ОТЗЫВ

официального оппонента на диссертационную работу
Шмакова Сергея Анатольевича на тему «Разработка биоинформатического подхода для поиска новых CRISPR-Cas систем», представленную на соискание ученой степени кандидата биологических наук
по специальности 03.01.09 – математическая биология, биоинформатика

Актуальность темы исследования

CRISPR-Cas являются адаптивными иммунными системами, защищающими прокариотические организмы от мобильных генетических элементов. Данный механизм защиты состоит из следующих этапов: а) набор новых фрагментов в CRISPR кассеты из чужеродных генетических последовательностей (адаптация); б) экспрессия CRISPR кассеты и её процессинг в небольшие РНК фрагменты (crРНК) и в) деградация чужеродных элементов путём распознавания комплексом эффектор-crРНК участков чужеродной ДНК или РНК, комплементарных crРНК и гидролиза соответствующей мишени (интерференция). В течение последних нескольких лет CRISPR-Cas системы активно исследуются, что связано, в том числе и с их применением в биотехнологии и медицине в качестве инструментов для редактирования геномов.

До проведения данной работы выделяли 2 класса, 5 типов и 16 подтипов разных CRISPR-Cas систем, что указывает на их большую вариативность как в строении, так и в функциях. Системы второго класса, а именно большие мультидоменные и многофункциональные белки выполняющие функции процессинга pre-crРНК и распознавания/уничтожения цели, используются как эффективные инструменты для редактирования геномов как прокариот, так и эукариот. Данная эффективность достигается за счет высокоспецифичного узнавания участков ДНК, комплементарных crРНК и относительной простоты использования данных систем. При этом у них есть и недостатки, связанные в первую очередь с доставкой, выбором подходящей цели и не 100% специфичностью их действия. Для преодоления этих недостатков и/или улучшения эффективности CRISPR-Cas систем предпринимаются подходы по их оптимизации (с применением методов искусственного мутагенеза), или поиск ранее неизвестных природных вариантов, наподобие недавно найденного белка Cpf1 или систем, обнаруженных в ходе данной работы.

Научная новизна исследования

Научная новизна данной работы заключается в следующем:

a) Был разработан и применен на практике алгоритм поиска новых CRISPR-Cas систем второго класса;
b) был проведен наиболее полный на сегодняшний день анализ разнообразия CRISPR-Cas систем второго класса по доступным (мета)геномным данным бактерий и архей, который позволил оценить разнообразие и распространение данных систем среди репрезентативных микроорганизмов, а также выявить новые системы (см. ниже);
c) Были обнаружены новые типы CRISPR-Cas систем: подтипы V-B, V-C и предполагаемый подтип Type V-U; тип VI, включающий 3 подтипа VI-A, VI-B и VI-C.

Помимо этого, следует отметить, что некоторые результаты данной работы уже успели получить экспериментальные подтверждения, в частности другими авторами были получены 3D структуры подтипов V-B и VI-A, а для подтипов VI-A и VI-B было экспериментально показано таргетирование РНК.

Структура и объем работы

Диссертационная работа С.А. Шмакова изложена на 104 страницах и имеет классическую структуру: введение, обзор литературы, материалы и методы, результаты и обсуждение, заключение и выводы, последние почему-то называются «результаты исследования». При этом диссертация содержит дополнительные материалы, описание и ссылки на которые представлены в разделе «Приложение», что, насколько я знаю, является довольно необычным для диссертационных работ. Однако такая практика используется практически во всех современных научных журналах, и была разумно аргументирована автором, поэтому никаких нареканий это не вызвала.

Обзор литературы в достаточной степени раскрывает тематику исследования CRISPR-Cas систем, методическая часть также содержит всю необходимую информацию, хотя, пожалуй, написана несколько более сжато, чем хотелось бы. Структура самой работы и ее обсуждения ясна и логична и включает всю необходимую для понимания алгоритма действий и полученных результатов информацию. Имеющиеся рисунки хорошо дополняют текст и, за рядом незначительных исключений, расположены логичным образом и согласно контексту. Частичными исключениями являются рисунки 10 и 13, оба состоят из двух частей, которые недостаточным образом визуально разделены друг от друга.

Несмотря на то, что глобальная структура текста выглядит хорошо, вызывают нарекания структуры некоторых предложений и абзацев (см. ниже).

Достоверность полученных результатов, степень обоснованности результатов и выводов

Достоверность полученных результатов не вызывает сомнения, что также подтверждается тем, что защищаемая работа была опубликована в таких журналах как Nature Reviews Microbiology, Science и Molecular Cell, которые характеризутсяя в высшей степени высоким качеством peer review процесса. К слову, используя представленные в приложении материалы, я повторил некоторые расчеты и получил схожие результаты, например, филогенетический анализ, представленный на рисунке 15.

Выводы, сделанные по результатам исследования полностью обоснованы, хотя есть небольшие замечания к их формулировкам, в первую очередь, стилистические (см. ниже).

Результаты представлены на 3 международных конференциях и опубликованы в 4 рецензируемых журналах: Nature Reviews Microbiology, Molecular Cell и Science (названия говорят сами за себя), причем в двух случаях С.А. Шмаков был первым автором. Таким образом, количество и качество публикаций соответствуют требованиям ВАК для кандидатских диссертаций.

Вопросы и замечания

Серьезных замечаний к идее и реализации данной работы нет. Главное замечание – это стиль изложения и многочисленные орфографические и пунктуационные ошибки. Работа написана крайне сухим языком из-за чего некоторые вещи понимаешь только в том случае, если ты их заранее знаешь или после дополнительного чтения литературы. Помимо этого, как уже говорилось выше, внутренние структуры некоторых предложений и/или целых абзацев не выверены. Все это крайне затрудняет чтение, в некоторых случаях приходится буквально продираться сквозь текст. Также в работе иногда применяются ненужные англицизмы или неудачные, с моей точки зрения, термины, такие как «**имплементирован**» (применен), «энзимы» (ферменты), «**билобный**» (сдвоенный/двойной/двудоменный), «**частота пропуска**» (к параметрам множественного выравнивания, обозначает, по всей видимости, «соотношение гапов и неоднозначных остатков к общему числу остатков» и можно было бы заменить, например, на «пропорция неопределенных признаков/аминокислот/нуклеотидов»), «**виновность по ассоциации**» (в русском нет такого определения, и надо было бы привести английский вариант в скобках), «**затравка**» (лучше «запрос» т.к. у слова «затравка» есть устойчивое значение в биологии

– это синоним праймера в ПЦР). Стоит, однако, отметить, что многое (но не все) из выше перечисленного больше обусловлено тем, что русскоязычная литература не поспевает (и никогда не будет поспевать) за англоязычной. Встречались и забавные случаи, например, «**позитивно** заряженные основания» (положительно) или «доставка **по средствам** наночастиц», «эффектор становится **неразборчивой** РНКазой» (неспецифичная/неизбирательная/широкоспецифичная) и одна довольно непростительная ошибка для человека, претендующего на степень кандидата биологических наук, а именно «гл**ю**тамат» (глутамат). Довольно много примеров несоответствия чисел, склонений, родов, падежей, и т.д., что также невероятно затрудняет чтение, особенно учитывая то, что многие предложения составные и длинные. Например, «для которых была показан**о** или предсказа**н** РНКазн**ая**», ««HEPN домен 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». Это, к сожалению, относится и к выводам.

Вывод № 3: «Обновлённая классификация CRISPR-Cas систем 2 класса **была предложена, которая включает** 6 новых обнаруженных систем»

Вывод № 4: «Исчерпывающая оценка разнообразия была сделана для CRISPR-Cas систем 2 класса, **показывающая** распространение»

Вывод № 6: «Были предложены возможные биотехнологические применения, включая редактирование геномов, **для новых CRISPR-Cas систем**».

Еще одно общее замечание, являющееся на самом деле следствием серьезного научного уровня и высокой актуальности данной работы, заключается в том, что автор не достаточно четко разделил свою часть и последующие работы, его цитирующие. Из-за этого я только ближе к концу текста окончательно понял, что экспериментальная проверка новых систем является не куском данной работы, а ее следствием, и была выполнена другими исследователями.

Среди частных вопросов и замечаний можно выделить следующие основные:

Стр. 6. «*...но цепи разрезаются в разных местах*». Что это значит? создаются липкие концы или она режет вообще в стороне от сайта таргетирования?

Стр. 27. «*...происхождение было найдено по гомологии...*». Предсказано, все-таки, по сходству. А происхождение по гомологии - это масло масляное т.к. гомология - это сходство обусловленное происхождением от общего предка.

Стр. 35. Раздел Кластеризация и филогенетический анализ, 1 предложение. Здесь слова "идентичность" и "схожесть" обозначают одно и то же (просто в разных алгоритмах

априори используются разные слова и автор просто скопировал это) или есть смысловая нагрузка в их различии? Для амк последовательности similarity может включать в себя не идентичные, но однотипные амк, например, +заряженные, или ароматические.

Стр. 35. Там же 2е предложение. «*Сайты с частотой пропуска > 0.5 и гомогенностью < 0.1*". Частота пропуска - это количество гапов на позицию? не маловато ли? часто позиции с гапами вообще выбрасывают либо делают отсев по 0,95. Хотя тут конечно слишком много сиквенсов и мало консервативности. Гомогенность – это количество идентичных аминокислот на позицию?

Стр. 35. Там же 3е предложение. *«...с 20 категориями»*. А зачем так много? Действительно ли ожидается, что столько позиций в одном белке будут сильно различно эволюционировать? Я, откровенно говоря, ни разу не видел деревьев по одному белку, где было бы больше 8, а обычно вообще 4. Нет, хуже, наверное, не будет (не считая времени обсчета), но мне не понятно - это специально так сделано или, грубо говоря, столько было по дефолту - столько и оставили?

Стр. 42. Подпись к Рис. 8. «*...помечены голубыми кругами (для тех систем, которые были обнаружены по ассоциации с cas1) и красными кругами (системы, обнаруженные по ассоциации с CRISPR кассетами)*». А что - перекрытия (и красный и синий кружочек одновременно) не получилось? WGS база использовалась и там и там, т.е. теоретически одни и те же кассеты могли найтись при помощи обоих запросов-затравок.

Стр. 42. Там же. «*...инактивация RuvC подобного нуклеазного домена обозначена перекрестьем*». То же на стр. 56 «*каталитически не активный гомолог*». А как определили то, что он неактивный? Свдиг рамки? Проблемы с каталитическими мотивами?

Стр. 44. Рис. 9. «*Доменная архитектура CRISPR-Cas эффекторных белков 2 класса систем*». А не пробовали сравнить 3D структуры? Возможно, обнаружатся какие-нибудь закономерности?

Стр. 47. «*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]*». Оборот «во время» несколько путает, лучше «позднее определенная другими авторами». На рисунке 9 нет никаких кристаллических структур – только доменная огранизация.

Стр. 49. Подпись к Рис. 11. «*Плохо выравненные последовательности между хорошо выравненными блоками, показаны числами*». Что эти числа обозначают? количество аминокислот? тогда так и надо было написать: неконсервативные участки показаны числами, которые соответствуют количеству аминокислот в данном участке.

Стр. 50. «*…позитивно заряженную, длинную α-спираль являющуюся двойником спирального мостика…*». Что значит двойником? Аналогом?

Стр. 50. «*Таким образом предсказывается, что, схожие с 2 классом эффекторов, TnpB белки связываются с РНК.*». Не понял почему. Во-первых, не вижу в данном абзаце доказательств связывания с РНК. Во-вторых, не вижу написанного логического объяснения, зачем ей связываться с РНК - У Cas9, Cfp1, C2c1 - у них у всех мишени - ДНК. У TnpB по определению мишень - ДНК, зачем ей связываться с РНК совсем не понятно. Если же речь о crРНК, то это а) надо указать, б) доказать/объяснить и в) объяснить/предположить зачем это TnpB как нуклеазе транспозона.

Стр. 55. Подпись к Рис. 12. «*Прерывистая линия обозначает произвольную отсечку ~2 (в размерности единиц дистанции, показанной линией масштаба ниже дерева), которая, эмпирически*». Почему произвольную если эмпирически подобрана и как эмпирически?

Стр. 57. «*глютамат, встроенный в хорошо предсказанную длинную α-спираль и соответствующий схожим мотивам HEPN доменов*». Что значит "хорошо предсказанную"? Как это "глутамат <...> соответствующий мотивам HREPN? Аминокислота соответствует мотивам? Или в каждом из этих мотивов как раз должен быть глутамат?

Стр. 59. «*…которые содержат предсказанные трансмембранные домены; в случае VI-B1 белок кодирует 4 таких домена и в VI-B2 кодируется только один*». Это не домены, а трансмембранные регионы/спирали. В случае V1-B1 возможно весь участок, включающий 4 ТМ региона, является трансмембранным доменом. В V1-B2 - точно нет - там есть HEPN домен и N-концевой трансмембранный регион. Кстати, важно, что TMHMM часто ложно предсказывает N-концевые ТМ-регионы, которые на самом деле свидетельствуют о наличии сигнального пептида. Не думаю, что здесь такой случай т.к. не понятно зачем бы этому белку быть внеклеточным или периплазматическим, но уточнить, возможно, стоило бы. Я попытался сделать это сам, используя номера, указанные в выравнивании на рисунке S6, но ни в генбанке, ни в юнипроте по таким query ничего не нашлось. Попробовал сделать BLAST по куску верхнего сиквенса из выравнивания S6, получил несколько хитов, но попытки скачать их последовательности не увенчались успехом.

Стр. 59. «*…что VI-B1 и VI-B2 могут регулировать РНК интерференцию (VI-B1 подавляет и VI-B2 улучшает)*». В смысле? Выше сказано, что весь VI-B обладает РНКазной активностью, и эта активность является конечным пунктом механизма интерференции, как тогда VI-B1 может подавлять интерференцию? Или это в случае если они работают в тандеме с другими системами (т.к. у VI-B нет cas1, они по идее должны полагаться на

другие системы) и соответственно подавляют их интерференцию? Хотелось бы более четкого разъяснения этого момента подавления и улучшения.

Стр. 66. «*...что Cas1 белки из различных подтипов II типа и V типа гомологичны различным подтипам I типа*». А не все cas1 белки гомологичны? Тут, наверное, имеется в виду то, что некоторые Cas1 I типа филогенетически ближе Cas1 II и V типов? Судя по рисунку 15. V-B и V-C тип произошел от различных I, V-A от III, при этом III сами произошли в р-те нескольких гор. переносов из I. В то же время II наследовался вертикально, а вот VI уже перенесся из него горизонтально (и возможно от III тоже, судя по цвету, хотя там не подписано, что эти 2 сиквенса из VI системы).

Стр. 68. Подпись к Рис. 15. «...множественное выравнивание 1498 Cas1 *последовательностей, которые содержат 304 <...> позиций*». Должно быть "которое", потому что выравнивание содержит 304 информативных позиции.

Стр. 68. «*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*». Поскольку дерево безкорневое, то и утверждение о том, что TnpB имеет предковый статус тоже, по-моему некорректно. Налицо множественные гор. переносы, но без достоверного корня и при том, что гор. переносы происходили в глубоких ветках мы не будем знать наверняка кто от кого произошел.

Стр. 72. «... *для специфичной геномной интеграции...*» Что такое геномная интеграция? имеется в виду вставка каких-то кусков в геном?

Стр. 73-74. «*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*». Из текста не понятно, что автор имеет в виду говоря «*исчерпывающе нашёл варианты CRISPR-Cas систем*» - представленность или разнообразие? Скорее всего речь идет о том, что новые варианты могут быть найдены (т.е. разнообразие еще не исчерпано), но их представленность будет крайне мала. Тогда «исчерпывающе» скорее будет относиться к представленности. В общем, это надо было бы написать четче.

Все вышеперечисленные вопросы и замечания не снижают научного значения данного исследования.

Заключение

Нет никаких сомнений, что диссертационная работа Шмакова С.А. на тему «Разработка биоинформатического подхода для поиска новых CRISPR-Cas систем» является актуальной, и содержит новые научные результаты. Диссертация представляет собой научно-квалификационную работу, в которой содержится решение научной задачи, имеющей значение для развития не только биоинформатики, но и многих других областей биологии, таких как микробиология, эволюционная биология, молекулярная биология, а также медицины и биотехнологии. Диссертация соответствует требованиям, изложенным в действующем «Положении о присуждении ученых степеней» (Постановление Правительства Российской Федерации №842 от 24.09.2013), а ее автор, Шмаков Сергей Анатольевич, заслуживает присуждения ученой степени кандидата биологических наук по специальности 03.01.09 – математическая биология, биоинформатика.

Официальный оппонент,
заведующий лабораторией метаболизма
экстремофильных прокариот,
ФИЦ Биотехнологии РАН
кандидат биологических наук Илья Валерьевич Кубланов

26 сентября 2017 г.

Федеральное государственное учреждение «Федеральный исследовательский центр «Фундаментальные основы биотехнологии» Российской академии наук» (ФИЦ Биотехнологии РАН)
119071, г. Москва, Ленинский проспект, д. 33, стр. 2,
тел. +7 (495) 954-52-83, факс +7 (495) 954-27-32.
e-mail: kublanov.ilya@gmail.com