



# БИОМИКА/BIOMICS

<http://biomics.ru>



## БИОИНФОРМАТИЧЕСКИЕ РЕСУРСЫ ДЛЯ ПОИСКА *IN SILICO* CRISPR-ЛОКУСОВ В ГЕНОМАХ ПРОКАРИОТ

Баймиев Ан.Х.<sup>1</sup>, Чемерис Д.А.<sup>1</sup>, Кирьянова О.Ю.<sup>2,3</sup>, Матниязов Р.Т.<sup>1</sup>,  
Валеев А.Ш.<sup>1</sup>, Баймиев Ал.Х.<sup>1</sup>, Губайдуллин И.М.<sup>2,3</sup>, Чемерис А.В.<sup>1</sup>

<sup>1</sup>Институт биохимии и генетики Уфимского научного центра Российской академии наук, Уфа, chemeris@anrb.ru

<sup>2</sup>Институт нефтехимии и катализа Российской академии наук, Уфа

<sup>3</sup>Уфимский государственный нефтяной технический университет, Уфа

### Резюме

Дана краткая характеристика CRISPR-локусов, встречающихся приблизительно у половины бактерий и у большинства архей. Показана их типичная организация, важным элементом которой служат CRISPR-кассеты, содержащие уникальные спейсеры, перемежающиеся одинаковыми прямыми повторами. Кратко рассмотрены специализированные программы поиска CRISPR-кассет в секвенированных геномах микроорганизмов и в метагеномных данных путем выявления в них повторяющихся участков. Приведены актуальные web-страницы таких программ и в табличной форме указаны их предназначения и возможности. Отмечены базы данных по CRISPR-локусам с указанием их web-адресов. Проведен анализ практически всей имеющейся литературы по данному вопросу и соответствующие интернет-ресурсы.

**Ключевые слова:** CRISPR/Cas система, CRISPR-локус, CRISPR-кассеты, спейсер, квазитандемные повторы, компьютерная программа, web-ресурс, база данных

### Содержание

	Стр.
Введение	229
Организация CRISPR локусов в геномах прокариот	230
Выявление CRISPR локусов с помощью программ поиска повторяющихся элементов геномов	232
Выявление CRISPR локусов с помощью специализированных программ	232
Выявление CRISPR локусов с помощью специализированных программ в метагеномных данных	236
Алгоритмы специализированных программ	236
Базы данных по CRISPR локусам	238
Заключение	241
Благодарности	241
Литература	242

### Введение

Историю изучения CRISPR-локусов бактерий принято отсчитывать с 1987 г.<sup>1</sup>, когда японскими исследователями у кишечной палочки *E.coli* рядом с геном *iap* были обнаружены некие

особые последовательности нуклеотидов [Ishino et al., 1987]. В 1989 г. эта же группа авторов опубликовала статью, в заголовке которой уже звучали слова о необычной организации повторяющихся последовательностей у *E.coli* («Unusual nucleotide arrangement with repeated sequences...»), причем подобные повторы блот-гибридизацией были обнаружены ими тогда еще и у двух других видов энтеробактерий [Nakata et al., 1989]. Свое продолжение история изучения CRISPR-

<sup>1</sup> Которые как CRISPR тогда еще не называли, поскольку они получили это наименование только в 2002 г. [Jansen et al., 2002].

локусов получила в Испании, где сначала в геноме археи *Haloferox mediterranei* были обнаружены почти совершенные повторы длиной 30 п.н., разделенные спейсерами сходного размера [Mojica et al., 1993], а затем похожие на них элементы генома и у другого вида архей *H. volcanii* [Mojica et al., 1995]. Последняя работа фактически положила начало *in silico* анализу CRISPR-локусов. Был сделан вывод, что такие последовательности весьма важны для прокариот, поскольку имеются как у архей, так и у бактерий, характеризуюсь при этом общей структурой, хотя и отличаясь по последовательности. Продолжая анализировать становящиеся известными полные геномы различных микроорганизмов, F.Mojica с соавт. обнаружили, что подобные повторы имеются у большого числа видов микроорганизмов, однако их функция оставалась неизвестной [Mojica et al., 2000]. В те годы для поиска в геномах разнообразных повторяющихся элементов использовались различные пакеты прикладных программ (**PC/Gene**, **DNASIS** и др.), с помощью которых могло производиться построение графов точечных матриц, позволяющих выявлять совпадающие (повторяющиеся) мотивы в анализируемых нуклеотидных последовательностях. Так, в статье 1995 г. [Mojica et al., 1995] для поиска таких повторов говорится, например, о применении программного обеспечения **PC/Gene** фирмы *Intelligenetics, Inc.* (США), а в статьях 2000 г. и 2005 г. [Mojica et al., 2000; 2005] уже упоминается специальная компьютерная программа, но без приведения ее характеристик, а также отмечается использование ресурса **BLASTn**.

В 2002 г. была опубликована статья, в которой голландские авторы [Jansen et al., 2002], согласовав с F.Mojica, назвали такие элементы прокариотических геномов как **C**lustered **R**egularly **I**nterspaced **S**hort **P**alindromic **R**epeats (CRISPR) и этот акроним быстро стал общепринятым. В этой же статье было показано, что рядом с CRISPR повторами имеются еще так называемые CRISPR-ассоциированные или Cas (CRISPR associated) гены. Но функция этих CRISPR-локусов по-прежнему оставалась загадочной. В 2005 г. тремя группами авторов были высказаны предположения, что CRISPR-локусы могут служить прокариотам в качестве некоей иммунной системы [Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005]. К этому времени уже появилось немало специализированных программ поиска повторяющихся элементов геномов, которые стали применять и для обнаружения CRISPR-локусов, но прежде чем перейти к описанию таких программных продуктов необходимо уделить некоторое внимание организации этих элементов

прокариотических геномов у разных микроорганизмов.

Цель написания данной статьи заключалась в том, чтобы собрать в одной публикации сведения о существующих компьютерных программах поиска в геномных ДНК микроорганизмов CRISPR-локусов и о базах данных по таким элементам геномов с указанием их актуальных web-страниц, поскольку подобные обзоры в мировой литературе отсутствуют. Есть немало обзорных работ<sup>2</sup>, посвященных дизайну гидРНК, причем в них упоминаются и некоторые программы поиска CRISPR-кассет, но приведенная там информация очень отрывочна. При этом обнаружение с помощью специализированных программ поиска в полногеномных или в метагеномных данных новых CRISPR/Cas систем у микроорганизмов в дополнение к двум десяткам известных представляет серьезный интерес, поскольку могут быть выявлены несколько иные типы устройства и функционирования этих оригинальных систем иммунной защиты у прокариот, способных оказаться не менее, если не более пригодными для целей как геномного редактирования, так и для других применений, что описано нами в еще одной статье данного выпуска журнала [Кулуев и др., 2017a].

#### Организация CRISPR локусов в геномах прокариот

CRISPR-локусы в геномах микроорганизмов устроены довольно сложно. Помимо самих CRISPR-повторов (CRISPR-кассет), представляющих собой непростые элементы, состоящие из двух компонентов – самих повторов, которые могут быть названы квазитандемными, и перемежающих их уникальных спейсеров, в типичный CRISPR-локус входят также ассоциированные с ними гены ряда Cas ферментов и другие участки, например, так называемая лидерная последовательность в виде АТ-богатого участка, содержащего промотор, обычно длиной от 100 до 500 п.н. [Jansen et al., 2002]. При этом было обнаружено, что если какой-либо микроорганизм содержит несколько CRISPR-локусов, то в случае присутствия в них одинаковых повторов, Cas-гены могут иметься только у одного такого локуса, который формально (условно) можно считать полноценным или основным. Другие локусы (точнее CRISPR-касеты) для своего функционирования пользуются ферментами, нарабатываемыми с генов, расположенных в другом месте генома. Однако если разные CRISPR-локусы несут в своем составе разные типы тандемных

<sup>2</sup> Отмеченных нами в предыдущей статье этого выпуска журнала [Чемерис и др., 2017].

повторов, разделяющих уникальные спейсеры, то тогда им, как правило, требуются «свои» Cas-гены, расположенные неподалеку.

CRISPR-локусы присущи почти половине всех видов бактерий и подавляющему большинству архей. Причем около половины несущих CRISPR-локусы микроорганизмов имеют в геномах их множественные копии, число которых, например, у архей *Methanocaldococcus jannaschii* достигает 18 [Grissa et al., 2007]. Другой характеристикой CRISPR-локусов являются обнаруживаемые в одной CRISPR-кассете уникальные спейсеры, перемежаемые одинаковыми тандемными повторами, причем количество этих элементов может различаться в довольно больших пределах – от единичных копий до очень больших величин. Так, например, у микроорганизма *Haliangium ochraceum* DSM 14365 в одной CRISPR-кассете таких выявлено 588 штук [Drevet, Pourcel, 2012]. Впрочем, и размер генома у этой океанической бактерии тоже весьма внушителен – около 10 млн.п.н. [Ivanova et al., 2010]. Еще более крупная величина в 1371 повтор в составе одной CRISPR-касеты без указания вида бактерии, которой он принадлежит, упоминается в другой работе [Lange et al., 2013].

Обратившись к базе данных **CRISPR-Exposed** (о которой еще пойдет речь), мы обнаружили в ней имеющуюся информацию по анализу генома *Haliangium ochraceum* DSM 14365, подтвердившую, что таких повторов/спейсеров в одной из CRISPR-кассет – 588 [Drevet, Pourcel, 2012], при этом в остальных восьми найдены следующие количества таких элементов – 5; 3; 8; 189; 36; 6; 2 и 15. Самые большие CRISPR-касеты (из 588, 189 и 36 элементов) имеют одинаковый тип повторов – GTTCCACGACCCAACAGTCGTGGCCTCATTGAA GC и расположены неподалеку друг от друга, на расстояниях 3 – 4 т.п.н., тогда как между остальными CRISPR-касетами у этого вида бактерий – миллионы пар нуклеотидов и никакой гомологии в прямых повторах. Забегая вперед, скажем, что информация по CRISPR-локусам этой бактерии представлена и в других базах данных, что дало возможность ее сопоставить, но об этом будет говориться ниже.

Цифровые данные о содержании в известных геномах микроорганизмов CRISPR-локусов по

состоянию на 09.05.2017 г. (дата последнего обновления данных) приведены на специальном сервере **CRISPRs web server** (<http://crispr.i2bc.paris-saclay.fr>), где указано, что из 232 проанализированных геномов архей у 202 из них (87%) CRISPR-локусы имеются, причем таковых выявлено 870. Для бактерий указано, что из 6782 секвенированных геномов CRISPR-локусы имеются у 3059 (45%), а их общее количество составляет 8069. Таким образом, в среднем одна архея, имеющая CRISPR/Cas систему иммунной защиты, несет около 4,3 CRISPR-локусов/касеты. А одна бактерия с такой же иммунной защитой – немного меньше – приблизительно 2,5 CRISPR-локусов/касеты. Но это в среднем, а как уже отмечалось выше, среди микроорганизмов имеются свои «рекордсмены». Здесь можно также привести сведения из другой базы данных **CRISPI** (<http://crispi.genouest.org>), последнее обновление которой сделано 21 марта 2017 г., где сообщается, что из включенных в их базу данных 137 видов архей (секвенированных геномов – 168) в 113 видах имеются CRISPR-локусы, что составляет около 82%, из 1259 проанализированных видов бактерий (2644 геномов) такие локусы содержат 753 вида или 60%.

По данным, составленным на основе анализов CRISPR-локусов у более чем 2500 микроорганизмов, оказывается, что у архей в CRISPR-касетах число повторов варьирует от 3 до 190 со средним количеством, составляющим 15 штук, а у бактерий эти показатели составляют – от 3 до 1371 повтора (в среднем 12 повторов на CRISPR-касету) [Lange et al., 2013]. Также в этой работе приводятся размеры спейсеров и прямых повторов у этих групп микроорганизмов. Так, для архей типичны длины повторов от 20 до 44 п.н. (в среднем 29 п.н.), а размеры спейсеров – 20 – 50 п.н. со средней длиной 38 п.н. Для бактерий эти характеристики довольно близки, но все же несколько отличаются – прямые повторы имеют размеры от 19 до 48 п.н. (в среднем 30 п.н.), а спейсеры могут быть от 19 до 70 нуклеотидов со средней длиной равной 35 п.н. Упрощенная схема «полноценного» функционального CRISPR-локуса приведена на рис. 1.



Рис. 1. Упрощенная схема организации функционального CRISPR-локуса (*масштаб не соблюден*)

Сейчас CRISPR-локусы благодаря развитию полногеномного секвенирования стали обнаруживать, в том числе, и у неподдающихся культивированию микроорганизмов, поэтому легко представить, что установленное разнообразие таких CRISPR систем в прокариотическом мире еще вырастет. Но пока принято считать, что по особенностям организации CRISPR-локусов с Cas-генами они укладываются в 19 подтипов, относящихся к 6 типам и двум классам. Предпринятая ранее попытка классифицировать CRISPR-локусы по последовательностям, а также по вторичным структурам их прямых повторов, входящих в состав 561 CRISPR-кассеты, найденной у 195 видов микроорганизмов (44% от всех исследованных), позволила уложить их в 12 кластеров [Kunin et al., 2007]. Позже другими авторами на большем объеме исследуемого материала (279 геномов архей и 2289 геномов бактерий) были обнаружены 40 консервативных групп таких повторов [Lange et al., 2013].

Не имея в данной статье возможности рассматривать подробно все разнообразие CRISPR-локусов, включая ассоциированные с ними гены ряда необходимых для функционирования этой системы белков, считаем, что будет правильным отослать заинтересованных читателей к ряду обзорных статей на этот счет [Koonin et al., 2017; Barrangou, Horvath, 2017], включая нашу в данном номере журнала [Кулуев и др., 2017]. А здесь мы решили ограничиться этой краткой информацией и приведением общей организации CRISPR-локусов (рис. 1).

#### Выявление CRISPR локусов с помощью программ поиска повторяющихся элементов геномов

В уже упоминавшейся статье 2005 г. F. Mojica с соавторами в 67 штаммах, относящихся к 36 родам бактерий и архей, нашли около 4500 различных CRISPR спейсеров [Mojica et al., 2005]. С учетом высказанных в это же время предположений об участии этих элементов прокариотических геномов в организации иммунной защиты микроорганизмов интерес к CRISPR-локусам стал быстро расти в связи с чем возросла потребность в компьютерных программах поиска повторяющихся последовательностей в геномах [Ussery et al., 2004]. Поскольку специализированных программ именно для обнаружения CRISPR-локусов еще не было, то их выявление осуществляли с помощью программ поиска любых повторяющихся элементов, таких как **PatScan** [Dsouza et al., 1997], **Tandem Repeat Finder** [Benson, 1999], **REPuter** [Kurtz, Schleiermacher, 1999; Kurtz et al., 2001], **RepeatFinder** [Volfovsky et al., 2001], **RECON** [Bao, Eddy, 2001], **FORRepeats** [Lefebvre et al., 2003], **RepeatGluer** [Pevzner et al.,

2004], **PILER** [Edgar, Myers, 2005], **RepeatScout** [Price et al., 2005], **PYGRAM** [Durand et al., 2006] **TRAP** [Sobreira et al., 2006] и некоторых других [Gusfield, Stoye, 2004], включая ресурс **BLAST**. Так, например, проведенный с помощью программы **PatScan** анализ 370 геномов бактерий и архей разных групп (мезофиллы, термофилы, аэробы, анаэробы, гетеротрофы, фототрофы и др.) позволил авторам сделать вывод о горизонтальном переносе этих локусов между микроорганизмами [Godde, Bickerton, 2006].

Однако все эти программы не были нацелены непосредственно на поиск повторяющихся элементов в CRISPR локусах и с легкостью могли их пропустить.

#### Выявление CRISPR локусов с помощью специализированных программ

В 2007 г. можно сказать произошел некий прорыв в написании компьютерных программ, ориентированных именно на поиск CRISPR-локусов, поскольку таковых появилось сразу три. Первой из них стала специализированная версия известной программы **PILER** [Edgar, Myers, 2005], получившая название **PILER-CR (CRISPR Repeats)** [Edgar, 2007]. Причем на анализ прокариотического генома размером около 5 млн.п.н. на предмет наличия в нем CRISPR-кассет сообщалось, что программе **PILER-CR** было достаточно всего 5 секунд работы обычного настольного компьютера (346 прокариотических геномов были проанализированы за 15 минут с помощью процессора с тактовой частотой 2 ГГц). Версия **1.06** этой программы, являющейся частью семейства программ **PILER**, может быть загружена со страницы <http://drive5.com/pilercr/><sup>3</sup>. Там же находится образец выдаваемого отчета при поиске CRISPR-локусов. Для обнаружения CRISPR-кассет данной программой по умолчанию устанавливаются минимальные / максимальные длины повторов и спейсеров, составляющие 16/64 и 8/64 нуклеотида соответственно. В этих пределах пользователям допускается их менять. Минимально допустимое число повторов (включая спейсеры) в одной CRISPR-кассете для поиска равно трем. При этом отмечается, что если сделать его равное двум, то возрастает вероятность получения ложно-положительных результатов, а при увеличении их минимального числа появляется риск пропуска коротких блоков таких повторов, что будет являться уже ложно-

<sup>3</sup> Работающего в сети Интернет оригинального варианта самой программы **PILER-CR** не предусмотрено, однако в некоторых базах данных она используется в составе других программ для поиска CRISPR-кассет в режиме on-line.

негативным результатом. Авторы отмечают, что с их установленными по умолчанию параметрами чувствительность **PILER-CR** составляет 100%, а специфичность также весьма высока и достигает 94%.

Второй такой программой в 2007 г. стала **CRISPR Recognition Tool (CRT)**, ведущая анализ геномов путем сравнения так называемых *k-mer* мотивов [Bland et al., 2007]. Она может быть скачана с сайта <http://www.room220.com/crt/><sup>4</sup>. Авторы провели сравнение ее возможностей с программой **PILER-CR**, а также с программой **PatScan** и обнаружили, что программа **CRT** по скорости работы сравнима с **PILER-CR**, превосходя ее по некоторым параметрам, а перед **PatScan** имеет огромные преимущества, ввиду того, что последняя требует значительных усилий по ручной обработке получаемых результатов. С установленными минимальными и максимальными пороговыми значениями длин повторов и спейсеров в 19/50 и 19/60 нуклеотидов соответственно при минимально допустимом числе повторов со спейсерами равном трем чувствительность и специфичность **CRT** для обоих показателей для отдельных микроорганизмов составила 99%.

Здесь можно заметить, что в США в Department of Energy Joint Genome Institute при описании геномов микроорганизмов в качестве стандарта для выявления CRISPR-локусов принято пользоваться программами **CRT** и **PILER-CR** [Huntmann et al., 2015].

Третьей программой поиска CRISPR-локусов, написанной в 2007 г., стала **CRISPRfinder** [Grissa et al., 2007a]. Данный web-ресурс находится по адресу <http://crispr.i2bc.paris-saclay.fr/Server/>, являющемуся частью **CRISPRs web server** (<http://crispr.i2bc.paris-saclay.fr>), где помимо программы **CRISPRfinder** представлены и другие web-ресурсы (**CRISPRsdb**, **CRISPRcompar**, **CRISPRtionary**, **FlankAlign**), о которых будет говориться ниже. Программа **CRISPRfinder** имеет весьма удобный интерфейс, позволяющий пользователю легко устанавливать минимальные и максимальные размеры повторов и спейсеров. По умолчанию они составляют для повторов от 23 до 55 нуклеотидов, а для спейсеров имеют уменьшающиеся и увеличивающиеся их размер коэффициенты в пределах от 0,6 до 2,5. Для повторов допускается установить одно допустимое неспаривание. Имеется и ряд

других опций. После этого в соответствующее окно помещаются последовательности ДНК в FASTA формате (максимально длиной до 67 млн. нуклеотидов) или загружаются соответствующие файлы, после чего запускается поиск. Допускается использование вместо неизвестных нуклеотидов символа «N», однако прочие обозначения одновременно нескольких нуклеотидов типа «Y», «R» для пиримидинов и пуринов и им подобные автоматически превращаются в «N» и рассматриваются как неспаривания. Причем на первом этапе поиска CRISPR-локусов при выявлении повторяющихся элементов генома работает программа **Vmatch** (<http://www.vmatch.de>), являющаяся усовершенствованной версией программы **REPuter**. Важным отличием **CRISPRfinder** от других программ поиска CRISPR-локусов является то, что она способна обнаружить состоящие всего из одного-двух мотивов CRISPR спейсеров и повторов. Еще одним важным моментом является предоставление информации о фланкирующих последовательностях, что позволяет устанавливать лидерный участок. Данная программа самостоятельно запускает поиск выявленных CRISPR спейсеров по **GenBank** с помощью программы **BLAST**, а также проверяет присутствие в других прокариотических геномах похожих прямых повторов. Выдаваемые **CRISPRfinder** результаты поиска и проведенного анализа оформляются в виде таблицы, сопровождаемые цветным графическим отображением спейсеров (в широкой цветовой гамме) и прямых повторов, которые всегда одинаково маркируются желтым цветом. Образец проведенного поиска CRISPR-локусов по геному *Aquifex aeolicus* VF5 представлен на странице [http://crispr.i2bc.paris-saclay.fr/crispr/HelpTopics/examples\\_CRISPRdatabase.html](http://crispr.i2bc.paris-saclay.fr/crispr/HelpTopics/examples_CRISPRdatabase.html). На основе программы **CRISPRfinder** подготовлено руководство для поиска CRISPR-кассет в секвенированных нуклеотидных последовательностях [Drever, Pourcel, 2012].

Вышеупомянутые программы только находят CRISPR-локусы, не устанавливая цепь ДНК, кодирующую некодирующую РНК. Web-ресурс **CRISPRDetect** / **CRISPRDirection** ([http://bioanalysis.otago.ac.nz/CRISPRDetect/predict\\_crispr\\_array.html](http://bioanalysis.otago.ac.nz/CRISPRDetect/predict_crispr_array.html)) [Biswas et al., 2014; 2016] сначала с помощью встроенных программ **CRT**, **PILER-CR** находит CRISPR-кассеты (причем ряд опций повышает вероятность их обнаружения), удаляя в том числе обычные повторяющиеся элементы генома, а затем с помощью дополнительного анализа фланкирующих последовательностей с достоверностью в 94% устанавливается точное направление транскрипции и определяется какой

<sup>4</sup> Работающего в сети Интернет оригинального варианта самой программы **CRT** (как и **PILER-CR**) не предусмотрено, однако в некоторых базах данных она используется в составе других программ для поиска CRISPR-кассет в режиме on-line.

может быть спейсерная РНК. Последовательность АТТГААН около 3'-конца некоторых<sup>5</sup> прямых повторов также указывает направление транскрипции, что важно знать при поиске мишеней (протоспейсеров) для CRISPR/Cas-систем, для чего этими же авторами написана программа **CRISPRTarget**, куда могут экспортироваться полученные данные и о которой будет говориться ниже. Вслед за программой **CRISPRDirection** другими авторами была разработана программа **CRISPRstrand** (<http://rna.informatik.uni-freiburg.de/CRISPRmap/>), которая также позволяет устанавливать направление транскрипции [Alkhnabashi et al., 2014]. Направление транскрипции важно также знать для нахождения лидерной последовательности, для чего этими же авторами была разработана специальная программа **CRISPRleader** [Alkhnabashi et al. 2016].

Недавно разработана еще одна компьютерная программа поиска CRISPR-локусов **CRISPRdigger** [Ge et al., 2016]. Она может быть загружена с сайта <http://www.healthinformatics.org/supp/resources.php>. Особенностью этой программы является то, что для поиска повторов в геноме она использует другую программу **RepeatScout** [Price et al., 2005], а затем за «дело» принимается программа **RepeatMasker** [<http://repeatmasker.org>]. Вводимые последовательности должны иметь FASTA формат. Проведенное разработчиками **CRISPRdigger** сравнение эффективности поиска CRISPR повторов с помощью четырех программ (**CRISPRdigger**, **CRISPRfinder**, **PILER-CR** и **CRT**) в том числе только секвенированных геномов микроорганизмов показало, что последние две программы уступают первым двум, среди которых в ряде случаев преимущество имела программа **CRISPRfinder**, а в других случаях – **CRISPRdigger**. При сравнении затрачиваемого этими программами времени на поиск CRISPR-локусов в геномах нескольких микроорганизмов (размерами от 2,207 до 4,777 млн.п.н.) программы **CRISPRfinder**, **PILER-CR** и **CRT** затрачивали от 1 до 7 секунд, тогда как программе **CRISPRdigger** требовалось от двух с половиною минут до почти четырех. Однако считать это серьезным недостатком данной программы вряд ли стоит.

Недавно внимание CRISPR/Cas сообщества исследователей было обращено на факт ложного выявления CRISPR-локусов в геномах многих бактерий и архей [Zhang, Ye, 2017]. Этими авторами

был подготовлен специальный web-ресурс **CRISPRone** (<http://omics.informatics.indiana.edu/CRISPRone/>), позволяющий заново искать в полных геномах любых прокариотических организмов CRISPR-локусы, отсекая «false-CRISPR» элементы, что достигается подключением программы **FragGeneScan**, ищущей близости от CRISPR повторов Cas-гены. В данной работе было продемонстрировано, что ложные CRISPR-повторы можно отнести к четырем категориям – обычные тандемные повторы; простые повторы; STAR-подобные (*Staphylococcus aureus*) элементы; прочие, которые пока невозможно отнести к каким-либо конкретно.

Несколько обособленно выглядит программа **CRISPRTarget** ([http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr\\_analysis.html](http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html)) [Biswas et al., 2013; 2015], нацеленная на поиск мишеней (протоспейсеров) для которых микроорганизмы и создают свои CRISPR-локусы. Можно сказать, что это также программа выявления CRISPR-локусов, причем делает это она в режиме online, используя, в том числе программы **CRT**, **PILER-CR**, не имеющие собственных интернет-версий, а также еще одну популярную программу поиска **CRISPRfinder**. После обнаружения потенциальных CRISPR-кассет с уникальными спейсерами запускается поиск по одной из предлагаемых на выбор нескольких **RefSeq** баз данных с последовательностями вирусов, плазмид и др. Важной особенностью выдаваемых программой **CRISPRTarget** результатов служат найденные PAM (Protospacer Adjacent Motif) последовательности в проанализированных геномах и фланкирующие найденные протоспейсеры участки геномов. Причем программа **CRISPRTarget** берет в анализ обе цепи ДНК, если не известно направление CRISPR-локусов, которое способна предсказать уже упоминавшаяся программа этих же авторов **CRISPRDirection** [Biswas et al., 2014].

Недавно китайскими авторами опубликована статья [Mai et al., 2016], в которой они сообщают, что, по их мнению, программный поиск CRISPR-локусов в геномах микроорганизмов с помощью программ **CRT**, **PILER-CR**, **CRISPRfinder** должен сопровождаться ручной «доводкой» результатов и приводят пример, что из обнаруженных в восьми полностью секвенированных геномах бактерий рода *Thermoanaerobacter* из 95 CRISPR-кассет 59 представляют собой участки геномов, нарушенные транспозонами.

Все вышеупомянутые программы ориентированы на поиск прямых повторов и спейсеров в составе CRISPR-локусов, оставляя Cas гены практически без внимания за исключением web-ресурса **CRISPRone**, но данная задача для этой

<sup>5</sup> Последовательность АТТГААН обнаружена в прямых повторах у 1070 CRISPR-кассет [Biswas et al., 2016].

программы является дополнительной. Не так давно была разработана компьютерная программа **MacSyFinder** (**Macromolecular System Finder**), нацеленная именно на обнаружение Cas белков

[Abby et al., 2014]. Эта программа не имеет web-сервера и может быть скачана с этого сайта <https://github.com/gem-pasteur/macsyfinder>.

Таблица 1.

Программы для обнаружения CRISPR-локусов в геномах прокариот

Название программного ресурса	URL-адрес	Функциональные возможности	Тип поддерживаемой ОС	Тип приложения
<b>PILER / PILER-CR</b>	<a href="http://drive5.com/pilercr/">http://drive5.com/pilercr/</a>	<ul style="list-style-type: none"> <li>Идентификация и анализ CRISPR-кассет</li> </ul>	Windows / Linux	Работа при установке на ПК
<b>CRISPR Recognition Tool</b>	<a href="http://www.rooim220.com/crt/">http://www.rooim220.com/crt/</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> <li>Анализ геномов (сравнение <i>k-mer</i> мотивов)</li> </ul>	Кроссплатформенное приложение	Работа при установке на ПК
<b>CRISPRfinder</b>	<a href="http://crispr.i2bc.paris-saclay.fr/Server/">http://crispr.i2bc.paris-saclay.fr/Server/</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> <li>Предоставление информации о фланкирующих последовательностях</li> <li>Проверка присутствия в других прокариотических геномах похожих прямых повторов</li> </ul>	-	Web-ресурс
<b>Vmatch</b>	<a href="http://www.vmatch.de">http://www.vmatch.de</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> </ul>	-	Web-ресурс
<b>CRISPRDetect / CRISPRDirection</b>	<a href="http://bioanalysis.otago.ac.nz/CRISPRDetect/predict_crispr_array.html">http://bioanalysis.otago.ac.nz/CRISPRDetect/predict_crispr_array.html</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> <li>Определение направления транскрипции</li> </ul>	-	Web-ресурс
<b>CRISPRmap CRISPRstrand</b>	<a href="http://rna.informatik.uni-freiburg.de/CRISPRmap/">http://rna.informatik.uni-freiburg.de/CRISPRmap/</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> <li>Определение направления транскрипции</li> </ul>	-	Web-ресурс, возможна работа при установке на ПК
<b>CRISPRleader</b>	<a href="http://www.bioinf.uni-freiburg.de/Software/CRISPRleader/">http://www.bioinf.uni-freiburg.de/Software/CRISPRleader/</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> <li>Определение направления транскрипции</li> <li>Определение лидерной последовательности</li> </ul>	Linux	Работа при установке на ПК
<b>CRISPRdigger</b>	<a href="http://www.althinformatics.org/supp/resources.php">http://www.althinformatics.org/supp/resources.php</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-кассет</li> </ul>	Linux	Работа при установке на ПК
<b>CRISPRone</b>	<a href="http://omics.informatics.indiana.edu/CRISPRone/">http://omics.informatics.indiana.edu/CRISPRone/</a>	<ul style="list-style-type: none"> <li>Поиск в полных геномах любых прокариотических организмов CRISPR-локусов с учетом выявления ложных CRISPR-локусов</li> </ul>	-	Web-ресурс
<b>CRISPRTarget</b>	<a href="http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html">http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-локусов</li> <li>Поиск мишеней (протоспейсеров), для которых микроорганизмы и создают свои CRISPR-локусы</li> </ul>	-	Web-ресурс
<b>MacSyFinder</b>	<a href="https://github.com/gem-pasteur/macsyfinder">https://github.com/gem-pasteur/macsyfinder</a>	<ul style="list-style-type: none"> <li>Поиск CRISPR-локусов</li> <li>Поиск Cas белков</li> </ul>	Кроссплатформенное приложение	Работа при установке на ПК

В табл. 1 представлена краткая информация о вышеописанных программах, которая может быть полезна конечным пользователям при выборе программного приложения для проведения анализа геномов на наличие в них CRISPR-локусов.

#### Выявление CRISPR локусов с помощью специализированных программ в метагеномных данных

В связи с бурным развитием полногеномного секвенирования новых поколений вслед за геномами конкретных видов прокариот стали определять нуклеотидные последовательности и у популяций микроорганизмов, получая так называемые метагеномные данные, в которых обнаружение CRISPR-локусов также представляет значительный интерес, тем более, если принять во внимание, что около 95% микроорганизмов не поддаются культивированию в лабораторных условиях. Подобные работы проводятся уже довольно давно, производя поиск CRISPR-локусов в метагеномных данных, полученных из разных сред, например – из горячих источников в Йеллоустонском национальном парке [Heidelberg et al., 2008], океана (из 57 различных его участков) [Sorokin et al., 2010], из подводных гидротермальных источников в Тихом океане [Anderson et al., 2011], микробиома кишечника человека [Stern et al., 2012; Gogleva et al., 2014] и из других мест, однако в этих работах использовались уже известные программы поиска CRISPR-локусов – **CRT**, **PILER-CR** и **CRISPRfinder**. Недавно сообщено об исследовании метавирома<sup>6</sup>, в ходе которого было секвенировано более трех тысяч образцов из окружающей среды из разных мест Планеты, а также принадлежащих различным биологическим хозяевам [Paez-Espino et al., 2016]. При этом уделялось внимание и обнаружению CRISPR-локусов, проводившемуся с помощью модифицированной программы **CRT**, поскольку это дает возможность установить связь микроорганизмов с атакующими их вирусами, оставившими свои следы в виде CRISPR-спейсеров. В другой работе при исследовании микробиома ротовой полости человека путем метагеномного секвенирования для поиска CRISPR-локусов была применена программа **CRT**, которая была несколько изменена и получила название **metaCRT** (<https://omictools.com/metacrt-tool>), причем в паре с ней использовалась программа **CRISPRAlign** (<https://omictools.com/crispralign-tool>) [Rho et al.,

2012]. При этом обе программы тоже, как и сама **CRT**, требуют установки на компьютер пользователя.

Особенностью метагеномных данных является довольно ограниченная протяженность доступных для анализа нуклеотидных последовательностей. Поэтому для эффективного поиска среди CRISPR-локусов необходимо использование специально предназначенных для этого программ и таковая названная **Crass** (**The CRISPR Assembler**)<sup>7</sup> в 2013 г. появилась [Skennerton et al., 2013]. Она также требует установки на компьютер пользователя и может быть скачана с сайта <http://ctskennerton.github.io/crass/>. В опубликованной статье говорится, что эта программа предназначена для выявления в первичных данных после секвенирования shotgun фрагментов ДНК с помощью платформ Illumina, Ion Torrent, Roche 454, а также в результате секвенирования по Сэнгеру. Позже для поиска CRISPR-локусов в метагеномных данных были написаны еще две программы **CrisprDetector** [Ben-Bassat, Chor, 2015; 2016] и **metaCRISPR** [Lei, Sun, 2016], которые можно скачать с сайтов [http://www.cs.tau.ac.il/~bchor/CRISPR\\_JCB/](http://www.cs.tau.ac.il/~bchor/CRISPR_JCB/) и <https://github.com/hangelwen/metacrispr> соответственно.

Здесь также можно заметить, что в США в Department of Energy Joint Genome Institute при описании метагеномов микроорганизмов как и для описания геномов конкретных микроорганизмов (о чем уже упоминалось выше) в качестве стандарта для выявления CRISPR-локусов принято пользоваться программами **CRT** и **PILER-CR** [Huntemann et al., 2016].

#### Алгоритмы специализированных программ

Фактически работу специализированных программ поиска CRISPR-локусов объединяет общая задача – нахождение одинаковых повторяющихся подпоследовательностей в некоторой последовательности данных. В дальнейшем проводится анализ выявленных повторяющихся подпоследовательностей и формируется результат. Для решения данной задачи применяются эвристические алгоритмы и методы, применяемые в задачах поиска отдельных фрагментов массивов, разделения строк на подстроки. Так, в программе **ForRepeats** разработан собственный эвристический метод, названный FORRepeats, основанный на конечном автомате (так называемый «фактор оракула» [Lefebvre et al., 2003]). На первом этапе происходит обнаружение точных повторов в

<sup>6</sup> Метавиrom – совокупность нуклеотидных последовательностей различных вирусов, населяющих Землю, в виде относительно небольших фрагментов нуклеиновых кислот.

<sup>7</sup> Не путать с программой crAss (**Cross-Assembly of Metagenomes**) [Dutilh et al., 2012], ставящей задачей анализ и сборку метагеномных данных прокариот.

большой последовательности. Затем, на втором этапе вычисляются приблизительные повторы и выполняется парное сравнение.

Некоторые методы (более ранние программы, такие как **Tandem Repeat Finder**, **RepeatFinder**) основаны на так называемом методе тандемных повторов (Model of Tandem Repeats). В этих программах задается ряд статистических критериев, согласно которым происходит анализ последовательности (максимальная длина повторяющегося участка, максимальный период повторения, параметры обнаружения). Критерий оценки подпоследовательностей основан на схеме Бернулли. Вероятность успешности события в этом случае означает средний процент идентичности между копиями. По умолчанию, вероятность успешности (совпадения копий) равна 80% (то есть 80% компонентов сравниваемых подпоследовательностей будут совпадать). Алгоритм ищет соответствующие нуклеотиды, разделенные общим расстоянием  $d$ , которое заранее не указано. Происходит поиск так называемых  $k$ -кортежей.  $K$ -кортеж представляет собой окно из  $k$  последовательных символов нуклеотидной последовательности. Соответствующие  $k$ -кортежи – это два окна с одинаковым содержанием. Основная операция обнаружения одинаковых участков выглядит следующим образом. Пусть  $S$  – нуклеотидная последовательность. Пользователь выбирает небольшое целое число  $k$  для определения размера кортежа (окна) (например,  $k = 5$ ). Программа сохраняет список всех возможных строк длины  $k$  (для алфавита ДНК {A, C, G, T}), которые условно можно назвать пробниками. Путем продвижения окна по последовательности определяется подстрока размера  $k$  в каждой позиции  $i$  в  $S$ . Для каждой подстроки сохраняется список истории позиций, в которых происходит совпадение. Когда позиция  $i$  добавляется, то просматриваются все ранние вхождения данной подстроки. Пусть одно раннее вхождение будет в позиции  $j$ . Так как  $i$  и  $j$  – индексы совпадающих  $k$ -кортежей, то расстояние  $d = i - j$  является возможным расстоянием для тандемного повтора. Поэтому для проверки также необходима информация о других совпадениях  $k$ -кортежей на том же расстоянии  $d$ , где ведущий кортеж встречается в последовательности между  $j$  и  $i$ . Список расстояний хранит эту информацию. Его можно рассматривать как скользящее окно длины  $d$ , которое отслеживает позиции совпадений и их общее количество. Список обновляется при появлении нового совпадения. Далее программа переходит к анализу длины расстояний между совпадающими участками, оценки вероятности совпадений, размера кортежа. Если информация в списке расстояний проходит анализ,

шаблон-кандидат, состоящий из положений  $j + 1 \dots i$ , выбирается из нуклеотидной последовательности и выравнивается с окружающей последовательностью. Если, по меньшей мере, две копии шаблона выровнены с последовательностью, сообщается о тандемном повторе. С более подробной информацией о данном методе можно ознакомиться по адресу <https://tandem.bu.edu/trf/trfdesc.html>.

Программа **PILER-CR** основывается на методах, разработанных в семействе алгоритмов **PILER** для анализа повторов. Основными параметрами алгоритмов являются: минимальная длина повтора (16)<sup>8</sup>, максимальная длина повтора (64), минимальная длина спейсера (8), максимальная длина спейсера (64), минимальное количество повторов в массиве (3). Алгоритм состоит из 8 шагов. На первом этапе происходит поиск локальных повторов при передвижении окна фиксированной длины по последовательности генома. Программа фиксирует координаты начала повтора. Далее происходит формирование «стопок» – повторяющихся участков, которые между собой разделены уникальными участками.

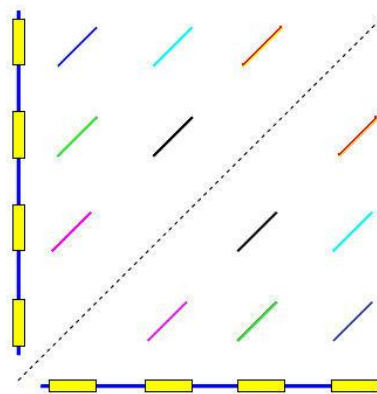


Рис. 2. Формирование графа автомодельности

На третьем шаге формируется граф автомодельности, схема которого представлена на рис. 2. Каждое совпадение соединяет две стопки. Используя совпадения как ребра и стопки в качестве узлов, создается граф, и компоненты графа идентифицируются.



Рис.3. Формирование первичного массива

<sup>8</sup> Длины указаны в нуклеотидах.

Далее формируется первичный массив, который в дальнейшем будет оптимизироваться. Каждый найденный совпадающий компонент является кандидатом для хранения одного или нескольких массивов CRISPR. Это значительно уменьшает пространство поиска для следующего шага.

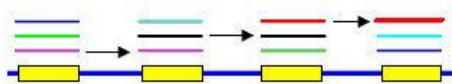


Рис. 4. Слияние смежных участков

Для объединения двух массивов должен выполняться следующий критерий: соотношения длины повторения и длины спейсеров для двух массивов должны быть  $\geq 0,95$ . Часто бывает, что повторы в разных массивах имеют схожий состав. Поэтому проводится кластеризация массивов и выравнивание последовательностей с использованием быстрых опций алгоритма MUSCLE [Edgar, 2004]. После чего программа формирует отчет о проведенном анализе.

Во многих программах применяются алгоритмы, основанные на понятии суффиксного дерева, а также суффиксного массива для более оптимальной и быстрой работы программы. Например, метод поиска в программе **Vmatch** основан на структуре суффиксного дерева [Abouelhoda et al., 2004].

На данный момент для многих разработчиков помимо улучшения качества анализа данных важным критерием работы программы является ее скорость. Поэтому разрабатываются современные алгоритмы, которые используют динамические массивы (массивы, размер которых меняется в процессе работы программы) для более быстрого доступа к информации, а также методы, которые позволяют хранить как можно меньше информации в процессе анализа последовательности.

#### Базы данных по CRISPR локусам

Подготовив для поиска CRISPR-локусов в геномах микроорганизмов web-ресурс **CRISPRfinder** [Grissa et al., 2007a], его авторы сразу же создали постоянно обновляемую специализированную базу данных **CRISPRsdb** [Grissa et al., 2007], разместив их на портале **CRISPRs web server** (<http://crispr.i2bc.paris-saclay.fr>) с локализацией во Франции, где имеются и другие web-ресурсы (**CRISPRcompar**, **CRISPRtionary**, **FlankAlign**). На данном портале размещено также подробное руководство для работы с базой данных, программами и остальными утилитами

([http://crispr.i2bc.paris-saclay.fr/crispr/HelpTopics/help\\_CRISPRdatabase.html](http://crispr.i2bc.paris-saclay.fr/crispr/HelpTopics/help_CRISPRdatabase.html)).

База данных **CRISPRsdb** по состоянию на 09.05.2017 г. (дата последнего обновления информации) содержит сведения о 7014 проанализированных прокариотических геномах (232 археи и 6782 бактерии), в которых найдено 8939 CRISPR-кассет. Для их просмотра необходимо выбрать конкретный геном того или иного микроорганизма из списка на данной странице <http://crispr.i2bc.paris-saclay.fr/crispr/>, которая несколько дольше грузится, поскольку на ней сразу предоставляется информация обо всех штаммах, содержащихся в этой базе данных. Причем все штаммы показаны на соответствующем фоне: розовый – CRISPR-кассеты выявлены, сероватый – CRISPR-кассеты имеются предположительно, желтоватый – CRISPR-кассеты у таких штаммов отсутствуют. Для получения более детальной информации о конкретном штамме следует кликнуть на его название, после чего откроется окно сначала с краткими сведениями, где необходимо поставить галочку в нужное окошко и запросить информацию обо всех CRISPR-локусах этого микроорганизма. В сводной таблице на розовом фоне будет дана информация о подтвержденных CRISPR-локусах, а на желтом фоне – будут приведены данные о предположительных CRISPR-локусах, если таковые в анализируемом геноме имеются. Ниже сервер приводит подробную информацию о каждом CRISPR повторе и спейсере, включая их нуклеотидные последовательности и места локализации в геноме, выделяя при этом их разными цветами, где желтым цветом всегда указаны прямые повторы. Там же можно запустить поиск **BLAST**, запросить для конкретной CRISPR-кассеты фланкирующие его последовательности. Имеются и другие возможности.

Помимо публичной базы данных **CRISPRsdb** каждому пользователю предоставляется возможность создания собственной базы данных **My CRISPRsdb**, в которой может храниться информация о CRISPR-кассетах, обнаруженных во введенных исследователем нуклеотидных последовательностях. Для организации такой базы данных требуется указать свой адрес электронной почты и придумать пароль. Форма хранения сведений по CRISPR-кассетам в этой частной базе данных аналогична таковой в базе данных **CRISPRsdb**. С помощью программы **CRISPRcompar** [Grissa et al., 2008] имеется возможность разностороннего сравнения данных по CRISPR-локусам в базе данных автора с хранящимися в основной базе данных. При выполнении опции сравнения спейсеров активируется утилита **CRISPRtionary**, создающая некое подобие словаря (dictionary) характерных

мотивов нуклеотидных последовательностей. Также после завершения сравнения различных CRISPR-кассет у разных штаммов с помощью утилиты **FlankAlign** можно увидеть выравненные фланкирующие нуклеотидные последовательности у проанализированных микроорганизмов.

В качестве примера хранимых сведений о CRISPR-локусах бактерий и архей в базе данных **CRISPRsdb** обратимся к уже упоминавшемуся микроорганизму *Haliangium ochraceum*, «богату» на CRISPR-касеты. Так, выведенная информация по этой бактерии (полученная с помощью программы **CRISPRfinder**) свидетельствует, что у ней имеется только три подтвержденных участка CRISPR-локусов с числом повторов – 587, 189 и 36 единиц, выделенных розовым фоном, а остальные 14 CRISPR-локусов с числом повторов от одного до четырех, выделенных желтым фоном, относятся к разряду неподтвержденных, что несколько отличается от данных по этому штамму, упомянутым нами выше из базы данных **CRISPR Exposed**.

Также уже довольно давно (с конца 2008 г.) во Франции существует интерактивная периодически обновляемая база данных **CRISPI** – **CRISPR Interactive database** (<http://crispi.genouest.org>) [Nicolas et al., 2008; Rousseau et al., 2009]. На момент последнего обновления (21.03.2017 г.) в **CRISPI** содержалась информация о 168 геномах архей, относящихся к 137 видам, с обнаруженными в них 531 CRISPR-локусе, и о 2644 геномах бактерий (1259 видов бактерий), несущих 2808 CRISPR-локусов. Для облегчения работы пользователя с базой данных **CRISPI** на сайте имеется подробное руководство, доступное в html или pdf виде, а также video инструкция.

На основной странице базы данных **CRISPI** пользователь должен выбрать одну из трех предоставляемых возможностей: 1) провести поиск известных геномов архей и бактерий, для которых уже имеется информация о содержащихся в них CRISPR-касетах, а также Cas-генах; 2) сопоставить имеющуюся у пользователя последовательность нуклеотидов с базой данных **CRISPI** на предмет гомологии с CRISPR-локусами; 3) идентифицировать CRISPR-касеты в геноме какого-либо микроорганизма, вводимого пользователем в соответствующее окно в FASTA формате или загружаемого в виде файла.

В первом случае геном интересующего микроорганизма может быть найден или по Accession number или по названию микроорганизма, в том числе выбирая из открывающегося списка всех видов, или воспользовавшись таксономическим браузером, либо напрямую выбрать конкретный геном из предлагаемого списка. После чего

необходимо нажать кнопку «Display Crisprs» для получения в табличной форме и в графическом виде исчерпывающей информации о CRISPR-локусах в выбранном геноме, включая сведения о CRISPR-касетах и Cas-генах. Можно ввести запрос о предоставлении всех характеристик, и появятся развернутые формы таблиц, однако и в них CRISPR повторы приведены только в виде консенсуса с указанием их размеров, а уникальные последовательности спейсеров можно увидеть, кликнув на той или иной последовательности консенсусного повтора, после чего результаты представляются в графическом виде, включая отображение консенсусного повтора в WebLogo формате. Можно также запросить последовательности уникальных спейсеров из данной кассеты и фланкирующие последовательности, причем можно менять их длины слева и справа (по умолчанию по 500 нуклеотидов). Эти результаты формируются в виде отдельных файлов с расширением \*.data.

При сопоставлении пользователем некоей нуклеотидной последовательности с базой данных **CRISPI** имеется ряд опций, которые рекомендуется менять с осторожностью, поскольку по умолчанию они как раз оптимизированы для поиска и анализа разных CRISPR-кассет. Для обнаружения CRISPR-локусов в геноме какого-либо микроорганизма, еще не включенного в базу данных необходимо ввести такую последовательность. Результаты анализа будут предоставлены в табличной форме, а введенный пользователем геном будет храниться на конфиденциальной web-странице в базе данных **CRISPI** в течение еще 10 дней, после чего будет удален.

Выбранный нами в качестве демонстрации геном бактерии *Haliangium ochraceum* DSM 14365 по данным базы данных **CRISPI** содержит 9 CRISPR-кассет и 6 Cas-генов, о чем сообщается в первоначальной краткой информации в табличной форме. При этом девять CRISPR-кассет данного штамма содержат соответственно 6, 6, 5, 106, 203, 5, 216, 160 и 36 повторов/спейсеров, что отличается от сведений в других базах данных.

Еще один крупный портал, находящийся в Новой Зеландии и названный **CRISPRSuite** (<http://bioanalysis.otago.ac.nz/CRISPRTarget/CRISPRSuite.html>), содержит различные инструменты: **CRISPRDetect** ([http://brownlabtools.otago.ac.nz/CRISPRDetect/predict\\_crispr\\_array.html](http://brownlabtools.otago.ac.nz/CRISPRDetect/predict_crispr_array.html)) для поиска CRISPR-локусов и установления направления транскрипции, **CRISPRBank** (<http://bioanalysis.otago.ac.nz/CRISPRBank/>), несущий информацию о CRISPR-локусах, включая сведения о Cas-генах и CRISPR-касетах для большого числа

видов микроорганизмов с секвенированными геномами, а также программный продукт **CRISPRTarget** ([http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr\\_analysis.html](http://bioanalysis.otago.ac.nz/CRISPRTarget/crispr_analysis.html)), предназначенный для поиска протоспейсеров в соответствующих RefSeq базах данных. Полезной чертой базы данных **CRISPRBank** служит информация о гомологии прямых повторов в каждом CRISPR-кассет, отражаемая в % и демонстрируемая в виде замен нуклеотидов в конкретной последовательности.

При поиске в **CRISPRBank** информации о CRISPR-локусах все того же микроорганизма *Haliangium ochraceum* сначала предстояло из открывающего списка выбрать геном штамма DSM 14365 данной бактерии, после чего пользователю предоставляется большой объем информации, включающей в левой колонке сведения о повторах (длины и их последовательности), Cas-генах. Справа приводится информация о гомологии (в %) повторов, даются последовательности спейсеров и фланкирующие последовательности справа и слева по 500 нуклеотидов. По версии ресурса **CRISPRBank** бактерия *Haliangium ochraceum* содержит следующие CRISPR-касеты с разным количеством повторов/спейсеров – 588, 190, 37, 3, 3, 3. Всего шесть блоков, что также не совпадает с данными в других базах данных.

Имеется еще один своеобразный интернет-проект **CRISPR-Exposed**, подготовленный командой The Omicians, находящийся по адресу <http://crispex.net>. Его авторы отмечают, что их интерактивная база данных сродни таким ресурсам как **CRISPRdb** и **CRISPI**, но пока она находится на стадии развития (указанная на сайте дата проекта – 2015 г.) и предлагают всем заинтересованным лицам, если необходимо связываться с ними по электронной почте, указанной на сайте. Визуализация характеристических особенностей CRISPR-кассет различных групп микроорганизмов, представленная на их сайте, дает в настоящее время минимум информации. Эта база данных **CRISPR-Exposed** имеет поисковый инструмент в виде web-версии известной программы **CRT**, с помощью которой можно проводить анализ геномов из ниспадающего большого списка на предмет наличия в них CRISPR-кассет. При этом сначала требуется выбрать или латинское название бактерии, или археи или целую группу микроорганизмов, произвести поиск и для окончательного анализа уже выбрать известные для них те или иные полные геномы. Также можно вводить любую последовательность в FASTA формате или самостоятельно загружать готовый файл. Однако пользователю не допускается менять параметры поиска, принимаемые по умолчанию такими – минимальное число повторов – 3; диапазон

длины прямых повторов – от 19 до 38; диапазон длин спейсеров – от 19 до 48 нуклеотидов. Некоторым недостатком этой базы данных является отсутствие информации о Cas-генах.

В начале данной статьи на примере ресурса **CRISPR-Exposed** нами были приведены сведения о содержании в геноме бактерии *Haliangium ochraceum* CRISPR-кассет. Напомним, что CRISPR-локусов в ее геноме найдено 9 со следующими количествами повторов/спейсеров – 5; 3; 8; 588, 189; 36; 6; 2 и 15. При этом минимальные и максимальные выявленные размеры повторов и спейсеров составили соответственно 19/36 и 19/64 п.н., что превышает установленные программой пределы при поиске, но необходимо заметить, что в той кассете, где выявился спейсер со столь крупным размером, повторяющихся элементов «повтор/спейсер» содержится 15 штук и при одинаковой длине повторов в 22 п.н., размеры спейсеров оказались весьма переменными – от 19 до 64 п.н. при том, что основная их масса имеет длину 22 п.н. Таким образом, данная поисковая программа не проигнорировала эту кассету, несмотря на то, что в нем есть такой длинный спейсер, и нашла его, поскольку «увидела» повторяющиеся элементы, укладывающиеся в установленные рамки.

Подводя некий итог сравнению информации о CRISPR-касетах в геноме бактерии *Haliangium ochraceum* штамма DSM 14365, представленной в четырех базах данных (**CRISPRsdb**, **CRISPI**, **CRISPRBank** и **CRISPR-Exposed**) необходимо заметить, что между некоторыми из них имеются довольно серьезные отличия, в том числе в числе как самих CRISPR-локусов (CRISPR-кассет), так и в количествах повторов/спейсеров в них, возникающие из-за того, что для нахождения CRISPR-локусов были использованы различные программы поиска CRISPR-кассет со своими алгоритмами, при этом также могли задаваться разные параметры поиска при использовании одинаковых программ. Так, например, CRISPR-кассета этой бактерии из 587/588 повторов и спейсеров в базе данных **CRISPI** оказался разделен, если так можно выразиться, на субкасеты из 106, 203, 5 и из 216 повторяющихся блоков, что произошло ввиду того, что используемая программа поиска сочла, что некоторые повторы внутри этого большой кассеты имеют длину не 36 нуклеотидов, как все остальные, а 41 нуклеотид, добавив к ним на 5'- и 3'-концы по несколько нуклеотидов из соседних с ними спейсеров (которые внутри этой CRISPR-кассеты имеют некоторую переменность по длине, что вполне типично и для прочих CRISPR-кассет у различных микроорганизмов) при практически 100%-ной гомологии остальной части всех этих повторов (размером 36 нуклеотидов) из всего данной кассеты.

Лишь три таких прямых повтора в этой кассете имеют замены одиночных нуклеотидов на их 3'-конце.

Портал **Freiburg RNA Tools CRISPRmap – CRISPR Conservation** (<http://rna.informatik.uni-freiburg.de/CRISPRmap/Input.jsp>) расположен в Германии и содержит такой инструмент как **CRISPRmap**, объединенный с программой **CRISPRstrand**, предсказывающей ориентацию tandemных повторов с целью установления какая цепь ДНК кодирует крПНК, чтобы на основании архитектуры CRISPR повторов отнести их к конкретному типу [Lange et al., 2013]. Данный ресурс содержит информацию о 4719 консенсусных CRISPR повторах, формирующих 24 эволюционно консервативных семейств таких повторов. Более ранняя версия содержала меньшее число консенсусных повторов (3527), но распределенных по большому числу групп – 40. Пользователям дозволяется одновременно вводить до 400 последовательностей CRISPR повторов (длиной до 50 нуклеотидов) в FASTA формате как прописными, так и строчными буквами (ACGTUacgtu). Затем предлагается выбрать вариант поиска путем сравнения с 24 или 40 группами. По завершению анализа введенных последовательностей повторов строится филогенетическое древо и определяется, к какой группе относятся те или иные анализируемые CRISPR повторы. В случае, если явное родство с известными группами повторов не обнаружится, то такие неохарактеризованные CRISPR повторы тем не менее займут соответствующее место на построенном древе. Также будет выдана информация о потенциальной вторичной структуре молекул РНК, транскрибируемых с CRISPR повторов, а также изготовлено WebLogo изображение предварительно выравненной последовательности повтора той или иной группы, но уже с учетом анализируемого экспериментатором повтора.

Стоит отметить, что практически все базы данных имеют схожую программную структуру и организацию. В основном они основаны на реляционной и постреляционной модели данных и разработаны с применением MySQL, Oracle (для больших баз данных). Связанные web-сервисы с базами данных реализованы на языке Perl с применением специализированной библиотеки BioPerl, или JavaScript. В основном, web-сервисы рассчитаны на решение двух задач: поиск и добавление информации в базе данных; проведение анализа геномных последовательностей. Разработанные web-сервисы позволяют пользователю проводить поиск данных в базе, добавлять новые данные (после проверки информации на корректность). Базы данных содержат взаимозависимую информацию о геномах и CRISPR-локусах.

### Заключение

Исходя из того, что биоразнообразие бактерий и архей на планете Земля достаточно велико, можно не сомневаться в том, что еще далеко не все типы CRISPR-локусов известны и их обязательно будут продолжать находить, в том числе благодаря бурно развивающемуся полногеномному секвенированию новых поколений, включая метагеномное. Дополнительным основанием к росту такого интереса служит та заметная роль, которую CRISPR/Cas системы уже стали выполнять в генно-инженерных манипуляциях и она непременно будет расти, имея хорошие перспективы, что нашло отражение в других наших статьях в этом номере журнала [Баймиев и др., 2017; Кулуев и др., 2017]. Для успешного редактирования геномов эукариотических организмов необходимо свое биоинформационное сопровождение и свои специализированные программы дизайна молекул РНК, обеспечивающих нахождение точного места внесения мутаций, а также внедрения новых генов, что рассмотрено в другой нашей статье [Чемерис и др., 2017]. Что касается широкомасштабного поиска у различных микроорганизмов с полностью или частично секвенированными геномами (включая метагеномы) различных CRISPR элементов в виде одинаковых прямых повторов, перемежающихся уникальными спейсерами, то для этого уже существуют немало специализированных компьютерных программ и web-ресурсов, а также соответствующих баз данных, описанных в этой статье. Тем не менее, стоят задачи по недопущению появления при компьютерном поиске ложно-положительных сигналов при обнаружении CRISPR-кассет у различных микроорганизмов и исключению получения ложно-негативных результатов в виде пропуска таких элементов геномов при их наличии. Продемонстрированные разночтения в сведениях о CRISPR-локусах по одним и тем же геномам микроорганизмов в разных базах данных свидетельствуют, что далеко не все вопросы с выявлением *in silico* этих важным элементов геномов прокариот еще решены.

### Благодарности

Думаем, что в данной статье нам удалось собрать сведения о практически всех компьютерных программах, предназначенных для поиска и анализа CRISPR-кассет, а также о базах данных по ним, однако не исключаем, что какие-то могли ускользнуть от нашего внимания. Поэтому будем крайне признательны за информацию о таких пропущенных нами программах и базах данных, а их разработчикам заранее приносим свои извинения.

Неизвестному рецензенту за сделанные ценные замечания выражаем свою признательность.

## Литература

1. Баймиев Ан.Х., Кулуев Б.Р., Вершинина З.Р., Князев А.В., Чемерис Д.А., Геращенко Г.А., Баймиев Ал.Х., Чемерис А.В. CRISPR/Cas редактирование геномов (растений) и общество // Биомика. 2017. Т.9. С.183-202.
2. Кулуев Б.Р., Геращенко Г.А., Рожнова Н.А., Баймиев Ан.Х., Вершинина З.Р., Князев А.В., Матниязов Р.Т., Гумерова Г.Р., Никоноров Ю.М., Чемерис Д.А., Баймиев Ал.Х., Чемерис А.В. CRISPR/Cas редактирование геномов растений // Биомика. 2017. Т.9. С.155-182.
3. Кулуев Б.Р., Баймиев Ан.Х., Чемерис Д.А., Матниязов Р.Т., Геращенко Г.А., Никоноров Ю.М., Баймиев Ал.Х., Чемерис А.В. Применение CRISPR-локусов не для редактирования геномов // Биомика. 2017а. Т.9. С.271-283.
4. Чемерис Д.А., Кирьянова О.Ю., Губайдуллин И.М., Чемерис А.В. Дизайн праймеров для полимеразной цепной реакции. (Краткий обзор компьютерных программ и баз данных) // Биомика. 2016. Т.8. С.215-238.
5. Чемерис Д.А., Кирьянова О.Ю., Геращенко Г.А., Кулуев Б.Р., Рожнова Н.А., Матниязов Р.Т., Баймиев Ан.Х., Баймиев Ал.Х., Губайдуллин И.М., Чемерис А.В. Биоинформатические ресурсы для CRISPR/Cas редактирования геномов // Биомика. 2017. Т.9. С.203-228.
6. Abby S.S., Néron B., Ménager H., Touchon M., Rocha E.P. MacSyFinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems // PLoS One. 2014. V.9(10):e110726.
7. Abouelhoda M.I., Kurtz S., Ohlebusch E. Replacing suffix trees with enhanced suffix arrays // J. Discrete Algorithms. 2004. V.2. P.53–86.
8. Alkhnbashi O.S., Costa F., Shah S.A., Garrett R.A., Saunders S.J., Backofen R. CRISPRstrand: predicting repeat orientations to determine the crRNA-encoding strand at CRISPR loci // Bioinformatics. 2014. V.30. P.489-496.
9. Alkhnbashi O.S., Shah S.A., Garrett R.A., Saunders S.J., Costa F., Backofen R. Characterizing leader sequences of CRISPR loci // Bioinformatics. 2016. V.32. P.i576-i585.
10. Anderson R.E., Brazelton W.J., Baross J.A. Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage // FEMS Microbiol. Ecol. 2011. V.77. P.120-133.
11. Barrangou R., Horvath P. A decade of discovery: CRISPR functions and applications // Nat. Microbiol. 2017. V.2:17092.
12. Bao Z., Eddy S.R. Automated de novo identification of repeat sequence families in sequenced genomes // Genome Res. 2002. V.12. P.1269-1276.
13. Ben-Bassat I., Chor B. CRISPR Detection from Short Reads Using Partial Overlap Graphs // Intern. Conf. Res. Comput. Mol. Biol. RECOMB 2015: Research in Computational Molecular Biology. P.16-27.
14. Ben-Bassat I., Chor B. CRISPR detection from short reads using partial overlap graphs // J. Comput. Biol. 2016. V.23. P.461-471.
15. Benson G. Tandem repeats finder: a program to analyze DNA sequences // Nucleic Acids Res. 1999. V.27. P.573-580.
16. Biswas A., Gagnon J.N., Brouns S.J., Fineran P.C., Brown C.M. CRISPRTarget: bioinformatic prediction and analysis of crRNA targets // RNA Biol. 2013. V.10. P.817-827.
17. Biswas A., Fineran P.C., Brown C.M. Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs // Bioinformatics. 2014. V.30. P.1805-1813.
18. Biswas A., Fineran P.C., Brown C.M. Computational Detection of CRISPR/crRNA Targets // Methods Mol. Biol. 2015. V.1311. P.77-89.
19. Biswas A., Staals R.H., Morales S.E., Fineran P.C., Brown C.M. CRISPRDetect: A flexible algorithm to define CRISPR arrays // BMC Genomics. 2016. V.17:356.
20. Bland C., Ramsey T.L., Sabree F., Lowe M., Brown K., Kyripides N.C., Hugenholtz P. CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats // BMC Bioinformatics. 2007. V.8:209.
21. Bolotin A., Quinquis B., Sorokin A., Ehrlich S.D. Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin // Microbiology. 2005. V.151. P.2551–2561.
22. Drevet C., Pourcel C. How to identify CRISPRs in sequencing data // Methods Mol. Biol. 2012. V.905. P.15-27.
23. Dsouza M., Larsen N., Overbeek R. Searching for patterns in genomic data // Trends Genet. 1997. V.13. P.497-498.
24. Durand P., Mahé F., Valin A.S., Nicolas J. Browsing repeats in genomes: Pygram and an application to non-coding region analysis // BMC Bioinformatics. 2006. V.7:477.
25. Dutilh B.E., Schmieder R., Nulton J., Felts B., Salamon P., Edwards R.A., Mokili J.L. Reference-independent comparative metagenomics using cross-assembly: crass // Bioinformatics. 2012. V.28. P.3225-3231.
26. Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity // BMC Bioinformatics. 2004. V.5: 113.

27. Edgar R.C. PILER-CR: fast and accurate identification of CRISPR repeats // *BMC Bioinformatics*. 2007. V.8:18.
28. Edgar R.C., Myers E.W. PILER: identification and classification of genomic repeats // *Bioinformatics*. 2005. V.21. Suppl 1:i152-8.
29. Ge R., Mai G., Wang P., Zhou M., Luo Y., Cai Y., Zhou F. CRISPRdigger: detecting CRISPRs with better direct repeat annotations // *Sci. Rep.* 2016. V.6:32942.
30. Godde J.S., Bickerton A. The repetitive DNA elements called CRISPRs and their associated genes: evidence of horizontal transfer among prokaryotes // *J. Mol. Evol.* 2006. V.62. P.718-729.
31. Gogleva A.A., Gelfand M.S., Artamonova I.I. Comparative analysis of CRISPR cassettes from the human gut metagenomic contigs // *BMC Genomics*. 2014. V.15:202.
32. Grissa I., Vergnaud G., Pourcel C. CRISPRfinder: a web tool to identify clustered regularly interspaced short palindromic repeats // *Nucl. Acids Res.* 2007. V.35. W52-57.
33. Grissa I., Vergnaud G., Pourcel C. The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats // *BMC Bioinformatics*. 2007. V.8:172.
34. Grissa I., Vergnaud G., Pourcel C. CRISPRcompar: a website to compare clustered regularly interspaced short palindromic repeats // *Nucl. Acids Res.* 2008. V.36. W145-148.
35. Gusfield D., Stoye J. Linear time algorithms for finding and representing all the tandem repeats in a string // *J. Computer and System Sciences*. 2004. V.69. P.525-546.
36. Heidelberg J.F., Nelson W.C., Schoenfeld T., Bhaya D. Germ warfare in a microbial mat community: CRISPRs provide insights into the co-evolution of host and viral genomes // *PLoS One*. 2009. V.4(1):e4169.
37. Huntemann M., Ivanova N.N., Mavromatis K., Tripp H.J., Paez-Espino D., Palaniappan K., Szeto E., Pillay M., Chen I.M., Pati A., Nielsen T., Markowitz V.M., Kyrpides N.C. The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4) // *Stand Genomic Sci.* 2015. V.10:86.
38. Huntemann M., Ivanova N.N., Mavromatis K., Tripp H.J., Paez-Espino D., Tennessen K., Palaniappan K., Szeto E., Pillay M., Chen I.M., Pati A., Nielsen T., Markowitz V.M., Kyrpides N.C. The standard operating procedure of the DOE-JGI Metagenome Annotation Pipeline (MAP v.4) // *Stand Genomic Sci.* 2016. V.11:17.
39. Jansen R., Embden J.D., Gaastra W., Schouls L.M. Identification of genes that are associated with DNA repeats in prokaryotes // *Mol Microbiol.* 2002. V. 43. P. 1565–1575.
40. Ishino Y., Shinagawa H., Makino K., Amemura M., Nakata A. Nucleotide sequence of the *iap* gene, responsible for alkaline phosphatase isozyme conversion in *Escherichia coli*, and identification of the gene product // *J. Bacteriol.* 1987. V. 169. P. 5429–5433.
41. Ivanova N., Daum C., Lang E., Abt B., Kopitz M., Saunders E., Lapidus A., Lucas S., Glavina Del Rio T., Nolan M., Tice H., Copeland A., Cheng J.F., Chen F., Bruce D., Goodwin L., Pitluck S., Mavromatis K., Pati A., Mikhailova N., Chen A., Palaniappan K., Land M., Hauser L., Chang Y.J., Jeffries C.D., Detter J.C., Brettin T., Rohde M., Göker M., Bristow J., Markowitz V., Eisen J.A., Hugenholtz P., Kyrpides N.C., Klenk H.P. Complete genome sequence of *Haliangium ochraceum* type strain (SMP-2) // *Stand Genomic Sci.* 2010. V.2. P.96-106.
42. Koonin E.V., Makarova K.S., Zhang F. Diversity, classification and evolution of CRISPR-Cas systems // *Curr. Opin. Microbiol.* 2017. V.37. P.67-78.
43. Kunin V., Sorek R., Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats // *Genome Biol.* 2007. V.8(4):R61.
44. Kurtz S., Choudhuri J.V., Ohlebusch E., Schleiermacher C., Stoye J., Giegerich R. REPuter: the manifold applications of repeat analysis on a genomic scale // *Nucleic Acids Res.* 2001. V.29. P.4633-4642.
45. Kurtz S., Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes // *Bioinformatics*. 1999. V.15. P.426-427.
46. Lange S.J., Alkhnbashi O.S., Rose D., Will S., Backofen R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems // *Nucleic Acids Research*. 2013. V.41. P.8034-8044.
47. Lei J., Sun Y. Assemble CRISPRs from metagenomic sequencing data // *Bioinformatics*. 2016. V.32. P.i520-i528.
48. Lefebvre A., Lecroq T., Dauchel H., Alexandre J. FORRepeats: detects repeats on entire chromosomes and between genomes // *Bioinformatics*. 2003. V.19. P.319-326.
49. Mai G., Ge R., Sun G., Meng Q., Zhou F. A Comprehensive Curation Shows the Dynamic Evolutionary Patterns of Prokaryotic CRISPRs // *Biomed. Res. Int.* 2016;2016:7237053.
50. Mojica F.J., Juez G., Rodríguez-Valera F. Transcription at different salinities of *Haloferax mediterranei* sequences adjacent to partially modified *Pst*I sites // *Mol Microbiol.* 1993. V.9. P.613–621.

51. Mojica F.J., Ferrer C., Juez G., Rodríguez-Valera F. Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloflex mediterranei* and *Haloflex volcanii* and could be involved in replicon partitioning // *Mol Microbiol.* 1995. V. 17. P. 85–93.
52. Mojica F.J., Díez-Villaseñor C., Soria E., Juez G. Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria // *Mol Microbiol.* 2000. V. 36. P. 244–246.
53. Mojica F.J., Díez-Villaseñor C., García-Martínez J., Soria E. Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements // *J Mol Evol.* 2005. V. 60. P. 174–182.
54. Nakata A., Amemura M., Makino K. Unusual nucleotide arrangement with repeated sequences in the *Escherichia coli* K-12 chromosome // *J Bacteriol.* 1989. V. 171. P. 3553–3556.
55. Nicolas J., Rousseau C., Siegel A., Peterlongo P., Coste F., Durand P., Tempel S., Valin A-S., Mahe F. Modeling local repeats on genomic sequences // *Research Report RR-6802, INRIA.* 2008. pp.43.
56. Paez-Espino D., Eloie-Fadrosh E.A., Pavlopoulos G.A., Thomas A.D., Huntemann M., Mikhailova N., Rubin E., Ivanova N.N., Kyrpides N.C. Uncovering Earth's virome // *Nature.* 2016. V.536. P.425-430.
57. Pevzner P.A., Tang H., Tesler G. De novo repeat classification and fragment assembly // *Genome Res.* 2004. V.14. P.1786-1796. Erratum in: *Genome Res.* 2004. V.14. P.2510.
58. Pourcel C., Salvignol G., Vergnaud G. CRISPR elements in *Yersinia pestis* acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies // *Microbiology.* 2005. V. 151. P. 653–663.
59. Price A.L., Jones N.C., Pevzner P.A. De novo identification of repeat families in large genomes // *Bioinformatics.* 2005. V.21. Suppl 1:i351-358.
60. Rho M., Wu Y.W., Tang H., Doak T.G., Ye Y. Diverse CRISPRs evolving in human microbiomes // *PLoS Genet.* 2012. V.8(6):e1002441.
61. Rousseau C., Gonnet M., Le Romancer M., Nicolas J. CRISPI: a CRISPR interactive database // *Bioinformatics.* 2009. V.25. P.3317–3318.
62. Skennerton C.T., Imelfort M., Tyson G.W. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data // *Nucleic Acids Res.* 2013. V.41(10):e105.
63. Sobreira T.J., Durham A.M., Gruber A. TRAP: automated classification, quantification and annotation of tandemly repeated sequences // *Bioinformatics.* 2006. V.22. P.361-362.
64. Stern A., Mick E., Tirosh I., Sagy O., Sorek R. CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome // *Genome Res.* 2012. V.22. P.1985-1994.
65. Ussery D.W., Binnewies T.T., Gouveia-Oliveira R., Jarmer H., Hallin P.F. Genome update: DNA repeats in bacterial genomes // *Microbiology.* 2004. V.150. P.3519-3521.
66. Volfovsky N., Haas B.J., Salzberg S.L. A clustering method for repeat analysis in DNA sequences // *Genome Biol.* 2001. V.2(8):RESEARCH0027
67. Zhang Q., Ye Y. Not all predicted CRISPR-Cas systems are equal: isolated cas genes and classes of CRISPR like elements // *BMC Bioinformatics.* 2017. V.18(1):92.

#### BIOINFORMATIC RESOURCES FOR *IN SILICO* SEARCH OF THE CRISPR LOCI IN THE GENOMES OF PROKARYOTES

Baymiev An.Kh.<sup>1</sup>, Chemeris D.A.<sup>1</sup>, Kiryanova O.Yu.<sup>2,3</sup>, Matniyazov R.T.<sup>1</sup>,  
Valeev A.Sh.<sup>1</sup>, Baymiev Al.Kh.<sup>1</sup>, Gubaydullin I.M.<sup>2,3</sup>, Chemeris A.V.<sup>1</sup>

<sup>1</sup>Institute of Biochemistry and Genetics, Ufa Scientific Center of RAS, Ufa, Russia, [chemeris@anrb.ru](mailto:chemeris@anrb.ru)

<sup>2</sup>Institute of Petrochemistry and Catalysis of Russian Academy of Sciences, Ufa, Russia

<sup>3</sup>Ufa State Petroleum Technological University, Ufa, Russia

#### Resume

Brief characteristics of CRISPR loci which found in approximately half of the bacteria and most archaea are given. Their typical organization, an important element of which serve CRISPR-cassette that contains unique spacers alternating with identical direct repeats are shown. Specialized search programs for CRISPR-cassettes in the sequenced genomes of microorganisms and metagenomic data by identifying of repeating sections in them are briefly considered. The web pages of these programs and their purpose and capabilities are shown in tabular form. Databases for CRISPR-loci showing their web addresses are described. Almost all available literature on the matter and relevant Internet resources are analyzed.

**Keywords:** CRISPR, CRISPR/Cas system, CRISPR-locus, CRISPR-cassette, spacer, quasi-tandem repeat, protospacer, software, web-resource, database