На правах рукописи

Шмаков Сергей Анатольевич

Разработка биоинформатического подхода для поиска новых CRISPR-Cas систем

Специальность 03.01.09 математическая биология, биоинформатика

АВТОРЕФЕРАТ

диссертации на соискание ученой степени

кандидата биологических наук

Москва 2017

Работа выполнена в центре по системной биомедицине и биотехнологии автономной некоммерческой образовательной организации высшего образования “Сколковский институт науки и технологий”.

Научные руководители:

Северинов Константин Викторович, доктор биологических наук, профессор, Автономная некоммерческая образовательная организация высшего образования «Сколковский институт науки и технологий», центр по системной биомедицине и биотехнологии.

Кунин Евгений Викторович, кандидат биологических наук, ведущий научный сотрудник, центр Биотехнологической Информации Национальных Институтов Здравоохранения США

Официальные оппоненты:

Озолинь Ольга Николаевна, доктор биологических наук, профессор, Федеральное государственное бюджетное учреждение науки «Институт биофизики клетки» Российской академии наук

Кубланов Илья Валерьевич, кандидат биологических наук, Федеральное государственное учреждение Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук

Ведущая организация:

Федеральное государственное бюджетное учреждение науки «Институт химической биологии и фундаментальной медицины» Сибирского отделения Российской академии наук

Защита диссертации состоится 16 октября 2017 года в 15 часов на заседании Диссертационного совета Д 002.077.04 при Федеральном государственном бюджетном учреждении науки Институте проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: 143025, Москва, ул. Нобеля, д.3, Инновационный центр Сколково.

С диссертацией можно ознакомиться в библиотеке и на сайте Федерального государственного бюджетного учреждения науки Института проблем передачи информации им. А.А. Харкевича Российской академии наук по адресу: http://iitp.ru/upload/content/1340/ShSdisser.pdf

Автореферат разослан «_____»_____ 2017 года.

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

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

CRISPR-Cas являются разнообразными защитными системами бактерий и архей, которые способны «запоминать» предыдущие вирусные инфекции путем встраивания фрагментов вирусной ДНК в свой геном, и использовать эту информацию для специфичной защиты от будущих инфекций за счет специфического узнавания вирусной ДНК клеточным РНК-белковым комплексом. Данный процесс интересен с точки зрения того, что он является единственным известным примером адаптивного иммунитета у прокариот, и представляет собой «Ламарковский» механизм эволюции (E. V Koonin, Y. I. Wolf 2016). Также эти системы нашли широкое применение для задач редактирования геномов прокариот и эукариот (Jinek M. et al. 2013; Ran F.A. 2013; Barrangou R. et al. 2016).

Эффекторные комплексы CRISPR-Cas систем, такие как crРНК, направляемая нуклеаза Cas9, совершили революцию в редактировании геномов за счет увеличения эффективности, специфичности и простоты внесения двуцепочечных разрывов (Ran F.A. et al. 2013; Hsu P.D. 2014; Mohanraju P. et al. 2016). Но свойства задействованных CRISPR-Cas белков не являются оптимальными для всех случаев (Slaymaker I.M. et al. 2016), например, их специфичность к определённого рода PAM последовательностям, ограничивает выбор цели; не специфичное разрезание не позволяет надёжно использовать их для чувствительных задач; размер использующихся CRISPR-Cas белков накладывает ограничения на средства доставки эффекторных комплексов в клетки. Это означает, что есть необходимость в новых CRISPR-Cas эффекторных комплексах, которые могли бы расширить и обогатить возможности CRISPR-Cas инструментов, как, например, Cpf1, который был недавно охарактеризован (Zetsche B. et al. 2015) и показал уникальные свойства: большая специфичность, липкие концы после внесения разреза и возможность таргетировать сразу несколько целей. Этот пример показывает, что тщательное исследование геномных и метагеномных данных может дать новые варианты CRISPR-Cas эффекторных комплексов, которые могут обладать более оптимальными свойствами для определённых задач или открыть новые области применения. Исследование подобного рода было проведено для полных геномов (полностью отсеквенированных геномов) в 2015 году (Makarova K.S. et al. 2015), которое сделало подробную характеризацию систем на использовавшемся наборе данных, и показало, что основные варианты CRISPR-Cas систем уже известны, но редкие варианты всё же могут быть обнаружены. Однако, большая часть геномных данных, которая включает в себя не полные геномы (частично отсеквенированные) и

метагеномные базы данных не были покрыты в данном исследовании, таким образом, новые системы могут быть обнаружены данном, не покрытом ранее наборе данных.

В представленной работе была предпринята попытка оценки разнообразия, а также поиск новых CRISPR-Cas систем 2 класса в непокрытой ранее области геномных и метагеномных данных бактерий и архей. Полученные в результате данные могут лучше объяснить взаимодействия вирусов и их прокариотических хозяев. Оценка разнообразия или новые варианты систем 2 класса позволят получить более полное понимание эволюции систем защиты и CRISPR-Cas систем в частности. Новые варианты систем могут дать новые возможности для задач редактирования геномов или расширить область биотехнологического применения механизмов CRISPR-Cas систем.

Цели и задачи исследования

Целями настоящей работы были исчерпывающая оценка разнообразия CRISPR-Cas систем 2 класса и поиск новых вариантов эффекторных комплексов CRISPR-Cas систем 2 класса в доступных геномных и метагеномных данных.

Для достижения поставленных целей был разработан биоинформатический подход, который включает в себя решение следующих задач:

1. Нахождение всех Cas1 белок-кодирующих участков и CRISPR содержащих участков (затравок) в доступных геномных и метагеномных данных.
2. Аннотация прилегающих последовательностей всех затравок на предмет белок-кодирующих участков и типов CRISPR-Cas систем
3. Выделение белковых семейств путём кластеризации последовательностей и оценка их связанности с затравками
4. Автоматическое выделение перспективных кандидатов из найденных кластеров
5. Ручное курирование выделенных кандидатов с помощью чувствительных инструментов для определения белковых доменов
6. Выделение и биоинформатическая характеризация наиболее перспективных семейств

Научная новизна, теоретическая и практическая значимость

Главная научная ценность работы заключается в том, что в ней впервые была проведена исчерпывающая оценка разнообразия CRISPR-Cas систем 2

типа в доступных на 2016 год геномных и метагеномных данных. Данная оценка позволила обнаружить шесть новых подтипов CRISPR-Cas систем, включая один подтип содержащий несколько предполагаемых подсистем. Было описано три новых подтипа для V типа: Тип V-B, Тип V-C, предполагаемый подтип V-U содержащий RuvC нуклеазный домен. А также был обнаружен ранее не описанный VI тип CRISPR-Cas систем включающий в себя три подтипа: Тип VI-A, Тип VI-B и Тип VI-C; которые включают в себя два HEPN (higher eukaryotes and prokaryotes nucleotide-binding domains) домена и для которых было биоинформатически предсказано (а позже независимо экспериментально показано для Тип VI-B и Тип VI-C) таргетирование РНК.

Открытые системы могут использоваться как биотехнологические инструменты, в том числе, для задач редактирования геномов. Новые подтипы V типа отличаются от известных эффекторных комплексов следующими уникальными свойствами структурой (кристаллическая структура была независимо разрешена для V-B подтипа), PAM последовательностью и наличием tracrРНК для V-B подтипа. HEPN домены VI типа, могут быть использованы для таргетирования РНК молекул, что может быть использовано для детектирования РНК, изменения или измерения уровня экспрессий, визуализации молекул РНК и т.д. Системы предполагаемого V-U подтипа могут быть эффективно использованы для доставки с помощью вирусных векторов в клетки в связи с их маленьким размером (если их bona fide CRISPR активность будет доказана).

Недавние независимые исследования показали использование открытого типа VI-A (Cas13a) как очень высокоспецифичного механизма определения молекул нуклеиновых кислот, что было применено для детектирования различных вирусов, патогенных бактерий, детектирования раковых клеток и т.д.

Методология и методы работы

Работа выполнена с использованием современных программных средств и биоинформатических методологий и многопроцессорных вычислительных кластеров.

Основные положения, выносимые на защиту

1. Разработанный биоинформатический подход для поиска CRISPR-Cas эффекторных комплексов второго класса.
2. Открытие 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 открытых CRISPR-Cas систем.

Степень достоверности и апробация результатов работы

Цели, поставленные в работе, достигнуты, результаты приведенных в работе экспериментов грамотно интерпретированы, и сделанные в работе выводы обоснованы и независимо экспериментально подтверждены, их достоверность не вызывает сомнений. Результаты работы были опубликованы в рецензируемых научных журналах и представлены на международных конференциях.

Публикации

Результаты работы были представлены на трех научных конференциях и опубликованы в четырёх статьях в рецензируемых научных журналах (см. «Список работ, опубликованные по теме диссертации»).

Структура и объем работы

Диссертационная работа состоит из введения, обзора литературы, материалов и методов, результатов и их обсуждения, заключения, выводов, списка литературы и благодарностей. Работа изложена на 104 страницах машинописного текста, включая 16 рисунков и 2 таблиц. Список цитируемых литературных источников включает 211 наименований.

СОДЕРЖАНИЕ РАБОТЫ

1. Биоинформатический подход для поиска новых локусов CRISPR–Cas 2 класса.

Поиск новых CRISPR-Cas систем 2 класса был начат с разработки биоинформатического подхода для поиска этих систем и его имплементации в виде стека программ и подходов изображенного на Рисунке 1.

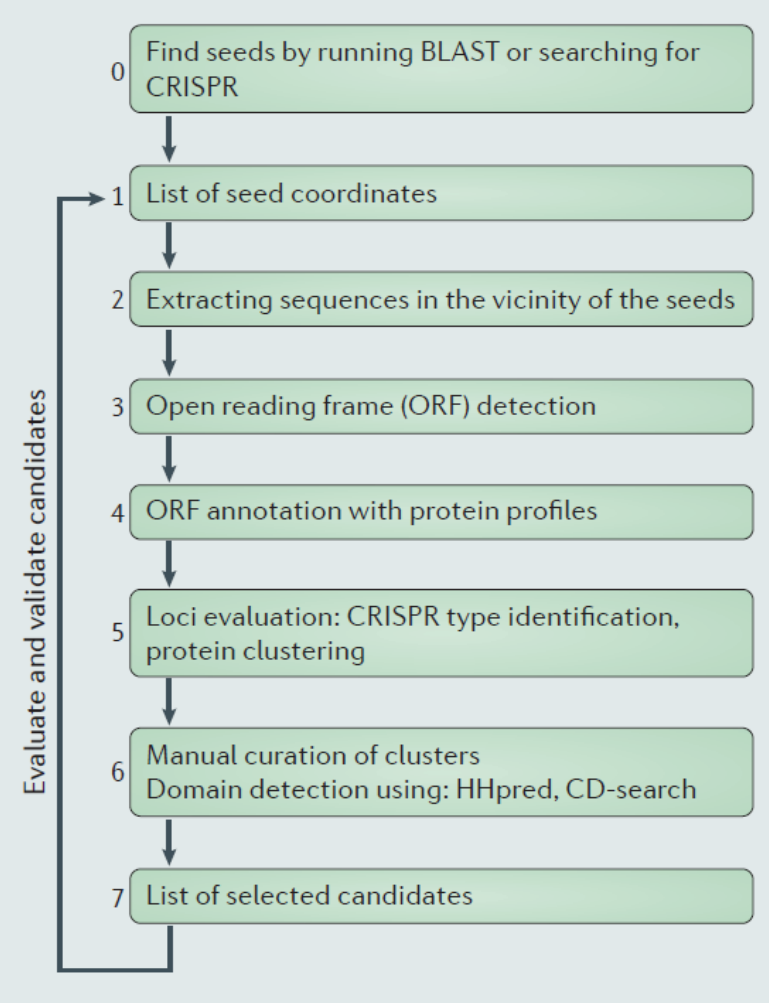


Рисунок 1. Схематическое изображение разработанного биоинформатического подхода для новых CRISPR-Cas систем (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)). Изображен программный комплекс и подходы для поиска новых эффекторных комплексов CRISPR-Cas систем 2 класса. Детали для показанных на рисунках шагов приведены ниже.

Для поиска новых CRISPR-Cas систем и исчерпывающей оценки разнообразия систем 2 класса был произведён поиск возможных CRISPR-Cas локусов в геномных и метагеномных базах данных с помощью двух затравок, представляющие ключевые элементы CRISPR-Cas систем: Cas1 белок-

кодирующие области, представляющие адаптационные модули (так как ген *cas1* является наиболее распространённым и консервативным среди всех CRISPR-Cas систем) и CRISPR кассеты являющиеся необходимым элементом для работы эффекторных комплексов CRISPR-Cas систем. Белок-кодирующие области для *cas1* были найдены с помощью белковых профилей и BLAST. CRISPR кассеты были найдены с помощью Piler-CR и CRISPRFinder (было взято две программы для обеспечения большей чувствительности), результаты их предсказаний были объединены в один набор, использовавшийся для поиска далее. Данная процедура нашла 47,174 CRISPR кассет, что почти вдвое больше обнаруженных Cas1 белков.

Далее все локусы содержащие затравки были проаннотированы на наличие открытых рамок считывания и была произведена их аннотация с помощью белковых профилей CDD и отдельных белковых Cas профилей описанных в литературе. Все белки в найденных локусах были прокластеризованы для определения белковых семейств. Для всех локусов был проаннотирован тип CRISPR-Cas системы и локусы содержащие известные, полные типы CRISPR-Cas были отфильтрованы, оставляя для дальнейшего анализа только локусы с неизвестными или не полными типами. Следующим этапом фильтрации была выборка больших белков стоящих рядом с затравками (Cas1 или CRISPR) по характерному размеру для белков 2 класса CRISPR-Cas систем. Для этих белков была установлена степень эволюционной связи с затравками, для устранения случайно (локусы содержащие защитные гены сильно вариативны) оказавшихся рядом с Cas1 или CRISPR больших белков. Для этих белков был произведён также высокочувствительный поиск белковых доменов. Белки содержащие домены не относящиеся к работе CRISPR-Cas систем (метаболические или трансмембранные транспортеры и т.д.) были отброшены, оставляя белки содержащие нуклеазные домены. Данные белки были дополнительно биоинформатически охарактеризованы: произведён многоитеративный поиск гомологов, для них установлена эволюционная связь с затравками. Наиболее перспективные белки (см. Рисунок 2, 3) были отправлены для экспериментальной валидации.

Данная процедура поиска CRISPR-Cas систем является исчерпывающей, так как все большие белки рядом с *cas1* или CRISPR были детально проанализированы. Только точность распознавания Cas1 или CRISPR и ограничение на размер белка являются ограничением для данного подхода.

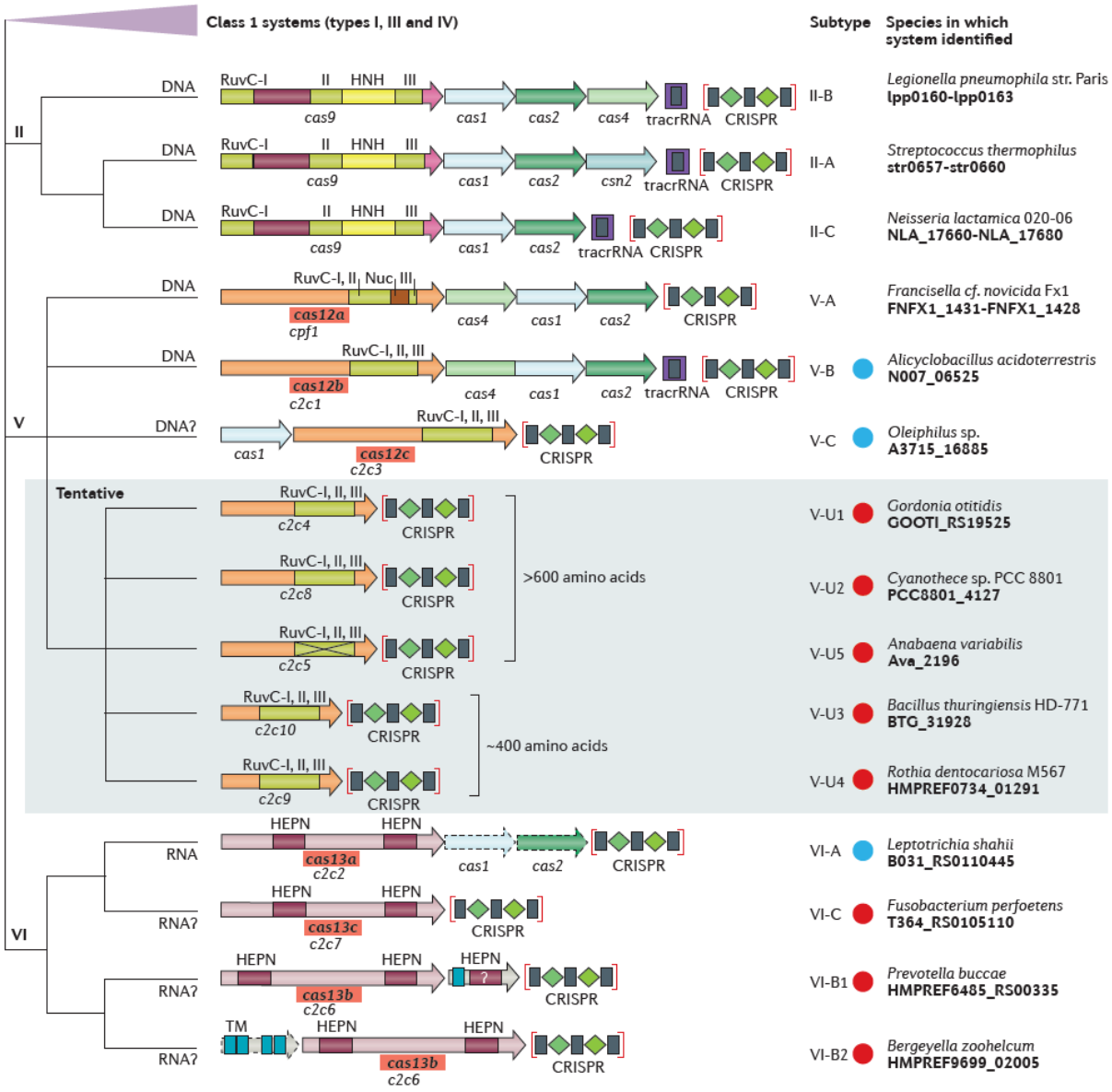

Class 1 systems (types I, III and IV)
Subtype
Species in which system identified
II
V
VI
DNA
DNA
DNA
DNA
DNA
DNA?
Tentative
RNA
RNA?
RNA?
RNA?
RuvC-I II HNH III
cas9
cas1 cas2 cas4 tracrRNA CRISPR
II-B
Legionella pneumophila str. Paris
lpp0160-lpp0163
RuvC-I II HNH III
cas9
cas1 cas2 csn2 tracrRNA CRISPR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

Рисунок 2. Обновлённая классификация для CRISPR–Cas систем 2 класса с учётом открытых систем (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)). Все системы первого класса схлопнуты в верху данной схемы; все остальные показанные системы принадлежат ко 2 классу. Новые системы, которые были обнаружены с помощью описываемого подхода (см. Рисунок 1), 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.

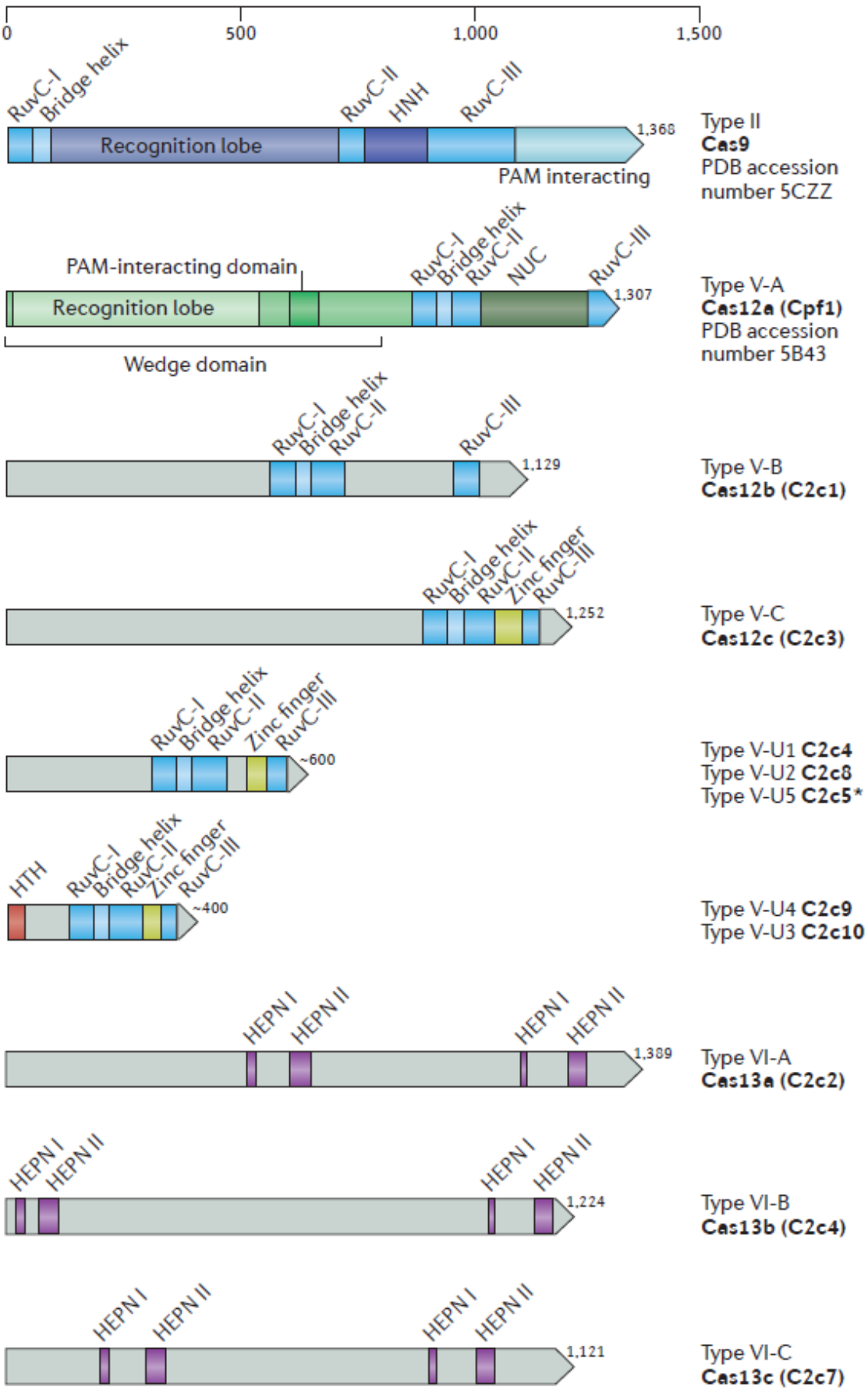

0
500
1,000
1,500
RuvC-I
Bridge helix
RuvC-II
HNH
RuvC-III
Recognition lobe
1,368
PAM interacting
Type II
Cas9
PDB accession
number 5CZZ
PAM-interacting domain
RuvC-I
Bridge helix
RuvC-II
NUC
RuvC-III
Recognition lobe
1,307
Wedge domain
Type V-A
Cas12a (Cpf1)
PDB accession
number 5B43
RuvC-I
Bridge helix
RuvC-II
RuvC-III
1,129
Type V-B
Cas12b (C2c1)
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)
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 3. Доменная архитектура CRISPR-Cas эффекторных белков 2 класса систем (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)). 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 предварительные имена указаны для не охарактеризованных вероятных эффекторов типа V-U. 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.

2. Подтипы V-B и V-C: большие мульти-доменные эффекторы

В результате описанного выше подхода были обнаружены эффекторы (используя *cas1* затравки), согласно структурным, доменным особенностям и сходству последовательностей, которые были отнесены к V типу CRISPR-Cas систем. Обе ранее известных группы эффекторов CRISPR-Cas систем II и V типа содержат RuvC подобный нуклеазный домен, но только Cas9 содержит HNH нуклеазный домен. Но, не смотря на то, что RuvC является единственным схожим по последовательности элементом между этими группами эффекторов, они имеют схожую форму согласно разрешённым кристаллическим структурам. Все эффекторы V типа, обнаруженные в данной работе, имеют размер в ~900 - ~1300 аминокислот и RuvC подобный нуклеазный домен (см. Рисунок 3), являющимся единственным известным нуклеазным доменом в данном белке. Данные особенности указывают на то, что все белки V типа образуют билобные структуры, способные захватывать crРНК и целевую ДНК, не смотря на возможное независимое происхождение этих систем. Данные белки содержат районы гомологичные TnpB – три RuvC каталитических мотива RuvC подобной нуклеазы (см. Рисунок 3), и

соединённый с ними район bridge helix, который отвечает за связвание с crРНК (как это было показано с Cas9). Консервативность внутри каталитических RuvC мотивов указывает на нуклеазную активность этих белков. N концевые участки обоих белков не имеют детектируемого сходства, но предсказание вторичной структуры показывает, что два региона принимают смешанную α/β конформацию. Таким образом общая доменная архитектура для найденных белков походит на Cpf1 (Тип V), но не на Cas9 (Тип II). Согласно этому, данным белкам были присвоены типы V-B и V-C а Cpf1 получил тип V-A.

Поиск гомологов для CRISPR-Cas систем 2 класса, показывает, что RuvC подобный нуклеазный домен имеет сходство с *tnpB* автономных и не автономных транспозаз. Изучение TnpB белков показало, что они могут связываться с РНК за счет длинной позитивно заряженной α-спирали, что схоже с эффекторами 2 класса. Поиск гомологов для Cas9 показал гомологичность к определённой группу TnpB подобных транспозонов IscB. Белки же V класса не показывают схожесть к какой-либо отдельной группе TnpB, что может указывать на независимое происхождение этих эффекторов.

Белок Cas12b (C2c1) из *Alicyclobacillus acidoterrestris ATCC 49025 (Aac)* был независимо экспериментально охарактеризован в лаборатории Фанг Жанга. Для этой системы было показано: активная транскрипция CRISPR кассеты. Транскрипция участка рядом с кассетой, которая имеет сходство с повтором кассеты, и было показано, что данная система способна к интерференции, то есть является bona fide CRISPR-Cas системой. Отдельная РНК молекула, закодированная в локусе, является tracrРНК, что было показано путём предсказания вторичной структуры crРНК и tracrРНК и экспериментально.

3. Подтип V-U: маленький возможный эффектор

Другой группой приписанной к V типу оказалось несколько TnpB подобных семейств (см. Рисунки 2,3) эволюционно стабильно связанных с CRISPR кассетами. Данные семейства были найдены с помощью CRISPR затравок. Из-за свойств описанных далее, данная группа семейств получила предварительный подтип V-U (где ‘U’ означает ‘uncharacterized’ – не охарактеризованный). Все найденные семейства объединяет две ключевые особенности, отличающие их от других эффекторных белков: их размер составляет от ~500 аминокислот (средний размер TnpB транспозонов) до ~700 аминокислот (размер между TnpB и bona fide эффекторами 2 класса CRIPSR-Cas систем); эти семейства показывают больший уровень сходства с TnpB чем эффекторы II и V типов. Второе справедливо для трёх подгрупп данного типа (обозначенные как: V-U1, V-U2 и VU-5), которые аннотируются TnpB профилями, но имеют эволюционную стабильность в рамках консервации

последовательностей, стойкую ассоциацию с CRISPR кассетами и присутствуют в различных группах бактерий.

Для расширения данной группы был запущен поиск (см. Рисунок 1) без фильтрации по длине эффектора, что дало большое количество TnpB подобных семейств белков. Но большинство из полученных локусов не были эволюционно консервативными, таким образом их связь с CRISPR кассетой сомнительна. Данный поиск дал две дополнительные группы небольших CRISPR ассоциированных TnpB (V-U3 и V-U4), имеющие схожие свойства с описанными выше семействами V-U. Также все эти группы имеют признаки стабилизирующего отбора на уровне последовательностей белков. 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.

Эволюционный анализ данных групп показывает, что только одно семейство V-U1 находится в различных типах бактерий, распространение остальных ограничены одним таксоном. Также филогенетический анализ указывает на то, что все системы V типа произошли независимо из различных семейств TnpB.

Маленький размер данных белков указывает на то, что они не могут принимать традиционную билобную структуру (захватывающую crРНК и целевую ДНК), таким образом их эффекторная функция маловероятна, без дополнительных белков (в локусах этих семейств других Cas белков не было найдено). Дополнительные белки могут приходить из других локусов присутствующих в геномах. Состав CRISPR кассет отличается между видами, что указывает на активный набор спэйсеров. Также для типа V-U3 были найдены вирусные протоспэйсеры. Это указывает на то, что данные системы имеют какую-то CRISPR связанную функцию, а некоторые могут быть функциональными CRISPR-Cas системами. Альтернативно, некоторые системы могут иметь регуляторную активность. Например, подгруппа V-U5 имеет инактивированный нуклеазный домен, это указывает на то, что данная подгруппа не способна разрезать ДНК.

4. Подтипы VI-A, VI-B и VI-C: РНК таргетирующие CRISPR–Cas многоблоковые эффекторы

Отличительной особенностью другой найденной группы белков, является наличие двух РНКазных доменов HEPN (Higher Eukaryotes and Prokaryotes Nucleotide-binding) (см. Рисунки 2, 3), что выделило их в отдельный тип VI. HEPN домен распространён среди различных систем защиты, таких как: токсины из прокариотических токсин-антитоксин систем или эукариотическая RNase L, которые имеют РНКазную активность. Таким образом для этих белков была предсказана РНКазаная активность. Поиски по

геномным базам данных не показали схожести этих белков к каким-либо известным семействам. Но множественное выравнивание показало, что среди этих групп есть два консервативных R(N)xxxH мотива, которые характеризуют HEPN домен, а также предсказывается длинная α-спираль, которая согласуется с HEPN фолдом. Предсказание вторичной структуры, также не дало сходства с известными системами. Три подсистемы, различаются между собой различным положением HEPN доменов в эффекторном комплексе, а так же структурой локусов – Тип VI-B имеет дополнительные белки (которые разделяют подтип на две подгруппы).

Два семейства из трёх было экспериментально охарактеризовано в лаборатории Фанг Жанга. Для системы получившей подтип VI-A, была подтверждена РНКазная активность для эффекторного комплекса Cas13a (или C2c2), на примере защиты от РНК бактериофага MS2. Для данного белка была найдена уникальная особенность: после специфической интерференции, данный комплекс становится неспецифичной РНКазой, что имеет токсичный эффект и подавляет рост бактерии. Так же было показано, что этот белок способен к процессингу pre-crРНК.

Система VI-B была второй экспериментально охарактеризованной системой. В локусах данной системы были обнаружены белки VI-B1 и VI-B2 (см. Рисунок 2). Данные белки имеют дополнительную функции и содержат трансмембранные домены: VI-B1 кодирует 4 таких домена и VI-B2 кодирует один. Филогенетический анализ дополнительных белков и эффекторных комплексов показывает, что эти варианты разошлись во время эволюции. Для VI-B2 также был предсказан один HEPN домен. Экспериментальная характеризация данной системы показала, что данный белок обладает РНКазной активностью, а дополнительные белки могут регулировать РНК интерференцию (VI-B1 подавляет и VI-B2 улучшает).

Все найденные подсистемы VI типа имеют размер, характерный для CRISPR-Cas систем 2 класса, что указывает на их возможную функциональность. Системы не имеющие адаптационного модуля (см. Рисунок 2), вероятно, полагаются на другие системы присутствующие в геноме.

5. Исчерпывающая оценка разнообразия CRISPR-Cas систем 2 класса в локусах бактерий и архей

Для оценки разнообразия CRISPR-Cas систем 2 класса, были созданы профиля для всех эффекторных белков, включая открытые в данной работе (за исключением V-U5; V-U3 и V-U4 в связи с их схожестью с TnpB). Далее был выполнен поиск этих профилей в прокариотической базе данных, содержащих 4,961 полностью отсеквенированных геномов и 43,599 частично отсеквенированных геномах, которые доступны в National Center for

Biotechnology Information (NCBI) базе данных. Подобный поиск определил все экземпляры эффекторов.

Найденные локусы были проверены на наличие CRISPR кассет и других *cas* генов. Данная оценка показала доминирование II типа систем во 2 классе, которые представляют 8% бактериальных геномов. Количество остальных типов оказалось на порядок меньше.

	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 1. Оценка разнообразия CRISPR–Cas систем 2класса в прокариотах (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)).

‡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.

Почти каждая из CRISPR-Cas систем 2 класса представляет таксономически разнообразные группы бактерий. Было показано, что филогенетические деревья эффекторных комплексов отличается от топологии видов, что может указывать на активную роль горизонтального переноса в эволюции CRISPR-Cas систем. Тип VI-B отличается тем, что он присутствует только в бактероидетах, что может вызвано особенной биологией этих бактерий.

Все CRISPR-Cas системы 2 класса, находятся только в бактериях (за исключением двух вариантов V-A в археях), археи же представлены только первым классом, что указывает на не ясное функциональное отличие двух классов CRISPR-Cas систем.

6. Обновлённая классификация CRISPR–Cas систем 2 класса

Новые найденные системы, потребовали обновления классификации для 2 класса CRISPR-Cas систем. К четырём известным ранее подтипам, данная работа добавила шесть новых (см. Рисунок 2). Тип V-U остаётся не охарактеризованным, но если появится подтверждение их CRISPR активности, то они получат отдельные подтипы в V типе. Исходя из подробности данного исследования, которое подтвердило предсказания сделанные ранее, новые CRISPR-Cas системы, если они будут найдены, будут ещё более редкими чем найденные или ограничены группами бактерий или архей, которые не представлены адекватно в текущих геномных базах данных.

Расширение классификации потребовало изменения в номенклатуре системных имён *cas* генов: эффекторы V-A (бывший V), V-B, V-C типов получили имена Cas12a, Cas12b, Cas12c; эффекторы VI-A, VI-B, VI-C получили имена Cas13a, Cas13b, Cas13c. Предполагаемый подтип V-U не получил имён, так как его bona fide CRISPR активность не была доказана. В случае доказательства этой активности для какой-либо из подгрупп, она будет относиться к Cas12.

7. Возникновение новых CRISPR-Cas систем 2 класса

Гипотеза о возникновении систем 2 класса из TnpB подобных элементов была предложена ранее. Данное исследование описывает новые системы в рамках этой теории. Дополнительно используется информация полученная из исследования V-U типа (см. Рисунок 4).

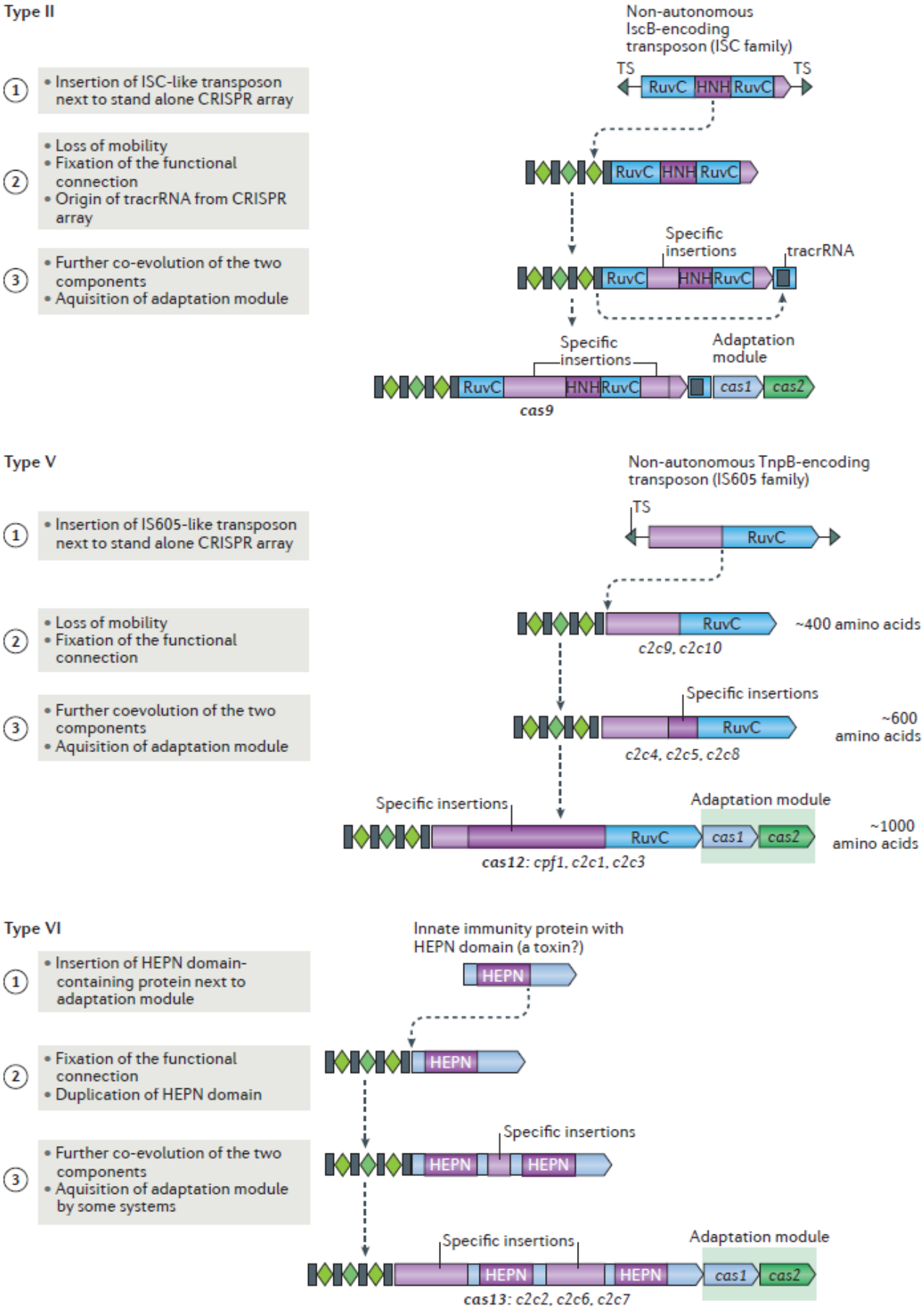

Type II
1
• Insertion of ISC-like transposon next to stand alone CRISPR array
2
• Loss of mobility
• Fixation of the functional connection
• Origin of tracrRNA from CRISPR array
3
• Further co-evolution of the two components
• Aquisition of adaptation module
Non-autonomous IscB-encoding transposon (ISC 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)
TS
RuvC
~400 amino acids
c2c9, c2c10
Specific insertions
~600 amino acids
c2c4, c2c5, c2c8
Specific insertions
Adaptation module
~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
Specific insertions
Adaptation module
cas1
cas2
cas13: c2c2, c2c6, c2c7

Рисунок 4. Возможные варианты эволюции CRISPR–Cas систем 2 класса (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)). Данный рисунок отображает трёх этапный путь эволюционного развития CRISPR-Cas 2 класса. Систематичные и/или изначальные названия для имён генов расположены ниже полноценных 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).

Пять вариантов V-U типа имеют схожесть к TnpB элементам, но в то же время стабильно ассоциированы с CRISPR кассетами, имея размер от стандартного для TnpB до размера близкого к эффекторам 2 класса. Данные группы могут представлять собой промежуточные звенья в становлении CRISPR-Cas системы 2 класса. Другие TnpB, найденные рядом с CRISPR повторами, не имеют стабильной ассоциации, но могут представлять зачаточные 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, представляющих первый этап, происходят путём случайной вставки TnpB-кодирующего элемента рядом с одиночной CRISPR кассетой (см. этап 1 на Рисунке 4). Следующий этап подразумевает фиксацию ассоциации между CRISPR кассетой и TnpB гомологом и его увеличением путём дупликаций или вставок (см. этап 2 на Рисунке 4). Данная фиксация подразумевает функцию, не известную на данный момент. Исследование V-U типа, может дать больше информации о данном этапе. Последний этап подразумевает дальнейший рост эффектора до полноценного размера, способного вмещать crРНК и цель (см. этап 3 на Рисунке 4). В некоторых случаях на последнем этапе, CRISPR-Cas системы 2 класса, могут приобретать адаптационные модули из других систем, что подтверждается фактом того, что Cas1 белки для различных подгрупп

подтипов II и V-A типов имеют происхождение из разных CRISPR-Cas систем 1 класса. Подобный сценарий можно применить к типу VI, за исключением того, что в данной работе не удалось проследить происхождение HEPN домена, который возможно пришёл от однго из Cas белков или HEPN содержащих токсинов.

Обратный сценарий, того, что TnpB деградировал из эффекторного комплекса маловероятен, в связи с тем, что: TnpB много больше распространены во всех видах бактерий и архей; размер и сложность эффекторов делает их маловероятными предковыми формами; TnpB могут встроиться около CRISPR кассет, а CRISPR-Cas системы не имеют мобильности; филогения TnpB белков и ветки их вариантов ассоциированных с CRISPR кассетами подразумевает предковый статус TnpB.

8. Возможные применения для новых CRISPR-Cas систем 2 класса

Открытые CRISPR-Cas системы 2 класса, вносят как больше разнообразия, так потенциально новые функции, которые могут быть применены для создания биотехнологических инструментов. Существующие инструменты - варианты Cas9 начали революцию в области редактирования геномов, но они не избавлены от проблем связанных со специфичностью, внецелевыми эффектами, ограниченностью связанную с PAM последовательностью.

Как открытые так и известные варианты эффекторных комплексов описанные в этой работе (см. Оценка разнообразия CRISPR-Cas систем) образуют большое разнообразие механизмов действия эффекторных комплексов (см. Рисунок 5): различные цели – ДНК или РНК, наличие tracrРНК, PAM последовательность, тип разрезания ДНК. Пример Cas12a (Cpf1) показывает, что новые эффекторные комплексы быстро начинают применяться в биотехнологии, что является стимулом для дальнейшей характеризации вариативности CRISPR-Cas систем.

Открытый и охарактеризованный эффекторный комплекс V-B подтипа, располагается между Cas9 и Cpf1, имея отличия в структуре и доменах V типа, но используя tracrРНК, и иной PAM мотив. Эта вариативность может получить свою собственную нишу. Белки подтипа V-U, если для них будет показана bona fide CRISPR активность, могут применяться для эффективной доставки в связи с их малым размером.

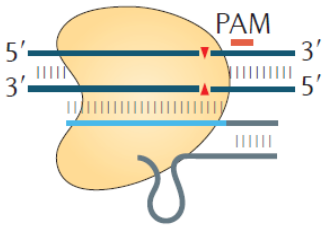
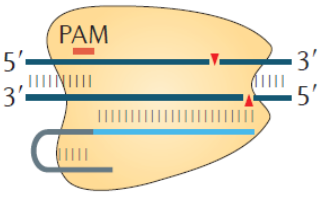
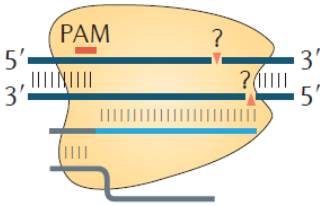
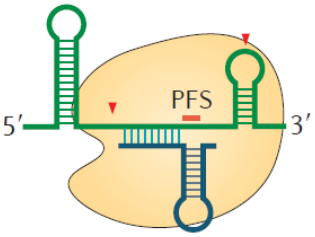
		Nuclease domains	tracrRNA	PAM	Substrate	Cleavage pattern
Type II Cas9		RuvC and HNH	Yes	3′, GC-rich	dsDNA	Blunt ends
Type V-A Cas12a (Cpf1)		RuvC and Nuc	No	5′, AT-rich	dsDNA	Staggered ends, 5′overhangs
Type V-B Cas12b (C2c1)		RuvC	Yes	5′, AT-rich	dsDNA	Staggered seven-nucleotide cut of target DNA
Type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

Рисунок 5. Функциональное разнообразие экспериментально охарактеризованных CRISPR–Cas систем 2 класса (воспроизведено с разрешения Nature reviews. Microbiology (Shmakov *et al.*, 2017)). 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).

Открытый VI тип CRISPR-Cas систем добавил новую уникальную возможность для технологий технологий РНК-направляемого РНК таргетирования, благодаря который возможно изменять, модулировать, модифицировать и отслеживать специфичные РНК транскрипты в клетках. Создание мутированного “dead” Cas13a с выключенной РНКазной

активностью, может дать механизмы для того чтобы наблюдать различные состояния клеток, манипулировать трансляцией, отслеживать уровни РНК локализации в живых клетках. Неспецифичная активность Cas13a может использоваться для селективного устранения или ингибирования клеток (например основываясь на уровне экспрессии).

Недавняя независимая работа показала применение открытого Cas13a для детектирования ДНК или РНК на аттомолярных уровнях. Что было показано на примере диагностики: определение типов вирусов, патогенных бактерий, детектирование клеток опухолей.

Основные результаты

1. В данной работе был разработан биоинформатический подход для нахождения новых CRISPR-Cas систем 2 класса и оценки разнообразия этого класса в прокариотах.
2. Было обнаружено шесть новых систем CRISPR-Cas 2 класса, включая один предполагаемый подтип 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. Подтипы V-B, VI-A, VI-B были независимо экспериментально охарактеризованы, подтверждая предсказания сделанные в данной работе.
3. Была обновлена классификация CRISPR-Cas систем, добавляя 6 новых подтипов и систематическое именование для них.
4. Выполнена исчерпывающая оценка разнообразия систем 2 класса показывающая распространение и дающая количественную оценку этих систем среди бактерий и архей.
5. Представлены гипотезы о происхождении CRISPR-Cas систем 2 класса.
6. Предложены варианты применения открытых CRISPR-Cas систем в качестве биотехнологических инструментов.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в научных журналах:

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.

Материалы конференций:

1. **Shmakov SA**, Makarova KS, Wolf YI. CRISPR Effector Discovery Pipeline. // CRISPR 2016. Rehovot, Israel, 2016.
2. **Shmakov SA**, Makarova KS, Wolf YI. CRISPR Effector Discovery Pipeline. // Genome Engineeng 4.0. Boston, USA, 2016.
3. **Shmakov SA**, Sitnik V, Makarova KS, Wolf YI, Severinov KV, Koonin EV. The CRISPR Spacerome. // CRISPR 2017. Big Sky, USA, 2017