На правах рукописи

Мазин Павел Владимирович

Анализ возрастных изменений альтернативного сплайсинга в коре головного мозга высших приматов

03.01.09 — математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени

кандидата биологических наук

Москва - 2019

Работа выполнена в Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук (ИППИ РАН).

Научный руководитель:	**Хайтович Филипп Ефимович** кандидат биологических наук Сколковский институт науки и технологий
Официальные оппоненты:	**Самсонова Мария Георгиевна** доктор биологических наук Федеральное государственное автономное образовательное учреждение высшего образования«Санкт-Петербургский политехнический университет Петра Великого»
	Ушаков Вадим Леонидович кандидат биологических наук Национальный исследовательский центр «Курчатовский институт»
Ведущая организация:	Федеральное государственноебюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук

Защита диссертации состоится ___ _____ 2019 года в ____ на заседании диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 127051, г. Москва, Большой Каретный переулок, д. 19, стр. 1.

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

Автореферат разослан «__» _____ 2019 г.

Ученый секретарь диссертационного совета

доктор биологических наук, профессор Рожкова Г. И.

Общая характеристика работы

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.

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.

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.

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.

Структура и объем диссертации. Диссертация состоит из введения, обзора литературы, трёх глав, выводов, библиографии и приложений. Общий объем диссертации 105 страниц, из них 90 страниц текста, включая 23 рисунка. Библиография включает 163 наименования на 12 страницах.

Содержание работы

Разработка метода анализа альтернативного сплайсинга. 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. 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, 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.

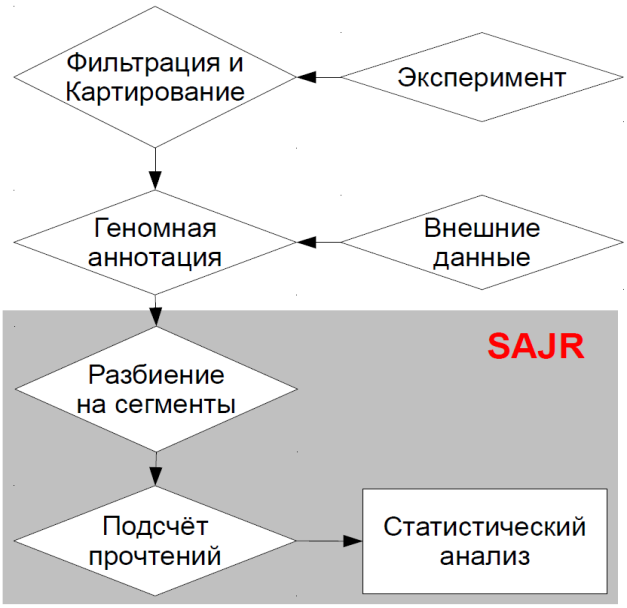

Рисунок 1.Схема анализа данных МСНП и метода SAJR. Полученные в результате эксперимента прочтения фильтруются и картируются на геном. Далее либо используется известная геномная аннотация, либо аннотация создаётся заново исходя из прочтений. Далее SAJR разбивает гены на сегменты, подсчитывает прочтения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

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.

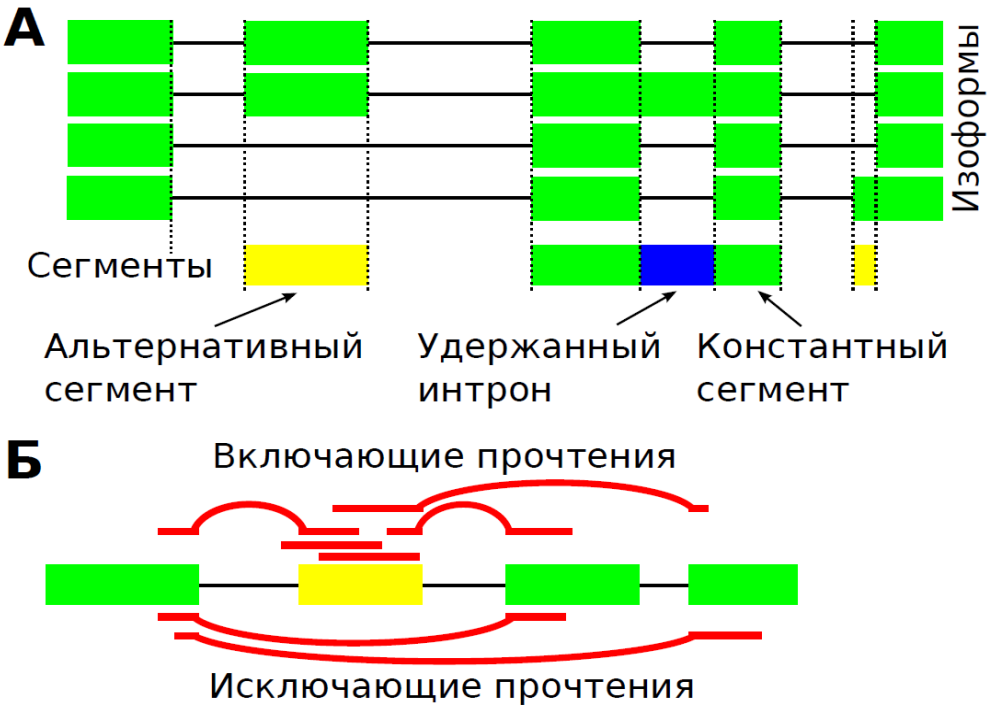

Рисунок 2.Разделение гена на сегменты (А) и подсчёт включающих и исключающих прочтений (Б). Экзоны показаны зелёными прямоугольниками, интроны показаны горизонтальными линиями, вертикальные пунктирные линии обозначают сайты сплайсинга. Прочтения обозначены красными линиями; горизонтальные участки изображают 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 реализован на java, и 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).

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) и тест на логарифм правдоподобия 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. При наличии достаточного числа образцов SAJR позволяет использовать квази-биномиальное распределение и тест на квази-правдоподобие для учёта биологической вариабельности.

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

$$p = logit^{-1}\left(\sum a_i \times x_i\right)$$

$$logit(x) = \log\left(\frac{x}{1-x}\right)$$

$$logit^{-1}(x) = \frac{e^x}{1+e^x}$$

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. 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:

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

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

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. 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.

Возрастные изменения сплайсинга в мозге человека. В данном исследовании были использованы два набора данных. 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. Второй набор данных (НД1.2) содержит 13 индивидуальных образцов префронтальной коры. В обоих наборах данных возрасты доноров покрывают всё протяжение жизни от рождения до старости. 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.

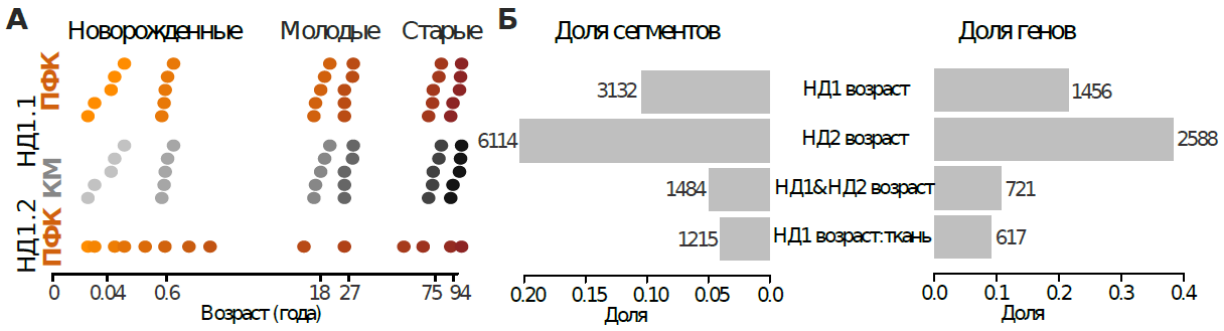

Рисунок 3. Возрастные изменения АС в мозгу человека. (А) Возрасты индивидуальных доноров; ПФК и КМ показаны серым и оранжевым соответственно, по горизонтальной оси отложен возраст доноров, в НД1.1 доноры, использованные для одного образца показаны вместе. (Б) Количества (и доля от общего числа тестированных) 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.

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 7), 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).

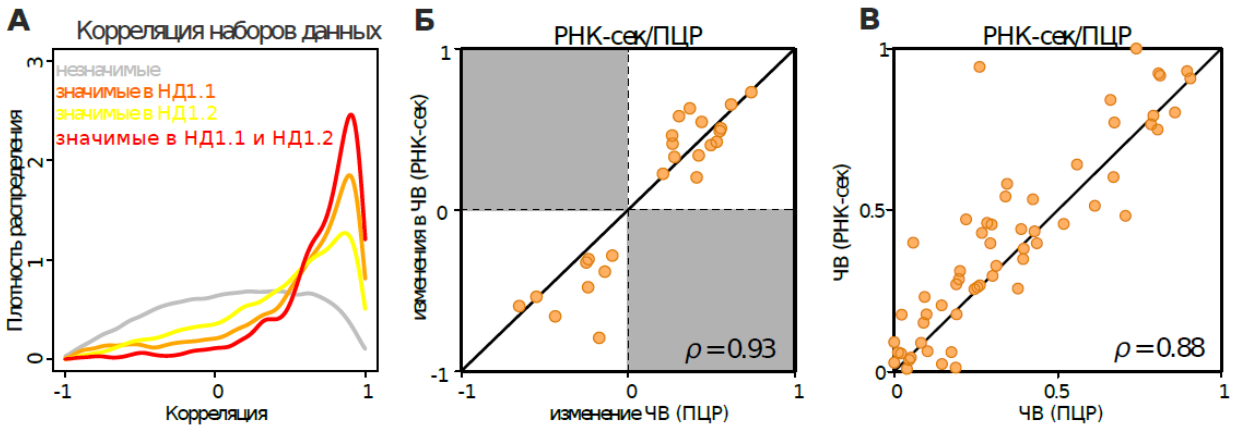

Рисунок 4. Подтверждение результатов РНК-Сек. (А) Распределение коэффициентов корреляции возрастных изменений АС между наборами данных; сегменты незначимые ни в одном наборе данных, значимые только в НД1.1, только в НД2.1 и в обоих наборах данных показаны серым, оранжевым, жёлтым и красным соответственно.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

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

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. 5). Для обозначения паттернов далее будут использоваться сокращения С1–С8. Как видно на рисунке 5, 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. 5), 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).

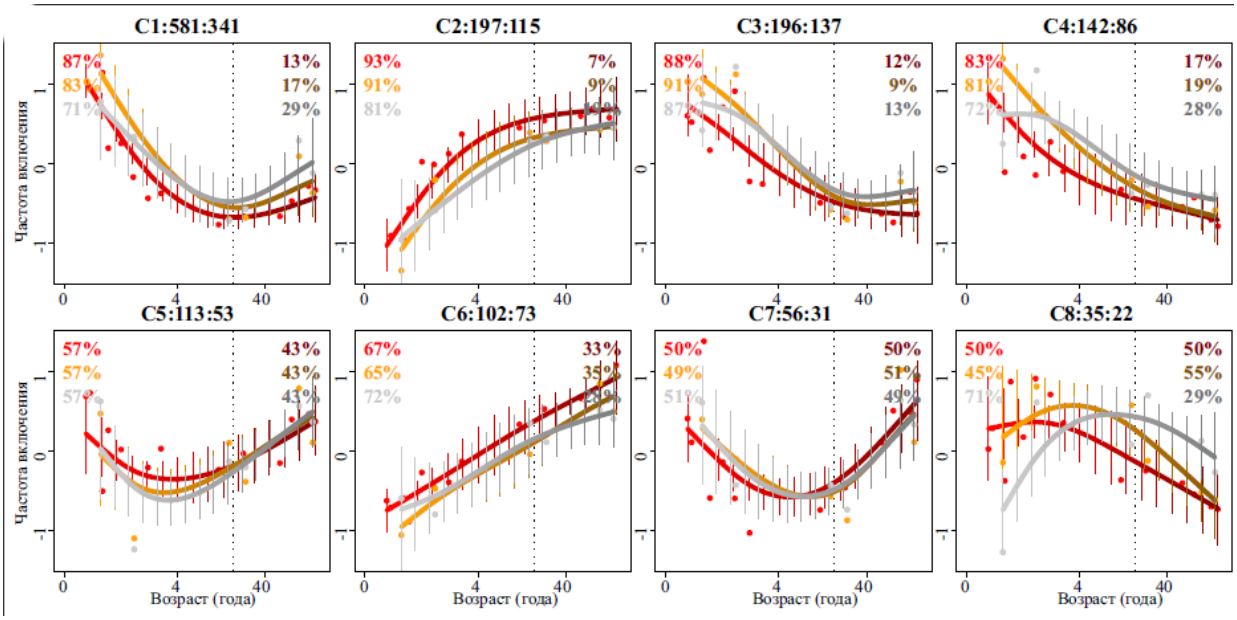

Рисунок 5. Паттерны возрастных изменений АС в ПФК и КМ человека. Все сегменты значимо меняющие АС с возрастом в НД1.1 и НД1.2 разбиты на 8 паттернов. Паттерны упорядочены по количеству сегментов которые ему следуют. По вертикальной оси отложена усреднённая по кластеру, нормализованная ЧВ, по горизонтальной 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.

Сегменты, следующие различным возрастным паттернам, различаются по биологическим свойствам. Функциональный анализ при помощи пакета GOstat 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). 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 C1, в то время как доля белок-кодирующих сегментов максимальна в паттернах в которых ЧВ растёт с возрастом, например в C2 (рис. 6). Удержанные интроны могут вызывать деградацию мРНК либо через ПСК-зависимый механизм либо при помощи ядерной экзосомы, поэтому сегменты из C1 (а так же C3 и C4) могут участвовать в регуляции клеточных концентраций мРНК.

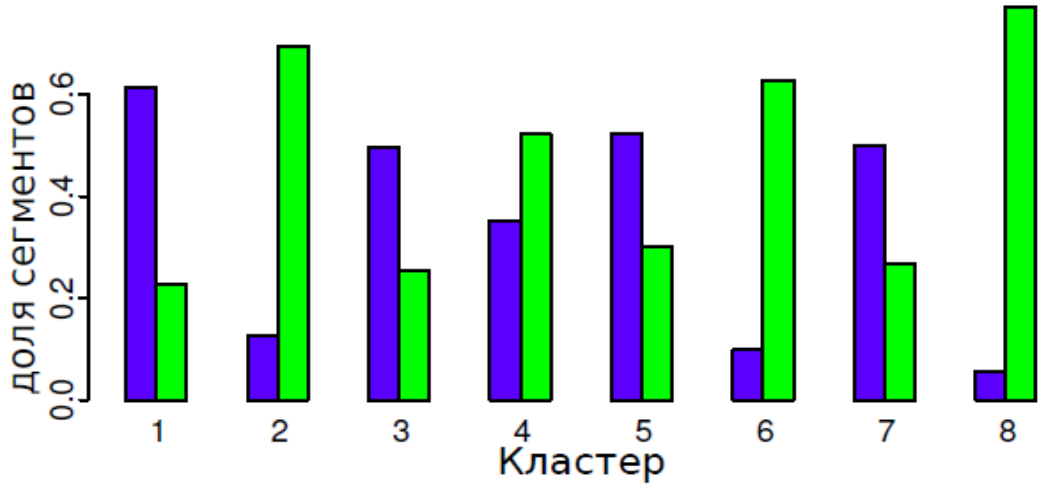

Рисунок 6. Доля удержанных интронов (синий) и белок-кодирующих сегментов (зеленый) в каждом паттерне.

Хотя возрастное изменение АС большинства сегментов схоже в ПФК и КМ (рис. 7А), 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; 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. 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. 7Б–В).

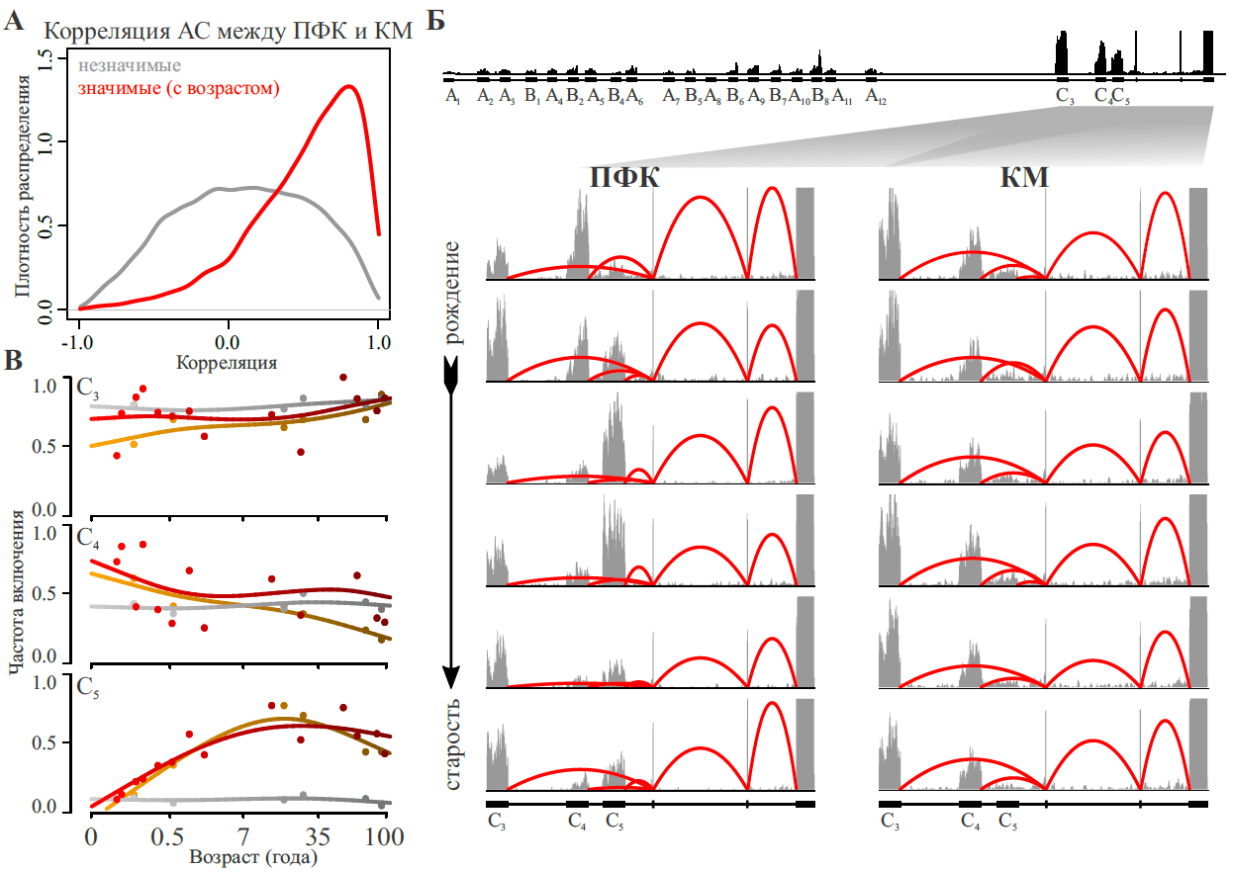

Рисунок 7. Возрастные изменения АС в ПФК и КМ человека. (А) Распределение коэффициента корреляции для ЧВ между ПФК и КМ для значимых и незначимых сегментов. (Б)–(В) Изменение частоты использования трёх основных альтернативных первых экзонов в протокадгерние гамма. Зависимость частоты включения от возраста 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).

Сравнительный анализ альтернативного сплайсинга в мозге высших приматов. Для анализа возрастных изменений АС в головном мозге человека (Homo sapiens), шимпанзе (Pan troglodytes) и макаки (Macaca mulatta) были использованы три набора данных. Первый набор данных (НД2.1) содержит 40, 39 и 40 индивидуальных образцов префронтальной коры человека, шимпанзе и макаки соответственно. 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. 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. 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 данных доминируют межвидовые отличия (рис. 8А). Если провести аналогичный анализ для отклонений ЧВ от среднего значения в данном виде, образцы упорядочиваются по возрасту (рис. 8Б), что указывает на эволюционную консервативность возрастных изменений сплайсинга в мозге приматов и самосогласованность наших данных.

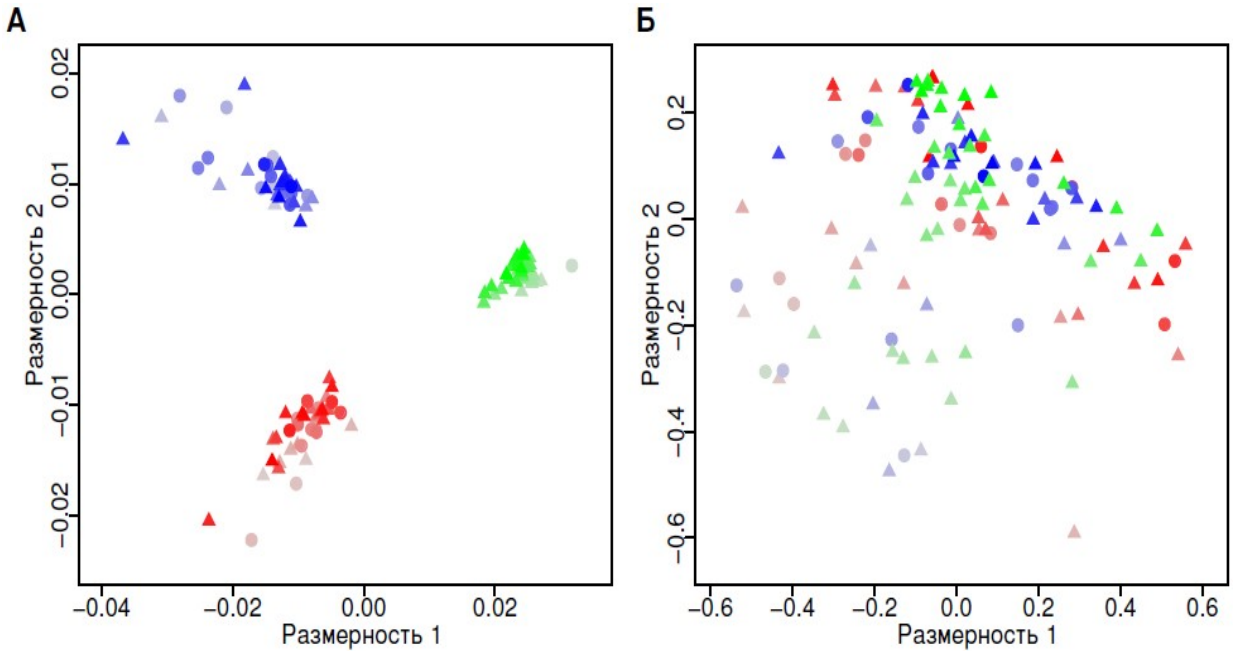

Рисунок 8. Многомерное шкалирование (к размерности два) образцов мозга на основании ЧВ для всех сегментов прошедших фильтрацию. Расстояние между образцами определялось как единица минус коэффициент корреляции Пирсона между векторами содержащими ЧВ (панель (А) или отклонение ЧВ от среднего (в данном виде 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.

Количества сегментов со значимыми различиями сплайсинга между видами, видоспецифичных сегментов и сегментов значимо меняющих сплайсинг с возрастом показаны на рисунке 9А. 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.

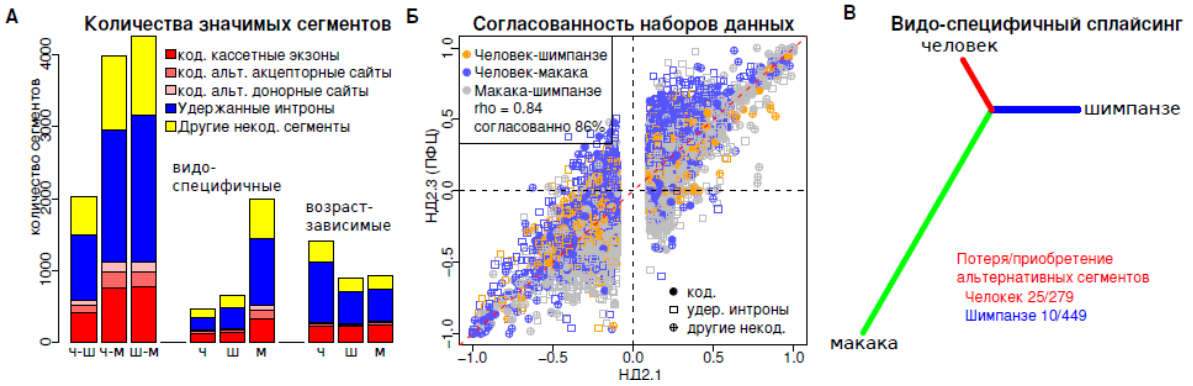

Рисунок 9: Видоспецифичные и возрастные изменения сплайсинга. (А) Высота столбцов показывает количество сегментов. значимо отличающихся между каждой из трёх пар видов (слева), видоспецифичных сегментов (посередине) и возраст-зависимых сегментов (справа). Тип сегментов показан цветом, виды обозначены 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.

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. 9Б). 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. 9В).

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 схожесть сайтов сплайсинга с консенсусной последовательностью

(сила сайта) для каждого сегмента была вычислена в каждом виде. 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. 10). 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. 11).

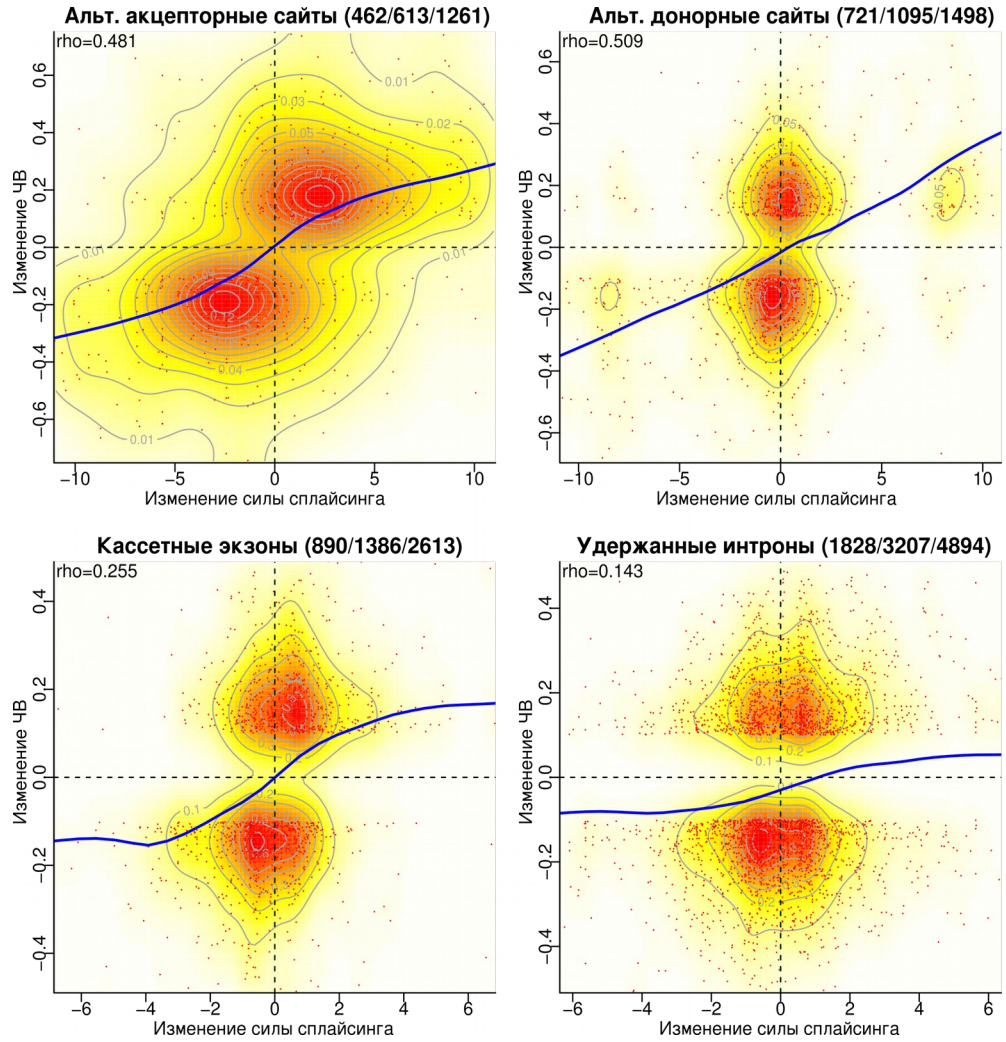

Рисунок 10: Зависимость межвидовых отличий ЧВ (вертикальная ось) от изменения силы сплайсинга сегмента (горизонтальная ось). Каждый тип сегментов показан на отдельной панели. Для каждой пары видов сегменты со значимыми отличиями ЧВ показаны точкой, если сила сплайсинга меняется более чем на один бит. 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 lowess). Коэффициент корреляции Пирсона указан в верхнем левом углу каждой панели, число сравнений с совпадающими направлениями изменений ЧВ и силы сплайсинга, с изменением силы сплайсинга более чем на единицу и общее число значимых сравнений указаны в названии каждой панели.

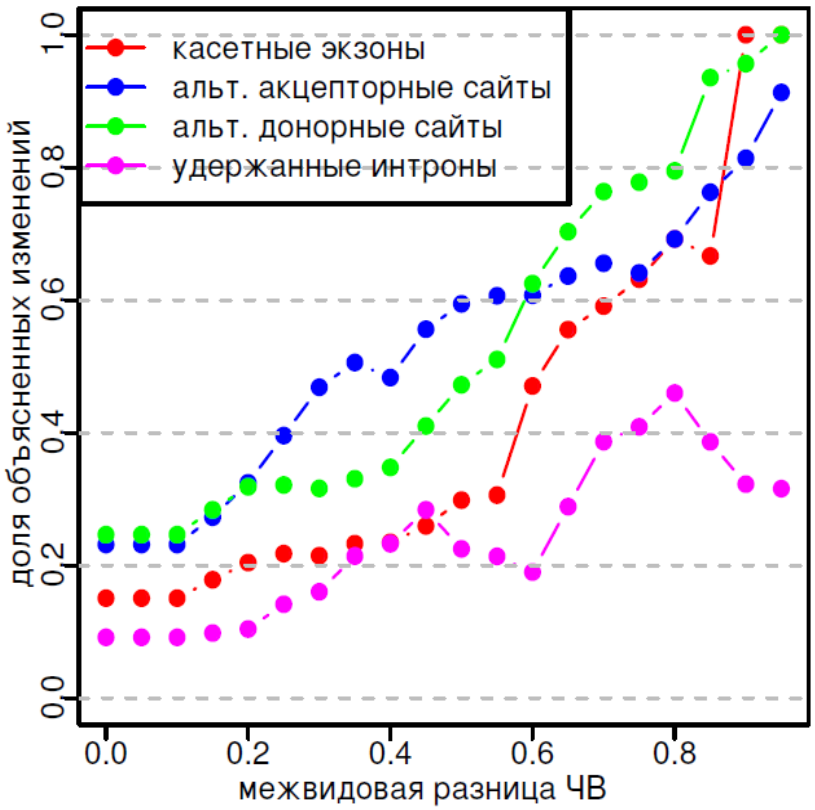

Рисунок 11. Зависимость доли межвидовых отличий объяснённых мутациями в сайтах сплайсинга, от амплитуды изменений ЧВ. Различные типы сегментов показаны разными цветами.

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. 19)

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. 12A). Это объясняется человеко-специфичным одно-нуклеотидным полиморфизмом (ОНП) в основном донорном сайте. В данном случае изменение AC еще не зафиксировались в популяции, однако частота альтернативного аллеля достигла 18%.

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. 12Б). В этом сегменте находится ОНП rs12898397 представленный в 39% популяции. Альтернативный аллель в данном ОНП создаёт динуклеотид GT внутреннего альтернативного сайта и, таким образом, скорее всего отвечает за АС данного сегмента.

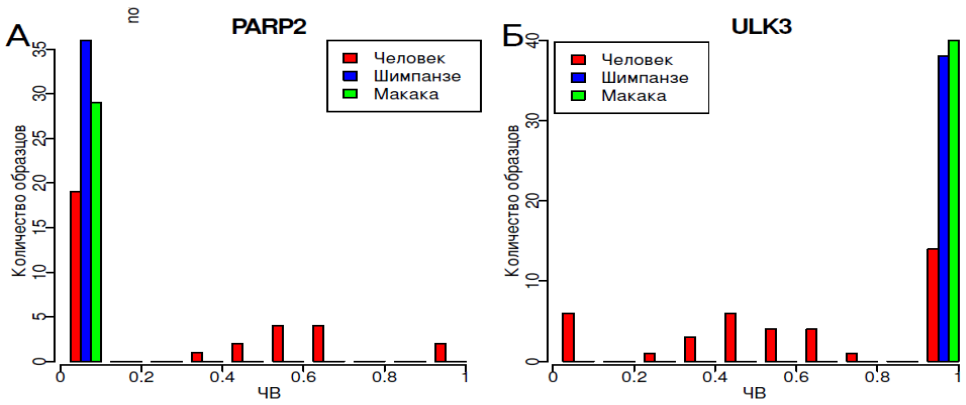

Рисунок 12. Альтернативный донорный сайт в гене PARP2 (А) и ULK3 (Б). Распределение образцов из НД2.1 по ЧВ соответствующего сегмента показано для человека, шимпанзе и макаки красным, синим и зелёным соответственно.

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. 13А). 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. 14) и происходят одинаково во всех видах (рис. 13Б). 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. 13В). Интересно, что у человека наблюдается в два раза больше возраст-зависимых удержанных интронов, чем у шимпанзе или макаки (рис. 14).

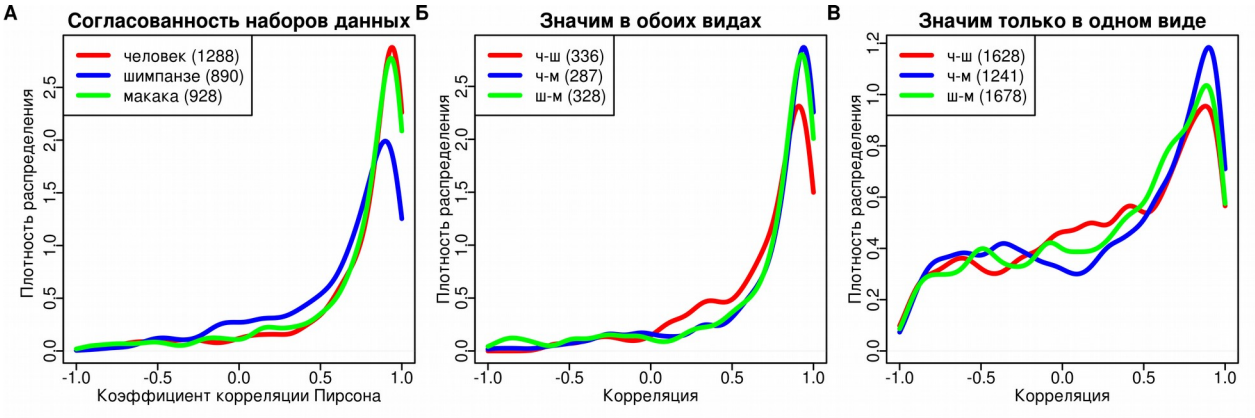

Рисунок 13. (А) Распределение коэффициента корреляции Пирсона между НД2.1 и НД2.2 для всех значимых сегментов. (Б) и (В) Распределение коэффициента корреляции Пирсона для пар видов в НД2.1. Использовались сегменты значимые в обоих видах (Б) или только в одном виде (В). Ч, ш и м обозначают человека, шимпанзе и макаку соответственно. Коэффициенты корреляции рассчитывались на основании корректированных возрастов.

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.

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, $p < 10^{-10}$).

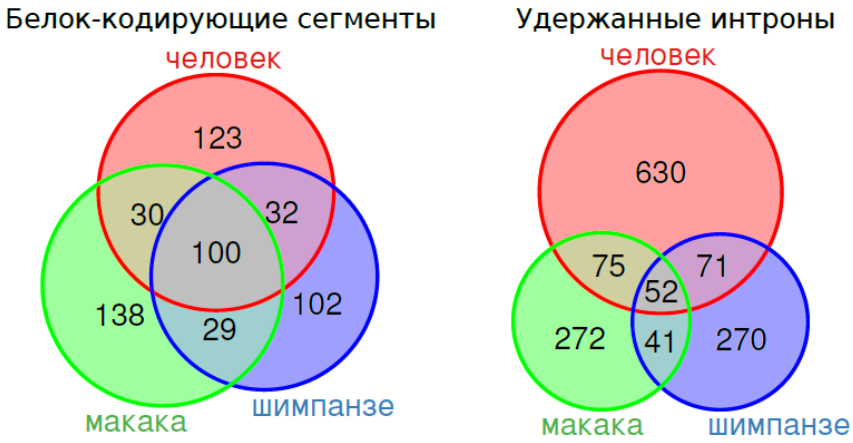

Рисунок 14: Диаграмма Венна для количества возраст-зависимых белок-кодирующих сегментов и удержанных интронов.

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), 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, так, чтобы максимизировать схожесть изменения между видами. Распределения этих коэффициентов для каждой пары видов имеют единственный максимум (рис. 15). 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. 13Б 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.

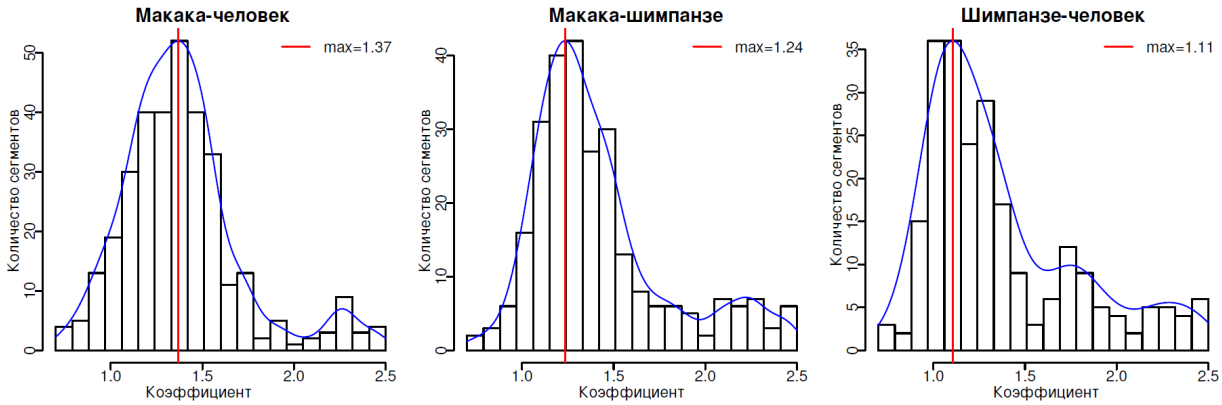

Рисунок 15. Распределение сегментов по оптимальным значениям коэффициента перевода возрастов между тремя парами видов (в шкале возраст$^{0.25}$). Мода распределения показана красной линией.

Чтобы лучше охарактеризовать разнообразие возраст-зависимых изменений АС возраст-зависимые сегменты разбиты на шесть кластеров (рис. 16, 17).

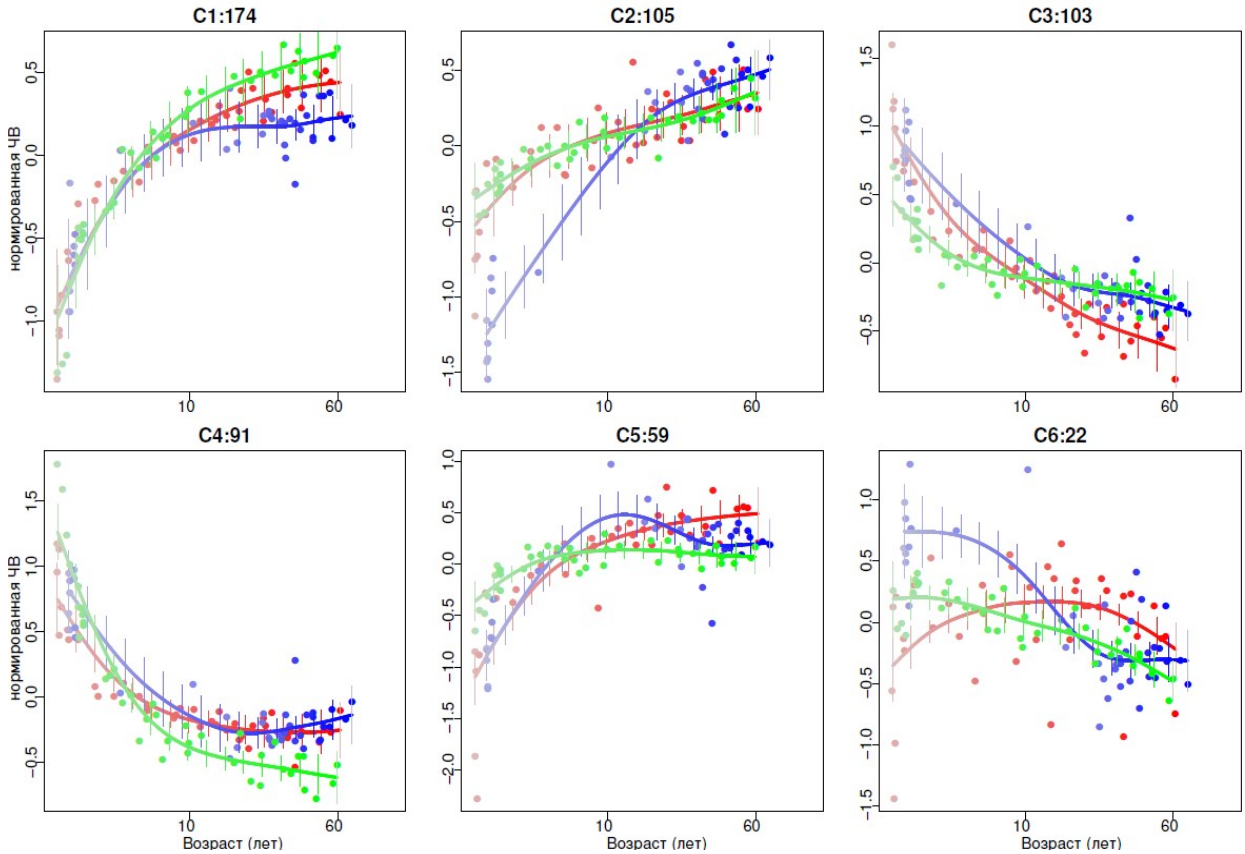

Рисунок 16. Разбиение возраст-зависимых белок-кодирующих сегментов на шесть кластеров. Каждый кластер показан на отдельной панели, кластеры упорядочены по числу сегментов, относящихся к ним (указано в названии панели). Каждая точка обозначает среднюю нормализованную ЧВ (вертикальная ось) в зависимости 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.

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. 17).

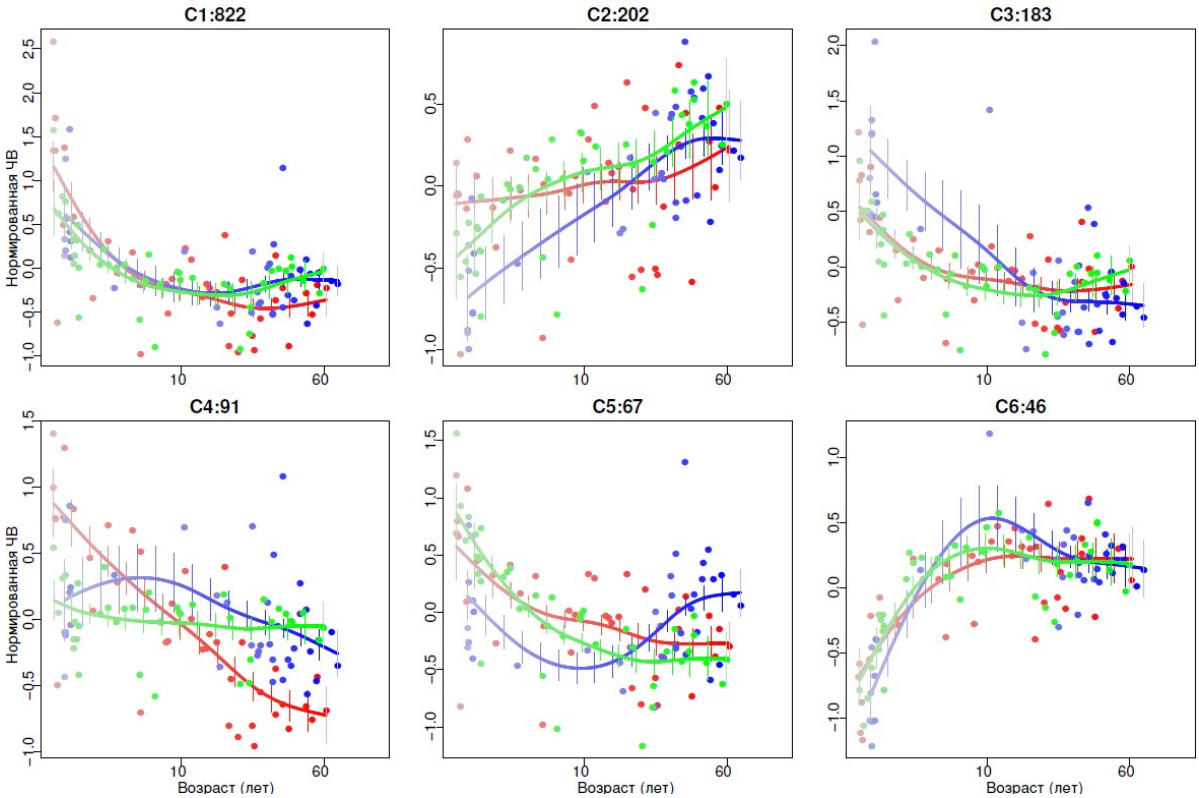

Рисунок 17. Разбиение возраст-зависимых удержанных интронов на шесть кластеров. Описание см. в подписи к рис. 16.

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. 18, 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. Ранее 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.

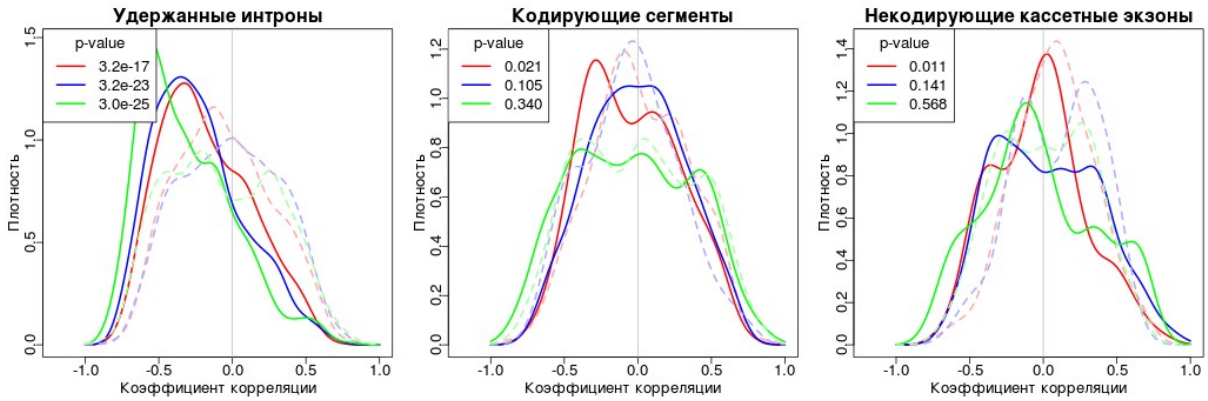

Рисунок 18. Распределение коэффициента корреляции Пирсона между ЧВ возраст-зависимых сегментов и уровнями экспрессии соответствующих (показано сплошной линией) или случайно выбранных (показано пунктиром) генов в трёх видах: человек (красный), шимпанзе (синий) и макака (зеленый). Значимость отличий распределений между правильными и случайными парами сегмент-ген показаны в легенде (*p* согласно тесту Вилкоксона).

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 19). 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, содержащая информацию о 219 мотивах связываемых 392 РНК-связывающими белками человека (экспрессия 315 из них была детектированна в данной работе). Были обнаружены 23 мотива, аффинность которых была значимо увеличена около 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; *RBM4*, регулирующий сплайсинг мРНК, кодирующей белок *tau*, вовлечённый в болезнь Альцгеймера; *YB-1*, подавление которго материнскими антителами в ходе эмбрионального развития связанно с аутизмом.

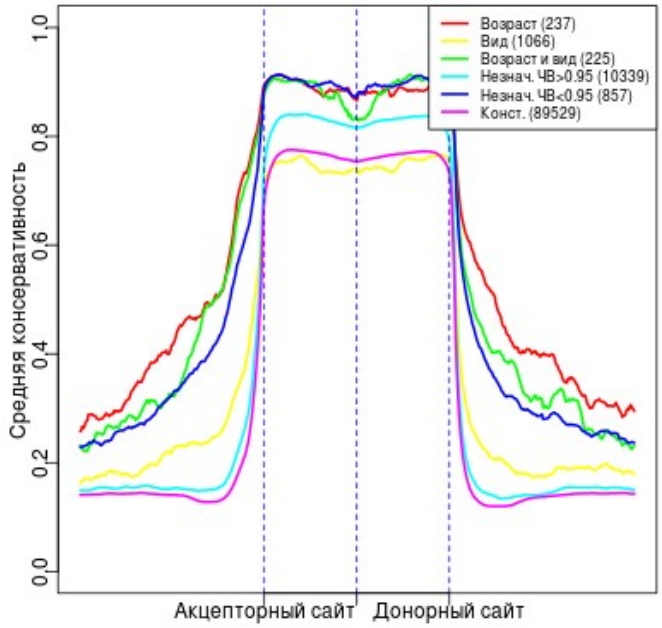

Рисунок 19: Средняя консервативность (phastcons для приматов) около кассетных экзонов (слева) и интронов (справа) для различных категорий сегментов: возраст-зависимых; значимо различающихся между видами; возраст-зависимых и значимо различающихся между видами; незначимых (отдельно показаны сегменты с ЧВ больше или меньше чем 0.95); и константных.

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. Чтобы избежать переобучения была использована L1 -регуляризация с весовым значением 0.01, соответствующем оптимальному значению коэффициента корреляции Пирсона в кросс-валидации.

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. 20). 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.

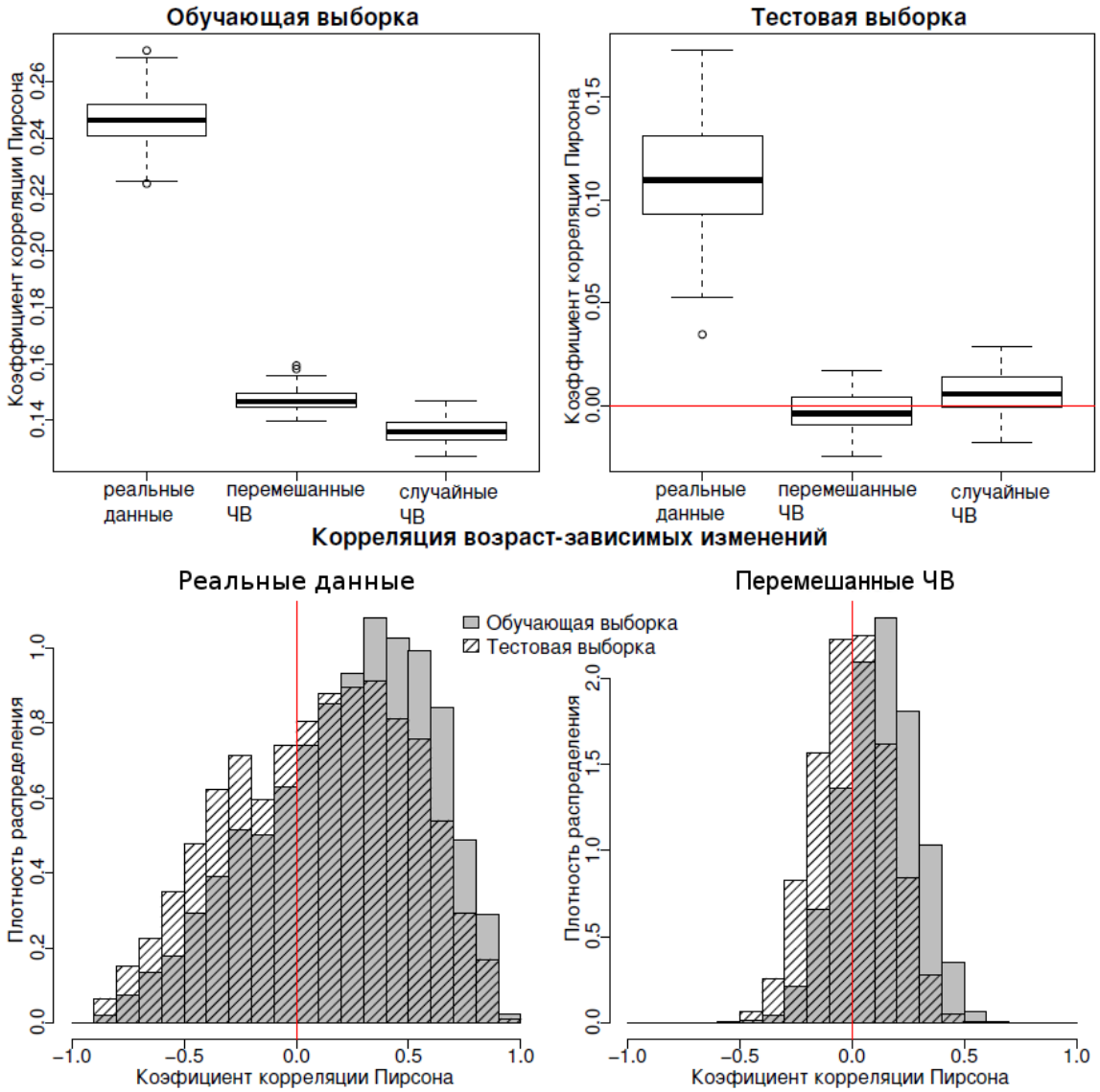

Рисунок 20: Моделирование ЧВ при помощи уровней экспрессии и аффинности факторов сплайсинга. Сверху показаны распределения коэффициентов корреляции Пирсона для реальных, перемешанных и случайных данных для обучающей и тестовой выборок. Здесь коэффициент корреляции рассчитывался между всеми ЧВ всех сегментов, 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.

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.

Список публикаций по теме диссертации

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.

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)