MARCOS DAMRLEY GALVÃO DA LUZ

ANÁLISE DA DIVERSIDADE DE CRISPR EM DADOS METAGENÔMICOS

BELÉM

2017

MARCOS DAMRLEY GALVÃO DA LUZ

ANÁLISE DA DIVERSIDADE DE CRISPR EM DADOS METAGENÔMICOS

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Bacharelado em Ciências Biológicas da Universidade Federal do Pará, como requisito parcial para a obtenção do grau de Bacharel em Biologia.

Orientador: Diego Assis das Graças. Laboratório de Genômica e Bioinformática – ICB – UFPA

BELÉM

2017

MARCOS DAMRLEY GALVÃO DA LUZ

ANÁLISE DA DIVERSIDADE DE CRISPR EM DADOS METAGENÔMICOS

Trabalho de Conclusão de Curso apresentado ao Colegiado do Curso de Bacharelado em Ciências Biológicas da Universidade Federal do Pará, como requisito parcial para a obtenção do grau de Bacharel em Biologia.

Orientador: Prof. Dr. Diego Assis das Graças

Laboratório de Genômica e Bioinformática, UFPA

Avaliador: Prof. Dr. Rafael Azevedo Baraúna

Laboratório de Genômica e Bioinformática, UFPA

Avaliador: Mestranda Roberta Rezende de Castro

Laboratório de Polimorfismo do DNA, UFPA

BELÉM

2017

Dedico este trabalho a um pedaço de mim que se foi no último mês de novembro. Muito obrigado por seu amor e sua alegria. Lembro cada conversa que tivemos, especialmente após minha aprovação na Biologia. Sentirei saudades e me apegarei sempre às boas lembranças, meu avô.

AGRADECIMENTOS

Agradeço primeiramente a minha família. Aos meus pais pela confiança, acima de tudo, e todo suporte que me deram. À minha irmã por sua companhia. À Roberta, meu amorzinho, que vejo crescer a cada dia, preenchendo nossas vidas com sua alegria e inteligência;

Ao professor Artur que me recebeu tão bem e me fez sentir à vontade no LPDNA (agora Laboratório de Genômica e Bioinformática), onde nos últimos dois anos e meio, tenho percorrido os primeiros passos da minha vida científica;

Ao Diego Assis pela orientação durante minha iniciação científica, pela paciência infinita, seriedade profissional e confiança.

Ao CNPq pela bolsa de iniciação científica concedida;

Aos colegas de laboratório. Principalmente às amizades que lá construí. Obrigado Ailton, Yan e Amália por nossas viagens e saídas, até as comemorações mais banais com vocês são imbatíveis;

Aos amigos da faculdade. Não consigo resumir quatro anos de aprendizado pessoal (de tão imenso) ao lado de vocês, Denyse, Joás, Caio, Carol... E agora nessa fase final do curso, o que seria de mim sem o Ericks e a Klyssia... E também o Danielson com muitos conselhos excelentes!

À Faculdade de Biologia, especialmente, aos professores que têm contribuído para formar biólogos mais críticos. À UFPA, por ser o ambiente onde tantas coisas boas aconteceram...

SUMÁRIO

LISTA DE FIGURAS E TABELAS

RESUMO

O sistema de defesa CRISPR-Cas presente em arqueias e bactérias é responsável pela eliminação de sequências exógenas invasoras de vírus e plasmídeos. Na natureza a corrida armamentista entre vírus e seus hospedeiros gera pressões evolutivas nas comunidades microbianas e modificações no meio ambiente. A identificação de sequências relacionadas à CRISPR-Cas é uma nova ferramenta que contribui para a compreensão da diversidade de ambientes, como rios e lagos de água doce. Através da abordagem metagenômica é possível acessar informações genéticas ambientais e, a partir de 2005, com o advento das tecnologias de Sequenciamento de Próxima Geração os dados provenientes de amostras ambientais foram entregues de maneira mais barato e factível. Neste trabalho foi realizado a coleta de água doce e sedimento no reservatório da usina hidrelétrica Curuá-Una, na Amazônia. O DNA total dessas amostras foi extraído, e sequenciado pela plataforma Ion Torrent PGM chip 318, de acordo com as instruções dos fabricantes. A análise e reconstrução de diferentes tipos de arranjos CRISPR foram feitas pelo algoritmo Crass. Os dados metagenômicos foram submetidos e analisados pelo *pipeline* do EBI Metagenomics. Os dados brutos de sequenciamento de todas as amostras foram de 6.113.450 leituras e 1,7 Gpb de rendimento. Após a análise de qualidade foram totalizadas 4.535.043 leituras e um rendimento de 1,1 Gpb de informação metagenômica. Em relação à atribuição taxonômica (análise de rRNA). Para a camada fótica de água doce, 78% de 620 leituras foram atribuídas ao domínio Bacteria. Na camada afótica, 48% de 588 leituras foram relacionadas ao domínio Bacteria. Em sedimento, 26% estavam associadas ao domínio Bacteria e 11% ao domínio Archaea. De acordo com a análise funcional foram identificadas 389 CDS preditas associadas à CRISPR-Cas, 88% destas na camada sedimentar, 8% na fótica e 4% na afótica. O Crass identificou um total de dezesseis grupos CRISPR distintos, com diferentes organizações dos espaçadores em arranjos. A curadoria manual revelou 27 espaçadores não identificados pelo Crass. Este trabalho mostrou um maior número de sequências relacionadas à CRISPR-Cas na camada sedimentar. E contribui para a caracterização da microbiota de ambientes de água doce, modificados pela ação humana.

PALAVRAS-CHAVE

CRISPR-Cas; Crass; metagenoma

1 INTRODUÇÃO

Os micro-organismos são ubíquos na superfície da Terra, estão presentes nos mais diversos ecossistemas, entre eles fontes termais, rios, florestas tropicais, estabelecendo algum tipo de relação ecológica por meio de associações entre suas populações, ou com animais e plantas [1]. Em ambientes aquáticos marinhos, o número total de células procarióticas somente é superado pela quantidade de vírus, a entidade biológica mais abundante, cuja maioria é de bacteriófagos (ou fagos) [2, 3]. A interação fago-hospedeiro é moldada por uma constante corrida armamentista evolutiva, na qual mutações selecionam bactérias aptas a escapar de infecções, em contrapartida, os fagos também modificam suas especificidades infecciosas [4, 5]. Nos últimos anos, foi descoberto em arqueias e bactérias, Repetições Palindrômicas Curtas Interespaçadas e Regularmente Agrupadas (CRISPR, do inglês Clustered Regularly Interspaced Short Palindromic Repeats), e proteínas associadas (Cas, do inglês CRISPR associated), que juntos constituem um sistema de defesa adaptativo, conhecido como CRISPR-Cas.

Esse sistema está presente em 90% dos genomas de arqueias e em metade dos genomas bacterianos [http://crispr.i2bc.paris-saclay.fr/crispr/ acessado em 09.01.2017], atribuindo resistência contra vírus e plasmídeos, através de pequenos RNA-guia e proteínas Cas [6, 7]. Em geral, seu *locus* consiste em um óperon de genes *cas* adjacente a uma série de repetições diretas (DR, do inglês Direct Repeats) que são interespaçadas por sequências variáveis de DNA exógeno, os chamados espaçadores [8]. Esses espaçadores são como um registro histórico e evolutivo de infecções prévias, utilizados para a eliminação de invasões

subsequentes [9]. CRISPR-Cas é constituído por três etapas - adaptação, transcrição e interferência. Na primeira etapa, novos espaçadores adquiridos, cujo comprimento pode variar de 24 a 48 nucleotídeos, são integrados ao arranjo CRISPR, na maioria dos casos, pelo complexo de proteínas Cas1-Cas2 [10, 11]. Anteriormente, ocorre o reconhecimento e a seleção de sequências potencialmente invasoras, conhecidas como protoespaçadores, geralmente pela identificação de um motivo adjacente ao protoespaçador (PAM, do inglês Protospacer Adjacent Motif) [12]. Na segunda etapa, apenas uma proteína (Cas9; Cpf1) ou conjunto proteico (*Cascade*; Cmr), processa uma molécula única de RNA, que contém todos os espaçadores, em pequenos RNA CRISPR (crRNA) [13]. Cada crRNA é uma unidade espaçadora flanqueada por DR. Na última etapa, as proteínas Cas guiadas por essas pequenas moléculas de RNA reconhecem e clivam as sequências invasoras [14].

Uma vez que a coevolução entre micróbios e vírus geram pressões nas comunidades microbianas, afetando a ciclagem de nutrientes e o clima global [4]. É de suma importância o conhecimento da microbiota e da composição gênica funcional, incluindo a análise de CRISPR-Cas, nos mais variados ecossistemas. Por meio da abordagem metagenômica temos acesso a essas informações genéticas ambientais, assim como, à estrutura, composição e relações filogenéticas de comunidades microbianas, sem a necessidade de cultivo ou isolamento de espécimes [15]. Recentemente, um trabalho realizado em diferentes ambientes identificou dois novos sistemas CRISPR-Cas, ainda não caracterizados [16]. Enquanto que na Amazônia alguns estudos realizados em rios e lagos de

água doce apontaram os grupos mais abundantes de bactéria e arqueia [17, 18] No entanto, nenhum até o presente momento havia investigado a diversidade de CRISPR-Cas. Atrelado também, às novas descobertas e ao avanço nas pesquisas de ecologia molecular, está o Sequenciamento de Próxima Geração (NGS, do inglês Next-Generation Sequencing). Uma plataforma de alto rendimento de NGS como o Ion Torrent Personal Genome Machine (PGM) da Thermo Fisher Scientific é capaz de gerar milhões de leituras de sequenciamento de maneira massivamente paralela [19]. Além disso, melhorias no rendimento de sequenciamento, no comprimento de leitura e acurácia estão permitindo uma melhor representação da diversidade ambiental, a um custo cada vez menor [20].

2 MATERIAL E MÉTODOS

2.1 Coleta e amostragem

Uma coleta foi realizada em fevereiro de 2016 no reservatório da usina hidrelétrica de Curuá-Una no Estado do Pará, Brasil. No total foram retiradas três amostras, em um único ponto geográfico (2° 53’ 32”S, 54° 24’ 49”O). Duas de água doce (1 L de água para cada uma), sendo uma amostra na camada fótica há 1 m de profundidade, e outra na camada afótica há 11,3 m de profundidade. A terceira foi de sedimento (50 g) na profundidade de 11,3 m. Água e sedimento foram coletados utilizando garrafa de van Dorn (Alfakit, Florianópolis, Brasil) e garrafa de van Veen, respectivamente. Duas etapas foram empregadas para a filtragem da água: primeiro utilizando um papel qualitativo com poros de 0,8 µm,

em seguida, uma membrana de nitrocelulose com poros de 0,22 µm (Whatman/GE Healthcare, Maidstone, Reino Unido). Os dois tipos de materiais amostrados foram armazenados em solução tampão STE (50 mM Tris-HCl, 500 mM NaCl, 125 mM EDTA pH 8,0) e mantidos conservados a -20 ºC até o procedimento de extração de DNA.

2.2 Extração e sequenciamento de DNA metagenômico

O DNA metagenômico total foi extraído através do UltraClean Microbial DNA Isolation Kit e UltraClean Mega Soil DNA Isolation Kit (MoBio/QIAGEN, Carlsbad, Estados Unidos), respectivamente, para água doce e sedimento, de acordo com as instruções do fabricante. O sequenciamento do DNA total extraído foi realizado pela plataforma Ion Torrent PGM, utilizando o Ion 318 Chip v2 com o Ion PGM Sequencing 400 Kit, seguindo as especificações do fabricante (Thermo Fisher Scientific, Waltham, Estados Unidos).

2.3 Análise de dados

2.3.1 Algoritmo Crass (CRISPR)

O Crass foi o algoritmo utilizado para localizar individualmente leituras e agrupá-las de acordo com o tipo de repetição direta, a partir dos dados brutos de sequenciamento [21]. A busca inicial identificou apenas leituras com o mínimo de duas repetições diretas. Para leituras longas (>176 pb) foi necessário no mínimo 3 repetições diretas identificadas para que elas fossem assumidas como parte de um arranjo CRISPR. Foram utilizados os arquivos de saída fasta, contendo as leituras com o tipo de CRISPR encontrado, e os arquivos crass.crispr, com informações

sobre repetições diretas, espaçadores, flanqueadores (sequências que indicam o término de um arranjo no Crass), cobertura etc., para todos os grupos CRISPR, obtidos em formato XML. Em todas as etapas do Crass foram utilizados os parâmetros padrões: limite inferior e superior para uma repetição direta de 23 e 47 pb, respectivamente; para espaçadores esses valores variaram de 26 a 50 pb. Ao invés de se utilizar os arquivos . gv para a reconstrução de arranjos CRISPR, optou-se pela curadoria manual através do BioEdit.

2.3.2 EBI Metagenomics

Foi realizada a submissão dos dados brutos de leituras de sequenciamento ao EBI Metagenomics, cujo pipeline (v.3.0) foi responsável pela análise automatizada das sequências metagenômicas. Esse procedimento incluiu a etapa de controle de qualidade (QC, do inglês Quality Control) para a remoção de sequências de curta ou baixa qualidade; e também a análise de diversidade taxonômica (baseada em rRNA) e análise funcional (InterPro), incluindo proteínas associadas ao sistema CRISPR-Cas [22].

3 RESULTADOS

3.1 Sequenciamento e análise de qualidade

Os dados brutos de sequenciamento das amostras de água doce e sedimento através da plataforma Ion Torrent PGM foram de 6.113.450 leituras e 1,7 Gpb de rendimento. Após a análise de qualidade realizada pelo EBI foram totalizadas

4.535.043 leituras e um rendimento de 1,1 Gpb de informação metagenômica. As camadas fótica, afótica e sedimentar têm suas especificações de sequenciamento e análise de qualidade apresentados nas Tabelas 1, 2 e 3, respectivamente.

TABELA 1 **S**equenciamento e análise de qualidade da camada fótica

Parâmetros	**Submetido**	**Análise de Qualidade**
Dados Brutos	723 Mpb	469,9 Mpb
Número de Leituras	2.491.500	1.862.354
Tamanho Médio das Leituras	290	252 ± 86
% Sequências Pós-Análise	74,7%	
% Leituras	36%	

TABELA 2 Sequenciamento e análise de qualidade da camada afótica

Parâmetros	**Submetido**	**Análise de Qualidade**
Dados Brutos	278 Mpb	188 Mpb
Número de Leituras	971.007	731.965
Tamanho Médio das Leituras	287 pb	258 ± 82 pb

% Sequências Pós-Análise	75%
% Leituras Proteicas	22%

TABELA 3 Sequenciamento e análise de qualidade do sedimento

Parâmetros	**Submetido**	**Análise de Qualidade**
Dados Brutos	774 Mpb	485 Mpb
Número de Leituras	2.650.943	1.940.724
Tamanho Médio das Leituras	292 pb	250 ± 88
% Sequências Pós-Análise	73%	
% Leituras Proteicas	42%	

3.2 Análise taxonômica, funcional e proteínas Cas

A abundância relativa da composição de todos os filos de arqueias e bactérias foi identificada pelo pipeline do EBI Metagenomics (dados de rRNA) para as camadas fótica, afótica e sedimentar (Fig. 1).

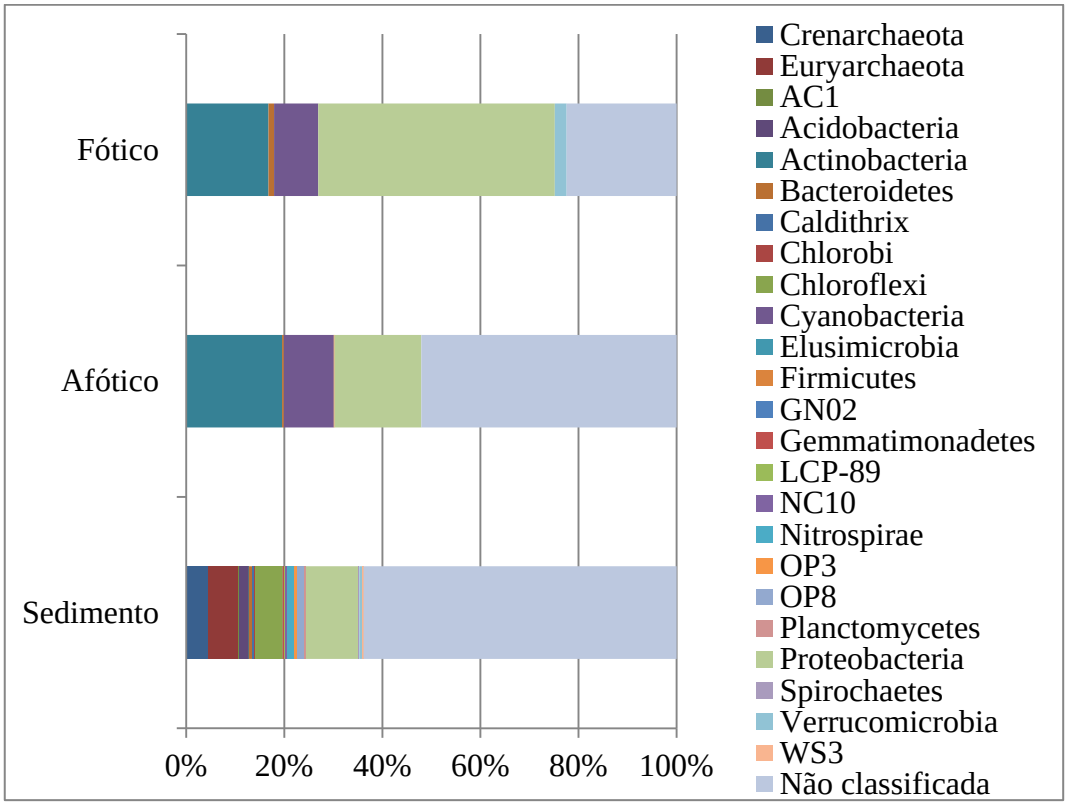

FIGURA 1 Composição de filos de arqueias e bactérias nas camadas fótica, afótica e sedimentar

Na camada fótica 78% de todas as 620 leituras foram atribuídas ao domínio Bacteria; os demais 22% não foram classificados. A camada afótica teve 52% das leituras não atribuídas a qualquer grupo taxonômico e 48% de um total de 588 leituras relacionadas ao domínio Bacteria. Para maior parte das 791 leituras encontradas no sedimento não foi possível atribuir taxonomia (64%), 26% estavam associadas ao domínio Bacteria e 11% ao domínio Archaea.

A análise funcional (InterPro) da camada fótica identificou 692.954 CDS preditas e 686.685 leituras com CDS preditas. Na camada afótica quanto às CDS preditas, foram identificadas 162.761; já as leituras com CDS preditas foram de 161.608. Para o sedimento foram identificadas 827.007 CDS preditas e 822.571 leituras com CDS preditas. Proteínas Cas e outras proteínas relacionadas ao sistema CRISPR-Cas foram identificadas, 389 CDS preditas no total, sendo 88% encontradas no sedimento, 8% na fótica e 4% na afótica (Anexo I).

3.3 Diversidade de arranjos CRISPR

O algoritmo Crass identificou um total de dezesseis grupos CRISPR distintos. Na camada fótica foram identificados oito grupos (G6, G15, G20, G29, G72, G77, G680, G689). No sedimento também foram identificados oito grupos (G6, G11, G17, G32, G38, G50, G56, G58). Para a camada afótica não foram localizadas leituras com repetições diretas que indicassem sequências relacionadas a arranjos CRISPR.

Por meio de curadoria manual (BioEdit) foi notado que cada grupo encontrado independentemente se relacionado à camada fótica ou ao sedimento apresentou repetições diretas, flanqueadores e espaçadores diferentes (Anexos II e III). Assim como a disposição dos espaçadores nos arranjos CRISPR foi distinta. No caso de G6 da camada fótica foram obtidas dez prováveis estruturas de organização das sequências espaçadoras (em 87 leituras). Na estrutura 1 (Fig. 2) chamada assim por estar contida na maior leitura (406 pb), foi possível organizar os espaçadores identificados (4) e a sequência flanqueadora entre as repetições

diretas, sem nenhum *gap*, diferente do que ocorreu para as demais estruturas (2 a 10), nas quais foram identificados um ou dois espaçadores, e repetições diretas que se alternavam com sequências não identificadas pelo Crass (Fig. 3). Os demais grupos da camada fótica tiveram três ou quatro espaçadores identificados em sequências que apresentaram ou não flanqueadores (Anexos II e III).

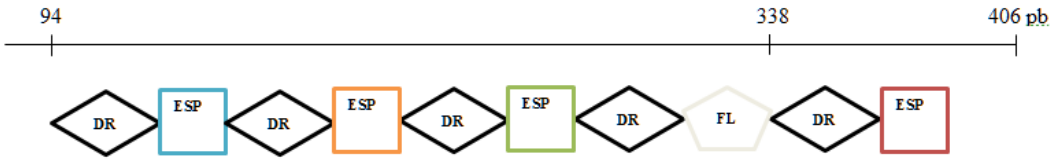

FIGURA 2 Arranjo CRISPR da estrutura 1 encontrada no G6 da camada fótica. DR, *direct repeat*; ESP, espaçador; FL, flanqueador

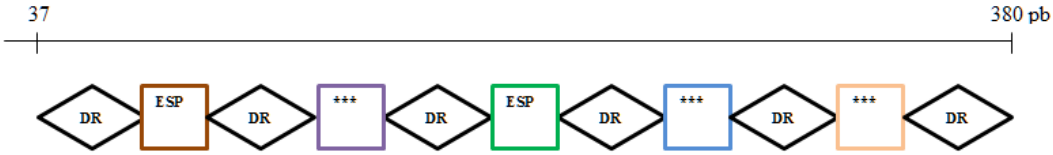

FIGURA 3 Arranjo CRISPR da estrutura 2 encontrada no G6 da camada fótica. DR, *direct repeat*; ESP, espaçador; ***, espaçador não identificado pelo Crass

Todos os grupos da camada fótica, com exceção de G6, e de sedimento apresentaram apenas uma possibilidade de organização dos espaçadores em um arranjo CRISPR. Por exemplo, na camada de sedimento o G58 apresentou três

espaçadores localizados posteriormente ao flanqueador em uma leitura de 313 pb [Fig. 4]. Nos grupos de sedimento houve ainda, trechos de sequências DR fora do comprimento da leitura de 384 pb (G38); espaçadores com comprimento médio de 15 pb (G50); neste último os *gaps* foram mais frequente. O número de espaçadores para as sequências de sedimento variou entre três e cinco.

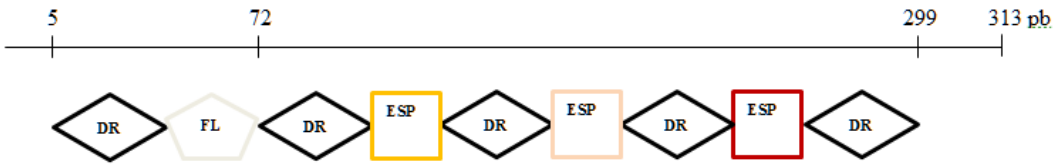

FIGURA 4 Arranjo CRISPR encontrado no G58 da camada sedimentar. DR, *direct repeat*; ESP, espaçador

4 DISCUSSÃO

Os dois filos mais abundantes, Proteobacteria e Actinobacteria, presentes na coluna d'água (camada fótica e afótica) foram apontados também como os principais grupos taxonômicos, baseado em dados metagenômicos, em um estudo realizado no rio Amazonas (o qual o rio Curuá-Una deste trabalho é afluente) [17]. Na camada considerada afótica Cyanobacteria se mostrou bem representativa possivelmente porque o reservatório de Curuá-Una manteve o aspecto de rio mesmo após a construção da usina hidrelétrica, ou seja, a água é muito movimentada e com muita vazão.

Com exceção da camada fótica, mais da metade das leituras para as demais camadas não foram atribuídas a nenhuma classificação taxonômica. Embora entre todas as leituras identificadas nenhuma tenha correspondido a organismos eucariotos ou vírus, supõe-se que grande parte dessas sequências seja de vírus, visto que, em amostras ambientais mais da metade das sequências encontradas em um viroma são desconhecidas [3], e também pela escassez de depósito de genomas virais nos bancos de dados. Além do mais já foi demonstrada uma alta abundância viral em ambientes marinho e de água doce, assim como, em sedimentos de ambos ambientes [23, 24]. O sedimento bentônico de água possui uma densidade de vírus geralmente maior, em média 34,2 x 10^8 g^{-1} (sedimento seco) [25]. É justamente na camada de sedimento que se encontrou um número maior de sequências desconhecidas. E também uma quantidade maior de leituras com CDS preditas relacionadas à CRISPR-Cas. Foram encontrados tipos proteicos e proteínas Cas, algumas das quais, suas sequências nucleotídicas são assinaturas gênicas para determinados subtipos de CRISPR-Cas ou compõem algum dos módulos funcionais desse sistema [26].

Em estudo metagenômico recente foi analisado um conjunto de mais de 155 milhões de genes codificantes, oriundos de comunidades microbianas de águas subterrâneas, sedimentos, biofilmes e intestinos, para identificar novos sistemas CRISPR-Cas de classe II. Eles se concentraram em genes proximais aos arranjos CRISPR e integrase Cas1. Ao fazer isso, o grupo identificou as primeiras proteínas Cas9 em espécies não bacterianas (em nanoarchaeas), chamados ARMAN-1 e ARMAN-4, demonstrando o enorme potencial da análise de

CRISPR em dados metagenômicos, pois estes micro-organismos não possuem representantes cultivados. Esse mesmo estudo também destaca que o novo sistema chamado CRISPR-Cas Y está presente em espécies bacterianas que ainda não tem representantes cultivados, o que justifica um foco maior nas análises metagenômicas [15].

Os resultados da curadoria das sequências demonstram que há 27 possíveis novos espaçadores não detectados pelo programa CRASS. Apesar de o programa CRASS ter fornecido um bom resultado, foi possível demonstrar que as análises ainda precisam de uma curadoria manual e ainda há necessidade de mais ferramentas para detecção de sequências CRISPR-Cas e que tenham bases de dados atualizadas. Os resultados sobre a detecção de sequências CRISPR encontradas nos dados metagenômicos analisados demonstram que na camada sedimentar do reservatório de Curuá-Una há uma quantidade maior de sequências atribuídas a este sistema. Existem pouquíssimos trabalhos que realizam detecção de CRISPR em dados metagenômicos, sendo este o primeiro em ambiente de reservatório e comparando camadas. Apesar das análises ainda não suportarem tal conclusão, nós especulamos que as camadas sedimentares de reservatórios devam possuir maior abundância e diversidade de sequências CRISPR, acompanhando métricas ecológicas e de diversidade taxonômica que já conhecemos para estes ambientes [27]. De fato, se o ambiente sedimentar é mais diverso que o da coluna d’ água, esperamos que a diversidade de CRISPR também seja maior nesta camada.

Nós destacamos que a análise de sequências CRISPR em dados metagenômicos se apresenta como uma nova fronteira de análise do sistema de defesa adaptativa de bactérias e arqueias, e que a enorme quantidade de dados disponíveis nos bancos de dados requer novas ferramentas que automatizem os processos, mas que a curadoria manual das detecções ainda é importante.

REFERÊNCIAS

[1] Madigan MT, Martinko JM, Bender KS, Buckley DH. Microbiologia de Brock. Porto Alegre: Artmed; 2016 . p. 598-602.

[2] Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. 2000;64:69-114.

[3] Rosario K, Breitbart M. Exploring the viral world through metagenomics. Curr. Opin. Virol. 2011;1:289-97. DOI: 10.1016/j.coviro.2011.06.004

[4] Stern A, Sorek R. The phage-host arms race: shaping the evolution of microbes. Bioessays 2010;33:43-51. DOI: 10.1002/bies.201000071

[5] Lopez-Pascua LDC, Buckling A. Increasing productivity accelerates host-parasite coevolution. J. Evol. Biol. 2008;21:853-60. DOI: 10.1111/j.1420-9101.2008.01501.x

[6] Brouns SJJ, Jore MM, Lundgren M, Westra ER. Small CRISPR RNAs guide antiviral defense in prokaryotes. Science 2008;321:960-4.

[7] Horvath P, Romero DA, Coûté-Monvoisin AC, Richards M. J. Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. J. Bacteriol. 2008;190:1401-12. DOI: 10.1128/JB.01415-07

[8] Mohanraju P, Makarova KS, Zetsche B, Zhang F. Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science 2016;353:556. DOI: 10.1126/science.aad5147

[9] Barrangou R, Fremaux C, Deveau H, Richards M. CRISPR provides acquired resistance against viroses in prokaryotes. Science 2007;315:1709-12.

[10] Hatoum-Aslan A, Samai P, Maniv I, Jiang W. A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J. Biol. Chem. 2013;288:27888-97. DOI: 10.1074/jbc.M113.499244

[11] Amitai G, Sorek R. CRISPR-Cas adaptation: insights into the mechanism of action. Nat. Rev. Microbiol. 2016;14:67-76.

[12] Mojica FJM, Díez-Villaseñor C, García-Martínez J, Almendros C. Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology 2009;155:733:40.

[13] Charpentier E, Richter H, van der Oost J, White MF. Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptative immunity. FEMS Microbiol. Rev. 2015;39:428-41. DOI: 10.1093/femsre/fuv023

[14] Marraffini LA, Sontheimer EJ. CRISPR interference limits horizontal gene transfer in Staphylococci by targeting DNA. Science 2008;322:1843-5.

[15] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl. Environ. Microbiol 2011;77:1153-61. DOI: 10.1128/AEM.02345-10

[16] Burstein D, Harrington LB, Strutt SC, Probst AJ. New CRISPR-Cas systems from uncultivated microbes. Nature 2017;542:237-41. DOI: 10.1038/nature21059

[17] Ghai R, Rodríguez-Valera F, McMahon KD, Toyama D. Metagenomics of the water column in the pristine upper course of the Amazon river. Plos One 2011;6:e23785. DOI: 10.1371/journal.pone.0023785

[18] Graças DA, Ramos RTJ, Sá PG, Baraúna RA. Semiconductor sequencing reveals the diversity of bacterial communities in an Amazonian reservoir. Aquat. Sci. Technol. 2014;3:18-32.

[19] Buermans HPJ, den Dunnen JT. Next generation sequencing technology: advances and applications. Biochim. Biophys. Acta 2014;1842:1932-41. DOI: 10.1016/j.bbadis.2014.06.015

[20] Shokralla S, Spall JL, Gibson JF, Hajibabaei M. Next-generation sequencing technologies for environmental DNA research. Mol. Ecol. 2012;21:1794-805.

[21] Skennerton CT, Imelfort M, Tyson GW. Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. Nucleic Acids Res. 2013;41:e105. DOI: 10.1093/nar/gkt183

[22] Mitchell A, Bucchini F, Cochrane G, Denise H. EBI metagenomics in 2016 – an expanding and envolving resource for the analysis and archiving of

metagenomic data. Nucleic Acids Res. 2016;44:D595-D603. DOI: 10.1093/nar/gkv1195

[23] Wommack KE, Colwell RR. Virioplankton: viruses in aquatic ecosystems. Microbiol. Mol. Biol. Rev. 2000;64:69-114.

[24] Middelboe M, Jacquet S, Weinbauer M. Viruses in freshwater ecosystems: an introduction to the exploration of viroses in new aquatic habitats. Freshwater Biol. 2008;53:1069-75.

[25] Donavaro R, Corinaldesi C, Filippini M. Viriobenthos in freshwater and marine sediments: a review 2008;53:1186-1213.

[26] Makarova KS, Wolf YI, Alkhnbashi OS, Costa F. An updated evolutionary classification of CRISPR-Cas systems. Nat. Rev. Microbiol. 2015;13:722-36.

[27] Lozupone CA, Knight R. Global patterns in bacterial diversity. Proc. Nati. Acad. Sci. USA 2007;104:11436-40.

ANEXO I

Proteínas Cas e outras proteínas preditas (389) pelo EBI Metagenomics, relacionadas ao sistema CRISPR-Cas, para as camadas fótica, afótica e sedimento.

IPR	Descrição	Fótica	Afótica	Sedimento
IPR002729	CRISPR-associated protein Cas1	1	0	125
IPR010144	CRISPR-associated protein, Csd1-type	8	0	30
IPR006482	CRISPR-associated protein Cas7, subtype I-B/I-C	0	0	22
IPR021124	CRISPR-associated protein, Cas5	1	0	22
IPR013422	CRISPR-associated protein Cas5, N-terminal	0	0	17
IPR005537	CRISPR type III-associated RAMP protein	0	12	14
IPR010148	CRISPR-associated protein, CT1975	13	0	11
IPR013381	CRISPR-associated protein Cse1	0	0	8
IPR013403	CRISPR-associated protein GSU0053	0	0	7
IPR019199	Virulence-associated protein D / CRISPR associated protein Cas2	1	0	7
IPR021127	CRISPR-associated endonuclease Cas2	5	0	7
IPR013412	CRISPR-associated RAMP Csm3	0	2	0
IPR019857	CRISPR-associated protein Cas1, YPEST subtype	1	0	0
IPR010155	CRISPR pre-crRNA endoribonuclease Cas5d	0	0	6
IPR019267	CRISPR-associated protein Cas6, C-terminal	0	0	6
IPR006483	CRISPR-associated Cas3-HD domain	0	0	5

IPR009260	CRISPR-associated exonuclease Csa1	0	0	5
IPR010179	CRISPR-associated protein Cse3	0	0	5
IPR013343	CRISPR-associated protein Cas4	0	0	4
IPR013389	CRISPR-associated protein TM1802	0	0	4
IPR019089	CRISPR-associated protein, GSU0054	0	0	3
IPR019117	CRISPR-associated protein, Cmr3	0	0	3
IPR024615	CRISPR-associated protein Crm2, N-terminal	0	0	3
IPR006474	Helicase Cas3, CRISPR-associated, core	0	0	2
IPR010147	CRISPR-associated protein, CasD	0	0	2
IPR010149	CRISPR-associated protein, Csm2 Type III-A	0	0	2
IPR010156	CRISPR-associated protein, Cas6	0	0	2
IPR010173	CRISPR-associated protein, TM1807	0	0	2
IPR013442	CRISPR-associated protein APE2256	0	0	2
IPR014174	CRISPR-associated protein, Cas6-related	0	0	2
IPR017574	CRISPR-associated protein Csc2	0	0	2
IPR019016	CRISPR-associated DxTHG protein	0	0	2
IPR019092	CRISPR-assoc protein, NE0113/Csx13	0	0	2
IPR010152	CRISPR-associated protein Cas2 subtype	0	0	1
IPR010154	CRISPR-associated protein Cas7/Cst2/DevR	0	0	1
IPR010160	CRISPR-associated protein, Cmr5	0	0	1
IPR013382	CRISPR-associated protein Cse2	0	0	1
IPR013383	CRISPR-associated protein DxTHG, conserved site	0	0	1
IPR013407	CRISPR-associated protein Crm2	0	0	1

IPR013409	CRISPR-associated protein Csx3	0	0	1
IPR019855	CRISPR-associated protein Cas1, NMENI subtype	0	0	1
IPR026483	CRISPR-associated protein Csx17, subtype Dpsyc	0	0	1
IPR031820	CRISPR-associated protein Csn2 subfamily St	0	0	1
IPR010180	CRISPR-associated protein, CXXC-CXXC	0	1	0

ANEXO II

Dados do Crass (crass.crispr) para a camada fótica. DR, *direct repeat*; ESP., espaçador; FL, flanqueador. T.m, tamanho médio; Co. m., cobertura média

GID	DR consenso	T.m. DR	Esp.	T.m. Esp.	Co. m. Esp.	FL	T.m. FL	Leituras
G6	AATTAACACATCTAGATAATCTTCCTATTAATCTTAC	37	9	23	2	1	27	87
G15	AATCCAAATGCCATTCATATCTTAGAAAAGAACTTGGAT	39	3	31	5	1	42	8
G20	AGTTTTAATAAAATAACAAGTATATGTCC	29	3	40	2	0	0	4
G29	AATTAACCCAATTGGATAATTTACCGCAAAGTCTTAAAA	39	3	21	3	1	24	5
G72	CTGCTATTAGTAATTTAAACGCTGC	25	4	38	5	0	0	8
G77	TATTGCAGCTGGATTTTCAGATAAA	25	4	41	1	0	0	2
G680	ACTTTGCGGTAAATTATCCAAATGAGTTAT	30	4	30	2	0	0	12
G689	TAACCCAATTGGATAATTTACCGCAAA	27	3	33	6	0	0	8

ANEXO III

Dados do Crass (crass.crispr) para a camada sedimento. DR, *direct repeat*; ESP., espaçador; FL, flanqueador. T.m, tamanho médio; Co. m., cobertura média

GID	DR consenso	T.m. DR	Esp.	T.m. Esp.	Co. m. Esp.	FL.	T.m. FL.	Leituras
G6	AAATCCAAATGCGATTCATTTATTAGAACAAA	32	3	33	1	0	0	1
G11	AATTAAAAGATCTTAAGGGAATAGAAAATCTTACCAACTTG	41	3	25	2	0	0	2
G17	GCTTCAATGTTTCCGCGTGCAATTGCACGCGGAAAC	36	3	36	1	0	0	1
G32	CGAGATTGTCGAGCCCCTCGATCTTCC	27	4	38	1	0	0	1
G38	ATTTCAATTCTCTTTAATGAGACTTATCCTTTGCAACG	38	5	37	1	0	0	1
G50	AATTTTTATAACTGAATTTGGCAAATGTTCAAAGGATACTATTTG	45	4	15	1	0	0	1
G56	CAAATAACTGATCAAGGTTTAAAACATCTCAAAGG	35	3	31	1	0	0	18
G58	AAAGTTTGAATAATACTATTAAGGATTAAAAGTA	34	3	30	1	1	34	1