На правах рукописи

Храмеева Екатерина Евгеньевна

Дальние взаимодействия в геномах эукариот и регуляция сплайсинга

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва – 2014

Работа выполнена в Федеральном государственном образовательном учреждении высшего профессионального образования "Московский государственный университет имени М.В. Ломоносова" и в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научные руководители:

Миронов Андрей Александрович, доктор биологических наук, профессор, ведущий научный сотрудник Учебно-научного центра "Биоинформатика" ИППИ РАН.

Гельфанд Михаил Сергеевич, доктор биологических наук, профессор, заведующий Учебно-научным центром "Биоинформатика" ИППИ РАН.

Официальные оппоненты:

Карягина-Жулина Анна Станиславовна, доктор биологических наук, профессор, главный научный сотрудник Федерального государственного бюджетного учреждения "Научно-исследовательский институт эпидемиологии и микробиологии имени почетного академика Н.Ф. Гамалеи" Министерства здравоохранения Российской Федерации.

Алексеевский Андрей Владимирович, кандидат физико-математических наук, ведущий научный сотрудник Научно-исследовательского института физико- химической биологии имени А.Н. Белозерского Федерального государственного образовательного учреждения высшего профессионального образования "Московский государственный университет имени М.В. Ломоносова".

Ведущая организация:

Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук.

Защита состоится 20 ноября 2014 года в 16:00 часов на заседании диссертационного совета Д.002.007.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук, расположенном по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, 19, стр. 1.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН), а также на сайте ИППИ РАН по адресу www.iitp.ru.

Автореферат разослан 10 октября 2014 года.

Ученый секретарь диссертационного совета,
доктор биологических наук, профессор  *Рожкова Г.И.*

Общая характеристика работы

Актуальность работы. Ни одна из естественных наук в настоящее время не обходится без применения компьютерных методов. Они позволяют хранить и обрабатывать большие объемы данных, моделировать природные процессы и системы, предсказывать их поведение. Так, на стыке биологии и компьютерных наук появилось новое самостоятельное научное направление — биоинформатика, которая использует компьютерные средства для решения биологических задач.

По мере развития экспериментальных методов, секвенирование геномов становится все более быстрым и дешевым процессом. Темпы секвенирования значительно опережают темпы экспериментального анализа геномов, и изучение структуры и функции ДНК, РНК и белков на всех этапах включает использование специальных вычислительных методов. Наличие большого количества расшифрованных геномов позволяет предсказывать функции генов и механизмы их регуляции в новых геномах по аналогии с уже изученными геномами. Задача изучения регуляции генов особенно интересна для многоклеточных эукариот, так как их гены имеют наиболее сложную структуру и считываемая с них пре-мРНК часто альтернативно сплайсируется.

Альтернативный сплайсинг — один из основных механизмов создания разнообразия белковых последовательностей. Альтернативный сплайсинг может вносить незначительные изменения в структуру и функцию белка, может резко изменять их, может приводить к формированию нетранслируемых изоформ. Альтернативный сплайсинг не только является одним из важных механизмов регуляции функционального состояния белков (а значит клеток и организмов в целом), но и сам подвержен сложной регуляции.

Известно, что у многих организмов вторичная структура РНК оказывает значительное влияние на процессы транскрипции и трансляции, однако влияние таких структур на процесс сплайсинга до сих пор систематически не изучено. Этот вопрос интересовал исследователей ещё со времен открытия сплайсинга, однако считалось, что основную роль в регуляции сплайсинга играют трансфакторы (элементы сплайсосомы, отдельные вспомогательные белки, различные регуляторные РНК). Недавние исследования показали, что механизмы сплайсинга, опосредованные вторичной структурой, могут быть более распространенными, чем предполагалось ранее [5]. Поэтому возникла необходимость систематического исследования окружений донорных и акцепторных сайтов сплайсинга на предмет наличия консервативных вторичных структур РНК, которые могли выступать в качестве регуляторов сплайсинга.

Вторичная структура РНК не является единственным механизмом регуляции альтернативного сплайсинга. Белковые факторы также способны оказывать влияние на сплайсинг, связываясь с особыми регуляторными последовательностями — энхансерами и сайленсерами. Энхансеры способствуют вырезанию интрона, а сайленсеры, напротив, противодействуют. Известно несколько

примеров регуляции сплайсинга белком hnRNPL, принадлежащим к семейству белков hnRNP — высоко экспрессированных в клетке белков, выполняющих разнообразные функции в метаболизме РНК. Функция белка hnRNPL до конца не ясна, поэтому, учитывая его широкую представленность в клетке, представляется интересным изучить механизмы регуляции альтернативного сплайсинга белком hnRNPL на полногеномном уровне.

Транс-сплайсинг — это особая форма процессинга РНК, в результате которой экзоны, находящиеся на двух разных молекулах РНК, соединяются и лигируются. Секвенирование транскриптомов различных организмов, от червей до человека, показало, что многие транскрипты состоят из сегментов последовательностей, которые не следуют друг за другом на хромосоме, а происходят из удаленных участков генома, и даже с разных хромосом. Некоторые из таких химерных транскриптов образуются в результате генетических перестроек у объекта секвенирования (чаще всего, в опухолевых клетках). Другие возникают в нормальных тканях пост-транскрипционно в результате транс-сплайсинга. Это указывает на то, что транс-сплайсинг у млекопитающих распространен существенно шире, чем считалось ранее, что делает задачу его изучения крайне важной.

Цель диссертационной работы состоит в изучении механизмов регуляции альтернативного сплайсинга вторичными структурами РНК и белковыми факторами, а также распространенности транс-сплайсинга, на полногеномном уровне.

Для достижения поставленной цели были решены следующие задачи:

1. Разработать метод поиска консервативных вторичных структур, ассоциированных со сплайсингом.

2. Исследовать окружения донорных и акцепторных сайтов сплайсинга на предмет наличия вторичных структур РНК, которые могли бы регулировать сплайсинг.

3. Изучить полногеномную карту позиций связывания белка hnRNPL с РНК в клетках человека HeLa.

4. Установить взаимосвязь между относительной позицией контакта белка hnRNPL с РНК и его активаторным или репрессирующим влиянием на сплайсинг.

5. Осуществить поиск химерных транскриптов, содержащих последовательности, соответствующие разным хромосомам.

6. Проверить, производят ли пространственно близкие участки генома большое количество химерных РНК по механизму транс-сплайсинга.

Научная новизна и практическая значимость. Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, — с помощью хэширования. Этот метод является альтернативой широко применяемому методу построения вторичных структур РНК с использованием множественного выравнивания [6]. Преимуществом метода хэширования является вычислительная быстрота, позволяющая осуществлять поиск вторичных структур РНК в полногеномном масштабе для многоклеточных эукариот без использования выравнивания. Применение множественного выравнивания ко вторичным структурам, определенным с помощью хэширования, позволяет уточнять предсказания, незначительно увеличивая время расчета.

В настоящее время экспериментально показано наличие около 15 случаев участия вторичной структуры в регуляции сплайсинга у различных организмов [7]: у вирусов (вирус гепатита B, аденовирус, вирус иммунодефицита человека типа 1, вирус саркомы Рауса), дрожжей, растений (*Nicotiana plumbaginifolia*), насекомых (*Drosophila*), а также у крыс и мышей. Известно, что в организме человека вторичные структуры могут оказывать влияние на эффективность распознавания сайтов сплайсинга и таким образом участвовать в формировании изоформ гормона роста, гена *tau*, гена *Hprt* и гена *hnRNPA1*. Ошибки сплайсинга, обусловленные влиянием вторичных структур РНК, приводят к таким патологиям человека, как мышечная дистрофия, кистозный фиброз и паркинсонизм. В настоящей работе идентифицировано несколько сотен генов, содержащих потенциальные консервативные вторичные структуры. Это наблюдение позволяет предположить, что механизмы сплайсинга, опосредованные вторичной структурой, могут быть более распространенными, чем предполагалось ранее.

Впервые получена и проанализирована точная полногеномная карта позиций связывания белка hnRNPL с РНК в клетках человека HeLa. Известно несколько случаев регуляции сплайсинга с помощью белка hnRNPL. Белок hnRNPL регулирует пропуск экзона в гене *CD45* в ответ на активацию Т-клеток. Он также оказывает влияние на альтернативный сплайсинг других генов посредством удерживания интрона, подавления включения нескольких экзонов и выбора альтернативного сайта полиаденилирования. Белок hnRNPL взаимодействует с 3' нетранслируемой областью (3'НТО) мРНК синтазы оксида азота и регулирует её стабильность. Белок hnRNPL принадлежит обширному семейству белков hnRNP, которые являются одними из наиболее высоко экспрессируемых в клетке и выполняют разнообразные функции в метаболизме пре-мРНК, среди которых упаковка только что синтезированных транскриптов, регуляция конститутивного и альтернативного сплайсинга, транспорт молекул мРНК и их локальная трансляция, регуляция стабильности мРНК, активация или репрессия трансляции. В настоящей работе показано, что участие белка hnRNPL в регуляции сплайсинга не ограничивается отдельными случаями, а носит полногеномный характер.

Транс-сплайсинг встречается не только у трипаносом и нематод, как счита-

лось ранее. Ранее были экспериментально показаны два случая транс-сплайсинга в клетках человека. Транс-сплайсинг может происходить между 5' экзонами гена JAZF1 на локусе 7p15 и 3' экзонами гена JJAZ1 на локусе 17q11, причем получившийся в результате химерный транскрипт транслируется в белок, препятствующий апоптозу. Транс-сплайсинг наблюдается и между генами SLC45A3 и ELK4, также с образованием функционального белка. Кроме того, данные высокопроизводительного секвенирования указывают на то, что транс-сплайсинг у млекопитающих — не редкое явление, как считалось ранее, а довольно распространенный механизм, что подтверждается результатами настоящей работы.

Впервые систематически изучено функциональное состояние часто контактирующих участков ДНК, находящихся на разных хромосомах. С помощью полногеномной карты частот контактов участков ДНК показано, что часто контактирующие фрагменты имеют сходный уровень экспрессии, модификаций хроматина, метилирования ДНК, чувствительности к ДНКазе, а также производят большое количество химерных РНК, большая часть которых, по-видимо, имеет пост-транскрипционное происхождение и образуется в результате транс-сплайсинга. Это наблюдение подтверждает существование транскрипционных фабрик, обогащенных факторами транскрипции и сплайсинга, в которых активно экспрессирующиеся ко-регулируемые гены могут образовывать межхромосомные контакты. Возможность организации генов в транскрипционные фабрики открывает новый, более сложный уровень регуляции генной активности и показывает, что современные представления о многокомпонентной системе регуляции экспрессии генов у многоклеточных эукариот являются лишь вершиной айсберга.

Апробация работы Материалы исследований по теме диссертации были представлены на международных конференциях: XV Международной конференции студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2008, диплом за лучший доклад), I Международном конкурсе научных работ молодых ученых в области нанотехнологии "Роснанотех" (Москва, 2008, призер конкурса), XVII Международной конференции студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2010), Bioinformatics after Next Generation Sequencing (Звенигород, 2010), 18th annual international conference on Intelligent Systems for Molecular Biology ISMB (Бостон, 2010), 24th International Mammalian Genome Conference (Крит, 2010), Albany 2011: The 17th Conversation (Олбани, 2011), Moscow Conference on Computational Molecular Biology MCCMB (Москва, 2011), 1st Cold Spring Harbor Asia conference on Bioinformatics of Human and Animal Genomes (Сучжоу, 2011), 16th Annual International Conference on Research in Computational Molecular Biology RECOMB (Барселона, 2012), Chromosomes, Stem Cells and Disease (Барселона, 2012), а также на конференциях "Информационные технологии и системы" ИТиС-31 (Геленджик, 2008), ИТиС-33 (Геленджик, 2010), ИТиС-34 (Геленджик, 2011), ИТиС-35 (Петрозаводск, 2012).

Структура и объем диссертации Диссертация состоит из введения, обзора литературы, 3 глав, заключения и библиографии. Общий объем диссер-

тации 108 страниц, из них 97 страницы текста, включая 42 рисунка и 3 таблицы. Библиография включает 87 наименований на 9 страницах.

Содержание работы

Введение Во введении обоснована актуальность диссертационной работы, сформулирована цель и аргументирована научная новизна исследований, показана практическая значимость полученных результатов, представлены выносимые на защиту научные положения.

Обзор литературы В разделе содержится мотивировка поставленных задач, а также аналитический обзор современной литературы по проблемам, рассмотренным в диссертации.

Глава 1. Поиск вторичных структур РНК, участвующих в регуляции сплайсинга Поиск консервативных вторичных структур был выполнен для группы плацентарных млекопитающих: *H. sapiens*, *R. macaque*, *P. troglodytes*, *M. musculus*, *R. norvegicus*, *C. familiaris*, *F. catus*, *B. taurus*, *E. caballus*, *M. domestica*. Для них была составлена база данных событий сплайсинга на основе данных RefSeq, полученных из UCSC genome browser database. Она содержала около 383 тыс. сайтов сплайсинга, 200 тыс. интронов и 213 тыс. экзонов.

Мы предполагаем, что вторичные структуры, расположенные вблизи сайтов сплайсинга, вероятнее ассоциированы со сплайсингом, чем удаленные структуры. Поэтому для поиска консервативных вторичных структур мы рассматривали только окружения донорных и акцепторных сайтов сплайсинга. При этом под окружением понимались 150 нуклеотидов внутрь интрона и 0 нуклеотидов внутрь экзона. Области внутри экзона не рассматривались, чтобы уменьшить процент ложных положительных предсказаний, поскольку вероятность нахождения консервативной вторичной структуры по случайным причинам выше в последовательностях с высокой средней консервативностью.

Для предсказания вторичных структур был разработан и применен метод хэширования [1]. Мы требовали присутствия хотя бы двух GC пар и консервативности затравки не менее, чем в 9 из 12 видов млекопитающих. При этом консервативность означала попарное сходство (с тремя или менее нуклеотидными заменами) затравок между всеми видами млекопитающих.

С использованием указанных ограничений был получен набор из 211 структур (Таблица 1). Для оценки уровня ложных положительных предсказаний применялся контрольный опыт, в котором каждой донорной части интрона сопоставлялась акцепторная часть чужого интрона. Приведенные оценки уровня ложных положительных предсказаний являются пессимистическими, поскольку вероятность образования вторичных структур повышена в последовательностях, которые содержат повторы, даже после перемешивания донорных и акцепторных частей интронов. Поэтому мы осуществили поиск вторичных структур в том же наборе последовательностей, но с маскированными повторами. Аб-

Таблица 1. Количество найденных вторичных структур и оценка уровня ложных положительных предсказаний при контроле без ограничения на GC состав или консервативность (Контроль), при ограничении на GC состав (Контроль+GC) и при ограничении на GC состав и консервативность (Контроль+GC+Cons). Указано среднее значение ± стандартное отклонение в 1000 повторений контрольного опыта.

Повторы	Структуры	Контроль	Контроль+GC	Контроль+GC+Cons
Не маскированы	211	60.1±4.2 (28%±2%)	61.6±4.3 (29%±2%)	76.0±4.1 (36%±2%)
Маскированы	167	47.4±3.1 (28%±2%)	43.8±3.2 (26%±2%)	50.6±3.0 (30%±2%)

солютное количество предсказанных вторичных структур уменьшилось, как и уровень ложных положительных предсказаний (Таблица 1)

Необходимо отметить, что консервативные вторичные структуры часто встречаются в альтернативно сплайсируемых генах. Рисунок 1 показывает, что 33% предсказанных структур соответствуют событиям альтернативного сплайсинга, аннотированным в RefSeq. При этом в контрольном наборе интронов (без предсказанных структур) того же размера ожидается только 10% структур, соответствующих событиям альтернативного сплайсинга. Кроме того, почти все подтипы альтернативного сплайсинга ассоциированы со вторичными структурами сильнее, чем ожидалось в контрольном наборе интронов.

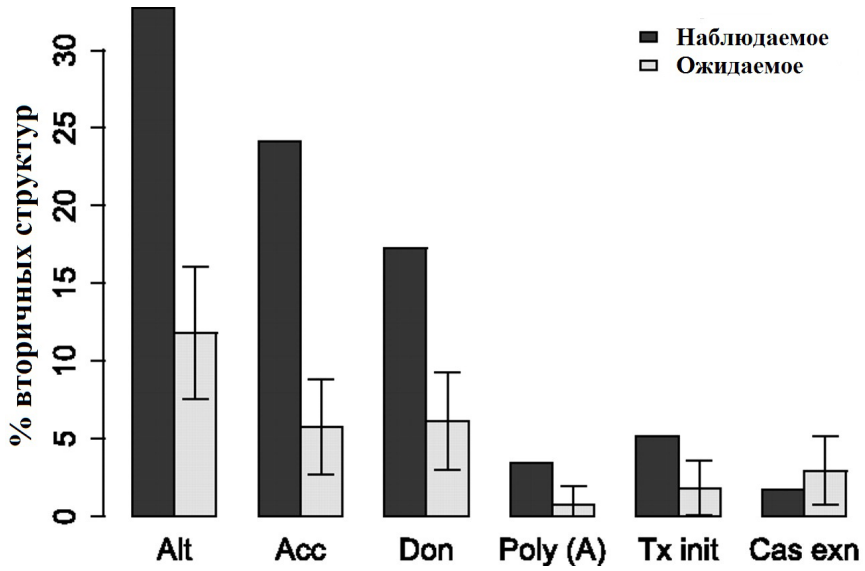

Рис. 1. Классы аннотированных событий сплайсинга, ассоциированных со вторичными структурами. Показаны пропорции относительно общего количества предсказанных вторичных структур. Также показаны ожидаемые значения и стандартные отклонения. Alt – альтернативный сплайсинг, Acc – альтернативный акцепторный сайт, Don – альтернативный донорный сайт, Poly(A) – альтернативный сайт полиаденилирования, Tx init – альтернативный сайт инициации транскрипции, Cas exn – интрон, содержащий один или несколько кассетных экзонов.

Можно представить два гипотетических механизма регуляции альтернативного сплайсинга посредством образования вторичных структур РНК: (1) блокирование сайта сплайсинга вторичной структурой, в результате чего сайт

сплайсинга оказывается недоступным для сплайсосомы, и используется альтернативный сайт (например, в гене *SLC39A7*); (2) вторичная структура скрепляет концы интрона и сближает сайты сплайсинга, что приводит к преимущественному использованию данного сайта сплайсинга по сравнению с альтернативным (например, в гене *SF1*).

В гене *SLC39A7* внутри интрона между экзонами 2 и 3 есть два альтернативных сайта сплайсинга – донорный и акцепторный. Найденная нами вторичная структура закрывает внутренние донорный и акцепторный сайты сплайсинга. Можно предположить, что регуляция в данном случае могла бы осуществляться следующим образом: когда участки вторичной структуры спарены, внутренние сайты сплайсинга оказывается недоступным для сплайсосомы, и сплайсинг идет по внешним сайтам. А если участки не спарены, то сплайсинг может идти по внутренним сайтам.

В гене *SF1* вблизи донорных сайтов экзонов 3 и 9 были найдены консервативные участки (C и E, соответственно), комплементарные одному и тому же участку (F) вблизи акцепторного сайта сплайсинга экзона 10. Предполагается, что при спаривании участков E-F экзоны 4-9 будут преимущественно включаться в процессированный транскрипт, а при спаривании C-F они включаться не будут. При этом важно, что участки C и E не могут одновременно связываться с участком F, и выбор пути сплайсинга будет зависеть от того, какой участок (C или E) взаимодействует с участком F в данный момент. Мы проанализировали публично доступные в базе NCBI Sequence Read Archive (SRA) данные по секвенированию транскриптома (RNA-seq) в клетках лимфатического узла (ERX011193) и щитовидной железы (ERX011194). Как и было предсказано, транскрипты, соответствующие пропуску экзонов 4-9, были найдены в обоих проанализированных наборах данных.

Кроме того, мы предполагаем, что интрон между экзонами 9 и 10 не вырезается, если участки E и F не спарены. Это предположение косвенно подтверждается тем, что данный интрон содержит стоп-кодон, и его удерживание привело бы к деградации мРНК гена *SF1*. Гипотеза была проверена экспериментально группой Петра Рубцова из Института молекулярной биологии им. В.А. Энгельгардта РАН [1]. Было показано, что нарушение комплементарности участков E и F приводит к вырезанию более длинного интрона из-за преимущественного использования внешнего акцепторного сайта сплайсинга вместо внутреннего акцепторного сайта. Таким образом, комплементарность участков E-F критически необходима для правильного вырезания интрона.

Глава 2. Регуляция сплайсинга с помощью белковых факторов
Основываясь на данных, полученных с помощью метода iCLIP, совмещенного с высокопроизводительным секвенированием, мы построили точную полногеномную карту позиций связывания белка hnRNPL с РНК в клетках человека (клеточная линия HeLa). Экспериментальные данные были предоставлены группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия (Prof. Dr. Albrecht Bindereif, Justus-Liebig-University-Giessen). Было выполнено

три независимых эксперимента и три контрольных экспермента (без обогащения антителами на белок hnRNPL). Экспериментальные процедуры и выравнивание прочтений выполнялись как описано в работе [8].

Анализ содержания пентамеров в последовательностях $\sim$ 1.1 млн. позиций связывания hnRNPL был выполнен, как описано в работе [9]. Рассматривались участки последовательности от -30 до -10 и от +10 до +30 по отношению к позиции связывания. В этих участках наблюдалось значительное обогащение CA-повторами и CA-богатыми мотивами: CA-повторы (ACACA, CACAC) были представлены чаще ожидаемого на 110-140%, а CA-богатые мотивы (ACAT, TACA) — на 80-95%. В контрольном эксперименте подобного обогащения не наблюдается.

Позиции связывания hnRNPL ($\sim$ 1.1 млн.) были кластеризованы и профильтрованы, и для дальнейшего функционального анализа было отобрано 622789 позиций связывания. Чтобы исследовать паттерны связывания hnRNPL вблизи сайтов сплайсинга, мы отобрали внутренние экзоны всех аннотированных в GENCODE V4 транскриптов белок-кодирующих генов. Их альтернативные и константные 3'- и 5'-сайты сплайсинга рассматривались отдельно (рис. 2A). Наблюдается значительно более высокая плотность позиций связывания hnRNPL вблизи альтернативных 5'-сайтов сплайсинга, как в экзонной (интервал от -30 до 0), так и в интронной (от 0 до +70) части. Кроме того, hnRNPL гораздо чаще связывается вблизи альтернативных 3'-сайтов сплайсинга (около позиции -20). Это наблюдение означает, что hnRNPL может регулировать эффективность использования сайтов сплайсинга посредством связывания вблизи 5'-сайтов сплайсинга, взаимодействуя с интронными или экзонными элементами, а также посредством связывания с полипиримидиновым трактом вблизи 3'-сайтов сплайсинга.

Далее, мы исследовали особенности связывания hnRNPL вблизи экзонов-мишеней hnRNPL, детектированных с помощью анализа на микрочипах. Экспериментальные данные были предоставлены группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия. Белок hnRNPL был подвергнут нокдауну с помощью РНК-интерференции в клетках HeLa, с последующей обработкой клеток циклогексимидом, чтобы подавить антисмысловой распад мРНК (NMD, или 'nonsense-mediated mRNA decay'). В результате анализа данных, полученных с помощью микрочипа, было отобрано 890 экзонов с пониженным уровнем включения после нокдауна hnRNPL (hnRNPL действует как активатор) и 574 экзона с повышенным уровнем включения (hnRNPL действует как репрессор).

Мы проанализировали плотность позиций связывания hnRNPL отдельно для L-активируемых и L-репрессируемых экзонов (Рисунок 2B). Их непосредственное сравнение показало, что белок hnRNPL репрессирует включение экзона, если связывается в интронной области, в непосредственной близости от 3'-сайта сплайсинга, и, напротив, активирует включение, если связывается в интронной области вблизи 5'-сайта сплайсинга (первые 200 нуклеотидов). Таким

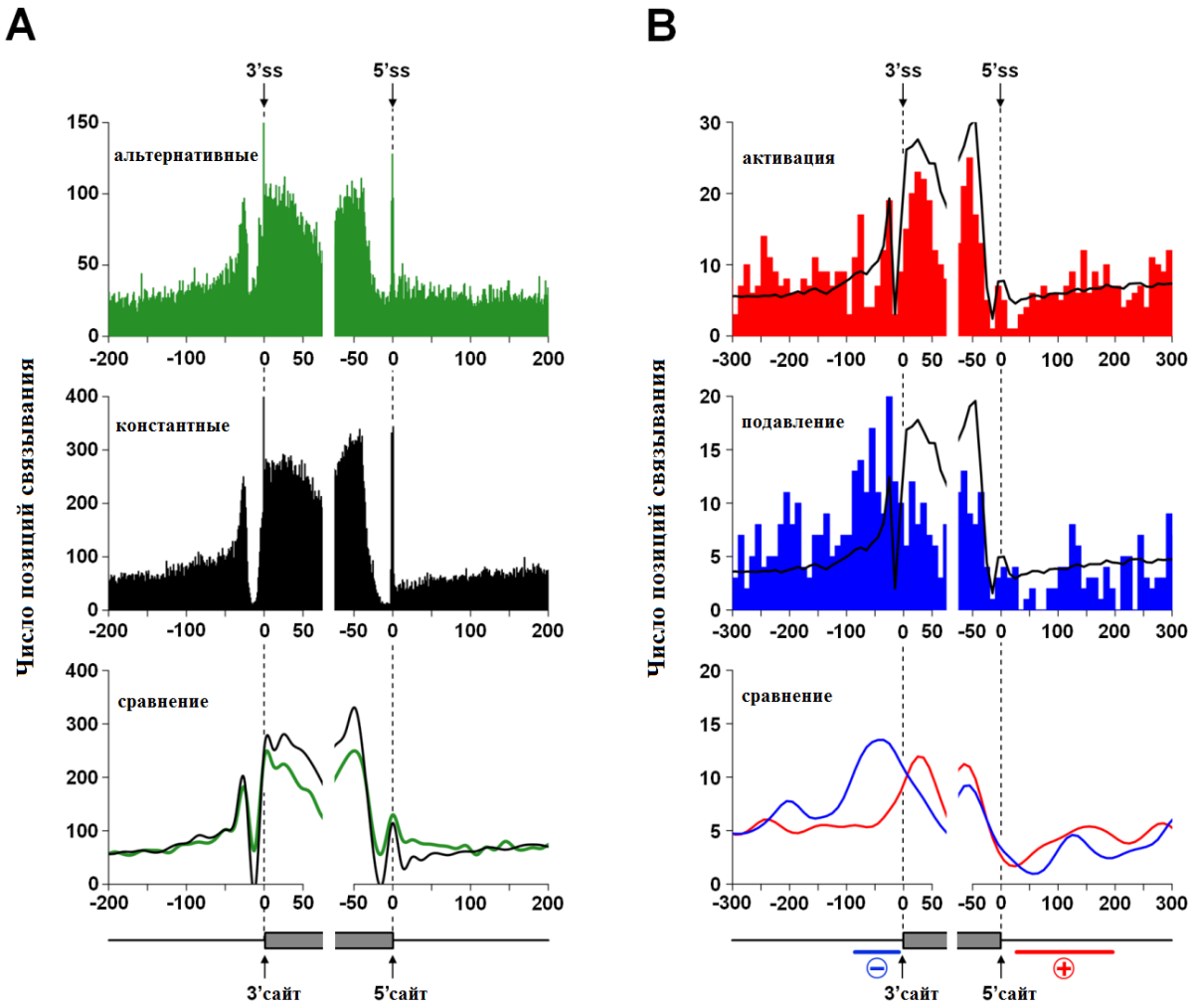

Рис. 2. (A) Количество позиций связывания hnRNPL вблизи 3'-сайтов сплайсинга (интервал от -200 до +75) и 5'-сайтов сплайсинга (интервал от -75 до +200) белок-кодирующих генов, отдельно для альтернативных и константных сайтов сплайсинга (верхняя и средняя панели, соответственно). На нижней панели, альтернативные (зеленые) и константные (черные) сайты сплайсинга сравниваются после нормализации и сглаживания. (B) Количество позиций связывания hnRNPL вблизи 3'-сайтов сплайсинга (интервал от -300 до +75) и 5'-сайтов сплайсинга (интервал от -75 до +300) отдельно для экзонов, включение которых активируется или репрессируется (верхняя и средняя панели, соответственно). Контрольный набор экзонов, не регулируемых белком hnRNPL, показан черной линией. На нижней панели, активируемые (красные) и репрессируемые (синие) экзоны сравниваются после нормализации и сглаживания.

образом, позиция связывания hnRNPL по отношению к регулируемому экзону определяет его активаторную или репрессорную функцию.

Чтобы изучить на полногеномном уровне способность белка hnRNPL регулировать подавляющее действие микроРНК на ген, мы использовали консервативные мишени микроРНК в области 3'НТО, предсказанные программой TargetScanHuman 5.1. Среди 3'НТО белок-кодирующих генов, аннотированных в GENCODE V4, было отобрано 5062 3'НТО, содержащих одновременно позиции связывания hnRNPL и предсказанные мишени микроРНК. Оказалось, что плотность позиций связывания hnRNPL значительно выше внутри мише-

нией микроРНК, чем в окружающих 3'НТО (тест Уилкоксона, P-значение $< 2.2e^{-16}$). Данное наблюдение может означать, что существует глобальный механизм конкуренции между hnRNPL и микроРНК, осуществляющий регуляцию активности транскриптов в цитоплазме.

Среди 349 аннотированных мякРНК (малые ядрышковые РНК, или snoRNA), 256 находятся внутри интронов. Большая часть этих интронов (237, или 93%) имеют длину менее 5 тыс. нуклеотидов. Именно для таких интронов мы проанализировали плотность позиций связывания hnRNPL. 106559 интронов длиной менее 5 тыс. нуклеотидов из 8376 экспрессированных мультиэкзонных белоккодирующих генов были использованы в качестве контроля. Интроны каждой группы были разбиты на 20 категорий в соответствии с их длиной, с интервалом в 250 нуклеотидов. Оказалось, что плотность позиций связывания hnRNPL в интронах, содержащих мякРНК, значительно выше. Данное наблюдение может свидетельствовать об участии белка hnRNPL в метаболизме мякРНК. Оно было подтверждено экспериментально группой Альбрехта Биндерейфа из Университета Юстуса-Либиха, г. Гиссен, Германия [2].

Глава 3. Транс-сплайсинг Используя данные секвенирования генома и транскриптома человека из 117 эксперментов, выполненных в 26 лабораториях по всему миру на разных платформах, мы вычислили профили покрытия генов, которые используются для решения таких популярных задач, как определение уровней экспрессии генов или включения экзонов. Были рассмотрены только одноэкзонные гены с высоким покрытием (178 генов), чтобы сделать профили покрытия генов сравнимыми между экспериментами по секвенированию мРНК в разных тканях, которые могут производить альтернативные изоформы в результате сплайсинга, а также между экспериментами по секвенированию геномов и транскриптомов.

Кластеризация экспериментов по профилям покрытия генов показала, что эксперименты по секвенированию транскриптома, сделанные в одной и той же лаборатории, имеют схожие профили покрытия, даже если были секвенированы такие разные ткани, как мозг и печень. Средний коэффициент корреляции между профилями покрытия одного и того же гена в разных экспериментах в одной и той же лаборатории составляет 0.46 ± 0.14 для секвенирования РНК на платформе Иллюмина. В то же время, эксперименты по секвенированию транскриптома, сделанные в разных лабораториях, имеют значительно менее похожие профили покрытия, даже если была секвенирована одна и та же ткань. Средний коэффициент корреляции в этом случае составляет 0.27 ± 0.10 [4].

Карта контактов участков ДНК в пространстве была построена с помощью методов высокопроизводительного секвенирования [10], и поэтому может быть загрязнена систематическими ошибками секвенирования, происходящими в результате экспериментальной процедуры (а именно, полимеразной цепной реакции) или неправильного выравнивания прочтений. Оба этих типа систематических ошибок происходят наиболее часто между участками генома с высоким уровнем сходства.

Значения пространственной близости были условно разделены на 29 интервалов. Аномально высокое содержание идентичных последовательностей наблюдается в геномных фрагментах со значениями пространственной близости выше 0.55. Кроме того, общее количество пар фрагментов в концевых интервалов существенно меньше, чем в центральных интервалах, и результаты для таких интервалов имеют низкую статистическую значимость. Поэтому далее мы рассматривали только интервалы с пространственной близостью от -0.3 до 0.55.

Мы осуществили поиск химерных РНК в трех образцах по секвенированию транскриптома человека – мозговой ткани и клеточных линиях GM12878 и K562. Для каждого интервала пространственной близости мы вычислили долю взаимодействующих фрагментов ДНК, между которыми наблюдается образование химерных РНК. Чтобы сделать разные наборы данных сравнимыми, это значение было в дальнейшем разделено на общее количество химерных РНК в образце. Контрольные наборы данных для каждого из трех образцов были получены путем составления искусственных пар прочтений со случайными прочтениями на другой хромосоме.

Данные для каждого из трех образцов показали значимые корреляции между частотой образования химерных РНК и пространственной близостью фрагментов (коэффициент корреляции Спирмена = 0.88, 0.94, 0.85, P-значение $< 2.2e^{-16}$, $1.7e^{-6}$, $2.2e^{-16}$, соответственно), по сравнению с контрольными наборами данных (Рисунок 3A).

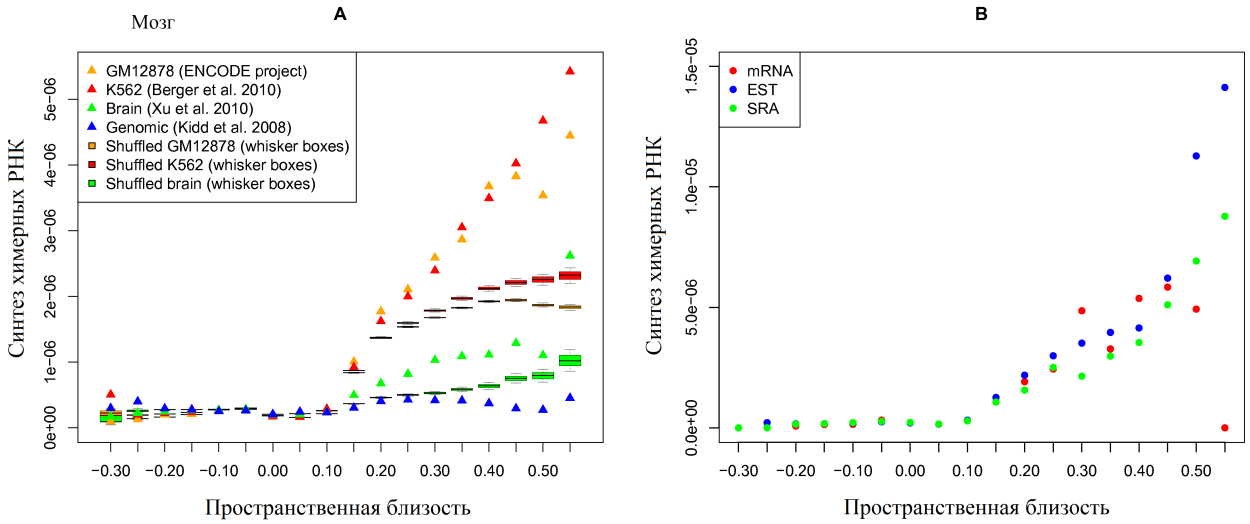

Рис. 3. Корреляция между производством химерных РНК и значениями пространственной близости для (A) клеточной линии K562, клеточной линии GM12878 и ткани мозга (красные, оранжевые и зеленые треугольники, соответственно); набора данных о геномных перестройках (показан синим); контрольных наборов данных K562, GM12878 и ткани мозга (красные, оранжевые и зеленые прямоугольники) и (B) трех наборов данных из базы ChimerDB: мРНК, EST и SRA (красные, синие и зеленые точки, соответственно).

Наблюдаемые корреляции во всех наборах данных могут быть вызваны, по крайней мере, двумя причинами: транс-сплайсингом и геномными переста-

новками. Чтобы выяснить, какая из причин оказывает наибольшее влияние, мы проанализировали данные о геномных перестройках из работы [11] и обнаружили, что повышения доли химерных РНК среди пространственно близких пар фрагментов в данном случае не наблюдается (Fig. 3A). Таким образом, найденные химерные транскрипты образуются, скорее всего, в результате транс-сплайсинга.

Мы также проанализировали данные из базы химерных РНК ChimerDB, которая содержит химерные транскрипты, собранные из различных публично доступных ресурсов. Рис. 3B показывает, что значения пространственной близости коррелируют с уровнем производства химерных РНК согласно всем трем источникам данных (экспрессированным коротким последовательностям, или EST, данным секвенирования, или SRA, и мРНК), доступным в ChimerDB (коэффициент корреляции Спирмена = 0.93, 0.97, 0.70, P-значение $< 2.2e^{-16}$, $7.8e^{-12}$, 0.001, соответственно).

Мы проверили, обладают ли пространственно близкие участки генома схожим уровнем модификаций гистонов, метилирования ДНК, чувствительности к ДНКазе и экспрессии. Для этого были использованы данные нескольких исследований. Все перечисленные свойства имеют одну и ту же структуру данных в виде маркеров, расположенных вдоль генома. После измерения каждый маркер характеризуется пиком определенной ширины и высоты, называемым сигналом. Чтобы усреднить силу сигнала $S(i)$ на фрагменте генома i длиной в 1Мб, мы умножили высоту каждого пика (H_k) на долю (W_k) фрагмента i, пересекающегося с данным пиком, а затем просуммировали результаты для всех пиков $1...n$ во фрагменте i:

$$S(i) = \sum_{k=1}^{n} (H_k \cdot W_k) \tag{1}$$

Разница $D(i,j)$ между силой сигнала в двух взаимодействующих фрагментах i и j была вычислена как:

$$D(i,j) = |\log \frac{S(i)}{S(j)}| \tag{2}$$

Для каждого из рассматриваемых интервалов пространственной близости в корреляционной матрице C была вычислена медиана $D(i,j)$. Оказалось, что эти медианы коррелируют со значениями пространственной близости в каждом типе данных, который был протестирован: уровень экспрессии (коэффициент корреляции Спирмена = -0.96, P-значение = $6.0e^{-6}$), модификации гистонов (средний коэффициент корреляции Спирмена = -0.85, среднее P-значение = $4.1e^{-4}$), метилирование ДНК (коэффициент корреляции Спирмена = -0.92, P-значение $< 2.2e^{-16}$), чувствительность к ДНКазе (коэффициент корреляции Спирмена = -0.93, P-значение $< 2.2e^{-16}$).

Далее мы проверили, распространяются ли наблюдаемые корреляции на функциональный уровень генов, согласно базе данных Gene Ontology. Для этого мы определили семантическое подобие терминов (GO-терминов), описывающих функции генов. Оказалось, что среднее семантическое подобие GO-терминов коррелирует со значениями пространственной близости для всех трех иерархий базы данных Gene Ontology 'Молекулярная функция', 'Биологический процесс' и 'Клеточный компонент' (коэффициент корреляции Спирмена = 0.78, 0.65, 0.98, P-значение = $2.1e^{-4}$, $4.3e^{-3}$, $8.4e^{-6}$, соответственно). Данное наблюдение позволяет говорить о том, что пространственно близкие участки генома обогащены генами со схожими функциями.

Мы также изучили ко-экспрессию генов в пространственно близких фрагментах. Данные были получены из базы COXPRESdb. Для каждого рассматриваемого интервала пространственной близости были вычислены медианы значений ко-экспрессии в двух взаимодействующих фрагментах генома, и для них наблюдается сильная корреляция со значениями пространственной близости (коэффициент корреляции Спирмена = 0.93, P-значение $< 2.2e^{-16}$) [3].

Затем мы рассмотрели так называемые состояния хроматина, биологически значимые комбинации эпигенетических маркеров [12]. Значения пространственной близости сильно коррелируют с подобием профилей состояний хроматина (коэффициент корреляции Спирмена = 0.99, P-значение = $9.6e^{-6}$) [3].

Наблюдаемые нами корреляции между пространственным расположением участков хроматина и подобием их состояний хорошо согласуются с теорией о транскрипционных фабриках [13]. Согласно этой теории, гены в транскрипционных фабриках характеризуются высоким уровнем экспрессии, обладают сходными эпигенетическими маркерами и выполняют похожие функции, а также часто ко-экспрессируются. Транскрипционные фабрики производят большое количество химерных РНК, в основном пост-транскрипционно, за счет трансплайсинга.

Выводы

1. Разработан новый метод поиска консервативных вторичных структур, ассоциированных со сплайсингом, – с помощью хэширования. С его помощью у млекопитающих идентифицировано несколько сотен генов, содержащих потенциальные консервативные вторичные структуры.

2. Консервативные вторичные структуры встречаются в альтернативно сплайсируемых генах чаще ожидаемого, что указывает на их участие в регуляции альтернативного сплайсинга. В частности, показано, что образование вторичной структуры необходимо для правильного вырезания интрона в гене *SF1*.

3. Построена полногеномная карта сайтов связывания белка hnRNPL с РНК в клетках человека HeLa. Показано, что сайты связывания hnRNPL обогащены CA-повторами и CA-богатыми мотивами.

4. Распределение позиций связывания hnRNPL вокруг 5'- и 3'-сайтов сплайсинга различается между альтернативными и константными экзонами, и между L-активируемыми и L-репрессируемыми экзонами. Позиция связывания белка hnRNPL определяет его активаторное или репрессирующее влияние на сплайсинг.

5. Белок hnRNPL часто связывается вблизи мишеней микроРНК в области 3'UTR и, возможно, регулирует стабильность мРНК за счет конкуренции с микроРНК.

6. Плотность позиций связывания hnRNPL в интронах, содержащих мякРНК, значительно выше, что может говорить об участии белка hnRNPL в биосинтезе мякРНК.

7. Анализ систематических ошибок секвенирования показал зависимость профилей покрытия генов от лаборатории в экспериментах по секвенированию мРНК. В данных о пространственной близости обнаружены и удалены систематические ошибки секвенирования.

8. Идентифицированы химерные РНК в трех наборах данных секвенирования транскриптома человека (ткань мозга, клеточная линия эритролейкемии K562, лимфобластоидная клеточная линия GM12878).

9. Пространственно близкие фрагменты ДНК образуют между собой больше химерных РНК, чем пространственно далекие, в основном за счет трансплайсинга.

10. Пространственно близкие фрагменты ДНК характеризуются схожими эпигенетическими маркерами и состояниями хроматина, гены в них функционально подобны и ко-экспрессируются, что хорошо согласуется с теорией о фабриках транскрипции.

Список публикаций по теме диссертации

Статьи в реферируемых журналах

1. Pervouchine D. D., Khrameeva E. E., Pichugina M. Y. et al. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures // RNA. 2012. — Jan. Vol. 18, no. 1. P. 1–15.

2. Rossbach O., Hung L. H., Khrameeva E. et al. Crosslinking-immunoprecipitation (iCLIP) analysis reveals global regulatory roles of hnRNP L // RNA Biol. 2014. — Feb. Vol. 11, no. 2. P. 146–155.

3. Khrameeva E. E., Mironov A. A., Fedonin G. G. et al. Spatial proximity and similarity of the epigenetic state of genome domains // PLoS One. 2012. Vol. 7, no. 4.

4. Khrameeva E. E., Gelfand M. S. Biases in read coverage demonstrated by inter-laboratory and interplatform comparison of 117 mRNA and genome sequencing experiments // BMC Bioinformatics. 2012. Vol. 13 Suppl 6.

Тезисы конференций

1. Khrameeva E.E., Mironov A.A., Gelfand M.S., Pervushin D.D. Secondary structures in Drosophila introns // XV Международная конференция студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2008)

2. Khrameeva E.E. Regulation of splicing by conserved RNA structures // I Международный конкурс научных работ молодых ученых в области нанотехнологии "Роснанотех" (Москва, 2008)

3. Khrameeva E.E. Long-range interactions in eukaryotic chromatin and their possible imprint on function and evolution // XVII Международная конференция студентов, аспирантов и молодых учёных "Ломоносов" (Москва, 2010)

4. Khrameeva E.E., Mironov A.A., Gelfand M.S. Long range correlations in the genome and chromatin // Bioinformatics after Next Generation Sequencing (Звенигород, 2010)

5. Khrameeva E.E., Mironov A.A., Khaitovich P., Gelfand M.S. Spatial proximity and similarity of functional states of genome domains // 18th annual international conference on Intelligent Systems for Molecular Biology ISMB (Бостон, 2010)

6. Khrameeva E.E., Mironov A.A., Gelfand M.S. Association between spatial proximity and functional similarity in human genome // 24th International Mammalian Genome Conference (Крит, 2010)

7. Khrameeva E.E., Mironov A.A., Gelfand M.S. The Impact of Interchromosomal Associations on the Functional State of the Human Genome // Albany 2011: The 17th Conversation (Олбани, 2011)

8. Khrameeva E.E., Gelfand M.S. Interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments // Moscow Conference on Computational Molecular Biology MCCMB (Москва, 2011)

9. Khrameeva E.E., Mironov A.A., Khaitovich P., Gelfand M.S. Spatial proximity and similarity of the epigenetic state of genome domains // 1st Cold Spring Harbor Asia conference on Bioinformatics of Human and Animal Genomes (Сучжоу, 2011)

10. Khrameeva E.E., Gelfand M.S. Biases in read coverage demonstrated by interlaborat and interplatform comparison of 117 mRNA and genome sequencing experiments // 16th Annual International Conference on Research in Computational Molecular Biology RECOMB (Барселона, 2012)

11. Khrameeva E.E. Spatial proximity and similarity of the epigenetic state of genome domain // Chromosomes, Stem Cells and Disease (Барселона, 2012)

12. Khrameeva E.E., Mironov A.A., Gelfand M.S., Pervushin D.D. Secondary structures in Drosophila introns // "Информационные технологии и системы" ИТиС-31 (Геленджик, 2008)

13. Khrameeva E.E., Mironov A.A., Gelfand M.S. Functional similarity and chimeric transcripts in spatially close genome domains // "Информационные технологии и системы" ИТиС-33 (Геленджик, 2010)

14. Khrameeva E.E., Gelfand M.S. Comparison of 117 mRNA and genome sequencing experiments between laboratories and platforms // "Информационные технологии и системы" ИТиС-34 (Геленджик, 2011)

15. Khrameeva E.E. Regulation of alternative splicing by hnRNPL protein // "Информационные технологии и системы" ИТиС-35 (Петрозаводск, 2012)

Цитированная литература

5. Raker V. A., Mironov A. A., Gelfand M. S., Pervouchine D. D. Modulation of alternative splicing by long-range RNA structures in Drosophila // Nucleic Acids Res. 2009. — Aug. Vol. 37, no. 14. P. 4533–4544.

6. Hofacker I. L., Fekete M., Stadler P. F. Secondary structure prediction for aligned RNA sequences // J Mol Biol. 2002. — Jun. Vol. 319, no. 5. P. 1059–1066.

7. Buratti E., Baralle F. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process // Mol. Cell. Biol. 2004. Vol. 24. P. 10505–10514.

8. König J., Zarnack K., Rot G. et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution // Nat Struct Mol Biol. 2010. — Jul. Vol. 17, no. 7. P. 909–915.

9. Wang Z., Kayikci M., Briese M. et al. iCLIP predicts the dual splicing effects of TIA-RNA interactions // PLoS Biol. 2010. Vol. 8, no. 10.

10. Lieberman-Aiden E., van Berkum N. L., Williams L. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome // Science. 2009. — Oct. Vol. 326, no. 5950. P. 289–293.

11. Kidd J. M., Cooper G. M., Donahue W. F. et al. Mapping and sequencing of structural variation from eight human genomes // Nature. 2008. — May. Vol. 453, no. 7191. P. 56–64.

12. Ernst J., Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome // Nat Biotechnol. 2010. — Aug. Vol. 28, no. 8. P. 817–825.

13. Gingeras T. R. Implications of chimaeric non-co-linear transcripts // Nature. 2009. — Sep. Vol. 461, no. 7261. P. 206–211.