185 нт (рис. 5.1А). Дистальный и прок-

симальный донорные сайты имеют почти одинаковую силу, но и в эндогенной пре-мРНК, и в транскриптах минигена в основном используется дистальный сайт (рис. 5.1Б).

При конструировании минигена в ККУ были введены четыре точечные мутации, которые разрушают комплементарное спаривание, но не изменяют консенсусную последовательность донорного сайта (рис. 5.1В). Внесение каждой из них по отдельности приводит к почти полному переключению сплайсинга на проксимальный донорный сайт, а двойная мутация, в которой структура РНК оказывается восстановлена, приводит к обратному переключению на дистальный донорный сайт, т.е., к восстановлению сплайсинга дикого типа (рис. 5.1Б). Таким образом, мы приходим к выводу, что именно формирование структуры РНК, а не сигналы в консервативных последовательностях являются основным фактором, определяющим использование донорного сайта. Следует также отметить, что в двойном мутанте расчетная энергия гибридизации оказалась на 4.3 ккал/моль выше, чем в диком типе, и при этом проксимальный донорный сайт оказался намного сильнее подавлен. Таким образом, более энергетически стабильная структура более эффективно подавляет активность альтернативного донорного сайта.

5.1.2 Выбор альтернативного акцепторного сайта в гене *Gug*

Ген атрофина (также называемый *Grunge* и *Gug*) кодирует корепрессор транскрипции с активностью гистондеацетилазы [449]. Проведенное в работе 2009 года исследование выявило наличие неаннотированного на момент публикации статьи акцепторного сайта, консервативной новой экзонной области и пары комплементарных последовательностей, одна из которых перекрывается со сплайс-сайтом (рис. 5.2А). ОТ-ПЦР-анализ эндогенной мРНК показал, что проксимальный акцепторный сайт действительно используется. Несмотря на то, что проксимальный акцептор сильнее дистального по консенсусной последовательности, как в транскриптах минигена, так и в эндогенной мРНК оба акцепторных сайта используются приблизительно в соотношении 1:1 (рис. 5.2Б). Как и в случае с геном *CG33298* (разд. 5.1.1), мутации внутри каждой из двух комплементарных последовательностей приводили к появле-

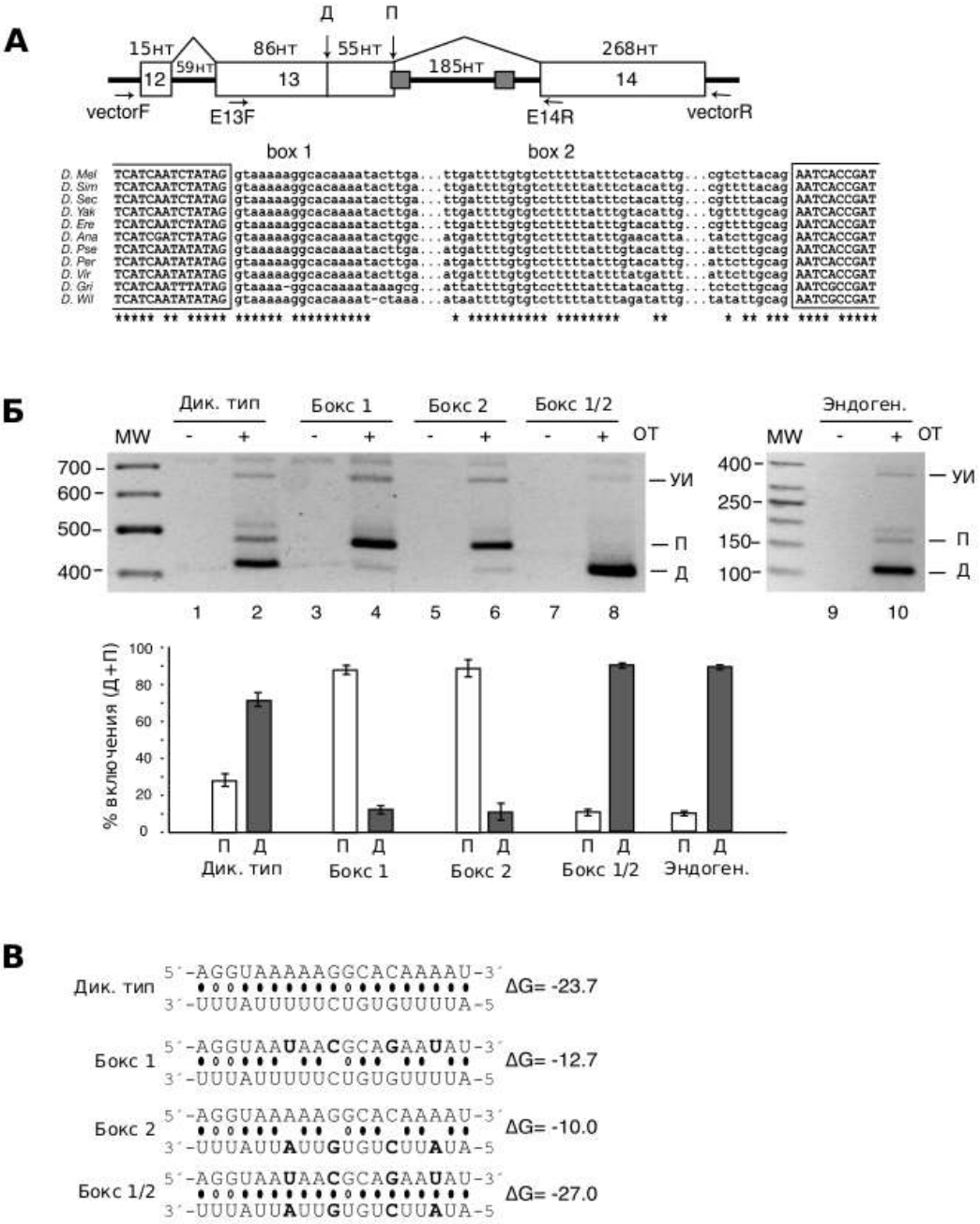


Рисунок 5.1 — Сплайсинг интрона с альтернативными донорными сайтами в гене *CG33298* регулируется структурой РНК. **(А)** Верхняя панель: схема минигена, содержащего фрагмент гена *CG33298*. ККУ (Бокс 1 и Бокс 2) выделены серым цветом. Стрелками показано расположение праймеров для ПЦР-амплификации мРНК минигена и эндогенных мРНК. Нижняя панель: множественное выравнивание фрагментов генома между экзонами 13 и 14. Д — дистальный, П —проксимальный донорные сайты. **(Б)** Вторичная структура, образованная бокс 1 и бокс 2, влияет на использование донорного сайта. Бокс 1 и бокс 2 соответствуют одиночным мутантам (см панель В), а бокс 1/2 — компенсаторному двойному мутанту. Показаны продукты амплификации транскриптов минигена (дорожки 1–8) и эндогенной мРНК (дорожки 9–10) с помощью ОТ-ПЦР и ОТ-ПЦР-РВ. Полосы на геле соответствуют двум донорным сайтам и удержанию интрона (УИ). Знаки + и − обозначают добавление или отсутствие обратной транскриптазы (ОТ) в реакцию. Столбики и усики на нижней панели показывают среднее трех реплик и стандартную ошибку. **(В)** Предсказанные структуры РНК для мутантов и дикого типа (точечные мутации показаны жирным шрифтом) и их равновесные свободные энергии (ΔG, ккал/моль). Сокращения в названиях видов как в [373].

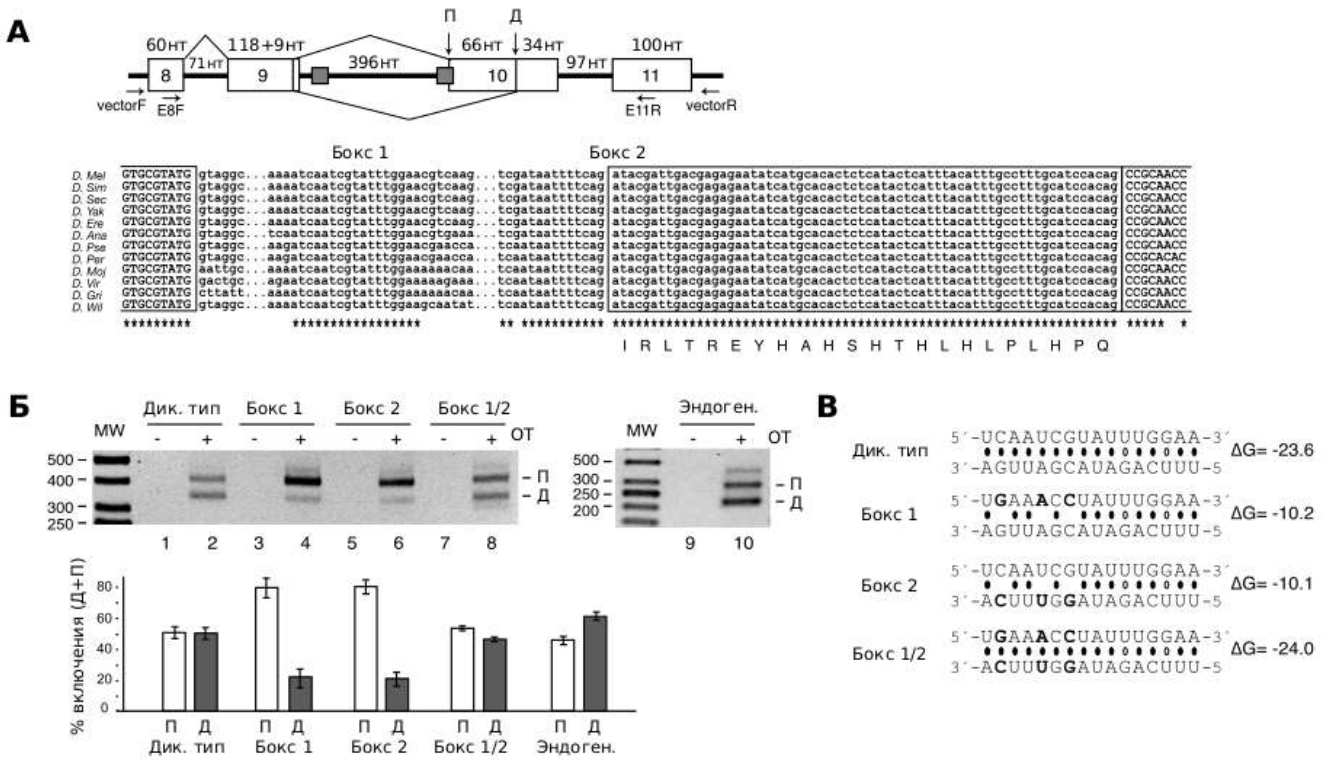


Рисунок 5.2 — Сплайсинг интрона в гене атрофина (*Gug*) с альтернативными акцепторными сайтами регулируется структурой РНК. **(А)** Верхняя панель: схема минигена, содержащего фрагмент последовательности между экзонами 8 и 11. ККУ (Бокс 1 и Бокс 2) выделены серым цветом. Нижняя панель: множественное выравнивание интронных последовательностей; под ней приводится аминокислотная последовательность альтернативной части экзона 10. Обозначения на панелях **(Б)** и **(В)** как на рис. 5.1.

нию продуктов сплайсинга с использованием главным образом проксимального акцепторного сайта, а при восстановлении структуры РНК в двойном мутанте наблюдалась картина сплайсинга дикого типа, причем со схожим соотношением изоформ. Это наблюдение хорошо согласуется с тем, что расчетные энергии гибридизации в диком типе и в двойном мутанте приблизительно одинаковы (рис. 5.2В). Из этого можно заключить, что структура РНК подавляет проксимальный акцепторный сайт, тем самым уравнивая два сплайс-сайта различной силы за счет включения более сильного сайта в структуру.

5.1.3 Пропуск терминального экзона и альтернативное полиаденилирование в гене *Nmnat*

Ген *Nmnat* дрозофилы кодирует никотинамидмононуклеотид-аденилтрансферазу, обратимо катализирующую последний этап пути восстановления никотинамидадениндинуклеотида (НАД) во всех живых организмах [450]. Он экспрессируется в виде двух сплайс-изоформ, которые транслируются в изоформы белка с разными нейропротекторными свойствами [451]. Одна из них локализована в ядре и обладает минимальной нейропротекторной способностью, а другая является цитоплазматической и обладает высокой нейропротекторной способностью. При стрессе в нейронах преимущественно образуется нейропротекторная изоформа.

Предсказание структуры РНК показало, что ген *Nmnat* содержит пару ККУ, которые окружают проксимальный альтернативный акцепторный сайт в интроне 4 (рис. 5.3А), использование которого приводит к включению в транскрипт терминального экзона с сайтом полиаденилирования, а использование дистального акцепторного сайта включает внутренний и несколько следующих за ним экзонов. Исследование изоформ, содержащих дистальный акцепторный сайт, который преимущественно используется в эндогенной мРНК (рис. 5.3Г) показало, что полная замена нуклеотидной последовательности бокса 1 или бокса 2, нарушающая комплементарность, резко снижает уровень включения экзона с дистальным акцептором (рис. 5.3Б), а восстановление структуры с помощью новой последовательности (рис. 5.3Д) обращает этот эффект.

ОТ-ПЦР изоформ, содержащих проксимальный акцепторный сайт с использованием праймера, специфичного для экзона 5 показывает, что уровень его использования увеличивается при разрушении комплементарности и снова уменьшается с образованием новой структуры (рис. 5.3В). Таким образом, вторичная структура РНК регулирует не только АС, но и альтернативное полиаденилирование. Механистически ее действие можно объяснить двояко. Во-первых, поскольку проксимальный акцепторный сайт является более сильным из двух сайтов, помещение в петлю делает его менее конкурентоспособным. Во-вторых, дистальный акцепторный сайт расположен более чем в 400 нт от проксимального, что делает длину сплайсируемого интрона намного большей, чем средняя длина интрона дрозофилы. Формирование структуры комплемен-

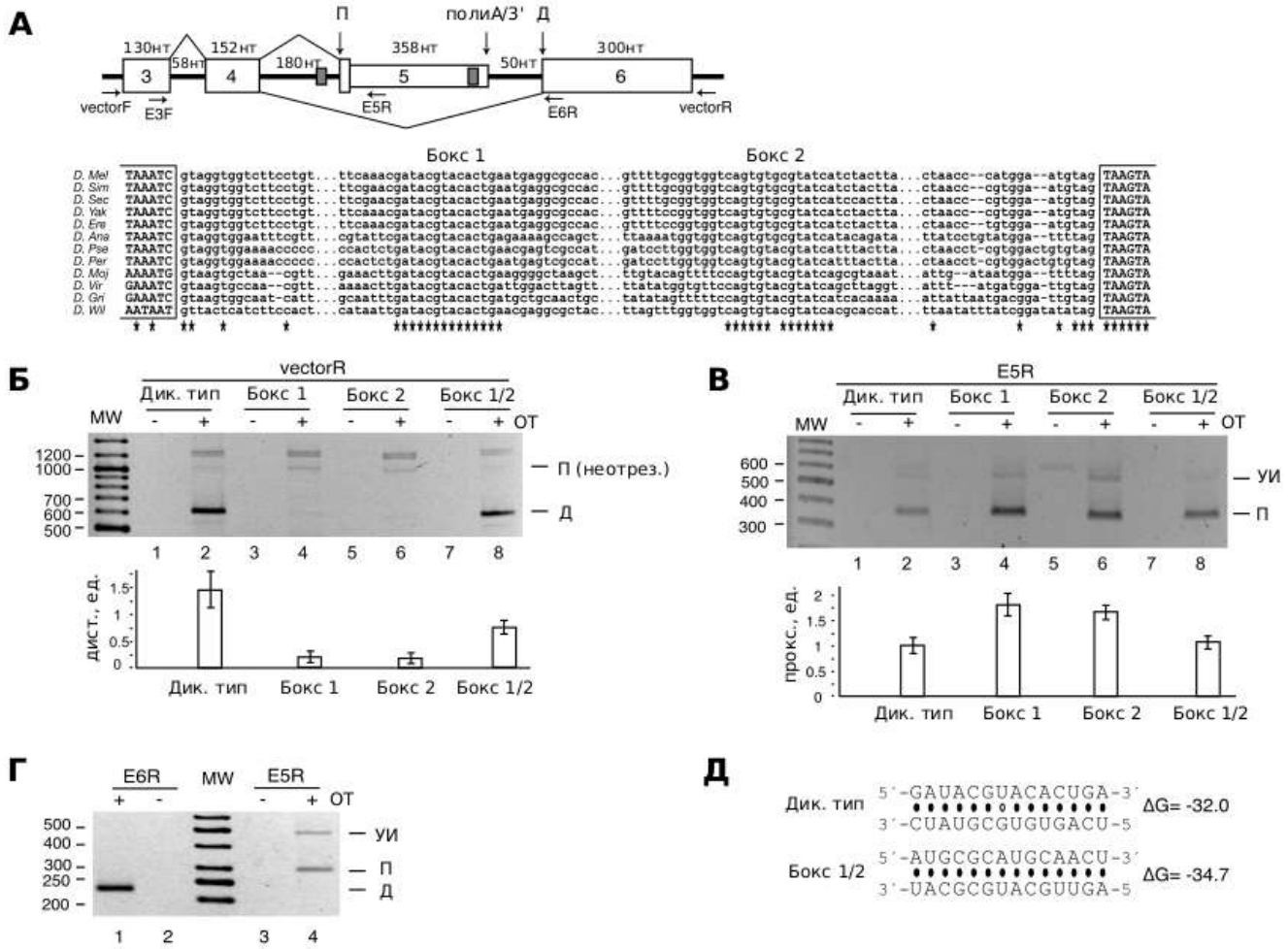


Рисунок 5.3 — АС и альтернативное полиаденилирование в гене *Nmnat* регулируются структурой РНК. **(А)** Верхняя панель: схема минигена, содержащего фрагмент последовательности между экзонами 4 и 6. Экзон 5 представляет собой внутренний терминальный экзон, который содержит сайт полиаденилирования/3'-процессинга. ККУ (Бокс 1 и Бокс 2) выделены серым цветом. Нижняя панель как на рис. 5.1. **(Б, В)** Продукты сплайсинга амплифицировали с помощью обратных праймеров vectorR и E5R для обнаружения изоформ, образованных в результате сплайсинга с дистальным акцептором (Д) или с проксимальным акцептором (П). **(Г)** Сплайсинг эндогенной мРНК, которую амплифицировали с помощью обратных праймеров в экзонах 5 и 6. **(Д** Предсказанные структуры РНК для дикого типа и двойного мутанта и их равновесные свободные энергии (ΔG, ккал/моль). Остальные обозначения как и на рис. 5.1.

тарными последовательностями, которые разделены примерно 350 нт, может физически приблизить дистальный сайт к донорному сайту и, тем самым, способствовать его использованию.

5.2 Экспериментальная валидация структур РНК в генах млекопитающих

В этом разделе приводятся результаты экспериментальной валидации влияния структуры РНК на АС в трех генах человека, а именно на сплайсинг кассетного экзона в гене *PHF20L1*, сплайсинг кассетного экзона в гене *CASK* и взаимоисключающее включение экзонов в гене *ATE1*. Для них были сконструированы минигены с разрушающими структуру одиночными и двойными компенсаторными мутациями (разд. 2.1.1), а также использованы блокирующие структуру АОН. Затем обсуждается влияние скорости элонгации транскрипции на АС через структуру РНК, а также перечисляются валидированные структуры РНК, не вошедшие в данную диссертационную работу.

5.2.1 Сплайсинг кассетного экзона в гене *PHF20L1*

Ген *PHF20L1* (plant homeodomain finger protein 20-like 1) кодирует фактор, считывающий состояния метилирования гистонов, который взаимодействует с моно- и диметилированными лизинами в H3K4me1, H4K20me1, H3K27me2, а также с эпигенетическими факторами, такими, как DNMT1 [452—454]. Он также участвует в поддержании стабильности метилированных белков SOX2 и pRb [455; 456]. Белок PHF20L1 важен для эпигенетического наследования у млекопитающих, плюрипотентности и дифференцировки клеток, а также для поддержания контрольной точки фазы G1-S, а нарушения его экспрессии характерны для рака молочной железы, колоректального рака и рака яичников [457—459].

Транскрипты гена *PHF20L1* существуют в виде двух сплайс-изоформ PHF20L1-a и PHF20L1-b, которые отличаются включением альтернативного кассетного экзона 6. Интроны, фланкирующие этот экзон, содержат пару ККУ, R1 и R2, которые могут образовывать стабильное комплементарное спаривание ($\Delta G = -23.6$ ккал/моль) и, тем самым, способствовать регуляции АС (рис. 5.4A). Эта пара ККУ была изначально найдена программой IRBIS (разд. 3.2), затем подтверждена PREPH (разд. 3.3), а также РНК-контактами,

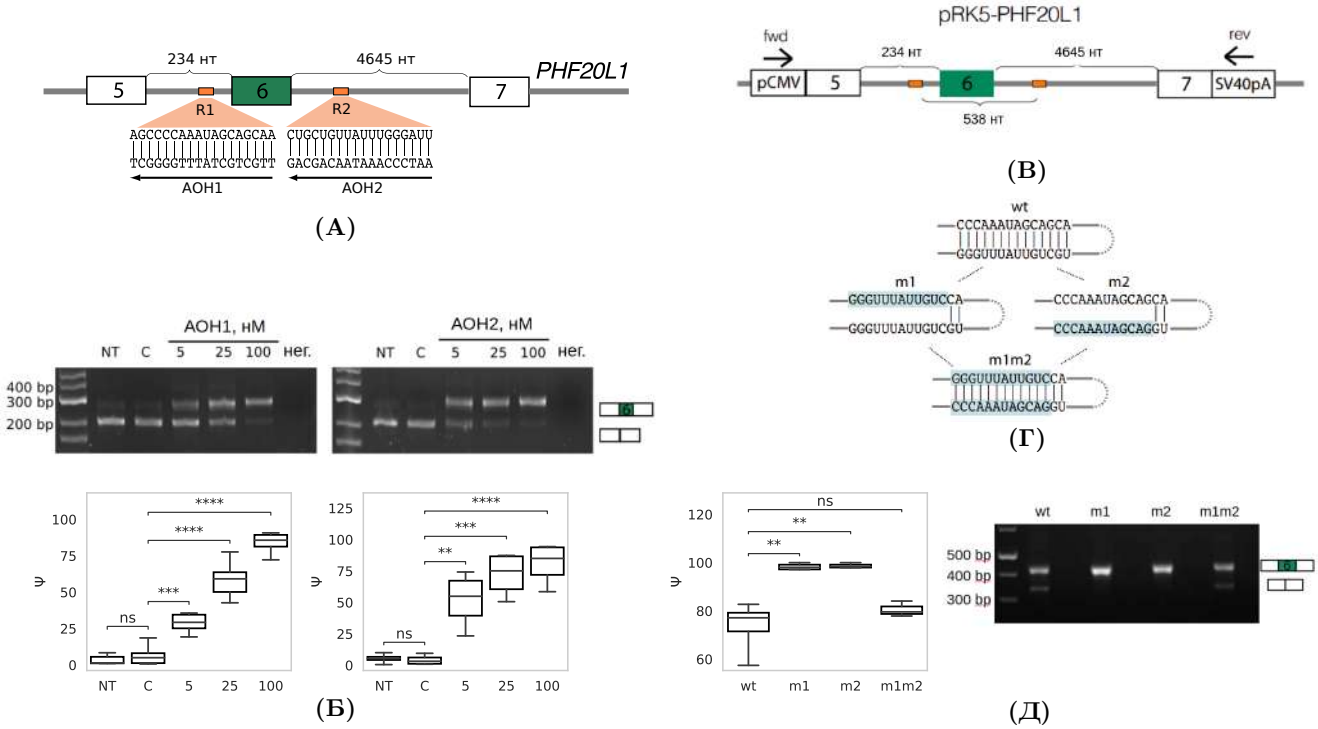


Рисунок 5.4 — Структура РНК регулирует сплайсинг кассетного экзона в гене *PHF20L1*. **(А)** Геномная организация экзонов 5–7 гена *PHF20L1*. Показаны комплементарные последовательности R1 и R2 и соответствующие им АОН (АОН1 и АОН2). **(Б)** Обработка АОН1 или АОН2 почти полностью подавляет пропуск экзона 6. NT – контроль без обработки АОН; C — контроль с обработкой АОН против гена люциферазы; нег. — отрицательный контроль. **(В)** Схема минигена, экспрессирующего фрагмент гена *PHF20L1*. Расположение праймеров указано стрелками. **(Г)** Схема мутагенеза минигена. wt — дикий тип. **(Г)** Пропуск экзона 6 у мутантов m1 и m2 подавлен, а у m1m2 сплайсинг возвращается к состоянию дикого типа. Символы **, *** и ns обозначают статистически различимые различия на уровне значимости 1%, 0.1%, и не значимые различия, соответственно.

наблюдаемыми в экспериментах RIC-seq в клеточных линиях HeLa и HepG2 (разд. 4.2.3).

Для того, чтобы проверить роль спаривания последовательностей R1/R2 в регуляции АС эндогенных транскриптов, были синтезированы АОН, комплементарные R1 и R2, называемые АОН1 и АОН2, соответственно. Качественный и количественный анализ ОТ-ПЦР показал, что уровень включения экзона 6 в ответ на увеличение концентрации АОН увеличивается, причем концентрация 5 нМ уже достаточна для существенного увеличения степени включения экзона 6 в эндогенном транскрипте (рис. 5.4Б).

Затем был сконструирован миниген, содержащий фрагмент геномной последовательности *PHF20L1* от экзона 5 до экзона 7 (рис. 5.4В), в который были введены мутации, разрушающие и восстанавливающие комплементарное спаривание оснований между R1 и R2 (рис. 5.4Г). Мутации, разрушающие спаривание

R1 с R2, называемые m1 и m2, способствовали включению экзона 6, а компенсаторный двойной мутант m1m2, в котором спаривание R1 с R2 восстановлено, по степени включения экзона 6 качественно и количественно совпадает с диким типом (рис. 5.4Д). Таким образом, ответ на действие комплементарных АОН и восстановление сплайсинга у компенсаторного двойного мутанта показывают, что включение экзона 6 контролируется спариванием оснований R1/R2.

5.2.2 Сплайсинг кассетного экзона в гене *CASK*

Ген *CASK* человека кодирует кальций/кальмодулин-зависимую сериновую протеинкиназу, однако этот белок не обладает киназной активностью [460] и функционирует как каркасный белок, участвующий в пресинаптической и постсинаптической передаче [461; 462]. У мышей делеция CASK приводит к летальному исходу [463], тогда как его инактивация в β-клетках поджелудочной железы влияет на гомеостаз глюкозы и чувствительность к инсулину [464]. Белок CASK взаимодействует с Tbr-1, транскрипционным фактором, участвующим в развитии переднего мозга [465] и, возможно, играет роль в установлении полярности эпителиальных клеток у млекопитающих [466].

Две альтернативные сплайс-изоформы *CASK* отличаются включением или пропуском кассетного экзона 19, который, как предполагается, модулирует связывание CASK с другими белками на разных стадиях развития и в клеточных популяциях с различной активностью нейронов [467; 468]. Фланкирующие экзон 19 интроны содержат пару ККУ, обозначаемых R3 и R4, которые расположены на расстоянии более 3000 нт друг от друга, а их взаимодействие потенциально выпетливает экзон 19 (рис. 5.5А). Как и в случае *PHF20L1*, эта пара ККУ была найдена программами IRBIS и PREPH, а также подтверждена РНК-контактами, наблюдаемыми в экспериментах RIC-seq в клеточных линиях IMR90, hNPC и GM12878 (разд. 4.2.3).

Степень включения экзона 19 существенно увеличивается при обработке клеток АОН3 и АОН4, комплементарными R3 и R4 (рис. 5.5Б). При внесении разрушающих структуру и компенсаторных мутаций в миниген, содержащий фрагмент *CASK* между экзонами 18 и 20 (рис. 5.5В,Г), одиночные мутации (m3 или m4) приводят к подавлению пропуска экзона 19, а в компенсаторном

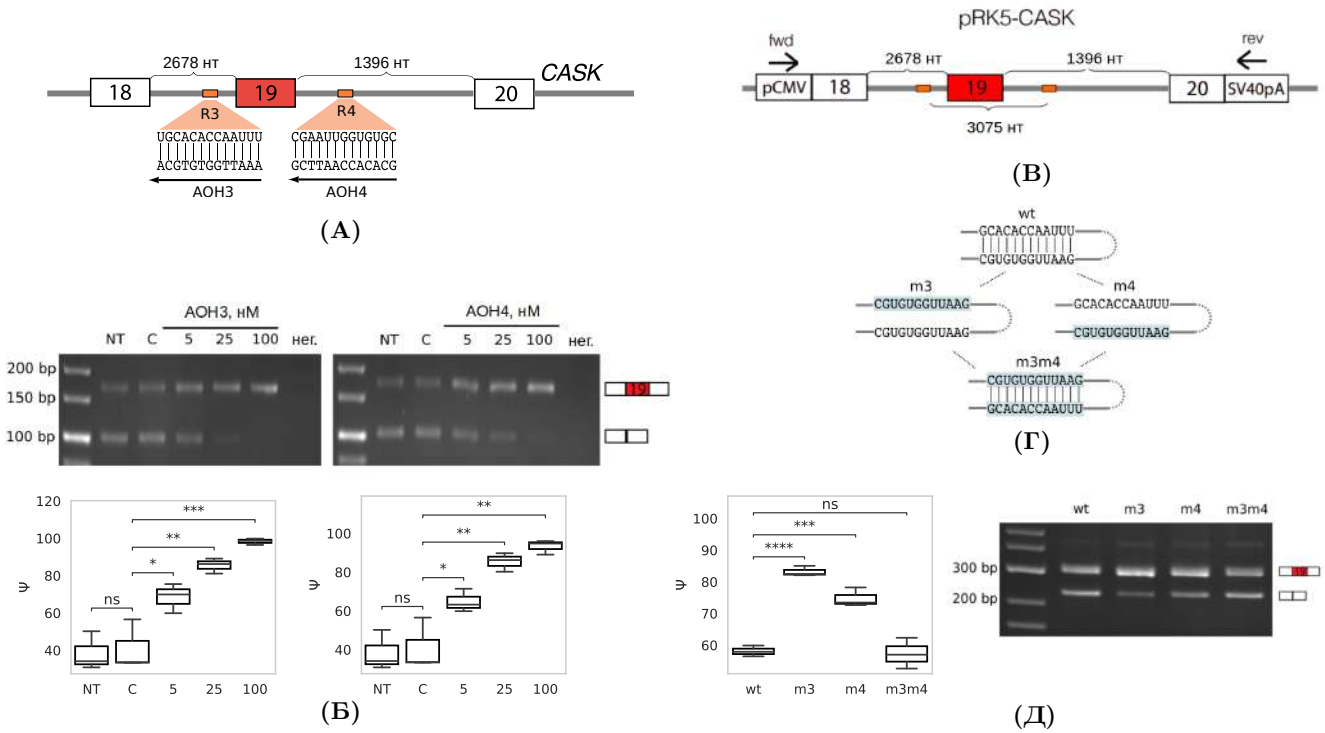


Рисунок 5.5 — Структура РНК регулирует сплайсинг кассетного экзона в гене *CASK*. **(А)** Фрагмент генома, содержащий экзоны 18–20 гена *CASK*. Показаны комплементарные последовательности R3 и R4 и соответствующие им АОН (АОН3 и АОН4). **(Б)** Добавление АОН3 или АОН4 подавляет пропуск экзона 19; NT — контроль без обработки АОН; C — контроль с АОН против гена люциферазы; нег. — отрицательный контроль. **(В)** Схема минигена, экспрессирующего фрагмент гена *CASK*. Стрелки обозначают положение праймеров. **(Г)** Схема мутагенеза минигена. wt — дикий тип. **(Д)** Пропуск экзона 19 в m3 и m4 качественно и количественно подавляется, а в m3m4 сплайсинг возвращается к состоянию дикого типа. Символы *, **, *** и ns обозначают статистически значимые различия на уровне значимости 5%, 1% и 0.1%, и не значимые различия, соответственно.

мутанте (m3m4) степень включения экзона 19 количественно восстанавливается до уровней дикого типа (рис. 5.5Д).

5.2.3 Регуляция взаимоисключающего сплайсинга в гене *ATE1*

Ген *ATE1* кодирует аргинил-трансферазу, которая участвует в посттрансляционной модификации белков, перенося аргинин к N-концевым остаткам на полипептидной цепи [469]. Белок *ATE1* является необходимым элементом во многих биохимических каскадах и участвует в регуляции протеолиза [470; 471], ответе на стресс и тепловой шок [472—474], эмбриогенезе [475—477], регенера-

тивных процессах [478—480] и старении [481; 482]. Дефицит *ATE1* приводит к эмбриональной смертности и тяжелым дефектам развития у мышей [476; 477; 483; 484].

Альтернативные сплайс-изоформы *ATE1* отличаются взаимоисключающим выбором двух соседних гомологичных экзонов (7a и 7b) длиной 129 п.о. каждый, а также альтернативным выбором первого экзона [485]. Двумя основными изоформами мРНК *ATE1* являются Ate1-1 (1b7a) и Ate1-2 (1b7b) [486]. 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 [485; 487; 488]. Ate1-2 локализуется исключительно в цитозоле, а Ate1-1 локализуется как в цитозоле, так и в ядре [485] и может специфически взаимодействовать с Liat1 — специфически экспрессирующимся в семенниках белком [489]. Инактивация экспрессии *ATE1* приводит к развитию опухолей в ксенографных мышиных моделях, причем рост опухоли можно частично остановить введением стабильно экспрессируемой изоформы Ate1-1, но не Ate1-2 [490]. Соотношение изоформ *ATE1*, содержащих экзоны 7a и 7b, существенно меняется во время мейоза у самцов мышей, что указывает на их роль в мейотическом цикле зародышевых клеток [488]. Таким образом, изоформы *ATE1* с экзонами 7a и 7b кодируют функционально различные аргинилтрансферазы.

До настоящего момента молекулярный механизм взаимоисключающего сплайсинга экзонов 7a и 7b гена *ATE1* не был изучен. Однако известно, что кластеры взаимоисключающих экзонов часто образуются при тандемных геномных дупликациях [124; 347; 491], причем многие из них регулируются конкурирующими структурами РНК [121]. Более того, конкурирующие структуры РНК могут возникать как естественный побочный продукт тандемных дупликаций, при которых удваивается одна из двух комплементарных последовательностей [125; 391]. Поскольку экзоны 7a и 7b *ATE1* гомологичны и схожи по длине, то естественно предположить, что их сплайсинг также регулируется конкурирующими структурами РНК.

Действительно, в геномном фрагменте между экзонами 6–8 гена *ATE1* методы IRBIS и PREPH идентифицировали несколько ККУ (рис. 5.6A). Две из них, называемые R1 и R4, перекрываются с акцепторными сайтами экзонов 7a и 7b и комплементарны последовательности R3, расположенной в интроне между ними. Интрон, соединяющий экзоны 7a и 7b, также содержит консервативную последовательность R2, которая комплементарна другой

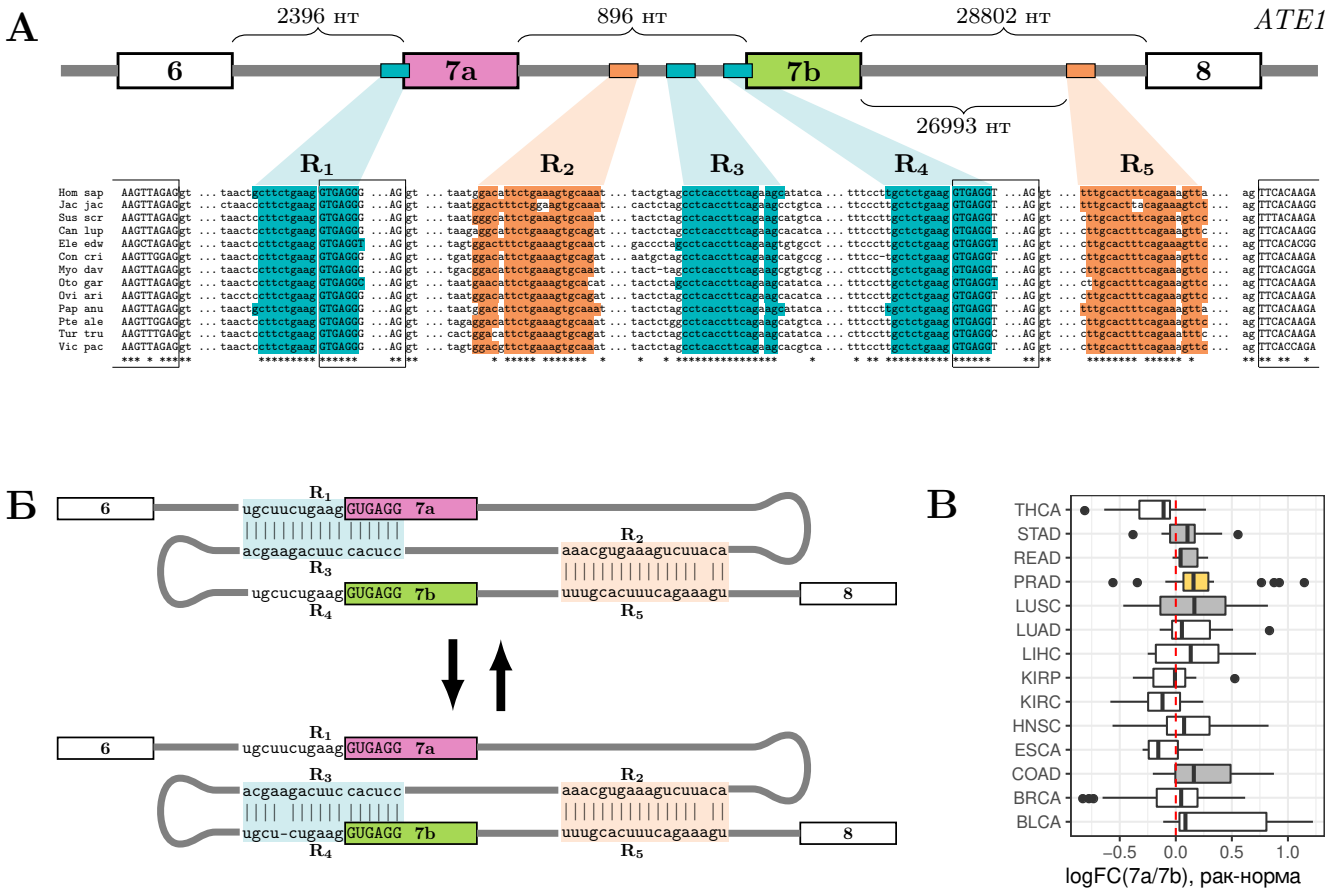


Рисунок 5.6 — Расположение ККУ в гене *ATE1* человека. **(А)** Геномная организация фрагмента гена *ATE1* между экзонами 6–8. Показаны пять эволюционно консервативных интронных элементов (R1–R5). Консервативные позиции отмечены звездочками. **(Б)** Последовательности R1 и R4 похожи между собой и конкурируют за спаривание оснований с R3; участок R2 комплементарен R5, который расположен в 30000 п.о. от него в сторону 3'-конца. **(В)** Относительная экспрессия изоформ экзона 7a/7b (разность $\log_{10} \frac{7a}{7b}$ между опухолью и нормальной тканью) значительно повышена в аденокарциноме желудка (STAD), прямой кишки (READ), толстой кишки (COAD), предстательной железы (PRAD), и плоскоклеточном раке легкого (LUSC); FWER < 0.05 – желтый, $P < 0.05$ – серый.

консервативной последовательности R5, расположенной в интроне между экзонами 7b и 8 на расстоянии около 30000 п.о. от R2 в направлении 3'-конца. Характер комплементарности между этими областями предполагает, что R1 и R4 могут конкурировать друг с другом за спаривание с R3, а вместе со спариванием R2 с R5 они образуют псевдоузел (рис. 5.6Б).

Экзоны 7a и 7b имеют широкий диапазон уровней включения в здоровых тканях человека (медианы 33% и 67%, соответственно) с наиболее заметным отклонением в семенниках (медианы 60% и 39%, соответственно), причем одновременного включения или одновременного пропуска обоих экзонов практически не наблюдается. Однако по сравнению с соответствующими нормальными тканями в образцах аденокарциномы простаты, а также других

эпителиальных опухолей, включая аденокарциному желудка, прямой кишки, толстой кишки и плоскоклеточный рак легких, наблюдается значительное увеличение соотношения сплайс-изоформ 7a/7b (рис. 5.6В).

Конкурирующие структуры РНК между R1, R3 и R4 контролируют взаимоисключающий сплайсинг

Для того, чтобы проверить, участвует ли структура РНК в регуляции сплайсинга *ATE1*, был сконструирован охватывающий экзоны 6–8 миниген, причем эндогенный интрон после экзона 7b был уменьшен в размере до 2 т.п.о. из-за очевидных ограничений при клонировании больших фрагментов (рис. 5.7А). Поскольку речь идет о конкурирующих структурах РНК (R1/R3 и R3/R4), для экспериментальной проверки требуются не двойные, а тройные мутанты, стратегия построения которых заключается в том, чтобы внести точечные мутации, которые нарушают структуру РНК по отдельности, но восстанавливают ее при внесении в различных комбинациях (рис. 5.7Б,В). Как показывает качественная ОТ-ПЦР (рис. 5.7Г), мутация m2, разрушающая спаривание оснований R1/R3, приводит к увеличению уровня включения экзона 7a, тогда как мутация m11, разрушающая спаривание оснований R3/R4, приводит к увеличению уровня включения экзона 7b. Мутация m1, разрушающая спаривание R3 и с R1, и с R4, приводит к увеличению доли транскриптов с двойными экзонами. При этом компенсаторная двойная мутация m1m2, восстанавливающая спаривание R1/R3, но разрушающая R3/R4, повышает уровень включения экзона 7b. Наоборот, компенсаторная двойная мутация m1m11, восстанавливающая спаривание R3/R4, но разрушающая спаривание R1/R3, приводит к увеличению уровня включения экзона 7a. Наконец, соотношение сплайс-изоформ в тройном мутанте m1m2m11, в котором восстановлена комплементарность как R1/R3, так и R3/R4, похоже на соотношение сплайс-изоформ в диком типе больше, чем в других мутантах.

Тернарная диаграмма количественно оцененных с помощью ОТ-ПЦР-РВ степеней включения экзона 7a, экзона 7b, а также двойных экзонов (7a7b) содержит четыре кластера в зависимости от того, комплементарность между какими последовательностями (R1/R3, R3/R4, ни одна, или обе) была

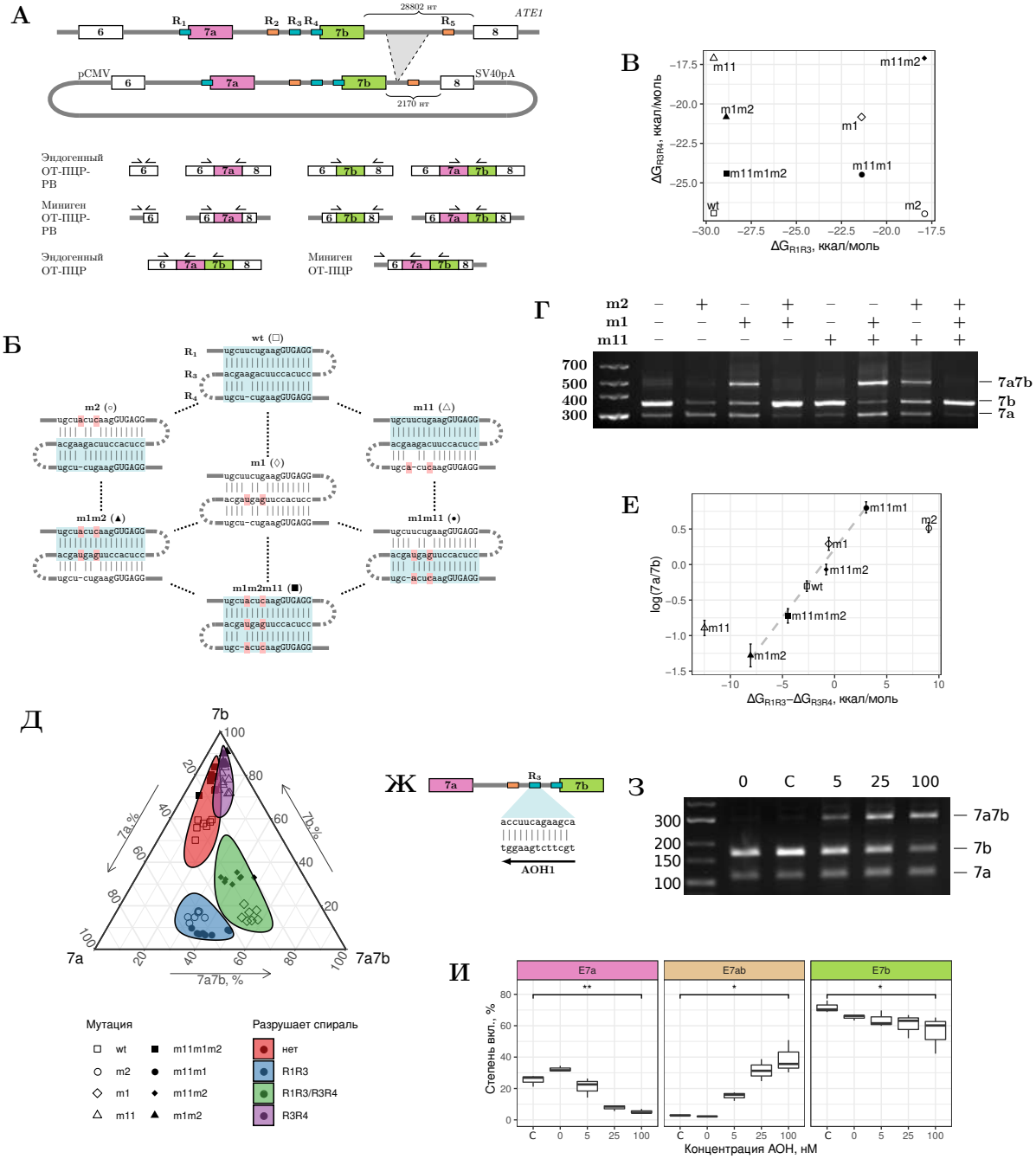


Рисунок 5.7 — Конкурирующие структуры РНК определяют АС в гене *ATE1*. **(А)** Интрон между экзоном 7b и экзоном 8 был уменьшен с 29000 п.о. в геноме до 2000 п.о. в минигене (вверху); праймеры для ОТ-ПЦР и ОТ-ПЦР-РВ (внизу). **(Б)** Разрушающие и компенсаторные мутации в R1, R3 и R4. **(В)** Предсказанные энергии гибридизации R1/R3 и R3/R4 для разрушающих и компенсаторных мутаций. **(Г)** Степень включения экзонов 7a, 7b и двойных экзонов (7a7b) меняется у одиночных и двойных мутантов, а сплайсинг дикого типа (wt) качественно восстанавливается у тройного мутанта m1m2m11. **(Д)** Тернарная диаграмма степеней включения экзонов 7a, 7b и 7a7b, измеренных с помощью ОТ-ПЦР-РВ. Цветные области обозначают 95% доверительные интервалы. **(Е)** Зависимость логарифма отношения степени включения экзонов 7a и 7b от разности термодинамических стабильностей R1/R3 и R3/R4. **(Ж)** AOH1 разрушает R1/R3 и R3/R4 посредством комплементарного спаривания с R3. **(З, И)** Уровень включения двойных экзонов в эндогенном транскрипте *ATE1*. C — контрольный AOH. Символы *, **, *** и ns обозначают статистически значимые различия на 10%, 5%, 1% и не значимые различия, соответственно.

нарушена (рис. 5.7Д). Доверительные интервалы, рассчитанные с помощью расстояния Махалнобиса для логарифмического преобразования пропорций [492], подтверждают, что степени включения экзонов значительно различаются, если разрушаются отдельные наборы спиралей. Несмотря на небольшие, но статистически значимые различия между отдельными мутантами внутри каждого кластера, картина сплайсинга в тройном мутанте (m1m2m11) с восстановленными R1/R3 и R3/R4 в целом ближе к дикому типу, чем к другим мутантам. Примечательно, что соотношение включения экзонов 7a/7b меняется пропорционально разнице термодинамической стабильности структуры, образованной R1/R3 и R3/R4 у всех мутантов, за исключением m2 и m11 (рис. 5.7Е).

Однако в клонированном фрагменте *ATE1* отсутствует значительная часть интрона 7, что может существенно влиять на сплайсинг. Для исследования роли R1, R3 и R4 в регуляции сплайсинга эндогенного транскрипта *ATE1* был использован АОН, комплементарный участку R3 (рис. 5.7Ж). Кроме того, R1 и R4 перекрываются с сайтами сплайсинга, поэтому использование комплементарных к ним АОН неизбежно приведет к изменению сплайсинга за счет блокировки распознавания сплайс-сайтов. ОТ-ПЦР показала, что трансфекция АОН1 в концентрации 5 нМ или больше индуцирует включение двойных экзонов и подавляет включение отдельных экзонов, в то время как трансфекция контрольного АОН не вызывает каких-либо изменений (рис. 5.7З). Количественная ОТ-ПЦР-РВ с праймерами, специфичными для изоформ, подтверждает эти наблюдения (рис. 5.7И). В совокупности мутагенез и действие АОН показывают, что функция конкурирующих структур РНК, образуемых участками R1, R3 и R4, заключается в поддержании взаимоисключающего сплайсинга экзонов 7a и 7b.

Дальние взаимодействия между R2 и R5 контролируют соотношение сплайс-изоформ

Для исследования функции двух других консервативных областей, R2 и R5, к минигену, несущему уменьшенный фрагмент *ATE1*, применялась аналогичная мутационная стратегия (рис. 5.8А). Как показывает ОТ-ПЦР, при раздельном введении разрушающих структуру мутаций m3 и m4 почти полно-

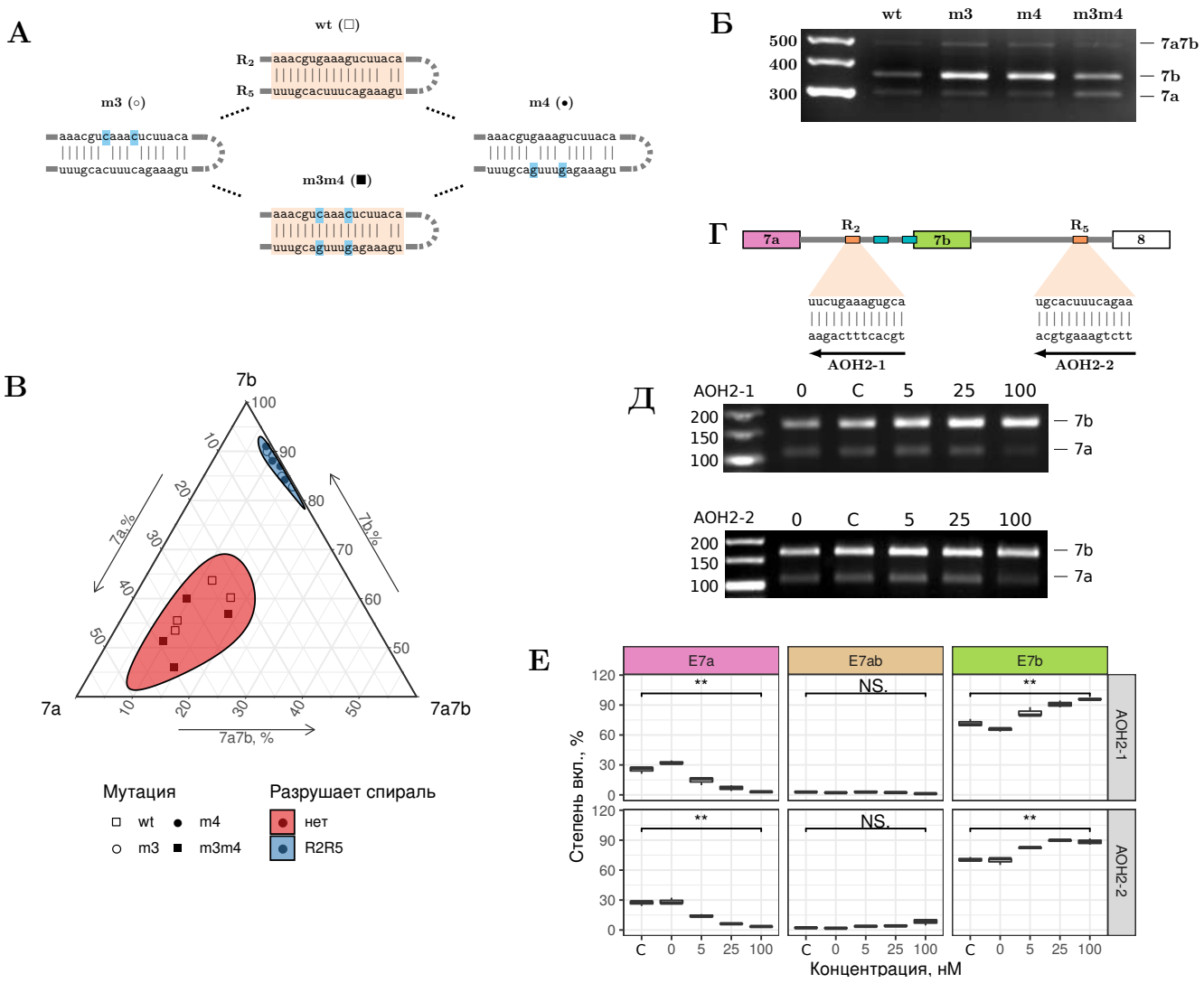


Рисунок 5.8 — Дальние взаимодействия между R2 и R5 контролируют соотношение сплайс-изоформ. **(А)** Стратегия мутагенеза для R2 и R5. **(Б)** Разрушающие комплементарность мутации (m3 и m4) усиливают включение экзона 7b, а компенсаторная мутация (m3m4) восстанавливает сплайсинг дикого типа (wt). **(В)** Тернарная диаграмма, отражающая результаты ОТ-ПЦР-РВ (см пояснения к рис. 5.7Д). **(Г)** AOH2-1 и AOH-2 разрушают структуру РНК, комплементарно спариваясь с R2 и R5. **(Д,Е)** Увеличение концентрации AOH2-1 и AOH2-2 подавляет включение экзона 7a и способствует включению экзона 7b в эндогенном транскрипте без появления двойных экзонов. C — контрольный AOH. Символы *, **, *** и ns обозначают статистически значимые различия на 10%, 5%, 1% уровне значимости и не значимые различия, соответственно.

стью подавляется включение экзона 7a и усиливается включение экзона 7b, а компенсаторный мутант m3m4 возвращает соотношение сплайс-изоформ к состоянию дикого типа (рис. 5.8Б). Количественно этот эффект подтверждается ОТ-ПЦР-РВ (рис. 5.8В). Следует отметить, что разрушающие мутации влияют только на соотношение изоформ с взаимоисключающими экзонами, но не вызывают появления двойных экзонов.

В эндогенной пре-мРНК R2 и R5 расположены на расстоянии около 30000 п.о. друг от друга, а в минигенной конструкции расстояние между ни-

ми было сокращено до 2000 п.о. Проверить влияние спаривания R2/R5 на сплайсинг эндогенного транскрипта можно с помощью АОН, комплементарных R2 и R5, называемых АОН2-1 и АОН2-2, соответственно (рис. 5.8Г). ОТ-ПЦР (рис. 5.8Д) и ОТ-ПЦР-РВ (рис. 5.8Е) показывают, что увеличение концентрации как АОН2-1, так и АОН2-2 приводит к уменьшению включения экзона 7a и увеличению включения экзона 7b без появления двойных экзонов. Из этих наблюдений можно заключить, что дальние взаимодействия между R2 и R5 регулируют соотношение сплайс-изоформ гена *ATE1*.

Взаимосвязь между локальными и дальними взаимодействиями в структуре РНК

Как было выяснено, вторичная структура пре-мРНК *ATE1* состоит из двух модулей, R1/R3/R4 и R2/R5, первый из которых контролирует взаимоисключающий сплайсинг экзонов 7a и 7b, а второй регулирует соотношение сплайс-изоформ. Для изучения взаимодействия этих модулей друг с другом было исследовано влияние АОН2-1, который блокирует дальние взаимодействия между R2 и R5, на уровни включения экзонов 7a и 7b в минигенах с разрушающими структуру R1/R3/R4 мутациями, а также у компенсаторных мутантов. Действие АОН2-1 было эквивалентно действию мутаций, которые разрушают взаимодействие между R2 и R5, независимо от мутаций, изменяющих комплементарность R1/R3 и R3/R4, т.е. АОН2-1 подавлял включение экзона 7a и способствовал включению экзона 7b, не изменяя долю двойных экзонов (рис. 5.9А). Таким образом, дальние взаимодействия между R2 и R5 играют доминирующую роль в выборе между экзонами 7a и 7b, при этом не влияя на их взаимоисключающий выбор.

Для исследования взаимодействия между R1/R3/R4 и R2/R5 в эндогенном транскрипте использовалось одновременное добавление АОН1 и АОН2-1. Клетки были обработаны комбинацией 25 нМ АОН1 и 25 нМ АОН2-1. Обработка только АОН1 увеличивала долю двойных экзонов на 28% без изменения степени включения экзона 7b, тогда как обработка только АОН2-1, наоборот, увеличивала степень включения экзона 7b на 31% без образования двойных экзонов. Одновременное добавление АОН1 и АОН2-1 привело к промежуточному

результату, при котором включение экзона 7b увеличилось на 16%, а включение двойных экзонов увеличилось на 7%. Этот эффект был аналогичен ответу мутанта m1 на обработку AOH2-1, при котором нарушалось взаимодействие между R1, R3 и R4 (рис. 5.9Б).

Эти результаты подтверждают предположение о функциональных различиях между двумя структурными модулями в пре-мРНК *ATE1*. Модуль конкурирующих локальных структур РНК (R1/R3/R4) отвечает за взаимное исключение экзонов 7a и 7b, а модуль дальних взаимодействий между R2 и R5 контролирует баланс сплайс-изоформ. Поскольку соотношение сплайс-изоформ различается в разных тканях и изменяется при заболеваниях (рис. 5.6В), естественно задать вопрос о том, какие факторы могут влиять на АС через структуру РНК.

Влияние скорости элонгации транскрипции на структуру РНК и АС

Как уже говорилось в разд. 3.3, скорость элонгации транскрипции оказывает значительное влияние на АС [493]. Несмотря на то, что медленная элонгация транскрипции открывает окно возможностей для распознавания пропускаемых в обычных условиях экзонов, в ряде случаев эффект может быть прямо противоположным [72; 77; 494]. Для изменения скорости элонгации транскрипции был выбран α-аманитин — селективный ингибитор, который взаимодействует с основной субъединицей RNAPII и переключает транскрипцию в «медленный режим» [495; 496]. После обработки α-аманитином соотношение 28S рРНК/GAPDH увеличилось почти вдвое, а также в некоторых генах были обнаружены характерные для замедления RNAPII продукты сплайсинга, известные из литературы [497]. Таким образом, было показано, что α-аманитин действительно ингибирует скорость элонгации RNAPII.

В эндогенном транскрипте *ATE1* воздействие α-аманитина приводит к заметному снижению уровня включения экзона 7b, увеличению уровня включения экзона 7a и незначительному увеличению уровня включения двойных экзонов (рис. 5.9В). Эта же закономерность наблюдается в данных секвенирования РНК для медленного мутанта RNAPII (R749H) [79]. При этом в минигенной конструкции соотношение изоформ с экзонами 7a и 7b под дей-

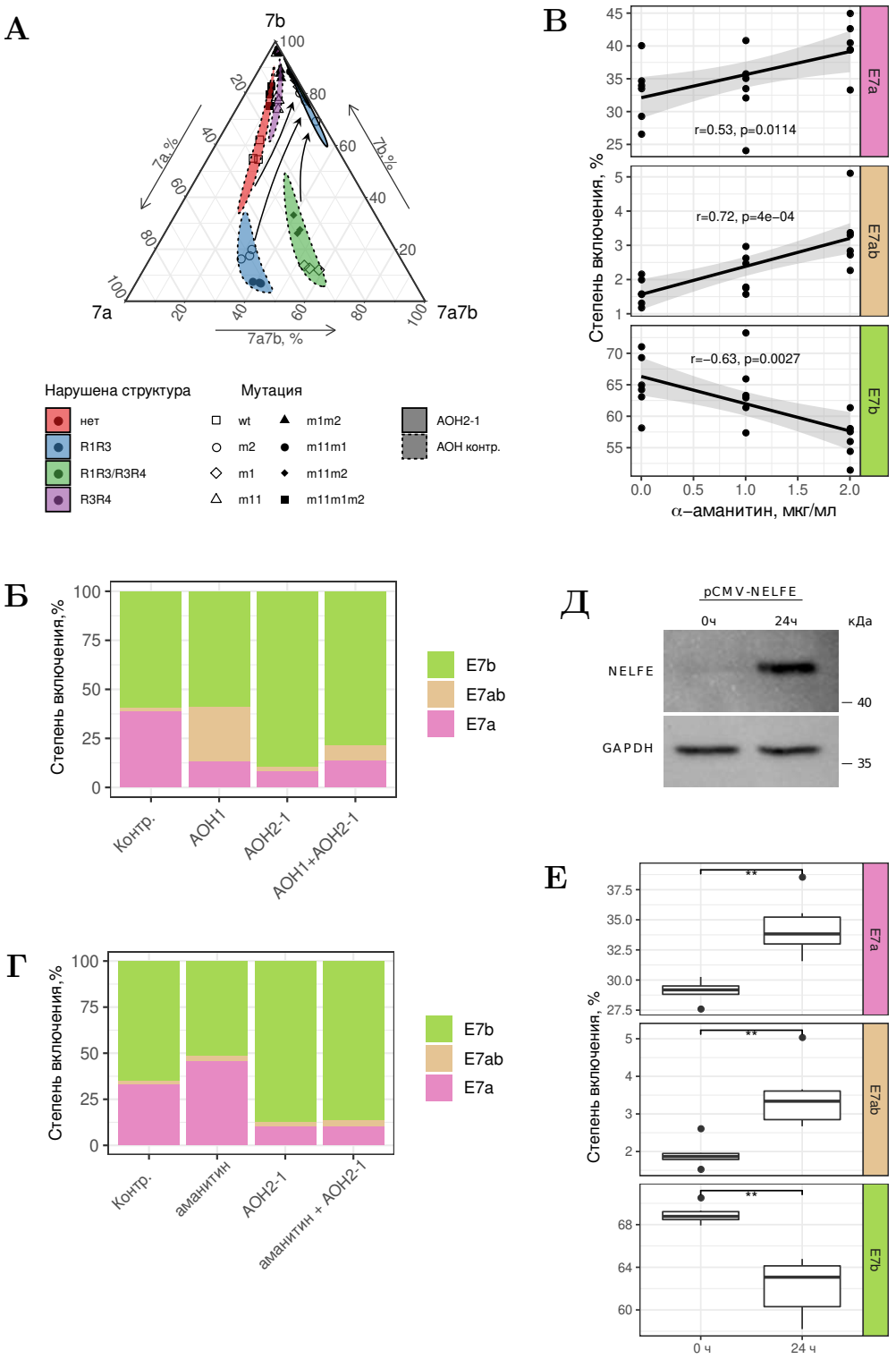


Рисунок 5.9 — Совместное влияние структуры РНК и скорости элонгации транскрипции на АС в гене *ATE1*. **(А)** Тернарная диаграмма степеней включения экзонов 7a, 7b и 7a7b при обработке AOH2-1. **(Б)** Разрушение конкурирующих структур R1/R3 и R3/R4 в эндогенном транскрипте при помощи AOH1 увеличивает долю двойных экзонов, не меняя уровня включения экзона 7b. **(В)** Уровень включения экзонов 7a, 7b, и 7a7b при воздействии α-аманитина. r — коэффициент корреляции Пирсона. **(Г)** Изменение уровня включения экзонов при совместной обработке AOH2-1 и α-аманитином. **(Д)** Вестерн-блот подтверждает суперэкспрессию белка NELFE. **(Е)** Изменение степени включения экзонов через 24 часа после суперэкспрессии NELFE. Символы *, **, *** и ns обозначают статистически значимые различия на 10%, 5%, 1% уровне значимости и не значимые различия, соответственно.

ствием α-аманитина существенно не меняется, что заставляет предположить, что отсутствие эффекта в минигене может быть связано с укорочением расстояния между R2 и R5.

Для того, чтобы исследовать совместное влияние замедления элонгации транскрипции и разрушения структуры РНК на АС в эндогенном транскрипте *ATE1*, использовалось одновременное добавление α-аманитина и обработка AOH2-1 в концентрации, блокирующей комплементарное спаривание R2/R5. Добавление α-аманитина в отсутствие AOH2-1 увеличивало уровень включения экзона 7a, а при разрушении спаривания R2/R5 при помощи AON2-1 добавление α-аманитина не приводило к подобному увеличению (рис. 5.9Г). Это указывает на то, что увеличение соотношения изоформ с экзонами 7a после обработки α-аманитином в эндогенном транскрипте с интактной структурой R2/R5 связано не с распознаванием сплайсосомой экзона 7a, а с более длительным временным окном, которое позволяет структуре РНК свернуться. Таким образом, образование дальних комплементарных взаимодействий между R2 и R5 зависит от скорости элонгации транскрипции и опосредует влияние замедления RNAPII на сплайсинг.

Один из механизмов замедления RNAPII в клетках млекопитающих задействует комплекс NELF, который функционирует не только в промоторных областях, но и взаимодействует с интеграторным комплексом, который специфически контролирует замедление и высвобождение RNAPII в кодирующих генах [498; 499]. Субъединица NELFE, связывание которой вызывает замедление RNAPII [500—502], высоко экспрессируется в семенниках, где уровень включения экзона 7a также является наибольшим. Было обнаружено, что инактивация NELFE при помощи микроРНК в клеточных линиях HepG2 и K562 приводит к значительному уменьшению включения экзона 7a с 94% до 48%, и что CUGAGG, канонический мотив NELFE у дрозофилы [503], встречается 14 раз в интроне между экзонами 7b и 8, в то время как по случайным причинам в среднем можно было бы ожидать только 6.5 таких случаев ($P = 0.007$). Это наблюдение заставляет предположить, что именно NELFE регулирует соотношение изоформ экзонов 7a/7b в семенниках, влияя на котранскрипционное сворачивание структуры R2/R5.

Для проверки этого предположения фактор NELFE был суперэкспрессирован под промотором CMV, а его суперэкспрессия была подтверждена с использованием соотношения NELFE/GAPDH, измеренного с помощью ОТ-

ПЦР-РВ и с помощью вестерн-блота (рис. 5.9Д). Измерение уровней включения экзонов 7a и 7b через 24 часа после трансфекции показало, что NELFE способствует включению экзона 7a и подавляет включение экзона 7b (рис. 5.9Е). Та же самая картина наблюдалась при обработке α-аманитином и в медленном мутанте RNAPII. Логично предположить, что специфичное для семенников включение экзона 7a может быть опосредовано структурой РНК и медленной элонгацией RNAPII, которая вызвана специфичной для семенников экспрессией NELFE.

5.3 Экспериментальная валидация РНК-структур в других генах

В этом разделе без иллюстраций перечисляются предсказанные структуры РНК, влияние которых на сплайсинг также было экспериментально показано, однако сами эксперименты во многом повторяют уже описанные и не заслуживают подробного обсуждения [108; 447; 504].

Ген *SF1* кодирует важнейший фактор сплайсинга, который участвует в АТФ-зависимом формировании сплайсосомного комплекса у позвоночных [504]. Пре-мРНК *SF1* состоит из четырнадцати экзонов, некоторые из которых подвергаются альтернативному сплайсингу. Согласно предсказаниям, интрон между экзонами 9 и 10, содержит пару ККУ, которые расположенны на расстоянии 260 п.о. друг от друга и потенциально могут способствовать сплайсингу, поскольку акцепторный сайт экзона 10 в является слабым, так как его последовательность (TAG) отличается от консенсусной (CAG), а бóльшая часть полипиримидинового тракта отсутствует. Однако, вопреки ожиданиям, при разрушении комплементарности между ККУ не происходит удержания интрона, а, наоборот, происходит вырезание более длинного интрона с дистальным акцепторным сайтом, расположенным на 21 нт в сторону 3'-конца по отношению к эндогенногму сайту. Оказалось, что при удержании интрона между экзонами 9 и 10 в рамке считывания образуется РТС и мРНК деградирует по пути NMD. Поскольку фактор *SF1* критически важен для выживания клетки, она предпочитает исключить семь аминокислот из последовательности белка за счет использования «аварийного» акцепторного сайта для того, чтобы избежать полной деградации мРНК этого гена [108].

Кроме этого примера, были изучены дальние взаимодействия в структуре РНК мышиных гомологов генов *CASK* и *PHF20L1* человека, *mCASK* и *mPHF20L1*. Оказалось, что ККУ, выпетливающие соответствующие кассетные экзоны, также регулируют АС в этих генах, т.е. они представляют собой эволюционно и функционально консервативный механизм контроля АС [504]. Выпетливающие кассетный экзон ККУ также были обнаружены в кодирующем член семейства серин/треониновых протеинкиназ гене *MARK2*, активность которого существенна при прогрессировании ряда опухолей, а повышение уровня экспрессии коррелирует с опухолевой трансформацией. *MARK2* может участвовать в активации клеточного цикла и репарации ДНК, а также влиять на лекарственную устойчивость [505] и эпителиально-мезенхимальный переход [506]. Вторичная структура, образуемая найденными ККУ, регулирует частоту включения экзона 17 и может модулироваться АОН. По результатам этого исследования получен патент на изобретение №2023124153/10(053096) «Система направленного изменения сплайсинга в гене MARK2», в основе которого лежит действие антисмысловых олигонуклеотидов на структуру РНК [447].

За рамками данной диссертационной работы экспериментальная валидация предсказанных структур РНК проводилась в нескольких международных исследовательских группах, из которых следует особенно отметить группу из Института наук о жизни Чжэцзянского университета под руководством проф. Юнфэн Джин. В 2016 году им была независимо обнаружена структура РНК (№39 в табл. 1 из [107]) в гене *Srp* дрозофилы, который играет важную роль в дифференцировке кишечника и кроветворении, и экспериментально валидировано ее влияние на АС [5], а также несколько других функционально важных РНК-структур.

5.4 Обсуждение результатов и выводы

5.4.1 Механизмы модуляции АС вторичной структурой РНК

Минигенные конструкции содержат участки генов, включающие в себя экзоны и фланкирующие интронные области с предполагаемыми регуляторными элементами. Использование данной модельной системы позволяет вносить мутации в элементы вторичной структуры РНК и оценивать их влияние на сплайсинг. Однако следует отметить, что в некоторых случаях минигенные конструкции технически не способны обнаружить изменения в событиях сплайсинга при разрушении структуры РНК. Например, было показано, что шпилечная структура влияет на сплайсинг пре-мРНК гена *Adh* дрозофилы и изменяет уровень экспрессии соответствующего белка, но изменения сплайсинга, наблюдаемые при разрушении шпильки *in vivo*, составляют только 6% [507]. Существенные различия в частоте включения кассетных экзонов также наблюдаются между сплайсингом в минигене и сплайсингом в эндогенном транскрипте. Эти различия следует учитывать при интерпретации результатов валидации структур РНК в минигенах и с помощью блокирующих АОН.

Как показывают полученные результаты, модуляция АС вторичной структурой РНК обеспечивает естественный механизм контроля за соотношением сплайс-изоформ и выбором сплайс-сайтов. В гене атрофина дрозофилы (разд. 5.1.2) два альтернативных акцепторных сайта используются с одинаковой частотой только тогда, когда вторичная структура частично блокирует более сильный из них. Таким образом, сбалансированное использование двух акцепторных сайтов обеспечивается за счет формирования вторичной структуры РНК без привлечения транс-действующих факторов. Некоторые природные системы, такие как ген *TAU* человека, также поддерживают баланс сплайс-изоформ с помощью структуры РНК, а изменения этого баланса вследствие нарушения структуры приводит к развитию патологий [508]. В гене *CG33298* вторичная структура полностью блокирует альтернативный донорный сайт, что дает пример структуры РНК, блокирующей вредоносный криптический сплайс-сайт (разд. 5.1.1). В гене *ATE1* человека конкурирующие спаривания между участками R1, R3 и R4 обеспечивают взаимоисключающий сплайсинг

экзонов 7a и 7b, причем соотношение сплайс-изоформ в одиночных и двойных мутантах прекрасно коррелирует с различием в термодинамической стабильности структур (разд. 5.2.3). Таким образом, результат сплайсинга определяется совокупным действием множества факторов, таких как сила сплайс-сайта, кинетика и термодинамика образования вторичной структуры.

Как видно на примере генов *Nmnat*, *CASK*, *PHF20L1*, и *ATE1*, дальние взаимодействия в структуре РНК могут простираться на десятки тысяч п.о., а образуемые ими петли могут изменять соотношение сплайс-изоформ, управлять включением кассетных экзонов и даже влиять на полиаденилирование. В гене *Nmnat* дрозофилы выпетливание необходимо для приближения дистального и снижения конкурентоспособности проксимального акцепторного сайта, причем разрушение структуры не только активирует дистальный сайт, но и запускает интронное полиаденилирование (разд. 5.1.3). Комплементарное спаривание между последовательностями R2 и R5 в гене *ATE1*, которое охватывает 30000 п.о. и является самым длинным из известных на сегодняшний день дальних взаимодействий, регулирует соотношение сплайс-изоформ с экзонами 7a и 7b в зависимости от скорости элонгации транскрипции. Элонгация транскрипции влияет на АС через структуру РНК, изменяя кинетику сворачивания, поскольку в присутствии комплементарного спаривания R2/R5 соотношение экзонов 7a/7b реагирует на замедление RNAPII, а при его разрушении этого не происходит. Как показано здесь, одним из факторов, регулирующих этот процесс, является NELFE — специфичная для семенников субъединица комплекса NELF, которая может способствовать включению экзона 7a в семенниках путем замедления элонгации транскрипции.

5.4.2 О происхождении конкурирующих структур РНК

Известно, что нуклеотидные последовательности экзонов, входящих в состав взаимоисключающих кластеров, часто имеют высокий уровень сходства, что указывает на их происхождение в результате тандемных геномных дупликаций [491]. Также известно, что взаимоисключающий сплайсинг экзонов часто регулируется конкурирующими структурами РНК — группами регуляторных элементов в пре-мРНК, называемых селекторными сайтами, которые конкури-

руют друг с другом за комплементарное спаривание с одним и тем же общим элементом — докерным сайтом.

В гене *ATE1* экзоны 7a и 7b гомологичны друг другу и схожи по длине, что согласуется с гипотезой об их происхождении в результате тандемной геномной дупликации [79]. Можно предположить, что она произошла после расхождения хордовых, поскольку гомолог экзона 7b отсутствует у беспозвоночных. Как было показано в разд. 5.2.3, селекторные сайты R1 и R4 конкурируют за спаривание с докерным сайтом R3, тем самым регулируя взаимоисключающий выбор экзонов, однако их происхождение невозможно проследить, поскольку на больших эволюционных расстояниях конкурирующие вторичные структуры РНК, как правило, не консервативны [124]. Следует также отметить, что длина интрона, следующего за экзоном 7b, уменьшается с увеличением эволюционного расстояния до человека. Комплементарное спаривание R2/R5 могло возникнуть уже после дупликации, чтобы противодействовать расширению этого интрона, содержащего длинные диспергированные повторы [79].

Однако взаимное расположение донорного и селекторных сайтов в *ATE1* отличается от расположения донорного и селекторных сайтов в гене *Dscam1*, для которого эта модель была впервые предложена [4]. В одной из моих работ было высказано предположение 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 [125].

Если дупликация (рис. 5.10) затрагивает область генома, содержащую экзон и часть левого фланкирующего интрона, а интрон, находящийся между экзонами 1 и 2, содержит пару комплементарных последовательностей, a и a', которые способны образовывать шпилечную структуру, то в результате дупликации, удваивающей только одну из двух комплементарных частей (a') две копии экзона 2 будут располагаться тандемно, а a' и его копия a'' будут комплементарны a. Такая дупликация создает пару конкурирующих структур, в которых a спаривается либо с a', либо с a'', причем в первом случае включается экзон 2.1, а во втором случае он выпетливается и пропускается [125]. Аналогичным образом дупликации могут создавать конкурирующие структуры $b - b' - b''$ с правым докерным сайтом (рис. 5.11).

В кластерах взаимоисключающих экзонов, в которых были обнаружены докерные и селекторные сайты, системы с правым докерным сайтом пре-

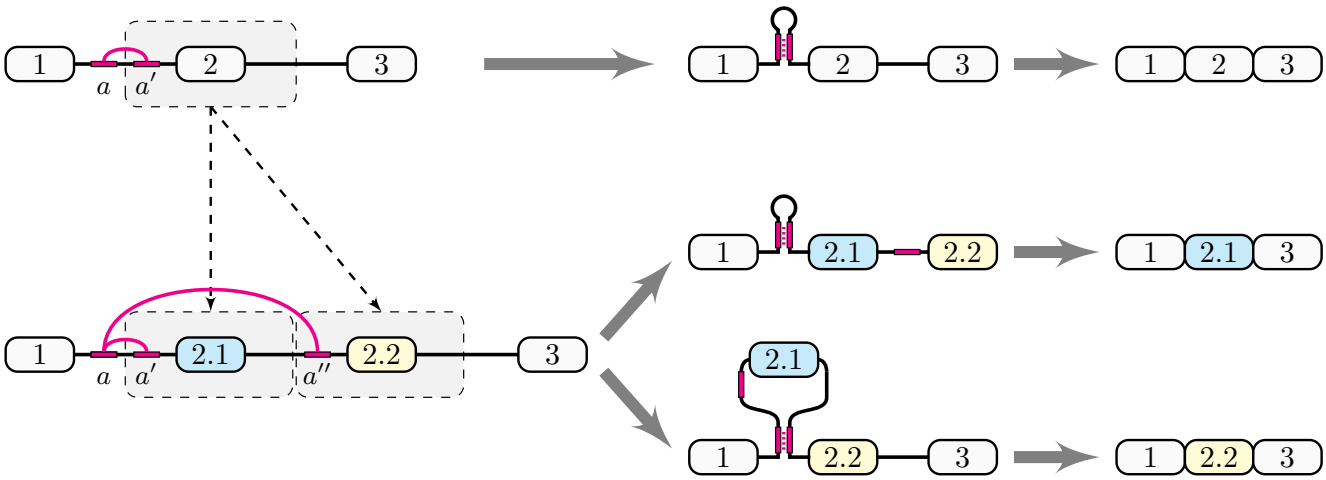


Рисунок 5.10 — Механизм образования конкурирующих структур РНК с левым докерным сайтом посредством дупликации. Если дупликация затрагивает экзон и одно плечо шпилечной структуры, расположенной в левом фланкирующем интроне, то образуется пара селекторных сайтов, которые могут конкурировать за один докерный сайт.

обладают над системами с левым докерным сайтом [125]. Фундаментальное различие ними заключается в регулируемости спаривания докерного сайта с селекторными последовательностями. Поскольку структура РНК формируется котранскрипционно, то спаривание $a' - a$ получает кинетическое преимущество по сравнению со спариванием $a - a''$, если докерный сайт расположен в левом интроне, а если докерный сайт расположен в правом интроне, то b'' и b' транскрибируются последовательно и получают равные шансы на спаривание с b.

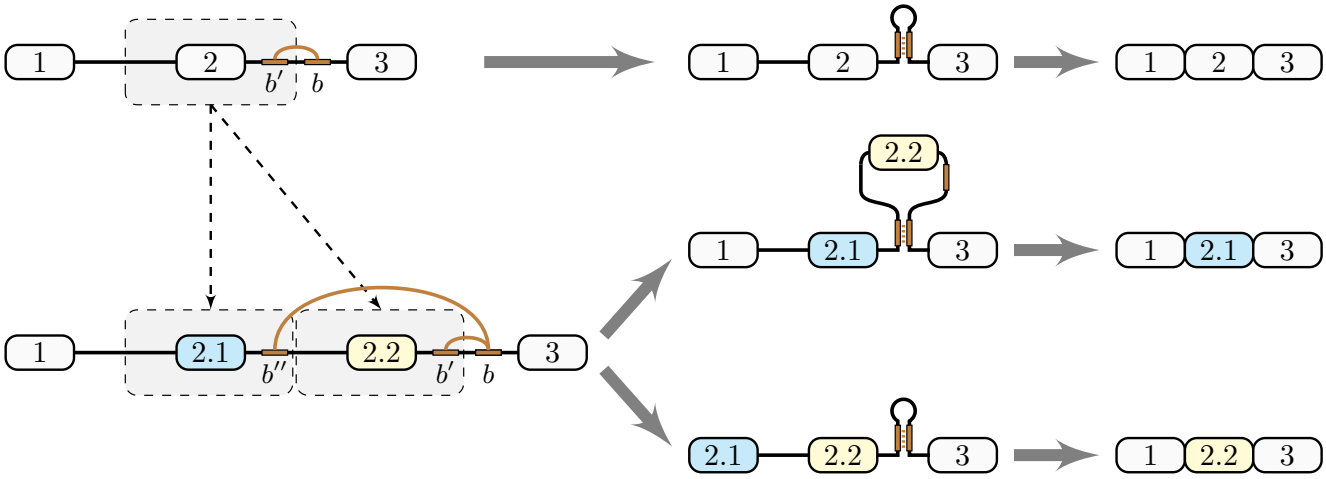


Рисунок 5.11 — Механизм образования конкурирующих структур РНК с правым докерным сайтом посредством дупликации (аналогично рис. 5.10).

Исследование структуры РНК в кластере экзонов 4 гена *srp* дрозофилы позволило предложить так называемую двунаправленную модель регуляции взаимоисключающего сплайсинга [5]. В гене *srp* две пары консервативных комплементарных интронных участков окружают два альтернативных экзо-

на, причем каждая из двух пар способствует активации одного экзона и инактивации второго. Их комплементарные спаривания не являются взаимоисключающими, но они образуют псевдоузел. Эта модель регуляции применима и к другим генам, таким как *RIC-3* и *MRP1*, кластеру экзонов 4 перепончатокрылых и кластеру экзонов 9 чешуекрылых генов семейства *Dscam* [5; 509].

Небольшая модификация сценария, показанного на рис. 5.10, может объяснить эволюционный механизм, который способен генерировать описанные структуры (рис. 5.12). А именно, в результате геномной дупликации, затрагивающей экзон и два его фланкирующих интрона, которые содержат две пары комплементарных последовательностей, $a - a'$ и $b' - b$, две копии экзона снова будут расположены тандемно, а также образуются две конкурирующие структуры РНК, в которых $a - a'$ конкурирует с $a - a''$, а $b' - b$ конкурирует с $b'' - b$. При этом они всегда будут расположены так, что b'' будет находиться 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, не все четыре комбинации равновероятны из-за псевдоузла. Если a взаимодействует с a', а b'' взаимодействует с b, то экзон 2.2 выпетливается и пропускается. Наоборот, если a взаимодействует с a'', а b' взаимодействует с b, то экзон 2.1 выпетливается и пропускается. В отличие от однонаправленной модели (рис. 5.10), двунаправленная модель объясняет механизм подавления включения одного экзона при включении другого.

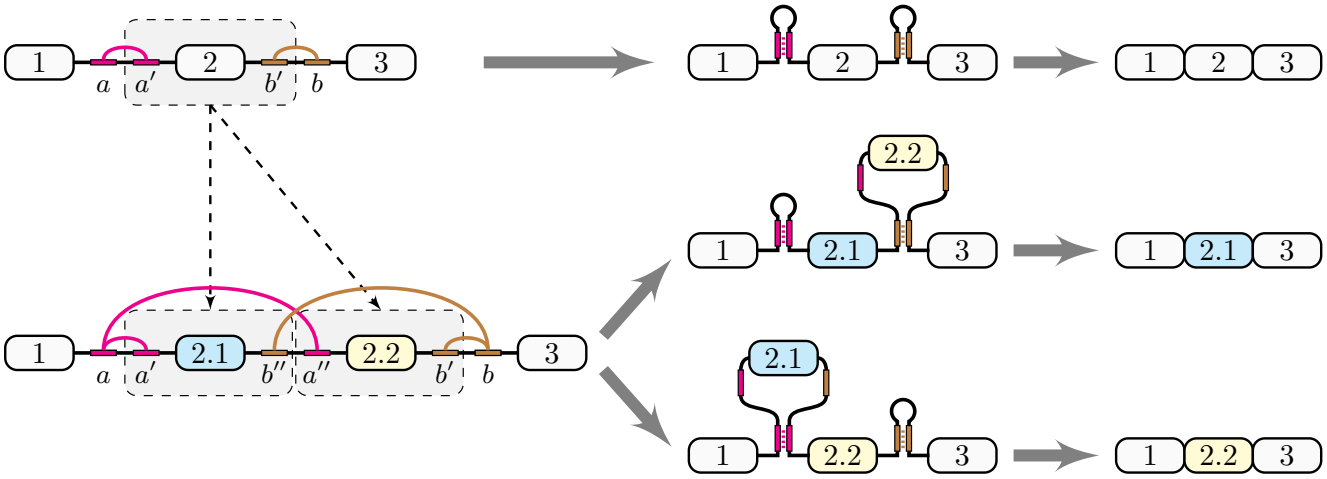


Рисунок 5.12 — Если тандемная дупликация затрагивает экзон и части двух его фланкирующих интронов, каждая из которых содержит шпилечную структуру, она может создать две пары конкурирующих комплементарных последовательностей, каждая из которых будет выпетливать один из экзонов 2.1 и 2.2. Расположение комплементарных частей в точности соответствует двунаправленной модели в гене *srp*.

Эти эволюционные модели дают объяснение тому, как последовательности докерных и селекторных сайтов могут самопроизвольно возникать в результате тандемных дупликаций. В гене *ATE1*, в котором докерный сайт R3 расположен между селекторными сайтами R1 и R4, мог иметь место аналогичный сценарий, при котором предковое состояние со шпилечной структурой вокруг донорного сайта экзона 7b было дуплицировано вместе с R4, но не R3 (рис. 5.6).

5.4.3 Использование АОН для терапевтической модуляции сплайсинга

Бóльшая часть технологий, связанных с направленным изменением сплайсинга в терапевтических целях, основана на использовании АОН, комплементарных участкам вокруг донорного или акцепторного сайтов сплайсинга, примыкающих к интересующему экзону. Такие олигонуклеотиды обычно блокируют распознавание сплайсосомой сайтов сплайсинга, и, таким образом, могут способствовать пропуску экзона или подавлению использования болезнетворного сайта сплайсинга. О терапевтических модифицированных АОН, активирующих, а не блокирующих сплайсинг, в настоящее время практически ничего не известно.

Примеры структур РНК в генах *CASK* и *PHF20L1* и их влияние на АС показывают, что использование двухцепочечных участков РНК в качестве мишеней АОН позволяет не только подавлять, но и активировать включение экзонов. Таким образом, разработка АОН, направленных на элементы вторичной структуры РНК в интронах, открывает новые терапевтические возможности для лечения заболеваний, вызываемых нарушением сплайсинга. При этом следует отметить, что АОН, блокирующие одну или другую последовательность в паре ККУ, могут иметь разную эффективность в отношении блокировки РНК-структур и, как следствие, по-разному влиять на АС. Такие различия, по всей видимости, связаны с котранскрипционным сворачиванием пре-мРНК, при котором 5'-ККУ транскрибируется первым и сразу вступает во взаимодействие с АОН, в то время как после появления 3'-ККУ конкурентное преимущество у АОН теряется. По своей сути это явление аналогично преобладанию систем с правым докерным сайтом над системами с левым докерным сайтом. Таким

образом, кинетика котранскрипционного сворачивания РНК является важным фактором, который необходимо учитывать при разработке терапевтических модифицированных АОН, влияющих на вторичную структуру РНК.

Глава 6. Регуляция непродуктивного сплайсинга РСБ и структурой РНК

В этой главе приводятся результаты, полученные мной в период с 2015 по 2024 год в сотрудничестве с лабораториями под руководством проф. Родрига Гигó (Центр Геномной Регуляции, г. Барселона), проф. Адама Франкиша (Европейский институт биоинформатики, г. Кембридж) и проф. О.А. Донцовой (Московский Государственный Университет им. М.В. Ломоносова) [173; 216; 351; 356; 357; 510].

6.1 Исследование ауторегуляторного непродуктивного сплайсинга

Изложенный в этом разделе метод обнаружения ауторегуляторных петель непродуктивного сплайсинга с отрицательными обратными связями основан на данных секвенирования РНК из общедоступных источников. Для нахождения непродуктивных событий используются данные по ответу транскриптома клеток HEK293 на совместную инактивацию двух компонентов системы NMD — хеликазы UPF1, обладающей активностью РНК-зависимой АТФазы, и 5'-3' экзорибонуклеазы XRN1 [367]. Предсказание регуляторных взаимодействий основано на данных по ответу транскриптома клеток K562 и HepG2 на инактивацию РСБ и данных eCLIP, полученных для этих же клеточных линий в рамках международного проекта ENCODE (Encyclopedia of DNA Elements) [110; 362]. Несмотря на то, что эти данные получены с использованием разных клеточных линий, их можно применить для обнаружения общих для клеточного гомеостаза регуляторных механизмов.

6.1.1 Ядовитые и необходимые экзоны

Количественная оценка изменения сплайсинга во всех экспериментах производилась с помощью метрики $\Delta\Psi$, которая равна разности степеней

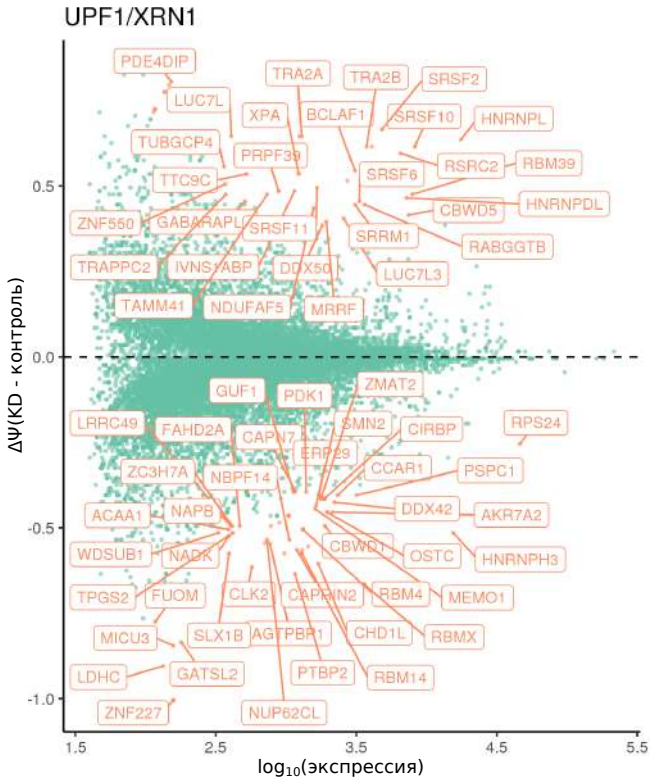


Рисунок 6.1 — Изменение степени включения экзонов при кодеплеции UPF1/XRN1 по сравнению с контролем (Ψ(KD-контроль)) в зависимости от среднего уровня экспрессии. Значимые на уровне 0.1% отклонения показаны оранжевым цветом.

включения экзона (Ψ) в эксперименте и в контроле. Метрика $\Delta\Psi$ принимает значения в интервале от -1 до 1 и может интерпретироваться как изменение процентного соотношения сплайс-изоформ, в которых экзон включается. Однако степень «экстремальности» значений $\Delta\Psi$ зависит от уровня экспрессии гена: для низкоэкспрессирующихся генов наблюдать большие по абсолютной величине изменения $\Delta\Psi$ намного менее удивительно, чем для высокоэкспрессирующихся генов. Для оценки статистической значимости $\Delta\Psi$ был разработан простой статистический метод, во многом аналогичный MA-диаграмме, которая обычно используются для представления изменений генной экспрессии по сравнению со средним значением между двумя образцами [511]. Суть метода состоит в оценке дисперсии распределения $\Delta\Psi$ в зависимости от уровня экспрессии гена, который можно оценить так же, как и Ψ, используя разрывные чтения, с последующим применением поправки на множественное тестирование для оценки доли ложноположительных предсказаний по методу q-значений [173; 377].

Совместное распределение $\Delta\Psi$ и логарифма уровня экспрессии имеет характерную форму равнобедренного треугольника (рис. 6.1), причем значимые изменения Ψ располагаются вдоль его боковых сторон. Среди генов,

содержащих реактивные к инактивации NMD экзоны, присутствуют основные сплайсосомные белки, SR-богатые белки, hnRNP и другие РНК-связывающие белки. Такая же картина наблюдается и при совместной деплеции XRN1 и SMG6 — другого ключевого фактора системы NMD, причем значения $\Delta\Psi$ в при кодеплеции UPF1/XRN1 и SMG6/XRN1 хорошо скоррелированы (коэффициент корреляции Пирсона $r = 0.85$) [173].

При инактивации системы NMD уровни включения аннотированных ядовитых экзонов, определяемых как кассетные экзоны, содержащие стоп кодон NMD-транскрипта, в среднем увеличиваются по сравнению с контрольной выборкой кодирующих кассетных экзонов (рис. 6.2А). Например, уровень включения известного ядовитого экзона, расположенного в 3'-UTR гена *SRSF3*, значительно усиливается ($\Delta\Psi = 0.34$). И наоборот, уровни включения необходимых экзонов, определяемых как экзоны длины $3n + 1$ или $3n + 2$, где n — целое число, в среднем уменьшаются по сравнению с контрольной выборкой экзонов длины $3n$ (рис. 6.2Б). Это определение корректно, так как экзоны длины $3n+1$ и $3n+2$ приводят к сдвигу рамки считывания, который почти наверняка вызывает PTC, тогда как большинство экзонов длины $3n$ не являются необходимыми. Так, пропуск необходимого экзона 10 в гене PTBP2 значительно подавляется при инактивации NMD ($\Delta\Psi = -0.64$).

Логика поиска заключается в том, что для ауторегуляции экспрессии через непродуктивный сплайсинг РСБ должен связывать свою собственную пре-мРНК, т. е. содержать свой собственный сигнал eCLIP. Во-вторых, в ответ на инактивацию системы NMD подавляемые в обычных условиях ядовитые экзоны должны повышать уровень включения ($\Delta\Psi > 0$), тогда как необходимые экзоны, пропуск которых обычно подавляется, должны, наоборот, его снижать ($\Delta\Psi < 0$). Наконец, ауторегуляция экспрессии РСБ может достигаться путем активации или подавления включения экзонов. В случае активации ядовитого экзона, инактивация экспрессии РСБ должна приводить к снижению уровня включения ($\Delta\Psi < 0$), и наоборот, в случае подавления необходимого экзона она должна приводить к повышению уровня включения ($\Delta\Psi > 0$). Таким образом, ядовитые и необходимые экзоны должны реагировать на инактивацию системы NMD и инактивацию гена-хозяина противоположным образом (рис. 6.2В).

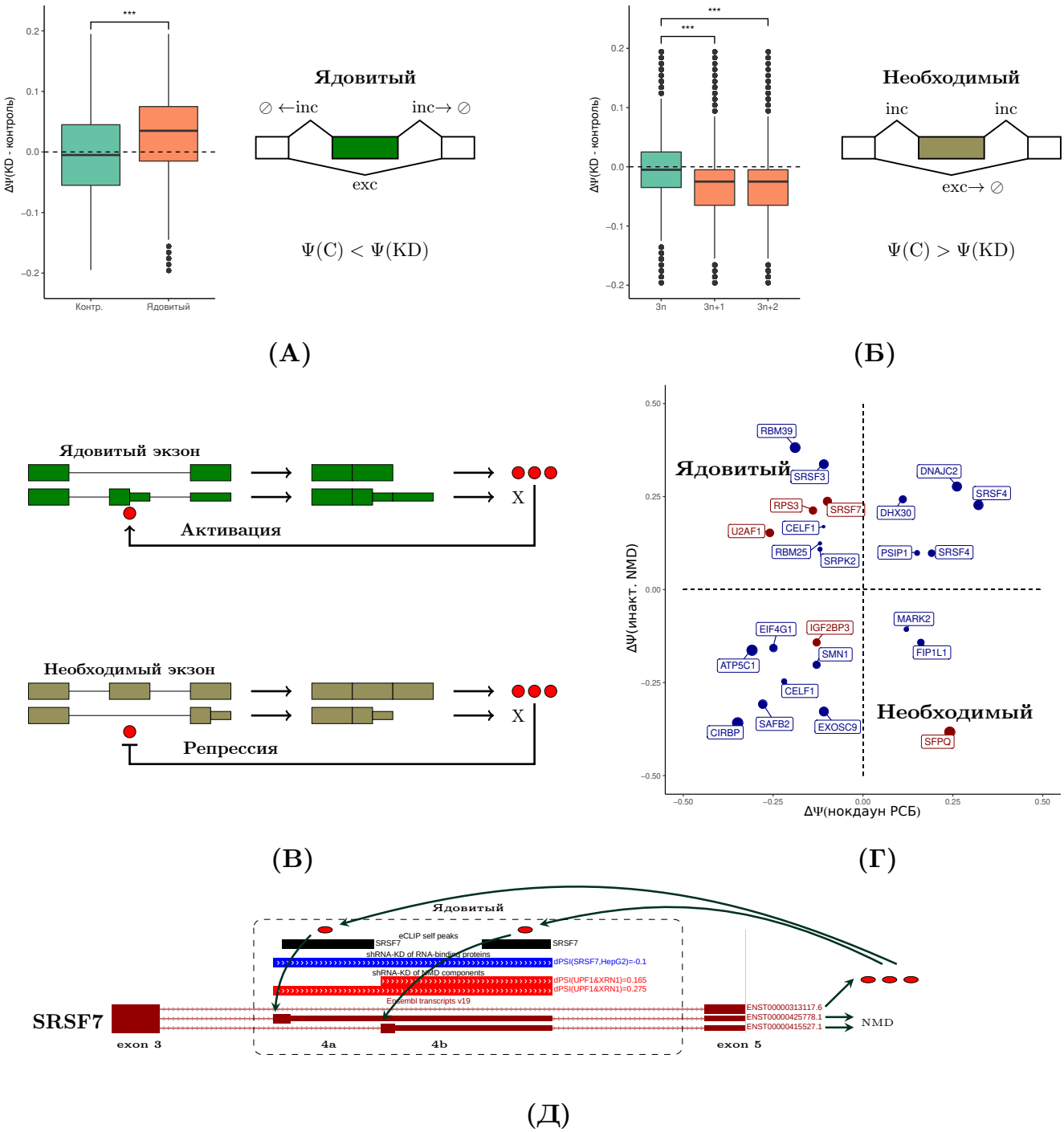


Рисунок 6.2 — Предсказание событий ауторегуляторного непродуктивного сплайсинга. **(А)** Частота включения ядовитых экзонов по сравнению с кассетными кодирующими экзонами при инактивации NMD. **(Б)** Частота включения необходимых экзонов ($3n+1$ и $3n+2$) по сравнению с экзонами, которые не являются необходимыми ($3n$). Символ *** обозначает статистически значимые различия на 1% уровне значимости. **(В)** Отрицательная обратная связь, обусловленная активацией включения ядовитого экзона (наверху) или подавлением пропуска необходимого экзона (внизу). **(Г)** Изменение степени включения экзонов при инактивации NMD (ось y) и изменение степени включения экзонов при инактивации РСБ (ось x). Показаны экзоны с $|\Delta\Psi| \geqslant 0.1$. Экзоны, содержащие в своей окрестности хотя бы один сигнал РСБ, показаны красным цветом. **(Д)** Ген *SRSF7* связывает свою собственную пре-мРНК, способствуя включению в нее ядовитых экзонов 4a и 4b. Показаны снимки из Геномного браузера UCSC.

6.1.2 Предсказание ауторегуляторного непродуктивного сплайсинга

В применении к известным примерам ауторегуляторного непродуктивного сплайсинга, данный подход показывает, что $\Delta\Psi$, как правило, имеют разные знаки при инактивации NMD и РСБ, однако эти изменения не всегда значимы. Например, пропуск альтернативного экзона 11 в мРНК гена *PTBP1*, должен приводить к деградации по NMD, причем бóльшая часть транскриптов *PTBP1* в клетках HeLa уничтожается именно таким образом [197]. Экзон 11 является необходимым, однако изменения степени его включения невелики по абсолютной величине ($\Delta\Psi_{UPF1} = -0.04$ и $\Delta\Psi_{PTBP1} = 0.03$) и статистически не значимы. Интересно отметить, что в окрестности экзона 11 наблюдается сигнал eCLIP, но два варианта экзона 9 реагируют на инактивацию *PTBP1* с гораздо более высоким $\Delta\Psi$, что позволяет предположить, что пропуск экзона 11 не является основным регуляторным событием. Аналогично, 3'-НТО гена *TARDBP* содержит несколько сигналов eCLIP, что подтверждает связывание TARDBP в его собственной 3'-НТО. В нем обнаруживается ядовитый экзон с $\Delta\Psi_{UPF1} = 0.25$, однако при инактивации самого *TARDBP* существенных изменений сплайсинга не происходит.

Поскольку несколько известных из литературы мишеней непродуктивного сплайсинга оказались ложноотрицательными, логичным решением представляется ограничиться нахождением нескольких ярких примеров, используя в качестве фильтра величину эффекта. При использовании порога $|\Delta\Psi| \geqslant 0.1$ как для инактивации NMD, так и РСБ, и накладывая условие, что пре-мРНК должна содержать хотя бы один сигнал eCLIP, расположенный в пределах 5000 п.о. от реактивного экзона, находятся три кандидатных ядовитых экзона в генах *SRSF7*, *U2AF1* и *RPS3*, а также один необходимый экзон в гене *SFPQ* (рис. 6.2Г). Ниже обсуждается лишь один из этих примеров (*SRSF7*), а обсуждение остальных можно найти в [173].

Ген *SRSF7*, также известный как фактор сплайсинга *9G8*, принадлежит семейству SR белков и играет важную роль в регуляции сплайсинга других генов, а также регуляции ядерного экспорта, трансляции, пролиферации раковых клеток и апоптоза [512—514]. Ранее было высказано предположение, что *SRSF7* может быть вовлечен в петлю отрицательной ауторегуляторной

обратной связи [178]. Сделанные наблюдения подтверждают существование двух ядовитых экзонов в *SRSF7*, причем включение ядовитых экзонов 4a и 4b существенно повышается при инактивации NMD ($\Delta\Psi_{UPF1} = 0.24$ и 0.14, соответственно), а инактивация самого *SRSF7* способствует пропуску этих экзонов ($\Delta\Psi_{SRSF7} = -0.10$ и $\Delta\Psi_{SRSF7} = -0.05$, соответственно). Это говорит о том, что включение ядовитых экзонов 4a и 4b активируется самим SRSF7. Кроме того, в обоих ядовитых экзонах расположены сигналы eCLIP белка SRSF7. В сумме все эти наблюдения дают достаточно оснований считать, что избыток белка SRSF7 связывает свою собственную пре-мРНК, способствуя включению ядовитых экзонов и, тем самым, понижает уровень своей экспрессии. Следует отметить, что предсказанный молекулярный механизм был экспериментально подтвержден в последующих работах других авторов [515]. Его детали оказались намного сложнее представленных здесь предсказаний. В частности, повышенные уровни экспрессии SRSF7 ингибируют систему NMD и способствуют трансляции двух половин белка, называемых разделенными рамками считывания, с бицистронного транскрипта.

6.2 Исследование кросс-регуляторного тканеспецифического непродуктивного сплайсинга

Представленная в разд. 6.1 линия рассуждений применима и к обнаружению событий кросс-регуляторного непродуктивного сплайсинга. Однако учитывая тот факт, что далеко не все известные из литературы мишени были найдены, а также то, что число попарных комбинаций экспериментов по инактивации РСБ очень велико, а при использовании поправки на множественное тестирование число ложноотрицательных предсказаний только увеличится, представляется целесообразным использовать больший объем данных секвенирования РНК. В этом разделе рассказывается об обнаружении тканеспецифических событий непродуктивного сплайсинга с использованием транскриптомов тканей человека из консорциума GTEx [356].

6.2.1 Тканеспецифически регулируемые события

Для предсказания тканеспецифических событий на основании базы данных GENCODE [371] и литературных сведений был составлен каталог событий непродуктивного сплайсинга. Среди них присутствовали события, для которых механизм непродуктивного сплайсинга был подтвержден экспериментально (валидированные) и те, для которых он был предсказан по аннотации (аннотированные). Из этого списка затем были извлечены события со значимой отрицательной взаимосвязью между уровнем экспрессии гена-мишени (e_g) и метрикой Ψ, далее называемые значимыми событиями. Отметим, что в этом разделе из соображений единообразия метрика Ψ определялась всегда по отношению к NMD-изоформе, т.е. $\Psi = 1$ для ядовитого экзона означает, что он всегда включается, а $\Psi = 1$ для необходимого экзона означает, что он всегда пропускается. Далее значимое событие классифицировалось как тканеспецифическое, если оно имело согласованные отклонения медианных значений e_g и Ψ в какой-либо ткани от соответствующих медианных значений по всем образцам. Такие события назывались тканеспецифическими. Затем для каждого потенциального РСБ-регулятора с тканеспецифической экспрессией находились события, значимо откликающиеся на инактивацию экспрессии РСБ и имеющие согласованное направление изменения экспрессии гена-мишени в тканях. Такие события назывались тканеспецифически регулируемыми. На эти результаты накладывались данные о сайтах связывания РСБ из базы данных POSTAR3 [516] и отбирались события двух типов: с сигналом CLIP в гене и сигналом CLIP в окрестности события. Такие события назывались имеющими поддержку CLIP в гене и имеющими локальную поддержку CLIP, соответственно.

Составленный каталог валидированных литературных событий и их регуляторов состоял из 48 РСБ и 57 экзонов, образующих 237 регуляторных пар РСБ-событие, в том числе 203 случая перекрестной регуляции и 34 случая ауторегуляции. Каталог аннотированных событий включал в себя 5271 событие (табл. 8). Для исследования взаимосвязи между уровнем экспрессии гена и метрикой Ψ были использованы 8500 образцов секвенирования РНК из GTEx. После исключения некоторых тканей, в которых действие системы NMD значительно отличается от действия в других тканях, и конститутивных

Таблица 8 — Число событий непродуктивного сплайсинга.

Группа	Валидированные	Аннотированные	Всего
Все	48	2754	2802
Значимые	11	568	579
Тканеспецифические	5	132	137
Тканеспецифически регулируемые	4	80	84
Из них с поддержкой CLIP в гене	4	46	50
Из них с локальной поддержкой CLIP	3	24	27

событий в списке осталось 2754 из 5271 аннотированных случаев и 48 из 57 валидированных случаев. Согласованные изменения между уровнем экспрессии и сплайсингом характеризовались разностью $\Psi_H - \Psi_L$ медианных значений Ψ и разностью Δe_g медианных значений уровня экспрессии гена-мишени между верхним и нижним квартилями распределения Ψ. Значимость изменений e_g между квартилями оценивалась с помощью z-значения из критерия Манна-Уитни. Было найдено 579 значимых событий с $z < -5$, среди которых было 11 валидированных событий, в большинстве из которых Ψ значительно изменялось между квартилями (рис. 6.3А).

Наиболее сильная связь между Ψ и e_g наблюдалась для хорошо известных мишеней фактора PTBP1, таких как *GABBR1* и *DLG4*, для которых разность $\Psi_H - \Psi_L$ была порядка 50% и сопровождалась более чем двукратным снижением уровня экспрессии (рис. 6.3Б) [517—519]. Другие примеры включали в себя события непродуктивного сплайсинга в гене *RBM10*, который является важным компонентом сплайсосомы и связан с наследственными и онкологическими заболеваниями, и гене *TRA2A* — онкогеном из семейства SR-богатых белков [520—522]. В *RBM10* пропуск необходимого экзона 6 приводит к снижению экспрессии, что способствует пролиферации опухолевых клеток при аденокарциноме легкого [189; 523]. В *TRA2A* подавление экспрессии происходит за счет стимуляции включения ядовитого экзона 2 продуктом гена TRA2A и его паралогом TRA2B [202]. Значительные тканеспецифические отклонения медианных значений Ψ и e_g наблюдались для 137 событий (табл. 8). Примечательно, что мозжечок характеризовался одновременным усилением экспрессии

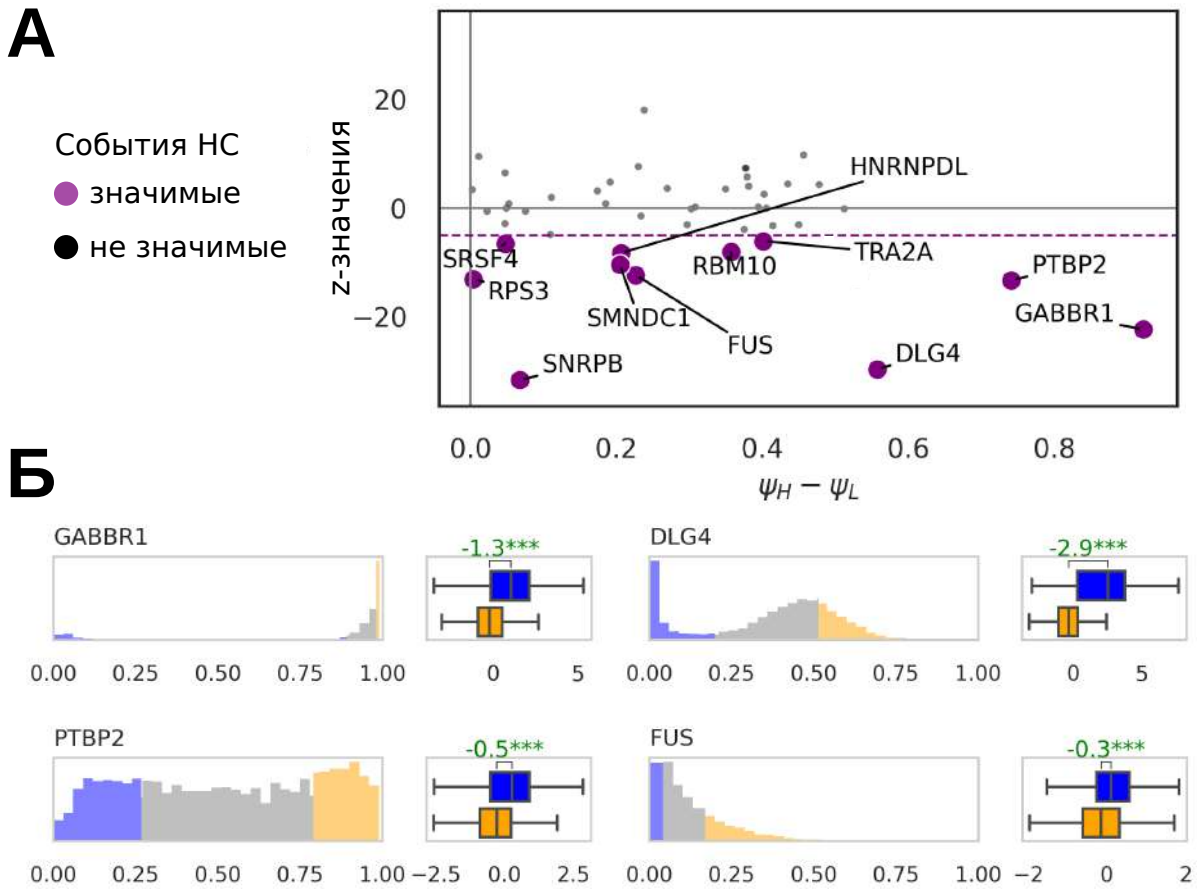


Рисунок 6.3 — Транскриптомные подписи валидированных событий кросс-регуляторного непродуктивного сплайсинга. **(А)** Значимые валидированные события характеризуются $z < -5$. $\Psi_H - \Psi_L$ обозначает разницу медиан в верхнем и нижнем квартилях распределения Ψ. **(Б)** Некоторые примеры распределения Ψ для предсказаний из (А). Верхние 25% и нижние 25% распределения значений Ψ показаны оранжевым и синим цветами. Ящичковые диаграммы показывают распределение значений e_g в верхнем и нижнем квартиле распределения Ψ. Зеленым цветом показаны значения Δe_g. Символ *** обозначает статистически различимые различия на уровне значимости 0.1%.

изоформ NMD и снижением экспрессии многих генов, тогда как в мозге, крови, мышцах и сердце картина была противоположной. Это наблюдение подтверждает особую роль NMD в организации транскрипционных программ мозжечка на фоне более высокого количества событий AC и их вклада в развитие этого отдела мозга [524; 525].

Затем оценивалась реактивность событий непродуктивного сплайсинга по отношении к инактивации РСБ из базы данных ENCODE аналогично тому, как это было сделано в разд. 6.1. Такая оценка была возможна в 124 из 203 подтвержденных случаев кросс-регуляции, среди которых было найдено 30 пар РСБ-событие с ожидаемым направлением регуляции и одна пара, в которой направление было противоположным известному из литературы. Применительно к тканеспецифическим событиям по крайней мере один регулятор был найден

в 113 случаях, 84 из которых были тканеспецифически регулируемыми, т.е. уровень экспрессии тканеспецифического РСБ сопровождался тканеспецифичными изменениями Ψ (табл. 8). Для выявления причинно-следственных связей список кандидатов был дополнительно ограничен требованием, чтобы РСБ связывались с пре-мРНК гена-мишени, используя эксперименты CLIP из базы данных POSTAR3 [516]. Для 50 событий был обнаружен хотя был один сигнал тканеспецифического регулятора РСБ в гене, а в 27 случаях наблюдалось связывание РСБ в непосредственной близости от самого события. Сопоставление тканей GTEx с тканями из базы данных Protein Atlas в ряде случаев позволило оценить изменения уровня экспрессии гена-мишени не только на уровне мРНК, но и на уровне белка. Однако ввиду разреженности протеомных данных эта информация использовалась только как консультативная и в фильтрации событий не участвовала.

6.2.2 Экспериментальная валидация непродуктивного сплайсинга в генах *DCLK2* и *IQGAP1*

Кластеризация 27 тканеспецифически регулируемых событий с локальной поддержкой CLIP выявила четыре группы событий, которые характеризуются снижением Ψ в мозге (кластер 1), увеличением Ψ в мозге (кластер 2), снижением Ψ в крови (кластер 3) и снижением Ψ в скелетных мышцах и сердце (кластер 4). Для этих событий была построена кросс-регуляторная сеть, в которой PTBP1 контролирует экспрессию семи различных мишеней (рис. 6.4). Отметим, что в ней присутствуют события в генах *DLG4* и *GABBR1*, которые подавляются в мозге и активируются в других тканях [517; 526]. Для экспериментальной валидации неизвестных ранее регуляторных событий были выбраны два гена, которые также являются мишенями PTBP1.

Первый из них, *DCLK2*, необходим для развития гиппокампа и регуляции роста дендритов [527; 528]. Закономерности в изменении уровня экспрессии *DCLK2* и частоты включения ядовитого экзона 16 указывают на то, что PTBP1 стимулирует включение этого экзона в неневральных тканях, что приводит к подавлению экспрессии *DCLK2* (рис. 6.5A, слева). Наличие сигналов CLIP как в направлении 5'-конца, так и в направлении 3'-конца от ядовитого экзона

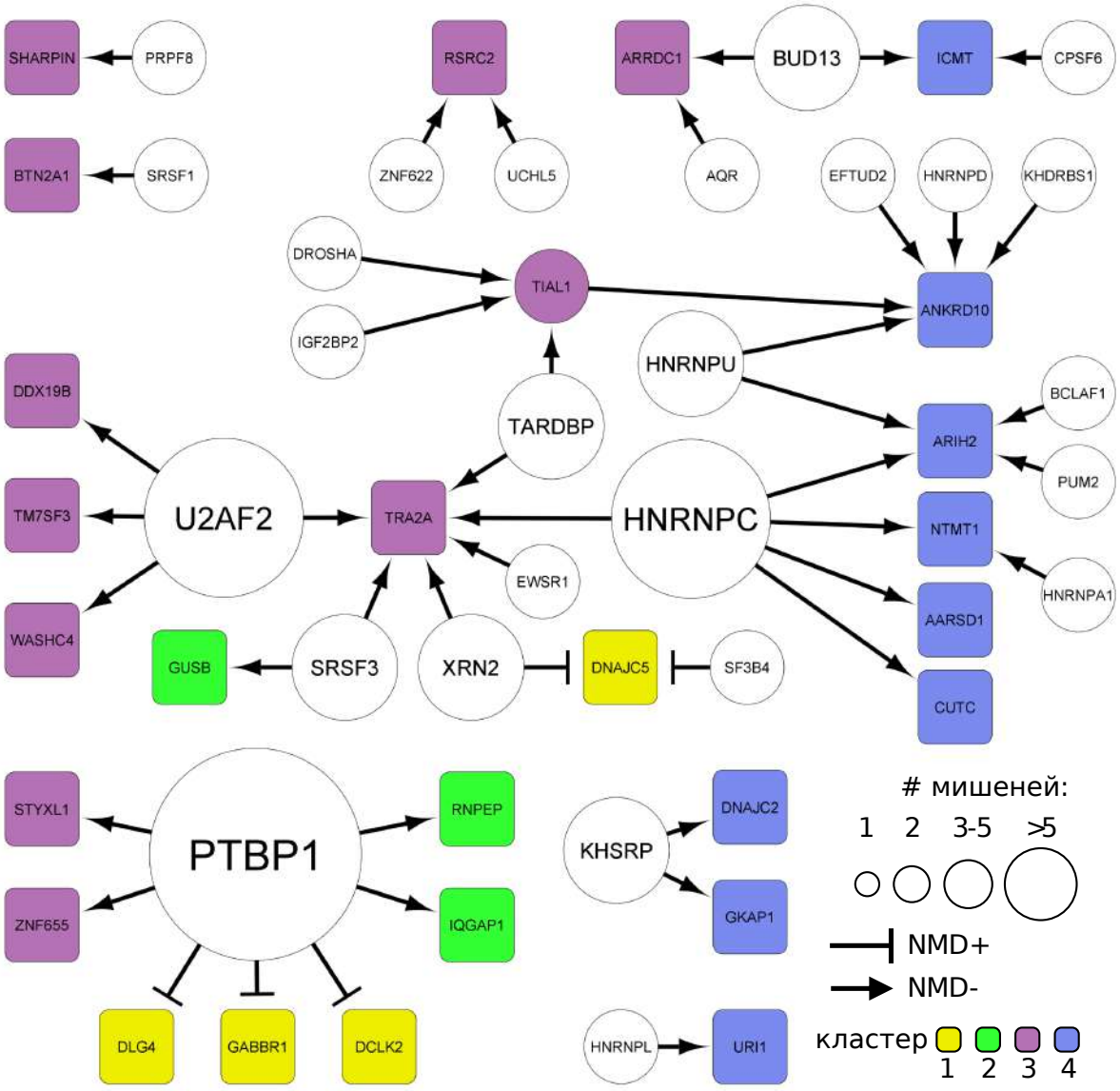


Рисунок 6.4 — Предсказанная сеть тканеспецифически регулируемых событий непродуктивного сплайсинга с локальной поддержкой CLIP. Вершины графа обозначают события и регуляторы. Ребра обозначают регуляторные связи, способствующие образованию (NMD+) и ингибирующие образование (NMD-) NMD-изоформы. События со снижением Ψ в мозге (кластер 1), с увеличением Ψ в мозге (кластер 2), со снижением Ψ в крови (кластер 3) и со снижением Ψ в скелетных мышцах и сердце (кластер 4) показаны цветами.

указывает на неканоническую роль PTBP1 как активатора сплайсинга. Это согласуется с известным из литературы двойственным действием PTBP1, которое зависит от положения его сайта связывания относительно экзона [215]. Для экспериментальной валидации была использована инактивация PTBP1 с помощью специфических микроРНК, которая была подтверждена вестерн-блотом [216]. При инактивации PTBP1 наблюдалось существенное (25%) снижение частоты включения экзона 16 как в исходной клеточной линии A549, так и в условиях, когда путь NMD был инактивирован циклогексимидом (рис. 6.5А, справа).

Второй пример, ген *IQGAP1*, экспрессируется на низком уровне в здоровых тканях головного мозга, а в опухолях его экспрессия увеличивается [529]. Низкий уровень экспрессии *IQGAP1* в мозге и частота включения ядовитого экзона 29, который почти полностью подавляется в неневральных тканях, позволяют предположить, что экспрессия *IQGAP1* в мозге специфически подавляется PTBP1 (рис. 6.5Б, слева). Действительно, наличие сигнала CLIP в направлении 5'-конца от события непродуктивного сплайсинга подтверждает то, что PTBP1, экспрессия которого в мозге подавлена, является супрессором включения экзона 29. В полном согласии с этим предсказанием при подавлении экспрессии PTBP1 наблюдается увеличение частоты включения экзона 29, причем с гораздо более сильным ответом в присутствии циклогексимида (рис. 6.5Б, справа).

6.3 Структура РНК и непродуктивный сплайсинг

Семейство BET, являющееся подгруппой суперсемейства бромодоменовых белков [530—532], состоит из четырех белков — *BRD2*, *BRD3*, *BRD4* и *BRDT*, считывающих хроматиновые метки и имеющих широкую специфичность в отношении активации транскрипции [533; 534]. Они содержат два гомологичных тандемных бромодомена, которые узнают и связывают ацетилированные остатки лизина в гистоновых белках, добавочный концевой (extra terminal, ET) домен и два небольших консервативных мотива, расположенные между двумя бромодоменами и между вторым бромодоменом и ET-доменом [535]. Члены семейства BET возникли в результате серии дупликаций, которые произошли до радиации позвоночных [536].

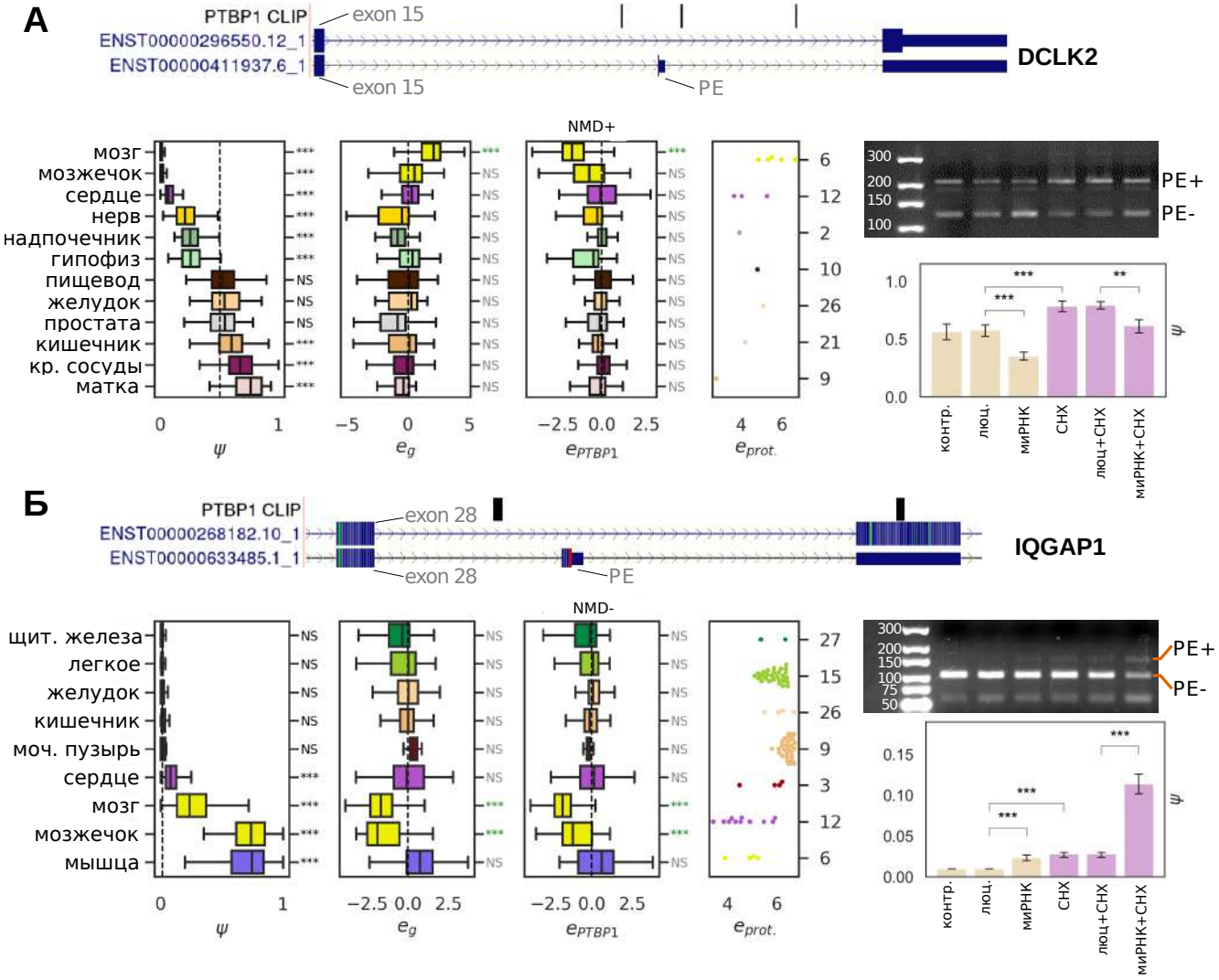


Рисунок 6.5 — Валидация тканеспецифически регулируемых событий непродуктивного сплайсинга в генах *DCLK2* (А) и *IQGAP1* (Б). Ящичковые диаграммы показывают (слева направо) распределение Ψ, уровень экспрессии гена-мишени (e_g), уровень экспрессии регулятора PTBP1 (e_{PTBP1}) и уровень экспрессии белка. Ткани окрашены цветовыми кодами GTEx и отсортированы в порядке возрастания Ψ. Идеограммы на каждой панели показывают положение сигналов CLIP для PTBP1. На панелях справа показаны результаты экспериментов ОТ-ПЦР (вверху) и ОТ-ПЦР-РВ (внизу). PE+ и PE- обозначают изоформы с ядовитым экзоном и без него, соответственно. Дорожки: (слева направо) необработанный контроль, обработка контрольной миРНК против гена люциферазы светлячка, микроРНК против PTBP1, циклогексимидом (CHX), CHX и контрольной микроРНК, CHX и микроРНК против PTBP1. Символы ***, **, * и NS обозначают статистически значимые различия на 0.1%, 1%, 5% уровне значимости и не значимые различия, соответственно.

Бромодоменовые белки и, в частности, члены семейства BET, содержат события непродуктивного сплайсинга. В *BRD2* включение в транскрипт ядовитого экзона 3b, окруженного консервативными интронными последовательностями, приводит к сдвигу рамки считывания и появлению PTC [219; 537]. Другим примером является ген *BRD9*, который содержит ядовитый экзон, включение которого запускает деградацию его мРНК в некоторых опухолях [273]. Как уже отмечалось, среди паралогов, часто содержащих ультраконсервативные элементы вокруг ядовитых экзонов, наблюдается тенденция к развитию множественных кросс-регуляторных сетей [175].

6.3.1 ККУ и ядовитые экзоны в генах *BRD2* и *BRD3*

Предыдущие исследования показали, что семейство BET позвоночных состоит из четырех отдельных ортологичных групп, соответствующих *BRD2*, *BRD3*, *BRD4* и *BRDT*, соответственно, причем наибольшая степень сходства наблюдается между *BRD2* и *BRD4* [536]. Однако исследование далеких гомологов *BRD2* в двусторонне-симметричных позвоночных с помощью базы данных eggNOG [538] подтверждает разделение на четыре ортологичные группы, но свидетельствует о последнем расхождении *BRD2* и *BRD3*, а не *BRD2* и *BRD4* (рис. 6.6А). Гипотеза о том, что ближайшим гомологом *BRD2* является *BRD3*, дополнительно подтверждается взаимосвязью между филогенией и профилями экспрессии этих генов [536], а также тем фактом, что *BRD2* и *BRD3* оба имеют более короткий C-концевой домен по сравнению с *BRD4* и *BRDT* [539].

Ген *BRD2* человека охватывает 13 экзонов, причем стартовый кодон расположен в экзоне 2. Включение экзона 3b длиной 92 нт, следующего за экзоном 3, вызывает сдвиг рамки считывания, который индуцирует PTC в следующем экзоне (рис. 6.6Б). Согласно предсказаниям PREPH, экзон 3b расположен внутри пары ККУ, обозначаемых R1 и R2, со свободной энергией гибридизации $\Delta G = -29.1$ ккал/моль. Хотя уровень изменчивости нуклеотидных последовательностей R1 и R2 недостаточен для оценки значимости компенсаторных замен в них, границы R1 и R2 хорошо коррелируют с профилем эволюционной консервативности, образуя два выраженных пика сигнала phastCons, которые заканчиваются там же, где заканчивается комплементарность. Эта

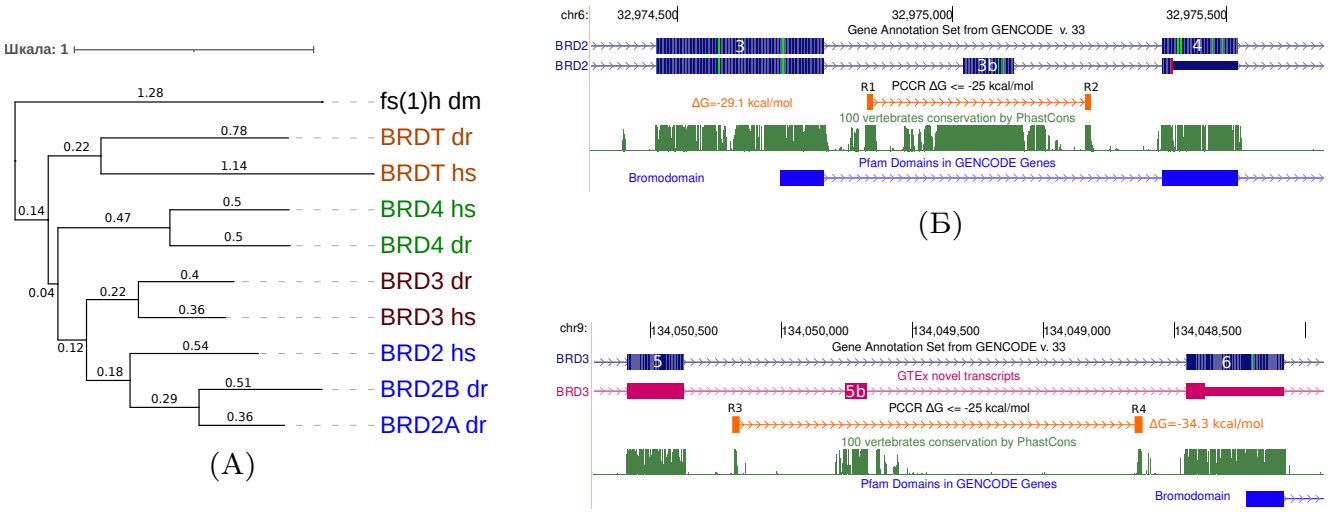


Рисунок 6.6 — Семейство BET-белков. **(А)** Филогенетическое дерево гомологов *BRD2* у позвоночных из базы данных eggNOG, ограниченное *H. Sapiens* и *D. rerio*. Белок *fs(1)h* дрозофилы использовался в качестве внешней группы. Числа на ветвях обозначают среднее количество аминокислотных замен на сайт. **(Б)** Включение экзона 3b в транскрипт *BRD2* приводит к образованию PTC в следующем экзоне. Экзон окружен парой ККУ (R1 и R2) с ΔG=-29.1 ккал/моль. Экзон 3b расположен внутри области, кодирующей бромодомен 1. Снимок из Геномного Браузера UCSC. **(В)** Криптический экзон 5b в *BRD3* не аннотирован и наблюдается в транскриптомах тканей человека (мышца и толстая кишка). Его включение также приводит к образованию PTC в следующем экзоне. Он окружен парой ККУ (R3 и R4) с ΔG=-34.3 ккал/моль. Экзон 5b расположен между бромодоменами 1 и 2. Снимок из Геномного Браузера UCSC.

закономерность была обнаружена ранее как характерное свойство многих функциональных структур РНК [391].

Естественно спросить не содержит ли ген *BRD3*, ближайший гомолог *BRD2*, ядовитые экзоны или другие события непродуктивного сплайсинга. При исследовании гомологичного интрона между экзонами 3 и 4 *BRD3* ядовитого экзона обнаружено не было, однако был найден длинный консервативный элемент в интроне между экзонами 5 и 6 (рис. 6.6Б). Как выяснилось, этот элемент представляет собой неаннотированный экзон 5b длиной 82 нт с каноническими консенсусными последовательностями сайтов сплайсинга GT/AG. Этот экзон обнаруживается в транскриптомах тканей человека (по данным GTEx), а именно в мышцах и толстой кишке. Его включение также приводит к сдвигу рамки считывания, который индуцирует PTC в следующем экзоне. Более того, он окружен парой ККУ, обозначаемых R3 и R4 и напоминающих R1 и R2 в гене *BRD2*. Исследование нуклеотидных последовательностей R3 и R4 показало, что они комплементарны друг другу с $\Delta G = -34.3$ ккал/моль, причем их консервативность резко снижается как только заканчивается комплементар-

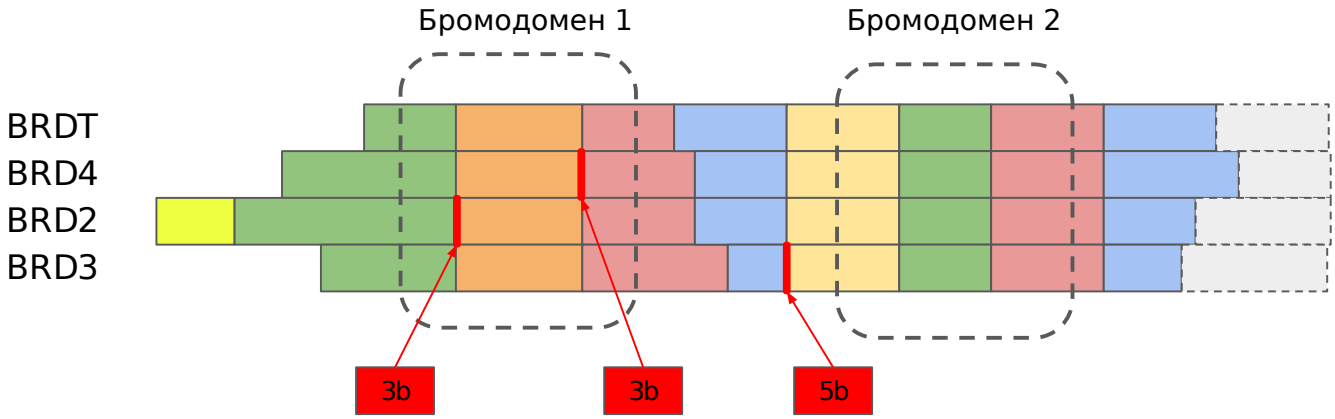


Рисунок 6.7 — Консервативность экзонных границ в семействе BET-белков. Показана схема множественного выравнивания последовательностей BET-белков человека. Цветные блоки обозначают разные экзоны. Области, кодирующие бромодомены 1 и 2, показаны пунктирными линиями. Красные вертикальные линии обозначают ЭЭС, соответствующие ядовитым экзонам (последние показаны ниже в виде красных прямоугольников).

ность. Интересно отметить, что эта пара ККУ была предсказана IRBIS, но не PREPH, поскольку один из рассматриваемых интервалов выходит за рамки консервативных элементов РНК [23; 372].

Как показывает множественное выравнивание аминокислотных последовательностей человеческих белков семейства BET, расположение границ экзонов практически не изменилось после их расхождения (рис. 6.7). При этом экзон 3b у *BRD2* и экзон 5b у *BRD3* расположены в негомологичных интронах, один из которых разделяет экзоны, кодирующие бромодомен, а другой расположен между бромодоменами. В гене *BRD4* обнаруживается ядовитый экзон 3b, но он расположен в еще одном, третьем негомологичном интроне. Последний, и самый отдаленный член семейства BET, *BRDT*, не имеет аннотированных NMD-изоформ и не содержит каких-либо экспрессируемых консервативных интронных элементов, которые могли бы представлять собой ядовитые экзоны. Интересно, что экзон 3b *BRD2* можно проследить до амфибий, тогда как ядовитые экзоны в *BRD3* и *BRD4* впервые появляются у млекопитающих.

6.3.2 Экспериментальная валидация влияния структур РНК на непродуктивный сплайсинг

Для проверки влияния структуры РНК на сплайсинг ядовитых экзонов в *BRD2* и *BRD3* использовалась та же стратегия, что и для других генов,

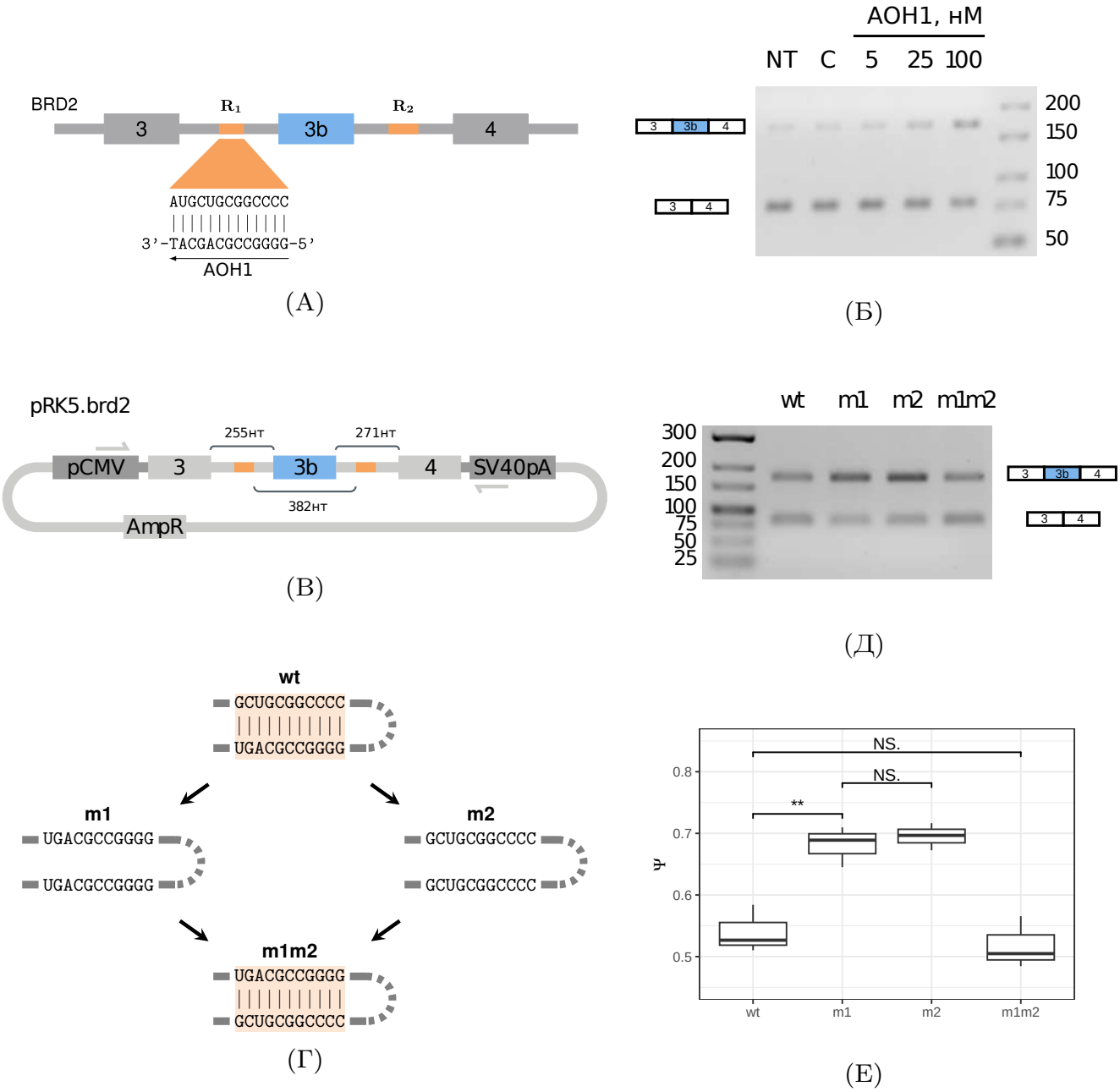


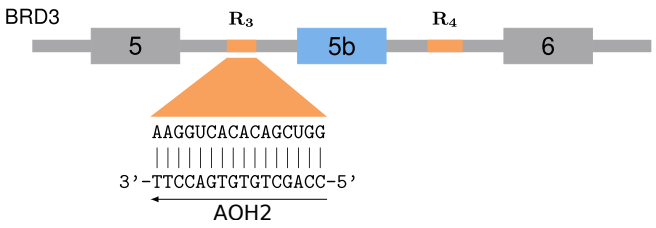
Рисунок 6.8 — Проверка влияния структуры РНК на включение ядовитого экзона 3b в гене *BRD2*. **(А)** AOH1 комплементарен последовательности R1. **(Б)** Частота включения экзона 3b увеличивается при разрушении комплементарности R1/R2 с помощью AOH1. NT — необработанный контроль, C – контрольный AOH против люциферазы. Концентрации AOH указаны в нМ. **(В)** Миниген pRK5.BRD2, содержащий фрагмент *BRD2* между экзонами 3 и 4. Показаны промотор pCMV и сигнал раннего полиаденилирования SV40. **(Г)** Стратегия внесения разрушающих структуру мутаций (m1 и m2) и двойной компенсаторной мутации (m1m2). **(Г)** Частота включения экзона 3b увеличивается при разрушении структуры, но восстанавливается у компенсаторного двойного мутанта. WT — дикий тип. Символы ** и NS обозначают статистически значимые различия на уровне значимости 1% и не значимые различия, соответственно.

т.е. сначала измерялось изменение сплайсинга в эндогенной пре-мРНК в ответ на блокировку структуры АОН, а затем конструировались минигены, несущие мутированные фрагменты генов, для исследования сплайсинга у одиночных и двойных мутантов. Обработка АОН, комплементарным последовательности R1 (рис. 6.8А), показала, что частота включения экзона 3b в *BRD2* заметно увеличивается даже при низкой концентрации АОН1 (рис. 6.8Б). В присутствии циклогексимида уровень включения экзона 3b после обработки АОН1 также значимо повышался [351]. В дополнение к этому был исследован ответ транскриптома клеточной линии HeLa на кодеплецию UPF1 и XRN1 [173; 367]. Оказалось, что частота включения экзона 3b *BRD2* увеличивается с 16% до 40% при инактивации NMD, что дополнительно подтверждает то, что экзон 3b является ядовитым.

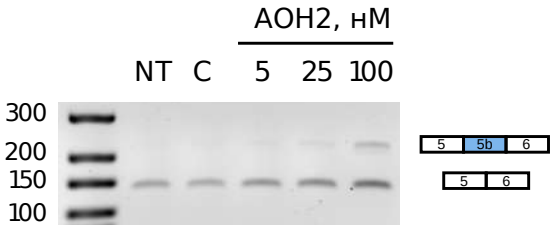
Затем был сконструирован миниген, содержащий фрагмент *BRD2* между экзонами 3 и 4 (рис. 6.8В) с одиночными мутациями, разрушающими структуру РНК и двойными компенсаторными мутациями (рис. 6.8Г). Трансфекция этих конструкций в клетки A549, ОТ-ПЦР и ОТ-ПЦР-РВ показали, что у мутантов m1 и m2 степень включения экзона 3b значительно увеличена, а у двойного мутанта m1m2 соотношение сплайс-изоформ возвращается к таковому у дикого типа (рис. 6.8Д,Е).

Согласно ОТ-ПЦР-РВ, в *BRD3* обработка АОН, комплементарным последовательности R3 (рис. 6.9А), приводила к большему включению экзона 5b (рис. 6.9Б), причем значительно большему в присутствии циклогексимида [351]. Трансфекция клеток A549 минигенными конструкциями, содержащими *BRD3* (рис. 6.9В) с разрушающими структуру (m3 и m4) и компенсаторными (m3m4) мутациями (рис. 6.9Г) показала, что уровень включения экзона 5b увеличивается при разрушении структуры и возвращается в исходное состояние при восстановлении структуры в двойном мутанте (рис. 6.9Д,Е). Следует отметить, что изоформы *BRD3*, содержащие криптический экзон 5b, имеют крайне низкие уровни экспрессии.

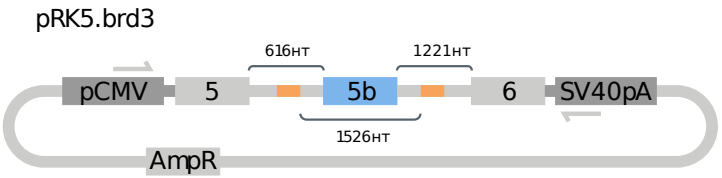
В совокупности эти эксперименты демонстрируют, что комплементарные спаривания вокруг экзона 3b гена *BRD2* и экзона 5b гена *BRD3* действительно регулируют непродуктивный сплайсинг.



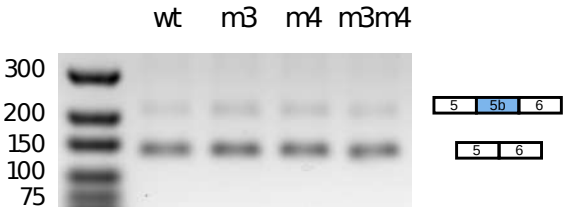
(А)



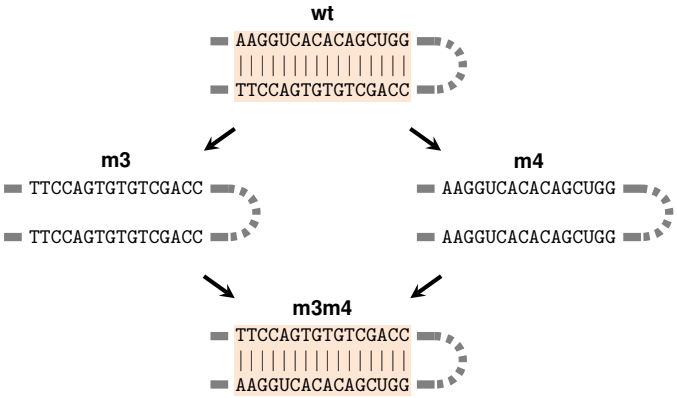
(Б)



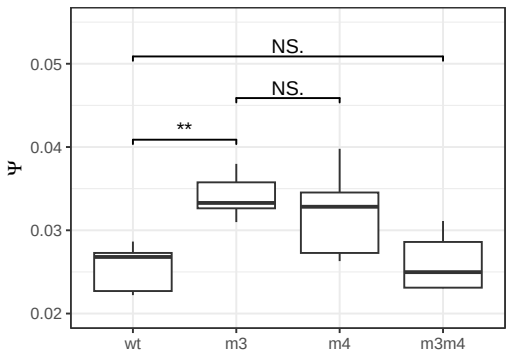
(В)



(Д)



(Г)



(Е)

Рисунок 6.9 — Проверка влияния структуры РНК на включение ядовитого экзона 5b в гене *BRD3*. **(А)** AOH2 комплементарен последовательности R3. **(Б)** Частота включения экзона 5b увеличивается при разрушении комплементарности R3/R4 с помощью AOH2. **(В)** Миниген pRK5.BRD3, содержащий фрагмент гена *BRD3* между экзонами 5 и 6. **(Г)** Стратегия внесения разрушающих структуру мутаций (m3 и m4) и двойной компенсаторной мутации (m3m4). **(Д)** Частота включения экзона 5b увеличивается при разрушении структуры, но восстанавливается у компенсаторного двойного мутанта. Остальные обозначения как на рис. 6.8.

6.3.3 Непродуктивный сплайсинг *BRD2* и *BRD3* в тканях и опухолях

Для характеризации взаимосвязи между экспрессией и непродуктивным сплайсингом генов *BRD2* и *BRD3* в тканях человека использовались данные секвенирования РНК из консорциума GTEx. Сравнение медианного уровня экспрессии ($\log_{10} TPM$) транскриптов *BRD2* в ткани с медианной частотой включения ядовитого экзона 3b (Ψ) обнаружило отрицательную корреляцию ($r_p = -0.34$, рис. 6.10А). Отметим, что высокий уровень экспрессии *BRD2* и низкая частота включения ядовитого экзона 3b наблюдались в семенниках, в согласии с литературными данными [540].

Аналогичная взаимосвязь наблюдалась при сравнении уровня экспрессии *BRD2* и частоты включения экзона 3b в транскриптомах опухолей из проекта TCGA (The Cancer Genome Atlas). Изменение медианного уровня экспрессии *BRD2* ($\Delta \log_{10} TPM$, опухоль по сравнению с нормальной тканью) отрицательно коррелирует с медианной частотой изменения степени включения экзона 3b ($\Delta\Psi$) для восемнадцати изученных типов опухолей ($r_p = -0.3$, рис. 6.10Б). Также отметим, что наибольшее увеличение степени включения экзона 3b, сопровождающееся существенным снижением экспрессии *BRD2*, наблюдалось при светлоклеточной карциноме почки (KIRC) и папиллярно-клеточной карциноме почки (KIRP) — двух типах рака, в которых снижение уровня *BRD2* ассоциировано с плохим прогнозом [541]. Аналогичное исследование для ядовитого экзона *BRD3* возможно только в части тканей, поскольку он экспрессируются на гораздо более низком уровне по сравнению с *BRD2*. Тем не менее, медианный уровень экспрессии и медианный уровень включения экзона 5b в тканях также оказались отрицательно связаны ($r_p = -0.25$).

Эти наблюдения согласуются с тем, что экзоны 3b и 5b являются ядовитыми, однако остается открытым вопрос о том, зачем нужно регулировать их уровень включения при помощи структуры РНК. Ответ на него дают эксперименты по изучению реакции транскриптома на замедление элонгации транскрипции: один при обработке α-аманитином (разд. 3.3), а другой — в медленном мутанте RNAPII (R749H) [363]. В обоих этих экспериментах наблюдалось значимое снижение частоты включения экзона 3b при замедлении элонгации транскрипции (рис. 6.10В,Г). В соответствии с пропуском ядовитого

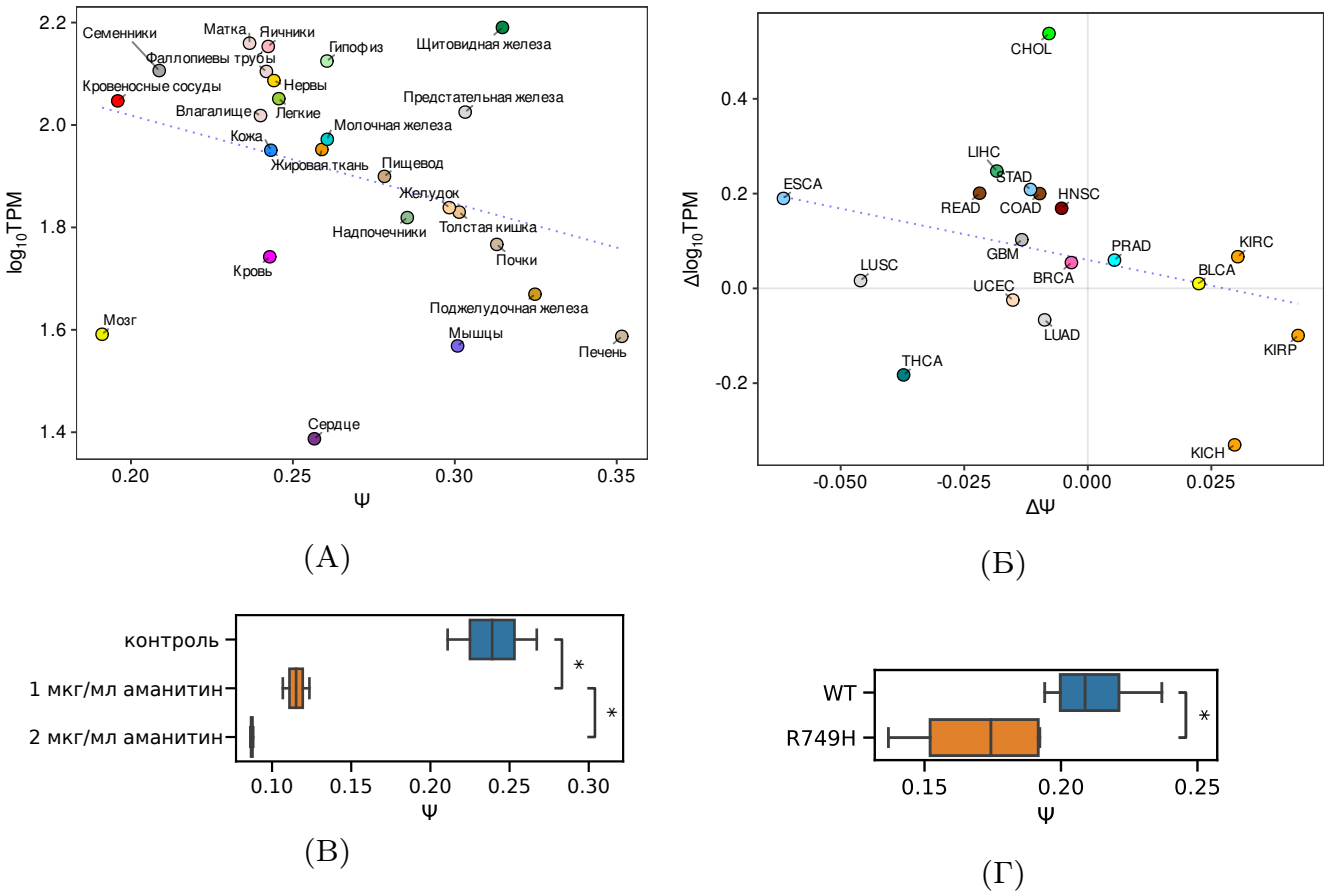


Рисунок 6.10 — Экспрессия и сплайсинг *BRD2* в тканях и опухолях. **(А)** Медианная степень включения экзона 3b (Ψ) в ткани отрицательно связана ($r_p = -0.34$) с $\log_{10}$-трансформированной медианой уровня экспрессии транскриптов *BRD2* ($\log_{10} TPM$). Цветовые коды тканей такие же, как в [216]. r_p — коэффициент корреляции Пирсона. **(Б)** Изменение средней частоты включения экзона 3b ($\Delta\Psi$, опухоль против нормальной ткани) отрицательно коррелирует ($r_p = -0.3$) с изменением уровня экспрессии транскриптов ($\Delta \log_{10} TPM$, опухоль против нормальной ткани). Цветовые коды и сокращения для опухолей приведены в [216]. **(В)** Изменение уровня включения экзона 3b (Ψ) в ответ на обработку α-аманитином. **(Г)** Изменение уровня включения экзона 3b в медленном мутанте RNAPII (R749H) по сравнению с диким типом (wt).

экзона уровень экспрессии *BRD2* ожидаемо увеличивается, однако это увеличение не было статистически значимым. Реакция ядовитых экзонов в *BRD3* и *BRD4* на замедление элонгации транскрипции не может быть достоверно оценена из-за низкого уровня экспрессии этих генов. Таким образом, как и в случае *ATE1*, структура РНК может регулировать уровни экспрессии этих генов через непродуктивный сплайсинг в зависимости от скорости элонгации транскрипции, чем может обуславливаться тканеспецифическая экспрессия в семенниках, где также экспрессируется и фактор NELFE (разд. 5.2.3).

6.4 Обсуждение результатов и выводы

6.4.1 Предсказание регуляции непродуктивного сплайсинга по транскриптомным данным

Регуляция экспрессии генов осуществляется множеством различных механизмов, которые определяют изменение количества мРНК в ответ на изменение внешних условий [152]. Все этапы генной экспрессии, от инициации транскрипции до посттрансляционной модификации белков, контролируются сложной регуляторной сетью, в которой один ген контролирует экспрессию и сам контролируется экспрессией множества других генов. Часто наблюдаемым элементарным звеном в таких сетях является ауторегуляция, которая обеспечивает простую и, пожалуй, самую надежную обратную связь, не требующую каких-либо промежуточных звеньев. Например, многие транскрипционные факторы у бактерий регулируют свою экспрессию, связываясь со своим собственным промотором и активируя или подавляя транскрипцию [542; 543].

Хотя некоторые эукариотические гены используют петли ауторегуляции на уровне транскрипции, их экспрессия также может регулироваться посттранскрипционно [544]. Например, связывание фактора YBX1 с регуляторным элементом в 3'-НТО его собственной мРНК специфически подавляет трансляцию этого гена [545]. Однако основной способ посттранскрипционной регуляции экспрессии эукариотических генов связан с изменением стабильности мРНК [546]. В частности, скоординированное взаимодействие между AC и NMD, приводящее к деградации мРНК и представляющее собой непродуктивный сплайсинг, широко распространено почти у всех эукариот [176; 219; 547; 548]. По сравнению с другими механизмами посттранскрипционной регуляции, такими как эндогенная РНК-интерференция [549—552] и контроль стабильности мРНК с помощью модификаций РНК [553—555], непродуктивный сплайсинг, по-видимому, действует повсеместно на уровне транскриптома, о чем свидетельствует тот факт, что почти треть генов, кодирующих человеческие белки, имеют по крайней мере одну аннотированную NMD изоформу, и многие из этих изоформ демонстрируют эволюционную консервативность [556].

Представленное исследование, основанное на изучении ответа транскриптома на инактивацию компонентов системы NMD, инактивацию экспрессии гена-хозяина и данных о связывании РСБ со своей собственной мРНК, позволило выявить новые случаи ауторегуляторного непродуктивного сплайсинга. Однако экзоны в генах с известными ауторегуляторными петлями, таких как *PTBP1* и *TARDBP*, продемонстрировали незначительные изменения сплайсинга, в числе прочего, из-за недостаточной эффективности подавления экспрессии этих РСБ. Поэтому следует ожидать, что представленный подход должен давать больше ложноотрицательных, чем ложноположительных предсказаний, поскольку в случае неэффективного подавления экспрессии $\Delta\Psi$ скорее недооценивается, чем переоценивается. Ложноположительные предсказания проистекают из оценок $\Delta\Psi$, обусловленных косвенными реакциями в регуляторных генных сетях, расхождениями между экспериментами в различных клеточных линиях, а также кросс-реактивностью РСБ в экспериментах eCLIP.

Следует отметить, что механизмы ауторегуляции могут действовать только в особом диапазоне клеточных концентраций РСБ, в котором может реализоваться нелинейный кооперативный механизм активации или репрессии. Например, если РСБ экспрессируется умеренно, а регуляторная петля обратной связи не активирована, то частота включения ядовитого экзона должна быть низкой, и дальнейшее снижение уровня экспрессии приведет к ее дальнейшему уменьшению. Это объясняет отсутствие ответа на инактивацию гена-хозяина у некоторых РСБ. Кроме этого, при ауторегуляции продукт гена, который содержит ядовитый экзон, противодействует своей собственной активации и приводит к подавлению NMD-изоформы. Деградация непродуктивных изоформ и постоянный приток пре-мРНК из-за продолжающейся транскрипции искажают реактивность экзонов на инактивацию РСБ, поскольку система NMD все еще остается активной [188].

Существование петель обратной связи и непрямых взаимодействий в сплайсинговых сетях является фундаментальной проблемой и при изучении кросс-регуляторных сетей. Например, *PTBP1* усиливает экспрессию *SRSF3*, который, в свою очередь, усиливает экспрессию *PTBP2*, но сам при этом *PTBP1* подавляет экспрессию *PTBP2*. Знак и абсолютная величина связи между $\Delta\Psi$ и экспрессией РСБ могут варьироваться в зависимости от связности в сети, что потенциально приводит как к ложноположительным, так и к ложноотрицательным предсказаниям. Непродуктивный сплайсинг может

быть неактивным в полностью дифференцированных тканях или действовать только в определенных условиях, таких как клеточная дифференцировка [209; 271], нейрогенез [519; 526] или гипоксия [212]. Следовательно, регуляторный потенциал непродуктивного сплайсинга требует относительно большой величины изменений AC. Все эти факторы ограничивают чувствительность предлагаемого метода.

Несмотря на все эти ограничения, представленная методология была успешно применена к идентификации тканеспецифичных событий непродуктивного сплайсинга и их регуляторов. Предсказаны 27 новых регуляторных событий непродуктивного сплайсинга, включая экспериментально валидированные мишени PTBP1 в генах *DCLK2* и *IQGAP1*. Эти результаты значительно расширяют текущие знания о тканеспецифической регуляции непродуктивного сплайсинга и открывают новые возможности для будущих исследований.

6.4.2 Конвергентная эволюция непродуктивного сплайсинга

Гены-паралоги, возникающие в результате дупликации и дивергенции, часто сохраняют значительную степень сходства не только между кодируемыми ими белками, но также и между цис-регуляторными элементами и механизмами регуляторного контроля [557]. В частности, это справедливо для многих факторов сплайсинга, многие из которых эволюционировали посредством дупликаций и содержат гомологичные цис-регуляторные элементы, которые связаны с непродуктивным сплайсингом [178]. Дальние взаимодействия в структуре РНК, обнаруженные в генах *BRD2* и *BRD3*, контролируют непродуктивный сплайсинг, что представляет собой важный пример того, что функция структуры РНК может быть связана с посттранскрипционной регуляцией экспрессии генов, а не с модуляцией AC и изменениями в результирующем белковом продукте.

Удивительно то, что в случае *BRD2* и *BRD3* регуляторные ядовитые экзоны и окружающие их комплементарные участки расположены в интронах, расположенных между негомологичными экзонами. Это говорит о том, что *BRD2* и *BRD3* либо приобрели их независимо, либо каждый независимо потерял по одному из них в процессе эволюции. Тот факт, что *BRD4* содержит

ядовитый экзон, но не имеет вокруг него структуры РНК, а также то, что *BRDT* и *fs(1)h*, гомолог белков BET семейства у беспозвоночных, не имеют аннотированных NMD-изоформ указывает на то, что имел место первый из этих двух сценариев. Более того, комплементарные участки, окружающие ядовитые экзоны в *BRD2* и *BRD3*, не имеют никакого сходства друг с другом по последовательности. Это означает, что регуляторные структуры РНК вокруг ядовитых экзонов являются результатом конвергентной эволюции. Быстрая потеря и повторное приобретение ядовитых экзонов [175] и независимое происхождение конкурирующих вторичных структур РНК, которые контролируют взаимоисключающий сплайсинг тандемно дуплицированных экзонов [509], указывают на то, что множество структур РНК, регулирующих непродуктивный сплайсинг, с большой долей вероятности выходит за рамки приведенных здесь примеров.

6.4.3 Структура РНК и регуляция непродуктивного сплайсинга

Можно предположить, что движущей силой конвергентной эволюции является способность структуры РНК регулировать непродуктивный сплайсинг через скорость элонгации транскрипции. Как было показано в разд. 5.2.3 на примере гена *ATE1*, замедление элонгации транскрипции дает достаточно времени для сворачивания структуры РНК, что, в свою очередь, способствует пропуску выпетливаемого экзона. Как и в случае *ATE1*, замедляющий RNAPII фактор NELFE [500—502] может определять пропуск ядовитого экзона и повышенную экспрессию *BRD2* и *BRD3* в семенниках. SR-богатые факторы сплайсинга часто реализуют петлю отрицательной обратной связи, которая подавляет продуктивную сплайс-изоформу в ответ на повышенную экспрессию генного продукта [173]. Аналогичная петля обратной связи может существовать и в *BRD2*, однако она действует посредством АС, а не при помощи замедления RNAPII, поскольку *BRD2* сам по себе не изменяет скорость элонгации транскрипции [558].

Вполне возможно, что регуляторная структура РНК в гене *BRD2* продолжается и вне описанных здесь ККУ, поскольку интрон в направлении 3'-конца от экзона 3b содержит полиморфный микросателлит, количество GT-повторов в котором отрицательно коррелирует с его степенью включения [537]. После-

довательности низкой сложности, такие как GT-тракты, способны влиять на АС из-за образования вторичной структуры РНК [559]. Хотя структуры РНК, образованные участками R1/R2 и R3/R4, были идентифицированы на основе эволюционной консервативности, пары оснований, важные для непродуктивного сплайсинга, вполне могут существовать и у других членов семейства BET, но за пределами консервативных областей. Следует также отметить тот факт, что ядовитый экзон 5b в гене *BRD3* не был аннотирован как экзон ни в одной базе данных и был обнаружен в данной работе из соображений консервативности нуклеотидной последовательности в результате исследования низкоэкспрессирующихся сплайс-изоформ в транскриптомах тканей человека. Это показывает, что важные события непродуктивного сплайсинга часто отсутствуют в базах данных из-за систематической недоаннотации NMD-мишеней.

Изучение механизмов, лежащих в основе специфических функций белков BET-семейства, имеет важное значение для разработки терапевтических решений, поскольку аномальная экспрессия этих белков приводит к онкологическим, метаболическим и сердечно-сосудистым заболеваниям [560—562]. Подход, основанный на применении синтетических АОН, является эффективным для модуляции непродуктивного сплайсинга у бромодоменовых белков. Например, АОН, которые способствуют пропуску ядовитого экзона 14a, что приводит к усилению деградации мРНК гена *BRD9* в опухолях, помогают восстановить уровни белка *BRD9* и остановить рост опухоли [273]. АОН, комплементарные элементам структуры РНК в генах *BRD2* и *BRD3*, предлагают уникальную возможность не только подавлять, но и терапевтически увеличивать степень включения ядовитых экзонов. Что еще более важно, доказательства конвергентной эволюции ядовитых экзонов и структур РНК в *BRD2* и в *BRD3* проливают свет на происхождение механизмов регуляции непродуктивного сплайсинга в этих и многих других паралогичных генах.

Заключение

В диссертационной работе разработаны новые методы предсказания дальних взаимодействий в структуре РНК. Сопоставление их предсказаний с экспериментальными данными, в частности, данными конформационного секвенирования РНК *in situ*, а также применение полученных результатов к исследованию влияния структуры РНК на альтернативный сплайсинг позволило сделать следующие **выводы**:

1. Элементы структуры РНК предпочтительно располагаются в интронах, подавляют использование криптических сплайс-сайтов и выпетливаемых экзонов, обогащены сайтами редактирования РНК и сайтами связывания РНК-связывающих белков, и поддерживаются данными конформационного секвенирования РНК *in situ*.
2. При замедлении элонгации транскрипции изменение частоты включения экзона зависит от структурированности предшествующего интрона.
3. На примере генов *CG33298*, *Gug*, *Nmnat*, *PHF20L1*, *CASK*, *ATE1*, *SF1* и *MARK2* показано, что дальние взаимодействия в структуре РНК могут регулировать все основные типы событий альтернативного сплайсинга и альтернативное полиаденилирование.
4. Структура РНК в гене *ATE1* состоит из двух функционально различных модулей, один из которых обеспечивает взаимоисключающий сплайсинг, а другой через дальние взаимодействия на расстоянии 30000 п.о. контролирует соотношение сплайс-изоформ в процессе котранскрипционного сворачивания пре-мРНК.
5. На примере генов *DCLK2*, *IQGAP1*, *BRD2* и *BRD3* показана регуляция непродуктивного сплайсинга фактором PTBP1 и дальними взаимодействиями в структуре РНК.

В целом диссертационная работа опровергает распространенное представление об эукариотических РНК как о длинных и неструктурированных молекулах, напоминающей спагетти, которые складываются в древовидные структуры, состоящие из шпилек, стеблей и внутренних петель. В действительности эукариотические РНК высокоструктурированы, а дальние взаимодействия в их структуре образуют псевдоузлы, предсказание которых

классическими методами невозможно. Приведенные в диссертации примеры показывают исключительную важность дальних взаимодействий для регуляции всех основных типов альтернативного сплайсинга и демонстрируют возможность воздействия на него через структуру РНК с помощью антисмысловых олигонуклеотидов, что имеет важное практическое значение. Суммарно полученные результаты показывают, что дальние взаимодействия в структуре РНК широко распространены в генах эукариот и координируют процессинг РНК во времени и в пространстве на больших расстояниях.

Дальнейшее развитие методов предсказания структуры РНК зависит от нескольких ключевых вопросов, из которых представляется важным отметить следующие. Во-первых, современные термодинамические модели структуры РНК основываются на оцененных в 1999 году энергетических параметрах, которые давно требуют пересмотра. Не исключено, что методы высокопроизводительного секвенирования в будущем смогут помочь измерить большее число таких параметров с большей точностью. Во-вторых, входными данными для филогенетических методов предсказания структуры РНК являются множественные выравнивания нуклеотидных последовательностей. Поскольку полногеномные множественные выравнивания по построению разрывны и не всегда однозначны, задача выравнивания последовательностей интронов должна основываться на построении ортологических рядов.

В заключение автор выражает благодарность и искреннюю признательность всем своим коллегам, которые прямо или косвенно принимали участие в этой работе: проф. О.А. Донцовой, к.х.н Д.А. Скворцову, проф. А.А. Миронову, проф. П.М. Рубцову, проф. М.С. Гельфанду, проф. Родригу Гигó, проф. Хуану Валкарселю, проф. Рори Джонсону, проф. Йонфенг Жин, проф. Юаньчао Сюэ, проф. Чанчан Цао, Марине Петровой, а также студентам и аспирантам: Светлане Калмыковой, Сергею Маргасюку, Марии Власенок, Льву Завилейскому, Ярославу Попову, Алексею Миронову, Марине Калининой и Маргарите Воробьевой. Автор также благодарит организаторов и участников международной конференции «Вычислительные подходы к структуре и функциям РНК», проводимой каждые три года в испанском городе Бенаске, за бесконечный источник мотивации и бесценные научные дискуссии. Автор выражает благодарность Сколковскому Институту Науки и Технологии, а также Московскому государственному университету им. М.В. Ломоносова и Центру Геномной Регуляции в г. Барселона, в которых на протяжении последних 15 лет выполнялась эта работа.

Список сокращений

АОН	антисмысловые (антисенс) олигонуклеотиды
АС	альтернативный сплайсинг
ВКК	вложенные кластеры контактов
мяРНК	малая ядерная РНК
мякРНК	малая ядрышковая РНК
мяРНП	малые ядерные рибонуклеопротеины
нт	нуклеотид
НТО	нетранслируемая область, untranslated region, UTR
кДНК	кодирующая ДНК
КИФ	консервативные интронные фрагменты
ККУ	консервативные комплементарные участки
КУ	кодирующий участок
поли(А)	полиаденилирование
п.о.	пара оснований
ОТ	обратная транскрипция
ПЦР	полимеразная цепная реакция
ПЦР-РВ	полимеразная цепная реакция в реальном времени
РСБ	РНК-связывающий белок
РСД	РНК-связывающий домен
ЭЭС	экзон-экзонное соединение
5'ss	5'-сайт сплайсинга (донорный сайт)
3'ss	3'-сайт сплайсинга (акцепторный сайт)
BPS	сайт ветвления, branch point sequence
GQ	G-квадруплекс
GTEx	консорциум Genotype Tissue Expression project
NMD	нонсенс-опосредованный распад, nonsense mediated decay
OR	отношение шансов, odds ratio
PPT	полипиримидиновый тракт,polypyrimidine tract
PTC	преждевременный стоп-кодон, premature termination codon
RNAPII	РНК-полимераза II, RNA polymerase II
TCGA	консорциум Атлас Ракового Генома, The Cancer Genome Atlas
SR	серин/аргинин богатый белок

Список литературы

1. *Singh, N. N.* How RNA structure dictates the usage of a critical exon of spinal muscular atrophy gene [текст] / N. N. Singh, R. N. Singh // Biochim Biophys Acta Gene Regul Mech. — 2019. — т. 1862, № 11/12. — с. 194403.
2. *Singh, N. N.* Modulating role of RNA structure in alternative splicing of a critical exon in the spinal muscular atrophy genes [текст] / N. N. Singh, R. N. Singh, E. J. Androphy // Nucleic Acids Res. — 2007. — т. 35, № 2. — с. 371—389.
3. hnRNP proteins and the biogenesis of mRNA [текст] / G. Dreyfuss [и др.] // Annu Rev Biochem. — 1993. — т. 62. — с. 289—321.
4. *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.
5. Long-range RNA pairings contribute to mutually exclusive splicing [текст] / Y. Yue [и др.] // RNA. — 2016. — янв. — т. 22, № 1. — с. 96—110.
6. The Short- and Long-Range RNA-RNA Interactome of SARS-CoV-2 [текст] / O. Ziv [и др.] // Mol Cell. — 2020. — дек. — т. 80, № 6. — с. 1067—1077.
7. 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.
8. *Cao, D.* Reverse complementary matches simultaneously promote both back-splicing and exon-skipping [текст] / D. Cao // BMC Genomics. — 2021. — авг. — т. 22, № 1. — с. 586.
9. *Nussinov, R.* Fast algorithm for predicting the secondary structure of single-stranded RNA [текст] / R. Nussinov, A. B. Jacobson // Proc Natl Acad Sci U S A. — 1980. — нояб. — т. 77, № 11. — с. 6309—6313.
10. *Zuker, M.* Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information [текст] / M. Zuker, P. Stiegler // Nucleic Acids Res. — 1981. — янв. — т. 9, № 1. — с. 133—148.

11. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs [текст] / T. Xia [и др.] // Biochemistry. — 1998. — окт. — т. 37, № 42. — с. 14719—14735.

12. *Rivas, E.* A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs [текст] / E. Rivas, J. Clements, S. R. Eddy // Nat Methods. — 2017. — янв. — т. 14, № 1. — с. 45—48.

13. *Rivas, E.* RNA structure prediction using positive and negative evolutionary information [текст] / E. Rivas // PLoS Comput Biol. — 2020. — окт. — т. 16, № 10. — e1008387.

14. UFold: fast and accurate RNA secondary structure prediction with deep learning [текст] / L. Fu [и др.] // Nucleic Acids Res. — 2022. — февр. — т. 50, № 3. — e14.

15. *Chen, C.-C.* REDfold: accurate RNA secondary structure prediction using residual encoder-decoder network [текст] / C.-C. Chen, Y.-M. Chan // BMC Bioinformatics. — 2023. — март. — т. 24, № 1. — с. 122.

16. *Shepard, P. J.* Conserved RNA secondary structures promote alternative splicing [текст] / P. J. Shepard, K. J. Hertel // RNA. — 2008. — авг. — т. 14, № 8. — с. 1463—1469.

17. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure [текст] / E. Torarinsson [и др.] // Genome Res. — 2006. — июль. — т. 16, № 7. — с. 885—889.

18. *Will, S.* Structure-based whole-genome realignment reveals many novel noncoding RNAs [текст] / S. Will, M. Yu, B. Berger // Genome Res. — 2013. — июнь. — т. 23, № 6. — с. 1018—1027.

19. RASP: an atlas of transcriptome-wide RNA secondary structure probing data [текст] / P. Li [и др.] // Nucleic Acids Res. — 2021. — янв. — т. 49, № D1. — с. D183—D191.

20. *Reuter, J. S.* RNAstructure: software for RNA secondary structure prediction and analysis [текст] / J. S. Reuter, D. H. Mathews // BMC Bioinformatics. — 2010. — март. — т. 11. — с. 129.

21. RIC-seq for global in situ profiling of RNA-RNA spatial interactions [текст] / Z. Cai [и др.] // Nature. — 2020. — июнь. — т. 582, № 7812. — с. 432—437.

22. IRIS: A method for predicting in vivo RNA secondary structures using PARIS data [текст] / J. Zhou [и др.] // Quant. Biol. — 2020. — т. 8, № 1. — с. 369—381.

23. Conserved long-range base pairings are associated with pre-mRNA processing of human genes [текст] / S. Kalmykova [и др.] // Nature Communications. — 2021. — апр. — т. 12, № 1. — с. 2300. — (1.96 п. л.; Вклад автора 75%; JIF=16.6 WoS).

24. RNA in situ conformation sequencing reveals novel long-range RNA structures with impact on splicing [текст] / S. Margasyuk [и др.] // RNA. — 2023. — сент. — т. 29, № 9. — с. 1423—1436. — (1.62 п. л.; Вклад автора 40%; JIF=3.9 WoS).

25. Long-range RNA structures in the human transcriptome beyond evolutionarily conserved regions [текст] / S. Margasyuk [и др.] // PeerJ. — 2023. — т. 11. — e16414. — (1.96 п. л.; Вклад автора 50%; JIF=2.7 WoS).

26. Expanded encyclopaedias of DNA elements in the human and mouse genomes [текст] / ENCODE Project Consortium [и др.] // Nature. — 2020. — июль. — т. 583, № 7818. — с. 699—710. — (1.39 п. л.; Работа в составе консорциума. Вклад автора менее 5%; JIF=64.8 WoS).

27. Principles of regulatory information conservation between mouse and human [текст] / Y. Cheng [и др.] // Nature. — 2014. — нояб. — т. 515, № 7527. — с. 371—375. — (0.58 п. л.; Работа в составе консорциума. Вклад автора менее 5%; JIF=64.8 WoS).

28. Comparative analysis of the transcriptome across distant species [текст] / M. B. Gerstein [и др.] // Nature. — 2014. — авг. — т. 512, № 7515. — с. 445—448. — (0.46 п. л.; Вклад автора 5%; JIF=64.8 WoS).

29. RNAget: an API to securely retrieve RNA quantifications [текст] / S. Upchurch [и др.] // Bioinformatics. — 2023. — апр. — т. 39, № 4. — btad126. — (0.23 п. л.; Работа в составе консорциума. Вклад автора менее 5%; JIF=5.8 WoS).

30. *Gilbert, W.* Why genes in pieces? [текст] / W. Gilbert // Nature. — 1978. — февр. — т. 271, № 5645. — с. 501.

31. *Will, C. L.* Spliceosome structure and function [текст] / C. L. Will, R. Lührmann // Cold Spring Harb Perspect Biol. — 2011. — июль. — т. 3, № 7.

32. *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.

33. *Yan, C.* Molecular Mechanisms of pre-mRNA Splicing through Structural Biology of the Spliceosome [текст] / C. Yan, R. Wan, Y. Shi // Cold Spring Harb Perspect Biol. — 2019. — янв. — т. 11, № 1.

34. *Shapiro, M. B.* RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression [текст] / M. B. Shapiro, P. Senapathy // Nucleic Acids Res. — 1987. — сент. — т. 15, № 17. — с. 7155—7174.

35. Mechanism of alternative splicing and its regulation [текст] / Y. Wang [и др.] // Biomed Rep. — 2015. — март. — т. 3, № 2. — с. 152—158.

36. *Nilsen, T. W.* Expansion of the eukaryotic proteome by alternative splicing [текст] / T. W. Nilsen, B. R. Graveley // Nature. — 2010. — янв. — т. 463, № 7280. — с. 457—463.

37. Alternative isoform regulation in human tissue transcriptomes [текст] / E. T. Wang [и др.] // Nature. — 2008. — нояб. — т. 456, № 7221. — с. 470—476.

38. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing [текст] / Q. Pan [и др.] // Nat Genet. — 2008. — дек. — т. 40, № 12. — с. 1413—1415.

39. *Grabowski, P. J.* Splicing regulation in neurons: tinkering with cell-specific control [текст] / P. J. Grabowski // Cell. — 1998. — март. — т. 92, № 6. — с. 709—712.

40. *Schwerk, C.* Regulation of apoptosis by alternative pre-mRNA splicing [текст] / C. Schwerk, K. Schulze-Osthoff // Mol Cell. — 2005. — июль. — т. 19, № 1. — с. 1—13.

41. *Baralle, F. E.* Alternative splicing as a regulator of development and tissue identity [текст] / F. E. Baralle, J. Giudice // Nat Rev Mol Cell Biol. — 2017. — июль. — т. 18, № 7. — с. 437—451.

42. *Ule, J.* Alternative Splicing Regulatory Networks: Functions, Mechanisms, and Evolution [текст] / J. Ule, B. J. Blencowe // Mol Cell. — 2019. — окт. — т. 76, № 2. — с. 329—345.

43. *Wahl, M. C.* The spliceosome: design principles of a dynamic RNP machine [текст] / M. C. Wahl, C. L. Will, R. Lührmann // Cell. — 2009. — февр. — т. 136, № 4. — с. 701—718.

44. *Wang, Z.* Splicing regulation: from a parts list of regulatory elements to an integrated splicing code [текст] / Z. Wang, C. B. Burge // RNA. — 2008. — май. — т. 14, № 5. — с. 802—813.

45. Deciphering the splicing code [текст] / Y. Barash [и др.] // Nature. — 2010. — май. — т. 465, № 7294. — с. 53—59.

46. *Gerstberger, S.* A census of human RNA-binding proteins [текст] / S. Gerstberger, M. Hafner, T. Tuschl // Nat Rev Genet. — 2014. — дек. — т. 15, № 12. — с. 829—845.

47. *Li, Q.* Neuronal regulation of alternative pre-mRNA splicing [текст] / Q. Li, J.-A. Lee, D. L. Black // Nat Rev Neurosci. — 2007. — нояб. — т. 8, № 11. — с. 819—831.

48. *Fu, X.-D.* Context-dependent control of alternative splicing by RNA-binding proteins [текст] / X.-D. Fu, M. Ares Jr // Nat Rev Genet. — 2014. — окт. — т. 15, № 10. — с. 689—701.

49. *Dvinge, H.* Regulation of alternative mRNA splicing: old players and new perspectives [текст] / H. Dvinge // FEBS Lett. — 2018. — сент. — т. 592, № 17. — с. 2987—3006.

50. *Zhou, Z.* Regulation of splicing by SR proteins and SR protein-specific kinases [текст] / Z. Zhou, X.-D. Fu // Chromosoma. — 2013. — июнь. — т. 122, № 3. — с. 191—207.

51. *Long, J. C.* The SR protein family of splicing factors: master regulators of gene expression [текст] / J. C. Long, J. F. Caceres // Biochem J. — 2009. — янв. — т. 417, № 1. — с. 15—27.

52. Genome-wide analysis reveals SR protein cooperation and competition in regulated splicing [текст] / S. Pandit [и др.] // Mol Cell. — 2013. — апр. — т. 50, № 2. — с. 223—235.

53. hnRNP proteins and splicing control [текст] / R. Martinez-Contreras [и др.] // Adv Exp Med Biol. — 2007. — т. 623. — с. 123—147.

54. Towards understanding pre-mRNA splicing mechanisms and the role of SR proteins [текст] / M. Sahebi [и др.] // Gene. — 2016. — авг. — т. 587, № 2. — с. 107—119.

55. *Liu, Y.* The roles of hnRNP A2/B1 in RNA biology and disease [текст] / Y. Liu, S.-L. Shi // Wiley Interdiscip Rev RNA. — 2021. — март. — т. 12, № 2. — e1612.

56. Genome-wide analysis of PTB-RNA interactions reveals a strategy used by the general splicing repressor to modulate exon inclusion or skipping [текст] / Y. Xue [и др.] // Mol Cell. — 2009. — дек. — т. 36, № 6. — с. 996—1006.

57. *Schaub, M. C.* Members of the heterogeneous nuclear ribonucleoprotein H family activate splicing of an HIV-1 splicing substrate by promoting formation of ATP-dependent spliceosomal complexes [текст] / M. C. Schaub, S. R. Lopez, M. Caputi // J Biol Chem. — 2007. — май. — т. 282, № 18. — с. 13617—13626.

58. *Caputi, M.* Determination of the RNA binding specificity of the heterogeneous nuclear ribonucleoprotein (hnRNP) H/H'/F/2H9 family [текст] / M. Caputi, A. M. Zahler // J Biol Chem. — 2001. — нояб. — т. 276, № 47. — с. 43850—43859.

59. Integrative modeling defines the Nova splicing-regulatory network and its combinatorial controls [текст] / C. Zhang [и др.] // Science. — 2010. — июль. — т. 329, № 5990. — с. 439—443.

60. Cooperative assembly of an hnRNP complex induced by a tissue-specific homolog of polypyrimidine tract binding protein [текст] / V. Markovtsov [и др.] // Mol Cell Biol. — 2000. — окт. — т. 20, № 20. — с. 7463—7479.

61. Essential roles for the splicing regulator nSR100/SRRM4 during nervous system development [текст] / M. Quesnel-Vallières [и др.] // Genes Dev. — 2015. — апр. — т. 29, № 7. — с. 746—759.

62. *Warf, M. B.* MBNL binds similar RNA structures in the CUG repeats of myotonic dystrophy and its pre-mRNA substrate cardiac troponin T [текст] / M. B. Warf, J. A. Berglund // RNA. — 2007. — дек. — т. 13, № 12. — с. 2238—2251.

63. Transcriptome-wide regulation of pre-mRNA splicing and mRNA localization by muscleblind proteins [текст] / E. T. Wang [и др.] // Cell. — 2012. — авг. — т. 150, № 4. — с. 710—724.

64. *Ladd, A. N.* The CELF family of RNA binding proteins is implicated in cell-specific and developmentally regulated alternative splicing [текст] / A. N. Ladd, N. Charlet, T. A. Cooper // Mol Cell Biol. — 2001. — февр. — т. 21, № 4. — с. 1285—1296.

65. Quaking and PTB control overlapping splicing regulatory networks during muscle cell differentiation [текст] / M. P. Hall [и др.] // RNA. — 2013. — май. — т. 19, № 5. — с. 627—638.

66. iCLIP predicts the dual splicing effects of TIA-RNA interactions [текст] / Z. Wang [и др.] // PLoS Biol. — 2010. — окт. — т. 8, № 10. — e1000530.

67. The pivotal roles of TIA proteins in 5' splice-site selection of alu exons and across evolution [текст] / N. Gal-Mark [и др.] // PLoS Genet. — 2009. — нояб. — т. 5, № 11. — e1000717.

68. ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism [текст] / S. Berto [и др.] // Hum Mol Genet. — 2016. — июнь. — т. 25, № 12. — с. 2451—2464.

69. RNA-binding proteins in neurological diseases [текст] / H. Zhou [и др.] // Sci China Life Sci. — 2014. — апр. — т. 57, № 4. — с. 432—444.

70. *Prashad, S.* RNA-binding proteins in neurological development and disease [текст] / S. Prashad, P. P. Gopal // RNA Biol. — 2021. — июль. — т. 18, № 7. — с. 972—987.

71. Functional coupling of RNAP II transcription to spliceosome assembly [текст] / R. Das [и др.] // Genes Dev. — 2006. — май. — т. 20, № 9. — с. 1100—1109.

72. 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.

73. *Bird, G.* RNA polymerase II carboxy-terminal domain phosphorylation is required for cotranscriptional pre-mRNA splicing and 3'-end formation [текст] / G. Bird, D. A. R. Zorio, D. L. Bentley // Mol Cell Biol. — 2004. — окт. — т. 24, № 20. — с. 8963—8969.

74. The C-terminal domain of the largest subunit of RNA polymerase II interacts with a novel set of serine/arginine-rich proteins [текст] / A. Yuryev [и др.] // Proc Natl Acad Sci U S A. — 1996. — июль. — т. 93, № 14. — с. 6975—6980.

75. *Misteli, T.* RNA polymerase II targets pre-mRNA splicing factors to transcription sites in vivo [текст] / T. Misteli, D. L. Spector // Mol Cell. — 1999. — июнь. — т. 3, № 6. — с. 697—705.

76. Transcription rate strongly affects splicing fidelity and cotranscriptionality in budding yeast [текст] / V. Aslanzadeh [и др.] // Genome Res. — 2018. — февр. — т. 28, № 2. — с. 203—213.

77. How slow RNA polymerase II elongation favors alternative exon skipping [текст] / G. Dujardin [и др.] // Mol Cell. — 2014. — май. — т. 54, № 4. — с. 683—690.

78. *Aitken, S.* Modelling reveals kinetic advantages of co-transcriptional splicing [текст] / S. Aitken, R. D. Alexander, J. D. Beggs // PLoS Comput Biol. — 2011. — окт. — т. 7, № 10. — e1002215.

79. Multiple competing RNA structures dynamically control alternative splicing in the human ATE1 gene [текст] / M. Kalinina [и др.] // Nucleic Acids Research. — 2021. — янв. — т. 49, № 1. — с. 479—490. — (1.39 п. л.; Вклад автора 40%; JIF=14.9 WoS).

80. Alternative RNA structures formed during transcription depend on elongation rate and modify RNA processing [текст] / T. Saldi [и др.] // Mol Cell. — 2021. — апр. — т. 81, № 8. — с. 1789—1801.

81. *Buratti, E.* Influence of RNA secondary structure on the pre-mRNA splicing process [текст] / E. Buratti, F. E. Baralle // Mol Cell Biol. — 2004. — дек. — т. 24, № 24. — с. 10505—10514.

82. Pre-mRNA secondary structures influence exon recognition [текст] / M. Hiller [и др.] // PLoS Genet. — 2007. — нояб. — т. 3, № 11. — e204.

83. Association of missense and 5'-splice-site mutations in tau with the inherited dementia FTDP-17 [текст] / M. Hutton [и др.] // Nature. — 1998. — июнь. — т. 393, № 6686. — с. 702—705.

84. Regulation of fibronectin EDA exon alternative splicing: possible role of RNA secondary structure for enhancer display [текст] / A. F. Muro [и др.] // Mol Cell Biol. — 1999. — апр. — т. 19, № 4. — с. 2657—2671.

85. *Damgaard, C. K.* hnRNP A1 controls HIV-1 mRNA splicing through cooperative binding to intron and exon splicing silencers in the context of a conserved secondary structure [текст] / C. K. Damgaard, T. O. Tange, J. Kjems // RNA. — 2002. — нояб. — т. 8, № 11. — с. 1401—1415.

86. *Rhodes, D.* G-quadruplexes and their regulatory roles in biology [текст] / D. Rhodes, H. J. Lipps // Nucleic Acids Res. — 2015. — окт. — т. 43, № 18. — с. 8627—8637.

87. RNA G-quadruplex secondary structure promotes alternative splicing via the RNA-binding protein hnRNPF [текст] / H. Huang [и др.] // Genes Dev. — 2017. — нояб. — т. 31, № 22. — с. 2296—2309.

88. Telomerase downregulation induced by the G-quadruplex ligand 12459 in A549 cells is mediated by hTERT RNA alternative splicing [текст] / D. Gomez [и др.] // Nucleic Acids Res. — 2004. — т. 32, № 1. — с. 371—379.

89. The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer [текст] / M.-C. Didiot [и др.] // Nucleic Acids Res. — 2008. — сент. — т. 36, № 15. — с. 4902—4912.

90. *Millevoi, S.* G-quadruplexes in RNA biology [текст] / S. Millevoi, H. Moine, S. Vagner // Wiley Interdiscip Rev RNA. — 2012. — т. 3, № 4. — с. 495—507.

91. G-quadruplex structures in TP53 intron 3: role in alternative splicing and in production of p53 mRNA isoforms [текст] / V. Marcel [и др.] // Carcinogenesis. — 2011. — март. — т. 32, № 3. — с. 271—278.

92. DeltaN-p53, a natural isoform of p53 lacking the first transactivation domain, counteracts growth suppression by wild-type p53 [текст] / S. Courtois [и др.] // Oncogene. — 2002. — окт. — т. 21, № 44. — с. 6722—6728.

93. *Wachter, A.* Riboswitch-mediated control of gene expression in eukaryotes [текст] / A. Wachter // RNA Biol. — 2010. — т. 7, № 1. — с. 67—76.

94. Long-range architecture in a viral RNA genome [текст] / E. J. Archer [и др.] // Biochemistry. — 2013. — май. — т. 52, № 18. — с. 3182—3190.

95. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing [текст] / S. Jacquenet [и др.] // Nucleic Acids Res. — 2001. — янв. — т. 29, № 2. — с. 464—478.

96. *Nicholson, B. L.* Functional long-range RNA-RNA interactions in positive-strand RNA viruses [текст] / B. L. Nicholson, K. A. White // Nat Rev Microbiol. — 2014. — июль. — т. 12, № 7. — с. 493—504.

97. *Miller, W. A.* Long-distance RNA-RNA interactions in plant virus gene expression and replication [текст] / W. A. Miller, K. A. White // Annu Rev Phytopathol. — 2006. — т. 44. — с. 447—467.

98. Circularization of an RNA template via long-range base pairing is critical for hepadnaviral reverse transcription [текст] / M.-K. Shin [и др.] // Virology. — 2008. — февр. — т. 371, № 2. — с. 362—373.

99. Conserved RNA secondary structures and long-range interactions in hepatitis C viruses [текст] / M. Fricke [и др.] // RNA. — 2015. — июль. — т. 21, № 7. — с. 1219—1232.

100. Overlapping local and long-range RNA-RNA interactions modulate dengue virus genome cyclization and replication [текст] / L. de Borba [и др.] // J Virol. — 2015. — март. — т. 89, № 6. — с. 3430—3437.

101. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation [текст] / J. G. Aw [и др.] // Mol Cell. — 2016. — май. — т. 62, № 4. — с. 603—617.

102. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure [текст] / Z. Lu [и др.] // Cell. — 2016. — май. — т. 165, № 5. — с. 1267—1279.

103. Global Mapping of Human RNA-RNA Interactions [текст] / E. Sharma [и др.] // Mol Cell. — 2016. — май. — т. 62, № 4. — с. 618—626.

104. COMRADES determines in vivo RNA structures and interactions [текст] / O. Ziv [и др.] // Nat Methods. — 2018. — окт. — т. 15, № 10. — с. 785—788.

105. A long-distance RNA-RNA interaction plays an important role in programmed -1 ribosomal frameshifting in the translation of p88 replicase protein of Red clover necrotic mosaic virus [текст] / Y. Tajima [и др.] // Virology. — 2011. — авг. — т. 417, № 1. — с. 169—178.

106. *Rüegsegger, U.* Block of HAC1 mRNA translation by long-range base pairing is released by cytoplasmic splicing upon induction of the unfolded protein response [текст] / U. Rüegsegger, J. H. Leber, P. Walter // Cell. — 2001. — окт. — т. 107, № 1. — с. 103—114.

107. Modulation of alternative splicing by long-range RNA structures in Drosophila [текст] / V. A. Raker [и др.] // Nucleic Acids Research. — 2009. — авг. — т. 37, № 14. — с. 4533—4544. — (1.39 п. л.; Вклад автора 75%; JIF=14.9 WoS).

108. 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. — (1.73 п. л.; Вклад автора 75%; JIF=3.9 WoS).

109. RNA secondary structure in mutually exclusive splicing [текст] / Y. Yang [и др.] // Nat Struct Mol Biol. — 2011. — февр. — т. 18, № 2. — с. 159—168.

110. A large-scale binding and functional map of human RNA-binding proteins [текст] / E. L. Van Nostrand [и др.] // Nature. — 2020. — июль. — т. 583, № 7818. — с. 711—719.

111. High-affinity hnRNP A1 binding sites and duplex-forming inverted repeats have similar effects on 5' splice site selection in support of a common looping out and repression mechanism [текст] / F.-U. H. Nasim [и др.] // RNA. — 2002. — авг. — т. 8, № 8. — с. 1078—1089.

112. *Miriami, E.* Conserved sequence elements associated with exon skipping [текст] / E. Miriami, H. Margalit, R. Sperling // Nucleic Acids Res. — 2003. — апр. — т. 31, № 7. — с. 1974—1983.

113. *Pervouchine, D. D.* IRBIS: a systematic search for conserved complementarity [текст] / D. D. Pervouchine // RNA. — 2014. — окт. — т. 20, № 10. — с. 1519—1531. — (1.50 п. л.; Вклад автора 100%; JIF=3.9 WoS).

114. *Wong, M. S.* Regulation of human telomerase splicing by RNA:RNA pairing [текст] / M. S. Wong, J. W. Shay, W. E. Wright // Nat Commun. — 2014. — февр. — т. 5. — с. 3306.

115. *Bernat, V.* RNA Structures as Mediators of Neurological Diseases and as Drug Targets [текст] / V. Bernat, M. D. Disney // Neuron. — 2015. — июль. — т. 87, № 1. — с. 28—46.

116. An intronic structure enabled by a long-distance interaction serves as a novel target for splicing correction in spinal muscular atrophy [текст] / N. N. Singh [и др.] // Nucleic Acids Res. — 2013. — сент. — т. 41, № 17. — с. 8144—8165.

117. *Singh, N. N.* Splicing regulation in spinal muscular atrophy by an RNA structure formed by long-distance interactions [текст] / N. N. Singh, B. M. Lee, R. N. Singh // Ann N Y Acad Sci. — 2015. — апр. — т. 1341. — с. 176—187.

118. PMD patient mutations reveal a long-distance intronic interaction that regulates PLP1/DM20 alternative splicing [текст] / J. R. Taube [и др.] // Hum Mol Genet. — 2014. — окт. — т. 23, № 20. — с. 5464—5478.

119. An RNA architectural locus control region involved in Dscam mutually exclusive splicing [текст] / X. Wang [и др.] // Nat Commun. — 2012. — т. 3. — с. 1255.

120. 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.

121. Mutually exclusive alternative splicing of pre-mRNAs [текст] / Y. Jin [и др.] // Wiley Interdiscip Rev RNA. — 2018. — май. — т. 9, № 3. — e1468.

122. Conservation and regulation of alternative splicing by dynamic inter- and intra-intron base pairings in Lepidoptera 14-3-3ξ pre-mRNAs [текст] / Y. Yang [и др.] // RNA Biol. — 2012. — май. — т. 9, № 5. — с. 691—700.

123. *Suyama, M.* Mechanistic insights into mutually exclusive splicing in dynamin 1 [текст] / M. Suyama // Bioinformatics. — 2013. — сент. — т. 29, № 17. — с. 2084—2087.

124. The landscape of human mutually exclusive splicing [текст] / K. Hatje [и др.] // Mol Syst Biol. — 2017. — дек. — т. 13, № 12. — с. 959.

125. *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. — 2018. — июль. — т. 9, № 7. — с. 356. — (1.50 п. л.; Вклад автора 75%; JIF=3.5 WoS).

126. *Welden, J. R.* Pre-mRNA structures forming circular RNAs [текст] / J. R. Welden, S. Stamm // Biochim Biophys Acta Gene Regul Mech. — 2019. — т. 1862, № 11/12. — с. 194410.

127. *Pervouchine, D. D.* Circular exonic RNAs: When RNA structure meets topology [текст] / D. D. Pervouchine // Biochimica et Biophysica Acta Gene Regulatory Mechanisms. — 2019. — т. 1862, № 11/12. — с. 194384. — (1.04 п. л.; Вклад автора 100%; JIF=4.7 WoS).

128. Effects of RNA secondary structure on alternative splicing of pre-mRNA: is folding limited to a region behind the transcribing RNA polymerase? [текст] / L. P. Eperon [и др.] // Cell. — 1988. — июль. — т. 54, № 3. — с. 393—401.

129. *Lunde, B. M.* RNA-binding proteins: modular design for efficient function [текст] / B. M. Lunde, C. Moore, G. Varani // Nat Rev Mol Cell Biol. — 2007. — июнь. — т. 8, № 6. — с. 479—490.

130. The tetranucleotide UCAY directs the specific recognition of RNA by the Nova K-homology 3 domain [текст] / K. B. Jensen [и др.] // Proc Natl Acad Sci U S A. — 2000. — май. — т. 97, № 11. — с. 5740—5745.

131. Sequence, Structure, and Context Preferences of Human RNA Binding Proteins [текст] / D. Dominguez [и др.] // Mol Cell. — 2018. — июнь. — т. 70, № 5. — с. 854—867.

132. *Witten, J. T.* Understanding splicing regulation through RNA splicing maps [текст] / J. T. Witten, J. Ule // Trends Genet. — 2011. — март. — т. 27, № 3. — с. 89—97.

133. RNA sequence context effects measured in vitro predict in vivo protein binding and regulation [текст] / J. M. Taliaferro [и др.] // Mol Cell. — 2016. — окт. — т. 64, № 2. — с. 294—306.

134. *Cusack, S.* RNA-protein complexes [текст] / S. Cusack // Curr Opin Struct Biol. — 1999. — февр. — т. 9, № 1. — с. 66—73.

135. Sequence-specific RNA binding by a Nova KH domain: implications for paraneoplastic disease and the fragile X syndrome [текст] / H. A. Lewis [и др.] // Cell. — 2000. — февр. — т. 100, № 3. — с. 323—332.

136. Protein-RNA and protein-protein recognition by dual KH1/2 domains of the neuronal splicing factor Nova-1 [текст] / M. Teplova [и др.] // Structure. — 2011. — июль. — т. 19, № 7. — с. 930—944.

137. Structure of a construct of a human poly(C)-binding protein containing the first and second KH domains reveals insights into its regulatory mechanisms [текст] / Z. Du [и др.] // J Biol Chem. — 2008. — окт. — т. 283, № 42. — с. 28757—28766.

138. Cooperation and competition by RNA-binding proteins in cancer [текст] / S. Nag [и др.] // Semin Cancer Biol. — 2022. — нояб. — т. 86, Pt 3. — с. 286—297.

139. *Schorr, A. L.* miRNA-Based Regulation of Alternative RNA Splicing in Metazoans [текст] / A. L. Schorr, M. Mangone // Int J Mol Sci. — 2021. — окт. — т. 22, № 21.

140. RNA editing in nascent RNA affects pre-mRNA splicing [текст] / Y.-H. E. Hsiao [и др.] // Genome Res. — 2018. — июнь. — т. 28, № 6. — с. 812—823.

141. Global regulation of alternative splicing by adenosine deaminase acting on RNA (ADAR) [текст] / O. Solomon [и др.] // RNA. — 2013. — май. — т. 19, № 5. — с. 591—604.

142. *Rueter, S. M.* Regulation of alternative splicing by RNA editing [текст] / S. M. Rueter, T. R. Dawson, R. B. Emeson // Nature. — 1999. — май. — т. 399, № 6731. — с. 75—80.

143. RNA-editing-mediated exon evolution [текст] / G. Lev-Maor [и др.] // Genome Biol. — 2007. — т. 8, № 2. — R29.

144. *Mazloomian, A.* Genome-wide identification and characterization of tissue-specific RNA editing events in D. melanogaster and their potential role in regulating alternative splicing [текст] / A. Mazloomian, I. M. Meyer // RNA Biol. — 2015. — т. 12, № 12. — с. 1391—1401.

145. Nuclear m(6)A Reader YTHDC1 Regulates mRNA Splicing [текст] / W. Xiao [и др.] // Mol Cell. — 2016. — февр. — т. 61, № 4. — с. 507—519.

146. N(6)-methyladenosine-dependent RNA structural switches regulate RNA-protein interactions [текст] / N. Liu [и др.] // Nature. — 2015. — февр. — т. 518, № 7540. — с. 560—564.

147. The protein factors MBNL1 and U2AF65 bind alternative RNA structures to regulate splicing [текст] / M. B. Warf [и др.] // Proc Natl Acad Sci U S A. — 2009. — июнь. — т. 106, № 23. — с. 9203—9208.

148. A network of DZF proteins controls alternative splicing regulation and fidelity [текст] / N. Haque [и др.] // Nucleic Acids Res. — 2023. — июль. — т. 51, № 12. — с. 6411—6429.

149. Intronic binding sites for hnRNP A/B and hnRNP F/H proteins stimulate pre-mRNA splicing [текст] / R. Martinez-Contreras [и др.] // PLoS Biol. — 2006. — февр. — т. 4, № 2. — e21.

150. hnRNP A1 and hnRNP H can collaborate to modulate 5' splice site selection [текст] / J.-F. Fisette [и др.] // RNA. — 2010. — янв. — т. 16, № 1. — с. 228—238.

151. An RNA map predicting Nova-dependent splicing regulation [текст] / J. Ule [и др.] // Nature. — 2006. — нояб. — т. 444, № 7119. — с. 580—586.

152. *Borbolis, F.* Cytoplasmic mRNA turnover and ageing [текст] / F. Borbolis, P. Syntichaki // Mech Ageing Dev. — 2015. — дек. — т. 152. — с. 32—42.

153. *Dassi, E.* Handshakes and Fights: The Regulatory Interplay of RNA-Binding Proteins [текст] / E. Dassi // Front Mol Biosci. — 2017. — т. 4. — с. 67.

154. *Lykke-Andersen, S.* Nonsense-mediated mRNA decay: an intricate machinery that shapes transcriptomes [текст] / S. Lykke-Andersen, T. H. Jensen // Nat Rev Mol Cell Biol. — 2015. — нояб. — т. 16, № 11. — с. 665—677.

155. Stabilization and ribosome association of unspliced pre-mRNAs in a yeast upf1- mutant [текст] / F. He [и др.] // Proc Natl Acad Sci U S A. — 1993. — авг. — т. 90, № 15. — с. 7034—7038.

156. The exon-exon junction complex provides a binding platform for factors involved in mRNA export and nonsense-mediated mRNA decay [текст] / H. Le Hir [и др.] // EMBO J. — 2001. — сент. — т. 20, № 17. — с. 4987—4997.

157. The spliceosome deposits multiple proteins 20-24 nucleotides upstream of mRNA exon-exon junctions [текст] / H. Le Hir [и др.] // EMBO J. — 2000. — дек. — т. 19, № 24. — с. 6860—6869.

158. *Nagy, E.* A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance [текст] / E. Nagy, L. E. Maquat // Trends Biochem Sci. — 1998. — июнь. — т. 23, № 6. — с. 198—199.

159. *Karousis, E. D.* Nonsense-mediated mRNA decay: novel mechanistic insights and biological impact [текст] / E. D. Karousis, S. Nasif, O. Mühlemann // Wiley Interdiscip Rev RNA. — 2016. — сент. — т. 7, № 5. — с. 661—682.

160. *Popp, M. W.* Leveraging Rules of Nonsense-Mediated mRNA Decay for Genome Engineering and Personalized Medicine [текст] / M. W. Popp, L. E. Maquat // Cell. — 2016. — июнь. — т. 165, № 6. — с. 1319—1322.

161. *Kurosaki, T.* Publisher Correction: Quality and quantity control of gene expression by nonsense-mediated mRNA decay [текст] / T. Kurosaki, M. W. Popp, L. E. Maquat // Nat Rev Mol Cell Biol. — 2019. — июнь. — т. 20, № 6. — с. 384.

162. *Isken, O.* Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function [текст] / O. Isken, L. E. Maquat // Genes Dev. — 2007. — авг. — т. 21, № 15. — с. 1833—1856.

163. *Loh, B.* The SMG5-SMG7 heterodimer directly recruits the CCR4-NOT deadenylase complex to mRNAs containing nonsense codons via interaction with POP2 [текст] / B. Loh, S. Jonas, E. Izaurralde // Genes Dev. — 2013. — окт. — т. 27, № 19. — с. 2125—2138.

164. *Unterholzner, L.* SMG7 acts as a molecular link between mRNA surveillance and mRNA decay [текст] / L. Unterholzner, E. Izaurralde // Mol Cell. — 2004. — нояб. — т. 16, № 4. — с. 587—596.

165. *Kurosaki, T.* Rules that govern UPF1 binding to mRNA 3' UTRs [текст] / T. Kurosaki, L. E. Maquat // Proc Natl Acad Sci U S A. — 2013. — февр. — т. 110, № 9. — с. 3357—3362.

166. *Hogg, J. R.* Upf1 senses 3'UTR length to potentiate mRNA decay [текст] / J. R. Hogg, S. P. Goff // Cell. — 2010. — окт. — т. 143, № 3. — с. 379—389.

167. *Singh, G.* A competition between stimulators and antagonists of Upf complex recruitment governs human nonsense-mediated mRNA decay [текст] / G. Singh, I. Rebbapragada, J. Lykke-Andersen // PLoS Biol. — 2008. — апр. — т. 6, № 4. — e111.

168. *Lopez, P. J.* Genomic-scale quantitative analysis of yeast pre-mRNA splicing: implications for splice-site recognition [текст] / P. J. Lopez, B. Séraphin // RNA. — 1999. — сент. — т. 5, № 9. — с. 1135—1137.

169. Quality control of transcription start site selection by nonsense-mediated-mRNA decay [текст] / C. Malabat [и др.] // Elife. — 2015. — апр. — т. 4.

170. *Maquat, L. E.* Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics [текст] / L. E. Maquat // Nat Rev Mol Cell Biol. — 2004. — февр. — т. 5, № 2. — с. 89—99.

171. *Lewis, B. P.* Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans [текст] / B. P. Lewis, R. E. Green, S. E. Brenner // Proc Natl Acad Sci U S A. — 2003. — янв. — т. 100, № 1. — с. 189—192.

172. The coupling of alternative splicing and nonsense-mediated mRNA decay [текст] / L. F. Lareau [и др.] // Adv Exp Med Biol. — 2007. — т. 623. — с. 190—211.

173. Integrative transcriptomic analysis suggests new autoregulatory splicing events coupled with nonsense-mediated mRNA decay [текст] / D. Pervouchine [и др.] // Nucleic Acids Research. — 2019. — июнь. — т. 47, № 10. — с. 5293—5306. — (1.62 п. л.; Вклад автора 75%; JIF=14.9 WoS).

174. *Nasif, S.* Beyond quality control: The role of nonsense-mediated mRNA decay (NMD) in regulating gene expression [текст] / S. Nasif, L. Contu, O. Mühlemann // Semin Cell Dev Biol. — 2018. — март. — т. 75. — с. 78—87.

175. *Lareau, L. F.* Regulation of splicing factors by alternative splicing and NMD is conserved between kingdoms yet evolutionarily flexible [текст] / L. F. Lareau, S. E. Brenner // Mol Biol Evol. — 2015. — апр. — т. 32, № 4. — с. 1072—1079.

176. *García-Moreno, J. F.* Perspective in Alternative Splicing Coupled to Nonsense-Mediated mRNA Decay [текст] / J. F. García-Moreno, L. Romão // Int J Mol Sci. — 2020. — дек. — т. 21, № 24.

177. Alternative splicing and nonsense-mediated decay modulate expression of important regulatory genes in Arabidopsis [текст] / M. Kalyna [и др.] // Nucleic Acids Res. — 2012. — март. — т. 40, № 6. — с. 2454—2469.

178. Unproductive splicing of SR genes associated with highly conserved and ultraconserved DNA elements [текст] / L. F. Lareau [и др.] // Nature. — 2007. — апр. — т. 446, № 7138. — с. 926—929.

179. *Carvill, G. L.* Poison exons in neurodevelopment and disease [текст] / G. L. Carvill, H. C. Mefford // Curr Opin Genet Dev. — 2020. — дек. — т. 65. — с. 98—102.

180. *Leclair, N. K.* 'Poisoning' of the transcriptome by ultraconserved elements [текст] / N. K. Leclair, O. Anczuków // Nat Rev Mol Cell Biol. — 2022. — дек. — т. 23, № 12. — с. 777.

181. Regulation of AUF1 expression via conserved alternatively spliced elements in the 3' untranslated region [текст] / G. M. Wilson [и др.] // Mol Cell Biol. — 1999. — июнь. — т. 19, № 6. — с. 4056—4064.

182. Autoregulation of the nonsense-mediated mRNA decay pathway in human cells [текст] / H. Yepiskoposyan [и др.] // RNA. — 2011. — дек. — т. 17, № 12. — с. 2108—2118.

183. RNA homeostasis governed by cell type-specific and branched feedback loops acting on NMD [текст] / L. Huang [и др.] // Mol Cell. — 2011. — сент. — т. 43, № 6. — с. 950—961.

184. Pan-cancer pervasive upregulation of 3' UTR splicing drives tumourigenesis [текст] / J. J. Chan [и др.] // Nat Cell Biol. — 2022. — июнь. — т. 24, № 6. — с. 928—939.

185. RBM47 inhibits hepatocellular carcinoma progression by targeting UPF1 as a DNA/RNA regulator [текст] / T. Guo [и др.] // Cell Death Discov. — 2022. — июль. — т. 8, № 1. — с. 320.

186. Transcriptome-wide identification of NMD-targeted human mRNAs reveals extensive redundancy between SMG6- and SMG7-mediated degradation pathways [текст] / M. Colombo [и др.] // RNA. — 2017. — февр. — т. 23, № 2. — с. 189—201.

187. Identification of hundreds of novel UPF1 target transcripts by direct determination of whole transcriptome stability [текст] / H. Tani [и др.] // RNA Biol. — 2012. — нояб. — т. 9, № 11. — с. 1370—1379.

188. Deep sequencing of pre-translational mRNPs reveals hidden flux through evolutionarily conserved alternative splicing nonsense-mediated decay pathways [текст] / C. Kovalak [и др.] // Genome Biol. — 2021. — май. — т. 22, № 1. — с. 132.

189. Autoregulation of RBM10 and cross-regulation of RBM10/RBM5 via alternative splicing-coupled nonsense-mediated decay [текст] / Y. Sun [и др.] // Nucleic Acids Res. — 2017. — авг. — т. 45, № 14. — с. 8524—8540.

190. Global analysis reveals SRp20- and SRp75-specific mRNPs in cycling and neural cells [текст] / M.-L. Ankö [и др.] // Nat Struct Mol Biol. — 2010. — авг. — т. 17, № 8. — с. 962—970.

191. SC35 autoregulates its expression by promoting splicing events that destabilize its mRNAs [текст] / A. Sureau [и др.] // EMBO J. — 2001. — апр. — т. 20, № 7. — с. 1785—1796.

192. SF2/ASF autoregulation involves multiple layers of post-transcriptional and translational control [текст] / S. Sun [и др.] // Nat Struct Mol Biol. — 2010. — март. — т. 17, № 3. — с. 306—312.

193. Human tra2-beta1 autoregulates its protein concentration by influencing alternative splicing of its pre-mRNA [текст] / P. Stoilov [и др.] // Hum Mol Genet. — 2004. — март. — т. 13, № 5. — с. 509—524.

194. *Hillman, R. T.* An unappreciated role for RNA surveillance [текст] / R. T. Hillman, R. E. Green, S. E. Brenner // Genome Biol. — 2004. — т. 5, № 2. — R8.

195. In vivo regulation of alternative pre-mRNA splicing by the Clk1 protein kinase [текст] / P. I. Duncan [и др.] // Mol Cell Biol. — 1997. — окт. — т. 17, № 10. — с. 5996—6001.

196. *Le Guiner, C.* TIA-1 or TIAR is required for DT40 cell viability [текст] / C. Le Guiner, M.-C. Gesnel, R. Breathnach // J Biol Chem. — 2003. — март. — т. 278, № 12. — с. 10465—10476.

197. Autoregulation of polypyrimidine tract binding protein by alternative splicing leading to nonsense-mediated decay [текст] / M. C. Wollerton [и др.] // Mol Cell. — 2004. — янв. — т. 13, № 1. — с. 91—100.

198. Alternative splicing of brain-specific PTB defines a tissue-specific isoform pattern that predicts distinct functional roles [текст] / L. Rahman [и др.] // Genomics. — 2002. — сент. — т. 80, № 3. — с. 245—249.

199. *Kemmerer, K.* Auto- and cross-regulation of the hnRNPs D and DL [текст] / K. Kemmerer, S. Fischer, J. E. Weigand // RNA. — 2018. — март. — т. 24, № 3. — с. 324—331.

200. Evolutionarily conserved autoregulation of alternative pre-mRNA splicing by ribosomal protein L10a [текст] / S. Takei [и др.] // Nucleic Acids Res. — 2016. — июль. — т. 44, № 12. — с. 5585—5596.

201. Alternative splicing and nonsense-mediated mRNA decay regulate mammalian ribosomal gene expression [текст] / M. Cuccurese [и др.] // Nucleic Acids Res. — 2005. — т. 33, № 18. — с. 5965—5977.

202. Poison Exon Splicing Regulates a Coordinated Network of SR Protein Expression during Differentiation and Tumorigenesis [текст] / N. K. Leclair [и др.] // Mol Cell. — 2020. — нояб. — т. 80, № 4. — с. 648—665.

203. The RNA-binding landscapes of two SR proteins reveal unique functions and binding to diverse RNA classes [текст] / M.-L. Änkö [и др.] // Genome Biol. — 2012. — т. 13, № 3. — R17.

204. *Spellman, R.* Crossregulation and functional redundancy between the splicing regulator PTB and its paralogs nPTB and ROD1 [текст] / R. Spellman, M. Llorian, C. W. J. Smith // Mol Cell. — 2007. — авг. — т. 27, № 3. — с. 420—434.

205. Rbfox2 controls autoregulation in RNA-binding protein networks [текст] / M. Jangi [и др.] // Genes Dev. — 2014. — март. — т. 28, № 6. — с. 637—651.

206. Auto- and cross-regulation of the hnRNP L proteins by alternative splicing [текст] / O. Rossbach [и др.] // Mol Cell Biol. — 2009. — март. — т. 29, № 6. — с. 1442—1451.

207. Regulation of the MID1 protein function is fine-tuned by a complex pattern of alternative splicing [текст] / J. Winter [и др.] // Hum Genet. — 2004. — май. — т. 114, № 6. — с. 541—552.

208. Smg1 is required for embryogenesis and regulates diverse genes via alternative splicing coupled to nonsense-mediated mRNA decay [текст] / D. R. McIlwain [и др.] // Proc Natl Acad Sci U S A. — 2010. — июль. — т. 107, № 27. — с. 12186—12191.

209. Orchestrated intron retention regulates normal granulocyte differentiation [текст] / J. J.-L. Wong [и др.] // Cell. — 2013. — авг. — т. 154, № 3. — с. 583—595.

210. The unfolded protein response is shaped by the NMD pathway [текст] / R. Karam [и др.] // EMBO Rep. — 2015. — май. — т. 16, № 5. — с. 599—609.

211. Inhibition of nonsense-mediated RNA decay by ER stress [текст] / Z. Li [и др.] // RNA. — 2017. — март. — т. 23, № 3. — с. 378—394.

212. *Gardner, L. B.* Hypoxic inhibition of nonsense-mediated RNA decay regulates gene expression and the integrated stress response [текст] / L. B. Gardner // Mol Cell Biol. — 2008. — июнь. — т. 28, № 11. — с. 3729—3741.

213. ALS-associated mutation FUS-R521C causes DNA damage and RNA splicing defects [текст] / H. Qiu [и др.] // J Clin Invest. — 2014. — март. — т. 124, № 3. — с. 981—999.

214. Long pre-mRNA depletion and RNA missplicing contribute to neuronal vulnerability from loss of TDP-43 [текст] / M. Polymenidou [и др.] // Nat Neurosci. — 2011. — апр. — т. 14, № 4. — с. 459—468.

215. *Hamid, F. M.* A mechanism underlying position-specific regulation of alternative splicing [текст] / F. M. Hamid, E. V. Makeyev // Nucleic Acids Res. — 2017. — дек. — т. 45, № 21. — с. 12455—12468.

216. Tissue-specific regulation of gene expression via unproductive splicing [текст] / A. Mironov [и др.] // Nucleic Acids Research. — 2023. — апр. — т. 51, № 7. — с. 3055—3066. — (1.39 п. л.; Вклад автора 40%; JIF=14.9 WoS).

217. Auto-regulatory feedback by RNA-binding proteins [текст] / M. Müller-McNicoll [и др.] // J Mol Cell Biol. — 2019. — окт. — т. 11, № 10. — с. 930—939.

218. *Moschall, R.* Promiscuity in post-transcriptional control of gene expression: Drosophila sex-lethal and its regulatory partnerships [текст] / R. Moschall, M. Gaik, J. Medenbach // FEBS Lett. — 2017. — июнь. — т. 591, № 11. — с. 1471—1488.

219. Ultraconserved elements are associated with homeostatic control of splicing regulators by alternative splicing and nonsense-mediated decay [текст] / J. Z. Ni [и др.] // Genes Dev. — 2007. — март. — т. 21, № 6. — с. 708—718.

220. *Graveley, B. R.* Alternative splicing: increasing diversity in the proteomic world [текст] / B. R. Graveley // Trends Genet. — 2001. — февр. — т. 17, № 2. — с. 100—107.

221. 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.

222. *Tress, M. L.* Most Alternative Isoforms Are Not Functionally Important [текст] / M. L. Tress, F. Abascal, A. Valencia // Trends Biochem Sci. — 2017. — июнь. — т. 42, № 6. — с. 408—410.

223. *Tress, M. L.* Alternative Splicing May Not Be the Key to Proteome Complexity [текст] / M. L. Tress, F. Abascal, A. Valencia // Trends Biochem Sci. — 2017. — февр. — т. 42, № 2. — с. 98—110.

224. Re-evaluating the impact of alternative RNA splicing on proteomic diversity [текст] / J. M. Manuel [и др.] // Front Genet. — 2023. — т. 14. — с. 1089053.

225. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene [текст] / M. Gonzàlez-Porta [и др.] // Genome Biol. — 2013. — июль. — т. 14, № 7. — R70.

226. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues [текст] / J. Merkin [и др.] // Science. — 2012. — дек. — т. 338, № 6114. — с. 1593—1599.

227. Drift and conservation of differential exon usage across tissues in primate species [текст] / A. Reyes [и др.] // Proc Natl Acad Sci U S A. — 2013. — сент. — т. 110, № 38. — с. 15377—15382.

228. Domain-based prediction of the human isoform interactome provides insights into the functional impact of alternative splicing [текст] / M. A. Ghadie [и др.] // PLoS Comput Biol. — 2017. — авг. — т. 13, № 8. — e1005717.

229. The evolutionary landscape of alternative splicing in vertebrate species [текст] / N. L. Barbosa-Morais [и др.] // Science. — 2012. — дек. — т. 338, № 6114. — с. 1587—1593.

230. Alternative RNA processing in calcitonin gene expression generates mRNAs encoding different polypeptide products [текст] / S. G. Amara [и др.] // Nature. — 1982. — июль. — т. 298, № 5871. — с. 240—244.

231. *Leff, S. E.* Splice commitment dictates neuron-specific alternative RNA processing in calcitonin/CGRP gene expression [текст] / S. E. Leff, R. M. Evans, M. G. Rosenfeld // Cell. — 1987. — февр. — т. 48, № 3. — с. 517—524.

232. 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.

233. MBNL proteins repress ES-cell-specific alternative splicing and reprogramming [текст] / H. Han [и др.] // Nature. — 2013. — июнь. — т. 498, № 7453. — с. 241—245.

234. *Marasco, L. E.* The physiology of alternative splicing [текст] / L. E. Marasco, A. R. Kornblihtt // Nat Rev Mol Cell Biol. — 2023. — апр. — т. 24, № 4. — с. 242—254.

235. Function of alternative splicing [текст] / O. Kelemen [и др.] // Gene. — 2013. — февр. — т. 514, № 1. — с. 1—30.

236. *Scotti, M. M.* RNA mis-splicing in disease [текст] / M. M. Scotti, M. S. Swanson // Nat Rev Genet. — 2016. — янв. — т. 17, № 1. — с. 19—32.

237. *Jiang, W.* Alternative splicing: Human disease and quantitative analysis from high-throughput sequencing [текст] / W. Jiang, L. Chen // Comput Struct Biotechnol J. — 2021. — т. 19. — с. 183—195.

238. *Tazi, J.* Alternative splicing and disease [текст] / J. Tazi, N. Bakkour, S. Stamm // Biochim Biophys Acta. — 2009. — янв. — т. 1792, № 1. — с. 14—26.

239. *Nikom, D.* Alternative splicing in neurodegenerative disease and the promise of RNA therapies [текст] / D. Nikom, S. Zheng // Nat Rev Neurosci. — 2023. — авг. — т. 24, № 8. — с. 457—473.

240. Genetic basis of variable exon 9 skipping in cystic fibrosis transmembrane conductance regulator mRNA [текст] / C. S. Chu [и др.] // Nat Genet. — 1993. — февр. — т. 3, № 2. — с. 151—156.

241. *Singh, N. N.* Alternative splicing in spinal muscular atrophy underscores the role of an intron definition model [текст] / N. N. Singh, R. N. Singh // RNA Biol. — 2011. — т. 8, № 4. — с. 600—606.

242. *Liu, F.* Tau exon 10 alternative splicing and tauopathies [текст] / F. Liu, C.-X. Gong // Mol Neurodegener. — 2008. — июль. — т. 3. — с. 8.

243. Transcriptome sequencing reveals aberrant alternative splicing in Huntington's disease [текст] / L. Lin [и др.] // Hum Mol Genet. — 2016. — авг. — т. 25, № 16. — с. 3454—3466.

244. A highly conserved program of neuronal microexons is misregulated in autistic brains [текст] / M. Irimia [и др.] // Cell. — 2014. — дек. — т. 159, № 7. — с. 1511—1523.

245. RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts [текст] / Y. I. Li [и др.] // Genome Res. — 2015. — янв. — т. 25, № 1. — с. 1—13.

246. *Chabot, B.* Defective control of pre-messenger RNA splicing in human disease [текст] / B. Chabot, L. Shkreta // J Cell Biol. — 2016. — янв. — т. 212, № 1. — с. 13—27.

247. A p120 catenin isoform switch affects Rho activity, induces tumor cell invasion, and predicts metastatic disease [текст] / M. Yanagisawa [и др.] // J Biol Chem. — 2008. — июнь. — т. 283, № 26. — с. 18344—18354.

248. The gene encoding the splicing factor SF2/ASF is a proto-oncogene [текст] / R. Karni [и др.] // Nat Struct Mol Biol. — 2007. — март. — т. 14, № 3. — с. 185—193.

249. RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF(V600E) [текст] / P. I. Poulikakos [и др.] // Nature. — 2011. — нояб. — т. 480, № 7377. — с. 387—390.

250. Intron retention is a widespread mechanism of tumor-suppressor inactivation [текст] / H. Jung [и др.] // Nat Genet. — 2015. — нояб. — т. 47, № 11. — с. 1242—1248.

251. Synonymous mutations frequently act as driver mutations in human cancers [текст] / F. Supek [и др.] // Cell. — 2014. — март. — т. 156, № 6. — с. 1324—1335.

252. Recurrent de novo point mutations in lamin A cause Hutchinson-Gilford progeria syndrome [текст] / M. Eriksson [и др.] // Nature. — 2003. — май. — т. 423, № 6937. — с. 293—298.

253. Electrophysiological and histopathological characteristics of progressive atrioventricular block accompanied by familial dilated cardiomyopathy caused by a novel mutation of lamin A/C gene [текст] / J. Otomo [и др.] // J Cardiovasc Electrophysiol. — 2005. — февр. — т. 16, № 2. — с. 137—145.

254. The novel MAPT mutation K298E: mechanisms of mutant tau toxicity, brain pathology and tau expression in induced fibroblast-derived neurons [текст] / M. Iovino [и др.] // Acta Neuropathol. — 2014. — февр. — т. 127, № 2. — с. 283—295.

255. Altered splicing of ATP6AP2 causes X-linked parkinsonism with spasticity (XPDS) [текст] / O. Korvatska [и др.] // Hum Mol Genet. — 2013. — авг. — т. 22, № 16. — с. 3259—3268.

256. RBM5, 6, and 10 differentially regulate NUMB alternative splicing to control cancer cell proliferation [текст] / E. G. Bechara [и др.] // Mol Cell. — 2013. — дек. — т. 52, № 5. — с. 720—733.

257. Cancer-Associated SF3B1 Hotspot Mutations Induce Cryptic 3' Splice Site Selection through Use of a Different Branch Point [текст] / R. B. Darman [и др.] // Cell Rep. — 2015. — нояб. — т. 13, № 5. — с. 1033—1045.

258. Aberrant splicing of U12-type introns is the hallmark of ZRSR2 mutant myelodysplastic syndrome [текст] / V. Madan [и др.] // Nat Commun. — 2015. — янв. — т. 6. — с. 6042.

259. The RNA-binding protein QKI suppresses cancer-associated aberrant splicing [текст] / F.-Y. Zong [и др.] // PLoS Genet. — 2014. — апр. — т. 10, № 4. — e1004289.

260. Exon-skipping antisense oligonucleotides for cystic fibrosis therapy [текст] / Y. J. Kim [и др.] // Proc Natl Acad Sci U S A. — 2022. — янв. — т. 119, № 3.

261. *Gong, Q.* Inhibition of nonsense-mediated mRNA decay by antisense morpholino oligonucleotides restores functional expression of hERG nonsense and frameshift mutations in long-QT syndrome [текст] / Q. Gong, M. R. Stump, Z. Zhou // J Mol Cell Cardiol. — 2011. — янв. — т. 50, № 1. — с. 223—229.

262. Eteplirsen for the treatment of Duchenne muscular dystrophy [текст] / J. R. Mendell [и др.] // Ann Neurol. — 2013. — нояб. — т. 74, № 5. — с. 637—647.

263. Open-Label Evaluation of Eteplirsen in Patients with Duchenne Muscular Dystrophy Amenable to Exon 51 Skipping: PROMOVI Trial [текст] / C. M. McDonald [и др.] // J Neuromuscul Dis. — 2021. — т. 8, № 6. — с. 989—1001.

264. *Aartsma-Rus, A.* The 10th Oligonucleotide Therapy Approved: Golodirsen for Duchenne Muscular Dystrophy [текст] / A. Aartsma-Rus, D. R. Corey // Nucleic Acid Ther. — 2020. — апр. — т. 30, № 2. — с. 67—70.

265. Upregulation of SYNGAP1 expression in mice and human neurons by redirecting alternative splicing [текст] / R. Yang [и др.] // Neuron. — 2023. — май. — т. 111, № 10. — с. 1637—1650.

266. Antisense oligonucleotide modulation of non-productive alternative splicing upregulates gene expression [текст] / K. H. Lim [и др.] // Nat Commun. — 2020. — июль. — т. 11, № 1. — с. 3501.

267. Aberrant Inclusion of a Poison Exon Causes Dravet Syndrome and Related SCN1A-Associated Genetic Epilepsies [текст] / G. L. Carvill [и др.] // Am J Hum Genet. — 2018. — дек. — т. 103, № 6. — с. 1022—1029.

268. Disrupted auto-regulation of the spliceosomal gene SNRPB causes cerebro-costo-mandibular syndrome [текст] / D. C. Lynch [и др.] // Nat Commun. — 2014. — июль. — т. 5. — с. 4483.

269. Regulating PCCA gene expression by modulation of pseudoexon splicing patterns to rescue enzyme activity in propionic acidemia [текст] / U. S. Spangsberg Petersen [и др.] // Mol Ther Nucleic Acids. — 2024. — март. — т. 35, № 1. — с. 102101.

270. Recurrent SRSF2 mutations in MDS affect both splicing and NMD [текст] / M. A. Rahman [и др.] // Genes Dev. — 2020. — март. — т. 34, № 5/6. — с. 413—427.

271. SRSF2 Mutations Contribute to Myelodysplasia by Mutant-Specific Effects on Exon Recognition [текст] / E. Kim [и др.] // Cancer Cell. — 2015. — май. — т. 27, № 5. — с. 617—630.

272. SF3B1 mutations in myelodysplastic syndromes: A potential therapeutic target for modulating the entire disease process [текст] / M. Jiang [и др.] // Front Oncol. — 2023. — т. 13. — с. 1116438.

273. Spliceosomal disruption of the non-canonical BAF complex in cancer [текст] / D. Inoue [и др.] // Nature. — 2019. — окт. — т. 574, № 7778. — с. 432—436.

274. N^6-Methyladenosine Modulates Nonsense-Mediated mRNA Decay in Human Glioblastoma [текст] / F. Li [и др.] // Cancer Res. — 2019. — нояб. — т. 79, № 22. — с. 5785—5798.

275. *Bennett, C. F.* Therapeutic Antisense Oligonucleotides Are Coming of Age [текст] / C. F. Bennett // Annu Rev Med. — 2019. — янв. — т. 70. — с. 307—321.

276. Nusinersen: A Treatment for Spinal Muscular Atrophy [текст] / M. K. Claborn [и др.] // Ann Pharmacother. — 2019. — янв. — т. 53, № 1. — с. 61—69.

277. Treatment of Symptomatic Spinal Muscular Atrophy with Nusinersen: A Prospective Longitudinal Study on Scoliosis Progression [текст] / H. N. H. Ip [и др.] // J Neuromuscul Dis. — 2024. — т. 11, № 2. — с. 349—359.

278. Intracellular localization and splicing regulation of FUS/TLS are variably affected by amyotrophic lateral sclerosis-linked mutations [текст] / Y. Kino [и др.] // Nucleic Acids Res. — 2011. — апр. — т. 39, № 7. — с. 2781—2798.

279. ALS-associated fused in sarcoma (FUS) mutations disrupt Transportin-mediated nuclear import [текст] / D. Dormann [и др.] // EMBO J. — 2010. — авг. — т. 29, № 16. — с. 2841—2857.

280. ALS-associated FUS mutations result in compromised FUS alternative splicing and autoregulation [текст] / Y. Zhou [и др.] // PLoS Genet. — 2013. — окт. — т. 9, № 10. — e1003895.

281. RNA buffers the phase separation behavior of prion-like RNA binding proteins [текст] / S. Maharana [и др.] // Science. — 2018. — май. — т. 360, № 6391. — с. 918—921.

282. A Liquid-to-Solid Phase Transition of the ALS Protein FUS Accelerated by Disease Mutation [текст] / A. Patel [и др.] // Cell. — 2015. — авг. — т. 162, № 5. — с. 1066—1077.

283. Molecular determinants and genetic modifiers of aggregation and toxicity for the ALS disease protein FUS/TLS [текст] / Z. Sun [и др.] // PLoS Biol. — 2011. — апр. — т. 9, № 4. — e1000614.

284. *Schneider-Poetsch, T.* Splicing modulators: on the way from nature to clinic [текст] / T. Schneider-Poetsch, J. K. Chhipi-Shrestha, M. Yoshida // J Antibiot (Tokyo). — 2021. — окт. — т. 74, № 10. — с. 603—616.

285. An orally available, brain penetrant, small molecule lowers huntingtin levels by enhancing pseudoexon inclusion [текст] / C. G. Keller [и др.] // Nat Commun. — 2022. — март. — т. 13, № 1. — с. 1150.

286. *Estevez-Fraga, C.* Huntington's Disease Clinical Trials Corner: November 2022 [текст] / C. Estevez-Fraga, S. J. Tabrizi, E. J. Wild // J Huntingtons Dis. — 2022. — т. 11, № 4. — с. 351—367.

287. *Paik, J.* Risdiplam: A Review in Spinal Muscular Atrophy [текст] / J. Paik // CNS Drugs. — 2022. — апр. — т. 36, № 4. — с. 401—410.

288. An alternative splicing modulator decreases mutant HTT and improves the molecular fingerprint in Huntington's disease patient neurons [текст] / F. Krach [и др.] // Nat Commun. — 2022. — нояб. — т. 13, № 1. — с. 6797.

289. *Williamson, J. R.* The ribosome at atomic resolution [текст] / J. R. Williamson // Cell. — 2009. — дек. — т. 139, № 6. — с. 1041—1043.

290. Cryo-EM Structure of Human Dicer and Its Complexes with a Pre-miRNA Substrate [текст] / Z. Liu [и др.] // Cell. — 2018. — май. — т. 173, № 5. — с. 1191—1203.

291. Cryo-EM advances in RNA structure determination [текст] / H. Ma [и др.] // Signal Transduct Target Ther. — 2022. — февр. — т. 7, № 1. — с. 58.

292. *Frank, J.* Single-Particle Reconstruction of Biological Molecules-Story in a Sample (Nobel Lecture) [текст] / J. Frank // Angew Chem Int Ed Engl. — 2018. — авг. — т. 57, № 34. — с. 10826—10841.

293. The Protein Data Bank [текст] / H. M. Berman [и др.] // Nucleic Acids Res. — 2000. — янв. — т. 28, № 1. — с. 235—242.

294. Advances and opportunities in RNA structure experimental determination and computational modeling [текст] / J. Zhang [и др.] // Nat Methods. — 2022. — окт. — т. 19, № 10. — с. 1193—1207.

295. Genome-wide measurement of RNA secondary structure in yeast [текст] / M. Kertesz [и др.] // Nature. — 2010. — сент. — т. 467, № 7311. — с. 103—107.

296. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features [текст] / Y. Ding [и др.] // Nature. — 2014. — янв. — т. 505, № 7485. — с. 696—700.

297. Genome-wide probing of RNA structure reveals active unfolding of mRNA structures in vivo [текст] / S. Rouskin [и др.] // Nature. — 2014. — янв. — т. 505, № 7485. — с. 701—705.

298. Structural imprints in vivo decode RNA regulatory mechanisms [текст] / R. C. Spitale [и др.] // Nature. — 2015. — март. — т. 519, № 7544. — с. 486—490.

299. *Watters, K. E.* Mapping RNA Structure In Vitro with SHAPE Chemistry and Next-Generation Sequencing (SHAPE-Seq) [текст] / K. E. Watters, J. B. Lucks // Methods Mol Biol. — 2016. — т. 1490. — с. 135—162.

300. Pervasive Regulatory Functions of mRNA Structure Revealed by High-Resolution SHAPE Probing [текст] / A. M. Mustoe [и др.] // Cell. — 2018. — март. — т. 173, № 1. — с. 181—195.

301. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding [текст] / A. Helwak [и др.] // Cell. — 2013. — апр. — т. 153, № 3. — с. 654—665.

302. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1 [текст] / Y. Sugimoto [и др.] // Nature. — 2015. — март. — т. 519, № 7544. — с. 491—494.

303. Higher-Order Organization Principles of Pre-translational mRNPs [текст] / M. Metkar [и др.] // Mol Cell. — 2018. — нояб. — т. 72, № 4. — с. 715—726.

304. *Ponting, C. P.* Evolution and functions of long noncoding RNAs [текст] / C. P. Ponting, P. L. Oliver, W. Reik // Cell. — 2009. — февр. — т. 136, № 4. — с. 629—641.

305. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression [текст] / T. Derrien [и др.] // Genome Res. — 2012. — сент. — т. 22, № 9. — с. 1775—1789.

306. RNA interference: biology, mechanism, and applications [текст] / N. Agrawal [и др.] // Microbiol Mol Biol Rev. — 2003. — дек. — т. 67, № 4. — с. 657—685.

307. An extensive microRNA-mediated network of RNA-RNA interactions regulates established oncogenic pathways in glioblastoma [текст] / P. Sumazin [и др.] // Cell. — 2011. — окт. — т. 147, № 2. — с. 370—381.

308. Emerging roles of RNA-RNA interactions in transcriptional regulation [текст] / D. Wang [и др.] // Wiley Interdiscip Rev RNA. — 2022. — сент. — т. 13, № 5. — e1712.

309. *Xue, Y.* Architecture of RNA-RNA interactions [текст] / Y. Xue // Curr Opin Genet Dev. — 2022. — февр. — т. 72. — с. 138—144.

310. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure [текст] / D. H. Mathews [и др.] // J Mol Biol. — 1999. — май. — т. 288, № 5. — с. 911—940.

311. *Zuker, M.* On finding all suboptimal foldings of an RNA molecule [текст] / M. Zuker // Science. — 1989. — апр. — т. 244, № 4900. — с. 48—52.

312. ViennaRNA Package 2.0 [текст] / R. Lorenz [и др.] // Algorithms Mol Biol. — 2011. — нояб. — т. 6. — с. 26.

313. *Herschlag, D.* RNA chaperones and the RNA folding problem [текст] / D. Herschlag // J Biol Chem. — 1995. — сент. — т. 270, № 36. — с. 20871—20874.

314. *Schroeder, R.* Strategies for RNA folding and assembly [текст] / R. Schroeder, A. Barta, K. Semrad // Nat Rev Mol Cell Biol. — 2004. — нояб. — т. 5, № 11. — с. 908—919.

315. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP) [текст] / E. L. Van Nostrand [и др.] // Nat Methods. — 2016. — июнь. — т. 13, № 6. — с. 508—514.

316. Use of Bru-Seq and BruChase-Seq for genome-wide assessment of the synthesis and stability of RNA [текст] / M. T. Paulsen [и др.] // Methods. — 2014. — май. — т. 67, № 1. — с. 45—54.

317. *Morgan, S.* Evidence for kinetic effects in the folding of large RNA molecules [текст] / S. Morgan, P. Higgs // The Journal of Chemical Physics. — 1996. — т. 105. — с. 7152.

318. *Lai, D.* On the importance of cotranscriptional RNA structure formation [текст] / D. Lai, J. R. Proctor, I. M. Meyer // RNA. — 2013. — нояб. — т. 19, № 11. — с. 1461—1473.

319. *Lyngsø, R. B.* RNA pseudoknot prediction in energy-based models [текст] / R. B. Lyngsø, C. N. Pedersen // J Comput Biol. — 2000. — т. 7, № 3/4. — с. 409—427.

320. *Rivas, E.* A dynamic programming algorithm for RNA structure prediction including pseudoknots [текст] / E. Rivas, S. R. Eddy // J Mol Biol. — 1999. — февр. — т. 285, № 5. — с. 2053—2068.

321. *Pervouchine, D. D.* IRIS: intermolecular RNA interaction search [текст] / D. D. Pervouchine // Genome Informatics. — 2004. — т. 15, № 2. — с. 92—101. — (1.04 п. л.; Вклад автора 100%).

322. *Umu, S. U.* A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life [текст] / S. U. Umu, P. P. Gardner // Bioinformatics. — 2017. — апр. — т. 33, № 7. — с. 988—996.

323. *Lai, D.* A comprehensive comparison of general RNA-RNA interaction prediction methods [текст] / D. Lai, I. M. Meyer // Nucleic Acids Res. — 2016. — апр. — т. 44, № 7. — e61.

324. Thermodynamics of RNA-RNA binding [текст] / U. Mückstein [и др.] // Bioinformatics. — 2006. — май. — т. 22, № 10. — с. 1177—1182.

325. *Wenzel, A.* RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model [текст] / A. Wenzel, E. Akbasli, J. Gorodkin // Bioinformatics. — 2012. — нояб. — т. 28, № 21. — с. 2738—2746.

326. Fast accessibility-based prediction of RNA-RNA interactions [текст] / H. Tafer [и др.] // Bioinformatics. — 2011. — июль. — т. 27, № 14. — с. 1934—1940.

327. Fast and effective prediction of microRNA/target duplexes [текст] / M. Rehmsmeier [и др.] // RNA. — 2004. — окт. — т. 10, № 10. — с. 1507—1517.

328. *John, B.* Prediction of human microRNA targets [текст] / B. John, C. Sander, D. S. Marks // Methods Mol Biol. — 2006. — т. 342. — с. 101—113.

329. Transcriptome-wide prediction of miRNA targets in human and mouse using FASTH [текст] / C. Ragan [и др.] // PLoS One. — 2009. — май. — т. 4, № 5. — e5745.

330. *Wiebe, N. J. P.* TRANSAT– method for detecting the conserved helices of functional RNA structures, including transient, pseudo-knotted and alternative structures [текст] / N. J. P. Wiebe, I. M. Meyer // PLoS Comput Biol. — 2010. — июнь. — т. 6, № 6. — e1000823.

331. *Gardner, P. P.* A comprehensive comparison of comparative RNA structure prediction approaches [текст] / P. P. Gardner, R. Giegerich // BMC Bioinformatics. — 2004. — сент. — т. 5. — с. 140.

332. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences [текст] / S. E. Seemann [и др.] // Bioinformatics. — 2011. — янв. — т. 27, № 2. — с. 211—219.

333. *Bindewald, E.* Computational detection of abundant long-range nucleotide covariation in Drosophila genomes [текст] / E. Bindewald, B. A. Shapiro // RNA. — 2013. — сент. — т. 19, № 9. — с. 1171—1182.

334. *Fricke, M.* Prediction of conserved long-range RNA-RNA interactions in full viral genomes [текст] / M. Fricke, M. Marz // Bioinformatics. — 2016. — окт. — т. 32, № 19. — с. 2928—2935.

335. A comparative method for finding and folding RNA secondary structures within protein-coding regions [текст] / J. S. Pedersen [и др.] // Nucleic Acids Res. — 2004. — т. 32, № 16. — с. 4925—4936.

336. An evolutionary model for protein-coding regions with conserved RNA structure [текст] / J. S. Pedersen [и др.] // Mol Biol Evol. — 2004. — окт. — т. 21, № 10. — с. 1913—1922.

337. *Fu, Y.* Dynalign II: common secondary structure prediction for RNA homologs with domain insertions [текст] / Y. Fu, G. Sharma, D. H. Mathews // Nucleic Acids Res. — 2014. — дек. — т. 42, № 22. — с. 13939—13948.

338. *Eddy, S. R.* RNA sequence analysis using covariance models [текст] / S. R. Eddy, R. Durbin // Nucleic Acids Res. — 1994. — июнь. — т. 22, № 11. — с. 2079—2088.

339. *Sun, E. I.* Computational analysis of riboswitch-based regulation [текст] / E. I. Sun, D. A. Rodionov // Biochim Biophys Acta. — 2014. — окт. — т. 1839, № 10. — с. 900—907.

340. Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions [текст] / S. E. Seemann [и др.] // Algorithms Mol Biol. — 2010. — май. — т. 5. — с. 22.

341. RNA-RNA interaction prediction based on multiple sequence alignments [текст] / A. X. Li [и др.] // Bioinformatics. — 2011. — февр. — т. 27, № 4. — с. 456—463.

342. *Knudsen, B.* Pfold: RNA secondary structure prediction using stochastic context-free grammars [текст] / B. Knudsen, J. Hein // Nucleic Acids Res. — 2003. — июль. — т. 31, № 13. — с. 3423—3428.

343. Identification and classification of conserved RNA secondary structures in the human genome [текст] / J. S. Pedersen [и др.] // PLoS Comput Biol. — 2006. — апр. — т. 2, № 4. — e33.

344. *Rivas, E.* A range of complex probabilistic models for RNA secondary structure prediction that includes the nearest-neighbor model and more [текст] / E. Rivas, R. Lang, S. R. Eddy // RNA. — 2012. — февр. — т. 18, № 2. — с. 193—212.

345. *Sankoff, D.* Simultaneous solution of the RNA folding, alignment and protosequence problems. [текст] / D. Sankoff // SIAM J. Appl. Math. — 1985. — т. 45, № 5. — с. 810—825.

346. *Havgaard, J. H.* Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix [текст] / J. H. Havgaard, E. Torarinsson, J. Gorodkin // PLoS Comput Biol. — 2007. — окт. — т. 3, № 10. — с. 1896—1908.

347. *Hatje, K.* Expansion of the mutually exclusive spliced exome in Drosophila [текст] / K. Hatje, M. Kollmar // Nat Commun. — 2013. — т. 4. — с. 2460.

348. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming [текст] / Y. Kato [и др.] // Bioinformatics. — 2010. — сент. — т. 26, № 18. — с. i460—466.

349. *Touzet, H.* CARNAC: folding families of related RNAs [текст] / H. Touzet, O. Perriquet // Nucleic Acids Res. — 2004. — июль. — т. 32, Web Server issue. — W142—145.

350. A predictive model for secondary RNA structure using graph theory and a neural network [текст] / D. R. Koessler [и др.] // BMC Bioinformatics. — 2010. — окт. — т. 11 Suppl 6, Suppl 6. — S21.

351. BRD2 and BRD3 genes independently evolved RNA structures to control unproductive splicing [текст] / M. Petrova [и др.] // NAR Genomics and Bioinformatics. — 2024. — март. — т. 6, № 1. — lqad113. — (1.16 п. л.; Вклад автора 40%; JIF=4.6 WoS).

352. SMN2 gene and restore SMN protein expression in type 1 SMA fibroblasts [текст] / A. Touznik [и др.] // Sci Rep. — 2017. — июнь. — т. 7, № 1. — с. 3672.

353. One signal stimulates different transcriptional activation mechanisms [текст] / M. Y. Mazina [и др.] // Biochim Biophys Acta Gene Regul Mech. — 2018. — февр. — т. 1861, № 2. — с. 178—189.

354. *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. — (0.23 п. л.; Вклад автора 75%; JIF=5.8 WoS).

355. Gene-specific patterns of expression variation across organs and species [текст] / A. Breschi [и др.] // Genome Biology. — 2016. — июль. — т. 17, № 1. — с. 151. — (1.50 п. л.; Вклад автора 20%; JIF=12.3 WoS).

356. Human genomics. The human transcriptome across tissues and individuals [текст] / M. Melé [и др.] // Science. — 2015. — май. — т. 348, № 6235. — с. 660—665. — (0.69 п. л.; Вклад автора 10%; JIF=56.9 WoS).

357. Enhanced transcriptome maps from multiple mouse tissues reveal evolutionary constraint in gene expression [текст] / D. D. Pervouchine [и др.] // Nature Communications. — 2015. — янв. — т. 6. — с. 5903. — (1.27 п. л.; Вклад автора 50%; JIF=16.6 WoS).

358. A comparative encyclopedia of DNA elements in the mouse genome [текст] / F. Yue [и др.] // Nature. — 2014. — нояб. — т. 515, № 7527. — с. 355—364. — (1.16 п. л.; Вклад автора 5%; JIF=64.8 WoS).

359. The effects of death and post-mortem cold ischemia on human tissue transcriptomes [текст] / P. G. Ferreira [и др.] // Nature Communications. — 2018. — февр. — т. 9, № 1. — с. 490. — (1.73 п. л.; Вклад автора 10%; JIF=16.6 WoS).

360. Landscape of transcription in human cells [текст] / S. Djebali [и др.] // Nature. — 2012. — сент. — т. 489, № 7414. — с. 101—108.

361. *GTEx Consortium.* Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans [текст] / GTEx Consortium // Science. — 2015. — май. — т. 348, № 6235. — с. 648—660.

362. *ENCODE Project Consortium.* An integrated encyclopedia of DNA elements in the human genome [текст] / ENCODE Project Consortium // Nature. — 2012. — сент. — т. 489, № 7414. — с. 57—74.

363. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate [текст] / N. Fong [и др.] // Genes Dev. — 2014. — дек. — т. 28, № 23. — с. 2663—2676.

364. RNAcontacts: A Pipeline for Predicting Contacts from RNA Proximity Ligation Assays [текст] / S. D. Margasyuk [и др.] // Acta Naturae. — 2023. — т. 15, № 1. — с. 51—57. — (0.81 п. л.; Вклад автора 50%; JIF=2.0 WoS).

365. STAR: ultrafast universal RNA-seq aligner [текст] / A. Dobin [и др.] // Bioinformatics. — 2013. — янв. — т. 29, № 1. — с. 15—21.

366. The Cancer Genome Atlas Pan-Cancer analysis project [текст] / J. N. Weinstein [и др.] // Nat Genet. — 2013. — окт. — т. 45, № 10. — с. 1113—1120.

367. Human nonsense-mediated RNA decay initiates widely by endonucleolysis and targets snoRNA host genes [текст] / S. Lykke-Andersen [и др.] // Genes Dev. — 2014. — нояб. — т. 28, № 22. — с. 2498—2517.

368. rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data [текст] / S. Shen [и др.] // Proc Natl Acad Sci U S A. — 2014. — дек. — т. 111, № 51. — E5593—5601.

369. ProteomicsDB: toward a FAIR open-source resource for life-science research [текст] / L. Lautenbacher [и др.] // Nucleic Acids Res. — 2022. — янв. — т. 50, № D1. — с. D1541—D1552.

370. The UCSC Genome Browser Database [текст] / D. Karolchik [и др.] // Nucleic Acids Res. — 2003. — янв. — т. 31, № 1. — с. 51—54.

371. GENCODE 2021 [текст] / A. Frankish [и др.] // Nucleic Acids Res. — 2021. — янв. — т. 49, № D1. — с. D916—D923.

372. Aligning multiple genomic sequences with the threaded blockset aligner [текст] / M. Blanchette [и др.] // Genome Res. — 2004. — апр. — т. 14, № 4. — с. 708—715.

373. The human genome browser at UCSC [текст] / W. J. Kent [и др.] // Genome Res. — 2002. — июнь. — т. 12, № 6. — с. 996—1006.

374. *Love, M. I.* Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2 [текст] / M. I. Love, W. Huber, S. Anders // Genome Biol. — 2014. — т. 15, № 12. — с. 550.

375. RNA-SeQC 2: efficient RNA-seq quality control and quantification for large cohorts [текст] / A. Graubert [и др.] // Bioinformatics. — 2021. — сент. — т. 37, № 18. — с. 3048—3050.

376. *Zhu, A.* Heavy-tailed prior distributions for sequence count data: removing the noise and preserving large differences [текст] / A. Zhu, J. G. Ibrahim, M. I. Love // Bioinformatics. — 2019. — июнь. — т. 35, № 12. — с. 2084—2092.

377. *Storey, J.* The positive false discovery rate: a Bayesian interpretation and the q-value [текст] / J. Storey // Ann. Statist. — 2003. — т. 31, № 6. — с. 2013—2035.

378. *Shapiro, B. A.* Comparing multiple RNA secondary structures using tree comparisons [текст] / B. A. Shapiro, K. Z. Zhang // Comput Appl Biosci. — 1990. — окт. — т. 6, № 4. — с. 309—318.

379. Local similarity in RNA secondary structures [текст] / M. chsmann [и др.] // Proc IEEE Comput Soc Bioinform Conf. — 2003. — т. 2. — с. 159—168.

380. *Gerlach, W.* GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing [текст] / W. Gerlach, R. Giegerich // Bioinformatics. — 2006. — март. — т. 22, № 6. — с. 762—764.

381. *Ma, B.* PatternHunter: faster and more sensitive homology search [текст] / B. Ma, J. Tromp, M. Li // Bioinformatics. — 2002. — март. — т. 18, № 3. — с. 440—445.

382. *Edgar, R. C.* MUSCLE: multiple sequence alignment with high accuracy and high throughput [текст] / R. C. Edgar // Nucleic Acids Res. — 2004. — т. 32, № 5. — с. 1792—1797.

383. *Tafer, H.* RNAplex: a fast tool for RNA-RNA interaction search [текст] / H. Tafer, I. L. Hofacker // Bioinformatics. — 2008. — нояб. — т. 24, № 22. — с. 2657—2663.

384. *Stanley, J. R.* Cell adhesion molecules as targets of autoantibodies in pemphigus and pemphigoid, bullous diseases due to defective epidermal cell adhesion [текст] / J. R. Stanley // Adv Immunol. — 1993. — т. 53. — с. 291—325.

385. Novel alternative splicings of BPAG1 (bullous pemphigoid antigen 1) including the domain structure closely related to MACF (microtubule actin cross-linking factor) [текст] / M. Okumura [и др.] // J Biol Chem. — 2002. — февр. — т. 277, № 8. — с. 6682—6687.

386. *Menhart, N.* Hybrid spectrin type repeats produced by exon-skipping in dystrophin [текст] / N. Menhart // Biochim Biophys Acta. — 2006. — июнь. — т. 1764, № 6. — с. 993—999.

387. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing [текст] / S. Kishore [и др.] // Hum Mol Genet. — 2010. — апр. — т. 19, № 7. — с. 1153—1164.

388. Human box C/D snoRNA processing conservation across multiple cell types [текст] / M. S. Scott [и др.] // Nucleic Acids Res. — 2012. — апр. — т. 40, № 8. — с. 3676—3688.

389. Analogues of artificial human box C/D small nucleolar RNA as regulators of alternative splicing of a pre-mRNA target [текст] / G. A. Stepanov [и др.] // Acta Naturae. — 2012. — янв. — т. 4, № 1. — с. 32—41.

390. *Kishore, S.* The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C [текст] / S. Kishore, S. Stamm // Science. — 2006. — янв. — т. 311, № 5758. — с. 230—232.

391. *Pervouchine, D. D.* Towards Long-Range RNA Structure Prediction in Eukaryotic Genes [текст] / D. D. Pervouchine // Genes. — 2018. — июнь. — т. 9, № 6. — с. 302. — (1.04 п. л.; Вклад автора 100%; JIF=3.5 WoS).

392. *Smith, T. F.* Identification of common molecular subsequences [текст] / T. F. Smith, M. S. Waterman // J Mol Biol. — 1981. — март. — т. 147, № 1. — с. 195—197.

393. *Mann, M.* IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions [текст] / M. Mann, P. R. Wright, R. Backofen // Nucleic Acids Res. — 2017. — июль. — т. 45, W1. — W435—W439.

394. Identification and characterization of multi-species conserved sequences [текст] / E. H. Margulies [и др.] // Genome Res. — 2003. — дек. — т. 13, № 12. — с. 2507—2518.

395. *Li, P.* icSHAPE-pipe: A comprehensive toolkit for icSHAPE data analysis and evaluation [текст] / P. Li, R. Shi, Q. C. Zhang // Methods. — 2020. — июнь. — т. 178. — с. 96—103.

396. *Chamary, J. V.* Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals [текст] / J. V. Chamary, L. D. Hurst // Genome Biol. — 2005. — т. 6, № 9. — R75.

397. A global reference for human genetic variation [текст] / A. Auton [и др.] // Nature. — 2015. — окт. — т. 526, № 7571. — с. 68—74.

398. Human Survival Motor Neuron genes generate a vast repertoire of circular RNAs [текст] / E. W. Ottesen [и др.] // Nucleic Acids Res. — 2019. — апр. — т. 47, № 6. — с. 2884—2905.

399. Comprehensive characterization of tissue-specific circular RNAs in the human and mouse genomes [текст] / S. Xia [и др.] // Brief Bioinform. — 2017. — нояб. — т. 18, № 6. — с. 984—992.

400. RNA editing by mammalian ADARs [текст] / M. Hogg [и др.] // Adv Genet. — 2011. — т. 73. — с. 87—120.

401. *Nishikura, K.* Functions and regulation of RNA editing by ADAR deaminases [текст] / K. Nishikura // Annu Rev Biochem. — 2010. — т. 79. — с. 321—349.

402. *Ramaswami, G.* RADAR: a rigorously annotated database of A-to-I RNA editing [текст] / G. Ramaswami, J. B. Li // Nucleic Acids Res. — 2014. — янв. — т. 42, Database issue. — с. D109—113.

403. REDIportal: a comprehensive database of A-to-I RNA editing events in humans [текст] / E. Picardi [и др.] // Nucleic Acids Res. — 2017. — янв. — т. 45, № D1. — с. D750—D757.

404. *Wu, X.* Widespread Influence of 3'-End Structures on Mammalian mRNA Processing and Stability [текст] / X. Wu, D. P. Bartel // Cell. — 2017. — май. — т. 169, № 5. — с. 905—917.

405. 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.

406. A quantitative atlas of polyadenylation in five mammals [текст] / A. Derti [и др.] // Genome Res. — 2012. — июнь. — т. 22, № 6. — с. 1173—1183.

407. Post-transcriptional processing generates a diversity of 5'-modified long and short RNAs [текст] / K. Fejes-Toth [и др.] // Nature. — 2009. — февр. — т. 457, № 7232. — с. 1028—1032.

408. *Vlasenok, M.* Transcriptome sequencing suggests that pre-mRNA splicing counteracts widespread intronic cleavage and polyadenylation [текст] / M. Vlasenok, S. Margasyuk, D. D. Pervouchine // NAR Genomics and Bioinformatics. — 2023. — июнь. — т. 5, № 2. — lqad051. — (1.73 п. л.; Вклад автора 50%; JIF=4.6 WoS).

409. *Tian, B.* Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing [текст] / B. Tian, Z. Pan, J. Y. Lee // Genome Res. — 2007. — февр. — т. 17, № 2. — с. 156—165.

410. *Ruskin, B.* An RNA processing activity that debranches RNA lariats [текст] / B. Ruskin, M. R. Green // Science. — 1985. — июль. — т. 229, № 4709. — с. 135—140.

411. *Mohanta, A.* Dbr1 functions in mRNA processing, intron turnover and human diseases [текст] / A. Mohanta, K. Chakrabarti // Biochimie. — 2021. — янв. — т. 180. — с. 134—142.

412. Alternative polyadenylation by sequential activation of distal and proximal PolyA sites [текст] / P. Tang [и др.] // Nat Struct Mol Biol. — 2022. — янв. — т. 29, № 1. — с. 21—31.

413. Reduced expression of the ATRX gene, a chromatin-remodeling factor, causes hippocampal dysfunction in mice [текст] / T. Nogami [и др.] // Hippocampus. — 2011. — июнь. — т. 21, № 6. — с. 678—687.

414. Tissue-selective restriction of RNA editing of CaV1.3 by splicing factor SRSF9 [текст] / H. Huang [и др.] // Nucleic Acids Res. — 2018. — авг. — т. 46, № 14. — с. 7323—7338.

415. Crystal structure of the two-RRM domain of hnRNP A1 (UP1) complexed with single-stranded telomeric DNA [текст] / J. Ding [и др.] // Genes Dev. — 1999. — май. — т. 13, № 9. — с. 1102—1115.

416. *Wang, X.* Nucleic acid-binding specificity of human FUS protein [текст] / X. Wang, J. C. Schwartz, T. R. Cech // Nucleic Acids Res. — 2015. — сент. — т. 43, № 15. — с. 7535—7543.

417. *Schor, I. E.* Coupling between transcription and alternative splicing [текст] / I. E. Schor, L. I. a, A. R. Kornblihtt // Cancer Treat Res. — 2013. — т. 158. — с. 1—24.

418. *Baralle, F. E.* RNA structure and splicing regulation [текст] / F. E. Baralle, R. N. Singh, S. Stamm // Biochim Biophys Acta Gene Regul Mech. — 2019. — т. 1862, № 11/12. — с. 194448.

419. Genetic mapping uncovers cis-regulatory landscape of RNA editing [текст] / G. Ramaswami [и др.] // Nat Commun. — 2015. — сент. — т. 6. — с. 8194.

420. RNA polymerase II kinetics in polo polyadenylation signal selection [текст] / P. A. Pinto [и др.] // EMBO J. — 2011. — май. — т. 30, № 12. — с. 2431—2444.

421. *Graveley, B. R.* RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor [текст] / B. R. Graveley, E. S. Fleming, G. M. Gilmartin // Mol Cell Biol. — 1996. — сент. — т. 16, № 9. — с. 4942—4951.

422. *Leppek, K.* Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them [текст] / K. Leppek, R. Das, M. Barna // Nat Rev Mol Cell Biol. — 2018. — март. — т. 19, № 3. — с. 158—174.

423. U1 snRNP telescripting regulates a size-function-stratified human genome [текст] / J. M. Oh [и др.] // Nat Struct Mol Biol. — 2017. — нояб. — т. 24, № 11. — с. 993—999.

424. Direct interactions between subunits of CPSF and the U2 snRNP contribute to the coupling of pre-mRNA 3' end processing and splicing [текст] / A. Kyburz [и др.] // Mol Cell. — 2006. — июль. — т. 23, № 2. — с. 195—205.

425. The 3'-end-processing factor CPSF is required for the splicing of single-intron pre-mRNAs in vivo [текст] / Y. Li [и др.] // RNA. — 2001. — июнь. — т. 7, № 6. — с. 920—931.

426. *Misra, A.* From polyadenylation to splicing: Dual role for mRNA 3' end formation factors [текст] / A. Misra, M. R. Green // RNA Biol. — 2016. — т. 13, № 3. — с. 259—264.

427. Capture RIC-seq reveals positional rules of PTBP1-associated RNA loops in splicing regulation [текст] / R. Ye [и др.] // Mol Cell. — 2023. — апр. — т. 83, № 8. — с. 1311—1327.

428. *Quinlan, A. R.* BEDTools: The Swiss-Army Tool for Genome Feature Analysis [текст] / A. R. Quinlan // Curr Protoc Bioinformatics. — 2014. — сент. — т. 47. — с. 1—34.

429. Global analysis of the RNA-protein interaction and RNA secondary structure landscapes of the Arabidopsis nucleus [текст] / S. J. Gosai [и др.] // Mol Cell. — 2015. — янв. — т. 57, № 2. — с. 376—388.

430. RNA structure maps across mammalian cellular compartments [текст] / L. Sun [и др.] // Nat Struct Mol Biol. — 2019. — апр. — т. 26, № 4. — с. 322—330.

431. *Zafrir, Z.* Nucleotide sequence composition adjacent to intronic splice sites improves splicing efficiency via its effect on pre-mRNA local folding in fungi [текст] / Z. Zafrir, T. Tuller // RNA. — 2015. — окт. — т. 21, № 10. — с. 1704—1718.

432. Mutations in GANAB, encoding the glucosidase IIα subunit, cause autosomal-dominant polycystic kidney and liver disease [текст] / B. Porath [и др.] // Am J Hum Genet. — 2016. — июнь. — т. 98, № 6. — с. 1193—1207.

433. A noncoding variant in GANAB explains isolated polycystic liver disease (PCLD) in a large family [текст] / W. Besse [и др.] // Hum Mutat. — 2018. — март. — т. 39, № 3. — с. 378—382.

434. ZNF655 accelerates progression of pancreatic cancer by promoting the binding of E2F1 and CDK1 [текст] / Z. Shao [и др.] // Oncogenesis. — 2022. — авг. — т. 11, № 1. — с. 44.

435. *Felsenstein, J.* A Hidden Markov Model approach to variation among sites in rate of evolution [текст] / J. Felsenstein, G. A. Churchill // Mol Biol Evol. — 1996. — янв. — т. 13, № 1. — с. 93—104.

436. *McCord, R. P.* Chromosome Conformation Capture and Beyond: Toward an Integrative View of Chromosome Structure and Function [текст] / R. P. McCord, N. Kaplan, L. Giorgetti // Mol Cell. — 2020. — февр. — т. 77, № 4. — с. 688—708.

437. *Jerkovic, I.* Understanding 3D genome organization by multidisciplinary methods [текст] / I. Jerkovic, G. Cavalli // Nat Rev Mol Cell Biol. — 2021. — авг. — т. 22, № 8. — с. 511—528.

438. Robust Hi-C Maps of Enhancer-Promoter Interactions Reveal the Function of Non-coding Genome in Neural Development and Diseases [текст] / L. Lu [и др.] // Mol Cell. — 2020. — авг. — т. 79, № 3. — с. 521—534.

439. Integrative Analysis of Zika Virus Genome RNA Structure Reveals Critical Determinants of Viral Infectivity [текст] / P. Li [и др.] // Cell Host Microbe. — 2018. — дек. — т. 24, № 6. — с. 875—886.

440. *Blanchette, M.* Modulation of exon skipping by high-affinity hnRNP A1-binding sites and by intron elements that repress splice site utilization [текст] / M. Blanchette, B. Chabot // EMBO J. — 1999. — апр. — т. 18, № 7. — с. 1939—1952.

441. Structure of PTB bound to RNA: specific binding and implications for splicing regulation [текст] / F. C. Oberstrass [и др.] // Science. — 2005. — сент. — т. 309, № 5743. — с. 2054—2057.

442. *Warf, M. B.* Role of RNA structure in regulating pre-mRNA splicing [текст] / M. B. Warf, J. A. Berglund // Trends Biochem Sci. — 2010. — март. — т. 35, № 3. — с. 169—178.

443. Transcriptional elongation and alternative splicing [текст] / G. Dujardin [и др.] // Biochim Biophys Acta. — 2013. — янв. — т. 1829, № 1. — с. 134—140.

444. Design of antisense oligonucleotides stabilized by locked nucleic acids [текст] / J. Kurreck [и др.] // Nucleic Acids Res. — 2002. — май. — т. 30, № 9. — с. 1911—1918.

445. Expanding the design horizon of antisense oligonucleotides with alpha-L-LNA [текст] / M. Frieden [и др.] // Nucleic Acids Res. — 2003. — нояб. — т. 31, № 21. — с. 6365—6372.

446. *Frieden, M.* Nuclease stability of LNA oligonucleotides and LNA-DNA chimeras [текст] / M. Frieden, H. F. Hansen, T. Koch // Nucleosides Nucleotides Nucleic Acids. — 2003. — т. 22, № 5—8. — с. 1041—1043.

447. *Патент на изобретение №2810907.* Система направленного изменения сплайсинга в гене MARK2 [текст] / Д. Д. Первушин [и др.] (Российская Федерация) ; Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий» ; патент. поверенный Егорова Г. Б. — № 2000108705/28 ; заявл. 03.01.2020 ;

опубл. 02.01.2020, Бюл. № 7 (I ч.) ; приоритет 01.01.2020, 09/289, 037 (Рос. Федерация). — 5 с. : ил.

448. *Ma, Z.* Membrane phospholipid asymmetry counters the adverse effects of sterol overloading in the Golgi membrane of Drosophila [текст] / Z. Ma, Z. Liu, X. Huang // Genetics. — 2012. — апр. — т. 190, № 4. — с. 1299—1308.

449. Histone deacetylase-associating Atrophin proteins are nuclear receptor corepressors [текст] / L. Wang [и др.] // Genes Dev. — 2006. — март. — т. 20, № 5. — с. 525—530.

450. *Zhai, R. G.* Nicotinamide/nicotinic acid mononucleotide adenylyltransferase, new insights into an ancient enzyme [текст] / R. G. Zhai, M. Rizzi, S. Garavaglia // Cell Mol Life Sci. — 2009. — сент. — т. 66, № 17. — с. 2805—2818.

451. Alternative splicing of Drosophila Nmnat functions as a switch to enhance neuroprotection under stress [текст] / K. Ruan [и др.] // Nat Commun. — 2015. — нояб. — т. 6. — с. 10057.

452. Tudor, MBT and chromo domains gauge the degree of lysine methylation [текст] / J. Kim [и др.] // EMBO Rep. — 2006. — апр. — т. 7, № 4. — с. 397—403.

453. Methyllysine reader plant homeodomain (PHD) finger protein 20-like 1 (PHF20L1) antagonizes DNA (cytosine-5) methyltransferase 1 (DNMT1) proteasomal degradation [текст] / P.-O. Estève [и др.] // J Biol Chem. — 2014. — март. — т. 89, № 12. — с. 8277—8287.

454. PHF20L1 as a H3K27me2 reader coordinates with transcriptional repressors to promote breast tumorigenesis [текст] / Y. Hou [и др.] // Sci Adv. — 2020. — апр. — т. 6, № 16. — eaaz0356.

455. Tudor-domain protein PHF20L1 reads lysine methylated retinoblastoma tumour suppressor protein [текст] / S. M. Carr [и др.] // Cell Death Differ. — 2017. — дек. — т. 24, № 12. — с. 2139—2149.

456. LSD1 demethylase and the methyl-binding protein PHF20L1 prevent SET7 methyltransferase-dependent proteolysis of the stem-cell protein SOX2 [текст] / C. Zhang [и др.] // J Biol Chem. — 2018. — март. — т. 293, № 10. — с. 3663—3674.

457. Identification of tumor suppressors and oncogenes from genomic and epigenetic features in ovarian cancer [текст] / K. O. Wrzeszczynski [и др.] // PLoS One. — 2011. — т. 6, № 12. — e28503.

458. Integrative genomic and transcriptomic analysis for pinpointing recurrent alterations of plant homeodomain genes and their clinical significance in breast cancer [текст] / H. Yu [и др.] // Oncotarget. — 2017. — февр. — т. 8, № 8. — с. 13099—13115.

459. Two precision medicine predictive tools for six malignant solid tumors: from gene-based research to clinical application [текст] / Z. Zhang [и др.] // J Transl Med. — 2019. — дек. — т. 17, № 1. — с. 405.

460. Human CASK/LIN-2 binds syndecan-2 and protein 4.1 and localizes to the basolateral membrane of epithelial cells [текст] / A. R. Cohen [и др.] // J Cell Biol. — 1998. — июль. — т. 142, № 1. — с. 129—138.

461. Structural basis for nucleotide-dependent regulation of membrane-associated guanylate kinase-like domains [текст] / Y. Li [и др.] // J Biol Chem. — 2002. — февр. — т. 277, № 6. — с. 4159—4165.

462. CASK participates in alternative tripartite complexes in which Mint 1 competes for binding with caskin 1, a novel CASK-binding protein [текст] / K. Tabuchi [и др.] // J Neurosci. — 2002. — июнь. — т. 22, № 11. — с. 4264—4273.

463. Transgenic mouse model of X-linked cleft palate [текст] / J. B. Wilson [и др.] // Cell Growth Differ. — 1993. — февр. — т. 4, № 2. — с. 67—76.

464. Cell CASK Deletion Reduces Hyperinsulinemia [текст] / X. Liu [и др.] // Diabetes. — 2021. — окт.

465. Identification of Tbr-1/CASK complex target genes in neurons [текст] / T. F. Wang [и др.] // J Neurochem. — 2004. — дек. — т. 91, № 6. — с. 1483—1492.

466. *Caruana, G.* Genetic studies define MAGUK proteins as regulators of epithelial cell polarity [текст] / G. Caruana // Int J Dev Biol. — 2002. — т. 46, № 4. — с. 511—518.

467. Alternative Splicing of a Novel Inducible Exon Diversifies the CASK Guanylate Kinase Domain [текст] / J. A. Dembowski [и др.] // J Nucleic Acids. — 2012. — т. 2012. — с. 816237.

468. Functional analysis of CASK transcript variants expressed in human brain [текст] / D. Tibbe [и др.] // PLoS One. — 2021. — т. 16, № 6. — e0253223.

469. 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.

470. 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.

471. *Solomon, V.* The N-end rule pathway catalyzes a major fraction of the protein degradation in skeletal muscle [текст] / V. Solomon, S. H. Lecker, A. L. Goldberg // J Biol Chem. — 1998. — сент. — т. 273, № 39. — с. 25216—25222.

472. 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.

473. *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.

474. 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.

475. 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.

476. Arginylation-dependent neural crest cell migration is essential for mouse development [текст] / S. Kurosaka [и др.] // PLoS Genet. — 2010. — март. — т. 6, № 3. — e1000878.

477. 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.

478. *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.

479. *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.

480. *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.

481. *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.

482. *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.

483. *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.

484. *Brower, C. S.* Ablation of arginylation in the mouse N-end rule pathway: loss of fat, higher metabolic rate, damaged spermatogenesis, and neurological perturbations [текст] / C. S. Brower, A. Varshavsky // PLoS One. — 2009. — нояб. — т. 4, № 11. — e7757.

485. *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.

486. *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.

487. 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.

488. 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.

489. 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.

490. 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.

491. *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.

492. *Reiser, B.* Confidence intervals for the Mahalanobis distance [текст] / B. Reiser // Communications in Statistics. Simulation and Computation. — 2001. — март. — т. 30.

493. RNA Polymerase II Elongation at the Crossroads of Transcription and Alternative Splicing [текст] / M. de la Mata [и др.] // Genet Res Int. — 2011. — т. 2011. — с. 309865.

494. A slow RNA polymerase II affects alternative splicing in vivo [текст] / M. de la Mata [и др.] // Mol Cell. — 2003. — авг. — т. 12, № 2. — с. 525—532.

495. *Gong, X. Q.* Alpha-amanitin blocks translocation by human RNA polymerase II [текст] / X. Q. Gong, Y. A. Nedialkov, Z. F. Burton // J Biol Chem. — 2004. — июнь. — т. 279, № 26. — с. 27422—27427.

496. *Kaplan, C. D.* The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin [текст] / C. D. Kaplan, K.-M. Larsson, R. D. Kornberg // Mol Cell. — 2008. — июнь. — т. 30, № 5. — с. 547—556.

497. Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation [текст] / J. Y. Ip [и др.] // Genome Res. — 2011. — март. — т. 21, № 3. — с. 390—401.

498. NELF and DSIF cause promoter proximal pausing on the hsp70 promoter in Drosophila [текст] / C.-H. Wu [и др.] // Genes Dev. — 2003. — июнь. — т. 17, № 11. — с. 1402—1414.

499. Integrator complex regulates NELF-mediated RNA polymerase II pause/release and processivity at coding genes [текст] / B. Stadelmayer [и др.] // Nat Commun. — 2014. — нояб. — т. 5. — с. 5531.

500. Architecture and RNA binding of the human negative elongation factor [текст] / S. M. Vos [и др.] // Elife. — 2016. — июнь. — т. 5.

501. Evidence that negative elongation factor represses transcription elongation through binding to a DRB sensitivity-inducing factor/RNA polymerase II complex and RNA [текст] / Y. Yamaguchi [и др.] // Mol Cell Biol. — 2002. — май. — т. 22, № 9. — с. 2918—2927.

502. Dynamics of human immunodeficiency virus transcription: P-TEFb phosphorylates RD and dissociates negative effectors from the transactivation response element [текст] / K. Fujinaga [и др.] // Mol Cell Biol. — 2004. — янв. — т. 24, № 2. — с. 787—795.

503. Defining NELF-E RNA binding in HIV-1 and promoter-proximal pause regions [текст] / J. M. Pagano [и др.] // PLoS Genet. — 2014. — янв. — т. 10, № 1. — e1004090.

504. Splicing factors SF1 and U2AF associate in extraspliceosomal complexes [текст] / J. Rino [и др.] // Mol Cell Biol. — 2008. — май. — т. 28, № 9. — с. 3045—3057.

505. Microtubule affinity-regulating kinase 2 is associated with DNA damage response and cisplatin resistance in non-small cell lung cancer [текст] / R. Hubaux [и др.] // Int J Cancer. — 2015. — нояб. — т. 137, № 9. — с. 2072—2082.

506. MARK2 inhibits the growth of HeLa cells through AMPK and reverses epithelial-mesenchymal transition [текст] / G. Xu [и др.] // Oncol Rep. — 2017. — июль. — т. 38, № 1. — с. 237—244.

507. *Chen, Y.* Compensatory evolution of a precursor messenger RNA secondary structure in the Drosophila melanogaster Adh gene [текст] / Y. Chen, W. Stephan // Proc Natl Acad Sci U S A. — 2003. — сент. — т. 100, № 20. — с. 11499—11504.

508. RNA Secondary Structure-Based Design of Antisense Peptide Nucleic Acids for Modulating Disease-Associated Aberrant Tau Pre-mRNA Alternative Splicing [текст] / A. A. L. Ong [и др.] // Molecules. — 2019. — авг. — т. 24, № 16.

509. 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.

510. Re-annotation of 191 developmental and epileptic encephalopathy-associated genes unmasks de novo variants in SCN1A [текст] / C. A. Steward [и др.] // NPJ Genomic Medicine. — 2019. — т. 4, № 1. — с. 31. — (1.27 п. л.; Вклад автора 10%; JIF=5.3 WoS).

511. Interpretation of differential gene expression results of RNA-seq data: review and integration [текст] / A. McDermaid [и др.] // Brief Bioinform. — 2019. — нояб. — т. 20, № 6. — с. 2044—2054.

512. *Bourgeois, C. F.* Broad specificity of SR (serine/arginine) proteins in the regulation of alternative splicing of pre-messenger RNA [текст] / C. F. Bourgeois, F. Lejeune, J. Stévenin // Prog Nucleic Acid Res Mol Biol. — 2004. — т. 78. — с. 37—88.

513. *Fu, Y.* SRSF7 knockdown promotes apoptosis of colon and lung cancer cells [текст] / Y. Fu, Y. Wang // Oncol Lett. — 2018. — апр. — т. 15, № 4. — с. 5545—5552.

514. Serine/arginine-rich splicing factor 7 regulates p21-dependent growth arrest in colon cancer cells [текст] / S. Saijo [и др.] // J Med Invest. — 2016. — т. 63, № 3/4. — с. 219—226.

515. SRSF7 maintains its homeostasis through the expression of Split-ORFs and nuclear body assembly [текст] / V. Königs [и др.] // Nat Struct Mol Biol. — 2020. — март. — т. 27, № 3. — с. 260—273.

516. POSTAR3: an updated platform for exploring post-transcriptional regulation coordinated by RNA-binding proteins [текст] / W. Zhao [и др.] // Nucleic Acids Res. — 2022. — янв. — т. 50, № D1. — с. D287—D294.

517. PSD-95 is post-transcriptionally repressed during early neural development by PTBP1 and PTBP2 [текст] / S. Zheng [и др.] // Nat Neurosci. — 2012. — янв. — т. 15, № 3. — с. 381—388.

518. *Zheng, S.* Alternative splicing and nonsense-mediated mRNA decay enforce neural specific gene expression [текст] / S. Zheng // Int J Dev Neurosci. — 2016. — дек. — т. 55. — с. 102—108.

519. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing [текст] / E. V. Makeyev [и др.] // Mol Cell. — 2007. — авг. — т. 27, № 3. — с. 435—448.

520. TRA2A binds with LncRNA MALAT1 to promote esophageal cancer progression by regulating EZH2/β-catenin pathway [текст] / X. Zhao [и др.] // J Cancer. — 2021. — т. 12, № 16. — с. 4883—4890.

521. TRA2A promotes proliferation, migration, invasion and epithelial mesenchymal transition of glioma cells [текст] / Y. Tan [и др.] // Brain Res Bull. — 2018. — окт. — т. 143. — с. 138—144.

522. Meta-analysis of gene expression profiles indicates genes in spliceosome pathway are up-regulated in hepatocellular carcinoma (HCC) [текст] / W. Xu [и др.] // Med Oncol. — 2015. — апр. — т. 32, № 4. — с. 96.

523. Functional analysis reveals that RBM10 mutations contribute to lung adenocarcinoma pathogenesis by deregulating splicing [текст] / J. Zhao [и др.] // Sci Rep. — 2017. — янв. — т. 7. — с. 40488.

524. NMD abnormalities during brain development in the Fmr1-knockout mouse model of fragile X syndrome [текст] / T. Kurosaki [и др.] // Genome Biol. — 2021. — нояб. — т. 22, № 1. — с. 317.

525. *Xu, Q.* Genome-wide detection of tissue-specific alternative splicing in the human transcriptome [текст] / Q. Xu, B. Modrek, C. Lee // Nucleic Acids Res. — 2002. — сент. — т. 30, № 17. — с. 3754—3766.

526. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons [текст] / P. L. Boutz [и др.] // Genes Dev. — 2007. — июль. — т. 21, № 13. — с. 1636—1652.

527. Mice lacking doublecortin and doublecortin-like kinase 2 display altered hippocampal neuronal maturation and spontaneous seizures [текст] / G. Kerjan [и др.] // Proc Natl Acad Sci U S A. — 2009. — апр. — т. 106, № 16. — с. 6766—6771.

528. Doublecortin-like kinase enhances dendritic remodelling and negatively regulates synapse maturation [текст] / E. Shin [и др.] // Nat Commun. — 2013. — т. 4. — с. 1440.

529. IQGAP1 and IGFBP2: valuable biomarkers for determining prognosis in glioma patients [текст] / K. L. McDonald [и др.] // J Neuropathol Exp Neurol. — 2007. — май. — т. 66, № 5. — с. 405—417.

530. *Jones, M. H.* Identification and characterization of BRDT: A testis-specific gene related to the bromodomain genes RING3 and Drosophila fsh [текст] / M. H. Jones, M. Numata, M. Shimane // Genomics. — 1997. — нояб. — т. 45, № 3. — с. 529—534.

531. The yeast BDF1 gene encodes a transcription factor involved in the expression of a broad class of genes including snRNAs [текст] / Z. Lygerou [и др.] // Nucleic Acids Res. — 1994. — дек. — т. 22, № 24. — с. 5332—5340.

532. A bromodomain protein, MCAP, associates with mitotic chromosomes and affects G(2)-to-M transition [текст] / A. Dey [и др.] // Mol Cell Biol. — 2000. — сент. — т. 20, № 17. — с. 6537—6549.

533. The BET family in immunity and disease [текст] / N. Wang [и др.] // Signal Transduct Target Ther. — 2021. — янв. — т. 6, № 1. — с. 23.

534. Comparative structure-function analysis of bromodomain and extraterminal motif (BET) proteins in a gene-complementation system [текст] / M. T. Werner [и др.] // J Biol Chem. — 2020. — февр. — т. 295, № 7. — с. 1898—1914.

535. *Sanchez, R.* The role of human bromodomains in chromatin biology and gene transcription [текст] / R. Sanchez, M.-M. Zhou // Curr Opin Drug Discov Devel. — 2009. — сент. — т. 12, № 5. — с. 659—665.

536. Bromodomain testis-specific protein is expressed in mouse oocyte and evolves faster than its ubiquitously expressed paralogs BRD2, -3, and -4 [текст] / A. Paillisson [и др.] // Genomics. — 2007. — февр. — т. 89, № 2. — с. 215—223.

537. The bromodomain-containing gene BRD2 is regulated at transcription, splicing, and translation levels [текст] / E. Shang [и др.] // J Cell Biochem. — 2011. — окт. — т. 112, № 10. — с. 2784—2793.

538. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses [текст] / J. Huerta-Cepas [и др.] // Nucleic Acids Res. — 2019. — янв. — т. 47, № D1. — с. D309—D314.

539. *Belkina, A. C.* BET domain co-regulators in obesity, inflammation and cancer [текст] / A. C. Belkina, G. V. Denis // Nat Rev Cancer. — 2012. — июнь. — т. 12, № 7. — с. 465—477.

540. Chromosomal localization, gene structure and transcription pattern of the ORFX gene, a homologue of the MHC-linked RING3 gene [текст] / K. L. Thorpe [и др.] // Gene. — 1997. — окт. — т. 200, № 1/2. — с. 177—183.

541. Gene Signature Associated With Bromodomain Genes Predicts the Prognosis of Kidney Renal Clear Cell Carcinoma [текст] / J. Lu [и др.] // Front Genet. — 2021. — т. 12. — с. 643935.

542. *Hermsen, R.* Combinatorial gene regulation using auto-regulation [текст] / R. Hermsen, B. Ursem, P. R. ten Wolde // PLoS Comput Biol. — 2010. — июнь. — т. 6, № 6. — e1000813.

543. *Hanamura, A.* Molecular mechanism of negative autoregulation of Escherichia coli crp gene [текст] / A. Hanamura, H. Aiba // Nucleic Acids Res. — 1991. — авг. — т. 19, № 16. — с. 4413—4419.

544. *Bateman, E.* Autoregulation of eukaryotic transcription factors [текст] / E. Bateman // Prog Nucleic Acid Res Mol Biol. — 1998. — т. 60. — с. 133—168.

545. Interplay between Y-box-binding protein 1 (YB-1) and poly(A) binding protein (PABP) in specific regulation of YB-1 mRNA translation [текст] / D. N. Lyabin [и др.] // RNA Biol. — 2011. — т. 8, № 5. — с. 883—892.

546. *Schoenberg, D. R.* Regulation of cytoplasmic mRNA decay [текст] / D. R. Schoenberg, L. E. Maquat // Nat Rev Genet. — 2012. — март. — т. 13, № 4. — с. 246—259.

547. *Mitrovich, Q. M.* Unproductively spliced ribosomal protein mRNAs are natural targets of mRNA surveillance in C. elegans [текст] / Q. M. Mitrovich, P. Anderson // Genes Dev. — 2000. — сент. — т. 14, № 17. — с. 2173—2184.

548. Genome-wide identification of alternative splice forms down-regulated by nonsense-mediated mRNA decay in Drosophila [текст] / K. D. Hansen [и др.] // PLoS Genet. — 2009. — июнь. — т. 5, № 6. — e1000525.

549. *Wight, M.* The functions of natural antisense transcripts [текст] / M. Wight, A. Werner // Essays Biochem. — 2013. — т. 54. — с. 91—101.

550. Genome-wide survey of tissue-specific microRNA and transcription factor regulatory networks in 12 tissues [текст] / Z. Guo [и др.] // Sci Rep. — 2014. — июнь. — т. 4. — с. 5150.

551. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes [текст] / T. Watanabe [и др.] // Nature. — 2008. — май. — т. 453, № 7194. — с. 539—543.

552. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes [текст] / O. H. Tam [и др.] // Nature. — 2008. — май. — т. 453, № 7194. — с. 534—538.

553. Dynamic m6A modification regulates local translation of mRNA in axons [текст] / J. Yu [и др.] // Nucleic Acids Res. — 2018. — февр. — т. 46, № 3. — с. 1412—1423.

554. A mRNA modifications are deposited in nascent pre-mRNA and are not required for splicing but do specify cytoplasmic turnover [текст] / S. Ke [и др.] // Genes Dev. — 2017. — май. — т. 31, № 10. — с. 990—1006.

555. Inosine induces context-dependent recoding and translational stalling [текст] / K. Licht [и др.] // Nucleic Acids Res. — 2019. — янв. — т. 47, № 1. — с. 3—14.

556. *Mudge, J. M.* Functional transcriptomics in the post-ENCODE era [текст] / J. M. Mudge, A. Frankish, J. Harrow // Genome Res. — 2013. — дек. — т. 23, № 12. — с. 1961—1973.

557. Paralogous Hox genes: function and regulation [текст] / M. Maconochie [и др.] // Annu Rev Genet. — 1996. — т. 30. — с. 529—556.

558. The C-terminal domain of Brd2 is important for chromatin interaction and regulation of transcription and alternative splicing [текст] / J. Hnilicová [и др.] // Mol Biol Cell. — 2013. — нояб. — т. 24, № 22. — с. 3557—3568.

559. A variable dinucleotide repeat in the CFTR gene contributes to phenotype diversity by forming RNA secondary structures that alter splicing [текст] / T. W. Hefferon [и др.] // Proc Natl Acad Sci U S A. — 2004. — март. — т. 101, № 10. — с. 3504—3509.

560. *Sarnik, J.* BET Proteins as Attractive Targets for Cancer Therapeutics [текст] / J. Sarnik, T. Popławski, P. Tokarz // Int J Mol Sci. — 2021. — окт. — т. 22, № 20.

561. *Wu, D.* Roles of Bromodomain Extra Terminal Proteins in Metabolic Signaling and Diseases [текст] / D. Wu, Q. Duan // Pharmaceuticals (Basel). — 2022. — авг. — т. 15, № 8.

562. *Borck, P. C.* BET Epigenetic Reader Proteins in Cardiovascular Transcriptional Programs [текст] / P. C. Borck, L.-W. Guo, J. Plutzky // Circ Res. — 2020. — апр. — т. 126, № 9. — с. 1190—1208.

563. *Zavileyskiy, L. G.* Post-transcriptional Regulation of Gene Expression via Unproductive Splicing [текст] / L. G. Zavileyskiy, D. D. Pervouchine // Acta Naturae. — 2024. — т. 16, № 1. — с. 4—13. — (1.16 п. л.; Вклад автора 50%; JIF=2.0 WoS).

564. Functional identification of cis-regulatory long noncoding RNAs at controlled false discovery rates [текст] / B. Dhaka [и др.] // Nucleic Acids Research. — 2024. — апр. — т. 52, № 6. — с. 2821—2835. — (1.73 п. л.; Вклад автора 10%; JIF=14.9 WoS).

565. *Vorobeva, M. A.* Cooperation and Competition of RNA Secondary Structure and RNA-Protein Interactions in the Regulation of Alternative Splicing [текст] / M. A. Vorobeva, D. A. Skvortsov, D. D. Pervouchine // Acta Naturae. — 2023. — т. 15, № 4. — с. 23—31. — (1.04 п. л.; Вклад автора 40%; JIF=2.0 WoS).

566. Transcriptome analysis reveals high tumor heterogeneity with respect to re-activation of stemness and proliferation programs [текст] / A. Baranovsky [и др.] // PLoS One. — 2022. — т. 17, № 5. — e0268626. — (2.66 п. л.; Вклад автора 50%; JIF=3.7 WoS).

567. *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. — (1.04 п. л.; Вклад автора 75%; JIF=2.0 WoS).

568. An extended catalogue of tandem alternative splice sites in human tissue transcriptomes [текст] / A. Mironov [и др.] // PLoS Computational Biology. — 2021. — апр. — т. 17, № 4. — e1008329. — (3.47 п. л.; Вклад автора 40%; JIF=4.3 WoS).

569. A limited set of transcriptional programs define major cell types [текст] / A. Breschi [и др.] // Genome Research. — 2020. — июль. — т. 30, № 7. — с. 1047—1059. — (1.50 п. л.; Вклад автора 10%; JIF=7.0 WoS).

570. Perspectives on ENCODE [текст] / ENCODE Project Consortium [и др.] // Nature. — 2020. — июль. — т. 583, № 7818. — с. 693—698. — (0.69 п. л.; Работа в составе консорциума. Вклад автора менее 5%; JIF=64.8 WoS).

571. A benchmark for RNA-seq quantification pipelines [текст] / M. Teng [и др.] // Genome Biology. — 2016. — апр. — т. 17. — с. 74. — (1.39 п. л.; Вклад автора 10%; JIF=12.3 WoS).

572. 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, № 8. — S2. — (1.27 п. л.; Вклад автора 10%; JIF=4.4 WoS).

573. RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA [текст] / L. V. Danilova [и др.] // Journal of Bioinformatics and Computational Biology. — 2006. — апр. — т. 4, № 2. — с. 589—596. — (0.92 п. л.; Вклад автора 25%; JIF=1.0 WoS).

574. Engineered riboregulators enable post-transcriptional control of gene expression [текст] / F. J. Isaacs [и др.] // Nature Biotechnology. — 2004. — июль. — т. 22, № 7. — с. 841—847. — (0.92 п. л.; Вклад автора 25%; JIF=46.9 WoS).

575. *Pervouchine, D. D.* On the normalization of RNA equilibrium free energy to the length of the sequence [текст] / D. D. Pervouchine, J. H. Graber, S. Kasif // Nucleic Acids Research. — 2003. — май. — т. 31, № 9. — e49. — (0.69 п. л.; Вклад автора 90%; JIF=14.9 WoS).

Список рисунков

Список таблиц