ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ОБЩЕЙ ГЕНЕТИКИ им. Н.И. ВАВИЛОВА
РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

ГОГЛЕВА АННА АНАТОЛЬЕВНА

ИССЛЕДОВАНИЕ CRISPR-СИСТЕМ ПРОКАРИОТИЧЕСКОГО ИММУНИТЕТА МЕТОДАМИ СРАВНИТЕЛЬНОЙ ГЕНОМИКИ

03.01.09 - математическая биология, биоинформатика

Диссертация на соискание учёной степени
кандидата биологических наук

Научный руководитель:
к.б.н.
Артамонова Ирена Игоревна

Москва – 2016

Оглавление

Введение

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 [1] , их иммунная функция была установлена только в 2005 г [2]–[4]. 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 [5]. Изучение микробиома представляет интерес, поскольку он оказывает существенное влияние на здоровье человека. 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 — метагеном [6]. На основании метагеномных 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 [7]. 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 метагеномного подхода реконструирован ряд природных сообществ населяющих тело человека [8]–[11]. Основная цель таких исследований – понять роль эндогенной микрофлоры в развитии заболеваний и поддержании здоровья человека. 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) [12]–[14]. 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 [15]–[17]. 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 [18].

Методология и методы исследования

Для решения поставленных задач применялись методы сравнительной геномики. Использовались все существующие алгоритмы предсказания CRISPR-кассет, а также методы кластеризации повторов. Процедура реконструкции структуры CRISPR-кассет впервые применена к трем метагеномным коллекциям кишечника человека. Для определения уровня значимости полученных результатов проводились симуляции по методу Монте-Карло.

Основные положения, выносимые на защиту

1. Большая часть контигов, содержащих CRISPR-кассеты, отнесена к типу *Firmicutes*.

2. Сравнение обнаруженных спейсеров с известными виромами человека, коллекцией NR базы данных GenBank и полными вирусными геномами выявило лишь небольшое число совпадений (протоспейсеров). Большая часть простопейсеров 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);
- Русско-немецком семинаре «Regulation and Evolution of Cellular Systems» (RECESS, Регуляция и эволюция клеточных процессов) (Венеция, май 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.

Личный вклад

Все основные результаты, включённые в диссертацию, получены лично соискателем. Обсуждение и интерпретация результатов осуществлялись совместно с научным руководителем.

Объём и структура работы

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

Глава 1. Обзор литературы

1.1 CRISPR-Cas системы

1.1.1 Основные элементы CRISPR-кассет

CRISPR — это система адаптивного иммунитета прокариот. Впервые необычные повторяющиеся последовательности равной длины, перемежающиеся уникальными участками, были описаны в 1987 г. в геноме *E.coli* рядом с геном *iap* [1]. Позднее подобные структуры были обнаружены в геномах многих видов прокариот [19]–[21]. На данный момент известно, что CRISPR очень широко распространены: кассеты встречаются в геномах 90% архей и 60% бактерий [22], [23].

CRISPR-система состоит из двух принципиальных компонентов: CRISPR-кассет и Cas-белков (от англ. CRISPR-associated proteins). Каждая функциональная кассета содержит элементы трёх типов: лидерную последовательность, спейсеры и повторы (**Рисунок 1**).

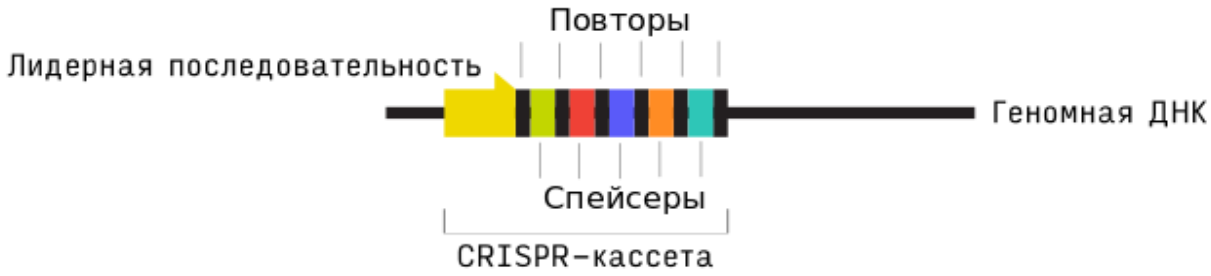


Рис.1. Структура CRISPR-кассеты.

Повторы

Средняя длина повтора составляет около 30 пар оснований. Повторы в пределах одной кассеты, как правило, идентичны между собой по последовательности и длине, реже — могут различаться по одному-двум, часто концевым, нуклеотидам. У многих видов последний повтор кассеты не совпадает в нескольких концевых позициях с канонической последовательностью остальных повторов в той же кассете [24].

Для повторов характерна частичная диадная симметрия, то есть часть последовательности в начале повтора обратно комплементарна участку последовательности соответствующей длины в конце повтора. При транскрипции CRISPR-кассет концы повторов могут комплементарно взаимодействовать между собой благодаря диадной симметрии с образованием устойчивых вторичных структур, прежде всего, различных шпилек [25]. Шпильки необходимы для взаимодействия с Cas-белками [26].

Спейсеры

Между повторами располагаются спейсеры (**Рисунок 1**). Длина спейсеров совпадает в пределах кассеты и примерно равна длине повторов. Чаще всего все спейсеры в кассете имеют различную последовательность. Набор спейсеров в штаммах одного вида, как правило, сильно различен. Благодаря высокой вариабельности CRISPR-локусы используются для быстрого типирования бактериальных штаммов, например, *Mycobacterium tuberculosis* [27], *Yersinia pestis* [28] *Streptococcus pyogenes* [4], *Corynebacterium diphtheriae* [29] и *Campylobacter jejuni* [30]. Сравнение последовательностей спейсеров с известными нуклеотидными последовательностями показало, что некоторые спейсеры совпадают с участками вирусных и плазмидных геномов [2]–[4]. Это впоследствии позволило доказать иммунную роль CRISPR.

Лидерная последовательность

В начале CRISPR-кассеты располагается лидерная последовательность. Она задает направление транскрипции кассеты (**Рисунок 1**). Длина лидерной последовательности значительно больше длины повторов и спейсеров, и составляет в среднем 400 пар оснований. Установлено [31], что лидерные последовательности не содержат открытых рамок считывания и, как правило, АТ-богаты. Двухцепочечная ДНК в АТ-богатых участках плавится при более мягких условиях [32]. Кроме того, в АТ-богатых регионах малая бороздка ДНК имеет меньшую ширину — такая топология служит характерным местом посадки для многих белков, взаимодействующих с ДНК [32], [33]. АТ-богатые участки часто встречаются в различных регуляторных последовательностях (например, промоторах), а также точках начала репликации. Предполагают, что лидерная последовательность регулирует транскрипцию CRISPR-кассет, а следовательно и функционирование всей системы [24]. Для ряда организмов наличие промоторов в лидерной области было подтверждено экспериментально [34].

Cas-белки

Рядом с CRISPR-кассетами располагаются локусы *cas*-генов. Cas-белки многочисленны и разнообразны, обеспечивают молекулярные механизмы CRISPR-опосредованного иммунитета. Они содержат функциональные домены, участвующие в различных взаимодействиях с нуклеиновыми кислотами [35].

Четыре гена: *cas1-cas4* часто располагаются в непосредственной близости от кассет[36]. Наиболее часто эти гены собраны в локус вида *cas3-cas4-cas1-cas2* и транскрибируются совместно [31]. Cas1 находят в геномах всех без исключения организмов, содержащих

CRISPR, поэтому данный ген является универсальным маркером системы. Для белка Cas1 характерен выраженный положительный заряд, который может способствовать электростатическому взаимодействию с отрицательно заряженным сахаро-фосфатным остовом ДНК. Функции белка Cas2 долго оставались неизвестными, лишь в 2012 была показана его эндорибонуклеазная активность [37]. В некоторых случаях функциональные домены Cas2 и Cas3 транслируются как единый белок [36]. Cas3 обладает хеликазной активностью, Cas4 — сходен с экзонуклеазами семейства RecB и содержит структурный мотив, богатый остатками цистеина [38], что может говорить о его ДНК-связывающей активности.

Первая классификация Cas-белков построена в результате анализа 200 полных геномов прокариот и содержит 45 семейств, подразделённых на 8 подтипов [35]. Позднее выделено большее число подтипов Cas-белков на основании филогенетической классификации систем из 703 полных геномов архей и бактерий [36]. Вероятно, классификация будет развиваться по мере описания новых Cas-белков.

1.1.2 Гипотезы о роли CRISPR-систем в клетках прокариот

На основе функций Cas-белков была выдвинута гипотеза о связи CRISPR-системы с процессами перестройки ДНК. В частности, было выдвинуто предположение о участии CRISPR в репарации ДНК у термофильных бактерий и архей [39]. Термофильные микроорганизмы обладают высокой устойчивостью к воздействию различных факторов, повреждающих ДНК, таких как ионизирующее и ультрафиолетовое излучение, а также химические мутагены. Тем не менее, систем репарации, сходных с уже описанными, у этих организмов не обнаружено.

В пользу этой гипотезы также говорило сходство некоторых Cas-белков с эндонуклеазами RecB, принадлежащих RecBCD — основной системе рекомбинационной репарации *E.coli* [40]. Ряд Cas-белков содержит домены, сходные с каталитическими доменами ДНК- и РНК-полимераз (например, Cas10), а также хеликаз (например, Cas3), участвующих в репарации ДНК [36]. Тем не менее, функции многих Cas-белков оставались неизвестными. Эти белки были названы RAMP-белками (**R**epeat **A**ssociated **M**ysterious **P**roteins — загадочными белками, ассоциированными с повторами).

Предполагали, что RAMP могли служить дополнительными регуляторными ДНК-связывающими субъединицами репарационных комплексов, или из этих белков могли быть построены «скользящие зажимы» (sliding clamps) [39] ДНК-полимераз.

Согласно другой гипотезе, CRISPR-система участвует в сегрегации репликонов. В пользу этой гипотезы говорит сходство повторов кассет и итеронов parC области [41]. Показано, что введение дополнительных кассет в составе плазмид в клетки *Haloferax volcanii* и *Haloferax mediterranei* снижает жизнеспособность клеток и часто приводит к отклонениям в распределении генетического материала при делении. CRISPR-кассеты *H. volcanii* и *H. mediterranei* находятся в наиболее крупных репликонах — мегаплазмидах и собственно хромосомной ДНК. Такое расположение указывает на то, что CRISPR может выступать в роли системы сегрегации и отвечать за правильное распределение генов наиболее крупных репликонов по дочерним клеткам, в то время как распределение плазмид меньшего размера с несущественными для выживания генами является до определённой степени стохастическим [21]. Кроме того, было предположено что повторы могут служить мишенью для рекомбинации, тем самым обеспечивать механизм генерации дополнительной изменчивости в геномах прокариот [42].

1.1.3 Доказательство иммунной функции CRISPR

В 2005 г. было показано, что последовательности спейсеров кассет *Streptococcus thermophilus* и *Streptococcus vestibularis* часто совпадали с участками генов бактериофагов, специфичных к стрепококкам, или плазмид *S. thermophilus* и *Lactococcus lactis* [2]. Кроме того, последовательности некоторых спейсеров совпадали с последовательностями бактериальных геномов [2]–[4]. При этом ряд последовательностей комплементарен фрагментам профагов, интегрированных в бактериальный геном.

Впоследствии внехромосомное происхождение спейсеров было показано для ряда бактерий [43] и архей [44]. Спейсеры обладают высокой вариабельностью, и соответствуют случайным участкам вирусных или плазмидных геномов (протоспейсерам). Связи между расположением спейсеров в кассете и протоспейсеров в геноме не обнаружено.

Продукты генов, содержащих протоспейсеры, задействованы в процессах репликации ДНК, сборки вирусных частиц, интеграции умеренных бактериофагов в геном клетки хозяина, а также и их обратной активации, сегрегации репликонов. Эти функции являются необходимыми для поддержания мобильных элементов, т.е., проникновения в клетки прокариот, размножения, а также дальнейшего распространения [2], [45].

Одновременно с установлением внехромосомного происхождения спейсеров, накопился ряд важных наблюдений о деятельности и функциях CRISPR-систем:

- показана корреляция между числом спейсеров в CRISPR-кассете и устойчивостью к фаговым инфекциям (у *S.thermophilus*) [2];

- механизм образования CRISPR-кассет тесно связан с *cas*-генами [31].
- CRISPR-кассеты транскрибируются (у *Archeoglobus fulgidus* и *Sulfolobus sulfoctaricus*) с образованием малых РНК. Таким образом, CRISPR-кассеты являются активными компонентами генома [46], [47].

Была сформулирована гипотеза о том, что CRISPR-система защищает прокариот от вирусных и плазмидных инвазий [2]. Было высказано предположение, что включение участков геномов мобильных элементов в виде спейсеров является генетически наследуемой иммунной памятью. CRISPR-система широко распространилась среди прокариот благодаря тому, что обеспечивает существенный выигрыш в приспособленности.

Впоследствии гипотеза об иммунной функции CRISPR получила экспериментальное подтверждение. Показано, что после заражения *S. thermophilus* бактериофагами, в состав CRISPR-кассет выживших клонов добавляется 1-4 новых спейсера, рядом с лидерной последовательностью [45]. Более того, при повторном заражении, большая часть клонов выживала. Новые спейсеры были комплементарны участкам генома заразившего бактерию фага.

В результате искусственной интеграции участков фагового генома в CRISPR-кассеты, штаммы приобретали устойчивость к исходному вирусу. Более того, удаление спейсеров приводило к потере устойчивости [45].

Непосредственно после заражения последовательность спейсера полностью совпадает с протоспейсером. Со временем между спейсером и протоспейсером накапливаются различия, в силу высокой скорости мутагенеза вирусов, и бактерии становятся менее устойчивыми к последующим заражениям. Неэффективные спейсеры могут быть утеряны в результате выщепления из-за рекомбинации между повторами кассеты [43], [48].

В результате накопления точечных мутаций в последовательности протоспейсера, вирусы избегают действия CRISPR-системы. Позиции внутри спейсера не эквивалентны по чувствительности к мутациям, единичные замены определённых (якорных) областей спейсера [49] сразу лишают клетку устойчивости к вирусу, замены в других областях — менее опасны, так как делают CRISPR-иммунитет только менее эффективным [50].

Стоит отметить, что новые спейсеры встраиваются в кассеты только рядом с лидерной последовательностью (**Рисунок 2**). Направленное включение новых спейсеров в состав CRISPR-кассет позволяет реконструировать историю взаимоотношений прокариот и их вирусов на определённом эволюционном промежутке.

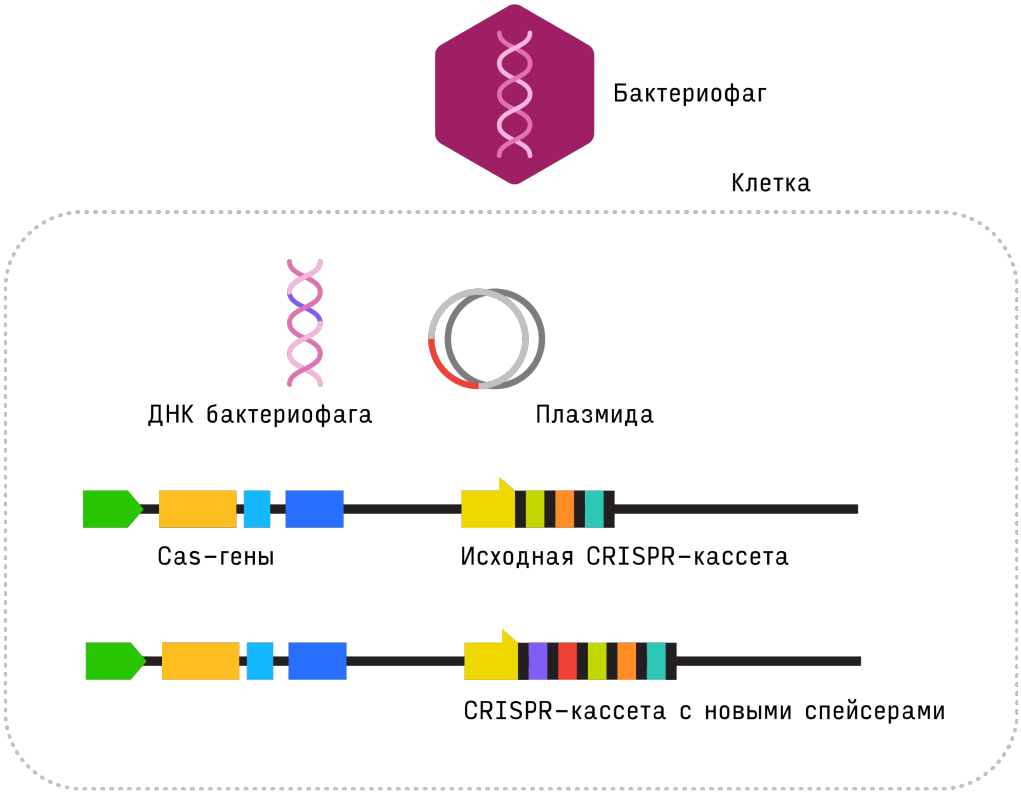


Рис.2 Включение новых спейсеров в CRISPR-кассету.

1.1.4 Механизм работы

CRISPR-системы широко распространены среди прокариот и крайне разнообразны. Тем не менее, механизм действия CRISPR-систем различных видов прокариот имеет общие черты. CRISPR-системы часто называют CRISPR-Cas системами, так как для их работы необходимы сложные комплексы из Cas-белков. CRISPR-cas-система работает в три стадии:

1) адаптация, то есть приобретение новых спейсеров;

2) созревание эффекторных комплексов;

3) иммунный ответ — разрушение чужеродной ДНК или РНК.

1.1.4.1 Адаптация

В ходе адаптации при заражении клетки фрагмент генома вируса встраивается в кассету в качестве спейсера (**Рисунок 2**) [45]. При повторных заражениях тем же агентом данный спейсер повышает вероятность приобретения дополнительных спейсеров против этого агента. Этот эффект назван праймированием [49]. Увеличение числа спейсеров повышает шансы справиться с инфицирующим агентом, даже при неполном совпадении последовательностей спейсеров и их протоспейсеров.

Добавление новых спейсеров всегда происходит рядом с лидерной последовательностью. Было показано, что с ней взаимодействуют белковые комплексы, обеспечивающие включение новых спейсеров, в то время как её удаление приводит к невозможности вставки [51].

Специфических вирусных и плазмидных генов, служащих источником протоспейсеров, не обнаружено, тем не менее, это не совсем случайные последовательности. Как правило, протоспейсеры расположены рядом с короткими мотивами длиной 2-3 нуклеотида, т. н., PAM-последовательностями (от **P**rotospacer **A**djacent **M**otif) [52]. По-видимому, наличие PAM необходимо для связывания белков, выщепляющих протоспейсер. Кроме того, предполагают, что PAM могут служить маркерами, позволяющими CRISPR-системе отличить собственную ДНК от чужеродной [53].

Белки Cas1 и Cas2 выполняют ключевую роль на этапе адаптации [54]. В некоторых случаях, Cas1 и Cas2 даже слиты в единый белок. Экспериментально установлено, что белок Cas1 является эндонуклеазой и способен вносить разрывы в молекулы одноцепочечной ДНК, одноцепочечной РНК, а также двухцепочечной ДНК [55]. Cas1 вносит разрыв в первый повтор кассеты и, возможно, осуществляет предварительные манипуляции с будущей последовательностью спейсера.

Cas2 обладает эндорибонуклеазной активностью и может вносить внутренние разрывы в молекулы РНК [37]. Однако роль эндорибонуклеазной активности в процессе приобретения новых спейсеров не вполне очевидна. У ряда организмов (*T. thermophilis* и *B. halodurans*) Cas2 вносит разрывы в молекулы двухцепочечной ДНК с образованием фрагментов длиной около 120 пар оснований. Вероятно, Cas2 также участвует в предварительном процессинге последовательностей спейсеров и обмене CRISPR-локусами между разными организмами [37]. Общая последовательность молекулярных событий на стадии адаптации к новым вирусам такова:

- Белковый комплекс, включающий в себя Cas1 и Cas2, сканирует последовательность инфицирующей ДНК.
- После обнаружения PAM-мотива, происходит выщепление преспейсера — участка, содержащего протоспейсер, а также фрагмент PAM-мотива.
- Белковый комплекс распознает лидерную последовательность CRISPR-кассеты, и вставляет новый спейсер рядом с лидерной последовательностью.

Детали механизма включения новых спейсеров остаются неизученными, в частности, неизвестен механизм дупликации последовательности первого повтора.

1.1.4.2 Созревание эффекторных комплексов

В ходе защиты от вирусной инфекции происходит перевод пассивной иммунной памяти в активный ответ. Этот процесс состоит из нескольких этапов (**Рисунок 3**).

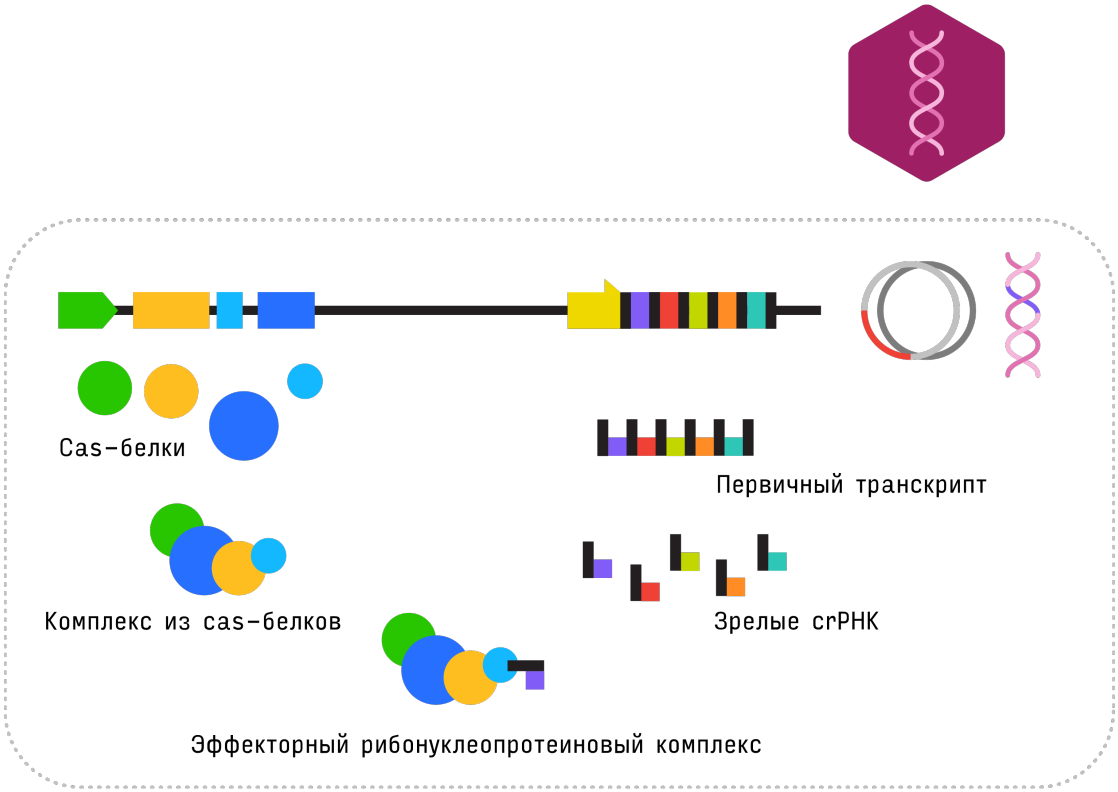


Рис.3. Созревание эффекторных комплексов.

На первом этапе происходит экспрессия CRISPR-кассеты с образованием длинной молекулы первичного РНК-предшественника. Предполагается, что запуск экспрессии происходит с использованием промоторной области в составе лидерной последовательности. Одновременно начинается экспрессия *cas*-генов, продукты которых впоследствии войдут в состав эффекторного комплекса. У *E.coli* эффекторный комплекс носит название Cascade [56] и состоит из пяти белков: Cse1, Cse2, Cas7, Cas5 и Cas6e в стехиометрическом соотношении 1:2:6:1:1. Собранный Cascade комплекс по форме напоминает морского конька (**Рисунок 4**). Похожие белковые комплексы, обнаружены также у ряда прокариот, включая филогенетически далёкие виды с другими типами CRISPR-систем [57].

Cas6e обладает эндорибонуклеазной активностью и разрезает длинный первичный РНК-предшественник, транскрибированный с CRISPR-кассеты. В результате работы Cas6e белка образуется более короткая РНК (crРНК) длиной 61 нуклеотид, соответствующая последовательности спейсера с двумя фланкирующими последовательностями разной длины (8 нт с 5’-конца и 21 нт с 3’-конца). Фланкирующие последовательности являются участками повторов [56].

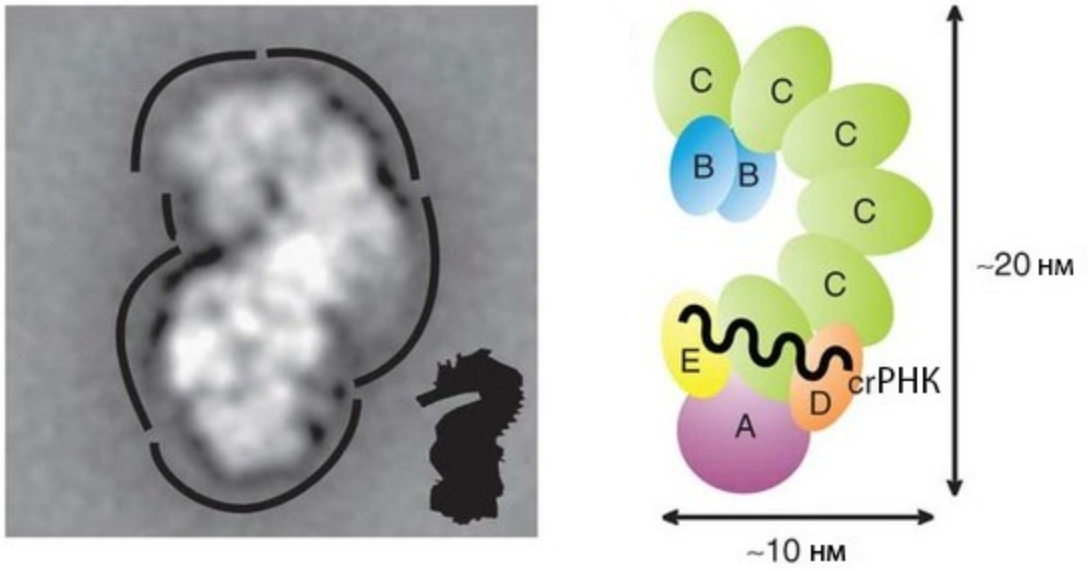


Рис.4. Структура Cascade-комплекса, адаптировано из [56].

При помощи рентгено-структурного анализа установлено, что скелет Cascade-комплекса образован шестью молекулами Cas7 (**Рисунок 4**). Комплекс содержит спиральную бороздку, в нее помещается зрелая молекула crРНК. 5’-конец crРНК непосредственно связан с Cse1 и/или Cas7 и/или Cas5-белками. Более длинный 3’-фланк за счёт частично палиндромной природы последовательности повтора формирует шпильку, которая помещается в выемку на поверхности Cas6e. Таким образом, процесс созревания эффекторных комплексов заканчивается формированием Cascade (или Cascade-подобного) комплекса, содержащего короткую crРНК.

1.1.4.3 Иммунный ответ

Экспериментально показано, что CRISPR-системы нейтрализуют фаговые инфекции, предотвращают трансформацию плазмидами, нарушают процесс лизогенизации, а также индукцию профагов [24]. Предполагалось, что CRISPR-системы эффективны только против чужеродной ДНК, позже были открыты CRISPR-Cas системы архей, направлено уничтожающие РНК-вирусы [58].

При выполнении иммунной функции CRISPR, происходит узнавание и уничтожение чужеродной ДНК (или РНК). CrРНК в составе эффекторного комплекса может комплементарно взаимодействовать со своим прототипом (протоспейсером), расположенным в вирусной или плазмидной ДНК. У *E.coli* белок Cas3 обладает эндонуклеазной активностью и разрезает молекулы одноцепопчечной и двухцепочечной ДНК, и обеспечивает деградацию чужеродной ДНК. Показано, что Cas3 эффективно разрезает ДНК, но при этом обладает

низкой специфичностью. Cascade-комплекс, напротив, благодаря комплементарному взаимодействию crРНК с последовательностью протоспейсера является высокоспецифичным. После узнавания последовательности протоспейсера Cascade комплекс рекрутирует белок Cas3 и ориентирует его правильным образом [58].

Молекулярный сценарий разрушения чужеродной ДНК при помощи CRISPR-Cas системы складывается из следующих этапов (**Рисунок 5**):

1. Рибонуклеопротеиновый комплекс (Cascade + crРНК) сканирует чужеродную ДНК в поисках PAM-мотивов и якорных областей [59]. Для узнавания протоспейсера необходим ряд взаимодействий: 1) PAM-мотив рядом с протоспейсером связывается неструктурированной петлёй белка Cse1; 2) 7 нуклеотидов 3’-конца протоспейсера комплементарно взаимодействуют с соответствующим участком crРНК. Такой регион спейсера/протоспейсера называют якорной областью. Если в этой области между спейсером и протоспейсером есть различия хотя бы в один нуклеотид, то CRISPR-опосредованный иммунитет не сработает [50]. Отличия между спейсером и протоспейсером за пределом якорной области допускаются, но они снижают эффективность CRISPR-иммунитета.

2. Cascade-комплекс, связанный с чужеродной ДНК-мишенью, меняет свою конформацию и рекрутирует белок Cas3.

3. Cas3 вносит одноцепочечные разрывы в двухцепочечную ДНК-мишень [59], одновременно с этим сродство Cascade комплекса к ДНК снижается. Через некоторое время комплекс распадается на субъединицы. Cas3 вносит большое число разрывов в чужеродную ДНК, начиная с места своей посадки и далее в более или менее случайной манере — до полного её уничтожения.

Фрагменты чужеродной ДНК могут быть впоследствии использованы в качестве новых спейсеров. Такая преемственность между циклами работы CRISPR-Cas систем по-видимому объясняет недавно обнаруженный эффект праймирования, т.е. ускоренного включения нескольких спейсеров против одного и того же чужеродного репликона [49].

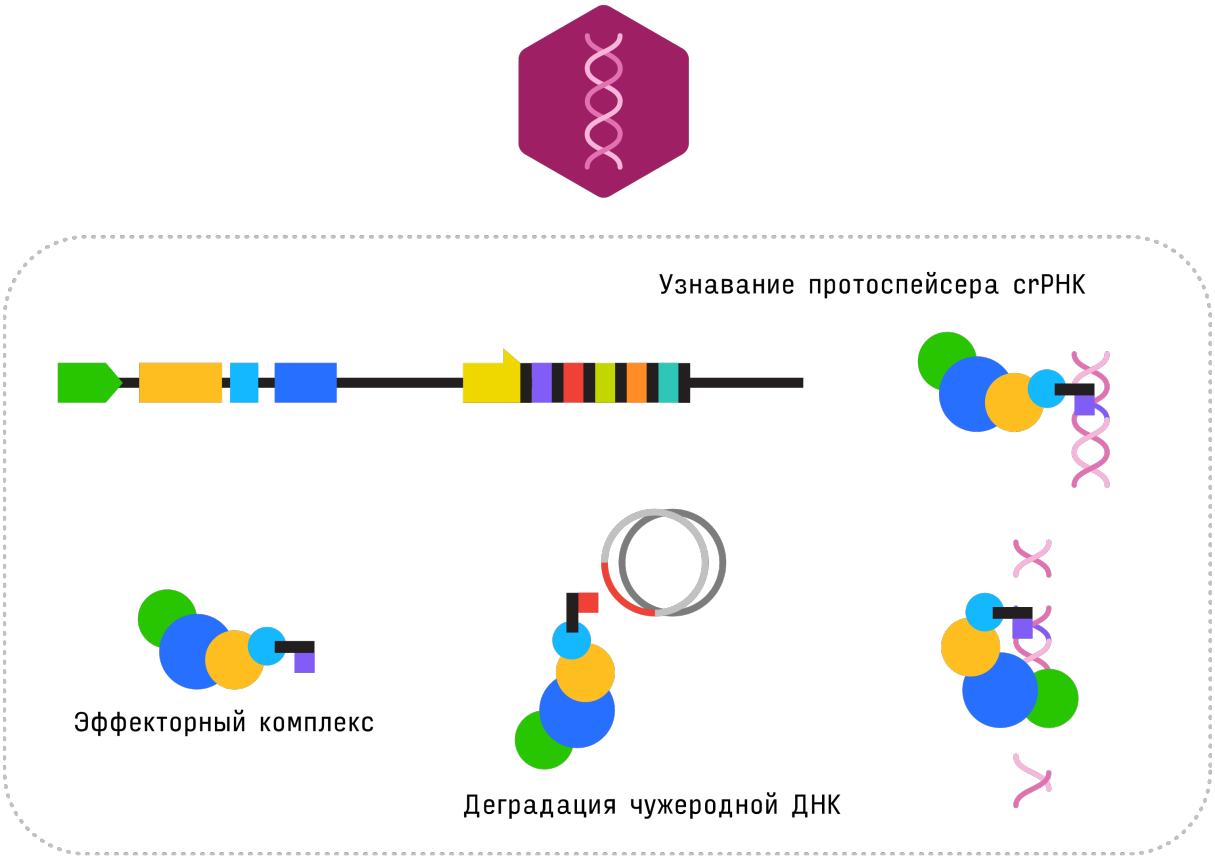


Рис.5. Деградация чужеродной ДНК эффекторными комплексами CRISPR-систем.

1.1.5 Типы CRISPR-Cas систем

Согласно современной классификации, CRISPR-Cas системы подразделяются на три основных типа (I, II, III) — по составу *cas*-локусов [36]. Разные наборы *cas*-генов обуславливают разные детали механизма CRISPR-опосредованного иммунитета.

Гены *cas1* и *cas2* являются универсальными, то есть присутствуют во всех предположительно активных CRISPR-Cas системах и составляют функциональный блок, задействованный в процессе интеграции новых спейсеров [60]. Гены *cas1* и *cas2* в большинстве случаев располагаются в непосредственной близости друг от друга. *Cas2* кодирует гомолог токсина — мРНК интерферазу [61]. Вероятно, изначально Cas1-Cas2 модуль функционировал по типу автономной токсин-антитоксин системы [61]–[63].

Функциональные блоки CRISPR-Cas систем, необходимые для адаптации и реализации иммунного ответа, очень разнообразны (**Рисунок 6**). Большинство CRISPR-Cas локусов можно однозначно отнести к одному из трех основных типов на основании характеристических тип- или подтип-специфичных генов. Тем не менее, существует ряд организмов, *cas*-локусы которых не вписываются в текущую классификацию, например, *Acidithiobacillus ferrooxidans* str. ATCC 23270; таким локусам присваивается тип «U» [36].

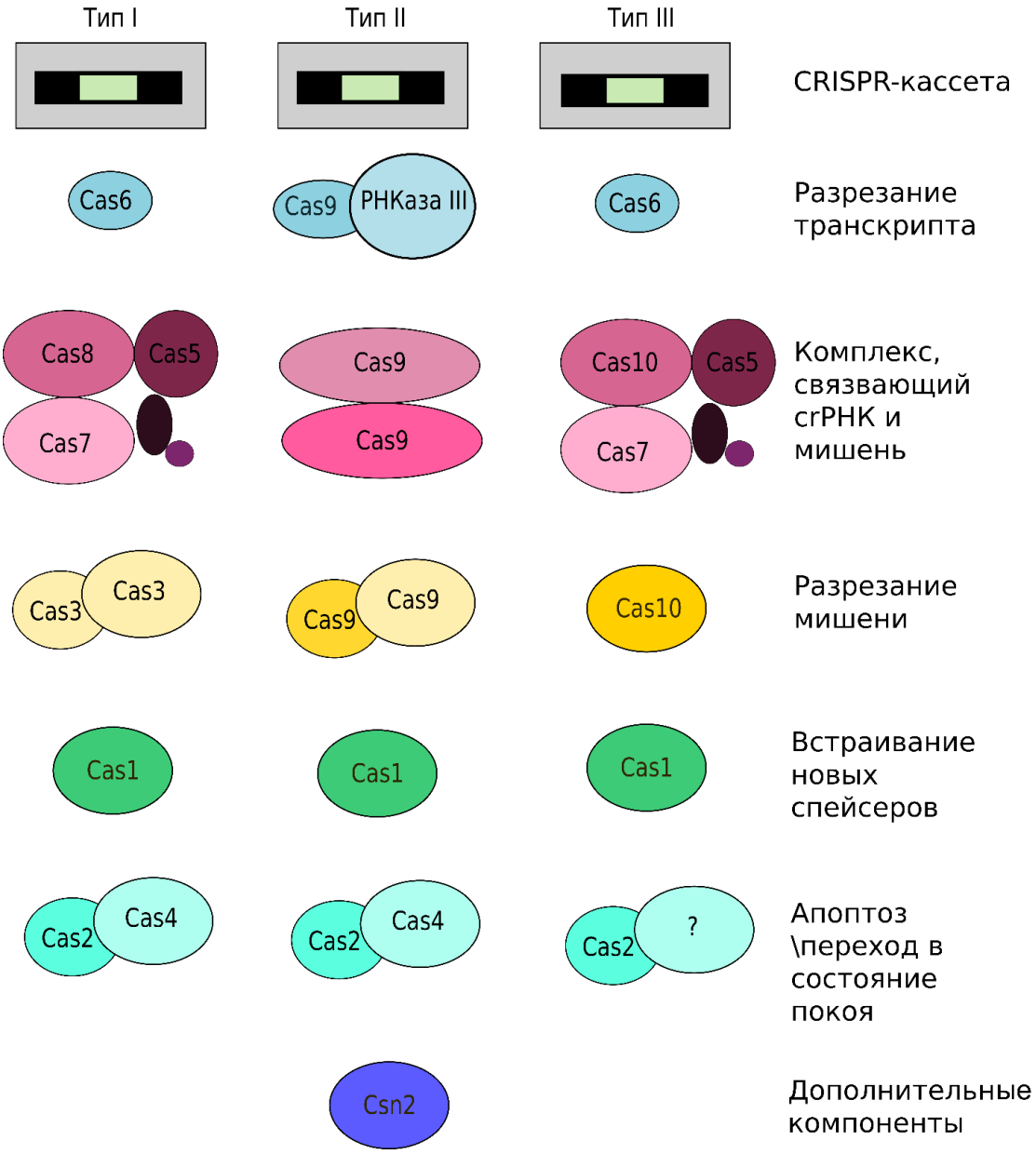


Рис.6. Функциональный состав *cas*-локусов CRISPR-Cas систем I, II и III типов.

Показано, что PAM-последовательности необходимы для работы CRISPR-Cas-систем I и II типов, системы III типа обходятся без них [23], [64]. Для разных типов и подтипов характерны свои PAM-последовательности. По-видимому, PAM эволюционно взаимосвязаны с типом cas1 и лидерной последовательностью [65].

Системы II типа были обнаружены исключительно в геномах бактерий, тогда как системы III типа в большей степени характерны для архей. Стоит отметить, что CRISPR-Cas системы чаще встречаются у архей, чем у бактерий [66], [67]. Часто геномы архей содержат несколько CRISPR-локусов различного типа.

1.1.5.1 CRISPR-Cas системы I типа

Типичный *cas*-локус I типа содержит ген *cas3*, кодирующий хеликазу [68], а также гены, кодирующие Cascade-подобные комплексы различного состава [69], [70]. Эффекторные комплексы построены из белков суперсемейства RAMP, а именно — семейств Cas5, Cas6 и Cas7 [39]**.** Помимо субъединиц, определяющих иммунную функцию, эффекторные комплексы I типа могут содержать крупные белки (такие как Cse1 и BH0338), а также маленькие белки, преимущественно состоящие из α-спиралей (Cse2).

Основным компонентом Cascade-подобного комплекса является эндорибонуклеаза RAMP суперсемейства, катализирующая превращение молекулы первичного РНК-предшественника в зрелые молекулы crРНК [69]. В большинстве случаев RAMP белки систем I типа (Cas6, Cas6e, Cas6f) не принадлежат к наиболее распространённым семействам Cas5 и Cas7, и зачастую закодированы на периферии соответствующих оперонов. Исключение составляют CRISPR-Cas локусы подтипа I-C, кодирующие Cas5 и Cas7 эндорибонуклеазы.

Мишенью для CRISPR-Cas систем I типа служит ДНК. HD-нуклеазный домен белка Cas3 катализирует рестрикцию чужеродной ДНК. В нескольких подтипах RecB нуклеазный домен белка Cas4 вместе с Cas1 образуют слитный белок. Возможно, Cas4 принимает участие в интеграции новых спейсеров [36].

1.1.5.2 CRISPR-Cas системы II типа

CRISPR-Cas системы II типа характерны для бактерий рода *Sterptococcus*, а также для *Neisseria meningitidis.* Характерным белком *cas*-локусов этого типа является белок Cas9, участвующий в созревании crРНК, а также разрезании чужеродной ДНК. Cas9 содержит, по меньшей мере, два нуклеазных домена: N-концевой RuvC-подобный домен и HNH (или McrA-подобный) нуклеазный домен — внутри белка. HNH-домены характерны для многих эндонуклеаз рестрикции [71].

Для CRISPR-Cas системы II типа *S. thermophilus* активность в отношении плазмидной и фаговой ДНК продемонстрирована *in vivo* [72]. Показано, что инактивация Cas9 нарушает интерференцию [45].

Системы II типа реализуют необычный механизм разрезания предшественника crРНК, в ходе которого формируется дуплекс между короткой трэйсерной РНК (tracРНК) и частью повтора внутри предшественника crРНК. Первый разрез при созревании crРНК приходится на область повтора. Эту реакцию катализирует белок домашнего хозяйства РНКаза III в присутствии белка Cas9 [22].

1.1.5.3 CRISPR-Cas системы III типа

CRISPR-Cas системы III типа содержат гены полимераз и RAMP белков, которые участвуют в созревании crРНК, аналогично комплексам типа Cascade, характерных для I-E подтипа. Системы III типа подразделяют на два основных подтипа: III-A и III-B. Мишенями для CRISPR-Cas систем III-A типа может служить плазмидная ДНК. Это продемонстрировано *in vivo* на примере *S. epidermidis* [23]. Для CRISPR-Cas систем III-B подтипа *P. furiousus* продемонстрирована активность в отношении РНК-мишеней [58].

Эндорибонуклеаза Cas6 является маркерным белком всех *cas*-локусов III типа. Многие cas-опероны III типа не содержат генов *cas1-cas2*. Во всех подобных случаях в геноме присутствует дополнительный CRISPR-локус (I или II типа), содержащий данные гены. Другие системы III типа, напротив, содержат гены *cas1* и *cas2*, организованные в один оперон вместе с RAMP-белками. Локусы такой структуры типичны для *S.epidermidis* и *Mycobacterium tuberculosis* (III-A подтип), *Halorhodospira halophils* (III-B подтип). У этих организмов в геноме присутствуют только такие *cas*-локусы III типа [36].

1.1.6 Аутоиммунитет

Довольно часто находят сходство между последовательностями спейсеров с участками бактериальных геномов, не несущих CRISPR-кассет. Среди всех известных спейсеров доля спейсеров с протоспейсерами в клеточной ДНК составляет всего ~0.4%. Тем не менее, у ~18% организмов, имеющих CRISPR-систему, содержится по меньшей мере одна кассета с хотя бы одним таким спейсером. Такие спейсеры широко распространены в различных филогенетических линиях микроорганизмов [53].

Наличие спейсеров к геномной ДНК хозяина может представлять опасность для клетки, так как заставляет CRISPR-систему работать по аутоиммунному сценарию, то есть разрушать собственную ДНК. Такие спейсеры могут появляться в результате ошибок работы белковых комплексов, обеспечивающих добавление новых спейсеров. Предполагают, что CRISPR-система может отличать свою ДНК от чужеродной благодаря PAM-последовательностям: протоспейсеры в вирусной (или плазмидной) ДНК расположены рядом с PAM-мотивами, в геномной ДНК клетки в норме PAM-мотивов содержаться не должно [23].

Большая часть спейсеров с протоспейсерам в собственном геноме расположена в начале кассеты (1-2 позиции после лидерной последовательности)[53]. Согласно принципу построения CRISPR-кассет, расположение спейсеров в начале свидетельствует о том, что они были добавлены недавно. Возможно, такие аутоиммуные спейсеры имеют очень короткое

время жизни: кратковременно оказав негативное воздействие на клетку, они столь же быстро элиминируются из состава кассеты.

При попадании в кассету аутоиммунных спейсеров, CRISPR-система должна быть быстро инактивирована. Например, у *Lactobacillus acidophilus* NCFM обнаружен спейсер против гена 16S рРНК [53]. Такой спейсер в составе активной CRISPR-кассеты означает немедленную гибель для клетки. Примечательно, что у *L.acidophilicus* практически полностью отсутствуют *cas*-гены, то есть CRISPR-система, судя по всему, нефункциональна [5]Возможно, что негативные эффекты, обусловленные аутоиммунными спейсерами, объясняют широкое распространение очень деградированных CRISPR-систем.

1.2 Микробиом человека

Человек — это суперорганизм и конгломерат. Число клеток прокариот, постоянно ассоциированных с телом человека превышает число собственных эукариотических клеток по меньшей мере в десять раз [73] Всю совокупность комменсальных, патогенных и симбиотических микроорганизмов, местообитанием которых служит тело человека, называют микробиомом. Эта концепция была впервые сформулирована Ледербергом [5]. Отдельные исследователи присуждают микробиому статус «нового» или «забытого» органа [74], так как его существование и значение недооценивали вплоть до 1990-х г.г. Сейчас становится понятно, что микробиом вносит существенный вклад в функционирование и поддержание здорового состояния организма человека.

Микробиом человека синтезирует витамины [75], участвует в метаболизме сложных полисахаридов [76], контролирует деление клеток эпителия кишечника посредством синтеза короткоцепочечных жирных кислот [77], а также оказывает влияние на формирование и развитие иммунной системы [78], [79] Несмотря на важность микробиома для поддержания здоровья человека, долгое время он оставался плохо исследованным, так как большая часть микроорганизмов, ассоциированных с телом человека плохо культивируется в стандартных лабораторных условиях [5]. Изучить истинное видовое разнообразие микробиома стало возможно только с развитием метагеномного подхода и техник высокопроизводительного секвенирования. Метагеномика предполагает изучение генетического материала, полученного напрямую из образцов окружающей среды [6]. Эти методы позволили выявить роль микробиома в развитии различных физиологических состояний, таких как ожирение [80], [81], воспалительные заболевания кишечника [82] и появление камней в почках [83]. Наиболее крупномасштабные исследования микробиома ведутся в рамках метагеномных

проектов MetaHIT (изучение микробиома кишечника [84] и Human Microbiome Project (проект «Микробиом Человека» [5]).

1.2.1 Видовой состав микробиома человека. Энтеротипы

Видовое разнообразие в пределах конкретного местообитания определяется числом и степенью представленности различных групп организмов. Видовое разнообразие микробиома человека напрямую ассоциировано с развитием ряда заболеваний и патологических состояний человека. С низким микробным разнообразием кишечника связывают возникновение воспалительных заболеваний кишечника и склонность к полноте [84], [85], а повышенное микробное разнообразие часто характерно для развития бактериального вагиноза [86].

Оценка истинного видового разнообразия микробиома человека является нетривиальной задачей. По данным проекта HMP, микробиом человека содержит от 3’500 до 35’000 операционных таксономических единиц (OTU), в зависимости от выбранных параметров [9] что соответствует примерно 600 различным родам. Обнаружено только несколько родов, характерных почти для всех (>95%) исследованных индивидов [87]. Часто представители таких универсальных родов доминируют в пределах определённых локусов тела человека. Помимо небольшого числа универсальных родов, обнаружено большое число очень разнообразных специфических минорных таксонов [9].

В ходе проекта 'Микробиом человека' обнаружен ряд неизвестных ранее родов, а также, предположительно, новое семейство внутри порядка *Clostridiales* [73] Представители новых таксонов не очень многочисленны (<2%), тем не менее присутствуют в составе микробиомов большого числа индивидов [9]. Большая часть новых видов принадлежит плохо описанному роду *Barnesiella*, а также выделенным новым родам *Dorea*, *Oscillibacter* и *Desulfovibrio.* Эти таксоны ассоциированы с развитием рака прямой кишки и ряда оппортунистических инфекций [73]. Минорные таксоны могут оказывать влияние на общее состояние организма, имея даже небольшое представительство в микробиоме [88].

Тело человека не является однородным местообитанием, в его пределах можно выделить ряд локальных экологических ниш, существенно различающихся по физико-химическим параметрам. Эти различия влекут за собой и различия в составе микробных сообществ, их населяющих. Наибольшее микробное разнообразие характерно для желудочно-кишечного тракта: приблизительно 150 родов[10]. Большая часть обнаруженных видов, ассоциированных с кишечником человека, принадлежит двум типам: *Bacteroidetes* и *Firmicutes*, их соотношение отличается между индивидами. Видовой состав кишечного

микробиома довольно сильно различается у отдельных индивидов. Предположено, что кишечные микробиомы можно подразделить на несколько энтеротипов. Под энтеротипом понимают сходные дискретные стабильные сообщества микроорганизмов [89]. Гипотеза энтеротипов послужила предметом активных дискуссий, изучалась стабильность энтеротипов во времени и их взаимосвязь с диетой [90]. Сейчас большинство микробиомных данных говорит скорее в пользу существования градиента видового состава сообществ, чем дискретных энтеротипов [91]. Изначально, на выборке, содержащей представителей шести различных наций, выделено три основных энтеротипа, в зависимости от преобладания представителей родов *Prevotella*, *Ruminococcus* или *Bacteroidetes* [89]. Впоследствии подтверждено существование стабильного энтеротипа «Prevotella». Типы «Ruminococcus» и «Bacteroidetes», по-видимому, не являются дискретными и соответствуют крайним точкам в континууме сообществ от тех, в которых с преобладают *Bacteroidetes*, до сообществ, в которых доминируют представители типа *Firmicutes* [92].

В ходе исследования микробиомов людей с контролируемым рационом питания обнаружили взаимосвязь между энтеротипом «Prevotella» и диетой богатой углеводами, а энтеротип «Bacteroidetes» более характерен для людей преимущественно потребляющих белки и жиры животного происхождения. Кратковременные переходы с одного типа диеты на другой приводят к небольшим изменениям видового разнообразия, но в целом видовой состав микробиомов конкретных индивидов относительно стабилен [90].

1.2.2 Вариации видового состава микробиома человека

Микробиомы людей очень разнообразны, тем не менее существуют характерные типы сообществ, специфичные для детей и пожилых людей. Также наблюдаются различия в составе микробиомов географически удалённых популяций [91].

Наибольший интерес представляет кишечный микробиом грудных детей, так как в течение первого и нескольких последующих лет жизни происходит становление приобретённого иммунитета, а также происходит глобальное изменение рациона питания [93]. Микроорганизмы колонизируют кишечник новорожденных непосредственно в момент рождения, а состав первичной микрофлоры зависит от способа родов [15]. Индивидуальные различия между микробиомами детей как правило гораздо более выражены, чем у взрослых. Для микробиомов новорожденных характерно отчётливое доминирование рода *Bifidobacterium*, обусловленное диетой первых месяцев жизни [91].

Микробиом детей постепенно стабилизируется и начинает напоминать микробиом взрослых в концу второго-третьего года жизни [93]. С функциональной точки зрения, переход

от микробиома новорожденных к микробиому взрослых сопряжён с увеличением доли микроорганизмов, осуществляющих анаэробное брожение, расщепление сложных углеводов и одновременным сокращением доли микроорганизмов, осуществляющих транспорт и метаболизм простых сахаров (лактозы, глюкозы, сахарозы) [93]. С возрастом происходит постепенное снижение видового разнообразия, поэтому состав микробиомов пожилых людей становится нестабильным. Снижение разнообразия сопровождается увеличением доли Bacteriodetes и уменьшением доли Firmicutes. Это приводит к снижению продукцию глутарата и короткоцепочечных жирных кислот, обладающих противовоспалительным действием [94].

1.2.3 Функциональное содержание микробиома человека

Несмотря на значительные индивидуальные различия видового состава микробиомов, метаболические пути, ассоциированные с микробиомом, стабильны и универсальны [9], [84], [85], [95]. Как правило, видовой состав разных локальных мест обитания в пределах тела человека (кишечника, ротовой полости и т. д.) сильно варьирует между разными индивидами. Тем не менее, для каждой субнишы характерны два относительно стабильных универсальных набора метаболических путей. Первый набор содержит гены домашнего хозяйства, обеспечивающие набор минимальных функций прокариот: основной метаболизм нуклеиновых кислот и синтез белка. Второй набор содержит специфические функции, необходимые для совместной жизни с человеком: синтез незаменимых для человека витаминов и короткоцепочечных жирных кислот [84].

Определённые наборы функциональных генов в пределах каждого локального местообитания остаются стабильными в независимости от вариаций видового состава. Так, например, кишечный микробиом всегда обогащён генами, связанными с деградацией сложных углеводов, независимо от соотношения таксономических групп *Bacteroidetes* и *Firmicutes*. Для этого местообитания характерны гены, необходимые для продукции сероводорода и деградации метионина [9].

Примечательно, что для каждой локальной микрониши в пределах микробиома человека характерны свои наборы ферментов, утилизирующих сахара. Например, микробиом ротовой полости оптимизирован для первичного метаболизма сахаров, в особенности расщепления крахмала. И, наоборот, вагинальный микробиом обогащён генами, отвечающими за метаболизм гликогена и деградацию пептидогликанов [96].

Функциональный состав индивидуальных микробимов человека подвержен вариациям, так же как и видовой состав. Основные различия наблюдаются в составе слабо

представленных генов и метаболических путей, которые, как правило, персонализированы. Большая часть низко представленных генов микробиома человека плохо аннотирована и не ассоциирована с конкретной метаболической функцией, особенно много таких генов в кишечном микробиоме [9].

Вариации видового и функционального состава могут отражать индивидуальные особенности архитектуры микробных сообществ и специфические метаболические пути, необходимые для поддержания их целостности [9]. Сообщества, в свою очередь, формируются в контексте уникальных для каждого индивида рациона питания и статуса иммунной системы.

1.2.4 Исследования CRISPR-систем микробиомов человека

Многие из представителей микробиома человека плохо изучены, а около 60% микробного разнообразия не культивируется в лабораторных условиях [5]. Только недавно, благодаря усилиям проекта 'Микробиом человека', стали появляться полные геномы многих плохо изученных, но важных для здоровья человека микроорганизмов [97]. На настоящий момент для изучения состава и динамики сложных сообществ микроорганизмов с большим успехом используется метагеномный подход. Метагеномика позволяет получить полный «снимок» сосуществующих микроорганизмов в пределах одной экологической ниши: прокариот и их вирусов. Исследование CRISPR-систем метагеномных данных является эффективным подходом для изучения динамики таких эволюционных взаимоотношений.

CRISPR-системы проанализированы в метагеномах природных сообществ, населяющих кислые шахтные воды [98], горячие источники национального парка Йеллоустоун [99], гипер-солёное озеро Тиррелл в Австралии [100], мировой океан [101] и рубец жвачных животных [102].

Получен ряд метагеномов индивидуальных микробиомов человека. Пристальное внимание уделяется изучению микробных и вирусных сообществ кишечника [76], [103]–[108]. Микробиом человека подразделяется на ряд локальных ниш, которые заселены специфическими сообществами микроорганизмов. CRISPR-кассеты были описаны для разных локальных экологических ниш в рамках нескольких независимых метагеномных проектов [109]–[111], а также для данных проекта «Микробиом Человека» [12], [14].В упомянутых работах, из всего объёма сырых метагеномных данных отбирали последовательности, содержащие известные CRISPR-повторы. После этого, изучали спейсеры, заключённые между двумя идентифицированными повторами [12], [14]. Такой

подход позволяет идентифицировать большое число спейсеров, хотя и ограничен кассетами с уже известными последовательностями повторов.

Помимо поиска CRISPR-кассет, метагеномные данные проекта 'Микробиом человека' использованы для идентификации неизвестных ранее cas-генов [112]. В результате анализа генного окружения CRISPR-кассет и сравнения с известными *cas*-генами, было выделено 24 новых *cas* семейства. Таким образом, анализ метагеномных данных микробиома человека позволяет расширить спектр известных компонентов CRISPR-Cas систем.

Глава 2. Данные и алгоритмы

2.1 Метагеномные данные

2.1.1 Микробиомы человека

Для исследования CRISPR-систем использованы данные трёх метагеномных коллекций микробиома человека, доступных на момент начала исследования:

1) проект «Микробиом человека», **HMP** (**H**uman **M**icrobiome **P**roject) [10];

2) метагеном 13 здоровых японцев, **JPN** (Healthy Human Gut Metagenomes) [11];

3) биом нисходящей ободочной кишки человека, **DG** (Human **D**istal **G**ut biome project) [8].

Метагеномные данные кишечных образцов **HMP** загружены в виде сборки, состоящей из 1'889'651 контигов (http://public.genomics.org.cn/BGI/gutmeta/UniSet/). Суммарная длина контигов составила 3'732 мегабаз (Мб). Образцы были собраны у 124 взрослых европейцев (Германия, Дания, Испания) разного возраста (от 18 до 69 лет) и отсеквенированы на платформе Illumina GA [10].

Контиги метагенома **JPN** загружены с веб-сайта CAMERA (ftp://portal.camera.calit2.net/ftplinks/cam_datasets/projects/assemblies/CAM_PROJ_HumanGut.asm.fa.gz). Этот набор данных содержал 353'805 контигов, суммарная длина которых составила 463 Мб. Образцы были собраны у взрослых и детей, в том числе грудных младенцев: возрастной диапазон индивидов — от 6 месяцев до 45 лет. Выборка состояла из членов двух семей (3 и 4 человека, соответственно) и шести независимых индивидов. Метагеном был отсеквенирован по Сэнгеру на платформе MegaBACE4500 sequencer (GE Healthcare) [11]

Метагеномные контиги **DG** загружены из архива базы данных NCBI (http://www.ncbi.nlm.nih.gov/Traces/wgs/?val=AAQK01). Сборка состояла из 22'508 контигов, суммарной длиной 336 Мб. Образцы были собраны у двух здоровых взрослых добровольцев. Последовательности получены на платформе ABI 3730xl DNA analyzer [8].

2.1.2 Виромы человека

Для поиска протоспейсеров использовали данные двух виромов человека. Виром преимущественно некультивируемых форм вирусов (РНК-вирусов) (Uncultured Human Fecal Virus Metagenome) кишечника человека загружен из базы данных CAMERA (ftp://portal.camera.calit2.net/ftp-links/cam_datasets/projects/read/CAM_PROJ_HFVirus.read.fa.gz). Общее число последовательностей составило 36'769. Данный метагеном получен на кишечных образцах двух взрослых здоровых добровольцев из Сан-Диего.

Последовательности отсеквенированы на платформе BI3730 DNA analyzer (Applied Biosystems) [114].

Виром кишечника человека (Virome of human gut) загружен из архива базы данных NCBI (http://www.ncbi.nlm.nih.gov/sra?term=Cafe%20and%20FSM%3A%20Virome%20of%20human%20gut). Данный метагеном содержит последовательности вирусов и вирусоподобных частиц, отсеквенированные на платформе 454 Life Sciences (Roche) GS FLX Titanium. Общий размер метагенома составил 936'213 последовательностей со средней длиной 359 нуклеотидов и суммарной длиной 336 Мб [115].

Основные параметры исследованных метагеномов суммированы в **Таблице 1**.

Таблица 1. Характеристика проанализированных наборов метагеномных данных

Метагеномный проект	**Число контигов/ чтений**	**Общая длина**	**Источник образцов**	**Выборка индивидов**	**Платформа для секвениро-вания**	**Программа для сборки**
Проект микробиом человека (The Human Microbiome Project, **HMP**)	1'889'651 контиг	3'732 Мб	Кишечные образцы	124 европейца; возраст от 18 до 69 лет	Illumina GA	MetaMos
Микробиом кишечника здоровых японцев (Healthy Human Gut Metagenomes, **JPN**	353'805 контигов	463 Мб	Кишечные образцы	13 японцев; возраст от 6 месяцев до 45 лет. Две семьи (3 и 4 члена) и 6 независимых индивидов	Mega BACE4500 sequencer (GE Healthcare)	PCAP
Биом нисходящей ободочной кишки человека (Distal gut metagenomic project, **DG**)	22'508 контигов	336 Мб	Кишечные образцы	2 здоровых добровольца	ABI 3730x1 DNA analyzer	Celera Assembler
РНК-виром кишечника человека (Uncultured Human Fecal Virus Metagenome)	36'769 чтений	47 Мб	Кишечные образцы	2 здоровых добровольца, Сан Диего	ABI 3730 DNA analyzer	PHRAP
Виром кишечника человека (Virome of human gut)	636'213 чтений	336 Мб	Кишечные образцы	9 доброволь-цев, диета с низким или высоким содержанием жира	454 Life Sciences (Roche) GS FLX Titanium	Newbler

2.2 Идентификация и анализ CRISPR-кассет

2.2.1 Идентификация CRISPR-кассет, процедура фильтрации

Для построения надежного набора CRISPR-кассет, для каждой коллекции метагеномных данных, мы использовали три алгоритма: PILER-CR (PIL) [116], CRISPR Recognition Tool (CRT) [117] и CRISPRFinder (CFI) [118] и разработанную ранее процедуру фильтрации [101] (**Рисунок 7**). Фильтрация необходима по двум причинам. Во-первых, из-за большого объёма и высокой степени фрагментации метагеномных данных стандартные алгоритмы поиска кассет зачастую дают ложноположительные результаты. Во-вторых, результаты работы разных алгоритмов часто не совпадают из-за различий в способах поиска повторов и определения границ кассет.

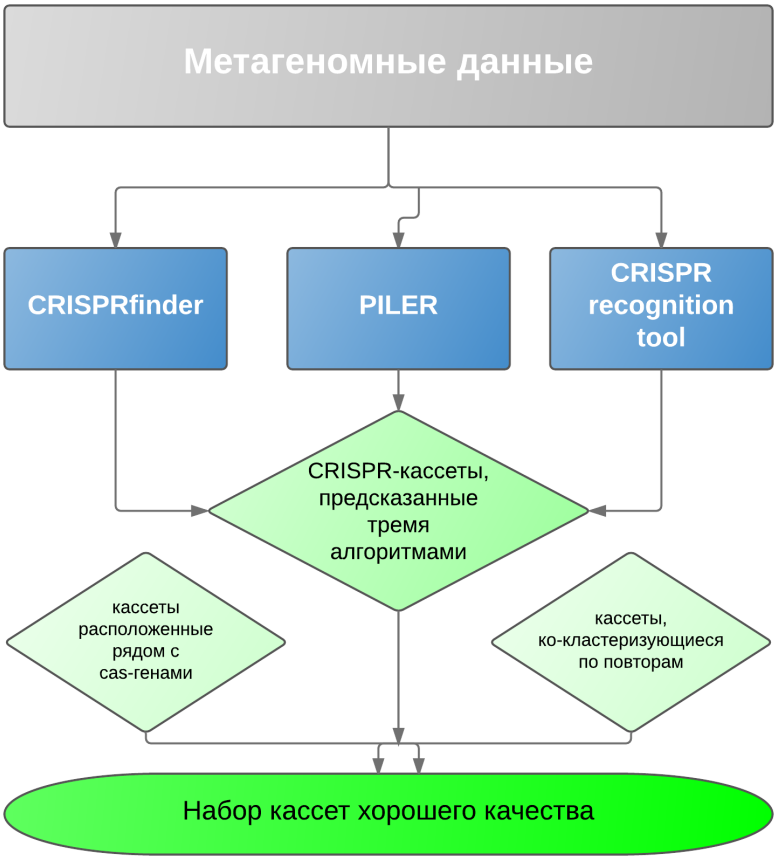


Рис.7. Схема процедуры фильтрации.

Алгоритм **PILER-CR** строит локальные выравнивания заданной последовательности самой на себя. Каждое короткое совпадение (хит) двух расположенных рядом участков считается выравниванием двух соседних повторов. Алгоритм ищет дальнейшие совпадения только в пределах главной диагонали матрицы выравнивания (dot plot), действуя согласно

парадигме динамического программирования и учитывая детерминированную структуру кассет. После идентификации первичных границ повторов, они последовательно (итеративно) уточняются [116].

Алгоритм **CRT** ищет повторы непосредственно в заданной последовательности без использования выравниваний. Алгоритм ориентирован на поиск серий коротких повторов заданной длины. Он ищет точные вхождения k-меров — последовательностей из k нуклеотидов; затем достраивает k-меры, пока позволяет заданный порог на число замен между повторами кассеты [117].

Алгоритм **CRISPRFinder** ищет повторы, используется суффиксное дерево. После нахождения первичных повторов, их границы последовательно уточняются [118].

Программы CRT и PILER-CR загружены с официальных сайтов и применялись с параметрами по умолчанию. Алгоритм CRISPRFinder изначально был доступен только в качестве веб-сервиса (http://crispr.u-psud.fr/Server/), поэтому написана программа, позволяющая автоматически загружать на сервер большое количество fasta-файлов с метагеномными контигами, запускать алгоритм предсказания кассет с параметрами по умолчанию, а затем собирать и обрабатывать текстовую выдачу.

Кассеты, полученные в результате работы каждого алгоритма, считали кандидатными. Списки кандидатных кассет сравнивали. Надежно предсказанными считались те кассеты, которые были предсказанные всеми тремя алгоритмами. Кассеты, предсказанные менее, чем тремя алгоритмами, могли быть добавлены к надежному списку кассет, если удовлетворяли одному из дополнительных требований:

1) последовательности, фланкирующие кассету, содержали *cas*-гены;

2) кассеты содержали повторы, совпадающие с повторами кассет, уже участвующих в списке надежно предсказанных, или ко-кластеризующиеся с ними.

Помимо стандартных алгоритмов PILER-CR, CRT и CRISPRFinder мы попробовали применить недавно опубликованный алгоритм Crass [119] с параметрами по умолчанию для сборки CRISPR-кассет из метагеномных чтений. Результаты сборки кассет для метагенома биома нисходящей ободочной кишки человека (DG) были, в целом, сопоставимы с результатами процедуры фильтрации, описанной выше. При помощи алгоритма Crass не удалось собрать ни одной CRISPR-кассеты из метагеномных чтений HMP. После снижения порогов на число повторов и длину повторов и спейсеров (-n 2 -w 6 -s 20 -S 55) удалось собрать только одну CRISPR-кассету на метагеномных данных HMP. Так как чтения метагеномной коллекции JPN не были доступны, ассемблер Crass в данном случае неприменим. Из соображений единообразия предсказаний, результаты работы алгоритма Crass далее не учитывали.

2.2.2 Предсказание *cas*-генов

Для идентификации *cas*-генов, применяли поиск с помощью программы blastx (http://blast.ncbi.nlm.nih.gov/Blast.cgi) [120] для последовательностей, фланкирующих CRISPR-кассеты, против невырожденной коллекции белковых последовательностей (NR) базы данных GenBank [121] с порогом на e-value 0.01. Текстовую выдачу программы обрабатывали автоматически, а затем выбирали хиты, содержащие в полях описания ключевые слова «cas» и/или «crispr». Отобранные хиты далее оценивали вручную.

2.2.3 Определение таксономии контигов, содержащих CRISPR-кассеты

Для последовательностей, фланкирующих CRISPR-кассеты, провели blastx поиск (http://blast.ncbi.nlm.nih.gov/Blast.cgi) против невырожденной коллекции белковых последовательностей (NR) базы данных GenBank с порогом на e-value $1e^{-6}$. Таксономические группы приписывали вручную на основании степени согласованности таксономического положения лучших хитов. Таксономическую группу на уровне типа присуждался контигу, если, по меньшей мере, десять верхних хитов принадлежали к одному и тому же типу. Таксономические группы на уровне класса, семейства и рода присуждались в случае, если 30 верхних хитов принадлежали к таксону этого уровня. Если таксономическое положение верхних хитов различалось, контигу присуждалась неспецифическая таксономическая группа (например, «Бактерии»).

Таксономия контигов могла быть не определена в нескольких случаях:

1) CRISPR-кассета занимает всю длину контига;

2) CRISPR-кассета фланкирована участками, содержащими только универсальные *cas*-гены (эти гены являются частой мишенью для горизонтальных переносов между геномами прокариот, поэтому их филогения может не соответствовать таксономии [35]);

3) последовательности, фланкирующие кассету, не содержат генов или содержат гены без значимого сходства хотя бы с одной из последовательностей невырожденной белковой коллекции GenBank.

2.2.4 Определение происхождения спейсеров (поиск протоспейсеров)

Для определения происхождения спейсеров (т.е. поиска протоспейсеров) мы применили blastn поиск против данных трёх типов. Во-первых, мы сравнили последовательности спейсеров со всеми известными вирусными последовательностями, в том числе последовательностями полных вирусных геномов базы данных GenBank. Во-вторых, мы

сравнили наборы спейсеров с собственно метагеномными данными микробиомов человека, полагая, что эти данные могут содержать последовательности фагового, профагового или плазмидного происхождения, даже после фильтрации от малых частиц (согласно протоколу выделения метагеномной ДНК, [122]. В-третьих, мы сравнили наборы спейсеров с последовательностями двух доступных виромов кишечника человека. Выравнивания между последовательностями спейсеров и протоспейсеров, как правило, очень короткие (в среднем — 30 нт), и, зачастую, пары спейсер-протоспейсер, различающиеся по внутренним позициям, могут быть выравнены алгоритмом blastn только частично. Чтобы избежать потери таких выравниваний и, соответственно, кандидатных протоспейсеров, все полученные хиты подвергали отдельной обработке. Если выравнивание оказывалось короче, чем исходная последовательность спейсера, недостающие фланкирующие участки достраивали с одного или обоих концов так, чтобы полученная последовательность соответствовала полноразмерной последовательности спейсера. Для полученных полноразмерных выравниваний между спейсером и кандидатным протоспейсером подсчитывали число замен. Протоспейсерами считали кандидатные последовательности, имеющие не более четырех замен по сравнению с соответствующим спейсером.

Для проверки того, что протоспейсеры не являются спейсерами неидентифицированной CRISPR-кассеты, проводили параллельный blastn поиск для последовательностей повторов соответствующих кассет против тех же наборов данных.

Таксономическое положение контигов, содержащих протоспейсеры, определяли согласно процедуре описанной ранее (см. Определение таксономии контигов, содержащих CRISPR-кассеты). Таксономическую группу контига затем переносили на спейсер. Если протоспейсер имел фаговое или плазмидное происхождение, использовали информацию о соответствующем организме-хозяине. Если протоспейсер был обнаружен в последовательности бактериального происхождения, таксономическое положение контига определяли, как описано ранее. В случае, если спейсер имел несколько таксономических групп, их сравнивали.

Для оценки значимости сходства между спейсерами и соответствующими базами данных последовательностей (вирусные последовательности GenBank и метагеномы микробиомов), строили наборы случайных «псевдоспейсеров»: каждый спейсер был заменён случайным фрагментом той же длины, предпочтительно из того же самого контига. Случайные последовательности выбирались только из фрагментов, не содержащих CRISPR-кассет. Если кассета занимала почти весь контиг, т. е., обе фланкирующих последовательности были короче 100 нт, выбирался фрагмент той же длины что и спейсер из случайно выбранного

контига, принадлежащего тому же индивидуальному метагеному, но не содержащего предсказанных CRISPR-кассет.

Описанная процедура не вполне застрахована от получения ложных результатов в силу гомологии генов, т. е. значимого сходства последовательностей, затрагивающего не только область между спейсером и протоспейсером, но и прилегающие области. Чтобы исключить ошибки такого рода провели дополнительную проверку. Для каждого псевдоспейсера извлекали фланкирующие последовательности, длина которых совпадала с длиной повтора из настоящей кассеты. Следуя введенной терминологии, будем называть такие последовательности псевдоповторами. Для полученного набора псевдоповторов проводили blastn поиск против тех же наборов данных, что использовали для псевдоспейсеров. Пару псевдоспейсер-псевдопротоспейсер принимали в расчет, только если ни один из псевдоповторов не имел сходства с той же последовательностью, что и псевдоспейсер.

2.2.5 Построение кластеров повторов

Консенсусные повторы кассет кластеризовали с помощью стандартной процедуры blastclust с параметрами L 0.5 -S 50 -e F -p F -W 15, согласно рекомендациям разработчиков (http://www.ncbi.nih.gov/IEB/ToolBox/C_DOC/lxr/source/doc/blast/blastclust.html). Выдачу программы обрабатывали, и кластеры, состоящие из более, чем одной последовательности, рассматривали далее. Последовательности каждого кластера выравнивали при помощи программы MUSCLE [123] с параметрами по умолчанию. В дальнейшем, повторы считали похожими, если они принадлежали к одному и тому же кластеру.

2.2.6 Поиск PAM-последовательностей

Для поиска PAM-последовательностей (**p**rotospacer **a**djacent **m**otifs – короткая последовательность, расположенная рядом с протоспейсером) анализировали участки, фланкирующие протоспейсеры длиной 10 нт, с обеих сторон [52].

2.2.7 Определение ориентации CRISPR-кассет

В случаях, когда это было возможно, ориентацию CRISPR-кассет определяли на основании положения и направления транскрипции ассоциированного локуса *cas*-генов. Конец кассеты считался лидерным, если прилежащая последовательность содержала *cas*-ген в нужной ориентации. Кассеты, фланкированные короткими последовательностями (менее 100 нуклеотидов), ориентировали в соответствии с направлением кассет с похожими

повторами, ориентацию которых уже удалось определить по ассоциированным *cas*-генам. Исходили из положения, что кассеты, повторы которых принадлежат одному кластеру, должны быть закодированы на одной и той же цепи.

Определить расположение лидерного и терминального конца было невозможно для кассет, повторы которых не ко-кластеризовались с повторами кассет с известной ориентацией. При дальнейшем анализе свойств кассет, зависящих от направления, такие кассеты не учитывали.

2.2.8 Определение сдвига спейсеров в кассете

Спейсерами с мишенями называли спейсеры, имеющие, по меньшей мере, один достоверный протоспейсер в том же самом индивидуальном метагеноме. *Общими спейсерами* называли спейсеры, обнаруженные в двух или более индивидуальных метагеномах.

Для оценки значимости смещения спейсеров с мишенями к лидерному концу CRISPR-кассет, а общих спейсеров — к дистальному концу, применяли симуляцию по методу Монте-Карло. Рассматривали только полные кассеты (т. е. фланкированные последовательностями не короче 50 нт) с установленной ориентацией. Спейсеры в каждой кассете получали порядковые номера, начиная со спейсера, рядом с лидерным концом. Для каждой кассеты порядковые номера всех спейсеров с мишенями суммировали. Полученные таким образом статистики суммировали для всего набора наблюдаемых (т. е. надежно предсказанных) кассет и получали интегральное значение статистики, описывающей относительное расположение всех спейсеров с мишенями.

Затем спейсеры в каждой кассете перемешивали случайным образом. Для вновь полученного набора симулированных кассет вычисляли значение статистики расположения спейсеров с мишенями. По такому принципу получили 100'000 наборов псевдокассет и для каждого раунда пермутаций подсчитали значение порядковой статистики расположения спейсеров с мишенями и построили распределение этих статистик. Значение статистики для наблюдаемых кассет сопоставили с полученным распределением для симулированных псевдокассет и на основании этого рассчитали значение p-value.

Аналогичным образом оценили смещение общих спейсеров к дистальному концу CRISPR-кассет.

2.2.9 Определение колокализации спейсеров и протоспейсеров

Для проверки гипотезы о наличии связи между распределением спейсеров и протоспейсеров в индивидуальных метагеномах, мы разработали статический тест, аналогичный критерию Кохрана-Мантеля-Ханцеля (Cochran–Mantel–Haenszel), CMH [124]. Для полного набора уникальных протоспейсеров из всех описываемых метагеномных коллекций и для каждого индивидуального метагенома строили таблицу сопряженности 2x2. Для заполнения таблицы все протоспейсеры классифицировали на основе присутствия или отсутствия протоспейсеров и их спейсеров в данном индивидуальном метагеноме (см. **Таблица 2**): 1) метагеном содержит как спейсер, так и протоспейсер (клетка **‘A’**); 2) в метагеноме обнаружен только протоспейсер (клетка **‘B’**); 3) в данном метагеноме обнаружен только спейсер (из спейсеров, имеющих протоспейсеры в полном наборе уникальных протоспейсеров, клетка **‘C’**); 4) оставшиеся протоспейсеры из набора уникальных протоспейсеров, не встреченные в данном метагеноме, ни в форме спейсера, ни в форме протоспейсеры (клетка **‘D’**). В клетки вносили число событий соответствующего типа.

	С+	**С-**
П+	**A** «С+П+» Число пар спейсер-протоспейсер	**B** «С-П+» Число протоспейсеров без спейсера
П-	**C** «С+П-» Число спейсеров «чужих» протоспейсеров	**D** «С-П-» Число протоспейсеров, не «упоминаемых» в данном метагеноме

Таблица 2. Схема заполнения таблицы сопряженности для индивидуального метагенома. Сокращения: «С» = спейсер; «П» = протоспейсер; «+» = присутствует; «-» = отсутствует

Для полученного набора таблиц сопряженности вычислили суммарную статистику. Для определения уровня значимости нулевое распределение рассчитывали следующим образом: протоспейсеры перемешивали между индивидуальными метагеномами таким образом, что

число протоспейсеров для каждого индивида оставалось неизменным. Такая симуляция была проделана 100'000 раз. Для каждого раунда пермутаций, т.е. нового набора таблиц сопряжённости, рассчитывали значение статистики CMH при помощи функции mantelhaen.test из R пакета stats (https://stat.ethz.ch/R-manual/R-devel/library/stats/html/mantelhaen.test.html). Значение статистики, полученное на реальных данных, сравнивали с нулевым распределением и рассчитывали p-value.

2.2.10 Определение типа CRISPR-Cas систем

Тип CRISPR-системы определяли двумя способами: 1) по характеристическим белкам ассоциированного *cas*-локуса в соответствии с принятой классификацией [36]; 2) на основании классификации повторов при помощи алгоритма CRISPRmap (http://rna.informatik-freiburg.de/CRISPRmap/Input.jsp) [125]. В случаях, когда тип кассеты можно было определить двумя способами, полученные классификации сравнивали и анализировали.

Глава 3. Результаты и обсуждение

3.1 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**) [10]. Кроме того, получен метагеном биома нисходящей ободочной кишки человека («Human Distal Gut Biome project», **DG**) [113] и метагеномы кишечника 13 здоровых японцев («13 Healthy Human Gut Metagenomes», **JPN**) [11]. Перечисленные наборы метагеномных данных использовали для предсказания CRISPR-кассет при помощи существующих алгоритмов — CRT, PILER-CR, CRISPRFinder. Каждый из них в качестве кандидатных CRISPR-кассет определяет последовательности коротких прямых повторов, разделённых уникальными последовательностями, однако детали поиска повторов и дальнейшего уточнения границ кассет различаются. Соответственно, сильно различаются и результаты предсказаний каждой из программ.

Кроме того, поиск CRISPR-кассет в метагеномных данных является нетривиальной задачей сам по себе. Во-первых, из-за большого объёма анализируемых данных, а во-вторых, из-за высокой степени их фрагментации. Например, метагеномные контиги, получаемые при секвенировании по методу дробовика (shotgun sequencing), довольно часто содержат короткие участки непрочитанных нуклеотидов ('NNNN'), фланкированные сходными последовательностями длиной 20-40 нт. Подобные участки формально удовлетворяют требованиям алгоритмов поиска повторов и поэтому могут быть ошибочно распознаны как CRISPR-кассеты. Особенно ложноположительные предсказания такого рода характерны для алгоритма CRISPRFinder, который находит большое число коротких недостоверных кандидатных кассет со структурой типа «повтор-спейсер-повтор»[101]. При использовании алгоритмов CRT и PILER-CR, участки геномных повторов, а также участки низкой сложности часто попадали в список кандидатных CRISPR-кассет.

В целях борьбы с ложноположительными предсказаниями и для получения надежного набора кассет мы использовали следующую процедуру фильтрации [101]. Надежно предсказанными считали кассеты нескольких типов:

1) кассеты, одновременно предсказанные тремя алгоритмами;

2) кассеты, предсказанные одним или двумя алгоритмами, но фланкированные *cas*-генами;

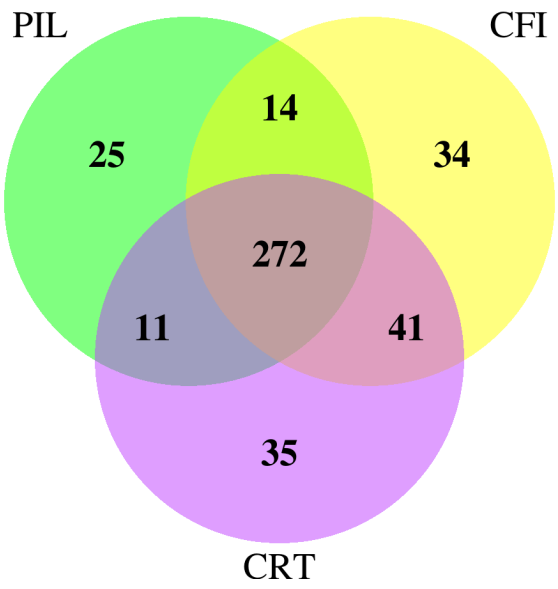
3) кассеты, предсказанные менее чем тремя алгоритмами, но содержащие повторы, сходные с консенсусной последовательностью повтора одной из кассет, уже добавленных к набору надежно предсказанных.

Кассеты, расположенные рядом с *cas*-генами (**Таблица 2**), могли не попасть в список кассет, предсказанных всеми тремя программами, в силу ряда причин: 1) располагались на границе контига, из-за чего были очень короткими; 2) содержали сильно различающиеся по длине спейсеры; 3) повторы кассет были сильно дивергированными. Кассеты с описанными отклонениями структуры могли быть не распознаны отдельными алгоритмами, но, в силу структуры CRISPR-локуса, действительно являются CRISPR-кассетами. Поэтому (согласно процедуре фильтрации) мы добавляли их к списку надежно предсказанных кассет на втором этапе процедуры формирования надежного списка кассет.

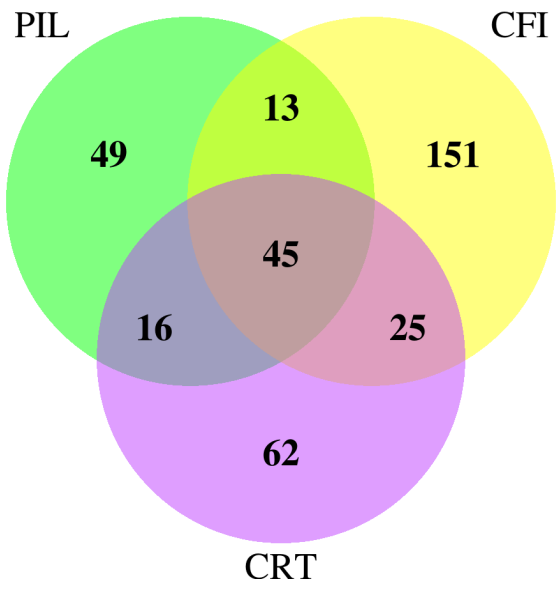
Третий этап процедуры фильтрации предполагал добавление кассет к списку надежно предсказанных на основании сходства последовательности повтора и повторов кассет уже отнесенных к категории надежно предсказанных. Мы руководствовались тем фактом, что повторы CRISPR-кассет образуют кластеры в пространстве последовательностей (Kunin, 2007); с большой вероятностью кандидатная кассета, содержащая достоверный повтор может также считаться достоверной, даже если ее границы предсказаны не совсем верно. Тем не менее, в данном исследовании на третьем этапе мы не добавили ни одной кассеты к списку надежно предсказанных, что свидетельствует в пользу надежности и достаточности критериев (1) и (2).

Наборы идентифицированных кассет проиллюстрированы на **Рисунке 8,** основные параметры кассет суммированы в **Таблице 3**. 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.

A) JPN



B) HMP



C) DG

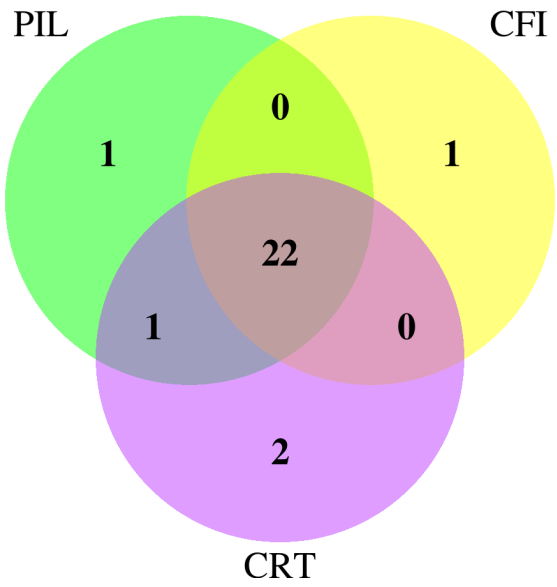


Рис.8. Диаграммы Венна, иллюстрирующие число CRISPR-кассет, предсказанных алгоритмами CRISPRFinder (CFI), PILER-CR (PIL) и CRT в трёх метагеномных наборах данных микробиомов человека.

Примечательно, что из 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. Кандидатных кассет, содержащих повторы, похожие на повторы надежно предсказанных кассет, в анализируемых коллекциях обнаружено не было.

Это наблюдение говорит в пользу устойчивости полученного набора надежно предсказанных кассет.

Окончательный набор надежно предсказанных кассет состоял из 296, 78 и 24 CRISPR-кассет, обнаруженных в метагеномных коллекциях JPN, HMP и DG, соответственно. Ни одна из кассет не была встречена одновременно в двух разных метагеномных коллекциях. Значительная часть кассет располагалась рядом с *cas*-генами. Было обнаружено 70, 56 и 6 таких кассет в метагеномных коллекциях JPN, HMP и DG, соответственно, то есть, 24%, 71% и 25% от всех надежно предсказанных кассет.

296 надежно предсказанных кассет, идентифицированных в метагеномной коллекции JPN, содержали 3410 спейсеров, 2992 из которых были уникальными. В 78 кассетах метагенома HMP содержалось 378 спейсеров, из них — 352 уникальных. Среди 344 спейсеров кассет метагенома DG, лишь один спейсер встретился дважды. Число уникальных последовательностей повторов составило 170, 74 и 19 для всех кассет метагеномных коллекций JPN, HMP и DG, соответственно. (**Таблица 3**).

Таблица 3. 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
тремя программами одновременно	272	45	22
1 или 2 программами, но расположены рядом с *cas*-генами	24	33	2
Набор надежно предсказанных кассет			
Общее число	296	78	24
Кассеты рядом с *cas*-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 [125]	122	18	11

Получив набор надежно предсказанных CRISPR-кассет (далее именуемых просто CRISPR-кассеты), мы сравнили последовательности их консенсусных повторов с повторами уже известных CRISPR-кассет, задепонированных в базу данных CRISPRdb [126]. Всего лишь 23 повтора из 263 оказались похожими на повторы из базы данных CRISPRdb. Все эти повторы обнаружены в кассетах метагеномной коллекции JPN и объединялись в 17 кластеров. Столь незначительное пересечение с базой данных CRISPRdb указывает на то, что подавляющая часть обнаруженных нами кассет не была известна ранее.

Число CRISPR-кассет в индивидуальных метагеномах сильно различалось. Мы не обнаружили взаимосвязи между числом идентифицированных кассет и средней длиной контига или числом контигов, приходящихся на образец (данные не представлены). Возможно, причинами наблюдаемых отличий между разными метагеномными коллекциями являются особенности технологий секвенирования и алгоритмов последующей сборки. С другой стороны, число CRISPR-кассет, приходящихся на индивидуальный микробиом, может коррелировать с распределением основных таксономических категорий в пределах микробиома человека и, возможно, косвенно отражать энтеротип конкретного индивида [90].

3.2. 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 (см. Данные и алгоритмы).

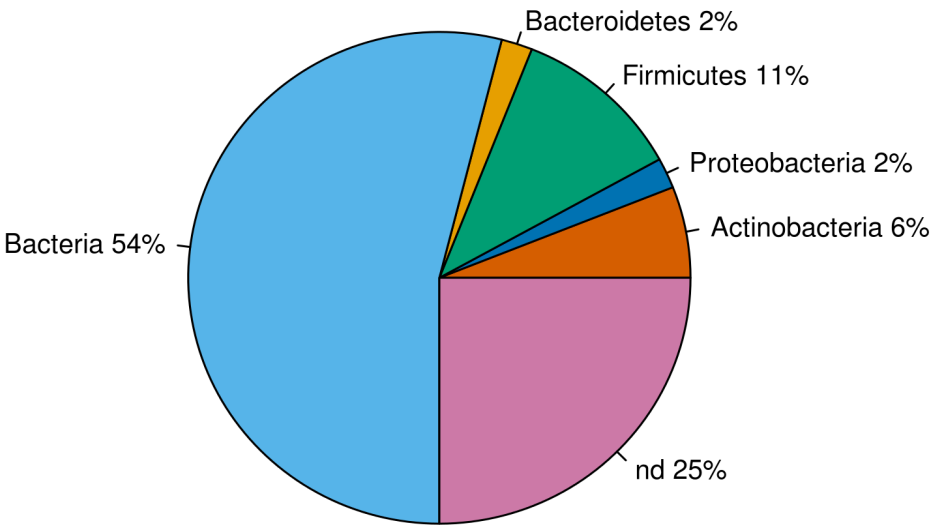
Зачастую, установление таксономического положения CRISPR-содержащих контигов является непростой задачей. Во-первых, в силу небольшой длины контигов. Во-вторых, из-за подверженности *cas*-генов частым горизонтальным переносам [36]. Короткие фланкирующие последовательности могут вмещать только *cas*-гены, таксономия которых часто не отражает филогенетического родства, и поэтому непригодна для определения таксономического положения соседних кассет. Таксономическое положение, по меньшей мере, на уровне домена удалось определить для 73 из 296 кассет метагеномной коллекций JPN (25%), 69 из 78 кассет коллекции HMP (88%) и 12 из 24 кассет коллекции DG (50%).

Несмотря на то, что общее число кассет, идентифицированных в каждой из трёх метагеномных коллекций, существенно различалось, превалирующие таксономические категории, к которым были отнесены CRIRSPR-содержащие контиги, совпадали (**Рисунок 9**, **Рисунок 10**). Большая часть контигов с установленной таксономией принадлежала к типу *Firmicutes*. Было обнаружено 33 и 43 таких контигов, в наиболее крупных метагеномных

коллекциях JPN (20 — в образцах взрослых и 13 — в образцах детей) и HMP; и восемь — в метагеноме DG. Основная часть контигов отнесённых к типу *Firmicutes* принадлежала бактериям родов *Bacilli* и *Clostridia*.

Значительная доля CRISPR-содержащих контигов метагеномной коллекции JPN была отнесена к типу *Actinobacteria*. В общей сложности идентифицировано 19 актинобактериальных контигов: 5 — в образцах взрослых и 14 — в образцах детей. Большая часть таких контигов отнесена к роду *Bifidobacterium*. Контиги, принадлежащие к типу *Proteobacteria* и группе *Bacteroidetes/Chlorobii*, составили только по 2% каждый. Контиги протеобактериального происхождения в основном принадлежали семейству *Enterobacteriaceae*, преимущественно — *E. coli*. Для четверти (25%) CRISPR-содержащих контигов метагеномной коллекции JPN не удалось определить таксономическое положение ниже уровня домена, поэтому они были отнесены к неспецифической таксономический группе 'Бактерии' (**Рисунок 9, A, B**).

A) JPN



B) JPN, дети

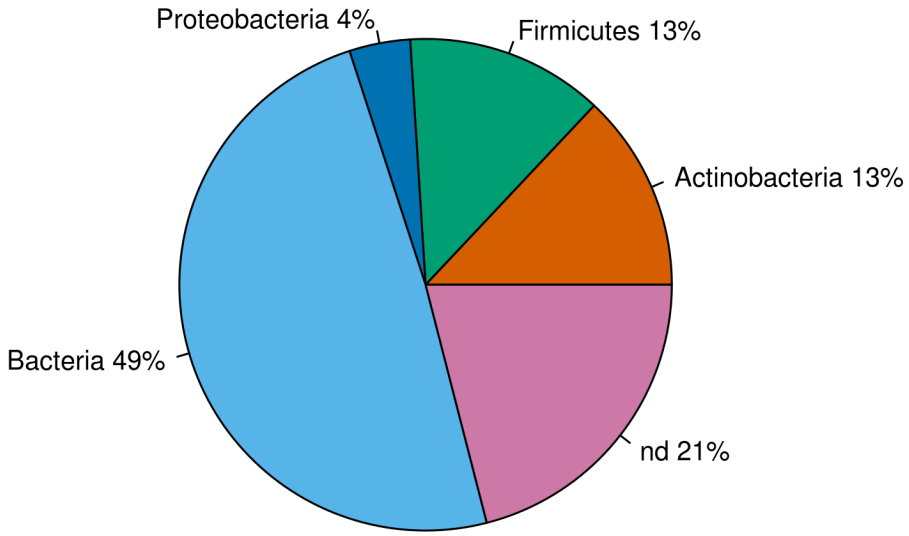


Рис.9. Таксономическое распределение метагеномных контигов коллекции JPN, содержащих CRISPR-кассеты. Условные обозначения: **nd** – таксономическое положение контига не установлено.

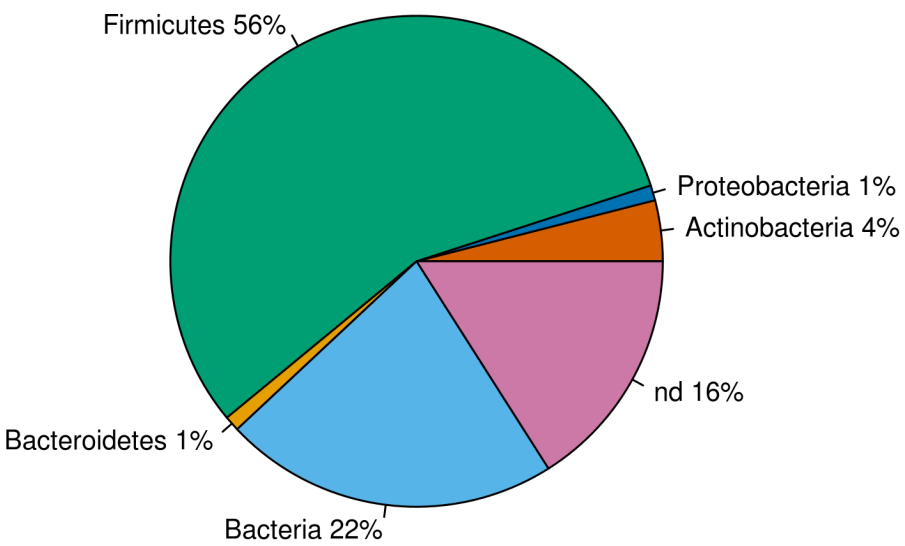
Согласно ранее опубликованным исследованиям метагеномной коллекции JPN [11], в микробиомах взрослых и детей доминирующие микробные таксоны различаются. По результатам нашей работы, метагеномные контиги с CRISPR-кассетами из образцов, взятых у детей, отнесены к тем же доминирующим таксонам, что и образцы взрослых. Однако, доля актинобактериальных контигов и, соответственно, кассет, у детей значительно больше, чем у взрослых (13%) (**Рисунок 9, А**).

Как уже было отмечено раньше, наиболее заметная доля контигов с CRISPR-кассетами в метагеномной коллекции HMP отнесена к типу *Firmicutes* (**Рисунок 10, А**). 22% контигов коллекции HMP имели очевидное бактериальное происхождение, но отнести их к какому-либо конкретному таксону более низкого порядка оказалось затруднительно. Для 16% контигов таксономическое положение не идентифицировано, так как сходных последовательностей в коллекции NR базы данных GenBank обнаружено не было. Минорная фракция CRISPR-содержащих HMP-контигов отнесена к типам *Actinobacteria*, *Bacteroidetes* и *Proteobacteria*. Суммарная доля таких контигов составила 6%.

Помимо кассет, отнесённых к типу *Firmicutes*, в метагеномной коллекции DG удалось определить таксономическое положение ещё для четырёх кассет (**Рисунок 10, В**). Две кассеты принадлежали типу *Actinobacteria*, одна — к типу *Bacteroidetes.* Оставшаяся кассета была архейного происхождения, единственная среди всех идентифицированных кассет. Последовательности, фланкирующие данную кассету, сходны с последовательностями *Methanobrevibacter smithii* — доминирующего архейного таксона, обитающего в кишечнике человека [127].

В целом стоит отметить, что распределение основных таксономических категорий контигов, содержащих CRISPR-кассеты, отличается от распределения основных таксономических категорий всех контигов по данным анализа генов 16S рРНК тех же метагеномных данных. Например, по данным 16S рРНК для метагеномной коллекции JPN, преобладающими таксонами в микробиомах взрослых и детей были представители типа *Bacteroidetes*, несколько родов типа *Firmicutes* и представители рода *Bifidobacterium.* Однако в микробиомах грудных детей преобладали представители рода *Bifidobacterium* и виды семейства *Enterobacteriaceae* [11]. Для CRISPR-содержащих контигов по нашим данным доминирующим таксоном для взрослых и детей был тип *Firmicutes*.

A) HMP



B) DG

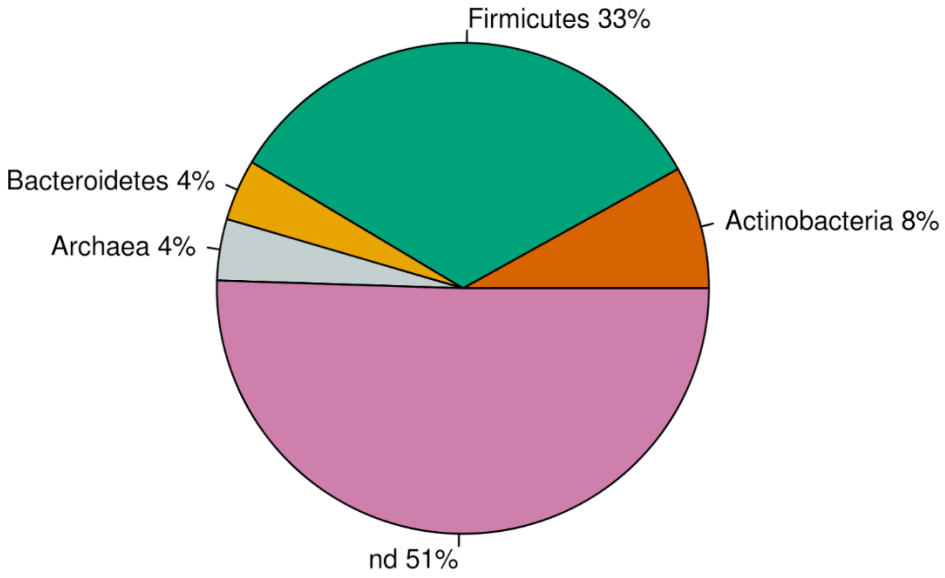


Рис.10. Таксономическое распределение метагеномных контигов коллекций HMP и DG, содержащих CRISPR-кассеты. Условные обозначения: **nd** – таксономическое положение контига не установлено.

В метагеноме HMP на основании анализа представленности генов 16S рРНК, доминирующим таксоном является тип *Firmicutes* [10]. Меньшие, но примерно равные доли занимают представители типов *Bacteroidetes*, *Actinobacteria* и *Proteobacteria*. Это означает,

что наблюдаемое распределение таксонов по 16S рРНК очень похоже на распределение таксономических категорий контигов с CRISPR-кассетами, полученное нами.

По данным анализа генов 16S рРНК метагеномной коллекции DG, значительная часть контигов была отнесена к типу *Firmicutes*, меньшая доля — к типу *Actinobacteria* [113]. Нам также удалось обнаружить CRISPR-содержащие контиги, которые можно отнести к упомянутым типам. По нашим данным, один контиг можно отнести к типу *Bacteroidetes*, но согласно анализу 16S рРНК этот таксон отсутствует в метагеномной коллекции DG. Возможной причиной наблюдаемых различий могут быть ошибки в численной оценке преобладающих таксонов, возникающие из-за различий копийности гена 16S рРНК (число копий гена может быть от 1 до 15 в зависимости от вида [128]) и того факта, что гены 16S рРНК некоторых таксонов могут не амплифицироваться при использовании универсальных праймеров [129]. Кроме того, несовпадение таксономических оценок может быть обусловлено различным числом CRISPR-кассет в разных типах бактерий.

3.3 Типы CRISPR-Cas систем

Активно функционирующие иммунные системы CRISPR-Cas состоят из CRISPR-кассет и ассоциированных с ними *cas*-генов [130]. Мы классифицировали идентифицированные кассеты по типу последовательности повтора и на основании состава *cas*-локусов там, где это было возможно.

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 [36], 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 11**).

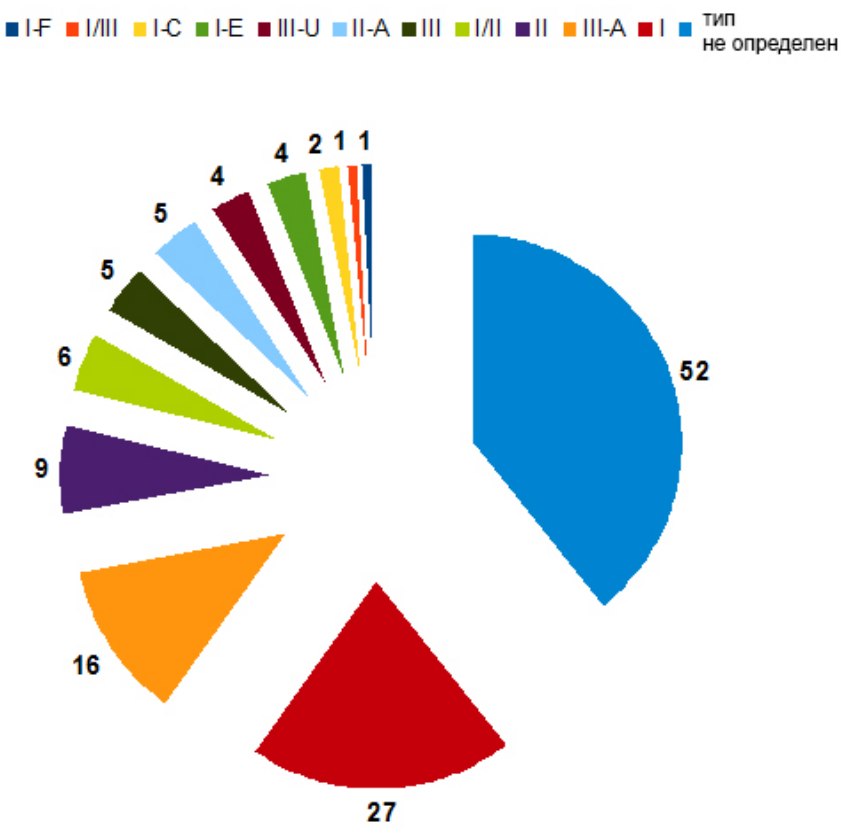


Рис.11. Распределение типов CRISPR-Cas систем среди идентифицированных кассет. Классификация основана на составе сцепленных *cas*-локусов. Голубым цветом («тип не определен») показаны кассеты, для которых *cas*-локус содержал только универсальный ген *cas1*, то есть конкретный тип CRISPR-Cas системы по *cas*-локусу определить было нельзя.

Последовательности повторов CRISPR-кассет можно разделить на несколько типов по сходству и способности формировать стабильные вторичные структуры [25], [36]. Как правило, тип последовательности повтора ассоциирован с определёнными *cas*-генами, поэтому CRISPR-кассеты можно классифицировать, исходя из типа повтора. Это удобно в случаях, когда информация об ассоциированных *cas*-генах недоступна. Например, в силу недостаточной длины участков метагеномных контигов, фланкирующих CRISPR-кассеты.

Недавно разработан алгоритм CRISPRmap, позволяющий автоматически классифицировать повторы CRISPR-кассет [125]. CRISPRmap используется для классификации всех известных (т.е. опубликованных) CRISPR-кассет исключительно на основании параметров последовательностей повторов. Согласно классификации CRISPRmap, все известные на настоящий момент повторы CRISPR-кассет можно подразделить на шесть основных суперклассов: A-F.

Мы применили алгоритм CRISPRmap для классификации идентифицированных кассет в различных метагеномных коллекциях микробиомов человека. По результатам его работы,194 уникальных повтора, соответствующих 236 CRISPR-кассетам, отнесены к одному из шести суперклассов (A-F) (**Приложение 1**). В анализируемой выборке повторов обнаружены представители каждого из шести суперклассов. Наиболее распространёнными оказались повторы суперклассов F, E и D. Необходимо отметить, что именно эти суперклассы содержат слабо консервативные последовательности повторов [125].

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 [36]. Это наблюдение говорит о том, что многие кассеты и повторы, и даже их типы, не были известны ранее, и, соответственно, о неполноте существующей классификации.

Среди всех идентифицированных нами CRISPR-кассет, 82 можно было классифицировать как на основании состава локуса ассоциированных генов, так и при помощи CRISPRmap — по типу повтора. Мы сравнили результаты двух способов классификации. Противоречия обнаружены в 16 случаях, для трёх суперклассов повторов (F, C и D) (**Таблица 4**). Возможно, такое наблюдение указывает на несовершенство принятого в настоящее время принципа соответствия между составом *cas*-локуса и типом повтора и необходимость пересматривать и развивать текущую классификацию.

Таблица 4. Расхождения в способах классификации CRISPR-кассет по типу повтора и по составу сцепленного *cas*-локуса.

Консенсусная последовательность повтора	Супер-класс повтора	Повтор ассоциирован с локусами типа(ов)	Тип *cas*-локуса
GCTTTGGAACCATAAAAAATTACA	F	I-A, I-B, I-D, III-A, III-B	II
CTTGGTTTACGGTACTTCCGAAAC	F	I-A, I-B, I-D, III-A, III-B	II
GGTGAACTACTGCTTGATCTACG	F	I-A, I-B, I-D, III-A, III-B	II
ATTTCAATCCACTCCGCTATCGCTAGCAGAGAC	D	I-C	I-F
GTCGCCCTCTTCACGAGGGCGTGGATTGAAAT	D	I-C	II-A
CCCCGCGAGGGGACGGTTACATACCCTTCA	D	I-C	III-A
TTTCCGTCCCCTTCCCGGGGATCTTATTTCTCAAT	D	I-C	III-A
TTTCCGTCCCCACGAAGGGGACCCATCTTCTCTAC	D	I-C	III-A
GTCGCCCCCGCAAGGGGGCGTGGATTGAAAT	D	I-C	III-U
GTTTTCGTCCCCATTAGGGGATTTTGTTTTTTATAT	D	I-C	III-A
CAGCTTATCCCCGCGAGGGGACGGTTAC	C	I-E, I-F	III
GTTTCAATTCCCGCAGTTGCGGGTAAGATACAT	C	I-E, I-F	III-U
GTTTCCGTCCCCTTGCGGGGATAAATTGCAAT	C	I-E, I-F	III-U
TTTCGCCCCCTTACGGGGATTGTA	C	I-E, I-F	III-A
TTTCCGTCCCCTAAGTTGGGGTTATCTCTAAAAT	C	I-E, I-F	III-A
GAGTTTCCGTCCCCTTGCGGGGTGTA	C	I-E, I-F	III-A

3.4 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, исходя из предпосылки, что они должны быть обогащены последовательностями вирусов, обычно населяющих микробиомы кишечника 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 5**). 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. Для 35 (15%) спейсеров из метагеномной коллекции HMP мы обнаружили 89 пар спейсер-протоспейсер. Только для шести спейсеров, предсказанных в метагеномной коллекции DG, нашелся протоспейсер. Все они происходили из одной кассеты, имеющей архейное происхождение (*Methanobrevibacter smithii*), а соответствующие им протоспейсеры приходятся на один метагеномный контиг (gi106748134|gb|AAQK01000324.1), отнесенный при аннотации к тому же таксону.

Таблица 5. 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
Виромы 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. Сведения о происхождении протоспейсеров, найденных в коллекции NR собраны в **Таблице 6**.

Таблица 6. Протоспейсеры, найденные в коллекции NR базы данных GenBank. Для спейсеров метагеномной коллекции DG не было найдено протоспейесров в коллекции NR.

Число спейсеров	Число соответствующих им протоспейсеров	Источник происхождения протоспейсеров
Пары спейсер-протоспейсер, найденные для метагеномной коллекции JPN:		
4	11	Полные геномы бактериофагов VT2-Sakai, epsilon15, Sf6
7	7	Последовательности неклассифицированных бактериофагов, изолированных из образцов кишечной микрофлоры человека
3	3	Геном Bifidobacterium longum subsp. infantis 157F
3	3	Плазмиды распространенные в энтеробактериях: *Escherichia coli*, *Salmonella enterica* и *Klebsiella pneumoniae*
Пары спейсер-протоспейсер, найденные для метагеномной коллекции HMP:		
5	5	Геном *Faecalibacterium prausnitzi*
3	3	Последовательности некультивируемых микроорганизмов (клоны LM0ABA27ZF12FM1 и VC1A546TR), изолированных из образцов кишечной микрофлоры человека
1	1	Геном *Bifidobacterium longum subsp. Longum*

Примечательно, что для одного из спейсеров коллекции JPN удалось найти идентичные между собой протоспейсеры, с четырьмя заменами по отношению к спейсеру, в пяти разных фагах энтеробактерий: фаге VT2-Sakai (инфицирующий род *Enterobacteria*), фаге Sf6 (инфицирующий род *Enterobacteria*), Stx1 конвертирующем бактериофаге (инфицирующем род *Enterobacteria*), Stx2 конвертирующем бактериофаге II (инфицирующем род *Enterobacteria*) и бактериофаге SE1 (инфицирующем *Salmonella enterica*). Данный протоспейсер соответствовал наиболее консервативному участку гена, кодирующего белок, подобный белку Ea22 фага лямбда, присутствующего в геномах всех перечисленных бактериофагов (**Рисунок 12**). Ea22 — это ранний белок фага лямбда [131]. Можно предположить, что такой спейсер служит залогом CRISPR-опосредованного иммунитета сразу против группы родственных энтеробактериальных бактериофагов по типу множественной устойчивости, как, например, описано для CRISPR-систем *Clostridium difficile* [132].

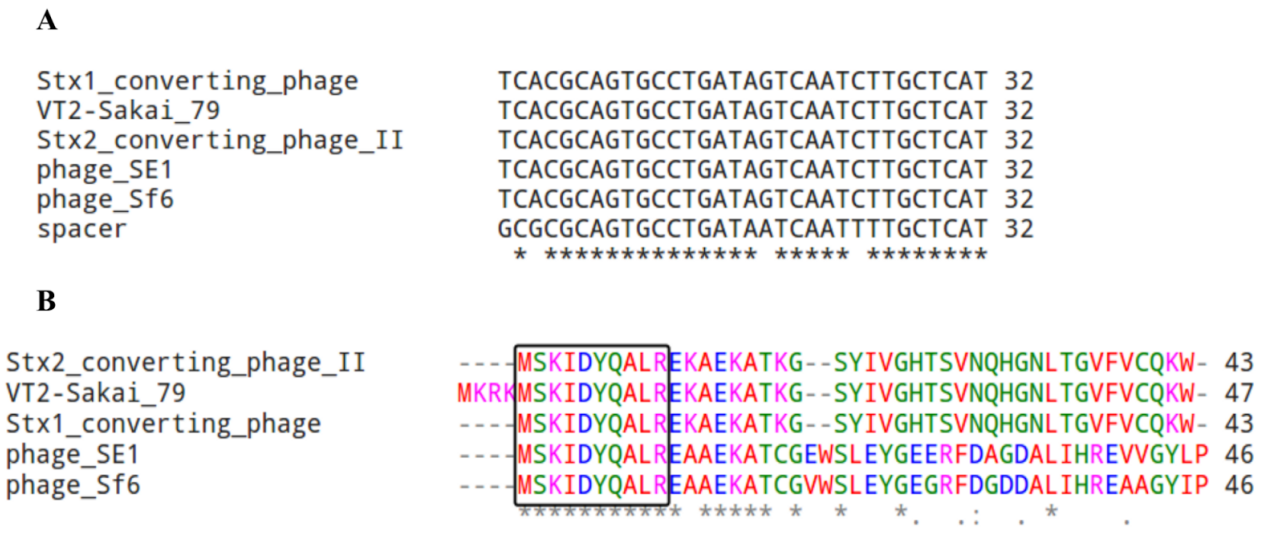


Рис.12. Расположение протоспейсера, соответствующего наиболее консервативной части гена**, кодирующего белок, подобный белку Ea22 фага лямбда, в пяти родственных бактериофагах энтеробактерий.** (**A**) Выравнивание нуклеотидных последовательностей спейсера и соответствующих протоспейсеров. (**B**) Аминокислотное выравнивание соответствующих протоспейсеру участков белков. Позиция спейсера показана рамкой.

В целом, наблюдаемое небольшое число совпадений между спейсерами и последовательностями полных и частичных фаговых геномов (задепонированных в GenBank), возможно, отражает тот факт, что пространство вирусных последовательностей до сих пор слабо изучено.

В силу небольшой длины спейсеров можно ожидать нахождение сходства с последовательностями столь большой базы данных, как NR коллекция GenBank, в силу чисто случайных причин. Что бы проверить, что обнаруженные протоспейсеры не являются случайными, мы, провели аналогичный поиск для симулированных последовательностей спейсеров (псевдоспейсеров) против той же коллекции последовательностей (см. Данные и алгоритмы). Для 2992 псевдоспейсеров, сконструированных на основе набора спейсеров JPN, мы обнаружили 66 хитов (в основном, относящиеся к различным штаммам *E.coli*), которые соответствовали 10 парам псевдоспейсер-псевдопротоспейсер. В отличие от спейсеров набора JPN, псевдоспейсеры, сконструированные на их основе, по большей части находили протоспейсеры, попадающие на участки полных геномов различных бактерий, чаще всего — межгенные. Только в трех случаях псевдоспейсеры были подобны последовательностям мобильных генетических элементов и генов, ассоциированных с вирусами или профагами. На основании такой симуляции, мы можем заключить, что найденные протоспейсеры не являются случайными совпадениями, возникшими в силу небольшой длины анализируемых

последовательностей и большого размера базы данных для сравнения. Симуляция для наборов спейсеров HMP и DG дала аналогичные результаты.

Помимо собственных метагеномных последовательностей и известных последовательностей вирусного происхождения базы данных GenBank мы сравнили полученные списки спейсеров с доступными данными виромных проектов микробиома человека [114], [115]. Ни одного достоверного протоспейсера среди последовательностей ДНК-вирусов микробиома человека («Virome of human gut») обнаружено не было. Однако нам удалось обнаружить один гипотетический протоспейсер для спейсера из метагеномной коллекции JPN (HumanGut_CONTIG_00008549_spacer_5) в вироме преимущественно некультивируемых вирусов (РНК-вирусов) кишечника человека[114]. Выравнивание спейсера и кандидатного протоспейсера содержало четыре нуклеотидные замены, что формально соответствовало нашим критериям для идентификации протоспейсеров. Спейсер происходил из кассеты, отнесённой к роду *Bacillus*.

Данное наблюдение примечательно, так как известно, что мишенями для бактериальных CRISPR-систем служит чужеродная ДНК, а не РНК [24]. С другой стороны, мишенью для crРНК CRISPR-системы археи *Pyrococus furiosus* может служит матричная РНК [58]. Кроме того, отмечены частые совпадения между CRISPR-спейсерами и последовательностями РНК-виромов, изолированных из горячих источников [133]. Таким образом, наличие CRISPR-опосредованного иммунитета против РНК-вирусов в архейных системах не подлежит сомнению, но достоверных примеров такового в бактериальных системах описано не было.

В ходе дальнейшего анализа выяснилось, что виромная последовательность (HFVirus_READ_00009708), включающая обсуждаемый протоспейсер, имеет сходство с вирусным белком. Но, кроме этого, она содержит в середине участок, сходный с последовательностью вектора для клонирования, и, вероятно, является химерным. Так как участок, подобный гену вирусного белка, и протоспейсер лежат по разные стороны от векторный вставки, нельзя с достаточной долей уверенности считать, что данный протоспейсер имеет вирусное происхождение.

Известно, что РНК служит мишенью для CRISPR-Cas систем типа III-B (Staals et al., 2013). Метагеномный контиг с анализируемым спейсером не содержал *cas*-генов, поэтому мы классифицировали кассету на основании структуры и последовательности повтора при помощи алгоритма CRISPRmap [125]. Повтор отнесён к семейству структурных мотивов 23 (motif 23), которое преимущественно ассоциировано с генами CRISPR-Cas систем III-A и III-B подтипов: *csm2*, *csm3*, *csm5*, *cmr6*, *cmr1*, *cmr4*. Мишенью для CRISPR-Cas систем III-B подтипа может служить РНК, в то время как III-A подтип нацелен на ДНК [134], [135]. Если

обнаруженный протоспейсер действительно имеет вирусное происхождение, он может служить мишенью для CRISPR-Cas системы III-B-подобного типа.

Несмотря на то, что короткие мотивы, расположенные рядом с протоспейсерами (PAM) сейчас считаются частым и почти неотъемлемым элементов различных CRISPR-Cas систем [52], мы не обнаружили ни одного достоверного PAM-мотива, ни для кассет, для нескольких спейсеров которых найдены протоспейсеры, ни после кластеризации протоспейсеров по повторам соответствующих кассет.

3.5 Таксономия протоспейсеров в сравнении с таксономией CRISPR-кассет

Таксономическое положение метагеномных контигов, содержащих CRISPR-кассеты можно определять, как на основании анализа последовательностей, фланкирующих CRISPR-кассеты, так и на основании информации о происхождении протоспейсеров. Для некоторых кассет удалось получить информацию о таксономии как на основании фланкирующих последовательностей, так и протоспейсеров. Интересно сравнить, совпадают ли в этих случаях приписанные таксономические группы.

Из 296 метагеномных контигов коллекции JPN, содержащих CRISPR-кассеты, для 73 таксономическое положение определили по фланкирующим последовательностям, для 13 контигов таксономическое положение определено на основании источника протоспейсеров. Только для семи метагеномных контигов таксономическое положение определено и тем, и другим способом. В пяти случаях таксономические группы разного происхождения хорошо согласовались между собой, и, по меньшей мере, совпадали на уровне типа. В двух оставшихся случаях, на основании фланкирующих последовательностей установлена неспецифическая таксономическая группа только на уровне домена «Бактерии», в то время как таксономическое происхождение, установленное на основании анализа протоспейсеров, было более специфическим.

Из 78 контигов с CRISPR-кассетами, идентифицированными в метагеномной коллекции HMP, для 48 таксономическое положение можно было определить на основании последовательностей, фланкирующих CRISPR-кассеты, и для шести контигов — на основании анализа происхождения протоспейсеров. Только трем контигам были приписаны таксономические группы обоих типов, и во всех случаях таксономическое положение, установленное на основании фланкирующих последовательностей и на основании протоспейсеров — в целом не противоречили друг другу. Таким образом, два способа определения таксономии CRISPR-кассет, как правило, дополняют друг друга.

3.6 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 (соседних в сравниваемых кассетах), одновременно присутствующих в наборах данных HMP и JPN. Контиги, содержащие этот участок кассеты, также перекрывались на протяжении небольшой фланкирующей последовательности (длиной 134 нуклеотида). Таксономическое положение указанных контигов, независимо определённое для разных метагеномов, совпадало. Контиги принадлежали типу *Firmicutes.*

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. Особенно много общих спейсеров обнаружено для пар индивидов F2X-F2Y (брат и сестра из одной семьи) и F2X-INM (не связанные между собой мальчик трех лет и девочка четырех месяцев) – 44 и 18 общих спейсеров соответственно. Между двумя парами индивидуальных метагеномов (INE-INB и F2W-INA) найдены целиком общие CRISPR-кассеты, вместе с фланкирующими последовательностями. Общие спейсеры происходили из CRISPR-кассет с идентичными последовательностями повторов. (**Рисунок 13**).

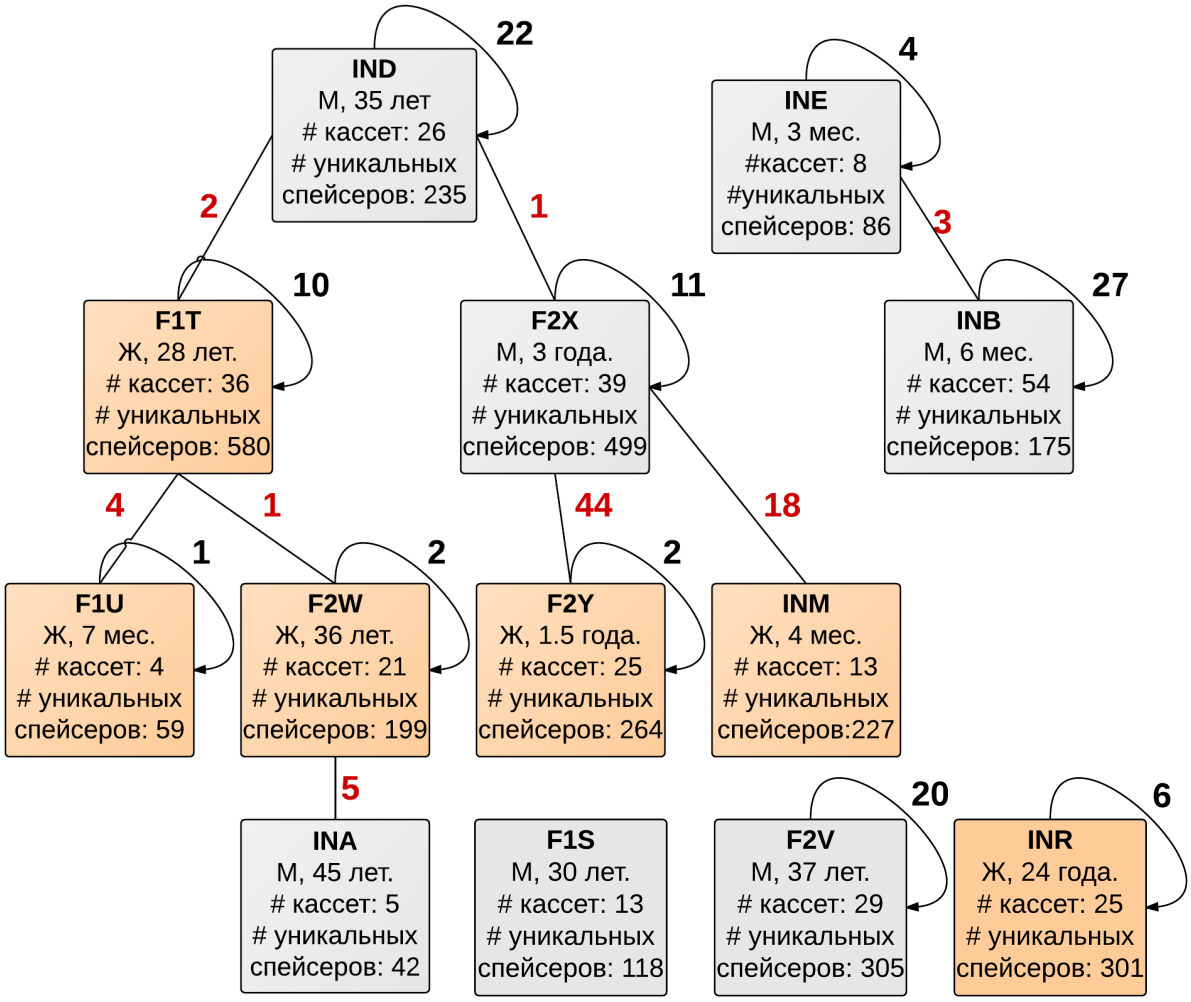


Рис.13. Общие спейсеры в индивидуальных метагеномах коллекции JPN.

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 (см. Данные и алгоритмы). Согласно этой процедуре мы перемешали кассеты между индивидами

случайным образом, так что число кассет, принадлежащих каждому индивиду, оставалось неизменным. Пермутации по описанной схеме были проделаны 100'000 раз. Для каждого раунда пермутаций считали число общих спейсеров. Среднее число общих спейсеров по результатам симуляций составило 127, распределение представлено на **Рисунке 14**.

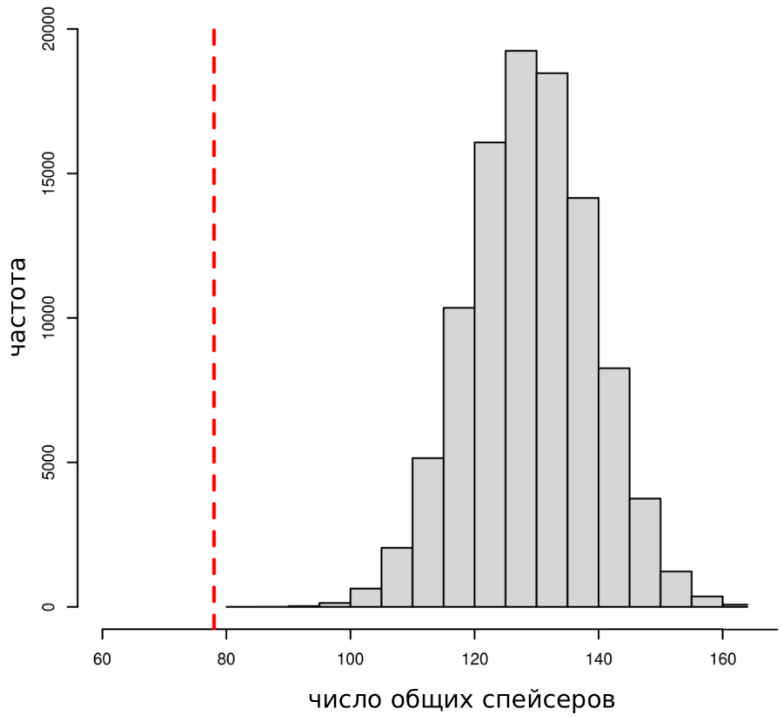


Рис.14. Распределение числа общих спейсеров между индивидами для 100'000 случайных пермутаций. Красная пунктирная линия соответствует наблюдаемому числу общих спейсеров в кассетах метагеномной коллекции JPN.

Число общих спейсеров для реальных данных было всегда меньше, чем мы наблюдали для симулированных данных (p-value < 10^{-5}). Таким образом, небольшое число общих спейсеров, которое мы видим между индивидуальными метагеномами, нельзя считать случайным. Здесь мы наталкиваемся на парадокс. Ранее было показано, что, микробиомы членов одной семьи и даже домашних животных могут значительно перекрываться [136], поэтому можно было бы ожидать наличие общих вирусов и, соответственно, спейсеров. Вопреки этому, мы наблюдаем небольшое число общих спейсеров между индивидами, даже принадлежащими к одной семье. 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.

Сравнивая распределение типов повторов (кластеров повторов) по индивидуальным метагеномам, мы не обнаружили универсальных, широко распространённых кластеров. Подавляющее большинство кластеров повторов оказались очень специфичны и были ассоциированы только с одним конкретным индивидуальным метагеномом. Тем не менее, 24 кластера повторов встречены в, по-меньшей мере, двух индивидуальных метагеномах (**Таблица 7**), что свидетельствует о наличии похожих или родственных кассет. Самый распространённый кластер повторов обнаружен в CRISPR-кассетах пяти разных индивидов, из одной и той же метагеномной коллекции JPN. Четыре кластера повторов обнаружены в индивидуальных микробиомах разных метагеномных коллекций (HMP и JPN). Эти повторы принадлежат CRISPR-кассетам индивидов, географически удалённых популяций: Япония и Европа. Тот факт, что такие общие повторы всё-таки удалось обнаружить, указывает на возможность существования широко распространенных бактериофагов.

Таким образом, отсутствие универсальных повторов и спейсеров говорит в первую очередь о высокой динамичности CRISPR-Cas систем в микробиомах кишечника человека. С ростом числа микробиомных данных и непрерывным улучшением технологий секвенирования, о наличии или отсутствии универсальных кассет можно будет судить с большей уверенностью.

Таблица 7. Общие кластеры повторов, т.е. кластеры, представленные, по меньшей мере, в двух разных индивидуальных метагеномах.

Кластер повторов	Индивиды с соответствующим повтором	Число различных индивидов
CLU1	F1S F2W F2X F2X IND INR INR INR INR	5
CLU2	F2W F2V F2V F2V IND IND IND IND INM	4
CLU3	F1T F1T INA IND MH0073	4
CLU4	F2X F2X F2X F2Y IND	3
CLU5	F2X F2Y O2.UC-13	3
CLU6	F2X F2X F2Y F2Y F2Y F2Y F2Y	2
CLU7	F2V F2V F2X F2Y	3
CLU8	F1T F1T F2V F2V	2
CLU9	IND IND INR	2
CLU10	F2V F2V MH0035	2
CLU11	F2X INE INE	2
CLU12	F2X INM INM	2
CLU13	F2Y F2Y INR	2
CLU14	F2Y INR INR	2
CLU15	AAQK AAQL	2
CLU16	F1S F2Y	2
CLU17	F1T F2W	2
CLU18	F2W F2X	2
CLU19	F2X IND	2
CLU20	MH0006 V1.UC-9	2
CLU21	MH0024 MH0041	2
CLU22	F1S IND	2
CLU23	F2W INA	2
CLLU24	INA INM	2

3.7 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. Для этой цели мы проанализировали объединённый (по данным трёх метагеномных проектов) 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. Очевидная колокализация спейсеров и протоспейсеров в рамках одного и того же индивидуального метагенома обнаружена только для трёх индивидов: F2Y, INB и INR. В указанных индивидах число идентифицированных пар спейсер-протоспейсер было особенно велико: 26, 49 и 14, соответственно. Вопреки ожиданиям, между метагеномами сиблингов (F2X и F2Y) пересекающихся пар спейсер-протоспейсер обнаружено не было.

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 15**). Возможное объяснение такого результата предоставляет протокол очистки метагеномной ДНК проекта HMP [10], согласно которому были отфильтрованы вирусные частицы и, следовательно, вирусные последовательности. Это объясняет то небольшое число протоспейсеров для HMP спейсеров, которое удалось обнаружить в том же наборе данных (по сравнению с числом протоспейсеров в метагеномной коллекции JPN).

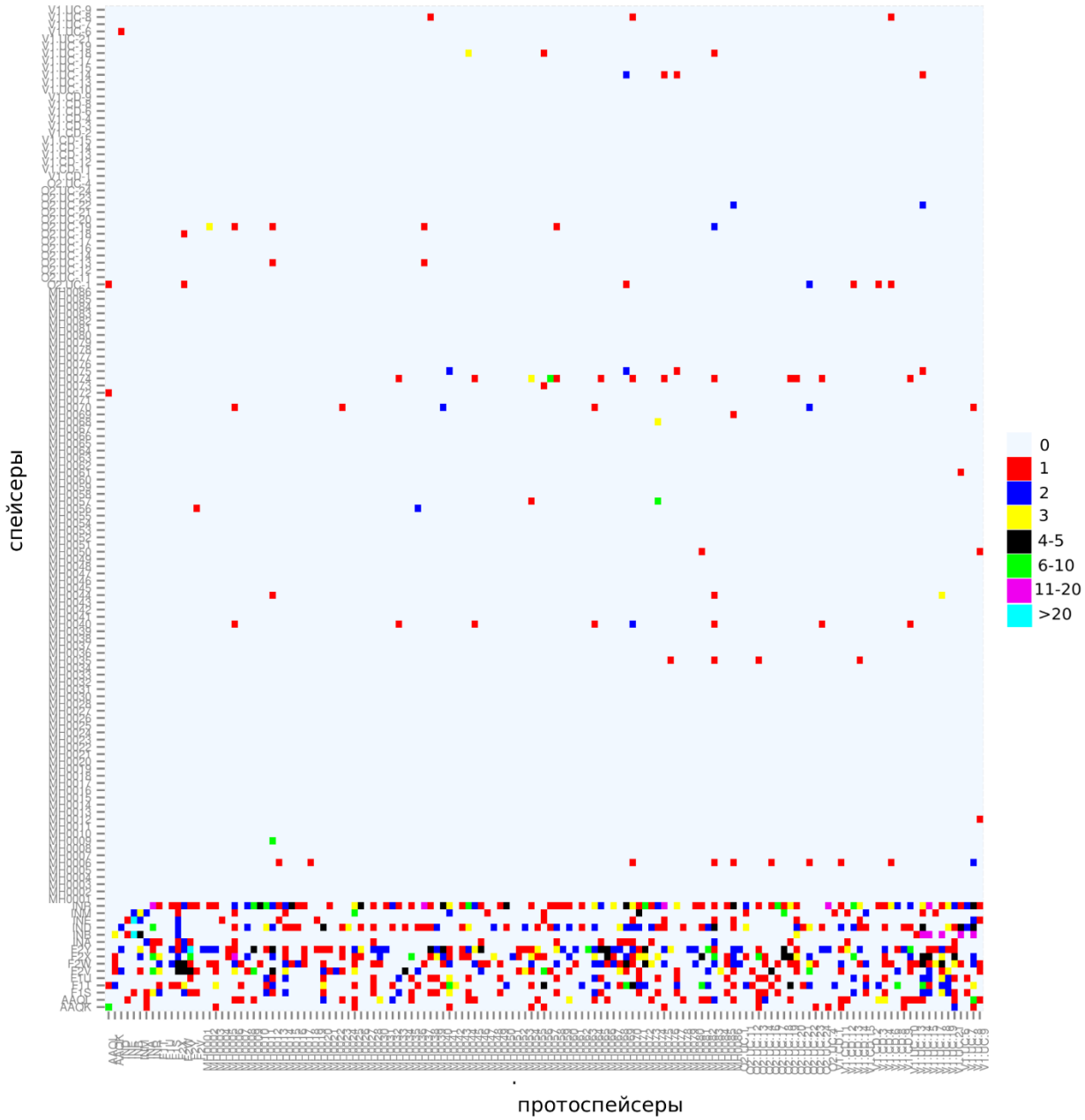


Рис.15. «Тепловая» карта, демонстрирующая распределение пар спейсер-протоспейсер по индивидуальным микробиомам человека. Цвета отражают число обнаруженных пар (расшифровано на рисунке).

Таким образом, мы обнаружили большое число пар спейсер-протоспейсер, разнесённых по разным индивидуальным метагеномам (**Рисунок 15**). Этот факт может говорить о том что, вирусы, ассоциированные с микробиомом кишечника человека, могут быть, в какой-то мере, универсальными. Более того, 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 CMH для наблюдаемых и симулированных наборов данных (см. Данные и алгоритмы). Значение критерия CMH для наблюдаемого набора данных составило 5,22. Распределение значений критерия CMH, рассчитанных для симулированных данных, приведено на **Рисунке 16**.

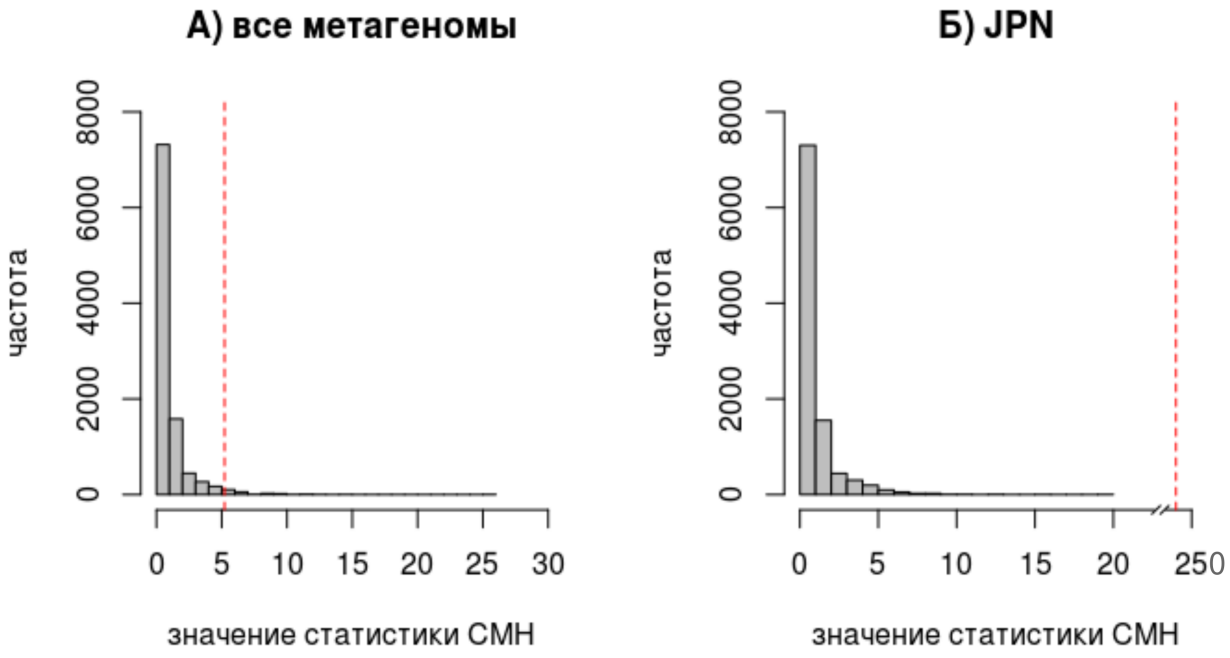


Рис.16. Распределение CMH-статистики, рассчитанное для проверки гипотезы о независимости распределения спейсеров и протоспейсеров по индивидуальным метагеномам для 10,000 случайных пермутаций спейсеров по индивидуальным метагеномам. А — для всех метагеномов; **Б** — только для метагеномов JPN**.** Красная пунктирная линия соответствует реальному значению статистики, полученному для наблюдаемых пар спейсер-протоспейсер в идентифицированных кассетах.

Как видно из рисунка, сравнение статистики CMH с нулевым распределением позволяет отклонить нулевую гипотезу при уровне значимости 0,05 (p-value = 0,0178) в пользу альтернативной. Что бы убедиться, что полученный результат не является следствием фильтрации от мелких вирусных частиц при приготовлении библиотеки, мы провели сходный анализ только для индивидуальных метагеномов коллекции JPN (нулевое распределение показано на рисунке **16 Б**). При этом значение статистики CMH составило 237 (p-value < 10^{-4}), что позволяет нам принять альтернативную гипотезу.

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 [12]. В ряде других работ, посвящённых анализу динамики CRISPR-систем в микробиоме ротовой полости человека [109]–[111] 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.

По сравнению с характером колокализации CRISPR спейсеров и протоспейсеров в других экосистемах (например, океаническом метагеноме) [101], микробиом человека является гораздо более однородным по составу CRISPR-элементов: для некоторых спейсеров удаётся обнаружить протоспейсеры в индивидуальных метагеномах, принадлежащих географически удалённым популяциям. Возможной причиной является бóльшая стабильность и единообразие физико-химических параметров в микробиоме человека в силу поддержания гомеостаза живым организмом (pH, температура, солёность, и т.д.) по сравнению с физико-химическими параметрами разных областей океана. Физико-химические параметры в пробах океанических образцов могут различаться очень сильно, что обуславливает различия микробного состава и, как следствие, CRISPR-систем данных мета-сообществ.

3.8 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-value < 0.0002) (см. Данные и алгоритмы) (**Рисунок 17, A)**. В тоже время общие спейсеры (т.е. спейсеры, встреченные в двух и более индивидуальных метагеномах) располагаются ближе к дистальному концу кассет (p-value < 0.001) (**Рисунок 17, B**). 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-кассет. Например, показано, что в ответ на фаговую инфекцию, *Streptococcus thermophilus* меняет состав CRISPR-кассеты: происходит добавление нового спейсера рядом с лидерной последовательностью [45], [48] и клетка становится устойчивой к этому фагу. Реконструкция CRISPR-кассет экстремофильных архей (I-plasma) показала, что спейсеры, находящиеся рядом с лидерной последовательностью, как правило, очень разнообразны, в то время как спейсеры, расположенные ближе к дистальному концу, более однородны и даже клональны в пределах одной популяции [137].

Наблюдаемую клональность состава спейсеров, расположенных ближе к дистальному концу кассеты, можно объяснить последовательным выметанием отбором (selective sweeps) [138]. Эффект выметания отбором заключается в снижении разнообразия последовательностей ДНК в окрестности мутации в результате недавнего положительного отбора [139]. 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. В последствии эти спейсеры будут сдвинуты ближе к дистальному концу кассеты и могут утратить мишени, если соответствующие вирусы окажутся элиминироваными из среды, в том числе в результате эффективной работы CRISPR-Cas системы.

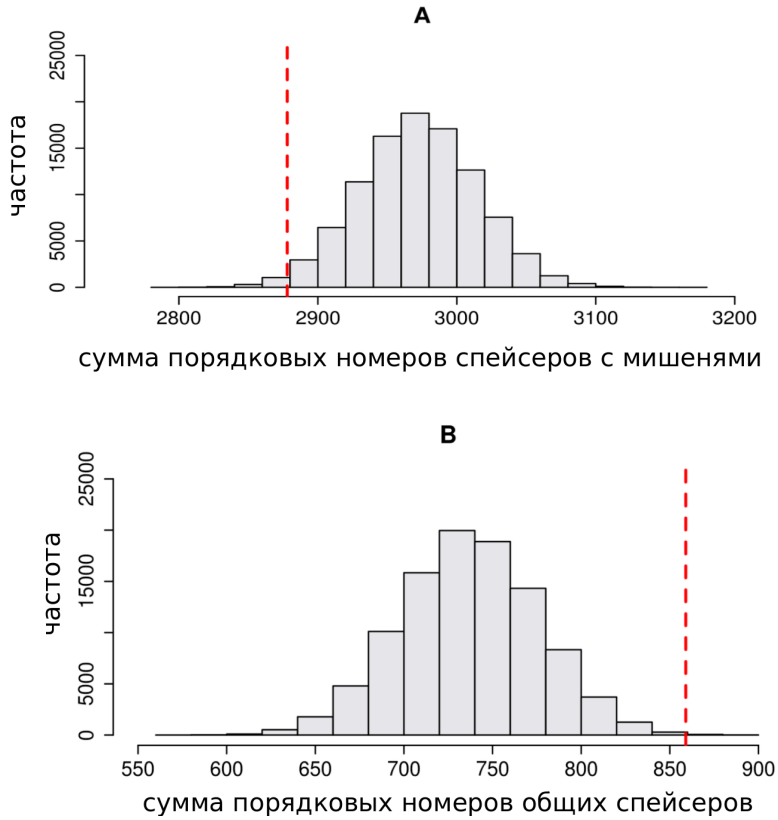


Рис.17. Сдвиг функционально значимых спейсеров по сравнению с расположением всех остальных спейсеров в кассете: (**A**) спейсеры с мишенями; (**B**) общие спейсеры. Представлены распределения сумм порядковых номеров функциональных спейсеров для симулированных наборов кассет (100'000 пермутаций). Красной пунктирной линией показано значение той же статистики расположения функционально значимых спейсеров в наблюдаемом наборе кассет (см. Данные и алгоритмы).

Глава 4. Гипотезы и перспективы

4.1. 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 [140], изучать совокупные наборы генов и геномов всех обитателей природного сообщества — метагеномы. Именно метагеномный подход позволил оценить не только масштаб истинного микробного разнообразия, но и определить метаболические возможности микробных сообществ, населяющих как глобальные экологические ниши, такие, как мировой океан, почву, тело человека, так и локальные, такие как горячие источники парка Йеллоустоун и многие другие [8-9,11][141]–[143] .

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. Примечательно, что большая часть видов (500-1000) принадлежит всего к двум основным типам — *Bacteroidetes* и *Firmicutes* [84]. Баланс такого сложного сообщества необходим для поддержания нормального функционирования организма человека [144].

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 [115]. Неограниченному размножению вирусов наряду с менее специфическими механизмами, такими как система абортивной инфекции (Abi) [145] и системы рестрикции-модификации [146], препятствуют CRISPR-Cas системы адаптивного прокариотического 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 — метагеном, можно попытаться смоделировать состав микробиома в прошлом, глядя только на структуру CRISPR-кассет его обитателей.

Кроме того, реконструируя CRISPR-кассеты из метагеномных данных, можно узнать больше о самих CRISPR-Cas системах, их биологии и разнообразии. На основании анализа доступных полных геномов прокариот, известно, что CRISPR-Cas системы широко распространены как среди бактерий, так и среди архей [36]. Отталкиваясь от положения о высоком видовом разнообразии микробиома, можно рассчитывать обнаружить большое разнообразие CRISPR-кассет и их типов в микробиомных данных. Учитывая, что микробиом человека во многом состоит из новых и мало исследованных видов [9], о геномах и CRISPR-Cas системах которых до сих пор известно немного, велик шанс найти новые кассеты и даже типы CRISPR-Cas систем в метагеномных данных микробиомов человека.

CRISPR-кассеты в метагеномных 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 [12]. 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 метагеномных 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 довольно 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 довольно 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 [8,9,11], отвечали этим критериям на момент начала исследования. Реконструируя CRISPR-

кассеты 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 [9]), 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-кассеты сами по себе сложно, так как они содержат повторы. Собирать CRISPR-кассеты 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) [11], [148]. Образцы проекта HMP секвенировали на платформе нового поколения, Illumina. Полученные короткие чтения (100 нт) собирали при помощи специализированного метагеномного ассемблера Metamos [147]. Алгоритм Metamos разбивает короткие чтения на еще более короткие k-меры, и использует их для построения графов Де Брейна [149]. 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.[11] 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.

Важно понимать, что технологии секвенирования и алгоритмы сборки геномных и метагеномных данных развиваются очень стремительно. Чтения становятся все длиннее, алгоритмы сборки — точнее и быстрее [150]. Все более и более популярной для секвенирования метагеномов природных сообществ становится технология ультра длинных чтений PacBIO [151]. Недостатком этой технологии пока остается довольное большое число ошибок [152]. Однако, в сочетании с короткими более точными чтениями, технология PacBIO позволила бы снизить долю неуверенности и неправильной сборки насыщенных повторами участков геномов, таких как CRISPR-кассеты.

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 [112], — актуальная задача, направленная на

оптимизацию существующих протоколов редактирования геномов и расширение спектра сходных задач (таких как глобальная регуляция транскрипции или направленная инициация кроссинговера [153], [154]. Например, нуклеазу cpf1 нашли в результате исключительно *in silico* поиска [155]. Cpf1 кодирует фермент, который компактнее и проще по сравнению с широко используемой нуклеазой Cas9, может узнавать разные PAM-последовательности и осуществляет разрез ДНК-мишени с образованием липких концов. Эти свойства позволяют, во-первых, расширить спектр возможных мишеней, а во-вторых — повысить эффективность рекомбинации после внесения разреза [156].

Таким образом, поиск CRISPR-кассет в метагеномных данных микробиомов человека, несмотря на технические сложности и ограничения, в целом оправдал себя. Работу в этом направлении стоит продолжать, особенно принимая во внимание появляющиеся новые технологии секвенирования и высокую актуальность поиска ассоциированных с кассетами *cas*-белков с новыми уникальными свойствами.

4.2. 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. Но насколько такой подход оправдан и уместен, насколько 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.

Начнем с того, что определение таксономии CRISPR-содержащих контигов — нетривиальная задача (см **Глава 3, Данные и Алгоритмы**). Для этой цели годятся лишь кодирующие участки контигов, не занятые ни кассетой, ни *cas*-генами. Однако в большом

числе случаев, даже если контиг и содержит достаточно длинные фланкирующие последовательности, определить таксономическую принадлежность на уровне рода, семейства и даже порядка невозможно, так как таксономическое положение лучших хитов различается. В таких случаях контигу с кассетой можно присвоить только неспецифическую категорию, объединяющую все лучшие хиты – общий тип или класс. Но часто такой категорией является только не очень информативная категория «Бактерии». Кроме того, зачастую мы попросту не находим хитов или находим хиты в плохо аннотированных последовательностях, многие из которых, как и наши данные, получены из природных изолятов. Таковы издержки работы с данными неизвестных организмов. В силу описанных сложностей, в нашей работе только 40% контигов с CRISPR-кассетами удалось приписать хоть какую-то таксономическую категорию.

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 [157]. Данному наблюдению можно привести несколько объяснений. Во-первых, не исключено, что численность известных представителей этого типа, оцененная на основании анализа последовательностей 16S рРНК [158], может быть завышена в силу вариабельности числа копий гена 16S рРНК. Во-вторых, бактерии типа *Bacteroidetes* возможно частзащищаются от вирусов при помощи альтернативных механизмов: систем рестрикции-модификации, за счет модификации поверхностных рецепторов, синтеза экзополисахаридных капсул или задействуя специфические противовирусные белки [159]. 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. И, наконец, третье объяснение, чисто

технического характера, заключается в том, что какая-то доля CRISPR-содержащих контигов, для которых удалось определить только неспецифическое происхождение, на самом деле может принадлежать типу *Bacteroidetes*. И это, возможно, выправило бы перекос и избавило бы нас от спекуляций.

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. Во-первых, косвенное и самое базовое представление о типах вирусов, распространённых 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 [160].

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 [160], 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 [114]. Последние попадают в микробиом человека с пищей, особенно если он придерживается растительного рациона. По-видимому, растительные РНК-вирусы являются транзиентными 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 вообще известно довольно мало: официально признано только два семейства — *Leviviridae* (хозяином служат энтеробактрии) и

Cystoviridae с одним единственным видом, инфицирующим *Pseudomonas* sp [161]. О РНК-бактериофагах микробиома человека известно еще меньше. 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 [162]. Оказалось, что, по крайней мере, у макак, РНК-бактериофаги не являются стабильными компонентами микробиома, для них характерны острые вспышки 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 [115], 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 виромными 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 (зонд) для поиска неизвестных ранее вирусов в непосредственно

микробиомных данных. Такой подход позволяет найти протоспейсеры уже для большего числа (**5.5%**) спейсеров. Что ожидаемо, так как метагеномные данные содержат как сами последовательности кассет, так и последовательности живущих тут же вирусов, с протоспейсерами в них. Однако происхождение подавляющего большинства спейсеров остается неизвестным. Этому может быть несколько объяснений. Во-первых, часть протоспейсеров не удается найти, так как часто при приготовлении библиотек для секвенирования образцов мелкие вирусные частицы отфильтровывают – 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. Реконструируя CRISPR-кассеты в микробиомных данных, сложно сказать, активны они или нет. Неактивные кассеты с неактуальными спейсерами могут сохраняться в геномах прокариот какое-то время [34].

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.

4.3 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 [9], [87]. 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 удается найти – это, в большинстве случаев, единичные общие спейсеры в кассетах с одинаковыми или очень похожими повторами, то есть в родственных кассетах в пределах каждого конкретного индивидуального микробиома. 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. Теоретически эту особенность роста CRISPR-кассет можно использовать даже в криминалистических целях.

Небольшое число общих спейсеров между индивидуальными метагеномами, в свою очередь, свидетельствует о высокой динамике CRISPR-кассет. Возможно, что 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 удается застать многие общие элементы. Сравнительное исследование CRISPR-кассет на протяжении длительного промежутка времени при контролируемой (или известной) вирусной нагрузке, возможно, позволило бы обнаружить большее сходство между индивидами на уровне CRISPR-кассет.

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. Как уже обсуждалось в предыдущем разделе, в микробиомных данных удается 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 «по дороге».

Глядя на распределение спейсеров и протоспейсеров по индивидуальным метагеномам, мы наблюдаем интересный эффект отталкивания: чаще всего в системе удается 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 рРНК: большая часть контигов (21%), содержащих CRISPR-кассеты отнесена к типу *Firmicutes*.

2. Сравнение обнаруженных спейсеров с известными виромами 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 колокализуются 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 сокращений и условных обозначений

CRISPR	clustered regularly interspaced short palindromic repeats; короткие палиндромные повторы, регулярно расположенные группами
Cas	CRISPR-associated; ассоциированный с CRISPR
RAMP	repeat associated mysterious proteins; загадочные белки, ассоциированные с повторами
crРНК	CRISPR РНК
мРНК	матричная РНК
рРНК	рибосомальная РНК
tracРНК	трейсерная РНК
PAM	protospacer-adjacent motif; мотив, расположенный рядом с протоспейсером
CMH	Cochran–Mantel–Haenszel test; тест Кохрана-Мантеля-Ханцеля
DG	Distal gut metagenomic project; Биом нисходящей ободочной кишки человека
HMP	Human Microbiome Project; проект «Микробиом человека»
JPN	Metagenome of 13 healthy japanese individuals; метагеном 13 здоровых японцев
VLP	virus-like particles; вирусоподобные частицы
PIL	программа PILER-CR
CFI	программа CRISPRFinder
CRT	программа CRISPR recognition tool
CRISPRdb	CRISPR database; база данных CRISPR
OTU	операционная таксономическая единица
NCBI	The National Center for Biotechnology Information
MetaHit	metagenomics of the human intestinal tract; метагеномный проект кишечного тракта человека

Список литературы

[1] Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, and a Nakata, “Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product.,” *J. Bacteriol.*, vol. 169, no. 12, pp. 5429–33, Dec. 1987.

[2] A. Bolotin, B. Quinquis, A. Sorokin, and S. D. Ehrlich, “Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin.,” *Microbiology*, vol. 151, no. Pt 8, pp. 2551–61, Aug. 2005.

[3] C. Pourcel, G. Salvignol, and G. Vergnaud, “CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies.,” *Microbiology*, vol. 151, no. Pt 3, pp. 653–63, Mar. 2005.

[4] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and E. Soria, “Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements.,” *J. Mol. Evol.*, vol. 60, no. 2, pp. 174–82, Feb. 2005.

[5] J. Peterson, S. Garges, M. Giovanni, P. McInnes, L. Wang, J. a Schloss, V. Bonazzi, J. E. McEwen, K. a Wetterstrand, C. Deal, C. C. Baker, V. Di Francesco, T. K. Howcroft, R. W. Karp, R. D. Lunsford, C. R. Wellington, T. Belachew, M. Wright, C. Giblin, H. David, M. Mills, R. Salomon, C. Mullins, B. Akolkar, L. Begg, C. Davis, L. Grandison, M. Humble, J. Khalsa, a R. Little, H. Peavy, C. Pontzer, M. Portnoy, M. H. Sayre, P. Starke-Reed, S. Zakhari, J. Read, B. Watson, and M. Guyer, “The NIH Human Microbiome Project.,” *Genome Res.*, vol. 19, no. 12, pp. 2317–23, Dec. 2009.

[6] P. D. Schloss and J. Handelsman, “Metagenomics for studying unculturable microorganisms: cutting the Gordian knot.,” *Genome Biol.*, vol. 6, no. 8, p. 229, Jan. 2005.

[7] J. S. Weitz and S. W. Wilhelm, “Ocean viruses and their effects on microbial communities and biogeochemical cycles.,” *F1000 Biol. Rep.*, vol. 4, no. September, p. 17, Jan. 2012.

[8] S. Gill, M. Pop, R. DeBoy, and P. Eckburg, “Metagenomic analysis of the human distal gut microbiome,” *Science (80-.).*, vol. 312, no. 5778, pp. 1355–1359, 2006.

[9] C. Huttenhower, D. Gevers, R. Knight, S. Abubucker et al, “Structure, function and diversity of the healthy human microbiome,” *Nature*, vol. 486, no. 7402, pp. 207–214, 2012.

[10] K. Li, M. Bihan, S. Yooseph, and B. A. Methe, “Analyses of the Microbial Diversity across the Human Microbiome,” vol. 7, no. 6, 2012.

[11] K. Kurokawa, T. Itoh, T. Kuwahara, K. Oshima, H. Toh, A. Toyoda, H. Takami, H. Morita, V. K. Sharma, T. P. Srivastava, T. D. Taylor, H. Noguchi, H. Mori, Y. Ogura, D. S. Ehrlich, K. Itoh, T. Takagi, Y. Sakaki, T. Hayashi, and M. Hattori, “Comparative metagenomics revealed

commonly enriched gene sets in human gut microbiomes.," *DNA Res.*, vol. 14, no. 4, pp. 169–81, Aug. 2007.

[12] A. Stern, E. Mick, I. Tirosh, O. Sagy, and R. Sorek, "CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome.," *Genome Res.*, vol. 22, no. 10, pp. 1985–94, Oct. 2012.

[13] E. Mick, A. Stern, and R. Sorek, "Holding a grudge: persisting anti-phage CRISPR immunity in multiple human gut microbiomes.," *RNA Biol.*, vol. 10, no. 5, pp. 900–6, 2013.

[14] M. Rho, Y.-W. Wu, H. Tang, T. G. Doak, and Y. Ye, "Diverse CRISPRs evolving in human microbiomes.," *PLoS Genet.*, vol. 8, no. 6, p. e1002441, Jan. 2012.

[15] M. G. Dominguez-Bello, E. K. Costello, M. Contreras, M. Magris, G. Hidalgo, N. Fierer, and R. Knight, "Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 26, pp. 11971–5, Jun. 2010.

[16] M. M. Harrison, B. V. Jenkins, K. M. O'Connor-Giles, and J. Wildonger, "A CRISPR view of development," *Genes Dev.*, vol. 28, no. 17, pp. 1859–1872, 2014.

[17] F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. a Scott, and F. Zhang, "Genome engineering using the CRISPR-Cas9 system.," *Nat. Protoc.*, vol. 8, no. 11, pp. 2281–308, Nov. 2013.

[18] A. S. Nilsson, "Phage therapy—constraints and possibilities," *Ups. J. Med. Sci.*, vol. 119, no. 2, pp. 192–198, 2014.

[19] P. M. Groenen, a E. Bunschoten, D. van Soolingen, and J. D. van Embden, "Nature of DNA polymorphism in the direct repeat cluster of Mycobacterium tuberculosis; application for strain differentiation by a novel typing method.," *Mol. Microbiol.*, vol. 10, no. 5, pp. 1057–65, Dec. 1993.

[20] N. P. Hoe, K. Nakashima, S. Lukomski, D. Grigsby, M. Liu, P. Kordari, S. J. Dou, X. Pan, J. Vuopio-Varkila, S. Salmelinna, a McGeer, D. E. Low, B. Schwartz, a Schuchat, S. Naidich, D. De Lorenzo, Y. X. Fu, and J. M. Musser, "Rapid selection of complement-inhibiting protein variants in group A Streptococcus epidemic waves.," *Nat. Med.*, vol. 5, no. 8, pp. 924–9, Aug. 1999.

[21] F. J. Mojica, C. Ferrer, G. Juez, and F. Rodríguez-Valera, "Long stretches of short tandem repeats are present in the largest replicons of the Archaea Haloferax mediterranei and Haloferax volcanii and could be involved in replicon partitioning.," *Mol. Microbiol.*, vol. 17, no. 1, pp. 85–93, Jul. 1995.

[22] E. Deltcheva, K. Chylinski, C. M. Sharma, and K. Gonzales, "Europe PMC Funders Group Europe PMC Funders Author Manuscripts CRISPR RNA maturation by trans -encoded small RNA and host factor RNase III," vol. 471, no. 7340, pp. 602–607, 2011.

[23] L. A. Marraffini and E. J. Sontheimer, “CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea,” *Nat Rev Genet*, vol. 11, no. 3, pp. 181–190, 2010.

[24] R. Sorek, V. Kunin, and P. Hugenholtz, “CRISPR--a widespread system that provides acquired resistance against phages in bacteria and archaea.,” *Nat. Rev. Microbiol.*, vol. 6, no. 3, pp. 181–6, Mar. 2008.

[25] V. Kunin, R. Sorek, and P. Hugenholtz, “Evolutionary conservation of sequence and secondary structures in CRISPR repeats.,” *Genome Biol.*, vol. 8, no. 4, p. R61, Jan. 2007.

[26] R. N. Jackson, S. M. Golden, P. B. G. van Erp, J. Carter, E. R. Westra, S. J. J. Brouns, J. van der Oost, T. C. Terwilliger, R. J. Read, and B. Wiedenheft, “Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli.,” *Science (80-.).*, vol. 345, no. 6203, pp. 1473–1479, 2014.

[27] Jeffrey R. Driscoll, “Spoligotyping for Molecular Epidemiology of the Mycobacterium tuberculosis Complex,” *Mol. Epidemiol. Microorg.*, vol. 551, pp. 117–128, 2009.

[28] Y. R. Vergnaud G, Li Y, Gorgé O, Cui Y, Song Y, Zhou D, Grissa I, Dentovskaya SV, Platonov ME, Rakin A, Balakhonov SV, Neubauer H, Pourcel C, Anisimov AP, “Analysis of the three Yersinia pestis CRISPR loci provides new tools for phylogenetic studies and possibly for the investigation of ancient DNA.,” *Adv Exp Med Biol.*, no. 603, pp. 327–38., 2007.

[29] N. O. Mokrousov I, Limeschenko E, Vyazovaya A, “Corynebacterium diphtheriae spoligotyping based on combined use of two CRISPR loci,” *Biotechnol J*, vol. 2, no. 7, pp. 901–6, 2007.

[30] L. M. Schouls, S. Reulen, B. Duim, A. Jaap, R. J. L. Willems, K. E. Dingle, M. Frances, J. D. A. Van Embden, J. A. Wagenaar, and F. M. Colles, “Comparative Genotyping of Campylobacter jejuni by Amplified Fragment Length Polymorphism , Multilocus Sequence Typing , and Short Repeat Sequencing : Strain Diversity , Host Range , and Recombination,” *J. Clin. Microbiol.*, vol. 41, no. 1, pp. 15–26, 2003.

[31] R. Jansen, J. D. a Van Embden, W. Gaastra, and L. M. Schouls, “Identification of genes that are associated with DNA repeats in prokaryotes.,” *Mol. Microbiol.*, vol. 43, no. 6, pp. 1565–75, Mar. 2002.

[32] J. SantaLucia and D. Hicks, “The thermodynamics of DNA structural motifs.,” *Annu. Rev. Biophys. Biomol. Struct.*, vol. 33, pp. 415–40, Jan. 2004.

[33] R. Rohs, S. M. West, A. Sosinsky, P. Liu, R. S. Mann, and B. Honig, “The role of DNA shape in protein-DNA recognition.,” *Nature*, vol. 461, no. 7268, pp. 1248–53, Oct. 2009.

[34] K. Pougach, E. Semenova, E. Bogdanova, K. a Datsenko, M. Djordjevic, B. L. Wanner, and K. Severinov, “Transcription, processing and function of CRISPR cassettes in Escherichia

coli.,” *Mol. Microbiol.*, vol. 77, no. 6, pp. 1367–79, Sep. 2010.

[35] D. H. Haft, J. Selengut, E. F. Mongodin, and K. E. Nelson, “A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes.,” *PLoS Comput. Biol.*, vol. 1, no. 6, p. e60, Nov. 2005.

[36] K. S. Makarova, D. H. Haft, R. Barrangou, S. J. J. Brouns, E. Charpentier, P. Horvath, S. Moineau, F. J. M. Mojica, Y. I. Wolf, A. F. Yakunin, J. van der Oost, and E. V Koonin, “Evolution and Classification of the CRISPR–Cas Systems,” *Nat. Rev. Microbiol.*, vol. 9, no. 6, pp. 467–477, 2011.

[37] K. H. Nam, F. Ding, C. Haitjema, Q. Huang, M. P. DeLisa, and A. Ke, “Double-stranded endonuclease activity in Bacillus halodurans clustered regularly interspaced short palindromic repeats (CRISPR)-associated Cas2 protein.,” *J. Biol. Chem.*, vol. 287, no. 43, pp. 35943–52, Oct. 2012.

[38] S. Lemak, B. Nocek, N. Beloglazova, T. Skarina, R. Flick, G. Brown, a. Joachimiak, a. Savchenko, and a. F. Yakunin, “The CRISPR-associated Cas4 protein Pcal_0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster: crystal structure and nuclease activity,” *Nucleic Acids Res.*, vol. 42, no. 17, pp. 11144–11155, Sep. 2014.

[39] K. S. Makarova, L. Aravind, N. V Grishin, I. B. Rogozin, and E. V Koonin, “A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis.,” *Nucleic Acids Res.*, vol. 30, no. 2, pp. 482–96, Jan. 2002.

[40] M. S. Dillingham and S. C. Kowalczykowski, “RecBCD enzyme and the repair of double-stranded DNA breaks.,” *Microbiol. Mol. Biol. Rev.*, vol. 72, no. 4, pp. 642–71, Table of Contents, Dec. 2008.

[41] R. B. Jensen, R. Lurz, and K. Gerdes, “Mechanism of DNA segregation in prokaryotes: replicon pairing by parC of plasmid R1.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 15, pp. 8550–5, Jul. 1998.

[42] S. Gudbergsdottir, L. Deng, Z. Chen, J. V. K. Jensen, L. R. Jensen, Q. She, and R. a Garrett, “Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers.,” *Mol. Microbiol.*, vol. 79, no. 1, pp. 35–49, Jan. 2011.

[43] P. Horvath, D. a Romero, A.-C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, and R. Barrangou, “Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus.,” *J. Bacteriol.*, vol. 190, no. 4, pp. 1401–12, Feb. 2008.

[44] A. Brodt, M. N. Lurie-Weinberger, and U. Gophna, “CRISPR loci reveal networks of gene exchange in archaea,” *Biol Direct*, vol. 6, no. 1, p. 65, 2011.

[45] R. Barrangou, C. Fremaux, H. Deveau, M. Richards, Patrick Boyaval, S. Moineau, D.

Romero, and P. Horvath, “CRISPRProvides Acquired Resistance Against Viruses in Prokaryotes,” *Science (80-.).*, vol. 315, no. 2007, pp. 1709–1712, 2007.

[46] T.-H. Tang, J.-P. Bachellerie, T. Rozhdestvensky, M.-L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Hüttenhofer, “Identification of 86 candidates for small non-messenger RNAs from the archaeon Archaeoglobus fulgidus.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 99, no. 11, pp. 7536–41, May 2002.

[47] T.-H. Tang, N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J. P. Bachellerie, and A. Hüttenhofer, “Identification of novel non-coding RNAs as potential antisense regulators in the archaeon Sulfolobus solfataricus.,” *Mol. Microbiol.*, vol. 55, no. 2, pp. 469–81, Jan. 2005.

[48] H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. a Romero, P. Horvath, and S. Moineau, “Phage response to CRISPR-encoded resistance in Streptococcus thermophilus.,” *J. Bacteriol.*, vol. 190, no. 4, pp. 1390–400, Feb. 2008.

[49] K. a Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, and E. Semenova, “Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system.,” *Nat. Commun.*, vol. 3, no. May, p. 945, Jan. 2012.

[50] E. Semenova, M. M. Jore, K. a Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, and K. Severinov, “Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108, no. 25, pp. 10098–103, Jun. 2011.

[51] R. K. Lillestøl, S. a Shah, K. Brügger, P. Redder, H. Phan, J. Christiansen, and R. a Garrett, “CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties.,” *Mol. Microbiol.*, vol. 72, no. 1, pp. 259–72, Apr. 2009.

[52] F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, and C. Almendros, “Short motif sequences determine the targets of the prokaryotic CRISPR defence system.,” *Microbiology*, vol. 155, no. Pt 3, pp. 733–740, Mar. 2009.

[53] A. Stern, L. Keren, O. Wurtzel, G. Amitai, and R. Sorek, “Self-targeting by CRISPR: Gene regulation or autoimmunity?,” *Trends Genet.*, vol. 26, no. 8, pp. 335–340, 2010.

[54] I. Yosef, M. G. Goren, and U. Qimron, “Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli.,” *Nucleic Acids Res.*, vol. 40, no. 12, pp. 5569–76, Jul. 2012.

[55] Y. A. F. Mohan Babu, Natalia Beloglazova, Robert Flick, Chris Graham, Tatiana Skarina, Boguslaw Nocek, Alla Gagarinova, Oxana Pogoutse1, Greg Brown1, Andrew Binkowski, Sadhna Phanse, Andrzej Joachimiak, Eugene V. Koonin3 Alexei Savchenko, Andrew Emili1, Jack Green, “A dual function of the CRISPR-Cas system in bacterial antivirus immunity and DNA repair,” *Mol. Microbiol.*, vol. 412, no. 2, pp. 426–434, 2011.

[56] M. M. Jore, M. Lundgren, E. van Duijn, J. B. Bultema, E. R. Westra, S. P. Waghmare, B. Wiedenheft, U. Pul, R. Wurm, R. Wagner, M. R. Beijer, A. Barendregt, K. Zhou, A. P. L. Snijders, M. J. Dickman, J. a Doudna, E. J. Boekema, A. J. R. Heck, J. van der Oost, and S. J. J. Brouns, “Structural basis for CRISPR RNA-guided DNA recognition by Cascade.,” *Nat. Struct. Mol. Biol.*, vol. 18, no. 5, pp. 529–36, May 2011.

[57] J. Reeks, J. H. Naismith, and M. F. White, “CRISPR interference: a structural perspective,” *Biochem. J.*, vol. 453, no. 2, pp. 155–166, 2013.

[58] C. R. Hale and M. O. Duff, “RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex,” *Cell*, vol. 139, no. 5, pp. 945–956, 2009.

[59] D. G. Sashital, B. Wiedenheft, and J. a Doudna, “Mechanism of foreign DNA selection in a bacterial adaptive immune system.,” *Mol. Cell*, vol. 46, no. 5, pp. 606–15, Jun. 2012.

[60] E. V Koonin and K. S. Makarova, “CRISPR-Cas: evolution of an RNA-based adaptive immunity system in prokaryotes.,” *RNA Biol.*, vol. 10, no. 5, pp. 679–86, 2013.

[61] Y. Yamaguchi and M. Inouye, *Chapter 12 mRNA Interferases, Sequence-Specific Endoribonucleases from the Toxin-Antitoxin Systems*, 1st ed., vol. 85, no. C. Elsevier Inc., 2009.

[62] K. S. Makarova, Y. I. Wolf, and E. V Koonin, “Comprehensive comparative-genomic analysis of type 2 toxin-antitoxin systems and related mobile stress response systems in prokaryotes.,” *Biol. Direct*, vol. 4, no. 1, p. 19, 2009.

[63] L. Van Melderen and M. Saavedra De Bast, “Bacterial toxin-antitoxin systems: more than selfish entities?,” *PLoS Genet.*, vol. 5, no. 3, p. e1000437, Mar. 2009.

[64] J. van der Oost, E. R. Westra, R. N. Jackson, and B. Wiedenheft, “Unravelling the structural and mechanistic basis of CRISPR-Cas systems.,” *Nat. Rev. Microbiol.*, vol. 12, no. 7, pp. 479–92, Jul. 2014.

[65] S. A. Shah, S. Erdmann, F. J. M. Mojica, and R. A. Garrett, “Protospacer recognition motifs: mixed identities and functional diversity.,” *RNA Biol.*, vol. 10, no. 5, pp. 891–9, 2013.

[66] M. F. White, “Structure, function and evolution of the XPD family of iron-sulfur-containing 5’-->3' DNA helicases.,” *Biochem. Soc. Trans.*, vol. 37, no. Pt 3, pp. 547–51, Jun. 2009.

[67] K. S. Makarova, N. V Grishin, S. A. Shabalina, Y. I. Wolf, and E. V Koonin, “A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action.,” *Biol. Direct*, vol. 1, no. 1, p. 7, 2006.

[68] T. Sinkunas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, and V. Siksnys, “Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system.,” *EMBO J.*, vol. 30, no. 7, pp. 1335–42, Apr. 2011.

[69] S. J. J. Brouns, M. M. Jore, M. Lundgren, E. R. Westra, R. J. H. Slijkhuis, A. P. L. Snijders, M. J. Dickman, K. S. Makarova, E. V Koonin, and J. van der Oost, “Small CRISPR RNAs guide antiviral defense in prokaryotes.,” *Science*, vol. 321, no. 5891, pp. 960–4, Aug. 2008.

[70] R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, and J. a Doudna, “Sequence- and structure-specific RNA processing by a CRISPR endonuclease.,” *Science*, vol. 329, no. 5997, pp. 1355–8, Sep. 2010.

[71] A. Jakubauskas, J. Giedriene, J. M. Bujnicki, and A. Janulaitis, “Identification of a single HNH active site in type IIS restriction endonuclease Eco31I.,” *J. Mol. Biol.*, vol. 370, no. 1, pp. 157–69, Jun. 2007.

[72] J. E. Garneau, M.-È. Dupuis, M. Villion, D. a Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadán, and S. Moineau, “The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA.,” *Nature*, vol. 468, no. 7320, pp. 67–71, Nov. 2010.

[73] K. M. Wylie, R. M. Truty, T. J. Sharpton, K. a Mihindukulasuriya, Y. Zhou, H. Gao, E. Sodergren, G. M. Weinstock, and K. S. Pollard, “Novel bacterial taxa in the human microbiome.,” *PLoS One*, vol. 7, no. 6, p. e35294, Jan. 2012.

[74] A. A. Saei and A. Barzegari, “The microbiome: the forgotten organ of the astronaut’s body--probiotics beyond terrestrial limits.,” *Future Microbiol.*, vol. 7, no. 9, pp. 1037–46, Sep. 2012.

[75] J. G. LeBlanc, C. Milani, G. S. de Giori, F. Sesma, D. van Sinderen, and M. Ventura, “Bacteria as vitamin suppliers to their host: a gut microbiota perspective.,” *Curr. Opin. Biotechnol.*, vol. 24, no. 2, pp. 160–8, Apr. 2013.

[76] D. a Ravcheev, A. Godzik, A. L. Osterman, and D. a Rodionov, “Polysaccharides utilization in human gut bacterium Bacteroides thetaiotaomicron: comparative genomics reconstruction of metabolic and regulatory networks.,” *BMC Genomics*, vol. 14, p. 873, Jan. 2013.

[77] a Nishimura, M. Fujimoto, S. Oguchi, R. D. Fusunyan, R. P. MacDermott, and I. R. Sanderson, “Short-chain fatty acids regulate IGF-binding protein secretion by intestinal epithelial cells.,” *Am. J. Physiol.*, vol. 275, no. 1 Pt 1, pp. E55–63, Jul. 1998.

[78] L. V. Hooper, D. R. Littman, and a. J. Macpherson, “Interactions Between the Microbiota and the Immune System,” *Science (80-.).*, vol. 336, no. 6086, pp. 1268–1273, 2012.

[79] C. L. Maynard, C. O. Elson, R. D. Hatton, and C. T. Weaver, “Reciprocal interactions of the intestinal microbiota and immune system.,” *Nature*, vol. 489, no. 7415, pp. 231–41, Sep. 2012.

[80] M. M. Finucane, T. J. Sharpton, T. J. Laurent, and K. S. Pollard, “A taxonomic signature of obesity in the microbiome? Getting to the guts of the matter.,” *PLoS One*, vol. 9, no. 1, p.

e84689, Jan. 2014.

[81] T. E. Sweeney and J. M. Morton, “The Human Gut Microbiome,” *JAMA Surg.*, vol. 148, no. 6, pp. 1–7, 2013.

[82] X. C. Morgan, T. L. Tickle, H. Sokol, D. Gevers, K. L. Devaney, D. V Ward, J. a Reyes, S. a Shah, N. LeLeiko, S. B. Snapper, A. Bousvaros, J. Korzenik, B. E. Sands, R. J. Xavier, and C. Huttenhower, “Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment.,” *Genome Biol.*, vol. 13, no. 9, p. R79, 2012.

[83] D. Mafra, J. C. Lobo, A. F. Barros, L. Koppe, N. D. Vaziri, and D. Fouque, “Role of altered intestinal microbiota in systemic inflammation and cardiovascular disease in chronic kidney disease,” *Future Microbiol.*, vol. 9, pp. 399–410, 2014.

[84] J. Qin, R. Li, J. Raes, M. Arumugam, K. S. Burgdorf, C. Manichanh, T. Nielsen, N. Pons, F. Levenez, T. Yamada, D. R. Mende, J. Li, J. Xu, S. Li, D. Li, J. Cao, B. Wang, H. Liang, H. Zheng, Y. Xie, J. Tap, P. Lepage, M. Bertalan, J.-M. Batto, T. Hansen, D. Le Paslier, A. Linneberg, H. B. Nielsen, E. Pelletier, P. Renault, T. Sicheritz-Ponten, K. Turner, H. Zhu, C. Yu, S. Li, M. Jian, Y. Zhou, Y. Li, X. Zhang, S. Li, N. Qin, H. Yang, J. Wang, S. Brunak, J. Doré, F. Guarner, K. Kristiansen, O. Pedersen, J. Parkhill, J. Weissenbach, P. Bork, S. D. Ehrlich, and J. Wang, “A human gut microbial gene catalogue established by metagenomic sequencing.,” *Nature*, vol. 464, no. 7285, pp. 59–65, Mar. 2010.

[85] P. J. Turnbaugh, M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan, R. E. Ley, M. L. Sogin, W. J. Jones, B. A. Roe, J. P. Affourtit, M. Egholm, B. Henrissat, A. C. Heath, R. Knight, and J. I. Gordon, “LETTERS A core gut microbiome in obese and lean twins,” *Nature*, vol. 457, no. 7228, pp. 480–484, 2009.

[86] D. N. Fredricks, T. L. Fiedler, and J. M. Marrazzo, “Molecular identification of bacteria associated with bacterial vaginosis.,” *N. Engl. J. Med.*, vol. 353, no. 18, pp. 1899–911, Nov. 2005.

[87] S. M. Huse, Y. Ye, Y. Zhou, and A. a Fodor, “A core human microbiome as viewed through 16S rRNA sequence clusters.,” *PLoS One*, vol. 7, no. 6, p. e34242, Jan. 2012.

[88] G. Hajishengallis, S. Liang, M. a Payne, A. Hashim, R. Jotwani, M. a Eskan, M. L. McIntosh, A. Alsam, K. L. Kirkwood, J. D. Lambris, R. P. Darveau, and M. a Curtis, “Low-abundance biofilm species orchestrates inflammatory periodontal disease through the commensal microbiota and complement.,” *Cell Host Microbe*, vol. 10, no. 5, pp. 497–506, Nov. 2011.

[89] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez,

C. Manichanh, H. B. Nielsen, T. Nielsen, N. Pons, and J. Poulain, “Enterotypes of the human gut microbiome,” *Nature*, vol. 12, no. 4737346, pp. 174–180, 2011.

[90] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, R. Sinha, E. Gilroy, K. Gupta, R. Baldassano, L. Nessel, and H. Li, “Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes,” *Science (80-.).*, vol. 334, no. October, pp. 105–109, 2011.

[91] Yatsunenko, T, F. Rey, M. Manary, I. Trehan, M. Dominguez-Bello, M. Contreras, M. Magris, G. Hidalgo, R. Baldassano, A. Anokhin, A. Heath, B. Warner, J. Reeder, J. Kuczynski, J. Caporaso, C. Lozupone, C. Lauber, J. Clemente, D. Knights, and R. Knight, “Human gut microbiome viewed across age and geography.,” *Nature*, vol. 486, no. 7402, pp. 222–227, 2012.

[92] N. P. McNulty, T. Yatsunenko, A. Hsiao, J. J. Faith, B. D. Muegge, A. L. Goodman, B. Henrissat, R. Oozeer, S. Cools-Portier, G. Gobert, C. Chervaux, D. Knights, C. a Lozupone, R. Knight, A. E. Duncan, J. R. Bain, M. J. Muehlbauer, C. B. Newgard, A. C. Heath, and J. I. Gordon, “The impact of a consortium of fermented milk strains on the gut microbiome of gnotobiotic mice and monozygotic twins.,” *Sci. Transl. Med.*, vol. 3, no. 106, p. 106ra106, Oct. 2011.

[93] J. E. Koenig, A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh, R. Knight, L. T. Angenent, and R. E. Ley, “Succession of microbial consortia in the developing infant gut microbiome.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 108 Suppl , pp. 4578–85, Mar. 2011.

[94] M. J. Claesson, I. B. Jeffery, S. Conde, S. E. Power, E. M. O’Connor, S. Cusack, H. M. B. Harris, M. Coakley, B. Lakshminarayanan, O. O’Sullivan, G. F. Fitzgerald, J. Deane, M. O’Connor, N. Harnedy, K. O’Connor, D. O’Mahony, D. van Sinderen, M. Wallace, L. Brennan, C. Stanton, J. R. Marchesi, A. P. Fitzgerald, F. Shanahan, C. Hill, R. P. Ross, and P. W. O’Toole, “Gut microbiota composition correlates with diet and health in the elderly.,” *Nature*, vol. 488, no. 7410, pp. 178–84, Aug. 2012.

[95] C. Burke, P. Steinberg, D. Rusch, S. Kjelleberg, and T. Thomas, “Bacterial community assembly based on functional genes rather than species,” 2011.

[96] B. L. Cantarel, V. Lombard, and B. Henrissat, “Complex carbohydrate utilization by the healthy human microbiome.,” *PLoS One*, vol. 7, no. 6, p. e28742, Jan. 2012.

[97] A. a Fodor, T. Z. DeSantis, K. M. Wylie, J. H. Badger, Y. Ye, T. Hepburn, P. Hu, E. Sodergren, K. Liolios, H. Huot-Creasy, B. W. Birren, and A. M. Earl, “The ‘most wanted’ taxa from the human microbiome for whole genome sequencing.,” *PLoS One*, vol. 7, no. 7, p. e41294, Jan. 2012.

[98] A. F. Andersson and J. F. Banfield, “Virus population dynamics and acquired virus resistance

in natural microbial communities.,” *Science*, vol. 320, no. 5879, pp. 1047–50, May 2008.

[99] B. Bolduc, D. P. Shaughnessy, Y. I. Wolf, E. V Koonin, F. F. Roberto, and M. Young, “Identification of novel positive-strand RNA viruses by metagenomic analysis of archaea-dominated Yellowstone hot springs.,” *J. Virol.*, vol. 86, no. 10, pp. 5562–73, May 2012.

[100] J. B. Emerson, K. Andrade, B. C. Thomas, A. Norman, E. E. Allen, K. B. Heidelberg, and J. F. Banfield, “Virus-host and CRISPR dynamics in Archaea-dominated hypersaline Lake Tyrrell, Victoria, Australia.,” *Archaea*, vol. 2013, p. 370871, Jan. 2013.

[101] V. a Sorokin, M. S. Gelfand, and I. I. Artamonova, “Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome.,” *Appl. Environ. Microbiol.*, vol. 76, no. 7, pp. 2136–44, Apr. 2010.

[102] M. E. Berg Miller, C. J. Yeoman, N. Chia, S. G. Tringe, F. E. Angly, R. a Edwards, H. J. Flint, R. Lamed, E. a Bayer, and B. a White, “Phage-bacteria relationships and CRISPR elements revealed by a metagenomic survey of the rumen microbiome.,” *Environ. Microbiol.*, vol. 14, no. 1, pp. 207–27, Jan. 2012.

[103] M. Breitbart, M. Haynes, S. Kelley, F. Angly, R. a Edwards, B. Felts, J. M. Mahaffy, J. Mueller, J. Nulton, S. Rayhawk, B. Rodriguez-Brito, P. Salamon, and F. Rohwer, “Viral diversity and dynamics in an infant gut.,” *Res. Microbiol.*, vol. 159, no. 5, pp. 367–73, Jun. 2008.

[104] F. H. Karlsson, V. Tremaroli, I. Nookaew, G. Bergström, C. J. Behre, B. Fagerberg, J. Nielsen, and F. Bäckhed, “Gut metagenome in European women with normal, impaired and diabetic glucose control.,” *Nature*, vol. 498, no. 7452, pp. 99–103, Jun. 2013.

[105] E. Le Chatelier, T. Nielsen, J. Qin, E. Prifti, F. Hildebrand, G. Falony, M. Almeida, M. Arumugam, J.-M. Batto, S. Kennedy, P. Leonard, J. Li, K. Burgdorf, N. Grarup, T. Jørgensen, I. Brandslund, H. B. Nielsen, A. S. Juncker, M. Bertalan, F. Levenez, N. Pons, S. Rasmussen, S. Sunagawa, J. Tap, S. Tims, E. G. Zoetendal, S. Brunak, K. Clément, J. Doré, M. Kleerebezem, K. Kristiansen, P. Renault, T. Sicheritz-Ponten, W. M. de Vos, J.-D. Zucker, J. Raes, T. Hansen, P. Bork, J. Wang, S. D. Ehrlich, O. Pedersen, E. Guedon, C. Delorme, S. Layec, G. Khaci, M. van de Guchte, G. Vandemeulebrouck, A. Jamet, R. Dervyn, N. Sanchez, E. Maguin, F. Haimet, Y. Winogradski, A. Cultrone, M. Leclerc, C. Juste, H. Blottière, E. Pelletier, D. LePaslier, F. Artiguenave, T. Bruls, J. Weissenbach, K. Turner, J. Parkhill, M. Antolin, C. Manichanh, F. Casellas, N. Boruel, E. Varela, A. Torrejon, F. Guarner, G. Denariaz, M. Derrien, J. E. T. van Hylckama Vlieg, P. Veiga, R. Oozeer, J. Knol, M. Rescigno, C. Brechot, C. M’Rini, A. Mérieux, and T. Yamada, “Richness of human gut microbiome correlates with metabolic markers.,” *Nature*, vol. 500, no. 7464, pp. 541–6, 2013.

[106] S. Nakamura, N. Maeda, I. M. Miron, M. Yoh, K. Izutsu, C. Kataoka, T. Honda, T. Yasunaga, T. Nakaya, J. Kawai, Y. Hayashizaki, T. Horii, and T. Iida, “Metagenomic diagnosis of bacterial infections.,” *Emerg. Infect. Dis.*, vol. 14, no. 11, pp. 1784–6, Nov. 2008.

[107] S. J. Song, C. Lauber, E. K. Costello, C. a Lozupone, G. Humphrey, D. Berg-Lyons, J. G. Caporaso, D. Knights, J. C. Clemente, S. Nakielny, J. I. Gordon, N. Fierer, and R. Knight, “Cohabiting family members share microbiota with one another and with their dogs.,” *Elife*, vol. 2, p. e00458, Jan. 2013.

[108] D. Willner, M. Furlan, M. Haynes, R. Schmieder, F. E. Angly, J. Silva, S. Tammadoni, B. Nosrat, D. Conrad, and F. Rohwer, “Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals.,” *PLoS One*, vol. 4, no. 10, p. e7370, Jan. 2009.

[109] D. T. Pride, C. L. Sun, J. Salzman, N. Rao, P. Loomer, G. C. Armitage, J. F. Banfield, and D. a Relman, “Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time.,” *Genome Res.*, vol. 21, no. 1, pp. 126–36, Jan. 2011.

[110] D. T. Pride, J. Salzman, and D. a Relman, “Comparisons of clustered regularly interspaced short palindromic repeats and viromes in human saliva reveal bacterial adaptations to salivary viruses.,” *Environ. Microbiol.*, vol. 14, no. 9, pp. 2564–76, Sep. 2012.

[111] R. Robles-Sikisaka, M. Ly, T. Boehm, M. Naidu, J. Salzman, and D. T. Pride, “Association between living environment and human oral viral ecology.,” *ISME J.*, vol. 7, no. 9, pp. 1710–24, Sep. 2013.

[112] Q. Zhang, T. G. Doak, and Y. Ye, “Expanding the catalog of cas genes with metagenomes.,” *Nucleic Acids Res.*, vol. 42, no. 4, pp. 2448–59, Feb. 2014.

[113] S. R. Gill, M. Pop, R. T. Deboy, P. B. Eckburg, P. J. Turnbaugh, B. S. Samuel, J. I. Gordon, D. a Relman, C. M. Fraser-Liggett, and K. E. Nelson, “Metagenomic analysis of the human distal gut microbiome.,” *Science*, vol. 312, no. 5778, pp. 1355–9, Jun. 2006.

[114] T. Zhang, M. Breitbart, W. H. Lee, J.-Q. Run, C. L. Wei, S. W. L. Soh, M. L. Hibberd, E. T. Liu, F. Rohwer, and Y. Ruan, “RNA viral community in human feces: prevalence of plant pathogenic viruses.,” *PLoS Biol.*, vol. 4, no. 1, p. e3, Jan. 2006.

[115] S. Minot, R. Sinha, J. Chen, H. Li, S. a Keilbaugh, G. D. Wu, J. D. Lewis, and F. D. Bushman, “The human gut virome: inter-individual variation and dynamic response to diet.,” *Genome Res.*, vol. 21, no. 10, pp. 1616–25, Oct. 2011.

[116] R. C. Edgar, “PILER-CR: fast and accurate identification of CRISPR repeats.,” *BMC Bioinformatics*, vol. 8, p. 18, Jan. 2007.

[117] C. Bland, T. L. Ramsey, F. Sabree, M. Lowe, K. Brown, N. C. Kyrpides, and P. Hugenholtz,

"CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats.," *BMC Bioinformatics*, vol. 8, p. 209, Jan. 2007.

[118] I. Grissa, G. Vergnaud, and C. Pourcel, "CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats," *Nucleic Acids Res.*, vol. 35, no. Web Server issue, pp. 52–57, 2007.

[119] C. T. Skennerton, M. Imelfort, and G. W. Tyson, "Crass: identification and reconstruction of CRISPR from unassembled metagenomic data.," *Nucleic Acids Res.*, vol. 41, no. 10, p. e105, May 2013.

[120] S. Altschul, W. Gish, W. Miller, E. Myers, and D. Lipman, "Basic local alignment search tool," *Mol Biol.*, vol. 215, no. 3, pp. 403–10, 1990.

[121] D. a Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank.," *Nucleic Acids Res.*, vol. 41, no. Database issue, pp. D36–42, Jan. 2013.

[122] C. Manichanh, L. Rigottier-Gois, E. Bonnaud, K. Gloux, E. Pelletier, L. Frangeul, R. Nalin, C. Jarrin, P. Chardon, P. Marteau, J. Roca, and J. Dore, "Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.," *Gut*, vol. 55, no. 2, pp. 205–11, Mar. 2006.

[123] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput.," *Nucleic Acids Res.*, vol. 32, no. 5, pp. 1792–7, Jan. 2004.

[124] W. Janet and S. Wallenstein, "The Power of the Mantel-Haenszel Test," *J. Am. Stat. Assos.*, vol. 82, no. 400, pp. 1104–1109, 1987.

[125] S. J. Lange, O. S. Alkhnbashi, D. Rose, S. Will, and R. Backofen, "CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems.," *Nucleic Acids Res.*, vol. 41, no. 17, pp. 8034–44, Sep. 2013.

[126] I. Grissa, G. Vergnaud, and C. Pourcel, "The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats.," *BMC Bioinformatics*, vol. 8, p. 172, Jan. 2007.

[127] C. Bang, K. Weidenbach, T. Gutsmann, H. Heine, and R. a Schmitz, "The intestinal archaea Methanosphaera stadtmanae and Methanobrevibacter smithii activate human dendritic cells.," *PLoS One*, vol. 9, no. 6, p. e99411, Jan. 2014.

[128] C. J. Smith, D. B. Nedwell, L. F. Dong, and a M. Osborn, "Evaluation of quantitative polymerase chain reaction-based approaches for determining gene copy and gene transcript numbers in environmental samples.," *Environ. Microbiol.*, vol. 8, no. 5, pp. 804–15, May 2006.

[129] T. Matsuki, K. Watanabe, J. Fujimoto, Y. Miyamoto, T. Takada, K. Matsumoto, H. Oyaizu, and R. Tanaka, "Development of 16S rRNA-gene-targeted group-specific primers for the

detection and identification of predominant bacteria in human feces," *Appl. Environ. Microbiol.*, vol. 68, no. 11, pp. 5445–5451, 2002.

[130] K. S. Makarova, Y. I. Wolf, and E. V Koonin, "Comparative genomics of defense systems in archaea and bacteria.," *Nucleic Acids Res.*, vol. 41, no. 8, pp. 4360–77, Apr. 2013.

[131] D. Schindler and H. Echols, "Retroregulation of the int gene of bacteriophage lambda: control of translation completion.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 78, no. 7, pp. 4475–9, Jul. 1981.

[132] K. R. Hargreaves, C. O. Flores, T. D. Lawley, and M. R. J. Clokie, "Abundant and Diverse Clustered Regularly Interspaced Short Palindromic Repeat Spacers in Clostridium difficile Strains and Prophages Target Multiple Phage Types within This Pathogen.," *MBio*, vol. 5, no. 5, Jan. 2014.

[133] N. L. Held, A. Herrera, H. C. Quiroz, and R. J. Whitaker, "CRISPR associated diversity within a population of Sulfolobus islandicus," *PLoS One*, vol. 5, no. 9, 2010.

[134] R. H. J. Staals, Y. Agari, S. Maki-Yonekura, Y. Zhu, D. W. Taylor, E. van Duijn, A. Barendregt, M. Vlot, J. J. Koehorst, K. Sakamoto, A. Masuda, N. Dohmae, P. J. Schaap, J. a Doudna, A. J. R. Heck, K. Yonekura, J. van der Oost, and A. Shinkai, "Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus.," *Mol. Cell*, vol. 52, no. 1, pp. 135–45, Oct. 2013.

[135] G. Vestergaard, R. a Garrett, and S. a Shah, "CRISPR adaptive immune systems of Archaea.," *RNA Biol.*, vol. 11, no. 2, pp. 156–67, Feb. 2014.

[136] G. J. Lax S, Smith DP, Hampton-Marcell J, Owens SM, Handley KM, Scott NM, Gibbons SM, Larsen P, Shogan BD, Weiss S7 Metcalf JL, Ursell LK, Vázquez-Baeza Y, Van Treuren W, Hasan NA, Gibson MK, Colwell R, Dantas G, Knight R, "Longitudinal analysis of microbial interaction between humans and the indoor environment," *Science (80-.).*, vol. 29, no. 345, pp. 1048–52, 2014.

[137] A. D. Weinberger, C. L. Sun, M. M. Pluciński, V. J. Denef, B. C. Thomas, P. Horvath, R. Barrangou, M. S. Gilmore, W. M. Getz, and J. F. Banfield, "Persisting viral sequences shape microbial CRISPR-based immunity.," *PLoS Comput. Biol.*, vol. 8, no. 4, p. e1002475, Jan. 2012.

[138] K. S. Korolev, M. J. I. Müller, N. Karahan, A. W. Murray, O. Hallatschek, D. R. Nelson, M. J. I. Muller, N. Karahan, A. W. Murray, O. Hallatschek, and D. R. Nelson, "Selective sweeps in growing microbial colonies," *Phys. Biol.*, vol. 9, no. 2, pp. 1–38, 2013.

[139] C.-J. Rubin, M. C. Zody, J. Eriksson, J. R. S. Meadows, E. Sherwood, M. T. Webster, L. Jiang, M. Ingman, T. Sharpe, S. Ka, F. Hallböök, F. Besnier, O. Carlborg, B. Bed'hom, M. Tixier-Boichard, P. Jensen, P. Siegel, K. Lindblad-Toh, and L. Andersson, "Whole-genome

resequencing reveals loci under selection during chicken domestication.,” *Nature*, vol. 464, no. 7288, pp. 587–91, Mar. 2010.

[140] M. S. Rappé and S. J. Giovannoni, “The uncultured microbial majority.,” *Annu. Rev. Microbiol.*, vol. 57, pp. 369–94, Jan. 2003.

[141] D. B. Rusch, A. L. Halpern, G. Sutton, K. B. Heidelberg, S. Williamson, S. Yooseph, D. Wu, J. A. Eisen, J. M. Hoffman, K. Remington, K. Beeson, B. Tran, H. Smith, H. Baden-Tillson, C. Stewart, J. Thorpe, J. Freeman, C. Andrews-Pfannkoch, J. E. Venter, K. Li, S. Kravitz, J. F. Heidelberg, T. Utterback, Y. H. Rogers, L. I. Falcón, V. Souza, G. Bonilla-Rosso, L. E. Eguiarte, D. M. Karl, S. Sathyendranath, T. Platt, E. Bermingham, V. Gallardo, G. Tamayo-Castillo, M. R. Ferrari, R. L. Strausberg, K. Nealson, R. Friedman, M. Frazier, and J. C. Venter, “The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific,” *PLoS Biol.*, vol. 5, no. 3, pp. 0398–0431, 2007.

[142] M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, A. Kara, B. A. Lynch, I. A. Macneil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Handelsman, R. M. Goodman, M. R. Rondon, P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. N. A. M. A. C. Neil, C. Minor, C. L. A. I. Tiong, M. Gilman, M. S. Osburne, J. O. N. Clardy, and J. O. Handelsman, “Cloning the Soil Metagenome : a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms Cloning the Soil Metagenome : a Strategy for Accessing the Genetic and Functional Diversity of Uncultured Microorganisms,” *Appl. Environ. Microbiol.*, vol. 66, no. 6, pp. 2541–2547, 2000.

[143] T. Schoenfeld, M. Patterson, P. M. Richardson, K. E. Wommack, M. Young, and D. Mead, “Assembly of viral metagenomes from Yellowstone hot springs,” *Appl. Environ. Microbiol.*, vol. 74, no. 13, pp. 4164–4174, 2008.

[144] M. C. Cénit, V. Matzaraki, E. F. Tigchelaar, and a Zhernakova, “Rapidly expanding knowledge on the role of the gut microbiome in health and disease.,” *Biochim. Biophys. Acta*, May 2014.

[145] R. L. Dy, R. Przybilski, K. Semeijn, G. P. C. Salmond, and P. C. Fineran, “A widespread bacteriophage abortive infection system functions through a Type IV toxin-antitoxin mechanism,” *Nucleic Acids Res.*, vol. 42, no. 7, pp. 4590–4605, 2014.

[146] I. Kobayashi, “Behavior of restriction-modification systems as selfish mobile elements and their impact on genome evolution,” *Nucleic Acids Res.*, vol. 29, no. 18, pp. 3742–3756, 2001.

[147] J. Qin, “A human gut microbial gene catalogue established by metagenomic sequencing: Commentary,” *Inflamm. Bowel Dis. Monit.*, vol. 11, no. 1, p. 28, 2010.

[148] X. Huang, J. Wang, S. Aluru, S. P. Yang, and L. Hillier, “PCAP: a whole-genome assembly

program,” *Genome Res.*, vol. 13, pp. 2164–2170, 2003.

[149] T. J. Treangen, S. Koren, D. D. Sommer, B. Liu, I. Astrovskaya, B. Ondov, A. E. Darling, A. M. Phillippy, and M. Pop, “MetAMOS: a modular and open source metagenomic assembly and analysis pipeline.,” *Genome Biol.*, vol. 14, no. 1, p. R2, 2013.

[150] S. Goodwin, J. D. McPherson, and W. R. McCombie, “Coming of age: ten years of next-generation sequencing technologies,” *Nat. Rev. Genet.*, vol. 17, no. 6, pp. 333–351, 2016.

[151] A. Rhoads and K. F. Au, “PacBio Sequencing and Its Applications,” *Genomics, Proteomics Bioinforma.*, vol. 13, no. 5, pp. 278–289, 2015.

[152] D. Laehnemann, A. Borkhardt, and A. C. McHardy, “Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction,” *Brief. Bioinform.*, vol. 17, no. 1, pp. 154–179, 2016.

[153] M. J. Sadhu, J. S. Bloom, L. Day, and L. Kruglyak, “CRISPR-directed mitotic recombination enables genetic mapping without crosses,” *bioRxiv*, vol. 5124, p. 19, 2016.

[154] A. A. Dominguez, W. A. Lim, and L. S. Qi, “Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation.,” *Nat. Rev. Mol. Cell Biol.*, vol. 17, no. 1, pp. 5–15, 2015.

[155] B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, A. Regev, E. V Koonin, F. Zhang, T. Pam, B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, and I. M. Slaymaker, “Cpf1 Is a Single RNA-Guided Endonuclease of a Class 2 CRISPR-Cas System,” *Cell*, pp. 1–13, 2015.

[156] D. Kim, J. Kim, J. Hur, K. W. Been, S. Yoon, and J.-S. Kim, “Genome-wide target specificities of Cpf1 nucleases in human cells,” *Nat. Biotechnol.*, vol. 9, no. February 2015, pp. 1–7, 2016.

[157] R. Ley, P. Turnbaugh, S. Klein, and J. Gordon, “Microbial ecology: human gut microbes associated with obesity.,” *Nature*, vol. 444, no. 7122, pp. 1022–3, 2006.

[158] H. Hayashi, M. Sakamoto, and Y. Benno, “Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods.,” *Microbiol. Immunol.*, vol. 46, no. 8, pp. 535–548, 2002.

[159] H. M. Wexler, “Bacteroides: The good, the bad, and the nitty-gritty,” *Clin. Microbiol. Rev.*, vol. 20, no. 4, pp. 593–621, 2007.

[160] M. Kazlauskiene, G. Amulaitis, G. Kostiuk, Č. Venclovas, and V. Siksnys, “Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition,” *Cell*, 2016.

[161] J. Bollback and J. Huelsenbeck, “Phylogeny, genome evolution, and host specificity of single-stranded RNA bacteriophage (family Leviviridae).,” *J Mol Evol*, vol. 52(2), pp. 117–28, 2001.

[162] S. R. Krishnamurthy, A. B. Janowski, G. Zhao, D. Barouch, and D. Wang, “Hyperexpansion of RNA Bacteriophage Diversity.,” *PLoS Biol.*, vol. 14, no. 3, p. e1002409, 2016.

Список иллюстративного материала

Глава 1

Глава 2

Глава 3

Таблица 6. Протоспейсеры, найденные в коллекции NR базы данных GenBank.

Рис.12. Расположение протоспейсера, соответствующего наиболее консервативной части гена, кодирующего белок, подобный белку Ea22 фага лямбда, в пяти родственных бактериофагах энтеробактерий.

Рис.13. Общие спейсеры в индивидуальных метагеномах коллекции JPN.

Рис.14. Распределение числа общих спейсеров между индивидами для 100'000 случайных пермутаций.

Таблица 7. Общие кластеры повторов, т. е. кластеры, представленные, по меньшей мере, в двух разных индивидуальных метагеномах.

Рис.15. «Тепловая» карта, демонстрирующая распределение пар спейсер-протоспейсер по индивидуальным микробиомам человека.

Рис.16. Распределение CMH-статистики, рассчитанное для проверки гипотезы о независимости распределения спейсеров и протоспейсеров по индивидуальным метагеномам для 100,000 случайных пермутаций спейсеров по индивидуальным метагеномам.

Рис.17. Сдвиг функционально значимых спейсеров по сравнению с расположением всех остальных спейсеров в кассете: (A) спейсеры с мишенями; (B) общие спейсеры.

Благодарности

Хочу выразить благодарность своему научному руководителю Ирене Игоревне Артамоновой, Михаилу Сергеевичу Гельфанду и Якову Давыдову, а также коллегам УНЦ «Биоинформатика» за ценные советы и помощь в выполнении работы.

Отдельно хочу поблагодарить Антона Уткина за помощь в подготовке иллюстраций.

Хочу также поблагодарить свою семью, друзей и кота, группу Dr Sebastian Schornack и участников журнального клуба «Bioinformatics without tears» за терпение, поддержку и сочувствие при подготовке диссертации.