Федеральное государственное бюджетное учреждение науки
Институт проблем передачи информации им. А.А. Харкевича
Российской академии наук

На правах рукописи

Мазин Павел Владимирович

Анализ возрастных изменений альтернативного сплайсинга в коре головного мозга высших приматов

03.01.09 — математическая биология, биоинформатика

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
PhD Хайтович Филипп Ефимович

Москва — 2018

Содержание

1. Введение

1.1. Актуальность работы

Понимание молекулярных механизмов функционирования живых организмов — основное направление современной биологии. В связи с развитием высокопроизводительных методов, таких как методы секвенирования нового поколения, все большее значение в биологии приобретают вычислительные методы обработки данных. С применением этих методов за последние годы было показано, что у эукариот многие транскрипты подвержены альтернативному сплайсингу (АС). То есть вырезание интронов (участков не включающихся в состав мРНК и, тем самым, не участвующих в кодировании белка) и сшивание остающихся кусков РНК, — экзонов, может происходить у одного гена не единственным способом. По современным представлениям большинство генов человека подвержено АС. Известно, что АС часто регулируется тканеспецифично и играет важную роль как в нормальном развитии тканей, так и во многих заболеваниях.

Одной из тканей с наиболее специфичным АС является нервная ткань. Известно, что при некоторых заболеваниях мозга, таких как аутизм или болезнь Альцгеймера, могут происходить изменения АС, что указывает на возможную роль АС в развитии этих патологий. Однако на настоящий момент не существует ни одного полноценного исследования АС в ходе нормального развития мозга, что затрудняет понимание роли АС в патологиях.

Человека от общего предка с шимпанзе отделяет всего шесть миллионов лет эволюции, однако когнитивные способности и социальное поведение человека (функции, за выполнение которых отвечает головной мозг) резко отличаются от таковых у шимпанзе. Хотя отличия в уровнях экспрессии генов между человеком и

другими приматами изучены относительно неплохо, опубликовано лишь несколько работ посвящённых АС, и ни в одной из них не производится сравнения возрастной регуляции АС в мозге приматов.

Таким образом, изучение возрастных АС в мозге приматов с эволюционной точки зрения может помочь лучше понять патогенез различных заболеваний мозга и пролить свет на эволюцию мозга человека.

1.2. Цели и задачи исследования

Целью данной работы являлось изучение и сравнение возрастных изменений АС в мозге человека и других высших приматов. Для этого были поставлены следующие задачи:

1. Разработать метод анализа АС в одном виде, исходя из данных массового секвенирования РНК (РНК-Сек).
2. Исследовать изменения альтернативного сплайсинга в головном мозге человека в ходе развития и старения.
3. Найти возможные регуляторные механизмы ответственные за наблюдаемые возрастные изменения
4. Разработать методы сопоставления экзонов между несколькими видами.
5. Сравнить возрастные изменения АС в мозге человека, шимпанзе и макаки
6. Проанализировать межвидовые отличия АС у человека, шимпанзе и макаки. Найти возможные причины этих отличий.

1.3. Научная новизна и практическая значимость

С методической точки зрения, новизна работы состоит в разработке и программной реализации нового алгоритма анализа АС, исходя из данных

массового секвенирования РНК. Этот алгоритм устойчив к экспериментальным артефактам, таким как избыточная амплификация библиотеки, и пригоден для анализа всех типов АС и для обработки результатов экспериментов со сложным дизайном. На данный момент подобных программ не существует.

Систематические исследования возрастных изменений сплайсинга в мозге человека и сравнение таковых с возрастными изменениями АС в мозге приматов также ранее не проводилось. Изучение нормального развития мозга является первым шагом к изучению патологий мозга и, таким образом, может иметь практическое значение для медицины.

1.4. Основные результаты и положения, выносимые на защиту

1. Разработан метод SAJR для количественного анализа альтернативного сплайсинга. Программа позволяет определить частоты включения альтернативных сегментов в мРНК и сравнить частоты включения между несколькими образцами. Программа обладает рядом преимуществ по сравнению с другими методами: позволяет учитывать биологическую вариабельность, подходит для сложных экспериментальных дизайнов, использует информацию о прочтениях РНК, которые картируются на экзон-экзонную границу. Показано, что результаты SAJR устойчивы и воспроизводятся на независимо полученных данных РНК-Сек, а также данных, полученных принципиально другим экспериментальным методом — полуколичественной полимеразной цепной реакцией.
2. При помощи SAJR показано, что в ходе развития мозжечка и префронтальной коры мозга человека меняется альтернативный сплайсинг сотен генов. При этом, хотя большая часть изменений происходит одинаково в обеих областях мозга, около 15% генов с возраст-зависимым сплайсингом

ведут себя по-разному в коре и мозжечке.

3. Разработан сбалансированный метод, позволяющий сравнивать альтернативный сплайсинг в нескольких видах. Показано, что межвидовые отличия в последовательностях сайтов связывания сплайсосомы (сайтов сплайсинга) объясняют около 20% всех межвидовых отличий, а в случае высоко-амплитудных изменений, эта доля увеличивается до 80%.
4. Показана консервативность возрастных изменений альтернативного сплайсинга белок-кодирующих сегментов в мозге человека, шимпанзе и макаки. Найдены вероятные регуляторные мотивы и идентифицированы связывающиеся с ними факторы сплайсинга. На основании найденных мотивов и факторов сплайсинга построена модель возрастной регуляции альтернативного сплайсинга в мозгу человека.
5. Показано, что частота удержания интронов падает с возрастом в префронтальной коре всех трёх изучаемых видов. Отрицательная корреляция между частотой включения интрона и экспрессией всего гена указывает на существенную роль удержания интронов в возраст-зависимой регуляции уровня экспрессии генов.

1.5. Публикации. Степень достоверности и апробация результатов

По материалам диссертации опубликованы две статьи, результаты работы представлены на международных (MCCMB'09, MCCMB'11, MCCMB'13, Postgenome'14, IMGC'15) и российских (ИТиС'11, ИТиС'12) конференциях.

1.6. Структура и объем диссертации

Диссертация состоит из введения, обзора литературы, двух глав, выводов, библиографии и приложений. Общий объем диссертации 102 страниц, из них 88

страниц текста, включая 23 рисунка. Библиография включает 163 наименования на 12 страницах.

1.7. Список используемых сокращений и обозначений

АС — альтернативный сплайсинг

АТФ — аденозинтрифосфат

ДНК — дезоксирибонуклеиновая кислота

мРНК — матричная РНК

мяРНК — малая ядерная РНК

гяРНП — гетерогенный ядерный рибонуклеопротеин

нт — нуклеотид

МСНП — методы секвенирования нового поколения

кДНК — комплементарная ДНК

РНК — рибонуклеиновая кислота

ССТФ — сайт связывания ТФ

ТФ — транскрипционный фактор

ФС — фактор сплайсинга

ЧП — число прочтений

ОЛМ — обобщённая линейная модель

НТО — нетранслируемая область

ПСК — преждевременный стоп-кодон

ЧВ — частота включения

ПФК — префронтальная кора (головного мозга)

КМ — кора мозжечка

РНК-Сек — секвенирование РНК при помощи МСНП

НД — набор данных

ПЦР — полимеразная цепная реакция

ПВМ — позиционная весовая матрица

2. Обзор литературы

2.1. Регуляция биосинтеза мРНК у эукариот

Информация об аминокислотной последовательности всех белков организма закодирована последовательностью нуклеотидов в хромосомной ДНК [Crick, 1970]. В ядре эукариотической клетки ДНК не существует в свободном виде, а связана различными белками в структуру, называемую хроматином. Основными белками, образующими хроматин, являются гистоны. Хроматин не только механически организует ДНК, но и участвует во всех процессах, затрагивающих ДНК, в первую очередь в репликации и транскрипции. Активно транскрибируемые участки хроматина развёрнуты и называются эухроматином, в то время как участки, на которых транскрипция подавлена, компактно свёрнуты и называются гетерохроматином [Orphanides, Reinberg, 2002; Svejstrup, 2004].

У эукариот реализация генетической информации, заложенной в ядерной ДНК, происходит с помощью транскрипции, процессинга (кэпирования, полиаденилирования и сплайсинга), РНК-редактирования, ядерно-цитоплазматического транспорта и трансляции. Эти процессы сложно взаимодействуют друг с другом, и каждый из них может регулироваться клеткой [Orphanides, Reinberg, 2002]. Использование альтернативных промоторов при

транскрипции [Singer и др., 2008], альтернативных сайтов полиаденилирования [Miura и др., 2013] и альтернативного сплайсинга [Chow и др., 1977; Early и др., 1980] позволяют одному гену производить несколько мРНК (или изоформ) и, соответственно, несколько различных белков. Для количественного описания работы гена принято использовать (с некоторыми вариациями) две меры: уровень экспрессии гена (определяется как сумма концентраций всех кодируемых геном мРНК [Wagner, Kin, Lynch, 2012]) и частота изоформы (определяется как концентрация данной мРНК, делённая на уровень экспрессии всего гена [Katz и др., 2010]). Клетка контролирует уровень экспрессии гена при помощи регуляции скорости синтеза и/или деградации мРНК. Частоты изоформ контролируются клеткой за счёт регуляции альтернативного сплайсинга, использования альтернативных промоторов и/или сайтов полиаденилирования, изоформ-специфичной деградации.

2.1.1. Регуляция транскрипции

У эукариот транскрипция всех белок-кодирующих генов, а также многих некодирующих, осуществляется РНК-полимеразой II. РНК-полимераза II состоит из 10-12 субъединиц и способна к ДНК-зависимому синтезу РНК, однако не способна распознавать промоторные последовательности. Распознавание промотора происходит при помощи дополнительных белков. Промоторы белок-кодирующих генов у эукариот могут быть грубо разделены на две группы [Carninci и др., 2006]. К первой относятся сайты содержащие TATA-бокс (AT-богатый участок), такие промоторы эволюционно консервативны и инициируют транскрипцию с точно (в пределах четырёх нуклеотидов) фиксированного сайта. Лишённые TATA-бокса промоторы обогащены динуклеотидами ЦГ, быстрее эволюционируют и их сайты инициации транскрипции длиннее. TATA-бокс связывается TATA-связывающим белком, который способен взаимодействовать с

РНК-полимеразой II и привлекать её к промотору. Дополнительные регуляторные последовательности связываются специальными белками — транскрипционными факторами (ТФ), способными прямо или через ко-регуляторы взаимодействовать с РНК-полимеразой II [Lee, Young, 2000].

Транскрипция может быть разбита на восемь стадий: 1) декомпактизация хроматина и обеспечение доступа к промотору; 2) привлечение РНК-полимеразы II, основных факторов транскрипции и сборка пре-инициаторного комплекса; 3) раскручивание ДНК и начало транскрипции; 4) освобождение промотора полимеразой и задержка полимеразы, кэпирование; 5) фосфорилирование С-терминального домена полимеразы и переход к элонгации; 6) элонгация транскрипции; 7) терминация транскрипции; 8) ре-инициация транскрипции (быстрый переход РНК-полимеразы II с терминатора транскрипции на промотор) [Fuda, Ardehali, Lis, 2009]. Фактически каждая из описанных стадий может регулироваться, активироваться или подавляться клеткой в ответ на различные стимулы.

В основе регуляции транскрипции лежит способность ТФ специфически связывать определённые последовательности ДНК [Neph и др., 2012; Wang и др., 2012]. ТФ могут быть как активаторами (способствуют транскрипции гена), так и репрессорами (подавляют транскрипцию гена), при этом один и тот же ТФ может быть активатором для одного гена и репрессором для другого [Huang и др., 1995]. ТФ могут связываться с ДНК как непосредственно около промотора, так и в энхансерах или сайленсерах — участках ДНК, способных активировать или подавлять экспрессию генов, чьи промоторы находятся на расстоянии в десятки тысяч пар нуклеотидов [Lee, Young, 2000]. У эукариот сайты связывания ТФ (ССТФ), как правило, короткие (5—10 нт) и вырожденные [Matys и др., 2006], поэтому в геноме существует множество мест, где данный ТФ мог бы связаться.

Специфичность связывания ТФ достигается, по всей вероятности, за счёт кооперации между несколькими молекулами ТФ, как одного, так и разных типов, связывающихся в одном месте [Zinzen и др., 2009]. Регуляция транскрипции может приводить как к общему изменению уровня экспрессии гена, так и к изменению соотношения изоформ за счёт активации альтернативных промоторов [Singer и др., 2008].

В большинстве случаев для активации или подавления транскрипции ТФ нуждаются в ко-факторах — в ко-активаторах или ко-репрессорах. ТФ и их ко-факторы используют множество механизмов для регуляции транскрипции: так, они могут непосредственно взаимодействовать с РНК-полимеразой II или основными ТФ, могут модифицировать (например, фосфорилировать) полимеразу или другие ТФ, модифицировать (метилировать, ацетилировать, убиквитинировать и фосфорилировать, а также деметилировать [Kooistra, Helin, 2012] и деацетилировать [Ruijter de и др., 2003]) гистоны или катализировать их диссоциацию от ДНК. Кроме того, роль ТФ и ко-факторов может состоять в привлечении других белков, осуществляющих какую-либо из перечисленных выше функций.

При активации гена, находящегося в состоянии гетерохроматина, ТФ должен быть способным связываться с конденсированным хроматином и переводить его в эухроматин (как, например, глюкокортикоидный рецептор [Orphanides, Reinberg, 2002]). Далее ТФ (сам или через ко-факторы) привлекает РНК-полимеразу II и способствует созданию пре-инициаторного комплекса. ТФ также влияют на инициацию элонгации, привлекая ко-активаторы, модифицирующие гистоны и катализирующие их диссоциацию с ДНК. После начальной инициации полимераза, как правило, останавливается через 30-100 нт после старта транскрипции [Kwak и др., 2013]. Во время этой паузы происходит кэпирование

транскрипта. Продолжение элонгации не происходит автоматически после кэпирования и, как правило, требует дополнительной активации. Для человека и плодовой мушки показано, что многие гены регулируются именно на этой стадии. В этом случае наиболее стохастическая стадия — сборка пре-инициаторного комплекса, требующая взаимодействия десятков белков, — уже пройдена и активация происходит быстрее и синхроннее. Считается, что таким образом достигается большая скоординированность активации набора генов в группах клеток. В случае регуляции на этой стадии активаторы способствуют фосфорилированию С-терминального домена РНК-полимеразы II [Kobor, Greenblatt, 2002] или фактора подавления элонгации (NELF). ТФ также могут способствовать элонгации посредством модификации хроматина. Так, некоторые модификации (H3K36me3, H3K4me2, H3K4me3 и H3K9ac) связывают с активацией транскрипции, в то время как другие (H3K9me2, H3K9me3 и H3K27me3) — с подавлением [Kornblihtt и др., 2013]. Кроме того, ТФ могут способствовать реинициации транскрипции, удерживая основные факторы инициации транскрипции на промоторе [Weake, Workman, 2010].

Для того, чтобы ТФ могли регулировать транскрипцию в ответ на стимул, необходимо, чтобы стимул менял активность ТФ. Это может достигаться за счёт активации экспрессии самого ТФ, за счёт активации ТФ при помощи связывания его с лигандом или его ковалентной модификации - фосфорилирования или убиквитинирования (интересно, что в последнем случае модификация ограничивает время действия активатора, направляя его на деградацию), либо изменением локализации ТФ с цитоплазменной на ядерную, что также может достигаться за счёт фосфорилирования [Orphanides, Reinberg, 2002].

Кроме активации ТФ могут обеспечивать также и репрессию. Репрессор может подавлять активатор (например, у дрожжей Gal80 связывается с

активатором Gal4 и подавляет его взаимодействие с ко-факторами [Weake, Workman, 2010]) или связываться с участком ДНК, пересекающим сайт связывания активатора (например, Acr1). Репрессор может связываться с TATA-связывающим белком и вызывать его диссоциацию (например, Mot1) или способствовать репрессивным модификациям хроматина (например, комплекс mSin3A или белок NuRD деацетилируют гистоны) [Lee, Young, 2000].

Кроме белков, в регуляции транскрипции могут участвовать некодирующие РНК. Так, например, инактивация одной из копий Х-хромосомы у млекопитающих происходит при помощи длинной некодирующей РНК Xist [Chow и др., 2005]). Длинные некодирующие РНК могут участвовать в модификациях хроматина, активации транскрипции через взаимодействие с ТФ и/или РНК-полимеразой II или в посттранскрипционной регуляции [Mercer, Dinger, Mattick, 2009]. Было также показано, что короткие (около 20 нт) двуцепочечные РНК, комплементарные промоторным участкам или сайтам внутри гена, могут вызывать как активацию транскрипции, так и её подавление через Ago2-зависимую модификацию хроматина [Li и др., 2006].

Последней стадией транскрипции является терминация, которая тесно связана с формированием 3'-конца транскрипта. Зрелые мРНК большинства белок-кодирующих генов получаются за счёт обрезания транскрипта и его полиаденилирования. Обрезание пре-мРНК происходит между консервативным гексануклеотидом AAУAAA и УГ-богатым участком, как правило, по динуклеотиду ЦА. Хотя точный механизм терминации транскрипции неизвестен, считается, что он связан с полиаденилированием [Proudfoot, Furger, Dye, 2002]. Многие гены имеют несколько сайтов полиаденилирования, и их выбор может регулироваться [Miura и др., 2013] белковыми факторами, вторичными

структурами в пре-мРНК и модификациями хроматина [Di Giammartino, Nishida, Manley, 2011; Luco и др., 2011]. Однако на данный момент регуляция полиаденилирования изучена слабо.

2.1.2. Регуляция альтернативного сплайсинга

У эукариот полученная в результате транскрипции пре-мРНК, как правило, состоит из экзонов — участков пре-мРНК, которые войдут в зрелую мРНК, и интронов — участков, которые должны быть вырезаны перед экспортом мРНК в цитоплазму. Процесс вырезания интронов и сшивания экзонов называется сплайсингом [Chow и др., 1977]. Границы интронов определяются сайтами сплайсинга — консервативными нуклеотидными последовательностями. 5'-конец интрона ограничен донорным сайтом, а 3'-конец интрона — акцепторным. Донорный и акцепторный сайты имеют следующие консенсусные последовательности: (А/Ц)АГ|ГТ(А/Г)АГТ и $(Т/Ц)_{н}$НЦАГ|Г, где '|' обозначает экзон-интронную границу [Mount, 1982]. Донорному сайту предшествует полипиримидиновый тракт $(Т/Ц)_{n}$. Ближе к акцепторному сайту внутри интрона располагается точка ветвления, она содержит аденин, осуществляющий нуклеофильную атаку на первый нуклеотид интрона, с которой начинается вырезание интрона.

Если сплайсинг одной пре-мРНК может идти несколькими путями, то говорят об альтернативном сплайсинге (АС). АС очень распространён у высших эукариот, 95% генов человека подвержены АС [Pan и др., 2008]. Выделяют четыре простых типа АС: альтернативный донорный или акцепторный сайт (простое удлинение/укорочение интрона за счёт выбора между одним из двух альтернативных сайтов), кассетный экзон (может либо включаться в мРНК, либо исключаться вместе с фланкирующими интронами) и удержанный интрон (интрон который может вырезаться или не вырезаться). Более сложные типы АС являются

комбинациями простых.

Сплайсинг пре-мРНК осуществляется сплайсосомой — молекулярной машиной, состоящей из пяти малых ядерных РНК (мяРНК) и более чем 200 белков. Сборка сплайсосомы происходит одновременно с распознаванием интрона непосредственно на пре-мРНК. На первой стадии мяРНК U1, фактор сплайсинга (ФС) 1 и вспомогательный белок U2AF связываются с донорным сайтом, сайтом ветвления и с акцепторным сайтом сайтом соответственно. Получившаяся структура называется Е-комплексом. На этой стадии отдельные компоненты сплайсосомы ещё не взаимодействуют друг с другом и в основном сконцентрированы около экзонов (поскольку, во всяком случае у многоклеточных животных, экзоны гораздо короче интронов), поэтому на этой стадии происходит так называемое распознавание экзонов [Chen, Manley, 2009]. Чтобы перейти к сплайсингу, компонентам сплайсосомы необходимо соединиться таким образом, чтобы между ними оказался интрон, который будет впоследствии вырезан. Этот переход называется распознаванием интрона и осуществляется за счёт замены ФС1 на мяРНК U2 (переход к А-комплексу) и сопряжённому с присоединением трёх мяРНК (U4, U5 и U6) переходу к В-комплексу. На этой стадии интрон, который будет вырезан, уже окончательно определён. Далее В-комплекс претерпевает конформационные изменения и переходит в каталитически активный С-комплекс [Chen, Manley, 2009].

Считается, что регуляция сплайсинга в основном осуществляется либо на стадии распознавания сайтов сплайсинга, либо на стадии определения интрона. Наиболее изученным механизмом регуляции сплайсинга являются регуляция при помощи ФС, способных специфически связываться с определёнными регуляторными последовательностями на молекуле пре-мРНК [Irimia, Blencowe, 2012]. Регуляторные последовательности в зависимости от их положения и

способности активировать или подавлять данный сайт сплайсинга делят на четыре группы: экзонные или интронные энхансеры или сайленсеры. На данный момент известно несколько классов ФС. Во-первых, это серин-аргинин-содержащие белки (SR-белки). SR-белки в основном связываются с определёнными последовательностями РНК в экзонных энхансерах и, взаимодействуя со сплайсосомой своим RS-доменом, повышают включение данного экзона в мРНК. Вторым классом ФС являются гетерогенные ядерные РНК-белковые комплексы — рибонуклеопротеины (гяРНП). гяРНП, как правило, связываются с сайленсерами (интронными и экзонными) и подавляют сплайсинг либо за счёт конкуренции с SR-белками и компонентами сплайсосомы за сайты связывания, либо за счёт изменения конформации пре-мРНК. Некоторое количество ФС (такие, как NOVA, FOX1 и FOX2, nPTB и другие) не относятся ни к одному из перечисленных выше классов [Chen, Manley, 2009]. Интересно, что эффект многих ФС зависит от положения сайта связывания относительно экзона. Так, связывание NOVA, гяРНП C, L и H, Fox, PTB и Mbnl1 в экзоне или в интроне до него приводит к подавлению включения, а связывание NOVA, гяРНП L, Fox, PTB, Mbnl1 и TIA в интроне после экзона — к увеличению частоты включения экзона [Ule и др., 2006; Witten, Ule, 2011; Zhang и др., 2010]. Позиционная зависимость может объясняться тем, что эти факторы, связываясь внутри альтернативного экзона, маркируют его как интрон.

Многочисленные исследования указывают на то, что сборка сплайсосомы и, в меньшей степени, вырезание интронов происходят ко-транскрипционно, то есть пока пре-мРНК еще не отделилась от хроматина [Tilgner и др., 2012; Luco и др., 2011]. Было показано, что РНК-полимераза II необходима для нормального сплайсинга, а сплайсинг мРНК, синтезированных другими полимеразами, существенно подавлен [Luco и др., 2011]. Считается, что ФС связываются с С-концевым доменом РНК-полимеразы II и таким образом получают возможность

взаимодействовать с сайтами сплайсинга непосредственно сразу после их синтеза. В экспериментах с ядерными экстрактами было показано, что SR-белки активируют сплайсинг, если он идёт ко-транскрипционно, но не в том случае, если пре-мРНК добавлена в экстракт извне. Считается, что это связано с неспецифическим взаимодействием РНК с гяРНП, которые подавляют связывание мРНК с компонентами сплайсосомы [Kornblihtt и др., 2013].

Ко-транскрипционная природа сплайсинга позволяет предположить связь между структурой хроматина и регуляцией АС. Действительно, было показано, что нуклеосомы (а также некоторые их модификации и метилирование ДНК) чаще встречаются в экзонах, чем в интронах, и что состояние хроматина влияет на сплайсинг [Andersson и др., 2009; Irimia, Blencowe, 2012; Luco и др., 2011]. Существуют две модели, объясняющие влияние хроматина на сплайсинг: привлечение ФС и кинетическая модель. Согласно первой, модифицированные гистоны прямо или опосредованно взаимодействуют с ФС и направляют таким образом АС. Известно, например, что триметилированный по 36-ому лизину третий гистон (H3K36me3) взаимодействует с PTB через белок MRG15; гистоны H3K4me3 и H3K9me3 привлекают мяРНК U2 и гяРНП через вспомогательные белки CHD1 и HP1 соответственно; а ацетилирование третьего гистона вызывает его связывание с мяРНК U2 через белок Gcn5 [Luco и др., 2011]. Согласно кинетической модели, влияние хроматина на сплайсинг осуществляется через модуляцию скорости элонгации РНК-полимеразы II. Медленно двигающаяся РНК-полимераза II даёт больше времени только что синтезированному акцепторному сайту на сборку сплайсосомы до того, как следующий, возможно конкурентный, сайт будет синтезирован. Это приводит к тому, что вырезается наиболее короткий интрон из возможных, а частота включения экзонов повышается. Многочисленные эксперименты с замедленной РНК-полимерзой II (при помощи мутации, УФ-обработки или ингибиторов) действительно показывают увеличение частоты

включения альтернативных экзонов в зрелые мРНК [Kornblihtt и др., 2013]. Вероятно, что оба варианта взаимодействия структуры хроматина и сплайсинга играют роль в регуляции АС. Интересно, что в некоторых случаях было показано, что не только хроматин влияет на сплайсинг, но и сплайсинг может влиять на структуру хроматина. Например, АС привлекает H3K36-метилтрансферазу (SETD2), а связывание белка Hu с пре-мРНК вызывает гиперацетилирование гистонов [Irimia, Blencowe, 2012].

Считается, что тканеспецифичная регуляция АС происходит в основном за счёт разных уровней экспрессии ФС в различных тканях. Так, например, было показано, что два органа с максимальным разнообразием АС, семенники и головной мозг, имеют наиболее специфические паттерны экспрессии ФС [Grosso и др., 2008]. В некоторых случаях регуляция ФС так же, как и ТФ, может достигаться за счёт пост-трансляционных модификаций и/или смены локализации белка в клетке. Интересно, что изменение уровней экспрессии не только ФС, но и базовых элементов сплайсосомы (таких как мяРНК) может регулировать тканеспецифичный сплайсинг [Chen, Manley, 2009].

Регуляция при помощи ФС осуществляется за счёт их связывания с регуляторными последовательностями на пре-мРНК, это позволяет использовать алгоритмы машинного обучения для предсказания тканеспецифичного сплайсинга, исходя из наличия или отсутствия тех или иных нуклеотидных мотивов около альтернативного экзона. Несмотря на использование более чем 1000 различных характеристик РНК, подобные алгоритмы позволяют предсказать только направление изменения, причём со сравнительно небольшой чувствительностью — не более 60% [Barash и др., 2010]. В другой работе авторам удалось разработать метод, позволяющий количественно оценить эффект мутаций на АС в человеке [Xiong и др., 2014]. Однако этот метод не позволяет понять принцип регуляции

АС, так как он использует эволюционную консервативность пре-мРНК — информацию, недоступную сплайсосоме — в качестве одной из характеристик. Таким образом, существующие модели не позволяют объяснить регуляцию АС при помощи только регуляторных последовательностей, расположенных в непосредственной близости от альтернативного экзона. Вероятно, другие факторы, такие как состояние хроматина или связывание белков на существенном удалении от альтернативного экзона, также играют существенную роль в регуляции АС. Это косвенно подтверждается тем, что на данный момент известно всего 416 кодируемых в геноме человека РНК-связывающих белков [Cook и др., 2010] и только 72 из них способны селективно связываться с определёнными нуклеотидными последовательностями [Giulietti и др., 2012]. Это значительно ниже числа известных ТФ (около 2500) [Chen, Manley, 2009].

Хотя считается, что ФС и, в меньшей степени, хроматин, являются основными регуляторами АС, недавно было показано участие вторичных структур РНК [Pervouchine и др., 2012], а также некодирующих РНК [Khanna, Stamm, 2010] в АС. В некоторых случаях выбор альтернативного промотора определяет выбор донорного сайта, например, в случае гамма-протокадгеринов [Morishita, Yagi, 2007]. В некоторых случаях, если сайт полиаденилирования находится в интроне, возможна конкуренция между сплайсингом и терминацией транскрипции, в которой ключевую роль может играть скорость элонгации.

2.1.3. Регуляция деградации мРНК

Внутриклеточная концентрация молекул мРНК определяется скоростью не только их синтеза, но и их разложения. Деградация мРНК в клетке идёт не самопроизвольно, а осуществляется специальными ферментами — РНКазами, которые разделяются на два класса: эндо-РНКазы — ферменты, способные расщеплять фосфодиэфирную связь, находящуюся в молекуле РНК далеко от края,

и экзо-РНКазы — ферменты, способные отщеплять крайний нуклеотид от молекулы РНК. Последние делятся на два типа в зависимости от того, с какого конца молекулы РНК они способны отщеплять нуклеотид: 3'–5' и 5'–3'. Система деградации РНК служит многим целям. В первую очередь, это гидролиз нефункциональных РНК (например, вырезанных интронов), который необходим для вторичного использования нуклеотидов и РНК-связывающих белков, и защита от вирусов [Houseley, Tollervey, 2009]. Система деградации РНК чрезвычайно эффективна, любая незащищённая молекула РНК в клетке быстро гидролизуется. Поэтому молекулы мРНК имеют специальную защиту: 3'-конец мРНК защищён от 3'–5' экзоРНКаз полиаденинoвым хвостом, который связывается белком PABP, 5'-конец мРНК защищён кэпом. Многочисленные исследования показывают, что время жизни мРНК зависит от ее типа, а также от состояния клетки и внешних стимулов [Sharova и др., 2009], соответственно скорость деградации мРНК может регулироваться клеткой и зависит от последовательности самой мРНК. Скорость деградации различных мРНК одного гена может регулироваться по-разному, изменяя таким образом соотношения изоформ [Salomonis и др., 2010].

Хотя в некоторых исследованиях было показано, что мРНК может подвергаться гидролизу в ядре и что этот процесс важен для регуляции экспрессии генов при созревании нейронов [Yap и др., 2012], считается, что в большинстве случаев разложение мРНК происходит в цитоплазме. Деградация мРНК требует снятия с неё защиты, это может осуществляться либо за счёт разрезания мРНК эндо-РНКазой с образованием двух незащищенных концов, либо за счёт удаления полиаденинового хвоста и/или за счёт удаления кэпа. Незащищенный 5'-конец РНК гидролизуется экзо-РНКазой XRN1, а 3'-конец разрушается экзосомой — белковым комплексом, обладающим экзо-РНКазной активностью [Nicholson и др., 2009]. Наиболее изученными являются два механизма регуляции деградации мРНК: микроРНК-зависимый и разложение, вызванное преждевременным стоп-

кодоном.

2.1.3.1. микроРНК

МикроРНК — это короткие (22 нт) молекулы РНК, способные комплементарно взаимодействовать с мишенями — молекулами мРНК, — подавляя их трансляцию и/или вызывая деградацию. МикроРНК синтезируются в клетке из длинных молекул-предшественников, которые, в свою очередь, синтезируются РНК-полимеразой II. Молекула-предшественник может содержать одну или несколько шпилек (участков двухцепочечной РНК), эти шпильки вырезаются при помощи эндо-РНКаз Drosha и/или Dicer, далее одна из двух цепей связывается с одним из белков семейства AGO, образуя активный комплекс RISC [Huntzinger, Izaurralde, 2011].

Механизм действия микроРНК основан на комплементарном связывании с молекулой мРНК. У многоклеточных животных сайт связывания микроРНК, как правило, находится в 3'-нетранслируемой области (НТО) и комплементарен первым 5–7 нт микроРНК. Хотя ранее считалось, что у животных микроРНК в основном действует через подавление трансляции, на данный момент показано, что в большинстве случаев микроРНК вызывает деградацию мРНК-мишени. Если сайт связывания комплементарен микроРНК вплоть до двенадцатой позиции, возможно эндонуклеазное разрезание молекулы мРНК белком AGO2. Однако у многоклеточных животных этого, как правило, не происходит из-за низкой комплементарности [Wu, Belasco, 2008]. В этом случае деградация мРНК происходит по другому механизму. Связавшись с мишенью, комплекс RISC взаимодействует с белком PABP через белок Gw182 и вызывает деаденилирование мРНК при помощи белков PAN2/PAN3 и CCR4/CAF1 и удаление кэпа при помощи белков DCP1/DCP2. Далее молекула мРНК подвергается гидролизу с обоих концов [Huntzinger, Izaurralde, 2011].

Считается, что в геноме человека содержится около 500 микроРНК, причём их количество сильно выросло в ходе эволюции приматов, и что около половины всех генов находится под непосредственной регуляцией микроРНК [Meunier и др., 2013]. Исследования показали, что уровни экспрессии микроРНК антикоррелируют с уровнями экспрессии их мишеней. Было показано участие микроРНК в нормальном развитии организма и в болезнях, например, в раке [Somel и др., 2011]. Концентрации самих микроРНК могут регулироваться при помощи регуляции РНК-полимеразы II, описанной выше, альтернативного сплайсинга молекул-предшественников [Chang и др., 2015], или за счёт регуляции белков, участвующих в созревании микроРНК [Yang, Wang, 2011]. Кроме того, недавно было показано, что микроРНК могут регулироваться другим классом молекул РНК — циклическими РНК, образующимися в результате нелинейного сплайсинга. Такие РНК устойчивы к деградации, так как не имеют свободного конца. Было показано, что циклическая РНК CDR1as содержит 74 сайта связывания микроРНК miR-7 и является её антагонистом: miR-7, связываясь с CDR1as, теряет возможность вызывать деградацию своих мишеней [Hansen и др., 2013]. Биоинформатический анализ позволил обнаружить около 2000 циклических РНК в транскриптоме человека [Memczak и др., 2013].

2.1.3.2. Разложение мРНК, вызванное преждевременным стоп-кодоном

Многочисленные исследования показали, что появление в мРНК преждевременного стоп-кодона (ПСК) может приводить к деградации мРНК [Chang, Imam, Wilkinson, 2007; Karam и др., 2013; Nicholson и др., 2009]. ПСК может появляться в мРНК как вследствие изменений генома (точечных мутаций, вставок или удалений нуклеотидов, перестроек), так и в результате АС. Считается, что до 30% транскриптов альтернативно сплайсируемых генов содержат ПСК и подвергаются деградации. Распознавание ПСК связано с трансляцией. Считается,

что основной фактор ПСК-зависимой деградации, белок UPF1, вместе с киназой SMG1 непосредственно связывается с факторами терминации трансляции eRF1 и eRF3 и образует комплекс SURF. При этом UPF1 конкурирует с PABP, поэтому большое расстояние между стоп-кодоном и полиадениновым хвостом является одним из факторов, способствующих деградации мРНК. Далее SMG1 фосфорилирует С-концевой серин-глутаминовый мотив белка UPF1. Считается, что фосфорилирование сильно ускоряется белками UPF2 и UPF3, хотя показано, что оно может происходить и в их отсутствии. Белки UPF2 и UPF3 способны связываться с белковым комплексом, который остаётся на месте вырезанного интрона и состоит из белков Y14, MAGOH, eIF4A3 и Barentsz, поэтому наличие экзон-экзонной границы после ПСК стабилизирует взаимодействие комплекса SURF с белками UPF2 и UPF3 и способствует деградации мРНК. Фосфорилирование UTF1 вызывает его диссоциацию с eRF3 и делает UTF1 способным связываться либо с комплексом SMG5/SMG7, либо с белком SMG6, который, в свою очередь, вызывает дефосфорилирование UTF1 при помощи белка PP2A. Дальнейший механизм деградации зависит от того, какой из SMG-белков связался с UTF1. Белок SMG6 является эндонуклеазой, его связывание с UTF1 приводит к разрезанию мРНК и последующей деградации обоих кусков от точки разреза. Комплекс SMG5/SMG7 привлекает CCR4/CAF1 и DCP1/DCP2, которые деаденилируют и декэпируют мРНК, делая ее чувствительной к РНазам [Nicholson и др., 2009].

ПСК-зависимая деградация мРНК участвует в регуляции экспрессии генов. Управление этим процессом может осуществляться за счёт регуляции уровней экспрессии факторов ПСК-зависимой деградации. Например, было показано, что miR-128 повышает концентрации мРНК генов, связанных с функционированием нервной системы, за счёт подавления экспрессии гена UTF1. Кроме того, к появлению ПСК может приводить АС. Таким образом, АС может вызывать

деградацию транскрипта. Интересно, что так регулируются многие ФС. Например, пропуск кассетного экзона в ФС PTBP1 приводит к появлению ПСК и деградации соответствующей мРНК [Karam и др., 2013]. Внутриклеточные концентрации мРНК факторов ПСК-зависимой деградации также регулируются за счёт обратной связи, так как сами являются мишенями деградации.

ПСК-зависимая деградация играет роль в развитии многих болезней, таких, как рак, мышечная дистрофия и талассемия. При этом в различных случаях ПСК-зависимая деградация может как усугублять болезнь, разрушая мРНК, способные кодировать хотя бы частично функциональный белок, так и облегчать симптомы болезни, предотвращая синтез ядовитых полипептидов. ПСК-зависимая деградация также участвует в разложении мРНК, кодирующих белки, содержащие селено-цистеин. Кодон УГА, кодирующий в некоторых случаях селено-цистеин, при недостатке последнего распознаётся как ПСК [Nicholson и др., 2009].

Таким образом, ПСК-зависимая деградация участвует в регуляции внутриклеточных концентраций мРНК как при болезни, так и в норме, в ответ на внешние стимулы, и взаимосвязана с процессами микроРНК-зависимой деградации и АС.

2.2. Методы массового анализа транскриптома

Развитие технологий в течение последних десятков лет позволило создать методы, позволяющие определять концентрации фактически всех мРНК в образце ткани или клеточной линии. Эти методы можно разделить на две группы: микрочипы и методы секвенирования нового поколения. Для обработки результатов применения этих методов были созданы многочисленные статистические подходы, позволяющие определять гены, которые имеют различные уровни экспрессии или сплайсинга в различных биологических

образцах.

2.2.1. Экспериментальные методы

За последние двадцать лет были разработаны многочисленные методы массового определения концентраций молекул ДНК заданной последовательности в образце. Хотя большинство этих методов может быть применено для анализа как генома, так и транскриптома, в данном обзоре акцент сделан в основном на последнем.

Все современные методы анализа транскриптома включают в себя стадию выделения РНК из образца. В зависимости от цели исследования используются те или иные фракции РНК. Они могут отличаться по локализации (вся клетка, цитоплазма, полисомы, ядро, хроматин и так далее) или по структуре самой РНК (длина молекулы РНК, наличие полиаденилирования). Для анализа уровней экспрессии белок-кодирующих генов обычно используют полиаденилированную РНК, выделенную из всей клетки (или из цитоплазматической фракции) [Dunham и др., 2012]. Из выделенной РНК получают комплементарную ДНК (кДНК) с помощью реакции обратной транскрипции. В дальнейшем анализе используется кДНК.

Исторически первым методом массового анализа транскриптома были микрочипы [Schena и др., 1995]. Принцип действия этого метода основывается на комплементарном взаимодействии между молекулой кДНК, полученной из анализируемого образца и помеченной флуоресцентной меткой, и молекулами ДНК, называемыми пробами, закреплёнными на твёрдой подложке. Определённые участки подложки заполнены пробами одного типа. Количество кДНК, связавшегося с данным участком подложки, может быть определено по интенсивности флуоресцентного свечения. Для определения уровней экспрессии

тех или иных мРНК используются пробы, комплементарные строго данной мРНК. Хотя в качестве проб могут использоваться молекулы ДНК, амплифицированные при помощи полимеразной цепной реакции (ПЦР) с геномной ДНК или кДНК, в большинстве случаев пробами являются короткие (менее 100 нт) синтезированные олигонуклеотиды [Katagiri, Glazebrook, 2009]. В зависимости от набора проб, микрочипы могут быть разделены на несколько типов: генные чипы, имеющие несколько проб к одному гену и позволяющие определять уровень экспрессии всего гена; экзонные чипы, содержащие несколько проб к каждому экзону и позволяющие оценивать альтернативный сплайсинг; и черепичные (tiling) чипы, пробы которых покрывают весь геном и позволяют анализировать уровни и экспрессии и сплайсинга даже в отсутствие аннотации [Hoheisel, 2006]. Кроме того, были разработаны специальные чипы, содержащие пробы, комплементарные экзон-экзонным границам и позволяющие проводить более точный анализ альтернативного сплайсинга [Fehlbaum и др., 2005].

Хотя микрочипы широко используются для анализа транскриптомов, чаще для определения экспрессии генов и реже для исследования АС, они обладают рядом недостатков. Так, микрочипы обладают сравнительно низкой чувствительностью и высоким уровнем шума, а кросс-гибридизация затрудняет анализ схожих мРНК — разных изоформ одного гена или мРНК гомологичных генов. Поэтому методы секвенирования нового поколения (МСНП) по мере уменьшения себестоимости постепенно вытесняют микрочипы из транскриптомных исследований [Metzker, 2009].

Основой МСНП является одновременное секвенирование миллионов кДНК. Как правило, кДНК фрагментируется, отдельные фрагменты изолируются на твёрдой подложке или в каплях эмульсии на микрогранулах и амплифицируются при помощи ПЦР, в результате чего копии одного фрагмента кДНК оказываются

изолированы в пространстве. Нуклеотидная последовательность фрагмента определяется при помощи синтеза комплементарной цепи, который может осуществляться как при помощи лигирования динуклеотидов (технология SOLiD), так и при помощи присоединения отдельных нуклеотидов ДНК-полимеразой (технологии Illumina, Roche/454, IonTorrent и Helicos). Контроль за синтезом осуществляется при помощи использования блокирующих групп на нуклеотидах (Illumina, Helicos и SOLiD) или добавлением нуклеотидов только одного типа (Roche/454 и IonTorrent). Добавление нуклеотида (и его тип) детектируется при помощи флуоресцентной метки (Illumina, Helicos и SOLiD) или по выделению молекулы пирофосфата (Roche/454 и IonTorrent) [Metzker, 2009]. Результатом применения МСПН являются миллионы последовательностей фрагментов кДНК с длиной от десятков (SOLiD, ранние версии Illumina) до сотен (Roche/454) нуклеотидов, называемых прочтениями (reads). МСНП существенно превосходят микрочипы по чувствительности, специфичности, отношению сигнал/шум и способности детектировать неизвестные ранее транскрипты. Применение МСНП для секвенирования клеточной РНК обычно называют секвенированием РНК или РНК-Сек. Описанные выше методы позволяют получать прочтения, которые существенно короче, чем средняя длина мРНК (не более 500 нт у IonTorrent, и не более 300 нт у Illumina). Поэтому в общем случае эти методы не позволяют восстановить полную последовательность транскрипта.

В последнее время получают распространение новые методы, такие как PacBio [Gao и др., 2016] и Oxford Nanopore [Loose, Malla, Stout, 2016], позволяющие получать более длинные прочтения — до нескольких тысяч нт. Однако оба этих метода обладают существенными недостатками: низкой точностью у Oxford Nanopore (более 30% нуклеотидов читаются неправильно [Laver и др., 2015]) и высокой ценой в случае PacBio.

2.2.2. Вычислительные методы

В результате применения МСНП или микрочипов получаются огромные массивы информации, обработать которые можно только при помощи компьютера. Анализ микрочипов включает в себя нормализацию сигнала, определение уровней экспрессии генов и/или экзонов и поиск генов и/или экзонов, значимо меняющих уровни экспрессии между образцами. Как правило, статистический анализ осуществляется в логарифмической шкале при помощи линейной модели и дисперсионного анализа [Smyth, 2005].

В случае МСНП первой стадией обработки данных является картирование — соотнесение прочтений с определёнными позициями генома. На данный момент создано множество методов картирования, часть из них позволяет находить вхождения прочтения только в геноме, в то время как другие дают возможность картировать прочтения и на экзон-экзонные границы. В последнем случае программы либо используют известные заранее координаты интронов, либо предсказывают их сами, исходя только из последовательности генома и набора прочтений [Garber и др., 2011]. На данный момент наиболее широко используемыми программами являются bowtie, картирующая прочтения на геном [Langmead и др., 2009], и tophat картирующая прочтения и на геном, и на экзон-экзонные границы. Tophat использует bowtie и может использовать как известные экзон-экзонные границы, так и предсказывать их de novo [Trapnell, Pachter, Salzberg, 2009]. Недавно авторами программы tophat была предложена новая более быстрая программа hisat2. hisat2 обладает существенными преимуществами по сравнению с tophat: кроме существенно большей производительности (hisat2 примерно в десять раз быстрее) hisat2 позволяет учитывать информацию о полиморфизмах при картировании [Kim, Langmead, Salzberg, 2015].

После картирования для каждого элемента генома (гена или экзона)

вычисляется число прочтений (ЧП), пересекающихся с ним, и эта величина используется для оценки уровня экспрессии.

2.2.2.1. Методы анализа экспрессии

Поскольку, в отличие от микрочипов, где использование линейных моделей и дисперсионного анализа является устоявшейся практикой [Smyth, 2005], МСНП является сравнительно молодой областью, я остановлюсь здесь в основном на методах анализа уровней экспрессии генов, исходя из данных МСНП.

Существуют два подхода к анализу экспрессии по данным МСНП. Во-первых, ЧП могут непосредственно использоваться для анализа. В этом случае используются дискретные распределения, например пуассоновское, обратное биномиальное или бета-биномиальное [Soneson, Delorenzi, 2013]. Во втором случае, ЧП нормируются на общее число картируемых прочтений в библиотеке и на длину гена. Полученные величины пропорциональны концентрациям мРНК в исходном образце [Wagner, Kin, Lynch, 2012] и могут использоваться методами, изначально разработанными для анализа микрочипов.

Методы, использующие непосредственно ЧП, такие, как DEseq [Anders, Huber, 2010], EdgeR [Robinson, McCarthy, Smyth, 2010] и другие, основываются на предположении, что прочтения являются случайной выборкой фрагментов мРНК из общего пула, представленного в образце и, соответственно, ЧП одного гена должны следовать распределению Пуассона. Действительно, дисперсия ЧП, наблюдаемая в технических повторах (секвенировании разных аликвот одного и того же образца), совпадает со средним, что подтверждает Пуассоновский характер распределения ЧП [Oberg и др., 2012]. Однако при сравнении образцов, полученных из различного биологического материала (например, разных особей одного вида), наблюдается так называемая избыточная дисперсия (превышение

дисперсией среднего). Избыточная дисперсия объясняется фенотипической вариабельностью доноров и наблюдается в большинстве реальных исследований. Использование распределения Пуассона для моделирования ЧП в случае наличия избыточной дисперсии ведёт к большому количеству ложно положительных предсказаний [Soneson, Delorenzi, 2013], поэтому возникает необходимость использования дискретных распределений с дисперсионным параметром. Так, три наиболее популярных метода (DEseq, EdgeR, cuffdiff 2 [Trapnell и др., 2012]) используют обратное биномиальное распределение. Все три метода используют обобщённую линейную модель (ОЛМ) и тест на лог-правдоподобие для оценки значимости изменений уровней экспрессии.

2.2.2.2. Методы анализа альтернативного сплайсинга

Полногеномный анализ альтернативного сплайсинга стал возможен только с появлением МСНП. Поэтому, в отличие от методов анализа экспрессии генов, которые разрабатывались ещё для микрочипов, количество методов вычислительного анализа АС не так велико. На данный момент предложено около двух десятков различных методов анализа дифференциального АС при помощи данных РНК-Сек, но многие из них либо не были реализованны в виде программного продукта, либо созданные программы не поддерживаются авторами и фактически не работают. Все известные мне работающие программы представлены в приложении 1. Использование двух из этих методов — Alexa-seq и juncBASE — затруднено, так как они реализованы в виде десятков отдельных подпрограмм.

Одним из важных признаков, различающих существующие методы, является объект исследования. Часть алгоритмов (например, MISO и cuffdiff 2) сравнивают целые транскрипты, в то время как в других методах сравниваются только альтернативные участки (метод JuncBASE), экзоны (методы MATS, Alexa-seq) или

даже части экзонов (методы DEXseq и JunctionSeq).

Использование транскриптов в качестве объекта исследования имеет два существенных недостатка. Во-первых, даже установив, что соотношение изоформ некого гена различно в различных тканях (или условиях), достаточно сложно определить, использование какого именно экзона изменилось, так как различные изоформы могут отличаться друг от друга на несколько экзонов. Однако для изучения последствий изменения альтернативного сплайсинга (например, доменного состава белка или наличия сайтов фосфорилирования в белке), или для поиска регуляторных последовательностей, ответственных за регуляцию АС данного экзона, это знание может быть необходимо. Во-вторых, из-за относительно маленьких длин прочтений данные МСНП не позволяют в общем случае определить структуру транскриптов. Например, ген с двумя альтернативными экзонами может породить четыре изоформы мРНК. Однако, если альтернативные экзоны находятся на достаточно большом расстоянии друг от друга (так, что одно прочтение не может пересечь оба из них), восстановить частоты всех четырёх изоформ невозможно [Trapnell и др., 2010]. Поэтому некоторые методы отказываются от использования транскриптов в качестве объекта исследования и используют экзоны или так называемые сегменты — участки гена между двумя ближайшими сайтами сплайсинга. В этом случае вместо частот транскриптов используются частоты включения сегментов (или экзонов) — доля транскриптов, включающих данный сегмент (или экзон), среди всех транскриптов гена. Существует два подхода к вычислению частоты включения: либо среднее покрытие (среднее число прочтений, картирующихся на один нуклеотид) сегмента прочтениями делится на среднее покрытие гена (DEXseq и Alexa-seq), либо число прочтений, подтверждающих включение данного экзона в транскрипт (то есть пересекающихся с экзоном), сравнивается с числом прочтений, подтверждающих исключение экзона из транскрипта (то есть

картирующихся на такую экзон-экзонную границу, так что экзон оказывается внутри вырезанного интрона). Такой подход используется в методах MATS и juncBASE. Поскольку в реальности прочтения выбираются из пула фрагментов мРНК неслучайно и их распределение зависит от экспериментальной процедуры и сильно варьирует между различными лабораториями [Khrameeva, Gelfand, 2012], первый подход представляется неоптимальным. К другим недостаткам существующих методов относится то, что многие из них не позволяют учитывать биологическую вариабельность (Alexa-seq, MISO, MATS и juncBASE) и/или могут осуществлять только парные сравнения (Alexa-seq, MISO, MATS, juncBASE и cuffdiff 2) и не пригодны для более сложного анализа (например, временных рядов).

Таким образом, все существующие методы анализа альтернативного сплайсинга обладают существенными недостатками, поэтому разработка нового метода, позволяющего работать с сегментами, учитывать биологическую вариабельность, использовать прочтения, картирующиеся на экзон-экзонные границы, и тестировать сложные модели, является актуальной задачей. Кроме того, следует отметить, что все упомянутые методы были опубликованы после начала данной работы, поэтому разработка нового метода анализа АС была необходимой частью данной работы.

2.3. Современная транскриптомика

Широкое применение микрочипов и МСНП позволило сделать измерение уровней экспрессии всех генов в образце рутинной процедурой и дало начало новому разделу вычислительной биологии — транскриптомике. Полногеномные сравнения уровней экспрессии генов в различных тканях млекопитающих показали, что межтканевые отличия в экспрессии консервативны и доминируют над межвидовыми различиями [Brawand и др., 2011; Chan и др., 2009]. Интересно,

что при аналогичном сравнении АС межвидовые различия доминируют над различиями между тканями [Barbosa-Morais и др., 2012; Merkin и др., 2012]. Это может быть связано с тем, что многие изоформы, получающиеся в результате АС, нефункциональны и, соответственно, их регуляция находится под ослабленным отбором [Pickrell и др., 2010; Sorek, Shamir, Ast, 2004].

Межвидовые сравнения уровней экспрессии генов в фиксированной ткани показали, что экспрессионная вариабельность соответствует филогенетической: деревья, построенные на основе схожести экспрессионных профилей (измеренных в одних и тех же тканях разных видов), фактически идентичны филогенетическим деревьям. При этом скорость эволюции уровней экспрессии может существенно различаться между тканями. Так, уровни экспрессии генов эволюционируют быстрее всего в семенниках и наиболее консервативны в мозге, особенно в коре [Brawand и др., 2011]. Сравнение профилей экспрессии в разных тканях человека и мыши показало, что разные ткани экспрессируют разное количество генов. Около 8000 генов экспрессируются повсеместно, при этом в скелетной мускулатуре и печени экспрессируется сравнительно немного генов (около 12000), а семенники и головной мозг лидируют по этому показателю (16000 и 14000 генов, соотстветственно) [Ramsköld и др., 2009]. Интересно, что те же самые ткани (семенники и головной мозг) лидируют по частоте альтернативного сплайсинга [Yeo и др., 2004].

Транскриптомные исследования позволили установить связь между уровнями экспрессии генов и/или АС и многими заболеваниями, такими, как рак [Vijver van de и др., 2002; David, Manley, 2010; Venables и др., 2009], диабет [Voisine и др., 2004], аутизм [Voineagu и др., 2011], шизофрения [Guillozet-Bongaarts и др., 2013] и многими другими. Совмещение транскриптомных исследований с генотипированием на больших выборках позволяет обнаружить нуклеотидные

полиморфизмы (различия в нуклеотидной последовательности ДНК у особей одного вида), связанные с уровнями экспрессий определённых генов. В некоторых случаях такие исследования позволяют найти участки ДНК, ответственные за регуляцию экспрессии, и установить механизм по которому та или иная мутация оказывает влияние на фенотип носителя [Franke, Jansen, 2009].

Совмещение измерений экспрессии генов с определением позиций связывания ТФ позволяют находить мишени ТФ и восстанавливать регуляторные сети [Gao, Foat, Bussemaker, 2004], а одновременное измерение концентраций микроРНК позволяет, с последующим использованием корреляционного анализа, улучшить предсказание мишеней микроРНК [Gennarino и др., 2009].

Особый интерес представляет такое активно развивающееся в последние годы направление, как транскриптомика отдельных клеток. В обычных транскриптомных экспериментах для получения достаточного количества РНК используют миллионы клеток. В результате измеряемые уровни экспрессии являются усреднением и не содержат информации о межклеточной вариабельности в концентрациях мРНК. Использование новых технологий, основанных на секвенировании и позволяющих определять уровни экспрессии генов в отдельной клетке, позволили обнаружить бимодальную экспрессию и сплайсинг во внешне гомогенной популяции эукариотических клеток [Shalek и др., 2013]. Секвенирование мРНК отдельных клеток, выделенных из сложных гетерогенных тканей, позволяют находить новые клеточные типы и/или определять пути клеточной дифференцировки [Jang и др., 2017; Karaiskos и др., 2017]. Совместное измерение концентраций белков и соответствующих мРНК в одиночных бактериальных клетках при помощи флуоресцентных меток подтвердило стохастичность процессов транскрипции и трансляции и показало отсутствие корреляции между концентрациями мРНК и белков в одной клетке

[Taniguchi и др., 2010].

Кроме работ, осуществляемых одной лабораторией, за последние десять лет были начаты (и достигли существенного успеха) несколько крупных межлабораторных проектов, поставивших себе целью систематически и максимально полно охарактеризовать экспрессию генов и ее регуляцию в некотором выбранном объекте. К таким проектам следует в первую очередь отнести ENCODE, направленный на изучение экспрессии, структуры хроматина и профилей связывания факторов транскрипции в различных клеточных линиях, полученных из человека [Dunham и др., 2012], а также modENCODE — аналог проекта ENCODE, направленный на изучение двух модельных организмов: плодовой мушки (*Drosophila melanogaster*) и нематоды (*Caenorhabditis elegans*) [Elsner, Mak, 2011]. Проект ENCODE позволил приписать функции более чем 80% человеческого генома, определить, в некоторых случаях с однонуклеотидным разрешением, сайты связывания более чем ста факторов транскрипции, обнаружить взаимодействующие регуляторные участки ДНК, выделить, на основании данных о модификациях гистонов и открытости хроматина, семь основных типов состояния хроматина (таких, как энхансер, промотор, репрессированный участок и т. д.). Хотя впоследствии заявление о функциональности 80% генома человека было подвергнуто критике на основании того, что (а) не более 15% генома человека эволюционно консервативно, а функциональность подразумевает консервативность и (б) наличие воспроизводимого биохимического свойства у данного участка ДНК (транскрипция, модификация гистонов, связывание ТФ и т. д.) ещё не обозначает функциональности [Doolittle, 2013; Graur и др., 2013], проект ENCODE существенно расширил наши знания о молекулярном строении тела человека.

Другим, не менее амбициозным проектом является Allen Brain Atlas [Sunkin и

др., 2012], поставивший себе целью составить полную пространственно-временную карту экспрессии генов головного мозга мыши и человека, совместив её с гистологическими и анатомическими данными. Первоначально проект основывался на гибридизации in situ гистологических срезов, другие методы, такие, как микрочипы и МСНП, были добавлены позже. Проект Allen Brain Atlas показал, что 90% генов экспрессируются различно в разных областях мозга и/или в разных точках развития, и что большинство различий проявляются до рождения, в то время как экспрессия во взрослом мозге более равномерна [Kang и др., 2011]. Были найдены гены с измененной экспрессией у больных шизофренией [Guillozet-Bongaarts и др., 2013], было показано, что гены со схожими пространственными паттернами экспрессии в головном мозге, как правило, имеют связанные функции [Liu и др., 2007]. В последней работе, опубликованной в рамках проекта Allen Brain Atlas, при помощи микрочипов были изучены изменения экспрессии генов в ходе пренатального (1212 образца) и постанатального (724 образца) развития мозга шимпанзе. В работе рассматривались пять отделов мозга (префронтальная кора, миндалина, полосатое тело, гипокамп и первичная зрительная кора), при этом для отделов коры были получены данные для отдельных слоёв; десять возрастов: пять до рождения (начиная с пятидесятого эмбрионального дня) и пять после рождения (до четырёх лет). Результаты показали, что экспрессионные паттерны слоёв коры головного мозга сильно отличаются друг от друга, однако происходящие в них возрастные изменения скоординированы. Большая часть возрастных изменений происходит до рождения, а в момент рождения происходит резкая смена регуляторной программы, при этом кора приобретает взрослые черты медленнее, чем остальные части мозга, что указывает на возможную роль внешних стимулов в процессе созревания коры. Авторы показали, что гены, мутации в которых повышают риск аутизма, экспрессируются в постмитотических нейронах как до, так и после рождения, а гены, связанные с шизофренией, — только после

рождения [Bakken и др., 2016].

Проект GTEx направлен на поиск генетических вариантов, оказывающих влияние на экспрессию генов в различных тканях человека. Для этого образцы десятков тканей, полученные от сотен доноров, подвергаются РНК-Сек и одновременно секвенируются геномы доноров [eGTEx Project, 2017].

2.3.1. Транскриптомика головного мозга человека

Недавнее широкомасштабное исследование префронтальной коры (ПФК) головного мозга методом микрочипов, охватывающее 269 доноров возрастом от 14 недель после зачатия до 80 лет, позволило восстановить возрастную динамику экспрессии большинства генов, транскрибирующихся на детектируемом уровне в данной ткани [Colantuoni и др., 2011]. Результаты показали, что изменения экспрессии, происходящие до рождения и в первые шесть месяцев жизни, превосходят по амплитуде изменения, наблюдаемые в последующие периоды, в более чем десять раз. Хотя направление изменений экспрессии многих генов совпадает до и после рождения (например, уровни экспрессии генов, вовлечённых в клеточный цикл, падают, а в синаптогенез — растут), экспрессия большинства генов достигает экстремума в момент рождения. Так, экспрессия генов, связанных с функционированием аксонов, растёт в ходе пренатального развития и начинает падать вскоре после рождения. Авторы исследования предполагают, что такие изменения связаны с созреванием мозга: удалением излишних аксонов и фиксацией синапсов. Экспрессия генов, вовлечённых в синтез АТФ, связывание кальция и работу ионных каналов, следует противоположному паттерну и достигает минимума в момент рождения, что, по всей вероятности, связано с растущими энергетическими потребностями развивающегося мозга. Интересно, что многие изменения экспрессии генов, наблюдаемые непосредственно после рождения, в значительной степени инвертируются в ходе старения, после 50 лет.

Другое исследование, с меньшим числом доноров, однако покрывающее 27 различных областей головного мозга, включая 11 зон неокортекса, показало, что во всех изученных зонах так же, как и в префронтальной коре, большинство изменений уровней экспрессии генов происходит до рождения. Хотя более 70% генов имеют значимо различную между какими-либо двумя областями мозга экспрессию (в один возрастной период), с возрастом эти различия убывают, особенно в неокортексе [Kang и др., 2011]. Эти результаты показывают важность изучения возрастной динамики экспрессии генов.

2.3.2. Сравнительная транскриптомика головного мозга приматов

Несмотря на большое анатомическое и генетическое сходство человека и высших приматов, таких, как шимпанзе, люди сильно отличаются от последних в социальном поведении и когнитивных способностях — функциях, за выполнение которых отвечает кора мозга [Klein, 2009]. Наравне с эволюцией белковых последовательностей, эволюция регуляции биосинтеза и деградации мРНК и белков играет существенную роль в видообразовании и представляет большой интерес [Khaitovich и др., 2006]. Поэтому сравнительный анализ уровней экспрессии и сплайсинга генов в мозге человека и других приматов является хорошим инструментом для понимания его природы и может помочь понять природу различных нервных расстройств, например, шизофрении [Crespi, Summers, Dorus, 2007].

Исследования мозга взрослых людей и других высших приматов показали, что с точки зрения экспрессии генов мозг является одной из самых консервативных тканей. Тем не менее, межвидовые отличия превосходят внутривидовую вариабельность, что позволяет обнаружить гены со специфичными для человека экспрессионными изменениями. Однако количество таких генов невелико и примерно равно числу генов с шимпанзе-специфичными

изменениями экспрессии. Хотя некоторые из генов с человеко-специфичными изменениями экспрессии, по всей вероятности, связаны с развитием когнитивных способностей (например, ТФ *FOXP2*, связанный с речью) и их регуляция, вероятно, эволюционировала под действием положительного отбора, большинство наблюдаемых изменений являются нейтральными [Somel, Liu, Khaitovich, 2013]. Таким образом, различия в экспрессии генов во взрослом мозге, по всей вероятности, не могут объяснить отличий человека от близкородственных приматов.

Однако при сравнении возрастных паттернов изменения экспрессии ситуация меняется. Оказывается, что возрастная регуляция экспрессии генов в коре мозга эволюционировала в линии человека гораздо быстрее, чем в линии шимпанзе, после их расхождения [Somel и др., 2011]. Как правило, изменения экспрессии, наблюдаемые в обезьянах, происходят в человеке с задержкой, что согласуется с неотенической теорией эволюции человека, согласно которой онтогенетическое развитие человека отстаёт от такового у обезьян [Somel и др., 2009]. Размер задержки различен у разных генов и в некоторых случаях может объясняться различиями в продолжительности жизни. Одним из ярких примеров транскрипционной неотении являются гены, связанные с развитием синапсов. Было показано, что у человека уровни экспрессии этих генов достигают максимума примерно в пять лет, в то время как у шимпанзе и макаки максимум приходится на первые месяцы жизни. Такой сдвиг в несколько раз превосходит ожидаемый, исходя из различий в продолжительности жизни. Интересно, что плотность синаптических контактов в мозгу этих трёх видов следует примерно такому же паттерну: растёт у человека вплоть до пяти лет, но падает в мозге обезьян практически с самого рождения [Liu и др., 2012]. Таким образом, возрастная регуляция экспрессии генов, а не экспрессия во взрослом состоянии, отличает мозг человека от мозга других приматов.

Так как МСНП появились сравнительно недавно, а микрочипы не позволяют анализировать АС с необходимой точностью, большинство транскриптомных исследований было посвящено экспрессии генов, и только в последние годы в печати начали появляться статьи, посвящённые полногеномному анализу АС. Так, в двух работах, опубликованных в журнале Science в конце 2012 года, был проведён сравнительный анализ АС в девяти тканях различных позвоночных, от лягушки до приматов [Barbosa-Morais и др., 2012; Merkin и др., 2012]. Результаты показали, что, в отличие от экспрессии генов, АС существенно варьирует между видами, и при этом межвидовые отличия обычно доминируют над межтканевыми. Эти результаты делают АС привлекательным объектом для поиска человеко-специфичных изменений, однако такая низкая консервативность ставит вопрос о функциональности большей части АС. К сожалению, этот анализ был ограничен только взрослыми особями. Более позднее мета-исследование АС хоть и включает в себя некоторое количество эмбриональных образцов, не покрывает постнатальное развитие и в основном базируется на образцах тканей, полученных от взрослых доноров [Tapial и др., 2017].

Лишь небольшое количество работ посвящено полногеномному анализу возрастных изменений АС в мозге млекопитающих. Например, при анализе мозга мыши было найдено 387 экзонов с частотами включения, значимо отличающимися между эмбрионом и взрослой особью, и показано, что гены, содержащие такие экзоны, связаны с цитоскелетом и передачей нервного импульса [Dillman и др., 2013]. В другом исследовании были показаны многочисленные, связанные с падением активности ФС PTB, изменения АС в ходе старения. Там же было показано, что нейродегенеративные заболевания сопровождаются изменениями АС, связанными с падением концентрации ФС NOVA [Tollervey и др., 2011]. В многочисленных исследованиях, посвящённых отдельным генам, была показана роль АС в нормальном (например, PTB-зависимый АС в ФС nPTB [Boutz и др.,

2007]) и патологическом (например, нарушение работы ФС MBNL1 при миотонической дистрофии, приводящее, в том числе, к патологическим изменениям АС в головном мозге [Charizanis и др., 2012]) развитии мозга, старении и нейродегенеративных заболеваниях (например, ген MAPT, вовлечённый в болезнь Альцгеймера [Niblock, Gallo, 2012]). Кроме того, известно, что АС играет роль в развитии и других органов, например сердца, скелетной мускулатуры, семенников и иммунной системы [Baralle, 2017].

Однако на данный момент не было проведено ни одного полногеномного исследования возрастной динамики АС на всей протяжённости жизни ни для человека, ни для других приматов.

3. Разработка метода анализа альтернативного сплайсинга

Так как все существующие на данный момент методы анализа АС на основе данных МСНП обладают определёнными недостатками, в настоящей работе был разработан новый алгоритм SAJR (Splicing Analyser by Java&R). Объектом анализа SAJR является сегмент — участок гена между двумя ближайшими сайтами сплайсинга, или между сайтом сплайсинга и сайтом инициации транскрипции или сайтом полиаденилирования. Сегменты, ограниченные двумя сайтами сплайсинга называются внутренними, сегменты, одной из границ которых являются сайт инициации транскрипции или сайт полиаденилирования, называются, соответственно, первыми или последними. При анализе сегмента SAJR не использует информацию о покрытии прочтениями всего гена и поэтому мало зависит от неравномерности покрытия. Благодаря использованию обобщённых линейных моделей (ОЛМ) с квази-биномиальным распределением и тестом на логарифм правдоподобия, SAJR позволяет проводить анализ сложных моделей и учитывать биологическую вариабельность.

SAJR состоит из двух частей. Во-первых, это java-приложение, позволяющее

найти число прочтений, пересекающихся с данным геном или сегментом или картировавшихся на данную экзон-экзонную границу. Вторым компонентом SAJR является пакет, написанный на языке R [R Core Team, 2013], позволяющий произвести статистический анализ. SAJR свободно доступен, его можно скачать с веб-страницы, расположенной по адресу http://storage.bioinf.fbb.msu.ru/~mazin/index.html. В общей совокупности SAJR содержит более 4500 строк программного кода. Схема анализа данных МСНП представлена на рисунке 1.

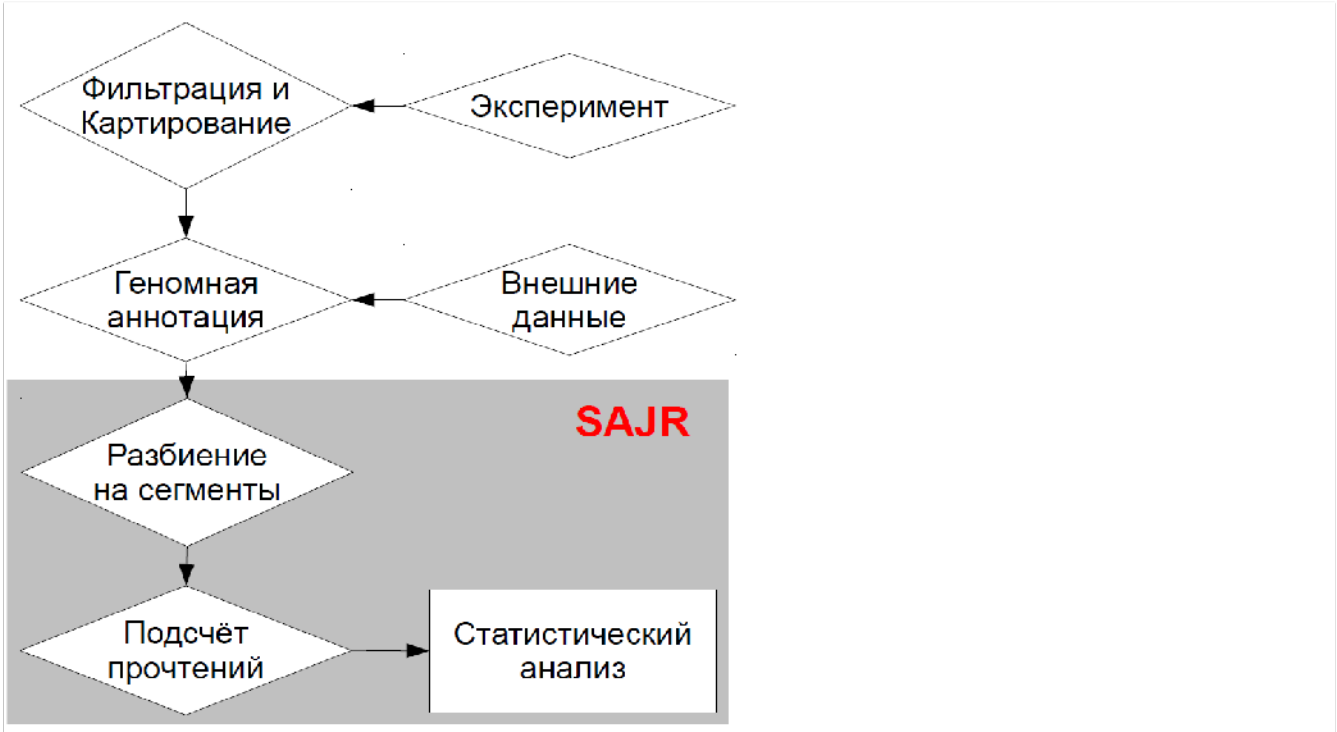

Рисунок 1.Схема анализа данных МСНП и метода SAJR. Полученные в результате эксперимента прочтения фильтруются и картируются на геном. Далее либо используется известная геномная аннотация, либо аннотация создаётся заново исходя из прочтений. Далее SAJR разбивает гены на сегменты, подсчитывает прочтения и проводит статистический анализ.

3.1. Подсчёт прочтений

Перед началом анализа все гены разбиваются на сегменты. Сегменты разделяются на три типа: константные сегменты, которые включаются во все изоформы, проходящие через них; альтернативные сегменты, которые включаются только в часть изоформ; и удержанные интроны, которые являются альтернативными сегментами, полностью совпадающими по геномным координатам с интроном (рис. 2 А). Первые (последние) сегменты считаются альтернативными, если в гене присутствует более одного первого (последнего) сегмента.

Для каждого сегмента вычисляется число прочтений, подтверждающих включение данного сегмента в транскрипт (то есть пересекающих его хотя бы по одному нт), и число прочтений, подтверждающих исключение сегмента из транскрипта (то есть картирующихся на границу между экзоном, находящимся до данного сегмента, и экзоном, находящимся после него). Далее такие прочтения будут называться включающими и исключающими, соответственно (рис. 2 Б). В случае первых (последних) сегментов включающими считались прочтения, картирующиеся на экзон-экзонную границу, соединяющую данный сегмент с остальным геном, а исключающими считались прочтения, являющиеся включающими для остальных первых (последних) сегментов Для анализа уровня экспрессии генов подсчитываются все прочтения, пересекающие хотя бы один константный экзон.

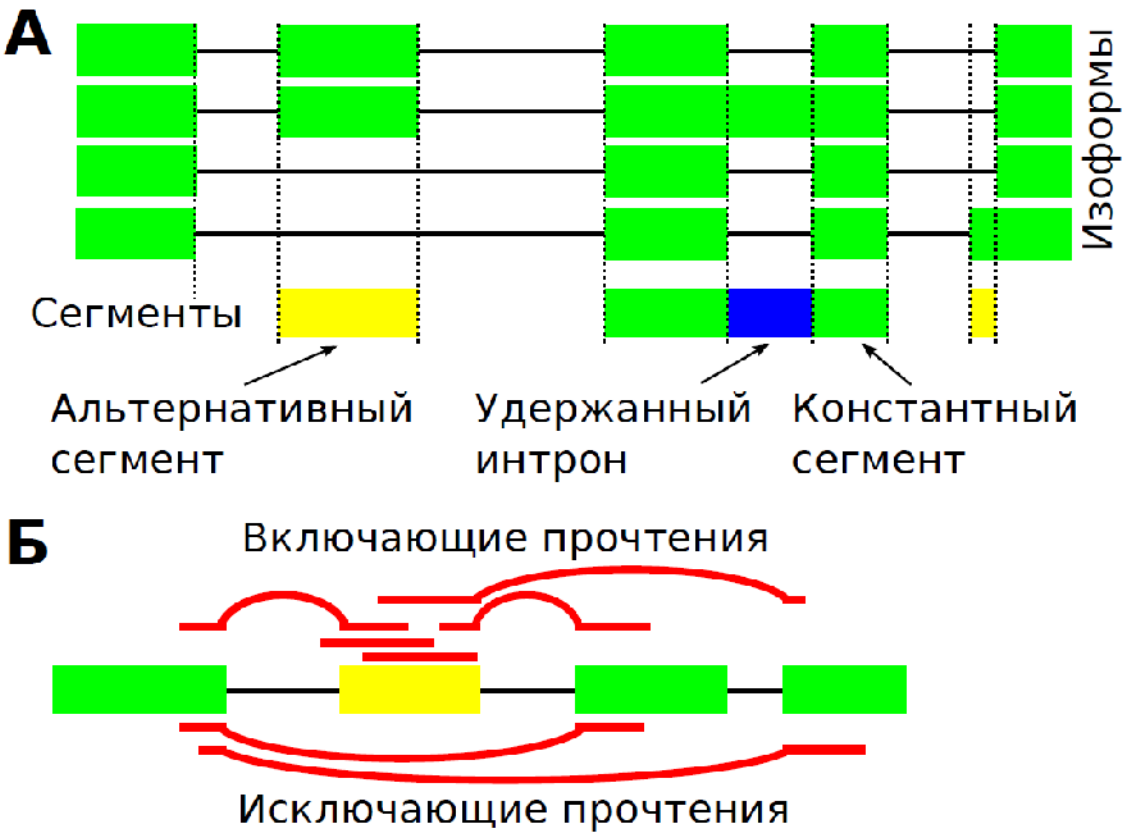

Рисунок 2.Разделение гена на сегменты (А) и подсчёт включающих и исключающих прочтений (Б). Экзоны показаны зелёными прямоугольниками, интроны показаны горизонтальными линиями, вертикальные пунктирные линии обозначают сайты сплайсинга. Прочтения обозначены красными линиями; горизонтальные участки изображают части прочтения, картировавшиеся непосредственно на геномную последовательность; дугами обозначены прочтения, картировавшиеся на экзон-экзонные границы.

Исходя из количеств включающих и исключающих прочтений, для каждого сегмента можно рассчитать частоту включения (ЧВ) — долю транскриптов, содержащих данный сегмент, по формуле:

$$ЧВ = \frac{\frac{в}{дс + дп - 1}}{\frac{в}{дс + дп - 1} + \frac{и}{дп - 1}}$$

где *в* и *и* обозначают количество включающих и исключающих прочтений, а

дс и *дп* обозначают длину сегмента и прочтения, соответственно.

SAJR имеет большое количество настроек, позволяющих фильтровать прочтения по степени надёжности. В зависимости от настроек, SAJR может использовать или не использовать прочтения, картирующиеся на несколько позиций генома, а также парные прочтения, из которых картировалось только одно или позиции картирования которых не образуют корректную пару. Так как в большинстве экспериментов цитоплазму не отделяют от ядер, среди секвенируемой РНК может оказаться недопроцессированная пре-мРНК, и прочтения, полученные с такой пре-мРНК, могут исказить результаты. SAJR позволяет не использовать прочтения, пересекающиеся с константными или удержанными интронами, при подсчёте включающих прочтений для альтернативных сегментов (но не удержанных интронов). В большинстве случаев использование только тех прочтений, которые картируются в одно место генома, не пересекаются с интронами и образуют корректные пары (в случае парных прочтений), является предпочтительным. Однако в некоторых случаях, например, при сравнении нескольких видов, у которых могут быть различные наборы паралогов, может быть полезно использование прочтений, картирующихся в несколько позиций генома.

Так как SAJR реализован на java, он может быть использован под любой операционной системой, поддерживающей java 8. SAJR использует в качестве входных данных файл, содержащий геномные позиции прочтений в формате bam или sam, которые являются стандартными для большинства программ, осуществляющих картирование прочтений, и геномную аннотацию в формате gff, gff3 или gtf. SAJR обрабатывает прочтения по одному, загружая в оперативную память только геномную аннотацию, поэтому он не требует большого количества компьютерных ресурсов (от 100 до 1000 мегабайт оперативной памяти в

зависимости от размера аннотации) и способен обсчитывать 10 миллионов прочтений за одну минуту, на обработку одного образца РНК-Сек уходит около 10 минут на обычном рабочем компьютере (один поток 3460МГц, 4Г RAM).

3.2. Статистический анализ

Количества включающих и исключающих прочтений могут быть рассмотрены как результат биномиальных испытаний, при этом частоты успеха монотонно связаны с ЧВ. SAJR использует обобщённые линейные модели (ОЛМ) [Nelder, Wedderburn, 1972] и тест на логарифм правдоподобия [Wilks, 1938] для оценки значимости изменений ЧВ. Дисперсия биномиального распределения зависит от среднего и числа испытаний, что не позволяет учитывать биологическую вариабельность и может приводить к большому числу ложноположительных результатов [Oberg и др., 2012]. При наличии достаточного числа образцов SAJR позволяет использовать квази-биномиальное распределение и тест на квази-правдоподобие для учёта биологической вариабельности [Wedderburn, 1974].

В биномиальной ОЛМ логистически преобразованная вероятность успеха моделируется линейной комбинацией независимых переменных:

$$p = logit^{-1}(\sum a_i \times x_i)$$
$$logit(x) = \log(\frac{x}{1-x})$$
$$logit^{-1}(x) = \frac{e^x}{1+e^x}$$

где p — вероятность успеха, x_i — независимые переменные, a_i — параметры модели.

Для построения ОЛМ в SAJR используется функция glm из статистического пакета R. Эта функция максимизирует правдоподобие (вероятность получить

наблюдаемые значения при условии данной модели) при помощи итеративно перевзвешиваемого метода наименьших квадратов [Dobson, 2002]. Для учёта биологической вариабельности используется так называемый тест на квази-правдоподобие. При этом правдоподобие делится на дисперсионный параметр, который вычисляется для каждого сегмента, исходя из отклонений наблюдений от модели [McCullagh, Nelder, 1989]:

$$od = \frac{\sum_i pr_i^2}{residual.df}$$

$$pr_i = \frac{(y_i - n_i \times f_i)}{\sqrt{(n_i \times f_i \times (1 - f_i))}}$$

где od — дисперсионный параметр, pr_i — пирсоновское остаточное отклонение для образца i, $residual.df$ — остаточное число степеней свободы модели (число наблюдений минус число независимых параметров в модели), y_i — число включающих прочтений в образце i, n_i — общее число прочтений (включающих и исключающих) в образце i и f_i — вероятность получения включающего прочтения, предсказанная моделью [Wedderburn, 1974]. Иногда вычисленный таким образом дисперсионный параметр может оказаться меньше единицы, в таких случаях он принимается равным единице.

Для поправки на множественное тестирование SAJR использует процедуру Бенджамини-Хохберга [Benjamini, Hochberg, 1995].

4. Возрастные изменения сплайсинга в мозге человека

4.1. Материалы и методы

4.1.1. Образцы ткани

Для анализа возрастных изменений АС в головном мозге человека были использованы два набора данных, полученных в лаборатории Филиппа Хайтовича

(Лаборатория сравнительной биологии, CAS-MPG PICB, Шанхай, Китай). Первый набор данных (НД1.1) состоит из 12 образцов, по 6 для префронтальной коры (ПФК) и коры мозжечка (КМ). Для снижения биологической вариабельности каждый образец был получен смешением равных количеств РНК, выделенной из 5 доноров примерно одного возраста; для получения образцов коры и мозжечка использовались одни и те же доноры. Возрасты доноров покрывают всё протяжение жизни от рождения до старости, подробная информация о НД1.1 представлена в приложении 2. Второй набор данных (НД1.2) содержит 13 индивидуальных (то есть каждый получен от одного донора) образцов префронтальной коры. Возрасты доноров так же находятся в интервале от рождения до старости, подробная информация о НД1.2 представлена в приложении 3 и на рис. 3А.

Все образцы были взяты посмертно у доноров, не страдавших нейродегенеративными заболеваниями, не претерпевших длительной агонии, информированное согласие было получено у родственников.

4.1.2. Секвенирование

Выделение и секвенирование полиаденилированной фракции РНК было произведено лабораторией Филиппа Хайтовича (Лаборатория сравнительной биологии, CAS-MPG PICB, Шанхай, Китай) на секвенаторе Illumina Genome Analyzer II system. НД1.1 был секвенирован по протоколу для парных прочтений длины 75 нт, для НД1.2 были получены непарные прочтения длиной 100 нт. Полученные прочтения были загружены в архив коротких прочтений [Wheeler и др., 2008] под идентификатором SRP005169.

4.1.3. Картирование прочтений

Для картирования прочтений была создана библиотека экзон-экзонных

границ. Для этого, у каждого гена из базы данных Ensembl [Flicek и др., 2012], были рассмотрены все пары экзонов таких, что первый экзон заканчивается до того, как начинается второй (в координатах транскрипта). Каждая такая пара образует потенциальную экзон-экзонную границу. Чтобы получить соответствующую ей нуклеотидную последовательность, такие экзоны были соединены вместе и продлены далее в транскрипт, так, чтобы суммарная расстояние от экзон-экзонной границы до края фрагмента была не менее 100 нт. В результате было получено 302179 известных ранее экзон-экзонных границ и 2267677 новых. Этот набор содержит все потенциально возможные границы между аннотированными экзонами.

Прочтения были картированы на геном человека версии hg19 и базу экзон-экзонных границ при помощи программы bowtie [Langmead и др., 2009], допуская не более трёх замен. Такая процедура выравнивания была использована, потому что на момент исследования не существовало специализированных программ для картирования данных РНК-Сек, таких как tophat.

4.1.4. Статистический анализ

Статистический анализ проводился в пакете R [R Core Team, 2013], поиск сегментов с значимыми изменениями АС производился при помощи программы SAJR, описанной выше. Для НД1.1 и НД2.1 использовались следующие модели соответственно:

$$ЧВ \sim возраст + возраст^2 + ткань + возраст{:}ткань + возраст^2{:}ткань$$

$$ЧВ \sim возраст + возраст^2$$

где *возраст* это логарифм возраста донора. *P*-значения были откорректированы для учёта множественного тестирования методом Бенджамини-Хохберга [Benjamini, Hochberg, 1995]. Сегменты с корректированным $p<0.05$

считались значимыми. Сегменты, у которых либо линейный, либо квадратичный возрастной член модели был значим, считались значимо меняющими АС с возрастом. Дисперсионный параметр для всех сегментов принимался равным единице, для учёта биологической вариабельности в анализе использовались только сегменты, значимо меняющие АС с возрастом в обоих наборах данных.

ЧВ для каждого сегмента вычислялись при помощи программы SAJR, описанной выше. Если сумма включающих и исключающих прочтений была меньше 5, ЧВ считалась неопределённой.

4.1.5. Подтверждение изменений АС при помощи ПЦР

Для подтверждения изменений АС при помощи ПЦР были отобраны 30 сегментов. Для каждого сегмента были выбраны два образца из НД1.1 (ПФК) с максимальным различием в ЧВ. Для каждого образца была синтезирована кДНК. Для каждого сегмента была подобрана пара праймеров, комплементарных соседним константным сегментам. После 30 циклов ПЦР, 5 мкл продукта реакции было использовано для электрофореза на 2% агарозном геле. Определение концентрации продуктов ПЦР производилось по интенсивности свечения при помощи программы Quantity One. ЧВ вычислялась как отношение концентрации длинного (т. е. соответствующего включению сегмента) продукта ПЦР к сумме концентраций обоих (длинного и короткого) продуктов. Эксперименты проводились сотрудниками лаборатории Филиппа Хайтовича в институте CAS-MPG PICB в Шанхае, не являются частью данной диссертации и приводятся здесь для только в качестве подтверждения результатов метода SAJR.

4.1.6. Подсчёт корреляции между наборами данных

Для сравнения возрастных изменений наблюдаемых в НД1.1 и НД2.1, возрастные изменения в НД2.1 были аппроксимированы кубическим сплайном с

тремя степенями свободы (возраст рассматривался в логарифмической шкале). На основания сплайна ЧВ были предсказаны в возрастах соответствующих средним возрастам доноров образцов в НД1.1. Далее коэффициент корреляции Пирсона вычислялся между предсказанными ЧВ и ЧВ вычисленными на основании образцов из ПФК из НД1.1.

4.1.7. Разбиение на кластеры

Для разбиения сегментов на кластеры была использована иерархическая кластеризация (функция hclust пакета R). В качестве расстояния между сегментами использовалась единица минус коэффициент корреляции Пирсона. Использовались только 1422 сегмента, которые имели определённые ЧВ (то есть сумма включающих и исключающих прочтений была более пяти) во всех образцах и значимо меняли ЧВ с возрастом в обоих наборах данных. Полученное в результате кластеризации дерево было разрезано на уровне 1.5, выбранном вручную, в результате получилось 8 паттернов.

4.2. Результаты

В совокупности в результате секвенирования было получено 181555729 пар прочтений в НД1.1 и 274927771 прочтений в НД1.2, из которых 64% удалось картировать, из них 93% картируются внутрь границ генов.

Анализ показал, что 22% генов с достаточным покрытием прочтениями в НД1.1 (3132 сегмента из 1456 генов) и 38% генов в НД1.2 (6114 сегментов из 2588 генов) значимо меняют АС с возрастом (рис. 3Б). 1484 сегментов из 721 генов значимы в обоих наборах данных, только эти сегменты использовались в последующем анализе.

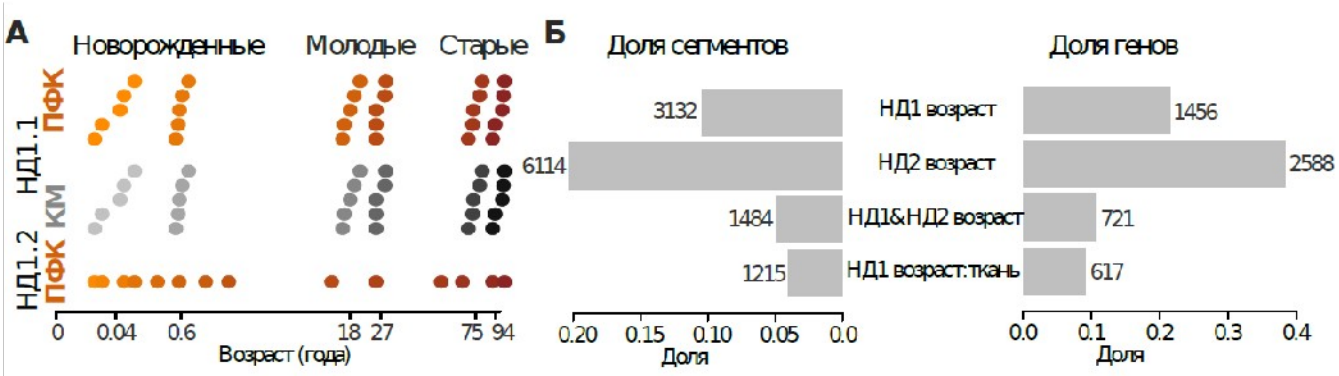

Рисунок 3. Возрастные изменения АС в мозгу человека. (А) Возрасты индивидуальных доноров; ПФК и КМ показаны серым и оранжевым соответственно, по горизонтальной оси отложен возраст доноров, в НД1.1 доноры, использованные для одного образца показаны вместе. (Б) Количества (и доля от общего числа тестированных) сегментов и генов значимо меняющих сплайсинг с возрастом в НД1.1, НД1.2, в обоих наборах данных и имеющих значимо различные возрастные паттерны в ПФК и КМ. Рисунок взят с изменениями из [Mazin и др., 2013].

Пересечение значимых сегментов в обоих наборах данных, хотя и не очень велико, однако статистически значимо больше, чем ожидаемое случайно (тест Фишера, $p<0.05$). Однако даже сегменты, возрастные изменения ЧВ которых значимы только в одном наборе данных, имеют высокую корреляцию ЧВ между наборами данных (рис. 4А). Чтобы получить независимое подтверждение обнаруженных возрастных изменений АС, были выбраны 30 сегментов, и ЧВ в них было измерено при помощи ПЦР. Для 24 сегментов были получены ПЦР-продукты ожидаемого размера, для всех из них направление изменений совпало с вычисленным, исходя из данных РНК-Сек (рис. 4Б и 5), коэффициент корреляции Пирсона между изменениями ЧВ, измеренными при помощи РНК-Сек и ПЦР составил 0.93. Более того, не только изменения, но и абсолютные значения ЧВ, определённые при помощи ПЦР, показывают высокую корреляцию (0.88) с ЧВ, определёнными при помощи РНК-Сек (рис. 4В).

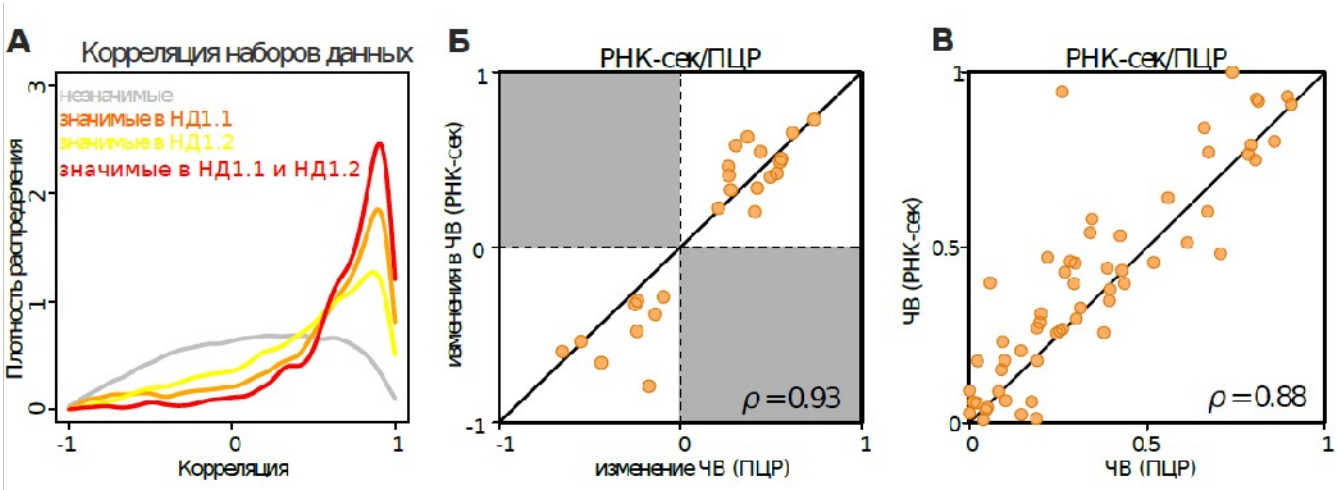

Рисунок 4. Подтверждение результатов РНК-Сек. (А) Распределение коэффициентов корреляции возрастных изменений АС между наборами данных; сегменты незначимые ни в одном наборе данных, значимые только в НД1.1, только в НД2.1 и в обоих наборах данных показаны серым, оранжевым, жёлтым и красным соответственно. (Б) и (В) Сравнение ЧВ (В), и их изменения (Б) с возрастом, вычисленных на основании данных РНК-Сек и при помощи ПЦР. Рисунок взят с изменениями из [Mazin и др., 2013].

Два примера возрастного изменения АС показаны на рис. 5. ЧВ восьмого экзона гена *PALM*, участвующего в перестройках цитоплазматической мембраны в нейронах [Kutzleb и др., 1998], растёт с возрастом, так же как ЧВ третьего и десятого экзонов гена *MAPT*, вовлечённого в нейродегенеративные заболевания, такие как болезнь Альцгеймера. Было показано, что нарушение регуляции сплайсинга десятого экзона гена *MAPT* часто связанно с нейродегенеративными заболеваниями [Liu, Gong, 2008].

В совокупности это указывает на то, что обнаруженные нами возрастные изменения ЧВ в ПФЦ человека могут быть воспроизведены на независимом наборе данных и при помощи различных экспериментальных процедур.

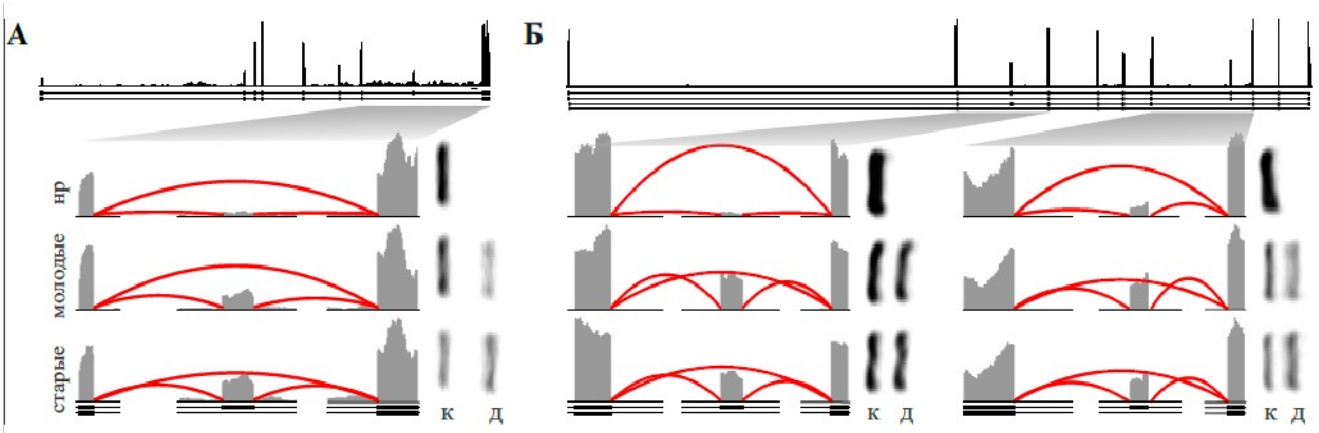

Рисунок 5. Примеры возрастных изменений АС в генах PALM (А) и MART (Б) в ПФК. На верхних панелях показано покрытие всего гена прочтениями (все образцы НД1.1) и экзонная структура основных изоформ. На нижних панелях показаны увеличенный альтернативный и два ближайших константных экзона, для новорожденных (нр), молодых и старых доноров (в каждом случае два образца схожего возраста из НД1.1 объединены вместе). Красными дугами показаны прочтения, картирующиеся на экзон-экзонные границы. Справа от схем показаны результаты ПЦР (фото геля), короткая и длинная изоформы обозначены буквами к и д. Рисунок взят с изменениями из [Mazin и др., 2013].

Чтобы более детально охарактеризовать изменения АС с возрастом, 1422, сегмента значимых в обоих наборах данных и имеющих достаточное покрытие во всех образцах, были разбиты на восемь паттернов при помощи иерархической кластеризации (рис. 6). Для обозначения паттернов далее будут использоваться сокращения С1–С8. Как видно на рисунке 6, паттерны хорошо воспроизводятся как между наборами данных, так и между двумя регионами мозга. Интересно, что, хотя большая часть изменений АС происходит в ходе развития (до 20 лет, пунктирная линия на рис. 6), около 30% изменений (26% в НД1.1 и 33% в НД1.2) изменений приходятся на старение.

Наиболее популярным паттерном является убывание ЧВ с возрастом (паттерны С1, С3 и С4, 62% сегментов), следующим по популярности является монотонное возрастание ЧВ с возрастом (паттерны С2 и С6, 21% сегментов).

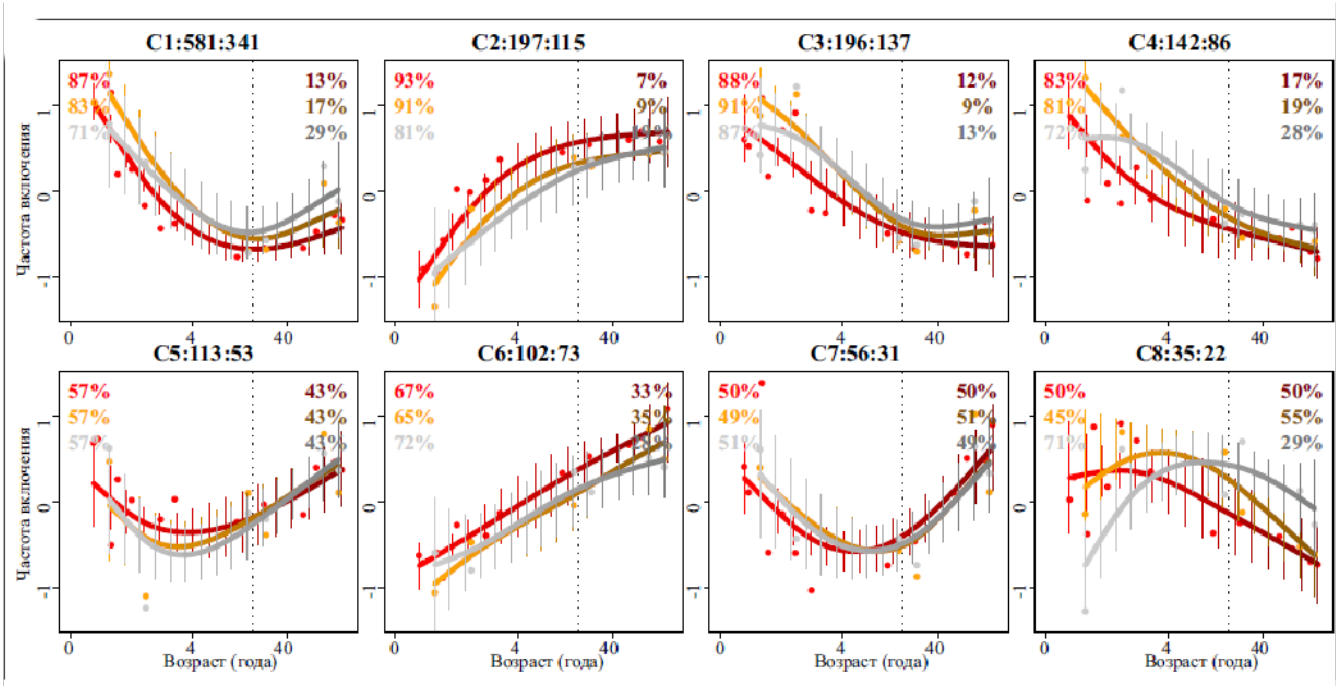

Рисунок 6. Паттерны возрастных изменений АС в ПФК и КМ человека. Все сегменты значимо меняющие АС с возрастом в НД1.1 и НД1.2 разбиты на 8 паттернов. Паттерны упорядочены по количеству сегментов которые ему следуют. По вертикальной оси отложена усреднённая по кластеру, нормализованная ЧВ, по горизонтальной оси отложен возраст в годах. НД1.1 ПФК, НД1.1 КМ и НД1.2 ПФК показаны красным, серым и оранжевым соответственно; точки показывают усреднённую ЧВ, их аппроксимация при помощи кубических сплайнов с тремя степенями свободы показаны кривой. Вертикальная пунктирная линия разделяет развитие и старение, доли вариабельности АС, приходящаяся на каждый из периодов указаны на графике. Номер паттерна, число сегментов и число генов указаны в названии каждого графика. Рисунок взят с изменениями из [Mazin и др., 2013].

Сегменты, следующие различным возрастным паттернам, различаются по биологическим свойствам. Функциональный анализ при помощи пакета GOstat [Falcon, Gentleman, 2007] показал, что наборы генов, содержащих сегменты из паттернов C6 и C8, значимо обогащены функциями связанными с развитием нервной системы и синапсов, поведением и обучением (C6); цитоскелетом и апоптозом (C8) (приложение 4). Кроме того, различные типы сегментов неравномерно распределены среди паттернов. Например, удержанные интроны

чаще встречаются в паттернах, в которых ЧВ убывает с возрастом, и особенно в С1, в то время как доля белок-кодирующих сегментов максимальна в паттернах в которых ЧВ растёт с возрастом, например в С2 (рис. 7). Удержанные интроны могут вызывать деградацию мРНК либо через ПСК-зависимый механизм либо при помощи ядерной экзосомы [Yap и др., 2012], поэтому сегменты из С1 (а так же С3 и С4) могут участвовать в регуляции клеточных концентраций мРНК.

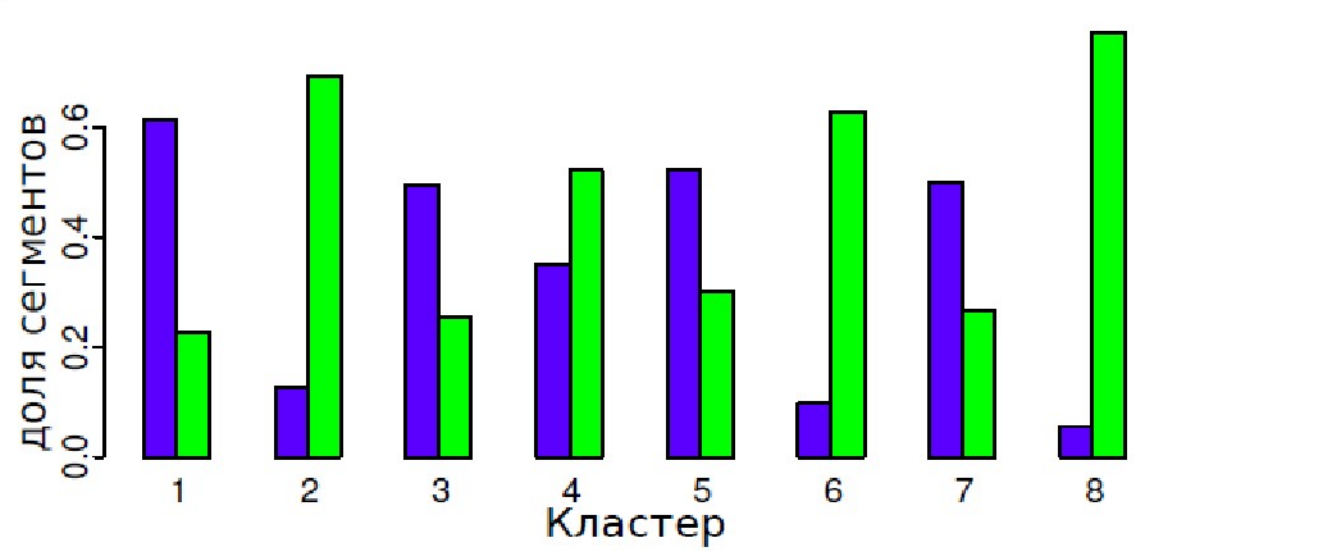

Рисунок 7. Доля удержанных интронов (синий) и белок-кодирующих сегментов (зеленый) в каждом паттерне.

Хотя возрастное изменение АС большинства сегментов схоже в ПФК и КМ (рис. 8А), около 15% сегментов (из 3132 возраст-зависимых в НД1.1 сегментов) имеют статистически значимые различия в возрастной регуляции АС между двумя регионами мозга. Например, такими генами являются предшественник амилоида бета (*APP*), связанный с болезнью Альцгеймера [O'Brien, Wong, 2011]; ген *BIN1*, который кодирует белок, участвующий в эндоцитозе синаптических везикул; или ген протокадгерина гамма (*PCDHG*), кодирующий при помощи набора альтернативных первых экзонов 22 белка, участвующих в образовании специфических клеточных контактов [Lefebvre и др., 2012]. В случае

протокадгерина наши результаты показывают, что, в то время как в ПФК ЧВ трёх основных первых экзонов сильно меняются с возрастом, в КМ изменений фактически не наблюдается (рис. 8Б–В). Обнаруженные нами возрастные изменения ЧВ первых экзонов в протокадгерине гамма у человека напоминают обнаруженные ранее возрастные изменения в мозге мыши [Frank и др., 2005], что указывает на эволюционную консервативность регуляции выбора первого экзона в гене *PCDHG*.

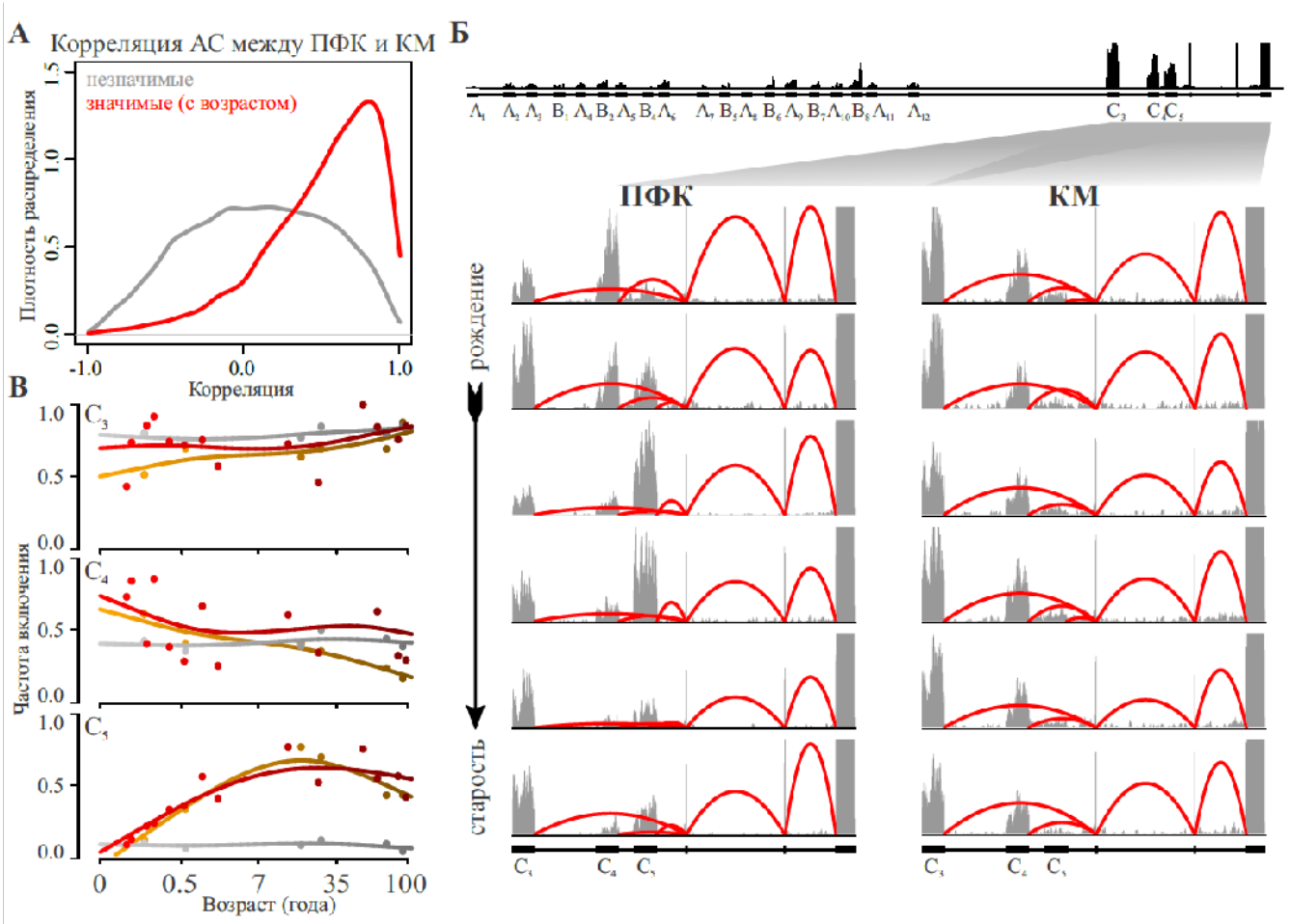

Рисунок 8. Возрастные изменения АС в ПФК и КМ человека. (А) Распределение коэффициента корреляции для ЧВ между ПФК и КМ для значимых и незначимых сегментов. (Б)–(В) Изменение частоты использования трёх основных альтернативных первых экзонов в протокадгерние гамма. Зависимость частоты включения от возраста показана на панели (В), ПФК из НД1.1, КМ из НД1.1 и ПФК из НД1.2 показаны оранжевым, серым и красным соответственно. Покрытие основных экзонов прочтениями из НД1.1 показано на панели (Б). Рисунок взят с изменениями из [Mazin и др., 2013].

4.3. Выводы

В этой части работы были получены следующие основные результаты.

1. Показано, что разработанный нами метод SAJR позволяет получать

воспроизводимые (на различных наборах данных) оценки частоты включения, которые хорошо соотносятся с результатами других экспериментальных методов, таких как ПЦР.

2. Показано, что значительная доля генов (более 10%), экспрессируемых в мозге человека, меняет сплайсинг в ходе постнатального развития.
3. Сегменты со значимыми возрастными изменениями частоты включения формируют несколько различных паттернов зависимости включения от возраста, при этом у большинства сегментов частота включения падает с возрастом, и эти сегменты, как правило обогащены удержанными интронами.
4. Хотя большинство изменений альтернативного сплайсинга приходится на развитие, 30% изменений происходит в ходе старения.
5. Около 15% генов имеют значимо различную возрастную регуляцию альтернативного сплайсинга в префронтальной коре и в коре мозжечка.

5. Сравнительный анализ альтернативного сплайсинга в мозге высших приматов

5.1. Материалы и методы

5.1.1. Образцы ткани

Для анализа возрастных изменений АС в головном мозге человека (Homo sapiens), шимпанзе (Pan troglodytes) и макаки (Macaca mulatta) были использованы три набора данных, полученных в лаборатории Филиппа Хайтовича (Лаборатория сравнительной биологии, CAS-MPG PICB, Шанхай, Китай).

Первый набор данных (НД2.1) содержит 40, 39 и 40 индивидуальных (каждый образец получен от одного донора) образцов префронтальной коры человека,

шимпанзе и макаки соответственно (приложение 5). Второй набор данных (НД2.2) содержит 13, 15 и 15 индивидуальных образцов префронтальной коры человека, шимпанзе и макаки, соответственно. Образцы человеческой ткани из НД2.2 совпадают с образцами из НД1.2 (приложение 6). Третий набор данных (НД2.3) содержит 12 образцов ткани человека из НД1.1 и по 4 образца для шимпанзе и макаки: два участка мозга (ПФК и КМ) и два возраста (новорожденные и молодые). Так же, как и образцы из НД1.1, образцы обезьян из НД2.3 были получены смешением равных количеств РНК выделенной из образцов пяти доноров примерно одного возраста, для получения образцов коры и мозжечка использовались одни и те же доноры (приложение 7).

Все образцы человеческих тканей были взяты посмертно у доноров, не страдавших нейродегенеративными заболеваниями, не претерпевших длительной агонии; информированное согласие было получено у родственников. Обезьяны, чьи ткани использовались в работе, погибли без продолжительной агонии, по причинам, не связанным с их участием в данном исследовании.

5.1.2. Секвенирование

Образцы из НД2.3 были секвенированы так же как образцы из НД1.1. Секвенирование образцов человеческой ткани из НД2.2 описано выше в разделе 4.1. Образцы тканей шимпанзе и макаки из НД2.2, а так же все образцы из НД2.1 были секвенированы по протоколу для непарных, не цепь-специфичных прочтений длины 100 нт на секвенаторе Illumina Genome Analyzer II system.

5.1.3. Картирование прочтений

Все прочтения были картированы программой tophat [Trapnell, Pachter, Salzberg, 2009] на соответствующие геномы (версии hg38, panTro4 и rheMac3 для человека, шимпанзе и макаки, соответственно) с не более чем тремя заменами на

прочтение, с использованием экзон-экзонных границ из аннотации, полученной на основании НД2.1 (см. ниже).

5.1.4. Экзон-интронная аннотация геномов

Для создания аннотации использовался набор НД2.1, так как он содержит наибольшее количество образцов.

Аннотация создавалась по следующей процедуре.

1. Все прочтения, относящиеся к данному виду, картировались на соответствующий геном при помощи программы tophat, позволяющей находить интроны при помощи только данных РНК-Сек и генома. Для увеличения чувствительности все образцы, относящиеся к данному виду, картировались вместе.
2. Определённые tophat границы интронов (сайты сплайсинга) выравнивались между видами при помощи программы liftOver [Hinrichs и др., 2006] и попарных геномных выравниваний, полученных из базы данных ucsc Genome Browser [Rosenbloom и др., 2015]. Только интроны, обе границы которых могут быть однозначно выравнены между всеми тремя видами (ортологичные интроны), использовались в дальнейшем анализе.
3. Данные заново картировались при помощи tophat с использованием только интронов полученных на шаге 2 (без предсказания новых интронов). Эта процедура гарантирует, что в анализе будут использоваться только интроны, ортологичные во всех трёх видах.
4. Участок генома между двумя ближайшими границами сайтами сплайсинга, со средним покрытием прочтениями (хотя бы в одном виде) более единицы, и с пропусками в покрытии не более пяти нуклеотидов, считался сегментом. Покрытие определялось как количество прочтений из всех образцов данного

вида, проходящих через данную позицию генома.

5. Далее для каждого вида создавался граф сплайсинга, вершинами которого являются сайты сплайсинга, а рёбрами — интроны, полученные на шаге 2, и сегменты, полученные на шаге 4. Связные компоненты в графе сплайсинга считались генами. В дальнейшем анализе использовались только гены, в которых последовательность ортологичных сайтов вдоль генома была идентичной во всех трёх видах.

Так как в предыдущей главе было показано, что АС различных типов сегментов (например, удержанных интронов и кодирующих сегментов) регулируется с возрастом по-разному, сегменты, полученные на шаге 4, были разбиты на четыре группы:

1. Константные сегменты (не пересекаются с интронами, 155244 сегментов)
2. Альтернативные кодирующие (пересекаются с интронами и входят в белок-кодирующую последовательность согласно базе данных Ensembl 77, 32213 сегментов).
3. Удержанные интроны (52118)
4. Другие типы некодирующих альтернативных сегментов (35730).

Для некоторых типов анализа сегменты были разбиты на группы по типам сайтов сплайсинга которые их ограничивают:

1. Кассетный экзон: сегмент между акцепторным и донорным сайтами
2. Альтернативный донорный сайт: сегмент между двумя донорными сайтами
3. Альтернативный акцепторный сайт: сегмент между двумя акцепторными сайтами
4. Удержанный интрон: сегмент между донорным и акцепторным сайтом.

В анализе были использованы все типы сегментов, кроме константных.

5.1.5. Статистический анализ

Анализ АС проводился на основании НД2.1, так как он содержит максимальное количество образцов, остальные наборы данных использовались только для подтверждения результатов, полученных на основании НД2.1. Анализ проводился при помощи описанной выше программы SAJR с использованием двух моделей:

$$ЧВ \sim возраст + возраст^2$$

$$ЧВ \sim вид + возраст + возраст^2 + вид{:}возраст + вид{:}возраст^2$$

Первая модель применялась для каждого вида, вторая — для каждой пары видов. В качестве возраста использовался корень четвёртой степени из возраста (в днях) с момента зачатия. Логарифмическое преобразование, использованное в предыдущей главе, смещает фокус исследования на ранние стадии развития, что в данном случае было нежелательно, так как в этой главе для адекватного сравнения видов возраст отсчитывался не от рождения, а от зачатия, но большинство образцов было постнатальными. Срок беременности принимался равным 266, 237 и 164 дням для человека, шимпанзе и макаки, соответственно. Для определения значимости изменений применялся тест на лог-правдоподобие и поправка на множественное тестирование Бенджамини-Хохберга. Сегменты, у которых хотя бы один из двух членов в первой модели был значим (корректированное $p<0.05$) и амплитуда изменений превышала 0.1. считались значимо меняющими АС с возрастом в данном виде. Для вычисления амплитуды возрастных изменений ЧВ сегмента упорядочивались по убыванию и бралась разница второго и предпоследнего значения.

Для определения значимости различий в среднем уровне ЧВ и в возрастных паттернах изменений ЧВ между парой видов использовалась вторая модель. В случае отличий в средних уровнях ЧВ, дополнительно требовали, чтобы разница

ЧВ между видами превышала 0.1.

Многомерное шкалирование к двум размерностям проводилось функцией cmdscale в R, с использованием единицы минус коэффициент корреляции Пирсона в качестве меры расстояния.

5.1.6. Определение видоспецифичных изменений

Считалось, что сегмент имеет специфичное для вида *x* изменение средней ЧВ если:

1. Отличия в средних уровнях между *x* и обоими оставшимися видами *a* и *б* значимы. При этом различия между *a* и *б* должны быть незначимы.
2. Средняя ЧВ в *x* должна быть больше или меньше средний ЧВ в *a* и *б*.
3. $\frac{min(|ЧВ(x)-ЧВ(a)|,|ЧВ(x)-ЧВ(б)|)}{max(|ЧВ(x)-ЧВ(a)|,|ЧВ(x)-ЧВ(б)|)}>0.7$, где ЧВ(*x*) — частота включения в виде *x*.

5.1.7. Определение направления видоспецифичных изменений

Сегменты с видоспецифичными изменениями классифицировали по двум признакам на четыре группы:

1. По средней (по всем видам) ЧВ на мажорные (более 0.5) и минорные (менее 0.5).
2. По направлению изменения на увеличившие ЧВ (ЧВ в виде *x* больше, чем в *a* и *б*) и уменьшившие ЧВ.

Минорные сегменты с увеличивающейся ЧВ и мажорные сегменты с уменьшающейся ЧВ соответствуют альтернификации (приобретению или усилению альтернативности) сегмента в ходе эволюции.

5.1.8. Выравнивание возрастных паттернов АС

Пусть $f(a)$ и $g(a)$ обозначают зависимость ЧВ от возраста a в двух видах. В простейшем случае задача выравнивания может быть сформулирована как поиск коэффициента x, минимизирующего расстояние между кривыми, которое может быть вычислено как интеграл $\int |f(x \times a) - g(a)| da$, где интегрирование ведётся в общем для обоих видов интервале возрастов (с учётом x). Чтобы избежать влияния межвидовых отличий в средней ЧВ перед подсчётом интеграла из функций $f(a)$ и $g(a)$ вычитались их средние значения на указанном интервале. Для поиска оптимального x (для фиксированных сегмента и пары видов) использовалась следующая процедура:

1. Рассматривались только сегменты у которых ЧВ значимо менялась с возрастом в обоих видах.
2. Для каждого вида строилась аппроксимация кубическим сплайном с четырьмя степенями свободы (ЧВ ~ возраст$^{0.25}$).
3. Функция optimize из пакета R, использующая комбинацию метода золотого сечения и последовательной кубической интерполяции [Brent, 1973] использовалась для поиска оптимального значения x, минимизирующего описанный выше интеграл.

5.1.9. Анализ эволюции сайтов сплайсинга

Для сравнения сил сайтов сплайсинга для акцепторного и донорного сайта была построена позиционная весовая матрица (ПВМ). Для этой цели использовались последовательности всех сайтов сплайсинга человека, определённых в этой работе. Вес нуклеотида n в позиции сайта i определялся как $\log_2(\frac{f_i(n)}{f(n)})$, где $f_i(n)$ — частота нуклеотида n в позиции сайта i, а $f(n)$ — частота

нуклеотида *n* во всем геноме. Длина донорного сайта составляла 9 нт (3 нт в экзоне и 6 нт в интроне), длина акцепторного сайта составляла 25 нт (24 нт в интроне и 1 нт в экзоне). Сила сайта сплайсинга определялась как сумма весов нуклеотидов последовательности данного сайта.

Сила сплайсинга сегмента определялась как линейная комбинация весов сайтов, ограничивающих данный сегмент. Для кассетного сегмента веса складывались, в случае альтернативного донорного или акцепторного сайта вес внутреннего сайта вычитался из веса внешнего, для удержанных интронов бралась отрицательная сумма весов. Таким образом, сегменты с большей силой сплайсинга должны иметь большую вероятность включения. Для оценки доли межвидовых изменений АС объясняемой отличиями в последовательностях сайтов сплайсинга, для разных порогов на межвидовое отличие ЧВ были посчитаны: (а) общее количество изменений (*N*), (б) количество изменений, у которых направление изменения ЧВ совпадает (*agree*) и (в) не с совпадает с направлением изменения силы сплайсинга (*disagree*). Доля объяснённых изменений вычислялась как отношение разницы *agree* и *disagree* к *N*.

5.1.10. Функциональный анализ сегментов

Поиск биологических функций, значимо перепредставленных среди определённого набора генов (например значимо меняющих сплайсинг с возрастом), осуществлялся при помощи покета goseq [Young и др., 2010].

Все перепредставленные функции с корректированным $p<0.2$ (поправка Бенджамини-Хохберга) рассматривались как значимо обогащённые.

5.1.11. Поиск мотивов связывания факторов сплайсинга

ПВМ для 219 мотивов РНК-связывающих белков человека были загружены из

базы данных cisbp-rna [Ray и др., 2013]. Для анализа были отобраны кассетные экзоны с длиной не менее 50 нт, и длиной фланкирующих интронов не менее 100 нт. Сто нт до начала экзона, 50 первых нт экзона, 50 последних нт экзона и 100 нт после экзона были разбиты на шесть интервалов по 50 нт. Аффинность каждого мотива (ПВМ) к каждому интервалу была вычислена как описано в [Lee, Bussemaker, 2010]. Далее, для каждого интервала были найдены мотивы со значимо большей аффинностью у возраст-зависимых экзонов, по сравнению с остальными альтернативными экзонами (тест Вилкоксона, поправка Бенджамини-Хохберга, корректированное $p<0.05$).

5.1.12. Разбиение на кластеры

Для разбиения сегментов на кластеры была использована иерархическая кластеризация (функция hclust пакета R). В качестве расстояния между сегментами использовалось евклидово расстояние между нормализованными аппроксимированными ЧВ. Сперва ЧВ была аппроксимирована в 15 равномерно распределенных возрастных точках при помощи кубического сплайна с четырьмя степенями свободы. Потом из аппроксимированных ЧВ было вычтено среднее значение ЧВ данного сегмента в данном виде. Далее ЧВ данного сегмента во всех видах были объединены и поделены на среднеквадратичное отклонение ЧВ данного сегмента. Евклидово расстояние рассчитывалось между векторами, содержащими ЧВ всех трёх видов. Число кластеров (шесть) было выбрано вручную, на основании того, что при увеличении их числа получалось несколько кластеров с похожими паттернами возрастных изменений ЧВ.

5.1.13. Определение уровня экспрессии генов

Для каждого гена вычислялось число прочтений (из каждого образца), пересекающихся с хотя бы одним константным сегментом гена. Полученное число

прочтений делилось на общую длину всех константных сегментов данного гена и на общее число прочтений, отнесённых к генам в данном образце. Полученная величина использовалась в качестве оценки уровня экспрессии гена. Для анализа изменений уровня экспрессии генов использовалась линейная модель (полином от возраста второй степени) и дисперсионный анализ. Все гены, у которых хотя бы один из возрастных членов был значим (поправка Бенджамини-Хохберга, корректированное $p<0.05$), считались значимыми.

5.1.14. Моделирование возрастных изменений ЧВ

Для моделирования была использована следующая линейная модель:

$$ЧВ(э,о) \sim \sum_{у} \sum_{м} \sum_{фс} \log(афф(у,э,м) \times \frac{экспрессия(о,фс)}{max(экспрессия(о,фс))})$$, где э — экзон, *о* — образец, *у* — один из шести участков последовательности РНК около экзона, *м* — один из мотивов, перепредставленных около возраст-зависимых экзонов, *фс* — один из факторов сплайсинга, связывающий хотя бы один из значимых мотивов; *афф*(*у*,*э*,*м*) — аффинность мотива *м* в участке *у* сегмента *э*, *экспрессия*(*о*,*фс*) — уровень экспрессии данного фактора сплайсинга в данном образце. Из всех предикторов были убраны такие, у которых частота самого частого значения превышала 90%; дополнительно предикторы были профильтрованы так, чтобы не осталось ни одной пары предикторов с коэффициентом корреляции Пирсона выше 0.9 (функция findCorrelation из пакета caret статистического пакета R). Перед моделированием ЧВ были нормированы к среднему ноль и единичной дисперсии.

5.2. Результаты и обсуждение

Из 1219708998 прочтений из НД2.1, 85% картируется на соответствующие геномы. В анализе использовались только сегменты, у которых для каждого вида было не менее 20 образцов с покрытием (сумма количеств включающих и

исключающих прочтений) не менее десяти прочтений: 11193 кодирующих сегментов, 21625 удержанных интронов и 12753 некодирующих сегментов. Многомерное масштабирование (к размерности два) показывает, что в наших данных доминируют межвидовые отличия (рис. 9А). Если провести аналогичный анализ для отклонений ЧВ от среднего значения в данном виде, образцы упорядочиваются по возрасту (рис. 9Б), что указывает на эволюционную консервативность возрастных изменений сплайсинга в мозге приматов и самосогласованность наших данных.

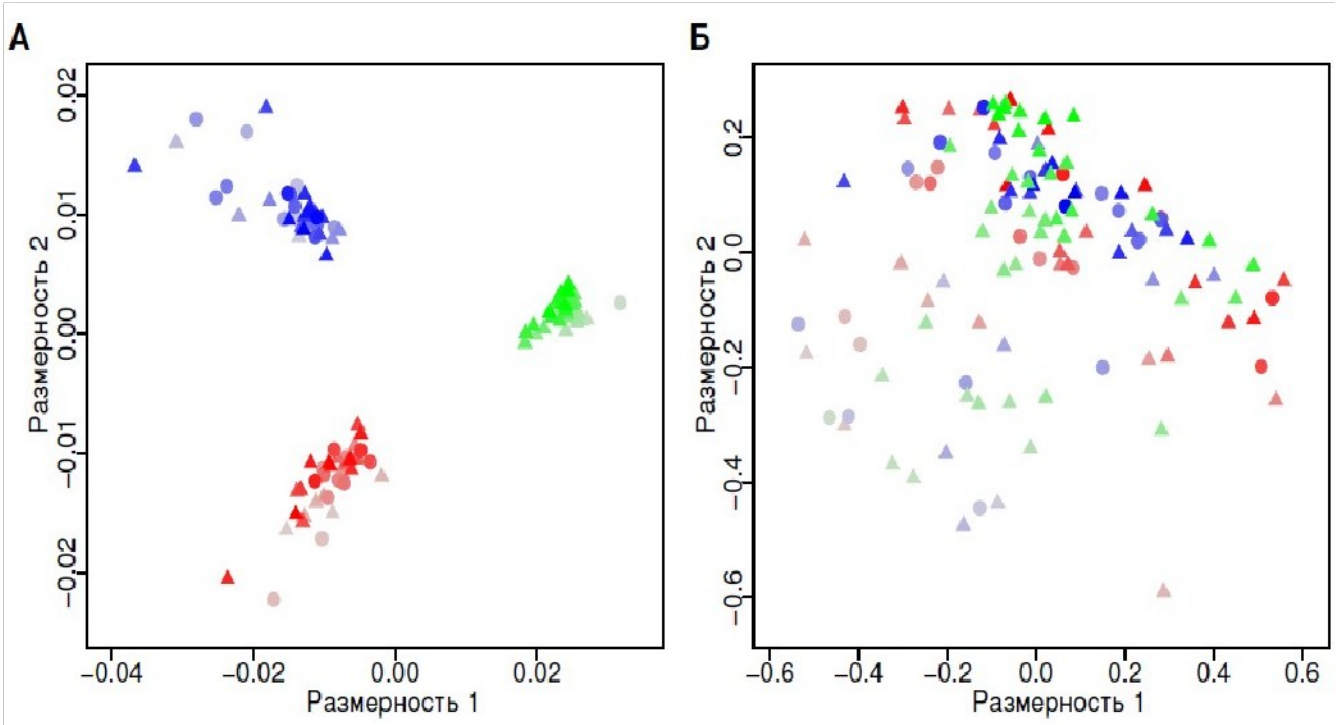

Рисунок 9. Многомерное шкалирование (к размерности два) образцов мозга на основании ЧВ для всех сегментов прошедших фильтрацию. Расстояние между образцами определялось как единица минус коэффициент корреляции Пирсона между векторами содержащими ЧВ (панель (А) или отклонение ЧВ от среднего (в данном виде для данного сегмента, панель (Б) всех сегментов в данном образце. Каждый образец показан отдельной точкой, самцы и самки обозначены треугольниками и кружками, соответственно. Образцы ткани человека, шимпанзе и макаки обозначены красным, синим и зелёным, соответственно. Яркость цвета возрастает с возрастом донора. Рисунок взят с изменениями из [Mazin и др., 2018].

Количества сегментов со значимыми различиями сплайсинга между видами, видоспецифичных сегментов и сегментов значимо меняющих сплайсинг с возрастом показаны на рисунке 10А. Количество сегментов со значимыми различиями сплайсинга между видами примерно в два раза превышает число возраст-зависимых сегментов. При этом число макака-специфичных сегментов более чем в два раза больше числа человек- и шимпанзе-специфичных сегментов, что хорошо согласуется с эволюционной историей этих трёх видов.

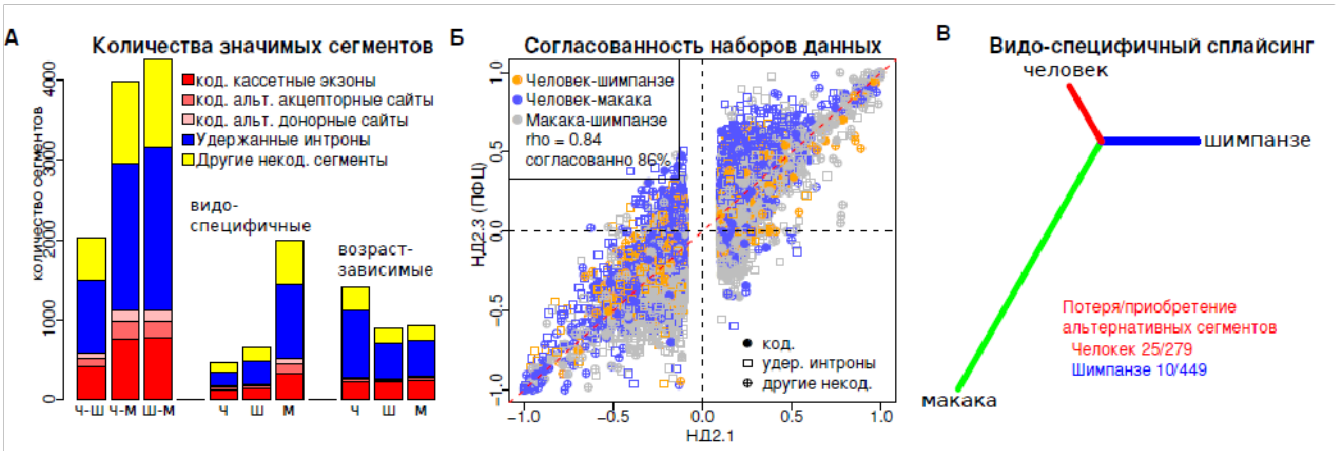

Рисунок 10: Видоспецифичные и возрастные изменения сплайсинга. (А) Высота столбцов показывает количество сегментов. значимо отличающихся между каждой из трёх пар видов (слева), видоспецифичных сегментов (посередине) и возраст-зависимых сегментов (справа). Тип сегментов показан цветом, виды обозначены буквами ч, ш и м для человека, шимпанзе и макаки соответственно. (Б) Согласованность межвидовых отличий сплайсинга в НД2.1 и НД2.2. Одна точка обозначает один сегмент в данной паре видов, пары видов обозначены цветом, форма значка обозначает тип сегмента. Коэффициент корреляции Пирсона и доля сегментов, меняющих сплайсинг в одном направлении в обоих наборах данных, показаны в легенде. Показаны только сравнения, значимые в НД2.1. (В) Дерево видов построенное на основании сходства (корреляции Пирсона) средних профилей альтернативного сплайсинга. Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.1. Различия в средних уровнях частоты включения

Чтобы проверить воспроизводимость полученных результатов, разницы между ЧВ в двух видах, вычисленные на основе НД2.1, были сравнены с

разницами, вычисленными на основе НД2.3 (рис. 10Б). Для всех типов сегментов и пар сравниваемых видов были получены высокие значения коэффициента корреляции Пирсона (более 0.75) и согласованность в направлении изменений (более 80%). Таким образом, полученные нами результаты хорошо воспроизводятся на независимых наборах данных с использованием различных протоколов секвенирования.

Анализ видоспецифичных сегментов показал, что средняя частота включения сегмента связана с направлением изменения ЧВ в ходе эволюции: мажорные сегменты как правило уменьшают, а минорные увеличивают ЧВ в ходе эволюции. Таким образом, основным направлением эволюции средних значений ЧВ в мозге приматов является увеличение альтернативности: частоты включения смещаются от нуля и единицы в направлении 0.5 (рис. 10В).

Эволюционные изменения ЧВ могут быть объяснены либо цис-эффектом (изменением регуляторных последовательностей, энхансеров или сайленсеров сплайсинга, непосредственно около альтернативного сегмента) или транс-эффектом (изменением уровней экспрессии и/или специфичностей факторов сплайсинга). Изучение транс-эффекта существенно сложнее, так как требует определения факторов сплайсинга, регулирующих каждый данный сегмент, что является трудноразрешимой задачей, так как мотивы связывания факторов плохо изучены и вырождены, а связывание факторов часто происходит кооперативно. Поэтому в данной части работы мы остановились на цис-эффекте. Для этой цели сила сайтов сплайсинга для каждого сегмента была вычислена в каждом виде как описано выше. 61% сегментов с значимыми межвидовыми изменениями сплайсинга имеют межвидовые отличия в нуклеотидных последовательностях сайтов сплайсинга. Для большинства из этих сегментов (57–75% в зависимости от типа сегмента) изменения в силе сайтов сплайсинга соответствует изменению

частоты включения (рис. 11). Хотя в общем мутации в последовательностях сайтов сплайсинга могут объяснить всего 20% межвидовых отличий, эта доля возрастает до 80% если рассматривать только высокоамплитудные изменения (рис. 12).

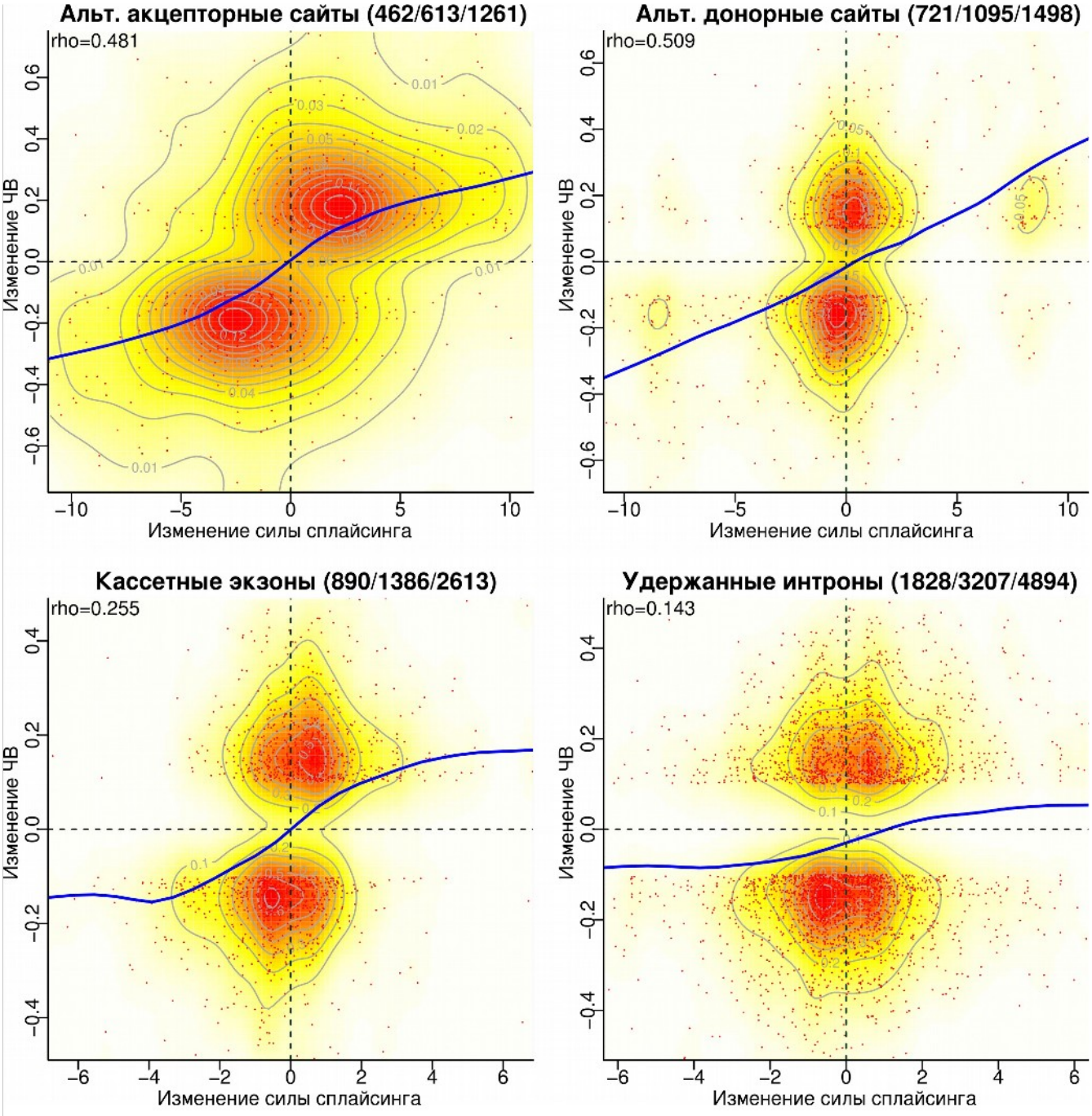

Рисунок 11: Зависимость межвидовых отличий ЧВ (вертикальная ось) от изменения силы сплайсинга сегмента (горизонтальная ось). Каждый тип сегментов показан на отдельной панели. Для каждой пары видов сегменты со значимыми отличиями ЧВ показаны точкой, если сила сплайсинга меняется более чем на один бит. Двумерная плотность показана градиентом цвета (от белого к красному), серым показаны линии с постоянным значением плотности. Синей линией показана аппроксимация локально-взвешенной полиномиальной регрессией (функция

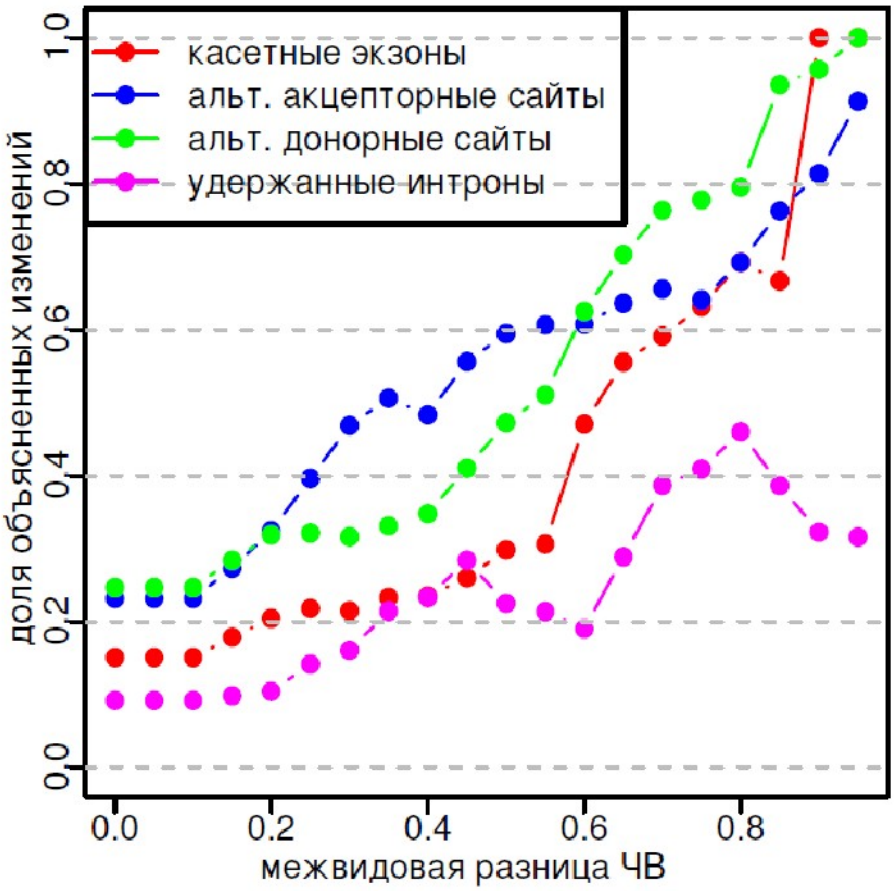

Рисунок 12: Зависимость доли межвидовых отличий объяснённых мутациями в сайтах сплайсинга, от амплитуды изменений ЧВ. Различные типы сегментов показаны разными цветами. Рисунок взят с изменениями из [Mazin и др., 2018].

В случаях, когда межвидовые изменения ЧВ не объясняются эволюцией сайтов сплайсинга, роль могут играть дополнительные регуляторные последовательности, что подтверждается более низкой эволюционной консервативностью нуклеотидных последовательностей сегментов с межвидовыми отличиями ЧВ, в сравнении с другими альтернативными сегментами (рис. 22). Рисунок взят с изменениями из [Mazin и др., 2018].

Интересным примеров эволюции АС является некодирующий ген SNHG11, содержащий ген нуклеолярной РНК. Этот ген содержит человеко- специфичные

кассетный экзон и альтернативный донорный сайт и интрон, вырезающийся только у макаки (рис. 13). Хотя функция этого гена не установлена, он экспрессируется на значительном уровне во всех трёх видах на всех стадиях развития мозга.

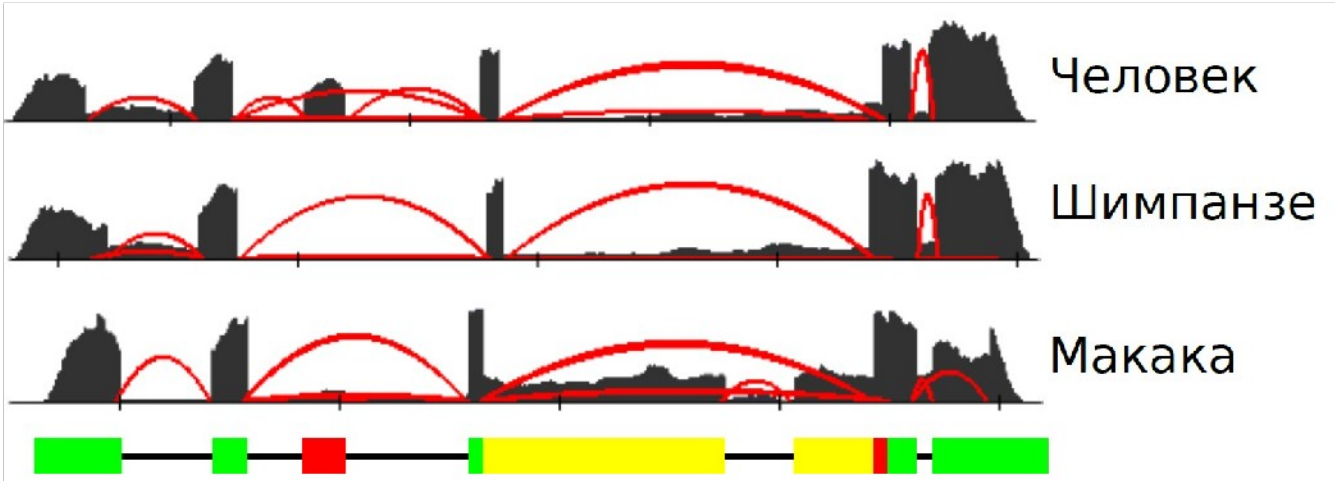

Рисунок 13. Видоспецифичные изменения сплайсинга в гене SNHG11. Показано покрытие прочтениями (все образцы из НД2.1) соответствующих участков геномов в трёх видах, высота закрашенной серой области пропорциональна числу прочтений, картирующихся на данный участок генома, красными дугами показаны прочтения, картирующиеся на экзон-экзонные границы, высота дуг пропорциональна числу прочтений. Внизу показана схема гена: сегменты, использующиеся во всех трёх видах, показаны зелёным, человек- и макака-специфичные сегменты показаны красным и жёлтым, соответственно. Рисунок взят с изменениями из [Mazin и др., 2018].

Ещё одним интересным примером человеко-специфичного сплайсинга является ген PARP2. Второй экзона этого гена содержит человеко-специфичный донорный сайт, благодаря которому экзон у человека иногда оказывается на 39 нт длиннее. Интересно, что ЧВ этого альтернативного сегмента в человеке принимает дискретные значения: либо 0, либо 1, либо около 0.5 (рис. 14А). Это объясняется человеко-специфичным одно-нуклеотидным полиморфизмом (ОНП) в основном донорном сайте [Coulombe-Huntington и др., 2009]. В данном случае изменение АС еще не зафиксировались в популяции, однако частота альтернативного аллеля

достигла 18% [Consortium, 2012].

Мы попробовали найти другие белок-кодирующие сегменты, сплайсинг которых зависит от ОНП. Для этой цели были отобраны все сегменты, удовлетворяющие следующим требованиям: а) есть хотя бы по одному образцу с ЧВ меньше 0.1, больше 0.9 и в интервале от 0.25 до 0.75; б) к каждому образцу было приписано ближайшее из 0, 0.5 и 1 значение, среднеквадратичное расстояние от реальных ЧВ до приписанных должно быть меньше 0.01. В результате этой процедуры был обнаружен ещё один сегмент: альтернативный донорный сайт четырнадцатого экзона гена ULK3 — серин-треаниновой киназы участвующей в регуляции эмбрионального развития и аутофагии (рис. 14Б). В этом сегменте находится ОНП rs12898397 представленный в 39% популяции. Альтернативный аллель в данном ОНП создаёт динуклеотид ГТ внутреннего альтернативного сайта и, таким образом, скорее всего отвечает за АС данного сегмента.

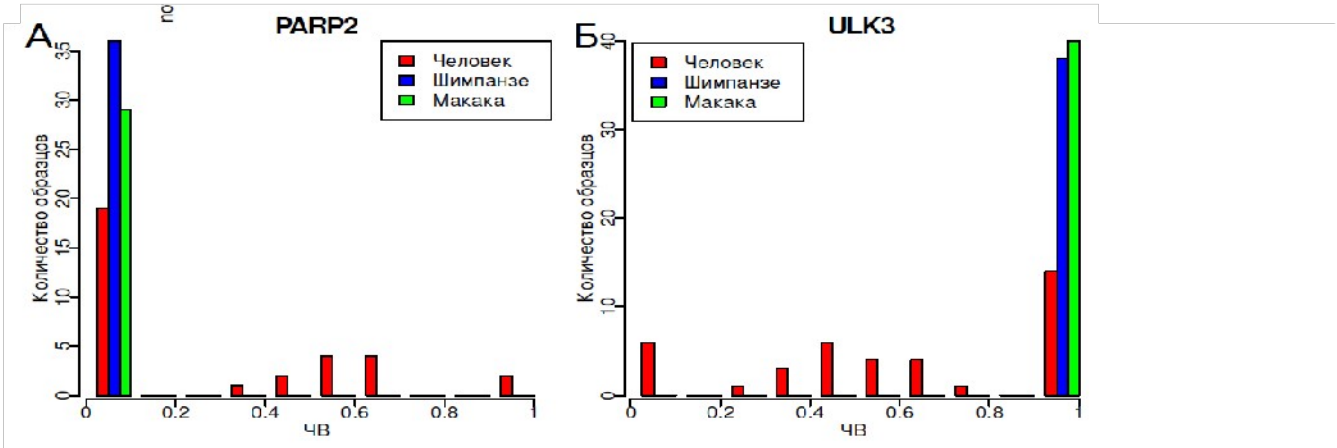

Рисунок 14. Альтернативный донорный сайт в гене PARP2 (А) и ULK3 (Б). Распределение образцов из НД2.1 по ЧВ соответствующего сегмента показано для человека, шимпанзе и макаки красным, синим и зелёным соответственно. Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.2. Возрастные изменения АС в мозге высших приматов

Сотни сегментов значимо меняют сплайсинг с возрастом в каждом из видов, и

эти изменения хорошо воспроизводятся в разных наборах данных (рис. 15А). В случае белок-кодирующих сегментов количество возрастных изменений примерно одинаково во всех трёх видах и изменения часто наблюдаются в одних и тех же сегментах (рис. 16) и происходят одинаково во всех видах (рис. 15Б). Даже в тех случаях, когда возрастные изменения ЧВ сегмента значимы только в одном из видов, как правило, все равно наблюдается положительная корреляция между видами (рис. 15В). Интересно, что у человека наблюдается в два раза больше возраст-зависимых удержанных интронов, чем у шимпанзе или макаки (рис. 16).

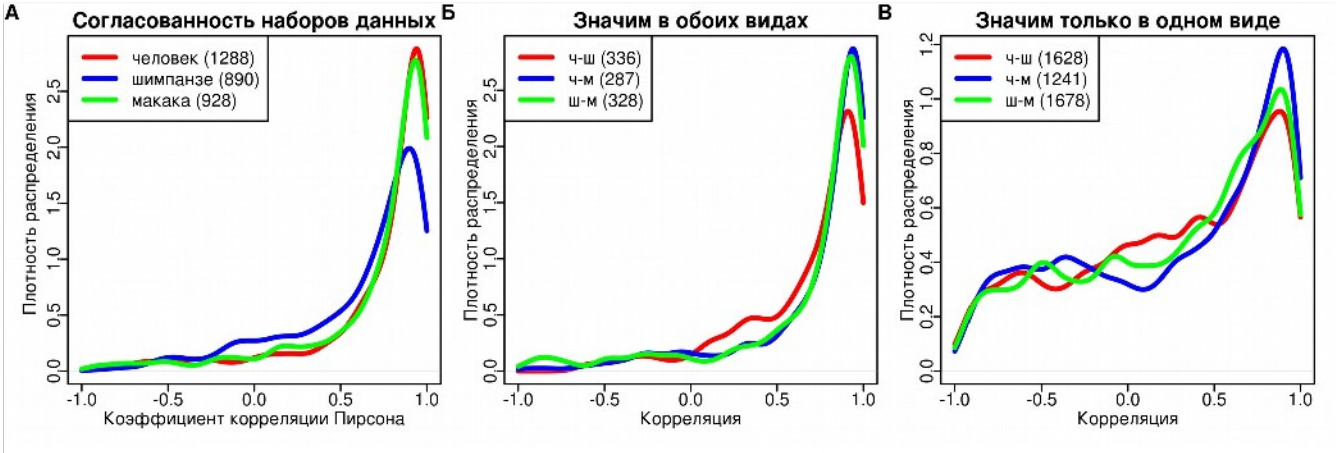

Рисунок 15. (А) Распределение коэффициента корреляции Пирсона между НД2.1 и НД2.2 для всех значимых сегментов. (Б) и (В) Распределение коэффициента корреляции Пирсона для пар видов в НД2.1. Использовались сегменты значимые в обоих видах (Б) или только в одном виде (В). Ч, ш и м обозначают человека, шимпанзе и макаку соответственно. Коэффициенты корреляции рассчитывались на основании корректированных возрастов. Рисунок взят с изменениями из [Mazin и др., 2018].

Функциональный анализ показывает, что гены с возраст-зависимым АС вовлечены во многие функции, связанные с развитием мозга. Например, гены, у которых сплайсинг белок-кодирующих сегментов меняется с возрастом хотя бы в одном из видов, связаны с такими функциями и клеточными структурами как клеточная адгезия, нейрогенез, дифференцировка нейронов, передача нервного импульса, синапс, аксон, ионные каналы и многими другими (приложение 8).

Недавно было показано что экзоны не длиннее 27 нт (микроэкзоны) специфично используются в нервной ткани [Irimia и др., 2014]. Среди сегментов, прошедших фильтрацию, были обнаружены 176 микроэкзонов. Гены, содержащие микроэкзоны, значимо ассоциированы с развитием нервной системы и перепредставлены среди возраст-зависимых сегментов (тест Фишера, $p < 10^{-10}$).

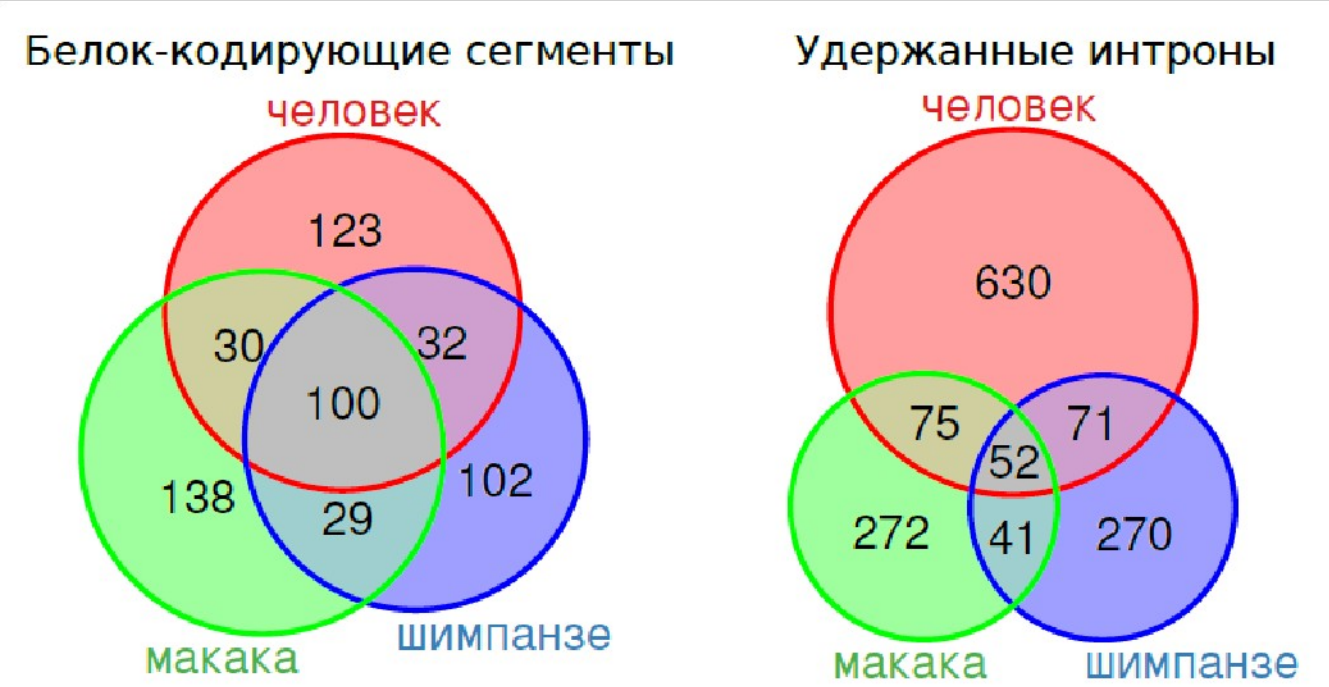

Рисунок 16: Диаграмма Венна для количества возраст-зависимых белок-кодирующих сегментов и удержанных интронов. Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.2.1. Соотнесение возрастов между видами

Так как рассматриваемые в настоящей работе виды сильно различаются по продолжительности жизни, это должно быть учтено при межвидовом сравнении возрастных паттернов АС. Однако, использовать максимальную продолжительность жизни для корректировки возрастов затруднительно, так как для человека она непропорционально высока (122.5 года) по сравнению с шимпанзе и макакой (59.4 и 40 лет соответственно [Tacutu и др., 2013]), вероятно, из-за наличия гораздо большего числа наблюдений для человека по сравнению с

обезьянами. Поэтому в данной работе поправочные коэффициенты для возрастов были вычислены на основании возрастных изменений АС (см методы, раздел 5.1.8). Распределения этих коэффициентов для каждой пары видов имеют единственный максимум (рис. 17). Для перевода возрастов обезьян в возраст человека были использованы моды этих распределений. Соответственно, возраст макаки был умножен примерно на 3.5, а коэффициент для перевода возраста шимпанзе в шкалу человека был равен примерно 1.5. Интересно, что поскольку возраст считали с момента зачатия, в соответствии с этими коэффициентами новорожденная макака соответствует десятимесячному, а новорождённый шимпанзе трёхмесячному человеческому ребёнку. Далее в работе все возрасты были приведены к возрастам человека при помощи указанных выше коэффициентов. С учётом такой коррекции возрастов большинство сегментов показывают высокую корреляцию возрастных изменения сплайсинга между видами (рис. 15Б и В). Таким образом, с учётом различий в продолжительности жизни, возрастная регуляция АС в мозгах высших приматов достаточно консервативна для всех типов сегментов. Коррекция возрастов также может частично объяснить отличия в поведении удержанных интронов у человека и обезьян. Действительно, основные возрастные изменения удержания интронов в человеке происходят на ранних стадиях постнатального развития (смотри ниже). Однако в соответствии с описанной выше возрастной коррекцией, этот период соответствует пренатальному развитию у обезьян, который фактически не покрыт в данном исследовании.

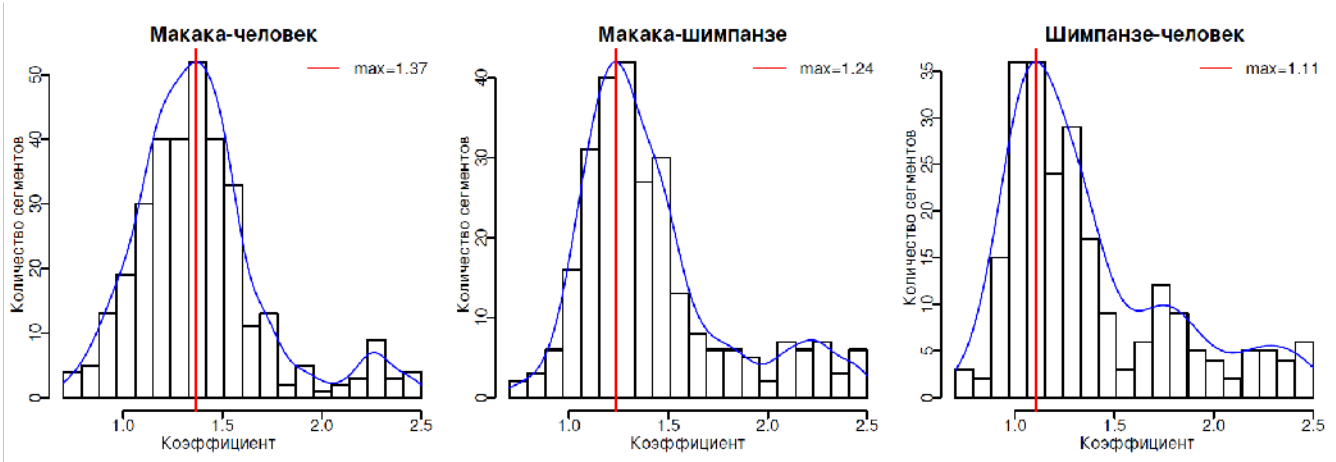

Рисунок 17. Распределение сегментов по оптимальным значениям коэффициента перевода возрастов между тремя парами видов (в шкале возраст$^{0.25}$). Мода распределения показана красной линией. Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.2.2. Кластерный анализ возраст-зависимых сегментов

Чтобы лучше охарактеризовать разнообразие возраст-зависимых изменений АС возраст-зависимые сегменты каждого типа были разбиты на шесть кластеров (рис. 18, 19).

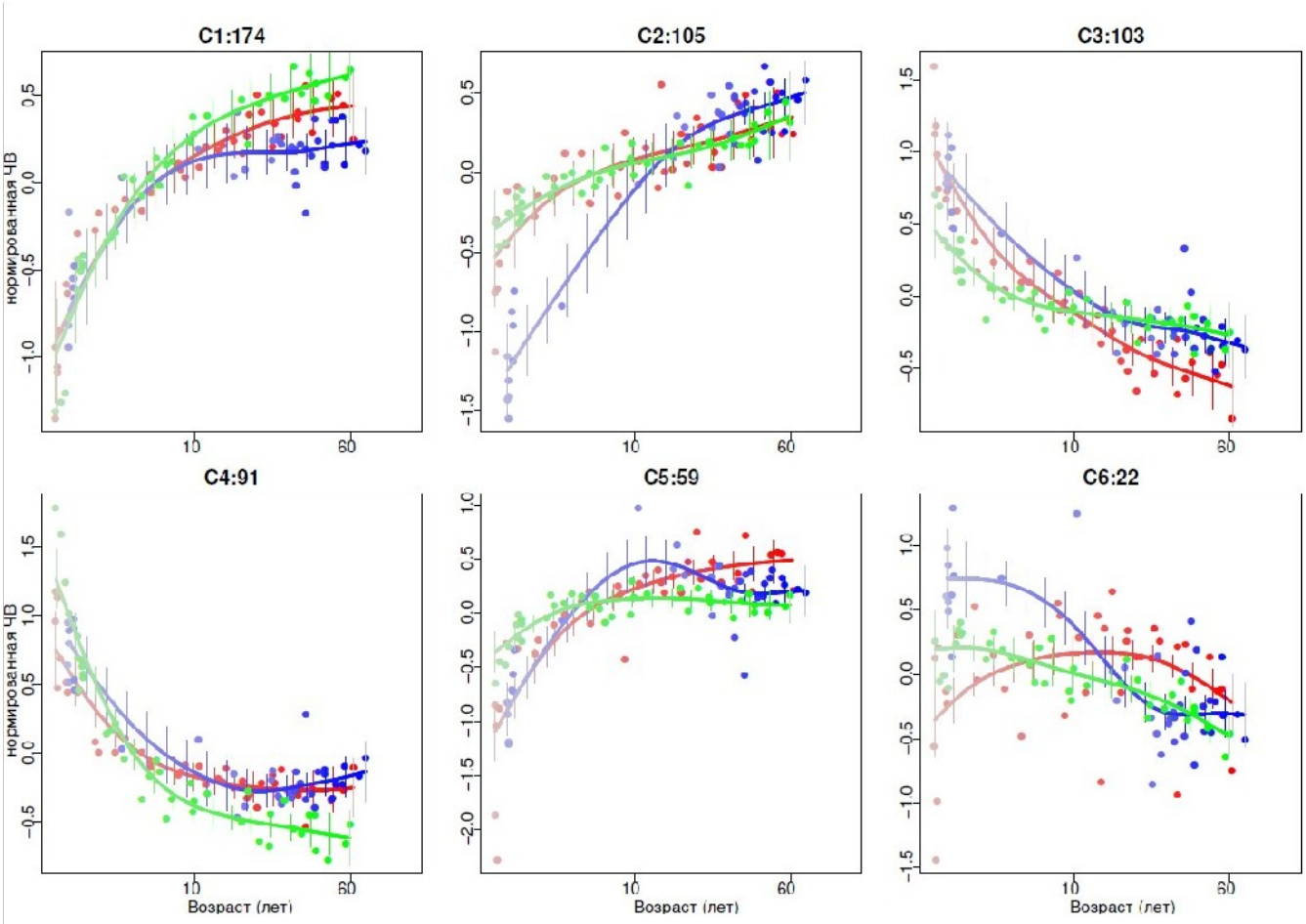

Рисунок 18. Разбиение возраст-зависимых белок-кодирующих сегментов на шесть кластеров. Каждый кластер показан на отдельной панели, кластеры упорядочены по числу сегментов, относящихся к ним (указано в названии панели). Каждая точка обозначает среднюю нормализованную ЧВ (вертикальная ось) в зависимости от возраста (горизонтальная ось), кривыми показана аппроксимация кубическим сплайном с четырьмя степенями свободы. Вертикальные линии показывают стандартное отклонение линии аппроксимации. Человек, шимпанзе и макака показаны красным, синим и зелёным, соответственно. Возрасты шимпанзе и макаки пересчитаны в шкалу человека. Рисунок взят с изменениями из [Mazin и др., 2018].

Основной характеристикой всех кластеров является то, что большая часть изменений происходит в первые годы жизни. Интересно, что белок-кодирующие сегменты меняют сплайсинг во всех возможных направлениях и примерно одинаково во всех трёх видах, в то время как ЧВ большинства удержанных интронов падает с возрастом, при этом у 52% интронов это падение проявляется

сильнее у человека чем у других видов (первый кластер на рис. 19).

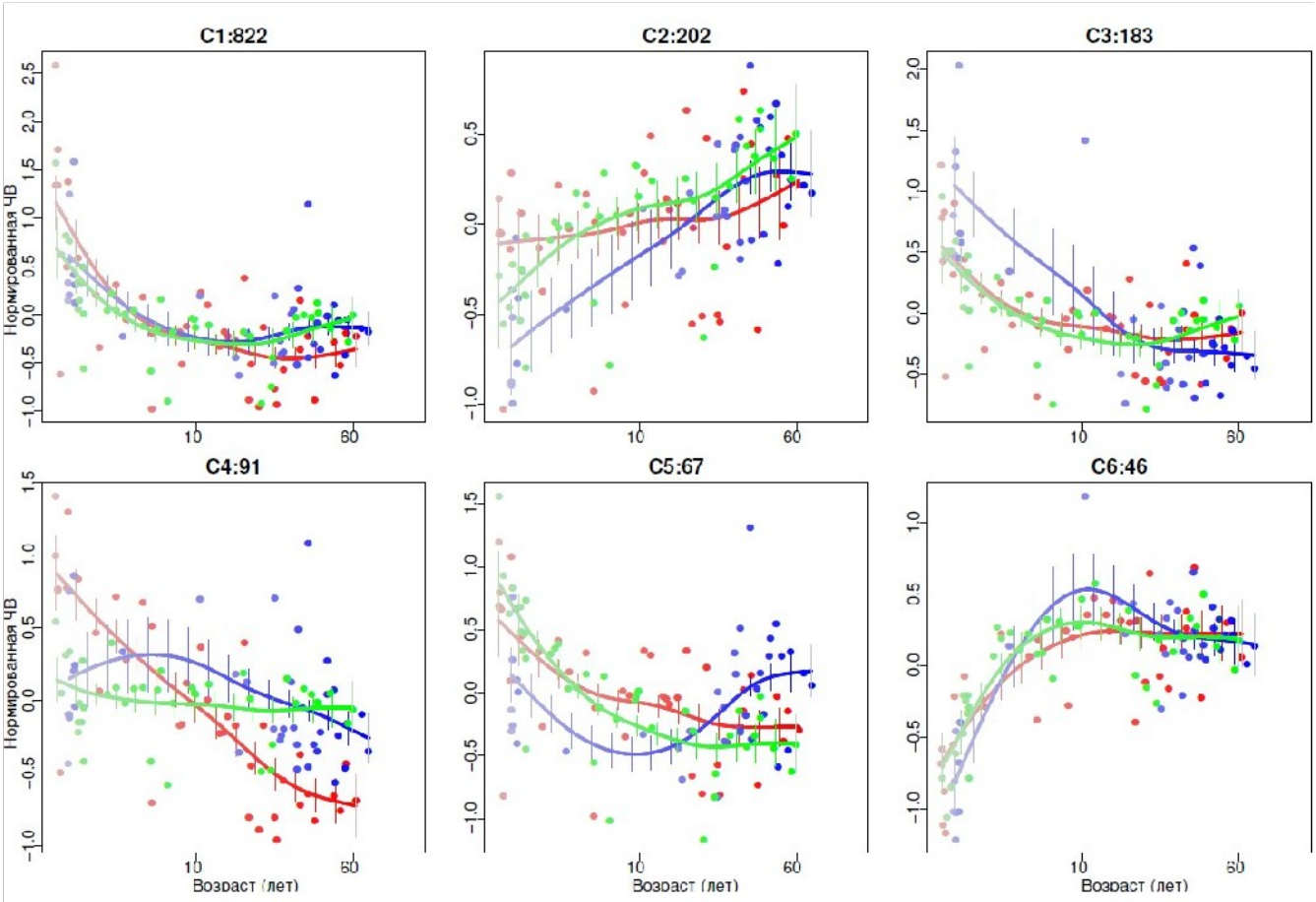

Рисунок 19. Разбиение возраст-зависимых удержанных интронов на шесть кластеров. Описание см. в подписи к рис. 18. Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.2.3. Удержанные интроны

Удержание интронов может играть роль в регуляции концентраций мРНК, вызывая ее деградацию. Чтобы проверить эту гипотезу, были вычислилены коэффициенты корреляции Пирсона между ЧВ удержанного интрона и уровнем экспрессии соответствующего гена и, в качестве контроля, произвольно выбранного гена. Дополнительно эти расчёты были повторены для других типов альтернативных сегментов. Как видно на рис. 20, во всех трёх видах между ЧВ удержанных интронов и уровнями экспрессии соответствующих генов, но не случайных генов, наблюдается значительная отрицательная корреляция. Для

возраст-зависимых сегментов других типов такой корреляции не наблюдается. В [Yap и др., 2012] было показано, что кроме обычной ПСК-зависимой деградации, в ходе развития головного мозга у мышей удержание последнего интрона может приводить к деградации мРНК в ядре. У человека, но не у обезьян, возраст-зависимые удержанные интроны значимо чаще оказываются последним интроном гена, чем удержанные интроны, у которых ЧВ не меняется с возрастом (тест Фишера, $p<0.027$, отношение шансов = 1.28). Таким образом, механизм ПСК-независимой ядерной деградации мРНК может работать в раннем постнатальном развитии мозга человека, но не обезьян. Таким образом, обнаруженные нами возрастные изменения ЧВ удержанных интронов могут играть роль в регуляции экспрессии генов в ходе развития головного мозга приматов. В [Liu и др., 2012] было показано, что максимум экспрессии, генов участвующих в синаптогенезе, у человека наблюдается существенно позже, чем у шимпанзе и макаки. По нашим данным, 24 из 184 генов с таким поведением содержат возраст-зависимые интроны, что значимо больше, чем можно ожидать случайно (тест Фишера, $p=0.0005$).

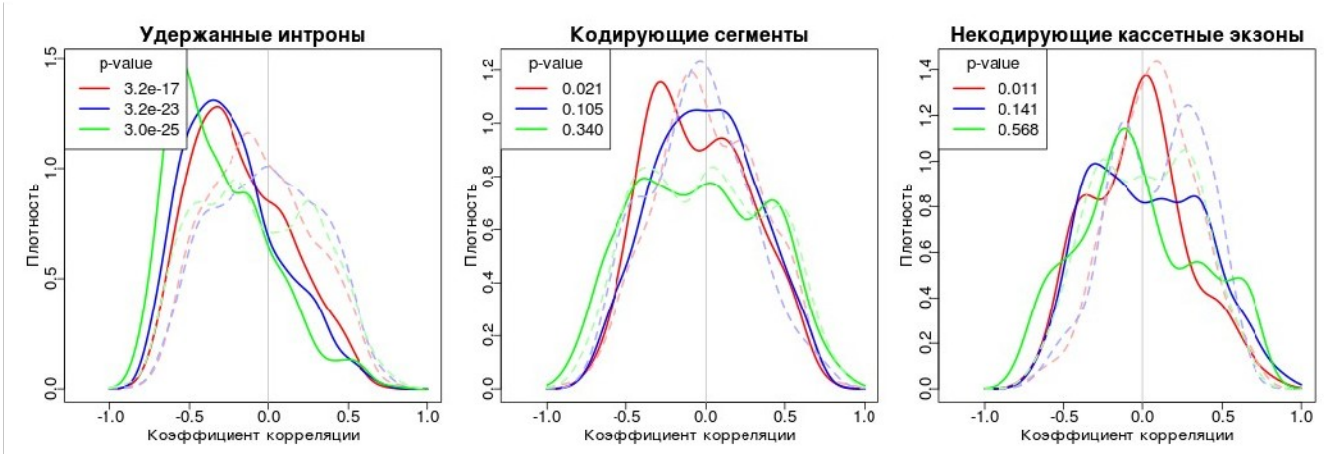

Рисунок 20. Распределение коэффициента корреляции Пирсона между ЧВ возраст-зависимых сегментов и уровнями экспрессии соответствующих (показано сплошной линией) или случайно выбранных (показано пунктиром) генов в трёх видах: человек (красный), шимпанзе (синий) и макака (зеленый). Значимость отличий распределений между правильными и случайными парами сегмент-ген показаны в легенде (*p* согласно тесту Вилкоксона). Рисунок взят с изменениями из [Mazin и др., 2018].

5.2.2.4. Возрастная регуляция альтернативного сплайсинга

Для исследования возрастной регуляции АС мы решили сфокусироваться на одном типе сегментов — кассетных экзонах. Если их регуляция осуществляется за счёт специфического связывания факторов сплайсинга с РНК в непосредственной близости от альтернативного экзона, то стабилизирующий отбор должен действовать сильнее на последовательность в непосредственной близости от регулируемого альтернативного экзона, чем около константного. Наш анализ показал, что сами возраст-зависимые экзоны, а так же фланкирующие их участки ДНК более консервативны, чем константные или альтернативные, но не возраст-зависимые экзоны (рис 21). Чтобы определить, какие непосредственно факторы сплайсинга могут быть связаны с обнаруженными нами возрастными изменениями АС была использована база данных сайтов связывания факторов сплайсинга CISBP-RNA [Ray и др., 2013], содержащая информацию о 219 мотивах

связываемых 392 РНК-связывающими белками человека (экспрессия 315 из них была детектированна в данной работе). Были выделены шесть участков по 50 нт до, внутри и после каждого кассетного сегмента и вычисленна средняя аффинность (по методике, описанной в [Ray и др., 2013]) внутри каждого участка для каждого сегмента и мотива из базы данных (см методы, раздел 5.1.11). Были обнаружены 23 мотива (приложение 9), аффинность которых была значимо увеличена в хотя бы одном из шести участков возраст-зависимых кассетных экзонов. Двадцать шесть факторов сплайсинга связывают хотя бы один из этих мотивов и экспрессируются на детектируемом в наших данных уровне. Двадцать три из них значимо меняют экспрессию с возрастом хотя бы в одном из трёх видов, что в более чем три раза чаще, чем можно ожидать случайно (тест Фишера, $p<0.025$). Шесть из этих факторов значимо меняют экспрессию с возрастом во всех трёх видах. Как минимум четыре из них связаны с функционированием головного мозга: *MBNL2* и *MBNL1*, вовлечённые в развитие миотонической дистрофии, и связанных с ней нарушений в работе мозга [Charizanis и др., 2012]; *RBM4*, регулирующий сплайсинг мРНК, кодирующей белок *tau,* вовлечённый в болезнь Альцгеймера [Kar и др., 2006, с. 4]; *YB-1*, подавление которго материнскими антителами в ходе эмбрионального развития связанно с аутизмом [Braunschweig и др., 2013]. RBFOX2, вовлечённый в развитие головного мозга [Gehman и др., 2012], и *RBM8A*, связанный с аутизмом, шизофренией и микроцефалией [Zou и др., 2015] также связывают обогащённые мотивы и значимо меняют экспрессию с возрастом в части видов (приложение 9).

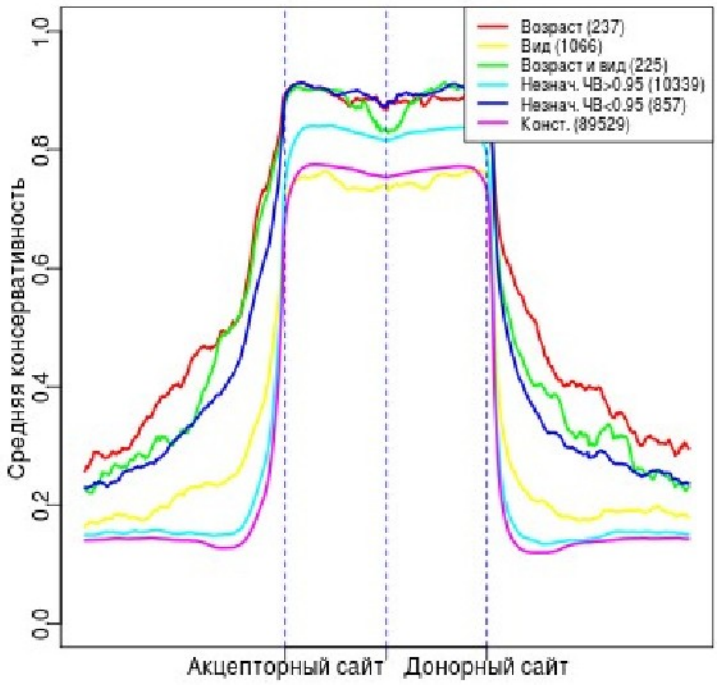

Рисунок 21: Средняя консервативность (phastcons для приматов) около кассетных экзонов (слева) и интронов (справа) для различных категорий сегментов: возраст-зависимых; значимо различающихся между видами; возраст-зависимых и значимо различающихся между видами; незначимых (отдельно показаны сегменты с ЧВ больше или меньше чем 0.95); и константных. Рисунок взят с изменениями из [Mazin и др., 2018].

Наличие мотивов, значимо часто встречающихся около возраст-зависимых экзонов и связанных с ними факторов сплайсинга с возраст-зависимой экспрессией, позволяет построить простую механистическую модель для предсказания возрастных изменений ЧВ, предположив, что они пропорциональны линейной комбинации произведений уровней экспрессии факторов сплайсинга и аффинностей соответствующих мотивов (см методы 5.1.14). Чтобы избежать переобучения была использована L1 -регуляризация с весовым значением 0.01, соответствующем оптимальному значению коэффициента корреляции Пирсона в кросс-валидации.

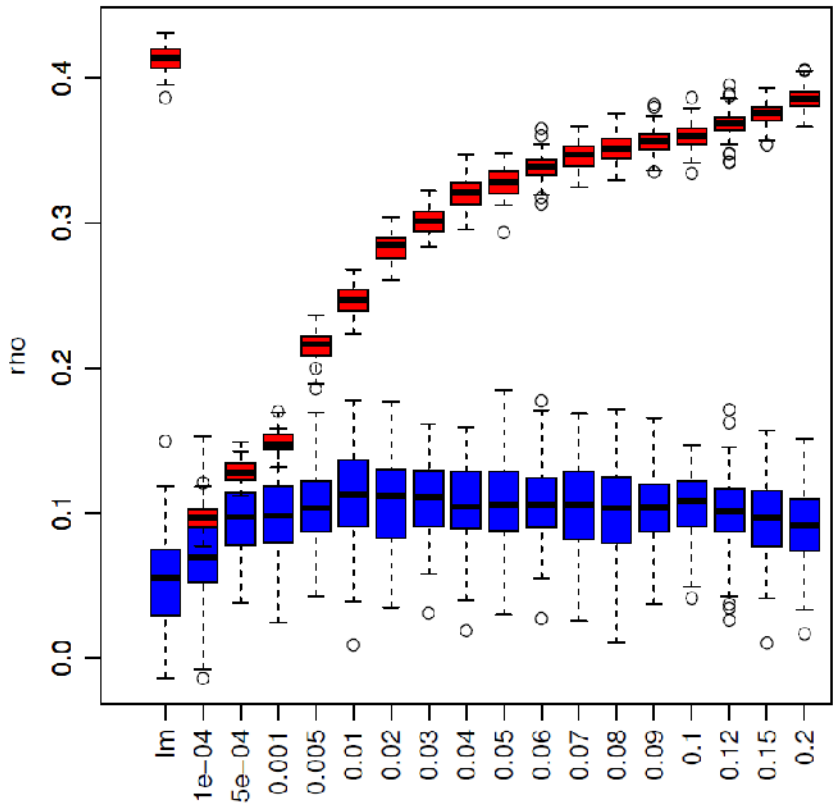

Рисунок 22: Зависимость коэффициента корреляции Пирсона для обучающей (красная) и тестовой (синяя) выборок при предсказании ЧВ по аффинностям мотивов и уровням экспрессии связывающих их генов в зависимости от весового значения L1 регуляризации. Крайнее слева распределение соответствует линейной модели без регуляризации. Распределения для каждого случая были получены в результате 100 случайных разбиений всех возраст-зависимых экзонов на обучающую (70%) и тестовую (30%) выборки. Рисунок взят с изменениями из [Mazin и др., 2018].

Медиана распределения коэффициента корреляции Пирсона между реальными и предсказанными моделью и ЧВ равна 0.11, что значимо больше нуля (тест Вилкоксона, $p<4\times10^{-18}$). Чтобы дополнительно верифицировать результаты, моделирование было повторено с использованием либо перемешанных относительно предикторов ЧВ, либо ЧВ, сгенерированных случайным образом (нормальное распределение с нулевым средним, и единичной дисперсией). В обоих случаях в кросс-верификации наблюдается коэффициент корреляции Пирсона, не отличающийся от нуля (рис. 23). Эти результаты дополнительно

подтверждают, что обнаруженные в настоящей работе мотивы и факторы действительно отвечают за возрастные изменения ЧВ, а изменения АС можно предсказать, исходя из простых начальных принципов.

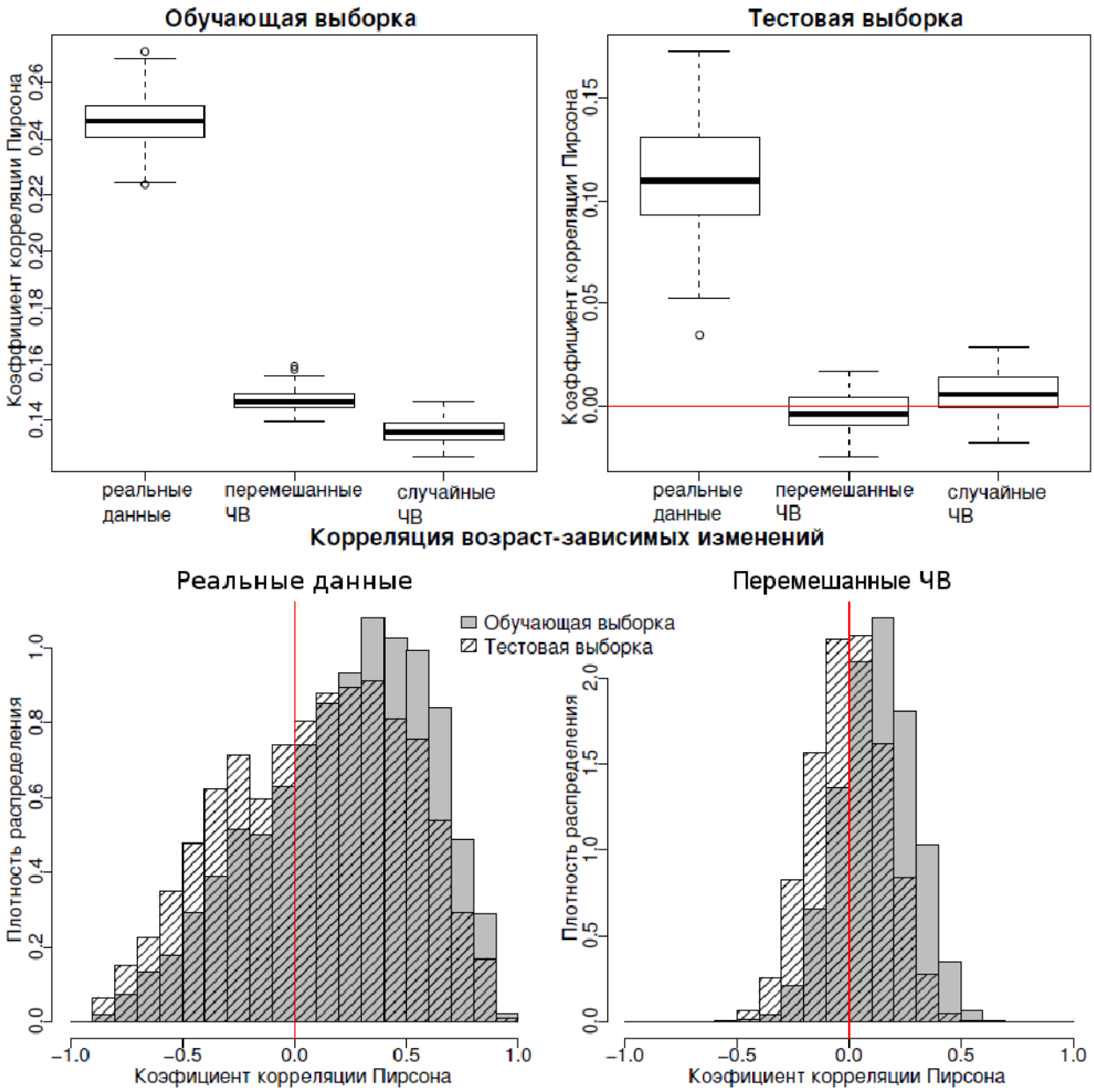

Рисунок 23: Моделирование ЧВ при помощи уровней экспрессии и аффинности факторов сплайсинга. Сверху показаны распределения коэффициентов корреляции Пирсона для реальных, перемешанных и случайных данных для обучающей и тестовой выборок. Здесь коэффициент корреляции рассчитывался между всеми ЧВ всех сегментов, попавших в данную выборку. Чтобы оценить качество предсказаний для отдельных сегментов, был посчитан коэффициент корреляции для индивидуальных сегментов, распределения таких коэффициентов для реальных (слева) и перемешанных (справа) данных показаны внизу. Рисунок взят с изменениями из [Mazin и др., 2018].

5.3. Выводы

В этой части работы были получены следующие основные результаты.

1. При сравнении возрастных изменений альтернативного сплайсинга в мозге приматов межвидовые отличия доминируют над возрастными.
2. На основании возрастных паттернов альтернативного сплайсинга были найдены оптимальные коэффициенты пересчёта возрастов трёх изучаемых видов. Полученные коэффициенты согласуются с данными по продолжительности жизни. Согласно этим коэффициентам, новорожденные макаки и шимпанзе соответствуют десяти- и трехмесячным детям человека, соответственно.
3. Возрастная регуляция сплайсинга кодирующих и кассетных не кодирующих белок сегментов консервативна в изучаемых видах. В то же время, количество возраст-зависимых удержанных интронов существенно больше у человека, чем у других видов.
4. Изменения частот включения удержанных интронов с возрастом могут играть роль в регуляции концентраций мРНК.
5. Существенная часть межвидовых изменений АС может объясняться эволюцией сайтов сплайсинга.
6. Для возраст-зависимых кассетных экзонов были найдены предположительные регуляторы (факторы сплайсинга) и их мотивы.

6. Общие выводы

1. Разработан, реализован и валидирован метод анализа альтернативного сплайсинга на основании данных РНК-Сек.
2. Межвидовые отличия альтернативного сплайсинга доминируют над

возрастными.

3. Удержанные интроны и белок-кодирующие сегменты существенно различно регулируются с возрастом. У большинства удержанных интронов частота включения падает в ходе развития, в то время как у белок-кодирующих сегментов частота включения меняется в обе стороны примерно с равной частотой.
4. Изменения альтернативного сплайсинга в ходе развития мозга очень схожи у приматов.
5. Большинство высокоамплитудных межвидовых отличий альтернативного сплайсинга объясняются изменениями нуклеотлидных последовательностей сайтов сплайсинга.

7. Список публикаций по теме диссертации

7.1. Статьи в научных журналах

1. Mazin P. и др. Widespread splicing changes in human brain development and aging // Mol. Syst. Biol. 2013. Т. 9. С. 633.
2. Mazin P.V. и др. Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques // RNA. 2018. Т. 24. № 4. С. 585–596.

7.2. Тезисы конференций

1. Mazin P. и др. Investigation of age related alternative splicing changes in human brain using solexa sequencing // Moscow Conference on Computational Molecular Biology `09 (постерный доклад)
2. Mazin P. и др. Splicing changes in human brain over the course of lifespan // CSH-Asia Conference “Computational Biology”. Suzhou, China, 27.09-1.10.2010 (постерный доклад)
3. Mazin P. и др. Splicing differences in primate brain development // Moscow Conference on Computational Molecular Biology `11 (постерный доклад)

4. Мазин П.В. и др. Изменения сплайсинга в ходе развития мозга у приматов. Информационные технологии и системы (ИТиС) // Октябрь 2 – 7, 2011, Геленджик, Россия (устный доклад)

5. Mazin P. и др. Age changes and tissue differences of alternative splicing in the primate brain // The Eighth Winter Symposium on Chemometrics, Moscow, Febrary 27 - March 2, 2012 (устный доклад)

6. Mazin P. и др. Widespread differences in age-related splicing patterns between higher primates // Information Technologies and Systems (ITaS). August 19 – 25, 2012, Petrozavodsk, Russia (постерный доклад)

7. Mazin P. и др. Widespread splicing changes in human brain development and aging // ECCB'12 - European Conference on Computational Biology 2012, 9-12 September 2012, Basel Switzerland.

8. Mazin P. и др. Conserved age-related splicing regulation in primate brains // Moscow Conference on Computational Molecular Biology `13 (устный доклад)

9. Mazin P. Quantification of alternative splicing using RNA-seq // Postgenome-2014. 29.10-01.11.2014 Kazan, Russia (устный доклад)

10. Mazin P. и др. Conservation and evolution of splicing patterns during postnatal development of prefrontal cortex in primates // IMGC 2015 Yokohama, Japan. 8-11.11.2015 (устный доклад)

1. Список литературы

1. Anders S. и др. Detecting differential usage of exons from RNA-Seq data // Nature Precedings. 2012.

2. Anders S., Huber W. Differential expression analysis for sequence count data // Genome Biology. 2010. Т. 11. № 10. С. R106.

3. Andersson R. и др. Nucleosomes are well positioned in exons and carry characteristic histone modifications // Genome Res. 2009.

4. Bakken T.E. и др. A comprehensive transcriptional map of primate brain development // Nature. 2016. Т. 535. № 7612. С. 367–375.

5. Baralle F.E. Alternative splicing as a regulator of development and tissue identity // Nat Rev Mol Cell Biol. 2017. Т. 18. № 7. С. 15.

6. Barash Y. и др. Deciphering the splicing code // Nature. 2010. Т. 465. № 7294. С. 53–59.

7. Barbosa-Morais N.L. и др. The evolutionary landscape of alternative splicing in vertebrate species // Science. 2012. Т. 338. № 6114. С. 1587–1593.

8. Benjamini Y., Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing // Journal of the Royal Statistical Society. Series B (Methodological). 1995. Т. 57. № 1. С. 289–300.

9. Boutz P.L. и др. A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons // Genes Dev. 2007. Т. 21. № 13. С. 1636–1652.

10. Braunschweig D. и др. Autism-specific maternal autoantibodies recognize critical proteins in developing brain // Transl Psychiatry. 2013. Т. 3. № 7. С. e277.

11. Brawand D. и др. The evolution of gene expression levels in mammalian organs // Nature. 2011. Т. 478. № 7369. С. 343–348.

12. Brent R.P. Algorithms for Minimization Without Derivatives. Dover Publications, 1973. 208 С.

13. Brooks A.N. и др. Conservation of an RNA regulatory map between Drosophila and mammals // Genome Res. 2011. Т. 21. № 2. С. 193–202.

14. Carninci P. и др. Genome-wide analysis of mammalian promoter architecture and evolution // Nature Genetics. 2006. Т. 38. № 6. С. 626–635.

15. Chan E.T. и др. Conservation of core gene expression in vertebrate tissues // J Biol. 2009. Т. 8. № 3. С. 33.

16. Chang T.-C. и др. Genome-wide annotation of microRNA primary transcript structures reveals novel regulatory mechanisms // Genome Research. 2015. Т. 25. № 9. С. 1401–1409.

17. Chang Y.-F., Imam J.S., Wilkinson M.F. The Nonsense-Mediated Decay RNA Surveillance Pathway // Annual Review of Biochemistry. 2007. Т. 76. № 1. С. 51–74.

18. Charizanis K. и др. Muscleblind-like 2-mediated alternative splicing in the developing brain and dysregulation in myotonic dystrophy // Neuron. 2012. Т. 75. № 3. С. 437–450.

19. Chen M., Manley J.L. Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches // Nature Reviews Molecular Cell Biology. 2009.

20. Chow J.C. и др. Silencing of the Mammalian X Chromosome // Annual Review of Genomics and Human Genetics. 2005. Т. 6. № 1. С. 69–92.

21. Chow L.T. и др. An amazing sequence arrangement at the 5’ ends of adenovirus 2 messenger RNA // Cell. 1977. Т. 12. № 1. С. 1–8.

22. Colantuoni C. и др. Temporal dynamics and genetic control of transcription in the human prefrontal cortex // Nature. 2011. Т. 478. № 7370. С. 519–523.

23. Consortium T. 1000 G.P. An integrated map of genetic variation from 1,092 human genomes // Nature. 2012. Т. 491. № 7422. С. 56–65.

24. Cook K.B. и др. RBPDB: a database of RNA-binding specificities // Nucleic Acids Research. 2010. Т. 39. № Database. С. D301–D308.

25. Coulombe-Huntington J. и др. Fine-Scale Variation and Genetic Determinants of Alternative Splicing across Individuals // PLoS Genet. 2009. Т. 5. № 12. С. e1000766.

26. Crespi B., Summers K., Dorus S. Adaptive evolution of genes underlying schizophrenia // Proc. R. Soc. B. 2007. Т. 274. № 1627. С. 2801–2810.

27. Crick F. Central dogma of molecular biology // Nature. 1970. Т. 227. № 5258. С. 561–563.

28. David C.J., Manley J.L. Alternative pre-mRNA splicing regulation in cancer: pathways and programs unhinged // Genes & Development. 2010. Т. 24. № 21. С. 2343–2364.

29. Di Giammartino D.C., Nishida K., Manley J.L. Mechanisms and Consequences of Alternative Polyadenylation // Molecular Cell. 2011. Т. 43. № 6. С. 853–866.

30. Dillman A.A. и др. mRNA expression, splicing and editing in the embryonic and adult mouse cerebral cortex // Nature Neuroscience. 2013. Т. 16. № 4. С. 499–506.

31. Dobson A.J. An introduction to generalized linear models. Boca Raton: Chapman & Hall/CRC, 2002.

32. Doolittle W.F. Is junk DNA bunk? A critique of ENCODE // PNAS. 2013. Т. 110. № 14. С. 5294–5300.

33. Dunham I. и др. An integrated encyclopedia of DNA elements in the human genome // Nature. 2012. Т. 489. № 7414. С. 57–74.

34. Early P. и др. Two mRNAs can be produced from a single immunoglobulin μ gene by alternative RNA processing pathways // Cell. 1980. Т. 20. № 2. С. 313–319.

35. eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease // Nat. Genet. 2017. Т. 49. № 12. С. 1664–1670.

36. Elsner M., Mak H.C. A modENCODE snapshot // Nat Biotech. 2011. Т. 29. № 3. С. 238–240.

37. Falcon S., Gentleman R. Using GOstats to test gene lists for GO term association // Bioinformatics. 2007. Т. 23. № 2. С. 257–258.

38. Fehlbaum P. и др. A microarray configuration to quantify expression levels and relative abundance of splice variants // Nucleic Acids Res. 2005. Т. 33. № 5. С. e47.

39. Flicek P. и др. Ensembl 2013 // Nucleic Acids Research. 2012. Т. 41. № D1. С. D48–D55.

40. Frank M. и др. Differential expression of individual gamma-protocadherins during mouse brain development // Molecular and Cellular Neuroscience. 2005. Т. 29. № 4. С. 603–616.

41. Franke L., Jansen R.C. eQTL analysis in humans // Methods Mol. Biol. 2009. Т. 573. С. 311–328.

42. Fuda N.J., Ardehali M.B., Lis J.T. Defining mechanisms that regulate RNA polymerase II transcription in vivo // Nature. 2009. Т. 461. № 7261. С. 186–192.

43. Gao F., Foat B.C., Bussemaker H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data // BMC Bioinformatics. 2004. Т. 5. № 1. С. 31.

44. Gao S. и др. PacBio full-length transcriptome profiling of insect mitochondrial gene expression // RNA Biol. 2016. С. 1–6.

45. Garber M. и др. Computational methods for transcriptome annotation and quantification using RNA-seq // Nature Methods. 2011. Т. 8. № 6. С. 469–477.

46. Gehman L.T. и др. The splicing regulator Rbfox2 is required for both cerebellar development and mature motor function // Genes Dev. 2012. Т. 26. № 5. С. 445–460.

47. Gennarino V.A. и др. MicroRNA target prediction by expression analysis of host genes // Genome Res. 2009. Т. 19. № 3. С. 481–490.

48. Giulietti M. и др. SpliceAid-F: a database of human splicing factors and their RNA-binding sites // Nucleic Acids Research. 2012. Т. 41. № D1. С. D125–D131.

49. Graur D. и др. On the Immortality of Television Sets: “Function” in the Human Genome According to the Evolution-Free Gospel of ENCODE // Genome Biol Evol. 2013. Т. 5. № 3. С. 578–590.

50. Griffith M. и др. Alternative expression analysis by RNA sequencing // Nature Methods. 2010. Т. 7. № 10. С. 843–847.

51. Grosso A.R. и др. Tissue-specific splicing factor gene expression signatures // Nucleic Acids Research. 2008. Т. 36. № 15. С. 4823–4832.

52. Guillozet-Bongaarts A.L. и др. Altered gene expression in the dorsolateral prefrontal cortex of individuals with schizophrenia // Mol Psychiatry. 2013.

53. Hansen T.B. и др. Natural RNA circles function as efficient microRNA sponges // Nature. 2013. Т. 495. № 7441. С. 384–388.

54. Hartley S.W., Mullikin J.C. Detection and visualization of differential splicing in RNA-Seq data with JunctionSeq // Nucl. Acids Res. 2016. С. gkw501.

55. Hinrichs A.S. и др. The UCSC Genome Browser Database: update 2006 // Nucleic Acids Res. 2006. Т. 34. № Database issue. С. D590-598.

56. Hoheisel J.D. Microarray technology: beyond transcript profiling and genotype analysis // Nature Reviews Microbiology. 2006. Т. 7. № 3. С. 200–210.

57. Houseley J., Tollervey D. The Many Pathways of RNA Degradation // Cell. 2009. Т. 136. № 4. С. 763–776.

58. Huang J.D. и др. Binding sites for transcription factor NTF-1/Elf-1 contribute to the ventral repression of decapentaplegic. // Genes Dev. 1995. Т. 9. № 24. С. 3177–3189.

59. Huntzinger E., Izaurralde E. Gene silencing by microRNAs: contributions of translational repression and mRNA decay // Nature Reviews Genetics. 2011. Т. 12. № 2. С. 99–110.

60. Irimia M. и др. A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains // Cell. 2014. Т. 159. № 7. С. 1511–1523.

61. Irimia M., Blencowe B.J. Alternative splicing: decoding an expansive regulatory layer // Current Opinion in Cell Biology. 2012. Т. 24. № 3. С. 323–332.

62. Jang S. и др. Dynamics of embryonic stem cell differentiation inferred from single-cell transcriptomics show a series of transitions through discrete cell states // eLife. 2017. Т. 6. С. e20487.

63. Kang H.J. и др. Spatio-temporal transcriptome of the human brain // Nature. 2011. Т. 478. № 7370. С. 483–489.

64. Kar A. и др. RBM4 Interacts with an Intronic Element and Stimulates Tau Exon 10 Inclusion // J. Biol. Chem. 2006. Т. 281. № 34. С. 24479–24488.

65. Karaiskos N. и др. The Drosophila embryo at single-cell transcriptome resolution // Science. 2017. Т. 358. № 6360. С. 194–199.

66. Karam R. и др. Regulation of nonsense-mediated mRNA decay: Implications for physiology and disease // Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms. 2013.

67. Katagiri F., Glazebrook J. Overview of mRNA Expression Profiling Using DNA Microarrays // Current Protocols in Molecular Biology / под ред. F.M. Ausubel и др. Hoboken,

NJ, USA: John Wiley & Sons, Inc., 2009.

68. Katz Y. и др. Analysis and design of RNA sequencing experiments for identifying isoform regulation // Nature Methods. 2010. Т. 7. № 12. С. 1009–1015.

69. Khaitovich P. и др. Evolution of primate gene expression // Nature Reviews Genetics. 2006. Т. 7. № 9. С. 693–702.

70. Khanna A., Stamm S. Regulation of alternative splicing by short non-coding nuclear RNAs // RNA Biol. 2010. Т. 7. № 4. С. 480–485.

71. Khrameeva E.E., Gelfand M.S. Biases in read coverage demonstrated by interlaboratory and interplatform comparison of 117 mRNA and genome sequencing experiments // BMC Bioinformatics. 2012. Т. 13 Suppl 6. С. S4.

72. Kim D., Langmead B., Salzberg S.L. HISAT: a fast spliced aligner with low memory requirements // Nature Methods. 2015. Т. 12. № 4. С. 357–360.

73. Klein R.G. The human career: human biological and cultural origins. Chicago [etc.]: University of Chicago Press, 2009.

74. Kobor M.S., Greenblatt J. Regulation of transcription elongation by phosphorylation // Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression. 2002. Т. 1577. № 2. С. 261–275.

75. Kooistra S.M., Helin K. Molecular mechanisms and potential functions of histone demethylases // Nat Rev Mol Cell Biol. 2012. Т. 13. № 5. С. 297–311.

76. Kornblihtt A.R. и др. Alternative splicing: a pivotal step between eukaryotic transcription and translation // Nature Reviews Molecular Cell Biology. 2013. Т. 14. № 3. С. 153–165.

77. Kutzleb C. и др. Paralemmin, a prenyl-palmitoyl-anchored phosphoprotein abundant in neurons and implicated in plasma membrane dynamics and cell process formation // J. Cell Biol. 1998. Т. 143. № 3. С. 795–813.

78. Kwak H. и др. Precise Maps of RNA Polymerase Reveal How Promoters Direct Initiation and Pausing // Science. 2013. Т. 339. № 6122. С. 950–953.

79. Langmead B. и др. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome // Genome Biol. 2009. Т. 10. № 3. С. R25.

80. Laver T. и др. Assessing the performance of the Oxford Nanopore Technologies MinION // Biomolecular Detection and Quantification. 2015. Т. 3. С. 1–8.

81. Lee E., Bussemaker H.J. Identifying the genetic determinants of transcription factor activity // Mol. Syst. Biol. 2010. Т. 6. С. 412.

82. Lee T.I., Young R.A. Transcription of eukaryotic protein-coding genes // Annu. Rev. Genet. 2000. Т. 34. С. 77–137.

83. Lefebvre J.L. и др. Protocadherins mediate dendritic self-avoidance in the mammalian nervous system // Nature. 2012. Т. 488. № 7412. С. 517–521.

84. Li L.-C. и др. Small dsRNAs induce transcriptional activation in human cells // Proc. Natl. Acad. Sci. U.S.A. 2006. Т. 103. № 46. С. 17337–17342.

85. Liu F., Gong C.-X. Tau exon 10 alternative splicing and tauopathies // Molecular Neurodegeneration. 2008. Т. 3. № 1. С. 8.

86. Liu X. и др. Extension of cortical synaptic development distinguishes humans from chimpanzees and macaques // Genome Research. 2012. Т. 22. № 4. С. 611–622.

87. Liu Z. и др. Study of gene function based on spatial co-expression in a high-resolution mouse brain atlas // BMC Systems Biology. 2007. Т. 1. № 1. С. 19.

88. Loose M., Malla S., Stout M. Real-time selective sequencing using nanopore technology // Nat Meth. 2016. Т. advance online publication.

89. Luco R.F. и др. Epigenetics in Alternative Pre-mRNA Splicing // Cell. 2011. Т. 144. № 1. С. 16–26.

90. Matys V. и др. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes // Nucleic Acids Res. 2006. Т. 34. № Database issue. С. D108-110.

91. Mazin P. и др. Widespread splicing changes in human brain development and aging // Mol. Syst. Biol. 2013. Т. 9. С. 633.

92. Mazin P.V. и др. Conservation, evolution, and regulation of splicing during prefrontal cortex development in humans, chimpanzees, and macaques // RNA. 2018. Т. 24. № 4. С. 585–596.

93. McCullagh P., Nelder J.A. Generalized Linear Models, Second Edition. Chapman and Hall/CRC, 1989. Вып. 2. 532 С.

94. Memczak S. и др. Circular RNAs are a large class of animal RNAs with regulatory potency // Nature. 2013. Т. 495. № 7441. С. 333–338.

95. Mercer T.R., Dinger M.E., Mattick J.S. Long non-coding RNAs: insights into functions // Nature Reviews Genetics. 2009. Т. 10. № 3. С. 155–159.

96. Merkin J. и др. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues // Science. 2012. Т. 338. № 6114. С. 1593–1599.

97. Metzker M.L. Sequencing technologies — the next generation // Nature Reviews Genetics. 2009. Т. 11. № 1. С. 31–46.

98. Meunier J. и др. Birth and expression evolution of mammalian microRNA genes // Genome Res. 2013. Т. 23. № 1. С. 34–45.

99. Miura P. и др. Widespread and extensive lengthening of 3’ UTRs in the mammalian brain // Genome Research. 2013.

100. Morishita H., Yagi T. Protocadherin family: diversity, structure, and function // Current Opinion in Cell Biology. 2007. Т. 19. № 5. С. 584–592.

101. Mount S.M. A catalogue of splice junction sequences // Nucl. Acids Res. 1982. Т. 10. № 2. С. 459–472.

102. Nelder J.A., Wedderburn R.W.M. Generalized Linear Models // Journal of the Royal Statistical Society. Series A (General). 1972. Т. 135. № 3. С. 370.

103. Neph S. и др. An expansive human regulatory lexicon encoded in transcription factor footprints // Nature. 2012. Т. 489. № 7414. С. 83–90.

104. Niblock M., Gallo J.-M. Tau alternative splicing in familial and sporadic tauopathies // Biochem. Soc. Trans. 2012. Т. 40. № 4. С. 677–680.

105. Nicholson P. и др. Nonsense-mediated mRNA decay in human cells: mechanistic insights, functions beyond quality control and the double-life of NMD factors // Cellular and Molecular Life Sciences. 2009. Т. 67. № 5. С. 677–700.

106. Oberg A.L. и др. Technical and biological variance structure in mRNA-Seq data: life in the real world // BMC Genomics. 2012. Т. 13. № 1. С. 304.

107. O’Brien R.J., Wong P.C. Amyloid precursor protein processing and Alzheimer’s disease // Annu. Rev. Neurosci. 2011. Т. 34. С. 185–204.

108. Orphanides G., Reinberg D. A unified theory of gene expression // Cell. 2002. Т. 108. № 4. С. 439–451.

109. Pan Q. и др. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing // Nature Genetics. 2008. Т. 40. № 12. С. 1413–1415.

110. Pervouchine D.D. и др. Evidence for widespread association of mammalian splicing and conserved long-range RNA structures // RNA. 2012. Т. 18. № 1. С. 1–15.

111. Pickrell J.K. и др. Noisy splicing drives mRNA isoform diversity in human cells // PLoS Genet. 2010. Т. 6. № 12. С. e1001236.

112. Proudfoot N.J., Furger A., Dye M.J. Integrating mRNA processing with transcription // Cell. 2002. Т. 108. № 4. С. 501–512.

113. R Core Team. R: A Language and Environment for Statistical Computing. , 2013.

114. Ramsköld D. и др. An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data // PLoS Comput Biol. 2009. Т. 5. № 12. С. e1000598.

115. Rasche A. и др. ARH-seq: identification of differential splicing in RNA-seq data // Nucleic Acids Research. 2014.

116. Ray D. и др. A compendium of RNA-binding motifs for decoding gene regulation // Nature. 2013. Т. 499. № 7457. С. 172–177.

117. Robinson M.D., McCarthy D.J., Smyth G.K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data // Bioinformatics. 2010. Т. 26. № 1. С. 139–140.

118. Rosenbloom K.R. и др. The UCSC Genome Browser database: 2015 update // Nucl. Acids Res. 2015. Т. 43. № D1. С. D670–D681.

119. Ruijter A.J.M. de и др. Histone deacetylases (HDACs): characterization of the classical HDAC family // Biochem. J. 2003. Т. 370. № Pt 3. С. 737–749.

120. Salomonis N. и др. Alternative splicing regulates mouse embryonic stem cell pluripotency and differentiation // Proceedings of the National Academy of Sciences. 2010. Т. 107. № 23. С. 10514–10519.

121. Schena M. и др. Quantitative monitoring of gene expression patterns with a complementary DNA microarray // Science. 1995. Т. 270. № 5235. С. 467–470.

122. Shalek A.K. и др. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells // Nature. 2013. Т. advance online publication.

123. Sharova L.V. и др. Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells // DNA Res. 2009. Т. 16. № 1. С. 45–58.

124. Shen S. и др. MATS: a Bayesian framework for flexible detection of differential alternative splicing from RNA-Seq data // Nucl. Acids Res. 2012. Т. 40. № 8. С. e61–e61.

125. Singer G.A. и др. Genome-wide analysis of alternative promoters of human genes using a custom promoter tiling array // BMC Genomics. 2008. Т. 9. № 1. С. 349.

126. Smyth G.K. Limma: linear models for microarray data // Bioinformatics and computational biology solutions using R and Bioconductor. Springer, 2005. С. 397–420.

127. Somel M. и др. Transcriptional neoteny in the human brain // Proceedings of the National Academy of Sciences. 2009. Т. 106. № 14. С. 5743–5748.

128. Somel M. и др. MicroRNA-Driven Developmental Remodeling in the Brain Distinguishes Humans from Other Primates // PLoS Biology. 2011. Т. 9. № 12. С. e1001214.

129. Somel M., Liu X., Khaitovich P. Human brain evolution: transcripts, metabolites and their regulators // Nature Reviews Neuroscience. 2013. Т. 14. № 2. С. 112–127.

130. Soneson C., Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data // BMC Bioinformatics. 2013. Т. 14. № 1. С. 91.

131. Sorek R., Shamir R., Ast G. How prevalent is functional alternative splicing in the human genome? // Trends in Genetics. 2004. Т. 20. № 2. С. 68–71.

132. Sunkin S.M. и др. Allen Brain Atlas: an integrated spatio-temporal portal for exploring the central nervous system // Nucleic Acids Research. 2012. Т. 41. № D1. С. D996–D1008.

133. Svejstrup J.Q. The RNA polymerase II transcription cycle: cycling through chromatin // Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression. 2004. Т. 1677. № 1–3. С. 64–73.

134. Tacutu R. и др. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing // Nucleic Acids Res. 2013. Т. 41. № Database issue. С. D1027-1033.

135. Taniguchi Y. и др. Quantifying E. coli Proteome and Transcriptome with Single-Molecule Sensitivity in Single Cells // Science. 2010. Т. 329. № 5991. С. 533–538.

136. Tapial J. и др. An atlas of alternative splicing profiles and functional associations reveals new regulatory programs and genes that simultaneously express multiple major isoforms // Genome Research. 2017. С. gr.220962.117.

137. Tilgner H. и др. Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs // Genome Res. 2012. Т. 22. № 9. С. 1616–1625.

138. Tollervey J.R. и др. Analysis of alternative splicing associated with aging and neurodegeneration in the human brain // Genome Research. 2011. Т. 21. № 10. С. 1572–1582.

139. Trapnell C. и др. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation // Nature Biotechnology. 2010. Т. 28. № 5. С. 511–515.

140. Trapnell C. и др. Differential analysis of gene regulation at transcript resolution with RNA-seq // Nature Biotechnology. 2012. Т. 31. № 1. С. 46–53.

141. Trapnell C., Pachter L., Salzberg S.L. TopHat: discovering splice junctions with RNA-Seq // Bioinformatics. 2009. Т. 25. № 9. С. 1105–1111.

142. Ule J. и др. An RNA map predicting Nova-dependent splicing regulation // Nature. 2006. Т. 444. № 7119. С. 580–586.

143. Venables J.P. и др. Cancer-associated regulation of alternative splicing // Nature Structural & Molecular Biology. 2009. Т. 16. № 6. С. 670–676.

144. Vijver M.J. van de и др. A gene-expression signature as a predictor of survival in breast cancer // N. Engl. J. Med. 2002. Т. 347. № 25. С. 1999–2009.

145. Voineagu I. и др. Transcriptomic analysis of autistic brain reveals convergent molecular pathology // Nature. 2011. Т. 474. № 7351. С. 380–384.

146. Voisine P. и др. Differences in Gene Expression Profiles of Diabetic and Nondiabetic Patients Undergoing Cardiopulmonary Bypass and Cardioplegic Arrest // Circulation. 2004. Т. 110. № 11 suppl 1. С. II-280- II–286.

147. Wagner G.P., Kin K., Lynch V.J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples // Theory in Biosciences. 2012. Т. 131. № 4. С. 281–285.

148. Wang J. и др. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors // Genome Research. 2012. Т. 22. № 9. С. 1798–1812.

149. Weake V.M., Workman J.L. Inducible gene expression: diverse regulatory mechanisms // Nature Reviews Genetics. 2010. Т. 11. № 6. С. 426–437.

150. Wedderburn R.W. Quasi-likelihood functions, generalized linear models, and the Gauss—Newton method // Biometrika. 1974. Т. 61. № 3. С. 439–447.

151. Wheeler D.L. и др. Database resources of the National Center for Biotechnology Information // Nucleic Acids Res. 2008. Т. 36. № Database issue. С. D13-21.

152. Wilks S.S. The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses // The Annals of Mathematical Statistics. 1938. Т. 9. № 1. С. 60–62.

153. Witten J.T., Ule J. Understanding splicing regulation through RNA splicing maps // Trends in Genetics. 2011. Т. 27. № 3. С. 89–97.

154. Wu J. и др. SpliceTrap: a method to quantify alternative splicing under single cellular conditions // Bioinformatics. 2011. Т. 27. № 21. С. 3010–3016.

155. Wu L., Belasco J.G. Let Me Count the Ways: Mechanisms of Gene Regulation by

miRNAs and siRNAs // Molecular Cell. 2008. Т. 29. № 1. С. 1–7.

156. Xiong H.Y. и др. The human splicing code reveals new insights into the genetic determinants of disease // Science. 2014. С. 1254806.

157. Yang Z., Wang L. Regulation of microRNA expression and function by nuclear receptor signaling // Cell & Bioscience. 2011. Т. 1. № 1. С. 31.

158. Yap K. и др. Coordinated regulation of neuronal mRNA steady-state levels through developmentally controlled intron retention // Genes & Development. 2012. Т. 26. № 11. С. 1209–1223.

159. Yeo G. и др. Variation in alternative splicing across human tissues // Genome Biol. 2004. Т. 5. № 10. С. R74.

160. Young M.D. и др. Gene ontology analysis for RNA-seq: accounting for selection bias // Genome Biol. 2010. Т. 11. № 2. С. R14.

161. Zhang C. и др. Integrative Modeling Defines the Nova Splicing-Regulatory Network and Its Combinatorial Controls // Science. 2010. Т. 329. № 5990. С. 439–443.

162. Zinzen R.P. и др. Combinatorial binding predicts spatio-temporal cis-regulatory activity // Nature. 2009. Т. 462. № 7269. С. 65–70.

163. Zou D. и др. A critical role of RBM8a in proliferation and differentiation of embryonic neural progenitors // Neural Dev. 2015. Т. 10. С. 18.

8. Приложения

Приложение 1: Сравнение методов анализа альтернативного сплайсинга исходя из данных полученных МС

метода указаны объект с которым он работает, распределение и тест которые он использует, принимае

избыточную дисперсию (и.д.), доступность программы и ссылка на статью в которой он описан

Название	Объект анализа	Распределение	И.Д.	Статистический тест	Доступность программы
MISO	изоформа	Распределение Пуассона	нет	фактор Баеса (аналог теста на лог-правдоподобие)	доступна в виде единой программы
cuffdiff 2	изоформа	бета обратное биномиальное	да	расхождение Дженсона-Шаннона, оценка значимости методом Монте-Карло	доступна в виде единой программы
DEXseq	сегмент	обратное биномиальное	да	Обобщённая линейная модель, тест на лог-правдоподобие	доступна в виде единой программы
MATS	альтернатива	биномиальное	нет	Постериорная вероятность наблюдать отличие больше заданного, оценка значимости методом Монте-Карло	доступна в виде единой программы
Alexa-seq	экзон	Распределение Пуассона	нет	тест Фишера, фиксированные пороги на размер эффекта	доступна в виде набора подпрограмм
JuncBASE	альтернатива	Распределение Пуассона	нет	тест Фишера	доступна в виде набора подпрограмм
SpliceTrap	экзон	нет	нет	Нет. Используется просто порог на разницу частот включения. ЧВ оценивается при помощи Баесовского подхода.	доступна в виде набора подпрограмм
ARHseq	изоформа	Распределение Вейбулла	да	Эмпирический	В статье нет ссылки на программу
JunctionSeq	сегмент	обратное биномиальное	да	Обобщённая линейная модель, тест на лог-правдоподобие	доступна в виде единой программы

Приложение 2 Набор данных НД1.1. В таблице представлена информация о индивидуальных донорах и о то

объединены в образцы

Образец	Идентификатор донора в банке тканей	Возраст		Пол	ИПС[1]	УЦР[2]		Источник	Этническая группа	Причина смерти
		Лет	Дней			ПФК	КМ			
1	Maryl_447	0	2	М	3	8	7.2	NICHD[3]	Европеоидная	Осложнения при
	Maryl_779	0	5	М	5	8.8	7.8	NICHD	Афроамериканец	Врождённый пор
	Maryl_398	0	16	Ж	3	9.1	8.3	NICHD	Афроамериканец	Осложнения при
	Maryl_1157	0	20	Ж	14	7.1	7.5	NICHD	Европеоидная	Ассоциированная пневмония
	Maryl_759	0	35	М	7	7.9	6.9	NICHD	Европеоидная	Спонтанное лёго
2	Maryl_1325	0	182	Ж	1	8.4	8.1	NICHD	Афроамериканец	Синдром внезапн
	Maryl_131	0	198	Ж	24	7.8	7.9	NICHD	Европеоидная	Синдром внезапн
	Maryl_1281	0	206	М	6	8.4	7.2	NICHD	Афроамериканец	Синдром внезапн
	Maryl_121	0	224	М	20	6.7	6.8	NICHD	Европеоидная	Синдром внезапн
	Maryl_435	0	274	М	10	7.5	6.2	NICHD	Европеоидная	Менингит
3	4669	16	125	М	16	8.3	8.2	NICHD	Европеоидная	Повреждение гол
	4848	16	271	М	15	9.1	7.6	NICHD	Европеоидная	Несчастный случ
	1409	18	38	М	6	7.2	6.8	NICHD	Европеоидная	Несчастный случ
	1011	19	69	Ж	7	6.5	7	NICHD	Европеоидная	Несчастный случ
	933	20	255	М	12	8.7	8.9	NICHD	Европеоидная	Несчастный случ
4	1455	25	149	Ж	7	7.4	8.1	NICHD	Европеоидная	Множественные
	605	25	152	М	19	9.2	8.9	NICHD	Афроамериканец	Астма
	602	27	42	М	15	8.8	8.9	NICHD	Афроамериканец	Астма
	1026	28	131	М	6	8.1	7.5	NICHD	Европеоидная	Врожденный пор
	1365	28	239	М	17	8.2	7.4	NICHD	Европеоидная	Несчастный случ
5	S96/206	70	0	Ж	11	8	7.6	NBB[4]	Европеоидная	Рак груди
	4735	73	184	М	21	7.5	6	NICHD	Европеоидная	Хроническая обс
	S01/322	73	0	Ж	14	7.6	6.4	NBB	Европеоидная	Дыхательная нед

	S00/059	78	0	Ж	7	8.3	7.9	NBB	Европеоидная	Сердечная недос
	S04/057	80	0	М	7	8.6	7	NBB	Европеоидная	Фибриляция жел
6	S01/118	88	0	М	7	7.7	7.3	NBB	Европеоидная	Эвтаназия
	S96/297	90	0	Ж	6	7.8	7.6	NBB	Европеоидная	Остановка сердц
	S03/084	96	0	М	6	7.3	7.1	NBB	Европеоидная	Сердечная недос
	S03/119	97	0	Ж	5	8.4	8	NBB	Европеоидная	Сердечная астма
	S00/047	98	0	М	9	7.3	7.5	NBB	Европеоидная	Расслоение аорты

[1]ИПС: интервал после смерти (в часах)
[2]УЦР: уровень целостности РНК (измерено прибором Agilent Bioanalyzer)
[3]NICHD: Банк образцов мозга и тканей для изучения пороков развития в Университете Мэриленда, США (B
Developmental Disorders at the University of Maryland)
[4]NBB: Нидерландский банк образцов мозга, Амстердам, Нидерланды. (Netherlands Brain Bank)

Приложение 3: Набор данных НД1.2.

Идентификатор донора в банке тканей	Возраст		Пол	ИПС[1]	УЦР[2]	Источник	Этническая группа	Причина смерти
	Лет	Дней						
Maryl_779	0	5	М	5	8.8	NICHD[3]	Афроамериканец	Врождённый порок сердца
Maryl_1157	0	20	Ж	14	7.1	NICHD	Европеоидная	Ассоциированная с вдыханием мекo
Maryl_759	0	35	М	7	7.9	NICHD	Европеоидная	Спонтанное лёгочное кровоизлияние
Maryl_1055	0	94	М	12	7.7	NICHD	Европеоидная	Бронхопневмония
Maryl_1281	0	204	М	6	8.4	NICHD	Афроамериканец	Синдром внезапной детской смерти
1453	1	78	М	19	7.6	NICHD	Афроамериканец	Астма
1275	2	57	Ж	21	7.5	NICHD	Афроамериканец	Острый Миокардит
1908	13	360	М	13	8.3	NICHD	Европеоидная	Повешенье
605	25	152	М	19	9.2	NICHD	Афроамериканец	Астма
1496	53	112	М	17	8.3	NICHD	Европеоидная	Атеросклероз
S06/117	66	0	М	10	8.6	NBB[4]	Европеоидная	Разрыв аневризмы брюшной аорты
S01/118	88	0	М	7	7.7	NBB	Европеоидная	Эвтаназия
S00/047	98	0	М	9	7.3	NBB	Европеоидная	Расслоение аорты

[1]ИПС: интервал после смерти (в часах)
[2]УЦР: уровень целостности РНК (измерено прибором Agilent Bioanalyzer)
[3]NICHD: Банк образцов мозга и тканей для изучения пороков развития в Университете Мэриленда, США (B
Developmental Disorders at the University of Maryland)
[4]NBB: Нидерландский банк образцов мозга, Амстердам, Нидерланды. (Netherlands Brain Bank)]

Приложение 4. Функциональный анализ генов в различных возрастных паттернах изменений АС. Показаны хотя бы одной значимо обогащённой функцией.

Паттерн	Функции	Наблюдение[1]	Всего[2]	p-value
CL6	nervous system development	16	64	3.98E-003
	mating	3	0	1.06E-003
	synaptogenesis	4	4	5.32E-003
	neuromuscular process	4	4	5.32E-003
	suckling behavior	3	0	1.06E-003
	mating behavior	3	0	1.06E-003
	behavior	7	13	2.39E-003
	learning	3	1	3.93E-003
	actin filament-based movement	3	1	3.93E-003
	feeding behavior	3	1	3.93E-003
	anatomical structure development	23	107	2.51E-003
CL8	cell adhesion	5	43	1.66E-002
	actin cytoskeleton organization	5	38	1.04E-002
	cytoskeleton organization	6	55	1.08E-002
	interspecies interaction between organisms	4	15	2.55E-003
	cell death	6	45	4.35E-003
	apoptosis	6	34	1.17E-003
	anti-apoptosis	3	5	1.61E-003
	negative regulation of apoptosis	4	11	9.80E-004
	death	6	45	4.35E-003
	regulation of apoptosis	5	28	3.19E-003
	regulation of synaptic transmission	3	12	1.12E-002
	negative regulation of programmed cell death	4	11	9.80E-004
	regulation of programmed cell death	5	28	3.19E-003

	regulation of cellular component organization	5	47	2.31E-002
	regulation of neurogenesis	3	10	7.37E-003
	actin filament-based process	5	40	1.26E-002
	programmed cell death	6	35	1.34E-003
	negative regulation of cellular process	7	68	7.36E-003
	biological adhesion	5	43	1.66E-002

[1]Наблюдение: количество генов обладающих данной функцией и относящихся к данному паттерну
[2]Всего: общее количество генов с данной функцией среди 721 гена со значимыми возрастными изменениями АС.

Приложение 5: Набор данных НД2.1

Вид	Идентификатор донора в банке тканей	Возраст		Пол
		Лет	Дней	
человек	447	0	2	М
человек	779	0	5	М
человек	398	0	16	Ж
человек	1157	0	20	Ж
человек	759	0	35	М
человек	1055	0	96	М
человек	5183	0	107	М
человек	1303	0	212	М
человек	1453	1	78	М
человек	814	1	123	М
человек	1275	2	57	Ж
человек	1791	2	286	Ж
человек	6736	4	0	Ж
человек	1185	4	258	М
человек	4907	4	274	Ж
человек	4898	7	272	М
человек	1860	8	2	М
человек	1706	8	214	Ж
человек	5161	10	262	Ж
человек	М3228	11	294	Ж
человек	1908	13	360	М

человек	M3830	14	250	M
человек	5242	15	119	M
человек	7387	17	0	M
человек	5251	19	0	M
человек	4548	20	63	Ж
человек	1846	20	221	Ж
человек	1442	22	322	M
человек	7738	24	0	M
человек	602	27	42	M
человек	B-24	28	0	Ж
человек	1502	29	363	M
человек	7369	36	0	M
человек	7344	36	0	Ж
человек	7561	39	0	M
человек	1134	41	241	M
человек	6259	50	0	M
человек	1578	53	112	M
человек	6860	56	0	M
человек	4263	61	187	M
шимпанзе	11454	0	0	M
шимпанзе	13302	0	1	Ж
шимпанзе	58	0	7	Ж
шимпанзе	13308	0	7	M
шимпанзе	13311	0	8	M
шимпанзе	60	0	32	Ж

шимпанзе	59	0	34	М
шимпанзе	13309	0	39	М
шимпанзе	13304	0	45	Ж
шимпанзе	11452	1	160	Ж
шимпанзе	13312	6	123	Ж
шимпанзе	57	10.4	0	М
шимпанзе	1560	10.9	0	М
шимпанзе	Japie	12	35	М
шимпанзе	37	16.6	0	Ж
шимпанзе	38	16.8	0	Ж
шимпанзе	1480	17.8	0	Ж
шимпанзе	40	18	0	Ж
шимпанзе	CAO127	18.5	0	Ж
шимпанзе	31	20.2	0	Ж
шимпанзе	32	20.5	0	Ж
шимпанзе	30	21.3	0	М
шимпанзе	1465	21.4	0	Ж
шимпанзе	1437	21.5	0	Ж
шимпанзе	1275	22.7	0	М
шимпанзе	1365	23.9	0	М
шимпанзе	1358	24.1	0	М
шимпанзе	1028	25.5	0	М
шимпанзе	1351	26.4	0	М
шимпанзе	1266	27.1	0	М
шимпанзе	1342	28.9	0	М

шимпанзе	1167	29	0	М
шимпанзе	936	30.8	0	Ж
шимпанзе	1060	31.2	0	М
шимпанзе	53	32.4	0	М
шимпанзе	905	34.5	0	М
шимпанзе	Zurich	35	9	М
шимпанзе	5	39.9	0	Ж
шимпанзе	1045	42.6	0	М
макака	f0507910	0	-70	Ж
макака	f9604682	0	-56	М
макака	F0909	0	-42	М
макака	F0906	0	-30	М
макака	NB0903	0	0.5	М
макака	NB0901	0	1	М
макака	NB0905	0	7	М
макака	0705	0	16	М
макака	704	0	20	М
макака	0702	0	23	М
макака	0701	0	24	М
макака	070175	0	151	М
макака	70115	0	179	М
макака	70133	0	207	М
макака	6709	0	278	М
макака	06237	0	353	М
макака	61569	1	80	М

макака	051087	1	170	M
макака	051373	1	242	M
макака	051469	1	294	M
макака	051095	2	9	M
макака	05773	2	101	M
макака	4093	3	40	M
макака	050715	3	80	M
макака	04089	3	110	M
макака	3071	4	27	M
макака	B00051	6	165	M
макака	00135	7	15	M
макака	99057	8	16	M
макака	98145	9	37	M
макака	98073	9	104	M
макака	96007	10	328	M
макака	95001	11	346	M
макака	94051	13	17	M
макака	93041	14	21	M
макака	92095	14	349	M
макака	92107	15	3	M
макака	90049	17	22	M
макака	87015	20	91	M
макака	86023	21	8	M

[1]УЦР*:* уровень целостности РНК (измерено прибором Agilent Bioanalyzer)

Приложение 6: Набор данных НД2.2. НД2.2 так же включает образцы из НД1.2 (приложение 3).

вид	Идентификатор донора в банке тканей	Возраст		Пол	ИПС[1]	УЦР[2]	Причина смерти
		Лет	Дней				
шимпанзе	11454	0	0	М	НД	8.2	мертворождение
шимпанзе	13302	0	1	Ж	НД	8.5	мертворождение
шимпанзе	13308	0	7	М	НД	6.5	НД
шимпанзе	13311	0	8	М	НД	8.9	НД
шимпанзе	13309	0	39	М	НД	6.6	НД
шимпанзе	13304	0	45	Ж	НД	6.8	НД
шимпанзе	11452	1	160	Ж	НД	6.8	НД
шимпанзе	13312	6	123	Ж	НД	6.5	НД
шимпанзе	KOKO	8	НД	Ж	НД	8	НД
шимпанзе	Koos	11	346	М	НД	7.2	несчастный случай
шимпанзе	Herman	12	НД	М	НД	5.8	НД
шимпанзе	Japie	12	35	М	НД	8	шок от анастезии
шимпанзе	Mayumi	27	НД	Ж	НД	8.5	НД
шимпанзе	Zurich	35	9	М	НД	7.6	НД
шимпанзе	Reba	44	71	Ж	НД	8.9	анемия
макака	NB0902	0	1	М	0	8.9	лабораторное умертвление
макака	NB0901	0	1	М	0	8.3	лабораторное умертвление
макака	NB0905	0	7	М	0	8.7	лабораторное умертвление
макака	0705	0	16	М	0	9.1	лабораторное умертвление
макака	0703	0	22	М	0	9.9	лабораторное умертвление
макака	070141	0	153	М	0	9.3	лабораторное умертвление

макака	070133	0	207	М	0	9.1	лабораторное умертвление
макака	06711	0	310	М	0	9.8	лабораторное умертвление
макака	051095	2	9	М	0	9.5	лабораторное умертвление
макака	03071	4	27	М	0	9.7	лабораторное умертвление
макака	98073	9	104	М	0	9	лабораторное умертвление
макака	92107	15	3	М	0	9.5	лабораторное умертвление
макака	85091	22	74	М	0	9.2	лабораторное умертвление
макака	198100	26	28	М	0	8.6	лабораторное умертвление
макака	QDL II	28	НД	Ж	0	7.8	лабораторное умертвление

[1]ИПС: интервал после смерти (в часах)
[2]УЦР: уровень целостности РНК (измерено прибором Agilent Bioanalyzer)

Приложение 7: Набор данных НД2.3. НД2.3 так же включает образцы из НД1.1 (приложение 2)

образец	вид	Идентификатор донора в банке тканей	Возраст		Пол	ИПС[1]	
			Лет	Дней			
1	шимпанзе	11454	0	0	М	НД	
	шимпанзе	13302	0	1	Ж	НД	
	шимпанзе	13311	0	8	М	НД	
	шимпанзе	13309	0	39	М	НД	
	шимпанзе	13304	0	45	Ж	НД	
2	шимпанзе	13312	8	НД	Ж	НД	
	шимпанзе	Japie	12	НД	М	НД	
	шимпанзе	Herman	12	НД	М	НД	
	шимпанзе	Zurich	40	НД	М	НД	
	шимпанзе	Reba	44	НД	Ж	НД	
3	макака	NB0903	0	0.5	М	0	
	макака	NB0902	0	0.75	М	0	
	макака	NB0901	0	1	М	0	

	макака	NB0904	0	2	Ж	0	
	макака	NB0905	0	7	М	0	
4	макака	99057	8	16	М	0	
	макака	98073	9	104	М	0	
	макака	96007	10	328	М	0	
	макака	95001	11	346	М	0	
	макака	93041	14	21	М	0	

[1]ИПС: интервал после смерти (в часах)
[2]УЦР*:* уровень целостности РНК (измерено прибором Agilent Bioanalyzer)

Приложение 8: Функциональный анализ генов с возраст-зависимыми белок-кодирующими сегментами

GO категория	Корректированное p-значение	Онтология	Функция
GO:0007156	3.00E-07	BP	homophilic cell adhesion
GO:0007155	3.00E-07	BP	cell adhesion
GO:0022610	3.00E-07	BP	biological adhesion
GO:0008092	3.10E-06	MF	cytoskeletal protein binding
GO:0098609	1.70E-05	BP	cell-cell adhesion
GO:0005509	2.40E-05	MF	calcium ion binding
GO:0071944	1.40E-04	CC	cell periphery
GO:0005886	1.50E-03	CC	plasma membrane
GO:0030182	2.20E-02	BP	neuron differentiation
GO:0003779	1.10E-02	MF	actin binding
GO:0048666	3.00E-02	BP	neuron development
GO:0022008	4.80E-02	BP	neurogenesis
GO:0048699	5.20E-02	BP	generation of neurons
GO:0030276	4.20E-02	MF	clathrin binding
GO:0044459	3.80E-02	CC	plasma membrane part
GO:0032403	4.90E-02	MF	protein complex binding
GO:0005856	7.40E-02	CC	cytoskeleton
GO:0043194	7.60E-02	CC	axon initial segment
GO:0005905	7.80E-02	CC	coated pit
GO:0044304	7.80E-02	CC	main axon

GO:0042995	7.80E-02	CC	cell projection
GO:0097458	7.80E-02	CC	neuron part
GO:0045211	7.80E-02	CC	postsynaptic membrane
GO:0097060	7.80E-02	CC	synaptic membrane
GO:0030424	8.80E-02	CC	axon
GO:0045202	8.80E-02	CC	synapse
GO:0005911	8.80E-02	CC	cell-cell junction
GO:0014704	8.80E-02	CC	intercalated disc
GO:0030666	8.80E-02	CC	endocytic vesicle membrane
GO:0030118	9.20E-02	CC	clathrin coat
GO:0044430	9.30E-02	CC	cytoskeletal part
GO:0042641	9.30E-02	CC	actomyosin
GO:0043005	9.90E-02	CC	neuron projection
GO:0030132	1.00E-01	CC	clathrin coat of coated pit
GO:0031252	1.00E-01	CC	cell leading edge
GO:0015629	1.10E-01	CC	actin cytoskeleton
GO:0030017	1.10E-01	CC	sarcomere
GO:0030125	1.20E-01	CC	clathrin vesicle coat
GO:0001725	1.20E-01	CC	stress fiber
GO:0032432	1.20E-01	CC	actin filament bundle
GO:0030016	1.20E-01	CC	myofibril
GO:0030669	1.30E-01	CC	clathrin-coated endocytic vesicle membrane
GO:0031594	1.30E-01	CC	neuromuscular junction
GO:0045334	1.50E-01	CC	clathrin-coated endocytic vesicle
GO:0044291	1.50E-01	CC	cell-cell contact zone
GO:0043292	1.50E-01	CC	contractile fiber
GO:0005862	1.50E-01	CC	muscle thin filament tropomyosin
GO:0042383	1.50E-01	CC	sarcolemma
GO:0033268	1.50E-01	CC	node of Ranvier
GO:0044449	2.00E-01	CC	contractile fiber part
GO:0044433	2.00E-01	CC	cytoplasmic vesicle part

[1]Наблюдение: количество генов обладающих данной функцией и относящихся к данному паттерну

[2]Всего: общее количество генов с данной функцией среди 721 гена со значимыми возрастными изменениями АС.

Приложение 9: Мотивы связывания транскрипционных факторов значимо перепредставленные около возр
сегментов. Во второй колонке указаны последовательности около (1 и 2 — в интроне перед экзоном, 5 и 6 в
или внутри (3 и4) экзона.

Мотив	Участки последовательности в которых обогащён мотив	Факторы сплайсинга связывающие данный мотив	Факторы сплайсинга мотив и значимо мен возрастом хотя бы в
M043	123456	PCBP4,PCBP3,PCBP2	PCBP4,PCBP3,PCBP
M044	123456	PPRC1	PPRC1
M054	12456	RBM8A	RBM8A
M320	1256	MBNL3,MBNL2,MBNL1	MBNL2,MBNL1
M344	1235	RBMX,RBMXL1	RBMX,RBMXL1
M050	245	RBM4B,RBM4,RBM14	RBM4B,RBM4,RBM1
M177	125	PCBP4,PCBP3	PCBP4,PCBP3
M109	24	RBM4B,RBM4,RBM14	RBM4B,RBM4,RBM1
M207	15	PCBP4,PCBP3	PCBP4,PCBP3
M026	15	hnRNPK	hnRNPK
M298	56	A2BP1,RBFOX2,RBFOX3	A2BP1,RBFOX2,RBF
M297	56	RBFOX2,RBFOX3	RBFOX2,RBFOX3
M037	1	MBNL3,MBNL2,MBNL1	MBNL2,MBNL1
M081	5	CSDA,YB-1	CSDA,YB-1
M111	5	CSDA,YB-1	CSDA,YB-1
M188	1	PCBP4,PCBP3	PCBP4,PCBP3
M211	1	PCBP4,PCBP3	PCBP4,PCBP3
M262	1	QKI	
M227	1	PTBP1,PTBP2,ROD1	PTBP2,ROD1
M228	1	PTBP1,PTBP2,ROD1	PTBP2,ROD1
M077	1	U2AF2	U2AF2
M273	5	SRSF1	
M354	5	YTHDC1	