

На правах рукописи



Гоглева Анна Анатольевна

**Исследование CRISPR-систем прокариотического иммунитета методами
сравнительной геномики**

03.01.09 – математическая биология, биоинформатика

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата биологических наук

Москва 2016

Работа выполнена в группе биоинформатики Федерального государственного бюджетного учреждения науки Института общей генетики им. Н.И. Вавилова Российской академии наук.

Научный руководитель: кандидат биологических наук,
АРТАМОНОВА Ирена Игоревна,
Федеральное государственное бюджетное учреждение науки Институт общей генетики им. Н.И. Вавилова Российской академии наук, старший научный сотрудник, руководитель группы биоинформатики,
Федеральное государственное бюджетное учреждение науки Институт проблем передачи информации им. А.А. Харкевича Российской академии наук, старший научный сотрудник,

Официальные оппоненты: кандидат биологических наук,
САВИЦКАЯ Екатерина Евгеньевна
Сколковский Институт Науки и Технологий,
старший научный сотрудник

доктор биологических наук, профессор РАН
ИЛЬИНА Елена Николаевна
Федеральное государственное бюджетное учреждение Федеральный научно-клинический центр физико-химической медицины Федерального медико-биологического агентства, заместитель генерального директора по научной работе

Ведущая организация: Федеральное государственное автономное образовательное учреждение высшего образования «Балтийский федеральный университет имени Иммануила Канта»

Защита состоится 20 февраля 2017 г в 15.00 часов на заседании диссертационного совета Д 002.077.04 на базе ИППИ РАН Большой Каретный пер., д. 19, стр. 1, Москва, ГСП-4, 127994

С диссертацией можно ознакомиться в библиотеке ИППИ РАН.

Автореферат разослан _____

Ученый секретарь
диссертационного совета
кандидат физико-математических наук

Рожкова Г.И.

Актуальность темы

CRISPR (от англ. **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats, короткие палиндромные повторы, регулярно расположенные группами) — это иммунная система прокариот, обеспечивающая защиту от чужеродных репликонов, в первую очередь — вирусов и плазмид. Хотя CRISPR-системы были впервые описаны в 1987 г (Ishino et al. 1987), их иммунная функция была установлена только в 2005 г (Bolotin et al. 2005; Pourcel et al. 2005; Mojica et al. 2005). Устойчивость к повторным инфекциям приобретается в результате включения в состав CRISPR-кассет коротких последовательностей, спейсеров, комплементарных участкам соответствующих вирусных или плазмидных геномов. Рост CRISPR-кассет имеет направленный характер, а состав и порядок спейсеров является уникальным отпечатком эволюции взаимоотношений между прокариотами и их вирусами в определённых экосистемах.

Одним из важных биологических сообществ является совокупность микроорганизмов, населяющих тело человека — микробиом (Peterson et al. 2009). До недавнего времени бóльшая часть микробного разнообразия, ассоциированного с организмом человека представляла из себя «тёмную материю», недоступную для изучения стандартными микробиологическими методами. Прорыв произошёл благодаря развитию методик, позволяющих напрямую анализировать совокупную ДНК природных сообществ — метабеном (Schloss & Handelsman 2005). На основании метабеномных данных можно оценить таксономическое и функциональное разнообразие сообществ. Помимо бактерий, архей, простейших и микроскопических грибов, неотъемлемым компонентом микробиома человека являются вирусы. Они контролируют численность микроорганизмов, и за счёт этого поддерживают баланс в сложных сообществах (Weitz & Wilhelm 2012). CRISPR-системы — удобный инструмент для изучения динамики эволюционных взаимоотношений прокариот и их вирусов в микробиоме человека.

Степень разработанности темы

Довольно долго о CRISPR-системах микробиома человека было известно крайне мало, так как основные работы были сосредоточены на исследовании CRISPR-систем немногочисленных модельных организмов. В то же время, сам микробиом человека активно изучают. При помощи метабеномного подхода реконструирован ряд природных сообществ населяющих тело человека (Gill et al. 2006; Huttenhower et al. 2012; Li et al. 2012; Kurokawa et al. 2007). По сравнению с остальными участками тела, наиболее разнообразен видовой состав микробного сообщества кишечника, что делает микробиомы кишечника привлекательной моделью для изучения CRISPR-систем. Во время нашего исследования был опубликован ряд работ, где были охарактеризованы CRISPR-системы кишечных

метагеномов, полученных в рамках проекта «Микробиом человека» («Human Microbiome Project», HMP) (Stern et al. 2012; Mick et al. 2013; Rho et al. 2012). Эти работы фокусируются на исследовании состава спейсеров CRISPR-кассет, содержащих уже известные последовательности повторов, но не принимают во внимание структуру CRISPR-кассеты.

Цель и задачи исследования

Цель данной работы — изучить эволюцию и динамику CRISPR-систем в микробиоме человека. В ходе работы было необходимо решить следующие задачи:

1. Идентифицировать CRISPR-кассеты в трёх метагеномных коллекциях микробиома человека.
2. Установить таксономическую принадлежность идентифицированных CRISPR-кассет.
3. Определить тип CRISPR-Cas систем для идентифицированных кассет.
4. Определить источник происхождения спейсеров, т.е. найти протоспейсеры.
5. Сравнить наборы CRISPR-кассет, повторов и спейсеров между разными индивидуальными метагеномами и целыми метагеномными коллекциями микробиомов человека.
6. Исследовать динамику спейсеров и эволюцию CRISPR-кассет микробиома человека.

Научная новизна и практическая значимость

В ходе работы проанализирован состав CRISPR-кассет трёх метагеномных коллекций кишечника человека, двух из них — впервые. Большая часть идентифицированных CRISPR-кассет обнаружена впервые. Определены таксономическое положение и тип CRISPR-Cas систем для найденных кассет, а также идентифицированы протоспейсеры и проанализировано распределение спейсеров и протоспейсеров по индивидуальным метагеномам. Кроме того, исследована динамика функционально важных классов спейсеров.

Исследование CRISPR-систем в микробиоме само по себе является фундаментальной задачей, но оно имеет и прикладное значение. На настоящий момент на основании CRISPR-системы II типа разработана эффективная технология внесения направленных модификаций в геномы широкого спектра организмов, как прокариот, так и эукариот (Dominguez-Bello et al. 2010; Harrison et al. 2014; Ran et al. 2013). Изучение новых CRISPR-систем в метагеномных данных поможет выявить другие привлекательные системы, которые можно использовать в качестве инструментов в молекулярно-биологических исследованиях. Кроме того, изучение CRISPR-систем микробиома человека важно для разработки эффективных протоколов фаготерапии бактериальных инфекций человека (Nilsson et al, 2014).

Основные положения, выносимые на защиту

1. Большая часть контигов, содержащих CRISPR-кассеты, отнесена к типу *Firmicutes*.
2. Сравнение обнаруженных спейсеров с известными виروмами человека, коллекцией NR базы данных GenBank и полными вирусными геномами выявило лишь небольшое число совпадений (протоспейсеров). Большая часть простоспейсеров обнаружена в метагеномных данных микробиомов человека.
3. Состав CRISPR-кассет очень специфичен, лишь небольшое число спейсеров и повторов, встречается в двух и более индивидуальных метагеномах.
4. Спейсеры и соответствующие им протоспейсеры распределяются по индивидуальным метагеномам независимо.
5. Спейсеры, для которых найден протоспейсер в том же индивидуальном метагеноме, располагаются ближе к лидерному концу кассет и являются отпечатком недавних вирусных инфекций.
6. Спейсеры, общие для двух и более метагеномов, располагаются ближе к дистальному концу кассеты и соответствуют более древнему состоянию CRISPR-иммунитета.

Степень достоверности и апробация исследования

Полученные данные согласуются с известными литературными данными.

Основные результаты работы докладывались на:

- 34-й конференции молодых учёных и специалистов ИППИ РАН ИТиС'11 (Геленджик, октябрь 2011);
- 35-й конференции молодых учёных и специалистов ИППИ РАН ИТиС'12 (Петрозаводск, август 2012);
- Международной конференции CRISPR: Evolution, Mechanisms and Infection (St Andrews, University of St Andrews, UK, June 2013);
- 6-ой Московской конференции по вычислительной молекулярной биологии MCCMB'13 (Москва, июль 2013).

По материалам диссертации опубликовано семь печатных работ, из них три – статьи в журналах, рекомендованных ВАК, и четыре — тезисы в материалах конференций.

Объем и структура диссертации

Диссертация состоит из введения, четырех глав, заключения, выводов, благодарностей, списка литературы и 1 приложения. Полный объем диссертации составляет 101 страниц с 17 рисунками и 7 таблицами. Список литературы содержит 162 наименования.

Основное содержание работы

1. Характеристика идентифицированных CRISPR-кассет

На настоящий момент доступны данные нескольких метагеномных исследований микробиома человека, полученные независимыми научными коллективами. Самое крупномасштабное изучение микробиома осуществляется в рамках международного проекта «Human Microbiome Project» (**HMP**) (Li et al. 2012), получен метагеном биома нисходящей ободочной кишки человека («Human Distal Gut Biome project», **DG**) (Gill et al. 2006) и метагеномы кишечника 13 здоровых японцев («13 Healthy Human Gut Metagenomes», **JPN**) (Kurokawa et al. 2007). Мы использовали перечисленные наборы данных для предсказания CRISPR-кассет при помощи существующих алгоритмов — CRT, PILER-CR и CRISPRFinder. Поиск CRISPR-кассет в метагеномных данных является нетривиальной задачей, во-первых, из-за большого объёма анализируемых данных, а во-вторых — из-за высокой степени их фрагментации. В целях борьбы с ложноположительными мы использовали процедуру фильтрации

Наборы идентифицированных кассет проиллюстрированы на **Рисунке 1**, основные параметры кассет суммированы в **Таблице 1**. Наибольшее число CRISPR-кассет обнаружено в метагеноме JPN, примерно в четыре раза меньше кассет найдено в метагеноме HMP и меньше всего кассет предсказано в метагеноме DG. Результаты работы алгоритмов сильно различались для одних и тех же метагеномных данных.

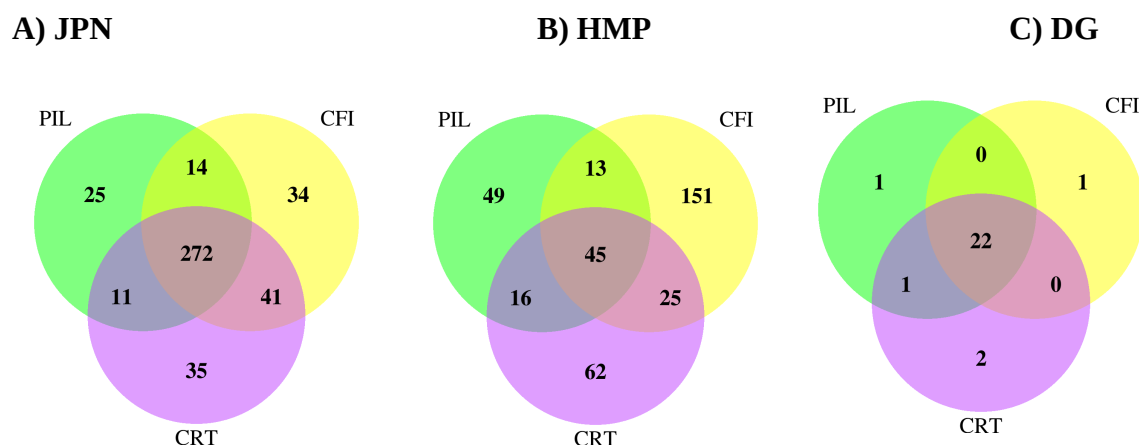


Рис.1. Диаграммы Венна, иллюстрирующие число CRISPR-кассет, предсказанных алгоритмами CRISPRFinder (CFI), PILER-CR (PIL) и CRT в трёх метагеномных наборах данных микробиомов человека.

Среди 134 кассет, расположенных рядом с *cas*-генами, 119 (88%) содержали повторы, не обнаруженные в полных геномах ранее и не представленные в базах данных элементов CRISPR-кассет. Окончательный набор надежно предсказанных кассет состоял из 296, 78 и 24

CRISPR-кассет, обнаруженных в метагеномных коллекциях JPN, HMP и DG, соответственно (Таблица 1).

Таблица 1. Основные характеристики идентифицированных CRISPR-кассет.
Столбцы соответствуют трём метагеномным коллекциям.

Метагеномная коллекция	JPN	HMP	DG
Кассеты, идентифицированные с помощью:			
PILER-CR	322	121	24
CRT	359	149	25
CRISPRFinder	361	235	23
три программы одновременно	272	45	22
1 или 2 программами, но расположены рядом с <i>cas</i> -генами	24	33	2
Набор надёжно предсказанных кассет			
Общее число	296	78	24
Кассеты рядом с <i>cas</i> -генами	70 (24%)	56 (71%)	6 (25%)
Кассеты, для которых было определено таксономическое положение	73 (25%)	69 (82%)	12 (50%)
Кассеты, для которых был определён тип CRISPR-Cas системы:			
I тип	18 (6%)	16 (20%)	1 (4 %)
II тип	9 (3%)	4 (5 %)	1 (4%)
III тип	6 (2%)	18 (23%)	1 (4%)
Спейсеры			
Общее число	3410	378	344
Уникальные спейсеры	2992	352	343
Спейсеры с протоспейсерами, обнаруженными в:			
той же метагеномной коллекции	136	59	6
коллекции NR базы данных GenBank	17	9	0
Повторы			
Уникальные повторы	170	74	19
Повторы, совпадающие с повторами базы данных CRISPRdb	23	0	0
Повторы, отнесённые к известным кластерам согласно алгоритму CRISPRmap	122	18	11

Получив набор надежно предсказанных CRISPR-кассет, мы сравнили последовательности их консенсусных повторов с повторами уже известных CRISPR-кассет, задепонированных в базу данных CRISPRdb (Grissa et al. 2007). Всего лишь 23 повтора из 263 оказались похожими на повторы из базы данных CRISPRdb. Незначительное пересечение с базой данных CRISPRdb указывает на то, что подавляющая часть обнаруженных нами кассет не была известна ранее.

1.2. Таксономия метагеномных контигов, содержащих CRISPR-кассеты

Таксономическую принадлежность контигов с CRISPR-кассетами определяли с помощью blastx-поиска соответствий между последовательностями, фланкирующими кассеты, и коллекцией NR базы данных GenBank.

Установление таксономического положения CRISPR-содержащих контигов является непростой задачей. Во-первых, в силу небольшой длины контигов. Во-вторых, из-за подверженности *cas*-генов частым горизонтальным переносам. Таксономическое положение, по меньшей мере, на уровне домена удалось определить для 73 из 296 кассет метагеномной коллекции JPN (25%), 69 из 78 кассет коллекции HMP (88%) и 12 из 24 кассет коллекции DG (50%). Хотя общее число кассет, идентифицированных в каждой из трёх метагеномных коллекций, существенно различалось, преобладающие таксономические категории, к которым были отнесены CRISPR-содержащие контиги, совпадали (**Рисунок 2, Рисунок 3**). Большая часть контигов с установленной таксономией принадлежала к типу *Firmicutes*. Основная часть контигов отнесённых к типу *Firmicutes* принадлежала бактериям родов *Bacilli* и *Clostridia*.

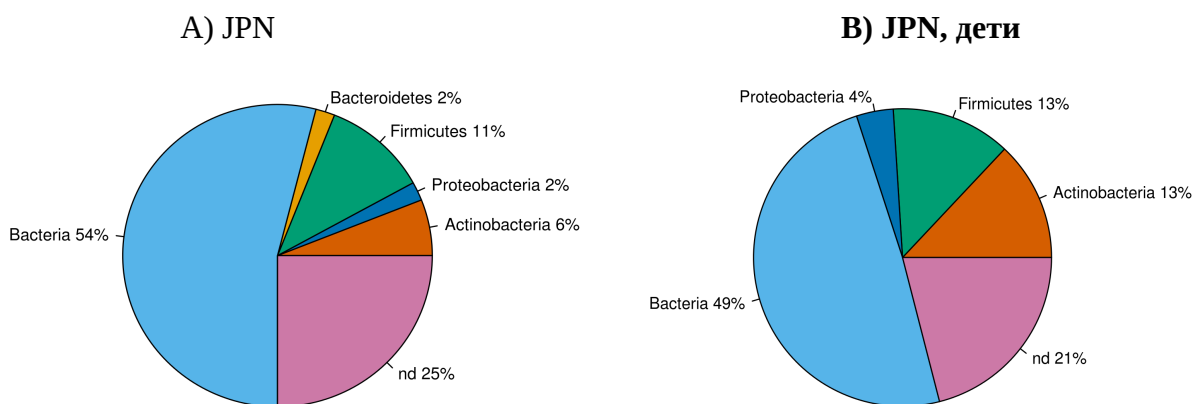


Рис.2. Таксономическое распределение метагеномных контигов коллекции JPN, содержащих CRISPR-кассеты. Условные обозначения: **nd** – таксономическое положение контига не установлено.

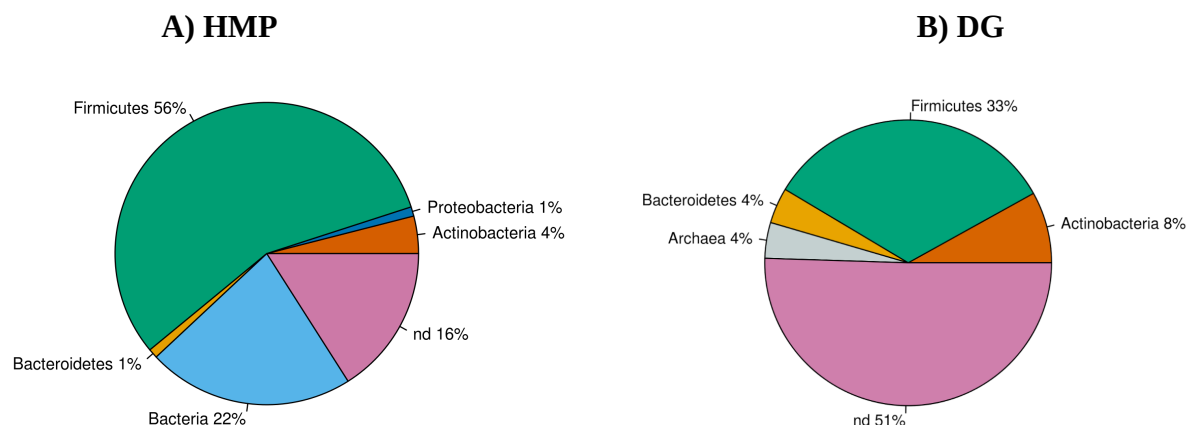


Рис.3. Таксономическое распределение метагеномных контигов коллекций HMP и DG, содержащих CRISPR-кассеты. Условные обозначения: **nd** — таксономическое положение контига не установлено.

1.3 Типы CRISPR-Cas систем

Активно функционирующие иммунные системы CRISPR-Cas состоят из CRISPR-кассет и ассоциированных с ними *cas*-генов (Makarova et al. 2013). Мы классифицировали идентифицированные кассеты по типу последовательности повтора и на основании состава *cas*-локусов там, где это было возможно (**Рисунок 4**).

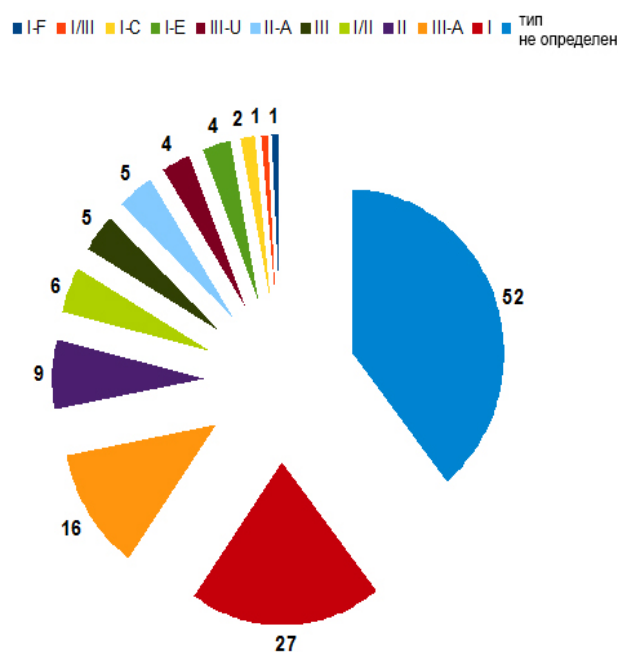


Рис.4. Распределение типов CRISPR-Cas систем среди идентифицированных кассет. Классификация основана на составе сцепленных *cas*-локусов. Голубым цветом («тип не определен») показаны кассеты, для которых *cas*-локус содержал только универсальный ген *cas1*, то есть конкретный тип CRISPR-Cas системы по *cas*-локусу определить было нельзя.

Cas-гены обнаружены рядом со 132 CRISPR-кассетами. В значительной доле случаев (52 кассеты, 39%), единственным *cas*-геном, который можно было идентифицировать, был ген *cas1*. *Cas1* является универсальным маркером CRISPR-Cas систем поэтому не может использоваться для определения типа и подтипа системы. Среди кассет, которые возможно отнести к конкретным типам CRISPR-Cas систем по составу ассоциированных *cas*-генов, 34 отнесены к CRISPR-Cas системам I типа; 25 кассет — к CRISPR-Cas системам III типа и 14 кассет — к CRISPR-Cas системам II типа. Для 29 кассет состав ассоциированного CRISPR-локуса был достаточно специфичен и позволил определить, в том числе, подтип соответствующих CRISPR-Cas систем (**Рисунок 4**).

Последовательности повторов CRISPR-кассет можно разделить на несколько типов по сходству и способности формировать стабильные вторичные структуры (Kunin et al. 2007; Makarova et al. 2011). Как правило, тип последовательности повтора ассоциирован с определёнными *cas*-генами. Недавно разработан алгоритм CRISPRmap, классифицирующий повторы CRISPR-кассет исключительно на основании параметров последовательностей (Lange et al. 2013). По результатам его работы, 194 уникальных повтора, соответствующих 236 идентифицированным нами CRISPR-кассетам, можно отнести к одному из шести суперклассов (A-F). Наиболее распространёнными оказались повторы суперклассов F, E и D. Именно эти суперклассы содержат слабо консервативные последовательности повторов (Lange et al. 2013).

Повторы 160 CRISPR-кассет не удалось соотнести ни с одним из известных суперклассов повторов, согласно классификации CRISPRmap. Для 50 из этих кассет тип CRISPR-Cas системы был определён по составу ассоциированного локуса *cas*-генов (Makarova et al. 2011). Это наблюдение говорит о том, что многие кассеты и повторы, и даже их типы, не были известны ранее, и, соответственно, о неполноте существующей классификации.

1.4 Поиск протоспейсеров

Протоспейсер — это комплементарный спейсеру участок чужеродной ДНК, его прототип. Для поиска протоспейсеров мы сопоставили полученные списки уникальных спейсеров с тремя наборами данных: 1) одноименными метагеномами, содержащими предсказанные кассеты, полагая что, помимо собственно бактериальных и архейных последовательностей, они могут содержать и последовательности вирусов и профагов 2) коллекцией NR базы данных GenBank, пытаясь найти протоспейсеры в известных вирусных последовательностях; 3) доступными метагеномами виромов человека.

В результате поиска протоспейсеров в одноименном наборе метагеномных данных, наибольшее число пар спейсер-протоспейсер (240 пар) было идентифицировано для

спейсеров из метагеномной коллекции JPN (Таблица 2). Чаще всего на метагеномный контиг приходился лишь один протоспейсер. Исключение составили только два контига (HumanGut_CONTIG_00179657 и HumanGut_CONTIG_00179696), содержащие 10 и 16 протоспейсеров соответственно. Столь высокую плотность протоспейсеров объясняет очевидное вирусное происхождение контигов: оба сходны с последовательностями phi29-подобных бактериофагов.

Таблица 2. Общие результаты поиска протоспейсеров.

	JPN	HMP	DG
Одноименная метагеномная коллекция:			
число пар спейсер-протоспейсер	240	35	6
число уникальных спейсеров в парах	136	89	6
число метагеномных контигов с протоспейсерами	165	59	1
Коллекция NR базы данных GenBank:			
число пар спейсер-протоспейсер	75	9	0
число уникальных спейсеров в парах	17	9	0
Вироми кишечника человека:			
число пар спейсер-протоспейсер	1	0	0
число уникальных спейсеров в парах	1	0	0
число метагеномных контигов с протоспейсерами	1	0	0

Сравнивая полученные списки спейсеров с известными вирусными последовательностями коллекции NR базы данных GenBank, мы нашли протоспейсеры к 17 и 9 спейсерам из коллекций JPN и HMP соответственно (Таблица 6). Для спейсеров из кассет метагеномной коллекции DG не было обнаружено ни одного достоверного протоспейсера в последовательностях NR коллекции. Подавляющее большинство найденных протоспейсеров, имели, согласно аннотации, фаговое или плазмидное происхождение.

1.5 Сходство состава спейсеров между метагеномами индивидуальных микробиомов человека

Сравнивая метагеномные коллекции по составу спейсеров, мы обнаружили, что они крайне непохожи. Мы выявили только два спейсера, одновременно присутствующих в наборах данных HMP и JPN. Контиги, содержащие этот участок кассеты, также перекрывались на протяжении небольшой фланкирующей последовательности (длиной 134 нуклеотида).

Сравнивая спейсеры в индивидуальных микробиомах, самое большое число общих спейсеров мы обнаружили в метагеномной коллекции 13 здоровых японцев (JPN).

Максимальное число попарно общих спейсеров приходится на CRISPR-кассеты, предсказанные в индивидуальных метагеномах детей (**Рисунок 5**).

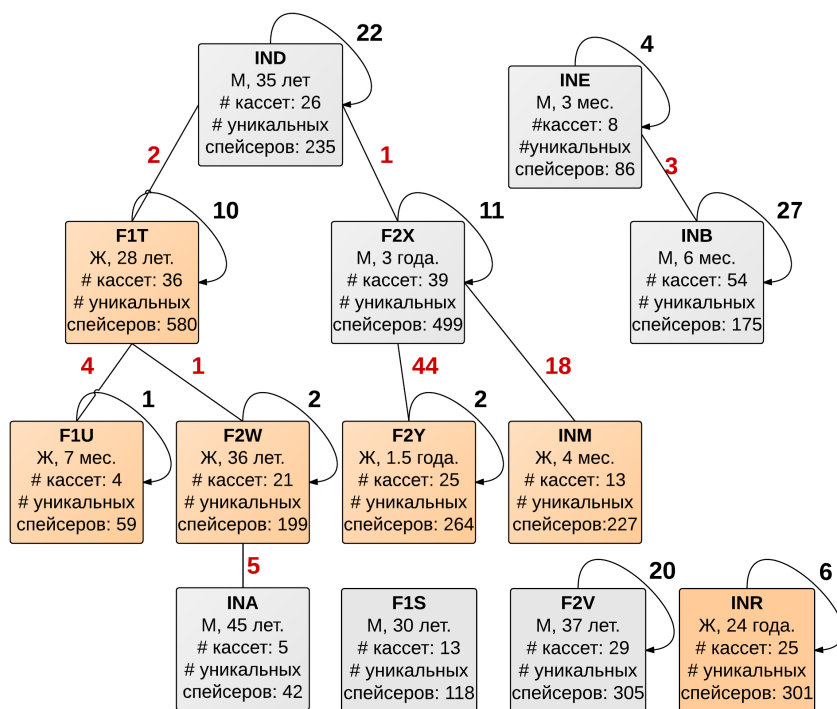


Рис.5. Общие спейсеры в индивидуальных метагеномах коллекции JPN.

Квадраты символизируют индивидуальные метагеномы (индивиды мужского пола обозначены серым, женского пола - оранжевым). Число идентифицированных кассет, уникальных спейсеров и возраст каждого индивида приведены в соответствующем квадрате. Идентификаторы каждого метагенома выделены жирным шрифтом. Идентификаторы, начинающиеся с 'F' (F1 и F2) соответствуют членам двух семей; идентификаторы, начинающиеся с 'IN' соответствуют независимым индивидам. Число общих спейсеров подписано на ребрах, соединяющих индивидуальные метагеномы; число повторяющихся спейсеров в каждом индивидуальном метагеноме подписано на направленных рёбрах.

Среди всех спейсеров CRISPR-кассет метагеномной коллекции JPN, 78 присутствовали в не менее чем двух индивидуальных метагеномах. Для оценки того, насколько значимо такое наблюдение отличается от ожидаемого, мы применили симуляцию на основе пермутаций. Число общих спейсеров для реальных данных было всегда меньше, чем мы наблюдали для симулированных данных ($p\text{-value} < 10^{-5}$), значит, небольшое число общих спейсеров, которое мы видим между индивидуальными метагеномами, нельзя считать случайным. Можно предположить, что даже при сходстве видового состава микробиомов: бактерий и их вирусов,

цепь событий приводящая к формированию набора спейсеров в CRISPR-кассете индивидуальна для каждого конкретного микробиома.

1.7 Колокализация спейсеров и протоспейсеров в индивидуальных метагеномах

После идентификации протоспейсеров мы проверили, склонны ли спейсеры и соответствующие им протоспейсеры находиться в одном и том же индивидуальном метагеноме. Мы проанализировали объединённый набор данных, который состоял из 139 индивидуальных метагеномов. Для спейсеров из 42 (30%) индивидуальных метагеномов мы идентифицировали протоспейсеры в метагеномах других индивидов. Для 38 (27%) таких индивидов мы обнаружили преимущественный метагеном другого человека, на долю которого приходилась бóльшая часть из найденных протоспейсеров.

Достаточно неожиданно, что мы обнаружили большое число протоспейсеров в метагеномных контигах проекта JPN, совпадающих с последовательностями спейсеров, идентифицированных в кассетах метагеномной коллекции HMP. Необходимо подчеркнуть, что число протоспейсеров для HMP спейсеров в JPN наборе данных было гораздо больше числа протоспейсеров, обнаруженных в данных своего метагеномного проекта (**Рисунок 6**).

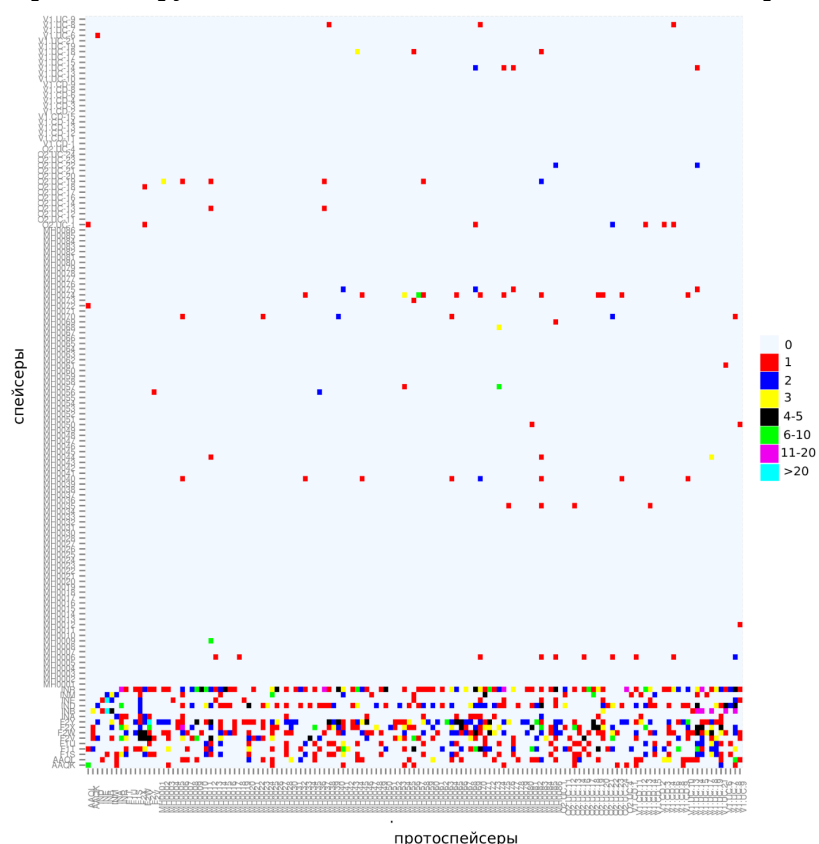


Рис.6 «Тепловая» карта, демонстрирующая распределение пар спейсер-протоспейсер по индивидуальным микробиомам человека. Цвета отражают число обнаруженных пар (расшифровано на рисунке).

Тот факт, что мы видим большое число пар спейсер-протоспейсер, разнесенных по разным индивидуальным метагеномам может говорить о том что, вирусы, ассоциированные с микробиомом кишечника человека, могут быть, в какой-то мере, универсальными. Среди них можно выделить категорию вездесущих вирусов, присутствующих у представителей разных популяций, несмотря на географическую разобщённость последних.

Для статистической проверки гипотезы о колокализации спейсеров и протоспейсеров в рамках одних и тех же индивидуальных метагеномов, мы рассчитали значение критерия корреляции СМН для наблюдаемых и симулированных наборов данных. Сравнение статистики СМН с нулевым распределением позволяет отклонить нулевую гипотезу при уровне значимости 0.05 ($p\text{-value} = 0.0178$) в пользу альтернативной. Альтернативная гипотеза постулирует, что спейсеры и протоспейсеры «отталкиваются», т.е., колокализуются в одном и том же индивидуальном метагеноме реже, чем если бы они были распределены по индивидуальным метагеномам случайно.

Этот факт может указывать на то, что CRISPR-системы большинства рассмотренных индивидуальных метагеномов активны и высоко эффективны против бактериофагов. Аналогичное наблюдение было сделано ранее (Stern et al. 2012). В ряде других работ, посвящённых анализу динамики CRISPR-систем в микробиоме ротовой полости человека (Pride et al. 2011; Pride et al. 2012; Robles-Sikisaka et al. 2013) спейсеры и протоспейсеры часто обнаруживали в метагеномных данных одного и того же индивида (в случаях, когда также были доступны виромные данные для тех же индивидов). Нельзя исключить, что при наличии комплементарных индивидуальных виромов (полученных для тех же самых индивидов), нам удалось бы обнаружить больше протоспейсеров в каждом конкретном индивиде и картина распределения спейсеров и протоспейсеров оказалась бы иной.

1.8 Положение спейсеров с мишенями и общих спейсеров в кассете

Мы проанализировали расположение функционально значимых спейсеров в кассетах. Оказалось, что спейсеры с мишенями (т.е. спейсеры с протоспейсерами в том же индивидуальном метагеноме) сдвинуты ближе к лидерному концу кассеты ($p\text{-value} < 0.0002$) (**Рисунок 7, А**). В тоже время общие спейсеры (т.е. спейсеры, встреченные в двух и более индивидуальных метагеномах) располагаются ближе к дистальному концу кассет ($p\text{-value} < 0.001$) (**Рисунок 7, В**). Как и ожидалось, общие спейсеры соответствуют более древнему состоянию CRISPR-кассеты. Эти наблюдения хорошо согласуются с более ранними сообщениями о динамике спейсеров в пределах CRISPR-кассы [137, 45, 48]. Например, показано, что в ответ на фаговую инфекцию, *Streptococcus thermophilus* меняет состав

CRISPR-кассеты: происходит добавление нового спейсера рядом с лидерной последовательностью и клетка становится устойчивой к этому фагу.

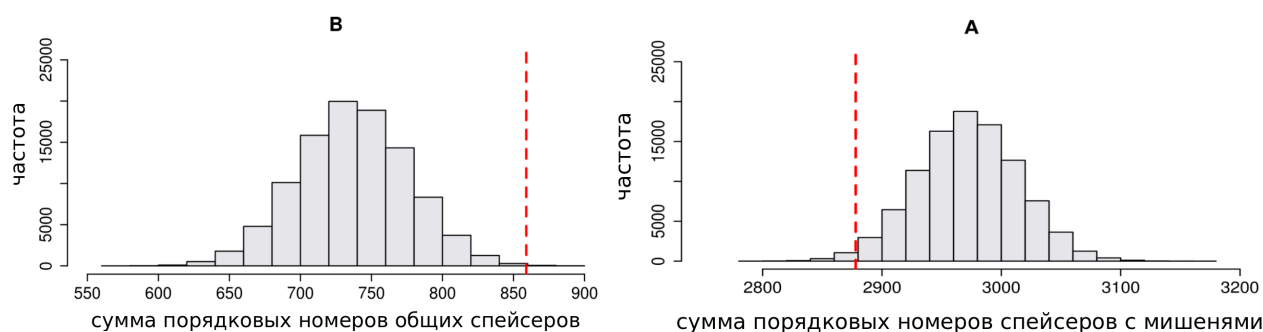


Рис.7. Сдвиг функционально значимых спейсеров по сравнению с расположением всех остальных спейсеров в кассете: (А) спейсеры с мишенями; (В) общие спейсеры. Представлены распределения сумм порядковых номеров функциональных спейсеров для симулированных наборов кассет (100'000 пермутаций). Красной пунктирной линией показано значение той же статистики расположения функционально значимых спейсеров в наблюдаемом наборе кассет.

Наблюдаемую клональность состава спейсеров, расположенных ближе к дистальному концу кассеты, можно объяснить последовательным выметанием отбором (selective sweeps) (Korolev et al. 2013). Эффект выметания отбором заключается в снижении разнообразия последовательностей ДНК в окрестности мутации в результате недавнего положительного отбора (Rubin et al. 2010). Если динамика CRISPR-кассет происходит согласно описанному сценарию, то можно ожидать бóльшую однородность состава спейсеров на дистальных, более древних, концах CRISPR-кассет в пределах всей популяции и одновременно с этим — повышенное разнообразие состава спейсеров на лидерном, более молодом, конце кассет. Именно такую картину мы наблюдаем в проанализированных CRISPR-кассетах микробиомов человека. Спейсеры с мишенями, сконцентрированные на лидерном конце кассеты, можно рассматривать как «горячую точку» CRISPR-Cas опосредованного иммунитета: клетка вырабатывает устойчивость к новому фагу за счет этих спейсеров непосредственно в момент наблюдения.

1.9 Поиск CRISPR-кассет в метагеномных данных. Успехи, сложности и перспективы

Логика построения CRISPR-кассет позволяет отслеживать динамику взаимоотношений прокариот и специфичных к ним вирусов в локальных экосистемах. Зная порядок расположения спейсеров в кассете и их происхождение, теоретически можно реконструировать историю вирусных или плазмидных инфекций конкретного клона бактерий

или архей или даже целого микробного сообщества. Насколько легко это сделать на практике?

С развитием технологий секвенирования стало возможно напрямую, минуя зачастую невозможную стадию культивирования в лаборатории (Rappé & Giovannoni 2003), изучать совокупные наборы генов и геномов всех обитателей природного сообщества — метабеномы. Среди всех природных микробных сообществ наибольшей клинической значимостью обладает микробное сообщество, ассоциированное с телом человека. Основное видовое разнообразие микробиома человека сконцентрировано в кишечнике, и составляет не менее 600 родов. Баланс такого сложного сообщества необходим для поддержания нормального функционирования организма человека (Cénit et al. 2014). Не до конца понятно, за счет чего достигается устойчивость микробиома человека – возможно, важную роль при этом играют вирусы бактерий и архей, препятствующие бесконтрольному размножению отдельных видов-оппортунистов (Minot et al. 2011). Неограниченному размножению вирусов наряду с менее специфическими механизмами препятствуют CRISPR-Cas системы адаптивного прокарриотического иммунитета.

CRISPR-кассеты служат удобным средством для изучения динамики сложных сообществ, так как сохраняют память о предшествующих инфекциях, записанную в форме последовательностей спейсеров. Имея в распоряжении лишь один моментальный снимок микробиома человека — метабеном, можно попытаться смоделировать состав микробиома в прошлом, глядя только на структуру CRISPR-кассет его обитателей. Кроме того, реконструируя CRISPR-кассеты из метабеномных данных, можно узнать больше о самих CRISPR-Cas системах, их биологии и разнообразии.

CRISPR-кассеты в метабеномных данных можно изучать как на уровне сырых чтений, так и на уровне собранных контигов. Разнообразие спейсеров, пожалуй, удобнее изучать непосредственно на уровне чтений, выбирая те из них, которые содержат последовательности уже известных повторов CRISPR-кассет (Stern et al. 2012). Такой подход эффективен для получения совокупного пула спейсеров, однако, не позволяет проследить хронологию включения спейсеров в кассету, определить таксономическую принадлежность кассет, к которым относятся спейсеры, а также ограничен довольно небольшим набором уже известных повторов. Реконструкция полноразмерных кассет на основании метабеномных контигов, напротив, позволяет сохранить порядок спейсеров в кассете и лучше подходит для поиска новых кассет, содержащих неизвестные ранее повторы. Поэтому именно такую тактику мы выбрали для этого исследования.

Для реконструкции полноразмерных CRISPR-кассет годятся не любые метагеномные данные, а отвечающие, по крайней мере, двум критериям. Во-первых, метагеномная сборка должна быть основана на достаточно большом числе сырых чтений. Микробиом человека содержит много минорных видов с небольшой численностью, а так же близких видов, геномы которых очень похожи. Для детекции таких видов и их CRISPR-кассет требуется достаточное покрытие. Во-вторых, метагеномные контиги должны быть достаточно длинными, так чтобы могла уместиться полная CRISPR-кассета, и, желательно, ассоциированный с кассетой локус *cas*-генов и фланкирующие последовательности, необходимые для определения таксономической принадлежности контига.

Из всех общедоступных проектов, посвященных микробиому человека, только три (Ken Kurokawa et al. 2007; Steven R Gill et al. 2006; Qin 2010) отвечали этим критериям на момент начала исследования. Реконструируя CRISPR-касеты в данных этих проектов, мы действительно находим кассеты во всех из них. Однако суммарное число обнаруженных кассет невелико (всего 398), гораздо меньше числа видов в тех же данных (около 5000 видов, по оценкам проекта HMP (Huttenhower et al. 2012)), и сильно различается между метагеномными коллекциями. Например, в данных проекта HMP, содержащих образцы 124 европейцев, найдено почти в четыре раза меньше кассет, чем в данных проекта JPN, полученных при секвенировании образцов от всего лишь 13 здоровых японцев. Метагеномная коллекция HMP содержит примерно в пять раз больше контигов, а их суммарная длина превосходит суммарную длину контигов коллекции JPN в 8 раз, хотя средние длины контигов в коллекциях HMP и JPN сопоставимы.

Вряд ли наблюдаемое различие числа предсказанных кассет можно объяснить вескими биологическими причинами – например, различными типами микробиомов (энтеротипами), характерными для географически разобщенных популяций людей. Скорее всего, данный эффект обусловлен техническими сложностями, связанными со сборкой CRISPR-содержащих последовательностей.

Собирать CRISPR-касеты сами по себе сложно, так как они содержат повторы. Собирать CRISPR-касеты на основании метагеномных данных еще сложнее, так как данные сильно фрагментированы и неоднородны. Кроме того, как уже упоминалось выше, микробиом человека населен большим числом близких видов, поэтому можно ожидать наличие кассет с похожими повторами. Таким образом, помимо самих повторяющихся последовательностей в составе CRISPR-кассет, при сборке метагенома зачастую приходится иметь дело со сходными, но не идентичными повторами, что повышает риск неправильной сборки или сборки химерных последовательностей.

Все три метагенома секвенировали при помощи разных технологий и, соответственно, собирали при помощи разных алгоритмов. Метагеномную коллекцию JPN секвенировали по технологии Сэнгера, с образованием длинных (до 900 нт), высоко точных чтений; контиги собирали при помощи геномного ассемблера PCAP, выравнивая полученные последовательности и ища перекрытия (алгоритм поиска наибольшей общей последовательности) (Huang et al. 2003; Ken Kurokawa et al. 2007). Образцы проекта HMP секвенировали на платформе нового поколения, Illumina. Полученные короткие чтения (100 нт) собирали при помощи специализированного метагеномного ассемблера Metamos (Qin 2010). Алгоритм Metamos разбивает короткие чтения на еще более короткие k-меры, и использует их для построения графов Де Брейна (Treangen et al. 2013). В простом случае поиск пути Эйлера по вершинам приводит к реконструкции индивидуальных контигов. Подход HMP, как с точки зрения секвенирования, так и последующей сборки, гораздо более производителен и удобен для анализа большого числа образцов. Однако, как оказалось, не вполне подходит для точной реконструкции CRISPR-кассет. Напротив, гораздо более трудоемкий (как с точки зрения секвенирования, так и с точки зрения скорости работы алгоритма сборки) подход Kurokawa et al. к сборке метагеномной коллекции JPN, позволяет гораздо более полно и аккуратно реконструировать CRISPR-кассеты. Небольшое число кассет в данных третьей метагеномной коллекции, DG, как и JPN, отсеквенированной по Сэнгеру, скорее всего объясняется небольшим числом исходных сырых данных.

Несмотря на существенные технические сложности и сравнительно небольшое число полных CRISPR-кассет, реконструированных из метагеномных данных микробиомов человека, полученные кассеты выглядят достоверно: имеют регулярную структуру, фиксированную в пределах кассеты длину повторов и спейсеров и, как правило, консервативную последовательность повторов. Сходство некоторых повторов с повторами ранее описанных кассет в полных геномах прокариот, а так же сходство последовательностей спейсеров с участками геномов бактериофагов и плазмид так же говорит в пользу достоверности предсказанных кассет. 33% кассет расположены в контигах, содержащих *cas*-гены, что хорошо согласуется со средней длиной контигов, часто недостаточной для того, чтобы вместить *cas*-локус целиком.

Как и ожидалось, большая часть реконструированных кассет из метагеномных данных микробиомов человека не была известна ранее. Сравнивая полученные повторы с повторами CRISPR-кассет, предсказанных в полных геномах прокариот, мы находим лишь единичные совпадения. Это не удивительно, так как микробиом человека содержит большое число

сложных для культивирования микроорганизмов, о геномах и, тем более, CRISPR-Cas системах которых до сих пор известно совсем немного (или ничего).

Конечно, было бы интересно провести экспериментальную валидацию предсказанных кассет и ассоциированных с ними *cas*-генов. Поиск новых типов *cas*-генов, наряду с рациональным дизайном уже известных (Zhang et al. 2014), — актуальная задача, направленная на оптимизацию существующих протоколов редактирования геномов и расширение спектра сходных задач (таких как глобальная регуляция транскрипции или направленная инициация кроссинговера (Sadhu et al. 2016; Dominguez et al. 2015)).

1.10. CRISPR-кассеты как редуцированное представление о микробном сообществе

Основная и довольно привлекательная идея — использовать CRISPR-кассеты как аппроксимацию, сжатое представление одновременно как о составе микробного сообщества, так и о его хронологии взаимодействия с вирусами. С исключительно практической точки зрения возможность секвенировать не весь микробиом, а только его CRISPR-кассеты, открывает перспективы для быстрого мониторинга состояния природных микробных сообществ. Насколько же полно множество CRISPR-кассет, реконструированное из доступных метагеномных данных, отражает состав и динамику микробиома человека?

Имея контиги с CRISPR-кассетами, можно извлечь данные двух типов, пригодные для оценки состава сообщества: 1) таксономия контигов должна отражать распределение основных таксонов сообщества, так как CRISPR-системы, хотя и не универсальны, но достаточно широко распространены среди прокариот; 2) множество спейсеров должно отражать разнообразие пула вирусов и плазмид, актуальных для рассматриваемого сообщества в данный момент времени или бывшего актуальным некоторое время назад.

Распределение таксонов контигов с кассетами качественно согласуется с распределением таксонов всех контигов микробиома. И в том и в другом случае можно выделить четыре основных типа: *Firmicutes*, *Actinobacteria*, *Proteobacteria* и *Bacteroidetes*. К трем последним типам суммарно отнесено только около 10% всех контигов с кассетами, что в несколько раз меньше, чем аналогичная доля от всех контигов микробиома. Результаты сходны для всех трех метагеномных коллекций. Исключение составляет только выборка детей из метагеномной коллекции JPN, у которых доля одних только актинобактериальных кассет достигает 13%, вдвое больше чем у взрослых из того же набора данных. Это не удивительно, так как в микробиомах детей преобладают *Bifidobacteria*, в силу преимущественно молочной диеты (Turroni F, 2012). Настораживает лишь то, что к типу *Bacteroidetes* удается отнести совсем немного (менее 5%) контигов с CRISPR-кассетами. Это несколько неожиданно, так

как наряду с представителями типа *Firmicutes*, *Bacteroidetes* — один из наиболее стабильных компонентов микробиома человека (Ley et al. 2006). Данному наблюдению можно привести несколько объяснений. Во-первых, не исключено, что численность известных представителей этого типа, оцененная на основании анализа последовательностей 16S рРНК (Hayashi et al. 2002), может быть завышена в силу вариабельности числа копий гена 16S рРНК. Во-вторых, бактерии типа *Bacteroidetes* возможно часто защищаются от вирусов при помощи альтернативных механизмов: систем рестрикции-модификации, за счет модификации поверхностных рецепторов, синтеза экзополисахаридных капсул или задействуя специфические противовирусные белки (Wexler 2007). С другой стороны, рассматривая представителей типа *Bacteroidetes* в контексте густо населенного микробного сообщества, можно предположить, что для защиты от вирусов с широким спектром хозяев, они могли бы полагаться на CRISPR-Cas системы других видов, обитающих тут же.

Таким образом, глядя только на совокупность контигов, содержащих CRISPR-кассеты, можно отследить глобальные изменения состава прокариотической составляющей микробиома человека, пусть и на довольно поверхностном уровне. При этом, как показывает пример с *Bacteroidetes*, важно делать поправку на не универсальную распространенность CRISPR-Cas систем среди бактерий.

Помимо данных о составе прокариотической компоненты микробиома, контиги с кассетами содержат информацию о разнообразии вирусов, циркулирующих в сообществе. Косвенное и самое базовое представление о типах вирусов, распространенных в системе, можно получить, всего лишь зная тип CRISPR-Cas систем, который можно определить на основании повтора и состава *cas*-локуса, ассоциированного с кассетой. Мишенями для разных типов и подтипов CRISPR-Cas систем могут служить как ДНК, так и РНК, то есть разные типы могут атаковать ДНК-вирусы, РНК-вирусы и РНК-стадии ДНК-вирусов (Kazlauskienė et al. 2016).

Среди CRISPR-кассет, найденных в микробиомах человека, представлены все три основных типа (I, II, III), при этом заметную долю составляют кассеты III-A-типа. Мишенью для них может служить как ДНК, так РНК (Kazlauskienė et al. 2016), в случае с РНК не совсем понятно, работают ли системы III-A типа против РНК-вирусов или мРНК ДНК-содержащих вирусов.

По данным независимых исследований, среди вирусов, ассоциированных с микробными сообществами кишечника человека, преобладают умеренные ДНК-вирусы (бактериофаги), однако заметную долю занимают РНК-вирусы растений (Zhang et al. 2006). Последние попадают в микробиом человека с пищей, особенно если он придерживается растительного

рациона. По-видимому, растительные РНК-вирусы являются транзистными компонентами микробиома. Они являются патогенами растений и не представляют угрозу для прокариот микробиома, поэтому вряд ли являются мишенью живых микробиомных CRISPR-Cas систем III-A типа, а мы исходим из предположения, что именно такие системы представляют основную часть наблюдаемого нами множества. Скорее всего, мишенями для CRISPR-кассет III-A типа в микробиоме человека служат ДНК-вирусы. В противном случае, это могут быть РНК-стадии ДНК вирусов и/или бактериофаги, геном которых представлен только РНК (например, как у бактериофага MS2). О РНК-бактериофагах микробиома человека известно довольно . Распространенность CRISPR-Cas систем III-A типа в микробиоме человека может указывать на то, что РНК-бактериофаги присутствуют или присутствовали в системе. Интересные данные получены при длительном скрининге кишечных образцов макак-резус на присутствие двух видов РНК-бактериофагов (Krishnamurthy et al. 2016). Оказалось, что, по крайней мере, у макак, РНК-бактериофаги не являются стабильными компонентами микробиома, для них характерны острые всплески численности. Такая динамика прямо противоположна стабильному присутствию ДНК-бактериофагов в микробиомах как макак, так и человека. Возможно, это объясняет, почему РНК-бактериофаги трудно детектировать при секвенировании микробиомов и виромов кишечника человека, а, возможно, единственное, хотя и только косвенное свидетельство, которым мы располагаем, — распространенность CRISPR-Cas систем III-A типа.

Так, зная только типы CRISPR-Cas систем, уже можно сделать интересные наблюдения о разнообразии вирусов, ассоциированных с микробиомом человека. Более точно видовое разнообразие циркулирующего пула вирусов можно определить, установив происхождение спейсеров в CRISPR-кассетах, то есть найти комплементарные участки вирусных или плазмидных геномов, послужившие их прототипом — протоспейсеры.

Сложность в том, что до сих пор известно довольно мало вирусных последовательностей, то есть полученный набор спейсеров попросту не с чем сравнить. Для ничтожной доли спейсеров (в нашей работе – 0.7%) удается найти протоспейсеры в известных последовательностях вирусного происхождения. Поиск протоспейсеров среди неизвестных последовательностей вирусного происхождения (виромов человека) тоже не очень успешен: нам удалось найти всего лишь один протоспейсер. Это несколько неожиданно, и, возможно, указывает на то, что вирусная составляющая микробиома человека, во-первых, может быть значительно менее стабильной, чем прокариотическая, а во-вторых, — более сложной для секвенирования. Между тем есть указания на то, что состав виромов более или менее устойчив в течение жизни (Minot et al. 2011), но крайне специфичен для каждого человека. В

связи с этим, наиболее перспективно с точки зрения поиска релевантных протоспейсеров было бы исследовать индивидуальные микробиомные данные одновременно со сцепленными с ними вирусными данными. Кроме того, принимая во внимания возможные вспышки численности отдельных вирусов — исследовать серию таких сцепленных данных.

В силу не очень результативного поиска протоспейсеров среди проаннотированных и не проаннотированных последовательностей вирусного происхождения, мы используем спейсеры как пробу для поиска неизвестных ранее вирусов в непосредственно микробиомных данных. Такой подход позволяет найти протоспейсеры уже для большего числа (5.5%) спейсеров. Однако происхождение подавляющего большинства спейсеров остается неизвестным. Во-первых, часть протоспейсеров не удастся найти, так как часто при приготовлении библиотек для секвенирования образцов мелкие вирусные частицы отфильтровываются — так, например, было в проекте HMP. Во-вторых, некоторые из этих спейсеров могут быть довольно древними, то есть комплементарными не существующим уже участкам вирусных геномов или несуществующим уже вирусам.

Итак, CRISPR-кассеты содержат довольно много информации о составе микробного сообщества — как его прокариотической компоненты, так и вирусной. Благодаря направленному росту и сохранению старых спейсеров, они теоретически могут служить основой для реконструкции генетического ландшафта сообщества в прошлом. Однако реконструкция состава природного сообщества по CRISPR-кассетам осложняется тем, что последовательности многих доминантных и минорных видов микробиома человека до сих пор неизвестны (или недостаточно проаннотированы). С накоплением геномных данных, изучение сообщества исключительно по составу его CRISPR-кассет может оказаться удобным средством для быстрого мониторинга состояния клинических образцов.

1.11 Динамика и эволюция CRISPR-кассет в индивидуальных микробиомах

Состав микробиома человека меняется в течение жизни, однако основные таксоны — *Firmicutes*, *Bacteroidetes*, *Actinobacteria* и *Proteobacteria* — как правило, присутствуют постоянно, хотя может колебаться их соотношение. Ряд метагеномных исследований указывает на то, что существует некоторое стабильное ядро доминантных видов прокариот в микробиомах человека, которые должны сосуществовать со специфичными к ним пулами вирусов (Huse et al. 2012; Huttenhower et al. 2012). Следовательно, имея условно сходные изначальные наборы видов прокариот и вирусов в разных индивидуальных метагеномах, можно ожидать, что их CRISPR-кассеты будут похожи. Однако этого не происходит. Сравнивая наборы кассет между 143 индивидуальными микробиомами, мы не находим целиком общих кассет, даже внутри одной и той же метагеномной коллекции, даже среди

кассет, принадлежащих членам одной семьи (коллекция JPN). Все, что удастся найти – это, в большинстве случаев, единичные общие спейсеры в кассетах с одинаковыми или очень похожими повторами. Отсутствие целиком общих кассет и блоков спейсеров, вероятно, отражает тот факт, что последовательность событий встреч с одними и теми же вирусами, а значит и последовательность включения спейсеров в кассету, уникальна для каждого конкретного клона бактерий в каждом конкретном индивидуальном микробиоме.

Небольшое число общих спейсеров между индивидуальными метагеномами, в свою очередь, свидетельствует о высокой динамике CRISPR-кассет. Вероятно, потеря и приобретение спейсеров активными кассетами, в масштабах всего периода жизни человека, за который сменяется большое число поколений обитателей микробиома, происходит довольно быстро и нам просто не удастся застать многие общие элементы.

Примечательно расположение общих спейсеров в кассетах. Они статистически значимо сдвинуты к дистальному концу кассет, а значит, соответствуют наиболее древнему состоянию CRISPR-иммунитета. Дистальный конец, где происходит накопление спейсеров, гораздо более стабилен, по сравнению с лидерным, что повышает шанс застать общие элементы именно в этой области кассеты. С эволюционной точки зрения сохранять некоторые старые спейсеры в составе кассет может быть выгодно, так как не исключено повторное заражение сходными вирусами в будущем. В случае, если в кассете уже есть подходящий спейсер, иммунный ответ сработает гораздо быстрее и эффективнее. Однако не до конца понятен молекулярный механизм, обеспечивающий удержание таких спейсеров в кассете, как и механизм, контролирующий рост кассет в целом.

С точностью до наоборот, лидерный конец кассеты отражает состояние CRISPR-иммунитета в момент его формирования. Здесь активно происходит включение новых спейсеров против вирусов, атакующих клетку в данный момент. В микробиомных данных удастся найти в целом довольно мало протоспейсеров, однако спейсеры с протоспейсерами в том же самом индивидуальном метагеноме преимущественно сосредоточены рядом с лидерным концом кассет. Иными словами, спейсеры, расположенные в начале кассеты, отражают ситуацию более вероятного сосуществования вируса и кассеты. Со временем, бактериальная клетка переживает новые вирусные инфекции и бывшие активными спейсеры сдвигаются ближе к дистальному концу кассеты, где становятся частью долговременного иммунитета, либо выщепляются из кассеты со временем.

Глядя на распределение спейсеров и протоспейсеров по индивидуальным метагеномам, мы наблюдаем интересный эффект отталкивания: чаще всего в системе удастся застать либо спейсер, либо соответствующий ему протоспейсер. CRISPR-системы и вирусы

находятся в состоянии постоянного противостояния. Если мы наблюдаем в системе только спейсер, это говорит о том, что CRISPR-иммунитет оказался эффективен. В противном случае, если мы видим протоспейсер, то есть вирус жив, CRISPR-система оказалась неэффективна, либо не была задействована в защите от этого вируса. Если отбросить обсуждавшиеся ранее технические сложности, сопряженные с поиском кассет и протоспейсеров, тот факт, что в целом удается найти довольно мало протоспейсеров в микробиомных данных, может свидетельствовать в пользу высокой эффективности CRISPR-Cas систем.

Итак, несмотря на относительную стабильность состава микробиома человека, CRISPR-кассеты изменяются очень быстро и согласованно с изменением состава циркулирующего пула вирусов. Совокупность вирусов прокариот, ассоциированных с микробиомом человека, судя по всему, гораздо менее стабильна по сравнению с прокариотической составляющей и/или в большей степени уникальна. При этом CRISPR-Cas системы, вероятно, один из основных механизмов, позволяющий поддерживать баланс между популяциями вирусов и прокариот в микробиоме человека.

Заключение

Мы проанализировали CRISPR-системы трёх метагеномных коллекций микробиомов человека. Для поиска кассет использовали собранные метагеномные контиги и специальную процедуру фильтрации, основанную на применении всех трех общедоступных программ для предсказания кассет. Такой подход позволил нам идентифицировать большое число кассет, неизвестных ранее, и охарактеризовать эволюционную динамику спейсеров.

Большая часть контигов, содержащих идентифицированные кассеты, была отнесена к типу *Firmicutes*. Была обнаружена только одна кассета в контиге архейного происхождения. Сравнение полученного множества спейсеров с известными полными геномами вирусов и вирусными последовательностями коллекции NR базы данных GenBank позволило выявить протоспейсеры для очень малой доли (0.7%) всех спейсеров. Этот результат свидетельствует о том, что глобальное пространство вирусных последовательностей до сих пор остаётся сильно неисследованным. Напротив, мы обнаружили на порядок больше совпадений между последовательностями спейсеров (5.5% всех спейсеров) и метагеномами микробиомов.

Мы показали незначительное перекрытие по спейсерам между индивидуальными метагеномами, принадлежащими разным метагеномным коллекциям. Для спейсеров из 30% индивидуальных метагеномов мы обнаружили протоспейсеры в других индивидуальных метагеномах, в том числе — других метагеномных коллекциях. Проанализировав характер

колокализации спейсеров и протоспейсеров в индивидуальных метагеномах, мы показали, что пары спейсер-протоспейсер, как правило, разнесены по разным индивидуальным метагеномам. Отсутствие протоспейсеров в том же индивидуальном метагеноме может являться следствием высокой эффективности CRISPR-систем против соответствующих вирусов и, как результат, их быстрой элиминации из микробных сообществ.

Кроме того, наличие протоспейсеров в индивидуальных метагеномах других индивидов, в том числе из географически удалённых популяций, может говорить в пользу существования некоторого общего пула широко распространённых (универсальных) вирусов, характерных для микробиома человека в целом.

Поиск кассет в метагеномных контигах позволил нам реконструировать порядок расположения спейсеров. Спейсеры с мишенями статистически значимо сдвинуты к лидерному концу CRISPR-кассет. Известно, что рядом с лидерной последовательностью располагаются новые спейсеры, поэтому спейсеры с мишенями являются отпечатком наиболее недавних фаговых инфекций. И наоборот, спейсеры, общие для нескольких индивидуальных метагеномов, были сдвинуты к дистальному концу кассет, то есть соответствовали более древнему состоянию CRISPR-опосредованного иммунитета.

Выводы

1. Таксономическое распределение CRISPR-содержащих контигов для двух метагеномов (HMP и DG) качественно совпадает с результатами анализа генов 16S рПНК: большая часть контигов (21%), содержащих CRISPR-кассеты отнесена к типу *Firmicutes*.

2. Сравнение обнаруженных спейсеров с известными виروмами человека, NR коллекцией базы данных GenBank и полными вирусными геномами выявило протоспейсеры лишь для 0.7% спейсеров.

3. Основная часть пар спейсер-протоспейсер (77%) приходится на протоспейсеры, обнаруженные в метагеномных данных микробиомов человека.

4 Состав CRISPR-кассет очень специфичен: только 2 % спейсеров и 15% кластеров повторов, встречаются в двух и более индивидуальных метагеномах.

5. Пары спейсер-протоспейсер «отталкиваются», то есть редко колакализируются в одном и том же индивидуальном метагеноме.

6. Спейсеры с мишенями статистически значимо сдвинуты к лидерному концу кассет.

7. Спейсеры, общие для двух и более метагеномов, статистически значимо сдвинуты к дистальному концу кассеты.

Список публикаций

Статьи из перечня рецензируемых журналов

1. Gogleva AA, Gelfand MS, Artamonova II. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs // BMC Genomics. – 2014. –Mar 17.–15:202.
2. Гоглева А.А, Артамонова И.И. CRISPR системы: структура и гипотетические функции // Природа. – 2014. – № 1186. – С. 16 - 21.
3. Гоглева А.А, Артамонова И.И. CRISPR системы: механизм действия и применения // Природа. – 2014. – № 1186. – С. 3 - 9.

Тезисы докладов

1. Gogleva A. Composition of CRISPR-cassettes reflects the virome diversity in the human gut microbiome / Gogleva A.A, Artamonova I.I. Труды 35-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'12), Петрозаводск, 2012. - С. 262 - 264.
2. Гоглева А. CRISPR-системы прокариотического иммунитета в микробиоме человека/ Гоглева А.А, Артамонова И.И. Труды 34-й конференции молодых ученых и специалистов ИППИ РАН (ИТиС'12), Геленджик, 2011. - С. 135 - 139.
3. Gogleva A. CRISPRome reflects the virome composition in the human gut microbiome / Gogleva A.A, Artamonova I.I. CRISPR: Evolution, Mechanisms and Infection University of St Andrews, UK, 2013. - P017, P. 12
4. Gogleva A. CRISPRome reflects the virome composition in the human gut microbiome / Proceedings of Russian-German seminar «Regulation and Evolution of Cellular Systems» (RECESS, Регуляция и эволюция клеточных процессов), Венеция, 2013.
5. Gogleva A. Artamonova I. CRISPR systems in microbiomes. Proceedings of Moscow Conference on Computational Molecular Biology (MCCMB'13), Москва, 2013.

Подписано в печать 19.12.2016г. Формат 60х84/16. Усл. печ. л. 1,63.
Тираж 100 экз. Заказ 262.
ИП Матика Виктор Евгеньевич 624020 Свердловская обл.
г. Сысерть ул. Гагарина 1-1