Федеральное государственное бюджетное учреждение науки

Институт проблем передачи информации им. А.А. Харкевича

Российской академии наук

На правах рукописи

Денисов Степан Владимирович

# Отбор и эпистаз в сайтах сплайсинга

03.01.09 – математическая биология, биоинформатика

Диссертация на соискание учёной степени

кандидата биологических наук

Научный руководитель:

доктор биологических наук, профессор М.С. Гельфанд

Москва – 2017

# Содержание

1. Введение ..... 3

1.1. Актуальность работы ..... 3

1.2. Цели и задачи исследования ..... 5

1.3. Научная новизна и практическая ценность ..... 6

1.4. Основные результаты и положения, выносимые на защиту ..... 6

1.4.1. Отбор в сайтах сплайсинга ..... 6

1.4.2. Коррелированная эволюция позиций в сайтах сплайсинга млекопитающих ..... 7

1.4.3. Консервативность цис-регулятора сплайсинга UGCAUG ..... 8

1.5. Структура и объем диссертации ..... 9

1.6. Список публикаций по теме диссертации ..... 9

2. Обзор литературы ..... 11

2.1. Альтернативный сплайсинг ..... 11

2.1.1. Что такое сплайсинг? ..... 11

2.1.2. Сайты сплайсинга ..... 12

2.1.2. Альтернативный сплайсинг и его регуляция ..... 18

2.2. Эволюция сплайсинга ..... 32

2.2.1. Макроэволюция сплайсинга ..... 32

2.2.2. Микроэволюция сплайсинга ..... 48

2.3. Отбор и эпистаз ..... 51

2.3.1. Положительный и отрицательный отбор ..... 51

2.3.2. Эпистаз ..... 59

3. Данные и методы ..... 70

3.1. Исходные данные ..... 70

3.1.1. Выборки конститутивных и кассетных экзонов и соответствующих сайтов сплайсинга ..... 70

3.1.2. Поиск ортологичных сайтов сплайсинга ..... 72

3.1.3. Данные по уровню экспрессии генов, уровню рекомбинации, консервативности и однонуклеотидным полиморфизмам ..... 73

3.2. Методы ..... 74

3.2.1. Построение матриц нуклеотидных замен методом парсимонии ..... 74

3.2.2. Сила сайта и её изменение ..... 75

3.2.3. Матрица ковариаций силы отдельных нуклеотидов в сайтах сплайсинга 76

3.2.4. Построение нейтральных контролей для оценки изменения силы позиций сайтов сплайсинга ..... 76

3.2.5. Оценка многовидовой консервативности ..... 78

3.2.6. Контроль на динуклеотидный состав для полипиримидиновых трактов акцепторных сайтов сплайсинга ..... 79

3.2.7. Статистические методы. Тестирование статистических гипотез и построение доверительных интервалов. ..... 80

4. Результаты и обсуждение ..... 81

4.1. Положительный и отрицательный отбор в сайтах сплайсинга ..... 81

4.1.1. Тест на положительный и отрицательный отбор в сайтах сплайсинга. Построение нейтральных контролей ..... 81

4.1.2. Отбор на консенсусные и неконсенсусные нуклеотиды ..... 82

4.1.3. Оценка силы отбора ..... 86

4.1.4. Сайт-специфический отбор на неконсенсусные нуклеотиды ..... 89

4.1.5. Отличия в силе отбора между разными классами сайтов сплайсинга ..... 93

4.1.6. Сильный положительный отбор в молодых сайтах сплайсинга, появившихся на линии *Homo sapiens* после расхождения с *Macaca mulatta* ... 96

4.1.7. Дрейфовый груз и отбор на уровне целого генома ..... 98

4.1.8. Свидетельства отбора на уровне однонуклеотидных полиморфизмов . 100

4.2. Коррелированная эволюция позиций в сайтах сплайсинга млекопитающих ..... 102

4.2.1. Метод восстановления матриц нуклеотидных замен в сайтах сплайсинга ..... 102

4.2.2. Оценка вероятностей последовательностей предковых сайтов сплайсинга в позиционно-независимой модели ..... 105

4.2.3. Изменение силы сайтов в ходе эволюции. Проверка гипотезы о миграции сигнала ..... 107

4.2.4. Нуклеотидные замены в индивидуальных позициях сайтов сплайсинга ..... 111

4.2.5. Сила нуклеотидов в различных позициях сайтов сплайсинга взаимно скоррелирована ..... 113

4.2.6. Меняются ли ковариации между позициями в ходе эволюции? Проверка гипотезы о независимой эволюции позиций сайтов сплайсинга ..... 124

4.3. Отбор в окрестности сайтов сплайсинга ..... 129

4.3.1. Консервативность цис-регулятора сплайсинга UGCAUG в геномах человека и мыши ..... 129

4.3.2. Тканевая специфичность экспрессии кассетных экзонов, потенциально регулируемых UGCAUG ..... 134

5. Основные результаты и выводы ..... 136

6. Приложение ..... 138

7. Благодарности ..... 148

8. Список литературы ..... 149

# 1. Введение

## 1.1. Актуальность работы

Прогресс современных методов секвенирования ДНК привел к прочтению полных геномов и транскриптомов многих высших эукариот. Сравнение геномов с транскриптомами позволяет установить экзон-интронную структуру генов, а значит и точно картировать сайты сплайсинга.

Несмотря на то, что структура и экспрессия мРНК генов человека и других высших эукариот исследованы достаточно подробно, вопрос о том, почему те или иные изоформы мРНК экспрессируются в одном физиологическом контексте (ткани, органе, на определенной стадии развития), но не экспрессируются в другом контексте, остаётся актуальным по настоящее время. По сути вопрос в следующем: как регулируется альтернативный сплайсинг? Для решения этого вопроса проведено множество исследований, в результате которых стало ясно, что на сплайсинг влияет множество факторов, таких как скорость транскрипции, расположение нуклеосом и эпигенетических маркеров, вторичная структура пре-мРНК и др. Однако важнейшими из этих факторов являются сила сайтов сплайсинга и наличие специальных последовательностей в пре-мРНК – энхансеров и сайленсеров сплайсинга (цис-элементов). С цис-элементами связываются специальные белки и РНК-белковые комплексы (транс-факторы). За счет тканеспецифичной и орган-специфичной экспрессии транс-факторов и осуществляется регуляция сплайсинга. В настоящей работе исследована одна из систем регуляции сплайсинга, основанная на цис-элементе UGCAUG, и разработан сравнительно-геномный метод предсказания функциональности энахансеров UGCAUG.

Наличие большого количества геномов, с другой стороны, открывает новые перспективы для эволюционной биологии. Сравнение геномов родственных видов позволяет строить филогенетические деревья, а также отвечать на более сложные вопросы эволюционной геномики.

Сайты сплайсинга представляют уникальный объект исследования в первую очередь из-за их значительного количества в геноме (несколько сотен тысяч), что позволяет исследовать тонкие эволюционные эффекты с разрешением вплоть до одного нуклеотида. В нашей работе был исследован слабый отбор, действующий на сайты сплайсинга. В частности впервые впрямую было показано наличие слабополезных мутаций при неизменном ландшафте приспособленности, что было предсказано теоретически. Кроме этого, были исследованы корреляции между различными позициями в сайтах сплайсинга и исследована роль эпистаза в эволюции сайтов сплайсинга.

## 1.2. Цели и задачи исследования

Целью работы было изучить, каким образом естественный отбор, генетический дрейф и эпистаз между позициями влияют на эволюцию сайтов сплайсинга, а также изучить консервативность регуляции кассетных экзонов на примере цис-регуляторного элемента UGCAUG.

Были поставлены следующие задачи.

1. Оценить направление и силу отбора, действующего на консенсусные и неконсенсусные нуклеотиды в сайтах сплайсинга в линиях *Homo sapiens* и *Drosophila melanogaster*.
2. Выяснить, какие характеристики сайтов сплайсинга влияют на силу отбора.
3. Оценить силу отбора, действующего на молодые сайты сплайсинга (то есть появившиеся на линии *H. sapiens*).
4. Оценить, как меняется сила отдельных позиций сайтов сплайсинга с течением эволюции.
5. Разработать метод, позволяющий проверять гипотезу о независимости эволюции позиций сайтов сплайсинга.
6. Изучить, как замены в одних позициях сайтов сплайсинга влияют на замены в других позициях, и дать эволюционную интерпретацию найденным закономерностям.

7. Изучить консервативность цис-регулятора сплайсинга UGCAUG, и проанализировать, как консервативность этого регулятора соотносится с его с функциональностью.

## 1.3. Научная новизна и практическая ценность

В работе рассмотрены актуальные вопросы современной эволюционной и сравнительной геномики. Впервые систематически исследован отбор, действующий на отдельные позиции сайтов сплайсинга. Подтверждено теоретически предсказанное существование слабополезных мутаций.

Разработан и программно реализован метод, позволяющий оценить влияние, оказываемое заменами в одних позициях сайтов сплайсинга на замены в других позициях. Этот метод потенциально применим к любым наборам последовательностей, на которые действует схожий отбор (сайты связывания транскрипционных факторов, цис-регуляторы сплайсинга и т.п.). Впервые выявлено, что эпистаз является важнейшим фактором эволюции позиций в донорных сайтах сплайсинга.

Впервые проанализирована консервативность цис-регулятора сплайсинга UGCAUG, на основе которой сделаны выводы о функциональности этих регуляторов. Дальнейшие экспериментальные исследования подтвердили многие из этих выводов.

## 1.4. Основные результаты и положения, выносимые на защиту

### 1.4.1. Отбор в сайтах сплайсинга

Замены из неконсенсусных (редко встречающихся) нуклеотидов в консенсусные (частые) фиксируются чаще, а из консенсусных в неконсенсусные – реже, чем ожидается при нейтральной эволюции. Это говорит о том, что на неконсенсусные нуклеотиды действует положительный, а на консенсусные, соответственно, отрицательный отбор.

Сила положительного и отрицательного отбора относительно низка ($1<|4N_e s|<4$). Иными словами, мутации из неконсенсусных нуклеотидов в консенсусные являются слабополезными, а обратные мутации – слабовредными. Существует круговорот слабовредных и слабополезных мутаций, благодаря совместному действию генетического дрейфа и отбора.

Это соотношение ($|s| \approx N_e$) сохраняется как на линии *H. sapiens*, так и на линии *D. melanogaster*, несмотря на разницу в эффективных численностях популяций этих видов.

В каждой позиции сила положительного отбора не отличается от силы отрицательного по порядку величины, что согласуется с теоретическими представлениями о том, что ландшафт приспособленности схож у всех рассматриваемых сайтов. У части сайтов сплайсинга, однако, существует отбор на неконсенсусные нуклеотиды.

В геномах человека и *D. melanogaster* содержатся сотни тысяч сайтов сплайсинга, каждый из которых содержит слабовредные неконсенсусные нуклеотиды. Учитывая коэффициенты отбора, действующего против неконсенсусных нуклеотидов, можно утверждать, что суммарный дрейфовый груз, вносимый сайтами сплайсинга огромен.

### 1.4.2. Коррелированная эволюция позиций в сайтах сплайсинга млекопитающих

Разработан метод восстановления вероятностей замен нуклеотидов на ветках филогенетического дерева (матрицы нуклеотидных замен) на примере сайтов сплайсинга в геномах человека мыши и собаки.

Разработан метод, позволяющий оценить вероятности предковых последовательностей сайтов сплайсинга, а также их средние веса в позиционно-независимой модели.

Показано, что изменение силы отдельных позиций сайтов сплайсинга в целом имеет сложный паттерн. Однако в донорных сайтах сплайсинга конститутивных экзонов наблюдается ослабление экзонной части и слабое увеличение силы интронной части сайтов, что согласуется с гипотезой о миграции сигнала из экзонной части в интронную.

Силы нуклеотидов в различных позициях сайтов сплайсинга часто взаимно скоррелированы. В донорных сайтах сплайсинга наблюдаются положительные корреляции между позициями внутри экзонной и внутри интронной частей и отрицательные – между экзонными и интронными частями. В полипиримидиновом тракте акцепторных сайтов сплайсинга наблюдается характерный чередующийся характер корреляций: соседние позиции скоррелированы отрицательно, через позицию – положительно, через две – отрицательно и т.д.

Отбор против динуклеотида AG есть основная причина корреляций, наблюдаемых в полипиримидиновом тракте акцепторных сайтов сплайсинга. Эпистаз – наиболее вероятная причина корреляций в донорном сайте сплайсинга.

Разработан метод, позволяющий проверять гипотезу о независимости эволюции позиций сайтов сплайсинга. Показано, что взаимодействие между различными позициями оказывает влияние на эволюцию сайтов сплайсинга, хотя и разнонаправленно. В донорных сайтах сплайсинга наблюдается увеличение по модулю положительных корреляций в течение эволюции внутри экзонных и интронных частей, так и отрицательных корреляций между ними. Таким образом, в донорных сайтах сплайсинга действует эпистатический отбор, который усиливает существующие корреляции между нуклеотидами.

### 1.4.3. Консервативность цис-регулятора сплайсинга UGCAUG

Была исследована выборка кассетных экзонов человека и мыши, которые по литературным данным экспрессируются специфически в тканях мозга. Мы роанализировали гексануклеотиды UGCAUG, встречающиеся в интронах в пределах 1000 нт после этих экзонов и обнаружили, что консервативность этих гексануклеотидов в геномах человека и мыши существенно выше средней консервативности интронов, что говорит о том, что гексануклеотиды выполняют некоторую важную функцию.

Анализ экспрессии кассетных экзонов в разных органах и тканях (по данным EST) показал, что экзоны включаются в зрелую мРНК не только в мозге, но и в других тканях (в частности, в мышечной ткани).

## 1.5. Структура и объем диссертации

Диссертация изложена на 166 страницах. Она состоит из восьми разделов: введение, обзор литературы, данные и методы, результаты и обсуждение, основные результаты и выводы, благодарности, приложение и список литературы. Работа содержит 27 рисунков и 5 таблиц. Список литературы содержит 218 наименований. Приложение содержит 11 рисунков и 2 таблицы.

## 1.6. Список публикаций по теме диссертации

По материалам диссертации опубликовано три статьи в рецензируемых научных журналах:

1. S. Denisov, G. Bazykin, A. Favorov, A. Mironov, M. Gelfand (2015) Correlated evolution of nucleotide positions within splice sites in mammals. PLOS ONE 10(12): e0144388.
2. S. Denisov, G. Bazykin, R. Sutormin, A. Favorov, A. Mironov, M. Gelfand, A. Kondrashov (2014) Weak negative and positive selection and the drift load at splice sites. Genome Biology and Evolution 6(6): 1437-1447.
3. S. Denisov, M.S. Gelfand (2003) Conservedness of the alternative splicing signal UGCAUG in the human and mouse genomes. Biophysics (Moscow) 48(1): 30-35

Результаты работы были представлены на международных и российских конференциях:

1. С.В. Денисов, Г.А. Базыкин, Влияние корреляций между позициями на эволюцию сайтов сплайсинга в геномах млекопитающих. Информационные технологии и системы (ИТиС'15), 7 - 11 сентября, 2015, Олимпийская деревня, Сочи, Россия
2. С.В. Денисов, Слабый отбор и дрейфовый груз в сайтах сплайсинга. Информационные технологии и системы (ИТиС'14), 1 - 5 сентября, 2014, Нижний Новгород, Россия
3. С.В. Денисов, Новорожденные сайты сплайсинга находятся под положительным отбором. Информационные технологии и системы (ИТиС'13), 1 - 6 сентября, 2013, Калининград, Россия
4. Stepan V. Denisov, Georgii A. Bazykin, Mikhail S. Gelfand, Alexey S. Kondrashov. Turnover of Slightly Deleterious and Slightly Advantageous Alleles in Splice Sites of Humans and Drosophila melanogaster. Society for Molecular Biology & Evolution Conference 2013 (SMBE 2013), July 7 – 11, 2013, Chicago, USA.
5. С.В. Денисов, Г.А. Базыкин. Роль неконсенсусных нуклеотидов в эволюции сайтов сплайсинга: данные по однонуклеотидным полиморфизмам (SNP) и дивергенции. Информационные технологии и системы (ИТиС'12), 19 - 25 августа, 2012, Петрозаводск, Россия
6. Stepan V. Denisov, Georgii A. Bazykin, Alexander V. Favorov, Andrey A. Mironov, Mikhail S. Gelfand. Positive and negative selection in splice sites. Society for Molecular Biology & Evolution Conference 2012 (SMBE 2012), June 23-26, 2012, Dublin, Ireland

# 2. Обзор литературы

## 2.1. Альтернативный сплайсинг

### 2.1.1. Что такое сплайсинг?

Большая часть эукариотических РНК, транскрибированных ДНК-зависимой РНК полимеразой II, перед тем как покинуть ядро и выйти в цитоплазму, подвергается процессингу (созреванию). Первичный транскрипт, с которым еще не произошел процессинг, называется пре-мРНК, а после созревания — зрелой мРНК, или просто мРНК.

Процессинг включает три основных события: кэпирование 5’-конца пре-мРНК, сплайсинг и полиаденилирование со стороны 3’-конца. Присоединение CAP сопряжено с транскрипцией: когда длина синтезируемого транскрипта достигает 25-30 нт, 7-метилгуанозин и другие компоненты CAP оказываются присоединенными к 5’-концу [1]. Сплайсинг – это процесс вырезания фрагментов из молекулы пре-мРНК с последующим сшиванием оставшихся фрагментов (в том же порядке, в котором они были в исходном транскрипте). Вырезанные участки называют интронами, а вошедшие в зрелую мРНК – экзонами. Сплайсинг, как и кэпирование, происходит котранскрипционно [2] – т.е. из молекулы РНК уже по мере ее синтеза вырезаются интроны.

Вскоре после открытия сплайсинга у аденовируса в 1977 году [1,2], стало ясно, что транскрипты некоторых генов способны сплайсироваться разными способами [3,4]. Такой процесс назван альтернативным сплайсингом. В противоположность этому конститутивный сплайсинг происходит всегда по одной и той же схеме. Ранее считалось, что лишь небольшая часть первичных транскриптов генов человека способны альтернативно сплайсироваться (около 5% согласно ранним оценкам [5]). Однако после массового секвенирования последовательностей генома и транскриптома оценки изменились. Картирование EST (коротких секвенированных фрагментов мРНК) на геном [6] и сравнение мРНК/EST между собой [7] показали,

что не менее 30% клеточных мРНК подвергается альтернативному сплайсингу. В зависимости от источника и метода подсчета эта цифра варьирует от 20 до 80%. Современные оценки склоняются в сторону больших значений (50% или более) [8,9]. Хотя вопрос о функциональности всех наблюдаемых изоформ мРНК остается открытым, стало очевидно, что альтернативный сплайсинг – важнейший источник белкового разнообразия у эукариот.

### 2.1.2. Сайты сплайсинга

#### *2.1.2.1. Структура сайтов сплайсинга и взаимодействие сплайсосомы с ними*

Сложный мультисубъединичный комплекс белков и малых ядерных РНК (мяРНК), осуществляющий сплайсинг, называется сплайсосомой. Сплайсосома опознает определенные последовательности на 5’ конце и на 3’ конце интрона. Эти последовательности называются донорный (5’) и акцепторный (3’) сайты сплайсинга. Кроме того, ближе к акцепторному сайту находится специальная последовательность, т. наз. сайт ветвления, важная для протекания реакции сплайсинга.

Сплайсосома катализирует две следующие друг за другом реакции трансэтерификации (рис. 1). На первом этапе 2’-гидроксил рибозы аденозина (этот аденозин лежит как раз в сайте ветвления) атакует фосфат в 5’-сайте (фосфат находится на границе экзона 1 и интрона). Это приводит к разрезанию молекулы РНК на границе экзон-интрон и присоединению 5’-конца интрона через фосфат к 2’-ОН аденозина. Таким образом, произошла первая реакция трансэтерификации. На следующем этапе гидроксил на 3’-конце первого экзона атакует фосфат на границе интрона и экзона 2. В результате сшиваются два экзона, а интрон высвобождается в форме лассо [10].

Сплайсосома состоит из пяти мяРНК и около 200 белков [11]. В клетках эукариот имеется два вида сплайсосом: U2-зависимые сплайсосомы и U12-зависимые сплайсосомы. В состав U2-зависимых сплайсосом входят следующие мяРНК: U1, U2, U4, U5, U6 и разнообразные белки. В U12-зависимых сплайсосомах роль U2

выполняет другая мяРНК –. U12, остальные мяРНК также заменены аналогами [12,13]. Большинство интронов в эукариотах (> 98%) вырезается с помощью U2-зависисмых сплайсосом [13].

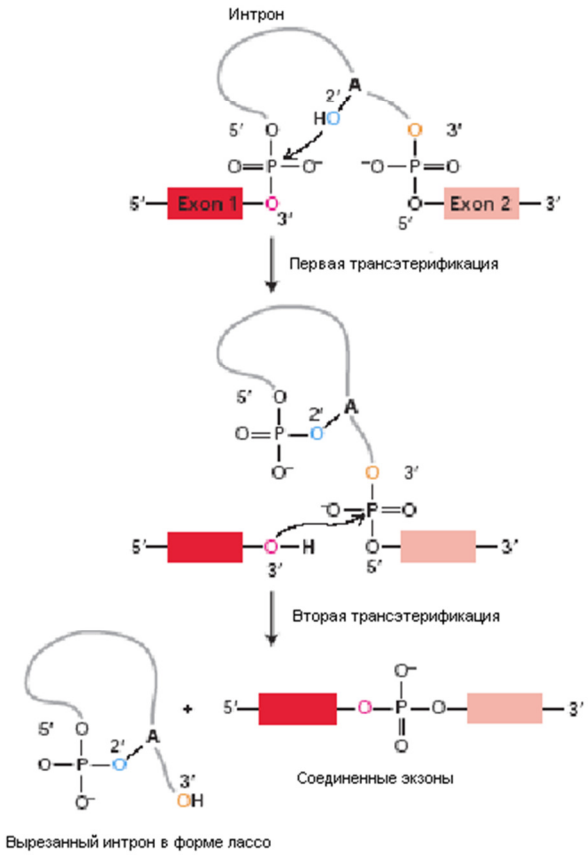


**Рисунок 1**. Химические реакции, происходящие при сплайсинге [10].

Для успешного вырезания интрона необходимо (хотя часто не достаточно) наличие донорного и акцепторного сайтов сплайсинга, а также сайта ветвления. Сайты сплайсинга разных экзонов и генов отличаются друг от друга и, соответственно, имеют разную энергию связывания со сплайсосомой [14,15]. Однако, существуют довольно четкие предпочтения в каждой позиции сайтов сплайсинга (рис. 2). Соответственно, можно разделить нуклеотиды в каждой позиции сайта на консенсусные (часто встречающиеся) и неконсенсусные (редкие). Консенсусные последовательности сайтов сплайсинга одинаковы по крайней мере для всех многоклеточных животных и очень схожи среди эукариот [16].

Рассмотрим структуру сайтов сплайсинга для U2-сплайсосомы (рис. 2). Донорный и акцепторный сайты лежат на границах экзона и интрона и имеют экзонную и интронную части. Интронная часть длиннее экзонной. Интронная часть донорного сайта состоит из практически инвариантного динуклеотида GU и еще четырех нуклеотидов. В структуре акцепторного сайта сразу перед экзонной частью (на конце интрона) почти всегда имеется динуклеотид AG, перед ним в большинстве случаев идет нуклеотид C, далее начинается полипиримидиновый тракт, где преобладают нуклеотиды U и C. (В соответствии с принятыми условностями, урацил (U), находящийся в мРНК часто обозначают соответствующим ему нуклеотидом на ДНК, т.е. как тимин (T). Так мы и будем делать далее по тексту.) Длина полипиримидинового тракта варьирует от сайта к сайту и составляет несколько десятков нуклеотидов [10]. Сайт ветвления, находящийся перед акцепторным сайтом в интроне, важен для сплайсинга, но не рассматривается в данной работе, в связи с тем, что на настоящее время сайты ветвления экспериментально идентифицированы только для ~20% интронов [17], а их последовательность весьма вырождена, что осложняет разработку надежного метода предсказания положения сайта ветвления в интроне [18,19].

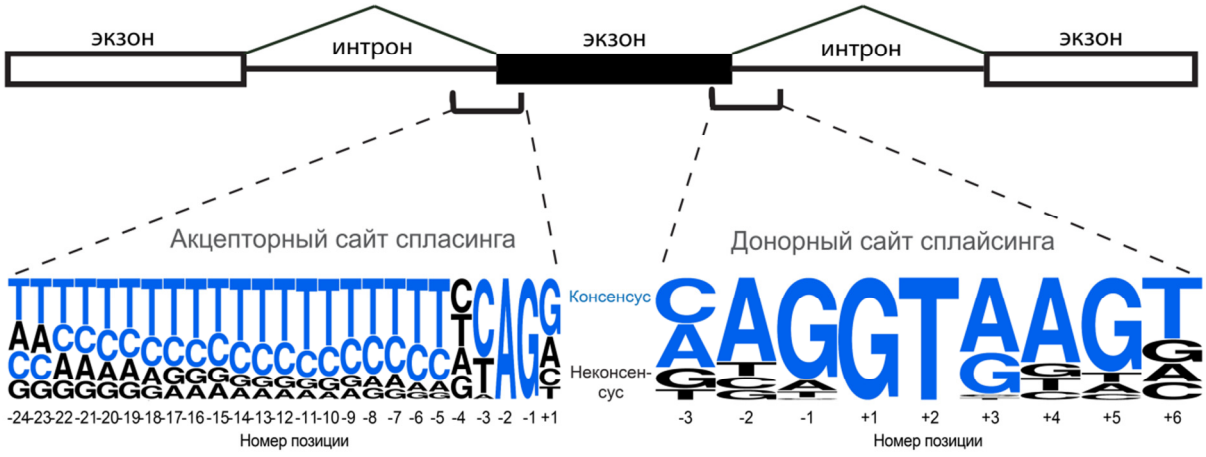


**Рисунок 2.** Консенсусные последовательности сайтов сплайсинга человека. Консенсусные нуклеотиды обозначены голубым цветом, неконсенсусные – черным. Высота буквы пропорциональна частоте ее встречаемости [20].

Динуклеотиды GT и AG в донорном и акцепторном сайтах сплайсинга, соответственно, имеются более, чем в 98,5% интронах эукариот [13]. Эти интроны обрабатываются U2-зависимой сплайсосомой. Существуют минорные классы интронов с неканоническими сайтами. Их можно разделить на U2-зависимые (с сайтами GC–AG и AT–AC) и U12-зависимые (AT–AC и GT–AG)[13]. Кроме того, известны случаи, когда TG выступает в качестве альтернативы AG в U2-зависимых интронах [21]. Мы в данной работе будем рассматривать только канонические GT-AG интроны.

Каким образом сплайсосома распознает сайты сплайсинга? Это достигается за счет образования пар между нуклеотидами мяРНК и нуклеотидами сайта сплайсинга, а также РНК-белковых взаимодействий (рис. 3). Компоненты сплайсосомы, участвующие в этих взаимодействиях, высококонсервативны среди эукариот, что свидетельствует о том, что сами взаимодействия схожи у разных видов [16].

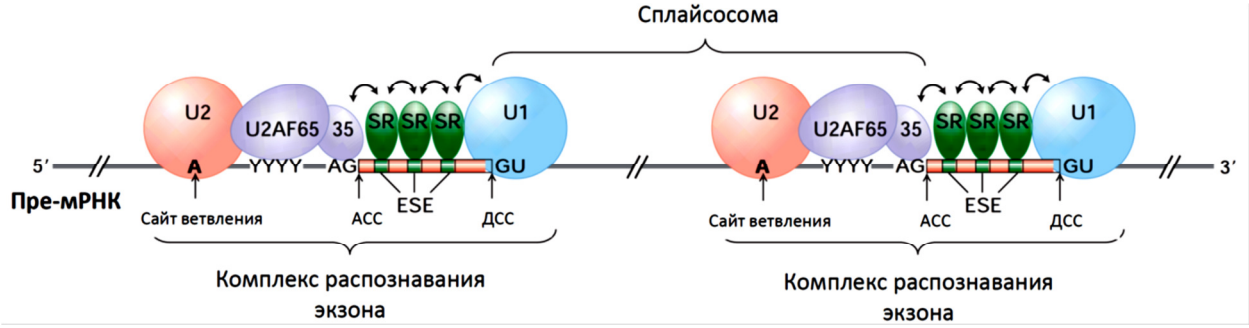


**Рисунок 3.** Взаимодействие сплайсосомы с пре-мРНК [10].

ДСС – донорный сайт сплайсинга, АСС – акцепторный сайт сплайсинга, ESE – экзонный энхансер сплайсинга.

5'-конец U1 мяРНК взаимодействует с донорным сайтом сплайсинга [22,23], что является ключевым фактором, влияющим на функциональность донорного сайта [15]. U1 может взаимодействовать с позициями донорного сайта с -3 до +6 [14], причем нуклеотиды в составе U1, которые образуют пары с позициями с -2 до +5 донорного сайта, консервативны в геномах всех эукариот [16,24], а нуклеотид G, входящий в U1, связывающийся с позицией -3(C) консервативен в геномах животных [16]. Первая стадия сборки сплайсосомы (формирование E-комплекса) зависит от связывания U1 с донорным сайтом сплайсинга [25–27]. Однако есть

свидетельства того, что сплайсинг в некоторых случаях может происходить в отсутствие U1, видимо, благодаря белку U1(C) [28]. Было показано, что человеческие белки U1(C) и U5(p220) участвуют во взаимодействии сплайсосомы с экзонной частью донорного сайта сплайсинга [29,30]. Далее в процессе сплайсинга на смену U1 приходят мяРНК U5 и U6. U5 связывается с позициями с -3 до +1, а U6 образует пары с нуклеотидами +5 и +6 донорного сайта [14], хотя по некоторым данным взаимодействие U5 с экзонными позициями сайта не очень существенно для протекания сплайсинга [31].

Специальные белки сплайсосомы взаимодействуют с акцепторным сайтом сплайсинга на стадии сборки E-комплекса. $U2AF^{65}$ связывается с полипиримидиновым трактом, а $U2AF^{35}$ взаимодействует с динуклеотидом AG. Белок SF1 связывается с сайтом ветвления, что в свою очередь облегчает взаимодействие $U2AF^{65}$ с ближайшим полипиримидиновым трактом [32,33]. $U2AF^{65}$ затем направляет U2 на связывание с сайтом ветвления [34]. У всех многоклеточных животных имеются функциональные гомологи этих трех белков. РНК-распознающие домены RRM1 и RRM2 белка $U2AF^{65}$ почти абсолютно консервативны среди позвоночных, уровень сходства этих мотивов между позвоночными и насекомыми больше 80% [16].

Считается, что сборка E-комплекса – это важнейшая стадия, на которой может происходить регуляция альтернативного сплайсинга, поскольку распознавание сайтов сплайсинга сплайсосомой происходит именно на этом этапе [32,35]. U1 и U2 играют здесь важную роль, хотя о сих пор не понятно, кто из них образует связь с пре-мРНК раньше [36]. Однако при любом сценарии другие мяРНК (U4, U5 и U6) привлекаются на более поздних стадиях сплайсинга [32]. Белки, ассоциированные с U5 мяРНК (образующие U5 мяРНП), связываются с полипиримидиновым трактом [37].

#### *2.1.2.2. Распознавание экзонов и интронов*

Важнейшей задачей, которая должна быть решена машиной сплайсинга, является четкое определение границ экзонов и интронов. Главными сигналами

распознавания являются собственно сайты сплайсинга. Их сила, т.е. приближенность соответствующих сайтов к консенсусу, играет важнейшую роль в распознавании [15]. Однако это необходимое условие, но недостаточное. Дело в том, что в геноме человека (и других позвоночных) экзоны представляют собой относительно короткие последовательности (средняя длина человеческого экзона – 150 нт), окруженные длинными интронами (средняя длина – около 3500 нт, хотя иногда они достигают 500000 нт) [38,39]. Сплайсосома должна четко опознать экзоны среди протяженных интронных последовательностей [40]. Определенные последовательности в сайтах сплайсинга и компоненты сплайсосомы, которые их специфически узнают, решают эту проблему лишь отчасти [38]. Существует значительное количество сайтов, которые сильно отклоняются от консенсуса, но тем не менее распознаются. Последовательности, которые соответствуют консенсусу сайтов сплайсинга часты в интронах, однако они в норме не используются для сплайсинга. Поэтому в интронах существуют т.н. псевдо-экзоны – последовательности, которые похожи на экзон как по размеру, так и по присутствию фланкирующих сайтов, но никогда не распознаваемые сплайсосомой как экзоны [19].

В 1990 г. Susan Berget предложила модель распознавания сайтов сплайсинга через экзон (exon definition). Согласно этой модели, части сплайсосомы, располагающиеся по разные стороны экзона, взаимодействуют друг с другом, что улучшает распознавание экзона. Было показано, что в таких комплексах распознавания экзона участвуют все пять сплайсосомных мяРНК [41]. Вероятными посредниками взаимодействия служат SR-белки, связанные с экзонными энхансерами сплайсинга (см. ниже) [14,39]. Таким образом, экзонные энхансеры важны не только в альтернативном сплайсинге, но и в конститутивном [14,38].

Размер экзона важен для эффективного сплайсинга. С одной стороны, он не может быть слишком маленьким в силу возникающих стерических затруднений. Так, было показано, что уменьшение изначально конститутивного экзона до размера менее чем 50 нт приводило к его пропуску [42]. Увеличение 18-нуклеотидного кассетного

экзона N1 гена c-src мыши до 109 нт приводило к эго конститутивному включению в транскрипт [43]. С другой стороны, слишком длинные экзоны могут некорректно сплайсироваться: увеличение внутреннего экзона более чем до 300 нт в *in vitro* эксперименте приводило либо к активации криптического сайта внутри экзона, либо к пропуску экзона целиком [39]. Учитывая, что большинство экзонов человека и других позвоночных короткие [38,39], следует полагать, что распознавание через экзон является преобладающим механизмом распознавания экзон-интронных границ. С этим согласуется наблюдение, что большинство мутаций в сайтах сплайсинга человека, вызывающих генетические заболевания, приводят чаще всего к пропуску экзона, реже к изменению длины экзона и реже всего к удержанию интрона [44].

У многих низших эукариот и у *Drosophila* наблюдается другая генная архитектура: короткие интроны чередуются с длинными экзонами. Например, у дрожжей *S. cerevisiae* интроны чаще всего короче 300 нт, а у D. melanogaster 50% интронов короче 100 нт [39]. В этом случае, распознавание через интрон (intron definition) является адекватной моделью: части сплайсосомы по разные стороны интрона взаимодействуют друг с другом. Эксперименты показали, что искусственное удлинение коротких экзонов у *D. melanogaster* приводит к различным дефектам сплайсинга (чаще всего, к удержанию интрона) [45,46].

### 2.1.2. Альтернативный сплайсинг и его регуляция

#### *2.1.2.1. Типы альтернатив, значение для организма*

Транскрипт одного гена может сплайсироваться множеством разных способов. Различные варианты мРНК, которые получаются в результате альтернативного сплайсинга транскрипта с одного и того же гена, называются изоформами (точнее мРНК-изоформами). Если альтернативный сплайсинг происходит в кодирующей области мРНК, в результате трансляции различных РНК-изоформ получаются различные белки – это белковые изоформы.

Разные гены способны давать различное количество изоформ. Известны примеры огромного числа вариантов мРНК, в частности ген *Dscam* у дрозофилы может продуцировать более 38000 разных изоформ [47]. У большинства же генов количество изоформ исчисляется единицами или десятками. Несмотря на все разнообразие изоформ, можно выделить определенные элементарные события альтернативного сплайсинга. Различают следующие элементарные альтернативы (рис. 4); (1) кассетный экзон – входит в одни РНК-изоформы и полностью отсутствует в других; (2) альтернативный акцепторный сайт, (3) альтернативный донорный сайт – приводят к удлинению или укорочению экзона; (4) удержанный интрон – в одном случае интрон вырезается, в другом – сплайсинга интрона вообще не происходит; (5) взаимоисключающие (чередующиеся) экзоны – в изоформу входит то один, то другой экзон. Отдельно выделяют альтернативные старты транскрипции и полиаденилирования.

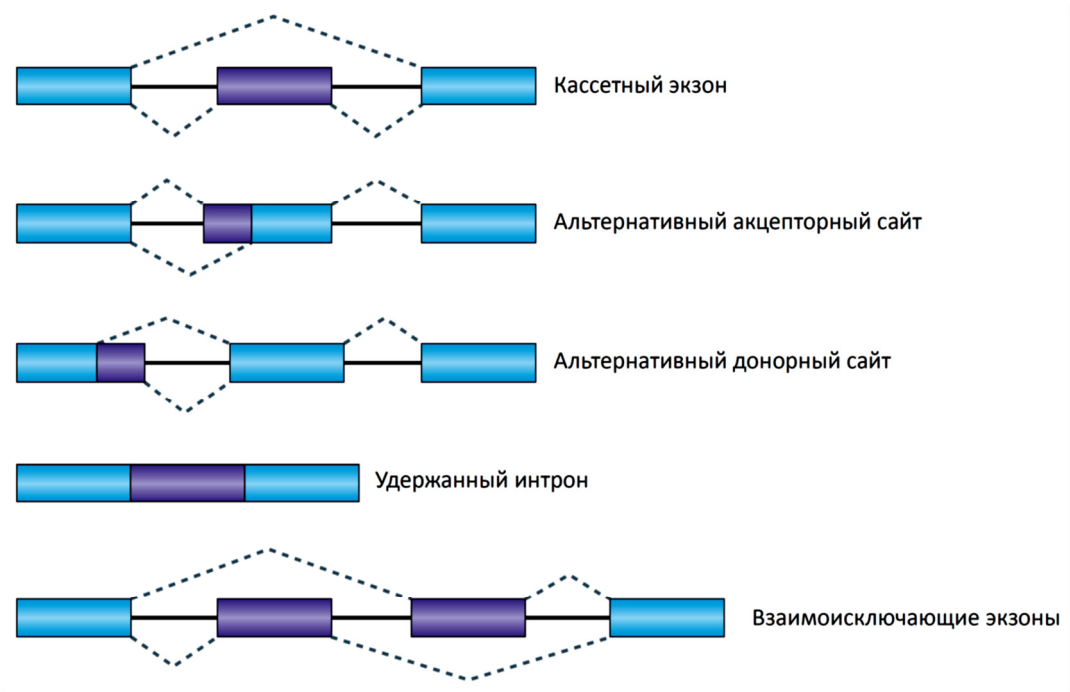


**Рисунок 4.** Типы элементарных альтернатив. Экзоны показаны прямоугольниками. Горизонтальные линии, соединяющие прямоугольники – интроны. Варианты вырезания интронов показаны ломаными пунктирными линиями. Из [48] с изменениями.

Существует определенный биологический контроль за тем, как сплайсинг должен происходить. Изоформы нередко являются специфичными для определенных

тканей и органов. Это подтверждено как конкретными примерами, так и полногеномными исследованиями [9]. Существуют механизмы регуляции альтернативного сплайсинга, которые позволяют определенным изоформам дифференциально экспрессироваться в том или ином физиологическом контексте (ткани, органы, стадии развития организма), однако целостная картина регуляции пока не ясна.

#### *2.1.2.2. Что регулирует альтернативный сплайсинг – основные факторы*

Какие механизмы позволяют избирательно использовать те или иные сайты сплайсинга (т.е. регулировать альтернативный сплайсинг)?

Мера приближенности сайта сплайсинга к консенсусу есть сила сайта сплайсинга. Сила сайтов сплайсинга существенно влияет на то, является ли сплайсинг альтернативным или конститутивным. В нескольких исследованиях было показано, что сайты альтернативных экзонов слабее сайтов конститутивных экзонов [49–53]. Сила сайта сплайсинга является важнейшим фактором для распознавания сплайсосомой [15]. Эти наблюдения согласуются с идеей, согласно которой относительно слабые сайты сплайсинга позволяют сделать сплайсинг более регулируемым, например, за счёт дополнительных энхансеров и сайленсеров (см. ниже).

Важна, по-видимому, длина экзона и интрона. У позвоночных, где преобладающим механизмом является распознавание через экзон [39], неоптимальная длина экзона способствует альтернативному сплайсингу. Однако у *Drosophila melanogaster*, где в основном происходит распознавание через интрон [39], важна длина интронов. Так, среди экзонов в геноме *D. melanogaster*, окруженных длинными интронами, доля кассетных экзоном более чем на порядок превышает среднюю. Альтернативные донорные и акцепторные сайты чаще встречаются на границе длинных интронов. В геноме человека ситуация иная: как и у *D. melanogaster*, кассетные экзоны чаще встречаются в окружении длинных интронов, но этот эффект гораздо слабее чем у *Drosophila*. Вероятность

альтернативных донорных и акцепторных сайтов в геноме человека и вовсе снижается с увеличением длины интрона [54].

Важнейшим компонентом регуляции сплайсинга являются специальные последовательностей в пре-мРНК и белковые или РНК-овые факторы, которые с ними взаимодействуют, направляя сплайсосому на распознавание того или иного сайта сплайсинга. Такие последовательности называются цис-элементами (цис-регуляторами), а соответствующие белки и РНК – транс-факторами. Среди цис-элементов различают энхансеры и сайленсеры сплайсинга. Энхансеры – это цис-элементы, которые позитивно влияют на распознавание данного сайта сплайсосомой, а сайленсеры – негативно [35]. Гены транс-факторов дифференциально экспрессируются в разных тканях, на разных стадиях развития, а также в различных физиологических условиях, что и делает возможным тот факт, что в разных клетках организма сплайсинг происходит по-разному.

Можно сказать, что сайты сплайсинга конкурируют за сплайсосому [55]. Если, к примеру, сайт сильно отклоняется от консенсуса по последовательности, либо его распознавание осложнено вторичной структурой РНК, или соответствующий экзон имеет неподходящий размер, это понижает вероятность того, что сайт будет распознан сплайсосомой. Однако позитивная или негативная регуляция со стороны цис-элементов может изменить эту ситуацию. Так, изначально слабый сайт, подверженный позитивной регуляции, может эффективно конкурировать с сильным. С другой стороны, сильный сайт под воздействием сайленсера сплайсинга может иметь сравнимую или даже меньшую вероятность распознавания сплайсосомой, чем слабый сайт. Сигналы, регулирующие сплайсинг, встречаются как в экзонах, так и в интронах. Соответственно различают экзонные и интронные энхансеры (ESE и ISE соответственно) и сайленсеры (ESS и ISS) сплайсинга. Так как последовательности ESE встречаются как в альтернативных, так и в конститутивных экзонах, считается, что они участвуют не только в регуляции альтернативного сплайсинга, но и в конститутивном сплайсинге [38,56].

Вторичная структура мРНК также может регулировать альтернативный сплайсинг. Элементы вторичной структуры могут закрывать сайты сплайсинга или цис-регуляторы сплайсинга, делая их недоступными для сплайсосомы или соответствующих транс-факторов. Кроме того возможна регуляция за счет уменьшения расстояния между ключевыми цис-элементами или сайтами сплайсинга [57].

На регуляцию сплайсинга существенное влияние оказывает то, что сплайсинг происходит котранскрипционно [58]. По-видимому, на сплайсинг влияют два основных фактора, связанных с транскрипцией: (1) фактор рекрутирования – транс-факторы сплайсинга садятся на соответствующие сайты на пре-мРНК в процессе транскрипции с участием РНК-полимеразы II; (2) скорость элонгации РНК-полимеразы II может регулировать альтернативный сплайсинг. Многие транс-факторы (SR-белки и hnRNP, см. ниже) иммунопреципитируются совместно с РНК-полимеразой II [59].

Эксперименты *in vitro* показали, что SR-белки эффективно усиливают сплайсинг, только если они были добавлены до начала транскрипции [58,59]. С-концевой домен РНК-полимеразы играет ключевую роль в этом процессе, во всяком случае для некоторых SR-белков. Представитель SR-белков SRSF3 (также известный как SRp20) в норме подавляет включение кассетного экзона 33 в мРНК гена фибронектина человека, однако удаление CTD, равно как и нокдаун гена SRSF3 приводят к включению экзона, причём этот эффект не определяется скоростью транскрипции [60]. Было также показано, что медиаторный комплекс, собирающийся на промоторе (тот, что взаимодействует с транскрипционными факторами) содержит субъединицу MED23, которая связывается с регулятором сплайсинга hnRNPL [58].

Действует также кинетический фактор. Скорость элонгации транскрипции влияет на сплайсинг, регулируя появление в новосинтезирующимся транскрипте сайтов сплайсинга и регуляторных цис-элементов, тем самым влияя на конкуренцию между сайтами сплайсинга за сплайсосому. Предполагаемый механизм состоит в

том, что при медленной элонгации донорный сайт интрона перед кассетным экзоном взаимодействует с ближайшим акцепторным сайтом через интрон, а не с более далёким сайтом, т.к. более далёкий сайт еще не появился в синтезирующемся транскрипте. Это и приводит к преимущественному включению кассетного экзона. Химические агенты, замедляющие РНК-полимеразу II, способствуют включению кассетных экзонов, а транскрипционные факторы, ускоряющие элонгацию, а также вещества, способствующие открытому хроматину, способствуют пропуску экзона. Аналогичная ситуация наблюдалась с использованием “медленных” мутантных полимераз [58]. Как уже упоминалось, вероятность того, что кассетный экзон включится в сплайсированную мРНК, зависит от длин окружающих интронов: чем длиннее окружающие интроны, чем чаще экзон включается. При этом гораздо сильнее выражена зависимость от длины интрона перед кассетным экзоном (а не после него) как в геноме человека, так и в геноме *D. melanogaster* [54], что подтверждает описанную модель.

Состояние хроматина также влияет на альтернативный сплайсинг. По-видимому, модификация гистонов и позиционирование нуклеосом являются определяющими факторами.

Говоря упрощенно, модификации гистонов можно разделить на те, что способствуют активной транскрипции (например, триметилирование лизина 36 гистона H3(H3K36me3), H3K4me2, H3K4me3, ацетилирование лизина 9 гистона H3 H3K9ac), и те, что подавляют транскрипцию (к примеру H3K9me2, H3K9me3 и H3K27me3) [58]. Некоторые типы гистоных меток (H3K36me3 и H3K27me2) перепредставлены в экзонах по сравнению с интронами [61]. Более того, экзоны, окруженные длинными интронами, обогащены метками, отличающими экзоны от интронов (H3K27me2 и H3K36me3), а также некоторыми другими (H3K4me3, H3K27me1 и H3K36me1)[61], что, вероятно, свидетельствует о том, что эти модификации способствуют распознаванию экзонов. На дрожжевых клетках было показано, что альтернативные экзоны обеднены H3K9me3 по сравнению с конститутивными [62], однако на человеческих клетках никакой разницы в

гистонных модификациях между конститутивными и альтернативными экзонами обнаружено не было [61]. Для гистонной модификации H3K36me3 показано участие в альтернативном сплайсинге транскриптов для ряда генов. Адапторный белок MRG15 взаимодействует с H3K36me3 и привлекает белок PTB, который подавляет включение кассетного экзона. Например, в гене FGFR2 экзон включается в транскрипт в эпителиальных клетках, где модификации H3K36me3 встречаются реже, а H3K4me3 – чаще, чем в других типах клеток. Гистонные модификации могут также влиять на скорость элонгации РНК-полимеразы II, которая в свою очередь влияет на сплайсинг. Например, деполяризация мембраны в нейронах вызывает ацетилирование H3K9ac непосредственно в гене NCAM, что приводит к локальному открытию хроматина [63]. Открытое состояние хроматина способствует быстрой скорости синтеза пре-мРНК и, соответственно, вырезанию кассетного экзона из транскрипта в согласии с кинетическим механизмом, описанным выше [58].

Полногеномные исследования показали, что нуклеосомы располагаются преимущественно в экзонах [61,64,65]. В интронах нуклеосомы тоже встречаются, но расположены хаотично, в то время как на один экзон в среднем приходится одна нуклеосома (средняя длина экзона человека в 150 нт [38] примерно равна одному обороту ДНК вокруг нуклеосомы) и расположение нуклеосом более регулярно [61,64]. Известно также, что РНК-полимераза II движется неравномерно по ДНК-матрице, замедляясь при прохождении через нуклеосому [58]. На основании этих наблюдений была предложена модель, согласно которой регулярное позиционирование нуклеосом на экзонах способствует включению экзонов в транскрипт, т.к. сплайсосоме предоставляется дополнительное время для распознавание акцепторного сайта сплайсинга. В пользу этой модели говорит то, что экзоны со слабыми сайтами сплайсинга обогащены нуклеосомами, причем отрицательная корреляция между силой сайта и частотой расположения нуклеосом сильнее для акцепторного сайта [61]. Этот механизм также может быть подвержен регуляции, о чем косвенно свидетельствует тот факт, что частота включения экзона коррелирует с частотой встречаемости нуклеосом [65].

### *2.1.2.3. Цис-элементы и транс-факторы, регулирующие сплайсинг*

Существуют две наиболее обширные группы белков, которые связываются с цис-элементами в экзонах и интронах, – это SR-белки и hnRNP. SR-белки необходимы для конститутивного сплайсинга (см. раздел 2.1), и они же участвуют в регуляции альтернативного сплайсинга [66]. Они высоко консервативны белков среди многоклеточных Эти белки имеют модульную структуру и состоят из одного или двух копий РНК-узнающего мотива (RNA-recognition motif, RRM) и С-концевого домена, богатого дипептидами Arg-Ser (RS-домен) [19]. RRM определяет связывание с определенными последовательностями пре-мРНК, а RS-домен отвечает за белок-белковые взаимодействия [67]. SR-белки могут взаимодействовать между собой и с другими белками, которые содержат RS-домены, например с регуляторами сплайсинга Tra и Tra2 и с компонентами сплайсосомы U1-70K и U2AF35. Важно, что для этих взаимодействий обязательно наличие RS-доменов в каждом из белков [67]. SR-белки принимают непосредственное участие в образовании комплекса распознавания экзона (рис. 3), а последовательности, с которыми они связываются, являются экзонными энхансерами и способствуют включению экзона в сплайсированную РНК.

Белки, принадлежащие к семейству гетерогенных рибонуклеопротеидов (hnRNP), представляют собой довольно разнородную группу. Эти белки выполняют множество различных функций: упаковка РНК, транспорт мРНК из ядра в цитоплазму, альтернативный сплайсинг [32]. Они содержат различные типы РНК-связывающих доменов. Некоторые белки содержат домен, богатый аргинином и глицином, который может как связывать РНК, так и участвовать в белок-белковых взаимодействиях [40]. hnRNP диффузно распределены по всему ядру, тогда как SR-белки совместно с другими факторами сплайсинга образуют компактные пятна локализации [40].

Кроме описанных групп белков существует множество других факторов, которые могут дифференциально экспрессироваться в разных тканях и на разных стадиях

развития организма и связываться с соответствующими цис-элементами (например, nPTB, PTB, NOVA [58]).

Так, на ряде примеров [69–72] было показано, что использование кассетных экзонов, экспрессирующихся преимущественно в мозговой ткани позвоночных, регулируется посредством цис-элемента UGCAUG – энхансера сплайсинга. Однако транс-фактор, взаимодействующий с этим элементом, был неизвестен.

Известно, что поли-G последовательности ($G_n$, где $n \geq 3$), часто расположенные кластерами в интронах, увеличивают вероятность распознавания ближайших к ним сайтов сплайсинга, как донорных, так и акцепторных (ISE) [72]. При этом G-триплеты, расположенные в непосредственной близости к донорному сайту сплайсинга, физически взаимодействуют с U1 мяРНК [73].

#### *2.1.2.4. Поиск новых цис-элементов*

Поиск новых цис-элементов представляет собой важную задачу, т.к. ясно, что эти элементы широко распространены, разнообразны и вносят важный вклад в регуляцию сплайсинга [58].

Исторически первым подходом к поиску цис-элементов являлось изучение отдельных генов и соответствующих событий сплайсинга. Так, например, энхансерная функция G-триплетов, расположенных вблизи донорных сайтов сплайсинга, была исследована на примере интрона гена альфа-глобина человека. Спаривание нуклеотидов G-триплета с соответствующими нуклеотидами U1 мяРНК (входящей в состав сплайсосомы) и его функциональность в качестве ISE были показаны методом сайт-направленного мутагенеза и введения компенсаторных замен [73]. Подобных примеров было описано достаточно много [32,35], но единой картины регуляции не вырисовывалось, вероятно, в связи с большим разнообразием цис-элементов и соответствующих транс-факторов.

Глобальный экспериментальный поиск сайтов связывания известных транс-факторов, а также новых энхансеров и сайленсеров спайсинга (вне привязки к конкретному гену) производился методом SELEX (systematic evolution of ligands by

exponential enrichment). Представим, что мы имеем дело с плазмидой, содержащей минигенную конструкцию с кассетным экзоном. Суть метода сводится к рандомизации последовательности внутри экзона с последующим тестированием экзона на включение в мРНК. На выходе имеется набор последовательностей длиной ~20 нт, про которые известно, что внутри них содержатся экзонные энхансеры (или сайленсеры). Анализ этих последовательностей позволяет выявить сходные мотивы – потенциальные ESE (ESS). Существует две группы исследований с использованием SELEX: (1) идентификация энхансерных последовательностей, которые являются потенциальными сайтами связывания для конкретных белков (например, SF2/ASF); (2) выявление экзонных энхансеров (или сайленсеров) как таковых без привязки к транс-факторам, которые с ними связываются;

Исследования первого рода проводятся *in vitro*. Реакцию сплайсинга проводят в т.наз. S100 экстрактах – это пост-ядерная и пост-рибосомальная клеточная фракция, которая может поддерживать сплайсинг *in vitro* только при добавлении в реакционную смесь одного или нескольких SR-белков [19]. Это удобная модельная бесклеточная система сплайсинга, позволяющая тестировать индивидуальные SR-белки на способность активировать сплайсинг. Liu с соавт. [74] использовали такой подход для анализа экзонных энхансеров, распознаваемых человеческими SR-белками SF2/ASF, SRp40 и SRp55. Были обнаружены следующие консенсусные мотивы – потенциальные сайты связывания: SRSASGA для SF2/ASF, ASDGS для SRp40 и USCGKM для SRp55. В этой же работе указывается, что известные сайты связывания для данных белков в конкретных генах часто противоречат найденному консенсусу. Стоит подчеркнуть, что реальные энхансеры могут отличаться от найденных таким путём из-за давления отбора на белок-кодирующую последовательность. Вероятно также, что правильная идентификация сайтов связывания сильно зависит от деталей экспериментальных процедур. *In vitro* системы часто не отражают реальных клеточных условий, т.к. сплайсинг проводится в условии избытка одних SR-белков и недостатка других. К тому же высокая аффинность не всегда коррелирует с лучшей функциональностью [19].

Исследования второго рода могут проводиться как *in vitro* [75], так и *in vivo* [76]. Стратегия *in vitro* использует клеточные экстракты, которые сами по себе могут эффективно стимулировать сплайсинг. Стратегия *in vivo* лишена вышеперечисленных недостатков, характерных для систем *in vitro*. Ранние исследования давали довольно общую картину консенсусных последовательностей: обнаруживались пурин-богатые мотивы [75] и A/C-богатые мотивы [76]. Wang с соавт. [77] применили SELEX *in vivo* для идентификации экзонных сайленсеров (ESS). Рандомизировалась последовательность из 10 нт. Кластерный анализ полученных декамеров и сравнение с известными ESS дали следующий результат. Две группы выявленных мотивов оказались сходны с известными сайтами связывания для hnRNPA1 и hnRNPH соответственно. Еще одна группа мотивов обнаруживает значительное сходство с консенсусом донорного сайта сплайсинга человека. Среди всех декамеров, обладающих сайленсерной активностью, был осуществлен поиск гексануклеотидов, представленных в избытке по сравнению со случайной последовательностью. Было найдено значительное количество таких гексамеров. Шесть гексамеров были протестированы в *in vivo* системе и четыре из них обладали сайленсерной активностью.

С появлением секвенированных геномов и транскриптомов эукариот стал возможен полногеномный поиск новых цис-элементов и соответствующих транс-факторов. Был предложен ряд биоинформатических, а также экспериментально-биоинформатических подходов.

Свойства найденных в SELEX-экспериментах потенциальных ISS и ESS были проверены на полногеномном уровне с использованием позиционных весовых матриц. Позиционная весовая матрица – это биоинформатический способ представления мотивов в нуклеотидных последовательностях, в частности, сайтов связывания белков с ДНК или РНК. Она строится на основании частот встречаемости нуклеотидов в каждой из позиций [78]. На основе данных SELEX были построены позиционные весовые матрицы для сайтов связывания SR-белков SF2/ASF, SC35, SRp40 и SRp55 и реализована программа ESEfinder [79],

осуществляющая поиск потенциальных сайтов связывания этих белков. Было показано, что предсказанные сайты связывания встречаются чаще в экзонах, чем в интронах. Известно, что синонимичные замены в белок-кодирующей области (т.е., не приводящие к изменению аминокислоты) могут существенно менять паттерн сплайсинга. Возможное объяснение состоит в том, что они разрушают энхансеры/сайленсеры сплайсинга [19]. Оказалось, что более половины точечных мутаций, приводящих к аномальному вырезанию экзона, попадают в область предсказанных энхансеров [79].

Wang с соавт. провели сравнительный анализ содержания потенциальных ESS-гексамеров и декамеров (найденных в SELEX эксперименте, см. выше) в следующих типах экзонов: (1) конститутивные экзоны, (2) кассетные экзоны, (3) псевдоэкзоны, (4) “строгие” экзоны – т.е. с близкими к консенсусу сайтами сплайсинга, и (5) “слабые” экзоны. Оказалось, что конститутивные экзоны обеднены потенциальными ESS как по сравнению с кассетными экзонами, так и по сравнению с псевдоэкзонами. Насыщенность псевдоэкзонов сайленсерами можно объяснить тем, что их необходимо репрессировать, чему и способствуют ESS. “Строгие” экзоны насыщены сайленсерами в сравнении со “слабыми” скорее всего потому, что в экзонах со слабыми сайтами идет негативная селекция против ESS, избыток которых может, грубо говоря, превратить экзон в псевдоэкзон. В кассетных экзонах больше ESS, чем в конститутивных, вероятно, потому, что они более подвержены разнообразной регуляции и поэтому должны содержать больше регуляторных элементов, в частности ESS. Однако это предположение не подтверждается для экзонных энхансеров (см. ниже).

Существуют также чисто биоинформатические подходы к поиску энхансеров и сайленсеров сплайсинга.

Некоторые гены не содержат интронов, следовательно, не сплайсируются вовсе. Поэтому, эти гены должны быть обеднены (или совсем не содержать) энхансеров и сайленсеров. В безинтронных генах содержится больше синонимичных однонуклеотидных полиморфизмов (SNP), чем в генах с интронами [80], что

согласуется с тем, что отрицательный отбор, действующий на потенциальные энхансеры/сайленсеры ослаблен или отсутствует. Олигонуклеотидный состав генов без интронов и генов с интронами в кодирующей области существенно различается [81], однако мотивов выявить не удалось, видимо, в связи с тем, что кодирование белка накладывает существенные ограничения на вариабельность последовательностей. Для того, чтобы избежать указанной проблемы Zhang с соавт. включили в анализ только некодирующие последовательности: они сравнили олигонуклеотидный состав некодирующих экзонов с псевдоэкзонами, а также с 5'-нетранслируемыми областями генов, не содержащих экзонов [82]. Было найдено несколько тысяч октамеров, пере- или недопредставленных в некодирующих экзонах. Перепредставленые последовательности представляют собой потенциальные энхансеры, а недопредставленные – потенциальные сайленсеры сплайсинга. Этот метод был назван PESX (Putative Exonic Splicing Enhancers/Silencers). Функциональность найденных октамеров была проверена в минигенных конструкциях [82], а также посредством мутационного анализа потенциальных энхансеров и сайленсеров, встречающихся в экзонах млекопитающих [83].

Существует гипотеза, что некоторые энхансеры усиливают слабые сайты сплайсинга (см. например, [55]). В работе [84] был разработан основанный на этой идее метод предсказания ESE, названный RESCUE-ESE (Relative Enhancer and Silencer Classification by Unanimous Enrichment). Авторы провели сравнительный анализ олигонуклеотидного состава трех групп экзонов: экзонов в целом, экзонов со слабым донорным сайтом и со слабым акцепторным сайтом. Гексануклеотиды, избыточно представленные в экзонах со слабыми сайтами по сравнению с экзонами в целом, были выявлены и кластеризованы по степени сходства. Было обнаружено пять кластеров для экзонов со слабым донорным сайтом и восемь кластеров для экзонов со слабым акцепторным сайтом. Интересно, что три кластера сходны между этими двумя группами экзонов, что говорит в пользу того, что некоторые энхансеры могут усиливать как слабый акцепторный сайт, так и слабый донорный сайт. Типичные представители из каждого кластера были

протестированы на способность усиливать сплайсинг *in vivo* с использованием минигенных конструкций. В качестве контроля использовались точечные мутанты тестируемых гексамеров. Все десять конструкций стимулировали сплайсинг, хотя и с разной эффективностью. Девять из десяти усиливали сплайсинг значительно сильнее, чем соответствующий точечный мутант. Один из наиболее часто встречающихся гексамеров, GAAGAA, был ранее найден во множестве экзонов, и для него была экспериментально показана активность в регуляции сплайсинга. Следует отметить, что вырожденность предсказанных энхансеров столь высока, что около 10% всех возможных гексамеров соответствуют консенсусу, и в среднем произвольный экзон человека содержит от трех до семи предсказанных сайтов. Кассетные и конститутивные экзоны не отличаются значительно по наличию предсказанных энхансеров, что согласуется с предположением о том, что конститутивные экзоны также подвержены положительной регуляции со стороны энхансеров. Анализ однонуклеотидных полиморфизмов в человеческой популяции, пересекающихся с предсказанными энхансерами, показал наличие в них отрицательного отбора [85].

Yeo с соавт. [86] исследовали экзонные и интронные энхансеры, а также сайты сплайсинга на предмет вариабельности между геномами позвоночных. Консенсусные последовательности сайтов сплайсинга значимо не отличаются как между разными млекопитающими, так и между млекопитающими и рыбами. Предсказанные программой RESCUE-ESE экзонные энхансеры также высоко консервативны. Интронные энхансеры, однако, значительно отличаются у рыб и млекопитающих. У млекопитающих часто встречаются GGG-триплеты, а у рыб – мотивы с повторами динуклеотидов AC и GT. Консервативность транс-факторов варьирует: SR-белки хорошо сохраняются среди всех позвоночных, hnRNP достаточно консервативны внутри млекопитающих, но различаются по доменной структуре и присутствию/отсутствию в геномах рыб и млекопитающих. Сказанное выше касается известных регуляторов. Что касается еще неидентифицированных цис-регуляторов в интронах, то косвенно о высокой их консервативности говорит тот факт, что между геномами человека и мыши интронные последовательности,

примыкающие к экзонам, более консервативны, чем последовательности в глубине интронов, причем кассетные экзоны окружены более протяженными участками консервативности, чем конститутивные экзоны [87]. О том же свидетельствует то, что распределение синонимичных SNP неравномерно вдоль экзонов, края экзонов обеднены SNP [80]. Т.к. цис-элементы и транс-факторы, влияющие на сплайсинг, высоко консервативны между млекопитающими, в то время как интронные последовательности в целом сильно различаются между разными млекопитающими (например, между человеком и мышью), существует возможность идентифицировать интронные регуляторные элементы, исходя из их консервативности.

Brudno с соавт. проанализировали последовательности в интронах, встречающиеся в окрестности кассетных экзонов, экспрессирующихся в нервной ткани человека и мыши, и обнаружили, что наиболее перепредставленным гексануклеотидом в интронах, следующих за кассетными экзонами (по сравнению с конститутивными экзонами), является UGCAUG. Мы в настоящей работе проанализировали консервативность этого элемента и сделали выводы о его функциональности.

## 2.2. Эволюция сплайсинга

### 2.2.1. Макроэволюция сплайсинга

#### *2.2.1.1. Макроэволюция интронов*

Ключевые компоненты сплайсосомы консервативны во всех хорошо охарактеризованных эукариотических организмах [88]. Филогенетические реконструкции показывают, что U2-сплайсосома присутствовала у общего предка всех эукариотических организмов (LECA, Last Eukaryotic Common Ancestor) [89].

В геномах бактерий и органелл встречаются самосплайсирующиеся интроны, хотя они и очень редки (т. наз. интроны типов I и II). Эти интроны вырезаются без помощи сплайсосомы за счёт рибозимной активности самих интронов,

образующих специальные структуры [90]. Сплайсосомные интроны (т.е. те, которые вырезаются с помощью сплайсосомы), равно как и компоненты сплайсосомы, в геномах архей и бактерий отсутствуют. У эукариот сплайсосомные интроны присутствуют, однако их количество в геноме значительно варьирует: от менее 100 у некоторых одноклеточных до сотен тысяч у позвоночных и растений [90]. Число интронов на тысячу нуклеотидов кодирующей последовательности варьирует от практически нулевых значений до 6. Средняя длина интронов также варьирует более чем на порядок [88]. Несмотря на большую вариабельность в частоте интронов на филогенетическом дереве, виды можно грубо разделить на богатые интронами (животные, растения и некоторые грибы) и бедные интронами — большинство одноклеточных эукариот (рис. 5).

Есть свидетельства того, что интроны типа II и сплайсосомные интроны имели общего предка. Механизм вырезания этой интронов типа II схож с таковым у сплайсосомных интронов. В частности, сплайсинг интронов типа II проходит через две реакции трансэтерификации с высвобождением интрона в форме лассо, при этом в сайте ветвления находится аденин [91]. Структурные исследования показали, что каталитический домен, осуществляющий первую реакцию трансэтерификации, схож со структурой, образуемой на донорном сайте сплайсинга с участием мяРНК U2 и U6 [92].

Эволюционная история интронов является предметом споров. Существует две крайние концепции — “ранние интроны” (introns early) и “поздние интроны” (introns late), а также ряд компромиссных точек зрения [88,90]. Согласно строгой версии концепции ранних интронов, интроны существовали у общего предка прокариот и эукариот, а затем были потеряны в прокариотах [93]. Согласно противоположной строгой концепции поздних интронов (introns late), прокариоты никогда не имели интронов, они возникли и распространились в процессе эукариотической эволюции [94]. В настоящее время, представители обоих лагерей смягчили позиции: сторонники концепции ранних интронов полагают, что лишь небольшая часть современных интронов произошла от общего предка прокариот и

эукариот [90], тогда как сторонники гипотезы поздних интронов считают, что эукариотические сплайсосомные интроны произошли от бактериальных самосплайсирующихся интронов типа II [95]. Консенсус также состоит в том, что достаточно большое количество интронов существовало в общем предке эукариотических организмов [90,96]. Большая вариабельность в числе интронов среди эукариот не содержит ярко выраженного паттерна: богатые и бедные интронами геномы перемешаны на филогенетическом дереве. В связи с этим стоит вопрос о том, насколько часто и на каких ветвях дерева интроны приобретались и терялись? Этот вопрос остаётся дискуссионным [90].

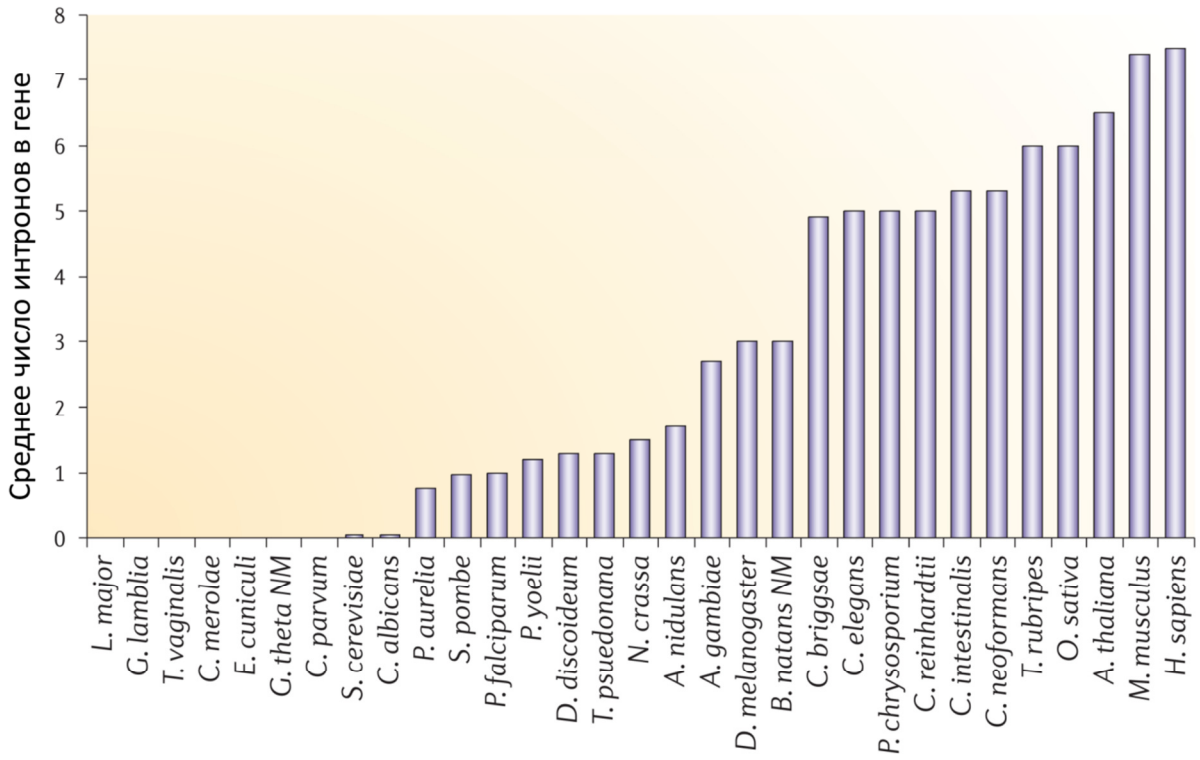


**Рисунок 5.** Среднее число интронов в гене у разных видов.
Полные названия видов в алфавитном порядке: *Anopheles gambiae; Arabidopsis thaliana; Aspergillus nidulans; Bigelowiella natans Nucleomorph; Caenorhabditis briggsae; Caenorhabditis elegans; Candida albicans;Chlamydomonas reinhardtii; Ciona intestinalis; Cryptococcus neoformans; Cryptosporidium parvum; Cyanidioschyzon merolae; Dictyostelium discoideum;Drosophila melanogaster; Encephalitozoon cuniculi; Giardia lamblia, Guillardia theta Nucleomorph; Homo sapiens; Leishmania major; Mus musculus; Neurospora crassa; Oryza sativa; Paramecium aurelia; Phanerochaete chrysosporium; Plasmodium falciparum; Plasmodium yoelii; Saccharomyces cerevisiae; Schizosaccharomyces pombe; Takifugu rubripes; Thalassiosira pseudonana; Trichomonas vaginalis* [90].

При исследовании выборки из 684 ортологичных кластеров генов из 8 полностью секвенированных геномов, оказалось, что 25% интронов человека находятся в тех же позициях (между гомологичными парами нуклеотидов в выравнивании), что и интроны *Arabidopsis thaliana*, а 40% интронных позиций *Schuzosaccharomices pombe* совпадают с таковыми из видов, не являющихся грибами. С другой стороны, 20-68% интронов специфичны для вида [97]. Эти наблюдения указывают на то, что в процессе эволюции эукариот происходили массовые потери и приобретения интронов. Очевидной интерпретацией совпадения позиций интронов является то, что они унаследованы от общего предка. Альтернативное объяснение - независимая вставка интронов в ортологичные позиции геномов – не может быть автоматически отвергнуто по следующим причинам. Неслучайность вставок интронов может иметь место в силу предпочтения определённых последовательностей (прото-сайтов сплайсинга, см. ниже) [98]. Поскольку в гене существует только несколько прото-сайтов сплайсинга, параллельная вставка интронов не кажется столь невероятной. Известны случаи параллельной вставки интронов в ортологичные места генов животных и растений [99,100]. Например, интрон в гене XDH наблюдается у двух близкородственных видов дрозофил, отсутствует у других представителей рода *Drosophila*. Более того, его нет у всех остальных видов животных, однако он обнаруживается в том же месте в геноме риса, *Arabidopsis thaliana* и хламидомонады [99]. Кажется очень маловероятным, что интрон мог независимо потеряться во многих линиях животных, поэтому более правдоподобной кажется параллельная вставка интронов в растения и предке двух близкородственных видов рода *Drosophila*.

В рамках гипотезы прото-сайтов сплайсинга, был смоделирован процесс вставок интронов в прото-сайты сплайсинга, учитывая наблюдаемые частоты этих сайтов и частоту интронов. Оказалось, что только 5-10% интронов, которые наблюдаются в ортологичных позициях у филогенетически далёких видов, могут быть объяснены параллельными вставками. Таким образом, большинство таких интронов унаследованы от общего предка [101].

Несколько групп построили реконструкции истории потерь и приобретений интронов на филогенетическом дереве эукариот [97,102–104]. Все они сходятся в том, что геном ближайший общего предка всех эукариот был богат интронами, а в некоторых линиях были массовые потери интронов (например, во многих грибах: *Schizosaccharomyces pombe*, *Neurospora crassa*, *Saccharomyces cerevisiae*), но детальные сценарии различаются. Отчасти эти различия обуславливаются использованными методами: так, в работе [97] использовался метод парсимонии Долло (Dollo parsimony), в других работах использовались вероятностные подходы [102–104]. Принцип парсимонии Долло исходит из того, что сложный признак (в данном случае интрон в конкретной позиции гена) возникает единожды в общем предке наиболее удалённых видов, а затем может один или несколько раз исчезать, при этом минимизируется число событий потерь на дереве. Даже если пренебречь возможностью независимых вставок интронов в ортологичные сайты геномов удалённых видов (что впрямую нарушает заложенный в основу метода принцип), возможное искажение может возникнуть, если в нескольких видах возникают параллельные потери, что выглядит как недавнее возникновение интрона на другой ветке (ветках), хотя он был унаследован от общего предка. Неясно, насколько часты параллельные потери интронов. От этого зависит применимость данного метода. Применение метода максимального правдоподобия к тому же набору исходных данных, что и в работе [97], привело к существенно отличным результатам [102]. В частности, метод максимального правдоподобия дал меньшее число приобретений интронов на конечных ветках и большее их число на внутренних. Стоит отметить, что вероятностный метод максимального правдоподобия также даёт различающиеся результаты [102,103], что говорит о том, что параметры этих моделей следует уточнить. Возможно, впрочем, эти различия обусловлены тем, что в работе [103] использовалось большее количество видов для сравнения. Наиболее полный анализ был сделан на 99 полных геномах и 245 кластерах ортологичных генов с использованием трех методов: Монте Карло с марковскими цепями (Markov Chain Monte Carlo), метода максимального правдоподобия и парсимонии Долло [104]. Все методы

предсказали богатого интронами общего предка всех эукариот. Методы Монте Карло с марковскими цепями и максимального правдоподобия дали очень похожие результаты, тогда как парсимония Долло недооценивала число предковых интронов по сравнению с этими двумя методами. В смоделированных примерах указанный перекос для парсимонии Долло сохранялся, что, вероятно, говорит о её ограниченной применимости для данной задачи.

Согласно реконструкции методом Монте Карло марковскими цепями наибольшее приобретение интронов в истории животных случилось при переходе к многоклеточности, после чего количество интронов плавно уменьшалось на линии позвоночных [104] (рис. 6). В целом можно сказать, что приобретения и потери интронов происходили в эволюции эукариот постоянно.

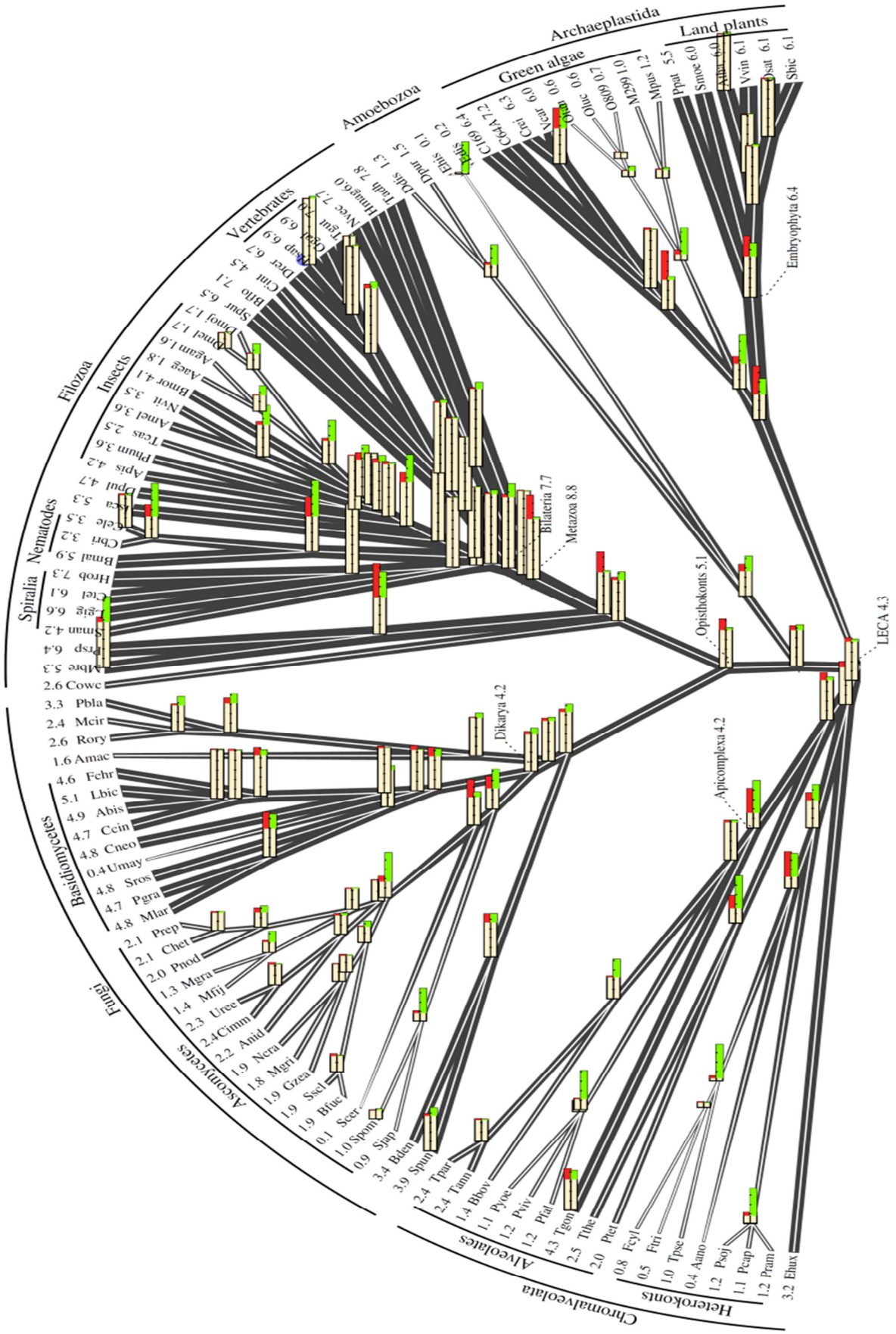


**Рисунок 6.** Реконструкция потерь и приобретений интронов в эволюции эукариот, а также частот интронов в предковых формах согласно [104].

Толщина ветвей пропорциональна частоте интронов (в расчёте на 1000 нт кодирующей последовательности), она же обозначена на каждой терминальной ветке (рядом с аббревиатурой вида) и у некоторых предков. Горизонтальные диаграммы показывают частоту интронов: верхняя полоска - в предковом узле, нижняя - в указанном узле. Потери и приобретения на линии (от предкового до текущего узла) показаны зелёным и красными цветами, соотнесено. Желтая часть полоски соответствует интронам, унаследованным от предкового до текущего узла [88].

Аббревиатуры видов: *Aureococcus anophagefferens (Aano), Aedesaegypti (Aaeg), Agaricusbisporus (Abis), Anopheles gambiae (Agam), Allomyces macrogynus (Amac), Apis mellifera (AmeI), Aspergillus nidulans (Anid),Acyrthosiphon pisum (Apis), Arabidopsis thaliana (Atha), Babesia bovis (Bbov), Batrachochytrium dendrobatidis (Bden), Branchiostoma floridae (Bflo),Botryotinia fuckeliana (Bfuc), Brugia malayi (Bmal), Bombyx mori (Bmor), Coccomyxa sp. C-169 (C169), Chlorella sp. NC64a (C64a), Caenorhabditisbriggsae (Cbri), Caenorhabditis elegans (Cele), Coprinopsis cinerea okayama (Ccin), Cochliobolus heterostrophus C5 (Chet), Coccidioides immitis(Cimm), Ciona intestinalis (Cint), Cryptococcus neoformans var. neoformans (Cneo), Chlamydomonas reinhardtii (Crei), Capitella teleta (Ctel),Capsaspora owczarzaki (Cowc), Dictyostelium discoideum (Ddis), Dictyostelium purpureum (Dpur), Drosophila melanogaster (Dmel), Drosophilamojavenis (Dmoj), Daphnia pulex (Dpul), Danio rerio (Drer), Entamoeba dispar (Edis), Entamoeba histolytica (Ehis), Emiliania huxleyi (Ehux),Fragilariopsis cylindrus (Fcyl), Phanerochaete chrysosporium (Fchr), Phaeodactylum tricornutum (Ftri), Gallus gallus (Ggal), Gibberella zeae (Gzea), Hydramagnipapillata (Hmag), Helobdella robusta (Hrob), Homo sapiens (Hsap), Ixodes scapularis (Isca), Laccaria bicolor (Lbic), Lottia gigantea (Lgig),Micromonas sp. RCC299 (M299), Monosiga brevicollis (Mbre), Mucor circinelloides (Mcir), Mycosphaerella fijiensis (Mfij), Mycosphaerella graminicola(Mgra), Magnaporthe grisea (Mgri), Melampsora laricis-populina (Mlar), Micromonas pusilla (Mpus), Neurospora crassa (Ncra), Nematostella vectensis(Nvec), Nasonia vitripennis (Nvit), Ostreococcus sp. RCC809 (O809), Ostreococcus lucimarinus (Oluc), Oryza sativa japonica (Osat), Ostreococcus taurii(Otau), Phytophthora capsici (Pcap), Plasmodium falciparum (Pfal), Puccinia graminis (Pgra), Pediculus humanus (Phum), Phaeosphaeria nodorum(Pnod), Physcomitrella patens subsp. patens (Ppat), Phytophthora ramorum (Pram), Pyrenophora tritici-repentis (Prep), Proterospongia sp. (Prsp),Phytophthora sojae (Psoj), Paramecium tetraurelia (Ptet), Plasmodium vivax (Pviv), Plasmodium yoelii yoelii (Pyoe), Rhizopus oryzae (Rory), Sorghumbicolor (Sbic), Saccharomyces cerevisiae (Scer), Schizosaccharomyces japonicus (Sjap), Schistosoma mansoni (Sman), Selaginella moellendorffii (Smoe),Schizosaccharomyces pombe (Spom), Spizellomyces punctatus (Spun), Strongylocentrotus purpuratus (Spur), Sporobolomyces roseus (Sros), Sclerotiniasclerotiorum (Sscl), Trichoplax adhaerens (Tadh), Theileria annulata (Tann), Tribolium castaneum (Tcas), Toxoplasma gondii (Tgon), Taenopygia guttata(Tgut), Theileria parvum (Tpar), Thalassiosira pseudonana (Tpse), Tetrahymena thermophila (Tthe), Ustilago maydis (Umay), Uncinocarpus reesii (Uree),Volvox carteri (Vcar), Vitis vinifera (Vvin).*

#### *2.2.1.2. Макроэволюционные изменения в сайтах сплайсинга*

Степень функциональной значимости каждой позиции сайта сплайсинга характеризуется силой позиции, а сила всего сайта или его части определяется как сумма соответствующих сил по позициям [50]. По сути, сила позиции отражает то, насколько часто нуклеотид, наблюдаемый в данной позиции, встречается среди всего набора сайтов сплайсинга (в соответствующей позиции). Чем чаще он встречается, тем выше сила позиции. Средняя сила набора сайтов сплайсинга эквивалентна информационному содержанию [105]. Как уже отмечалось, количество и частота интронов существенно варьируют среди эукариот, что свидетельствует об их масштабных приобретениях и потерях. Средняя сила сайтов сплайсинга также меняется между большими эукариотическими группами [88]. Так у голосумчатых грибов (гемиаскомицетов), а также у *Cryptosporidium parvum* донорные сайты существенно сильнее, чем у всех остальных клад эукариотических организмов (рис. 7). Эволюция акцепторных сайтов сплайсинга (точнее, полипиримидиновых трактов) обнаруживает иной паттерн: они слабы в большинстве грибов, средние по силе в растениях и некоторых одноклеточных эукариотах и монотонно усиливаются от нематоды к позвоночным [16]. Кроме того, наблюдается отрицательная корреляция между частотой интронов в геноме и силой донорного сайта сплайсинга [105] (рис. 8).

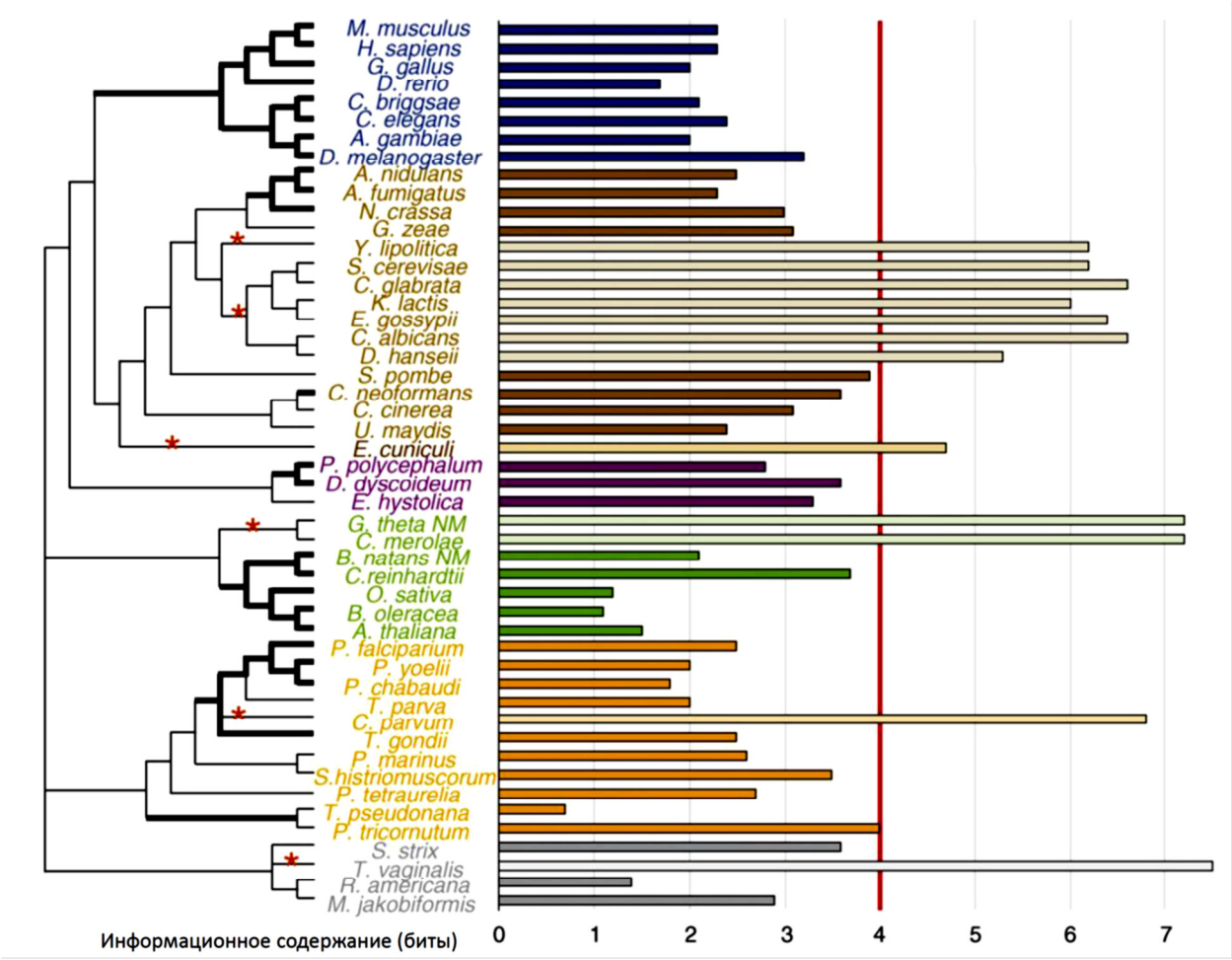


**Рисунок 7.** Сила донорного сайта сплайсинга на филогенетическом дереве видов.

Цвета соответствуют крупным таксономическим группам. Информационное содержание (в битах) донорных сайтов сплайсинга показано на диаграмме справа. Красная черта — условная граница, отделяющая с сильные донорные сайты от всех остальных. Линии, в которых произошло резкое усиление донорного сайта, помечены звёздочками. Бледным оттенком на диаграмме показаны те виды, у которых экстремально мало интронов (< 0.01 интронов на ген). Толстыми линиями на дереве отмечены виды, у которых есть альтернативный сплайсинг [105].

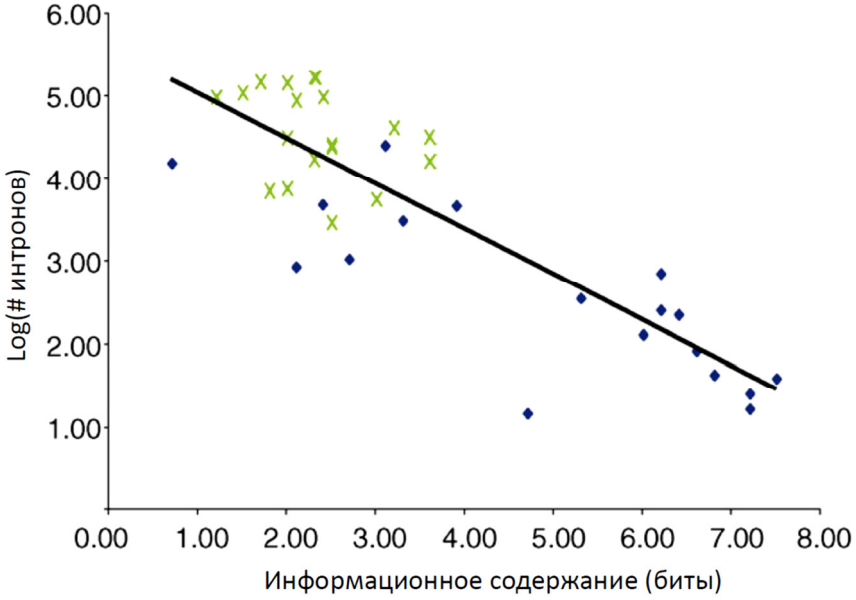


**Рисунок 8.** Зависимость между силой донорного сайта (информационное содержание) и числом интронов в геноме (в логарифмической шкале).

Виды, имеющие альтернативный сплайсинг, показаны крестиками, не имеющие — ромбиками [105].

Компоненты сплайсосомы высококонсервативны среди эукариот. Участок U1 мяРНК, связывающийся с донорным сайтом сплайсинга, одинаков у всех эукариот, за исключением нуклеотида, комплементарного позиции -3 донорного сайта, который отличается у некоторых грибов и растений [16,24] (рис. 9). При этом этом не прослеживается какой-либо корреляции между изменениями в U1 и донорными сайтами сплайсинга в соответствующем геноме. РНК-распознающие домены RRM1 и RRM2 белка U2AF65 (который связывается с полипиримидиновым трактом акцепторного сайта сплайсинга) почти не меняются у разных позвоночных и более чем на 80% идентичны между позвоночными и насекомыми [16]. Разница в силе сайтов в пределах позвоночных не может быть объяснена изменениями в сплайсосоме и соответствующей коэволюцией сайтов. Мы в настоящей работе изучали именно такие виды.

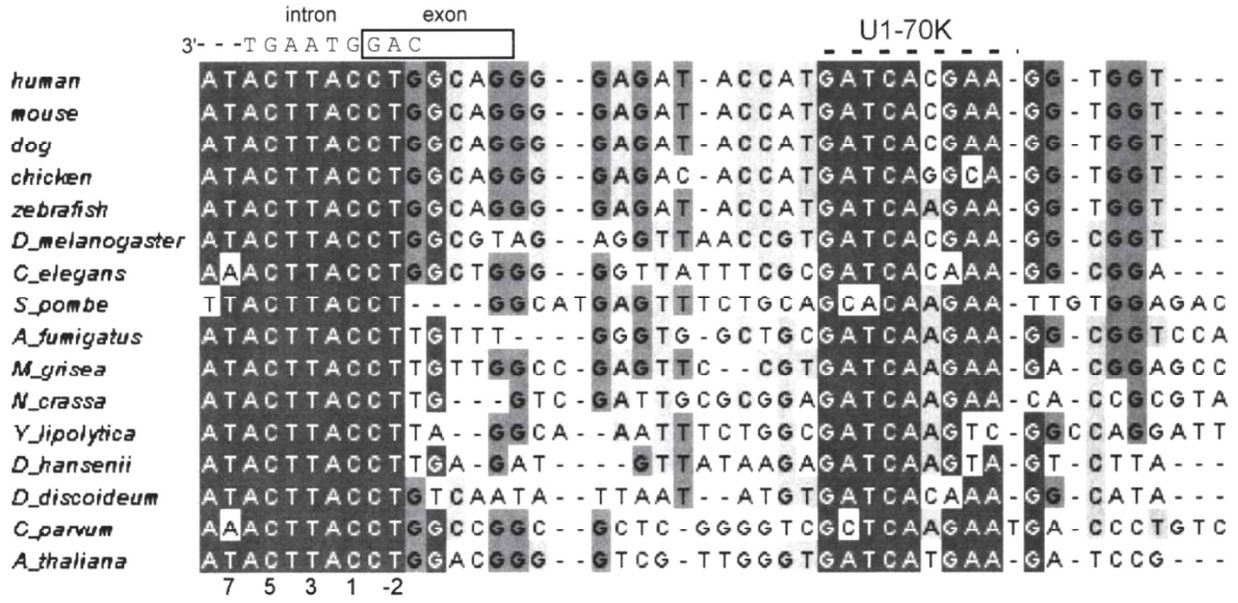


**Рисунок 9.** Множественное выравнивание 5’-концевого участка U1 мяРНК животных, растений и грибов.

Сверху указана консенсусная последовательность донорного сайта сплайсинга, взаимодействующего с U1 [16].

Если рассматривать не только позвоночных, а более далёкие виды, то прослеживается коэволюция сайтов сплайсинга и взаимодействующих с ней компонентов сплайсосомы. Для донорных сайтов коэволюция весьма слаба и найдена только для позиции -3 сайта сплайсинга и комплементарного нуклеотида в U1. Что же касается акцепторных сайтов, а также сайтов ветвления, то здесь коэволюционные изменения более обширны. Schwarz и соавт. исследовали замены, произошедшие в белках U2AF$^{35}$ (который связывается с AG), U2AF$^{65}$ (взаимодействует с полипиримидиновым трактом) и SF1 (связывается с сайтом ветвления) [16]. Исследованные организмы можно разделить на три группы: (1) похожие на человека (15 из 22 видов); (2) похожие на *S. cerevisiae* (4 вида) и (3) похожие на гемиаскомицетов (3 вида). В группу похожих на человека входят многоклеточные животные, растения, грибы (кроме гемиаскомицетов), а также простейшее *Dictyostelium discoideum.* В этой группе имеются ортологи всех трёх белков, поэтому распознавание в целом идёт по той же схеме, что и у человека. К группе похожих на *S. cerevisiae* относятся *S. cerevisiae*, *Candida glabrata*,

*Eremothecium gossypii* и *Kluyveromyces lactis*. В этих видах произошло неортологичное замещение белка U2AF$^{65}$ на белок MUD2 (так называется функциональный аналог U2AF$^{65}$ у *S. cerevisiae*), структура полипиримидинового тракта и механизм его распознавания похож на таковой у *S. cerevisiae*. Последняя группа включает гемиаскомицетов *Yarrowia lipolytica* и *Debaryomyces hansenii*, а также простейшее *Cryptosporidium parvum*. В этих организмах сохранились ортологи U2AF$^{35}$, U2AF$^{65}$ и SF1, однако U2AF$^{65}$ из-за произошедших замен утратил способность связываться с полипиримидиновым трактом. Способность к белок-белковым взаимодействиям между U2AF$^{65}$ и U2AF$^{35}$, а также между U2AF$^{65}$ и SF1 сохранилась, что говорит о том, что U2AF$^{65}$ выполняет роль медиатора, опосредующего взаимодействие U2AF$^{35}$ и SF1. Соответственно, у *Y. lipolytica* и *D. hansenii* отсутствует полипиримидиновый тракт. У *C. parvum* найдено множество мутаций во всех трех белках, поэтому механизм распознавания акцепторного сайта, вероятно, сильно отличается от человека, но детали не ясны.

Рассмотрим более подробно группу видов, похожих на человека, и проследим эволюцию U2AF$^{65}$. U2AF$^{65}$ содержит РНК-распознающие домены RRM1 и RRM2, которые связываются с полипиримидиновым трактом, а также мотив UHM, посредством которого U2AF$^{65}$ взаимодействует с SF1. Выравнивание RRM1 и RRM2 из разных видов обнаруживает большое количество произошедших замен, однако вопрос состоит в том, отражают ли эти замены исключительно процесс дивергенции между видами или там имеются следы коэволюции с соответствующими полипиримидиновыми трактами?

Рассмотрим позиции в белке, для которых ранее была показана ключевая роль во взаимодействии с полипиримидиновым трактом у человека [106]. Такие характеристики этих аминокислот, как полярность, заряд и ароматичность важны для описания взаимодействия с полипиримидиновым трактом. Поэтому резкие различия этих характеристик между видами должны приводить к изменению во взаимодействии с полипиримидиновым трактом. Среди многоклеточных животных такие различия практически отсутствуют, в то же время большое их

количество наблюдается между многоклеточными и грибами (не гемиаскомицетами). Соответственно, полипиримидиновые тракты у этих грибов значительно слабее, чем у многоклеточных. При этом довольно мало таких различий произошло между многоклеточными животными и *D. discoideum*, а также *A. thaliana*, несмотря на то, что указанные виды находятся на более далёком филогенетическом расстоянии от многоклеточных, чем грибы. Это согласуется с тем, что полипиримидиновые тракты у этих двух организмов сильнее, чем у не голосумчатых грибов. Таким образом, полипиримидиновый тракт и белок U2AF$^{65}$ коэволюционировали, что существенно на больших филогенетических расстояниях, и разница в силе полипиримидиновых трактов объясняется в первую очередь изменениями в U2AF$^{65}$. Однако в пределах позвоночных изменений произошло очень мало, и фактор коэволюции можно не учитывать.

Что касается сайта ветвления, то он отличается по силе в разных видах: самые сильные сайты ветвления встречаются у голосумчатых грибов, средние по силе у остальных грибов, растений и некоторых простейших, и самые слабые — у многоклеточных животных. Это в общих чертах напоминает картину, характерную для донорного сайта. Последовательность сайта ветвления (7 нт) не проявляет признаков коэволюции с U2 мяРНК, однако у некоторых видов (*N. crassa*, *Magnaporthe grisea*, *Aspergillus fumigatus* и в меньшей мере у *D. discoideum*, *D. hansenii*) ~6 нуклеотидов перед сайтом ветвления изменяются коррелируют с соответствующими нуклеотидами в U2, что говорит также о том, что у этих видов U2 связывается с более длинной последовательностью [16].

Важно заметить, что виды с небольшим эффективным размером популяции (например, позвоночные) чаще имеют альтернативный сплайсинг, содержат большое количество интронов, а средняя сила донорных сайтов, а также сайтов ветвления у них ниже, чем у видов с высоким эффективным размером популяции (например, различные одноклеточные эукариоты) [16,105]. Популяционно-генетическое объяснение этого феномена приведено ниже. Филогенетические реконструкции говорят, что общий предок всех эукариотических организмов имел

наряду с большим количеством интронов относительно слабые донорные сайты сплайсинга и сайты ветвления [16].

#### *2.2.1.3. Гипотеза о миграции сигнала*

Сила сайтов сплайсинга зависит от эволюционного возраста соответствующего экзона. Sverdlov и соавт. разделили интроны на две группы: “древние” интроны, которые консервативны в двух или более основных эукариотических группах, и “новые”, присутствующие только в одной из групп. Оказалось, что донорные сайты сплайсинга древних интронов имеют более сильную экзонную и более слабую интронную части по сравнению с новыми интронами. В акцепторных сайтах сплайсинга древних интронов экзонная часть также сильнее, чем у новых интронов, однако интронные части акцепторных сайтов древних и новых интронов статистически значимо не отличаются [107]. Формально, это наблюдение находится в согласии с гипотезой поздних интронов [88] (ближайший общий предок всех эукариотических организмов был беден интронами, которые вставлялись в гены в ходе последующей эволюции эукариот) и концепцией прото-сайтов сплайсинга. Эта концепция предполагает, что интроны вставлялись не в случайные места генов, а предпочитали определённые последовательности. Экзонные части донорного и акцепторного сайтов сплайсинга представляют собой реликты последовательностей, в которые вставлялись интроны [108,109]. Чтобы вставленный интрон эффективно сплайсировался, получившиеся сайты сплайсинга должны быть достаточно сильными. Так как вставленный интрон, вероятнее всего, представлял собой случайную последовательность, то и интронные части сайтов сплайсинга не могут быть заведомо сильными. Экзонные же части донорного и акцепторного сайтов, происходящие из прото-сайта, должны наилучшим образом соответствовать консенсусу. В ходе последующей эволюции сила экзонных частей уменьшалась, а интронных возрастала, т.е. происходила миграция сигнала из экзонной части в интронную.

Миграция сигнала, вероятно, была вызвана с одной стороны отбором на консенсусную последовательность (как в экзонной, так и в интронной частях сайта),

а с другой стороны — давлением отбора на кодирование белка (в экзонной части). Как следствие, интронная часть усиливалась, а в экзонной части возникал конфликт между отбором на эффективность сплайсинга (стремящимся приблизить последовательность к консенсусу) и отбором на определенную белковую последовательность [98]. В сайтах сплайсинга древних интронов этот конфликт разрешается следующим образом: они имеют сильную интронную часть (чтобы сплайсинг происходил эффективно) и слабую экзонную часть (вызванную интерференцией с кодированием определённой аминокислотной последовательности). Sverdlov с соавт. задались целью восстановить предполагаемые прото-сайты сплайсинга. Для этого они проанализировали последовательности вокруг мест соединения экзонов (теней интронов), которые пересекаются с триплетами, кодирующими аминокислоту, инвариантную у всех эукариот [110]. Такие аминокислоты поддерживаются отрицательным отбором, и, вероятнее всего, были унаследованы от общего предка всех эукариотических организмов; следовательно, здесь отсутствует конфликт между кодированием белка и связыванием со сплайсосомой. Стало быть, эти последовательности представляют собой “замороженные” прото-сайты сплайсинга. Было показано, что прото-сайты сплайсинга имели последовательность (A/C)AG||GT (положение интрона показано двумя вертикальными чертами), что совпадает с консенсусными последовательностями экзонных частей современных сайтов сплайсинга. Этот результат подразумевает, что прото-сайты сплайсинга были расположены не случайно в кодирующей последовательности, и что экзонные части сайтов сплайсинга были изначально сильными [98].

Описанный эволюционный механизм миграции сигнала может оказаться рабочим не только для интронов, вставленных на ранних стадиях эукариотической эволюции, но и для более поздних событий. Так, массовое приобретение интронов могло произойти при переходе к многоклеточности [104], а также либо при появлении двусторонне-симметричных животных [102], либо в линии хордовых [97] (результаты реконструкции варьируют в зависимости от применяемого метода). Так или иначе, если происходила вставка интронов, процесс миграции

сигнала можно попытаться непосредственно отследить даже на уровне относительно близких видов (таких как различные млекопитающие). Рассмотрение этого процесса на более далёких видах, по-видимому, не представляется возможным из-за проблемы множественных замен и, следовательно, невозможности адекватной реконструкции предковых последовательностей сайтов сплайсинга. В настоящей работе мы протестировали гипотезу миграции сигнала на ортологичных сайтах сплайсинга человека, мыши и собаки.

## 2.2.2. Микроэволюция сплайсинга

### *2.2.2.1. Микроэволюционные причины эволюции интронов и сайтов сплайсинга*

Почему бедные интронами виды представлены главным образом среди одноклеточных эукариот? Точного ответа на этот вопрос нет, однако вероятной причиной является эффективность отбора против интронов. Интроны в целом рассматриваются как слабовредные приобретения, поскольку они не кодируют белок, а на их репликацию и сплайсинг требуются дополнительные энергетические ресурсы. Поэтому против них должен действать слабый отрицательный отбор [111]. На эффективность отбора влияет эффективная численность популяции, а на его скорость — время жизни поколения: чем выше эффективная численность и время жизни поколения, тем эффективнее и быстрее из популяции удаляются слабовредные аллели. Приблизительные оценки эффективной численности популяций ($N_e$) качественно согласуются с этим объяснением: $N_e$ ~ $10^7$ – $10^8$ у одноклеточных эукариот, $N_e$ ~ $10^5$ – $10^6$ у беспозвоночных и ~ $10^4$ – $10^5$ у позвоночных [112,113]. Время жизни поколения у высших эукариот также обычно выше, чем у низших. Эффективная численность популяции варьирует вдоль генома в зависимости от уровня рекомбинации: чем ниже уровень рекомбинации, тем она меньше (эффект Хилла-Робертсона) [113]. В соответствии с теорией, частота и средний размер интронов выше в областях генома с низкой частотой рекомбинации (показано на геномах позвоночных и *D. melanogaster*), так как там отбор менее эффективен [114–116].

Однако частота интронов варьирует значительно даже среди многоклеточных эукариот [90], что может говорить как об изменении эффективной численности популяции в прошлом, так и о некоторых селективных преимуществах если не всех, то некоторых интронов (например, связанных с альтернативным сплайсингом). Тем не менее, большая часть исследователей сходится на том, что общий предок всех эукариот был богат интронами, и, видимо, общим трендом является скорее уменьшение их количества в некоторых линиях с высоким эффективным размером популяции [88]. Что же касается массового приобретения интронов, то такие события случались преимущественно на внутренних ветках в эпохи происхождения крупных групп, таких как многоклеточные, что могло совпадать с бутылочными горлышками, которые типичны для таких эпох [117]. Сильные донорные сайты сплайсинга и сильные сайты ветвления встречаются у одноклеточных эукариот, но не у многоклеточных животных или растений [16], что укладывается в популяционно-генетическое объяснение.

Стоит отметить, однако, что не все одноклеточные эукариоты имеют низкую частоту интронов и сильные сайты донорные сайты сплайсинга (рис. 7). Это показывает, что кроме эффективного размера популяции существуют и другие факторы, влияющие на эти характеристики. Однако, справедливым остаётся тот факт, что виды с экстремально сильными сайтами и малым количеством интронов встречаются исключительно среди одноклеточных. Также очевидно, что эволюция акцепторных сайтов сплайсинга не укладываются в столь простую схему: самые сильные акцепторные сайты сплайсинга обнаруживаются у позвоночных и они слабы в большинстве грибов, в том числе в голосумчатых. Однако, в связи с тем, что компоненты сплайсосомы, ответственные за связывание с акцепторным сайтом, существенно менялись в ходе эволюции [16], это могло приводить к изменению адаптивного ландшафта для акцепторных сайтов вплоть до того, что в некоторых видах полипиримидиновый тракт вовсе исчез (см. выше). Напротив, компоненты сплайсосомы, ответственные за связывание с донорным сайтом, показывают чрезвычайную эволюционную консервативность, и поэтому они

служат более адекватной моделью для понимания микроэволюционных причин эволюции сайтов сплайсинга.

Что касается альтернативного сплайсинга, то он чаще встречается у видов с высокой частотой интронов [105] и, соответственно, низким эффективным размером популяции (рис. 8). Вероятно, его появление является следствием того, что небольшой размер популяции не позволяет эффективно отсеивать ошибки сплайсинга. Некоторые из этих ошибок оказались полезными, были подхвачены отбором, что и привело к функциональному альтернативному сплайсингу.

#### *2.2.2.2. Микроэволюция сайтов сплайсинга*

Индивидуальные отличия между сайтами сплайсинга в рамках одного организма, а также различия между сайтами сплайсинга в ближайших видах могут объясняться различными причинами, например особенностями регуляции, наличием энхансеров и сайленсеров сплайсинга, отбором на другие функции, а также действием дрейфа.

Ключевой вопрос состоит в том, почему в геномах высших эукариот, таких как *Homo sapiens* и *Drosophila melanogasater,* многие сайты сплайсинга отклоняются от консенсуса, т.е. в сайтах сплайсинга существует множество неконсенсусных аллелей. Этому может быть несколько теоретических объяснений. Во-первых, эти аллели могу постоянно создаваться за счёт мутаций, они вредны, но скорость мутагенеза такова, что геном не успевает от них избавляться. Однако это объяснение можно сразу отмести, потому что современные оценки скорости точечного мутагенеза у человека и *Drosophila* противоречат этому [118,119]. Во-вторых, неконсенсусные аллели могут быть действительно полезными. Например, они важны для поддержания какой-либо иной функции, как связанной, так и не связанной со сплайсингом (пересекаться с сайтами посадки транкрипционных факторов и т.п.). Было показано, что альтернативные экзоны в среднем имеют более слабые сайты сплайсинга, чем конститутивные [49–53]. В некоторых случаях, отбор может предпочитать слабые сайты сильным [53]. Некоторые экспериментальные исследования говорят о том, что усиление слабых сайтов

альтернативных экзонов приводит к разрушению регуляции сплайсинга [14,120]. В-третьих, неконсенсусные аллели могут оказаться нейтральными, например потому, что сила сайта сплайсинга не столь важна, когда рядом находится эффективный энхансер сплайсинга. В-четвёртых, неконсенсусные аллели могут оказаться действительно вредными (точнее слабовредными), но в силу относительно небольшого размера популяции, отбор не в состоянии полностью их элиминировать, поэтому многие из них возникнув, фиксируются благодаря дрейфу [121].

Irimia с соавт. исследовали замены в донорных сайтах сплайсинга и получили кажущийся парадоксальным результат: консенсусные нуклеотиды находятся под отрицательным отбором, в то время как неконсенсусные нуклеотиды эволюционируют нейтрально [122]. Чтобы прояснить эту ситуацию, мы исследовали частоту замен, происходящих между консенсусными и неконсенсусными нуклеотидами как в донорных, так и в акцепторных сайтах сплайсинга и сравнили её со скоростью замен в нейтральных участках генома.

## 2.3. Отбор и эпистаз

### 2.3.1. Положительный и отрицательный отбор

#### *2.3.1.1. Что такое положительный и отрицательный отбор?*

С популяционно-генетической точки зрения, эволюция представляет собой процесс (1) появления новых аллелей за счёт мутаций и (2) изменения аллельных частот с течением времени. На динамику аллельных частот в популяции влияет множество факторов, важнейшие из которых: отбор, генетический дрейф, наличие/отсутствие полового размножения, (не)случайность скрещивания, неравномерный поток аллелей (изоляция, интрогрессия и т.п.) [123].

Приспособленность особи ($f$) в популяции можно определить как количество потомков, которое эта особь оставила. Это определение имеет популяционно-

генетический смысл, т.к. именно количество потомков определяет частоту сооответствующих аллелей в следующем поколении. Микроэволюцию же можно рассматривать как изменение частот аллелей в популяции (вплоть до полного вытеснения одних аллелей другими). Приспособленность (как частный случай фенотипа) определяется как генотипом особи, так и влиянием среды. Мы будем рассматривать только генетическую компоненту.

Пусть в гаплоидной популяции имеется два аллеля *A* и *a*. Особи, несущие гены *A* и *a,* могут отличаться по другим аллелям. Приспособленность аллеля *A*, *w(A)*, определяется как средняя приспособленность особей несущих, аллель *A* (т.е. усреднённая по всем генетическим фонам) [124].

Если в гаплоидной популяции аллели *A* и *a* имеют приспособленности *w(A)* и *w(a)* соответственно, причем *w(A) > w(a)*, то коэффициент отбора определяется как $s = \frac{w(A)-w(a)}{w(A)}$. Коэффициент отбора характеризует относительное увеличение приспособленности при замене *a→A*. Из определения следует, что *0 ≤ s ≤ 1*. Если *s = 0*, то селективное преимущество A над a отсутствует: $w(A) = w(a)$. Если $w(A) \gg w(a)$, то $s \approx 1$. Для диплоидной популяции коэффициент отбора определяется аналогично: $s = \frac{w(AA)-w(aa)}{w(AA)}$ [124,125].

Рассмотрим действие отбора и дрейфа. На изменение аллельных частот влияет то, насколько данный аллель вреден или полезен для организма (более точно насколько он влияет на приспособленность). Однако большинство аллелей нейтральные. Отбор определяет направленное изменение аллельных частот (вредные аллели понижают свою частоту, полезные — повышают), в то время как дрейф представляет собой стохастическое изменения частот аллелей в популяции. Если аллель нейтральный, то его судьба определяется только дрейфом [126]. Если аллель вредный или полезный, то его судьба определяется как дрейфом, так и отбором [121]. Одним из ключевых параметров, влияющих на соотношение дрейфа и отбора, является размер эффективный размер популяции [113]. Чем

меньше популяция, тем сильнее влияние дрейфа (и, соответственно, менее эффективен отбор).

Отрицательным (очищающим) отбором называется отбор против новых вредных аллелей в популяции. Он приводит к замедлению эволюции и сохранению признака, на который он действует. Например, консервативные области генома возникают в результате действия отрицательного отбора.

Положительным (движущим) отбором называется отбор, направленный на увеличение частоты (и фиксацию) полезных аллелей. Фиксация полезных аллелей происходит существенно быстрее, чем фиксация нейтральных аллелей за счет дрейфа [123]. На этом факте основан ряд тестов, позволяющих выявить действие отбора.

#### *2.3.1.2.Тестирование отбора на молекулярном уровне*

Сравнение геномов родственных видов, а также особей внутри популяции одного вида позволяет обнаружить отличия между геномами, что даёт возможность выявить эволюционные причины, лежащие в основе этих различий.

Существует ряд подходов, позволяющих выявить действие положительного или отрицательного отбора на те или иные последовательности генома. Практически все они основаны на сравнении процессов, происходящих в изучаемом участке генома с нейтральным контролем, т.е. таким участком генома, на который не действует отбор, из-за чего его эволюция определяется преимущественно генетическим дрейфом. Наибольший интерес представляет детекция положительного отбора, т.к. именно он ответственен за появление адаптивных изменений фенотипа. Глобально, тесты на наличие положительного отбора построены на двух биологических соображениях: (1) фиксация полезного аллеля происходит гораздо быстрее, чем фиксация нейтрального аллеля; (2) полезный аллель сцеплен с близлежайшими аллелями, которые фиксируются вместе с ним (эффект автостопа, hitchhiking), что формирует длинный гаплотип с пониженной вариабельностью (эффект выметания отбором, selective sweep) [123].

Отбор мог происходить в разные периоды времени; соответсвенно методы детекции недавно произошедшего положительного отбора отличаются от поиска отбора, произошедшего довольно давно [127]. Грубо методы детекции отбора можно разделить микроэволюционные (сравнение особей в рамках вида) и макроэволюционные (межвидовые сравнения).

Кратко рассмотрим идеи, лежащие в основе этих методов (в скобках указаны временные границы применимости методов для линии человека, согласно [127]).

**Макроэволюционные методы** (миллионы – десятки миллионов лет назад). Идея: большое количество замен, потенциально меняющих функцию по сравнению с нейтральными заменами. Тест $D_n/D_s$ (также известный как $K_a/K_s$, $K_n/K_s$ ω) предназначен для тестирования отбора в последовательностях, кодирующих белки. Он смотрит на соотношение числа несинонимичных замен $D_n$ (т.е. тех, которые приводят к замене аминокислоты) к числу синонимичных замен $D_s$ (не приводящих к изменению аминокислоты). Синонимичные замены рассматриваются как нейтральные в первом приближении. $D_n/D_s > 1$ говорит о наличии положительного отбора. Этот тест хорошо работает, если в тестируемом регионе генома присутствуют замены полезные и нейтральные. Присутствие вредных замен осложняет применение этого теста, т.к. они занижают $D_n$, что приводит к недопредсказанию положительного отбора [123]. Для того чтобы выявить положительный отбор в некоторых сайтах на фоне отрицательного отбора (в других сайтах рассматриваемого участка гена), применяется тест Макдональда-Крейтмана (McDonald-Kreitman test) [128]. В основе этого теста лежит следующее соображение: (сильно)вредные мутации быстро выметаются из популяции, поэтому они не будут присутствовать в популяции даже на уровне полиморфизмов. Значит, число несинонимичных полиморфизмов ($P_n$) будет меньше числа синонимичных полиморфизмов ($P_s$) ровно на число вредных полиморфизмов, отброшенных отбором. Поэтому можно посчитать отношение числа синонимичных и несинонимичных полиморфизмов в популяции ($P_n/P_s$) и сравнить его с $D_n/D_s$. Если $D_n/D_s > P_n/P_s$, имеет место положительный отбор. Однако

даже тест Макдональда-Крейтмана оказывается недостаточно чувствительным, вероятно из-за присутствия слабовредных аллелей, которые практически не влияют на $P_n$, но занижают $D_n$ [123].

**Микроэволюционные методы** (< 250 тысяч лет назад). Рассмотрим различные методы по мере уменьшения временных интервалов, на которых их можно применять.

Уменьшение генетического разнообразия вследствие выметания отбором (< 250 тысяч лет назад). При отборе на определённый аллель в некотором локусе генома, сцепленные аллели также повышают свою частоту, вытесняя все остальные. Таким образом, после фиксации полезного аллеля, вокруг соответствующего локуса формируется участок, на котором отсутствуют полиморфизмы. Постепенно этот участок насыщается новыми полиморфизмами из-за мутаций. Однако этот процесс медленный, поэтому вокруг локуса, бывшего под отбором, возникает зона с малым количеством полиморфизмов и со смещенным спектром аллельных частот (преобладают низкочастотные полиморфизмы). Длина участка с низкой вариабельностью зависит от силы отбора: чем он сильнее, тем быстрее происходит фиксация, тем длиннее получающийся участок [127]. Ряд тестов, таких как Ewens-Watterson test, Tajima's D и Fu and Li's D*, детектируют указанные подписи отбора [129]. Осложнением при детекции этой подписи отбора является популяционная история: резкое увеличение размера популяции приводит к увеличению доли редких аллелей.

Высокая частота производных полиморфизмов (< 80 тысяч лет назад). Если производный аллель находится под положительным отбором, но не достиг фиксации, в популяции будет наблюдаться повышенная частота этого и сцепленных с ним производных аллелей (по сравнению с нейтральными аллелями). В отличие от ситуации с пониженным генетическим разнообразием, эта ситуация наблюдается непродолжительный период времени — до фиксации аллеля. Однако длина этого периода зависит от силы отбора, действующего на аллель: если она относительно невелика, то положительный отбор можно

детектировать с большей вероятностью. Типичным тестом для выявления высокой частоты производных полиморфизмов служит Fay and Wu's H. Этот тест почти не чувствителен к увеличению размера популяции, однако чувствителен к разделению популяций [127].

Межпопуляционные различия частот аллелей (< 75 тысяч лет назад). Если две популяции оказываются разных условия среды и хотя бы частично репродуктивно изолированы, то они накапливают полезные аллели, повышающие среднюю приспособленность популяции к своей среде. Сравнение общей дисперсии частот аллелей с соответствующими дисперсиями внутри популяций позволяет выявить аллели под отбором. Типичной мерой для этого служит индекс фиксации Фишера (Fixation index, $F_{ST}$). Такие тесты как Lewontin-Krakauer test и Locus-specific branch length test позволяют делать межпопуляционные сравнения. Все методы из этой группы чувствительны к популяционной истории (уменьшение размера популяции усиливает дрейф, поэтому частоты между популяциями могут отличаться по случайным причинам) [127].

Длинные гаплотипы (< 30 тысяч лет назад). После фиксации полезный аллель некоторое время оказывается сцепленным с окружающими аллелями, формируя длинный гаплотип. С течением времени рекомбинация разрушает неравновесие по сцеплению, тем самым укорачивая этот гаплотип. Существует ряд тестов, позволяющий детектировать неравновесие по сцеплению, вызванное отбором на определённый аллель: Long-range haplotype test, Linkage disequilibrium decay, Identity-by-descent analysis и другие [129]. Этот подход способен детектировать гаплотипы, как находящиеся в процессе фиксации, так и недавно зафиксировавшиеся [127], причём можно анализировать относительно узкие геномные участки вплоть до одного гена. Ограничением этого подхода является то, что длинные гаплотипы распадаются довольно быстро. Например, в человеческой популяции в типичной хромосоме происходит более одного кроссовера на 100 тыс. нт за ~30000 лет, что делает гаплотипы слишком короткими, чтобы их можно было детектировать [127].

Мы в настоящей работе изучали отбор, действующий на сайты сплайсинга, используя аналог теста $D_n/D_s$, а также тестировали повышенную частоту производных аллелей.

#### *2.3.1.3. Слабый отбор. Слабовредные и слабополезные мутации*

Когда четыре нуклеотида (аллеля), которые могут находиться в определенном месте генома (локусе), имеют почти одинаковую приспособленность, отбор не действует на этот локус и его судьба определяется дрейфом (для диплоидной популяции: $s << 1/(4N_e)$). Напротив, когда один аллель имеет существенно большую приспособленность, чем все другие, этот аллель будет зафиксирован в популяции (вследствие положительного отбора), после чего отрицательный отбор будет предотвращать распространение альтернативных аллелей ($s >> 1/(4N_e)$) [124]. Однако, существует и промежуточная ситуация, когда приспособленности аллелей не идентичны, однако их различие не велико ($s \sim 1/(4N_e)$). В таком случае отбор не способен эффективно вычищать вредные мутации (они в данном случае называются слабовредными). Если существуют слабовредные мутации, то теоретически обратные мутации должны быть слабополезными [130]. Тогда, если скорость мутаций относительно невелика ($\mu << 1/N_e$), согласно теории в локусе происходят случайные фиксации слабовредных и слабополезных аллелей, даже если приспособленности этих аллелей не меняются со временем [121,131]. Если зафиксировался слабополезный аллель, мы имеем дело с слабым отрицательным отбором против альтернативных аллелей, когда фиксируется слабовредный аллель, мы имеем дело со слабым положительным отбором [130,132–134]. Таким образом, в дополнение к стандартной дарвиновской ситуации, когда положительный отбор вызывается изменением ландшафта приспособленности, слабый положительный отбор может присутствовать и при неизменном ландшафте из-за того, что дрейф постоянно фиксирует новые слабовредные мутации [121].

О существовании слабовредных мутаций говорит ряд свидетельств. Во-первых, полиморфизмы в ряде локусов генома сегрегируют с меньшей скоростью, чем

нейтральные полиморфизмы. Это касается несинонимичных полиморфизмов в белок-кодирующих генах [135] и полиморфизмов в консервативных некодирующих последовательностях [136]. Во-вторых, небольшой эффективный размер популяции способствует высокой скорости несинонимичных замен [137,138], а также замен в консервативных межгенных областях [139]. В-третьих, участки генома с локально низким эффективным размером популяции (половые хромосомы) также показывают также пониженные значения $D_n/D_s$ [140,141]. Наконец, в митохондриальных геномах некоторых видов $P_n/P_s$ превышает $D_n/D_s$, что совместимо с гипотезой о наличии большого числа слабовредных полиморфизмов [142].

Слабополезные мутации сложнее экспериментально наблюдать потому, что их эффект маскируется отрицательным отбором. Charlesworth с соавт. показали наличие слабополезных мутаций в популяциях, численность которых недавно резко увеличилась (хамелеоны, ящерицы рода *Anolis*, птицы рода *Monarcha*, растения рода *Coraria* и листовые лягушки рода *Eleutherodactylus*) [130].

Для того, чтобы напрямую показать наличие слабого положительного и слабого отрицательного отбора, необходимо наличие большого ансамбля локусов со схожим ландшафтом приспособленности. Такими локусами являются соответствующие позиции сайтов сплайсинга. Консервативность сайтов сплайсинга между видами [65,87], а также пониженный уровень внутривидового полиморфизма [143] свидетельствуют об отборе на эти последовательности. За исключением очень малого количества нетипичных сайтов сплайсинга (U12, U2 GC–AG и др., [13]), большинство сайтов сплайсинга распознаётся одной и той же сплайсосомой (AG-GT U2). Как у донорных, так и у акцепторных сайтов сплайсинга существует выраженный консенсус (рис. 2). Вышесказанное позволяет предположить, что соответствующие позиции в разных сайтах сплайсинга (локусы) имеют приблизительно одинаковый ландшафт приспособленности, где консенсусные нуклеотиды предпочитаются, а неконсенсусные – элиминируются отбором. Нашей задачей было подтвердить наличие слабовредных и

слабополезных мутаций и оценить силу отбора, действующего на соответствующие нуклеотиды.

## 2.3.2. Эпистаз

### *2.3.2.1. Что такое эпистаз? Виды эпистаза*

Значение термина “эпистаз” изменялось по мере развития науки. Впервые термин epistasis ввёл около 100 лет назад William Bateson [144], который заметил, что при некоторых скрещиваниях проявляются не все возможные фенотипические классы, и что некоторые комбинации аллелей приводят к проявлению новых фенотипов. R.A. Fisher впоследствии использовал производный термин — “epistasy” — для обозначения любых статистических отклонений от аддитивных эффектов двух аллелей на фенотип [145]. Дальнейшее развитие клеточной и молекулярной биологии привело к пониманию, какие именно функциональные феномены лежат в основе эпистаза, хотя много здесь еще предстоит исследовать [146].

В настоящее время эпистаз понимают как минимум в трех различных смыслах, которые не всегда однозначно соотносятся друг с другом.

Функциональный эпистаз – это любые функциональные взаимодействия между продуктами генов (и других значащих элементов генома), а также между отдельными позициями внутри генов и других значащих элементов генома. Под функциональными взаимодействиями белков подразумеваются как физическое взаимодействие с образованием комплексов, так и участие в одном процессе (например, ферменты одного биохимического пути) [147,148].

Композиционный эпистаз [149] — это модификация фенотипа одного аллеля действием другого аллеля. Это наиболее традиционное понимание, восходящее еще к Bateson [144]. Тестирование совместного действия аллелей происходит на некотором стандартном генетическом фоне, меняются только аллели в интересующих исследователя локусах. В некоторых случаях эпистатическому

взаимодействию аллелей (в смысле композиционного эпистаза) удаётся найти функциональную интерпретацию (т.е. связать с функциональным эпистазом). Например, исследованное еще Bateson и Pannet эпистатическое взаимодействие между аллелями, отвечающими за окраску цветка душистого горошка (*Lathyrus odoratus*), приводит к расщеплению во втором поколении 9:7. Впоследствии выяснилось, что эти гены кодируют два фермента, которые катализируют две последовательные реакции биосинтеза антоцианина — пигмента, отвечающего за сиреневую окраску цветка. Для того, чтобы антоцианин синтезировался, необходимо присутствие доминантных аллелей обоих генов, чем и объясняется расщепление 9:7 [149].

Статистический эпистаз относится к популяционной генетике и в своей основе имеет идеи Fisher [145] и Wright [150]. Согласно этому подходу, отклонение в фенотипе (например, приспособленности) при сочетании двух аллелей в разных локусах от ожидаемого считается как среднее по популяции, т.е. при наличии всех присутствующих в популяции генетических фонов. В этом ключевое отличие этого понимания эпистаза от композиционного, где эффекты совместного действия алллей оцениваются на некотором одном генетическом фоне. Можно понимать статистический эпистаз как усредненное по всем особям отклонение, которое вносится совместной заменой двух аллелей в различных локусах у случайно выбранной особи из популяции по сравнению с тем, что ожидается, если бы эти замены действовали независимо в той же особи [149]. Такой подход удобен для описания эволюции популяций. Формально, композиционный эпистаз можно считать частным случаем статистического, где частоты соответствующих аллелей равны 1 или 0. Понятно, что эффекты, описанные на уровне композиционного эпистаза, не всегда обобщаются на популяционный уровень. Мы будем использовать термин эпистаз именно в статистическом смысле. Далее, мы попытаемся дать более строгое определение с точки зрения эволюционной популяционной генетики.

Рассмотрим множество возможных генотипов особей. Каждому генотипу (точнее особи, несущей этот генотип) соответствует определенная приспособленность (см. раздел 2.3.1.1). Функция, отображающая пространство генотипов на приспособленность, называется ландшафтом приспособленности. Зачастую ландшафт приспособленности для наглядности рисуют в виде трехмерного графика, где в плоскости *XY* располагаются генотипы, а вдоль оси *Z* откладывается приспособленность (рис. 10). Соответственно ландшафт приспособленности изображается подобно горному ландшафту с локальными максимумами приспособленности (вершинами), седловинами и т.п. Однако стоит воспринимать такое изображение скорее как метафору, т.к. реальный ландшафт приспособленности многомерен (т.к. количество локусов в геноме огромно) и пространство возможных генотипов не может быть адекватно отображено на плоскости [150].

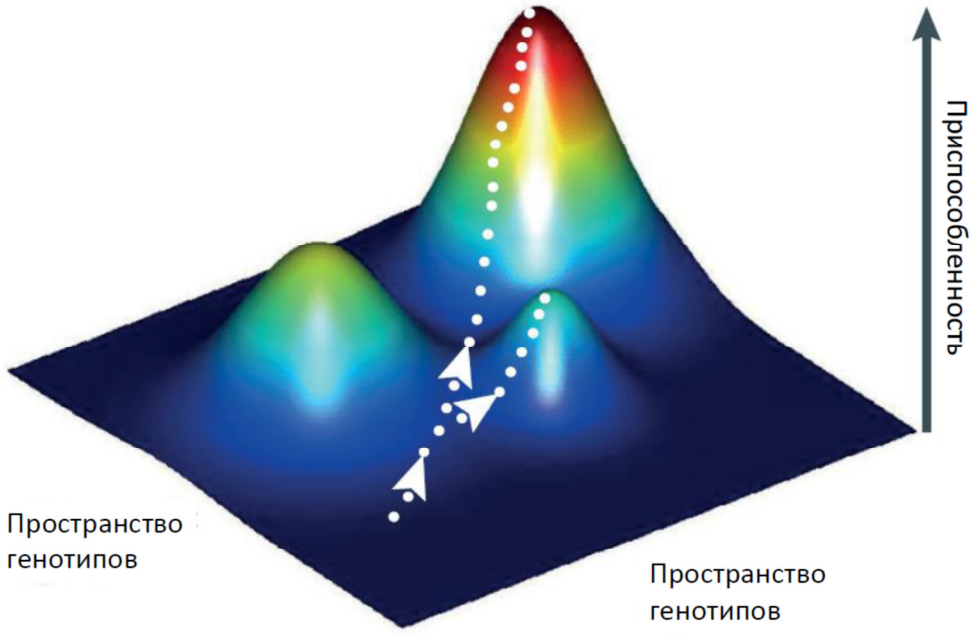


**Рисунок 10.** Ландшафт приспособленности [151].

В плоскости *XY* отложены генотипы, вдоль оси *Z* – приспособленность каждого генотипа.

Каждый из локусов генома вносит вклад в приспособленность особи в данной популяции. Если аллели в разных локусах воздействуют на приспособленность

независимо, то итоговая приспособленность особи пропорциональна произведению приспособленностей отдельных аллелей (т.е. приспособленность мультипликативна):

$$f = \prod_{k=1}^{L} w_k \quad (1)$$

где $k$ – номер локуса в геноме, $L$ – общее число локусов, $w_k$ – вклад в приспособленность соответствующего аллеля.

Тем самым, логарифм приспособленности аддитивен:

$$\log f = \sum_{k=1}^{L} \log w_k \quad (2)$$

В таком случае в ландшафте приспособленности существует один максимум, к которому популяция стремится. В реальности дело обстоит сложнее, т.к. аллели в различных локусах взаимодействуют друг с другом, и поэтому логарифм приспособленности неаддитивен.

Эпистаз – это феномен, когда вклад в приспособленность какого-то аллеля, находящегося в одном локусе (позиции) зависит от аллелей в других локусах (т.е. от генетического фона). Если рассматривать два аллеля в разных локусах, то в отсутствие эпистаза вклады каждого из аллелей в логарифм приспособленности организма аддитивны, т.е. если оба аллеля присутствуют одновременно, то их совместный вклад в логарифм приспособленности будет равен сумме вкладов, которые даёт каждый из аллелей. (Здесь и далее мы для простоты рассматриваем эпистаз в гаплоидных организмах. В диплоидных организмах ситуация аналогична, если принять возможность эпистатического взаимодействия между аллелями на гомологичных хромосомах одного локуса. Например, доминирование есть частный случай такого взаимодействия.)

При наличии эпистаза вклады аллелей неаддитивны, что в общем случае можно выразить следующим образом [152]:

$$\log f = \sum_k a_k^{(1)} \log w_k + \sum_{lm} a_{lm}^{(2)} \log w_l \log w_m + \sum_{lmn} a_{lmn}^{(3)} \log w_l \log w_m \log w_n + \ldots + a^{(L)} \log w_1 \log w_2 \ldots \log w_L \quad (3)$$

Здесь $\sum_k a_k^{(1)} \log w_k$ описывает аддитивную (неэпистатическую) компоненту, $\sum_{lm} a_{lm}^{(2)} \log w_l \log w_m$ – парные эпистатические взаимодействия, $\sum_{lmn} a_{lmn}^{(3)} \log w_l \log w_m \log w_n$ – тройные и т.д.

Вероятно, многие из коэффициентов $a^{(i)}$ равны нулю, однако пока до конца не понятно насколько распространён эпистаз высоких порядков [153].

Описание ландшафта приспособленности целиком на практике не представляется возможным в силу большого количества параметров. Современные экспериментальные работы обычно описывают взаимодействие нескольких локусов [154] или ограничиваются только парными взаимодействиями [155,156].

Зачастую влияние тех или иных замен в геноме на приспособленность организма сложно непосредственно измерить (за исключением экспериментов по экспериментальной эволюции микроорганизмов [155]), поэтому в качестве аналога приспособленности рассматривают какую-либо важную функциональную характеристику, например ферментативную активность [154] или свечение флуоресцентного белка [157]. Мы в данной работе в качестве такого приближения рассматриваем силу позиции сайта сплайсинга.

Рассмотрим два эпистатически взаимодействующих локуса *A* и *B*, с возможными аллелями $A_1/A_2$ и $B_1/B_2$ соответственно. (Вообще говоря, поскольку каждый локус могут занимать более двух аллелей, правильнее говорить не об эпистазе между локусами, а об эпистазе между заменами, см. ниже.) Рассмотрим замены $A_1 \rightarrow A_2$ и $B_1 \rightarrow B_2$ и соответствующие изменения логарифма приспособленности, вносимые этими заменами поодиночке $\Delta\log(w(A_1 \rightarrow A_2))$ *и* $\Delta\log(w(B_1 \rightarrow B_2))$, соответственно. Если замены происходят в том же геноме, соответствующее изменение логарифма приспособленности обозначим как $\Delta\log(f)$.

Если эпистаз отсутствует (рис. 11А), то $\Delta log(f) = \Delta log(w(A_1 \rightarrow A_2)) + \Delta log(w(B_1 \rightarrow B_2))$. Если $\Delta log(f)$, $\Delta log(w(A_1 \rightarrow A_2))$ и $\Delta log(w(B_1 \rightarrow B_2))$ одного знака, при этом нарушается вышеприведенное равенство, тогда эпистаз называется магнитудным (magnitude epistasis, рис. 11Б, В). Если же $\Delta log(w(A_1 \rightarrow A_2))$ и $\Delta log(w(B_1 \rightarrow B_2))$ имеют один знак, а $\Delta log(f)$ другой, то такой эпистаз называется знаковым (sign epistasis, рис. 11 Г). Таким образом, если эпистаз меняет лишь степень, но не направление изменения приспособленности, это – магнитудный эпистаз, если меняется направление – знаковый эпистаз [158].

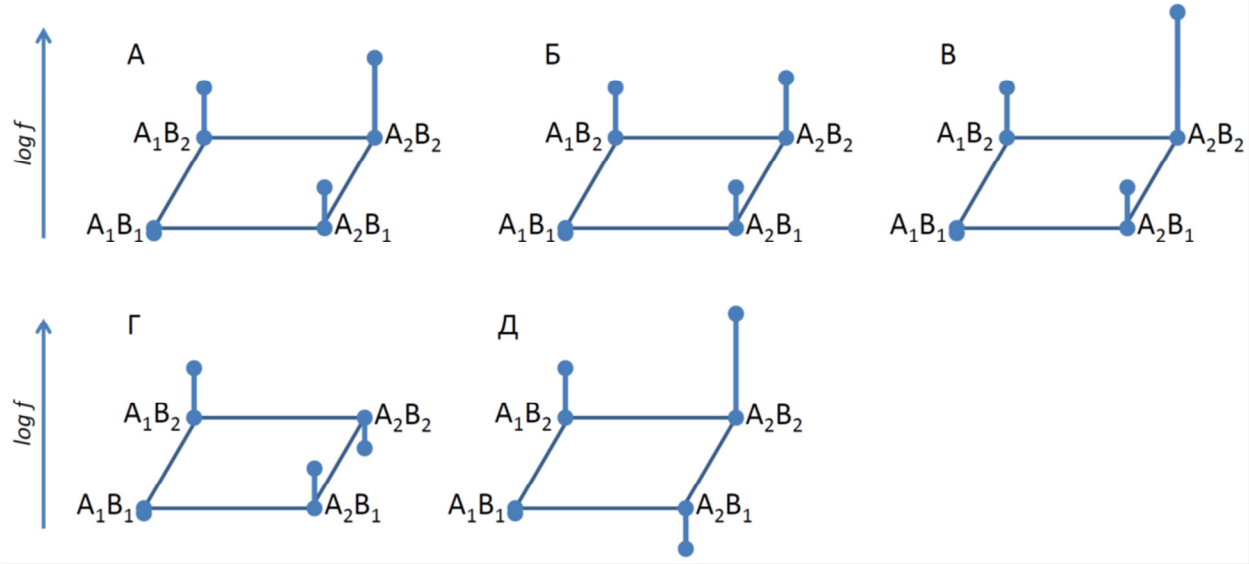


**Рисунок 11.** Виды эпистаза.

Рассматриваются виды эпистаза между заменами $A_1 \rightarrow A_2$ и $B_1 \rightarrow B_2$. В плоскости XY отображено пространство возможных генотипов (4 генотипа), вдоль оси Z отложен логарифм приспособленности. А – нет эпистаза, Б – отрицательный магнитудный эпистаз, В – положительный магнитудный эпистаз, Г – реципрокный знаковый эпистаз, Д – нереципрокный знаковый эпистаз (эквивалентен магнитудному).

Стоит отметить, что существует терминологическая путаница: то, что мы определили как знаковый эпистаз, некоторые авторы обозначают как реципрокный знаковый эпистаз (reciprocal sign epistasis), а под знаковым эпистазом понимается более широкий круг явлений: знак $\Delta log(f)$ должен быть противоположен знаку $\Delta log(w(A_1 \rightarrow A_2))$ **или** $\Delta log(w(B_1 \rightarrow B_2))$[151]. Так, если $\Delta log(w(A_1 \rightarrow A_2)) < 0$, $\Delta log(w(B_1 \rightarrow B_2)) > 0$, а $\Delta log(f) > 0$, то это знаковый эпистаз согласно такому определению. Однако, как несложно показать, если

переобозначить $A_1$ как $A_2$ и наоборот, этот эпистаз эквивалентен магнитудному (рис. 11, Д, Б).

Рассмотрим магнитудный эпистаз. Если $\Delta log(f) > \Delta log(w(A_1 \rightarrow A_2)) + \Delta log(w(B_1 \rightarrow B_2))$, такой эпистаз называется положительным (рис. 11В). Если $\Delta log(f) < \Delta log(w(A_1 \rightarrow A_2)) + \Delta log(w(B_1 \rightarrow B_2))$, то это – отрицательный эпистаз (рис. 11Б) [158]. Эта классификация не зависит от того, являются ли соответствующие замены полезными $\Delta log(w(A_1 \rightarrow A_2) > 0$ и $\Delta log(w(B_1 \rightarrow B_2)) > 0$) или вредными ($\Delta log(w(A_1 \rightarrow A_2)) < 0$ и $\Delta log(w(B_1 \rightarrow B_2)) < 0$). Это легко показать если переобозначить $A_1$ как $A_2$, $B_1$ как $B_2$ и наоборот (рис. 11Б, В).

Независимо от этого различают синергический и антагонистический эпистаз. Неформально, если эффект от двух мутаций более радикальный, чем ожидается, то это синергический эпистаз, если менее радикальный — антагонистический эпистаз [158]. Более формально, пусть $\Delta log(w(A_1 \rightarrow A_2))$ и $\Delta log(w(B_1 \rightarrow B_2))$ одного знака, тогда если $|\Delta log(f)| > |\Delta log(w(A_1 \rightarrow A_2))| + |\Delta log(w(B_1 \rightarrow B_2))|$, это синергический эпистаз; $|\Delta log(f)| < |\Delta log(w(A_1 \rightarrow A_2))| + |\Delta log(w(B_1 \rightarrow B_2))|$ — антагонистический эпистаз.

Стоит отметить, что понятия знаковый эпистаз (в смысле реципрокный знаковый), положительный и отрицательный магнитудный эпистаз относятся к паре замен $A_1 \rightarrow A_2$ и $B_1 \rightarrow B_2$. Однако если аллелей в каждом локусах $A$ и $B$ ровно по две, но не больше, то эти термины можно применять не к аллелям, а к локусам, т.к. направление замен не имеет значения (это несложно проверить, рис. 11Б, В, Г). Если в локусе больше двух аллелей, то между разные замены могу находиться в разных эпистатических взаимодействиях. Например, между заменами спаренных оснований в тРНК существует эпистаз. Рассмотрим, скажем, пару G≡C. Если заменить G на A, а C на T, между этими заменами наблюдается знаковый эпистаз (пара C=T даёт высокий вклад в приспособленность, в отличие от пар A~C и G~T) [156]. Если же производить замены G→T и C→T, то эпистаз будет другой по силе и, вероятно, положительный магнитудный. Аналогичная ситуация наблюдается для замен аминокислот в местах контактов взаимодействующих белков [154].

С геномной точки зрения эпистаз можно разделить на межгенный и внутригенный. Функциональной основой межгенного эпистаза являются взаимодействия между продуктами генов, генные сети — объекты, активно исследуемые системными биологами и генетиками [147]. Например, если два фермента катализируют две последовательные реакции в биосинтезе некоторого конечного продукта, то между вредными мутациями в соответствующих генах наблюдается положительный эпистаз, если же это две параллельные реакции (функциональная избыточность), то эпистаз будет отрицательным [146]. Если белки формируют мультибелковые комплексы, то эпистаз между заменами в местах белок-белковых контактов будет положительным или даже знаковым (в случае компенсирующих замен) [146,154].

Внутригенный эпистаз также весьма распространён как в белках [157,159,160], так и в РНК [155,156]. Мутации, нарушающие стабильность белка делают это не сразу, а только по достижении некоторого порога [159], что соответствует отрицательному эпистазу [161]. В snoRNA U3 отрицательные эпистатические взаимодействия между нуклеотидами равномерно распределены вдоль молекулы, количественно преобладают над положительными, что, по-видимому, свидетельствует о существовании порога стабильности [155]. Было показано существование знакового эпистаза в тРНК при замене одной комплементарной пары на другую [156]. Если белок приобретает новую функцию, такую как связывание нового лиганда, то кроме замен аминокислот, ответственных за взаимодействие с лигандом, необходимы соответствующие конформационные изменения [160], что можно назвать конформационным эпистазом. Более того, замены, приводящие к связыванию нового лиганда, могут иметь плейотропный эффект: связывание усиливается, а стабильность белка падает, поэтому происходит дополнительная замена, повышающая стабильность [162].

Указанные примеры функциональных основ эпистаза представляют далеко не полный список, для более подробного анализа см. [146].

#### *2.3.2.2. Эпистатический отбор*

Рассмотрим несколько локусов. Пусть в каждом из локусов происходят вредные замены, причем каждая замена уменьшает логарифм приспособленности приблизительно на одну и ту же величину. Если замены в этих локусах не взаимодействуют эпистатически, то зависимость логарифма приспособленности от количества замен линейна. Если между локусами (заменами) преобладает отрицательный эпистаз, то кривая зависимости будет выпуклой (каждая следующая замена вреднее предыдущей), если преобладает положительный эпистаз, то кривая будет вогнутой (рис. 12). Знаковый эпистаз приводит к немонотонности этой кривой. Если рассматривать не вредные, а полезные замены, то ситуация будет аналогичной, только зависимости будут не убывающие, а возрастающие.

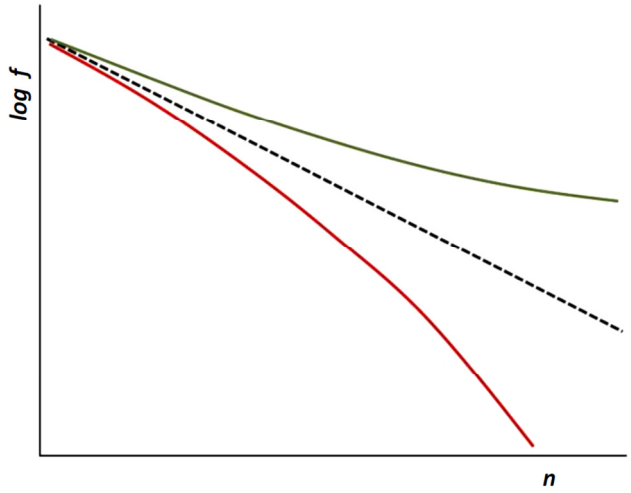


**Рисунок 12.** Значимость логарифма приспособленности (*log f*) от числа вредных мутаций (*n*).

В отсутствие эпистаза зависимость линейна (пунктирная линия), при положительном эпистазе кривая вогнута (верхняя зелёная линия), при отрицательном – выпукла (нижняя красная линия). Относительное положение кривых вдоль вертикальной оси условно, имеет значение только форма зависимости [158].

Как отбор действует на эпистатически взаимодействующие локусы? Рассмотрим два локуса *A* и *B* с аллелями *A/a* и *B/b* (*w(A) > w(a), w(B) > w(b)*). Пусть в состоянии мутационно-селекционного равновесия установились некоторые частоты

генотипов *AB*, *Ab*, *aB* и *ab* (рис. 13А). Рассмотрим несколько ситуаций: отсутствие эпистаза, положительный, отрицательный и знаковый эпистаз (рис. 13Б). Как эпистаз влияет на распределение этих частот (рис. 13А, В)? Отрицательный и знаковый эпистаз приводят к избытку генотипов с промежуточной приспособленностью (*Ab* и *aB*) по сравнению с ситуацией отсутствия эпистаза. Положительный эпистаз приводит к недостатку генотипов с промежуточной приспособленностью [163]. Таким образом, положительный/отрицательный эпистаз приводит к положительным/отрицательным корреляциям приспособленностей локусов *A* и *B*, соответственно (если за приспособленность локуса считать индивидуальный вклад соответствующего аллеля в предположении отсутствия эпистаза). Или, в других терминах, положительный/отрицательный эпистаз приводят положительному/отрицательному неравновесию по сцеплению, соответственно. Рекомбинация частично разрушает неравновесие по сцеплению (рис. 13В). Мы изучали эпистаз между позициями в сайтах сплайсинга и обнаружили как положительные, так и отрицательные корреляции между силами отдельных позиций, которые возникли в результате действия эпистатического отбора.

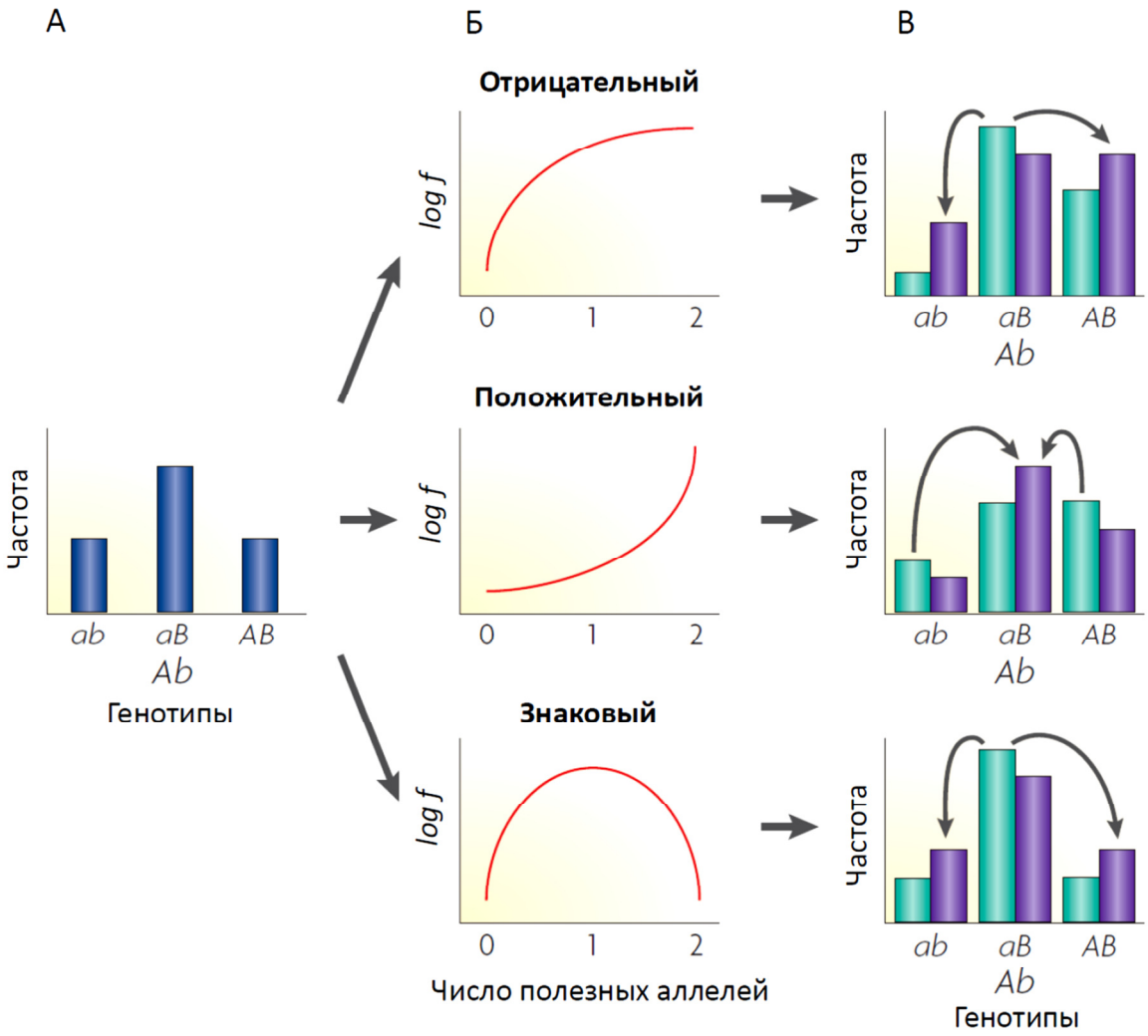


**Рисунок 13.** Эпистатический отбор.

А – частоты генотипов в равновесии при отсутствии эпистаза, Б – зависимость приспособленности от числа полезных аллелей (*A* или *B*) при разных типах эпистаза, В – частоты генотипов при наличии эпистаза и отсутствии рекомбинации (зелёные левые столбики), в результате рекомбинации (показано стрелками), неравновесие по сцеплению частично разрушается, что приводит к изменению частот генотипов (фиолетовые столбики справа) [152].

# 3. Данные и методы

## 3.1. Исходные данные

### 3.1.1. Выборки конститутивных и кассетных экзонов и соответствующих сайтов сплайсинга

Последовательности сайтов сплайсинга для изучения коррелированной эволюции позиций (раздел 4.2) были взяты из базы данных EDAS (http://edas2.bioinf.fbb.msu.ru/, [164]). Мы использовали только сайты с каноническими ключевыми динуклеотидами: GT для донорных сайтов и AG для акцепторных сайтов сплайсинга. Геномные координаты ортогогичных сайтов сплайсинга в геномах *Homo sapiens*, *Mus musculus* и *Canis familiaris* были получены как описано ранее [165,166]. На основании этих координат для каждого из трёх геномов мы получили последовательности сайтов и их геномного окружения: канонический динуклеотид, 18 ближайших нуклеотидов в интроне и 10 нуклеотидов в экзоне (суммарно 30 нт).

Мы рассматривали только сайты сплайсинга, фланкирующие кассетные или конститутивные экзоны (для классификации использовались транскрипты человека). Экзон считался конститутивным, если (1) он был поддержан как минимум 50 человеческими транскриптами (EST или мРНК), (2) он используется более, чем в 95% транскриптов, которые выравниваются на данный локус генома, т.е. частота включения экзона >95%. Экзон считался кассетным, если частота его включения между 5% и 95%. Экзоны, не определённые как конститутивные или альтернативные, были исключены из анализа. Мы рассматривали экзоны только внутри кодирующих последовательностей генов; кассетный экзон мог содержать стоп-кодон, но только если пропуск данного экзона удлинял открытую рамку считывания. На этом этапе было получено 12245 троек донорных сайтов и 12245 троек акцепторных сайтов сплайсинга.

Для того, чтобы протестировать устойчивость нашего анализа (см. раздел 4.2.3) мы использовали две выборки сайтов сплайсинга: (1) с использованием транскриптомных данных только для человека, как описано выше, и (2) накладывая дополнительное требование: каждый сайт сплайсинга должен иметь как минимум одну EST/мРНК в мыши, подтверждающую его функциональность. Ниже приводятся результаты, полученные с использованием второй выборки, сделанные выводы сохраняются и на первой выборке. Эта фильтрационная процедура примерно в три раза уменьшила размер выборки: осталось 4388 акцепторных сайтов и 3596 донорных сайтов сплайсинга, включая 3596 донорных сайтов конститутивных экзонов, 791 донорный сайт кассетных экзонов, 3601 акцепторный сайт конститутивных экзонов и 811 акцепторных сайтов кассетных экзонов. Кроме того, фильтрация незначительно сместила распределение частот включения кассетных экзонов в сторону низких значений (Приложение: табл. 1).

Для проверки гипотезы о миграции сигнала (см. раздел 4.2.3), конститутивные донорные сайты сплайсинга были разделены на низко- и высококонсервативные в соответствии с многовидовой консервативностью ключевого динуклеотида AG/GT (см. раздел 3.2.5). Донорные сайты сплайсинга, многовидовая консервативность ключевого динуклеотида которых (считая от генома *Mus musculus*) была < $1K_s$ (т.е. в пределах плацентарных млекопитающих) считались низкоконсервативными, сайты с консервативностью ≥ $1K_s$ рассматривались как высококонсервативные. Выборка конститутивных донорных сайтов сплайсинга была разделена на 1655 низко- и 1923 высококонсервативных сайтов. 19 сайтов были исключены, т.к. они неоднозначно картировались на геном человека.

Последовательности сайтов сплайсинга для изучения отбора в сайтах сплайсинга (раздел 4.1) были получены из аннотации GENCODE, версия 7 (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGencodeV7/) . Суммарно были получены данные по 255002 акцепторным сайтам и 261400 донорным сайтам сплайсинга человека и 46498 акцепторным и 48041 донорным сайтам сплайсинга *D. melanogaster*. Мы исключили из анализа все сайты

сплайсинга, в которых отсутствовал канонический динуклеотид (AG или GT) в геноме хотя бы одного из анализируемых видов: референтном виде, сестринской группе или во внешней группе (см. раздел 3.2.1). Для того чтобы минимизировать влияние отбора, связанного с регуляцией альтернативного сплайсинга, мы оставили в выборке только сайты, фланкирующие конститутивные экзоны в геноме референтного вида. В результате этой фильтрации осталось 159280 конститутивных акцепторных сайтов сплайсинга человека, 160234 конститутивных донорных сайтов сплайсинга человека, 40859 конститутивных донорных сайтов сплайсинга *D. melanogaster* и 41051 донорных сайтов сплайсинга *D. melanogaster*. Каждая из этих выборок разделялась на несколько подвыборок в зависимости от положения сайта в транскрипте (внутри кодирующей части гена или внутри нетранслируемых областей мРНК и в некодирующих РНК) и силы сайта сплайсинга (сильные, средние или слабые сайты). Сила сайта определялась как число консенсусных нуклеотидов; пороги для разделения были выбраны таким образом, чтобы приблизительно равное число сайтов попало в каждую из категорий. Из первоначальной выборки также были получены и отдельно проанализированы сайты сплайсинга, фланкирующие кассетные экзоны (накладывались те же ограничения на канонические динуклеотиды): 39758 (1376) донорных и 39644 (1353) акцепторных сайтов сплайсинга человека (*D. melanogaster*, соответственно).

Последовательности экзонов для исследования отбора в окрестностях сайтов сплайсинга были получены из дополнительных материалов к статье [167]. Для анализа консервативности использовались следующие сборки геномов: hg16 (июль 2003 г.) – геном *Homo sapiens*; mm4 (октябрь 2003 г.) – геном *Mus musculus*; rn3 (июнь 2003 г.) – геном *Rattus norvegicus*.

### 3.1.2. Поиск ортологичных сайтов сплайсинга

Выравнивания ортологичных сайтов сплайсинга из геномов млекопитающих, а также мух рода *Drosophila*, использованные для изучения отбора в сайтах сплайсинга (раздел 4.1), были получены из полногеномных выравниваний UCSC Genome Browser. Выравнивания, использованные для анализа коррелированной

эволюции сайтов сплайсинга (раздел 4.2), были получены как описано ранее [165,166].

Анализ консервативности цис-регулятора сплайсинга UGCAUG (раздел 4.3) производился с помощью UCSC Genome Browser (http://genome.ucsc.edu/) и программы BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat). Следует отметить, что изначальная выборка содержала как экзоны из генома человека, так и из генома мыши. В обоих случаях определялся ортологичный ген в оставшемся геноме и исследовалась консервативность UGCAUG. Таким образом, начальная гетерогенность выборки не оказывает существенного влияния на результат.

Процедура анализа консервативности регулятора сплайсинга UGCAUG в геномах человека и мыши включала следующую последовательность действий: (1) локализация экзона в соответствующем геноме с помощью программы BLAT; (2) поиск последовательности мРНК или белка, содержащей этот экзон с помощью UCSC Genome Browser (если таковой не находилось, то использовалась изоформа мРНК, содержащая соседние экзоны, исследуемый экзон вставлялся в эту мРНК и получившаяся искусственная изоформа транслировалась в белковую последовательность); (3) идентификация ортологичного гена в другом геноме путём выравнивания белковой последовательности с геномом другого вида (BLAT); (4) нахождение ортологичного экзона с использованием выравнивания BLAT и анализа EST/мРНК в UCSC Genome Browser; (5) поиск всех гексануклеотидов UGCAUG в пределах 1000 нт в интронах после рассматриваемых экзонов отдельно в геномах человека и мыши (если интрон был короче 1000 нт, то поиск производился только в интроне); (6) выравнивание интронных последовательностей с целью определить, насколько UGCAUG консервативен (мы делали локальное выравнивание найденных UGCACUG со 100-нт фланками с обеих сторон и интрона другого вида).

### 3.1.3. Данные по уровню экспрессии генов, уровню рекомбинации, консервативности и однонуклеотидным полиморфизмам

Данные по локальному уровню рекомбинации различных участков генома *D. melanogaster* были получены из работы Comeron с соавт. [168]. Для человеческого генома использовались усреднённые по полам данные по рекомбинации из статьи Kong с соавт. [169], скачанные с сайта UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/recombRate.txt.gz).

Данные по экспрессии генов человека были получены и обработаны как описано в [170].

Данные по phyloP score – мере консервативности каждого нуклеотида в геноме человека – рассчитанные по выравниванию геномов человека и приматов, были скачаны с сайта UCSC Genome Browser (http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP46way/primates/).

Координаты коротких интронов *D. melalanogaster* были получены из работы [171].

Данные по однонуклеотидным полиморфизмам (SNP) в геноме человека были взяты из проекта “1000 геномов” [172] (http://www.1000genomes.org/) и в геноме *D. melanogaster* — из проекта DGRP Freeze One Release [173] (https://www.hgsc.bcm.edu/arthropods/drosophila-genetic-reference-panel).

## 3.2. Методы

### 3.2.1. Построение матриц нуклеотидных замен методом парсимонии

При исследовании отбора, действующего на сайты сплайсинга (раздел 4.1), мы использовали метод парсимонии (наибольшей экономии) для восстановления предковых последовательностей и подсчёта числа нуклеотидных замен на ветках. Изучались события, произошедшие на линиях человека (*Homo sapiens*) и *Drosophila melanogaster* с момента их расхождения с линиями макаки (*Macaca mulatta*) и *Drosophila simulans*, соответственно. В качестве внешней группы использовались, соответственно, игрунка обыкновенная (*Callithrix jacchus*) и *Drosophila yakuba*.

Рассмотрим некоторую позицию в наборе сайтов сплайсинга данного вида (*H. sapiens* или *D. melanogaster*). Для каждой позиции каждого сайта сплайсинга мы восстанавливаем наиболее вероятный нуклеотид в геноме ближайшего общего предка данного вида и сестринской группы методом парсимонии. Таким образом для набора сайтов сплайсинга в каждой позиции мы имеем матрицу 4×4 количества замен каждого типа (A→A, A→C, …, T→T) на ветке человека или *D. melanogaster*.

Так как филогенетические расстояния между указанными видами относительно невелики [174,175], можно считать, что множественные замены в одном сайте практически отсутствуют, следовательно метод парсимонии должен работать хорошо. Однако, чтобы быть в этом до конца уверенными, мы повторили основные результаты по оценке силы отбора с использованием метода максимального правдоподобия (пакет PAML, модель нуклеотидных замен UNREST, [176]) и не получили существенных отличий (раздел 4.1.5).

### 3.2.2. Сила сайта и её изменение

Сила сайта сплайсинга [50] есть мера близости последовательности сайта к консенсусу. В настоящей работе мы используем не силу сайта целиком, а рассматриваем только силу отдельных позиций. Рассмотрим последовательность сайта сплайсинга *S*. В *i*-той позиции сайта находится нуклеотид *S(i)*. Сила позиции *S(i)* определятся как

$$W(S(i)) = \log \frac{p(S(i) \mid F)}{p(S(i) \mid B)} = \log \frac{p_F(S(i), i)}{p_B(S(i), i)}, \qquad (4)$$

где $p(S(i)|F)$ – вероятность нуклеотида *S(i)* в предположении, что *S(i)* стоит в *i*-той позиции сайта сплайсинга в основной модели (the **f**oreground model); $p(S(i)|B)$ вероятность нуклеотида *S(i)* в предположении, что этот нуклеотид *не* в сайте сплайсинга (фоновая модель, the **b**ackground model); $p_F(\alpha, i)$ – это вероятность нуклеотида $\alpha$ в позиции *i* в основной модели, посчитанная из частот встречаемости нуклеотидов в выборке сайтов сплайсинга соответствующего типа; $p_B(\alpha, i)$ – вероятность нуклеотида $\alpha$ в позиции *i* в фоновой модели.

Чтобы учесть возможность того, что наблюдаемые эффекты могут зависеть от выборки сайтов сплайсинга, используемой для построения основной модели, мы рассчитывали силу позиций сайтов как с использованием выборки сайтов сплайсинга из генома общего предка, так и с использованием всех сайтов из геномов потомков (человек, мышь и собака). Результаты анализа почти не отличались (данные не показаны), далее представлены результаты с использованием второго подхода

Мы использовали простейшую фоновую модель: $p_B(\alpha, i)= 0.25$ для всех $\alpha$ и для всех $i$. Поскольку в нашей работе мы изучаем разницу сил сайтов, выбор фоновой модели не влияет на результат.

Если нам известны частоты $q(\alpha, i)$ для всех нуклеотидов $\alpha$ и всех позиций $i$ в выборке сайтов сплайсинга, мы можем посчитать средний вес $i$-той позиции в данной выборке:

$$\hat{W}(i) = \sum_{\alpha} q(\alpha, i) \log \frac{p_F(\alpha, i)}{p_B(\alpha, i)}. \qquad (5)$$

### 3.2.3. Матрица ковариаций силы отдельных нуклеотидов в сайтах сплайсинга

Рассмотрим выборку сайтов сплайсинга одного вида (донорных или акцепторных). Тогда для каждой позиции каждого сайта сплайсинга мы можем оценить силу позиции с занимающим её нуклеотидом согласно формуле (4). Значит, каждому сайту сплайсинга можно сопоставить вектор сил соответствующих позиций. Выборке сайтов сплайсинга соответствует набор таких векторов. Имея набор векторов сил позиций, получаем матрицу ковариаций силы отдельных нуклеотидов.

### 3.2.4. Построение нейтральных контролей для оценки изменения силы позиций сайтов сплайсинга

Для каждой позиции *i* сайта сплайсинга мы определяем нейтральную матрицу (**E**xpected) нуклеотидных замен $P_E(i)$ предполагая отсутствие давления отбора в этой позиции (в противоположность реально наблюдаемой (**O**bserved) матрице замен $P_O(i)$). Для построения нейтральных матриц мы использовали экзонные и интронные позиции, расположенные поблизости от сайта сплайсинга, на которые почти не действует отбор. Для построения нейтральных матриц мы пользовались тем же алгоритмом, что и при построении наблюдаемых матриц (раздел 4.2.1.). Чтобы усреднить матрицы по нескольким позициям, мы суммировали матрицы совместных вероятностей, и затем нормировали их, чтобы получить нейтральную матрицу нуклеотидных замен ($P_E(i)$).

Для того чтобы учесть локальный контекст, мы использовали разные нейтральные контроли для различных позиций сайтов сплайсинга. Для интронной части донорного сайта сплайсинга, мы использовали ближайшие позиции в интроне (+7…+12). Это гарантирует отсутствие пересечения контрольных позиций с другими сайтами сплайсинга, а также минимизирует вероятность включения в контроль неаннотированных экзонов, поскольку маловероятно, что экзоны будут располагаться так близко к сайту сплайсинга. Поскольку полипиримидиновый тракт акцепторных сайтов сплайсинга представляет собой относительно длинную последовательность с плохо определенной дистальной границей (5’), мы не можем использовать в качестве нейтрального контроля интронные позиции вблизи акцепторного сайта. Вместо этого мы использовали в качестве контроля интронные позиции, следующие после донорного сайта сплайсинга того же экзона.

Для экзонных позиций сайтов сплайсинга мы строили нейтральные матрицы с учетом рамки считывания. В случае донорных сайтов сплайсинга мы использовали позиции -4 и -7 (-5 и -8; -6 и -9) для построения нейтральной матрицы для позиции -1 (позиций -2; -3, соответственно). Аналогично, позиции +4 и +7 (+5 и +8; +6 и +9) использовались для построения нейтральной матрицы для позиций +1 (соответственно, +2; +3) акцепторных сайтов сплайсинга.

Если для данной позиции *i* нам известен вектор вероятностей нуклеотидов в предковом геноме *p(ancestor, i)* и нейтральная матрица нуклеотидных замен $P_E(i)$, мы можем получить ожидаемый вектор вероятностей нуклеотидов в геноме потомка (человека, мыши или собаки) $p_E(descendant, i) = P_E(i) \times p(ancestor, i)$. Мы также знаем наблюдаемый вектор вероятностей нуклеотидов в геноме потомка $p_O(descendant, i)$ из выборки сайтов сплайсинга.

Ожидаемая средняя сила позиции *i* в геноме потомка (в предположении нейтральной эволюции) $\hat{W}_E(descendant, i)$ определяется как средняя сила позиции согласно уравнению (2), где частоты нуклеотидов берутся из вектора $p_E(descendant, i)$. Аналогично определяется наблюдаемая в сайте сплайсинга средняя сила позиции $\hat{W}_O(descendant, i)$, при этом частоты нуклеотидов берутся из вектора $p_O(descendant, i)$ Средняя сила позиции *i* в геноме общего предка – это средняя сила позиции, где частоты нуклеотидов берутся из вектора *p(ancestor, i)*.

Таким образом, для каждой позиции *i*, мы подсчитываем наблюдаемое изменение средней силы сайта $\Delta\hat{W}_O(i) = \hat{W}_O(descendant, i) - \hat{W}(ancestor, i)$ и ожидаемое при нейтральной эволюции изменение силы сайта $\Delta\hat{W}_E(i) = \hat{W}_E(descendant, i) - \hat{W}(ancestor, i)$.

### 3.2.5. Оценка многовидовой консервативности

Предположим, что у нас имеется последовательность из генома некоторого вида (будем называть этот вид референтным) и множественное выравнивание ортологичных ей последовательностей из других видов. Тогда для любой позиции из последовательности референтного вида можно вычислить многовидовую консервативность, которая является мерой того, насколько долго в эволюции этот нуклеотид остаётся неизменным.

Многовидовая консервативность определяется следующим образом. Будем измерять филогенетическое расстояние между видами как среднее число замен на один нуклеотидный сайт в нейтральных областях генома ($K_s$). Сначала мы вычисляем филогенетическое расстояние *L(r, i)* (в единицах $K_s$) между

референтным видом *r* и всеми видами *i*, представленными в выравнивании. Далее мы находим такой вид *x*, что он наиболее удалён от референтного вида и при этом нуклеотид (или набор нуклеотидов) в рассматриваемой позиции, наблюдаемый у референтного вида присутствует у всех видов *j* из выравнивания таких, что $L(r, j) < L(r, x)$. Величина $L(r, x)$ называется многовидовой консервативностью. Таким образом, многовидовая консервативность представляет собой максимальное наблюдаемое филогенетическое расстояние (в единицах $K_s$) от референтного вида до другого вида, где рассматриваемый нуклеотид (или набор нуклеотидов) остался неизменным, причём он также остался неизменным во всех видах, находящихся на меньшем расстоянии.

Если имеется выборка выравниваний последовательностей (сайтов сплайсинга), то для конкретной позиции мы можем усреднить $L(r, x)$ по всем выравниваниям (средняя многовидовая консервативность). Тогда для нуклеотида (или набора нуклеотидов) в рассматриваемой позиции мы можем сравнить наблюдаемую среднюю многовидовую консервативность (*obs*) с ожидаемой консервативностью в нейтрально эволюционирующих участках генома (*exp*), вычислив относительную консервативность $c = (obs - exp)/exp$. Если рассматривается один нуклеотид, то соответствующая консервативность обозначается $c_{nuc}$, если рассматривается набор нуклеотидов (например, все некосенсусные нуклеотиды), то соответствующая консервативность обозначается $c_{set}$.

Филогенетические расстояния между видами разными позвоночных (а также насекомых) получались из соответствующих деревьев путём суммирования длин соответствующих ветвей (деревья были скачаны с сайта UCSC Genome Browser, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/multiz100way/hg19.100way.nh , http://hgdownload.soe.ucsc.edu/goldenPath/dm6/multiz27way/dm6.27way.nh).

### 3.2.6. Контроль на динуклеотидный состав для полипиримидиновых трактов акцепторных сайтов сплайсинга

Для того, чтобы понять, вызван ли периодический паттерн корреляций сил позиций в позициях с -12 по -7 (раздел 4.2.5.1.) динуклеотидным составом, мы

сделали следующую процедуру. Мы получили выборку искусственно сгенерированных акцепторных сайтов сплайсинга того же размера, что и исходная выборка акцепторных сайтов, сохраняя без изменений последовательности до -12 и после -7, но заменив позиции -12…-7 искусственно сгенерированными последовательностями. Последовательности в позициях с -12 по -7 были получены с помощью цепи Маркова первого порядка, где исходные вероятности нуклеотидов были оценены из частот нуклеотидов, а переходные вероятности оценены из частот динуклеотидов на рассматриваемом участке. Полученные таким методом искусственные сайты сохраняют ковариации между соседними позициями на участке -12… -7, но более далёкие ковариации разрушаются.

### 3.2.7. Статистические методы. Тестирование статистических гипотез и построение доверительных интервалов.

Статистическая значимость результатов из раздела 4.1. оценивалась следующим образом. 95%-ные интервалы для частот замен были построены исходя из биномиального распределения с помощью метода Клоппера-Пирсона (функция binofit в пакете MATLAB). 95%-ные доверительные интервалы для $4N_es$ были посчитаны аналогично, предполагая, что ошибками в нейтральных контролях можно пренебречь. Для многовидовой консервативности 95%-ные доверительные интервалы рассчитывались с помощью бутстрепа колонок выравнивания (1000 итераций) [177].

Статистическая значимость результатов, представленных в разделе 4.2, оценивалась путём бутстрепа изначальной выборки троек сайтов (1500 итераций). Построение 95%-ных доверительных интервалов и проверка одно- и двухвыборочных статистических гипотез для различных статистик, производились как описано в [177].

# 4. Результаты и обсуждение

## 4.1. Положительный и отрицательный отбор в сайтах сплайсинга

### 4.1.1. Тест на положительный и отрицательный отбор в сайтах сплайсинга. Построение нейтральных контролей

Рассмотрим набор сайтов сплайсинга одного типа (донорных или акцепторных). Для каждой позиции сайта сплайсинга мы посчитали частоту встречаемости каждого нуклеотида. Один или два самых частых нуклеотида мы называли консенсусными, остальные – неконсенсусными (рис. 2). Показанные на рисунке 2 диаграммы Logo построены по выборке сайтов сплайсинга конститутивных экзонов человека (конститутивных сайтов сплайсинга). Частоты нуклеотидов для *D. melanogaster* очень близки к таковым у человека. Диаграммы строились с помощью программы WebLogo [178].

Для каждой позиции мы подсчитали наблюдаемую частоту замен между консенсусными (*Cn*) и неконсенсусными (*Nc*) нуклеотидами на линии *H. sapiens* (после расхождения с линией *M. mulatta*) и на линии *D. melanogaster* (после расхождения с линией *D. simulans*): $q_{obs}(Cn \to Nc)$ и $q_{obs}(Nc \to Cn)$:

$$q_{obs}(Cn \to Nc) = \frac{\sum_{Z \in Cn, X \in Nc} \#(Z \to X)}{\sum_{Z \in Cn} \#(Z \to Y)} \quad (6)$$

$$q_{obs}(Nc \to Cn) = \frac{\sum_{Z \in Nc, X \in Cn} \#(Z \to X)}{\sum_{Z \in Nc} \#(Z \to Y)} \quad (7)$$

где $\#(Z \to X)$ и $\#(Z \to Y)$ – числа замен между нуклеотидом *Z* в геноме предка и нуклеотидом *X* (или *Y*, соответственно) в геноме потомка.

Ожидаемые частоты замен $q_{exp}(Cn \to Nc)$ и $q_{exp}(Nc \to Cn)$ считались по этим же формулам, но вместо консенсусных/неконсенсусных нуклеотидов, определенных в конкретной позиции сайта сплайсинга, использовались формально те же

нуклеотиды в нейтрально эволюционирующем участке генома (нейтральный контроль). Мы использовали разные нейтральные контроли для разных позиций: для интронных позиций сайта сплайсинга брались примыкающие к сайту последовательности из того же интрона, а для экзонных – третьи позиции 4-вырожденных кодонов в примыкающем экзоне (см. раздел 3.2.4). Мы также отфильтровали около половины позиций в нейтральном контроле в человеческом геноме, у которых phyloP score < 0.6 [179], чтобы исключить возможность загрязнения нейтрального контроля последовательностями, находящимися под (слабым) отрицательным отбором. Для генома *D. melanogaster* данные по phyloP score отсутствуют, поэтому в качестве дополнительной проверки мы использовали короткие интроны – участки генома *D. melanogaster*, на которые действие отрицательного отбора минимально [171]. Полученные на коротких интронах ожидаемые частоты замен почти не отличались от таковых полученных способом, описанным для человеческого генома (без фильтра на phyloP score, данные не показаны).

Тест на отбор состоял в следующем. Рассмотрим замены из неконсенсусных нуклеотидов в консенсусные в данной позиции сайта сплайсинга. Нулевая гипотеза ($H_0$) — на неконсенсусные нуклеотиды отбор не действует, т.е. все замены происходят благодаря мутациям и генетическому дрейфу. Имеется две альтернативные гипотезы: $H_1$ — на неконсенсусные нуклеотиды действует отрицательный отбор; $H_2$ — на неконсенсусные нуклеотиды действует положительный отбор, стремящийся заменить их на консенсусные. Если $q_{obs}(Nc \rightarrow Cn) < q_{exp}(Nc \rightarrow Cn)$, то верна $H_1$; если $q_{obs}(Nc \rightarrow Cn) > q_{exp}(Nc \rightarrow Cn)$, то верна $H_2$. Если $q_{obs}(Nc \rightarrow Cn)$ статистически значимо не отличается от $q_{exp}(Nc \rightarrow Cn)$, то нельзя отвергнуть $H_0$. Для замен Cn → Nc тест полностью аналогичен. Этот тест является полным аналогом теста $D_n/D_s$ для белок-кодирующих последовательностей [123].

### 4.1.2. Отбор на консенсусные и неконсенсусные нуклеотиды

Рассмотрим замены на линии *H. sapiens*, произошедшие после расхождения с линией *M. mulatta*, и на линии *D. melanogaster* — после расхождения с линией *D.*

*simulans*. Мы не рассматривали позиции +1 и +2 (-1 и -2) донорных (соответственно, акцепторных) сайтов сплайсинга, так как по построению в них присутствуют инвариантные и консервативные динуклеотиды GT (AG), а также позицию -4 акцепторных сайтов сплайсинга, в связи с отсутствием в ней консенсуса.

На рисунках 14 и 15 представлены данные по частоте замен из консенсусных аллелей (нуклеотидов) в неконсенсусные и обратно в различных позициях сайтов сплайсинга в сравнении с соответствующими частотами для нейтрально эволюционирующих последовательностей генома на линии человека (рис. 14) и *D. melanogaster* (рис. 15). Мы отдельно изучали “некодирующие” сайты сплайсинга, находящиеся на экзон-интронных границах некодирующих сегментов РНК (UTR мРНК и некодирующие РНК), и “кодирующие” сайты сплайсинга, фланкирующие интроны, в кодирующих последовательностях мРНК, в связи с тем, что они показывают разные частоты замен.

Во всех позициях консенсусные нуклеотиды заменяются на неконсенсусные медленнее, чем в соответствующих нейтрально эволюционирующих последовательностях (p-значение < $10^{-18}$ по точному тесту Фишера, Приложение: табл. 2, рис. 14А-Г и 15А-Г). В большинстве позиций, неконсенсусные нуклеотиды заменяются на консенсусные быстрее, чем в нейтральном контроле ($p < 10^{-3}$ для всех выборок, кроме донорных сайтов *D. melanogaster*, где разница статистически не значима, Приложение: табл. 2, Рис. 14Д-З и 15Д-З). Это значит, что консенсусные нуклеотиды защищены отрицательным отбором, в то время как неконсенсусные нуклеотиды находятся под воздействием положительного отбора, стремящегося заменить их на консенсусные. В акцепторных сайтах сплайсинга сигнал как положительного, так и отрицательного отбора постепенно уменьшается в сторону 5'-конца интрона, становясь статистически неотличимым от нуля приблизительно в позиции -22 в человеческом геноме (в позиции -17 в геноме *D. melanogaster*), что, видимо, маркирует конец полипиримидинового тракта.

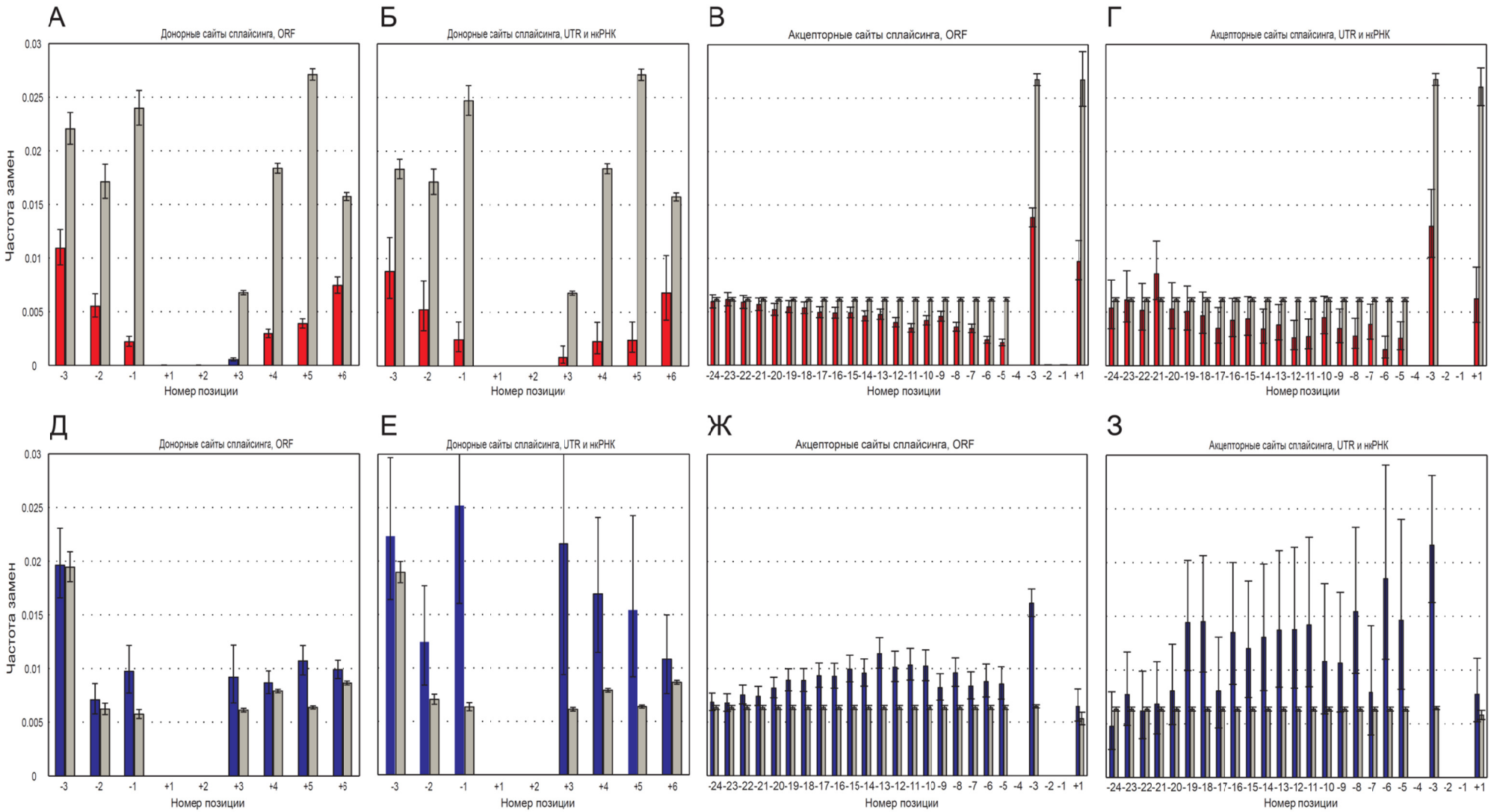


**Рисунок 14**. Частоты нуклеотидных замен в различных позициях сайтов сплайсинга на линии *Homo sapiens* после её дивергенции от линии *Macaca mulatta*.

По горизонтальной оси отложена позиция в сайте сплайсинга относительно инвариантного динуклеотида AG (GT). Верхний ряд (А–Г) – наблюдаемые (красные столбики) и ожидаемые (серые столбики) частоты замен *Cn*→*Nc*. Нижний ряд (Д–З) – наблюдаемые (красные столбики) и ожидаемые (серые столбики) частоты замен *Nc*→*Cn*. А, Б, Д, Е – донорные сайты; В, Г, Ж, З – акцепторные сайты. А, В, Д, Ж – кодирующие сайты; Б, Г, Е, З – некодирующие сайты. Усы – 95%-ные доверительные интервалы.

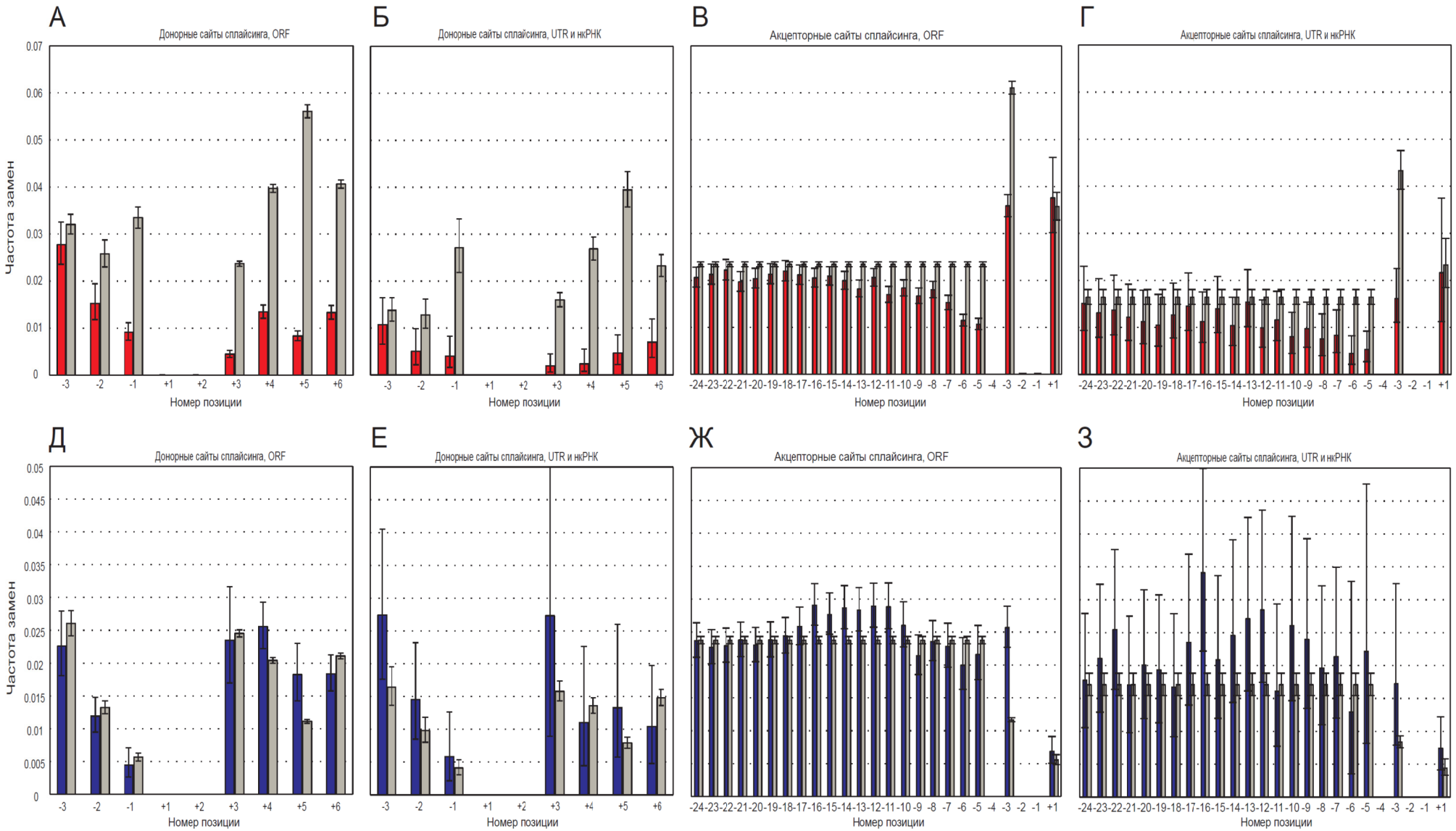


**Рисунок 15.** Частоты нуклеотидных замен в различных позициях сайтов сплайсинга на линии *D. melanogaster* после её дивергенции от линии *D. simulans*.

Обозначения те же, что на Рис. 14.

### 4.1.3. Оценка силы отбора

Вероятность фиксации единичного аллеля, на который действует отбор с коэффициентом отбора *s*, равна [126]

$$u(s) = \frac{1}{2N}\frac{4N_e s}{1 - e^{-4N_e s}}, \quad (8)$$

где *N* – реальная численность популяции, $N_e$ – эффективная численность популяции. Это вероятность равна $q_{obs}$.

Вероятность фиксации нейтрального аллеля равна $\frac{1}{2N}$. Откуда $q_{exp} = \frac{1}{2N}$. Таким образом, в предположении постоянства $4N_e s$ и скорости мутаций, одинаковой для сайта сплайсинга и нейтрального контроля, справедливо следующее соотношение:

$$\frac{q_{obs}}{q_{exp}} = \frac{4N_e s}{1 - e^{-4N_e s}}. \quad (9)$$

Зная $q_{obs}(Cn \rightarrow Nc)$ и $q_{exp}(Cn \rightarrow Nc)$ (см. формулы (6) и (7)), из этого уравнения можно вычислить силу (отрицательного) отбора (т.е. $4N_e s$), действующего на консенсусные нуклеотиды. Зная же $q_{obs}(Nc \rightarrow Cn)$ и $q_{exp}(Nc \rightarrow Cn)$, можно рассчитать силу (положительного) отбора, действующего на неконсенсусные нуклеотиды.

Если предположить, что соответствующие позиции всех сайтов сплайсинга имеют приблизительно одинаковый ландшафт приспособленности, мы ожидаем, что абсолютные значения $4N_e s$ для положительного отбора, действующего против некосенсусных нуклеотидов, и отрицательного отбора, сохраняющего консенсусные нуклеотиды, должны быть равны [130]. Эту гипотезу мы проверили в настоящей работе.

На рис. 16 представлены рассчитанные по формуле (9) значения $4N_e s$ для отбора, действующего на консенсусные и неконсенсусные нуклеотиды в разных позициях сайтов сплайсинга. Как и ожидалось, отбор во всех случаях слабый ($1 < |4N_e s| < 4$). Для акцепторных сайтов сплайсинга в целом сила отрицательного отбора, защищающего консенсусные нуклеотиды, приблизительно равна силе положительного отбора против неконсенсусных нуклеотидов. Это согласуется с

первоначальной гипотезой: слабый отбор действует на сайты с приблизительно одинаковым ландшафтом приспособленности. Слабовредные мутации (неконсенсусные нуклеотиды) фиксируются благодаря генетическому дрейфу, после чего на них начинает действовать слабый положительный отбор, стремящийся зафиксировать в популяции слабополезные мутации (консенсусные нуклеотиды). Возникает мутационно-селекционно-дрейфовое равновесие [133]. Однако во многих позициях сила отрицательного отбора превосходит силу положительного. Это верно для позиций -6, -5 и +1 акцепторных сайтов сплайсинга и почти для всех позиций донорных сайтов сплайсинга, в особенности для кодирующих сайтов в человеческой линии и для некодирующих сайтов сплайсинга в линии *D. melanogaster*.

Последнее наблюдение может означать, что по крайней мере в некоторых сайтах сплайсинга неконсенсусные нуклеотиды находятся под действием отрицательного отбора, и примесь таких сайтов снижает среднюю силу отбора, измеренную по общей выборке. Тем самым в некоторых сайтах максимальная сила не означает максимальную приспособленность. Это может происходить потому, что сильные сайты сплайсинга неэффективны (например, из-за нарушения регуляции [14,120,180]) или из-за конкурирующих селективных ограничений [181,182]. Внутри экзонных частей сайтов сплайсинга эти ограничения могут быть вызваны отрицательным отбором на частоту использования кодонов (codon usage); в интронных, равно как и в экзонных частях отрицательный отбор может быть вызван наличием сайтов связывания регуляторов сплайсинга или другими функциональными ограничениями, не связанными непосредственно с кодированием белка (например, [182]). При таком сценарии сила позиции конкретного сайта сплайсинга не может быть универсальным образом охарактеризована с использованием консенсуса, разные последовательности являются наиболее предпочтительными в разных сайтах сплайсинга.

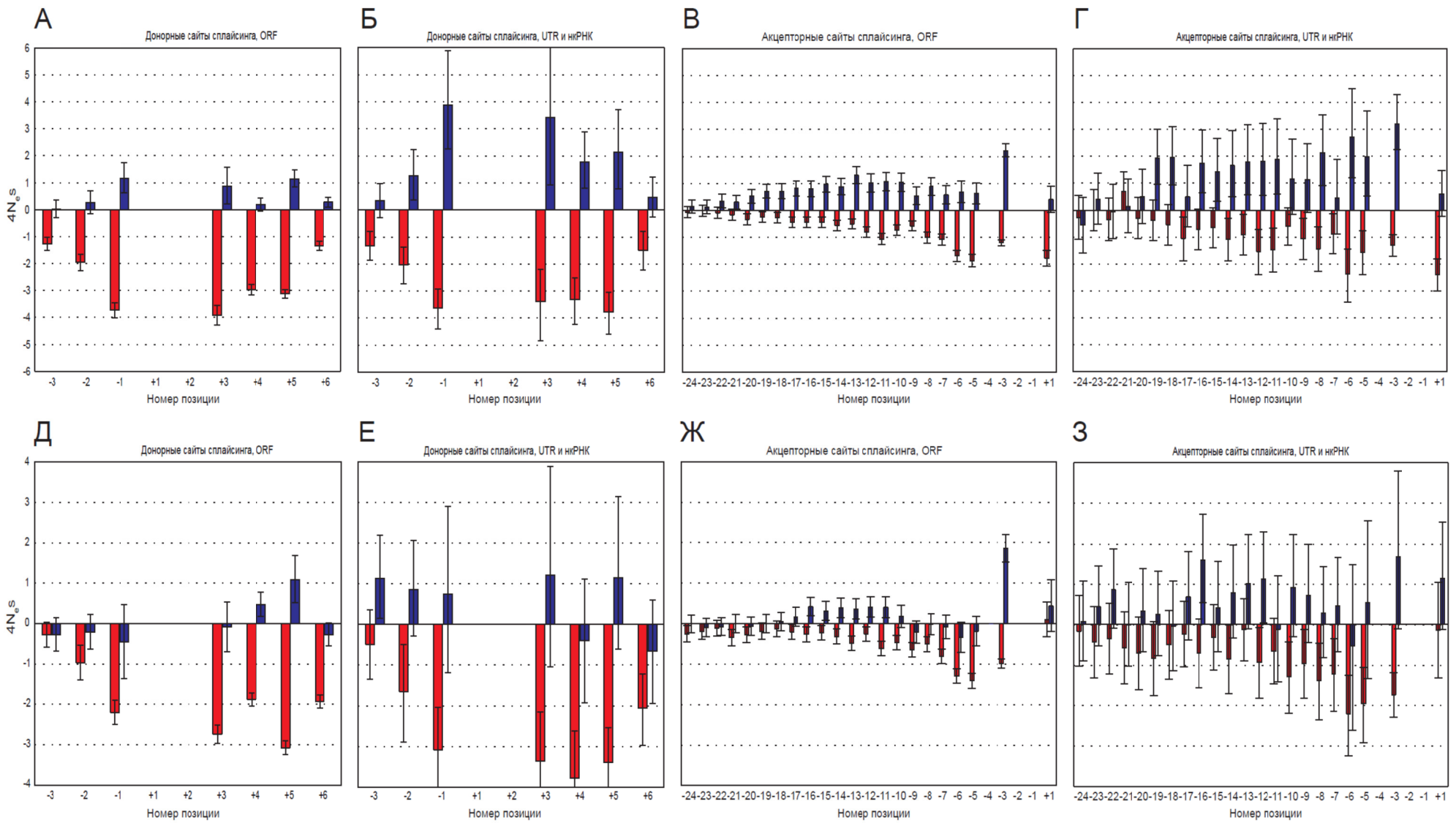


**Рисунок 16.** Сила отбора.

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены *Nc→Cn* (синие столбики) и на замены *Cn→Nc* (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А–Г – линия *H. sapiens*; Д–З – линия *D. melanogaster*. А, Б, Д, Е – донорные сайты; В, Г, Ж, З – акцепторные сайты. А, В, Д, Ж – кодирующие сайты; Б, Г, Е, З – некодирующие сайты. Усы – 95%-ные доверительные интервалы.

### 4.1.4. Сайт-специфический отбор на неконсенсусные нуклеотиды

Чтобы лучше разобраться, как отрицательный отбор воздействует на неконсенсусные и консенсусные нуклеотиды, мы сравнили их многовидовую консервативность с таковой для нейтрально эволюционирующих последовательностей. Мы оценивали консервативность нуклеотида, присутствующего в геноме *H. sapiens* или *D. melanogaster*, анализируя выравнивание с геномами 45 других видов позвоночных или 15 видов насекомых, соответственно (см. раздел 3.2.5).

Мы рассматривали относительную консервативность $c_{nuc}$ конкретного консенсусного (неконсенсусного) нуклеотида, присутствующего в данном сайте сплайсинга в геноме *H. sapiens* или *D. melanogaster,* а также относительную консервативность $c_{set}$ всех нуклеотидов, определяемых как консенсусные (или неконсенсусные).

Как и ожидалось, консенсусные нуклеотиды, наблюдаемые в геноме *H. sapiens* или *D. melanogaster* всегда более консервативны, чем нейтрально эволюционирующие последовательности. Более того, в большинстве позиций с двумя консенсусными нуклеотидами (см. рис 2), средняя консервативность $c_{set}$ двух консенсусных нуклеотидов выше, чем $c_{nuc}$ конкретного консенсусного нуклеотида, присутствующего в сайте сплайсинга, что говорит о том, что отбор предпочитает любой консенсусный нуклеотид, не пытаясь сохранить какой-то конкретный.

Для неконсенсусных нуклеотидов наблюдается более сложная картина. В позициях сайтов сплайсинга, занятых неконсенсусными нуклеотидами в человеческом геноме, этот нуклеотид в среднем консервативнее нейтральных последовательностей ($c_{nuc} > 0$, рис. 17А, 12Г). Это указывает на отрицательный отбор, действующий на неконсенсусные нуклеотиды. То же верно для неконсенсусных нуклеотидов в целом ($c_{set} > 0$); таким образом, даже переходы из неконсенсусных нуклеотидов в консенсусные бывают ограничены отбором. Однако отбор предпочитает конкретные неконсенсусные нуклеотиды, присутствующие в человеческом геноме, а не любой из них ($c_{nuc} > c_{set}$), что

справедливо для всех позиций и всех типов сайтов сплайсинга в человеческом геноме.

Следовательно, индивидуальные позиции донорных сайтов сплайсинга находятся под конфликтующими давлениями отбора: в большинстве случаев предпочитается консенсусный нуклеотид, хотя в части из них – неконсенсусный нуклеотид является предпочтительным. Если в человеческом геноме наблюдается неконсенсусный нуклеотид, то в большом числе таких случаев это нуклеотид защищен отрицательным отбором, что проявляется в том, что средняя относительная консервативность выше консервативности нейтральных последовательностей (рис. 17А-Г). Это противоречит нашему изначальному предположению, что соответствующие позиции во всех сайтах сплайсинга имеют одинаковый ландшафт приспособленности. Примесь сайтов, где на неконсенсусные нуклеотиды действует отрицательный отбор, приводит к несоответствию между силой отрицательного отбора, поддерживающего консенсусные нуклеотиды, и силой положительного отбора, стремящегося к их появлению (рис. 17 А, Б, Д, Е).

В геноме *D. melanogaster* $c_{nuc}$ и $c_{set}$ могут быть как положительными, так и отрицательными (рис. 17 Д-З), что свидетельствует о том, что универсальный отбор, поддерживающий мутации из неконсенсусных нуклеотидов в консенсусные, зачастую преобладает над сайт-специфическими предпочтениями неконсенсусных нуклеотидов. Об этом же свидетельствует большое количество отрицательных значений $c_{nuc}$ и $c_{set}$: замены происходят в среднем чаще, чем для нейтральных последовательностей, следовательно, большая доля сайтов находится под положительным отбором. Тем не менее, конкретные неконсенсусные нуклеотиды более консервативны, чем неконсенсусные нуклеотиды в целом ($c_{nuc} > c_{set}$), что говорит о том, что сайт-специфический отбор присутствует. Таким образом, в линии *H. sapiens*, сайт-специфический отбор, поддерживающий конкретные неконсенсусные нуклеотиды, встречается чаще, чем в линии *D. melanogaster*. Можно предположить, что это связано с более сложно устроенной регуляцией экспрессии генов у человека. Цис-элементы регуляции экспрессии генов на разных

уровнях (транскрипция, процессинг, трансляция, структура хроматина) могут пересекаться с сайтами сплайсинга, что вносит дополнительные селективные ограничения.

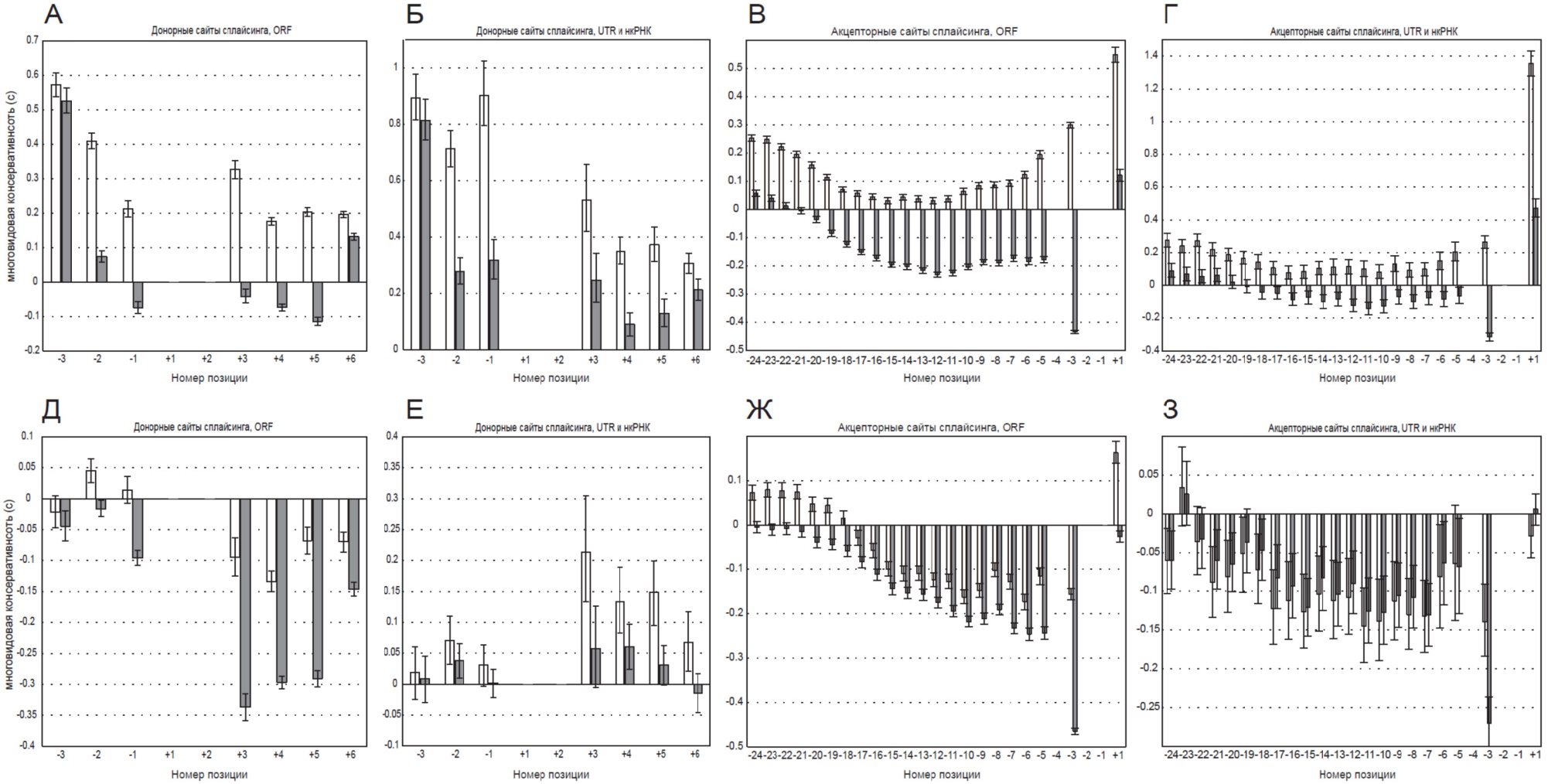


**Рисунок 17.** Многовидовая консервативность неконсенсусных нуклеотидов.

Вертикальная ось – величина *c=(obs-exp)/exp* для *Nc* нуклеотидов, где *obs* – наблюдаемая средняя многовидовая консервативность, а *exp* – ожидаемая средняя многовидовая консервативность (нейтральный контроль). Консервативность считалась как для конкретных *Nc* нуклеотидов, присутствующих в геноме *H. sapiens* или *D. melanogaster* ($c_{nuc}$, белые столбики), так и для всех нуклеотидов, определенных как *Nc* в данной позиции ($c_{set}$, серые столбики). Верхний ряд (А–Г) - *H. sapiens*; нижний ряд (Д–З) - *D. melanogaster*. А, Б, Д, Е – донорные сайты сплайсинга; В, Г, Ж, З - акцепторные сайты. А, В, Д, Ж – кодирующие сайты; Б, Г, Е, З – некодирующие сайты. Усы – 95%-ные доверительные интервалы.

Есть два (взаимно не исключающие) объяснения того, что на неконсенсусные нуклеотиды может действовать отрицательный отбор. Во-первых, может существовать отбор на то, чтобы сайт был слабым, например, для тонкой регуляции экспрессии гена. Сильные сайты сплайсинга (т.е. обогащенные консенсусными нуклеотидами) не всегда дают наибольший вклад в приспособленность. Например, в одном исследовании был сделан вывод, что на сайты сплайсинга кассетных экзонов действует отбор, направленный на поддержание их относительной слабости [180]. В нескольких экспериментальных работах показано, что мутации, усиливающие слабые сайты, могут вызывать потерю регуляции сплайлсинга [14,120]. Во-вторых, присутствующий неконсенсусный нуклеотид может быть предпочтителен из-за других селекционных требований, которые возможно не связаны со сплайсингом вовсе. Тот факт, что сила отбора, направленного на сохранение конкретных неконсенсусных нуклеотидов, выше таковой для неконсенсусных нуклеотидов в целом свидетельствует о том, что нуклеотидная идентичность, а не принадлежность к классу неконсенсусных нуклеотидов в первую очередь определяет консервативность. Таким образом, второе объяснение более вероятно, чем первое.

### 4.1.5. Отличия в силе отбора между разными классами сайтов сплайсинга

На наблюдаемые частоты замен нуклеотидов в сайтах сплайсинга и, следовательно, на оценку силы отбора могут влиять различные дополнительные факторы, кроме отбора непосредственно в сайтах, например скорость мутагенеза, отбор в сцепленных локусах и т.д. Мы оценили насколько это влияние существенно.

На соседние с сайтами локусы может действовать отрицательный отбор, что может уменьшать частоту замен в сайтах (фоновый отбор, background selection), либо положительный отбор, что приводит к уменьшению локального уровня полиморфизмов (выметание отбором, selective sweep). Мы разделили сайты на те, что находятся в участках генома с высоким и низким локальным уровнем

рекомбинации, и обнаружили, что сила отбора не зависит от уровня рекомбинации (Приложение: рис. 1), что говорит о том, что отбор в сцепленных локусах не оказывает существенного влияния на оценку силы отбора в сайтах сплайсинга.

Чтобы удостовериться, что мутационный контекст не влияет на оценку отбора, мы удалили из выборки все сайты, содержащие гипермутабильные динуклеотиды CpG [183]. Результаты, полученные на такой выборке, не отличались от исходных (Приложение: рис. 2). Не было найдено также зависимости силы отбора от уровня экспрессии соответствующего гена (Приложение: рис. 3) и от типа сплайсинга, т.е. между сайтами конститутивных и кассетных экзонов (Приложение: рис. 4).

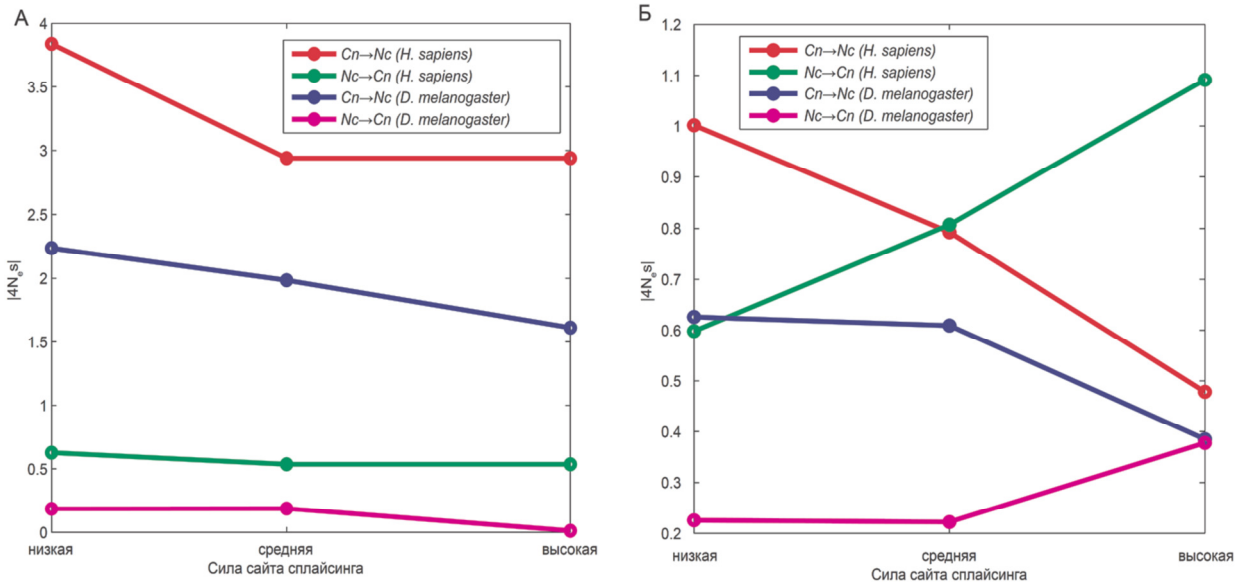


**Рисунок 18.** Зависимость силы отбора от силы сайта сплайсинга.

Сила отбора, действующего на замены *Cn→Nc* (красные и синии ломаные) и на замены *Nc→Cn* (зелёные и фиолетовые ломаные) в линии *H. sapiens* (красные и зелёные ломаные) и *D. melanogaster* (синие и фиолетовые ломаные), выраженная в абсолютных значениях $4N_e s$, усреднённых по всем позициям сайта сплайсинга. Все сайты сплайсинга были разделены на три класса по силе (низкая, средняя и высокая сила). А – донорные сайты сплайсинга; Б – акцепторные сайты сплайсинга.

Сила сайта (определяемая как количество консенсусных нуклеотидов в нём) сложным образом влияет на силу отбора (рис. 18). Единственной общей тенденцией (характерной для донорных и акцепторных сайтов как из генома *H.*

*sapiens*, так и из генома *D. melanogaster*) является отрицательная корреляция между силой сайта и силой отрицательного отбора, сохраняющего консенсусные нуклеотиды.

Наконец, результаты почти не зависели от того, каким методом, парсимонии или максимального правдоподобия, восстанавливались матрицы нуклеотидных замен (Приложение: рис. 5). В некоторых случаях, однако, метод максимального правдоподобия давал несколько более высокие значения $|4N_e s|$ для положительного отбора, действующего на неконсенсусные нуклеотиды и более низкие значения $|4N_e s|$ для отрицательного отбора, защищающего консенсусные нуклеотиды, по сравнению с методом парсимонии. Это, вероятно, связано с тем, что метод парсимонии имеет тенденцию недооценивать частоту замен из неконсенсусных нуклеотидов в консенсусные, что и приводит к переоценке силы отрицательного и недооценке силы положительного отбора.

Irimia с соавт. исследовали частоты переходов из консенсусных нуклеотидов в неконсенсусные и обратно [122]. Они обнаружили отрицательный отбор, действующий на консенсусные нуклеотиды, что согласуется с нашими данными. Однако в их исследовании частота переходов из неконсенсусных нуклеотидов в консенсусные не отличалась от таковой в нейтрально эволюционирующих участках генома. Существует ряд различий между работой Irimia с соавт. и нашим исследованием. Во-первых, они рассматривали только интронные части донорных сайтов сплайсинга, тогда как мы изучали как интронные, так и экзонные части донорных и акцепторных сайтов сплайсинга. Во-вторых, мы использовали для анализа только сайты конститутивных экзонов, а Irimia с соавт. не различали тип сплайсинга соответствующих экзонов. Хотя сайты сплайсинга конститутивных и альтернативных экзонов в нашей работе не показали статистически значимых различий в силе отбора (Приложение: рис. 4), часть сайтов альтернативных экзонов может находиться под отбором в связи с регуляцией сплайсинга, что уменьшает оценку силы положительного отбора. Кроме того, потенциальное загрязнение выборки внутренними донорными и акцепторными сайтами сплайсинга также

приводит к недооценке силы положительного отбора в связи с отбором, связанным с кодированием белка. Мы также рассматривали отдельно сайты сплайсинга, находящиеся в некодирующих сегментах РНК (UTR мРНК и некодирующие РНК) и сайты, фланкирующие интроны в кодирующей области. И наконец, мы использовали бóльшее количество сайтов сплайсинга из проекта GENCODE [184], что увеличивает статистическую значимость результатов.

### 4.1.6. Сильный положительный отбор в молодых сайтах сплайсинга, появившихся на линии *Homo sapiens* после расхождения с *Macaca mulatta*

Мы рассматриваем ситуации появления новых ключевых динуклеотидов на линии человека (*Homo sapiens*) с момента отделения её от линии макаки (*Macaca mulatta*). Наличие сайтов сплайсинга у человека определялось по аннотации GENCODE. Сайты сплайсинга в общем предке человека и шимпанзе восстанавливались парсимонией. В качестве внешней группы использовалась игрунка обыкновенная (*Callithrix jacchus*). Событием рождения нового сайта считалось появление соответствующего динуклеотида на линии человека. Мы обнаружили 1698 молодых донорных сайтов сплайсинга и 1470 - акцепторных сайтов. Часть из этих сайтов фланкирует конститутивные экзоны (914 донорных и 556 акцепторных сайтов), остальные – сайты альтернативно сплайсируемых сегментов (784 донорных и 914 акцепторных сайтов). Среди молодых сайтов альтернативно сплайсируемых сегментов значительную долю составляют сайты кассетных экзонов (543 донорных и 310 акцепторных сайтов).

Мы проанализировали отбор отдельно в донорных и акцепторных сайтах сплайсинга (конститутивные и альтернативные, а также из кодирующей и некодирующей областей были объединены в одной выборке для усиления статистического сигнала). Для каждой выборки мы оценили силу отбора, действующего на консенсусные и неконсенсусные нуклеотиды вокруг ключевых динуклеотидов, как описано в разделе 4.1.3. Как видно из рис. 19, на консенсусные нуклеотиды действует отрицательный отбор, а на некосенсусные – положительный, так же как и для старых сайтов сплайсинга, т. е. тех, у которых

ключевой динуклеотид консервативен в геномах *H. sapiens, M. mulatta* и *C. jacchus* (рис. 16). Сила отрицательного отбора на консенсусные нуклеотиды в молодых сайтах сплайсинга приблизительно совпадает с таковой в старых сайтах. Однако сила положительного отбора против неконсенсусных нуклеотидов у молодых сайтов существенно превышает таковую у старых сайтов сплайсинга, равно как и абсолютное значение силы отрицательного отбора в соответствующих позициях. Это верно для всех позиций донорных сайтов, для позиций -3, +1 акцепторных сайтов сплайсинга, однако эффект почти отсутствует в полипиримидиновом тракте акцепторных сайтов. Сила положительного отбора между старыми и молодыми сайтами может отличаться в ~10 раз (например, в позиции -3 акцепторных и +3 донорных сайтов сплайсинга).

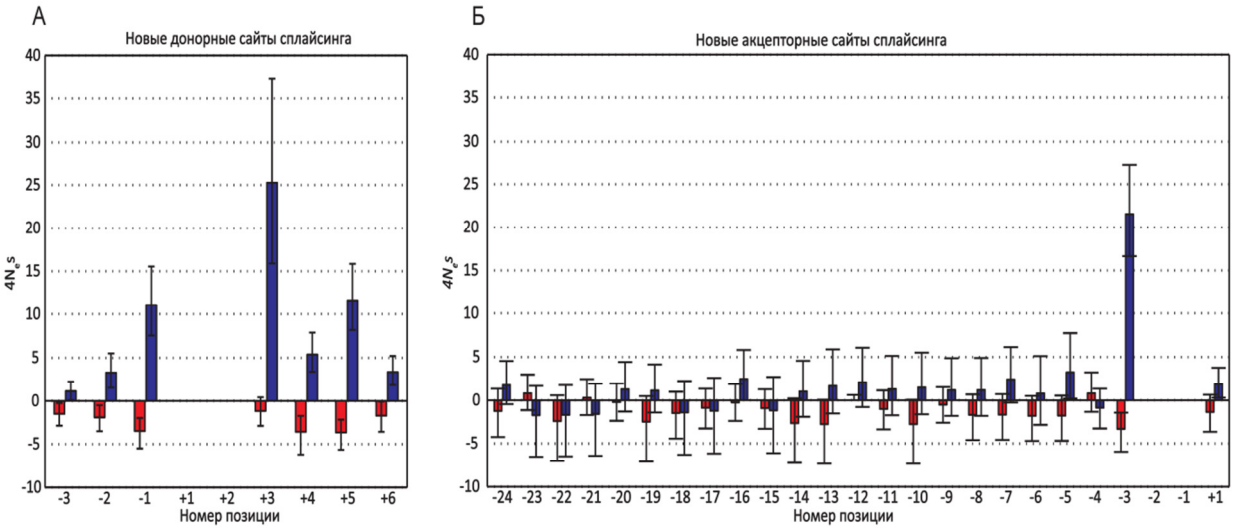


**Рисунок 19.** Сила отбора в молодых сайтах сплайсинга.

Обозначения те же, что на рис. 16. А – молодые донорные сайты сплайсинга, Б – молодые акцепторные сайты.

Таким образом, событие появления динуклеотида связано с повышенным уровнем положительного отбора на остальные нуклеотиды сайта. Дисбаланс между силой положительного отбора против неконсенсусных нуклеотидов и силой отрицательного отбора, поддерживающего консенсусные нуклеотиды, говорит о том, что молодые сайты еще не достигли мутационно-селекционно-дрейфового равновесия [133].

### 4.1.7. Дрейфовый груз и отбор на уровне целого генома

Как в геноме *H. sapiens*, так и в геноме *D. melanogaster* сила положительного и отрицательного отбора оценивается как $1 < |4N_e s| < 4$ (рис. 16). Однако эффективная численность ($N_e$) человеческой популяции составляет приблизительно $1 \times 10^4$, а популяции *D. melanogaster* — $1 \times 10^6$ [113]. Следовательно, коэффициент отбора на консенсусные нуклеотиды можно приблизительно оценить как $2 \times 10^{-5}$ для человеческой популяции и $2 \times 10^{-7}$ для популяции *D. melanogaster*. На скорость фиксаций слабовредных мутаций может влиять отбор на сцепленные локусы (эффект автостопа и фоновый отбор) [185], который, в свою очередь, зависит от эффективной численности популяции. Однако на отбор на сцепленные локусы должна оказывать существенное влияние локальная скорость рекомбинации. Мы не видим разницы в силе отбора между сайтами сплайсинга, находящимися в районах с высокой и низкой локальной скоростью рекомбинации, ни в геноме *D. melanogaster*, ни в геноме *H. sapiens* (Приложение: рис. 1).

Вероятной причиной отличия в коэффицентах отбора является эффект Ли-Акаши: функциональная единица накапливает вредные мутации при наличии отрицательного эпистаза до тех пор, пока эффект очередной вредной мутации станет достаточным, чтобы отбор мог эффективно ей противодействовать, что происходит, когда $|N_e s| \approx 1$ [132,134,186,187], т.е. при существенно меньшем коэффициенте отбора у *D. melanogast*er, чем у *H. sapiens*. Если функция, которая связывает силу отбора и текущую приспособленность, схожа в двух видах, и если эффект Ли-Акаши присутствует, мы ожидаем, что человеческие сайты сплайсинга должны быть в среднем слабее, чем у *D. melanogaster*. В обоих видах мы также ожидаем, что для слабых сайтов сплайсинга потеря консенсусных нуклеотидов более критична, чем для сильных, а приобретение консенсусных нуклеотидов должно находится под более сильным положительным отбором у слабых сайтов по сравнению с сильными (рис. 20).

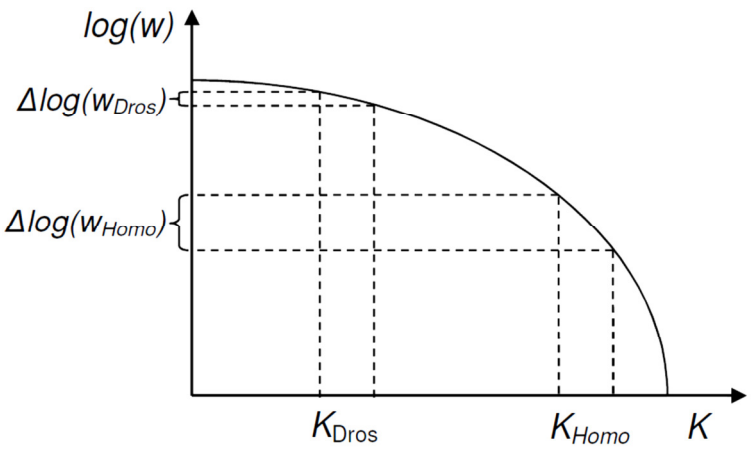


**Рисунок 20.** Гипотетическая кривая зависимости логарифма приспособленности (вертикальная ось) от числа *Nc* нуклеотидов *K* в сайте сплайсинга при наличии отрицательного эпистаза. Одна замена *Nc→Cn* у человека (Homo) ведёт к большему уменьшению приспособленности, чем такая же замена у *Drosophila*.

Действительно, средняя доля позиций акцепторных сайтов сплайсинга, занятых консенсусными нуклеотидами, ниже в геноме *H. sapiens* (77.8%), чем в геноме *D. melanogaster* (90.1%), однако для донорных сайтов эта величина приблизительно одинакова в обоих организмах (78.7% против 78.6%). В соответствии с предсказанием, сила отрицательного отбора против потери консенсусных нуклеотидов отрицательно коррелирует с силой сайта сплайсинга, как для донорных, так и для акцепторных сайтов (рис. 18А). Сила положительного отбора на приобретение консенсусных нуклеотидов сильнее у слабых акцепторных сайтов по сравнению с сильными, однако для донорных сайтов такая зависимость отсутствует (рис. 18Б). Как видно из представленных данных, эффект Ли-Акаши работает для акцепторных сайтов сплайсинга, но не для донорных. Вероятным объяснением служит то, что изначальное предположение о преобладании отрицательного эпистаза может быть неверно для донорных сайтов сплайсинга, где большое влияние оказывает внутрисайтовый положительный эпистаз (см. раздел 4.2.5.1). Так или иначе, вопрос, почему коэффициенты отбора, действующего на консенсусные нуклеотиды сайтов сплайсинга на линиях *D.*

*melanogaster* и *H. sapiens*, отличаются на два порядка остаётся открытым, т.к. указанные выше соображения не могут на него полностью ответить.

Неконсенсусные нуклеотиды занимают значительную долю позиций сайтов сплайсинга, т.е. несут существенный дрейфовый груз. Суммарно в человеческом геноме имеется приблизительно 150000 конститутивных донорных и приблизительно 150000 конститутивных акцепторных сайтов сплайсинга (в соответствии с аннотацией GENCODE [184]). Средний донорный сайт содержит 2.92 неконсенсусных нуклеотида, а средний акцепторный – 7.10 неконсенсусных нуклеотидов. Т.к. коэффициент отбора, действующий на один локус, занятый неконсенсусным нуклеотидом, составляет $2 \times 10^{-5}$, то сумма коэффициентов отбора против всех неконсенсусных нуклеотидов (дрейфовый груз) составляет приблизительно 10 для донорных сайтов и 20 для акцепторных сайтов сплайсинга. Эти значения могут быть несколько переоценены, т.к. на часть неконсенсусных нуклеотидов положительный отбор не действует (см. раздел 4.1.4). Однако такой груз слишком велик и может быть объяснён тем, что глобально отбор является скорее эпистатическим, чем локус-независимым, и/или скорее стабилизирующим, чем очищающим [134,188]. Аналогичные расчёты для *D. melanogaster* позволяют оценить дрейфовый груз как $2 \times 10^{-2}$ и $7 \times 10^{-2}$ для конститутивных донорных и акцепторных сайтов, соответственно.

### 4.1.8. Свидетельства отбора на уровне однонуклеотидных полиморфизмов

Мы изучали сегрегирующие однонуклеотидные полиморфизмы в сайтах сплайсинга конститутивных экзонов в популяциях *H. sapiens* и *D. melanogaster*. Мы различали предковый и производный аллель. Предковый аллель (нуклеотид) определялся как совпадающий с ортологичным нуклеотидом в геноме организма, являющегося внешней группой (*Pan troglodytes* для популяции *H. sapiens* и *D. simulans* для популяции *D. melanogaster*). Мы исключили из выборки полиморфизмы с частотой производного аллеля больше 99% и меньше 1%. Аналогично подходу с дивергенцией, мы рассматривали переходы из неконсенсусных нуклеотидов (предковый аллель) в консенсусные (производный

аллель) и обратно. Мы сравнивали полученные значения с таковыми в нейтральном контроле (построение нейтральных контролей аналогично тому, что описано в разделе 4.1.1). Мы наблюдали пониженную частоту переходов из консенсусных нуклеотидов в неконсенсусные, что свидетельствует об отрицательном отборе. Обратные переходы в некоторых случаях быстрее, а в некоторых случаях не отличаются от нейтральных, что не противоречит теории, т.к. аллели под положительным отбором быстро фиксируются и долго не могут пребывать в сегрегирующем состоянии (*H. sapiens* - рис. 21, *D. melanogaster* – Приложение: рис. 6)

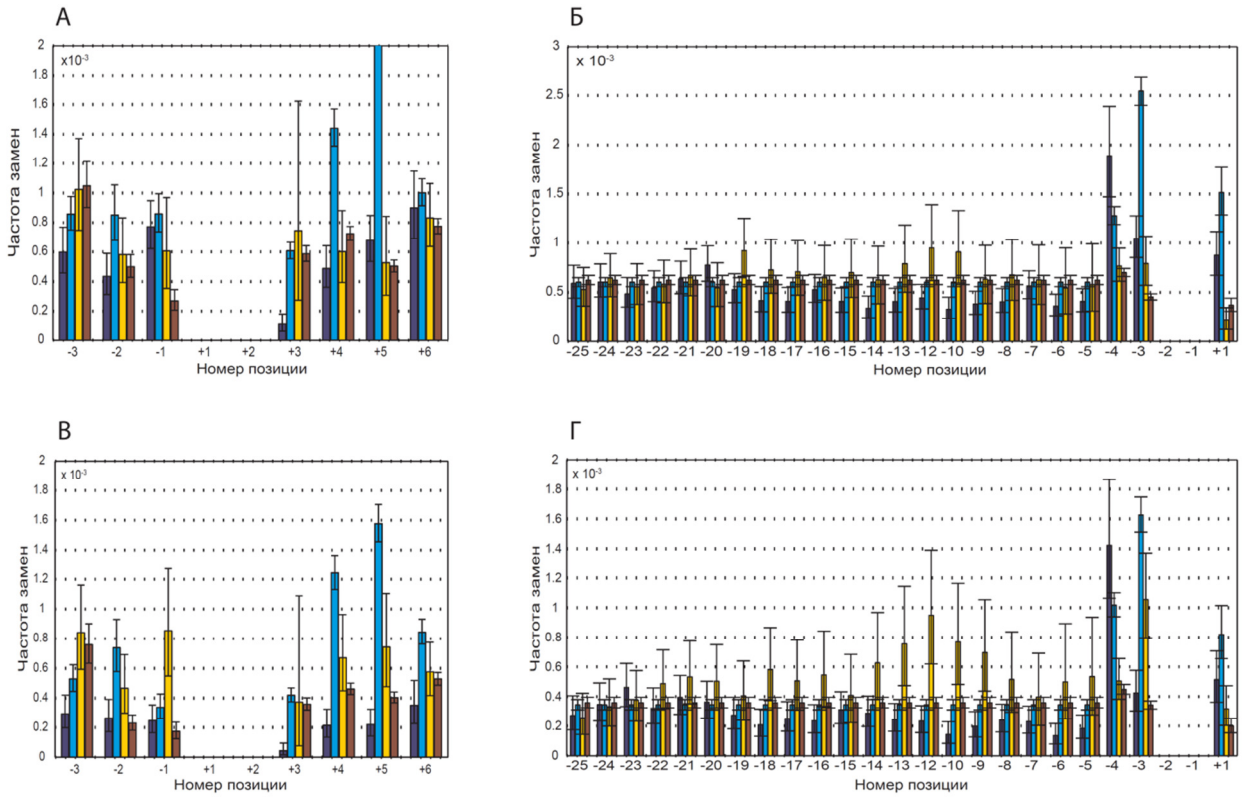


**Рисунок 21**. Сегрегирующие полиморфизмы в сайтах сплайсинга *H. sapiens*.

По горизонтальной оси – частота замен между предковыми и производными аллелями (рассматривались замены из *Cn* в *Nc* и обратно), по вертикальной оси – номер позиции в сайте сплайсинга. Изображена частота замен *Cn→Nc* в сайтах сплайсинга (синие столбики) и в близлежащих интронах (голубые столбики); замен *Nc→Cn* в сайтах сплайсинга (желтые столбики) и в близлежащих интронах (коричневые столбики). А,В – донорные сайты сплайсинга; Б, Г – акцепторные сайты. А, Б – частота производного аллеля в сайтах сплайсинга 1-10%; В, Г - частота производного аллеля в сайтах сплайсинга 25-99%;

Мы разделили все полиморфизмы на три класса в соответствии частотой производного аллеля: (1) 1–10%, (2) 10–25% и (3) 25–99%. Были посчитаны частоты переходов между консенсусными и неконсенсусными нуклеотидами отдельно в каждом из классов. Оказалось, что чем более высокочастотный класс мы рассматриваем, тем сильнее частота переходов из неконсенсусных нуклеотидов в консенсусные отличается от нейтральной (ср. рис. 21А и В, рис. 21Б и Г). Это верно как для популяции *H. sapiens* (рис. 21) так и *D. melanogaster* (Приложение: рис. 6). Данное наблюдение согласуется с гипотезой о положительном отборе, т.к. производные консенсусные аллели, находящиеся в процессе фиксации должны в среднем иметь более высокую частоту в популяции, чем нейтральные аллели.

Таким образом, наличие отрицательного отбора против неконсенсусных нуклеотидов и положительного отбора на вновь возникшие консенсусные нуклеотиды подтверждено как на межвидовом, так и на внутрипопуляционном уровне.

## 4.2. Коррелированная эволюция позиций в сайтах сплайсинга млекопитающих

Мы систематически изучили эволюцию сайтов сплайсинга в линиях млекопитающих, находящихся на средних филогенетических расстояниях друг от друга: человека (*Homo sapiens*), мыши (*Mus musculus*) и собаки (*Canis familiaris*).

### 4.2.1. Метод восстановления матриц нуклеотидных замен в сайтах сплайсинга

Мы разработали метод восстановления матриц нуклеотидных замен, который лишен некоторых недостатков, присущих методу максимальной парсимонии.

Описание метода. Внутренняя ветка на филогенетическом дереве млекопитающих, соединяющая общего предка *H. sapiens*, *M. musculus* и общего предка всех трёх организмов (*H. sapiens*, *M. musculus*, *C. familiaris*) относительно коротка (см. например, [174]). Поэтому мы упростили модель, приняв, что *H. sapiens*, *M.*

*musculus* и *C. familiaris* произошли от общего предка в результате трифуркации (рис. 22).

Для анализа эволюционного процесса каждой из трёх веток и каждой позиции сайта сплайсинга была поставлена в соответствие матрица нуклеотидных замен ***P*** размером 4×4. Элементы этой матрицы – это условные вероятности $p(X|Z)$ того, что нуклеотид $Z$ в данной позиции предкового генома заменился на нуклеотид $X$ в геноме потомка (*H. sapiens*, *M. musculus* или *C. familiaris*). В описании эволюционного процесса также присутствует вектор предковых вероятностей $\boldsymbol{p(Z)} = (p(Z=A), p(Z=C), p(Z=G), p(Z=T))$. Например, если мы знаем число замен в какой-то позиции сайта сплайсинга, скажем, из A в G, произошедших от общего предка до человека $\#(A \rightarrow G)$, и число аденинов $\#A$ в общем предке (а также $\#C$, $\#G$ и $\#T$), оценкой $p(G|A)$ будет величина $\#(A \rightarrow G) / \#A$, а оценкой $p(Z=A)$ будет $\#A/(\#A + \#C + \#G + \#T)$. В реальности число замен нам неизвестно. Но мы можем оценить $p(Z)$ и $P$ как описано ниже.

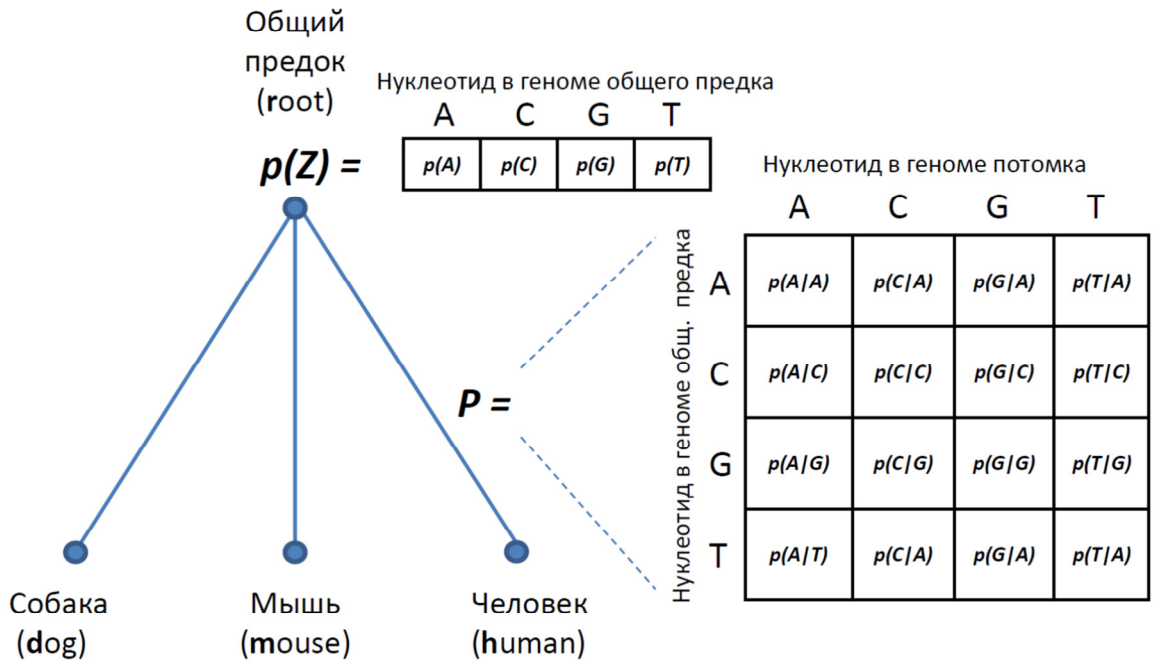


**Рисунок 22.** Матрица эволюционных замен и вектор предковых вероятностей.

Использовалось упрощенное филогенетическое дерево, где линии человека, мыши и собаки отходят одновременно от общего предка (трифуркация). Для каждой позиции и каждой ветки филогенетического дерева была реконструирована своя матрица ***P***. Строки соответствуют нуклеотидам в геноме общего предка, столбцы – в геноме потомка (на рисунке – человек как пример). Элементом матрицы ***P*** является условная вероятность *p(X|Z)* того, что нуклеотид *Z* в данной позиции в геноме общего предка был заменён на нуклеотид *X* в геноме потомка. Элементы вектора предковых вероятностей ***p(Z)*** – это вероятности каждого нуклеотида в геноме предка.

Для оценки ***p(Z)*** для каждой позиции в сайте сплайсинга и ***P*** для каждой позиции и на каждой ветки дерева, была использована модель нуклеотидных замен с максимальным возможным количеством параметров [189]; иными словами, не накладываются ограничения на относительные скорости нуклеотидных замен для разных пар нуклеотидов. Для каждой колонки в тройном выравнивании сайтов сплайсинга (соответствующей определенной позиции) нуклеотиды в геномах общего предка (**r**oot), человека (**h**uman), мыши (**m**ouse) и собаки (**d**og) были обозначены как *r*, *h*, *m*, *d*, соответственно. Были взяты колонки выравниваний всех сайтов сплайсинга, соответствующие определенной позиции *(hmd)* (здесь и далее, тройки), и для каждой из 64 возможный троек (AAA, AAC, …, TTT) рассчитывались их частоты q(AAA), q(AAC),…, q(TTT). Разработанный нами метод основан на минимизации разницы оцененных из модели вероятностей троек p(AAA), p(AAC),…, p(TTT) и соответствующих наблюденных частот троек q(AAA), q(AAC),…, q(TTT). В нашей модели вероятности троек оцениваются следующим образом.

$$\begin{aligned} p(AAA) &= \sum_r p(r)\,p(h=A\mid r)\,p(m=A\mid r)\,p(d=A\mid r) \\ p(AAC) &= \sum_r p(r)\,p(h=A\mid r)\,p(m=A\mid r)\,p(d=C\mid r) \\ &\dots \\ p(TTT) &= \sum_r p(r)\,p(h=T\mid r)\,p(m=T\mid r)\,p(d=T\mid r) \end{aligned} \qquad (10)$$

Таким образом, получается система из 64 уравнений с 39 независимыми переменными (три матрицы с 12 независимыми переменными в каждой и вектор предковых вероятностей с тремя независимыми переменными). Приближенное

решение этой переопределенной системы уравнений было получено методом наименьших квадратов с испольованием функции fmincon() в среде MATLAB. Мы минимизировали функцию $f(P, p(Z)) = (p(AAA)-q(AAA))^2 + (p(AAC)-q(AAC))^2 + … + = (p(TTT)-q(TTT))^2$. Мы начинали со случайных точек 10 раз и запоминали наименьший из полученных локальных минимумов. В большинстве случаев на каждой из 10 итераций мы получали тот же самый минимум.

Дополнительно, мы сравнивали результаты полученные нашим методом, с результатами применения стандартного метода максимальной экономии. Так как в методе максимальной экономии минимизируется число замен на дереве, во всех матрицах нуклеотидных замен, полученных этим методом, вероятность нуклеотида в какой-либо позиции сайта сплайсинга остаться неизменным в течение эволюции от предка к потомку почти всегда выше, чем при использовании нашего метода. По-видимому, это наблюдение объясняется тем, что наш метод, в отличие от метода максимальной экономии, способен учитывает двойные замены.

### 4.2.2. Оценка вероятностей последовательностей предковых сайтов сплайсинга в позиционно-независимой модели

Нами был разработан метод, посредством которого для каждой заданной тройки ортологичных сайтов сплайсинга (в геномах человека, мыши и собаки) можно оценить (1) вероятность любой заданной последовательности в геноме общего предка, от которой они произошли (см. данный раздел); (2) оценить среднюю силу сайта сплайсинга в геноме общего предка (см. раздел 4.2.3), предполагая позиционную независимость.

Предположим, что отдельные позиции сайтов сплайсинга в геноме общего предка взаимно независимы, тогда вероятность наблюдать последовательность *R* (**R**oot) в геноме общего предка, зная ортологичные последовательности *H*, *M* и *D* в геномах человека (**H**uman), мыши (**M**ouse) и собаки (**D**og), равна

$$p(R \mid HMD) = \prod_{i=1}^{L} p(r_i \mid h_i m_i d_i),$$

где $p(r_i|h_i m_i d_i)$ – вероятность нуклеотида $r_i$ в позиции $i$, при условии наблюдения нуклеотидов $h_i$, $m_i$ и $d_i$ в соответствующих позициях геномов человека, мыши и собаки.

Так как эволюция на отдельных ветвях происходит независимо,

$$p(r_i h_i m_i d_i) = p(r_i) p(h_i \mid r_i) p(m_i \mid r_i) p(d_i \mid r_i).$$

По теореме Байеса, имеем:

$$p(r_i \mid h_i m_i d_i) = \frac{p(r_i h_i m_i d_i)}{\sum_{x \in \{A,C,G,T\}} p(x h_i m_i d_i)} = \frac{p(r_i) p(h_i \mid r_i) p(m_i \mid r_i) p(d_i \mid r_i)}{\sum_{x \in \{A,C,G,T\}} p(x) p(h_i \mid x) p(m_i \mid x) p(d_i \mid x)} \qquad (11)$$

Следовательно, вероятность наблюдать последовательность $R$ в геноме общего предка, зная для каждой позиции $i$ вектор предковых вероятностей (набор $p(r_i)$) и матрицы переходных вероятностей (наборы $p(h_i|x)$, $p(m_i|x)$, $p(d_i|x)$), равна

$$p(R \mid HMD) = \prod_{i=1}^{L} \frac{p(r_i) p(h_i \mid r_i) p(m_i \mid r_i) p(d_i \mid r_i)}{\sum_{x \in \{A,C,G,T\}} p(x) p(h_i \mid x) p(m_i \mid x) p(d_i \mid x)}. \qquad (12)$$

Рассмотрим тройку ортологичных сайтов сплайсинга. Если в геномах человека, мыши и собаки в $i$-той позиции наблюдаются соответственно нуклеотиды $h_i$, $m_i$ и $d_i$, то математическое ожидание силы данной позиции в общем предке равно

$$\overline{W}(i) = \sum_{r_i} W(r_i) p(r_i \mid h_i m_i d_i), \qquad (13)$$

где $r_i$ пробегает все возможные нуклеотиды, а $p(r_i|h_i m_i d_i)$ считается по формуле (11).

### 4.2.3. Изменение силы сайтов в ходе эволюции. Проверка гипотезы о миграции сигнала

Нуклеотидные замены в сайтах сплайсинга могут приводить к изменениям силы соответствующих позиций. Мы изучили, как изменяется средняя сила индивидуальных позиций сайтов на разных ветках филогенетического дерева.

Для каждой позиции *i* в сайте сплайсинга мы определили *наблюдаемое* (**O**bserved) среднее изменение силы $\Delta\hat{W}_O(i)$ как разницу между средней наблюдаемой силой в геноме потомка и средней силой в геноме общего предка (посчитанной по формуле (13)). Эту величину мы сравнивали с *ожидаемым* (**E**xpected) средним изменением силы в *i*-той позиции $\Delta\hat{W}_E(i)$, то есть изменением силы, которое произошло, если бы позиции эволюционировали нейтрально (см. раздел 3.2.4). Как и ожидалось, $\Delta\hat{W}_E(i)$ отрицательна для всех позиций донорных и акцепторных сайтов сплайсинга (Приложение: рис. 7, 8). Это означает, что если бы сайты эволюционировали нейтрально, они становились бы менее похожи на консенсусную последовательность. Очевидно, что наблюдаемая эволюция не нейтральная: сайты могут как приближаться к консенсусу ($\Delta\hat{W}_O(i) > 0$), так и удаляться от него ($\Delta\hat{W}_O(i) < 0$). Для того чтобы иметь возможность сравнивать изменения силы между позициями и выборками сайтов, мы определили относительное изменение силы позиции $\omega = \frac{\Delta\hat{W}_E - \Delta\hat{W}_O}{\Delta\hat{W}_E}$. Если сила позиции не меняется, то ω = 1; в случае нейтральной эволюции позиции ω = 0. Поскольку $\Delta\hat{W}_E(i) < 0$, то ω > 1 соответствует увеличению силы, а при ω < 1 сила позиции уменьшается.

В таблицах 1 и 2 представлены значения ω для донорных и акцепторных сайтах сплайсинга, соответственно. Положительные значения ω во всех позициях донорных и в большинстве позиций акцепторных сайтов сплайсинга свидетельствуют о том, что сила позиции находится под действием отрицательного (стабилизирующего) отбора. Сила многих позиций статистически значимо меняется в одной или нескольких филогенетических линиях. В целом картина изменений

веса весьма сложна: положительные и отрицательные значения изменения силы позиций (ω > 1 и ω < 1, соответственно) встречаются как в донорных, так и в акцепторных сайтах сплайсинга. В позиции -4 акцепторных сайтов встречается даже отрицательные значения ω, что говорит о том, что эта позиция изменятся быстрее, чем ожидается при нейтральности. Однако в донорных сайтах конститутивных экзонов наблюдается следующая тенденция: экзонная часть сайта в среднем уменьшает свою силу, тогда как интронная часть, напротив, усиливается.

**Таблица 1.** Относительное изменение веса ω для разных позиций донорных сайтов сплайсинга, в течение их эволюции от общего предка человека, мыши и собаки до каждого из трёх видов.

| Позиция | Конститутивные экзоны | | | Кассетные экзоны | | |
|---|---|---|---|---|---|---|
| | Собака | Человек | Мышь | Собака | Человек | Мышь |
| **-3** | 0.58 | 0.91 | 1.06 | 0.22 | 10.78 | 1.46 |
| **-2** | **0.80** | **<u>0.47</u>** | **<u>0.67</u>** | 1.00 | 1.00 | 0.99 |
| **-1** | 1.05 | 0.97 | 0.92 | 1.20 | 0.94 | 1.17 |
| **3** | 1.00 | **<u>1.05</u>** | 1.01 | 0.99 | 1.04 | **<u>0.95</u>** |
| **4** | 0.99 | **1.06** | 1.00 | 1.01 | **<u>0.82</u>** | 0.99 |
| **5** | 0.99 | 1.00 | **1.04** | 1.05 | 1.02 | **<u>1.09</u>** |
| **6** | 1.10 | 1.03 | 0.98 | **1.23** | **0.66** | **<u>0.79</u>** |

ω > 1 соответствует увеличению силы позиции, ω < 1 – уменьшению веса (см. основной текст). Значения в таблице выделены в соответствии с p-значением (отклонение от нулевой гипотезы: ω = 1): **жирный шрифт**, $p < 0.1$; **<u>жирный и подчеркнутый шрифт</u>**, $p < 0.05$.

Описанные результаты могут зависеть от того, насколько корректно были восстановлены ортологичные функциональные сайты сплайсинга в других видах. Возможной проблемой может являться то, что одни сайты сплайсинга в течение эволюции могут исчезнуть, другие — появиться, также возможно радикальное изменение действия отбора на них. Примесь нефункциональных сайтов в каких-то видах может быть источником систематической погрешности при оценке изменения силы.

**Таблица 2.** Относительное изменение веса ω для разных позиций акцепторных сайтов сплайсинга, в течении их эволюции от общего предка человека, мыши и собаки до каждого из трёх видов.

| Позиция | Конститутивные экзоны | | | Кассетные экзоны | | |
|---|---|---|---|---|---|---|
| | Собака | Человек | Мышь | Собака | Человек | Мышь |
| -13 | 1.14 | 0.95 | 0.97 | **0.25** | 1.03 | **0.74** |
| -12 | **0.75** | **0.67** | 0.92 | **0.59** | 0.55 | **0.71** |
| -11 | **0.84** | 0.97 | 0.94 | 0.91 | **0.61** | 1.15 |
| -10 | 0.92 | 0.92 | 1.04 | 1.11 | 1.04 | 0.76 |
| -9 | 1.01 | 0.87 | 1.02 | 0.90 | 0.66 | 1.01 |
| -8 | **0.85** | 1.08 | 1.05 | 0.91 | **0.39** | 1.22 |
| -7 | 1.01 | 1.04 | 0.98 | 0.62 | **0.21** | 1.15 |
| -6 | 0.95 | **1.16** | 1.00 | 0.97 | **0.45** | 1.12 |
| -5 | 0.97 | 1.10 | **1.13** | **0.78** | 0.75 | 0.91 |
| -4 | **-0.40** | **-1.30** | **0.21** | -4.09 | 1.87 | **-1.36** |
| -3 | **0.96** | **0.95** | 0.99 | 1.03 | 1.01 | **1.07** |
| 1 | **0.96** | 0.97 | 1.02 | **0.90** | 0.94 | 0.97 |
| 2 | 0.99 | **1.50** | **0.83** | **-0.38** | 2.22 | 0.81 |
| 3 | -1.97 | 0.98 | 1.01 | 2.62 | **-1.40** | **-1.76** |

Обозначения те же, что и в Таблице 1.

Чтобы избежать этой проблемы, мы старались включить в выборку только те сайты сплайсинга, которые функциональны во всех трёх видах. В частности, мы изучали только те тройки сайтов, в которых консервативен ключевой динуклеотид (AG или GT). Однако, возможна ситуация, когда ключевой нуклеотид сохранился, однако сайт нефункционален. Функциональность сайтов сплайсинга может быть формально доказана путём наблюдения соответствующих изоформ мРНК. Однако если требовать поддержки во многих видах, это приводит к систематической недопредставленности относительно молодых сайтов, большинство из которых используется редко [166] и, следовательно, могут быть потеряны в одной из выборок EST. Кроме того, количество доступных EST для собаки невелико, а значит, если требовать поддержку EST сразу в трёх видах, включая собаку, исследуемая выборка уменьшится на порядок. Поэтому мы использовали поддержку EST как доказательство функциональности сайта только для человека и мыши, именно на этой выборке были получены все результаты. Чтобы оценить, не привносим ли мы

существенных искажений в нашу выборку, когда требуем поддержки не только в человеке, но и в мыши, мы использовали ~ в 4 раза б**о**льшую выборку, где EST-поддержка требовалась только в человеческом геноме. Ключевые параметры выборки не сильно изменились: соотношение числа конститутивных и кассетных экзонов почти не поменялось, то же можно сказать о распределении уровней включения кассетных экзонов (Приложение: табл. 1). Повторив анализ изменения сил позиций на большой выборке, мы обнаружили систематическую погрешность: сила сайта в среднем убывала на линиях мыши и собаки и возрастала на линии человека (данные не показаны). Так как поддержка EST в большой выборке производилась только для человека, то соответствующие выборки сайтов сплайсинга для мыши и собаки скорее всего обогащены редко используемыми или вовсе нефункциональными сайтами, что и приводит к описанной погрешности. В целом при интерпретации результатов стоит подходить с большей осторожностью к тенденциям, которые наблюдаются только на одной ветке филогенетического дерева, чем к тем, что подтверждаются на всех трёх ветках.

Даже если функциональность сайта сплайсинга сохранилась, тип сплайсинга мог измениться (конститутивный экзон стал кассетным или наоборот) между исследуемыми видами. Однако доля сайтов сплайсинга, для которых поменялся тип сплайсинга между человеком и мышью низка. Если использовать транскрипционные данные мыши, а не человека для классификации типов сплайсинга, результаты почти не меняются (данные не показаны).

Возможной причиной описанного ослабления экзонной части и усиления интронной части является миграция сигнала из экзонной части донорного сайта сплайсинга в интронную в течении эволюции. Этот феномен был ранее описан для интронов, которые вставились до дивергенции основных эукариотических групп [107]. Мы имеем дело с более близкими эволюционными временами и не различаем очень древние и относительно новые интроны, так как большинство рассматриваемых интронов появилось до существования общего предка человека мыши и собаки. Однако довольно много интронов появилось при переходе к

многоклеточности [104] или, по другой версии, в линии позвоночных после их дивергенции от насекомых [97], поэтому значительная часть исследуемых интронов могут быть относительно новыми, а в их сайтах происходить миграция сигнала. В донорных сайтах кассетных экзонов, однако, признаков миграции сигнала не наблюдается. Это может объясняться недостаточным количеством данных или тем, что кассетные экзоны имеют более сложную систему регуляции [14], которая вносит дополнительные ограничения, не дающие изменяться сайтам значительно. Признаков миграции сигнала не наблюдается в акцепторных сайтах сплайсинга, что согласуется со сделанным ранее наблюдением, что интронные части акцепторных сайтов разного возраста существенно не отличаются по силе [107].

### 4.2.4. Нуклеотидные замены в индивидуальных позициях сайтов сплайсинга

Мы изучили, какие нуклеотидные замены привели к изменениям силы позиций, перечисленных в таблицах 1 и 2. На рисунках 8 и 9 приложения показаны нуклеотидные замены, которые происходят значимо чаще или реже ($p < 0.05$), чем ожидается при нейтральном режиме эволюции. Наблюдаемые изменения силы обусловлены позиционно-специфическими изменениями в частотах нуклеотидов. Стоит отметить, что статистически значимые изменения частот нуклеотидов не полностью объясняют все значимые изменения сил позиций, и обратно: некоторые изменения частот не приводят к значимому изменению силы соответствующих позиций. Большинство случаев усиления или ослабления позиции обусловлено изменением частоты одного или двух (как в полипиримидиновом тракте акцепторных сайтов сплайсинга) наиболее частых нуклеотидов.

**Донорные сайты сплайсинга.** Ослабление позиции -2(A) донорных сайтов сплайсинга, фланкирующих конститутивные экзоны, в линиях человека и мыши (табл. 1) происходит из-за того, что падает частота самого частого (сильного) нуклеотида A и увеличивается частота наиболее редкого нуклеотида G (см. Приложение: рис. 9). Уменьшение силы позиции +4 в донорных сайтах кассетных

экзонов (на линии человека) вызвано падением частоты самого частого нуклеотида A. Рост частоты самого сильного нуклеотида A, сопровождаемый падением численности самого слабого нуклеотида G в позиции +3 донорных сайтов конститутивных экзонов, ведёт к усилению этой позиции на линии человека. Усиление позиции +5 в линии мыши в основном вызвано уменьшением частоты наиболее частого нуклеотида G.

**Акцепторные сайты сплайсинга.** Уменьшение силы позиции -3(C) в линиях мыши и собаки в выборке конститутивных экзонов вызывается уменьшением частоты самого частого нуклеотида C и увеличением численности второго по частоте нуклеотида T (Приложение: рис. 10). Противоположная тенденция, усиление позиции -3, вызванная ростом частоты C, наблюдается на выборке акцепторных сайтов кассетных экзонов в линии мыши. В позиции -4(N) частоты нуклеотидов отличаются друг от друга не сильно, однако сила этой позиции сильно уменьшается (в выборке сайтов конститутивных экзонов на линиях мыши и собаки и в выборке сайтов кассетных экзонов на линии мыши). Это объясняется тем, что частоты нуклеотидов в этой позиции всё же не одинаковы: самым частым нуклеотидом является T (~30%), а самым редким G(~20%). Ослабление этой позиции вызвано в основном увеличением частоты G и снижением частоты T. Наиболее существенной тенденцией в полипиримидиновом тракте акцепторных сайтов сплайсинга является снижение частоты C и увеличение частоты T, таким образом, полипиримидиновый тракт обогащается тиминами. Однако из-за того, что T встречается лишь немного чаще, чем C, указанная тенденция не приводит к консистентным изменениям веса позиций в полипиримидиновом тракте.

В нескольких экспериментальных работах было показано, что в целом увеличение числа тиминов в полипиримидиновом тракте улучшает использование соответствующего акцепторного сайта сплайсинга [190,191]. В то же время, изолированные цитозины в T-богатом полипиримидиновом тракте могут увеличивать частоту использования акцепторного сайта [191,192], хотя несколько подряд идущих цитозинов могут полностью лишить акцепторный сайт способности

сплайсироваться [190,192]. Недавние структурные исследования взаимодействия U2AF65 с полипиримидиновым трактом несколько прояснили картину. U2AF65 содержит два РНК-распознающих мотива, которые связываются с полипиримидиновым трактом: RRM1 и RRM2. RRM2 взаимодействует с 5'-проксимальным участком полипиримидинового тракта и предпочитает связываться с тиминами. RRM2 в свою очередь взаимодействует с участком, находящемся ближе к 3'-концу полипиримидинового тракта и является менее избирательным: для него подходят как тимины, так и цитозины. Конформационные изменения U2AF65 могут также помочь ему связываться с полипиримидиновыми трактами с разным нуклеотидным составом. Таким образом, баланс нескольких давлений отбора приводит в итоге к накоплению нуклеотидов Т в полипиримидиновом тракте.

Итак, в большинстве случаев изменение силы вызывается увеличением или уменьшением частоты консенсусного нуклеотида в данной позиции (например, нуклеотида А в позиции -2 донорного сайта). В полипиримидиновых трактах акцепторных сайтов сплайсинга, однако, происходят замены между двумя самыми частыми нуклеотидами (C→T).

### 4.2.5. Сила нуклеотидов в различных позициях сайтов сплайсинга взаимно скоррелирована

Вероятность того, что данный нуклеотид находится в некоторой позиции сайта сплайсинга, может зависеть от того, какие нуклеотиды находятся в других позициях. Мы переформулировали это в терминах силы позиций: зависит ли сила одной позиции от силы другой позиции? Для того, чтобы ответить на этот вопрос, мы получили (отдельно для выборок донорных и акцепторных сайтов) ковариационные матрицы векторов, соответствующих сайтам сплайсинга, элементами такого вектора являются силы нуклеотидов в позициях сайта.

Ковариационные матрицы векторов сил позиций донорных и акцепторных сайтов сплайсинга конститутивных экзонов из генома мыши показаны на рисунках 23 и 24, соответственно (здесь и далее инвариантные динуклеотиды AG в акцепторных

сайтах и GT в донорных сайтах не рассматриваются). Аналогичные матрицы для геномов человека и собаки выглядят очень сходным образом (данные не показаны).

Теоретически, наблюдаемые ковариации могут вызываться несколькими причинами: (1) мутационные эффекты (мутации происходят в определенном контексте), (2) отборные эффекты, а именно эпистатический отбор (вклад в приспособленность нуклеотида, находящегося в одной позиции зависит от нуклеотида в другой позиции), (3) исторические причины (статистический эффект смешивания в выборке сайтов разных возрастов). Мы рассмотрели каждую из этих возможностей и выяснили, какая из них наилучшим образом согласуется со сделанными наблюдениями.

#### *4.2.5.1. Эпистаз – наиболее вероятная причина ковариаций в донорном сайте сплайсинга*

Сила каждой позиции в экзонной части донорных сайтов сплайсинга отрицательно коррелирует (ковариирует) с силой каждой позиции в интронной части. Это означает, что сайты с сильной экзонной частью имеют тенденцию иметь слабую интронную часть и наоборот. Напротив, позиции внутри экзонной части положительно скоррелированы друг с другом. Позиции интронной части донорных сайтов сплайсинга +4(A), +5(G) и +6(T) также положительно скоррелированы, однако между позицией +3 и другими позициями интронной части наблюдается отрицательная корреляция. Описанные наблюдения сохраняются и для сайтов кассетных экзонов, причем отрицательные корреляции между экзонной и интронной частями даже более выражены, чем на выборке донорных сайтов конститутивных экзонов.

| | поз | -3 | -2 | -1 | +3 | +4 | +5 | +6 |
|---|---|---|---|---|---|---|---|---|
| ковариация | -3 | 0.29 | 0.14 | 0.05 | -0.03 | -0.05 | -0.09 | -0.05 |
| | -2 | 0.14 | 1.36 | 0.36 | -0.09 | -0.30 | -0.36 | -0.23 |
| | -1 | 0.05 | 0.36 | 1.94 | -0.13 | -0.34 | -0.36 | -0.34 |
| | +3 | -0.03 | -0.09 | -0.13 | 1.06 | -0.01 | -0.20 | -0.11 |
| | +4 | -0.05 | -0.30 | -0.34 | -0.01 | 1.80 | 0.37 | 0.02 |
| | +5 | -0.09 | -0.36 | -0.36 | -0.20 | 0.37 | 1.97 | 0.21 |
| | +6 | -0.05 | -0.23 | -0.34 | -0.11 | 0.02 | 0.21 | 0.73 |
| p-значения | -3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | -2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | -1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 |
| | +4 | 0.00 | 0.00 | 0.00 | 0.33 | 0.00 | 0.00 | 0.15 |
| | +5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 |

**Рисунок 23.** Ковариационная матрица векторов сил позиций донорных сайтов сплайсинга конститутивных экзонов (*M. musculus*).

Сверху показана ковариационная матрица. Положительные значения окрашены зелёным цветом, отрицательные – красным. Чем насыщенней окраска, тем больше абсолютное значение ковариации. Снизу указаны p-значения того, данная ковариация значимо отличается от нуля, полученные бутстрепом; p-значения < 0.05 окрашены розовым.

Можно предположить, что отрицательные ковариации между экзонной и интронной частями и положительные ковариации внутри экзонной и интронной частей донорных сайтов сплайсинга могут быть результатом миграции сигнала из экзонной части в интронную (исторические причины, см. выше) [107]. Для иллюстрации этой идеи можно представить, что донорный сайт рождается с идеальной экзонной частью, которая медленно накапливает (слабо)вредные мутации (т.е. деградирует), в то время как интронная часть рождается слабой и постепенно усиливается. Предположим также, что изучаемая выборка состоит из донорных сайтов различных возрастов. Тогда “новые” сайты должны иметь сильную экзонную и слабую интронную части, а у “древних” сайтов, напротив, должна быть слабая экзонная и сильная интронная части. Смешивание в одной выборки сайтов разных возрастов может привести к наблюдаемым ковариациям. Чтобы протестировать эту гипотезу, мы разделили изучаемую выборку на две подвыборки (низко- и высококонсервативные сайты). Это деление было основано

на уровне многовидовой консервативности ключевого динуклеотида (GT). Донорные сайты с многовидовой консервативностью GT < 1 $K_s$ от генома *Mus musculus* (т.е. в пределах плацентарных млекопитающих) считались низкоконсервативными, а те, что имели консервативность динуклеотида > 1$K_s$ – высококонсервативными (см. разделы 3.1.1 и 3.2.5). Для каждой подвыборки мы построили ковариационные матрицы векторов сил, как описано выше разделе 3.2.3. Ковариационные матрицы низко- и высококонсервативных донорных сайтов оказались очень похожи между собой (Приложение: Рис. 11).  и на матрицу, построенную на объединённой выборке (рис. 23). Это наблюдение противоречит выдвинутому предположению о том, что наблюдаемый паттерн ковариаций объясняется миграцией сигнала. Однако, миграция сигнала происходит, что проявляется в ослаблении экзонной и усилении интронной частей донорных сайтов сплайсинга (см. раздел 4.2.3). Таким образом миграция сигнала не оказывает существенного влияния на наблюдаемый эффект.

Другая возможная причина – мутационные эффекты. Так как ковариации между позициями на противоположных концах донорного сайта сплайсинга столь же сильны (по модулю), что и между соседними, маловероятно и то, что наблюдаемый паттерн объясняется мутационными эффектами, так как известные мутационные контексты включают только соседние нуклеотиды [193].

Таким образом, единственная оставшаяся альтернатива такова: ковариации между силами позиций вызваны естественным отбором, причем не позиционно-независимым, а эпистатическим. Положительные ковариации соответствуют ситуации, когда две мутации, уменьшающие вес сайта, являются менее вредными, чем, если бы они действовали по отдельности, т.е. положительному эпистазу. Отрицательные ковариации соответствуют отрицательному эпистазу, т.е. когда две мутации вместе являются более вредными, чем по отдельности.

Рассмотрим возможные функциональные причины указанных эпистатических взаимодействий. Распознавание сайта сплайсинга сплайсосомой зависит от ряда факторов: наличия энхансеров и сайленсеров сплайсинга [32], локальной скорости

транскрипции [194,195], структуры хроматина [195–197], уровня метилирования ДНК [198] и т.д. Функциональность акцепторного сайта сплайсинга также зависит от распознавания близлежащего сайта ветвления (за счет взаимодействия с мяРНК U2, входящей в состав сплайсосомы) [26]. Более того, распознавание акцепторных и донорных сайтов взаимозависимо за счёт образования комплексов распознавания экзона и распознавания интрона [199]. Сила позиции в сайте сплайсинга отражает не только степень комплементарности между U1 и донорным сайтом сплайсинга или способность U2AF связываться с акцепторным сайтом, но и суммарное влияние всех вышеперечисленных факторов.

Несмотря на вышесказанное, главным фактором определяющим функциональность донорного сайта сплайсинга, является степень комплементарности к мяРНК U1, входящей в состав сплайсосомы. Консенсусные нуклеотиды – это как раз те нуклеотиды, которые обеспечивают максимальную комплементарность [15], поэтому сила позиции хорошо коррелирует с наличием/отсутствием комплементарности. Действительно, консенсусная последовательность донорного сайта сплайсинга 5’CAG|GURAGU3’ идеально соответствует последовательности U1, которая с ним связывается: 3’GUCCAΨΨCA5’. Аналогично, функциональность акцепторного сайта сплайсинга зависит от способности связываться с U2AF [10]. Как в донорных, так и в акцепторных сайтах сплайсинга консенсусные нуклеотиды более консервативны, чем неконсенсусные (см. раздел 4.1.2). Эти наблюдения позволяют заключить, что сила сайта сплайсинга (отражающая в первую очередь уровень комплементарности между U1 и донорным сайтом сплайсинга или способность акцепторного сайта связываться с U2AF) вносит свой вклад в суммарную приспособленность организма.

Возможной функциональной причиной наличия положительного эпистаза в донорных сайтах сплайсинга является кооперативное связывание с U1. Действительно, кооперативность может достигаться за счёт стэкинг-взаимодействий между идущими подряд азотистыми основаниями (мы наблюдаем положительные ковариации между каждым из нуклеотидов в

последовательностях экзонной и интронной частей донорных сайтов). Природа отрицательных ковариации между экзонной и интронной частями менее понятна. Возможно, для некоторой части сайтов сплайсинга максимальный вес не означает максимального вклада в приспособленность. Например, экспериментально было показано, что усиление слабого сайта приводит к потере регуляции сплайсинга [120]. В таких сайтах сильная (слабая) экзонная часть компенсирует слабую (сильную) интронную часть. О наличии таких сайтов сплайсинга также косвенно свидетельствует тот факт, что сила отбора, действующего на консенсусные нуклеотиды, слабее для сильных сайтов, чем для слабых (рис. 18) .

Наличие эпистатических взаимодействий подтверждается следующим наблюдением: число замен между ортологичными сайтами сплайсинга человека и мыши в парах позиций -1 и +5, -2 и +5 было меньше, чем ожидалось, если бы эти позиции эволюционировали независимо, тогда как для позиций +4 и +5, +5 и +6 оно было выше ожидания [201]. Однако отсюда не ясно, для какой пары позиций речь идёт о положительном эпистазе, а для какой – об отрицательном.

#### *4.2.5.2. Отбор против динуклеотида AG, как основная причина ковариаций, наблюдаемых в полипиримидиновом тракте акцепторных сайтов сплайсинга*

В акцепторных сайтах сплайсинга ковариационная матрица устроена более сложным образом (рис. 24). В ней встречаются как положительные, так и отрицательные ковариации. Некоторые позиции (-3, +1 и +2) отрицательно скоррелированы со всеми остальными позициями. В полипиримидиновом тракте (позиции -12…-7), наблюдается характерный периодический характер ковариаций: силы соседних позиций (на расстоянии в 1 нт) скоррелированы отрицательно, тогда как позиции на расстоянии 2 нт скоррелированы положительно, на расстоянии 3 нт – вновь отрицательно и т.д. (рис. 24А). Очень похожая картина наблюдается на выборках акцепторных сайтов из геномов человека и собаки, а также на выборке сайтов кассетных экзонов (данные не показаны).

Мы предположили, что наблюдаемая периодичность ковариации вызвана смещенным динуклеотидным составом в полипиримидиновом тракте. Для того,

чтобы проверить эту гипотезу, мы сгенерировали искусственные последовательности с тем же динуклеотидным составом, что и в оригинальных последовательностях полипиримидиновых трактов, с помощью марковской цепи первого порядка (см. раздел 3.2.6). Затем мы построили матрицы ковариаций на этих искусственных последовательностях и сравнили их с матрицами ковариаций, построенных на реальных акцепторных сайтах (рис. 24Б). Периодический паттерн ковариаций наблюдается и на искусственных последовательностях, хотя в целом ковариации выражены слабее.

По построению наблюдаемые дальние ковариации на выборке искусственных последовательностей (рис. 24Б) возникают в результате распространения динуклеотидных ковариаций на более далёкие расстояния. Наблюдаемый на выборке реальных сайтов (рис. 24А) аналогичный периодический паттерн с дальними ковариациями, по-видимому, говорит о том, что динуклеотидный состав является определяющим фактором и здесь.

Таким образом, наблюдаемый периодический характер ковариаций в полипиримидиновом тракте акцепторных сайтов сплайсинга определяется в основном динуклеотидным составом, более дальние взаимодействия (например, взаимодействия через нуклеотид и т.п) вносят скромный вклад в наблюдаемые ковариации.

Каковы причины появления ковариаций? Теоретически, он может возникать (1) из-за особенностей мутагенеза или (2) благодаря действию естественного отбора.

Для того, чтобы разобраться, как именно устроен динуклеотидный состав, мы упорядочили динуклеотиды по степени их недо- или перепредставленности по сравнению с тем, что ожидается исходя из частот встречаемости (моно)нуклеотидов. Мы также посчитали аналогичные частоты для нейтрально эволюционирующих последовательностей, взятых из соседнего интрона (нейтральный контроль).

А

| | поз | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | +1 | +2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ковариация | -14 | 0.725 | 0.012 | 0.050 | -0.035 | 0.030 | 0.012 | 0.001 | -0.009 | -0.023 | -0.056 | -0.002 | -0.034 | -0.024 | 0.001 |
| | -13 | 0.012 | 0.762 | 0.022 | 0.041 | -0.038 | 0.003 | 0.018 | 0.018 | -0.003 | -0.047 | -0.001 | -0.012 | -0.021 | -0.017 |
| | -12 | 0.050 | 0.022 | 0.966 | -0.054 | 0.035 | -0.065 | 0.011 | -0.007 | 0.007 | -0.010 | 0.007 | -0.040 | -0.030 | -0.013 |
| | -11 | -0.035 | 0.041 | -0.054 | 1.057 | -0.052 | 0.081 | -0.046 | 0.002 | -0.002 | 0.015 | 0.002 | -0.056 | -0.018 | -0.007 |
| | -10 | 0.030 | -0.038 | 0.035 | -0.052 | 0.949 | -0.102 | 0.042 | -0.022 | 0.014 | 0.021 | 0.006 | 0.001 | -0.021 | 0.001 |
| | -9 | 0.012 | 0.003 | -0.065 | 0.081 | -0.102 | 0.753 | -0.053 | 0.033 | -0.019 | -0.044 | 0.004 | -0.029 | -0.016 | 0.005 |
| | -8 | 0.001 | 0.018 | 0.011 | -0.046 | 0.042 | -0.053 | 0.739 | -0.024 | 0.026 | -0.021 | -0.003 | 0.003 | -0.011 | -0.003 |
| | -7 | -0.009 | 0.018 | -0.007 | 0.002 | -0.022 | 0.033 | -0.024 | 0.712 | 0.002 | -0.035 | -0.005 | -0.009 | -0.023 | -0.017 |
| | -6 | -0.023 | -0.003 | 0.007 | -0.002 | 0.014 | -0.019 | 0.026 | 0.002 | 0.952 | 0.116 | 0.009 | 0.000 | -0.040 | 0.010 |
| | -5 | -0.056 | -0.047 | -0.010 | 0.015 | 0.021 | -0.044 | -0.021 | -0.035 | 0.116 | 0.990 | -0.005 | -0.029 | -0.030 | 0.000 |
| | -4 | -0.002 | -0.001 | 0.007 | 0.002 | 0.006 | 0.004 | -0.003 | -0.005 | 0.009 | -0.005 | 0.079 | -0.015 | 0.002 | 0.000 |
| | -3 | -0.034 | -0.012 | -0.040 | -0.056 | 0.001 | -0.029 | 0.003 | -0.009 | 0.000 | -0.029 | -0.015 | 0.807 | -0.028 | -0.026 |
| | +1 | -0.024 | -0.021 | -0.030 | -0.018 | -0.021 | -0.016 | -0.011 | -0.023 | -0.040 | -0.030 | 0.002 | -0.028 | 0.590 | -0.015 |
| | +2 | 0.001 | -0.017 | -0.013 | -0.007 | 0.001 | 0.005 | -0.003 | -0.017 | 0.010 | 0.000 | 0.000 | -0.026 | -0.015 | 0.221 |
| p-значения | -14 | 0.00 | 0.18 | 0.00 | 0.01 | 0.02 | 0.17 | 0.47 | 0.22 | 0.04 | 0.00 | 0.28 | 0.00 | 0.02 | 0.44 |
| | -13 | 0.18 | 0.00 | 0.07 | 0.00 | 0.00 | 0.41 | 0.09 | 0.08 | 0.41 | 0.00 | 0.41 | 0.19 | 0.03 | 0.01 |
| | -12 | 0.00 | 0.07 | 0.00 | 0.00 | 0.02 | 0.00 | 0.22 | 0.29 | 0.32 | 0.27 | 0.08 | 0.00 | 0.01 | 0.05 |
| | -11 | 0.01 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.43 | 0.44 | 0.18 | 0.31 | 0.00 | 0.08 | 0.17 |
| | -10 | 0.02 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.06 | 0.19 | 0.10 | 0.11 | 0.47 | 0.05 | 0.45 |
| | -9 | 0.17 | 0.41 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.07 | 0.00 | 0.14 | 0.01 | 0.08 | 0.23 |
| | -8 | 0.47 | 0.09 | 0.22 | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 | 0.04 | 0.07 | 0.22 | 0.40 | 0.15 | 0.29 |
| | -7 | 0.22 | 0.08 | 0.29 | 0.43 | 0.06 | 0.00 | 0.02 | 0.00 | 0.45 | 0.01 | 0.14 | 0.21 | 0.02 | 0.01 |
| | -6 | 0.04 | 0.41 | 0.32 | 0.44 | 0.19 | 0.07 | 0.04 | 0.45 | 0.00 | 0.00 | 0.03 | 0.49 | 0.00 | 0.08 |
| | -5 | 0.00 | 0.00 | 0.27 | 0.18 | 0.10 | 0.00 | 0.07 | 0.01 | 0.00 | 0.00 | 0.11 | 0.03 | 0.01 | 0.48 |
| | -4 | 0.28 | 0.41 | 0.08 | 0.31 | 0.11 | 0.14 | 0.22 | 0.14 | 0.03 | 0.11 | 0.00 | 0.00 | 0.29 | 0.42 |
| | -3 | 0.00 | 0.19 | 0.00 | 0.00 | 0.47 | 0.01 | 0.40 | 0.21 | 0.49 | 0.03 | 0.00 | 0.00 | 0.01 | 0.00 |
| | +1 | 0.02 | 0.03 | 0.01 | 0.08 | 0.05 | 0.08 | 0.15 | 0.02 | 0.00 | 0.01 | 0.29 | 0.01 | 0.00 | 0.00 |
| | +2 | 0.44 | 0.01 | 0.05 | 0.17 | 0.45 | 0.23 | 0.29 | 0.01 | 0.08 | 0.48 | 0.42 | 0.00 | 0.00 | 0.00 |

Б

| | pos | -14 | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | +1 | +2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ковариация | -14 | 0.725 | 0.012 | 0.004 | 0.007 | 0.012 | -0.021 | -0.006 | -0.009 | -0.023 | -0.056 | -0.002 | -0.034 | -0.024 | 0.001 |
| | -13 | 0.012 | 0.762 | -0.019 | 0.006 | 0.002 | 0.027 | 0.004 | -0.014 | -0.003 | -0.047 | -0.001 | -0.012 | -0.021 | -0.017 |
| | -12 | 0.004 | -0.019 | 0.970 | -0.076 | -0.010 | 0.003 | -0.007 | 0.015 | -0.004 | 0.031 | -0.002 | 0.020 | 0.001 | 0.000 |
| | -11 | 0.007 | 0.006 | -0.076 | 1.082 | -0.100 | 0.013 | -0.012 | 0.010 | -0.022 | 0.008 | -0.004 | -0.004 | 0.005 | -0.015 |
| | -10 | 0.012 | 0.002 | -0.010 | -0.100 | 0.977 | -0.055 | 0.006 | -0.019 | 0.008 | -0.006 | 0.003 | 0.001 | -0.016 | -0.003 |
| | -9 | -0.021 | 0.027 | 0.003 | 0.013 | -0.055 | 0.730 | -0.057 | -0.002 | 0.018 | 0.018 | 0.002 | -0.024 | 0.003 | 0.001 |
| | -8 | 0.006 | 0.004 | -0.007 | -0.012 | 0.006 | 0.057 | 0.735 | 0.021 | -0.013 | 0.024 | 0.008 | -0.007 | 0.005 | 0.002 |
| | -7 | -0.009 | -0.014 | 0.015 | 0.010 | -0.019 | -0.002 | -0.021 | 0.712 | 0.011 | 0.007 | 0.003 | -0.007 | -0.002 | 0.004 |
| | -6 | -0.023 | -0.003 | -0.004 | -0.022 | 0.008 | 0.018 | -0.013 | 0.011 | 0.952 | 0.116 | 0.009 | 0.000 | -0.040 | 0.010 |
| | -5 | -0.056 | -0.047 | 0.031 | 0.008 | -0.006 | 0.018 | -0.024 | 0.007 | 0.116 | 0.990 | -0.005 | -0.029 | -0.030 | 0.000 |
| | -4 | -0.002 | -0.001 | -0.002 | -0.004 | 0.003 | 0.002 | 0.008 | 0.003 | 0.009 | -0.005 | 0.079 | -0.015 | 0.002 | 0.000 |
| | -3 | -0.034 | -0.012 | 0.020 | -0.004 | 0.001 | -0.024 | -0.007 | -0.007 | 0.000 | -0.029 | -0.015 | 0.807 | -0.028 | -0.026 |
| | +1 | -0.024 | -0.021 | 0.001 | 0.005 | -0.016 | 0.003 | 0.005 | -0.002 | -0.040 | -0.030 | 0.002 | -0.028 | 0.590 | -0.015 |
| | +2 | 0.001 | -0.017 | 0.000 | -0.015 | -0.003 | 0.001 | 0.002 | 0.004 | 0.010 | 0.000 | 0.000 | -0.026 | -0.015 | 0.221 |
| p-значения | -14 | 0.00 | 0.18 | 0.29 | 0.16 | 0.05 | 0.00 | 0.14 | 0.06 | 0.04 | 0.00 | 0.28 | 0.00 | 0.02 | 0.44 |
| | -13 | 0.18 | 0.00 | 0.00 | 0.22 | 0.40 | 0.00 | 0.26 | 0.01 | 0.41 | 0.00 | 0.41 | 0.19 | 0.03 | 0.01 |
| | -12 | 0.29 | 0.00 | 0.00 | 0.00 | 0.03 | 0.31 | 0.16 | 0.02 | 0.34 | 0.00 | 0.18 | 0.00 | 0.45 | 0.46 |
| | -11 | 0.16 | 0.22 | 0.00 | 0.00 | 0.00 | 0.15 | 0.07 | 0.07 | 0.00 | 0.19 | 0.06 | 0.31 | 0.22 | 0.00 |
| | -10 | 0.05 | 0.40 | 0.03 | 0.00 | 0.00 | 0.00 | 0.41 | 0.01 | 0.17 | 0.23 | 0.07 | 0.44 | 0.00 | 0.24 |
| | -9 | 0.00 | 0.00 | 0.31 | 0.15 | 0.00 | 0.00 | 0.00 | 0.16 | 0.00 | 0.01 | 0.19 | 0.00 | 0.26 | 0.36 |
| | -8 | 0.14 | 0.26 | 0.16 | 0.07 | 0.41 | 0.00 | 0.00 | 0.39 | 0.03 | 0.00 | 0.00 | 0.14 | 0.16 | 0.23 |
| | -7 | 0.06 | 0.01 | 0.02 | 0.07 | 0.01 | 0.16 | 0.39 | 0.00 | 0.06 | 0.16 | 0.09 | 0.14 | 0.38 | 0.10 |
| | -6 | 0.04 | 0.41 | 0.34 | 0.00 | 0.17 | 0.00 | 0.03 | 0.06 | 0.00 | 0.00 | 0.03 | 0.49 | 0.00 | 0.08 |
| | -5 | 0.00 | 0.00 | 0.00 | 0.19 | 0.23 | 0.01 | 0.00 | 0.16 | 0.00 | 0.00 | 0.11 | 0.03 | 0.01 | 0.48 |
| | -4 | 0.28 | 0.41 | 0.18 | 0.06 | 0.07 | 0.19 | 0.00 | 0.09 | 0.03 | 0.11 | 0.00 | 0.00 | 0.29 | 0.42 |
| | -3 | 0.00 | 0.19 | 0.00 | 0.31 | 0.44 | 0.00 | 0.14 | 0.14 | 0.49 | 0.03 | 0.00 | 0.00 | 0.01 | 0.00 |
| | +1 | 0.02 | 0.03 | 0.45 | 0.22 | 0.00 | 0.26 | 0.16 | 0.38 | 0.00 | 0.01 | 0.29 | 0.01 | 0.00 | 0.00 |
| | +2 | 0.44 | 0.01 | 0.46 | 0.00 | 0.24 | 0.36 | 0.23 | 0.10 | 0.08 | 0.48 | 0.42 | 0.00 | 0.00 | 0.00 |

**Рисунок 24** Ковариационные матрицы векторов сил позиций акцепторных сайтов сплайсинга конститутивных экзонов (*M. musculus*).

Обозначения те же, что на Рис. 23. Позиции, для которых наблюдается характерный периодический паттерн ковариаций (-12 … -7) обведены черной рамкой. А — реальные последовательности акцепторных сайтов сплайсинга. Б — искусственные последовательности акцепторных сайтов с тем же динуклеотидным составом, что и в реальных сайтах на участке (-12 … -7).

Наиболее недопредставленным динуклеотидом оказался AG, который встречается в ~30 раз реже, чем ожидается на основе частот мононулеотидов, при этом такой недопредставленности не наблюдается в нейтральном контроле (рис. 25). Нуклеотиды A и G встречаются редко в полипиримидиновом трате, а сильная недопредставленность динуклеотидов из двух слабых букв повышает вероятность динуклеотидов из букв существенно разной силы (как TG), что хотя бы частично объясняет периодический паттерн ковариаций. Динуклеотид AG недопредставлен в полипиримидиновом тракте, чтобы избежать сплайсинга по этому нуклеотиду (он может служить альтернативой существующему ключевому динуклеотиду и приводить к смещению сайта сплайсинга) [203]. Существование зоны избегания AG (AG exclusion zone, AGEZ) было показано ранее [203,204]. Кроме того AGEZ играет существенную роль в сплайсинге. В частности, удаление фактора U2AF35 (который распознаёт ключевой AG) вызывает увеличение частоты включения одних экзонов и уменьшение частоты включения других. Акцепторные сайты сплайсинга экзонов, увеличивших степень включения, имеют более длинную AGEZ, чем те, что уменьшили частоту включения, а также, чем сайты экзонов из контрольной выборки [204].

Следующий недопредставленный динуклеотид – это CG (он встречается в 4 раза реже, чем ожидается на основе частот нуклеотидов). В отличие от AG, динуклеотид CG также недопредставлен и в нейтральном контроле (рис. 25). Это наблюдение отражает хорошо известный феномен: геномы млекопитающих обеднены CpG [205] из-за высокой вероятности мутации метилированного цитозина в составе CpG в тимин (CpG→TpG) [183]. Прочие особенности динуклеотидного состава полипиримидинового тракта могут быть обусловлены широком набором факторов, таких как оптимальное связывание U2AF65 и регуляторных транс-факторов сплайсинга, так и быть побочным следствием недопредставленности AG (например, перепредставленность AA по сравнению с нейтральным контролем).

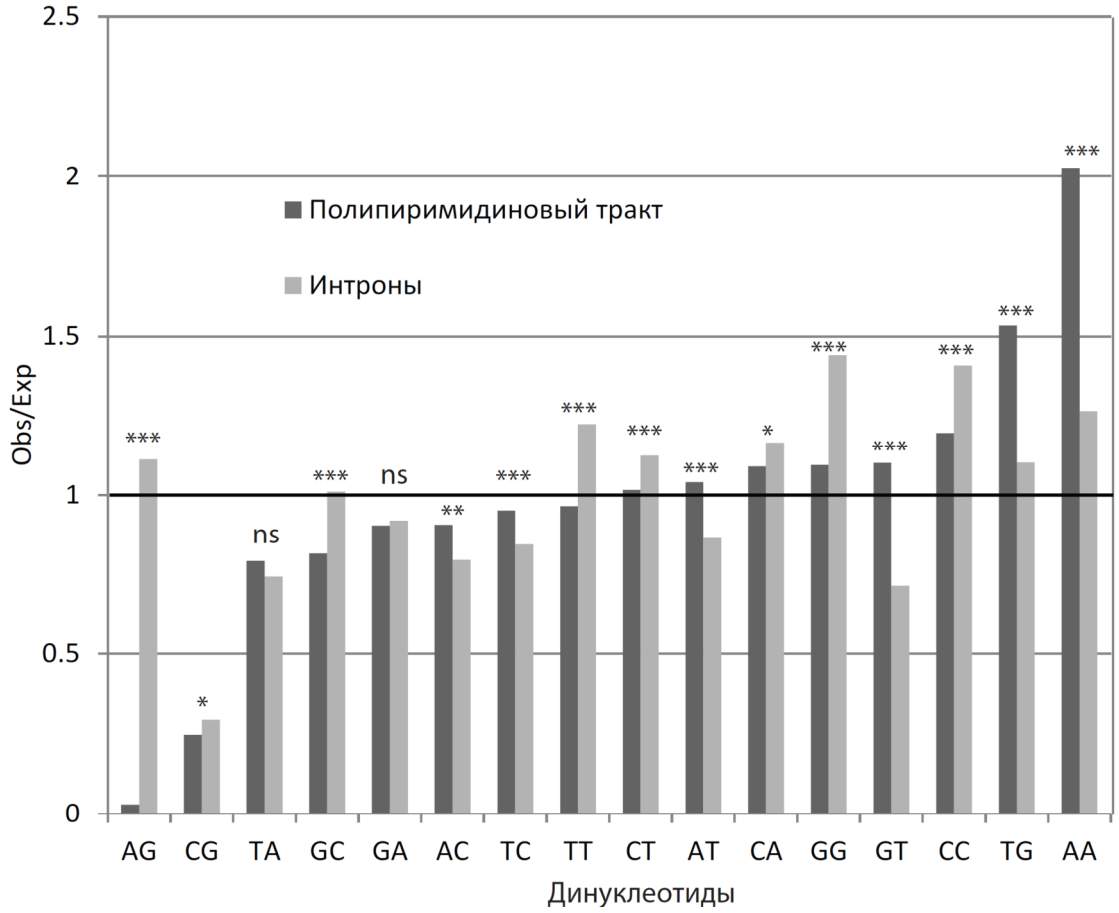


**Рисунок 25.** Динуклеотидный состав участка полипиримидинового тракта, где наблюдается периодический паттерн ковариаций (позиции – 12 … -7). По вертикальной оси отложено отношение наблюдаемой частоты динуклеотида к ожидаемой (*Obs/Exp*) на участке -12…-7 полипиримидинового тракта (черные столбики), а также в близлежащих интронах (серые столбики). Динуклеотиды упорядочены по возрастанию величины *Obs/Exp* на участке периодических ковариаций. Ожидаемая частота динуклеотидов (*Exp*) вычислялась исходя из мононуклеотидного состава. Показана статистическая значимость различий между величинами *Obs/Exp* на участке -12…-7 полипиримидинового тракта и близлежащих интронов: * $0.01 < p \leq 0.05$; ** $0.001 < p \leq 0.01$; *** $p \leq 0.001$; ns $p > 0.05$.

Периодический паттерн ковариаций на дальних расстояниях (рис. 24А) мог бы также формироваться трактами идентичных динуклеотидов, которые могут возникать из-за проскальзывания ДНК-полимеразы (особенности мутагенеза). Однако, распределения длин почти каждого из 16 типов трактов, состоящих из повторяющихся динуклеотидов $(AA)_n$, $(AC)_n$, …, $(TT)_n$ почти не отличаются от соответствующих распределений, построенных на сгенерированных

последовательностях (рис. 26). Два исключения составляют $(TC)_n$ и $(CT)_n$, длинные тракты здесь встречаются немного чаще, чем ожидается.

Итак, периодический характер ковариаций в полипиримидиновом тракте определяется в основном динуклеотидным составом, причем недопредставленность AG играет здесь ключевую роль. Формально говоря, отбор против AG является вырожденным случаем эпистатического отбора. Появление нуклеотида A (равно как и G) в результате мутации почти не уменьшает приспособленности, в то время как их совместное появление в виде нуклеотида AG существенно её снижает из-за нарушения сплайсинга.

Подводя итог, ковариации сил нуклеотидных позиций в сайтах сплайсинга могут вызывать несколько причин: (1) эпистаз между позициями внутри сайта сплайсинга; (2) статистический эффект смешивания в одной выборке сайтов сплайсинга разного возраста; (3) мутационные эффекты не связанные с отбором, такие как недопредставленность CpG; (4) отборные эффекты, например, избегание ApG в акцепторных сайтах сплайсинга. Ковариации, наблюдаемые в акцепторных сайтах сплайсинга, в основном вызваны (4), тогда как ковариации в донорных сайтах вероятно отражают эпистаз между позициями (1).

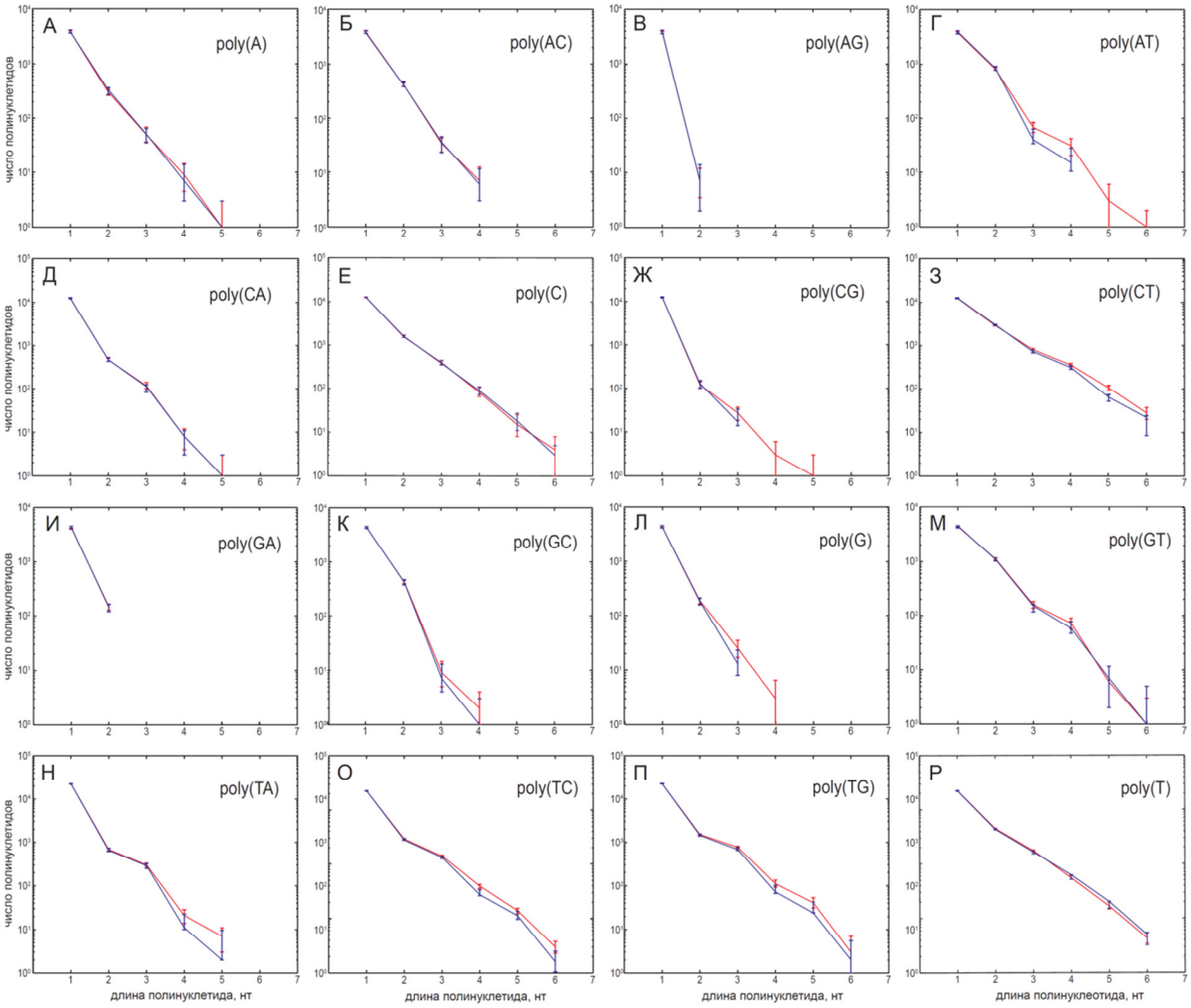


**Рисунок 26.** Распределение частот полинуклеотидов.

По горизонтальной оси отложена длина полинуклеотида, по вертикальной оси – число наблюденных нуклеотидов такой длины (в логарифмической шкале). Красная ломаная – наблюдаемое число полинуклеотидов, синяя линия – ожидаемое число полинуклеотидов. Панели соответствуют разным полинуклеотидам: А - polyA; Б - poly(AC); В - poly(AG), …, Р - polyT.

### 4.2.6. Меняются ли ковариации между позициями в ходе эволюции? Проверка гипотезы о независимой эволюции позиций сайтов сплайсинга

Наблюдаемые ковариации между силами позиций в сайтах сплайсинга – результат длительной эволюции, из которой мы наблюдаем финальную стадию: от общего предка человека мыши и собаки до каждого из этих видов. Естественным является следующий вопрос: влияют ли зависимости между позициями на эволюцию сайтов

сплайсинга на рассматриваемом финальном этапе или этим влиянием можно пренебречь; если таковое влияние есть, то на каких характеристиках сайтов сплайсинга оно отражается (средняя сила сайтов, матрица ковариаций)? Более конкретно, эпистатический отбор должен приводит к усилению (по модулю) соответствующих ковариаций (см. раздел 2.3.2.2). Верно ли это для сайтов сплайсинга?

Для ответа на эти вопросы мы разработали метод, основанный на компьютерном моделировании процесса эволюции в предположении независимости позиций внутри сайтов сплайсинга. Идея метода основана на сравнении результатов смоделированной позиционно-независимой эволюции с результатами реальной эволюции.

Далее при описании метода мы будем использовать как пример сайты сплайсинга в геноме человека; для сайтов в геномах мыши и собаки, вычисления полностью аналогичны. Мы сравниваем две выборки сайтов сплайсинга: (1) реальные сайты, наблюдаемые в геноме человека, и (2) искусственные сайты сплайсинга, полученные при моделировании эволюции в предположении независимости позиций. Искусственные сайты были получены следующим образом.

Рассмотрим любую данную тройку ортологичных сайтов сплайсинга *HMD* и множество возможных предковых сайтов *{R}*. Предполагая независимость эволюции отдельных позиций, вероятность искусственного сайта сплайсинга *H** в геноме человека, при условии наблюдения в геномах человека, мыши и собаки реальных сайтов *H*, *M* и *D*, равна

$$p(H^* \mid HMD) = \prod_{i=1}^{L} p(h_i^* \mid h_i m_i d_i), \tag{14}$$

где *i* – номер позиции в сайте, а

$$p(h_i^* \mid h_i m_i d_i) = \sum_{r_i \in \{A,C,G,T\}} p(h_i^* \mid r_i, h_i m_i d_i)\, p(r_i \mid h_i m_i d_i). \tag{15}$$

Вероятность порождения нуклеотида $h_i*$ зависит только от соотетствующего нуклеотида в сайте общего предка $r_i$,,и рассчитывается с использованием той же матрицы переходных вероятностей, что и $p(h_i|r_i)$:

$$p(h_i * | r_i, h_i m_i d_i) = p(h_l * | r_l) \quad (16)$$

Из уравнений (11), (15), (16) имеем:

$$p(h_i * | h_i m_i d_i) = \sum_{r_i \in \{A,C,G,T\}} \frac{p(h_i * | r_i)\, p(r_i)\, p(h_i | r_i)\, p(m_i | r_i)\, p(d_i | r_i)}{\sum_{x \in \{A,C,G,T\}} p(x)\, p(h_i | x)\, p(m_i | x)\, p(d_i | x)} \quad (17)$$

Мы произвели симуляции Монте-Карло, используя формулу (17), получив для каждой тройки реальных сайтов *HMD* из выборки тройку искусственно полученных сайтов *H*M*D**. Сравнение свойств выборок троек реальных сайтов и соответствующих с троек искусственных сайтов позволяет судить как зависимость между позициями влияет на эволюцию сайтов. В частности мы сравнивали средние силы сайтов и ковариационные матрицы векторов сил сайтов (см. Материалы и методы, разделы 3.2.2 и 3.2.3).

Средняя сила искусственных сайтов не сильно отличалась от средней силы реальных сайтов (данные не показаны). Таким образом, можно сказать, что ковариации не сильно влияют на изменение силы сайта. Сравнение ковариационных матриц позволяет ответить на вопрос, как влияет зависимость между позициями на ковариации в ныне существующих сайтах: они усиливаются, ослабляются или остаются неизменными (по сравнению с позиционно-независимой эволюцией). В таблице 3 показаны результаты этого сравнения.

Как в донорных сайтах, так и в акцепторных сплайсинга было найдено несколько статистически значимых различий между ковариационными матрицами. Общая тенденция состоит в том, что отрицательные ковариации усиливаются по модулю. Положительные ковариации также усиливаются, однако статистически значимо это происходит только на выборке акцепторных сайтов.

**Таблица 3.** Сравнение ковариационных матриц реальных сайтов и сайтов, полученных при воспроизведении позиционно-независимой эволюции.

| Сайты сплайсинга | Тип сплайсинга | Организм | Знак ковариации | Знак изменения ковариации | | p-значение | |
|---|---|---|---|---|---|---|---|
| | | | | + | - | Бином. тест | Точный тест Фишера |
| **Донорные** | **Конститутивные экзоны** | **Человек** | **+** | 3 | 3 | 6.56E-01 | 1.98E-01 |
| | | | **-** | 3 | 12 | **1.76E-02** | |
| | | **Мышь** | **+** | 5 | 1 | 1.09E-01 | **5.55E-03** |
| | | | **-** | 2 | 13 | **3.69E-03** | |
| | | **Собака** | **+** | 5 | 1 | 1.09E-01 | **1.39E-02** |
| | | | **-** | 3 | 12 | **1.76E-02** | |
| | **Кассетные экзоны** | **Человек** | **+** | 3 | 4 | 7.73E-01 | 2.99E-01 |
| | | | **-** | 3 | 11 | **2.87E-02** | |
| | | **Мышь** | **+** | 3 | 3 | 6.56E-01 | 6.33E-01 |
| | | | **-** | 7 | 8 | 5.00E-01 | |
| | | **Собака** | **+** | 7 | 0 | **7.81E-03** | **6.81E-03** |
| | | | **-** | 5 | 9 | 2.12E-01 | |
| **Акцепторные** | **Конститутивные экзоны** | **Человек** | **+** | 16 | 18 | 6.96E-01 | 5.45E-02 |
| | | | **-** | 16 | 41 | **6.32E-04** | |
| | | **Мышь** | **+** | 28 | 7 | **2.54E-04** | **3.75E-08** |
| | | | **-** | 12 | 44 | **1.04E-05** | |
| | | **Собака** | **+** | 28 | 8 | **5.97E-04** | **2.22E-06** |
| | | | **-** | 15 | 40 | **5.08E-04** | |
| | **Кассетные экзоны** | **Человек** | **+** | 17 | 27 | 9.52E-01 | 7.23E-01 |
| | | | **-** | 20 | 27 | 1.91E-01 | |
| | | **Мышь** | **+** | 26 | 14 | **4.03E-02** | **3.44E-04** |
| | | | **-** | 14 | 37 | **8.85E-04** | |
| | | **Собака** | **+** | 35 | 16 | **5.49E-03** | **9.58E-05** |
| | | | **-** | 11 | 29 | **3.21E-03** | |

Для каждой ковариационной матрицы векторов сил позиций реальных сайтов сплайсинга **R** (см. Рис. 23, 24А) подсчитывалось количество элементов в ней, расположенных выше диагонали, где ковариация больше нуля и меньше либо равна нулю (знак ковариации “+” или “-”, соответственно). Была также посчитана ковариационная матрица искусственных сайтов сплайсинга **A,** полученных в результате симуляции позиционно-независимой эволюции в каждой из ветвей дерева (человек, мышь, собака). Далее рассматривалась матрица разницы **D = A – R**. Знак изменения ковариации – это число элементов этой матрицы, расположенных выше диагонали, где значения больше нуля и меньше либо равна нулю (знак ковариации “+” или “–”, соответственно). Рассматривалась таблица сопряженности, где строкам соответвуют “+” и “–“ значения ковариации, а строкам — “+” и ”–“ значения изменения ковариации. Точный тест Фишера показывает, насколько независимо распределены элементы в этой таблице сопряженности. Биномиальный тест показывает насколько количество наблюденных “+” и “–“ элементов матрицы D отличается от случайного (для каждой строки). **Жирным шрифтом** выделены p-значения < 0.05

Сайты кассетных экзонов проявляют в целом те же свойства, что и сайты конститутивных экзонов, однако из-за меньшего объёма выборки статистическая достоверность результатов ниже. В линии мыши все тенденции проявляются наиболее контрастно в силу большей длины соответствующей ветви дерева (соответственно, успело произойти больше эволюционных событий).

Кроме того, имеются некоторые специфические для конкретных филогенетических линий и позиций паттерны. Отрицательные ковариации в интронной части акцепторных сайтов сплайсинга как конститутивных, так и кассетных экзонов усиливаются в линиях мыши и человека. С другой стороны, в экзонных частях сайтов сплайсинга отрицательные ковариации ослабляются (в линиях человека и мыши для конститутивных экзонов и в линии человека для кассетных экзонов). В экзонных частях кассетных экзонов в линии мыши происходит ослабление положительных ковариаций. Усиливаются отрицательные ковариации между экзонными и интронными частями сайтов сплайсинга конститутивных экзонов

(человек, мышь), чего не наблюдается для кассетных экзонов (данные не показаны).

Таким образом, как в донорных, так и в акцепторных сайтах сплайсинга наблюдается увеличение по модулю как отрицательных, так и положительных ковариаций, что является следствием действия эпистатического отбора. Отбор против динуклеотида AG в акцепторных сайтах сплайсинга является частным (вырожденным) случаем эпистатического отбора (см. выше).

## 4.3. Отбор в окрестности сайтов сплайсинга

### 4.3.1. Консервативность цис-регулятора сплайсинга UGCAUG в геномах человека и мыши

Brudno с соавт. [167] проанализировали выборку кассетных экзонов, специфично экспрессирующихся в мозгу, с интронными последовательностями вокруг них. Было показано, что в интронах, следующих за кассетными экзонами, мотив UGCAUG представлен в избытке по сравнению с контрольной выборкой, содержащей конститутивные экзоны с последующими интронами. Ранее считалось, что UGCAUG – сайт регуляции альтернативного сплайсинга специфичных для мозга экзонов, однако в той же работе [167] указывается, этот гексануклеотид также представлен и в ряде интронов, следующих за экзонами, специфичными для мышечных клеток.

Часть из найденных Brudno с соавт. гексануклеотидов UGCAUG могли встретиться случайно. Наша задача состояла в том, чтобы путем сравнительного анализа геномов человека и мыши отделить случайные гексануклеотиды от функциональных, т.е. принимающих участие в регуляции альтернативного сплайсинга.

Экзон-интронная структура генов меняется в ходе эволюции [85], [86], [87], [88], поэтому существование каждого экзона в геномах человека и мыши верифицировалось наличием соответствующих мРНК/EST. В некоторых случаях не находилось таких мРНК/EST, что может означать, что либо этот экзон

действительно не существует, либо в база данных мРНК/EST на момент исследования была не полна (не содержала изоформ, включающих этот экзон, хотя в действительности он существует). Все такие случаи представлены в таблице 4. Высокий уровень консервативности последовательности предполагаемых экзонов, сохранение ключевых динуклеотидов в сайтах сплайсинга и наличие функциональных ортологов в родственных организмах подтверждают их функциональность во всех случаях кроме двух. Этими двумя исключениями являются (1) ген мыши ANK2, где выравнивание очень плохое, (2) ген мыши KOR-3, где соответствующий предполагаемый экзон невозможно найти (отсутствует в таблице). Эти два случая далее не рассматривались. Экзон гена GRIN1 в геноме мыши существенно длиннее, чем ортологичный экзон в геноме человека. Это можно объяснить заменой, произошедшей в ключевом динуклеотиде донорного сайта сплайсинга (GT→GC). Это привело к тому, что сплайсосома перестала распознавать этот сайт, а вместо этого стала использовать расположенный далее сайт сплайсинга с динуклеотидом GT.

**Таблица 4. Экзоны, существование которых не подтверждено EST/мРНК**.

| Название гена | Выравни-вание | AG/GT | Genscan | У кого есть гомологичный экзон? |
|---|---|---|---|---|
| ***Человек*** | | | | |
| CACNA1B | + | ag/gt | - | крыса |
| NF1, экзон 9a | + | ag/gt | - | крыса |
| ACVR2A | + | ag/gt | - | крыса |
| SRC, экзон N | + | ag/gt | - | мышь, крыса |
| ***Мышь*** | | | | |
| ANK 2 | - | ag/gt | + | человек |
| ATP2B4 | + | ag/gt | - | человек, крыса, корова, собака |
| CACNA1B | + | ag/gt | - | человек, крыса |
| GRIN1, экзон 5 | + | ag/gt | 5'-удл. | крыса |
| KSR1 | + | ag/gt | + | человек |
| DLG1 | + | ag/gt | - | человек |
| LOC193091 | + | ag/gC | - | человек, крыса, корова (не удлинненный) |
| NF1, экзон 9a | + | ag/gt | - | крыса |
| ACVR2A | + | ag/gt | - | крыса |

Выравнивание: "+" – качественное, "–" – некачественное (с большим количеством пропусков, со сдвигом рамки считывания и т.п.). Genscan: "+" – экзон предсказан genscan, "–" – экзон не предсказан, "5'-удл." – предсказанный экзон длиннее с 5'-конца, чем у гомолога.

Мы проанализировали все находки гексануклеотида UGCAUG в пределах 1000 нт после кассетных экзонов из рассматриваемой выборки как в геноме человека, так и в геноме мыши. В 24 парах ортологичных генов, 28 мотивов UGCAUG найдены в интронах человека и 20 – в интронах мыши. Данные по консервативности этих мотивов представлены в таблице 5. Четырнадцать мотивов консервативны (не содержат ни одной замены) в геномах человека и мыши. То есть уровень консервативности составляет 14/28 = 50% для мотивов человека и 14/20 = 70% для мотивов мыши.

В двух случаях функциональность гексануклеотида не удалось установить, используя геномы человека и мыши, поэтому был также использован геном крысы. В гене ATP2B4 сигнал консервативен в геномах человека и крысы (геном мыши в этом месте не секвенирован). Поэтому этот элемент скорее всего является функциональным. UGCAUG в гене KSR1 превратился в UGCgUG у мыши, однако у крысы этот мотив сохранился. Следовательно, нельзя сделать однозначных выводов о функциональности этого элемента.

Чтобы понять, насколько вероятно то, что гексануклеотид UGCAUG консервативен по случайным причинам, мы приблизительно оценили вероятность того, что в нём произойдет хотя бы одна замена между геномами человека и мыши. Пусть средняя частота замен в нейтрально эволюционирующих последовательностях равна $q$. Тогда произвольный нуклеотид консервативен с вероятностью $1 - q$, и, следовательно, вероятность консервативности гексануклеотида равна $(1 - q)^6$. Частота замен в нейтральных регионах между геномами человека и мыши оценивается как 0.51 замены на нуклеотид [210]. Таким образом, если $q = 0.51$, то вероятность консервативности гексануклеотида равна $(1 - q)^6 = 1.8$ %, т.е. большинство гексануклеотидов должны содержать замены.

**Таблица 5. Консервативность UGCAUG между человеком и мышью.**

| № | Название гена | Источник / организм | Координаты UGCAUG | | Выравни-вание | Истинность сигнала |
|---|---|---|---|---|---|---|
| | | | Человек | Мышь | | |
| **01** | ANK 2 | AC073240.2/H. | ---------- | 1. 226 | есть | ложный |
| **02** | FHL1 | AL078638.9/H. | | | | |
| **03** | ATP2B4 | AL356980.4/H. | 1. 22 | ------- | нет | ложный |
| **04** | SCN8A | AF050730/H. | 1. 29 | 1. 29 | есть | истинный |
| 05 | BIN1 | AC012508/H. | ---------- | -------- | | |
| **06** | CACNA1B | AC020707.4/H. | ---------- | -------- | | |
| **07** | GRIN1, экзон 5 | Z32773/H. | 1. 8<br>2. 257<br>3. 492 | 1. 8<br>2. 262<br>3. 496 | есть<br>есть<br>есть | истинный<br>истинный<br>истинный |
| **08** | CLTB | AC010297.3/H. | ---------- | -------- | | |
| **09** | MAG, экзон 12 | AC002132.1/H. | 1. 62 | 1. 62<br>2. 598 | есть<br>нет | истинный<br>ложный |
| **10** | EPB41, экзон 15 | AL357500.6/H. | ---------- | -------- | | |
| **11** | KSR1 | AC015688.3/H. | 1. 12 | -------- | есть | ? |
| **12** | EPB41L1 | AL121895.21/ H. | 1. 5 | 1. 5 | есть | истинный |
| **13** | EPB41L3, экзон 15 | AC007445/H. | 1. 285<br>2. 343 | 1. 270<br>2. 327 | есть<br>есть | истинный<br>истинный |
| **14** | DLG1 | AC011322.3/H. | 1. 242<br>2. 275 | 1. 242<br>2. 274 | есть<br>есть | истинный<br>истинный |
| **15** | OPRL1 | U32929.1/M. | | 1. 304 | исключен из выборки | |
| **16** | AGRN, экзон 33 | M92657.1/M. | ---------- | -------- | | |
| **17** | LOC193091 | AC011061.4/H. | 1. 243 | -------- | нет | ложный |
| **18** | NF1, экзон 9a | AC004526.1/H. | 1. 56<br>2. 116 | 1. 56<br>2. 117 | есть<br>есть | истинный<br>истинный |
| **19** | PTPRF | AL158083.3/H. | 1. 884<br>2. 888 | 1. 961<br>-------- | нет<br>нет | ложный<br>ложный |
| **20** | AGRN, экзон 32 | M92657.1/M. | 1. 247<br>2. 251<br>3. 364<br>4. 445<br>5. 620<br>6. 645<br>7. 734<br>8. 885 | 1. 257<br>2. 265 | нет<br>нет | ложный<br>ложный<br>ложный<br>ложный<br>ложный<br>ложный<br>ложный<br>ложный |
| **21** | ACVR2A | AC009480/H. | ---------- | -------- | | |
| **22** | GABRG2 | AF165124.1/H. | 1. 30 | 1. 30 | есть | истинный |
| **23** | SRC, экзон N | AL133293/H. | 1. 66<br>2. 289 | 1. 57<br>2. 622 | есть<br>нет | истинный<br>ложный |
| **24** | AGRN, экзон 28 | AL390719.3/H. | ---------- | 1. 876 | нет | ложный |
| **25** | APBB1 | AF029234/H. | ---------- | -------- | | |

Как сказано выше в реальности уровень консервативности найденных UGCAUG существенно выше: 50% для мотивов человека и 70% для мотивов мыши. Поэтому консервативность гексануклеотида означает действие отрицательного отбора, т.е. его функциональную значимость. С другой стороны, мы не можем исключить возможности того, что часть неконсервативных мотивов функциональны. Регуляция альтернативного сплайсинга меняется в течение эволюции, и часть гексануклеотидов могли перестать выполнять прежнюю функцию в одной из линий. Мы не утверждаем, что нашли все функциональные UGCAUG, однако мы считаем, что консервативные гексануклеотиды скорее всего функциональны.

Все найденные консервативные гексануклеотиды UGCAUG окружены длинными консервативными участками с обеих сторон (пример показан на рис. 27А), причем часто консервативные участки примыкают к экзону (рис. 27Б, табл. 5). Это указывает на то, что регуляция альтернативного сплайсинга – довольно сложный процесс с участием многих белков, а исследуемый мотив – лишь часть системы цис-регуляторных элементов. Консервативность гексануклеотида UGCAUG в геномах позвоночных, а также его склонность находиться в консервативном окружении была позже подтверждена в работе [211]. В последующих работах было показано, что с гексануклеотидом UGCAUG специфически связываются транс-факторы сплайсинга Fox1 и Fox2 [212,213]. Гены этих факторов присутствуют в во многих эукариотических геномах от *C. elegans* до *H. sapiens* [214]. Было показано, что система UGCAUG-Fox1/2 участвует в тканеспецифичной регуляции альтернативных экзонов нескольких генов человека: EPB41 [213,215], CGRP [216], ATP5C1 [217], FGFR2 [218]. Кроме того, был произведён глобальный анализ сплайсинга, регулируемой системой UGCAUG-Fox1/2 [214]. Авторами исследования было показано, что гиперэкспрессия и нокдаун Fox1/2 приводит к изменению сплайсинга тех экзонов, в которых находятся UGCAUG-элементы с предсказанной функциональностью. Кроме того на нескольких примерах был показан механизм как именно связывание Fox1/2 с UGCAUG влияет на сплайсинг. На примере экзона 16 гена EPB41 было показано, что Fox2 связывается с энхансером находящимся в интроне после экзона, взаимодействует с U1C, что приводит к стабилизации

комплекса U1 с мРНК и активации донорного сайта сплайсинга [215]. Белки Fox-1 и Fox-2 связываются с UGCAUG, находящимися в интроне до экзона 4 гена кальцитонина, и регулируют сплайсинг путём блокирования связывания U2AF65 с акцепторным сайтом экзона 4[216]. In vitro эксперименты показали, что Fox1 препятствует формированию раннего комплекса (E-комплекса) на интроне 9 гена ATP5C1[217].

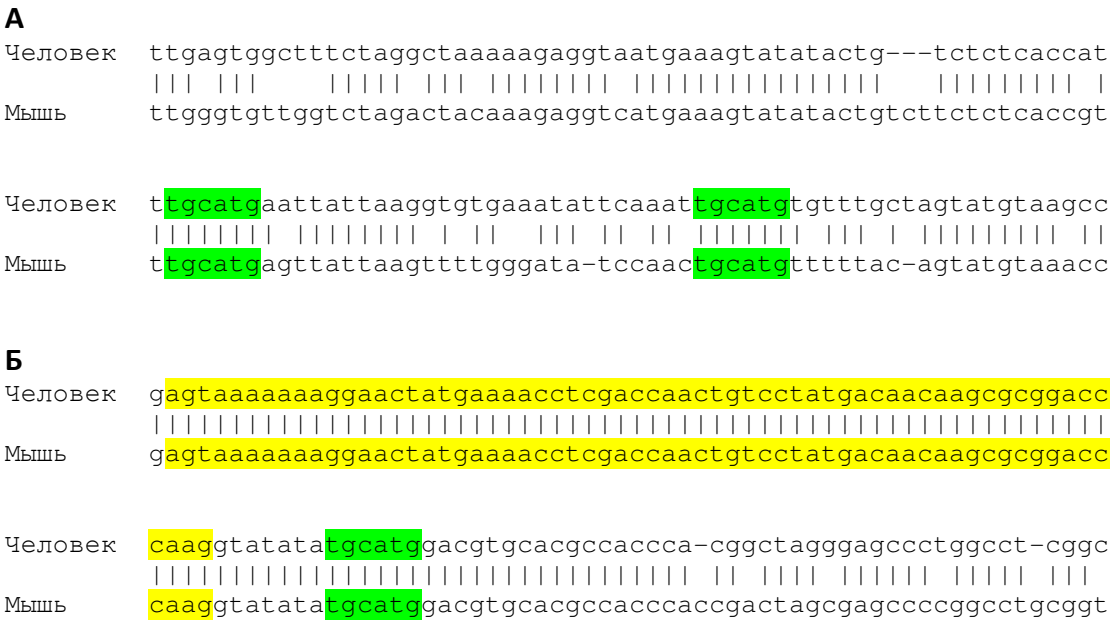


**Рисунок 27.** Гексануклеотиды находятся в консервативном окружении.

А — два гексануклеотида гена HDlg в консервативном окружении. Б — экзон (выделен жёлтым) и часть проксимальной консервативной последовательности, содержащей UGCAUG (NMDA-R1 exon 5).

### 4.3.2. Тканевая специфичность экспрессии кассетных экзонов, потенциально регулируемых UGCAUG

Ранее считалось, что все экзоны из нашей выборки являются мозгоспецифичными [167]. Мы произвели анализ распределения изоформ с кассетным экзоном по тканям и органам и выяснили, что строгой мозгоспецифичности не наблюдается. По меньшей мере 10 экзонов человека (№ 2,3,5,7,8,11,13,17,19,25 в Табл. 5) и 4 экзона мыши (№ 2,8,10,24 в Табл. 5) не являются мозгоспецифичными. Таким образом, возникло предположение, что UGCAUG не является строго мозгоспецифичным энхансером альтернативного сплайсинга. Стоит отметить, что

Brudno с соавт. [167] отмечали, что UGCAUG также избыточно представлен в небольшой группе мышцеспецифичных экзонов.

Анализ включения девяти экзонов из этой выборки в различных тканях мыши с помощью RT-PCR подтвердил, что часть из них экспрессируются также в скелетных мышцах и в лёгких, при этом все из них экспрессируются в мозге [211]. Было показано, что UGCAUG регулирует сплайсинг не только в нервной и мышечной тканях [213]. Тем не менее более позднее полногеномное исследование показало, что гены, сплайсинг которых регулируется системой UGCAUG-Fox1/2, чаще других выполняют функции связанные с нервной и мышечной системами [214].

# 5. Основные результаты и выводы

**Отбор в сайтах сплайсинга**

1. На неконсенсусные нуклеотиды действует слабый положительный, а на консенсусные, соответственно, слабый отрицательный отбор ($1<|4N_es|<4$). У части сайтов сплайсинга, однако, существует отбор на неконсенсусные нуклеотиды. Сила сайта сплайсинга влияет на силу отбора: на консенсусные нуклеотиды слабых сайтов сплайсинга действует более сильный отрицательный отбор.
2. На молодые сайты сплайсинга, появившиеся на линии человека после расхождения с макакой, действует отбор в 2-10 раз более сильный отбор, чем на старые сайты сплайсинга (т.е. общие для человека, макаки и игрунки).

**Коррелированная эволюция позиций в сайтах сплайсинга млекопитающих**

3. В донорных сайтах сплайсинга конститутивных экзонов наблюдается ослабление экзонной части и слабое увеличение силы интронной части сайтов, что согласуется с гипотезой о миграции сигнала из экзонной части в интронную.
4. Силы нуклеотидов в различных позициях сайтов сплайсинга часто взаимно скоррелированы. В донорных сайтах сплайсинга наблюдаются положительные ковариации между позициями внутри экзонной и внутри интронной частей и отрицательные – между экзонными и интронными частями. В полипиримидиновом тракте акцепторных сайтов сплайсинга наблюдается характерный чередующийся характер ковариаций: соседние позиции ковариируют отрицательно, через позицию – положительно, через две – отрицательно и т.д.
5. Отбор против динуклеотида AG есть основная причина ковариаций, наблюдаемых в полипиримидиновом тракте акцепторных сайтов сплайсинга. Эпистаз – наиболее вероятная причина ковариаций в донорном сайте сплайсинга.
6. Как в донорных, так и в акцепторных сайтах сплайсинга действует эпистатический отбор, который усиливает существующие ковариации между нуклеотидами.

**Консервативность цис-регулятора сплайсинга UGCAUG**

7. Консервативность гексануклеотидов UGCAUG, встречающиеся в интронах после кассетных экзонов в геномах человека и мыши существенно выше средней консервативности интронов, что говорит об их вероятной функциональности.

# 6. Приложение

**Таблица 1.** Распределение частоты использования экзонов до и после фильтрации.

| Тип сайтов | Частота включения экзоны | Число сайтов | | % | |
|---|---|---|---|---|---|
| | | До фильтрации | После фильтрации | До фильтрации | После фильтрации |
| **Донорные** | **5-10%** | 115 | 36 | 3 | 5 |
| | **10-30%** | 318 | 125 | 8 | 16 |
| | **30-70%** | 825 | 215 | 21 | 27 |
| | **70-95%** | 2671 | 415 | 68 | 52 |
| | **95-100%, >50 EST** | 8016 | 3597 | | |
| **Акцепторные** | **5-10%** | 115 | 35 | 3 | 4 |
| | **10-30%** | 318 | 129 | 8 | 16 |
| | **30-70%** | 825 | 223 | 20 | 28 |
| | **70-95%** | 2971 | 423 | 70 | 52 |
| | **95-100%, >50 EST** | 8016 | 3601 | | |

В каждой категории указано число сайтов кассетных экзонов и соотв. доля.

**Таблица 2.** Сравнение частот переходов из неконсенсусных нуклеотидов в консенсусные и обратно в сайтах сплайсинга и нейтральных контролях.

| Линия | Расположение сайтов сплайсинга | Тип сайтов | Cn->Nc | | | Nc->Cn | | |
|---|---|---|---|---|---|---|---|---|
| | | | Частота набл.[5] | Частота ожид.[6] | p-значение[7] | Частота набл. | Частота ожид. | p-значение |
| Человек | Код. посл.[1] | Дон.[3] | 0.0035 | 0.0154 | 0.00E+00 | 0.0099 | 0.0074 | 2.06E-25 |
| | | Акц.[4] | 0.0052 | 0.0069 | 3.75E-143 | 0.0090 | 0.0064 | 8.28E-126 |
| | Некод. посл.[2] | Дон. | 0.0037 | 0.0162 | 8.84E-90 | 0.0159 | 0.0082 | 1.09E-15 |
| | | Акц. | 0.0048 | 0.0078 | 1.35E-34 | 0.0109 | 0.0067 | 8.64E-24 |
| D. mel | Код. посл. | Дон. | 0.0104 | 0.0357 | 0.00E+00 | 0.0179 | 0.0180 | 8.29E-01 |
| | | Акц. | 0.0196 | 0.0246 | 9.71E-102 | 0.0240 | 0.0225 | 2.99E-06 |
| | Некод. посл. | Дон. | 0.0049 | 0.0222 | 5.29E-55 | 0.0142 | 0.0120 | 1.44E-01 |
| | | Акц. | 0.0113 | 0.0172 | 1.84E-19 | 0.0195 | 0.0160 | 3.85E-04 |

Код. посл.[1] - кодирующие последовательности белок-кодирующих генов; Некод. посл.[2] - некодирующие последовательности (5’ и 3’- нетранслируемые области белок-кодирующих генов и некодирующие РНК); Дон.[3] – донорные сайты сплайсинга; Акц.[4] – акцепторные сайты сплайсинга; Частота набл.[5] – частота переходов в сайтах сплайсинга (наблюдаемая); Частота ожид.[6] – частота переходов в нейтральном контроле (ожидаемая); p-значение[7] – p-значения точного теста Фишера (сравнение наблюдаемых и ожидаемых частот переходов).

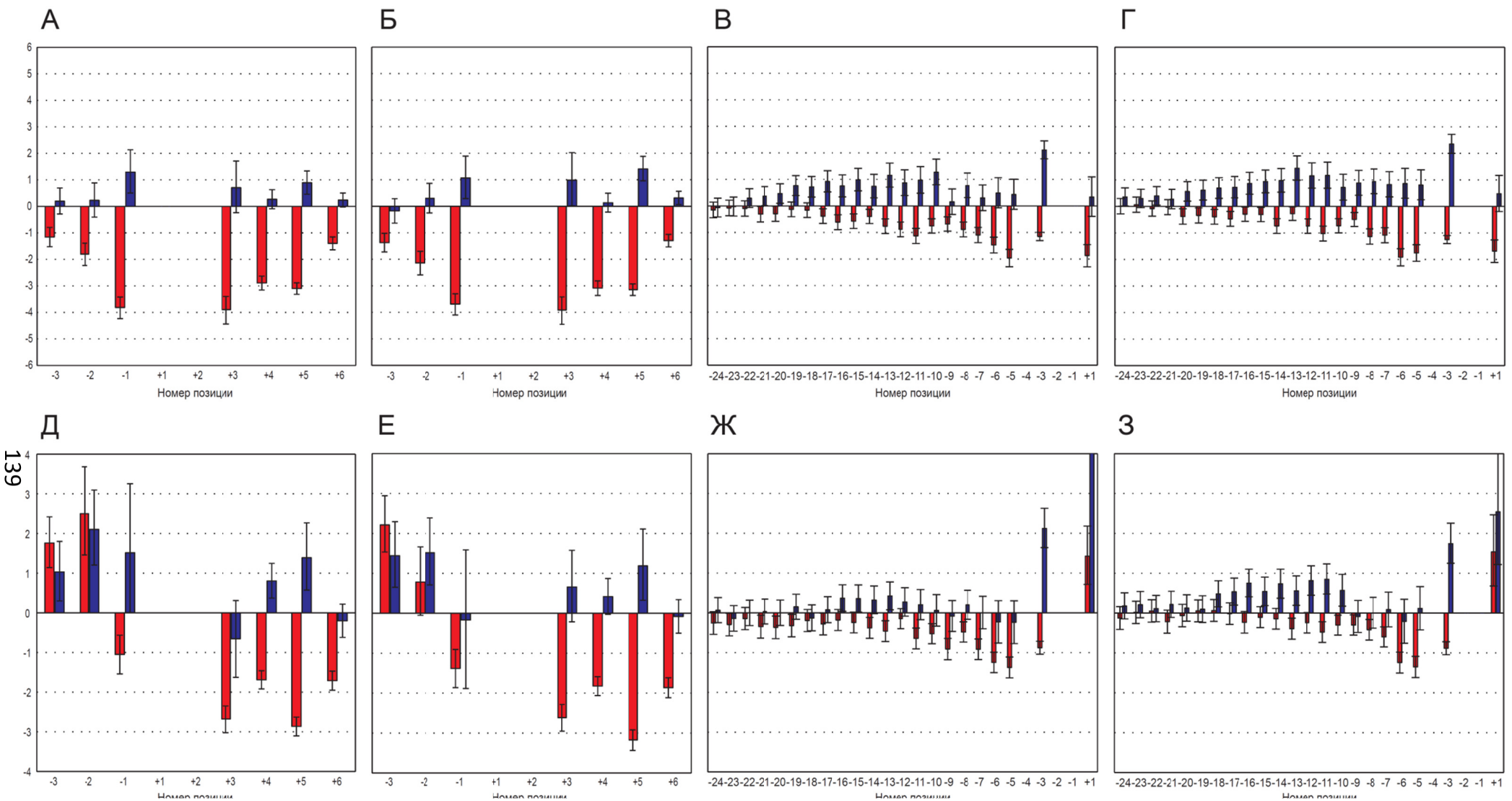


**Рисунок 1.** Сила отбора в районах генома с разным уровнем рекомбинации (кодирующие области генов).

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены *Nc→Cn* (синие столбики) и на замены *Cn→Nc* (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А–Г – линия *H. sapiens*; Д–З – линия *D. melanogaster*. А, Б, Д, Е – донорные сайты; В, Г, Ж, З – акцепторные сайты. А, В, Д, Ж – низкий уровень рекомбинации; Б, Г, Е, З – высокий уровень рекомбинации. Усы – 95%-ные доверительные интервалы.

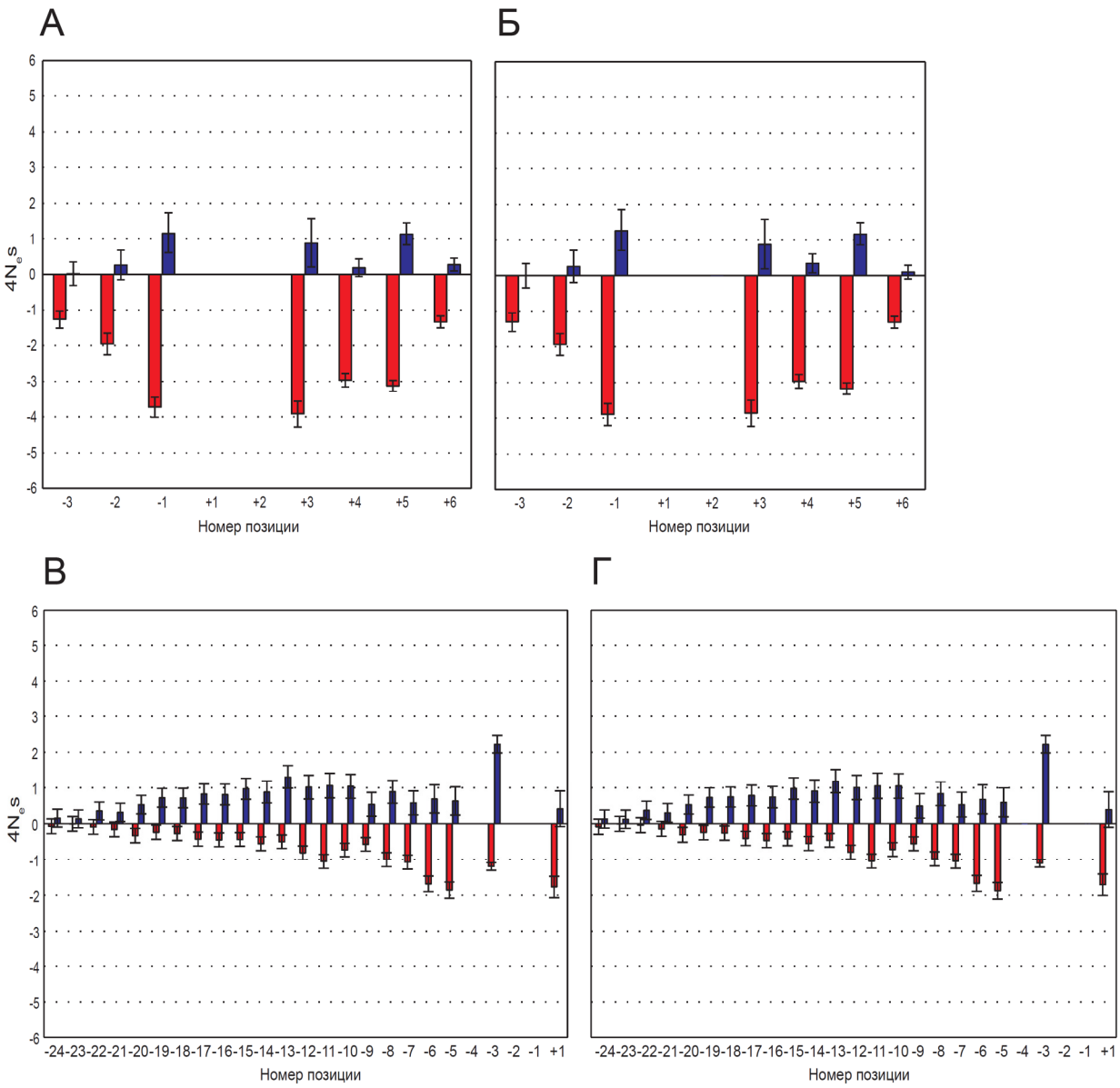


**Рисунок 2.** Сила отбора, действующего на позиции сайтов сплайсинга, в которых присутствуют и отсутствуют CpG (линия *H. sapiens,* кодирующие области генов).

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены Nc→Cn (синие столбики) и на замены Cn→Nc (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А, Б – донорные сайты; В,Г – акцепторные сайты. А, В – содержат CpG; Б, Г – не содержат CpG.

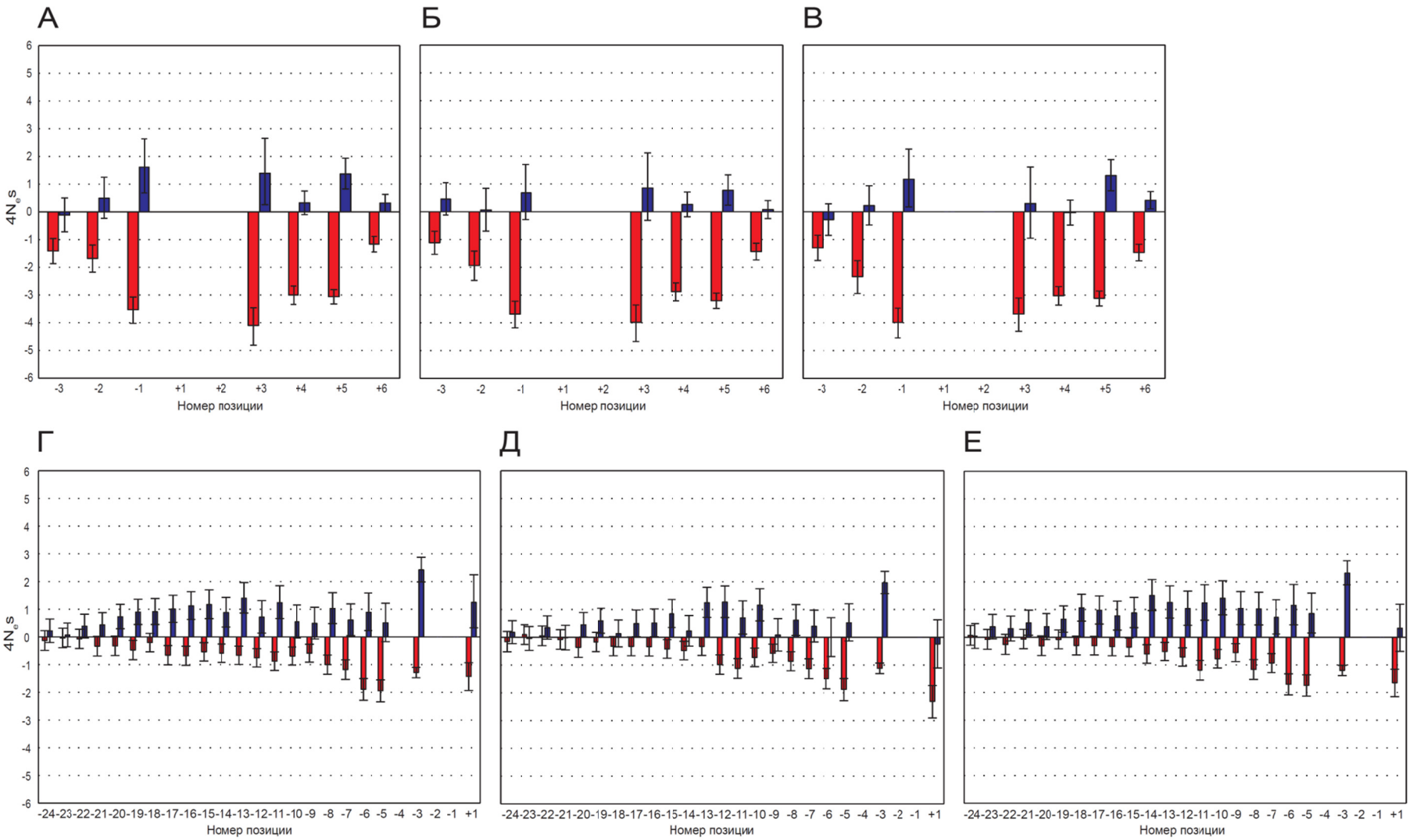


**Рисунок 3.** Сила отбора в сайтах сплайсинга, располагающихся в генах с разным средним уровнем экспрессии (линия *H. sapiens,* кодирующие области генов).

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены *Nc→Cn* (синие столбики) и на замены *Cn→Nc* (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А–В – донорные сайты; Г–Е – акцепторные сайты. А, Г – гены с низким уровнем экспрессии; Б, Д – гены со средним уровнем экспрессии; В, Е – гены с высоким уровнем экспрессии. Усы – 95%-ные доверительные интервалы.

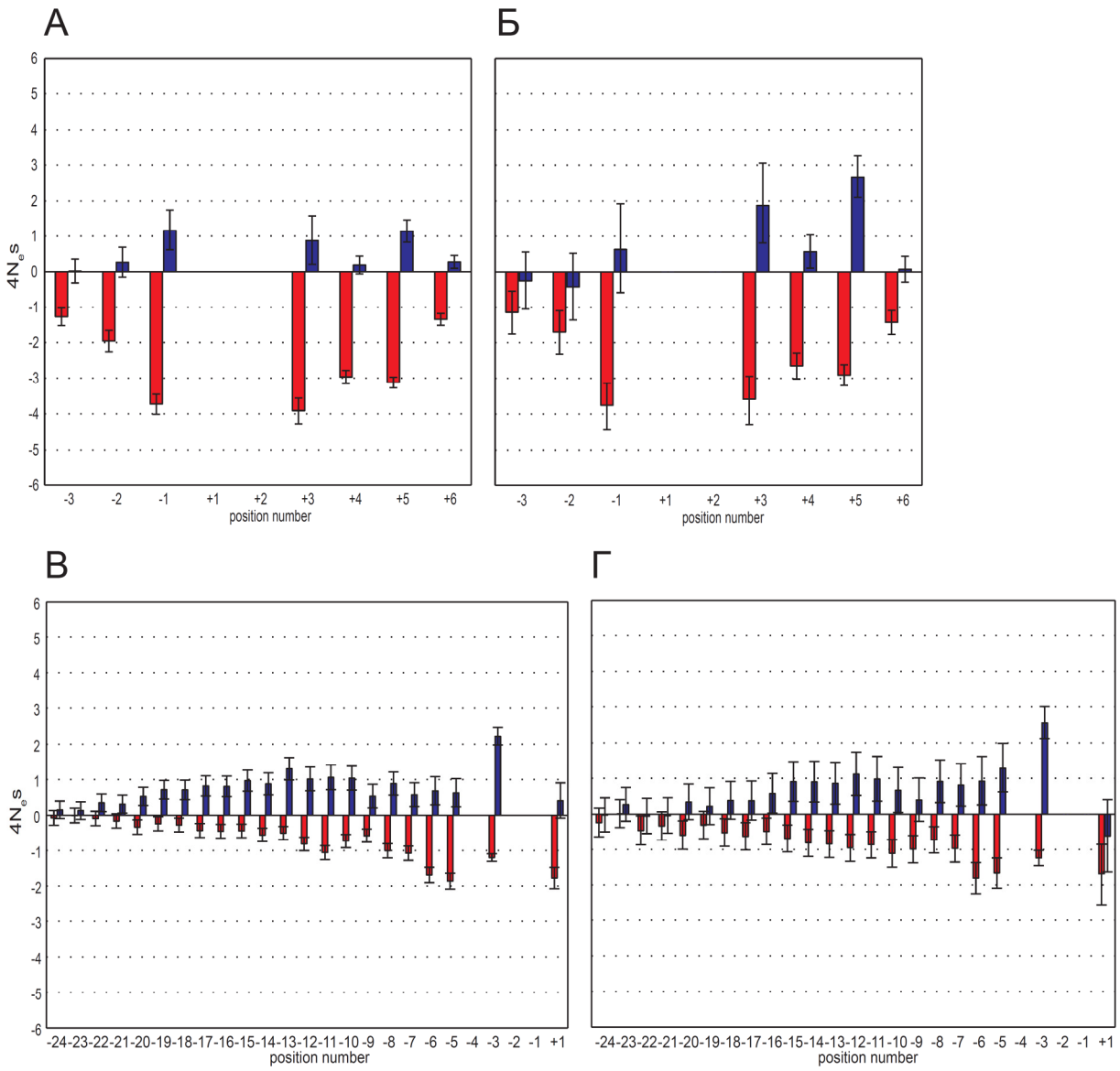


**Рисунок 4**. Сила отбора, действующего сайты сплайсинга конститутивных и кассетных экзонов (линия *H. sapiens*, кодирующие области генов).

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены *Nc→Cn* (синие столбики) и на замены *Cn→Nc* (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А, Б – донорные сайты; В, Г – акцепторные сайты. А, В – сайты сплайсинга конститутивных экзонов; Б, Г – сайты сплайсинга кассетных экзонов.

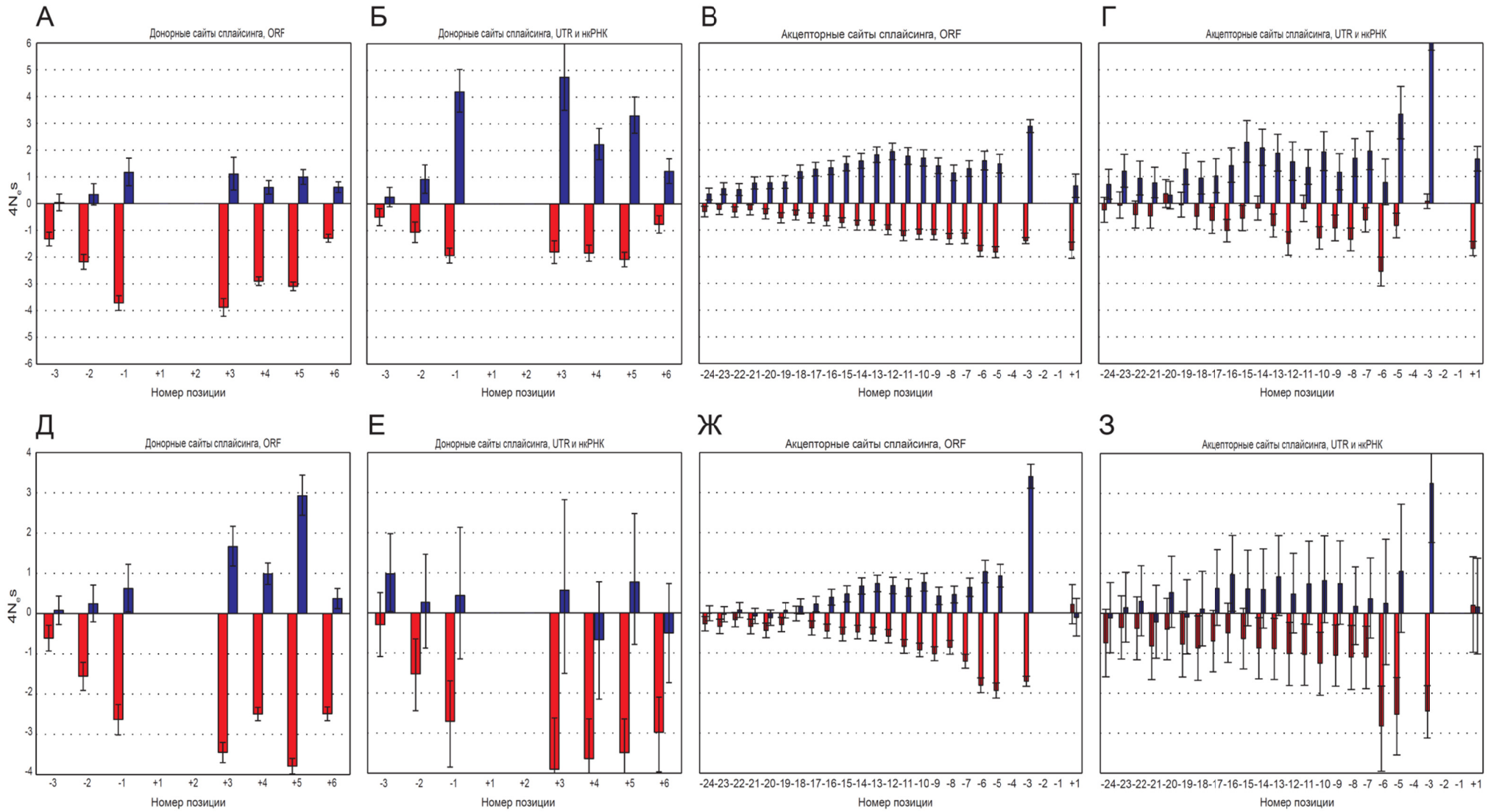


**Рисунок 5.** Сила отбора, измеренная методом измеренная методом максимального правдоподобия

Сила отбора, измеренная в единицах $4N_es$ (вертикальная ось), действующего на замены *Nc→Cn* (синие столбики) и на замены *Cn→Nc* (красные столбики). Положительные и отрицательные значения $4N_es$ соответствуют положительному и отрицательному отбору, соответственно. А–Г – линия *H. sapiens*; Д–З – линия *D. melanogaster*. А, Б, Д, Е – донорные сайты; В, Г, Ж, З – акцепторные сайты. А, В, Д, Ж – кодирующие сайты; Б, Г, Е, З – некодирующие сайты. Усы – 95%-ные доверительные интервалы.

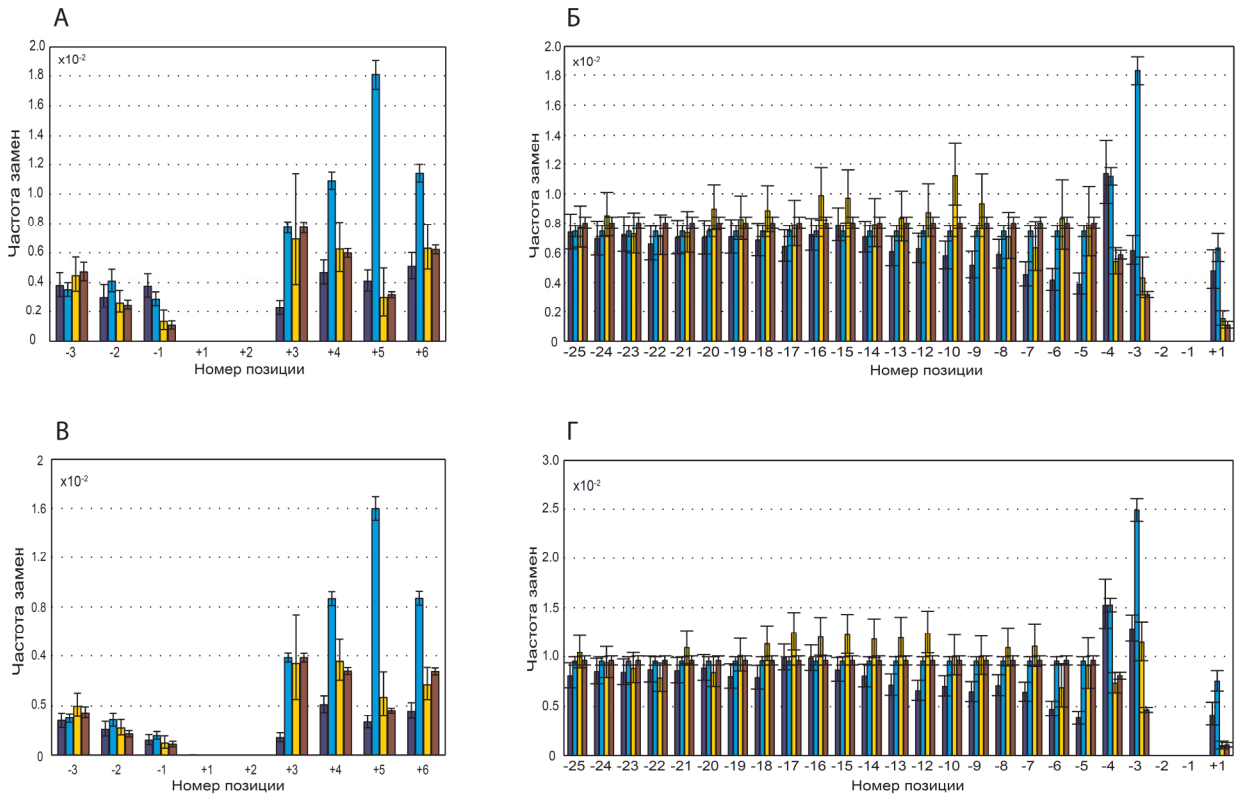


**Рисунок 6**. Сегрегирующие полиморфизмы в сайтах сплайсинга *D. melanogaster*.

По горизонтальной оси – частота замен между предковыми и производными аллелями (рассматривались замены из *Cn* в *Nc* и обратно), по вертикальной оси – номер позиции в сайте сплайсинга. Изображена частота замен *Cn→Nc* в сайтах сплайсинга (синие столбики) и в близлежащих интронах (голубые столбики); замен *Nc→Cn* в сайтах сплайсинга (желтые столбики) и в близлежащих интронах (коричневые столбики). А,В – донорные сайты сплайсинга; Б, Г – акцепторные сайты. А, Б – частота производного аллеля в сайтах сплайсинга 1-10%; В, Г - частота производного аллеля в сайтах сплайсинга 25-99%;

| | Конститутивные экзоны | | | Кассетные экзоны | | |
|---|---|---|---|---|---|---|
| Позиция | Собака | Человек | Мышь | Собака | Человек | Мышь |
| -3 | **-0.01** | **-0.01** | **-0.02** | -0.01 | 0.00 | -0.02 |
| -2 | **-0.04** | **-0.02** | **-0.08** | -0.05 | -0.01 | **-0.06** |
| -1 | -0.06 | **-0.09** | **-0.10** | -0.04 | -0.06 | **-0.13** |
| +3 | **-0.48** | **-0.32** | **-0.86** | **-0.29** | **-0.17** | **-0.47** |
| +4 | **-0.46** | **-0.29** | **-0.82** | **-0.17** | -0.09 | **-0.29** |
| +5 | **-0.63** | **-0.47** | **-1.04** | **-0.28** | **-0.22** | **-0.39** |
| +6 | **-0.11** | **-0.06** | **-0.22** | **-0.05** | **-0.04** | **-0.14** |

**Рисунок 7.** Ожидаемые средние изменения весов в каждой позиции донорных сайтов сплайсинга.

Отрицательные значения показаны красным, глубина цвета отражает абсолютное значение $\Delta\hat{W}_E(i)$. Значения в таблице выделены в соответствии с p-значением (отклонение от нулевой гипотезы: $\Delta\hat{W}_E(i) = 0$): **жирный шрифт**, $p < 0.01$; **<u>жирный и подчеркнутый шрифт</u>**, $p < 0.001$.

| | Конститутивные экзоны | | | Кассетные экзоны | | |
|---|---|---|---|---|---|---|
| Позиция | Собака | Человек | Мышь | Собака | Человек | Мышь |
| -13 | **<u>-0.11</u>** | **<u>-0.07</u>** | **<u>-0.23</u>** | **<u>-0.05</u>** | **<u>-0.04</u>** | **<u>-0.12</u>** |
| -12 | **<u>-0.15</u>** | **<u>-0.09</u>** | **<u>-0.30</u>** | **<u>-0.05</u>** | **<u>-0.05</u>** | **<u>-0.11</u>** |
| -11 | **<u>-0.18</u>** | **<u>-0.11</u>** | **<u>-0.34</u>** | **<u>-0.07</u>** | **<u>-0.06</u>** | **<u>-0.15</u>** |
| -10 | **<u>-0.17</u>** | **<u>-0.11</u>** | **<u>-0.32</u>** | **<u>-0.05</u>** | **<u>-0.04</u>** | **<u>-0.11</u>** |
| -9 | **<u>-0.13</u>** | **<u>-0.08</u>** | **<u>-0.24</u>** | **<u>-0.05</u>** | **<u>-0.04</u>** | **<u>-0.10</u>** |
| -8 | **<u>-0.13</u>** | **<u>-0.08</u>** | **<u>-0.24</u>** | **<u>-0.04</u>** | **<u>-0.03</u>** | **<u>-0.08</u>** |
| -7 | **<u>-0.13</u>** | **<u>-0.08</u>** | **<u>-0.24</u>** | **<u>-0.04</u>** | **<u>-0.04</u>** | **<u>-0.09</u>** |
| -6 | **<u>-0.22</u>** | **<u>-0.14</u>** | **<u>-0.39</u>** | **<u>-0.07</u>** | **<u>-0.05</u>** | **<u>-0.13</u>** |
| -5 | **<u>-0.23</u>** | **<u>-0.14</u>** | **<u>-0.42</u>** | **<u>-0.08</u>** | **<u>-0.06</u>** | **<u>-0.16</u>** |
| -4 | **-0.01** | 0.00 | **-0.02** | 0.00 | 0.00 | -0.01 |
| -3 | **<u>-0.70</u>** | **<u>-0.53</u>** | **<u>-1.14</u>** | **-0.31** | **-0.24** | **-0.50** |
| +1 | **<u>-0.11</u>** | **<u>-0.08</u>** | **<u>-0.19</u>** | **<u>-0.06</u>** | **-0.05** | **<u>-0.08</u>** |
| +2 | **<u>-0.02</u>** | -0.01 | **<u>-0.06</u>** | -0.01 | 0.00 | **-0.03** |

**Рисунок 8.** Ожидаемые средние изменения весов в каждой позиции акцепторных сайтов сплайсинга.

Отрицательные значения показаны красным, глубина цвета отражает абсолютное значение $\Delta\hat{W}_E(i)$. Значения в таблице выделены в соответствии с p-значением (отклонение от нулевой гипотезы: $\Delta\hat{W}_E(i) = 0$): **жирный шрифт**, $p < 0.01$; **<u>жирный и подчеркнутый шрифт</u>**, $p < 0.001$.

| Тип сплайсинга | Линия | -3 | -2 | -1 | +3 | +4 | +5 | +6 |
|---|---|---|---|---|---|---|---|---|
| Конститутивные экзоны | Собака | T+ | | | | | C+ | G- |
| | Человек | | G+ T+ A- C- | | A+ G- | | | G- |
| | Мышь | | C+ G+ A- T- | | A+ C+ G- | | C+ G+ A- T- | |
| Кассетные экзоны | Собака | | | | | | | |
| | Человек | | | | | A- | | |
| | Мышь | A- C+ | | A- | | | G+ | C+ |

**Рисунок 9.** Изменения частот нуклеотидов в различных позициях дононых сайтов сплайсинга.

Нуклеотиды, статистически значимо изменившие частоту ($p < 0.05$) в каждой позиции сайта сплайсинга показаны на каждой ветви филогенетического дерева и для разных типов сплайсинга. “+” обозначает увеличение частоты, “–“ — увеличение частоты соответствующего нуклеотида. Ячейки таблицы, где произошло статистически значимое изменение веса, раскрашены в соответствии с направлением изменения веса и соответствующим p-значением: красный — уменьшение веса, $p \leq 0.05$; оранжевый — уменьшение веса, $0.05 < p \leq 0.1$; зелёный — увеличение веса, $p \leq 0.05$; салатовый — увеличение веса, $0.05 < p \leq 0.1$.

| Тип сплайсинга | Линия | -13 | -12 | -11 | -10 | -9 | -8 | -7 | -6 | -5 | -4 | -3 | +1 | +2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Конститутивные экзоны | Собака | A+ C- | G+ | | | | | T+ C- | A+ C- | G+ | G+ T- | T+ C- | | |
| | Человек | A+ C- | A+ | A+ | C- | | C- | T+ C- | T+ C- | | | | A+ G- | T+ G- |
| | Мышь | A+ C- | G+ | | | | T+ C- | G+ T+ C- | | T+ A- | G+ T- | T+ C- | | G+ T- |
| Кассетные экзоны | Собака | | G+ T- | | | | | | | | | | G- | C+ T- |
| | Человек | C+ | G+ | | | | A+ C- | | G+ C- | | | | | |
| | Мышь | G+ | | | G+ | C+ | | | A- | | G+ A- T- | C+ | C+ T- | |

**Рисунок 10.** Изменения частот нуклеотидов в различных позициях акцепторных сайтов сплайсинга.

Нуклеотиды, статистически значимо изменившие частоту ($p < 0.05$) в каждой позиции сайта сплайсинга показаны на каждой ветви филогенетического дерева и для разных типов сплайсинга. “+” обозначает увеличение частоты, “–“ — увеличение частоты соответствующего нуклеотида. Ячейки таблицы, где произошло статистически значимое изменение веса, раскрашены в соответствии с направлением изменения веса и соответствующим p-значением: красный — уменьшение веса, $p \leq 0.05$; оранжевый — уменьшение веса, $0.05 < p \leq 0.1$; зелёный — увеличение веса, $p \leq 0.05$; салатовый — увеличение веса, $0.05 < p \leq 0.1$.

А

| | поз | -3 | -2 | -1 | +3 | +4 | +5 | +6 |
|---|---|---|---|---|---|---|---|---|
| ковариация | -3 | 0.26 | 0.11 | 0.06 | -0.03 | -0.04 | -0.07 | -0.05 |
| | -2 | 0.11 | 1.29 | 0.38 | -0.11 | -0.29 | -0.36 | -0.20 |
| | -1 | 0.06 | 0.38 | 1.81 | -0.13 | -0.35 | -0.35 | -0.33 |
| | +3 | -0.03 | -0.11 | -0.13 | 1.16 | -0.04 | -0.24 | -0.12 |
| | +4 | -0.04 | -0.29 | -0.35 | -0.04 | 1.87 | 0.44 | 0.02 |
| | +5 | -0.07 | -0.36 | -0.35 | -0.24 | 0.44 | 2.04 | 0.22 |
| | +6 | -0.05 | -0.20 | -0.33 | -0.12 | 0.02 | 0.22 | 0.74 |
| p-значения | -3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | -2 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | -1 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +3 | 0.00 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 |
| | +4 | 0.00 | 0.00 | 0.00 | 0.15 | 0.00 | 0.00 | 0.27 |
| | +5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.27 | 0.00 | 0.00 |

Б

| | поз | -3 | -2 | -1 | +3 | +4 | +5 | +6 |
|---|---|---|---|---|---|---|---|---|
| ковариация | -3 | 0.31 | 0.16 | 0.03 | -0.02 | -0.06 | -0.11 | -0.05 |
| | -2 | 0.16 | 1.35 | 0.30 | -0.05 | -0.29 | -0.35 | -0.25 |
| | -1 | 0.03 | 0.30 | 2.07 | -0.12 | -0.33 | -0.40 | -0.37 |
| | +3 | -0.02 | -0.05 | -0.12 | 0.98 | 0.00 | -0.18 | -0.12 |
| | +4 | -0.06 | -0.29 | -0.33 | 0.00 | 1.74 | 0.29 | 0.02 |
| | +5 | -0.11 | -0.35 | -0.40 | -0.18 | 0.29 | 1.90 | 0.21 |
| | +6 | -0.05 | -0.25 | -0.37 | -0.12 | 0.02 | 0.21 | 0.74 |
| p-значения | -3 | 0.00 | 0.00 | 0.05 | 0.10 | 0.00 | 0.00 | 0.00 |
| | -2 | 0.00 | 0.00 | 0.00 | 0.02 | 0.00 | 0.00 | 0.00 |
| | -1 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +3 | 0.10 | 0.02 | 0.00 | 0.00 | 0.47 | 0.00 | 0.00 |
| | +4 | 0.00 | 0.00 | 0.00 | 0.47 | 0.00 | 0.00 | 0.25 |
| | +5 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| | +6 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.00 | 0.00 |

**Рисунок 11**. Ковариационные матрицы высоко- (А) и низкоконсервативных (Б) донорных сайтов сплайсинга.

# 8. Благодарности

Автор благодарит родных и друзей за веру в лучшее, научного руководителя за предоставленные возможности, долготерпение и существенную редакторскую правку, Егора Базыкина, внесшего неоценимый вклад в формулировку идей и написание статей, а также уважаемых коллег за важные и конструктивные обсуждения.

# 8. Список литературы


1. Chow LT, Gelinas RE, Broker TR, Roberts RJ. An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA. Cell. 1977;12: 1–8. doi:10.1016/0092-8674(77)90180-5

2. Berget SM, Moore C, Sharp PA. Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. PNAS. 1977;74: 3171–3175. doi:10.1073/pnas.74.8.3171

3. Chow LT, Broker TR. The spliced structures of adenovirus 2 fiber message and the other late mRNAs. Cell. 1978;15: 497–510. doi:10.1016/0092-8674(78)90019-3

4. Nevins JR, Darnell JE. Steps in the processing of Ad2 mRNA: Poly(A)+ Nuclear sequences are conserved and poly(A) addition precedes splicing. Cell. 1978;15: 1477–1493. doi:10.1016/0092-8674(78)90071-5

5. Wolfsberg TG, Landsman D. A Comparison of Expressed Sequence Tags (ESTs) to Human Genomic Sequences. Nucl Acids Res. 1997;25: 1626–1632. doi:10.1093/nar/25.8.1626

6. Mironov AA, Fickett JW, Gelfand MS. Frequent Alternative Splicing of Human Genes. Genome Res. 1999;9: 1288–1293. doi:10.1101/gr.9.12.1288

7. Brett D, Hanke J, Lehmann G, Haase S, Delbrück S, Krueger S, et al. EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. FEBS Letters. 2000;474: 83–86. doi:10.1016/S0014-5793(00)01581-7

8. Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, et al. Novel RNAs Identified From an In-Depth Analysis of the Transcriptome of Human Chromosomes 21 and 22. Genome Res. 2004;14: 331–342. doi:10.1101/gr.2094104

9. Artamonova II, Gelfand MS. Comparative Genomics and Evolution of Alternative Splicing: The Pessimists' Science. ChemInform. 2007;38: no–no. doi:10.1002/chin.200746273

10. Lodish H, Berk A, Matsudaira P, Zipursky SL, Baltimore D, Darnell J. Molecular Cell Biology. Macmillan Higher Education; 1995.

11. Collins L, Penny D. Complex Spliceosomal Organization Ancestral to Extant Eukaryotes. Molecular Biology and Evolution. 2005;22: 1053–1066. doi:10.1093/molbev/msi091

12. Patel AA, Steitz JA. Splicing double: insights from the second spliceosome. Nature Reviews Molecular Cell Biology. 2003;4: 960–970. doi:10.1038/nrm1259

13. Sheth N, Roca X, Hastings ML, Roeder T, Krainer AR, Sachidanandam R. Comprehensive splice-site analysis using comparative genomics. Nucleic Acids Research. 2006;34: 3955–3967. doi:10.1093/nar/gkl556

14. Ast G. How did alternative splicing evolve? Nat Rev Genet. 2004;5: 773–782. doi:10.1038/nrg1451

15. Roca X, Sachidanandam R, Krainer AR. Determinants of the inherent strength of human 5′ splice sites. RNA. 2005;11: 683–698. doi:10.1261/rna.2040605

16. Schwartz S, Silva J, Burstein D, Pupko T, Eyras E, Ast G. Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Research. 2008;18: 88–103. doi:10.1101/gr.6818908

17. Mercer TR, Clark MB, Andersen SB, Brunck ME, Haerty W, Crawford J, et al. Genome-wide discovery of human splicing branchpoints. Genome Res. 2015; gr.182899.114. doi:10.1101/gr.182899.114

18. Breathnach R, Chambon P. Organization and expression of eucaryotic split genes coding for proteins. Annual review of biochemistry. 1981;50: 349–383.

19. Cartegni L, Chew SL, Krainer AR. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet. 2002;3: 285–298. doi:10.1038/nrg775

20. Denisov SV, Bazykin GA, Sutormin R, Favorov AV, Mironov AA, Gelfand MS, et al. Weak Negative and Positive Selection and the Drift Load at Splice Sites. Genome Biol Evol. 2014;6: 1437–1447. doi:10.1093/gbe/evu100

21. Szafranski K, Schindler S, Taudien S, Hiller M, Huse K, Jahn N, et al. Violating the splicing rules: TG dinucleotides function as alternative 3′ splice sites in U2-dependent introns. Genome Biol. 2007;8: R154. doi:10.1186/gb-2007-8-8-r154

22. Michaud S, Reed R. An ATP-independent complex commits pre-mRNA to the mammalian spliceosome assembly pathway. Genes Dev. 1991;5: 2534–2546.

23. Seraphin B, Rosbash M. Identification of functional U1 snRNA-pre-mRNA complexes committed to spliceosome assembly and splicing. Cell. 1989;59: 349–358.

24. Kretzner L, Rymond BC, Rosbash M. S. cerevisiae U1 RNA is large and has limited primary sequence homology to metazoan U1 snRNA. Cell. 1987;50: 593–602.

25. Lerner MR, Boyle JA, Mount SM, Wolin SL, Steitz JA. Are snRNPs involved in splicing? Nature. 1980;283: 220–224.

26. Zhuang Y, Weiner AM. A compensatory base change in U1 snRNA suppresses a 5′ splice site mutation. Cell. 1986;46: 827–835.

27. Séraphin B, Kretzner L, Rosbash M. A U1 snRNA:pre-mRNA base pairing interaction is required early in yeast spliceosome assembly but does not uniquely define the 5’ cleavage site. EMBO J. 1988;7: 2533–2538.

28. Du H, Rosbash M. The U1 snRNP protein U1C recognizes the 5’ splice site in the absence of base pairing. Nature. 2002;419: 86–90. doi:10.1038/nature00947

29. Wyatt JR, Sontheimer EJ, Steitz JA. Site-specific cross-linking of mammalian U5 snRNP to the 5’ splice site before the first step of pre-mRNA splicing. Genes & Development. 1992;6: 2542–2553. doi:10.1101/gad.6.12b.2542

30. Rossi F, Forné T, Antoine E, Tazi J, Brunel C, Cathala G. Involvement of U1 Small Nuclear Ribonucleoproteins (snRNP) in 5’ Splice Site-U1 snRNP Interaction. Journal of Biological Chemistry. 1996;271: 23985–23991. doi:10.1074/jbc.271.39.23985

31. Segault V, Will CL, Polycarpou-Schwarz M, Mattaj IW, Branlant C, Luhrmann R. Conserved Loop I of U5 Small Nuclear RNA Is Dispensable for Both Catalytic Steps of Pre-mRNA Splicing in HeLa Nuclear Extracts. Mol Cell Biol. 1999;19: 2782–2790.

32. Smith CWJ, Valcárcel J. Alternative pre-mRNA splicing: the logic of combinatorial control. Trends in Biochemical Sciences. 2000;25: 381–388. doi:10.1016/S0968-0004(00)01604-2

33. Manceau V, Swenson M, Le Caer J, Sobel A, Kielkopf CL, Maucuer A. Major phosphorylation of SF1 on adjacent Ser-Pro motifs enhances interaction with U2AF65. FEBS Journal. 2006;273: 577–587. doi:10.1111/j.1742-4658.2005.05091.x

34. Valcárcel J, Gaur RK, Singh R, Green MR. Interaction of U2AF65 RS Region with Pre-mRNA of Branch Point and Promotion Base Pairing with U2 snRNA. Science. 1996;273: 1706–1709. doi:10.1126/science.273.5282.1706

35. Black DL. Mechanisms of alternative pre-messenger RNA splicing. Annu Rev Biochem. 2003;72: 291–336. doi:10.1146/annurev.biochem.72.121801.161720

36. Shcherbakova I, Hoskins AA, Friedman LJ, Serebrov V, Corrêa IR, Xu M-Q, et al. Alternative spliceosome assembly pathways revealed by single-molecule fluorescence microscopy. Cell Rep. 2013;5: 151–165. doi:10.1016/j.celrep.2013.08.026

37. Chiara MD, Palandjian L, Feld Kramer R, Reed R. Evidence that U5 snRNP recognizes the 3’ splice site for catalytic step II in mammals. EMBO J. 1997;16: 4746–4759. doi:10.1093/emboj/16.15.4746

38. Wang Z, Burge CB. Splicing regulation: From a parts list of regulatory elements to an integrated splicing code. RNA. 2008;14: 802–813. doi:10.1261/rna.876308

39. De Conti L, Baralle M, Buratti E. Exon and intron definition in pre-mRNA splicing. WIREs RNA. 2013;4: 49–60. doi:10.1002/wrna.1140

40. Maniatis T, Tasic B. Alternative pre-mRNA splicing and proteome expansion in metazoans. Nature. 2002;418: 236–243. doi:10.1038/418236a

41. Schneider M, Will CL, Anokhina M, Tazi J, Urlaub H, Lührmann R. Exon Definition Complexes Contain the Tri-snRNP and Can Be Directly Converted into B-like Precatalytic Splicing Complexes. Molecular Cell. 2010;38: 223–235. doi:10.1016/j.molcel.2010.02.027

42. Dominski Z, Kole R. Selection of Splice Sites in Pre-Messenger-Rnas with Short Internal Exons. Mol Cell Biol. 1991;11: 6075–6083.

43. Black D. Does Steric Interference Between Splice Sites Block the Splicing of a Short C-Src Neuron-Specific Exon in Nonneuronal Cells. Genes Dev. 1991;5: 389–402. doi:10.1101/gad.5.3.389

44. Nakai K, Sakamoto H. Construction of a novel database containing aberrant splicing mutations of mammalian genes. Gene. 1994;141: 171–177. doi:10.1016/0378-1119(94)90567-3

45. Guo M, Lo PC, Mount SM. Species-specific signals for the splicing of a short Drosophila intron in vitro. Mol Cell Biol. 1993;13: 1104–1118. doi:10.1128/MCB.13.2.1104

46. Talerico M, Berget SM. Intron definition in splicing of small Drosophila introns. Mol Cell Biol. 1994;14: 3434–3445. doi:10.1128/MCB.14.5.3434

47. Schmucker D, Clemens JC, Shu H, Worby CA, Xiao J, Muda M, et al. Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. Cell. 2000;101: 671–684. doi:10.1016/S0092-8674(00)80878-8

48. Keren H, Lev-Maor G, Ast G. Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet. 2010;11: 345–355. doi:10.1038/nrg2776

49. Stamm S, Zhang MQ, Marr TG, Helfman DM. A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Research. 1994;22: 1515–1526. doi:10.1093/nar/22.9.1515

50. Stamm S, Zhu J, Nakai K, Stoilov P, Stoss O, Zhang MQ. An alternative-exon database and its statistical analysis. DNA and Cell Biology. 2000;19: 739–756.

51. Clark F, Thanaraj TA. Categorization and characterization of transcript-confirmed constitutively and alternatively spliced introns and exons from human. Human Molecular Genetics. 2002;11: 451.

52. Baek D, Green P. Sequence conservation, relative isoform frequencies, and nonsense-mediated decay in evolutionarily conserved alternative splicing.

Proceedings of the National Academy of Sciences of the United States of America. 2005;102: 12813.

53. Zheng CL, Fu X-D, Gribskov M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA. 2005;11: 1777–1787. doi:10.1261/rna.2660805

54. Fox-Walsh KL, Dou Y, Lam BJ, Hung S, Baldi PF, Hertel KJ. The architecture of pre-mRNAs affects mechanisms of splice-site pairing. PNAS. 2005;102: 16176–16181. doi:10.1073/pnas.0508489102

55. Lopez AJ. Alternative Splicing Of Pre-mRNA: Developmental Consequences and Mechanisms of Regulation. Annual Review of Genetics. 1998;32: 279–305. doi:10.1146/annurev.genet.32.1.279

56. Matlin AJ, Clark F, Smith CWJ. Understanding alternative splicing: towards a cellular code. Nat Rev Mol Cell Biol. 2005;6: 386–398. doi:10.1038/nrm1645

57. Buratti E, Baralle FE. Influence of RNA Secondary Structure on the Pre-mRNA Splicing Process. Mol Cell Biol. 2004;24: 10505–10514. doi:10.1128/MCB.24.24.10505-10514.2004

58. Kornblihtt AR, Schor IE, Alló M, Dujardin G, Petrillo E, Muñoz MJ. Alternative splicing: a pivotal step between eukaryotic transcription and translation. Nat Rev Mol Cell Biol. 2013;14: 153–165. doi:10.1038/nrm3525

59. Das R, Yu J, Zhang Z, Gygi MP, Krainer AR, Gygi SP, et al. SR Proteins Function in Coupling RNAP II Transcription to Pre-mRNA Splicing. Molecular Cell. 2007;26: 867–881. doi:10.1016/j.molcel.2007.05.036

60. de la Mata M, Kornblihtt AR. RNA polymerase II C-terminal domain mediates regulation of alternative splicing by SRp20. Nat Struct Mol Biol. 2006;13: 973–980. doi:10.1038/nsmb1155

61. Spies N, Nielsen CB, Padgett RA, Burge CB. Biased Chromatin Signatures around Polyadenylation Sites and Exons. Molecular Cell. 2009;36: 245–254. doi:10.1016/j.molcel.2009.10.008

62. Kolasinska-Zwierz P, Down T, Latorre I, Liu T, Liu XS, Ahringer J. Differential chromatin marking of introns and expressed exons by H3K36me3. Nat Genet. 2009;41: 376–381. doi:10.1038/ng.322

63. Schor IE, Rascovan N, Pelisch F, Alló M, Kornblihtt AR. Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing. PNAS. 2009;106: 4325–4330. doi:10.1073/pnas.0810666106

64. Andersson R, Enroth S, Rada-Iglesias A, Wadelius C, Komorowski J. Nucleosomes are well positioned in exons and carry characteristic histone modifications. Genome Res. 2009;19: 1732–1741. doi:10.1101/gr.092353.109

65. Schwartz S, Meshorer E, Ast G. Chromatin organization marks exon-intron structure. Nat Struct Mol Biol. 2009;16: 990–995. doi:10.1038/nsmb.1659

66. Cáceres JF, Kornblihtt AR. Alternative splicing: multiple control mechanisms and involvement in human disease. Trends in Genetics. 2002;18: 186–193. doi:10.1016/S0168-9525(01)02626-9

67. Graveley BR. Sorting Out the Complexity of SR Protein Functions. RNA. 2000;6: 1197–1211. doi:null

68. Kawamoto S. Neuron-specific Alternative Splicing of Nonmuscle Myosin II Heavy Chain-B Pre-mRNA Requires a Cis-acting Intron Sequence. J Biol Chem. 1996;271: 17613–17616.

69. Modafferi EF, Black DL. A complex intronic splicing enhancer from the c-src pre-mRNA activates inclusion of a heterologous exon. Mol Cell Biol. 1997;17: 6537–6545. doi:10.1128/MCB.17.11.6537

70. Huh GS, Hynes RO. Regulation of alternative pre-mRNA splicing by a novel repeated hexanucleotide element. Genes Dev. 1994;8: 1561–1574. doi:10.1101/gad.8.13.1561

71. Hedjran F, Yeakley JM, Huh GS, Hynes RO, Rosenfeld MG. Control of alternative pre-mRNA splicing by distributed pentameric repeats. PNAS. 1997;94: 12343–12347.

72. McCullough AJ, Berget SM. G triplets located throughout a class of small vertebrate introns enforce intron borders and regulate splice site selection. Mol Cell Biol. 1997;17: 4562–4571. doi:10.1128/MCB.17.8.4562

73. McCullough AJ, Berget SM. An Intronic Splicing Enhancer Binds U1 snRNPs To Enhance Splicing and Select 5′ Splice Sites. Mol Cell Biol. 2000;20: 9225–9235. doi:10.1128/MCB.20.24.9225-9235.2000

74. Liu H-X, Zhang M, Krainer AR. Identification of functional exonic splicing enhancer motifs recognized by individual SR proteins. Genes Dev. 1998;12: 1998–2012. doi:10.1101/gad.12.13.1998

75. Tian H, Kole R. Selection of novel exon recognition elements from a pool of random sequences. Mol Cell Biol. 1995;15: 6291–6298. doi:10.1128/MCB.15.11.6291

76. Coulter LR, Landree MA, Cooper TA. Identification of a new class of exonic splicing enhancers by in vivo selection. Mol Cell Biol. 1997;17: 2143–2150. doi:10.1128/MCB.17.4.2143

77. Wang Z, Rolish ME, Yeo G, Tung V, Mawson M, Burge CB. Systematic Identification and Analysis of Exonic Splicing Silencers. Cell. 2004;119: 831–845. doi:10.1016/j.cell.2004.11.010

78. Stormo GD. DNA binding sites: representation and discovery. Bioinformatics. 2000;16: 16–23. doi:10.1093/bioinformatics/16.1.16

79. Cartegni L, Wang J, Zhu Z, Zhang MQ, Krainer AR. ESEfinder: a web resource to identify exonic splicing enhancers. Nucl Acids Res. 2003;31: 3568–3571. doi:10.1093/nar/gkg616

80. Majewski J, Ott J. Distribution and Characterization of Regulatory Elements in the Human Genome. Genome Res. 2002;12: 1827–1836. doi:10.1101/gr.606402

81. Fedorov A, Saxonov S, Fedorova L, Daizadeh I. Comparison of intron-containing and intron-lacking human genes elucidates putative exonic splicing enhancers. Nucl Acids Res. 2001;29: 1464–1469. doi:10.1093/nar/29.7.1464

82. Zhang XH-F, Chasin LA. Computational definition of sequence motifs governing constitutive exon splicing. Genes Dev. 2004;18: 1241–1250. doi:10.1101/gad.1195304

83. Zhang XH-F, Kangsamaksin T, Chao MSP, Banerjee JK, Chasin LA. Exon Inclusion Is Dependent on Predictable Exonic Splicing Enhancers. Mol Cell Biol. 2005;25: 7323–7332. doi:10.1128/MCB.25.16.7323-7332.2005

84. Fairbrother WG, Yeh R-F, Sharp PA, Burge CB. Predictive Identification of Exonic Splicing Enhancers in Human Genes. Science. 2002;297: 1007–1013. doi:10.1126/science.1073774

85. Fairbrother WG, Holste D, Burge CB, Sharp PA. Single Nucleotide Polymorphism–Based Validation of Exonic Splicing Enhancers. PLOS Biol. 2004;2: e268. doi:10.1371/journal.pbio.0020268

86. Yeo G, Hoon S, Venkatesh B, Burge CB. Variation in sequence and organization of splicing regulatory elements in vertebrate genes. PNAS. 2004;101: 15700–15705. doi:10.1073/pnas.0404901101

87. Sorek R, Ast G. Intronic Sequences Flanking Alternatively Spliced Exons Are Conserved Between Human and Mouse. Genome Research. 2003;13: 1631–1637. doi:10.1101/gr.1208803

88. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct. 2012;7: 1–28. doi:10.1186/1745-6150-7-11

89. Russell AG, Charette JM, Spencer DF, Gray MW. An early evolutionary origin for the minor spliceosome. Nature. 2006;443: 863–866. doi:10.1038/nature05228

90. William Roy S, Gilbert W. The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet. 2006;7: 211–221. doi:10.1038/nrg1807

91. Lambowitz AM, Zimmerly S. Group II Introns: Mobile Ribozymes that Invade DNA. Cold Spring Harb Perspect Biol. 2011;3: a003616. doi:10.1101/cshperspect.a003616

92. Toor N, Keating KS, Taylor SD, Pyle AM. Crystal Structure of a Self-Spliced Group II Intron. Science. 2008;320: 77–82. doi:10.1126/science.1153803

93. Gilbert W. The Exon Theory of Genes. Cold Spring Harb Symp Quant Biol. 1987;52: 901–905. doi:10.1101/SQB.1987.052.01.098

94. Logsdon Jr JM. The recent origins of spliceosomal introns revisited. Current Opinion in Genetics & Development. 1998;8: 637–648. doi:10.1016/S0959-437X(98)80031-2

95. Cavalier-Smith T. Intron phylogeny: a new hypothesis. Trends in Genetics. 1991;7: 145–148. doi:10.1016/0168-9525(91)90377-3

96. Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biology Direct. 2006;1: 22. doi:10.1186/1745-6150-1-22

97. Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV. Remarkable Interkingdom Conservation of Intron Positions and Massive, Lineage-Specific Intron Loss and Gain in Eukaryotic Evolution. Current Biology. 2003;13: 1512–1517. doi:10.1016/S0960-9822(03)00558-X

98. Gelfand MS. Statistical analysis and prediction of the exonic structure of human genes. J Mol Evol. 1992;35: 239–252. doi:10.1007/BF00178600

99. Tarrío R, Rodríguez-Trelles F, Ayala FJ. A new Drosophila spliceosomal intron position is common in plants. PNAS. 2003;100: 6580–6583. doi:10.1073/pnas.0731952100

100. Hankeln T, Friedl H, Ebersberger I, Martin J, Schmidt ER. A variable intron distribution in globin genes of Chironomus: evidence for recent intron gain. Gene. 1997;205: 151–160. doi:10.1016/S0378-1119(97)00518-0

101. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Conservation versus parallel gains in intron evolution. Nucl Acids Res. 2005;33: 1741–1748. doi:10.1093/nar/gki316

102. Roy SW, Gilbert W. Complex early genes. PNAS. 2005;102: 1986–1991. doi:10.1073/pnas.0408355101

103. Carmel L, Wolf YI, Rogozin IB, Koonin EV. Three distinct modes of intron dynamics in the evolution of eukaryotes. Genome Res. 2007;17: 1034–1044. doi:10.1101/gr.6438607

104. Csuros M, Rogozin IB, Koonin EV. A Detailed History of Intron-rich Eukaryotic Ancestors Inferred from a Global Survey of 100 Complete Genomes. PLOS Comput Biol. 2011;7: e1002150. doi:10.1371/journal.pcbi.1002150

105. Irimia M, Penny D, Roy SW. Coevolution of genomic intron number and splice sites. Trends in Genetics. 2007;23: 321–325. doi:10.1016/j.tig.2007.04.001

106. Sickmier EA, Frato KE, Shen H, Paranawithana SR, Green MR, Kielkopf CL. Structural Basis for Polypyrimidine Tract Recognition by the Essential Pre-mRNA Splicing Factor U2AF65. Molecular Cell. 2006;23: 49–59. doi:10.1016/j.molcel.2006.05.025

107. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Evidence of Splice Signal Migration from Exon to Intron during Intron Evolution. Current Biology. 2003;13: 2170–2174. doi:10.1016/j.cub.2003.12.003

108. Dibb NJ, Newman AJ. Evidence that introns arose at proto-splice sites. EMBO J. 1989;8: 2015–2021.

109. Sadusky T, Newman AJ, Dibb NJ. Exon junction sequences as cryptic splice sites: implications for intron origin. Curr Biol. 2004;14: 505–509. doi:10.1016/j.cub.2004.02.063

110. Sverdlov AV, Rogozin IB, Babenko VN, Koonin EV. Reconstruction of Ancestral Protosplice Sites. Current Biology. 2004;14: 1505–1508. doi:10.1016/j.cub.2004.08.027

111. Lynch M. Intron Evolution as a Population-Genetic Process. Proceedings of the National Academy of Sciences of the United States of America. 2002;99: 6118–6123.

112. Lynch M, Conery JS. The Origins of Genome Complexity. Science. 2003;302: 1401–1404. doi:10.1126/science.1089370

113. Charlesworth B. Effective population size and patterns of molecular evolution and variation. Nature Reviews Genetics. 2009;10: 195–205. doi:10.1038/nrg2526

114. Carvalho AB, Clark AG. Genetic recombination: Intron size and natural selection. Nature. 1999;401: 344–344. doi:10.1038/43827

115. Comeron JM, Kreitman M. The Correlation Between Intron Length and Recombination in Drosophila: Dynamic Equilibrium Between Mutational and Selective Forces. Genetics. 2000;156: 1175–1190.

116. Duret L, Mouchiroud D, Gautier C. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. J Mol Evol. 1995;40: 308–317. doi:10.1007/BF00163235

117. The Logic of Chance: The Nature and Origin of Biological Evolution - Eugene V. Koonin - Google Books [Internet]. [cited 24 Apr 2016]. Available: https://books.google.ru/books?hl=en&lr=&id=fvmv2kU6PrYC&oi=fnd&pg=PR5&dq=koonin+logic+of+chance&ots=yPPs8CKytg&sig=6fC1-ByGae6p4NkZuquA5BZ7w0g&redir_esc=y#v=onepage&q=koonin%20logic%20of%20chance&f=false

118. Scally A, Durbin R. Revising the human mutation rate: implications for understanding human evolution. Nat Rev Genet. 2012;13: 745–753. doi:10.1038/nrg3295

119. Keightley PD, Ness RW, Halligan DL, Haddrill PR. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. Genetics. 2014;196: 313–320. doi:10.1534/genetics.113.158758

120. Zheng Z-M, Quintero J, Reid ES, Gocke C, Baker CC. Optimization of a Weak 3' Splice Site Counteracts the Function of a Bovine Papillomavirus Type 1 Exonic Splicing Suppressor In Vitro and In Vivo. J Virol. 2000;74: 5902–5910. doi:10.1128/JVI.74.13.5902-5910.2000

121. Ohta T. The Nearly Neutral Theory of Molecular Evolution. Annual Review of Ecology and Systematics. 1992;23: 263–286.

122. Irimia M, Roy SW, Neafsey DE, Abril JF, Garcia-Fernandez J, Koonin EV. Complex selection on 5' splice sites in intron-rich organisms. Genome Research. 2009;19: 2021–2027. doi:10.1101/gr.089276.108

123. Losos JB. The Princeton Guide to Evolution. Princeton University Press; 2014.

124. Charlesworth B. Elements of Evolutionary Genetics. Roberts and Company Publishers; 2010.

125. Gillespie JH. Population Genetics: A Concise Guide. JHU Press; 2010.

126. Kimura M. The neutral theory of molecular evolution. Cambridge [Cambridgeshire]; New York: Cambridge University Press; 1983.

127. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive Natural Selection in the Human Lineage. Science. 2006;312: 1614–1620. doi:10.1126/science.1124309

128. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351: 652–654. doi:10.1038/351652a0

129. Vitti JJ, Grossman SR, Sabeti PC. Detecting Natural Selection in Genomic Data. Annual Review of Genetics. 2013;47: 97–120. doi:10.1146/annurev-genet-111212-133526

130. Charlesworth J, Eyre-Walker A. The other side of the nearly neutral theory, evidence of slightly advantageous back-mutations. PNAS. 2007;104: 16992–16997. doi:10.1073/pnas.0705456104

131. Kimura M, Ōta T. Theoretical Aspects of Population Genetics. Princeton University Press; 1971.

132. Li W-H. Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. Journal of Molecular Evolution. 1987;24: 337–345. doi:10.1007/BF02134132

133. Bulmer M. The selection-mutation-drift theory of synonymous codon usage. Genetics. 1991;129: 897–907.

134. Kondrashov AS. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? Journal of Theoretical Biology. 1995;175: 583–594. doi:10.1006/jtbi.1995.0167

135. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, et al. Characterization of single-nucleotide polymorphisms in coding regions of human genes. Nat Genet. 1999;22: 231–238. doi:10.1038/10290

136. Drake JA, Bird C, Nemesh J, Thomas DJ, Newton-Cheh C, Reymond A, et al. Conserved noncoding sequences are selectively constrained and not mutation cold spots. Nat Genet. 2006;38: 223–227. doi:10.1038/ng1710

137. Ohta T. Amino acid substitution at the Adh locus of Drosophila is facilitated by small population size. PNAS. 1993;90: 4548–4551. doi:10.1073/pnas.90.10.4548

138. Johnson KP, Seger J. Elevated Rates of Nonsynonymous Substitution in Island Birds. Mol Biol Evol. 2001;18: 874–881.

139. Keightley PD, Kryukov GV, Sunyaev S, Halligan DL, Gaffney DJ. Evolutionary constraints in conserved nongenic sequences of mammals. Genome Res. 2005;15: 1373–1378. doi:10.1101/gr.3942005

140. Berlin S, Ellegren H. Fast Accumulation of Nonsynonymous Mutations on the Female-Specific W Chromosome in Birds. J Mol Evol. 2005;62: 66–72. doi:10.1007/s00239-005-0067-6

141. Bartolomé C, Charlesworth B. Evolution of Amino-Acid Sequences and Codon Usage on the Drosophila miranda Neo-Sex Chromosomes. Genetics. 2006;174: 2033–2044. doi:10.1534/genetics.106.064113

142. Gerber AS, Loggins R, Kumar S, Dowling TE. Does Nonneutral Evolution Shape Observed Patterns of DNA Variation in Animal Mitochondrial Genomes? Annual Review of Genetics. 2001;35: 539–566. doi:10.1146/annurev.genet.35.102401.091106

143. Lomelin D, Jorgenson E, Risch N. Human genetic variation recognizes functional elements in noncoding sequence. Genome Res. 2010;20: 311–319. doi:10.1101/gr.094151.109

144. Bateson W. Mendel's Principles of Heredity. Cosimo, Inc.; 2007.

145. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Earth and Environmental Science Transactions of the Royal Society of Edinburgh. 1919;52: 399–433. doi:10.1017/S0080456800012163

146. Lehner B. Molecular mechanisms of epistasis within and between genes. Trends in Genetics. 2011;27: 323–331. doi:10.1016/j.tig.2011.05.007

147. Costanzo M, Baryshnikova A, Bellay J, Kim Y, Spear ED, Sevier CS, et al. The Genetic Landscape of a Cell. Science. 2010;327: 425–431. doi:10.1126/science.1180823

148. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet. 2007;8: 437–449. doi:10.1038/nrg2085

149. Phillips PC. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9: 855–867. doi:10.1038/nrg2452

150. Wright S. The roles of mutation, inbreeding, crossbreeding, and selection in evolution [Internet]. na; 1932. Available: http://www.esp.org/books/6th-congress/facsimile/contents/6th-cong-p356-wright.pdf

151. Poelwijk FJ, Kiviet DJ, Weinreich DM, Tans SJ. Empirical fitness landscapes reveal accessible evolutionary paths. Nature. 2007;445: 383–386. doi:10.1038/nature05451

152. de Visser JAGM, Krug J. Empirical fitness landscapes and the predictability of evolution. Nat Rev Genet. 2014;15: 480–490. doi:10.1038/nrg3744

153. Weinreich DM, Lan Y, Wylie CS, Heckendorn RB. Should evolutionary geneticists worry about higher-order epistasis? Curr Opin Genet Dev. 2013;23: 700–707. doi:10.1016/j.gde.2013.10.007

154. Podgornaia AI, Laub MT. Pervasive degeneracy and epistasis in a protein-protein interface. Science. 2015;347: 673–677. doi:10.1126/science.1257360

155. Puchta O, Cseke B, Czaja H, Tollervey D, Sanguinetti G, Kudla G. Network of epistatic interactions within a yeast snoRNA. Science. 2016;352: 840–844. doi:10.1126/science.aaf0965

156. Meer MV, Kondrashov AS, Artzy-Randrup Y, Kondrashov FA. Compensatory evolution in mitochondrial tRNAs navigates valleys of low fitness. Nature. 2010;464: 279–282. doi:10.1038/nature08691

157. Sarkisyan KS, Bolotin DA, Meer MV, Usmanova DR, Mishin AS, Sharonov GV, et al. Local fitness landscape of the green fluorescent protein. Nature. 2016;advance online publication. doi:10.1038/nature17995

158. Visser JAGM de, Cooper TF, Elena SF. The causes of epistasis. Proceedings of the Royal Society of London B: Biological Sciences. 2011;278: 3617–3624. doi:10.1098/rspb.2011.1537

159. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. Nature. 2006;444: 929–932. doi:10.1038/nature05385

160. Ortlund EA, Bridgham JT, Redinbo MR, Thornton JW. Crystal Structure of an Ancient Protein: Evolution by Conformational Epistasis. Science. 2007;317: 1544–1548. doi:10.1126/science.1142819

161. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Current Opinion in Structural Biology. 2009;19: 596–604. doi:10.1016/j.sbi.2009.08.003

162. Wang X, Minasov G, Shoichet BK. Evolution of an Antibiotic Resistance Enzyme Constrained by Stability and Activity Trade-offs. Journal of Molecular Biology. 2002;320: 85–95. doi:10.1016/S0022-2836(02)00400-X

163. de Visser JAGM, Elena SF. The evolution of sex: empirical insights into the roles of epistasis and drift. Nature Reviews Genetics. 2007;8: 139–149. doi:10.1038/nrg1985

164. Nurtdinov RN, Neverov AD, Mal'ko DB, Kosmodem'yanskii IA, Ermakova EO, Ramenskii VE, et al. EDAS—A database of alternatively spliced human genes. BIOPHYSICS. 2006;51: 523–526. doi:10.1134/S0006350906040026

165. Nurtdinov R, Neverov A, Favorov A, Mironov A, Gelfand M. Conserved and species-specific alternative splicing in mammalian genomes. BMC Evolutionary Biology. 2007;7: 249. doi:10.1186/1471-2148-7-249

166. Nurtdinov RN, Mironov AA, Gelfand MS. Rodent-specific alternative exons are more frequent in rapidly evolving genes and in paralogs. BMC Evol Biol. 2009;9: 142–142. doi:10.1186/1471-2148-9-142

167. Brudno M, Gelfand MS, Spengler S, Zorn M, Dubchak I, Conboy JG. Computational analysis of candidate intron regulatory elements for tissue-specific alternative pre-mRNA splicing. Nucl Acids Res. 2001;29: 2338–2348. doi:10.1093/nar/29.11.2338

168. Comeron JM, Ratnappan R, Bailin S. The Many Landscapes of Recombination in Drosophila melanogaster. PLoS Genet. 2012;8: e1002905. doi:10.1371/journal.pgen.1002905

169. Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, et al. A high-resolution recombination map of the human genome. Nature Genetics. 2002; 241–247. doi:10.1038/ng917

170. Kurmangaliyev YZ, Sutormin RA, Naumenko SA, Bazykin GA, Gelfand MS. Functional implications of splicing polymorphisms in the human genome. Hum Mol Genet. 2013;22: 3449–3459. doi:10.1093/hmg/ddt200

171. Parsch J, Novozhilov S, Saminadin-Peter SS, Wong KM, Andolfatto P. On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in Drosophila. Mol Biol Evol. 2010;27: 1226–1234. doi:10.1093/molbev/msq046

172. Consortium T 1000 GP. A map of human genome variation from population-scale sequencing. Nature. 2010;467: 1061–1073. doi:10.1038/nature09534

173. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482: 173–178. doi:10.1038/nature10811

174. Prasad AB, Allard MW, Green ED. Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Data Sets. Molecular Biology and Evolution. 2008;25: 1795–1808. doi:10.1093/molbev/msn104

175. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature. 2007;450: 203–218. doi:10.1038/nature06341

176. Yang Z. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol. 2007;24: 1586–1591. doi:10.1093/molbev/msm088

177. Efron B, Tibshirani RJ. An introduction to the bootstrap. Boca Raton, Florida: CRC Press; 1993.

178. Crooks GE, Hon G, Chandonia JM, Brenner SE. Weblogo: A sequence logo generator. Genome research. 2004;14: 1188–1190.

179. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20: 110–121. doi:10.1101/gr.097857.109

180. Garg K, Green P. Differing patterns of selection in alternative and constitutive splice sites. Genome Research. 2007;17: 1015–1022. doi:10.1101/gr.6347907

181. Kotelnikova EA, Makeev VJ, Gelfand MS. Evolution of transcription factor DNA binding sites. Gene. 2005;347: 255–263. doi:10.1016/j.gene.2004.12.013

182. Stergachis AB, Haugen E, Shafer A, Fu W, Vernot B, Reynolds A, et al. Exonic Transcription Factor Binding Directs Codon Choice and Affects Protein Evolution. Science. 2013;342: 1367–1372. doi:10.1126/science.1243490

183. Scarano E, Iaccarino M, Grippo P, Parisi E. The heterogeneity of thymine methyl group origin in DNA pyrimidine isostichs of developing sea urchin embryos. Proc Natl Acad Sci U S A. 1967;57: 1394–1400.

184. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, et al. The GENCODE v7 catalog of human long noncoding RNAs: Analysis of their gene structure, evolution, and expression. Genome Res. 2012;22: 1775–1789. doi:10.1101/gr.132159.111

185. Gillespie JH. Is the population size of a species relevant to its evolution? Evolution. 2001;55: 2161–2169.

186. Akashi H. Inferring Weak Selection from Patterns of Polymorphism and Divergence at ``silent'' Sites in Drosophila DNA. Genetics. 1995;139: 1067–1076.

187. Goldstein RA. Population size dependence of fitness effect distribution and substitution rate probed by biophysical model of protein thermostability. Genome Biol Evol. 2013;5: 1584–1593. doi:10.1093/gbe/evt110

188. Charlesworth B. Why We Are Not Dead One Hundred Times Over. Evolution. 2013;67: 3354–3361. doi:10.1111/evo.12195

189. Barry D, Hartigan JA. Statistical Analysis of Hominoid Molecular Evolution. Statist Sci. 1987;2: 191–207. doi:10.1214/ss/1177013353

190. Coolidge CJ, Seely RJ, Patton JG. Functional Analysis of the Polypyrimidine Tract in pre-mRNA Splicing. Nucleic Acids Research. 1997;25: 888–896. doi:10.1093/nar/25.4.888

191. Bouck J, Litwin S, Skalka AM, Katz RA. In vivo selection for intronic splicing signals from a randomized pool. Nucl Acids Res. 1998;26: 4516–4523. doi:10.1093/nar/26.19.4516

192. Roscigno RF, Weiner M, Garcia-Blanco MA. A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing. J Biol Chem. 1993;268: 11222–11229.

193. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500: 415–421. doi:10.1038/nature12477

194. Kornblihtt AR, Mata MDL, Fededa JP, Muñoz MJ, Nogués G. Multiple links between transcription and splicing. RNA. 2004;10: 1489–1498. doi:10.1261/rna.7100104

195. Naftelberg S, Schor IE, Ast G, Kornblihtt AR. Regulation of alternative splicing through coupling with transcription and chromatin structure. Annu Rev Biochem. 2015;84: 165–198. doi:10.1146/annurev-biochem-060614-034242

196. Agirre E, Bellora N, Alló M, Pagès A, Bertucci P, Kornblihtt AR, et al. A chromatin code for alternative splicing involving a putative association between CTCF and HP1α proteins. BMC Biol. 2015;13: 31. doi:10.1186/s12915-015-0141-5

197. Gonzalez I, Munita R, Agirre E, Dittmer TA, Gysling K, Misteli T, et al. A lncRNA regulates alternative splicing via establishment of a splicing-specific chromatin signature. Nat Struct Mol Biol. 2015;22: 370–376. doi:10.1038/nsmb.3005

198. Lev Maor G, Yearim A, Ast G. The alternative role of DNA methylation in splicing regulation. Trends Genet. 2015;31: 274–280. doi:10.1016/j.tig.2015.03.002

199. Ram O, Ast G. SR proteins: a foot on the exon before the transition from intron to exon definition. Trends in Genetics. 2007;23: 5–7. doi:10.1016/j.tig.2006.10.002

200. Nelson KK, Green MR. Mechanism for cryptic splice site activation during pre-mRNA splicing. Proceedings of the National Academy of Sciences of the United States of America. 1990;87: 6253–6257.

201. Carmel I, Tal S, Vig I, Ast G. Comparative analysis detects dependencies among the 5’ splice-site positions. RNA. 2004;10: 828–840. doi:10.1261/rna.5196404

202. Ohno K, Brengman JM, Felice KJ, Cornblath DR, Engel AG. Congenital end-plate acetylcholinesterase deficiency caused by a nonsense mutation and an A-->G splice-donor-site mutation at position +3 of the collagenlike-tail-subunit gene (COLQ): how does G at position +3 result in aberrant splicing? Am J Hum Genet. 1999;65: 635–644. doi:10.1086/302551

203. Gelfand MS. Statistical analysis of mammalian pre-mRNA splicing sites. Nucl Acids Res. 1989;17: 6369–6382. doi:10.1093/nar/17.15.6369

204. Kralovicova J, Knut M, Cross NCP, Vorechovsky I. Identification of U2AF(35)-dependent exons by RNA-Seq reveals a link between 3’ splice-site organization

and activity of U2AF-related proteins. Nucl Acids Res. 2015;43: 3747–3763. doi:10.1093/nar/gkv194

205. Rollins RA, Haghighi F, Edwards JR, Das R, Zhang MQ, Ju J, et al. Large-scale structure of genomic methylation patterns. Genome Res. 2006;16: 157–163. doi:10.1101/gr.4362006

206. Modrek B, Lee CJ. Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nat Genet. 2003;34: 177–180. doi:10.1038/ng1159

207. Lev-Maor G, Sorek R, Shomron N, Ast G. The Birth of an Alternatively Spliced Exon: 3' Splice-Site Selection in Alu Exons. Science. 2003;300: 1288–1291. doi:10.1126/science.1082588

208. Alekseyenko AV, Kim N, Lee CJ. Global analysis of exon creation versus loss and the role of alternative splicing in 17 vertebrate genomes. RNA. 2007;13: 661–670. doi:10.1261/rna.325107

209. Merkin JJ, Chen P, Alexis MS, Hautaniemi SK, Burge CB. Origins and Impacts of New Mammalian Exons. Cell Reports. 2015;10: 1992–2005. doi:10.1016/j.celrep.2015.02.058

210. Chinwalla AT, Cook LL, Delehaunty KD, Fewell GA, Fulton LA, Fulton RS, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002;420: 520–562. doi:10.1038/nature01262

211. Minovitsky S, Gee SL, Schokrpur S, Dubchak I, Conboy JG. The splicing regulatory element, UGCAUG, is phylogenetically and spatially conserved in introns that flank tissue-specific alternative exons. Nucl Acids Res. 2005;33: 714–724. doi:10.1093/nar/gki210

212. Jin Y. A vertebrate RNA-binding protein Fox-1 regulates tissue-specific splicing via the pentanucleotide GCAUG. The EMBO Journal. 2003;22: 905–912. doi:10.1093/emboj/cdg089

213. Ponthier JL, Schluepen C, Chen W, Lersch RA, Gee SL, Hou VC, et al. Fox-2 Splicing Factor Binds to a Conserved Intron Motif to Promote Inclusion of Protein 4.1R Alternative Exon 16. J Biol Chem. 2006;281: 12468–12474. doi:10.1074/jbc.M511556200

214. Zhang C, Zhang Z, Castle J, Sun S, Johnson J, Krainer AR, et al. Defining the regulatory network of the tissue-specific splicing factors Fox-1 and Fox-2. Genes Dev. 2008;22: 2550–2563. doi:10.1101/gad.1703108

215. Huang S-C, Ou AC, Park J, Yu F, Yu B, Lee A, et al. RBFOX2 Promotes Protein 4.1R Exon 16 Selection via U1 snRNP Recruitment. Mol Cell Biol. 2012;32: 513–526. doi:10.1128/MCB.06423-11

216. Zhou H-L, Baraniak AP, Lou H. Role for Fox-1/Fox-2 in Mediating the Neuronal Pathway of Calcitonin/Calcitonin Gene-Related Peptide Alternative RNA Processing. Mol Cell Biol. 2007;27: 830–841. doi:10.1128/MCB.01015-06

217. Fukumura K, Kato A, Jin Y, Ideue T, Hirose T, Kataoka N, et al. Tissue-specific splicing regulator Fox-1 induces exon skipping by interfering E complex formation on the downstream intron of human F1γ gene. Nucl Acids Res. 2007;35: 5303–5311. doi:10.1093/nar/gkm569

218. Baraniak AP, Chen JR, Garcia-Blanco MA. Fox-2 Mediates Epithelial Cell-Specific Fibroblast Growth Factor Receptor 2 Exon Choice. Mol Cell Biol. 2006;26: 1209–1222. doi:10.1128/MCB.26.4.1209-1222.2006