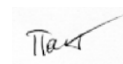


На правах рукописи



ПАНЧИН АЛЕКСАНДР ЮРЬЕВИЧ

ИССЛЕДОВАНИЕ ОБЩИХ ЗАКОНОМЕРНОСТЕЙ ЭВОЛЮЦИИ ГЕНОМА ЧЕЛОВЕКА
ПРИ ДУПЛИКАЦИИ ГЕНОВ И ТОЧЕЧНОМ МУТАГЕНЕЗЕ

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание степени

кандидата биологических наук

Москва 2011

Работа выполнена на Факультете биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова.

Научный руководитель:
кандидат биологических наук
Артамонова Ирена Игоревна

Официальные оппоненты:
кандидат физико-математических наук, доктор биологических наук, профессор
Миронов Андрей Александрович
Факультет биоинженерии и биоинформатики Московского государственного университета им. М.В. Ломоносова

доктор биологических наук
Спиридонов Сергей Эдуардович
Институт проблем экологии и эволюции им. А.Н. Северцова РАН

Ведущая организация:
ФГУП Государственный НИИ генетики и селекции промышленных микроорганизмов

Защита состоится 19 декабря 2011 года в 16.00 часов на заседании диссертационного совета Д.002.077.04 при Учреждении Российской академии наук Институте проблем передачи информации им. А.А. Харкевича РАН по адресу: 127994, г. Москва, ГСП-4, Большой Каретный переулок, д. 19, стр.1.

С диссертацией можно ознакомиться в библиотеке Учреждения Российской академии наук Института проблем передачи информации им. А.А. Харкевича РАН

Автореферат разослан __ ноября 2011 года.

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



ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность проблемы.

Активное развитие технологий секвенирования генетических последовательностей сделало возможными широкомасштабные проекты по чтению геномов. В 1995-ом году был опубликован первый полный геном – геном гемофильной палочки *Haemophilus influenzae*, а уже в 2001 году – геном человека. На сегодняшний день прочитаны геномы представителей практически всех известных крупных таксономических групп эукариот. За последние десять лет разработаны такие методы, как SOLiD, метод пиросеквенирования (технология 454), SOLEXA и Ion Torrent. Интенсивно разрабатываются методы секвенирования последовательностей ДНК, позволяющие анализировать одиночные молекулы, например, технологии Helicos и SMART. Стали доступны не только последовательности генов и полных геномов огромного количества живых организмов, но и, в ряде случаев, данные по полиморфизмам в этих последовательностях, данные по транскрипции (например, EST – expressed sequence tags), сплайсингу, наличию белковых продуктов, функциональной аннотации генов, а также профили метилирования ДНК в различных тканях и типах клеток.

Геном человека является удобной моделью для анализа закономерностей эволюции и мутагенеза благодаря высокому качеству сборки и постоянному притоку большого количества сопряженных данных. Наряду с полностью прочитанным геномом человека, также прочитаны геномы ископаемых представителей рода *Homo*: неандертальца *Homo neanderthalensis* и Денисовского человека *Denisova hominin*, а также геномы родственных приматов, таких как шимпанзе *Pan troglodytes*, орангутанов *Pongo abelii*, *Pongo pygmaeus* и макаки *Macaca mulata*. Большой интерес представляют как закономерности точечного мутагенеза (например, зависимость мутационных процессов от нуклеотидного контекста), так и закономерности эволюционных событий большего масштаба, например, дупликаций генов и последующего изменения их моделей сплайсинга.

Цели и задачи исследования:

Целью данного исследования является выявление закономерностей эволюции генетических последовательностей человека. Особенное внимание здесь уделено

закономерностям точечного мутагенеза и эволюции генов после дупликаций. При этом были поставлены следующие задачи:

- Создание коллекции “мусорных” фрагментов транскриптов (EST) человека. Анализ источников загрязнений.
- Анализ встречаемости 1-7 нуклеотидных слов в 139 полных геномах эукариот.
- Установление предковых состояний одиночных нуклеотидных полиморфизмов в геноме человека и поиск ранее не описанных мутационных контекстов, в которых наблюдается избыток или недостаток мутаций.
- Анализ эволюции генов человека, возникших в результате недавних дупликаций.
- Анализ закономерностей эволюции сайтов сплайсинга в генах человека после дупликаций.
- Исследование взаимосвязи между степенью метилирования цитозинов и частотой CpG>TpG мутаций в CpG островах и в остальных областях генома человека.

Научная новизна и практическая значимость.

В ходе данной работы мы использовали наиболее современные (на момент исследования) базы данных и провели анализ, который раньше был невозможен в силу недоступности или недостаточности экспериментальных данных. Так, в сравнительный анализ частот нуклеотидных слов в геномах эукариот нами было включено 139 полностью прочитанных геномов (включая геном человека), покрывающих значительную часть филогенетического древа эукариот. Анализ подобного масштаба позволил выявить ряд интересных закономерностей, которые могут помочь дальнейшему исследованию механизмов мутагенеза.

Для анализа частот мутаций в геноме человека мы использовали выборку из нескольких миллионов полиморфизмов в геноме человека, что позволило оценить частоты мутаций в мутационных контекстах длиной до 4-х букв, исключив при этом возможный статистический шум. Эти мутационные данные, наряду с появившимися совсем недавно данными по полногеномному анализу метилирования в эмбриональных стволовых клетках человека, были также использованы для исследования причин пониженной частоты CpG>TpG мутаций в CpG островах.

Новизна исследований, посвященных эволюции недавно дублированных генов человека, прежде всего, заключается в предложенном методологическом подходе – мы использовали семейства паралогичных генов для реконструкции предковых состояний в узлах филогенетических деревьев. Нам удалось подтвердить, что после дубликации две копии гена в значительной части случаев накапливают замены с разной скоростью и показать, что в быстро эволюционирующих копиях недавно дублированных генов человека среди замен, меняющих кодируемый белок, перепредставлены замены, влияющие на функцию белкового продукта. Нами также было показано, что после дубликаций генов появление сайтов сплайсинга преобладает над исчезновением, а также, что конститутивные сайты сплайсинга реже становятся альтернативными, чем наоборот.

Результаты исследований могут быть полезны в области медицинской генетики.

Публикации и апробация работы.

По материалам диссертации опубликовано 4 статьи в международных рецензируемых научных журналах, индексируемых в PubMed (две в журнале Journal of Bioinformatics and Computational Biology, одна в BMC Bioinformatics и одна в Biology Direct). Материалы работы были представлены на конференциях: Московской конференции по вычислительной молекулярной биологии (МССМВ 2007; МССМВ 2009; МССМВ 2011, Москва), конференции “Информационные технологии и системы” (ИТиС 2008; ИТиС 10; ИТиС 11, Геленджик; ИТиС 2009, Москва), “Ломоносов-2008” (Москва) и на Русско-Швейцарском семинаре “Сравнительная геномика и разнообразие транскриптома” (Женева, 2010).

Структура и объем работы.

Диссертация изложена на 192 страницах машинописного текста и содержит следующие разделы: введение, обзор литературы, а также шесть глав, каждая из которых содержит свое введение, методы, результаты, обсуждение (в главах 1, 2 и 5 результаты и обсуждение объединены вместе) и выводы. В конце приведены выводы, а затем список литературы. Материал иллюстрирован 35 рисунками. В работе 15 таблиц. Библиографический указатель включает 238 ссылок.

СОДЕРЖАНИЕ РАБОТЫ

1. Анализ “загрязнений” транскрипционных данных человека

Одной из наиболее полных баз транскрипционных данных, используемых в биоинформатических исследованиях (в том числе – в данной работе), является база данных dbEST. EST (expressed sequence tags) – это последовательности длиной порядка 500 нуклеотидов, полученные путем секвенирования клонов кДНК из кДНК-библиотек. Эти библиотеки создаются с помощью процесса обратной транскрипции молекул РНК, выделенных из образцов (тканей, клеточных культур), и представляют собой фрагменты существующих в образце РНК. В связи с тем, что все клеточные РНК образуются путем транскрипции на матрице геномной ДНК (для человека, геномы ядра или митохондрий), можно было бы ожидать, что все EST человека будут соответствовать геномным последовательностям человека. Однако последовательности многих EST не имеют сходства ни с какими геномными последовательностями.

Нами создана и проанализирована коллекция “мусорных” EST человека: последовательностей, которые, вопреки обозначенному источнику происхождения, не имеют сходства с какими-либо последовательностями в геноме человека. На рисунке 1 представлена схема построения коллекции “мусорных” EST.

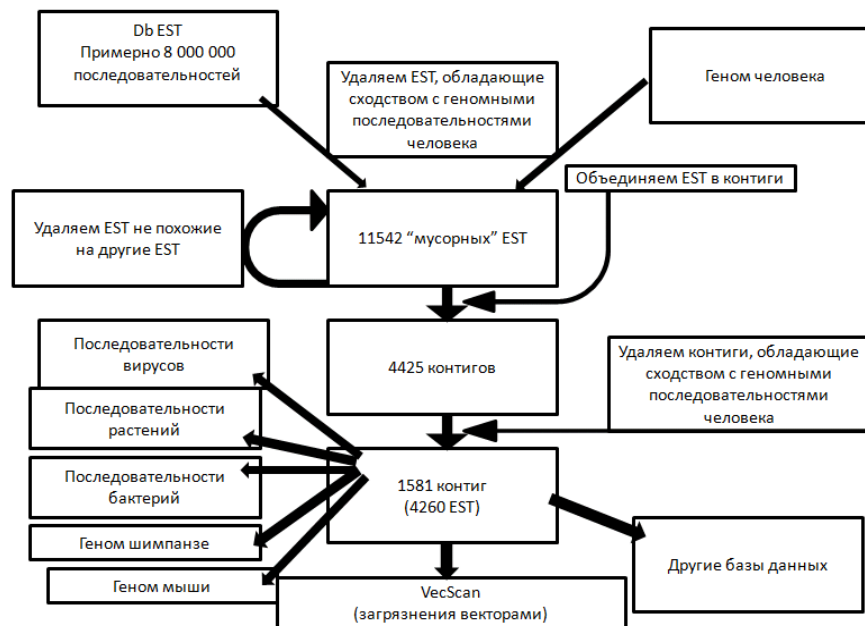


Рисунок 1. Создание коллекции “мусорных” EST. Последовательности базы данных EST человека (dbEST) сравнивались с геномом человека. EST, которые не имели сходства с геномом, отмечались как “мусорные”. Те мусорные EST, которые не имели сходства с другими мусорными EST, удалялись из выборки. Оставшиеся собирались в контиги с помощью программы SeqMap. Контиги вновь сравнивались с геномом человека и сохранялись в коллекции только, если сходства найдено не было.

Всего на момент исследования было обнаружено 11542 “мусорных” EST, которые были объединены в 4425 контигов. Эти контиги находятся в открытом доступе. Классификация EST по потенциальному источнику происхождения представлена на рисунке 2.

Рисунок 2. Классификация “мусорных” EST человека по источнику происхождения. Большинство “мусорных” EST не имели сходства ни с какими последовательностями из доступных на момент исследования разделов базы данных GeneBank. Распространены загрязнения последовательностями бактерий, других млекопитающих, вирусов, а также растений и грибов.



Наш анализ показал, что в базе данных EST человека есть последовательности, которые имеют наибольшее сходство с последовательностями растений (*Viridiplantae*). В некоторых случаях найденные последовательности практически идентичны растительным генам, участвующим в фотосинтезе. Поскольку по аннотации не ясно, каким образом, эти последовательности попали в базу данных EST человека, мы исследовали их подробнее.

Количество обнаруженных EST, имеющих высокий уровень сходства с растительными последовательностями, но не имеющих сходства с геномом человека велико (по крайней мере, 36 EST). Большая часть этих EST попала в три контига. Два контига соответствуют хлорофиллу а/б-связывающему белку (15 EST), а третий соответствует гену малой субъединицы рибулозобисфосфаткарбоксилазы/оксигеназы (RuBisCo, 6 EST).

Наиболее простое объяснение этого наблюдения заключается в том, что на определенном этапе секвенирования произошло загрязнение ряда образцов или была допущена ошибка при аннотации последовательностей. Тем не менее, есть ряд аргументов против такого объяснения:

1. Подобные растительные загрязнения были обнаружены в нескольких лабораториях независимо и в независимых библиотеках кДНК.
2. Подобные растительные загрязнения содержатся и в других базах данных EST млекопитающих, например, *Mus musculus*, *Bos taurus*.
3. Последовательности RuBisCo из базы данных EST *Mus musculus* были проанализированы с помощью ДНК-микрочипов. Анализ базы данных GEO показал наличие сильного сигнала, интенсивность которого варьировала между различными мышинными тканями.

Хотя мы понимаем, что, скорее всего растительные EST попали в базу данных EST человека как случайные загрязнения, можно выдвинуть альтернативную гипотезу. РНК растительного происхождения могли попасть в ткани животных из растительной пищи. Передача двухцепочечной РНК из пищи в ткани хорошо изучена на примере круглого червяка *C. elegans*. Механизм транспорта двухцепочечной РНК описан не полностью, но известно, что в нем принимает участие канал под названием Sid1. Гомологичный этому каналу белок кодируется и в геноме человека, а так же в геномах ряда других животных, причем гиперэкспрессия Sid1 ведет к увеличению чувствительности клеток человека к малым интерферирующим РНК.

Один из важных выводов данного анализа заключается в том, что в базах транскрипционных данных любого биологического вида может присутствовать значительное количество загрязнений (преимущественно, нуклеотидных последовательностей других видов). Мы делаем особый акцент на этом предположении в ходе дальнейшей работы. Например, при определении сайтов сплайсинга мы следим, чтобы сайты подтверждались EST из нескольких библиотек.

2. Сравнительных анализ частот нуклеотидных слов в геноме человека, а также в геномах других эукариот

Мы проанализировали частоты 1-7 буквенных нуклеотидных слов в последовательностях в трех выборках: 139 полных геномов эукариот, включая геном человека; 33 генома, в которых известные повторы заменены на N; кодирующие области 35 геномов.

В этой задаче нас интересовали систематически недопредставленные или перепредставленные (по сравнению с предсказаниями статистических моделей) нуклеотидные слова. Известно, что в ряде случаев отклонение частот слов от ожидаемых частот ассоциировано с существованием контекст-зависимых мутационных механизмов, поэтому такой анализ может помочь поиску механизмов мутагенеза. В качестве меры недопредставленности или перепредставленности нуклеотидных слов мы используем величину «частотность».

Частотностью слова $W = (w_1 \dots w_n)$ (где каждое w_i – это символ из алфавита А, Т, G, С) назовем величину $C = ((\text{Obs}(W) - \text{Exp}(W))/\text{Exp}(W)) * 100\%$, где $\text{Obs}(W)$ – число слов W в геноме, а $\text{Exp}(W)$ – ожидаемое число слов W в соответствии с моделью. Иными словами, «частотность» отражает степень недопредставленности или перепредставленности слова в процентах.

Мы использовали следующие критерии:

- $C(W) < -10\%$ означает, что слово W существенно недопредставлено в геноме;
- $C(W) > 10\%$ означает, что слово W существенно перепредставлено в геноме.

В Таблице 1 приведены примеры наиболее интересных случаев недопредставленности или перепредставленности нуклеотидных слов и их распределение по различным таксономическим группам. Значения частотности всех 1-7 буквенных слов для всех 139 геномов опубликованы и находятся в открытом доступе.

Самое недопредставленное слово в проанализированных геномах – ТА. Оно недопредставлено в 130 из 139 изученных геномов и во всех 33 геномах с отфильтрованными повторами. В тех геномах, где ТА недопредставлено, его частотность варьирует от -56% (*Sporobolomyces roseus*) до -12% (*Theileria parva*). Среднее значение частотности слова ТА по всем геномам составляет -25%. Биологические виды, в которых ТА недопредставлено, входят в состав всех таксономических групп, включенных в исследование, кроме типа пластинчатые (*Placozoa*). Несмотря на то, что недопредставленность ТА наблюдается повсеместно, нам удалось обнаружить ряд важных исключений. В организме *Mycosphaerella fijiensis* (Fungi) ТА перепредставлено

(частотность +40%). В организмах *Myceliophthora thermophila*, *Sporotrichum thermophilum* и *Thielavia terrestris* (Fungi), *Acyrtosiphon pisum* (Arthropoda), *Plasmodium falciparum* (Alveolata) и *Trichoplax adhaerens* (Placozoa) частотность слов ТА колеблется в пределах от -10% до +10%. Изучение биохимических процессов, происходящих в этих организмах, может помочь понять, чем обусловлена недопредставленность слова ТА в большей части эукариотических геномов, в том числе в геноме человека.

Если ТА – наиболее часто недопредставленное слово, то CG недопредставлено наиболее сильно (хоть и в меньшем числе геномов). Например, средний уровень недопредставленности CG в 32 видах группы челюстноротых *Gnathostomata* (к которой относится человек) составляет -77%, в то время как средняя недопредставленность ТА в той же группе составляет лишь -29%. В таксономической группе членистоногих (Arthropoda) частотность варьирует в зависимости от конкретного организма и значения контраста меняются в диапазоне от -14% (*Drosophila pseudoobscura*) до +66% (*Apis mellifera*).

Частоты двухбуквенных слов в кодирующих частях геномов в большинстве случаев ведут себя сходным образом с тем, как они ведут себя в полных геномах. Например, слово ТА недопредставлено во всех изученных кодирующих частях геномов, а слово CG недопредставлено во всех изученных кодирующих частях геномов, за исключением двух организмов (*Drosophila melanogaster* и *Caenorhabditis elegans*). Однако мы обнаружили ряд расхождений между частотами слов в полных геномах и соответствующих им кодирующих областях. Прежде всего, существуют слова, значения частотности которых в кодирующих участках существенно отличаются от значений частотности комплементарных слов (эффект, которого мы не наблюдаем для полных геномов). Например, слово CC перепредставлено в кодирующих участках 25 анализированных видов из 35, в то время как GG перепредставлено только у одного вида. Во-вторых, есть слова, которые систематически более редки или менее редки в кодирующих участках по сравнению с полным геномом. Например, триплет TAG, который может служить стоп-кодоном, что не удивительно, реже встречается в кодирующих участках. С другой стороны триплет TGG, который может служить кодоном

триптофана, чаще встречается в кодирующих последовательностях, что не согласуется с редкой встречаемостью триптофана в аминокислотных последовательностях белков.

	CCC (GGG)	CTA (TAG)	AA (TT)	AC (GT)	CA (TG)	CC (GG)	CGCG	CG	TATA	TA
<i>Fungi</i> (38)	3:35:0	33:5:0	0:28:10	17:21:0	3:17:18	8:29:1	6:29:1	27:9:2	1:25:12	35:2:1
<i>Gnathostomata</i> (32)	8:24:0	0:32:0	0:6:26	27:5:0	0:0:32	0:6:26	1:5:26	32:0:0	1:31:0	32:0:0
<i>Arthropoda</i> (18)	4:14:0	12:6:0	0:1:17	13:5:0	1:5:12	2:14:2	0:15:3	3:11:4	0:18:0	17:1:0
<i>Nematoda</i> (7)	3:4:0	0:7:0	0:0:7	7:0:0	0:6:1	0:7:0	0:4:3	2:4:1	0:6:1	7:0:0
<i>Streptophyta</i> (7)	7:0:0	1:6:0	0:2:5	6:1:0	0:1:6	0:4:3	0:3:4	7:0:0	0:6:1	7:0:0
<i>Chlorophyta</i> (7)	0:7:0	7:0:0	0:3:4	1:6:0	4:0:3	4:3:0	0:7:0	3:0:4	2:5:0	4:1:0
<i>Alveolata</i> (5)	4:1:0	3:2:0	0:3:2	4:1:0	0:3:2	0:0:5	1:2:2	5:0:0	1:4:0	1:1:0
<i>Mollusca</i> (2)	0:2:0	0:2:0	0:1:1	0:2:0	0:0:2	0:2:0	0:0:2	2:0:0	0:2:0	2:0:0
<i>Bacillariophyta</i> (2)	0:2:0	2:0:0	0:1:1	0:2:0	0:1:1	1:1:0	0:2:0	1:1:0	1:1:0	2:0:0
<i>Cnidaria</i> (2)	0:2:0	1:1:0	0:0:2	0:2:0	0:1:1	0:0:2	0:0:2	2:0:0	0:2:0	2:0:0
<i>Dictyosteliida</i> (2)	2:0:0	2:0:0	0:0:2	2:0:0	0:0:2	0:0:2	1:1:0	2:0:0	0:2:0	2:0:0
<i>Kinetoplastida</i> (2)	0:2:0	2:0:0	0:1:1	0:2:0	0:0:2	1:1:0	0:2:0	0:2:0	0:1:1	2:0:0
<i>Peronosporales</i> (2)	0:2:0	2:0:0	0:2:0	0:2:0	0:0:2	2:0:0	0:2:0	0:2:0	0:2:0	2:0:0
<i>Rhodophyta</i> (1)	0:1:0	1:0:0	0:1:0	0:1:0	0:0:1	1:0:0	0:1:0	0:0:1	0:1:0	1:0:0
<i>Entamoeba</i> (1)	1:0:0	1:0:0	0:0:1	1:0:0	0:0:1	0:1:0	1:0:0	1:0:0	1:0:0	1:0:0
<i>Petromyzontidae</i> (1)	0:1:0	1:0:0	0:0:1	0:1:0	0:0:1	0:1:0	0:1:0	1:0:0	0:1:0	1:0:0
<i>Cephalochordata</i> (1)	0:1:0	0:1:0	0:1:0	0:1:0	0:0:1	0:1:0	0:0:1	1:0:0	0:1:0	1:0:0
<i>Tunicata</i> (1)	0:1:0	1:0:0	0:0:1	0:1:0	0:0:1	0:1:0	0:1:0	1:0:0	0:1:0	1:0:0
<i>Echinodermata</i> (1)	0:1:0	0:1:0	0:1:0	0:1:0	0:0:1	0:0:1	0:0:1	1:0:0	0:1:0	1:0:0
<i>Annelida</i> (1)	0:1:0	1:0:0	0:1:0	1:0:0	0:0:1	0:1:0	0:1:0	1:0:0	0:1:0	1:0:0
<i>Placozoa</i> (1)	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0	0:1:0
<i>Heterolobosea</i> (2)	1:1:0	1:1:0	0:1:1	2:0:0	0:0:2	0:2:0	0:1:1	2:0:0	1:1:0	2:0:0
<i>Isochrysidales</i> (1)	1:0:0	1:0:0	0:1:0	0:1:0	0:1:0	1:0:0	0:1:0	0:1:0	0:0:1	1:0:0
<i>Pelagophyceae</i> (1)	1:0:0	1:0:0	0:0:1	0:1:0	1:0:0	1:0:0	0:1:0	0:0:1	0:0:1	1:0:0
<i>Choanoflagellida</i> (1)	0:1:0	1:0:0	0:0:1	0:1:0	0:0:1	1:0:0	0:1:0	1:0:0	0:1:0	1:0:0

Таблица 1. Примеры систематически недопредставленных или перепредставленных слов. В названиях столбцов приведены изученные таксономические группы (в скобках указано количество изученных полных геномов в данной группе). В ячейках через двоеточие указано количество геномов, в которых данное слово недопредставлено, не выделяется или перепредставлено, соответственно. Светлые и темные ячейки соответствуют перепредставленности и недопредставленности слова в большей части геномов в данной таксономической группе, соответственно.

3. Исследование мутационных контекстов в геноме человека

Продолжением исследования описанного в предыдущей главе стал анализ зависимости частот мутаций от нуклеотидного контекста в геноме человека. Нам удалось восстановить предковые состояния для более чем трех миллионов би-аллельных (имеющих два установленных аллеля) полиморфизмов, расположенных в межгенных

участках генома человека путем сравнения геномов человека, шимпанзе и орангутана и составить выборку направленных мутаций. Эти данные находятся в свободном доступе.

Мы использовали величину, называемую “контраст”, чтобы оценить влияет ли добавление конкретного нуклеотида к 5’ или 3’ концам одно-, двух- или трехбуквенных слов на вероятность определенного типа мутации в конкретной позиции.

По сравнению со средней частотой мутаций из цитозина в тимин (C>T) в геноме человека, существует избыток в 5.1 раз мутаций из цитозина в тимин, если за цитозином следует гуанин. Мы описываем такой мутационный контекст как {C>T|1, CG} и указываем его контраст, который в данном случае равен 5.1 при сравнении с его подконтекстом {C>T|1, C}. При значениях контраста больших единицы можно говорить об избытке мутаций, в то время как при значениях контраста меньше единицы – о недостатке. Значения контрастов для пары мутационного контекста {mut|pos, W} и подконтекста {mut|pos', W'} вычисляются на основании частот слов P_W и $P_{W'}$ и числа мутаций, наблюдаемых в контекстах $N_{\{mut|pos, W\}}$ и $N_{\{mut|pos', W'\}}$. Здесь pos означает позицию в слове, в котором происходит мутация, а mut – тип мутации, например, C>T или A>G.

$$\text{Контраст}_{(\{mut|pos, W\}, \{mut|pos', W'\})} = \frac{N_{\{mut|pos, W\}}/P_W}{N_{\{mut|pos', W'\}}/P_{W'}}$$

Мы предлагаем оценивать отклонения частот мутаций двумя значениями контраста: “смещение частоты мутаций” и “минимальный контраст”. Значение минимального контраста – это наиболее близкое к 1 значение контраста, достигаемое при сравнении контекста со всеми возможными его подконтекстами. Например, контекст {C>T|2, ACG} имеет три подконтекста: {C>T|2, AC}, {C>T|1, CG} и {C>T|1, C} со значениями контраста 5.08, 1.08 и 5.48, соответственно. 1.08 - это значение минимального контраста для контекста {C>T|2, ACG}. Значения контраста, полученные по отношению к однобуквенному подконтексту, например, к {C>T|1, C} называются смещением частоты мутаций. Значение 5.48 это смещение частоты мутаций для контекста {C>T|2, ACG} потому, что есть избыток в 5.48 раз C>T мутаций во второй позиции слова ACG, по сравнению со средней частотой C>T мутаций в геноме человека.

На рисунке 3 изображено распределение значений минимального контраста и смещений частот мутаций. Данный анализ не позволяет нам установить, на какой из двух цепей ДНК произошла мутация, поэтому, каждый мутационный контекст имеет комплементарный контекст со схожими свойствами: точки расположены на диаграмме парами. Кроме большого кластера, включающего большинство мутационных контекстов можно увидеть еще три отчетливых кластера. Как и ожидалось, кластер, характеризующийся наибольшими значениями смещения частоты мутаций и минимального контраста, представлен контекстом $\{C>T|1, CG\}$ и комплементарным контекстом $\{G>A|2, CG\}$. Второй кластер характеризуется малыми значениями минимального контраста, но большими значениями смещения частоты мутаций – он состоит из всех таких и только таких контекстов для которых $\{C>T|1, CG\}$ или $\{G>A|2, CG\}$ являются подконтекстами. Наконец, есть еще один кластер, который выделяется как за счет больших значений смещения частот мутаций, так и больших значений минимального контраста. Он содержит три пары контекстов: $\{T>C|2, ATAG\}$ и комплементарный ему $\{A>G|3, CAAT\}$; $\{T>C|2, ATAG\}$ и $\{A>G|3, CTAAT\}$; и $\{A>C|1, ACAA\}$ и $\{T>G|4, TTGT\}$. Кроме этого можно заметить, что частоты мутаций в геноме варьируют в широком диапазоне в зависимости от контекста и могут отличаться более чем в 20 раз.

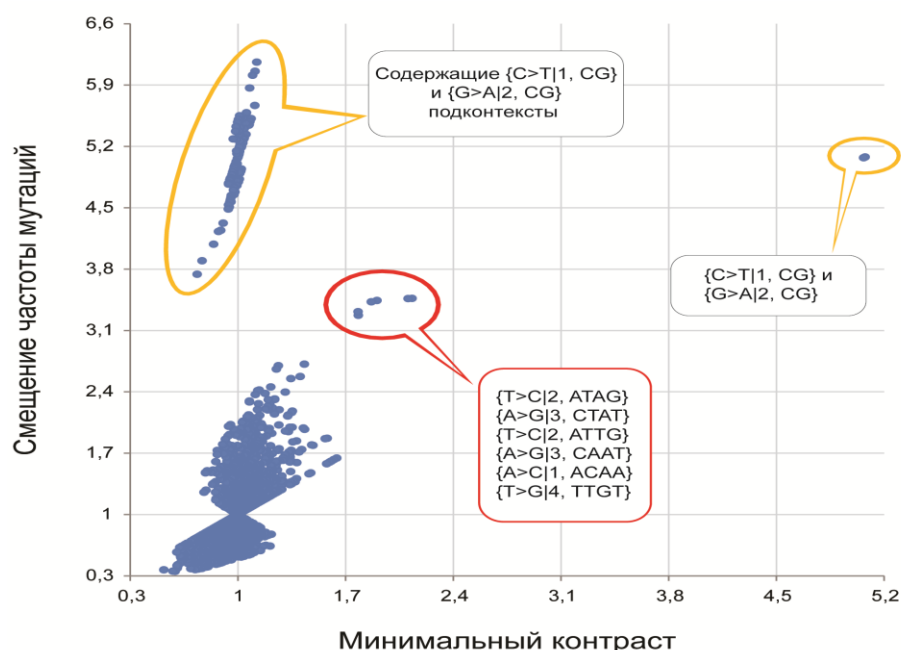


Рисунок 3. Распределение смещений частот мутаций и минимального контраста. Каждая точка представляет один мутационный контекст. Оранжевым выделены области, в которые попали уже известные мутационные контексты с высокой вероятностью мутировать. Красным выделена область, в которую попали обнаруженные нами мутационные контексты.

Мы обнаружили, что существует избыток в 3.5 и 3.3 раза T>C мутаций во второй позиции слов ATTG и ATAG, соответственно. Можно предположить, что существует мотив AT[A/T]G (ATWG) в геноме человека, в котором часто происходят мутации. Кроме того установлено, что существует избыток в 3.4 раза A>C мутаций в первой позиции слова АСAA. Наличие избыточных мутаций в описанных выше контекстах статистически значимо при сравнении с любыми из возможных подконтекстов (с учетом поправки на множественные сравнения, $P < 10^{-15}$).

4. Исследование эволюции недавно дублицированных генов человека

Мы составили выборку из 97 семейств паралогичных генов человека, содержащую 511 генов. Семейства были выбраны таким образом, чтобы в каждом насчитывалось не менее трех генов. С использованием этой выборки мы проанализировали, как меняется сила действия естественного отбора с момента дубликации генов, а также закономерности распределения замен между копиями генов.

Для оценки силы и направления естественного отбора, действующего на ген, можно использовать отношение dN/dS . dN - число несинонимичных (меняющих аминокислотную последовательность кодируемого белка) замен на несинонимичный сайт, dS - число синонимичных замен (не меняющих аминокислотную последовательность белка) на синонимичный сайт. $dN/dS < 1$ характеризует отрицательный отбор, $dN/dS > 1$ - положительный отбор, $dN/dS \sim 1$ свидетельствует о нейтральной эволюции.

Чтобы вычислить индивидуальные, специфичные для гена, значения dN и dS , мы выбрали ближайшего паралога (т.е. паралога с максимальным уровнем сходства) для каждого гена внутри семейства. Мы называли такой ген ближайшим соседом по отношению к исходному гену. Для выбранной пары генов мы находили следующий по сходству ген из того же семейства, расположенный на филогенетической ветви внешней по отношению к выбранной паре генов. Такой ген мы называем вторым ближайшим соседом.

Для каждого гена, имеющего второго ближайшего соседа, мы вычисляли индивидуальные значения dN (dN_i) и dS (dS_i), используя следующие формулы:

$$dN_i = (dN_{[i-k]} + dN_{[i-j]} - dN_{[j-k]})/2;$$

$$dS_i = (dS_{[i-k]} + dS_{[i-j]} - dS_{[j-k]})/2,$$

где индексы $[i-j]$, $[i-k]$, $[j-k]$ означают попарные значения dN или dS для гена i , его ближайшего соседа j и его второго ближайшего соседа k . Индивидуальные значения dN_i и dS_i , таким образом, отражают количество несинонимичных замен на несинонимичный сайт и синонимичных замен на синонимичный сайт, произошедших с момента последней дупликации данного гена.

Описанная процедура позволила нам сравнить скорость эволюции отдельных генов. Мы анализировали только гены, которые возникли в результате недавних дупликаций, то есть на генах с небольшими значениями индивидуальных dS (от 0.005 до 0.6).

Зависимость между индивидуальными значениями dN/dS (силой отбора) и dS (параметра, характеризующего время, прошедшее с момента последней дупликации, при условии нейтральной эволюции) показана на рисунке 4. Зависимость между индивидуальным dN и dS показана на рисунке 5. Мы наблюдаем уменьшение роста dN (уменьшение dN/dS) с ростом dS . Коэффициент линейной регрессии для зависимости dN/dS от dS составляет -1.37, а 95% доверительный интервал – (-2.43; -0.30). Это утверждение верно как для подвыборки одноэкзонных генов, так и для многоэкзонных генов. Поскольку dN/dS является мерой действия естественного отбора, а dS – оценкой времени с момента последней дупликации, мы можем утверждать, что паралоги, разошедшиеся раньше, эволюционируют под действием более сильного отрицательного отбора, в то время как более молодые паралоги эволюционируют под ослабленным отрицательным отбором, а в ряде случаев – под положительным.

Всего нашлось 144 пары генов, в которых каждый ген имел ближайшего соседа и второго ближайшего соседа. Мы посчитали количество аминокислотных замен в кодируемых этими генами белках, используя второго ближайшего соседа для установления аминокислоты предкового белка методом максимальной экономии (т.е.

минимизируя число замен между тремя белками). 26 пар генов из 144 (18%) были статистически значимо асимметричны по количеству аминокислотных замен в кодируемых белках ($P < 0.05$). Вероятность случайно наблюдать 26 и более успехов из 144 при испытаниях Бернулли с $p = 0.05$ равна 1.33×10^{-8} . Вычисленные значения позволяют нам говорить о том, что такое количество асимметричных пар генов не объясняется случайными причинами и описанное явление отражает тенденцию в эволюции паралогичных генов человека: примерно одна пятая часть от всех пар дублированных генов эволюционирует асимметрично.

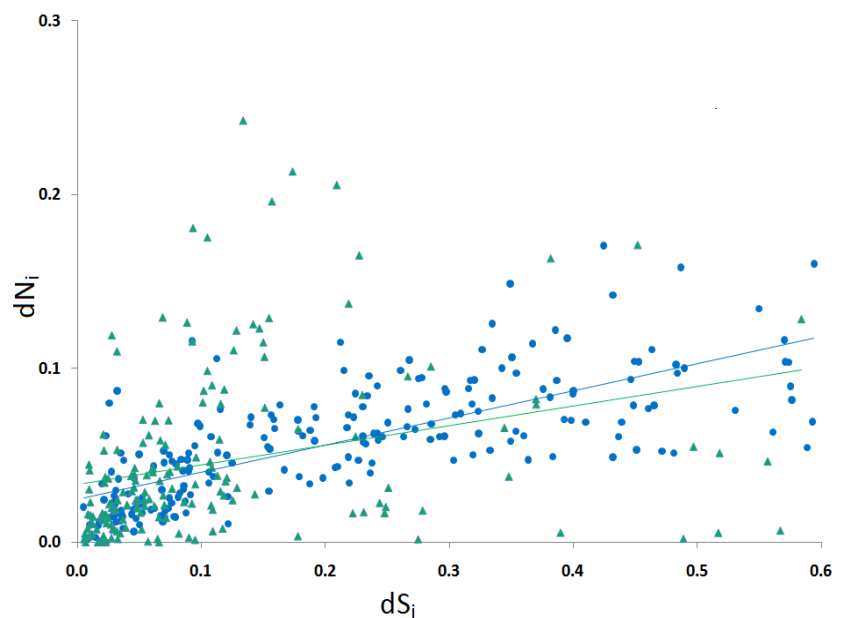
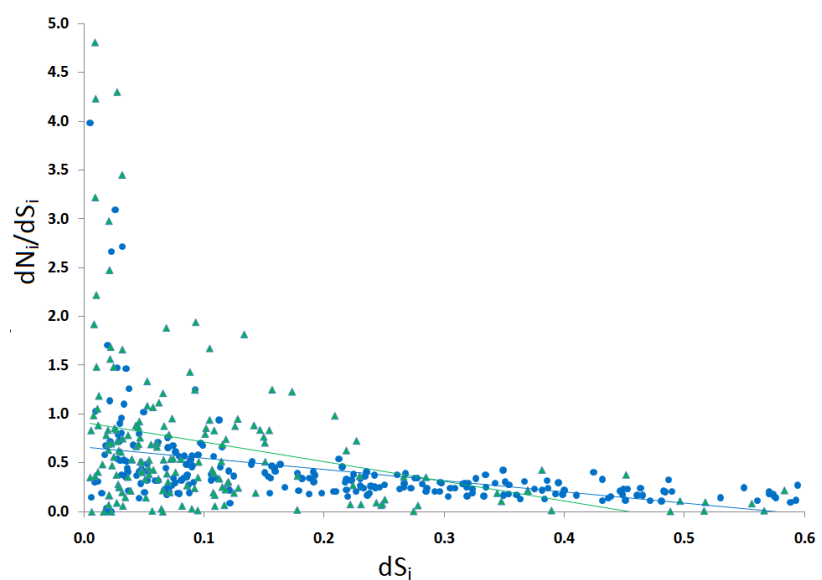


Рисунок 4. Зависимость между индивидуальными значениями dN/dS и dS генов. Одноэкзонные гены отмечены голубыми точками. Многоэкзонные гены отмечены зелеными треугольниками. Линии представляют соответствующие линейные регрессии. Отображены только гены с индивидуальными значениями $0.005 \leq dS \leq 0.6$

Рисунок 5. Зависимость между индивидуальными значениями dN и dS генов. Одноэкзонные гены отмечены голубыми точками. Многоэкзонные гены отмечены зелеными треугольниками. Линии представляют соответствующие линейные регрессии. Отображены только гены с индивидуальными значениями $0.005 \leq dS \leq 0.6$



С помощью программы PolyPhen мы оценили долю несинонимичных замен, потенциально влияющих на функционирование кодируемого белка, как для быстро и медленно эволюционирующих копий генов из асимметрично эволюционирующих пар паралогов, так и для генов из паралогичных пар эволюционирующих без наблюдаемой асимметрии. В 288 белках, кодируемых генами, которые использовались для исследования асимметрии, произошло 8104 аминокислотных замены. Эти замены классифицировались как “функциональные” или “нейтральные”. 542 из 8104 (6.7%) замен были охарактеризованы, как функциональные, из них 128 оказались в генах из асимметричных пар. Из этих 128 мутаций 127 оказались в быстро эволюционирующих копиях генов и лишь одна - в копиях, эволюционирующих медленно. Количество функциональных мутаций перепредставлено в быстро эволюционирующих копиях генов ($P < 0.001$) и недопредставлено в медленно эволюционирующих копиях ($P < 0.001$) по сравнению с выборкой неасимметричных пар генов.

При асимметричной эволюции одна из двух копий генов в асимметрично эволюционирующих парах набирает мутации очень быстро. Могло оказаться, что это связано с тем, что некоторые из таких генов являются псевдогенами и, поэтому, так быстро накапливают мутации. Чтобы исключить такую возможность следует предоставить свидетельства того, что быстро эволюционирующие копии генов из асимметричных пар не являются псевдогенами. Данные об экспериментально прочитанных фрагментах аминокислотных последовательностей белков, содержащиеся в базе данных Pride могут быть использованы для демонстрации существования белкового продукта гена. С использованием базы данных Pride мы установили, что для 23 из 26 быстро эволюционирующих генов из асимметричных пар прочитаны фрагменты кодируемых ими белков. Оставшиеся три гена и их ближайшие соседи вообще в базе данных Pride представлены не были. Этот анализ свидетельствует в пользу того, что наблюдаемая нами асимметрия генов после дупликации не связана с тем, что один из генов является псевдогеном.

5. Исследование закономерностей эволюции сайтов сплайсинга в семействах паралогичных генов человека

С использованием генов из базы данных RefSeq и программы Blat мы составили выборку семейств многоэкзонных паралогичных генов человека. Мы использовали данные о транскрипции из баз данных EST и мРНК, чтобы обнаружить сайты сплайсинга в исследуемых генах. Мы использовали только те сайты, которые подтверждаются последовательностями сплайсированных EST или мРНК из двух и более библиотек или курированной мРНК из RefSeq (с индикатором "NM").

Участки длиной в 100 нуклеотидов (по 50 нуклеотидов слева и справа от каждого сайта сплайсинга) каждого гена сравнивались со всеми остальными генами семейства, с помощью программы *discontiguous megablast* для выявления участков, гомологичных обнаруженным сайтам сплайсинга. Найденные гомологичные участки объединялись в кластеры. Кластеры, содержащие не менее трех генов и не более одной последовательности для каждого гена, выравнивались программой *ClustalW*. Таким образом, были получены множественные выравнивания участков паралогов, содержащих сайт сплайсинга, и гомологичных им участков.

Также с помощью программы *ClustalW* строились выравнивания белков, кодируемых генами, представленными в анализируемых кластерах. Для этих выравниваний с помощью программы *Phyml* были построены филогенетические деревья. Методом максимального правдоподобия (*maximum likelihood*) с помощью пакета программ *Raml* проводилась реконструкция предковых последовательностей во всех узлах деревьев для участков от -13 до +1 позиции акцепторных сайтов сплайсинга или от -3 до +7 позиции донорных сайтов сплайсинга. Эволюционная история сайтов сплайсинга тоже восстанавливалась методом максимального правдоподобия.

Подобный подход позволил выделить 4 типа ветвей филогенетических деревьев в зависимости от того, произошло ли появление или исчезновение сайта сплайсинга на данной ветви, и от наличия или отсутствия сайта в предковом узле дерева (сайт сохранился; исчез; появился; или не появился). Всего было проанализировано 8688

ветвей. Кроме того, сайты разделялись на альтернативные (используются не во всех транскриптах) и конститутивные (используются во всех транскриптах) и анализировались переходы между этими состояниями. Примерно в 70% рассмотренных семейств многоэкзонных генов было обнаружено хотя бы одно событие изменения модели сплайсинга гена (всего в нашей выборке удалось идентифицировать 399 событий появления или исчезновения сайтов сплайсинга).

В Таблице 3 приведены числа ветвей различного типа, обнаруженных в данном анализе. Общее количество появившихся сайтов сплайсинга (237) несколько превышает количество исчезнувших сайтов (162). В 22% случаев для акцепторных сайтов и в 25% случаев для донорных сайтов появление сайта сплайсинга согласовано с появлением ключевого динуклеотида. В 11% случаев для акцепторных и в 21% случаев для донорных сайтов исчезновение сайта согласовано с мутацией ключевого динуклеотида.

Акцепторные сайты		Потомок	
		Есть сайт	Нет сайта
Предок	Есть сайт	3051	89
	Нет сайта	123	969
Донорные сайты		Потомок	
		Есть сайт	Нет сайта
Предок	Есть сайт	3361	73
	Нет сайта	114	908

Таблица 4 (справа). Появление и исчезновение альтернативных и конститутивных сайтов сплайсинга. Показано число ветвей, на которых появился или исчез альтернативный или конститутивный акцепторный или донорный сайт.

Таблица 3 (слева). Появление и исчезновение сайтов сплайсинга. В ячейках представлены числа ветвей каждого из 4-ех типов (сайт сохранился; исчез; появился; не появился)

	Акцепторные сайты	
	Альтернативные	Конститутивные
Появление	103	20
Исчезновение	69	20
	Донорные сайты	
	Альтернативные	Конститутивные
Появление	94	20
Исчезновение	49	24

В Таблице 4 приведены числа случаев появления и исчезновения сайтов сплайсинга в зависимости от того, является ли сайт альтернативным или конститутивным. Альтернативные сайты появляются и исчезают намного чаще, чем конститутивные, несмотря на то, что конститутивных сайтов в нашей выборке в целом значительно больше.

В Таблице 5 приведено число переходов между альтернативными и конститутивными сайтами сплайсинга. Хотя конститутивных сайтов больше, чем альтернативных, альтернативные сайты чаще становятся конститутивными, чем наоборот (точный тест Фишера, $P < 0.001$), то есть в ходе эволюции происходит “конститутизация” имеющихся альтернативных сайтов сплайсинга.

Акцепторные сайты		Потомок	
		Альтернативный	Конститутивный
Предок	Альтернативный	1031	184
	Конститутивный	162	1674
Донорные сайты		Потомок	
		Альтернативный	Конститутивный
Предок	Альтернативный	1087	210
	Конститутивный	187	1877

Таблица 5. Переходы между альтернативными и конститутивными сайтами. В ячейках показано число ветвей, на которых наблюдается переход.

Мы проанализировали изменения веса сайтов сплайсинга на различных типах ветвей. Вес сайта – это его мера сходства с консенсусом сайтов сплайсинга, полученным на большой выборке подтвержденных сайтов. Вес сайта (H) подсчитывался по формуле:

$$H = \sum \ln(f_i/0.25);$$

Здесь суммирование идет по всем рассматриваемым позициям сайта, а f_i – частота наблюдаемого нуклеотида в i -ой позиции матрицы частот. На рисунках 6 и 7 показана зависимость веса сайта потомка от веса предкового сайта для четырех упомянутых выше типов ветвей. Несмотря на то, что в целом появление сайтов сплайсинга ассоциировано с увеличением их веса (точный тест Фишера, $P < 0.01$) в ряде случаев сайт появился без каких-либо изменений. В некоторых таких случаях (но не во всех), мы обнаружили изменения в сайте партнере (например, появление ключевого динуклеотида или вставку, содержащую “готовый” сайт). Интересно отметить, что наблюдались случаи (всего 24) возникновения и исчезновения экзонов без каких-либо изменений и в донорных, и в акцепторных сайтах.

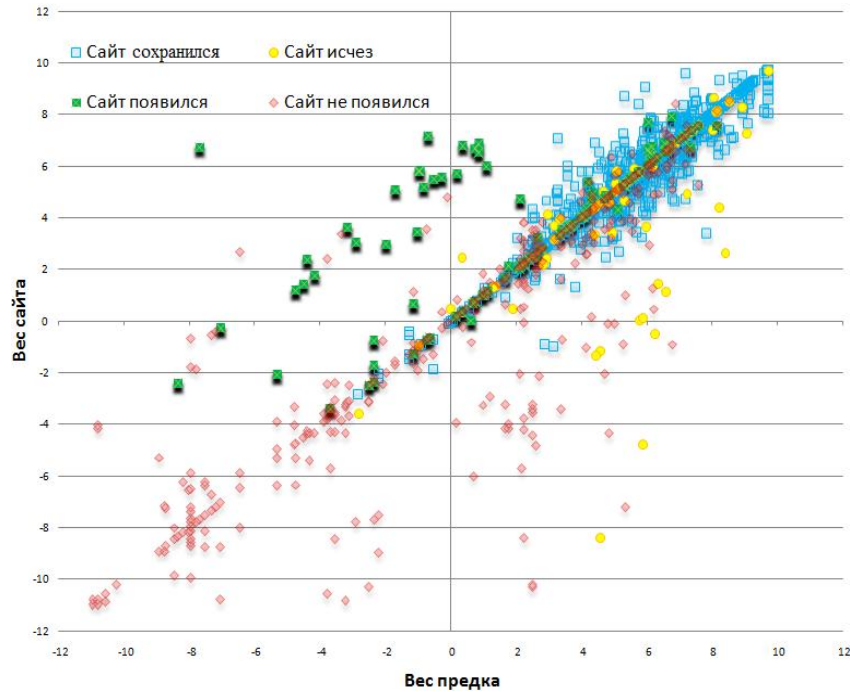
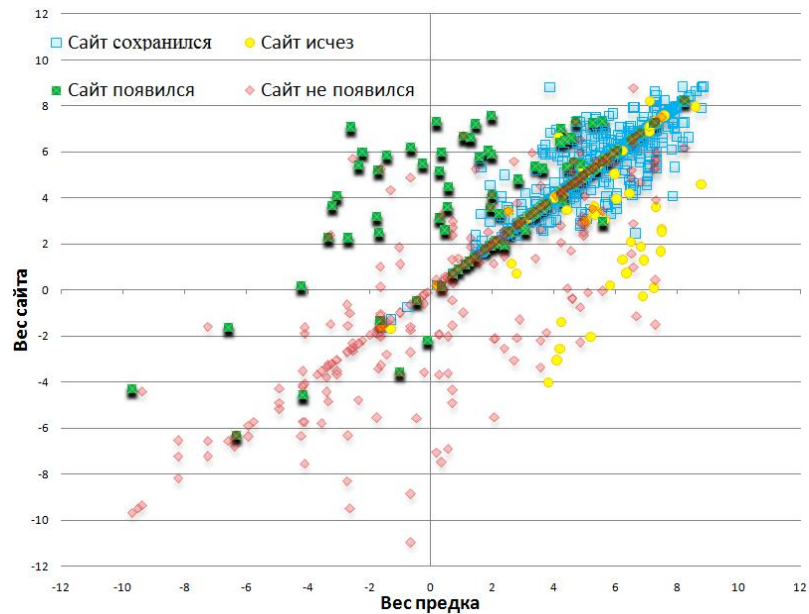


Рисунок 6. Зависимость веса акцепторных сайтов от веса их предков. Одна точка соответствует одной ветви на эволюционном дереве. Зеленым отмечены ветви, на которых сайт появился. Желтым – ветви, на которых сайт исчез. Красным – ветви, на которых сайт не появился. Синим – ветви, на которых сайт сохранился.

Рисунок 7. Зависимость веса донорных сайтов от веса их предков. Одна точка соответствует одной ветви на эволюционном дереве. Зеленым отмечены ветви, на которых сайт появился. Желтым – ветви, на которых сайт исчез. Красным – ветви, на которых сайт не появился. Синим – ветви, на которых сайт сохранился.



Кроме того, оказалось, что вес вновь возникших сайтов сплайсинга в среднем меньше, чем вес сайтов, которые являются потомками существовавших ранее сайтов (U-критерий Мана-Уитни, $P < 0.01$), как в случае донорных, так и в случае акцепторных сайтов. Мы предположили, что это может быть связано с тем, что чаще рождаются альтернативные сайты, вес которых в среднем меньше, чем вес конститутивных сайтов.

Поэтому мы повторно применили U-критерий Мана-Уитни, рассмотрев в качестве контроля только те сайты, которые были и остались альтернативными. Появившиеся альтернативные сайты по-прежнему имели значимо меньший вес ($P < 0.01$), чем сайты из контрольной выборки.

Мы также проанализировали как зависит количество наблюдаемых мутаций в позиции сайта сплайсинга от информационного содержания позиции и от типа рассматриваемой ветви филогенетического дерева (сайт появился, сохранился, исчез, не появился). Информационное содержание I позиции задается следующей формулой:

$$I = - \sum_{\alpha} f_{\alpha} (\log_2 P_{\alpha} - \log_2 f_{\alpha})$$

Здесь α – четыре разных нуклеотида, P_{α} - вероятность наблюдать нуклеотид α (можно принять за 0.25 для всех α), а f_{α} – частота нуклеотида α в данной позиции.

Для ветвей, на которых сайт не исчез, существует сильная отрицательная корреляция между информационным содержанием позиции и количеством мутаций в сайтах сплайсинга (-0,93 и -0,96 для донорных и акцепторных сайтов, соответственно; $P < 0.001$; критерий Пирсона). Для других типов ветвей наблюдается положительная корреляция, как для донорных, так и для акцепторных сайтов, но во всех случаях, кроме появления акцепторных сайтов эта корреляция статистически не значима ($P > 0,05$).

В таблице 6 показано число случаев увеличения веса сайта при переходе альтернативных сайтов в конститутивные. При таком переходе вес сайта чаще увеличивается, чем при сохранении сайта в состоянии альтернативного (точный тест Фишера, $P < 0.05$) как для донорных, так и для акцепторных сайтов.

Тип ветви	Акцепторные сайты		Донорные сайты	
	Вес сайта увеличился	Вес сайта не увеличился	Вес сайта увеличился	Вес сайта не увеличился
Альтернативный сайт переходит в конститутивный	24	160	17	193
Альтернативный сайт сохранился	61	970	42	1040

Таблица 6. Положительные изменения веса сайтов при превращении альтернативных сайтов в конститутивные и при сохранении альтернативности сайтов.

Дупликация генов является хорошо известным механизмом увеличения разнообразия транскриптов и белков. Кроме того, такое разнообразие может достигаться альтернативным сплайсингом. Существует представление, что дупликация генов и альтернативный сплайсинг могут конкурировать между собой: гипотеза о меньшем разнообразии альтернативно сплайсируемых изоформ дублицированных генов неоднократно обсуждалась в литературе. Результаты наших исследований согласуются с таким представлением, свидетельствуя в пользу конститутизации сайтов сплайсинга после дупликаций, однако противоречат ему при анализе появлений и исчезновений альтернативных сайтов сплайсинга – событий появления альтернативных сайтов заметно больше.

6. Метилированные цитозины реже мутируют в CpG островах

Как уже обсуждалось в разделе 2, CG - наиболее недопредставленный динуклеотид в геномах млекопитающих, в том числе в геноме человека. Есть основания полагать, что это связано с повышенной частотой мутаций метилированных CpG динуклеотидов. Однако в геноме человека существуют особые CpG-богатые участки, называемые CpG островами, в которых частота транзиций CpG>TpG примерно в 7 раз меньше, чем в остальном геноме. Считается, что этот эффект связан с пониженным уровнем метилирования CpG динуклеотидов в CpG островах.

Недавно были получены экспериментальные данные по полногеномному метилированию в эмбриональных стволовых клетках человека. Мутации, которые происходят в таких клетках, могут быть переданы в следующие поколения и есть

основания полагать, что профиль метилирования в таких клетках достаточно хорошо соответствует профилю метилирования в клетках зародышного пути в целом. Анализированные в разделе 3 полиморфизмы в геноме человека, для которых удалось реконструировать предковые состояния, можно рассмотреть как направленные мутации в клетках зародышного пути человека. Объединив описанные выше данные, мы сравнили частоту $^5\text{mCpG} > \text{TpG}$ мутаций в CpG островах человека и в остальном геноме, контролируя уровень метилирования.

Для каждого CpG динуклеотида мы фиксировали степень метилирования (доля экспериментальных чтений данного динуклеотида при которых цитозин оказался метилированным от общей доли чтений). Мы отобрали 149507 $^5\text{mCpG}$ динуклеотидов, расположенных внутри CpG островов. Для каждого $^5\text{mCpG}$ динуклеотида из CpG островов мы пытались найти пару – $^5\text{mCpG}$ динуклеотид со сходной степенью метилирования, но вне CpG островов. Если мы не могли найти подходящий контроль, то исходный CpG динуклеотид устранялся из образца. В силу особенностей наших мутационных данных, мы анализировали только те CpG динуклеотиды, которые достаточно хорошо выравнивались с геномами шимпанзе и орангутана, чтобы мы могли восстановить предковое состояние возможного полиморфизма в таком CpG динуклеотиде. Мы использовали три метода для формирования контрольной выборки. Эти методы описаны ниже.

Во всех случаях контрольная выборка формировалась случайным образом из областей генома, не принадлежащих CpG островам. CpG динуклеотиды контрольной выборки должны были находиться на той же хромосоме, иметь примерно такую же степень метилирования и глубину прочтения (число прочтений данного нуклеотида при составлении метилома), что и CpG динуклеотиды из образца ($\pm 10\%$). Таким образом, наша модель учитывает отличия в частотах мутаций между хромосомами. Первый («мягкий») метод формирования контрольной выборки не учитывал влияние локального C+G или CpG состава на частоту мутаций. Второй («умеренный») метод формирования контрольной выборки требовал, чтобы число цитозинов и гуанинов в окне размером в 21 нуклеотид (с центром в рассматриваемой позиции) между образцом и контролем

отличалось не больше, чем на 2, а число CpG динуклеотидов не больше, чем на 1. Третий (жесткий) метод требовал полного совпадения C+G и CpG состава в указанном окне. Мы считали, что CpG динуклеотид мутировал, если в нем наблюдался C/T или A/G полиморфизм с предковым состоянием C или G, соответственно. Сравнение частот мутаций между CpG островами и контролями представлено в Таблице 7.

Метод	Исключенные CpG	Всего CpG	CpG остров		Контроль		Контроль/остров
			Число мутаций	Частота мутаций	Число мутаций	Частота мутаций	
Мягкий	16	149491	949	0.0063	2033	0.0136	2.14
Умеренный	423	149084	944	0.0063	1953	0.0131	2.07
Жесткий	9405	140102	893	0.0064	1654	0.0118	1.85

Таблица 7. Сравнение частот ^{5m}CpG>TpG мутаций в CpG островах человека и в трех типах контролей. Во второй колонке показано число CpG динуклеотидов из CpG островов, для которых не был найден подходящий контрольный динуклеотид.

Мы заметили, что по мере увеличения «жесткости» контрольной выборки уменьшается частота ^{5m}CpG>TpG мутаций. Однако даже в жестком контроле частоты ^{5m}CpG>TpG мутаций примерно в 2 раза выше, чем в CpG островах. Этот эффект статистически значим (точный тест Фишера, $p < 10^{-100}$). Таким образом, различия в степени метилирования, не объясняют всю недопредставленность ^{5m}CpG>TpG мутаций в CpG островах.

Мы исследовали, какие другие факторы могут отвечать за наблюдаемые различия в частотах мутаций между CpG островами и последовательностями контролей. Известно, что при рекомбинации в случае когда между гомологичными хромосомами есть отличие – на одной хромосоме стоит C или G, а на другой A или T, более вероятно превращение A или T в C или G, чем наоборот (смещенная генная конверсия). Поскольку генная конверсия происходит при рекомбинации, можно предположить, что она будет более эффективной в участках генома с высокой частотой рекомбинаций. С использованием данных о частоте рекомбинаций в геноме человека из трех баз данных мы разделили все наши выборки на подвыборки из области с высокой и низкой частотой рекомбинации

(частота рекомбинации ниже или выше средней по трем источникам данных) и посчитали частоты мутаций в эти областях. Результаты представлены в Таблице 8.

Выборка и тип контроля		СрG острова			Контроль			Контроль /остров
		Число мутаций	Всего СрG	Частота мутаций	Число мутаций	Всего СрG	Частота мутаций	
СрG с низким уровнем рекомбинации	Мягкий	230	50896	0.0045	600	49303	0.0122	2.69
	Умеренный	228	50751	0.0045	554	48597	0.0114	2.54
	Жесткий	220	47587	0.0046	468	44911	0.0104	2.25
СрG с высоким уровнем рекомбинации	Мягкий	372	45470	0.0082	681	45124	0.0151	1.84
	Умеренный	370	45359	0.0082	681	45695	0.0149	1.83
	Жесткий	346	42565	0.0081	604	43770	0.0138	1.70

Таблица 8. Сравнение частот $^{5m}\text{CpG} > \text{TpG}$ мутаций в СрG островах человека при высокой и низкой степени рекомбинации.

Во всех сравнениях, частоты $^{5m}\text{CpG} > \text{TpG}$ мутаций в контроле остаются выше (точный тест фишера $P < 10^{-10}$), чем в СрG островах. В областях с низкой частотой рекомбинации отличия между частотами мутаций в СрG островах и вне островов возрастают. Мы предполагаем, что если бы рекомбинации не было вовсе, то разница между СрG островами и остальным геномом была бы существенно больше, чем наблюдаемая в данной работе. Таким образом, процессы, происходящие при рекомбинации, не могут быть причиной наблюдаемого снижения частот $^{5m}\text{CpG} > \text{TpG}$ мутаций в СрG островах.

Еще один эффект, который может вызывать снижение частот $^{5m}\text{CpG} > \text{TpG}$ мутаций в СрG островах, – естественный отбор. Чтобы исключить действие отбора хотя бы на уровне генов, мы сравнили частоты $^{5m}\text{CpG} > \text{TpG}$ мутаций в СрG островах и в контролях в межгенных областях генома и в окрестности генов отдельно. Анализ показал, что в обоих случаях частоты $^{5m}\text{CpG} > \text{TpG}$ существенно ниже в СрG островах (тест Фишера, $P < 10^{-100}$).

Можно предположить, что частота дезаминирования цитозинов в СрG контексте сама по себе ниже в СрG островах в силу пока неизвестных молекулярных механизмов, возможно в связи с наличием каких-нибудь ДНК-белковых взаимодействий в СрG островах или другого состояния хроматина.

ВЫВОДЫ

1. Создана коллекция фрагментов транскриптов человека, содержащая 4,260 последовательностей EST, которые, вопреки обозначенному источнику происхождения, не имеют сходства с какими-либо последовательностями в геноме человека. Анализ этой коллекции показал, что транскриптом человека загрязнен нуклеотидными последовательностями прокариот, вирусов и даже последовательностями других млекопитающих и растений.
2. Проведен анализ встречаемости 1-7 нуклеотидных слов в 139 полных геномах эукариот. Слово ТА является недопредставленным в наибольшем числе геномов. Описан ряд организмов, в геномах которых обнаружены отклонения от известных закономерностей недопредставленности и перепредставленности нуклеотидных слов и выявлен ряд новых закономерностей.
3. Установлено предковое состояние для примерно трех миллионов одиночных полиморфизмов человека. Выявлен более чем трехкратный избыток T>C мутаций во вторых позициях слов ATTG и ATAG, а также A>C мутаций в первой позиции слова ACAА.
4. Проведен анализ эволюции семейств паралогичных генов человека, содержащих более 500 генов. Примерно одна пятая часть пар ближайших паралогов человека накапливают несинонимичные замены с разной скоростью, при этом доля функционально важных несинонимичных замен существенно преобладает в быстро эволюционирующих копиях генов.
5. В 70% рассмотренных семейств многоэкзонных генов обнаружено хотя бы одно событие изменения модели сплайсинга гена после дупликации. Альтернативные сайты эволюционно подвижнее: они чаще рождаются и умирают, а также чаще переходят в конститутивные, чем наоборот.
6. Подтверждено, что частота мутаций метилированных цитозинов в CpG динуклеотидах внутри CpG островов понижена по сравнению с остальным геномом. Показано, что этот эффект не объясняется отличиями в частоте рекомбинаций или локальным нуклеотидным составом.

ПУБЛИКАЦИИ ПО ТЕМЕ РАБОТЫ

1. Panchin A.Y., Spirin S.A., Lukyanov S.A., Lebedev Y.B., Panchin Y.V. "Human trash ESTs - sequences from cDNA collection that are not aligned to genome assembly". // Journal of Bioinformatics and Computational Biology. – 2008 – V. 6 – P. 759-73.
2. Mitrofanov S.I., Panchin A.Y., Spirin S.A., Alexeevski A.V., Panchin Y.V. "Exclusive sequences of different genomes". // Journal of Bioinformatics and Computational Biology. – 2010 – V. 8 – P. 519-34.
3. Panchin A.Y., Gelfand M.S., Ramensky V.E., Artamonova I.I. Asymmetric and non-uniform evolution of recently duplicated human genes. // Biology Direct. – 2010 – V. 5 – P. 54.
4. Panchin A.Y., Mitrofanov S.I., Alexeevski A.V., Spirin S.A., Panchin Y.V. "New words in human mutagenesis". // BMC Bioinformatics. – 2011 – V. 12 – P. 268.
5. Panchin A.Y., Shustrova E.N., Artamonova I.I. "Recently duplicated human genes: Basics of evolution". // J Biomol Struct Dyn. – V. 28 – P. 1068
6. Panchin A.Y. et al. "Human trash EST study". // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'07) – 2007 – P. 124.
7. Панчин А. Ю., Артамонова И. И. "Эволюция недавно дуплицировавшихся паралогов генома человека". «Информационные технологии и системы» (ИТиС'08). Сборник тезисов. – 2008 – С. 308-313
8. Panchin A.Y., Artamonova I.I., Ramensky V.E., Gelfand M.S. "Major trends in the evolution of young human paralogs". // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'09). – 2009 – P. 281.
9. Mitrofanov S.I., Panchin A.Y., Alexeevski A.V., Spirin S.S., Panchin Y.V. "Exceptional nucleotide sequences in genomes of different organisms" // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'09). – 2009 – P. 240.
10. Митрофанов С.И., Панчин А.Ю., Спирин С.А., Панчин Ю.В., Алексеевский А.А. "Исключительные последовательности в геномах различных организмов". // «Информационные технологии и системы» (ИТиС'09). Сборник тезисов – 2009 – С. 351-353.
11. Панчин А.Ю., Артамонова И.И. "Эволюция моделей альтернативного сплайсинга паралогичных генов человека". // «Информационные технологии и системы» (ИТиС'09). Сборник тезисов. – 2009 – С. 347-350.
12. Панчин А.Ю., Митрофанов С.И., Алексеевский А.В., Спирин С.А., Панчин Ю.В. "Зависимость мутационных процессов в геноме человека от контекста". // «Информационные технологии и системы» (ИТиС'10). Сборник тезисов. – 2010 – С. 358-360.
13. Panchin A.Y., Mitrofanov S.I., Spirin S.A., Alexeevski A.V., Panchin Y.V. "Human mutagenesis in context". // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'11). – 2011 – P. 274-275
14. Panchin A.Y., Makeev V.Y., Medvedeva Y.A. "Decreased mutation rate of 5mCpG within CpG islands in the human genome. // Proceedings of the International Moscow Conference on Computational Molecular Biology (MCCMB'11). – 2011 – P. 228.
15. Панчин А.Ю., Артамонова И.И. "Эволюция сайтов сплайсинга в семействах недавно дуплицированных генов человека". // «Информационные технологии и системы» (ИТиС'11). Сборник тезисов. – 2011 – С. 107-111
16. Панчин А.Ю., Макеев В.Ю., Медведева Ю.А. "Снижение частоты мутаций метилцитозина в CpG островах". // «Информационные технологии и системы» (ИТиС'11). Сборник тезисов. – 2011 – С. 271-275
17. Panchin A.Y., Shustrova E.N., Gelfand M.S., Artamonova I.I. "Evolution of recently duplicated human genes." // Abstract of papers presented at the 2010 Cold Spring Harbor Asia Conference "Computational Biology" – P. 12