МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
имени М.В. ЛОМОНОСОВА

На правах рукописи

Первушин Дмитрий Давидович

Альтернативный сплайсинг и дальние взаимодействия в структуре эукариотических РНК

1.5.3. Молекулярная биология

Автореферат
диссертации на соискание учёной степени
доктора химических наук

Москва — 2024

Работа выполнена в Центре молекулярной и клеточной биологии автономной некоммерческой образовательной организации высшего образования «Сколковский Институт Науки и Технологии» (Сколтех).

Официальные оппоненты: **Юсупов Марат Миратович,** доктор химических наук, Федеральное государственное бюджетное учреждение науки Федеральный исследовательский центр «Казанский научный центр РАН», ведущий научный сотрудник лаборатории структурного анализа биомакромолекул

Шайтан Алексей Константинович, доктор физико-математических наук, член-корреспондент РАН, Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», биологический факультет, профессор кафедры биоинженерии

Кулаковский Иван Владимирович, доктор биологических наук, Федеральное государственное бюджетное учреждение науки «Институт белка Российской академии наук», ведущий научный сотрудник

Защита состоится 24 сентября 2024 г. в 16:00 часов на заседании диссертационного совета МГУ.014.2 Московского государственного университета имени М. В. Ломоносова по адресу: 119991, Москва, Ленинские горы д. 1, стр. 40, НИИФХБ, ауд. 501.

E-mail: agapkina@belozersky.msu.ru.

С диссертацией можно ознакомиться в отделе диссертаций научной библиотеки МГУ имени М. В. Ломоносова (Ломоносовский просп., д. 27) и на портале: https://dissovet.msu.ru/dissertation/3082.

Автореферат разослан ___ июля 2024 года.

Ученый секретарь
диссертационного совета МГУ.014.2,
кандидат химических наук Агапкина Ю. Ю.

Общая характеристика работы

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.

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 котранскрипционно, 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.

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 их результатов. 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]. 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

[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; *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.

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[2]. 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.[3]. 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[4]. 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

[2]hnRNP proteins and the biogenesis of mRNA [текст] / G. Dreyfuss [и др.] // Annu Rev Biochem. 1993. т. 62. с. 289—321.

[3]*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.

[4]Long-range RNA pairings contribute to mutually exclusive splicing [текст] / Y. Yue [и др.] // RNA. 2016. янв. т. 22, № 1. с. 96—110; 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.

также роли дальних взаимодействий в транс-сплайсинге появлялись в литературе[5], 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[6]. 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[7]. 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[8].

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[9]. 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

---

[5]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; *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.

[6]*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.

[7]*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; 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.

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

[9]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; *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.

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[10]. 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[11]. 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[12]. 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[13]. 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[14]. 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

[10] *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.

[11] 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.

[12] *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.

[13] 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; *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.

[14] 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.

ансамблей структур[15]. Таким образом, изучение 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.

[15]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.

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[16] 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

[16] Представлены в приложениях к [1—3].

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.
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 [A1—A4], 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.

Содержание работы

Во **введении** излагаются основные определения, описывается проблематика исследования структуры эукариотических РНК, обосновывается актуальность темы, формулируются цели и задачи работы и перечисляются основные положения, выносимые на защиту. Обсуждаются научная новизна, теоретическая и практическая значимость, объекты и методы исследования, степень достоверности результатов, апробация результатов и личный вклад автора. Описывается структура работы и приводится количество публикаций.

Первая глава посвящена обзору литературы по изучаемой проблеме. Обсуждается альтернативный сплайсинг (АС), его регуляция РНК-связывающими белками (РСБ) и структурой РНК. Приводятся примеры регуляторных механизмов, таких как блокировка, сближение и отдаление цис-регуляторных элементов, а также совместная регуляция АС структурой РНК и РСБ [4]. Далее обсуждаются система нонсенс-опосредованного распада и механизм пост-транскрипционного контроля экспрессии генов через непродуктивный сплайсинг (НС). Рассматриваются ауто- и кросс-регуляторные механизмы с участием активаторов и репрессоров сплайсинга и приводятся примеры. Затем описываются биологические функции и роль АС в заболеваниях человека, а также стратегии модуляции АС антисмысловыми олигонуклеотидами [5; 6].

Далее обсуждаются экспериментальные и вычислительные методы предсказания структуры РНК, объясняются их недостатки и ограничения в контексте проблемы исследования дальних взаимодействий. Выделяются две основные группы экспериментальных методов, которые различаются типами получаемой структурной информации: методы, основанные на футпринтинге, и методы, основанные на лигировании пространственно близких молекул. Вычислительные методы предсказания структуры РНК подразделяются, с одной стороны, на термодинамические (основанные на минимизации свободной энергии) и филогенетические (основанные на

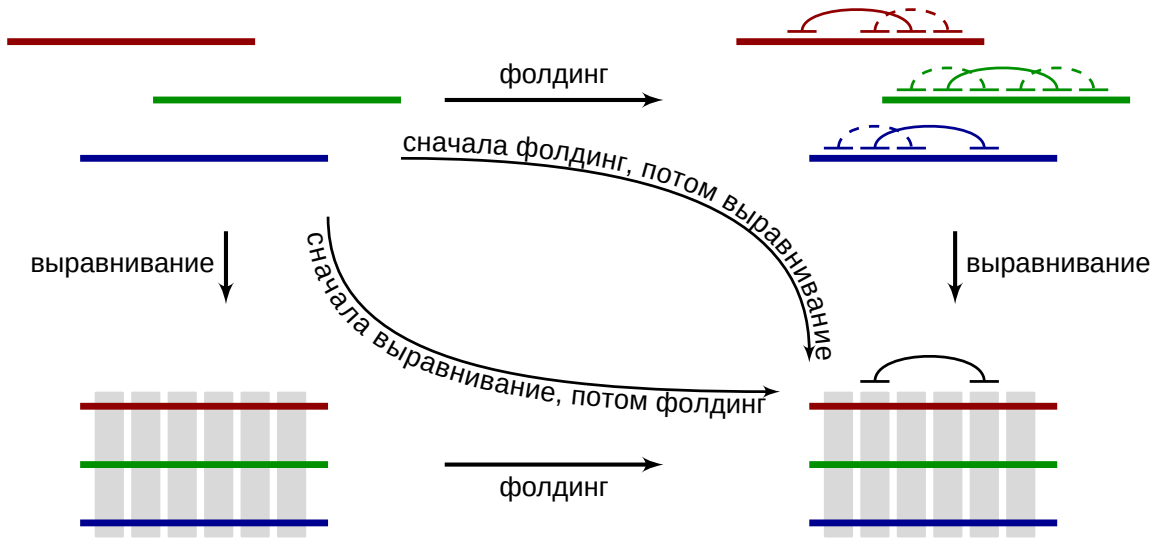


Рис. 1 — Диаграмма, описывающая одновременное выравнивание последовательностей и предсказание структуры. Вверху слева: невыровненные последовательности РНК. Внизу слева: их выравнивание без учета структуры; серым цветом показаны консервативные участки. Вверху справа: предсказанные элементы вторичной структуры для каждой последовательности показаны в виде дуг. Внизу справа: структуры сопоставляются либо в результате согласованного выравнивания, либо идентифицируются непосредственно в множественном выравнивании последовательностей.

поиске ковариаций) и, с другой стороны, на методы предсказания внутримолекулярных и межмолекулярных РНК-структур. Описывается основная проблема динамического программирования, которое за приемлемое вычислительное время может предсказать только структуры без псевдоузлов, что делает его неприменимым к предсказанию дальних взаимодействий.

Затем обсуждается алгоритм Санкова, который соединяет в себе термодинамический и филогенетический подходы для одновременного построения множественного выравнивания и предсказания структуры РНК[17]. 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, что выходит далеко за рамки современных вычислительных возможностей. Кроме того, алгоритм Санкова основан на динамическом программировании, поэтому его можно применить к поиску эукариотических РНК-структур только интерпретируя дальние взаимодействия как межмолекулярные, что еще больше усложняет задачу.

Далее рассматриваются две предельные реализации алгоритма Санкова, которые можно назвать «сначала выравнивание, потом фолдинг» и «сначала фолдинг, потом выравнивание» (рис. 1) [7]. В первом случае набор ортологичных последовательностей сначала выравнивается, а затем по полученному выравниванию предсказывается структура РНК. Несмотря

[17] *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.

на ограничения, связанные с качеством входного выравнивания, «сначала выравнивание, потом фолдинг» представляет собой наиболее простой, быстрый и мощный подход, который используется во многих современных сравнительных методах[18]. 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, на первый взгляд представляется нерациональной или даже невозможной, поскольку число структур для одной последовательности, возведенное в степень числа их комбинаций при множественном выравнивании, слишком велико. Утверждается, что ее можно реализовать благодаря существенному сокращению числа структур за счет рассмотрения только очень длинных и почти идеально комплементарных консервативных спариваний [8].

Во **второй главе** кратко перечисляются общие для всех глав материалы и методы. Перечисляются данные высокопроизводительного секвенирования, в том числе полученные при участии автора диссертации [A1; A3; A5; 9—11; A6], биоинформатические [A4; 12; 13], статистические и экспериментальные методы, такие как оценка уровней экспрессии генов и уровней включения экзонов, конструирование минигенов, трансфекция плазмидами и антисенс олигонуклеотидами, замедление элонгации транскрипции и др. Методы, разработанные в диссертации, излагаются в соответствующих главах и публикациях.

В **третьей главе** разрабатываются методы предсказания дальних взаимодействий в структуре РНК. Первой обсуждается стратегия «сначала фолдинг, потом выравнивание», называемая IRBIS, которая 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 комплементарности и с хэш-таблицами ортологов для обнаружения консервативности [8]. Ее преимущество заключается в том, что неконсервативные участки не приходится выравнивать. Входные данные состоят из набора невыровненных ортологичных сегментов последовательностей, например, интронов, соответствие между которыми устанавливается из соображений синтении. По построению не накладывается никаких ограничений ни на расстояние между парами оснований, ни на псевдоузлы.

Ключевым шагом этого метода является процедура предварительной фильтрации, называемая триммингом. Она использует тот факт, что хеш таблица, в которой хранятся позиции k-меров для каждого вида i, является упорядоченным массивом, а ее элементы можно за линейное время

[18]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; 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.

сравнить для всех i для того, чтобы заранее исключить из рассмотрения неконсервативные k-меры. Модификация этой процедуры позволяет учесть небольшое число G:U пар, а использование k-меров с пробелами позволяет рассматривать РНК-структуры с короткими внутренними петлями.

Затем оцениваются чувствительность и доля ложноположительных предсказаний. Показывается, что IRBIS имеет высокую чувствительность, но при этом высокую (не менее 15%) долю ложноположительных предсказаний. При жестких ограничениях на консервативность и длину спариваемой области метод применяется к интронам белок-кодирующих генов млекопитающих и дрозофил, в результате чего находятся 832 и 632 пары консервативных комплементарных участков (ККУ), соответственно, по одной структуре для каждой пары сегментов, и описываются их характеристики.

Общая тенденция в расположении найденных ККУ заключается в том, что они расположены неслучайно по отношению к сплайс-сайтам и событиям АС. У дрозофил часто наблюдаются структуры, выпетливающие кассетные экзоны, а среди альтернативно сплайсируемых интронов наблюдается обогащение интронами, которые содержат альтернативные акцепторные сайты [14]. Внутри области поиска (интронное окно 150 нт) ККУ также расположены не случайно, и предпочитают находиться на расстоянии от экзонов, избегая пересечений с полипиримидиновым трактом и сайтом ветвления. Тенденции в расположении интронных структур РНК у млекопитающих сходны с таковыми у дрозофил [15].

Далее приводится пример предсказания структуры РНК в гене дистонина человека. Экзоны 47–52 этого гена либо одновременно включаются, либо одновременно пропускаются. Они выпетливаются парой ККУ, которые располагаются во фланкирующих интронах на расстоянии около 10000 п.о. друг от друга. Однако, как и во многих других случаях, почти полное отсутствие в ККУ компенсаторных замен не позволяет установить наличие отбора на поддержание комплементарности.

Поскольку длинные некодирующие РНК могут регулировать экспрессию генов [16], метод IRBIS применяется к предсказанию РНК-РНК взаимодействий и приводится пример ложного предсказания комплементарности между длинной некодирующей РНК *RP11-439A17.4* и более чем 20 генами гистонов млекопитающих. Ген *RP11-439A17.4* находится в антисмысловой ориентации по отношению к соседнему гену *HIST2H2BA* и содержит в себе сайт связывания транскрипционного фактора MEF-2A, который также содержится во всех гистоновых генах, но на противоположной цепи. Эти сайты образуют пару «кажущихся» ККУ, которые не обусловлены РНК-РНК взаимодействиями.

Далее обсуждается стратегия «сначала выравнивание, потом фолдинг», в которой разреженное динамическое программирование (метод

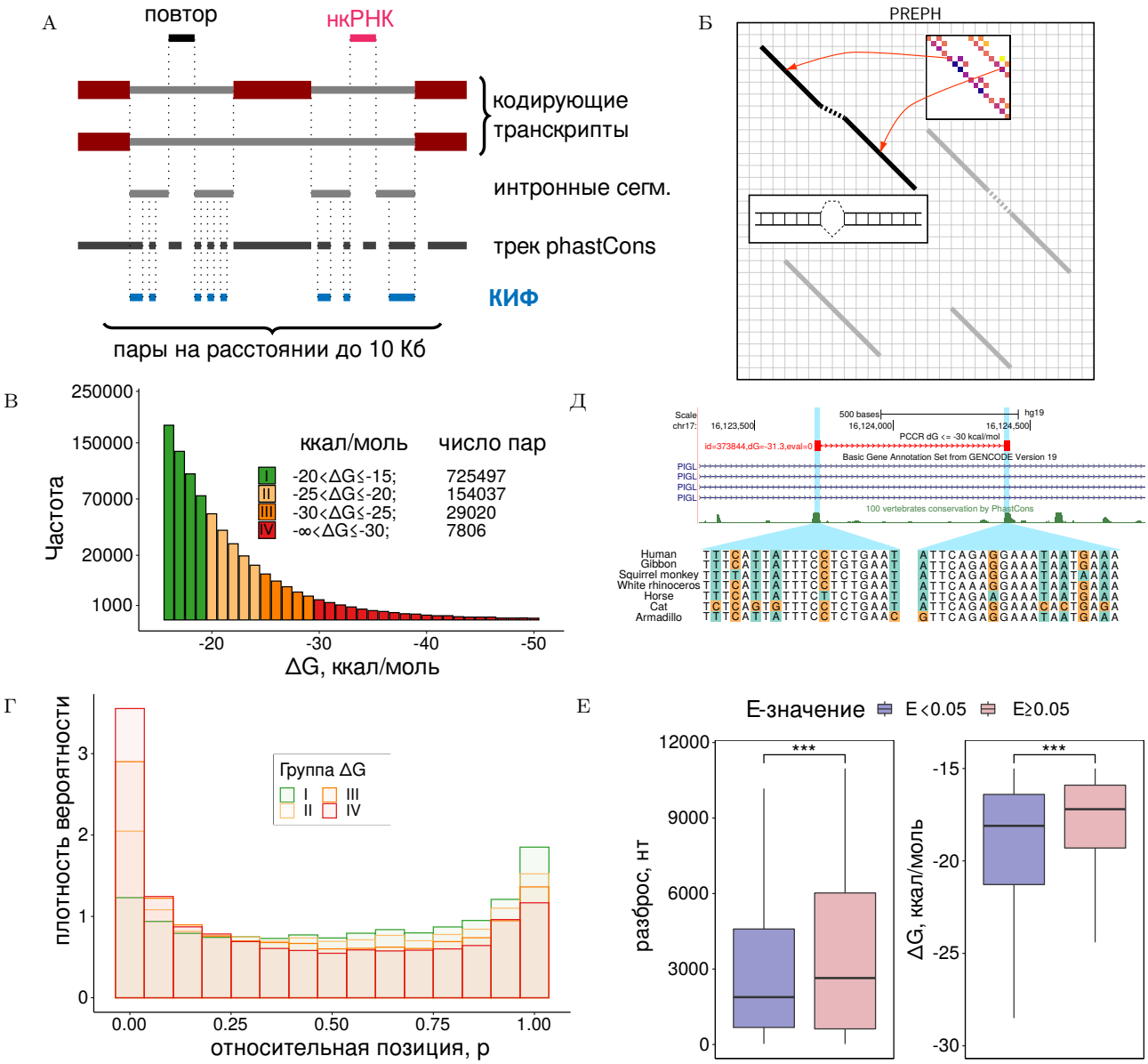


Рис. 2 — Характеристики консервативных комплементарных участков (ККУ). **(А)** Пары ККУ разыскиваются в консервативных интронных фрагментах (КИФ) на расстоянии не более 10000 нт друг от друга. **(Б)** Идея метода PREPH, использующего предварительно вычисленные энергии спаривания для всех k-меров (вставка). **(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 нуклеотидными ковариациями ($E < 0.05$, $n = 3204$) 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%.

PREPH) применяется к предсказанию комплементарности между участками интронов, рассматриваемыми как межмолекулярные РНК-РНК взаимодействия [1; 17; A7]. Этот подход позволяет охватить большее число структур и обойти ограничения первого метода путем исследования заранее вычисленных консервативных интронных фрагментов (КИФ) (рис. 2А). Во всех парных комбинациях КИФ, расположенных в не более чем $L = 10000$ нт друг от друга и принадлежащих одному и тому же гену, было найдено 916360 пар ККУ с длиной не менее 10 нт и свободной энергией $\Delta G \leqslant -15$ ккал/моль. Пары ККУ подразделяются на четыре группы по энергии (рис. 2В). Распределение значений метрики, оценивающей относительное положение пары ККУ внутри гена, имеет две выраженные моды на 5'- и 3'-конце (рис. 2Г).

Для исследования компенсаторных мутаций к фрагментам множественных выравниваний геномов позвоночных, соответствующих ККУ, применялся метод R-scape[19]. Отклонение от нулевой гипотезы о том, что парные ковариации в паре ККУ не обусловлены отбором на структуру РНК, оценивалось с помощью E-значений. Только 3204 пары ККУ имели E-значение менее 5%, причем они оказались в среднем более стабильными и имеющими меньший разброс[20] (рис. 2Е). В некоторых случаях структурное выравнивание поддерживалось ковариациями, как, например, спаривание, охватывающее 700 нт в гене *PIGL* (рис. 2Д). Однако E-значения известных из литературы структур были близки к единице, т.е. вариабельность нуклеотидных последовательностей ККУ в большинстве случаев недостаточна для оценки статистической значимости компенсаторных замен.

Затем предсказания сравнивались с экспериментальными данными, такими как данные о реактивности нуклеотидов по методу icSHAPE, данные псораленового анализа взаимодействий в структуре РНК и данные конформационного секвенирования РНК *in situ* (RIC-seq)[21]. Наилучшие показатели согласованности достигались по сравнению с RIC-seq[22]. Для оценки доли ложноположительных предсказаний использовалась процедура «пересоединения», в которой PREPH применялся к к химерным

[19] *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.

[20] Здесь и далее «разброс» — расстояние между ККУ в паре.

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

[22] Более подробно обсуждается в третьей главе.

последовательностям, которые были выбраны случайным образом из разных генов, но при этом имели такую же длину и динуклеотидный состав, как и исходные. Доля ложноположительных предсказаний варьировалась от 10% до 50% в зависимости от порогов на энергию структуры, разброс, GC состав и E-значение.

Затем исследовалась взаимосвязь между ККУ и цис-регуляторными элементами АС. Наблюдалось предпочтительное расположение ККУ внутри интронов, избегание выпетливания экзонов, уменьшение степени включения выпетливаемых экзонов с увеличением энергии структуры, избегание пересечения с активными сайтами сплайсинга, обогащение неактивными и криптическими сайтами сплайсинга, а также взаимосвязь с образованием кольцевых РНК [18; 19]. Обогащение сайтами редактирования РНК в ККУ значительно усиливается с ростом энергии структуры, что может рассматриваться как дополнительное подтверждение двухцепочечности предсказанных РНК-структур.

Далее показывается, что концы транскриптов значимо обогащены как в самих ККУ, так и в интервалах между ними. Строится каталог кластеров сайтов полиаденилирования по данным секвенирования РНК из консорциума GTEx, основанный на исследовании чтений, содержащих фрагменты поли(А)-хвоста мРНК [11]. С учетом нормировки показывается, что частота событий полиаденилирования в интронах в несколько раз больше, 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.

В связи с этим высказывается предположение о том, что структура РНК может подавлять преждевременное интронное полиаденилирование. А именно, его можно предотвратить путем котранскрипционного вырезания интрона, в то время как структура стабилизирует пре-мРНК посредством внутримолекулярных взаимодействий несмотря на разрезание основной цепи (рис. 3А). Этого не должно происходить в неструктурированных РНК, если сплайсинг происходит с задержкой по отношению к полиаденилированию (рис. 3Б). Обнаружение интронов, называемых сплайсированными полиаденилированными интронами, также подтверждает эту гипотезу [20].

Исследование взаимосвязи между ККУ и сайтами связывания РСБ, найденными по протоколу eCLIP[23], показало наличие предпочтений у некоторых факторов связываться с двухцепочечными участками. Одной из особенностей протокола eCLIP является то, что в процессе сшивания РСБ может оказаться связан с любой из двух цепей РНК, примыкающих к

[23] 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.

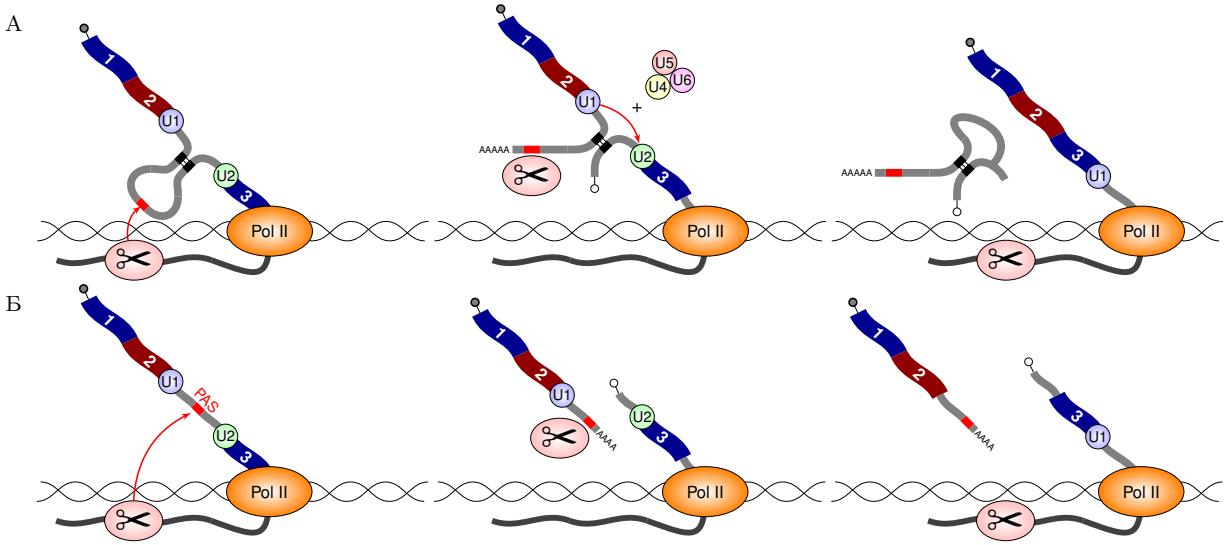


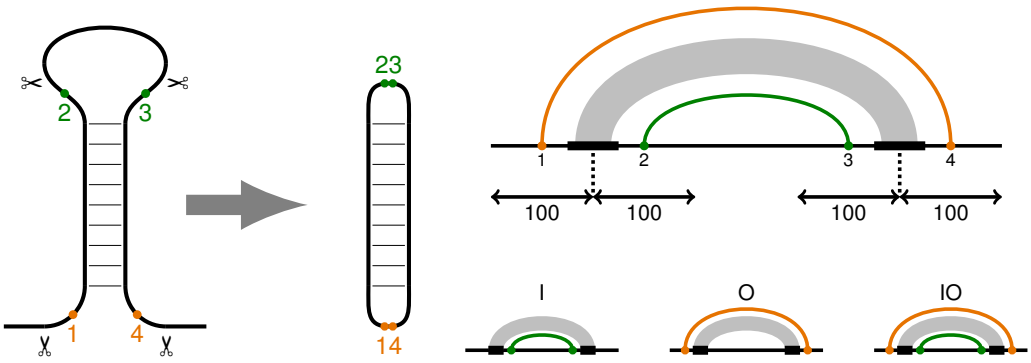
Рис. 3 — Гипотеза о котранскрипционном подавлении преждевременного полиаденилирования структурой РНК. **(А)** Если сплайсинг происходит вскоре после или одновременно с полиаденилированием, то структурированный интрон будет ко-транскрипционно вырезан несмотря на разрезание основной цепи благодаря внутримолекулярным 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, фолдинга и элонгации транскрипции. PAS — сайт полиаденилирования.

двухцепочечной области, что приводит к появлению так называемых раздвоенных сигналов eCLIP. Оказалось, что 64 из 74 исследованных РСБ имеют повышенную частоту раздвоенных сигналов eCLIP, что также подтверждает двухцепочечность ККУ.

Затем описывается эксперимент по изучению отклика транскриптома клеточной линии A549 на замедление элонгации транскрипции при помощи α-аманитина. Показывается, что при замедлении РНК-полимеразы (RNAPII) степени включения экзонов, следующих за короткими интронами, увеличиваются, а степени включения экзонов, следующих за длинными интронами, уменьшаются. Кинетика транскрипции также оказывает существенное влияние и на сворачивание РНК [21]. При замедлении RNAPII степени включения экзонов, которые следуют за интронами с ККУ, увеличиваются больше, чем степени включения экзонов, следующих за интронами без ККУ. Из этого делается вывод о том, что медленная элонгации транскрипции не только способствует распознаванию сайтов сплайсинга[24], но также дает достаточно времени для сворачивания структуры РНК в интронах, способствуя включению следующих за ними экзонов.

В заключение приводятся примеры структур РНК в консервативных областях, описывается протокол приоритизации РНК-структур и

[24]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.



(А) (Б)

Рис. 4 — 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.

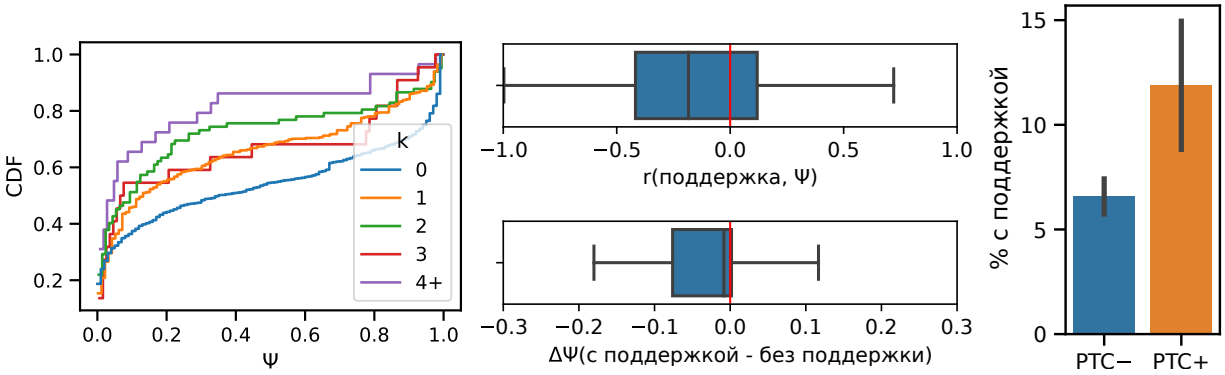
обсуждается взаимосвязь между структурой РНК, сплайсингом и полиаденилированием. Высказывается предположение о том, что эволюция позволяет создавать «разменные» сайты полиаденилирования в интронах, которые благодаря структуре РНК котранскрипционно вырезаются сплайсосомой и не вызывают преждевременной терминации транскрипции.

Четвертая глава посвящена предсказанию структуры РНК с использованием методов высокопроизводительного секвенирования. Сначала описываются метод RIC-seq и вычислительный конвейер «RNAcontacts», позволяющий картировать короткие чтения с двумя различными типами разрывов [22].

Затем предсказания PREPH сравниваются с результатами экспериментов RIC-seq в семи клеточных линиях человека [2]. Рассуждения основываются на предположении о том, что после лигирования рядом с двухцепочечными участками РНК должны возникать РНК-контакты, которые поддерживают пару ККУ с внешней или с внутренней стороны (рис. 4А). Для их нахождения вокруг центра каждого ККУ выбирается окно радиуса 100 нт, в результате чего набор пар ККУ разделяется на взаимоисключающие группы по поддержке внутри и снаружи (рис. 4Б).

Затем исследуются свойства РНК-структур в зависимости от уровня поддержки РНК-контактами. По мере увеличения k — числа клеточных линий, в которых структура поддерживалась как внутри, так и снаружи, наблюдалось значимое увеличение абсолютного значения свободной энергии, доли ККУ с сайтами редактирования РНК, доли ККУ с раздвоенными сигналами eCLIP и частот компенсаторных замен. Распределение средней степени включения (Ψ) экзонов, расположенных внутри ККУ, смещается в сторону меньших значений с увеличением k (рис. 5А).

Сравнение уровней включения экзонов, выпетливаемых ККУ, в экспериментах с поддержкой РНК-контактами и в экспериментах без поддержки



(А) **(Б)** **(В)**

Рис. 5 — Свойства экзонов, выпетливаемых ККУ с поддержкой РНК-контактами. **(А)** Распределение средней степени включения (Ψ) экзонов, выпетливаемых ККУ с поддержкой РНК-контактами 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 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+).

показывает, что разность уровней включения экзонов в клеточных линиях с поддержкой и без нее ($\Delta\Psi$) значительно смещена в сторону отрицательных значений (рис. 5Б), что позволяет предположить, что АС может регулироваться сборкой и разборкой структуры РНК в различных клеточных линиях. Выпетливаемые экзоны, содержащие преждевременные стоп-кодоны (premature termination codons, PTC), чаще поддерживается РНК-контактами (рис. 5В). Затем описывается классификатор на основе модели случайного леса, который предсказывает появление раздвоенных сигналов eCLIP в зависимости от числа контактов в окнах вокруг ККУ, и показывается, что наличие внутренних и внешних контактов в непосредственной близости от ККУ, является наиболее важным признаком для классификатора.

Затем описывается метод PHRIC, который по данным RIC-seq находит пары вложенных кластеров контактов (ВКК) на всей протяженности генов, включая экзоны и целые интроны, а затем выполняет сворачивание нуклеотидных последовательностей, заключенных между контактами [3]. Фактически он постулирует ситуацию, изображенную на рис. 5 для категории IO, для расширения предсказания структуры за пределы консервативных областей. По данным RIC-seq в семи клеточных линиях человека находятся около 29000 пар ВКК, по которым предсказываются 11998 РНК-структур с отсечением по свободной энергии $\Delta G \leqslant -15$ ккал/моль, большинство из которых (70%) располагаются вне консервативных областей. Их свойства во многом повторяют свойства ККУ, предсказанных PREPH. Показывается, что интронные структуры являются более стабильными по сравнению с экзонными и смешанными структурами, а также что

Таблица 1 — Подтверждение регуляции АС структурой РНК.

Ген	Событие АС	IRBIS	PREPH	RIC-seq	Раздел	Публ.
CG33298	Альт. 5'ss	+	+		5.1.1	[14]
Gug	Альт. 3'ss	+	+		5.1.2	[14]
Nmnat	Альт. термин. экзон	+	+		5.1.3	[14]
SF1	Удержание интрона	+	+		5.3	[15]
PHF20L1	Кассетный экзон	+	+	+	5.2.1	[2]
CASK	Кассетный экзон	+	+	+	5.2.2	[2]
ATE1	Взаимоискл. экзоны	+	+	+/−	5.2.3	[24]
MARK2	Кассетный экзон	+	+	+	5.3	[B1]
BRD2	Ядовитый экзон	+	+	+	6.3	[25]
BRD3	Ядовитый экзон	+	−	−	6.3	[25]

неконсервативные интронные РНК-структуры так же стабильны, как и консервативные.

В **Пятой главе** описывается экспериментальная валидация влияния РНК-структур на АС в нескольких генах. Обсуждаются десять примеров таких структур — три в генах дрозофилы и семь в генах человека. Список этих структур со ссылками на соответствующие разделы диссертации и публикации, а также указание методов, при помощи которых структура была предсказана и подтверждена, приводится в табл. 1. Структуры в генах *BRD2* и *BRD3* имеют отношение к непродуктивному сплайсингу и поэтому обсуждаются в шестой главе. Для валидации структур используется метод двойных мутантов, в котором фрагменты генов встраиваются в минигены с индуцируемым промотором, а также создаются несколько вариантов минигенов, в каждом из которых комплементарность нарушается, а при одновременном внесении нескольких мутаций — восстанавливается. Для проверки влияния РНК-структур на АС также используются блокирующие комплементарность антисмысловые олигонуклеотиды (АОН), изготовленные на основе LNA (locked nucleic acid) с заменой оснований ДНК в каждом втором нуклеотиде. Ранее идеи блокировки структуры РНК исследовались для бактериальных риборегуляторов [23].

Сначала описываются РНК-структуры в трех генах дрозофилы [14]. В гене *CG33298* один из ККУ перекрывается с проксимальным донорным сайтом экзона 13 (рис. 6А). Внесение дестабилизирующих точечных мутаций приводит к почти полному переключению сплайсинга на проксимальный донорный сайт, а двойная мутация, в которой структура восстанавливается, приводит к обратному переключению на дистальный донорный сайт. В двойном мутанте энергия гибридизации выше, чем в диком типе, и при этом проксимальный донорный сайт намного сильнее подавлен.

В гене атрофина (*Gug*) один из ККУ перекрывается с проксимальным акцепторным сайтом экзона 10 (рис. 6Б). В эндогенной мРНК оба акцепторных сайта используются в соотношении 1:1. При внесении точечных мутаций в ККУ сплайсинг переключается на проксимальный акцепторный

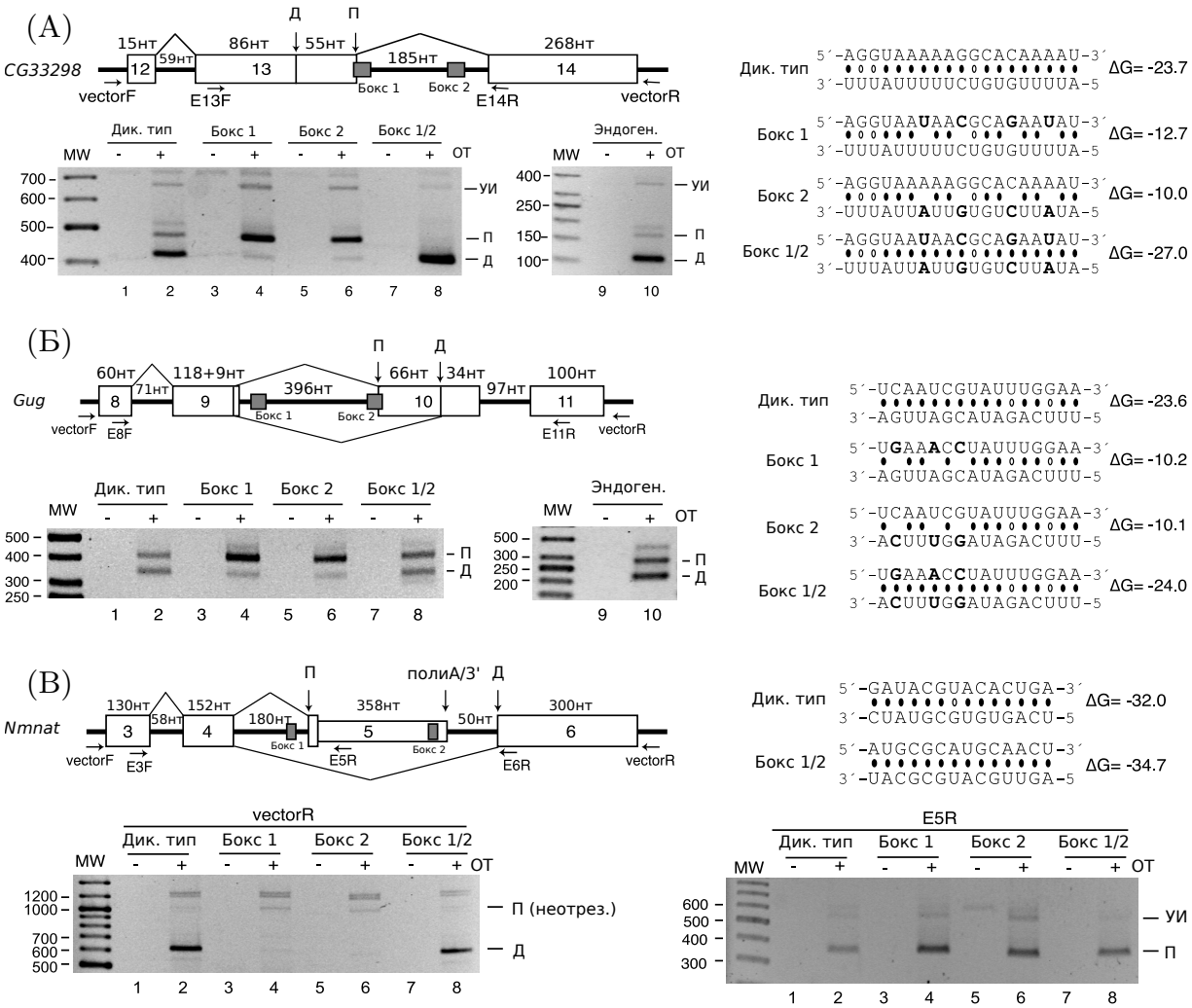


Рис. 6 — Вторичная структура РНК регулирует АС в генах *CG33298* (А), *Gug* (Б) и *Nmnat* (В) дрозофилы. На каждой панели показана схема минигена. Проксимальные и дистальные сайты сплайсинга обозначены П и Д, соответственно. События АС показаны ломаными линиями. ККУ (бокс 1 и бокс 2) обозначены серыми прямоугольниками. Справа показана схема мутагенеза (бокс 1, бокс 2 – одиночные мутанты, бокс 1/2 – двойной мутант). ΔG — свободная энергия гибридизации. Ниже показаны продукты ОТ-ПЦР транскриптов минигена и эндогенных транскриптов.

сайт, а при восстановлении структуры РНК в двойном мутанте наблюдается восстановление соотношения изоформ дикого типа.

Ген *Nmnat* содержит пару ККУ на расстоянии около 350 нт, которые окружают проксимальный акцепторный сайт в интроне 4, использование которого приводит к включению в транскрипт терминального экзона с сайтом полиаденилирования (рис. 6В). Разрушение структуры резко снижает уровень включения экзона с дистальным акцептором, а ее восстановление обращает этот эффект. Отмечается, что структура РНК влияет не только на АС, но и на альтернативное полиаденилирование.

Затем исследуются РНК-структуры в генах *PHF20L1*, *CASK* и *ATE1* человека. Все предсказанные структуры в достаточной мере поддерживаются РНК-контактами, наблюдаемыми в экспериментах RIC-seq [2].

Интроны, фланкирующие экзон 6 гена *PHF20L1*, содержат пару ККУ. С помощью ОТ-ПЦР и ОТ-ПЦР-РВ показывается, что в ответ на

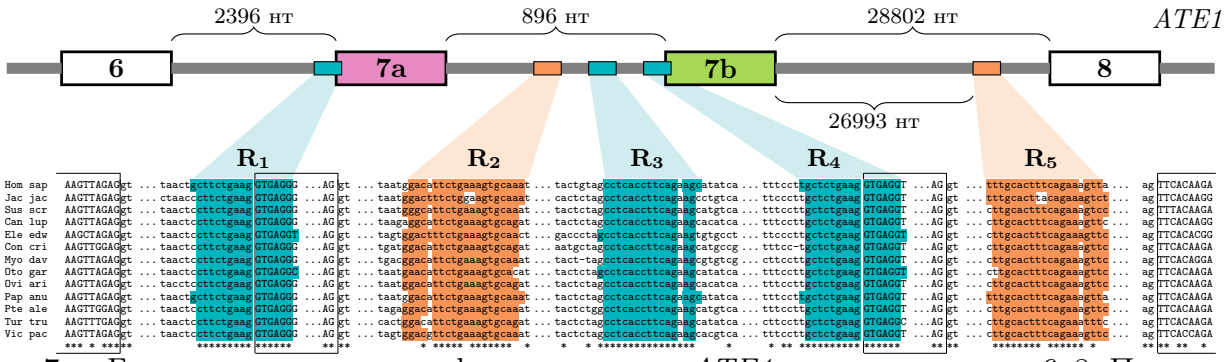


Рис. 7 — Геномная организация фрагмента гена *ATE1* между экзонами 6–8. Показаны пять эволюционно консервативных интронных элементов (R1–R5). Консервативные позиции отмечены звездочками. Участки R1 и R4 конкурируют за спаривание оснований с R3, а R2 комплементарен участку R5, который расположен на расстоянии 30000 п.о.

увеличение концентрации АОН частота включения этого экзона увеличивается. Мутации, разрушающие спаривание ККУ, способствуют включению экзона, а компенсаторные мутации возвращают соотношение изоформ к таковому у дикого типа. Аналогично показывается, что пара ККУ в фланкирующих экзон 19 гена *CASK* интронах регулирует частоту его включения. В обоих случаях механизм регуляции АС основан на выпетливании экзона.

Организация вторичной структуры РНК в гене *ATE1* намного сложнее [24]. Альтернативные сплайс-изоформы *ATE1* отличаются взаимоисключающим выбором двух соседних гомологичных экзонов (7a и 7b). С использованием данных GTEx и TCGA показывается, что экзоны 7a и 7b имеют широкий диапазон уровней включения в тканях человека с наиболее заметным отклонением в семенниках, причем одновременного включения или одновременного пропуска обоих экзонов не наблюдается. В образцах аденокарциномы простаты, а также других эпителиальных опухолей наблюдается значительное увеличение соотношения сплайс-изоформ 7a/7b.

Далее показывается, что между экзонами 6–8 этого гена расположены сразу несколько ККУ (рис. 7). Два из них, называемые R1 и R4, перекрываются с акцепторными сайтами экзонов 7a и 7b и комплементарны R3, расположенному в интроне между ними. Интрон, соединяющий экзоны 7a и 7b, содержит еще один консервативный участок R2, который комплементарен участку R5, расположенному в интроне между экзонами 7b и 8 на расстоянии около 30000 п.о. Характер комплементарности между этими участками предполагает, что R1 и R4 могут конкурировать друг с другом за спаривание с R3, а вместе со спариванием R2 с R5 они образуют псевдоузел.

Затем описываются эксперименты по сайт-направленному мутагенезу минигена, охватывающего фрагмент гена *ATE1* между экзонами 6 и 8, в котором эндогенный интрон после экзона 7b был уменьшен в размере до 2000 п.о. (рис. 8A). Для валидации конкурирующих структур РНК

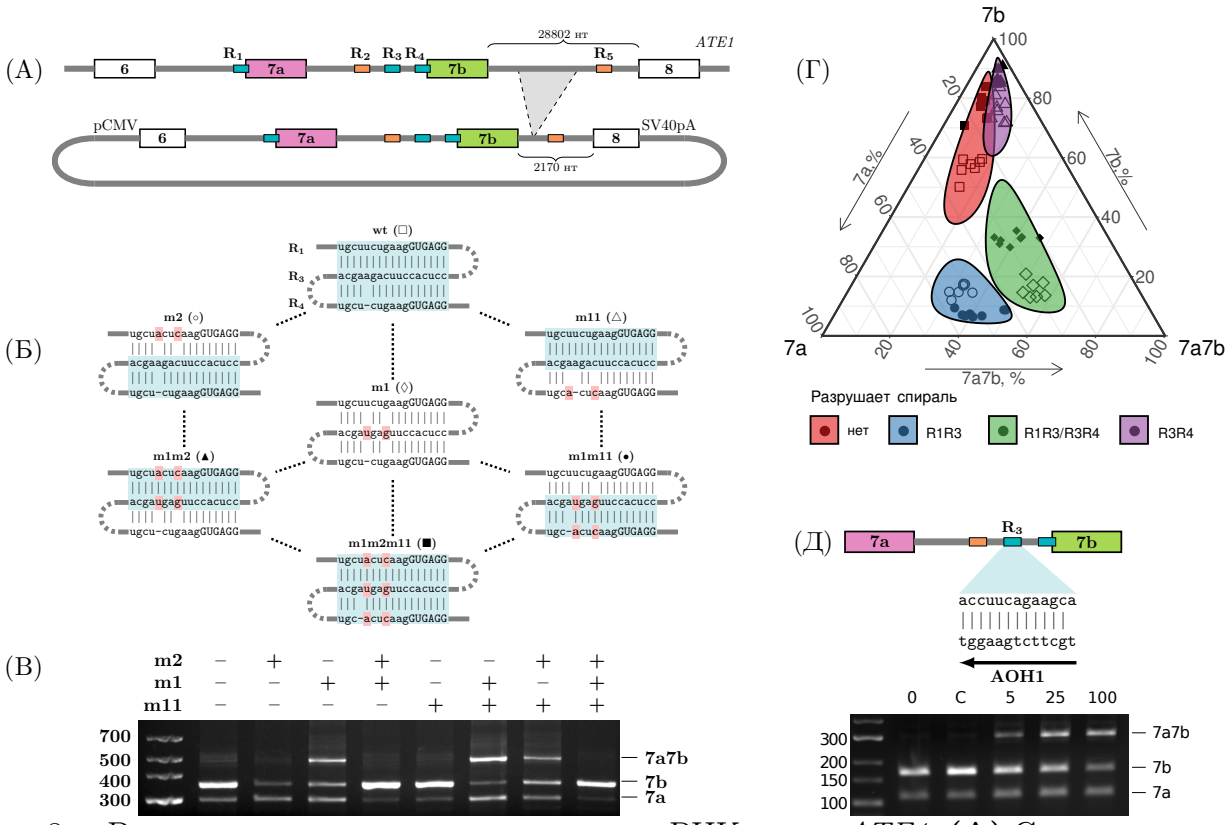


Рис. 8 — Валидация конкурирующих структур РНК в гене *ATE1*. **(А)** Схема минигена, экспрессирующего фрагмент гена *ATE1*. **(Б)** Схема мутагенеза участков R1, R3 и R4. **(В)** Оценка частоты включения экзонов 7a, 7b и двойных экзонов (7a7b) у одиночных, двойных и тройных мутантов с помощью ОТ-ПЦР. **(Г)** Тернарная диаграмма частот включения экзонов 7a, 7b и 7a7b, измеренных с помощью ОТ-ПЦР-РВ, при разрушении и восстановлении структуры R1/R3/R4. Цветные области обозначают 95% доверительные интервалы. **(Д)** В эндогенном транскрипте частота включения двойных экзонов увеличивается с увеличением концентрации АОН1 (в нМ); С — контрольный АОН.

(R1/R3/R4) используются тройные мутанты (рис. 8Б). С помощью ОТ-ПЦР и ОТ-ПЦР-РВ показывается, что одиночные мутации в R1 и R4, разрушающие спаривания R1/R3 и R3/R4, а также двойные компенсаторные мутации, восстанавливающие одно из этих спариваний, приводят к увеличению частоты включения соответствующих экзонов, а одиночная мутация в R3 приводит к увеличению доли транскриптов с двойными экзонами (рис. 8В). Соотношение сплайс-изоформ в тройном мутанте, в котором восстановлена комплементарность как R1/R3, так и R3/R4, ближе всего к соотношению изоформ в диком типе (рис. 8Г).

Поскольку в клонированном фрагменте отсутствует значительная часть интрона 7, далее с помощью АОН исследуется роль R1, R3 и R4 в регуляции сплайсинга эндогенного транскрипта *ATE1* (рис. 8Д). Обработка АОН, комплементарным R3, индуцирует включение двойных экзонов и подавляет включение отдельных экзонов, из чего делается вывод о том, что конкурирующие структуры РНК, образуемые участками R1, R3 и R4, поддерживают взаимоисключающий сплайсинг экзонов 7a и 7b.

Затем мутагенез (рис. 9А) применяется к R2 и R5, и показывается, что при раздельном введении разрушающих структуру мутаций почти

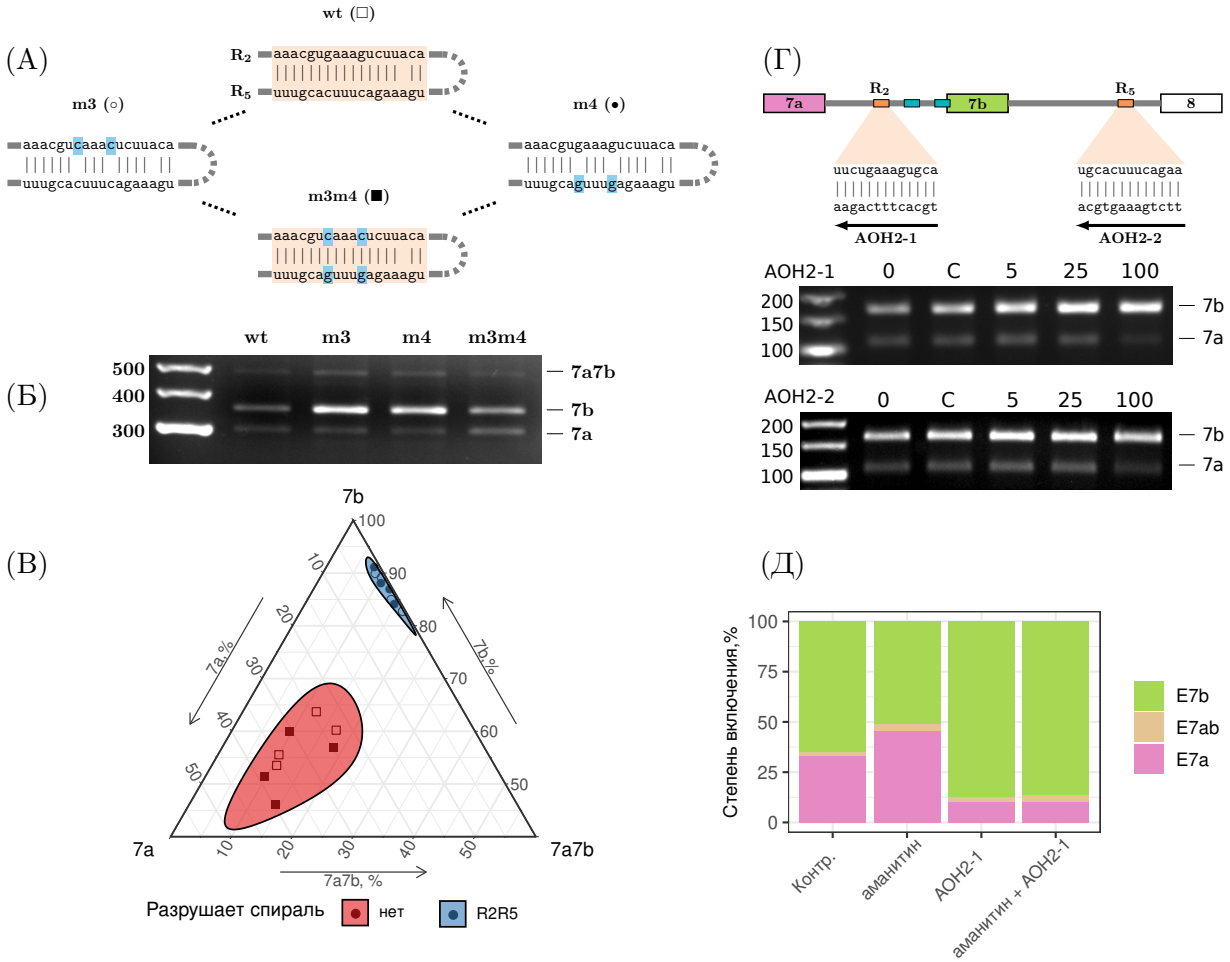


Рис. 9 — Валидация дальних взаимодействий в структуре РНК гена *ATE1*. **(А)** Схема мутагенеза участков R2 и R5. **(Б)** Оценка частоты включения экзонов 7a, 7b и двойных экзонов (7a7b) у одиночных и двойных мутантов с помощью ОТ-ПЦР. **(В)** Тернарная диаграмма, отражающая результаты ОТ-ПЦР-РВ (см пояснения к рис. 8). **(Г)** В эндогенном транскрипте увеличение концентрации AOH2-1 и AOH2-2 подавляет включение экзона 7a и способствует включению экзона 7b без появления двойных экзонов. C — контрольный AOH. **(Д)** Изменение уровня включения экзонов при совместной обработке AOH2-1, блокирующем структуру R2/R5, и α-аманитином. Добавление α-аманитина способствует включению экзона 7a только при наличии спаривания оснований R2/R5.

полностью подавляется включение экзона 7a, а компенсаторный мутант возвращает соотношение сплайс-изоформ к состоянию дикого типа без увеличения доли двойных экзонов (рис. 9Б и В). Разрушение спаривания R2/R5 в эндогенном транскрипте с помощью AOH приводит к такому же эффекту (рис. 9Г). Из этого делается заключение о том, что дальние взаимодействия между R2 и R5 регулируют соотношение сплайс-изоформ. С помощью AOH и мутагенеза исследуется взаимодействие между структурными модулями R1/R3/R4 и R2/R5 и показывается, что они независимы и функционально различны.

Затем изучается влияние скорости элонгации транскрипции на структуру РНК и АС в *ATE1*, для чего используется селективный ингибитор RNAPII α-аманитин[25]. В эндогенном транскрипте воздействие α-аманитина приводило к подавлению включения экзона 7b и усилению включения экзона 7a. В отсутствие АОН, блокирующего комплементарное спаривание R2/R5, добавление α-аманитина увеличивало уровень включения экзона 7a, а при добавлении АОН оно не приводило к подобному увеличению (рис. 9Д). Из этого делается вывод о том, что увеличение соотношения изоформ с экзонами 7a в эндогенном *ATE1* 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[26]. Субъединица NELFE, связывание которой вызывает замедление RNAPII[27], высоко экспрессируется в семенниках, где уровень включения экзона 7a также является наибольшим. Это заставляет предположить, что именно NELFE регулирует соотношение изоформ экзонов 7a/7b в семенниках, влияя на котранскрипционное сворачивание структуры R2/R5. При помощи ОТ-ПЦР-РВ показывается, что суперэкспрессия NELFE способствует включению экзона 7a и подавляет включение экзона 7b. Поскольку та же самая картина наблюдается при обработке α-аманитином, из этого делается вывод о том, что специфичное для семенников включение экзона 7a может быть вызвано экспрессией NELFE, который замедляет элонгацию транскрипции.

В заключение перечисляются предсказанные структуры РНК в других генах, влияние которых на АС было экспериментально подтверждено за рамками данной диссертации. Описывается пара ККУ, которая предотвращает удержание интрона в гене *SF1* [15], ККУ в транскриптах

---

[25] *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; *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.

[26] 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; 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.

[27] Architecture and RNA binding of the human negative elongation factor [текст] / S. M. Vos [и др.] // Elife. 2016. июнь. т. 5; 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; 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.

мышиных гомологов генов *CASK* и *PHF20L1* человека [A2], а также пара ККУ в гене *MARK2*, регулирующая частоту включения экзона 17 [B1]. Ряд полученных результатов подтвержден независимыми исследованиями других авторов[28]. В контексте регуляторных механизмов в гене *ATE1* обсуждается гипотеза о том, что тандемные дупликации, затрагивающие экзоны и части фланкирующих интронов, должны неизбежно приводить к образованию конкурирующих структур РНК и, как следствие, к взаимоисключающему типу сплайсинга [26; 27].

Регуляция экспрессии генов осуществляется множеством различных механизмов [28—31]. В последней, **шестой главе** изучается регуляция непродуктивного сплайсинга и структурой РНК. Первым излагается метод обнаружения ауторегуляторных петель непродуктивного сплайсинга с отрицательными обратными связями по данным секвенирования РНК [32]. Для нахождения непродуктивных событий используются данные по ответу транскриптома на инактивацию системы NMD[29]. Предсказание регуляторных взаимодействий основано на данных по ответу транскриптома на инактивацию РСБ и данных eCLIP, полученных в рамках международного проекта ENCODE[30].

Сначала показывается, что при инактивации системы NMD уровни включения ядовитых экзонов в среднем увеличиваются, а уровни включения необходимых (т.е. вызывающих NMD при пропуске) экзонов, наоборот, снижаются. Затем, после рассмотрения известных из литературы мишеней непродуктивного сплайсинга и оценки доли ложных предсказаний, выбирается консервативный подход, использующий в качестве фильтра величину эффекта ($\Delta\Psi$). При жестком ограничении $|\Delta\Psi| \geqslant 0.1$ находятся три кандидатных ядовитых экзона в генах *SRSF7*, *U2AF1* и *RPS3*, а также один необходимый экзон в гене *SFPQ*. Ген *SRSF7*, молекулярный механизм непродуктивного сплайсинга в котором был позднее подтвержден в работах других авторов[31], обсуждается детально.

---

[28]Long-range RNA pairings contribute to mutually exclusive splicing [текст] / Y. Yue [и др.] // RNA. 2016. янв. т. 22, № 1. с. 96—110; 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).

[29]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.

[30]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; *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.

[31]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.

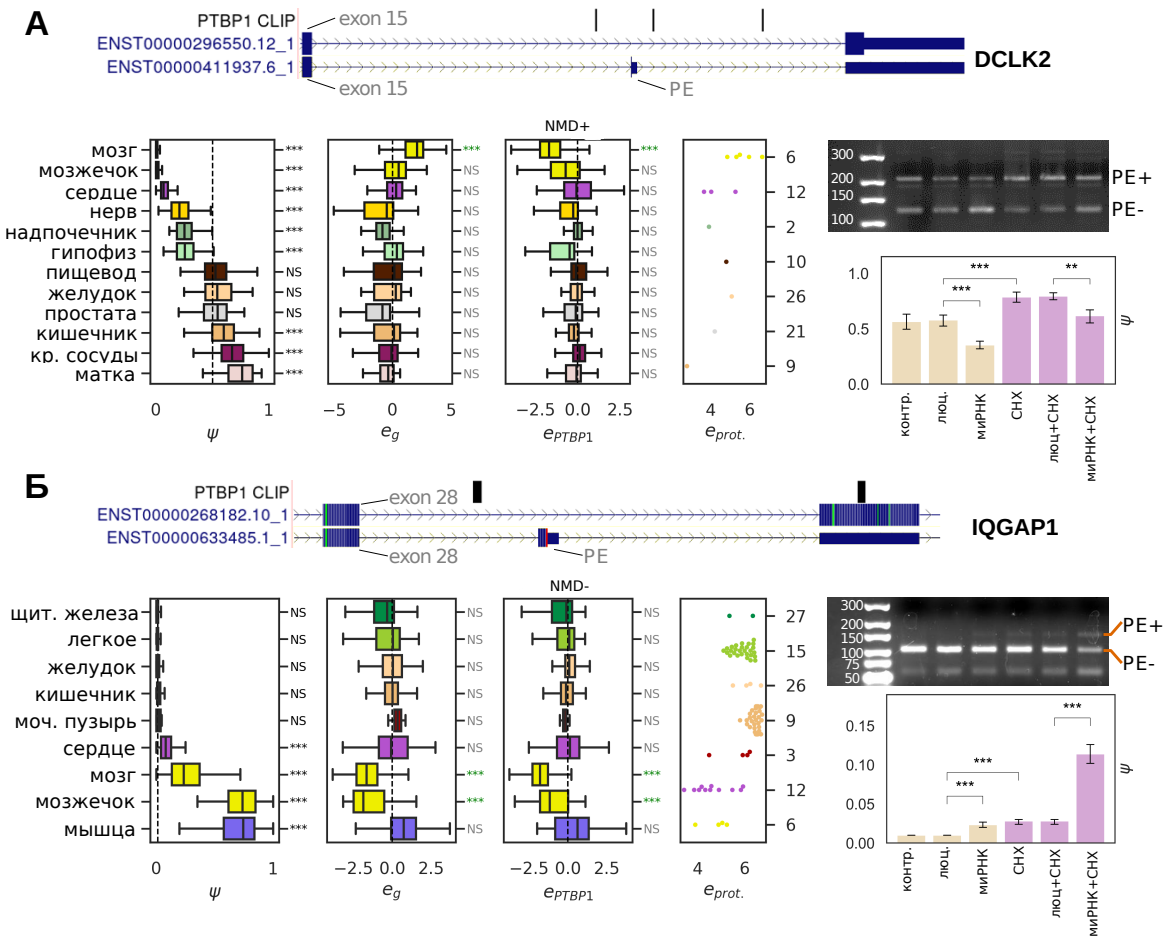


Рис. 10 — Валидация тканеспецифически регулируемых событий непродуктивного сплайсинга в генах *DCLK2* (А) и *IQGAP1* (Б). Ящичковые диаграммы показывают (слева направо) распределение Ψ, уровень экспрессии гена-мишени (e_g), уровень экспрессии регулятора PTBP1 (e_{PTBP1}) и уровень экспрессии белка. Идеограммы на каждой панели показывают положение сигналов CLIP для PTBP1. 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 микроРНК против гена люциферазы светлячка, микроРНК против PTBP1, циклогексимидом (СНХ), СНХ и контрольной микроРНК, СНХ и микроРНК против PTBP1. Символы *** и ** обозначают статистически значимые различия на 0.1% и 1% уровне значимости, соответственно.

Затем аналогичный подход применяется к предсказанию кросс-регуляторного тканеспецифического непродуктивного сплайсинга [33]. Составляются каталоги валидированных и аннотированных событий, а также и их регуляторов. После этого путем последовательной фильтрации среди них выделяются определенные группы событий, такие как значимо непродуктивные, тканеспецифические и тканеспецифически регулируемые, а в последней группе — события с поддержкой eCLIP. Удовлетворяющее наиболее жестким из этих критериев множество из 27 предсказанных тканеспецифически регулируемых событий с поддержкой eCLIP кластеризуется и для него строится кросс-регуляторная сеть. Для экспериментальной валидации из них выбираются две мишени фактора

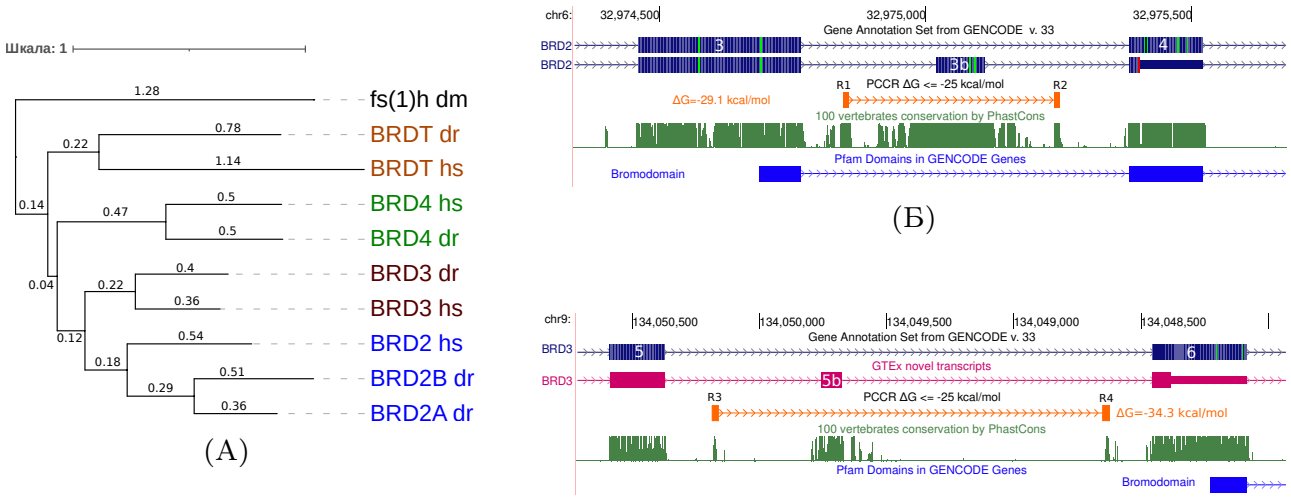


Рис. 11 — Семейство BET-белков. **(А)** Филогенетическое дерево гомологов *BRD2* у позвоночных из базы данных eggNOG, ограниченной *H. Sapiens* и *D. rerio*. Белок *fs(1)h* дрозофилы использовался в качестве внешней группы. Числа на ветвях обозначают среднее количество аминокислотных замен на сайт. **(Б)** 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.

PTBP1, *DCLK2* и *IQGAP1*, для которых регуляция подтверждается подавлением экспрессии PTBP1 с помощью микроРНК в клеточной линии A549 при инактивации системы NMD циклогексимидом (рис. 10).

Далее рассматривается влияние структуры РНК на непродуктивный сплайсинг в генах *BRD2* и *BRD3* из семейства BET-белков [25]. Строится филогенетическое дерево гомологов *BRD2* в двусторонне-симметричных позвоночных, которое подтверждает разделение на четыре ортологичные группы и свидетельствует о последнем расхождении *BRD2* и *BRD3* (рис. 11А).

В гене *BRD2* включение в транскрипт ядовитого экзона 3b, окруженного парой ККУ (R1 и R2), индуцирует появление PTC в следующем экзоне (рис. 11Б). Показывается, что в интроне между экзонами 5 и 6 его ближайшего гомолога *BRD3* содержится консервативный элемент, окруженный парой ККУ (R3 и R4), который является неаннотированным экзоном 5b (рис. 11В). Множественное выравнивание последовательностей белков семейства BET показывает, что расположение границ экзонов не изменилось после расхождения. При этом экзон 3b у *BRD2* и экзон 5b у *BRD3* расположены в негомологичных интронах, что указывает на их независимое приобретение. Далее производится экспериментальная валидация влияния структуры РНК на сплайсинг при помощи АОН, блокирующих спаривания R1/R2 и R3/R4, а также мутагенеза минигенов.

Затем характеризуется взаимосвязь между экспрессией и уровнями включения ядовитых экзонов *BRD2* и *BRD3* в тканях и опухолях по данным секвенирования РНК. Медианный уровень экспрессии транскриптов отрицательно коррелирует с медианной частотой включения ядовитых экзонов. Кроме того, частота включения экзона 3b *BRD2* снижается в ответ на замедление элонгации транскрипции α-аманитином. На основании этих наблюдений делается предположение о том, что структура РНК может регулировать уровни экспрессии *BRD2* и *BRD3* через непродуктивный сплайсинг в зависимости от скорости элонгации транскрипции, чем и обуславливается тканеспецифическая экспрессия этих генов в семенниках.

Далее обсуждаются свидетельства конвергентной эволюции, в ходе которой *BRD2* и *BRD3* независимо приобрели ядовитые экзоны и регулирующие их включение структуры РНК, а также регуляция непродуктивного сплайсинга структурой РНК.

В **заключении** приведены основные результаты работы, которые заключаются в следующем.

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.

Публикации автора по теме диссертации

Статьи в рецензируемых научных изданиях, индексируемых в базах данных Web of Science и Scopus, рекомендованных для защиты в диссертационном совете МГУ.014.2:

1. 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).
2. 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).

3. 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).

4. *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).

5. *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).

6. 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).

7. *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).

8. *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).

9. 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).

10. 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).

11. 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).

12. 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).

13. *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).

14. 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).

15. 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).

16. 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).

17. *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).

18. *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).

19. 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).

20. *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).

21. 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).

22. 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).

23. 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).

24. 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).

25. 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).

26. *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).

27. *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).

28. 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).

29. 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).

30. 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).

31. 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).

32. 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).

33. 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).

В прочих изданиях

A1. 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).

A2. 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).

A3. 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).

A4. 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).

A5. 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).

A6. 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).

A7. *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%).

Зарегистрированные патенты

B1. *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.

Первушин Дмитрий Давидович

Альтернативный сплайсинг и дальние взаимодействия в структуре эукариотических РНК

Автореф. дис. на соискание ученой степени докт. хим. наук

Подписано в печать 27.06.2024. Заказ № 001
Формат 60×90/16. Усл. печ. л. 1. Тираж 100 экз.
Типография «Реглет» г. Москва Ленинский пр-т. д. 67