Автономная некоммерческая образовательная организация высшего образования “Сколковский институт науки и технологий”

на правах рукописи

Шмаков Сергей Анатольевич

Разработка биоинформатического подхода для поиска новых CRISPR-Cas систем

Специальность 03.01.09 математическая биология, биоинформатика

Диссертация на соискание учёной степени

кандидата биологических наук

Научные руководители:

д.б.н., проф. Северинов Константин Викторович

к.б.н. Кунин Евгений Викторович

Москва – 2017

Оглавление

Список сокращений и условных обозначений

CRISPR - Clustered Regular Interspaced Short Palindromic Repeats, короткие палиндромные повторы, регулярно расположенные группами

CRISPR-Cas - CRISPR Associated, CRISPR ассоциированные

Спэйсер – последовательность между повторами в CRISPR кассете

Протоспэйсер – целевая последовательность комплементарная спэйсеру

PAM - Protospacer Adjacent Motif, последовательность фланкирующая протоспэйсер

crРНК – CRISPR РНК

crRNP – CRISPR ribonucleoprotein complex, CRISPR рибонуклеопротеиновый комплекс

pre-crРНК – pre-CRISPR РНК, РНК транскрипт CRISPR кассеты

tracrРНК – *trans*-acting CRISPR РНК, небольшая кодирующая область расположенная в отдалении от CRISPR

ORF – Open Reading Frame, белок-кодирующий участок

HEPN – Higher Eukaryotes and Prokaryotes Nucleotide-binding domains, домен связывающийся с ДНК присутствующий в эукариотах и прокариотах

NHEJ – non-homologous ends joining, негомологичное соединение концов

HR – homologous recombination, соединение концов ДНК путем гомологичной рекомбинации

DSB – double strand break, двунитевой разрыв ДНК

WGS – Whole Genome Shotgun projects database, база данных неполных геномов

sgРНК – single guide РНК, РНК последовательность, послученная путем слияния crРНК и tracrРНК

Введение

Актуальность работы

CRISPR-Cas это разнообразные адаптивные иммунные системы в бактериях и археях [9, 11, 111, 124, 128]. Эти системы недавно привлекли много внимания из-за их уникального “Ламарковского” поведения [98]: они сохраняют память (спэйсеры) от предыдущих инфекций, что дает специфичную защиту к этим же инфекциям через процесс использующий узнавание по РНК. Этот механизм был успешно и эффективно применён для редактирования геномов [158]. Механизм действия и структурные особенности CRISPR-Cas систем детально представлены в нескольких недавних обзорах [9, 124, 128, 145].

CRISPR-Cas системы обладают огромным разнообразием состава Cas белков, а также разновидной структурой геномных локусов [112, 114]. Но, несмотря на это разнообразие, CRISPR-Cas системы обладают одинаковым набором основных свойств, что может означать моно филетическое происхождение. Основные Cas белки могут быть сгруппированы в два основных функциональных модуля: адаптационный модуль, данный модуль добавляет новый генетический материал в CRISPR кассеты и эффекторный модуль, который уничтожает целевую последовательность. Адаптационные модули в основном сходны во всех CRISPR-Cas системах, и состоят из двух обязательных белков Cas1, Cas2 и иногда Cas4. Эффекторные модули же, напротив, показывают сильное разнообразие. Механизм действия CRISPR-Cas систем может быть разделён на три активности: адаптация, биогенезис crРНК и интерференция. Во время адаптации Cas1-Cas2 белковый комплекс (в некоторых случаях он может содержать дополнительные субъединицы) выделяет сегмент из целевой ДНК (этот сегмент называют протоспэйсером) и встраиваетяет этот сегмент в 5’ конец CRISPR кассеты между повторами, таким образом, добавляя новый спэйсер. На этапе экспрессии и обработки CRISPR кассеты (или биогенезис crРНК), кассета со спэйсерами транскрибируется в длинный транскрипт, известный как пре-CRISPR РНК (pre-crРНК),

который далее процессируется с помощью определённых Cas белков (которые, в отдельных случаях, требуют дополнительных белков и РНК молекул) с образованием коротких фрагментов CRISPR РНК (crРНК). Эффекторный модуль, направляемый crРНК, таргетирует (специфично распознает) и разрезает чужеродные молекулы нуклеиновых кислот [112, 116]. Во избежание самотаргетрования CRISPR-Cas системы учитывают мотив, расположенный рядом с протоспэйсером (protospacer adjacent motif (PAM)) – несколько нуклеотидов, располагающихся непосредственно перед протоспэйсером. Этот мотив одинаков для всех протоспэйсеров из одной CRISPR кассеты и определён эффекторным комплексом.

Последняя классификация CRISPR-Cas систем разделяет их на два класса, пять типов и шестнадцать подтипов, исходя из архитектуры эффекторных модулей [114]. Системы первого класса, которые включают в себя I и III типы, плюс предполагаемая система IV типа, обладают мульти-блоковыми эффекторными комплексами, которые составлены из нескольких Cas белков. Второй класс CRISPR-Cas систем, который включает в себя тип II и предполагаемый тип V, характеризуется наличием эффекторного комплекса, состоящего из одного большого блока – большого Cas белка.

Эффекторные комплексы первого типа CRISPR-Cas систем состоят из 4-7 блоков Cas белков в неравной стехиометрии, как показано на примерах CRISPR-ассоциированных комплексов для антивирусной защиты (Cascade) для систем первого типа [12, 22, 79, 84], и Csm–Cmr комплексов систем III типа [146, 163, 183, 189]. Для второго класса, это Cas9 - эффекторный белок второго типа CRISPR-Cas систем, это большая, мульти-доменная нуклеаза, которая варьируется по размеру, в зависимости от организма от ~950 до более чем 1,600 аминокислот и содержит два типа нуклеазных доменов: RuvC подобный домен (РНКаза H консервативный домен) и HNH (MrcA подобный консервативный домен) домен [111], которые используются для разрезания таргетируемой ДНК [11, 36, 51, 52, 82, 167]. Этот многофунцкиональный белок был использован как ключевой инструмент для редактирования геномов. Недавно, второй эффекторный белок второго класса – Cpf1, который содержит RuvC домен, но не HNH домен [114, 169], был обнаружен и было показано, что он также является РНК направляемой эндонуклеазой, которая разрезает таргетируемую ДНК, но цепи разрезаются в разных местах [205]. В связи со своей

уникальной архитектурой, Cpf1-содержащие системы были классифицированы как пятый тип CRISPR-Cas систем [114].

CRISPR-Cas белки, такие как crРНК направляемая нуклеаза cas9 [52], значительно улучшили эффективность редактирования геномов в связи с простотой их использования и высокой специфичности распознавания цели [28]. Однако, свойства существующих эффекторых комплексов не всегда оптимальны [177], что означает, что есть необходимость в новых белковых семействах, которые могли бы расширить и обогатить возможности CRISPR-Cas инструментов. Недавно обнаруженный Cpf1 эффекторный комплекс [205], который был позже использован для редактирования геномов благодаря его уникальным свойствам [206], показывает, что тщательное исследование геномных и метагеномных данных необходимо и ожидаемо. Подобное исследование было проведено для полных геномов (полностью отсеквенированных геномов) [114], которое показало, что основные варианты CRISPR-Cas систем уже известны, однако не полные геномы (частично отсеквенированные) и метагеномные базы данных не были покрыты, таким образом, новые редкие системы могут быть обнаружены в этих данных [114]. Механизм действия второго класса CRISPR-Cas систем достаточно хорошо охарактеризован [128], но их происхождение остается недостаточно объяснённым. Новые варианты CRISPR-Cas систем могут иметь иные свойства, что может дать новые данные для характеризации защитных систем в прокариотах, вирус-хост взаимодействиям, эволюции и функционированию CRISPR-Cas систем, а также новые варианты могут быть использованы как биотехнологические инструменты.

Цели и задачи исследования

Основной целью данного проекта была оценка разнообразия CRISPR-Cas систем второго класса – систем с большим (> 500 аминокислот) одноблоковым эффекторным комплексом и обнаружение новых CRISPR-*cas* генов второго класса или вариантов известных белков среди бактерий и архей, используя тщательное исследование доступных геномных и метагеномных данных. Для достижения разрешения

поставленных целей, был разработан биоинформатический подход, который использует ключевые компоненты CRISPR-Cas систем для поиска элементов ассоциированных с ними.

В геномных данных были определены и проанализированы два ключевых компонента, которые использовались в качестве затравки для поиска новых вариантов CRISPR-Cas систем:

1 – *cas1*: Данный компонент был выбран из-за того, что он является наиболее консервативным среди CRISPR-*cas* генов. Для данного гена были определены белковые профиля высокого качества [113, 187], которые могут быть использованы для поиска Cas1 в геномных данных. Факт того, что филогения *cas1* коррелирует с CRISPR-Cas типами [113] может быть использован для определения систем новых типов. Этот компонент является обязательным для многих CRISPR-Cas систем: чтобы быть адаптивной, CRISPR-Cas система должна иметь возможность встраивать новые спэйсеры, что является ключевой функцией *cas1* (большинство CRISPR-Cas локусов содержат этот ген [112, 114]). Таким образом, одиночные Cas1 ассоциированные с неизвестными белками являются целью данного исследования.

2 – CRISPR кассеты: это ключевой элемент любой CRISPR-Cas системы, который используется как база данных содержащая спэйсеры [132]. Большинство CRISPR-Cas локусов имеют кассеты рядом с *cas* генами [114], таким образом, новые системы могут быть определены по ассоциации с найденными CRISPR кассетами.

Известно, что есть Cas1 белки и CRISPR кассеты, в локусах которых не присутствует эффекторных белков [114], что может осложнить использование данного подхода путём добавления шума в набор соседских генов, но не смотря на это, основная часть Cas1 и CRISPR кассет ассоциированы с другими *cas* генами. Присутствие этих компонентов может указать на новый CRISPR-Cas локус, таким образом определив новые *cas* гены используя принцип “виновности по ассоциации”.

Для выполнения данного исследования, были определены следующие задачи:

1. Определение затравок – Cas1 или CRISPR позиции в бактериальных или архейных геномах и дальнейшее использование их в качестве якорных позиций в геноме. Данный поиск должен использовать следующие программные инструменты и базы данных: PSI-BLAST [3] для определения позиций Cas1; CRISPRFinder [59] и PILER-CR [46] для поиска CRISPR кассет; доступные открытые базы данных, такие как GenBank [13] и Whole Genome Shotgun projects database (WGS) [211] должны быть использованы в качестве пространства для поиска Cas1 и CRISPR кассет;
2. Определение открытых рамок считывания (ORFs) вокруг якорей используя GenMark [14] и их аннотация с использованием CDD [120, 121] белковых профилей и RPS-BLAST [123] для дальнейшей фильтрации;
3. Определение типов CRISPR-Cas систем для *cas* генов вокруг затравок (если такие присутствуют) используя подход предложенный Макаровой К.С. [113] для того, чтобы отфильтровать известные системы или выделить системы с не полным эффекторным комплексом;
4. Кластеризация всех белков используя UCLUST [47] для группировки белков и оценки разнообразия соседских генов;
5. Ручное курирование белковых кластеров используя чувствительные инструменты для определения белковых доменов таких как: HHpred [180] и CD-Search [120];
6. Выделение и биоинформатическая характеризация списка кандидатов путём определения белковых доменов в кандидатах, которые необходимы в CRISPR-Cas эффекторных комплексах (такие как нуклеазные домены или ДНК/РНК связывающие домены);

Для определения новых CRISPR-Cas списка кандидатов, которые были посланы на экспериментальную валидацию, эффекторных комплексов были выполнены шаги 1-6. Шаги 1-4 были достаточны для полной оценки CRISPR-Cas систем в доступных базах данных.

Научная новизна и практическая значимость работы

В данной работе была впервые проведена оценка разнообразия CRISPR-Cas систем на большом наборе прокариотических геномных и метагеномных данных доступных на март 2016 года, в то время как предыдущие работы были выполнены на более ограниченном наборе данных [114]. Данная оценка позволила обнаружить новые CRISPR-Cas системы: Тип V-B, Тип V-C, предполагаемый подтип Type V-U содержащий RuvC нуклеазный домен, таргетирование ДНК было экспериментально показано для типа V-B [173]. Также был обнаружен полностью новый Тип VI который включает HEPN (higher eukaryotes and prokaryotes nucleotide-binding domains) домен, в типе VI было выделено 4 подтипа: Тип VI-A, Тип VI-B и Тип VI-C, таргетирование РНК было экспериментально показано для типа VI-A [1] и типа VI-B [178].

Обнаруженные CRISPR-Cas системы могут использоваться в различных прикладных областях где специфичность к целевой ДНК или РНК необходима, например для задач редактирования геномов. Новые подтипы пятого типа отличаются от хорошо изученного CRISPR-Cas Тип II (Cas9) другой архитектурой, а также отличаются от недавно функционально охарактеризованного V-A типа (Cpf1) [205] за счет присутвия tracrРНК (transactivating crРНК) в V-B типе, что может помочь занять нишу среди инструментов для редактирования геномов. РНКазные домены в VI типе могут быть использованы для детектирования молекул РНК, также VI тип может быть использован для понижения экспрессии генов на подобии РНК интерференции. Маленький размер предсказанных эффекторных белков в V-U типе может позволить размещать из в более компактные средства доставки (например, вирусные капсиды), что даст определённое преимущество для редактирования геномов. Также все обнаруженные системы обладают уникальной специфичностью распознавания цели и разнообразным PAM мотивом.

Недавние исследования показали эффективное применение обнаруженного Cas13a (Тип VI-A) из *Leptotrichia wadei* для высокоспецифичного определения молекул нуклеиновых кислот [58]. Также было показано применение этого белка для детектирования вирусов, патогенных бактерий, определения раковых клеток и т.д.

Личный вклад соискателя

Большая часть исследования была проведена автором диссертации. Был разработан и имплементирован биоинформатический подход для оценки разнообразия в прокариотических геномных и метагеномных данных, что позволило обнаружить новые ранее неизвестные варианты CRISPR-Cas систем. Различные бактериальные и архейные базы данных были просканированы на наличие CRISPR кассет и Cas1 белков (затравок). Локусы вокруг затравок были проаннотированы и просканированы на наличие неизвестных белков, ассоциированных с компонентами CRISPR-Cas систем (Cas1 и CRISPR кассетами). Был определён список кандидатов в новые CRISPR-Cas системы. Новые системы были биоинформатически охарактеризованы и предложены для экспериментальной валидации.

Положения, выносимые на защиту

1. Биоинформатический подход для поиска CRISPR-Cas эффекторных комплексов второго класса.
2. Обнаружение шести новых CRISPR-Cas систем: Тип V-B, Тип V-C, Тип V-U, Тип VI-A, Тип VI-B, Тип VI-C.
3. Обновлённая классификация CRISPR-Cas систем с учётом шести новых подтипов.
4. Исчерпывающая оценка разнообразия CRISPR-Cas систем второго типа в прокариотических геномных данных.
5. Гипотеза возможного происхождения CRISPR-Cas систем второго класса.
6. Возможные варианты применения обнаруженных CRISPR-Cas систем.

Степень достоверности и апробация результатов

Результаты были представлены на трёх научных конференциях и опубликованы в рецензируемых научных журналах.

Публикации:

1. Shmakov S, Smargon A, Scott D, Cox D, Pyzocha N, Yan W, Abudayyeh OO, Gootenberg JS, Makarova KS, Wolf YI, Severinov K, Zhang F, Koonin EV. Diversity and evolution of Class 2 CRISPR-Cas systems. // Nature Reviews Microbiology. – 2017. – Т. 15. – № 3. – С. 169-182.
2. Smargon AA, Cox DB, Pyzocha NK, Zheng K, Slaymaker IM, Gootenberg JS, Abudayyeh OA, Essletzbichler P, Shmakov S, Makarova KS, Koonin EV, Zhang F. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. // Molecular cell. – 2017. – Т. 65. – № 4. – С. 618-630.
3. Abudayyeh OO, Gootenberg JS, Konermann S, Joung J, Slaymaker IM, Cox DB, Shmakov S, Makarova KS, Semenova E, Minakhin L, Severinov K, Regev A, Lander ES, Koonin EV, Zhang F. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. // Science. – 2016. – Т. 353. – С. 6299.
4. Shmakov S, Abudayyeh OO, Makarova KS, Wolf YI, Gootenberg JS, Semenova E, Minakhin L, Joung J, Konermann S, Severinov K, Zhang F, Koonin EV. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. // Molecular Cell – 2015. – Т. 60. – № 3. – С. 385-97.

Постер был представлен на следующих международных конференциях: Genome Engineering 4.0 (США, май 2016), CRISPR 2016 (Израиль, май 2016) и CRISPR 2017 (США, июнь 2017).

Обзор литературы

CRISPR-Cas системы

Механизм действия

CRISPR (clustered regularly interspaced short palindromic repeat - короткие палиндромные повторы, регулярно расположенные группами)-Cas (CRISPR ASsociated proteins - CRISPR ассоциированные) являются адаптивными иммунными системами в бактериях и археях [9, 11, 111, 124, 128]. Около 90% архей и 40% бактерий обладают этими защитными системами [15, 114]. Данные системы предоставляют защиту против вирусной ДНК [11, 111] и РНК [62] через трёх-этапный процесс состоящий из следующих механизмов: адаптация, биогенезис crРНК и интерференция. Схематичное представление этих процессов представлено на Рисунке 1. Данные процессы были тщательно представлены в недавних обзорах механизма действия CRISPR-Cas систем [124, 128].

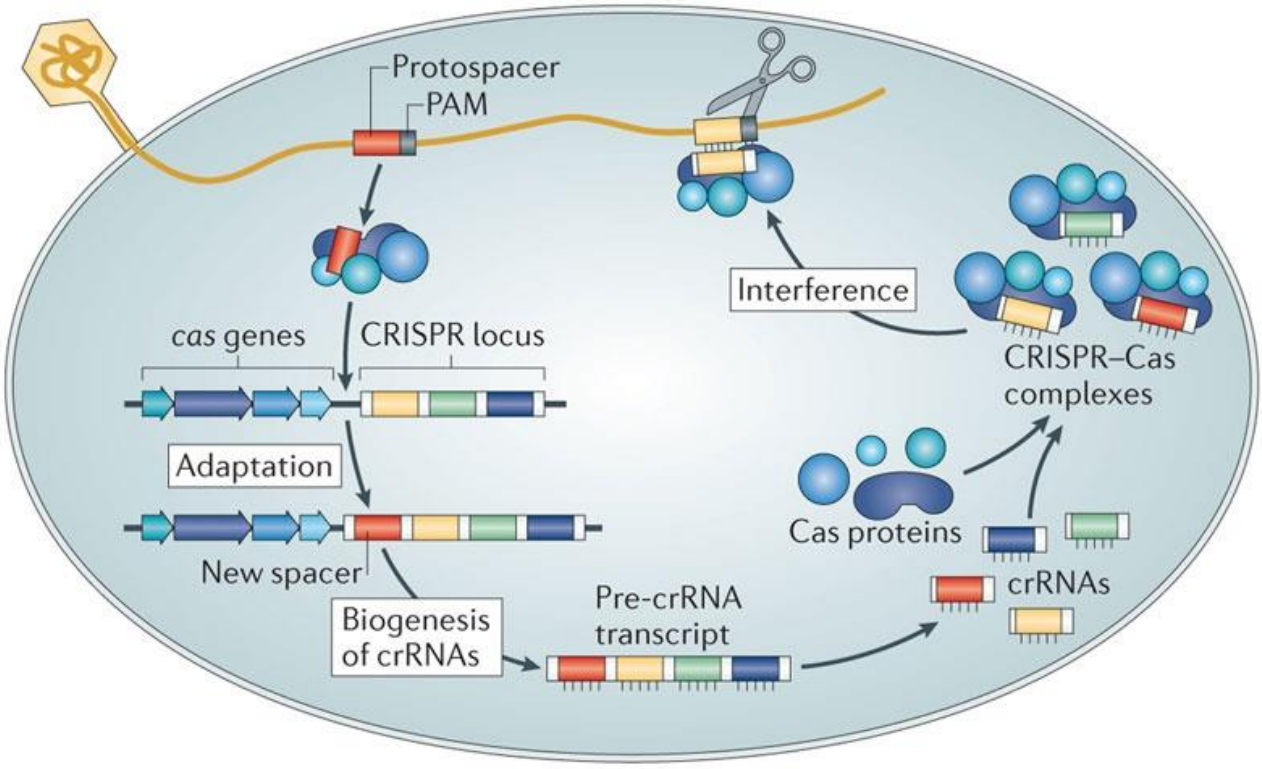


Рисунок 1. Схема функционирования CRISPR-Cas систем (воспроизведено с разрешения Nature Reviews Microbiology [166]). Три основные функции CRISPR-Cas систем представлены на рисунке: Адаптация (встраивание новых спэйсеров в CRISPR кассету), Биогенезис crРНКs (экспрессия и созревание crРНК) и CRISPR-Cas интерференция (распознавание и деградация чужеродной ДНК или РНК).

Адаптация – это процесс встраивания нового спэйсера в CRISPR кассету. Белки Cas1 и Cas2 (ядро адаптационного модуля) образуют структуру из двух димеров Cas1 и одного димера Cas2 [142], данный комплекс связывается с протоспэйсером [141, 192]. Недавние исследования выдвигают гипотезу, что протоспэйсеры появляются из остатков чужеродной ДНК после репарации с помощью механизма починки двухцепочечных разрывов (экзонуклеазный RecBCD комплекс [101]), который поставляет одноцепочечные фрагменты ДНК, которые располагались между chi (GCTGGTGG

мотив) сайтами на хромосоме или плазмиде [40]. Механизм вставки нового протоспэйсера варьируется между различными CRISPR-Cas типами [5].

Для I-E типа, Cas1-Cas2 комплекс в одиночку способен к встраиванию новых спэйсеров [202]. Данный комплекс работает как интеграза, которая делает одноцепочечные разрывы на концах первого повтора (повтор, который ближе всего к лидер последовательности – консервативная, АТ богатая область перед кассетой) в CRISPR кассете [143]. Для систем II-A типа, Cas9 также участвует в адаптации. Наличие эффекторного комплекса необходимо и возможно он отвечает за специфичность к PAM последовательности в рамках адаптационного процесса [66, 194]. Новые спэйсеры в основном встраиваются рядом концом лидер последовательности, было показано, что лидер узнается адаптационным комплексом [69][11, 51, 82].

Для типа I-E CRISPR-Cas систем было обнаружено два режима работы адаптационного модуля: наивный (naïve или не праймированный) и праймированный. Наивная адаптация, для встраивания новых спэйсеров в CRISPR кассету требует только адаптационного модуля. В случаях когда CRISPR кассета уже содержит спэйсер схожий к чужеродной последовательности, адаптация происходит намного более эффективно [35, 168, 174, 186]. Данный режим называется “праймированной адаптацией”. Cas3 (часть эффекторного комплекса) - нуклеазно-хеликазный белок, который разрезает чужеродную ДНК, и адаптационный модуль необходимы для данного режима адаптации [35, 168, 186]. Праймированная адаптация была обнаружена в CRISPR-Cas тип I-B [102], тип I-E [35, 168, 174, 186] и тип I-F [161] systems.

Биогенезис crРНК – это процесс экспрессии и созревания crРНК. Данный процесс отвечает за транскрипцию CRISPR кассеты в pre-crРНК (длинный транскрипт CRISPR кассеты), который далее режется на маленькие последовательности размером от 30 до 65 нуклеотидов, каждая из которых содержит спэйсер и фрагмент повтора с одной или с обоих сторон [22, 25]. Механизм генерации crРНК варьируется между различными CRISPR-Cas системами. Системы с *cas6* геном, кодирующим метал независимую эндорибонуклеазу, которая разрезает последовательности повторов в РНК транскрипте CRISPR кассеты [24, 127], имеют crРНК которая содержит 8 нуклеотидов повтора с 5’

конца спэйсера, сам спэйсер и часть повтора с 3’ конца, который может формировать шпильку в некоторых системах [84]. Также в некоторых системах Cas6 может продолжать быть связанным с транскриптом после разрезания или может быть ассоциированным с эффекторным комплексом [65, 162, 198]. Cas6 может работать интранс с CRISPR кассетами различных CRISPR-Cas систем расположенных в том же геноме [106]. В некоторых системах Cas5d подменяет Cas6 для выполнения тех же функций [135].

Системы, которые имеют одиночный блок эффекторного комплекса (Cas9 or Cpf1), процессируют транскрипт CRISPR кассеты используя свой эффекторный комплекс. Системы с Cas9 для биогенезиса crРНК используют хозяйский RNase III и tracrРНК (trans-acting CRISPR РНК, небольшая кодирующая РНК расположенная в отдалении от CRISPR) закодированный недалеко от CRISPR локуса. tracrРНК – небольшие РНК молекулы, которые имеют часть комплементарую повтору процессируемой CRISPR кассеты, которая формирует дуплекс с последовательностью повтора расположенного в pre-crРНК [25, 36, 82]. После разрезания pre-crРНК, комплекс из tracrРНК и crРНК связывается с Cas9, что вызывает конформационное изменение белка и позволяет комплексу искать и разрезать ДНК [36, 52, 82]. CRISPR-Cas системы с Cpf1 не имеют tracrРНК, вместо этого их эффекторные комплексы обладают РНКазной активностю, которая позволяет процессировать pre-crРНК создавая crРНК состоящую из спэйсера и части повтора, который формирует шпильку на 5’ конце crРНК [50].

Интерференция – это процесс внесения разрыва в чужеродную ДНК или РНК с помощью эффекторного комплекса связанного с crРНК с последующей деградацией цели [22, 51]. Для защиты от самоуничтожения (разрезание CRISPR кассеты содержащей спэйсер) CRISPR-Cas системы, которые таргетируют ДНК, в дополнении к комплементарности спэйсер-протоспэйсер последовательностей, требуют правильного PAM мотива (Protospacer Adjacent Motif) рядом с целью. Также было показано, что не все комплементарные позиции в паре спэйсер-протоспэйсер одинаково важны: в спэйсере существует Seed регион (8 нуклеотидов ближе к PAM), и было показано, что комплементарность в районе Seed’а наиболее важная в распознавании цели [170, 185,

197]. Детали распознавания цели, разрезания и деградации таргетируемой ДНК или РНК различаются в разных эффекторных комплексах.

Эффекторные комплексы I типа CRISPR-Cas систем, которые состоят из различных Cas белков в различных подтипах [114], ищут PAM последовательность, далее расплавляют ДНК в районе Seed'а crРНК, далее формируют начальную короткую R петлю с crРНК [84, 160]. В случае не полной комплементарности в районе Seed'а, формирование R петли останавливается, что предотвращает интерференцию путём запрещения докинга Cas3 [18]. В замен интерференции это может инициировать праймированную адаптацию за счёт формирования Cas1-Cas2, Cas3 комплекса [160]. В случае комплементарности, R петля распространяется на всю длинну спэйсера, разрешая докинг Cas3 к комплексу тем самым позволяя деградацию цели [22, 111, 164].

CRISPR-Cas тип III эффекторные комплексы могут таргетировать ДНК с помощью Csm белков (для III-A типа CRISPR-Cas систем) [125] и РНК, с помощью Cmr белков (для III-B типов систем) [64, 182, 204]. Csm и Cmr являются транскрипционно зависимыми нуклеазами [183, 188]; они распознают мРНК по комплементарности с crРНК, далее разрезают РНК и/или транскрибируемую ДНК [37, 48, 49, 55, 148, 165]. Csm3 и Cmr4 (аналоги Cas7 в I типе CRISPR-Cas систем) выстраивают скелет на протяжении последовательности спэйсера в crРНК [189] и разрезают таргетируемую РНК на 6 нуклеотидные фрагменты. Связывание Cmr комплекса с комплементарной crРНК ДНК активирует нуклеазную активность в Cas10 [48, 49, 165, 181, 189]. Основной отличительной особенностью III типа CRISPR-Cas систем является отсутствие требования комплементарности к PAM для интерференции. Было предложено, что для избежания самоуничтожения, Cmr эффекторные комплексы полагаются на распознавание CRISPR повторов, что блокирует самоинтерференцию [126, 145]. Однако, другое исследование показывает, что некоторое системы III типа нуждаются в РНК PAM'е для интерференции [48].

CRISPR-Cas тип II (*cas9* gene) и тип V (*cpf1* gene) – одноблоковые crРНК эффекторные комплексы. Cas9 белок связываясь с crРНК образует crRNP (CRISPR ribonucleoprotein complex – рибонуклеопротеиновый комплекс), который отвечает за распознавание и деградацию таргетируемой ДНК [52]. Различные части комплекса

отвечают за свои активности [139]. PAM на 3’ конце протоспэйсера необходим для расплавления вышестоящей ДНК, это позволяет формирование R петли и разрезание цели в случае спэйсер-протоспэйсер комплементарности [185] (недавно было показана возможность таргетирование одноцепочечной цели, которая не требует PAM последовательности [208]). Совпадение в районе Seed последовательности (12 нуклеотидов рядом с PAM), что позволяет дальнейшее расплавление и формирование crРНК-ДНК гетеродуплекса, активирует Cas9 нуклеазные сайты (HNH и RuvC нуклеазы) путём конформационного изменения белка. Это позволяет Cas9 разрезать две цепи ДНК в районе 3’ конца протоспэйсера, оставляя тупые концы в месте разреза [139, 184]. Cpf1 – другой одноблоковый эффекторный комплекс, который делает двуцепочечные разрывы в таргетируемой ДНК. Биоинформатическая характеризация белка показывает, что Cpf1 не имеет HNH нуклеазного домена (только RuvC), второй нуклеазный домен и механизм действия пока не охарактеризованы. Структура Cpf1 была недавно разрешена [199]. Уникальной отличительной чертой этого белка является отсутствие tracrРНК и то, что он делает двуцепочечные разрывы в ДНК вне протоспэйсера, оставляя липкие концы [199, 205].

Обнаружение и характеризация CRISPR-Cas систем

CRISPR-Cas системы были открыты достаточно недавно, но быстро набрали популярность, что позволило быстро и тщательно их функционально охарактеризовать. Первые наблюдения за необычной повторяющейся структурой в ДНК появились тридцать лет назад, но расцвет изучения CRISPR-Cas систем пришёлся на 2012-2013 года, кода было показано применение CRISPR-Cas эффекторных комплексов для задач редактирования геномов (см. Рисунок 2).

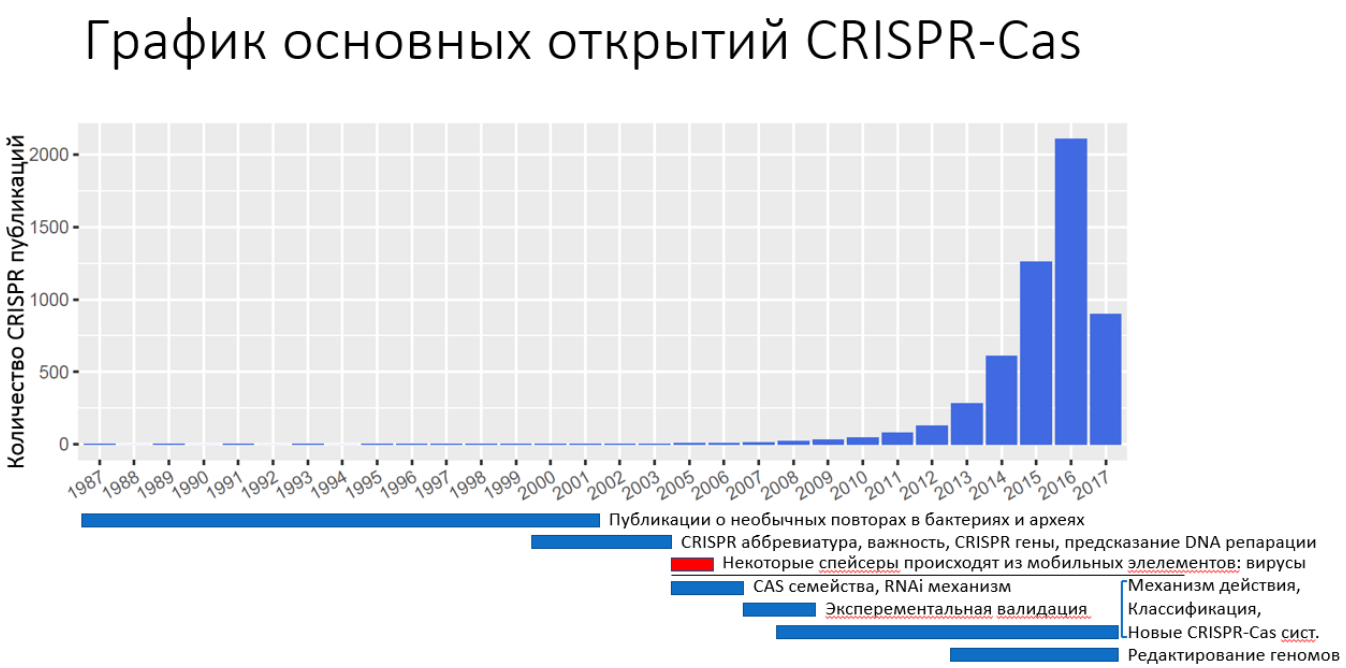


Рисунок 2. Количество публикаций в год упоминающих CRISPR. Эта гистограмма показывает количество научных статей в год содержащих CRISPR как ключевое слово (согласно данным NCBI Pubmed), а также ручного анализа статей, опубликованных до 2002 года. Основные события за основные периоды показаны ниже гистограммы. Период помеченный красным цветом отмечает ключевое исследование, объединившее предложенные ранее предсказания, показывающее, что CRISPR-Cas это антивирусная система в прокариотах.

Кластеры повторов разделённых спэйсерами были впервые замечены в *Escherichia coli* в 1987 году в статье описывающей *iap* ген [78] и структура CRISPR кассеты (сам термин был представлен позже) была описана в 1989 [134]. Однако, функции этих локусов не были описаны/предложены. Похожие нуклеотидные локусы были обнаружены и позже в различных бактериях и археях [23, 67, 71, 130], но эти наблюдения не вызывали большого интереса. Распространение и значимость CRISPR кассет была показана только в 2000 [129]. Два независимых исследования были проведены в 2002, одно описывало разнообразные распространённые семейства, предположительно связанные с репарацией ДНК [109] и другое показывающее связь этих генов с рядом стоящими CRISPR кассетами [81]. Исследование, произведённое позже ввело термин “CRISPR”. Увеличение количества прокариотических данных привело к ключевому

прорыву в понимании функций CRISPR-Cas систем. В 2005, научные группы обнаружили схожесть спэйсеров, фаговых и плазмидных последовательностей и предложили, что эта схожесть дает защиту прокариотам от инфекций [19, 132, 150]. Около 45 белковых семейств связанных с CRISPR были найдены в то же время [61]. Годом позже, в 2006, был предложен механизм РНК интерференции, как механизм действия CRISPR-Cas систем [111]. Доказательство предсказаний механизма действия CRISPR-Cas систем последовало в 2007 году [11]. Исследования появившиеся позже описывали организацию и детали действия CRISPR-Cas систем: что они действуют через РНК [22], что они таргетируют ДНК [125] и РНК [64], что необходима специальная последовательность рядом с протоспэйсером – PAM в некоторых системах [131]. Детали pre-crРНК транскрипции и процессинга также стали известны [65].

После 2010 настал период объединения данных и дальнейшая характеризация механизма действия основных блоков CRISPR-Cas систем. Первая классификация CRISPR-Cas систем была предложена в 2011 году [112]. Тем же годом крио-электронная микроскопия показала как блоки каскада расположены вдоль crРНК в crRNP [198]. В 2011 II тип CRISPR-Cas систем был далее охарактеризован с помощью открытия роли RNase III в процессинге pre-crРНК транскрипта [36]. Различные режимы адаптации были описаны в 2012 году [35], что позволило улучшить понимание кинетики инфекций в клетках обладающих CRISPR-Cas системами.

Исследования проведённые в 2012 году инициировали революцию в области редактирования геномов путём использования запрограммированного Cas9 для внесения двуцепочечного разрыва в целевом участке ДНК [82, 152]. За этими работами последовали другие, показывающие возможность редактирования генома человека [32, 83, 119], бактерий [52, 82] и дрожжей [39]. Эти результаты позволили инициировать первые клинические тесты с CRISPR-Cas генетически модифицированными клетками, которые были разрешены в 2016 году [33, 159].

Классификация CRISPR-Cas систем

CRISPR-Cas – это один из участников постоянной “гонки вооружений” между бактериальными/архейными хостами и вирусами, результатом этого является огромное разнообразие генов участвующих в этом процессе, в том числе CRISPR-Cas генов, локусов и механизмов действия CRISPR-Cas систем [114]. Но всё же, эти системы имеют общие особенности, используя которые было возможно сделать классификацию CRISPR-Cas систем [112, 114]. Было введено два уровня классификации: разделение по классам, разделение по типам/подтипам [114].

Разделение CRISPR-Cas систем по классам основано на структуре эффекторного комплекса (см. Рисунок 3).

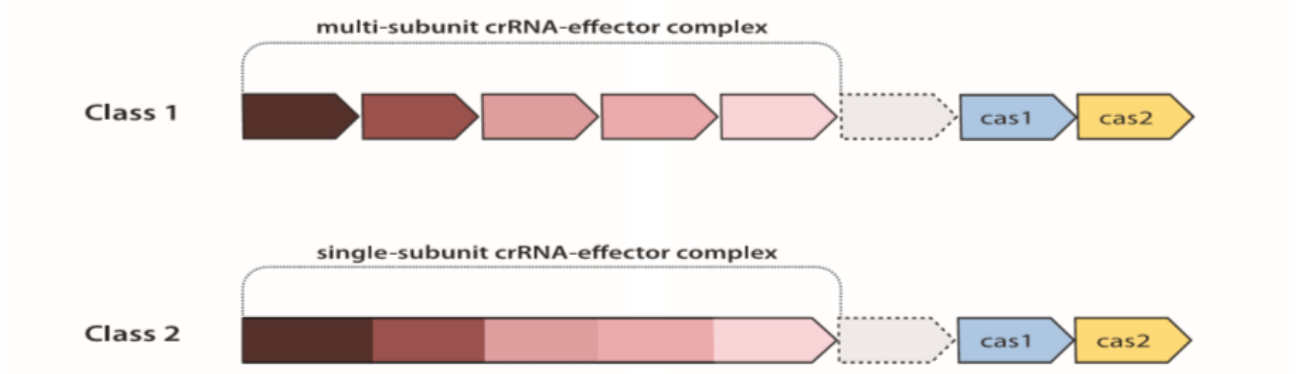


Рисунок 3. Разделение CRISPR-Cas систем на два класса (воспроизведено с разрешения Science [128]). Показано разделение CRISPR-Cas систем на два класса: Класс 1 (Class 1) содержит мульти-протеиновый (многоблоковый) эффекторный комплекс, тогда как в Класс 2 (Class 2) эффекторный комплекс состоит из одного большого белка (одноблоковый комплекс). Белки эффекторного комплекса показаны в красных тонах, градации красного представляют различные биохимические функции. Вспомогательные белки показаны с помощью прерывистой линии. Гены адаптационного модуля (*cas1*, *cas2*) расположены в конце локусов.

CRISPR–Cas системы первого класса, которые имеют многоблоковые crРНК-эффекторные комплексы (Тип I, Тип III, Тип IV), являются самым распространённым классом среди бактерий и архей (включая всех гипертермофилов), и составляют около

~90% от всего разнообразия CRISPR-Cas локусов [114]. Функциональные роли этого класса разделены между белками эффекторного комплекса. Оставшиеся ~10% всех найденных CRISPR-Cas локусов принадлежат второму классу CRISPR-Cas систем (которые используют Cas9 и Cpf1 эффекторные белки). Системы этого класса найдены только в бактериях (за исключением нескольких примеров Cpf1, которые были найдены в археях) и не были найдены в гипертермофилах [30, 114]. Все активности (за исключением адаптации, и транскрипции pre-crРНК во втором типе) выполняются всего одним белком (Cas9, Cpf1).

Второй уровень классификации разделяет CRISPR-Cas системы на типы, эта классификация использует сигнатурные гены и различает 5 типов, и подтипы, эта классификация использует уникальный набор генов, участвующих в эффекторном комплексе и выделяет 16 подтипов (см. Рисунок 4).

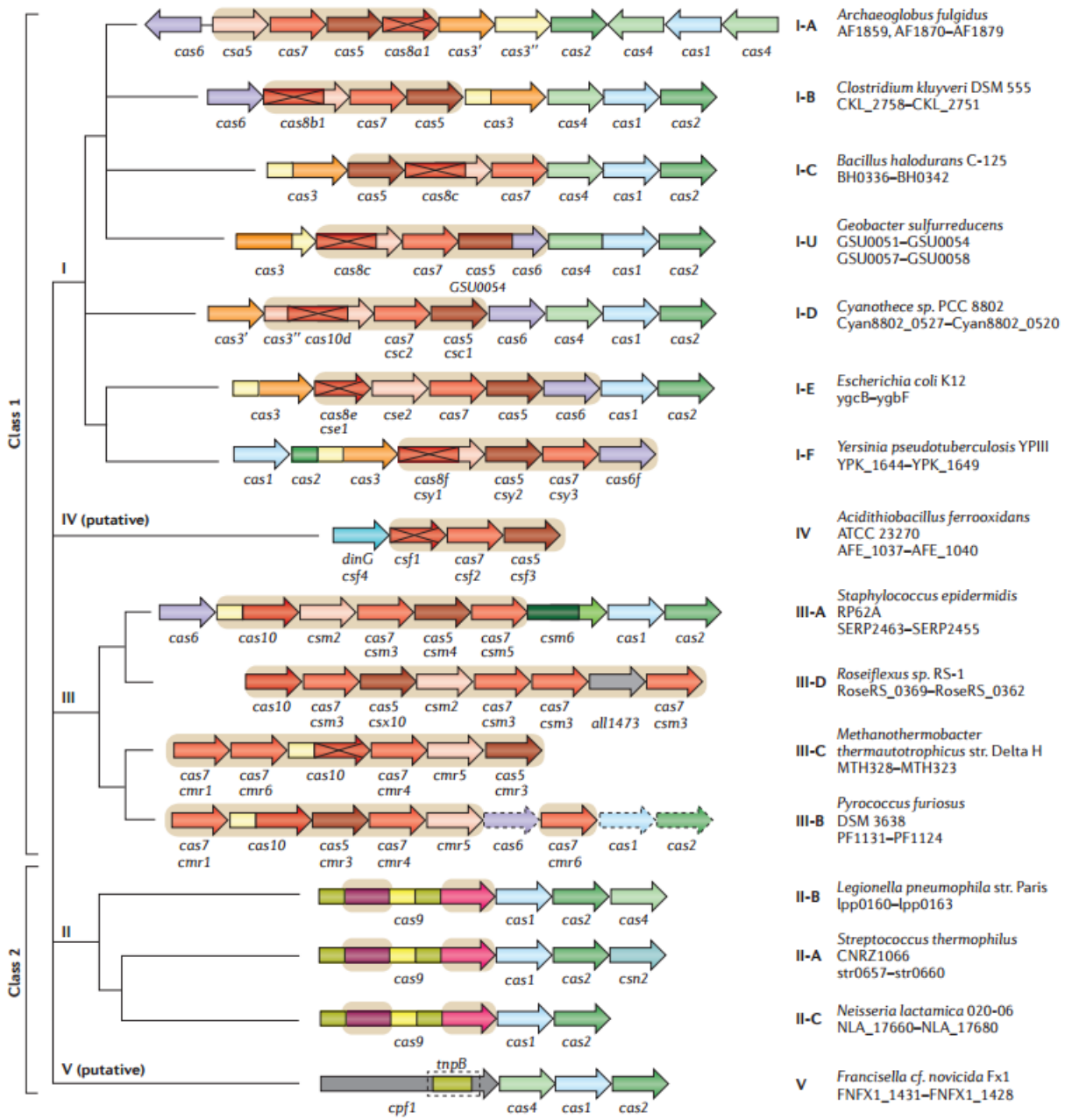


Рисунок 4. Классификация CRISPR-Cas систем на классы, типы и подтипы (воспроизведено с разрешения Nature Reviews Microbiology [114]). Этот рисунок показывает два класса CRISPR-Cas систем, разделённых на 5 типов и 16 подтипов. Горизонтальные линии показывают локусы для каждого подтипа, гены отображены в виде стрелок. Цвета обозначают гомологичные гены. Систематичные названия генов вместе с изначальными расположены под стрелками (например, *cas7* и *cmr1* для Тип III-B). Гены, помеченные перекрестиями, отображают инактивированные большие субъединицы эффекторных комплексов. Гены участвующие в интерференции имеют коричневый цвет на заднем плане. Гены, которые не присутствуют во всех экземплярах подтипов, помечены прерывистой линией.

Первый класс CRISPR-Cas систем разделён на Тип I, Тип III, Тип IV, которые включают в себя 12 подтипов. Сигнатурным геном в первом типе является ген *cas3*, который кодирует хеликазу второго суперсемейства (single-stranded DNA (ssDNA)-stimulated superfamily 2 helicase), которая ответственна за разворачивание двухцепочечной ДНК или РНК-ДНК дуплексов [57, 75, 176]. Cas3 может содержать HD нуклеазный домен (или этот домен может быть расположен в соседствующем белке), который отвечает за разрезание ДНК [133, 176]. В первом типе выделяется семь подтипов: от Тип I-A до Тип I-F и Тип I-U (U обозначает uncharacterized (неохарактеризованный); механизм действия этого типа пока не известен [114]). Все типы, показанные на Рисунке 4 назначены по уникальной комбинации Cas белков. В большинстве случаев, все гены в Тип I расположены в одном опероне генома, за исключением Тип I-A и Тип I-B [191]. Филогенетическое дерево Cas3 совпадает с классификацией первого типа [80], более детальное исследование предлагает гипотезу полифилитичное происхождение этих систем [114].

CRISPR-Cas системы III типа характеризуются присутствием Cas10 – мульти-доменный белок, который содержит домен распознающий РНК (Palm домен) и часто объединённый с HD нуклеазным доменом (отличительным от HD домена присутствующем в I типе [111]). Этот белок является самым большим среди Cas белков III типа и сильно разнообразен [114]. Cas5 и несколько вариантов Cas7 белков также присутствуют в III типе CRISPR-Cas систем. Выделятся 4 подтипа в III типе: от Тип III-A до Тип III-D – с уникальными наборами генов (см. Рисунок 4). Было показано, что все эти системы котранскрипционно таргетируют ДНК [37, 55, 125, 148, 165] и РНК [63, 64, 165, 181, 183, 188]. Филогенетический анализ *cas10* гена совпадает с классификацией сделанной в ([114]). В некоторых случаях Тип III не имеет адаптационного модуля, была предложена гипотеза, что эти системы могут использовать crРНК от других систем в клетке [114].

Тип IV является предсказанным и не был пока функционально охарактеризован. Данный тип не имеет *cas1* and *cas2*. Во многих случаях, не присутствует CRISPR кассета в локусах или даже в целых геномах. Локусы IV типа обычно кодируют Cas5, Cas7 и

сигнатурный белок Csf1 (см. Рисунок 4) [114]. Также было предсказано, что эта система может полагаться на crРНК произведённой из CRISPR кассет других систем.

Тип II второго класса CRISPR-Cas систем, выделяются наличием *cas9* гена, который является одноблоковым эффекторным комплексом. Относительно количества Cas генов в CRISPR локусах II типа систем, они являются наиболее простыми: присутствуют только *cas9* и адаптационный модуль (*cas1*, *cas2* и *cas4*). Cas9 таргетирует ДНК используя два разных нуклеазных домена: HNH и RuvC [82], также этот белок учувствует в адаптации [66, 194]. Уникальной особенностью систем II типа является наличие tracrРНК – короткой РНК, которая закодирована в локусе системы и формирует crРНК-tracrРНК комплекс за счет части комплементарной последовательности повтора CRISPR кассеты. Эта РНК необходима для процессинга pre-crРНК и она является частью эффекторного комплекса вместе с Cas9 [21, 30, 31]. Была предложена гипотеза, основанная на схожести последовательностей, что Cas9 произошёл от IscB транспозона [30]. Это осложняет поиск систем II типа. Принято, что рядом с Cas9 должен располагаться адаптационный модуль, для того чтобы классифицировать Cas9 как CRISPR-Cas систему. Во втором типе систем выделяют три подтипа: Тип II-A, Тип II-B и Тип II-C (наиболее всего распространённая CRISPR-Cas система среди бактерий [114]), эти системы отличаются уникальным составом локусов (см. Рисунок 4). Согласно анализу последовательностей, Тип II-A и Тип II-B являются монофилетичными, тогда как Тип II-C имеет отдельное происхождение [30, 114].

Тип V второго класса является недавно обнаруженной и охарактеризованной CRISPR-Cas системой [50, 169, 205] которая имеет сигнатурный ген *cpf1*. Обычный состав локуса включает *cpf1* и адаптационный модуль (*cas1* and *cas2*) (см. Рисунок 4). Данный тип выделен в отдельный тип (от II типа содержащего Cas9) согласно следующими особенностями: он имеет отдельное происхождение (Cpf1 гомологичен белкам семейства транспозонов IS605 [114]); Cpf1 имеет другую структуру белка (он имеет RuvC нуклеазный домен, но не имеет HNH нуклеазного домена [114, 199]). Также эта система не использует tracrРНК и процессирует pre-crРНК по другому [50]. В отличии от Тип II, эта система была также обнаружена в археях [191].

Происхождение CRISPR-Cas генов

Модульная структура CRISPR-Cas систем и изучение модулей по отдельности помогло выдвинуть гипотезу о происхождении CRISPR-*cas* генов. Адаптационный модуль, который является наиболее консервативным и распространенным и включает в себя *cas1*, *cas2* и (для некоторых систем) *cas4*, путём консервации с *cas10* геном, породило древнюю CRISPR-Cas систему первого класса [128] (см. Рисунок 5).

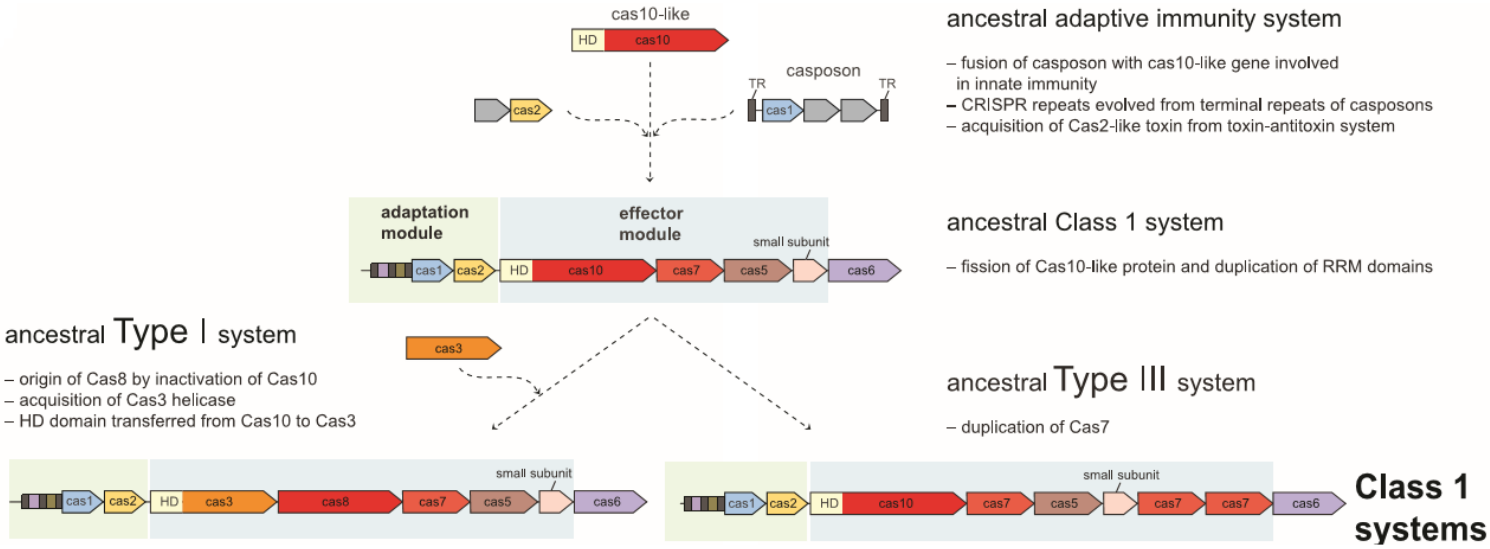


Рисунок 5. Предложенный эволюционный сценарий происхождения CRISPR-Cas систем (воспроизведено с разрешения Science, адаптировано из [128]). Данный рисунок показывает последнюю теорию о возможном происхождении CRISPR-Cas систем первого класса по средствам слияния с различными генетическими компонентами. Гены представлены стрелками; гены закрашенные серым предположительно были в изначальном локусе, но были потеряны в ходе эволюции CRISPR-Cas систем. Зелёный фон показывает CRISPR-Cas адаптационный модуль и синий фон представляет гены эффекторного комплекса. TR обозначает конечные повторы (terminal repeats), HD обозначает семейство HD эндонуклеаз.

Было предложено, что *cas1* ген, который имеет нуклеазную и интегразную функции (происхождение было найдено по гомологии с Cas'позонами (casposons), которые являются самосинтезирующимися транспозонами [68, 99, 100]), возможно вставился рядом с гомологом *cas10*, который содержал РНК связывающий нуклеазный домен а также работал иммунной системой зашиты клетки [128]. Было предсказано, что структуры похожие на CRISPR кассеты появились из инвертированных повторов транспозонов [97]. Ген *cas2*, который появился из токсин-антитоксин (как было предсказано по гомологии) комплекса [110, 116], был либо в локусе cas'позона или в локусе содержащем имунные белки (что являлось местом вставки cas'позона). Также было предсказано, что *cas10* произошёл из слияния Cas10-подобной нуклеазы и одного или более RRM (RNA Recognition Motif) консервативного домена полимеразы или

циклазы. Этот ген вместе с *cas1* и *cas2* дал начало предковой CRISPR-Cas системе [110, 116] (см. Рисунок 5).

Первый класс CRISPR-Cas систем, который является наиболее широко представленным в бактериях и археях [114], указывает на то, что изначальная CRISPR-Cas система имела схожую архитектуру [128]. CRISPR-Cas системы Тип I и Тип III были получены слиянием первоначальной системы с различными генами мобильных элементов. Предполагаемый сценарий о происхождении систем первого типа является: инактивация *cas10* (из которого произошёл *cas8*) и добавление Cas3-подобной хеликазы с приобритением HD нуклеазного домена. Тип III CRISPR-Cas систем, который появился под действием тех же факторов (рекомбинация и приобретение новых элементов из защитных островов [117]), имел дупликацию *cas7* гена в локусе.

Для систем II класса (все подтипы Типа II и Типа V) предсказано, что они появились в результате замены эффекторного модуля системы первого класса на нуклеазный белок, который происходит из различных мобильных элементов [128]. В недавних научных работах было показано, что некоторые части гена *cas9*, согласно гомологии происходят из IscB транспозаз [85], которые имеют RuvC и HNH нуклеазные домены, Тип V имеет только RuvC домен, но не HNH домен, также имеет родство к другому семейству транспозонов [205].

Прикладное применение CRISPR-Cas систем

CRISPR-Cas оказались высоко эффективными [28] и легко программируемыми инструментами для задач редактирования геномов в различных прокариотах и эукариотах включая растения и животных. Технологии редактирования геномов, использующие Cas белки, описаны в различных ревью [10, 73, 128, 158]. Первое применение эффекторных комплексов второго типа CRISPR-Cas систем для редактирования геномов клеток человека было показано в 2013 [29, 32, 83, 119]. В данных научных работах была показана возможность запрограммировать Cas9 с помощью sgРНК (single guide РНК – одно целевая РНК, которая является искусственной

РНК конструкцией: шпилька (для крепления в Cas9), терминальная последовательность и последовательность и спэйсера), чтобы внести двуцепочечный разрыв в ДНК, что может быть использовано для внесения инделей используя механизм репарации ДНК не гомологичного соединения концов или добавления нового материала ДНК используют ДНК шаблоны и механизм репарации по гомологии (см. Рисунок 6). Несмотря на доказанную высокую эффективность, Cas9 всё же имеет недостатки из-за спепецифичности к PAM и к последовательности протоспэйсера. Было изучено и использовано для редактирования геномов несколько вариантов Cas9 и показано, что они распознают различные PAM последовательности [38, 72, 207]. Также были представлены новые искусственно изменённые варианты Cas9, которые имеют улучшенную специфичность за счет уменьшения взаимодействия с нецелевыми участками ДНК [91, 177]. Была опубликована также другая возможность уменьшения таргетирования не целевых участков за счет использования димеров Cas9 [118, 157], в белке были промутированы нуклеазные сайты таким образом, что каждый димер мог раскусывать только одну цепь ДНК. Данный димер нуждается в двух целевых последовательностиях чтобы внести изменение в геном, таким образом увеличивая специфичность комплекса.

Основные операции по редактированию геномов, которые часто выполняются с помощью CRISPR-Cas эффекторных комплексов [20] включают в себя: нокаут генов за счет внесения двуцепочечного разрыва в ДНК белком Cas9, что вызывает сдвиг рамки считывания после ошибок механизма ДНК репарации негомологичного соединения концов; вставку нового ДНК материала за счет внесения двуцепочечного разрыва, но используя механизм починки рекомбинации по гомологии, для которого также дается шаблон, содержащий новую информацию. Эти операции схематично изображены на Рисунке 6.

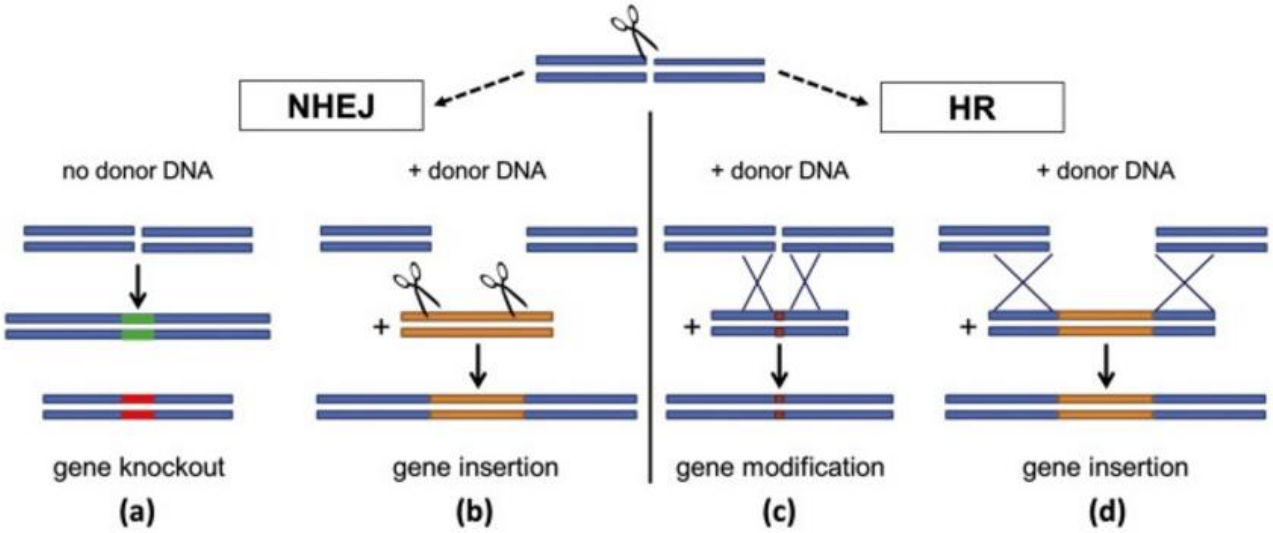


Рисунок 6. Основные методы редактирования геномов с Cas9 (воспроизведено с разрешения Biotechnology Advances [20]). На данном рисунке показано редактирование геномов с Cas9 (обозначенного ножницами). Выделено четыре сценария: a) нокаут генов за счет внесения инделей/сдвига рамки считывания; b) вставка новых генов; c) модификация существующих за счет починки по шаблону; d) вставка новых генов за счет добавления информации с ДНК шаблона. NHEJ означает негомологичное соединение концов (Non-Homologous Ends Joining), HR – починка по гомологии (Homologous directed Repair)

Общий подход к редактированию геномов включает в себя следующие операции:

1. Поиск последовательности в геноме с подходящей фланкирующей областью (PAM), которая может быть таргетирована используя выбранный вариант Cas9.
2. Минимизация не целевых воздействий эффекторного комплекса (за счет выбора уникальной последовательности в геноме и хорошей композиции этой последовательности, что наиболее всего подходит к выбранному варианту Cas9 [41]).
3. Разработка и синтез необходимой для данной цели sgРНК.
4. Производство комплекса Cas9/sgРНК.
5. Доставка комплекса в клетку или группу клеток.
6. Валидация результатов.

Однако, редактирование геномов является не единственным способом применения CRISPR-Cas эффекторных комплексов. Инактивированные варианты Cas9 или dCas9 (dead Cas9 – мёртвые Cas9), в которых разрушены HNN и RuvC нуклеазные домены, за счет чего этот белок не может вносить двуцепочечные разрывы в ДНК, могут быть использованы для специфичного распознавания цели без нуклеазной активности. Например: для активизации транскрипции или репрессии транскрипции [17, 54, 94, 153]; как флуоресцентный маркер [27]; для рекрутирования белков модификации гистонов [70, 87]. Недавние научные работы показывают возможности применения CRISPR-Cas эффекторных комплексов для создания логических схем, которые могут использоваться для создания каскадов активации/репрессии генов [88, 140], также для создания схем имитирующих логическую конъюнкцию, применение которых было показано на примере обнаружения раковых клеток в печени (за счет активизации сигнала люцефераз) [105]. Были показаны антимикробные методики специфичные к последовательности, основанные на доставке Cas9 вирусами, которые уничтожали плазмиды дававшие устойчивость клеткам [16], также были представлены антивирусные системы способные подавлять гепатит B или HIV-1 [44, 74, 155]. Другой областью применения CRISPR-Cas эффекторных комплексов является генетический скрининг на потерю функции для положительной и негативной селекции в клетках млекопитающих: производится большой набор sgРНК, таргетирующих интересующие регионы, который используется для генерации большого набора мутантов произведённых с помощью Cas9 [53, 95, 171, 193].

Недавно обнаруженный белок Cpf1 [205], другой представитель второго класса CRISPR-Cas, показал, что новые эффекторные комплексы могут дать новые возможности для редактирования геномов или других сфер применения нуклеаз специфичных к последовательности. Cpfl имеет такие же как и Cas9 или лучшие показатели по не целевым последовательностям [89, 92] и может быть использован, чтобы ещё больше упростить процесс редактирования геномов. Данный комплекс не требует tracrРНК [205], таким образом уменьшая количество необходимых элементов; он способен процессировать свою CRISPR кассету [205], что делает таргетирование множественных целей более простой задачей (не нужно создавать и доставлять большое количество

sgРНК, нужен лишь одна CRISPR кассета содержащая все цели в качестве спэйсеров); нуклеазная активность Cpf1 создает липкие концы [205], что может быть использовано для более эффективной вставки новой ДНК [128].

Основные проблемы, которые остается решить для полноценного применения CRISPR-Cas эффекторных комплексов,- это эффективная и тканеспецифичная доставка CRISPR-Cas эффекторов и этические вопросы относительно редактирования геномов в человеке [128]. Недавние достижения в области доставки, которые включат в себя: использование маленьких Cas9 белков [156], доставка по средствам наночастиц [149], доставка с использованием электропорации [154] и доставка за счет небольших везикулярных частиц (micropinocytosis) [34], показывают, что остается нужда в новых сайт специфичных нуклеазах. В связи с этим, эффекторные комплексы CRISPR-Cas систем первого класса, которые не используются для редактирования геномов сейчас, могут привлечь большее внимание в будущем.

Материалы и методы

Использовавшиеся базы данных

Поиск новых CRISPR-Cas систем был произведён на различных прокариотических наборах данных. В первой части исследования, поиск белков ассоциированных с *cas1* производился в WGS и NT NCBI базах данных [211]. Во второй части исследования, для поиска белков, ассоциированных с CRISPR кассетами, была собрана отдельная база данных. Архейные и бактериальные геномные последовательности были загружены с NCBI FTP портала (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/) в марте 2016. Для не полностью проаннотированных геномов (плотность кодирования менее чем 0.6 кодирующих участков на килобазу) ранее аннотированные открытые рамки считывания были игнорированы и заменены аннотацией полученной от Meta-GeneMark [14] используя стандартную модель MetaGeneMark_v1.mod (эвристическая модель для генетического кода 11 и GC 30). Полная база данных включает 4,961 полностью отсеквенированных и собранных геномов и 43,599 частично секвенированных геномов (в целом представляя 6,342,452 контигов, состоящих из 33,803 уникальных таксономических групп и 12,528 уникальных видов, кодируя 182,301,555 белков).

Стек программ для аннотации CRISPR-Cas локусов

Данное программное решение принимает на вход список позиций (координаты в соответствующей нуклеотидной последовательности) или затравок характеризуемых элементов (*cas1* или CRISPR). Два типа затравок было использовано: позиции Cas1 белков в базах данных NCBI: NR and WGS database [211] и позиции CRISPR кассет в WGS и прокариотической геномной базе данных описанной ранее. CRISPR и *cas1* наборы затравок не были объединены и анализировались по отдельности. TBLASTN

поиск [4] с E-value 0.01 фильтрацией и выключенным фильтром низкой сложности (low complexity) был выполнен против NR и WGS баз данных, используя профиля Cas1 [113] в качестве входных данных, результатом была идентификация 20,766 локусов. CRISPRfinder [59] и PILER-CR [46] программы были использованы (с параметрами по умолчанию) для поиска CRISPR кассет в WGS базе данных (47,174 локусов найдено) и в прокариотической геномной базе данных (45,373 локусов найдено). Все последовательности, включающие в себя области по 10 килобаз с обеих сторон от найденной затравки, были выделены и проанализированы в дальнейшем.

Аннотация открытых рамок считывания (Open Reading Frame (ORF)) была произведена с помощью Meta-GeneMark [210] используя стандартную модель MetaGeneMark_v1.mod (эвристическая модель для генетического кода 11 и GC 30). Все полученные рамки были далее аннотированы используя RPS-BLAST [122] и 30,953 белковых профилей (COG, pfam, cd) загруженные из NCBI CDD базы данных [123, 211], а также 217 отдельных белковых профилей [114]. Идентификация типов и подтипов CRISPR-Cas систем для всех локусов была выполнена согласно процедурам перечисленным в ранее описанной методике [113, 114].

Для *cas1* затравок из NR и WGS баз данных, все локусы, которые имели частичную или не известную аннотацию типа CRISPR-Cas системы и содержали белки размером более чем 500 аминокислот, были проанализированы по отдельности. В частности, для каждого большого белка из этих локусов был проведён поиск против NCBI NR базы данных используя PSI-BLAST [4], с фильтрацией по e-value 0.01 и выключенной фильтрацией по композиции (composition based-statistics) и низкой сложностью (low complexity). Для каждого белка с малым количеством хитов, обнаруженным в данном поиске, был произведён поиск в WGS базе данных используя TBLASTN (Altschul et al., 1997). Программа поиска доменов HHpred [180], была выполнена с настройками по умолчанию для того, чтобы установить дальнее сходство для всех белков найденных с помощью BLAST поиска. Множественные выравнивания для построения профилей белков были выполнены с помощью MUSCLE [45] и MAFFT [86].

Отдельно был выполнен следующий поиск: все кандидаты, найденные рядом с *cas1* и CRISPR затравками, были кластеризованы (см. Кластеризация). Потенциальные

кандидаты были выбраны среди всех не строгих кластеров, составленных из белков найденных в локусах с затравками, белки были отфильтрованы по размеру (> 500 аминокислот) и удалению от затравки (не далее, чем 4 рамки считывания от затравки, более близким белкам отдавался приоритет); Кластеры, которые содержали большее количество хитов вне локусов с затравками чем в сами локусы, были игнорированы. Дополнительное предсказание белковых доменов было выполнено с помощью CD-search [47] и HHpred [180].

Найденные кандидаты, полученные с помощью Cas1 затравок, были использованы для поиска с помощью PSI-BLAST в NCBI NR и NCBI WGS базах данных новых вариантов, NCBI WGS и прокариотическая база данных были использованы для кандидатов, полученных из локусов, содержащих CRISPR кассеты. Координаты новых белков были добавлены к списку затравок. Описанный поиск был повторён с новым списком затравок для расширения списка кандидатов.

Кластеризация и филогенетический анализ

Для получения репрезентативного не повторяющегося набора белковых последовательностей, была произведена строгая кластеризация с помощью NCBI BLASTCLUST program [196] (ftp://ftp.ncbi.nih.gov/blast/documents/blastclust.html) с параметрами отсечения по идентичности последовательностей 90% и длинны покрытия последовательностей 0.9. Самая длинная последовательность была выбрана как представитель кластера. Не строгая кластеризация была выполнена для получения белковых семейств, используя UCLUST [47], с порогом по схожести последовательностей 0.3 (sequence similarity threshold of 0.3).

Множественные выравнивания белковых последовательностей были созданы с помощью MUSCLE [45] и MAFFT [86] программ. Сайты с частотой пропуска > 0.5 и гомогенностью < 0.1 [203] были удалены из выравнивания. Филогенетический анализ был выполнен используя FastTree программу [151], с эволюционной моделью WAG и дискретной гамма моделью (discrete gamma model) с 20 категориями.

Взаимосвязи между различными семействами последовательностей были получены используя следующую процедуру: начальные кластеры были установлены используя UCLUST [47] с порогом схожести последовательностей 0.5; последовательности внутри кластера были выравнены с помощью MUSCLE [45]. Далее ранг кластер-кластер схожести был получен с помощью HHSEARCH [179] (включая тривиальные кластеры, содержащие по одной последовательности каждый) и UPGMA дендрограмма была построена из ранга попарной схожести. Кластеры с высокой схожестью (ранг попарной схожести / схожесть внутри кластера > 0.1) были выравнены между собой используя HHALIGN [179], эта процедура была повторена итеративно. На последнем шаге деревья, основанные на последовательностях, были построены из выравниваний кластеров используя FastTree программу [151], как объяснено выше, и укоренены по средней точке на длиннейшем пути дерева; далее эти деревья были соединены листьями UPGMA основанной на схожести профилей дендрограммы.

Анализ протоспэйсеров

Изначальный набор из 488,437 спэйсеров из найденных CRISPR кассет был уменьшен до набора 268,409 уникальных спэйсеров. MEGABLAST программа [209] с параметром word size = 18 была использована для поиска протоспэйсеров в вирусной части базы данных NR (TaxID:10239) и в прокариотической геномной базе данных. Максимальное количество несовпадений для спэйсера длинной l было ограничено функцией $x = \sqrt{\max(0,\ l - 22)}$. Все хиты MEGABLAST, которые попадали в CRISPR кассеты или эукариотические вирусы, были отфильтрованы. На выходе данная процедура дала 63,939 хита в прокариотическую базу данных и 5,095 в прокариотические вирусы. 33,480 открытые рамки считывания, которые содержали найденные протоспэйсеры, были использованы для поиска с помощью BLASTP в вирусной базе данных. Все белки с e-value < 10e-6 были классифицированы как белки из вирусов или провирусов.

Анализ синтении локусов в типе V-U

Белковые последовательности, закодированные в близости (±3 гена) от Тип V-U эффекторных комплексов, были собраны и прокластеризованы используя UCLUST [47] с порогом по схожести последовательностей равным 0.3. Гены были проаннотированы номерами кластеров; каждый локус был представлен как набором генов и неупорядоченными парами генов. Взвешенный коэффициент Жаккара был посчитан для всех пар в локусах по описанной ранее процедуре [114], граф схожести локусов был построен используя порог схожести 0.61 (e-0.5), связные компоненты (поднаборы сильно схожих локусов) были выделены для анализа.

Анализ роли селекции в эволюции эффекторных генов второго класса

Нуклеотидные и белковые последовательности эффекторных генов были собраны; Кластера белков с идентичными последовательностями были уменьшены до одного представителя; оставшиеся последовательности были прокластеризованы используя UCLUST [47] с порогом схожести последовательностей равным 0.67. Последовательности кластеров были выравнены и филогенетическое дерево было построено как описано выше и укоренено используя модифицированную процедуру средней точки. Выравнивания для последовательностей белков, относящаяся к поддеревьям с глубиной < 0.1, были собраны и переведены в нуклеотидные выравнивания последовательностей. Попарные dN, dS и dN/dS значения были получены с помощью codeml программы из пакета PAML [201]. Пары последовательностей с $0.0002 \leq dN \leq 1.0$ и $0.0002 \leq dS \leq 1.0$ были выбраны и значения dN/dS были посчитаны.

Результаты и обсуждение

Часть 1. Новые CRISPR-Cas системы 2 класса

Биоинформатический подход для поиска новых локусов CRISPR–Cas 2 класса

Был разработан стек программ для систематического обнаружения и характеризации CRISPR–Cas систем 2 класса (см. Рисунок 7).

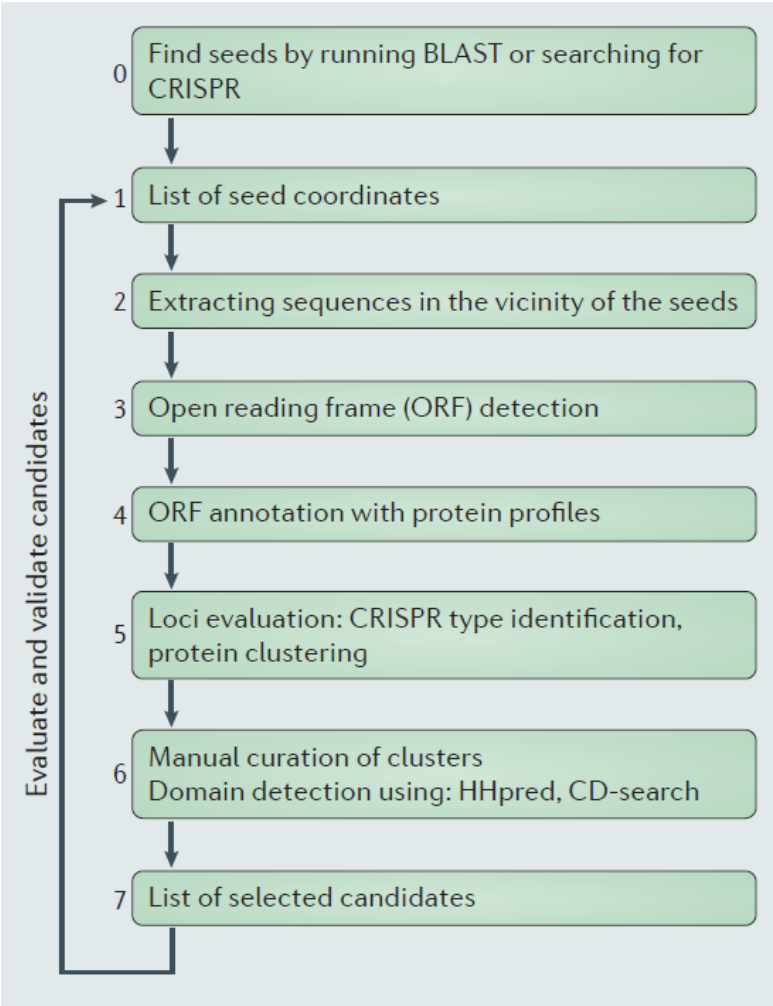


Рисунок 7. Схематическое изображение шагов для поиска белков, ассоциированных с CRISPR или Cas1 (воспроизведено с разрешения Nature reviews. Microbiology [175]). Изображен программный комплекс для обнаружения CRISPR-Cas локусов 2 класса. Шаги, выполненные в данном исследовании раскрыты в тексте ниже.

Процедура поиска начинается с идентификации затравок (Seed), которые увеличивают шанс нахождения CRISPR–Cas локуса в данной нуклеотидной последовательности (см. Рисунок 7; шаги в процедуре поиска пронумерованы в порядке, в котором они выполнялись). В данном исследовании новые CRISPR-Cas системы были найдены путём анализа доступных геномных баз данных (см. “Использовавшиеся базы данных” в материалах и методах). *cas1* был взят как затравка, так как он является наиболее распространённым белков среди всех CRISPR–Cas систем и имеющим наибольшую консервацию на уровне последовательности [187]. Для обеспечения максимальной эффективности обнаружения, данный поиск был выполнен путём сравнения профиля последовательностей Cas1 с протранслированными геномными и метагеномными последовательностями. После того, как все возможные последовательности *cas1* генов были обнаружены, их непосредственное окружение было проанализировано на наличие других *cas* генов путём поиска других Cas белков используя ~400 ранее созданных профилей и применяя ранее описанные критерии классификации CRISPR–Cas локусов [114]. В дополнение к этому, для увеличения пространства поиска, был применён комплементарный подход по идентификации CRISPR–Cas систем – та же процедура была повторена используя CRISPR кассеты в качестве затравок. Для обеспечения высокой чувствительности поиска CRISPR кассет, предсказания были сделаны двумя программами: Piler-CR [46] и CRISPRFinder [59], результаты были объединены и приняты за финальный набор CRISPR кассет (см. Рисунок 7). Данная процедура нашла 47,174 CRISPR кассет, что почти вдвое больше обнаруженных Cas1 белков, отображая факт того, что большое количество CRISPR–Cas локусов не имеют адаптационного модуля и того, что существует большое количество “одиноких” кассет, некоторые из которых возможно функциональны [2].

Все локусы, к которым были приписаны известные CRISPR–Cas подтипы, путём поиска белковых Cas профилей, были отфильтрованы из дальнейшего анализа, так как целью данного проекта было обнаружение новых подтипов. Среди оставшихся окрестностей Cas1 и CRISPR кассет, были тщательно проанализированы те, которые содержат большие белки (>500 аминокислот), это было сделано исходя из того, что Cas9

и Cpf1 являются большими белками (обычно >1000 аминокислот) и из-за того, что их белковые структуры указывают на то, что подобный большой размер необходим чтобы вместить в себя CRISPR РНК (crРНК) и таргетируемую ДНК [50, 138, 139]. Последовательности этих больших белков далее были проанализированы на наличие в них известных доменов, используя чувствительные методы поиска профилей, такие как HHpred [180], предсказание вторичной структуры и ручной анализ множественного выравнивания (см. Материалы и методы). Основываясь на предположении, что эффекторные белки 2 класса содержат нуклеазные домены, даже если они отдалённо схожи или не похожи на известные семейства нуклеаз, все белки, которые содержат домены не свойственные в контексте функций CRISPR-Cas систем (например, мембранные транспортеры или метаболические энзимы) были отброшены. Оставшиеся белки или содержали быстро определяемые, или полностью неизвестные нуклеазные домены. Последовательности этих белков были проанализированы с помощью наиболее чувствительных методов поиска доменов, таких как HHpred [180], используя курируемые множественные выравнивания соответствующих белков, используемых в качестве входных данных. Использования подобных чувствительных методов является обязательным из-за того, что белки вовлеченные в противовирусную защиту, Cas белки в частности, обычно эволюционируют очень быстро [115, 187].

Следует отметить, что данная процедура по поиску CRISPR–Cas систем 2 класса должна быть исчерпывающей, исходя из того, что все локусы, которые кодируют большие белки (возможные эффекторы 2 класса) в окрестностях *cas1* и/или CRISPR, были детально проанализированы. Предположение о структурном требовании к эффекторам 2 класса, которое определило размер белка, и точность детектирования *cas1* и CRISPR кассет - единственные ограничения данного подхода.

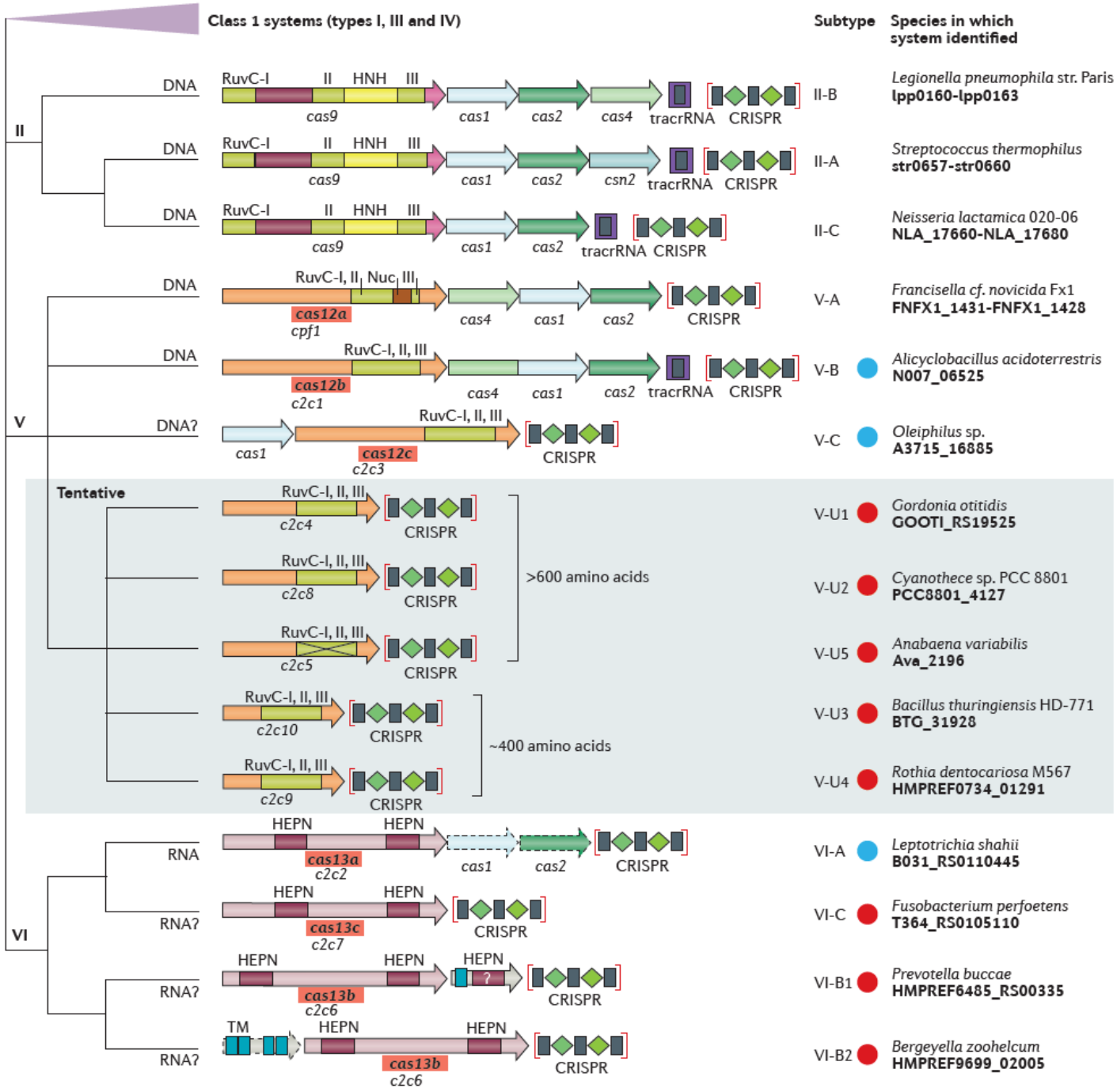

Class 1 systems (types I, III and IV)
Subtype
Species in which system identified
II
V
VI
Tentative
DNA
DNA
DNA
DNA
DNA?
RNA
RNA?
RNA?
RNA?
RuvC-I
II
HNH
III
cas9
cas1
cas2
cas4
csn2
tracrRNA
CRISPR
II-B
Legionella pneumophila str. Paris
lpp0160-lpp0163
II-A
Streptococcus thermophilus
str0657-str0660
II-C
Neisseria lactamica 020-06
NLA_17660-NLA_17680
RuvC-I, II, Nuc, III
cas12a
cpf1
V-A
Francisella cf. novicida Fx1
FNFX1_1431-FNFX1_1428
RuvC-I, II, III
cas12b
c2c1
V-B
Alicyclobacillus acidoterrestris
N007_06525
cas12c
c2c3
V-C
Oleiphilus sp.
A3715_16885
c2c4
V-U1
Gordonia otitidis
GOOTI_RS19525
c2c8
V-U2
Cyanothece sp. PCC 8801
PCC8801_4127
c2c5
V-U5
Anabaena variabilis
Ava_2196
>600 amino acids
c2c10
V-U3
Bacillus thuringiensis HD-771
BTG_31928
c2c9
V-U4
Rothia dentocariosa M567
HMPREF0734_01291
~400 amino acids
HEPN
cas13a
c2c2
VI-A
Leptotrichia shahii
B031_RS0110445
cas13c
c2c7
VI-C
Fusobacterium perfoetens
T364_RS0105110
cas13b
c2c6
VI-B1
Prevotella buccae
HMPREF6485_RS00335
TM
VI-B2
Bergeyella zoohelcum
HMPREF9699_02005

Рисунок 8. Схема обновлённой классификации для CRISPR–Cas систем 2 класса (воспроизведено с разрешения Nature reviews. Microbiology [175]). Все системы первого класса схлопнуты в верху данной схемы; все остальные показанные системы принадлежат ко 2 классу. Новые системы 2 класса, которые были обнаружены с помощью описываемого подхода (см. Рисунок 7), помечены голубыми кругами (для тех систем, которые были обнаружены по ассоциации с *cas1*) и красными кругами (системы, обнаруженные по ассоциации с CRISPR кассетами). Для каждого подтипа систем 2 класса, включая пять различных вариантов предположительного, не охарактеризованного V-U подтипа (V-uncharacterized (V-U)), схематически показана организация локуса и доменная архитектура эффекторов и вспомогательных белков. RuvC-I, RuvC-II и RuvC-III являются тремя различными мотивами, которые участвуют в каталитическом центре нуклеазы; номера на схеме соответствуют RuvC мотиву. Участки Cas9 белков, которые соответствуют распознавательной доле и домену ответственному за взаимодействие с PAM, показаны тёмно-красными и розовыми формами соответственно. Предложенные, новые системные названия генов показаны жирным шрифтом в красном прямоугольнике. Предварительные названия генов для эффекторных белков показаны ниже и расшифровываются как: C2c1–10, Class 2 candidate proteins 1–10 (Класс 2, кандидат 1-10); для подтипа указано V‑A ранее введённое общеупотребительное имя Cpf1. Для подтипа VI‑A, *cas1* и *cas2* показаны прерывистыми линиями, это означает, что только некоторые из них имеют адаптационный модуль. Для V-U5 варианта, инактивация RuvC подобного нуклеазного домена обозначена перекрестием. Названия штаммов бактерий, в которых эти системы встречаются, и названия локусов, где закодированы соответствующие гены, отмечены в правой части схемы. TM аббревиатура обозначает предсказанный transmembrane helix (транс мембранный домен). Предсказанный тип цели, ДНК или РНК, указано для каждого подтипа. Знак вопроса, расположенный за предсказанной целью, означает, что цель была только предсказана, но не продемонстрирована экспериментально. Цель не отмечена для типа V-U, так как их возможности к интерференции сомнительны, что дополнительно показано тёмным фоном.

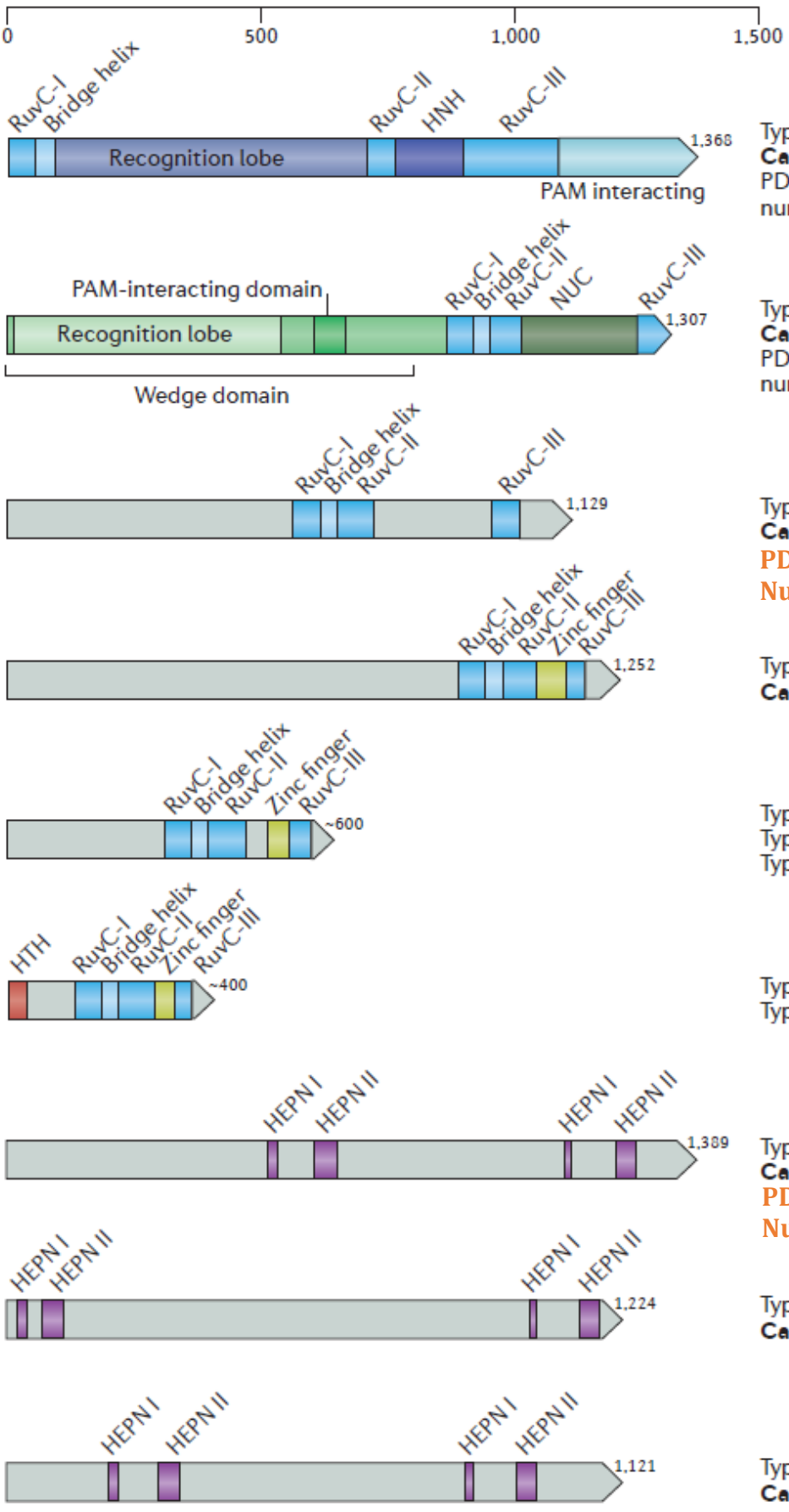

0
500
1,000
1,500
RuvC-I
Bridge helix
RuvC-II
HNH
RuvC-III
Recognition lobe
1,368
PAM interacting
Type II
Cas9
PDB accession
number 5CZZ
PAM-interacting domain
RuvC-I
Bridge helix
RuvC-II
NUC
RuvC-III
Recognition lobe
1,307
Wedge domain
Type V-A
Cas12a (Cpf1)
PDB accession
number 5B43
RuvC-I
Bridge helix
RuvC-II
RuvC-III
1,129
Type V-B
Cas12b (C2c1)
PDB accession
Number 5WQE, 5U34
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
1,252
Type V-C
Cas12c (C2c3)
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
~600
Type V-U1 C2c4
Type V-U2 C2c8
Type V-U5 C2c5*
HTH
RuvC-I
Bridge helix
RuvC-II
Zinc finger
RuvC-III
~400
Type V-U4 C2c9
Type V-U3 C2c10
HEPN I
HEPN II
HEPN I
HEPN II
1,389
Type VI-A
Cas13a (C2c2)
PDB accession
Number 5WTJ, 5WTK
HEPN I
HEPN II
HEPN I
HEPN II
1,224
Type VI-B
Cas13b (C2c4)
HEPN I
HEPN II
HEPN I
HEPN II
1,121
Type VI-C
Cas13c (C2c7)

Рисунок 9. Доменная архитектура CRISPR-Cas эффекторных белков 2 класса систем (воспроизведено с разрешения Nature reviews. Microbiology [175]. Для Тип II и подтипа V‑A эффекторов, кристаллическая структура (обозначенная здесь по их RCSB Protein Data Bank (PDB) номерам доступа (5CZZ и 5B43, соответственно)) доступна. Кристаллическая структура для некоторых новых эффекторов также доступна (PDB номера выделены оранжевым). Для оставшихся белков, серая область указывает на отсутствие структурной или функциональной информации. RuvC-I, RuvC-II и RuvC-III, как и HEPN I и HEPN II (Higher Eukaryotes and Prokaryotes Nucleotide-binding I и II), обозначают каталитические мотивы соответствующих нуклеазных доменов CRISPR эффекторов. “bridge helix” область соответствует аргинин-богатому региону, который следует RuvC-I мотиву. Остальные домены, указанные на схеме, означают следующее: “PAM interacting” – домен взаимодействующий с PAM последовательностью; HNH – HNH эндонуклеаза, zinc finger домен с CXXC..CXXC мотивом (точки означают вариабельную длину между двумя цистеинами); HTH - вероятный ДНК связывающий helix–turn–helix домен; NUC – нуклеазный домен. Белки и домены показаны в приблизительном масштабе. Для каждого белка указано соответствующее количество аминокислот (линейка на верху показывает масштаб). Для функционально охарактеризованных полноразмерных эффекторов, обозначена предложенная новая номенклатура (Cas12 и Cas13), тогда как для не охарактеризованных вероятных эффекторов типа V-U указаны предварительные имена. В том случае, если будет показано функционирование их как bona fide CRISPR эффекторов, они должны относиться к Cas12 белкам с соответствующей литерой. Предсказанные V-U1, V-U2 и V-U5 эффекторы больше чем типичные TnpB белки, тогда как V-U3 и V-U4 эффекторы совпадают по размеру с TnpB. Звёздочка в названии C2c5 указывает на то, что этот предсказанный эффекторный белок содержит замены в каталитическом центре RuvC-подобного нуклеазного домена и не содержит zinc finger.

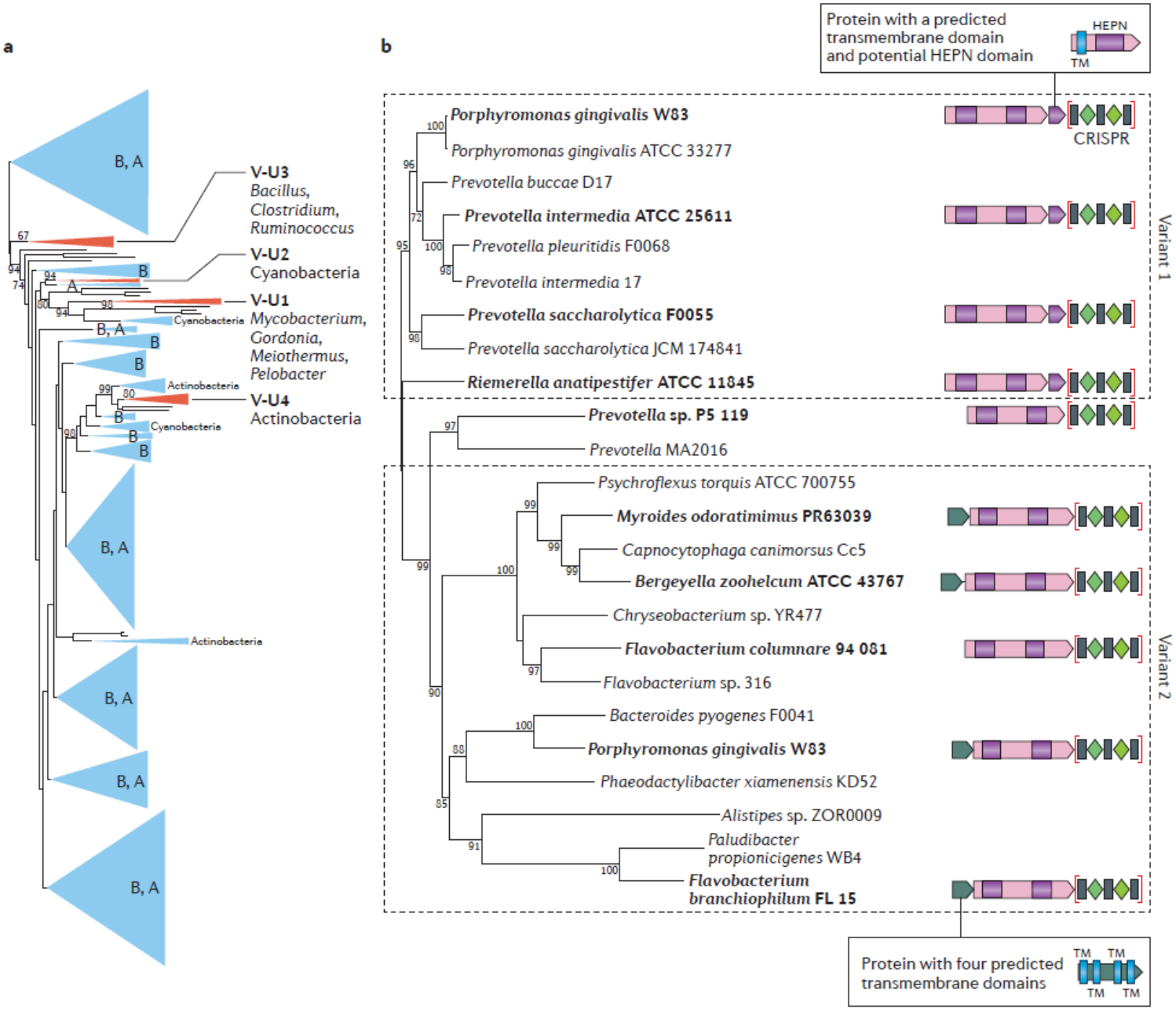

a
b
B, A
V-U3
Bacillus,
Clostridium,
Ruminococcus
V-U2
Cyanobacteria
V-U1
Mycobacterium,
Gordonia,
Meiothermus,
Pelobacter
Actinobacteria
V-U4
Actinobacteria
Cyanobacteria
Actinobacteria
Protein with a predicted transmembrane domain and potential HEPN domain
HEPN
TM
CRISPR
Porphyromonas gingivalis W83
Porphyromonas gingivalis ATCC 33277
Prevotella buccae D17
Prevotella intermedia ATCC 25611
Prevotella pleuritidis F0068
Prevotella intermedia 17
Prevotella saccharolytica F0055
Prevotella saccharolytica JCM 174841
Riemerella anatipestifer ATCC 11845
Prevotella sp. P5 119
Prevotella MA2016
Psychroflexus torquis ATCC 700755
Myroides odoratimimus PR63039
Capnocytophaga canimorsus Cc5
Bergeyella zoohelcum ATCC 43767
Chryseobacterium sp. YR477
Flavobacterium columnare 94 081
Flavobacterium sp. 316
Bacteroides pyogenes F0041
Porphyromonas gingivalis W83
Phaeodactylibacter xiamenensis KD52
Alistipes sp. ZOR0009
Paludibacter propionicigenes WB4
Flavobacterium branchiophilum FL 15
Variant 1
Variant 2
Protein with four predicted transmembrane domains
TM
TM
TM
TM

Рисунок 10. Филогения для эффекторов V и VI‑B типов (воспроизведено с разрешения Nature reviews. Microbiology [175]). a) Филогенетическое дерево TnpB нуклеаз, включающее предсказанные типы V-U эффекторов, с предсказанным RuvC доменом (см. Материалы и методы). Основные поддеревья транспозон-кодирующих TnpB белков схлопнуты и обозначены треугольниками; некоторые из этих больших групп включают в себя *tnpB* гены, которые стоят рядом с CRISPR кассетами, но они не показывают эволюционной стабильности, таким образом не могут быть идентифицированы как эффекторы. Четыре основные эволюционно стабильные группы CRISPR-ассоциированных TnpB приписанных к V-U показаны красными треугольниками. В целом это дерево включает в себя 1,770 уникальных TnpB последовательностей, 403 из которых этоTnpB белки, которые закодированы рядом с TnpA (автономный транспозон); 168 из *tnpB* генов стоят рядом с CRISPR кассетами и 49 из них приписаны к четырём вариантам подтипов V‑U (ни один из них не принадлежит к автономным транспозонам). Для поддеревьев, которые включают в себя варианты типа V-U, показаны bootstrap числа (проценты). Для каждого варианта V‑U указана доминирующая таксономическая группа. Доминирующая бактериальная или архейная группы указаны внутри треугольников (A, различные археи; B, различные бактерии). Для полного дерева и идентификаторов всех последовательностей см. сопроводительную информацию Box 2 (часть c и h).

b) Филогенетическое дерево для подтипа VI‑B Cas13b эффекторных белков. Данное дерево было построено по той же методике, что и дерево в части a и bootstrap значения (для >70%) обозначены на дереве. Типичная организация Cas13b локусов для выбранных представителей (особенно для тех, которые выделены жирным) схематично показана справа. Вариант 1 и вариант 2 соответствуют двум основным веткам дерева и отличаются доменной архитектурой второго не большого белка, закодированного в локусах; доменная архитектура этого предположительно вспомогательного белка показана выше (для варианта 1) и ниже (для варианта 2) для соответствующих локусов (указанных линией). CRISPR кассеты указаны схематично в скобках. TM обозначает transmembrane domain (трансмембранный домен), и показан синими прямоугольниками. Higher eukaryotes and prokaryotes nucleotide-связывающие (HEPN) домены показаны тёмно-красным.

Подтипы V-B и V-C обнаруженные с использованием cas1 затравки: большие мульти-доменные эффекторы

Отличительной особенностью Тип II и Тип V CRISPR–Cas последовательностей является наличие RuvC подобного нуклеазного домена в их мульти-доменных эффекторных белках [114]. Для эффекторов Cas9 во втором типе CRISPR-Cas систем RuvC подобный домен содержит добавленный HNH нуклеазный домен (см. Рисунки 8, 9). Кроме как RuvC подобного домена, эффекторные белки трёх подтипов V типа не содержат схожих последовательностей между собой или Cas9. Однако, только кристаллическая структура для эффекторов 2 класса (которая стала доступна во время данной научной работы), особенно для вариантов Cas9 и Cpf1, выявила схожую структуру (см. Рисунок 9) [42, 199]. Структуры обнаруженных, новых больших эффекторов V типа, обнаруженных с помощью *cas1* затравки, а именно эффекторы подтипов V-B и V-C, были не доступны на момент данной работы, позже для подтипа V-B эффектора C2c1, кристаллическая структура была разрешена [104, 200], также была показана сильная интерференционная активность [173]. Все эффекторы из V типа, которые были обнаружены в данной работе, имеют схожий большой, размер (обычно, 1,000–1,300 аминокислот) и схожий одиночный RuvC подобный эндонуклеазный домен (см. Рисунок 9), но с другой стороны схожесть последовательностей между эффекторными белками разных подтипов очень низкая. Вероятно, что все эффекторы V типа имеют схожие билобные структуры, позволяющие захватывать crРНК и таргетировать ДНК одновременно несмотря на то, что эффекторные белки разных подтипов, по всей видимости, не связаны на прямую.

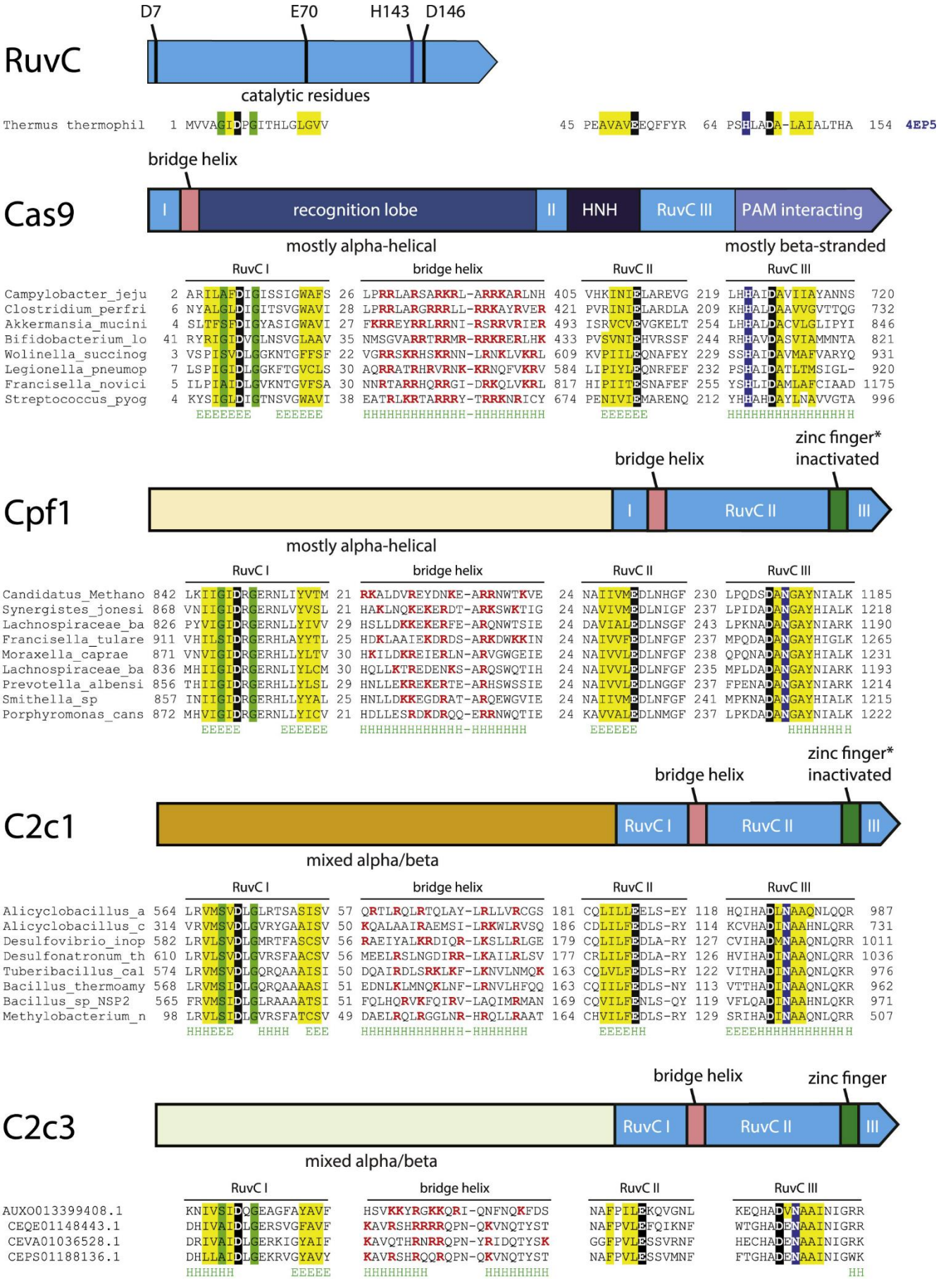

RuvC
D7
E70
H143
D146
catalytic residues
Thermus thermophil 1 MVVAGIDPGITHLGLGVV 45 PEAVAVEEQFFYR 64 PSHLADA-LAIALTHA 154 4EP5
Cas9
bridge helix
I
recognition lobe
II
HNH
RuvC III
PAM interacting
mostly alpha-helical
mostly beta-stranded
RuvC I
bridge helix
RuvC II
RuvC III
Campylobacter_jeju 2 ARILAFDIGISSIGWAFS 26 LPRRLARSARKRL-ARRKARLNH 405 VHKINIELAREVG 219 LHHAIDAVIIAYANNS 720
Clostridium_perfri 6 NYALGLDIGITSVGWAVI 28 LPRRLARGRRRLL-RRKAYRVER 421 PVRINIELARDLA 209 KHHAIDAAVVGVTTQG 732
Akkermansia_mucini 4 SLTFSFDIGYASIGWAVI 27 FKRREYRRLRRNI-RSRRVRIER 493 ISRVCVEVGKELT 254 LHHAIDACVLGLIPYI 846
Bifidobacterium_lo 41 RYRIGIDVGLNSVGLAAV 35 NMSGVARRTRRMR-RRKRERLHK 433 PVSVNIEHVRSSF 244 RHHAVDASVIAMMNTA 821
Wolinella_succinog 3 VSPISVDLGGKNTGFFSF 22 VGRRSKRHSKRNN-LRNKLVKRL 609 KVPIILEQNAFEY 229 SSHAIDAVMAFVARYQ 931
Legionella_pneumop 7 LSPIGIDLGGKFTGVCLS 30 AQRRATRHRVRNK-KRNQFVKRV 584 LIPIYLEQNRFEF 232 PSHAIDATLTMSIGL- 920
Francisella_novici 5 ILPIAIDLGVKNTGVFSA 30 NNRTARRHQRRGI-DRKQLVKRL 817 HIPIITESNAFEF 255 YSHLIDAMLAFCIAAD 1175
Streptococcus_pyog 4 KYSIGLDIGTNSVGWAVI 38 EATRLKRTARRRY-TRRKNRICY 674 PENIVIEMARENQ 212 YHHAHDAYLNAVVGTA 996
EEEEEEE EEEEEE HHHHHHHHHHHHH-HHHHHHHHH EEEEEE HHHHHHHHHHHHHHHH
Cpf1
zinc finger*
inactivated
bridge helix
I
RuvC II
III
mostly alpha-helical
RuvC I
bridge helix
RuvC II
RuvC III
Candidatus_Methano 842 LKIIGIDRGERNLIYVTM 21 RKALDVREYDNKE-ARRNWTKVE 24 NAIIVMEDLNHGF 230 LPQDSDANGAYNIALK 1185
Synergistes_jonesi 868 VNIIGIDRGERNLVYVSL 21 HAKLNQKEKERDT-ARKSWKTIG 24 NAVIVMEDLNIGF 237 LPIDADANGAYHIALK 1218
Lachnospiraceae_ba 826 PYVIGIDRGERNLLYIVV 29 HSLLDKKEKERFE-ARQNWTSIE 24 DAVIALEDLNSGF 243 LPKNADANGAYNIARK 1190
Francisella_tularе 911 VHILSIDRGERHLAYYTL 25 HDKLAAIEKDRDS-ARKDWKKIN 24 NAIVVFEDLNFGF 237 MPQDADANGAYHIGLK 1265
Moraxella_caprae 871 VNVIGIDRGERHLLYLTV 30 HKILDKREIERLN-ARVGWGEIE 24 NAIVVLEDLNFGF 238 QPQNADANGAYHIALK 1231
Lachnospiraceae_ba 836 MHIIGIDRGERNLIYLCM 30 HQLLKTREDENKS-ARQSWQTIH 24 NAIVVLEDLNFGF 235 MPLDADANGAYNIARK 1193
Prevotella_albensi 856 THIIGIDRGERHLLYLSL 29 HNLLEKREKERTE-ARHSWSSIE 24 NAIVVLEDLNGGF 237 FPENADANGAYNIARK 1214
Smithella_sp 857 INIIGIDRGERHLLYYAL 25 HNLLDKKEGDRAT-ARQEWGVIE 24 NAIIVMEDLNFGF 241 MPKNADANGAYHIALK 1215
Porphyromonas_cans 872 MHVIGIDRGERNLLYICV 21 HDLLESRDKDRQQ-ERRNWQTIE 24 KAVVAIEDLNMGF 237 LPKDADANGAYNIALK 1222
EEEEE EEEEEE HHHHHHHHHHHHH-HHHHHHH EEEEEE HHHHHHHH
C2c1
zinc finger*
inactivated
bridge helix
RuvC I
RuvC II
III
mixed alpha/beta
RuvC I
bridge helix
RuvC II
RuvC III
Alicyclobacillus_a 564 LRVMSVDLGLRTSASISV 57 QRTLRQLRTQLAY-LRLLVRCGS 181 CQLILLEELS-EY 118 HQIHADLNAAQNLQQR 987
Alicyclobacillus_c 314 VRVMSVDLGVRYGAAISV 50 KQALAAIRAEMSI-LRKWLRVSQ 186 CDLILLEDLS-RY 114 KCVHADINAAHNLQRR 731
Desulfovibrio_inop 582 LRVLSVDLGMRTFASCSV 56 RAEIYALKRDIQR-LKSLLRLGE 179 CQLILLEDLA-RY 127 CVIHADMNAAQNLQRR 1011
Desulfonatronum_th 610 LRVLSVDLGVRSFAACSV 56 MEELRSLNGDIRR-LKAILRLSV 177 CRLILLEDLA-RY 126 HVIHADINAAQNLQRR 1036
Tuberibacillus_cal 574 LRVMSVDLGQRQAAAISI 50 DQAIRDLSRKLKF-LKNVLNMQK 163 CQLVLLEDLS-RY 122 VITHADINAAQNLQKR 976
Bacillus_thermoamy 568 LRVMSIDLGQRQAAAASI 51 EDNLKLMNQKLNF-LRNVLHFQQ 163 CQIILLEDLS-NY 113 VTTHADINAAQNLQKR 962
Bacillus_sp_NSP2 565 FRVMSIDLGLRAAAATSI 51 FQLHQRVKFQIRV-LAQIMRMAN 169 CQVILLENLS-QY 119 VFLQADINAAHNLQKR 971
Methylobacterium_n 98 LRVLSIDLGVRSFATCSV 49 DAELRQLRGGLNR-HRQLLRAAT 164 CHVILLEDLS-RY 129 SRIHADINAAQNLQRR 507
HHHEEE HHHH EEE HHHHHHHHHHHHH-HHHHHHH EEEEHH EEEEHHHHHHHHHHHH
C2c3
bridge helix
zinc finger
RuvC I
RuvC II
III
mixed alpha/beta
RuvC I
bridge helix
RuvC II
RuvC III
AUXO013399408.1 KNIVSIDQGEAGFAYAVF HSVKKYRGKKQRI-QNFNQKFDS NAFPILEKQVGNL KEQHADVNAAINIGRR
CEQE01148443.1 DHIVAIDLGERSVGFAVF KAVRSHRRRRQPN-QKVNQTYST NAFPVLEFQIKNF WTGHADENAAINIGRR
CEVA01036528.1 DRIVAIDLGERKIGYAIF KAVQTHRNRRQPN-YRIDQTYSK GGFPVLESSVRNF HECHADENAAINIGRK
CEPS01188136.1 DHLLAIDLGEKRVGYAVY KAVRSHRQQRQPN-QKVNQTYST NAFPVLESSVMNF FTGHADENAAINIGWK
HHHHHH EEEEE HHHHHHHH HHHHHHHH HH

Рисунок 11. Доменная архитектура и консервативные мотивы эффекторов 2 класса V типа. (воспроизведено с разрешения Molecular Cell [173]) Тип II и V: TnpB-производные нуклеазы. Верхняя панель показывает RuvC нуклеазу из *Thermus thermophilus* (PDB: 4EP5) с обозначенными каталитическими аминокислотами. Далее показано выравнивание для консервативных мотивов в выбранных представителях для соответствующего семейства белков (и только одна последовательность для RuvC), которое подчеркивает доменную архитектуру для каждого семейства. Каталитические основания показаны белыми буквами на черном фоне; консервативные гидрофобные основания имеют желтый фон; консервативные не большие основания имеют зелёный фон; положительно заряженные основания в спиральных мостиках помечены красным фоном. Предсказание вторичной структуры показано ниже выравненных последовательностей: H обозначает α спираль, и E обозначает удлинённую коформацию (β тяж). Плохо выравненные последовательности между хорошо выравненными блоками, показаны числами.

Районы гомологичные TnpB в C2c1 и C2c3 содержат три каталитических мотивов RuvC подобной нуклеазы [7], этот регион соединён с аргинин богатым спиральным мостиком (bridge helix), который отвечает за связывание crРНК с Cas9 (в белке Cas9), и двойник Zn finger из TnpB (Zn связывающие цистеин основания консервативны в C2c3 но отсутствуют в большинстве Cpf1 и C2c1 белков; Cpf1 и C2c1 содержат множественные вставки и делеции в данном районе указывающие на функциональное разнообразие) (см. Рисунок 9, 11; сопроводительные материалы S1 и S4). Консервативность каталитических оснований подразумевает, что RuvC гомологичные домены из этих белков являются активными нуклеазами. N-концевой участок C2c1 и C2c3 не показывают никакого сходства между собой для всех найденных белков. Предсказание вторичной структуры указывает на то, что оба региона принимают смешанную α/β конформацию (см. сопроводительную информацию S1 и S4). Таким образом, общая домена архитектура C2c1 и C2c3, и в частности организация RuvC домена, походит на Cpf1, но отличается от Cas9 (см. Рисунок 11). Соответственно было предложено, что найденные C2c1 и C2c3 семейства образуют подтипы V-B и V-C, соответственно, а Cpf1 кодирующие локусы должны обозначаться как подтип V-A.

Система C2c1 из *Alicyclobacillus acidoterrestris ATCC 49025 (Aac)* была экспериментально охарактеризована в лаборатории Фанг Жанга [173]. Было показано, что CRISPR кассета активно транскрибируется в том же направлении, что и *cas* гены этого кластера, также было надёжно показан процессинг crРНКs, которая состоит из 34 оснований, включая участок повтора размером в 14 оснований на 5’ конце и 20 оснований принадлежащие спэйсеру. Также было показано, что 79-нт короткая РНК закодированная между *cas2* геном и CRISPR кассетой, транскрибируется в том же направлении, что и CRISPR кассета. Внутренняя часть этой РНК содержит последовательность комплементарную повтору процессируемой CRISPR кассеты (anti-repeat), это указывает на то, что данный транскрипт является tracrРНК. In silico сворачивание процессируемого участка повтора в 14 оснований и этой предположительной tracrРНК предсказывает стабильную вторичную структуру.

Поиск гомологов для эффекторов из Тип II и Тип V показывает, что RuvC-подобный нуклеазный домен имеет сходство с TnpB белками – сильно распространённое, но слабо охарактеризованное семейство нуклеаз, которое кодируется во многих автономных (т.е. кодирующих активную транспозазу и организовывающие свои собственные транспозиции) и ещё большем количестве не автономных (т.е. состоящих только из единственного *tnpB* гена и использующих транспозазы других элементов для транспозиции) бактериальных и архейных транспозонов [8, 85, 147] (см. Рисунок 10a). В дополнение к RuvC подобному нуклеазному домену TnpB белки содержат предсказанную, позитивно заряженную, длинную α-спираль являющуюся двойником спирального мостика, который является известным элементов в Cas9 и Cpf1 (см. Рисунок 9, 11). Таким образом предсказывается, что, схожие с 2 классом эффекторов, TnpB белки связываются с РНК. Более того, была опубликована информация о том, что TnpB белки из *haloarchaeon Halobacterium salinarum* связываются с короткими перекрывающими кодирующими транскриптами своего собственного гена [56]. Биохимическая и биологическая характеризация TnpB должна пролить свет на эволюцию функций CRISPR-Cas эффекторов 2 класса.

Ближайшие родственники и возможные предки Cas9 были определены за счет видимой схожести последовательностей и присутствия вставки HNH в RuvC подобный

нуклеазный домен определённого семейства TnpB белков, которые были обозначены как IscB (insertion sequences Cas9‑like protein B) [30, 85]. В текущих данных сложно проследить прямую связь между эффекторами V типа и определённой группой TnpB белков, так как эффекторы V типа показывают меньшее родство к TnpB белкам, нежели Cas9 показывает к IscB белкам. Тем не менее, эффекторы трёх подтипов V типа показывают сходство с различными семействами TnpB, что указывает на независимое происхождение разных подтипов V типа из набора *tnpB* генов.

Подтип V-U определённый с помощью CRISPR затравок: маленький возможный эффектор

Поиск CRISPR–Cas локусов, которые не имеют адаптационного модуля (т.е. те локусы, которые были обнаружены с помощью CRISPR затравки, но не с помощью *cas1* затравки; см. Рисунок 7) выявил несколько дополнительных вариантов возможных эффекторов систем V типа (см. Рисунки 8, 9, 10a), которые могут помочь объяснить, как CRISPR–Cas эффекторы эволюционировали из TnpB. Возможные эффекторные белки из этих локусов, которым был дан предварительный подтип V-U (где ‘U’ означает ‘uncharacterized’ – не охарактеризованный; см. далее), объединяют две ключевые особенности, которые отличают их от других эффекторов II типа и V типа, которые были найдены в локусах, содержащих Cas1 (см. Рисунок 8). Первое, эти белки значительно меньше чем другие эффекторы 2 класса, которые располагаются рядом с Cas1, их размер составляет от ~500 аминокислот (только немногим больше чем стандартный размер TnpB) до ~700 аминокислот (между размером TnpB и типичным размером bona fide эффекторов 2 класса). Второе, эти возможные эффекторы показывают много больший уровень сходства с TnpB белками, больший чем для белков эффекторов II типа и V типа (см. сопроводительные материалы S3). В частности, три группы TnpB гомологов, которые включены в подтип V-U (обозначенные как: V-U1, V-U2 и V-U5), показывают эволюционную стабильность в рамках консервации последовательностей, стойкую

ассоциацию с CRISPR кассетами и присутствуют в различных группах бактерий (см. Рисунок 8, 9; также см. далее). Более детальное исследование показывает, что среди каждой из этих групп соответственные локусы в близкородственных геномах полностью ортологичны согласно консервации синтении генов.

С точки зрения нахождения этих небольших CRISPR ассоциированных TnpB гомологов, был запущен разработанный поиск (см. Рисунок 7), но без фильтрации по минимальной длине белка стоящего рядом с CRISPR кассетой, результаты были изучены на наличие дополнительных TnpB гомологов. Различные CRISPR ассоциированные TnpB гомологи были обнаружены и имели размер в районе типичным для транспозон-кодирующих TnpB, что составляет ~400 аминокислот (см. сопроводительные материалы S2 (box), часть a). Большинство из этих локусов не были эволюционно консервативными, таким образом имея сомнительную функциональную связь с затравкой. Однако, две различные группы небольших CRISPR ассоциированных TnpB (V-U3 и V-U4) были дополнительно обнаружены и имели характеристики схожие к другим трём группам подтипа V-U среднеразмерных CRISPR ассоциированных TnpB (см. Рисунок 8, 9; Таблица 1). Данные гены предположительных эффекторов подтипа V-U показывают признаки стабилизирующего отбора на уровне последовательностей белков (по низким значениям не синонимичных к синонимичным нуклеотидным заменам, dN/dS; см. Таблица 1), который, как было найдено, особенно сильный для группы белков подтипа V-U3 (см. спороводительные материалы S2 (box), часть b, таблица 1). Объединяя, данные наблюдения указывают на то, что соответствующие TnpB гомологи имеют CRISPR-зависимые функции и, согласно данным наблюдениям, обосновывают обозначение соответствующих локусов как подтип V-U.

system	gene	no. of sequence pairs	dN/dS		
			1st quartile	median	3rd quartile
II-A	cas9	2239	0.12	0.19	0.25
II-B	cas9	67	0.21	0.32	0.64
II-C	cas9	2756	0.08	0.12	0.19
V-A	cas12a	48	0.04	0.13	0.21
V-B	cas12b	4	0.11	0.17	0.25
V-U1	c2c4	4	0.14	0.22	0.44
V-U2	c2c8	3	0.08	0.25	0.30
V-U3	c2c10	14	0.03	0.04	0.12
V-U4	c2c9	11	0.07	0.15	0.36
V-U5	c2c5	16	0.15	0.16	0.19
VI-A	cas13a1	8	0.27	0.39	0.41
VI-B	cas13b	515	0.34	0.39	0.46
VI-C	cas13a2	3	0.28	0.28	0.31

Таблица 1. Сила стабилизирующего отбора для эффекторных белковых семейств 2 класса (воспроизведено с разрешения Nature reviews. Microbiology [175]). Показаны три квартили с распределением dN/dS, рассчитанные для пар последовательностей с 0.0002 < dN < 1.0 и 0.0002 < dS < 1.0 (см. Материалы и методы). Цвет фона показывает разброс значений от низких (синий) до высоких (красный).

Для больших bona fide эффекторов V типа низкая консервативность последовательностей мешает провести надёжный филогенетический анализ, но для маленьких предположительных CRISPR ассоциированных эффекторных комплексов надёжное филогенетическое дерево, которое включает в себя транспозон-кодирующие TnpB, может быть построено (см. Материалы и методы, сопроводительная информация S2 (box), часть c). Топология данного дерева указывает на то, что четыре из пяти различных вариантов подтипа V-U (далее обозначенных как подтипы V-U1, V-U2, V-U3, V-U4 и V-U5) произошли из разных TnpB семейств (см. Рисунок 10a), что согласуется с гипотезой о независимой эволюции различных подтипов эффекторов 2 класса из транспозон-кодирующих нуклеаз. Пятый вариант (подтип V-U5), который был найден в различных цианобактериях, состоящий из различных дальних гомологов TnpB и имеющих несколько мутаций в каталитических мотивах их RuvC подобного домена, не

был включён в данное филогенетическое дерево. Из пяти стабильных вариантов только подтип V-U1 найден в различных бактериях, тогда как остальные подтипы в основном ограничены определённым бактериальным таксоном (см. Рисунок 10a; сопроводительная информация S2 (box), часть d). Далее этот эволюционный анализ был расширен, чтобы включать в себя все возможные подтипы эффекторов V типа, путём построения дендрограмы основанной на расстояниях, полученных из профиль-профиль сравнения для соответствующих белковых последовательностей (см. Материалы и методы). Результаты указывают на то, что эффекторы каждого подтипа, также пять различных вариантов в V-U подтипе, произошли независимо от различных TnpB семейств (см. Рисунок 12).

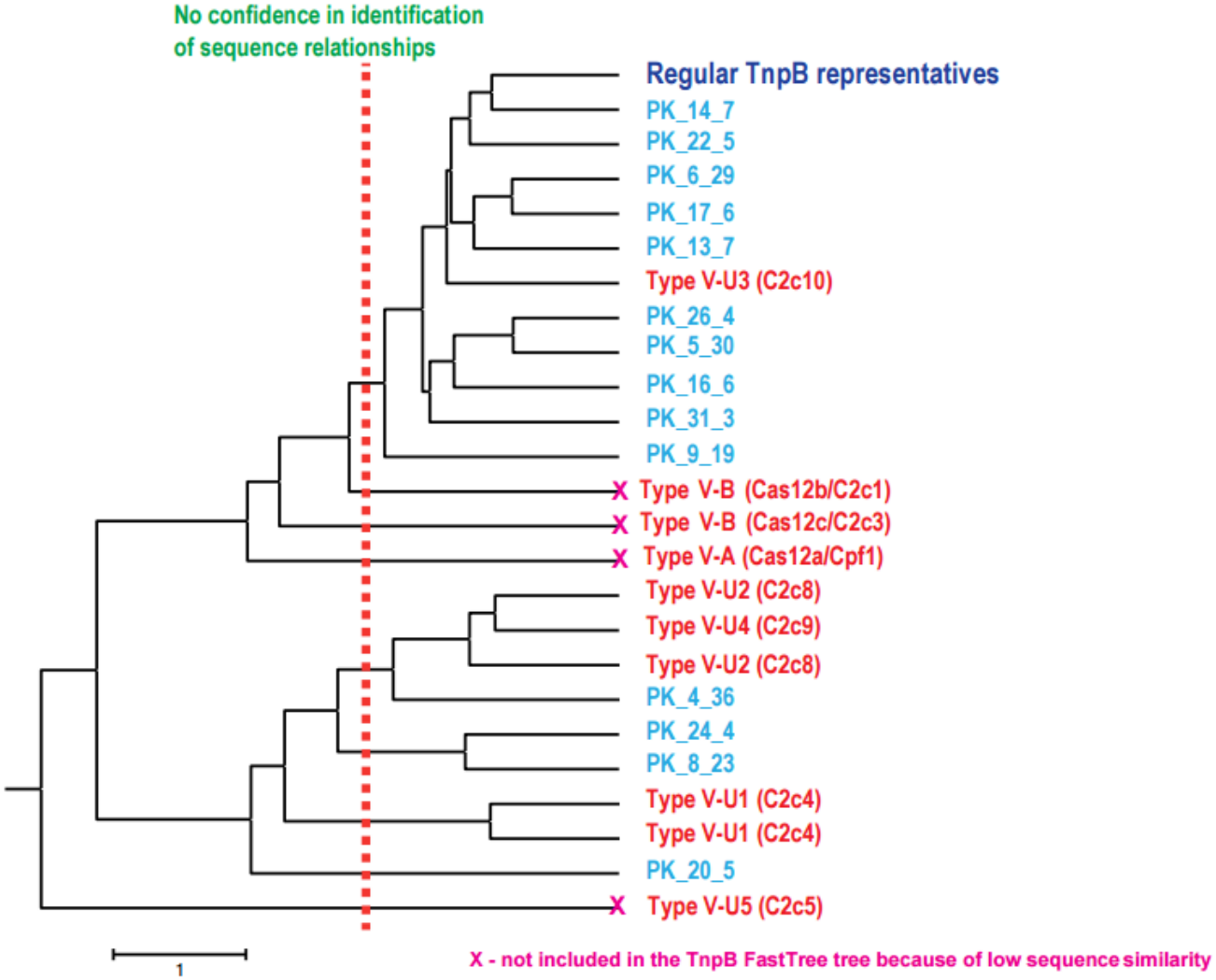


Рисунок 12. UPGMA дендрограмма схожести профилей белковых семейств (воспроизведено с разрешения Nature reviews. Microbiology [175]). Белковые профиля были построены для различных подсемейств систем V типа (красные) и семейства TnpB (синие). Профиля соответствуют кластерам, информация о которых приведена в сопроводительных материалах S2 (box, часть h). Дендрограма из профилей была построена на основе матрицы схожестей полученной из HHalign программы (см. детали в методах в сопроводительных материалах). Прерывистая линия обозначает произвольную отсечку ~2 (в размерности единиц дистанции, показанной линией масштаба ниже дерева), которая, эмпирически, обозначает предел надёжной идентификации между группами последовательностей (т.е. группы с права от прерывистой линии, предположительно, надёжно определены).

Подтип V-U TnpB подобные белки слишком маленькие, чтобы принимать билобную структуру необходимого размера, чтобы вмещать crРНК–целевая ДНК комплекс, которую принимают обычные эффекторные белки 2 класса, таким образом маловероятна их функция как эффекторных белков без дополнительных партнёров. Более того, локусы V-U подтипа не имеют других дополнительных *cas* генов (см. Рисунок 8), что, вместе со структурными соображениями приведёнными выше, предполагают, что нельзя сделать уверенное предсказание о том, что эти системы обладают полноценной CRISPR активностью. Тем не менее, эволюционно стабильная ассоциация с CRISPR кассетами, по крайней мере, в пяти различных вариантов подтипов V-U указывает на то, что некоторые из этих белков выполняют какую-то CRISPR зависимую биологическую функцию. Подобная функция может принимать типичную CRISPR активность, которая поддерживается другими Cas белками из других локусов и/или другими не Cas белками. Стоит отметить, что CRISPR кассеты ассоциированные с группой V-U3, которая найдена в основном в бациллах и в кластридиях, содержат несколько спэйсеров, которые совпадают с геномными последовательностями бактериофагов, которые инфицируют данные бактерии (см сопроводительные материалы S2 (box), часть e). Более того, наборы спэйсеров в каждой группе подтипов V-U отличны друг от друга, даже в близких бактериальных геномах (см. сопроводительная информация S2 (box), часть e), что указывает на активный набор спэйсеров. Разнообразие спэйсеров и присутствие спэйсеров таргетирующих фаги для подтипа V-U3 говорит о том, что по крайней мере

несколько из вариантов подтипа V-U являются функциональными CRISPR–Cas системами, которые вовлечены в антифаговый адаптивный иммунитет. Большое количество полных геномов, которые содержат локусы подгруппы V-U3 и подгруппы V-U4, не имеют других CRISPR–Cas систем (см. сопроводительная информация S2 (box), часть f), что оставляет не ясным механизм набора новых спэйсеров в CRISPR кассеты. Альтернативно, некоторые системы из подтипа V-U могут иметь определённую регуляторную активность, которая не требует формирования комплекса с crРНК и целевой ДНК; примеры подобных активностей не связанных с защитой были уже описаны [195]. Данный пример может подходить для подгруппы V-U5, которая, по всей видимости, представляет из себя каталитически не активный гомолог TnpB (см. Рисунок 9, где он обозначен как C2c5*; сопроводительная информация S3 (box)). Более того, в геномах, которые содержат подгруппы локусов V-U2 и V-U5 вместе с другими CRISPR–Cas системами, CRISPR последовательности (повторы) ассоциированные с V-U локусами уникальны (сопроводительная информация S2 (box), часть f), что указывает на то, что эти подтипы могут иметь иные функции.

Подтипы VI-A, VI-B и VI-C найденные с помощью cas1 и CRISPR затравок: РНК таргетирующие CRISPR–Cas многоблоковые эффекторы

Отличительной особенностью Тип VI систем является эффекторный белок, который содержит два HEPN домена (см. Рисунки 8, 9). HEPN домен (Higher Eukaryotes and Prokaryotes Nucleotide-binding) распространён среди различных систем защиты, среди них, которые были экспериментально охарактеризованы, токсины из многих прокариотических токсин-антитоксин систем или эукариотическая RNase L, все из них имеют РНКазную активность [6, 60, 108]. Таким образом, для первого предполагаемого Тип VI эффектора, обозначенного как C2c2, который был найден по ассоциации с *cas*1, была предсказана РНК направляемая РНКазная активность.

Поиски по геномным базам данных, показали отсутствие схожести последовательности C2c2 к другим известным белкам. Однако, изучение множественного выравнивания белков C2c2 выявило два консервативных R(N)xxxH мотива, которые являются характеристической особенностью HEPN домена [60]. В дополнение к этому, был обнаружен консервативный глютамат встроенный в хорошо предсказанную, длинную α-спираль и соответствующий схожим мотивам HEPN доменов (см. Рисунок 13).

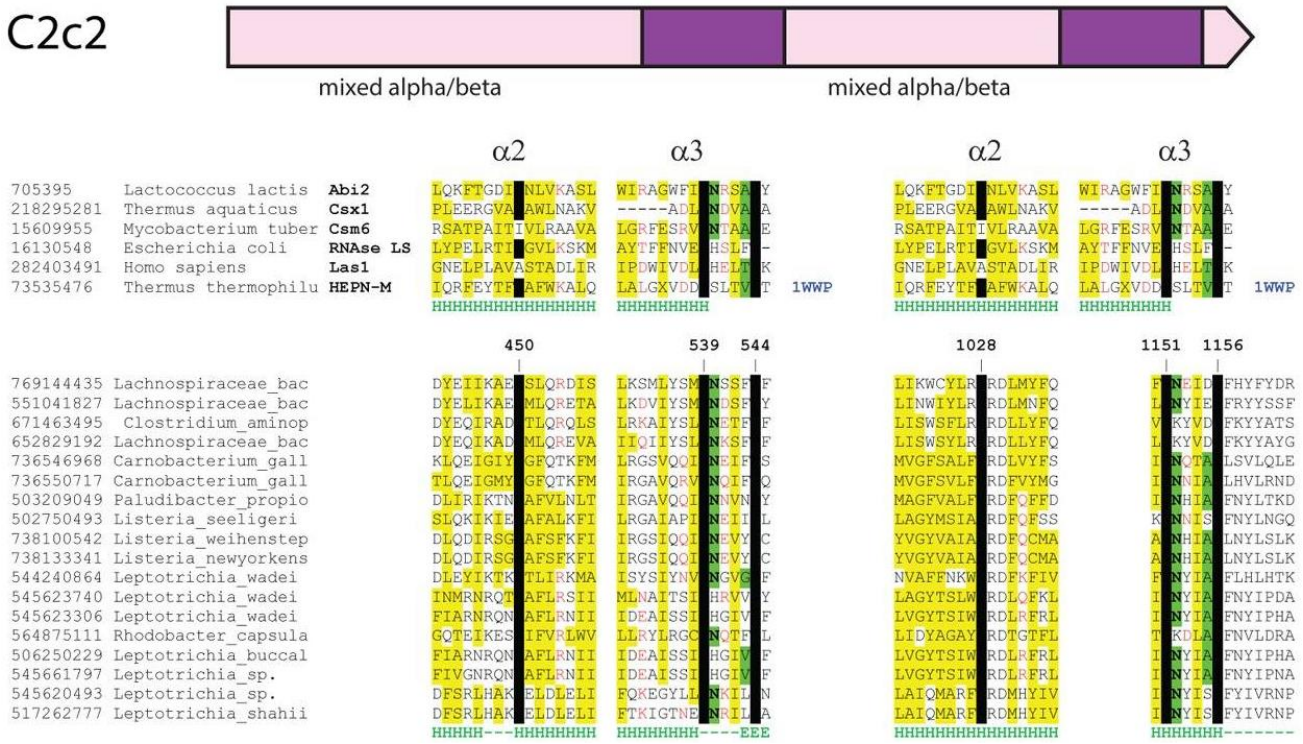


Рисунок 13. Тип VI: предсказанная РНКаза содержащая два HEPN домена (воспроизведено с разрешения Molecular Cell [173]). Верхний блок выравнивания включает выбранные HEPN домены, описанные выше и нижний блок выравнивания, включает каталитические мотивы из предсказанных Тип VI эффекторных белков. Каталитические основания показаны белыми буквами на черном фоне; консервативные гидрофобные основания помечены желтым фоном; консервативные малые основания помечены зелёным фоном; позитивно заряженные основания, в спиральном мостике, помечены красным цветом. Предсказание вторичной структуры показано ниже выравненных последовательностей: H обозначает α-спираль и E обозначает удлинённую конформацию (β-тяж). Плохо выравниваемые промежутки между выравниваниями показано цифрами.

HEPN суперсемейство включает в себя маленькие (~150 аминокислот) α-спиральные домены с сильно различающимися последовательностями, но сильно консервативными каталитическими мотивами, для которых была показано или предсказан РНКазная активность [6]. Поиск Pfam [122] в базах данных используя HHpred программу [180] и последовательности C2c2 как входные параметры, определило схожесть к HEPN домену для обоих предсказанных нуклеазных доменов C2c2, хотя и на очень значимом уровне. Однако важно, что эти выравнивания были единственными, где R(N)xxxH мотивы были консервативны. Идентификация HEPN доменов в C2c2 белках далее была поддержана предсказанием вторичной структуры, которая показала, что каждый мотив располагался среди совпадающих структурных контекстов, и предсказанная α-спиральная вторичная структура каждого домена согласуется с HEPN консервативным доменом (см. Рисунок 13). Вне этих HEPN доменов, C2c2 последовательность, согласно предсказаниям (см. Материалы и методы) принимает смешанную α/β структуру без какой-либо значимой схожести с известными белковыми консервативными доменами (см. сопроводительная информация S5). Основываясь на этих уникальных ключевых особенностях эффектора C2c2, данным системам была выделена отдельная Тип VI CRISPR-Cas система.

Впоследствии, таргетирование РНК было экспериментально проврено и показано, что Тип VI эффекторы защищают от РНК бактериофага MS2 [1]. В дополнении к этому, была показана новая уникальная особенность C2c2, что после нахождения цели, эффектор становится неразборчивой РНКазой, что имеет токсичный, подавляющий рост эффект на бактерию. Данные наблюдения демонстрируют связывание адаптивной иммунностью и запрограммированной смертью клетки (или впадение в спячку), что было ранее предсказано через сравнительный геномный анализ [107] и математическое моделирование [77]. Далее было показано, что C2c2 белок способен не только к интерференции, но также и к процессингу pre-crРНК [43].

Поиск CRISPR–Cas локусов используя CRISPR затравки позволил обнаружить два дополнительных семейства эффекторов, которым были назначены подтип VI-B и подтип VI-C, соответственно (согласно этому, C2c2 кодирующие локусы стали подтипом VI-A). Данная классификация Тип VI систем на разные подтипы была обоснована очень слабым

сходством на уровне последовательностей между этими тремя группами (которое только ограничено сходство в каталитических мотивах HEPN домена), различными позициями HEPN доменов в последовательности этих эффекторов и дополнительными особенностями строения локуса в случае подтипа VI-B (см. Рисунки 8, 9; сопроводительная информация S2 (box), часть d). В частности, было выделено два различных варианта подтипа VI-B (вариант VI‑B1 и VI‑B2), оба кодирующие дополнительные белки, которые содержат предсказанные трансмембранные домены; в случае VI‑B1 белок кодирует 4 таких домена и в VI‑B2 кодируется только один (см. Рисунок 10b; сопроводительная информация S2 (box), часть d). Филогенетический анализ эффекторных белков, указывает на то, что VI‑B1 и VI‑B2 варианты разошлись во время эволюции согласно их различным архитектурам ассоциированных предсказанных мембранных белков (см. Рисунок 10b; сопроводительная информация S2 (box), часть d). VI‑B1 системы, которые содержат несколько трансмембранных доменов, могут локализироваться на мембранах и, таким образом, быть мембранно-ассоциированными РНК-таргетирующими системами, что будет уникальным примером для биологии CRISPR–Cas систем. Однотрансмембранный белок в варианте VI‑B2 включает также в себя дополнительный HEPN домен, который является третьим в данной Тип VI системе (см. Рисунок 10b; сопроводительная информация S2 (box), часть d, и сопроводительная информация S6 (рисунок)).

Тип VI-B был экспериментально охарактеризован [178] и было показано, что он является функциональным и обладает РНКазной активностью. В этом исследовании было показано, что VI-B1 и VI-B2 могут регулировать РНК интерференцию (VI-B1 подавляет и VI-B2 улучшает).

Все найденные Тип VI эффекторы схожи по размеру с активными эффекторами 2 класса подтипа VI-A, что указывает на их возможную функциональность. Даже локусы, не имеющие Cas1, вероятно являются функционирующими CRISPR–Cas системами, которые полагаются на адаптационные модули из других локусов того же генома. Более того, учитывая, что РНК вирусы представляют только малую часть прокариотического вирома [96], системы VI типа могут показывать токсичное действие в качестве отклика

на активную транскрипцию чужеродной ДНК. Данный механизм, возможно, не ограничен только VI типом систем, учитывая присутствие HEPN доменов в плохо охарактеризованных Cas белках во многих других системах. Это подтверждает то, что РНКазная активность HEPN доменов в Csm6 и Csx1 была показана для систем III типа [137, 172], хотя их функции в биологии CRISPR-Cas систем остаются не известными.

Часть 2. Оценка разнообразия систем 2 класса и обновлённая классификация CRISPR-Cas систем

Оценка разнообразия CRISPR-Cas систем 2 класса в локусах бактерий и архей

Исчерпывающая оценка разнообразия всех типов и подтипов 2 класса CRISPR-Cas систем была сделана в этом исследовании на текущем наборе бактериальных и архейных геномов. Были созданы профили для всех эффекторов всех найденных подтипов 2 класса систем (два отдельных профиля были созданы для вариантов V-U1, V-U2 и V-U5; V-U3 и V-U4 варианты не были включены в данную оценку, так как в поиске по базе данных они трудно отличимы от транспозон-кодирующих TnpB) и сравнены с белками закодированными в 4,961 полностью отсеквенированных прокариотических геномах и 43,599 частичных прокариотических геномах, которые доступны в National Center for Biotechnology Information (NCBI) базе данных [211] (см. Материалы и методы). Данная процедура должна определить почти все экземпляры каждого эффектора, включая дальних родственников.

Окружение данных генов было проверено на наличие CRISPR кассет и дополнительных *cas* генов, как было описано ранее [114]. Наиболее интересным наблюдением, является доминирование II типа систем во 2 классе. Эти системы представляют около 8% бактериальных геномов (см. Таблица 2). Тип V и Тип VI вместе взятые являются менее представленными более чем на порядок, что согласуется с

ожиданием, что различные варианты CRISPR–Cas типов и подтипов которые будут найдены, будут являться редкими вариантами среди CRISPR-Cas эффекторов [114].

	Subtype						
	II	V-A	V-B	V-U*	VI-A	VI-B	VI-C
Effector‡	Cas9	Cas12a (Cpf1)	Cas12b (C2c1)	C2c4, C2c5; five distinct subgroups (V-U 1–5)	Cas13a (C2c2)	Cas13b (C2c6)	Cas13c (C2c7)
Number of loci in bacterial and archaeal genomes	• 3,822 in total • 2,109 II-A • 130 II-B • 1,573 II-C • 10 unassigned	70	18	92	30	94	6
Representation	Diverse bacteria	Diverse bacteria and two archaea	Diverse bacteria	Diverse bacteria	Diverse bacteria	Bacteroidetes	Fusobacteria and Clostridia
Other cas genes	85% *cas1* and *cas2*; 55% *csn2*; 3% *cas4*	70% *cas1* and *cas2*; 55% *cas4*	65% *cas1*, *cas2* and *cas4*	None	25% *cas1* and *cas2*	None	None
Percent of loci that contain CRISPR array	65%	68%	60%	~50%	73%	90%	83%

Таблица 2. Исчерпывающая оценка разнообразия CRISPR–Cas систем второго класса в бактериальных и архейных геномах (воспроизведено с разрешения Nature reviews. Microbiology [175]).

*Локусы подтипа V‑uncharacterized (V-U) были изначально обнаружены путём поиска *tnpB* генов рядом с CRISPR кассетами и их консервативной эволюционной связью. Далее, этот предполагаемый подтип 2 класса был расширен за счет поиска гомологов, соответствующих эффекторных белков, независимо от их близости к CRISPR кассетам. Таким образом, только половина V-U локусов включает CRISPR.

‡Предложенная система именования и изначальные имена генов используются для именования эффекторов, за исключением II типа эффекторов, которые имеют только систематические имена и V-U эффекторов, которые систематических имён не имеют.

Одним из интригующих вопросов является, предоставляют ли системы II типа существенное преимущество в приспособленности против других вариантов эффекторов 2 класса, будучи более эффективной системой в плане защиты и/или путём меньших затрат.

Большинство подтипов CRISPR-Cas систем 2 класса представлены в таксономически разнообразных групп бактерий, и, более того, для II типа и подтипа V-A, топология филогенетических деревьев эффекторов отличается от топологии видов, где они присутствуют [30, 205]. Данные наблюдения указывают на то, что горизонтальный перенос генов может играть ключевую роль в эволюции CRISPR–Cas систем. Однако, стоит заметить, что относительно широкий подтип VI-B ограничен только типом Бактероидетов, что, возможно, отражает уникальный аспект биологии данных бактерий. Похожая ситуация происходит с вариантом V-U5, что содержит инактивированный гомолог TnpB, который был найден только в Цианобактериях (смотри выше), и может использоваться в определённых цианобактериальных регуляторных путях. Как было замечено ранее [112, 114], и подчеркнуто данным расширением разнообразия 2 класса CRISPR-Cas систем, за исключением нахождения двух экземпляров подтипа V-A в мезофильных археях, 2 класс ограничен бактериями. Исключение второго класса из архей, в частности из гипертермофилов, в которых системы 1 класса широко распространены, подразумевает, что существует важное функциональное различие между двумя классами CRISPR–Cas систем, природа которого остается не известной.

Обновлённая классификация CRISPR–Cas систем 2 класса

Систематический поиск новых CRISPR–Cas локусов 2 класса, описанный в данной работе, привёл к важному расширению известного разнообразия этих систем. Вместо двух типов и четырёх подтипов, которые были включены в последнюю классификацию [114], стало три типа и как минимум 10 подтипов (см. Рисунок 8). Некоторая неясность остаётся в связи с недостатком функциональных данных о подтипе V-U, но очень вероятно, что эволюционно стабильные и видимо функциональные варианты, которые сгруппированы в этом предварительном подтипе, в частности V-U3, получат собственные подтипы в V типе. Функциональная характеризация вариантов V-U позволит сделать более точную классификацию, хотя вероятно, что многие V-U локусы не кодируют типичные активные CRISPR–Cas системы. Исходя из исчерпывающей

природы данного исследования (см. Рисунок 7), предполагается, что новые варианты будут ещё более редкими или ограничены в своем распространении определёнными группами бактерий или архей, которые не представлены адекватно в текущих геномных базах данных.

Данное расширение классификации CRISPR–Cas систем требует соответствующего изменения в номенклатуре, в которой новые гены, по крайней мере экспериментально охарактеризованные эффекторы и их гомологи, будут иметь имена соответствующие номерным Cas белкам (см. Рисунок 8; Таблица 2). Таким образом, эффекторы V типа названы Cas12a, Cas12b и Cas12c и эффекторы VI типа именованы Cas13a, Cas13b и Cas13c (продолжение нумерации после Cas9 не возможно, так как Cas10 и Cas11 уже использованы для других белков) [114]. Эффекторы предполагаемого подтипа V-U не именуется до тех пор, пока не будет показано, что они функционально обладают bona fide CRISPR характеристиками, в случае, если это будет показано, они будут относиться к Cas12 семейству.

Часть 3. Эволюционное возникновение новых CRISPR-Cas систем 2 класса

В дополнение к предыдущей гипотезе о независимом происхождении эффекторов разных типов и подтипов 2 класса CRISPR–Cas систем, была использована информация о не полных локусах V типа, чтобы предложить более детальный эволюционный сценарий (см. Рисунок 14).

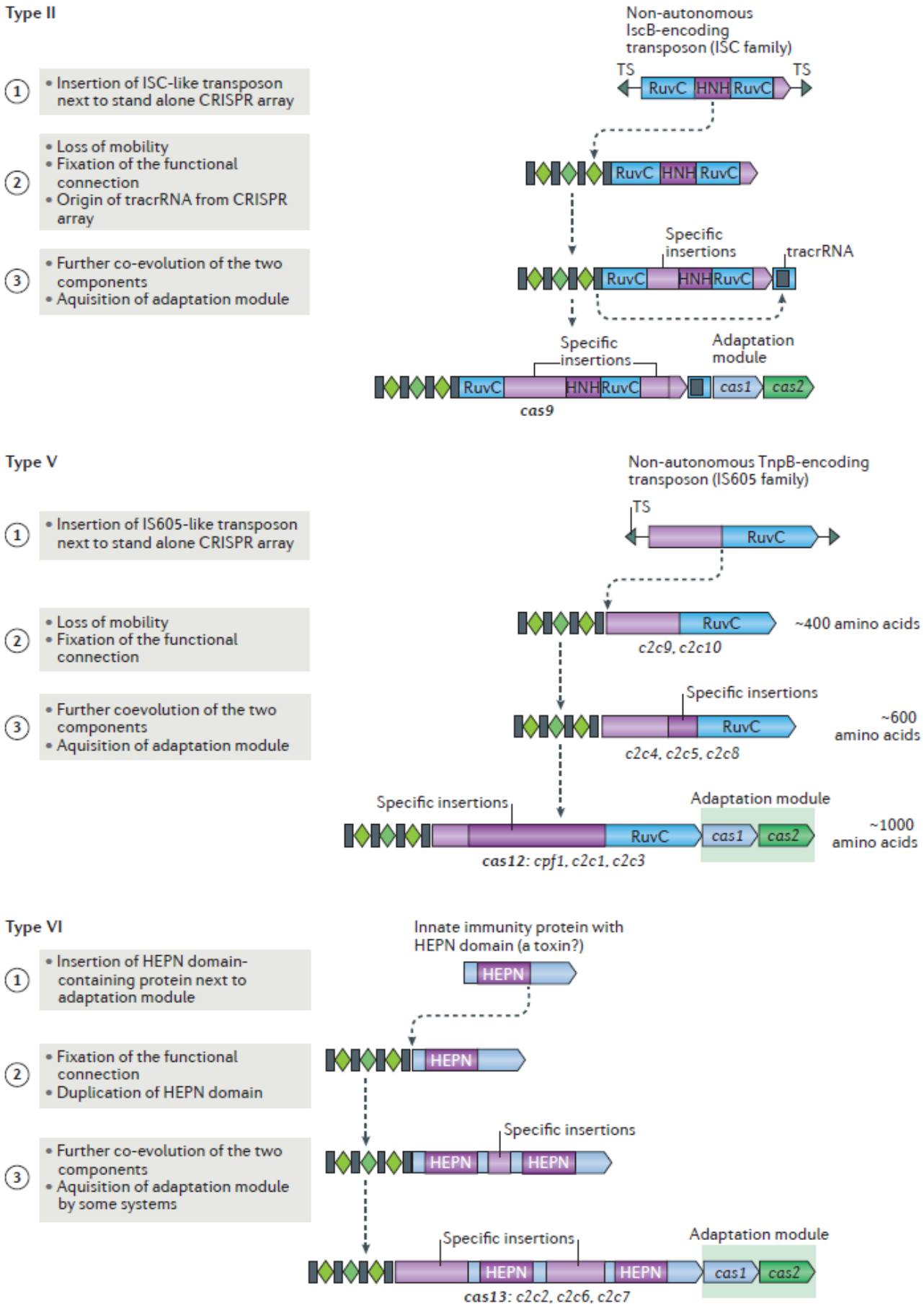

Type II
1
• Insertion of ISC-like transposon next to stand alone CRISPR array
2
• Loss of mobility
• Fixation of the functional connection
• Origin of tracrRNA from CRISPR array
3
• Further co-evolution of the two components
• Aquisition of adaptation module
Non-autonomous IscB-encoding transposon (ISC family)
TS
RuvC
HNH
RuvC
TS
Specific insertions
tracrRNA
Adaptation module
cas1
cas2
cas9
Type V
1
• Insertion of IS605-like transposon next to stand alone CRISPR array
2
• Loss of mobility
• Fixation of the functional connection
3
• Further coevolution of the two components
• Aquisition of adaptation module
Non-autonomous TnpB-encoding transposon (IS605 family)
~400 amino acids
c2c9, c2c10
~600 amino acids
c2c4, c2c5, c2c8
~1000 amino acids
cas12: cpf1, c2c1, c2c3
Type VI
1
• Insertion of HEPN domain-containing protein next to adaptation module
2
• Fixation of the functional connection
• Duplication of HEPN domain
3
• Further co-evolution of the two components
• Aquisition of adaptation module by some systems
Innate immunity protein with HEPN domain (a toxin?)
HEPN
cas13: c2c2, c2c6, c2c7

Рисунок 14. Предложенные варианты возникновения CRISPR–Cas систем 2 класса (воспроизведено с разрешения Nature reviews. Microbiology [175]). Данный рисунок отображает трёх этапный путь эволюционного “становления” для CRISPR-Cas систем II, V и VI типа. Систематичные и/или изначальные названия для имён генов расположены ниже “взрослых” эффекторных белков и предположительных промежуточных форм для систем V типа. Первый шаг включает случайную вставку TnpB кодирующей последовательности или последовательности IscB (Cas9-like protein B) – кодирующего транспозона или HEPN домена (higher eukaryotes and prokaryotes nucleotide-binding) РНКазы-кодирующего гена рядом с CRISPR кассетой для Тип II, Тип V и тип VI систем, соответственно. Во время второго шага, устанавливается функциональная связь между встроенным белком и CRISPR кассетой, далее начинается их коэволюция, в частности, в форме накопления вставок, которая содействует CRISPR РНК (crРНК) связыванию. Для систем V типа, промежуточные формы, советующие первому и последнему шагу обозначены как различные варианты V-uncharacterized (V-U) типа. Дополнительные компоненты системы, которые могли появиться во время 2 шага, такие как tracrРНК (*trans*-acting CRISPR РНК) в случае систем II типа. Во время третьего шага, дальнейшие вставки привели к улучшенной специфичности crРНК и связыванием с целью, и добавили возможность взаимодействия с вспомогательными белками, такими как Csn2 для II-A типа или белка с трансмембранным доменом (TM) для типа VI-B. Адаптационный модуль был встроен только для некоторых CRISPR-Cas систем 2 класса во время третьего шага. (TS) обозначает таргетируемую последовательность (Target Site).

Как описано выше, по крайней мере, пять различных вариантов внутри подтипа V-U показывают свойства эволюционной стабильности и стойкую связь с CRISPR кассетами, и обычно содержат TnpB гомологи имеющие средний размер между компактными транспозон-кодирующими TnpB белками и большими эффекторами 2 класса (см. Рисунок 9, 10b). Эти группы TnpB гомологов могут представлять собой промежуточные стадии в независимом процессе становления нового варианта CRISPR–Cas системы. Другие CRISPR–*tnpB* экземпляры не проявляют эволюционной стабильности и вероятно являются результатом более или менее случайной вставки *tnpB* генов рядом с CRISPR кассетой; некоторые из этих локусов могут представлять собой ранние этапы эволюции CRISPR–Cas систем.

Все локусы V-U подтипов не имеют адаптационного модуля, что указывает на то, что на ранние стадии эволюции новых CRISPR–Cas систем 2 класса происходят путём случайной вставки TnpB-кодирующего элемента рядом с одиночной CRISPR кассетой (см. Рисунок 14). На следующей стадии эволюции происходит фиксация ассоциации между CRISPR кассетой и TnpB гомологом в микробной популяции, предположительно в связи с появлением новой функции, точная природа которой пока не известна. Это может быть сопровождено увеличением размера белка через дупликацию внутренних участков и/или дополнительных вставок доменов (см. Рисунок 14). Последний этап включает в себя дальнейший рост эффекторного белка, достигая типичной билобной структуры и, в некоторых случаях, ассоциацией с адаптационным модулем путём рекомбинации с другими CRISPR–Cas локусами (см. Рисунок 14). Данный сценарий может быть подтверждён тем, что Cas1 белки из различных подтипов II типа и V типа гомологичны различным подтипам I типа (см. Рисунок 15). Факт того, что ни один локус V-U подтипа не содержит *cas1* и *cas2* гены, в то время как многие локусы кодирующие большие эффекторные белки содержат их, хорошо поддерживает гипотезу, что адаптационный модуль появляется последним.

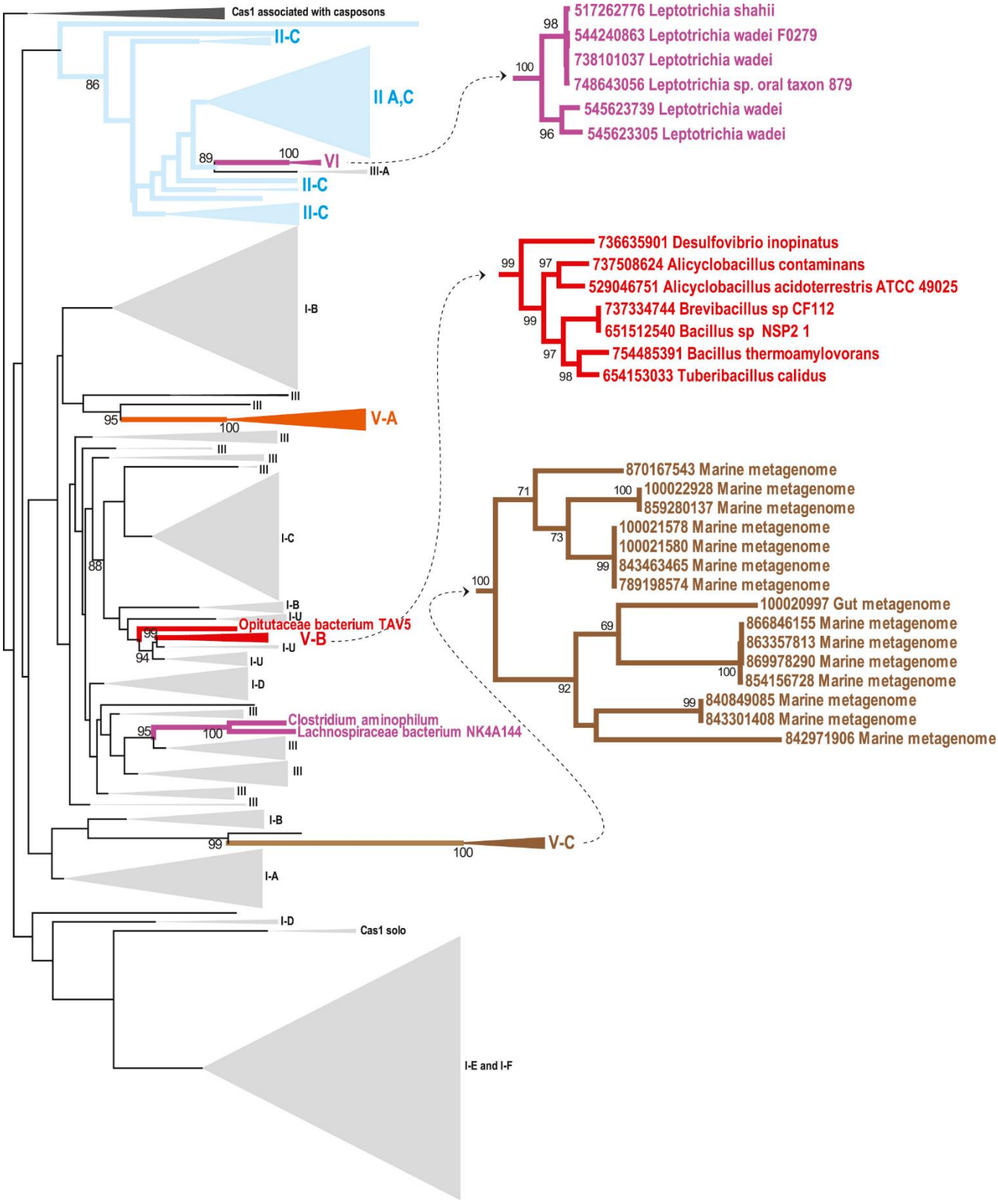

Cas1 associated with casposons
II-C
II A,C
VI
III-A
II-C
II-C
517262776 Leptotrichia shahii
544240863 Leptotrichia wadei F0279
738101037 Leptotrichia wadei
748643056 Leptotrichia sp. oral taxon 879
545623739 Leptotrichia wadei
545623305 Leptotrichia wadei
I-B
III
V-A
III
III
III
III
I-C
I-B
I-U
Opitutaceae bacterium TAV5
V-B
I-U
I-U
I-D
III
Clostridium aminophilum
Lachnospiraceae bacterium NK4A144
III
III
III
III
I-B
V-C
I-A
I-D
Cas1 solo
I-E and I-F
736635901 Desulfovibrio inopinatus
737508624 Alicyclobacillus contaminans
529046751 Alicyclobacillus acidoterrestris ATCC 49025
737334744 Brevibacillus sp CF112
651512540 Bacillus sp NSP2 1
754485391 Bacillus thermoamylovorans
654153033 Tuberibacillus calidus
870167543 Marine metagenome
100022928 Marine metagenome
859280137 Marine metagenome
100021578 Marine metagenome
100021580 Marine metagenome
843463465 Marine metagenome
789198574 Marine metagenome
100020997 Gut metagenome
866846155 Marine metagenome
863357813 Marine metagenome
869978290 Marine metagenome
854156728 Marine metagenome
840849085 Marine metagenome
843301408 Marine metagenome
842971906 Marine metagenome

Рисунок 15. Филогенетическое дерево Cas1 (воспроизведено с разрешения Molecular Cell [173]). Данное дерево было построено используя множественное выравнивание 1498 Cas1 последовательностей, которые содержат 304 филогенетически информативных позиций. Ветки, соответствующие 2 классу систем помечены голубым цветом; оранжевым для подтипа V-A; красным для подтипа V-B; коричневым для подтипа V-C; малиновым для типа VI. Выноски показывают раскрытые ветки для новых подтипов. Bootstrap значения показаны в процентах и отображены только для нескольких релевантных веток. Полное дерево в Newick формате с названиями видов и bootstrap значениями и множественное выравнивание, которое использовалось для постройки дерева доступно на ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Class2/. Также см. Материалы и методы.

Приведённый выше сценарий может быть оспорен с точки зрения направления эволюции: можно представить, что транспозон-кодирующий TnpB белок произошел из эффектора 2 класса. Но всё же, сценарий, когда транспозон-кодирующий TnpB является предком (см. Рисунок 14) является более предпочтительным по нескольким причинам. Первое, TnpB-кодирующие транспозоны (автономные и не автономные, включая те, что потеряли мобильность) много более распространены во всех видах бактерий и архей чем 2 класс CRISPR–Cas систем, которые относительно редки и ограничены в их распространении в определённом наборе некоторых типов бактерий (см. Оценка разнообразия 2 класса CRISPR-Cas систем; Таблица 2; сопроводительная информация S2 (box), часть d). Второе, возможно более важное, эффекторы 2 класса сильно больше и сложнее чем TnpB белки, что делает их маловероятными предковыми формами. Третье, TnpB белки закодированы в транспозонах, которые, в силу их мобильности, могут легко оказаться в окрестности CRISPR кассет; CRISPR–Cas системы, напротив, не имеют механизмов мобильности. Наконец, наблюдения, показанные здесь, о филогении TnpB белков, которые ассоциированы с CRISPR кассетами расположены между транспозон-кодирующими белками (см. Рисунок 10a), что подразумевает предковый статус за TnpB.

Гипотетически, похожий сценарий можно применить к типу VI CRISPR-Cas систем (см. Рисунок 14). Подробный поиск по HEPN домен-содержащих белков в геномных базах данных, которые закодированы рядом с CRISPR кассетами, не был способен найти эволюционно стабильные группы, которые могли бы быть аналогами подтипу V-U, тогда

как обнаружил многочисленные группы HEPN содержащих Cas белковых семейств, Csm6 и Csx1 (см. сопроводительную информацию S2 (box), часть g). Таким образом, возможно, что во время эволюции VI тип систем рекрутировал один из HEPN содержащих Cas белков, который после последовавшей дупликации HEPN домена и дальнейшего расширения белка до типичного размера 2 класса эффекторов, перерос в окончательный вариант VI (см. Рисунок 14). Однако, возможность того, что VI эффекторы происходят непосредственно от HEPN-содержащих токсинов, не может быть полностью исключена; дальнейшее исследование новых геномов и метагеномов на счет возможных предков двух-HEPN домен содержащих белков может указать место происхождения этих эффекторов более точно.

Часть 4. Возможные применения для новых CRISPR-Cas систем

Большинство методов применения CRISPR систем было сфокусировано на программируемой ДНК таргетирующей активности белка Cas9. Направленное внесение двухнитевых разрывов в ДНК с помощью Cas9 может быть использовано для редактирования геномов, включая нокаут генов и точное редактирование используя внутриклеточный механизм гомологичной рекомбинации. Каталитически неактивные варианты (‘dead’) Cas9 были использованы для контроля транскрипции [26], эпигенетической модуляции [190] и для внутриклеточной визуализации [27, 93, 136]. Не смотря на этот прогресс, Cas9 имеет свои недостатки, в связи с потенциальными вне целевыми эффектами, проблемами, ассоциированными с доставкой комплекса и сложностями таргетирования РНК вместо ДНК. Таким образом, альтернативные инструменты для редактирования геномов сильно ожидаемы.

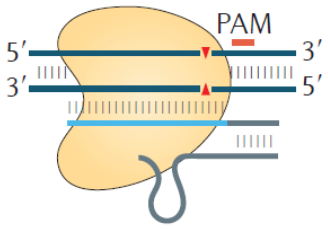
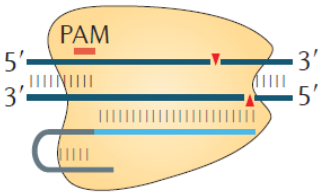
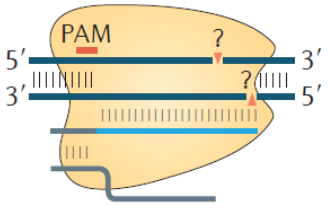
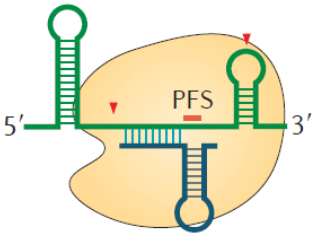
		Nuclease domains	tracrRNA	PAM	Substrate	Cleavage pattern
Type II Cas9		RuvC and HNH	Yes	3′, GC-rich	dsDNA	Blunt ends
Type V-A Cas12a (Cpf1)		RuvC and Nuc	No	5′, AT-rich	dsDNA	Staggered ends, 5′overhangs
Type V-B Cas12b (C2c1)		RuvC	Yes	5′, AT-rich	dsDNA	Staggered seven-nucleotide cut of target DNA
Type VI-A Cas13a (C2c2)		2 HEPN domains	No	5′, non-G PFS	ssRNA	Cleaves ssRNA near uracil and collateral activity

Рисунок 16. Функциональное разнообразие экспериментально охарактеризованных CRISPR–Cas систем 2 класса (воспроизведено с разрешения Nature reviews. Microbiology [175]). Для каждого типа CRISPR–Cas систем 2 класса (и двух подтипов в V типа), показано схематичное изображение комплексов с эффектором, целью, crРНК и в случае II типа и подтипа V-B систем, *trans*-acting CRISPR РНК (tracrРНК). Позиция PAM (protospacer adjacent motif) или PFS (protospacer flanking site) показана красной линией. Маленькие зелёные треугольники показывают место разреза или разрезов таргетируемой ДНК или РНК молекулы: dsДНК, двухцепочечная ДНК (double-stranded ДНК); ssРНК, одноцепочечная РНК (single-stranded РНК).

Прогресс в функциональной характеризации подтипов 2 класса, который далек от завершения, даже на данной стадии, показывает удивительное разнообразие механизмов эффекторов. Проявления этого разнообразия включают в себя: различные цели

эффекторов (двухцепочечные ДНК для II типа и V типа, но РНК для VI типа), наличие tracrРНК (для II типа и подтипа V-B, но не для подтипа V-A или типа VI), последовательность PAM'а и тип раскусывания цели на нуклеотидном уровне (см. Рисунок 16). Данное функциональное разнообразие является основным стимулом для дальнейшей характеризации различных систем 2 класса так как это создаёт возможности для расширения и экспансии этих возможностей редактирования геномов для исследований, биотехнологии и медицины [73]. Использование Cas12a (ранее известного как Cpf1) из подтипа V-A семейств эффекторов уже принесло более простой РНК направляемый и более специфичный чем Cas9 белок для приложений редактирования геномов [50, 76, 90, 92, 103, 205], также он предлагает альтернативный PAM который легче применять в AT-богатых геномах, как, например, *Plasmodium falciparum*.

Дальнейшее исследование разнообразия CRISPR эффекторов, таких как недавно охарактеризованный подтип VI-A эффектор Cas13a (ранее известный как C2c2) [1], также открыл двери для развития новых технологий РНК-направляемого РНК таргетирования, которые позволяют изменять, модулировать, модифицировать и отслеживать специфичные РНК транскрипты в клетках. Развитие эффективного программируемого РНК-связывающегося белка (например, 'dead' Cas13a, который имеет мутированные HEPN домены) может сильно улучшить существующее понимание РНК биологии. Подобный инструмент может позволить наблюдать различные состояния клеток, манипулировать трансляцией, отслеживать уровни РНК локализации в живых клетках. Не смотря на то, что Cas9 был модифицирован, чтобы иметь некоторые способности к РНК таргетированию [144], эта система требует доставки химически модифицированной ДНК, что сужает область применения, исключая широкогеномный скрининиг или доставку вирусами.

После связвывания с комплементарной РНК целью, Cas13a использует специфичную и не специфичную РНКазную активности, таким образом вызывая ингибирование роста в *Escherichia coli* [1]. Данная особенность усложняет использование Cas13a для специфичного РНК нокаута, но может быть потенциально использована для других приложений, таких как селективное устранение клеток, основанное на профилях экспрессии. Остаётся не выясненным можно ли инактивировать не специфичную РНК

активность в Cas13a или использовать независимо от его специфичной к цели активности и имеют ли другие эффекторы подтипов VI типа, такие как Cas13b, подобные свойства. Дальнейший поиск CRISPR–Cas систем или, в более широком плане, разнообразия бактериальных и архейных систем защиты и мобильных генетических элементов, может дать новые возможности и применения в биотехнологии. В частности, программируемые интегразы или транспозазы, которые ещё предстоит найти, могут быть мощными инструментами для специфичной геномной интеграции и перестроек.

Недавние исследования показывают примеры применения обнаруженного белка Cas13a [58]. В данном исследовании показано, что Cas13a может быть использован в специфичной диагностике, где он может детектировать ДНК или РНК на аттомолярных уровнях и уровнях с всего одной заменой. Этот метод был использован для определения специфичных штаммов эукариотических и прокариотических вирусов, различных видов опухолей, нахождения патогенных бактерий. Данные свойства можно применять в полевых условиях для определения инфекций, патогенов, определения количества ДНК/РНК и т.д.

Заключение

Геномный анализ, представленный в данной работе, увеличил разнообразие CRISPR–Cas систем 2 класса. В частности, поиск автономных и не автономных (не имеющих адаптационного модуля) CRISPR–Cas систем в геномных и метагеномных базах данных, привёл к обнаружению шести новых подтипов, увеличив количество подтипов во 2 классе CRISPR-Cas систем с 4 до 10. Более того, один из новых подтипов, V-U, являющийся набором различных вариантов маленьких потенциальных эффекторов, некоторые из которых ожидаемо могут стать полноценными подтипами V типа, после того как они будут полноценно охарактеризованы. Особенно примечательно, что обнаруженные новые системы 2 класса попадают в две ранее известные категории: эффекторы, которые раскусывают двуцепочечную ДНК используя RuvC подобную нуклеазу (для не целевой цепи); и эффекторы, которые атакуют РНК цели используя HEPN РНКазный домен. Повторяемое независимое появление этих систем отражает потребность прокариот в данном виде структур CRISPR–Cas вариантов, которые могут вместить crРНК и таргетируемую молекулу, для чего подходит не так много белковых структур.

Новые варианты 2 класса CRISPR-Cas систем имеют особенности, не встречавшиеся ранее: например, подтип V-B требует tracrРНК, тогда как V-A не требует этой молекулы, тогда как другие варианты, такие как подтип VI-A (и возможно все типы VI типа систем), таргетируют только РНК и, видимо, вызывают токсичный отклик в бактериальных клетках. Подтип V-U, как ожидается, может показать ещё более необычные свойства. Данное функциональное разнообразие обладает потенциалом для развития новых, разносторонних инструментов редактирования геномов и регуляторных механизмов. В данной работе, были показаны особенности, которые указывают на различное, независимое происхождение разных типов и подтипов CRISPR-Cas систем 2 класса из мобильных элементов, кодирующих разнородные TnpB белки (для II типа и V типа) или из HEPN домен-содержащих белков (для VI типа), которые изначально происходят из мРНК разрезающих токсинов. Несмотря на удивительное разнообразие найденных

систем, ожидается, что данный разработанный и применённый биоинформатический подход исчерпывающе нашёл варианты CRISPR-Cas систем 2 класса в доступных геномных и метагеномных данных. Новые варианты могут быть найдены, но они будут очень редки или ограничены не известным или плохо покрытым (в базах данных) типам бактерий. Однако, как показано на примере VI типа, не смотря на ограниченную представленность подобных вариантов их биологическая роль или особенности могут быть очень интересны и применимы для новых биотехнологических инструментов.

Результаты исследования

1. Был разработан и применён биоинформатический подход для поиска новых CRISPR-Cas систем 2 класса в геномных и метагеномных прокариотических базах данных.
2. Было обнаружено шесть новых CRISPR-Cas систем 2 класса, включая набор неизвестных вариантов типа V-U: подтипы V-B, V-C, V-U с RuvC подобными нуклеазными доменами; и VI-A, VI-B, VI-C с HEPN РНКазным доменом. Данные системы были биоинформатически охарактеризованы. Три подтипа были экспериментально проверены коллабораторами в других лабораториях и их результаты совпали со сделанными предсказаниями: Тип V-B является tracrРНК зависимым CRISPR-Cas эффекторным комплексом который предоставляет защиту от чужеродной ДНК [173]; типы VI-A и VI-B являются РНК направляемыми РНКазами и предоставляют защиту от РНК вирусов, а также вызывают ингибирование роста [1, 178].
3. Обновлённая классификация CRISPR-Cas систем 2 класса была предложена, которая включает 6 новых обнаруженных систем.
4. Исчерпывающая оценка разнообразия была сделана для CRISPR-Cas систем 2 класса, показывающая распространение и дающая количественную оценку этих систем среди бактерий и архей.
5. Были представлены гипотезы о возможных путях появления CRISPR-Cas систем 2 класса, показывающие развитие от транспозонов до развитых (или промежуточных для типа V-U) CRISPR-Cas систем.
6. Были предложены возможные биотехнологические применения, включая редактирование геномов, для новых CRISPR-Cas систем.

Список использованной литературы

1. Abudayyeh O.O. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. / O. O. Abudayyeh, J. S. Gootenberg, S. Konermann, J. Joung, I. M. Slaymaker, D. B. T. Cox, S. Shmakov, K. S. Makarova, E. Semenova, L. Minakhin, K. Severinov, A. Regev, E. S. Lander, E. V Koonin, F. Zhang // Science – 2016. – Т. 353 – № 6299– aaf5573c.

2. Almendros C. Anti-cas spacers in orphan CRISPR4 arrays prevent uptake of active CRISPR-Cas I-F systems. / C. Almendros, N. M. Guzmán, J. García-Martínez, F. J. M. Mojica // Nat. Microbiol. – 2016. – Т. 1 – № 8– 16081c.

3. Altschul S.F. Basic local alignment search tool. / S. F. Altschul, W. Gish, W. Miller, E. W. Myers, D. J. Lipman // J. Mol. Biol. – 1990. – Т. 215 – № 3– 403–10c.

4. Altschul S.F. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. / S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman // Nucleic Acids Res. – 1997. – Т. 25 – № 17– 3389–402c.

5. Amitai G. CRISPR-Cas adaptation: insights into the mechanism of action. / G. Amitai, R. Sorek // Nat. Rev. Microbiol. – 2016. – Т. 14 – № 2– 67–76c.

6. Anantharaman V. Comprehensive analysis of the HEPN superfamily: identification of novel roles in intra-genomic conflicts, defense, pathogenesis and RNA processing. / V. Anantharaman, K. S. Makarova, A. M. Burroughs, E. V Koonin, L. Aravind // Biol. Direct – 2013. – Т. 8– 15c.

7. Aravind L. SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families, phyletic distribution and evolutionary trajectories. / L. Aravind, K. S. Makarova, E. V Koonin // Nucleic Acids Res. – 2000. – Т. 28 – № 18– 3417–32c.

8. Bao W. Homologues of bacterial TnpB_IS605 are widespread in diverse eukaryotic transposable elements. / W. Bao, J. Jurka // Mob. DNA – 2013. – Т. 4 – № 1– 12c.

9. Barrangou R. CRISPR-Cas systems and RNA-guided interference. / R. Barrangou // Wiley

Interdiscip. Rev. RNA – 2013. – T. 4 – № 3– 267–78c.

10. Barrangou R. Applications of CRISPR technologies in research and beyond. / R. Barrangou, J. A. Doudna // Nat. Biotechnol. – 2016. – T. 34 – № 9– 933–941c.

11. Barrangou R. CRISPR provides acquired resistance against viruses in prokaryotes. / R. Barrangou, C. Fremaux, H. Deveau, M. Richards, P. Boyaval, S. Moineau, D. A. Romero, P. Horvath // Science – 2007. – T. 315 – № 5819– 1709–12c.

12. Beloglazova N. CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. / N. Beloglazova, K. Kuznedelov, R. Flick, K. A. Datsenko, G. Brown, A. Popovic, S. Lemak, E. Semenova, K. Severinov, A. F. Yakunin // Nucleic Acids Res. – 2015. – T. 43 – № 1– 530–43c.

13. Benson D.A. GenBank. / D. A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, E. W. Sayers // Nucleic Acids Res. – 2013. – T. 41– № Database issue– D36-42c.

14. Besemer J. GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. / J. Besemer, A. Lomsadze, M. Borodovsky // Nucleic Acids Res. – 2001. – T. 29 – № 12– 2607–18c.

15. Bhaya D. CRISPR-Cas systems in bacteria and archaea: versatile small RNAs for adaptive defense and regulation. / D. Bhaya, M. Davison, R. Barrangou // Annu. Rev. Genet. – 2011. – T. 45– 273–97c.

16. Bikard D. Exploiting CRISPR-Cas nucleases to produce sequence-specific antimicrobials. / D. Bikard, C. W. Euler, W. Jiang, P. M. Nussenzweig, G. W. Goldberg, X. Duportet, V. A. Fischetti, L. A. Marraffini // Nat. Biotechnol. – 2014. – T. 32 – № 11– 1146–50c.

17. Bikard D. Programmable repression and activation of bacterial gene expression using an engineered CRISPR-Cas system. / D. Bikard, W. Jiang, P. Samai, A. Hochschild, F. Zhang, L. A. Marraffini // Nucleic Acids Res. – 2013. – T. 41 – № 15– 7429–37c.

18. Blosser T.R. Two distinct DNA binding modes guide dual roles of a CRISPR-Cas protein complex. / T. R. Blosser, L. Loeff, E. R. Westra, M. Vlot, T. Künne, M. Sobota, C. Dekker, S.

J. J. Brouns, C. Joo // Mol. Cell – 2015. – T. 58 – № 1– 60–70c.

19. Bolotin A. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. / A. Bolotin, B. Quinquis, A. Sorokin, S. D. Ehrlich // Microbiology – 2005. – T. 151– № Pt 8– 2551–61c.

20. Bortesi L. The CRISPR/Cas9 system for plant genome editing and beyond. / L. Bortesi, R. Fischer // Biotechnol. Adv. – T. 33 – № 1– 41–52c.

21. Briner A.E. Guide RNA functional modules direct Cas9 activity and orthogonality. / A. E. Briner, P. D. Donohoue, A. A. Gomaa, K. Selle, E. M. Slorach, C. H. Nye, R. E. Haurwitz, C. L. Beisel, A. P. May, R. Barrangou // Mol. Cell – 2014. – T. 56 – № 2– 333–9c.

22. Brouns S.J.J. Small CRISPR RNAs guide antiviral defense in prokaryotes. / 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, J. van der Oost // Science – 2008. – T. 321 – № 5891– 960–4c.

23. Bult C.J. Complete genome sequence of the methanogenic archaeon, Methanococcus jannaschii. / C. J. Bult, O. White, G. J. Olsen, L. Zhou, R. D. Fleischmann, G. G. Sutton, J. A. Blake, L. M. FitzGerald, R. A. Clayton, J. D. Gocayne, A. R. Kerlavage, B. A. Dougherty, J. F. Tomb, M. D. Adams, C. I. Reich, R. Overbeek, E. F. Kirkness, K. G. Weinstock, J. M. Merrick, A. Glodek, J. L. Scott, N. S. Geoghagen, J. C. Venter // Science – 1996. – T. 273 – № 5278– 1058–73c.

24. Carte J. Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. / J. Carte, R. Wang, H. Li, R. M. Terns, M. P. Terns // Genes Dev. – 2008. – T. 22 – № 24– 3489–96c.

25. Charpentier E. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. / E. Charpentier, H. Richter, J. van der Oost, M. F. White // FEMS Microbiol. Rev. – 2015. – T. 39 – № 3– 428–41c.

26. Chavez A. Comparison of Cas9 activators in multiple species. / A. Chavez, M. Tuttle, B. W. Pruitt, B. Ewen-Campen, R. Chari, D. Ter-Ovanesyan, S. J. Haque, R. J. Cecchi, E. J. K. Kowal, J. Buchthal, B. E. Housden, N. Perrimon, J. J. Collins, G. Church // Nat. Methods –

2016. – T. 13 – № 7– 563–7c.

27. Chen B. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. / B. Chen, L. A. Gilbert, B. A. Cimini, J. Schnitzbauer, W. Zhang, G.-W. Li, J. Park, E. H. Blackburn, J. S. Weissman, L. S. Qi, B. Huang // Cell – 2013. – T. 155 – № 7– 1479–91c.

28. Chen L. Advances in genome editing technology and its promising application in evolutionary and ecological studies. / L. Chen, L. Tang, H. Xiang, L. Jin, Q. Li, Y. Dong, W. Wang, G. Zhang // Gigascience – 2014. – T. 3– 24c.

29. Cho S.W. Targeted genome engineering in human cells with the Cas9 RNA-guided endonuclease. / S. W. Cho, S. Kim, J. M. Kim, J.-S. Kim // Nat. Biotechnol. – 2013. – T. 31 – № 3– 230–2c.

30. Chylinski K. Classification and evolution of type II CRISPR-Cas systems. / K. Chylinski, K. S. Makarova, E. Charpentier, E. V Koonin // Nucleic Acids Res. – 2014. – T. 42 – № 10– 6091–105c.

31. Chylinski K. The tracrPHK and Cas9 families of type II CRISPR-Cas immunity systems. / K. Chylinski, A. Le Rhun, E. Charpentier // RNA Biol. – 2013. – T. 10 – № 5– 726–37c.

32. Cong L. Multiplex genome engineering using CRISPR/Cas systems. / L. Cong, F. A. Ran, D. Cox, S. Lin, R. Barretto, N. Habib, P. D. Hsu, X. Wu, W. Jiang, L. A. Marraffini, F. Zhang // Science – 2013. – T. 339 – № 6121– 819–23c.

33. Cyranoski D. CRISPR gene-editing tested in a person for the first time / D. Cyranoski // Nature – 2016. – T. 539 – № 7630– 479–479c.

34. D’Astolfo D.S. Efficient intracellular delivery of native proteins. / D. S. D’Astolfo, R. J. Pagliero, A. Pras, W. R. Karthaus, H. Clevers, V. Prasad, R. J. Lebbink, H. Rehmann, N. Geijsen // Cell – 2015. – T. 161 – № 3– 674–90c.

35. Datsenko K.A. Molecular memory of prior infections activates the CRISPR/Cas adaptive bacterial immunity system. / K. A. Datsenko, K. Pougach, A. Tikhonov, B. L. Wanner, K. Severinov, E. Semenova // Nat. Commun. – 2012. – T. 3– 945c.

36. Deltcheva E. CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. / E. Deltcheva, K. Chylinski, C. M. Sharma, K. Gonzales, Y. Chao, Z. A. Pirzada, M. R. Eckert, J. Vogel, E. Charpentier // Nature – 2011. – T. 471 – № 7340– 602–7c.

37. Deng L. A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. / L. Deng, R. A. Garrett, S. A. Shah, X. Peng, Q. She // Mol. Microbiol. – 2013. – T. 87 – № 5– 1088–99c.

38. Deveau H. Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. / H. Deveau, R. Barrangou, J. E. Garneau, J. Labonté, C. Fremaux, P. Boyaval, D. A. Romero, P. Horvath, S. Moineau // J. Bacteriol. – 2008. – T. 190 – № 4– 1390–400c.

39. DiCarlo J.E. Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. / J. E. DiCarlo, J. E. Norville, P. Mali, X. Rios, J. Aach, G. M. Church // Nucleic Acids Res. – 2013. – T. 41 – № 7– 4336–43c.

40. Dillingham M.S. RecBCD enzyme and the repair of double-stranded DNA breaks. / M. S. Dillingham, S. C. Kowalczykowski // Microbiol. Mol. Biol. Rev. – 2008. – T. 72 – № 4– 642–71, Table of Contentsc.

41. Doench J.G. Optimized sgPHK design to maximize activity and minimize off-target effects of CRISPR-Cas9. / J. G. Doench, N. Fusi, M. Sullender, M. Hegde, E. W. Vaimberg, K. F. Donovan, I. Smith, Z. Tothova, C. Wilen, R. Orchard, H. W. Virgin, J. Listgarten, D. E. Root // Nat. Biotechnol. – 2016. – T. 34 – № 2– 184–91c.

42. Dong D. The crystal structure of Cpf1 in complex with CRISPR RNA. / D. Dong, K. Ren, X. Qiu, J. Zheng, M. Guo, X. Guan, H. Liu, N. Li, B. Zhang, D. Yang, C. Ma, S. Wang, D. Wu, Y. Ma, S. Fan, J. Wang, N. Gao, Z. Huang // Nature – 2016. – T. 532 – № 7600– 522–6c.

43. East-Seletsky A. Two distinct RNase activities of CRISPR-C2c2 enable guide-RNA processing and RNA detection. / A. East-Seletsky, M. R. O'Connell, S. C. Knight, D. Burstein, J. H. D. Cate, R. Tjian, J. A. Doudna // Nature – 2016. – T. 538 – № 7624– 270–273c.

44. Ebina H. Harnessing the CRISPR/Cas9 system to disrupt latent HIV-1 provirus. / H. Ebina, N. Misawa, Y. Kanemura, Y. Koyanagi // Sci. Rep. – 2013. – T. 3– 2510c.

45. Edgar R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. / R. C. Edgar // Nucleic Acids Res. – 2004. – T. 32 – № 5– 1792–7c.

46. Edgar R.C. PILER-CR: fast and accurate identification of CRISPR repeats. / R. C. Edgar // BMC Bioinformatics – 2007. – T. 8– 18c.

47. Edgar R.C. Search and clustering orders of magnitude faster than BLAST. / R. C. Edgar // Bioinformatics – 2010. – T. 26 – № 19– 2460–1c.

48. Elmore J.R. Bipartite recognition of target RNAs activates DNA cleavage by the Type III-B CRISPR-Cas system. / J. R. Elmore, N. F. Sheppard, N. Ramia, T. Deighan, H. Li, R. M. Terns, M. P. Terns // Genes Dev. – 2016. – T. 30 – № 4– 447–59c.

49. Estrella M.A. RNA-activated DNA cleavage by the Type III-B CRISPR-Cas effector complex. / M. A. Estrella, F.-T. Kuo, S. Bailey // Genes Dev. – 2016. – T. 30 – № 4– 460–70c.

50. Fonfara I. The CRISPR-associated DNA-cleaving enzyme Cpf1 also processes precursor CRISPR RNA. / I. Fonfara, H. Richter, M. Bratovič, A. Le Rhun, E. Charpentier // Nature – 2016. – T. 532 – № 7600– 517–21c.

51. Garneau J.E. The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. / J. E. Garneau, M.-È. Dupuis, M. Villion, D. A. Romero, R. Barrangou, P. Boyaval, C. Fremaux, P. Horvath, A. H. Magadán, S. Moineau // Nature – 2010. – T. 468 – № 7320– 67–71c.

52. Gasiunas G. Cas9-crPHK ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. / G. Gasiunas, R. Barrangou, P. Horvath, V. Siksnys // Proc. Natl. Acad. Sci. U. S. A. – 2012. – T. 109 – № 39– E2579-86c.

53. Gilbert L.A. Genome-Scale CRISPR-Mediated Control of Gene Repression and Activation. / L. A. Gilbert, M. A. Horlbeck, B. Adamson, J. E. Villalta, Y. Chen, E. H. Whitehead, C. Guimaraes, B. Panning, H. L. Ploegh, M. C. Bassik, L. S. Qi, M. Kampmann, J. S. Weissman // Cell – 2014. – T. 159 – № 3– 647–61c.

54. Gilbert L.A. CRISPR-mediated modular RNA-guided regulation of transcription in eukaryotes. / L. A. Gilbert, M. H. Larson, L. Morsut, Z. Liu, G. A. Brar, S. E. Torres, N. Stern-

Ginossar, O. Brandman, E. H. Whitehead, J. A. Doudna, W. A. Lim, J. S. Weissman, L. S. Qi // Cell – 2013. – T. 154 – № 2– 442–51c.

55. Goldberg G.W. Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. / G. W. Goldberg, W. Jiang, D. Bikard, L. A. Marraffini // Nature – 2014. – T. 514 – № 7524– 633–7c.

56. Gomes-Filho J.V. Sense overlapping transcripts in IS1341-type transposase genes are functional non-coding RNAs in archaea. / J. V. Gomes-Filho, L. S. Zaramela, V. C. da S. Italiani, N. S. Baliga, R. Z. N. Vêncio, T. Koide // RNA Biol. – 2015. – T. 12 – № 5– 490–500c.

57. Gong B. Molecular insights into DNA interference by CRISPR-associated nuclease-helicase Cas3. / B. Gong, M. Shin, J. Sun, C.-H. Jung, E. L. Bolt, J. van der Oost, J.-S. Kim // Proc. Natl. Acad. Sci. U. S. A. – 2014. – T. 111 – № 46– 16359–64c.

58. Gootenberg J.S. Nucleic acid detection with CRISPR-Cas13a/C2c2. / J. S. Gootenberg, O. O. Abudayyeh, J. W. Lee, P. Essletzbichler, A. J. Dy, J. Joung, V. Verdine, N. Donghia, N. M. Daringer, C. A. Freije, C. Myhrvold, R. P. Bhattacharyya, J. Livny, A. Regev, E. V Koonin, D. T. Hung, P. C. Sabeti, J. J. Collins, F. Zhang // Science – 2017.

59. Grissa I. CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. / I. Grissa, G. Vergnaud, C. Pourcel // Nucleic Acids Res. – 2007. – T. 35– № Web Server issue– W52-7c.

60. Grynberg M. HEPN: a common domain in bacterial drug resistance and human neurodegenerative proteins. / M. Grynberg, H. Erlandsen, A. Godzik // Trends Biochem. Sci. – 2003. – T. 28 – № 5– 224–6c.

61. Haft D.H. A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. / D. H. Haft, J. Selengut, E. F. Mongodin, K. E. Nelson // PLoS Comput. Biol. – 2005. – T. 1 – № 6– e60c.

62. Hale C.R. Target RNA capture and cleavage by the Cmr type III-B CRISPR-Cas effector complex. / C. R. Hale, A. Cocozaki, H. Li, R. M. Terns, M. P. Terns // Genes Dev. – 2014. – T. 28 – № 21– 2432–43c.

63. Hale C.R. Essential features and rational design of CRISPR RNAs that function with the Cas RAMP module complex to cleave RNAs. / C. R. Hale, S. Majumdar, J. Elmore, N. Pfister, M. Compton, S. Olson, A. M. Resch, C. V. C. Glover, B. R. Graveley, R. M. Terns, M. P. Terns // Mol. Cell – 2012. – T. 45 – № 3– 292–302c.

64. Hale C.R. RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. / C. R. Hale, P. Zhao, S. Olson, M. O. Duff, B. R. Graveley, L. Wells, R. M. Terns, M. P. Terns // Cell – 2009. – T. 139 – № 5– 945–56c.

65. Haurwitz R.E. Sequence- and structure-specific RNA processing by a CRISPR endonuclease. / R. E. Haurwitz, M. Jinek, B. Wiedenheft, K. Zhou, J. A. Doudna // Science – 2010. – T. 329 – № 5997– 1355–8c.

66. Heler R. Cas9 specifies functional viral targets during CRISPR-Cas adaptation. / R. Heler, P. Samai, J. W. Modell, C. Weiner, G. W. Goldberg, D. Bikard, L. A. Marraffini // Nature – 2015. – T. 519 – № 7542– 199–202c.

67. Hermans P.W. Insertion element IS987 from Mycobacterium bovis BCG is located in a hot-spot integration region for insertion elements in Mycobacterium tuberculosis complex strains. / P. W. Hermans, D. van Soolingen, E. M. Bik, P. E. de Haas, J. W. Dale, J. D. van Embden // Infect. Immun. – 1991. – T. 59 – № 8– 2695–705c.

68. Hickman A.B. The casposon-encoded Cas1 protein from Aciduliprofundum boonei is a DNA integrase that generates target site duplications. / A. B. Hickman, F. Dyda // Nucleic Acids Res. – 2015. – T. 43 – № 22– 10576–87c.

69. Hille F. CRISPR-Cas: biology, mechanisms and relevance / F. Hille, E. Charpentier // Philos. Trans. R. Soc. B Biol. Sci. – 2016. – T. 371 – № 1707– 20150496c.

70. Hilton I.B. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. / I. B. Hilton, A. M. D'Ippolito, C. M. Vockley, P. I. Thakore, G. E. Crawford, T. E. Reddy, C. A. Gersbach // Nat. Biotechnol. – 2015. – T. 33 – № 5– 510–7c.

71. Hoe N. Rapid molecular genetic subtyping of serotype M1 group A Streptococcus strains. /

N. Hoe, K. Nakashima, D. Grigsby, X. Pan, S. J. Dou, S. Naidich, M. Garcia, E. Kahn, D. Bergmire-Sweat, J. M. Musser // Emerg. Infect. Dis. – T. 5 – № 2– 254–63c.

72. Horvath P. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. / P. Horvath, D. A. Romero, A.-C. Coûté-Monvoisin, M. Richards, H. Deveau, S. Moineau, P. Boyaval, C. Fremaux, R. Barrangou // J. Bacteriol. – 2008. – T. 190 – № 4– 1401–12c.

73. Hsu P.D. Development and applications of CRISPR-Cas9 for genome engineering. / P. D. Hsu, E. S. Lander, F. Zhang // Cell – 2014. – T. 157 – № 6– 1262–78c.

74. Hu W. RNA-directed gene editing specifically eradicates latent and prevents new HIV-1 infection. / W. Hu, R. Kaminski, F. Yang, Y. Zhang, L. Cosentino, F. Li, B. Luo, D. Alvarez-Carbonell, Y. Garcia-Mesa, J. Karn, X. Mo, K. Khalili // Proc. Natl. Acad. Sci. U. S. A. – 2014. – T. 111 – № 31– 11461–6c.

75. Huo Y. Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. / Y. Huo, K. H. Nam, F. Ding, H. Lee, L. Wu, Y. Xiao, M. D. Farchione, S. Zhou, K. Rajashankar, I. Kurinov, R. Zhang, A. Ke // Nat. Struct. Mol. Biol. – 2014. – T. 21 – № 9– 771–7c.

76. Hur J.K. Targeted mutagenesis in mice by electroporation of Cpf1 ribonucleoproteins. / J. K. Hur, K. Kim, K. W. Been, G. Baek, S. Ye, J. W. Hur, S.-M. Ryu, Y. S. Lee, J.-S. Kim // Nat. Biotechnol. – 2016. – T. 34 – № 8– 807–8c.

77. Iranzo J. Immunity, suicide or both? Ecological determinants for the combined evolution of anti-pathogen defense systems. / J. Iranzo, A. E. Lobkovsky, Y. I. Wolf, E. V Koonin // BMC Evol. Biol. – 2015. – T. 15– 43c.

78. Ishino Y. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. / Y. Ishino, H. Shinagawa, K. Makino, M. Amemura, A. Nakata // J. Bacteriol. – 1987. – T. 169 – № 12– 5429–33c.

79. Jackson R.N. Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. / 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, B. Wiedenheft // Science – 2014. – T. 345 – № 6203– 1473–9c.

80. Jackson R.N. Fitting CRISPR-associated Cas3 into the helicase family tree. / R. N. Jackson, M. Lavin, J. Carter, B. Wiedenheft // Curr. Opin. Struct. Biol. – 2014. – T. 24– 106–14c.

81. Jansen R. Identification of a novel family of sequence repeats among prokaryotes. / R. Jansen, J. D. A. van Embden, W. Gaastra, L. M. Schouls // OMICS – 2002. – T. 6 – № 1– 23–33c.

82. Jinek M. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. / M. Jinek, K. Chylinski, I. Fonfara, M. Hauer, J. A. Doudna, E. Charpentier // Science – 2012. – T. 337 – № 6096– 816–21c.

83. Jinek M. RNA-programmed genome editing in human cells. / M. Jinek, A. East, A. Cheng, S. Lin, E. Ma, J. Doudna // Elife – 2013. – T. 2– e00471c.

84. Jore M.M. Structural basis for CRISPR RNA-guided DNA recognition by Cascade. / 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, S. J. J. Brouns // Nat. Struct. Mol. Biol. – 2011. – T. 18 – № 5– 529–36c.

85. Kapitonov V. V ISC, a Novel Group of Bacterial and Archaeal DNA Transposons That Encode Cas9 Homologs. / V. V Kapitonov, K. S. Makarova, E. V Koonin // J. Bacteriol. – 2015. – T. 198 – № 5– 797–807c.

86. Katoh K. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. / K. Katoh, D. M. Standley // Mol. Biol. Evol. – 2013. – T. 30 – № 4– 772–80c.

87. Kearns N.A. Functional annotation of native enhancers with a Cas9-histone demethylase fusion. / N. A. Kearns, H. Pham, B. Tabak, R. M. Genga, N. J. Silverstein, M. Garber, R. Maehr // Nat. Methods – 2015. – T. 12 – № 5– 401–3c.

88. Kiani S. CRISPR transcriptional repression devices and layered circuits in mammalian cells.

/ S. Kiani, J. Beal, M. R. Ebrahimkhani, J. Huh, R. N. Hall, Z. Xie, Y. Li, R. Weiss // Nat. Methods – 2014. – T. 11 – № 7– 723–6c.

89. Kim D. Genome-wide analysis reveals specificities of Cpf1 endonucleases in human cells. / D. Kim, J. Kim, J. K. Hur, K. W. Been, S.-H. Yoon, J.-S. Kim // Nat. Biotechnol. – 2016. – T. 34 – № 8– 863–8c.

90. Kim Y. Generation of knockout mice by Cpf1-mediated gene targeting. / Y. Kim, S.-A. Cheong, J. G. Lee, S.-W. Lee, M. S. Lee, I.-J. Baek, Y. H. Sung // Nat. Biotechnol. – 2016. – T. 34 – № 8– 808–10c.

91. Kleinstiver B.P. High-fidelity CRISPR-Cas9 nucleases with no detectable genome-wide off-target effects. / B. P. Kleinstiver, V. Pattanayak, M. S. Prew, S. Q. Tsai, N. T. Nguyen, Z. Zheng, J. K. Joung // Nature – 2016. – T. 529 – № 7587– 490–5c.

92. Kleinstiver B.P. Genome-wide specificities of CRISPR-Cas Cpf1 nucleases in human cells. / B. P. Kleinstiver, S. Q. Tsai, M. S. Prew, N. T. Nguyen, M. M. Welch, J. M. Lopez, Z. R. McCaw, M. J. Aryee, J. K. Joung // Nat. Biotechnol. – 2016. – T. 34 – № 8– 869–74c.

93. Knight S.C. Dynamics of CRISPR-Cas9 genome interrogation in living cells. / S. C. Knight, L. Xie, W. Deng, B. Guglielmi, L. B. Witkowsky, L. Bosanac, E. T. Zhang, M. El Beheiry, J.-B. Masson, M. Dahan, Z. Liu, J. A. Doudna, R. Tjian // Science – 2015. – T. 350 – № 6262– 823–6c.

94. Konermann S. Optical control of mammalian endogenous transcription and epigenetic states. / S. Konermann, M. D. Brigham, A. E. Trevino, P. D. Hsu, M. Heidenreich, L. Cong, R. J. Platt, D. A. Scott, G. M. Church, F. Zhang // Nature – 2013. – T. 500 – № 7463– 472–6c.

95. Konermann S. Genome-scale transcriptional activation by an engineered CRISPR-Cas9 complex. / S. Konermann, M. D. Brigham, A. E. Trevino, J. Joung, O. O. Abudayyeh, C. Barcena, P. D. Hsu, N. Habib, J. S. Gootenberg, H. Nishimasu, O. Nureki, F. Zhang // Nature – 2015. – T. 517 – № 7536– 583–8c.

96. Koonin E. V Origins and evolution of viruses of eukaryotes: The ultimate modularity. / E. V Koonin, V. V Dolja, M. Krupovic // Virology – 2015. – T. 479–480– 2–25c.

97. Koonin E. V Evolution of adaptive immunity from transposable elements combined with innate immune systems. / E. V Koonin, M. Krupovic // Nat. Rev. Genet. – 2015. – T. 16 – № 3– 184–92c.

98. Koonin E. V Just how Lamarckian is CRISPR-Cas immunity: the continuum of evolvability mechanisms. / E. V Koonin, Y. I. Wolf // Biol. Direct – 2016. – T. 11 – № 1– 9c.

99. Krupovic M. Casposons: a new superfamily of self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. / M. Krupovic, K. S. Makarova, P. Forterre, D. Prangishvili, E. V Koonin // BMC Biol. – 2014. – T. 12– 36c.

100. Krupovic M. Recent Mobility of Casposons, Self-Synthesizing Transposons at the Origin of the CRISPR-Cas Immunity. / M. Krupovic, S. Shmakov, K. S. Makarova, P. Forterre, E. V Koonin // Genome Biol. Evol. – 2016. – T. 8 – № 2– 375–86c.

101. Levy A. CRISPR adaptation biases explain preference for acquisition of foreign DNA. / A. Levy, M. G. Goren, I. Yosef, O. Auster, M. Manor, G. Amitai, R. Edgar, U. Qimron, R. Sorek // Nature – 2015. – T. 520 – № 7548– 505–10c.

102. Li M. Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. / M. Li, R. Wang, D. Zhao, H. Xiang // Nucleic Acids Res. – 2014. – T. 42 – № 4– 2483–92c.

103. Li S.-Y. C-Brick: A New Standard for Assembly of Biological Parts Using Cpf1. / S.-Y. Li, G.-P. Zhao, J. Wang // ACS Synth. Biol. – 2016. – T. 5 – № 12– 1383–1388c.

104. Liu L. C2c1-sgPHK Complex Structure Reveals RNA-Guided DNA Cleavage Mechanism. / L. Liu, P. Chen, M. Wang, X. Li, J. Wang, M. Yin, Y. Wang // Mol. Cell – 2017. – T. 65 – № 2– 310–322c.

105. Liu Y. Synthesizing AND gate genetic circuits based on CRISPR-Cas9 for identification of bladder cancer cells. / Y. Liu, Y. Zeng, L. Liu, C. Zhuang, X. Fu, W. Huang, Z. Cai // Nat. Commun. – 2014. – T. 5– 5393c.

106. Majumdar S. Three CRISPR-Cas immune effector complexes coexist in Pyrococcus furiosus. / S. Majumdar, P. Zhao, N. T. Pfister, M. Compton, S. Olson, C. V. C. Glover, L.

Wells, B. R. Graveley, R. M. Terns, M. P. Terns // RNA – 2015. – T. 21 – № 6– 1147–58c.

107. Makarova K.S. Live virus-free or die: coupling of antivirus immunity and programmed suicide or dormancy in prokaryotes. / K. S. Makarova, V. Anantharaman, L. Aravind, E. V Koonin // Biol. Direct – 2012. – T. 7– 40c.

108. Makarova K.S. CARF and WYL domains: ligand-binding regulators of prokaryotic defense systems. / K. S. Makarova, V. Anantharaman, N. V Grishin, E. V Koonin, L. Aravind // Front. Genet. – 2014. – T. 5– 102c.

109. Makarova K.S. A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis. / K. S. Makarova, L. Aravind, N. V Grishin, I. B. Rogozin, E. V Koonin // Nucleic Acids Res. – 2002. – T. 30 – № 2– 482–96c.

110. Makarova K.S. Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. / K. S. Makarova, L. Aravind, Y. I. Wolf, E. V Koonin // Biol. Direct – 2011. – T. 6– 38c.

111. Makarova K.S. 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. / K. S. Makarova, N. V Grishin, S. A. Shabalina, Y. I. Wolf, E. V Koonin // Biol. Direct – 2006. – T. 1– 7c.

112. Makarova K.S. Evolution and classification of the CRISPR-Cas systems. / 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, E. V Koonin // Nat. Rev. Microbiol. – 2011. – T. 9 – № 6– 467–77c.

113. Makarova K.S. Annotation and Classification of CRISPR-Cas Systems. / K. S. Makarova, E. V Koonin // Methods Mol. Biol. – 2015. – T. 1311– 47–75c.

114. Makarova K.S. An updated evolutionary classification of CRISPR-Cas systems. / K. S. Makarova, Y. I. Wolf, O. S. Alkhnbashi, F. Costa, S. A. Shah, S. J. Saunders, R. Barrangou, S. J. J. Brouns, E. Charpentier, D. H. Haft, P. Horvath, S. Moineau, F. J. M. Mojica, R. M. Terns, M. P. Terns, M. F. White, A. F. Yakunin, R. A. Garrett, J. van der Oost, R. Backofen, E. V

Koonin // Nat. Rev. Microbiol. – 2015. – T. 13 – № 11– 722–36c.

115. Makarova K.S. Comparative genomics of defense systems in archaea and bacteria. / K. S. Makarova, Y. I. Wolf, E. V Koonin // Nucleic Acids Res. – 2013. – T. 41 – № 8– 4360–77c.

116. Makarova K.S. The basic building blocks and evolution of CRISPR-CAS systems. / K. S. Makarova, Y. I. Wolf, E. V Koonin // Biochem. Soc. Trans. – 2013. – T. 41 – № 6– 1392–400c.

117. Makarova K.S. Defense islands in bacterial and archaeal genomes and prediction of novel defense systems. / K. S. Makarova, Y. I. Wolf, S. Snir, E. V Koonin // J. Bacteriol. – 2011. – T. 193 – № 21– 6039–56c.

118. Mali P. Cas9 as a versatile tool for engineering biology. / P. Mali, K. M. Esvelt, G. M. Church // Nat. Methods – 2013. – T. 10 – № 10– 957–63c.

119. Mali P. RNA-guided human genome engineering via Cas9. / P. Mali, L. Yang, K. M. Esvelt, J. Aach, M. Guell, J. E. DiCarlo, J. E. Norville, G. M. Church // Science – 2013. – T. 339 – № 6121– 823–6c.

120. Marchler-Bauer A. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. / A. Marchler-Bauer, Y. Bo, L. Han, J. He, C. J. Lanczycki, S. Lu, F. Chitsaz, M. K. Derbyshire, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, F. Lu, G. H. Marchler, J. S. Song, N. Thanki, Z. Wang, R. A. Yamashita, D. Zhang, C. Zheng, L. Y. Geer, S. H. Bryant // Nucleic Acids Res. – 2017. – T. 45 – № D1– D200–D203c.

121. Marchler-Bauer A. CDD: a Conserved Domain Database for the functional annotation of proteins. / A. Marchler-Bauer, S. Lu, J. B. Anderson, F. Chitsaz, M. K. Derbyshire, C. DeWeese-Scott, J. H. Fong, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, J. D. Jackson, Z. Ke, C. J. Lanczycki, F. Lu, G. H. Marchler, M. Mullokandov, M. V Omelchenko, C. L. Robertson, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, N. Zhang, C. Zheng, S. H. Bryant // Nucleic Acids Res. – 2011. – T. 39– № Database issue– D225-9c.

122. Marchler-Bauer A. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. / A. Marchler-Bauer, A. R. Panchenko, B. A. Shoemaker, P. A. Thiessen, L. Y. Geer, S. H. Bryant // Nucleic Acids Res. – 2002. – T. 30 – № 1– 281–3c.

123. Marchler-Bauer A. CDD: conserved domains and protein three-dimensional structure. / A. Marchler-Bauer, C. Zheng, F. Chitsaz, M. K. Derbyshire, L. Y. Geer, R. C. Geer, N. R. Gonzales, M. Gwadz, D. I. Hurwitz, C. J. Lanczycki, F. Lu, S. Lu, G. H. Marchler, J. S. Song, N. Thanki, R. A. Yamashita, D. Zhang, S. H. Bryant // Nucleic Acids Res. – 2013. – T. 41– № Database issue– D348-52c.

124. Marraffini L.A. CRISPR-Cas immunity in prokaryotes. / L. A. Marraffini // Nature – 2015. – T. 526 – № 7571– 55–61c.

125. Marraffini L.A. CRISPR interference limits horizontal gene transfer in staphylococci by targeting DNA. / L. A. Marraffini, E. J. Sontheimer // Science – 2008. – T. 322 – № 5909– 1843–5c.

126. Marraffini L.A. Self versus non-self discrimination during CRISPR RNA-directed immunity. / L. A. Marraffini, E. J. Sontheimer // Nature – 2010. – T. 463 – № 7280– 568–71c.

127. Marraffini L.A. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. / L. A. Marraffini, E. J. Sontheimer // Nat. Rev. Genet. – 2010. – T. 11 – № 3– 181–90c.

128. Mohanraju P. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. / P. Mohanraju, K. S. Makarova, B. Zetsche, F. Zhang, E. V Koonin, J. van der Oost // Science – 2016. – T. 353 – № 6299– aad5147c.

129. Mojica F.J. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. / F. J. Mojica, C. Díez-Villaseñor, E. Soria, G. Juez // Mol. Microbiol. – 2000. – T. 36 – № 1– 244–6c.

130. Mojica F.J. 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. / F. J. Mojica, C. Ferrer, G. Juez, F. Rodríguez-Valera // Mol. Microbiol. – 1995. – T. 17 – № 1– 85–93c.

131. Mojica F.J.M. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. / F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, C. Almendros //

Microbiology – 2009. – T. 155 – № 3– 733–740c.

132. Mojica F.J.M. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. / F. J. M. Mojica, C. Díez-Villaseñor, J. García-Martínez, E. Soria // J. Mol. Evol. – 2005. – T. 60 – № 2– 174–82c.

133. Mulepati S. Structural and biochemical analysis of nuclease domain of clustered regularly interspaced short palindromic repeat (CRISPR)-associated protein 3 (Cas3). / S. Mulepati, S. Bailey // J. Biol. Chem. – 2011. – T. 286 – № 36– 31896–903c.

134. Nakata A. Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 chromosome. / A. Nakata, M. Amemura, K. Makino // J. Bacteriol. – 1989. – T. 171 – № 6– 3553–6c.

135. Nam K.H. Cas5d protein processes pre-crPHK and assembles into a cascade-like interference complex in subtype I-C/Dvulg CRISPR-Cas system. / K. H. Nam, C. Haitjema, X. Liu, F. Ding, H. Wang, M. P. DeLisa, A. Ke // Structure – 2012. – T. 20 – № 9– 1574–84c.

136. Nelles D.A. Programmable RNA Tracking in Live Cells with CRISPR/Cas9. / D. A. Nelles, M. Y. Fang, M. R. O'Connell, J. L. Xu, S. J. Markmiller, J. A. Doudna, G. W. Yeo // Cell – 2016. – T. 165 – № 2– 488–96c.

137. Niewoehner O. Structural basis for the endoribonuclease activity of the type III-A CRISPR-associated protein Csm6. / O. Niewoehner, M. Jinek // RNA – 2016. – T. 22 – № 3– 318–29c.

138. Nishimasu H. Crystal Structure of Staphylococcus aureus Cas9. / H. Nishimasu, L. Cong, W. X. Yan, F. A. Ran, B. Zetsche, Y. Li, A. Kurabayashi, R. Ishitani, F. Zhang, O. Nureki // Cell – 2015. – T. 162 – № 5– 1113–26c.

139. Nishimasu H. Crystal structure of Cas9 in complex with guide RNA and target DNA. / H. Nishimasu, F. A. Ran, P. D. Hsu, S. Konermann, S. I. Shehata, N. Dohmae, R. Ishitani, F. Zhang, O. Nureki // Cell – 2014. – T. 156 – № 5– 935–49c.

140. Nissim L. Multiplexed and programmable regulation of gene networks with an integrated RNA and CRISPR/Cas toolkit in human cells. / L. Nissim, S. D. Perli, A. Fridkin, P. Perez-

Pinera, T. K. Lu // Mol. Cell – 2014. – T. 54 – № 4– 698–710c.

141. Nuñez J.K. Foreign DNA capture during CRISPR-Cas adaptive immunity. / J. K. Nuñez, L. B. Harrington, P. J. Kranzusch, A. N. Engelman, J. A. Doudna // Nature – 2015. – T. 527 – № 7579– 535–8c.

142. Nuñez J.K. Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. / J. K. Nuñez, P. J. Kranzusch, J. Noeske, A. V Wright, C. W. Davies, J. A. Doudna // Nat. Struct. Mol. Biol. – 2014. – T. 21 – № 6– 528–34c.

143. Nuñez J.K. Integrase-mediated spacer acquisition during CRISPR-Cas adaptive immunity. / J. K. Nuñez, A. S. Y. Lee, A. Engelman, J. A. Doudna // Nature – 2015. – T. 519 – № 7542– 193–8c.

144. O'Connell M.R. Programmable RNA recognition and cleavage by CRISPR/Cas9. / M. R. O'Connell, B. L. Oakes, S. H. Sternberg, A. East-Seletsky, M. Kaplan, J. A. Doudna // Nature – 2014. – T. 516 – № 7530– 263–6c.

145. Oost J. van der Unravelling the structural and mechanistic basis of CRISPR-Cas systems. / J. van der Oost, E. R. Westra, R. N. Jackson, B. Wiedenheft // Nat. Rev. Microbiol. – 2014. – T. 12 – № 7– 479–92c.

146. Osawa T. Crystal structure of the CRISPR-Cas RNA silencing Cmr complex bound to a target analog. / T. Osawa, H. Inanaga, C. Sato, T. Numata // Mol. Cell – 2015. – T. 58 – № 3– 418–30c.

147. Pasternak C. ISDra2 transposition in Deinococcus radiodurans is downregulated by TnpB. / C. Pasternak, R. Dulermo, B. Ton-Hoang, R. Debuchy, P. Siguier, G. Coste, M. Chandler, S. Sommer // Mol. Microbiol. – 2013. – T. 88 – № 2– 443–55c.

148. Peng W. An archaeal CRISPR type III-B system exhibiting distinctive RNA targeting features and mediating dual RNA and DNA interference. / W. Peng, M. Feng, X. Feng, Y. X. Liang, Q. She // Nucleic Acids Res. – 2015. – T. 43 – № 1– 406–17c.

149. Platt R.J. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. / R. J. Platt, S. Chen, Y. Zhou, M. J. Yim, L. Swiech, H. R. Kempton, J. E. Dahlman, O. Parnas, T. M.

Eisenhaure, M. Jovanovic, D. B. Graham, S. Jhunjhunwala, M. Heidenreich, R. J. Xavier, R. Langer, D. G. Anderson, N. Hacohen, A. Regev, G. Feng, P. A. Sharp, F. Zhang // Cell – 2014. – T. 159 – № 2– 440–55c.

150. Pourcel C. CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. / C. Pourcel, G. Salvignol, G. Vergnaud // Microbiology – 2005. – T. 151– № Pt 3– 653–63c.

151. Price M.N. FastTree 2--approximately maximum-likelihood trees for large alignments. / M. N. Price, P. S. Dehal, A. P. Arkin // PLoS One – 2010. – T. 5 – № 3– e9490c.

152. Qi L. RNA processing enables predictable programming of gene expression. / L. Qi, R. E. Haurwitz, W. Shao, J. A. Doudna, A. P. Arkin // Nat. Biotechnol. – 2012. – T. 30 – № 10– 1002–6c.

153. Qi L.S. Repurposing CRISPR as an RNA-guided platform for sequence-specific control of gene expression. / L. S. Qi, M. H. Larson, L. A. Gilbert, J. A. Doudna, J. S. Weissman, A. P. Arkin, W. A. Lim // Cell – 2013. – T. 152 – № 5– 1173–83c.

154. Qin W. Efficient CRISPR/Cas9-Mediated Genome Editing in Mice by Zygote Electroporation of Nuclease. / W. Qin, S. L. Dion, P. M. Kutny, Y. Zhang, A. W. Cheng, N. L. Jillette, A. Malhotra, A. M. Geurts, Y.-G. Chen, H. Wang // Genetics – 2015. – T. 200 – № 2– 423–30c.

155. Ramanan V. CRISPR/Cas9 cleavage of viral DNA efficiently suppresses hepatitis B virus. / V. Ramanan, A. Shlomai, D. B. T. Cox, R. E. Schwartz, E. Michailidis, A. Bhatta, D. A. Scott, F. Zhang, C. M. Rice, S. N. Bhatia // Sci. Rep. – 2015. – T. 5– 10833c.

156. Ran F.A. In vivo genome editing using Staphylococcus aureus Cas9. / F. A. Ran, L. Cong, W. X. Yan, D. A. Scott, J. S. Gootenberg, A. J. Kriz, B. Zetsche, O. Shalem, X. Wu, K. S. Makarova, E. V Koonin, P. A. Sharp, F. Zhang // Nature – 2015. – T. 520 – № 7546– 186–91c.

157. Ran F.A. Double nicking by RNA-guided CRISPR Cas9 for enhanced genome editing specificity. / F. A. Ran, P. D. Hsu, C.-Y. Lin, J. S. Gootenberg, S. Konermann, A. E. Trevino, D. A. Scott, A. Inoue, S. Matoba, Y. Zhang, F. Zhang // Cell – 2013. – T. 154 – № 6– 1380–9c.

158. Ran F.A. Genome engineering using the CRISPR-Cas9 system. / F. A. Ran, P. D. Hsu, J. Wright, V. Agarwala, D. A. Scott, F. Zhang // Nat. Protoc. – 2013. – T. 8 – № 11– 2281–308c.

159. Reardon S. First CRISPR clinical trial gets green light from US panel / S. Reardon // Nature – 2016.

160. Redding S. Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. / S. Redding, S. H. Sternberg, M. Marshall, B. Gibb, P. Bhat, C. K. Guegler, B. Wiedenheft, J. A. Doudna, E. C. Greene // Cell – 2015. – T. 163 – № 4– 854–65c.

161. Richter C. Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. / C. Richter, R. L. Dy, R. E. McKenzie, B. N. J. Watson, C. Taylor, J. T. Chang, M. B. McNeil, R. H. J. Staals, P. C. Fineran // Nucleic Acids Res. – 2014. – T. 42 – № 13– 8516–26c.

162. Rollins M.F. Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. / M. F. Rollins, J. T. Schuman, K. Paulus, H. S. T. Bukhari, B. Wiedenheft // Nucleic Acids Res. – 2015. – T. 43 – № 4– 2216–22c.

163. Rouillon C. Structure of the CRISPR interference complex CSM reveals key similarities with cascade. / C. Rouillon, M. Zhou, J. Zhang, A. Politis, V. Beilsten-Edmands, G. Cannone, S. Graham, C. V Robinson, L. Spagnolo, M. F. White // Mol. Cell – 2013. – T. 52 – № 1– 124–34c.

164. Rutkauskas M. Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. / M. Rutkauskas, T. Sinkunas, I. Songailiene, M. S. Tikhomirova, V. Siksnys, R. Seidel // Cell Rep. – 2015.

165. Samai P. Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. / P. Samai, N. Pyenson, W. Jiang, G. W. Goldberg, A. Hatoum-Aslan, L. A. Marraffini // Cell – 2015. – T. 161 – № 5– 1164–74c.

166. Samson J.E. Revenge of the phages: defeating bacterial defences. / J. E. Samson, A. H. Magadán, M. Sabri, S. Moineau // Nat. Rev. Microbiol. – 2013. – T. 11 – № 10– 675–87c.

167. Sapranauskas R. The Streptococcus thermophilus CRISPR/Cas system provides immunity

in Escherichia coli. / R. Sapranauskas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, V. Siksnys // Nucleic Acids Res. – 2011. – Т. 39 – № 21– 9275–82с.

168. Savitskaya E. High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. / E. Savitskaya, E. Semenova, V. Dedkov, A. Metlitskaya, K. Severinov // RNA Biol. – 2013. – Т. 10 – № 5– 716–25с.

169. Schunder E. First indication for a functional CRISPR/Cas system in Francisella tularensis. / E. Schunder, K. Rydzewski, R. Grunow, K. Heuner // Int. J. Med. Microbiol. – 2013. – Т. 303 – № 2– 51–60с.

170. Semenova E. Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. / E. Semenova, M. M. Jore, K. A. Datsenko, A. Semenova, E. R. Westra, B. Wanner, J. van der Oost, S. J. J. Brouns, K. Severinov // Proc. Natl. Acad. Sci. U. S. A. – 2011. – Т. 108 – № 25– 10098–103с.

171. Shalem O. Genome-scale CRISPR-Cas9 knockout screening in human cells. / O. Shalem, N. E. Sanjana, E. Hartenian, X. Shi, D. A. Scott, T. S. Mikkelsen, D. Heckl, B. L. Ebert, D. E. Root, J. G. Doench, F. Zhang // Science – 2014. – Т. 343 – № 6166– 84–7с.

172. Sheppard N.F. The CRISPR-associated Csx1 protein of Pyrococcus furiosus is an adenosine-specific endoribonuclease. / N. F. Sheppard, C. V. C. Glover, R. M. Terns, M. P. Terns // RNA – 2016. – Т. 22 – № 2– 216–24с.

173. Shmakov S. Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems. / S. Shmakov, O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang, E. V Koonin // Mol. Cell – 2015. – Т. 60 – № 3– 385–97с.

174. Shmakov S. Pervasive generation of oppositely oriented spacers during CRISPR adaptation. / S. Shmakov, E. Savitskaya, E. Semenova, M. D. Logacheva, K. A. Datsenko, K. Severinov // Nucleic Acids Res. – 2014. – Т. 42 – № 9– 5907–16с.

175. Shmakov S. Diversity and evolution of class 2 CRISPR-Cas systems. / S. Shmakov, A. Smargon, D. Scott, D. Cox, N. Pyzocha, W. Yan, O. O. Abudayyeh, J. S. Gootenberg, K. S.

Makarova, Y. I. Wolf, K. Severinov, F. Zhang, E. V Koonin // Nat. Rev. Microbiol. – 2017. – T. 15 – № 3– 169–182c.

176. Sinkunas T. Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. / T. Sinkunas, G. Gasiunas, C. Fremaux, R. Barrangou, P. Horvath, V. Siksnys // EMBO J. – 2011. – T. 30 – № 7– 1335–42c.

177. Slaymaker I.M. Rationally engineered Cas9 nucleases with improved specificity. / I. M. Slaymaker, L. Gao, B. Zetsche, D. A. Scott, W. X. Yan, F. Zhang // Science – 2016. – T. 351 – № 6268– 84–8c.

178. Smargon A.A. Cas13b Is a Type VI-B CRISPR-Associated RNA-Guided RNase Differentially Regulated by Accessory Proteins Csx27 and Csx28. / A. A. Smargon, D. B. T. Cox, N. K. Pyzocha, K. Zheng, I. M. Slaymaker, J. S. Gootenberg, O. A. Abudayyeh, P. Essletzbichler, S. Shmakov, K. S. Makarova, E. V Koonin, F. Zhang // Mol. Cell – 2017. – T. 65 – № 4– 618–630.e7c.

179. Söding J. Protein homology detection by HMM-HMM comparison. / J. Söding // Bioinformatics – 2005. – T. 21 – № 7– 951–60c.

180. Söding J. HHsenser: exhaustive transitive profile search using HMM-HMM comparison. / J. Söding, M. Remmert, A. Biegert, A. N. Lupas // Nucleic Acids Res. – 2006. – T. 34– № Web Server issue– W374-8c.

181. Spilman M. Structure of an RNA silencing complex of the CRISPR-Cas immune system. / M. Spilman, A. Cocozaki, C. Hale, Y. Shao, N. Ramia, R. Terns, M. Terns, H. Li, S. Stagg // Mol. Cell – 2013. – T. 52 – № 1– 146–52c.

182. Staals R.H.J. Structure and activity of the RNA-targeting Type III-B CRISPR-Cas complex of Thermus thermophilus. / 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, A. Shinkai // Mol. Cell – 2013. – T. 52 – № 1– 135–45c.

183. Staals R.H.J. RNA targeting by the type III-A CRISPR-Cas Csm complex of Thermus

thermophilus. / R. H. J. Staals, Y. Zhu, D. W. Taylor, J. E. Kornfeld, K. Sharma, A. Barendregt, J. J. Koehorst, M. Vlot, N. Neupane, K. Varossieau, K. Sakamoto, T. Suzuki, N. Dohmae, S. Yokoyama, P. J. Schaap, H. Urlaub, A. J. R. Heck, E. Nogales, J. A. Doudna, A. Shinkai, J. van der Oost // Mol. Cell – 2014. – T. 56 – № 4– 518–30c.

184. Sternberg S.H. Conformational control of DNA target cleavage by CRISPR-Cas9. / S. H. Sternberg, B. LaFrance, M. Kaplan, J. A. Doudna // Nature – 2015. – T. 527 – № 7576– 110–3c.

185. Sternberg S.H. DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. / S. H. Sternberg, S. Redding, M. Jinek, E. C. Greene, J. A. Doudna // Nature – 2014. – T. 507 – № 7490– 62–7c.

186. Swarts D.C. CRISPR interference directs strand specific spacer acquisition. / D. C. Swarts, C. Mosterd, M. W. J. van Passel, S. J. J. Brouns // PLoS One – 2012. – T. 7 – № 4– e35888c.

187. Takeuchi N. Nature and intensity of selection pressure on CRISPR-associated genes. / N. Takeuchi, Y. I. Wolf, K. S. Makarova, E. V Koonin // J. Bacteriol. – 2012. – T. 194 – № 5– 1216–25c.

188. Tamulaitis G. Programmable RNA shredding by the type III-A CRISPR-Cas system of Streptococcus thermophilus. / G. Tamulaitis, M. Kazlauskiene, E. Manakova, Č. Venclovas, A. O. Nwokeoji, M. J. Dickman, P. Horvath, V. Siksnys // Mol. Cell – 2014. – T. 56 – № 4– 506–17c.

189. Taylor D.W. Structural biology. Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. / D. W. Taylor, Y. Zhu, R. H. J. Staals, J. E. Kornfeld, A. Shinkai, J. van der Oost, E. Nogales, J. A. Doudna // Science – 2015. – T. 348 – № 6234– 581–5c.

190. Thakore P.I. Editing the epigenome: technologies for programmable transcription and epigenetic modulation. / P. I. Thakore, J. B. Black, I. B. Hilton, C. A. Gersbach // Nat. Methods – 2016. – T. 13 – № 2– 127–37c.

191. Vestergaard G. CRISPR adaptive immune systems of Archaea. / G. Vestergaard, R. A. Garrett, S. A. Shah // RNA Biol. – 2014. – T. 11 – № 2– 156–67c.

192. Wang J. Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. / J. Wang, J. Li, H. Zhao, G. Sheng, M. Wang, M. Yin, Y. Wang // Cell – 2015. – T. 163 – № 4– 840–53c.

193. Wang T. Genetic screens in human cells using the CRISPR-Cas9 system. / T. Wang, J. J. Wei, D. M. Sabatini, E. S. Lander // Science – 2014. – T. 343 – № 6166– 80–4c.

194. Wei Y. Cas9 function and host genome sampling in Type II-A CRISPR-Cas adaptation. / Y. Wei, R. M. Terns, M. P. Terns // Genes Dev. – 2015. – T. 29 – № 4– 356–61c.

195. Westra E.R. CRISPR-Cas systems: beyond adaptive immunity. / E. R. Westra, A. Buckling, P. C. Fineran // Nat. Rev. Microbiol. – 2014. – T. 12 – № 5– 317–26c.

196. Wheeler D. BLAST QuickStart: example-driven web-based BLAST tutorial. (BLASTCLUST) / D. Wheeler, M. Bhagwat // Methods Mol. Biol. – 2007. – T. 395– 149–76c.

197. Wiedenheft B. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. / B. Wiedenheft, E. van Duijn, J. B. Bultema, J. Bultema, S. P. Waghmare, S. Waghmare, K. Zhou, A. Barendregt, W. Westphal, A. J. R. Heck, A. Heck, E. J. Boekema, E. Boekema, M. J. Dickman, M. Dickman, J. A. Doudna // Proc. Natl. Acad. Sci. U. S. A. – 2011. – T. 108 – № 25– 10092–7c.

198. Wiedenheft B. Structures of the RNA-guided surveillance complex from a bacterial immune system. / B. Wiedenheft, G. C. Lander, K. Zhou, M. M. Jore, S. J. J. Brouns, J. van der Oost, J. A. Doudna, E. Nogales // Nature – 2011. – T. 477 – № 7365– 486–9c.

199. Yamano T. Crystal Structure of Cpf1 in Complex with Guide RNA and Target DNA. / T. Yamano, H. Nishimasu, B. Zetsche, H. Hirano, I. M. Slaymaker, Y. Li, I. Fedorova, T. Nakane, K. S. Makarova, E. V Koonin, R. Ishitani, F. Zhang, O. Nureki // Cell – 2016. – T. 165 – № 4– 949–62c.

200. Yang H. PAM-Dependent Target DNA Recognition and Cleavage by C2c1 CRISPR-Cas Endonuclease. / H. Yang, P. Gao, K. R. Rajashankar, D. J. Patel // Cell – 2016. – T. 167 – № 7– 1814–1828.e12c.

201. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. / Z. Yang // Mol. Biol.

Evol. – 2007. – T. 24 – № 8– 1586–91c.

202. Yosef I. Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. / I. Yosef, M. G. Goren, U. Qimron // Nucleic Acids Res. – 2012. – T. 40 – № 12– 5569–76c.

203. Yutin N. The deep archaeal roots of eukaryotes. / N. Yutin, K. S. Makarova, S. L. Mekhedov, Y. I. Wolf, E. V Koonin // Mol. Biol. Evol. – 2008. – T. 25 – № 8– 1619–30c.

204. Zebec Z. CRISPR-mediated targeted mRNA degradation in the archaeon Sulfolobus solfataricus. / Z. Zebec, A. Manica, J. Zhang, M. F. White, C. Schleper // Nucleic Acids Res. – 2014. – T. 42 – № 8– 5280–8c.

205. Zetsche B. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. / B. Zetsche, J. S. Gootenberg, O. O. Abudayyeh, I. M. Slaymaker, K. S. Makarova, P. Essletzbichler, S. E. Volz, J. Joung, J. van der Oost, A. Regev, E. V Koonin, F. Zhang // Cell – 2015. – T. 163 – № 3– 759–71c.

206. Zetsche B. Erratum: Multiplex gene editing by CRISPR-Cpf1 using a single crPHK array. / B. Zetsche, M. Heidenreich, P. Mohanraju, I. Fedorova, J. Kneppers, E. M. DeGennaro, N. Winblad, S. R. Choudhury, O. O. Abudayyeh, J. S. Gootenberg, W. Y. Wu, D. A. Scott, K. Severinov, J. van der Oost, F. Zhang // Nat. Biotechnol. – 2017. – T. 35 – № 2– 178c.

207. Zhang Y. Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. / Y. Zhang, N. Heidrich, B. J. Ampattu, C. W. Gunderson, H. S. Seifert, C. Schoen, J. Vogel, E. J. Sontheimer // Mol. Cell – 2013. – T. 50 – № 4– 488–503c.

208. Zhang Y. DNase H Activity of Neisseria meningitidis Cas9. / Y. Zhang, R. Rajan, H. S. Seifert, A. Mondragón, E. J. Sontheimer // Mol. Cell – 2015. – T. 60 – № 2– 242–55c.

209. Zhang Z. A greedy algorithm for aligning DNA sequences. / Z. Zhang, S. Schwartz, L. Wagner, W. Miller // J. Comput. Biol. – T. 7 – № 1–2– 203–14c.

210. Zhu W. Ab initio gene identification in metagenomic sequences. / W. Zhu, A. Lomsadze, M. Borodovsky // Nucleic Acids Res. – 2010. – T. 38 – № 12– e132c.

211. Database resources of the National Center for Biotechnology Information / // Nucleic Acids Res. – 2016. – T. 44 – № D1– D7–D19c.

Приложение

Здесь представлены ссылки на приложения, в связи с их размером или неподходящим форматом (для текстового документа). Все файлы находятся на FTP сайте NCBI.

Приложение S2

Файлы находятся на FTP сайте по следующей ссылке: ftp://ftp.ncbi.nlm.nih.gov/pub/wolf/_suppl/CRISPRclass2NRM/

Описание файлов, расположенных на FTP сайте:

Supplementary information S2 (box, part a) (MS Excel):

Выходные данные из описанного стека программ: все белковые семейства, ассоциированные с CRISPR повторами. Кластера белков для всех рамок считывания в районе ±10 килобаз от затравки, их аннотация и последовательности представителей кластеров. Все кластеры отсортированы по относительной частоте встречаемости в CRISPR локусах.

Supplementary information S2 (box, part b) (MS Excel):

Локусы и их организация для всех вариантов систем V-U

Supplementary information S2 (box, part c):

Локусы систем 2 класса. Для каждого эффекторного гена показано окружение. Белок-кодирующие гены и CRISPR кассеты отображены. Гены, аннотированные в GenBank имеют GenBank locus tags; гены аннотированные de novo имеют идентификатор, состоящий из ID контига и номера гена.

Supplementary information S2 (box, part d):

Результат FastTree для TnpB семейств в newick формате. Полное дерево, отображенное на Рисунке 10a. Последовательности отмечены локальными GI

номерами, даны имена видов и виды, содержащие CRISPR кассеты помечены с “CRISPR” префиксом. Больше информации о последовательностях может быть найдено в supplementary information S2 (box, part g).

Supplementary information S2 (box, part e) (MS Excel):

Спэйсеры CRISPR кассет. Уникальные спэйсеры были получены из всех CRISPR кассет, находящихся в supplementary information S2 (box, part a). Поиск протоспэйсеров был выполнен используя MEGABLAST (см. Материалы и методы).

Supplementary information S2 (box, part f) (MS Excel):

CRISPR-Cas системы и CRISPR кассеты, найденные в геномах содержащих Type V-U системы. Для каждого полного генома, который содержит хотябы один V-U представитель, все CRISPR-Cas локусы, CRISPR кассеты и последовательности повторов приведены. Локусы аннотированы согласно классификации CRISPR-Cas систем. V-U гены отмечены.

Supplementary information S2 (box, part g) (MS Excel):

HEPN домен содержащие белки в окрестностях CRISPR затравок. Все белки содержащие HEPN домены известных семейств в окрестностях CRISPR кассет перечислены. Следующая информация приведена: ID гена и его координаты, HEPN семейство, тип CRISPR-Cas системы, ID кластера.

Supplementary information S2 (box, part h) (MS Excel):

Последовательности, использованные для анализа V типа систем и TnpB семейств. Для каждой последовательности, которые были использованы для построения филогенетического дерева (Рисунок 10a) и дендрограммы профилей (supplementary information S3, Рисунок 11) следующая информация приведена: ID последовательности TnpB и координаты в соответствующем геноме, ID кластера, описание подсемейства, ID генома и название вида ассоциированного с CRISPR кассетой.

Supplementary information S1: Multiple alignment of C2c1 protein family

Supplementary information S4: Multiple alignment of C2c3 protein family

Supplementary information S5: Multiple alignment of C2c2 protein family

Рисунки располагающиеся в 1 файле, см. S4, S5, S6 в:

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS1_4_5.pdf

Supplementary information S3: Multiple alignment of representatives from five V-U families.

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS3.pdf

Supplementary information S6 (figure): Membrane proteins associated with Cas13b genes

ftp://ftp.ncbi.nih.gov/pub/wolf/_suppl/Shmakov/SupplementS6.pdf

Благодарности

Автор хотел бы поблагодарить своих научных руководителей: Константина Викторовича Северинова и Евгения Викторовича Кунина, они не только инициировали данный проект и поделились огромным количеством знаний и опытом, но и организовали важные сотрудничества с различными экспериментальными лабораториями.

Автор хотел бы поблагодарить Киру Макарову и Юрия Вульфа, которые являлись прекрасными менторами, и кто внёс критический вклад в успех данного проекта.

Автор хотел бы поблагодарить сотрудников экспериментальных лабораторий в Broad Institute и Rutgers, которые верифицировали сделанные предсказания, сделав это очень качественно и в очень короткие сроки.

Автор хотел бы поблагодарить всех административных работников Сколковского института науки и технологий, которые оказали огромную административную поддержку, также хотел бы поблагодарить сотрудников NCBI и NIH за предоставленные ресурсы и техническую поддержку для данного огромного проекта.

Автор выражает благодарность всем аспирантам и постдокам в Сколковском институте науки и технологий и Национальных Институтах Здоровья США за их поддержку и важные замечания по проекту, за их дружбу которая помогала мне двигаться дальше.