A peer-reviewed version of this preprint was published in PeerJ on 7 September 2017.

View the peer-reviewed version (peerj.com/articles/3788), which is the preferred citable publication unless you specifically need to cite this preprint.

Moller AG, Liang C. 2017. MetaCRAST: reference-guided extraction of CRISPR spacers from unassembled metagenomes. PeerJ 5:e3788 https://doi.org/10.7717/peerj.3788

Category: Applications Note, Sequence Analysis

MetaCRAST: Reference-guided extraction of CRISPR spacers from unassembled metagenomes

Abraham Moller[1] and Chun Liang[1,*]

[1]Department of Biology, Miami University, Oxford, Ohio 45056

[*]Corresponding Author: Chun Liang, liangc@miamioh.edu

Abstract

Summary: Clustered regularly interspaced palindromic repeat (CRISPR) systems are prokaryotic adaptive immune systems against viral infection. CRISPR spacer sequences can provide valuable ecological insights by linking environmental viruses to microbial hosts. Despite this importance, metagenomic CRISPR detection remains a major challenge. Here we present a reference-guided CRISPR spacer detection tool (Metagenomic CRISPR Reference-Aided Search Tool - MetaCRAST) that constrains searches based on user-specified direct repeats (DRs). These DRs could be expected from assembly or taxonomic profiles of metagenomes. Our evaluation shows MetaCRAST improves CRISPR spacer detection in real metagenomes compared to *de novo* CRISPR detection methods. Simulations show it performs better than *de novo* tools for Illumina metagenomes.

Availability and implementation: MetaCRAST is implemented in Perl and takes metagenomic sequence reads and direct repeat queries (FASTA) as input. It is freely available for download at https://github.com/molleraj/MetaCRAST.

Contact: liangc@miamioh.edu or mollera2@miamioh.edu

Supplementary information: Supplementary data are available at Bioinformatics online.

Keywords

Metagenomics, CRISPR, microbial ecology, virus-host interactions, repetitive sequences

1. Introduction

Clustered regularly interspaced palindromic repeat (CRISPR) arrays, which are found in many archaeal and bacterial genomes, may help us better understand the virus-host interactions that mediate nutrient cycling in many ecosystems. Acting as adaptive immune systems against viral infection, CRISPRs incorporate short spacers cleaved from viral DNA into the host genomes, providing a record of past infections and thus associations between viruses and prokaryotic hosts (Sorek *et al.*, 2008; Makarova *et al.*, 2013). This power of CRISPR spacers to determine viruses' host specificity has recently been exploited using metagenomes from many ecosystems (Sanguino *et al.*, 2015; Anderson *et al.*, 2011; Edwards *et al.*, 2015). While many tools exist for detecting CRISPRs in assembled genomes (Rousseau *et al.*, 2009; Grissa *et al.*, 2007; Bland *et al.*, 2007; Edgar, 2007), few exist for CRISPR detection in metagenomic reads (Skennerton; Skennerton *et al.*, 2013; Rho *et al.*, 2012). The repetitive nature of CRISPRs makes them difficult to assemble from metagenomes, necessitating special tools to detect them in unassembled reads. The tool Minced, a modified version of CRT, detects CRISPR spacers (Skennerton, 2013), while the tool Crass detects and assembles CRISPR arrays (Skennerton *et al.*, 2013), both from raw reads. Both Minced and Crass do not rely on prior knowledge of direct repeat sequences, making them *de novo* detection methods. Instead, they use heuristics to determine whether detected repeats are indeed CRISPRs. Such heuristics include threshold array lengths to avoid short, spurious CRISPR arrays and threshold repeat-spacer similarities to avoid arrays where spacers are too similar to repeats (Skennerton *et al.*, 2013; Grissa *et al.*, 2007; Bland *et al.*, 2007), which might indicate microsatellites rather than CRISPRs.

In this work, we present Metagenomic CRISPR Reference-Aided Search Tool (MetaCRAST), a novel reference-guided tool to improve CRISPR spacer detection in unassembled metagenomic sequencing reads. Unlike Minced and Crass, MetaCRAST constrains spacer detection by searching metagenomes for direct repeats (DRs) that the user specifies. Relationships amongst these tools and such differences in use are further illustrated in Supplementary Figure S1. Such specified DRs may be selected based on assembly or taxonomic profiling of metagenomic reads. MetaCRAST improves CRISPR annotation by allowing users to control for the taxonomic composition of the metagenome. It also avoids the rejection of true CRISPRs that can occur due to the heuristics required for *de novo* detection methods. In addition, unlike Crass and Minced, MetaCRAST provides consistent performance over different read lengths of Illumina datasets.

2. Methods

2.1 Algorithm and implementation

MetaCRAST can constrain spacer detection by expected host species' DRs or DRs identified from assembly (Figure 1). It searches each read for DR sequences matching query DRs specified by the user. These DRs can be selected from CRISPR arrays detected with genomic CRISPR detection tools (e.g. PILER-CR, CRISPRFinder) in fully assembled microbial genomes or assembled metagenomic contigs. In the first step of the pipeline, reads containing DRs within a certain Levenshtein edit distance (i.e., number of insertions, deletions, or substitutions necessary to convert one sequence to another) of the query DRs are quickly identified using the Wu-Manber multi-pattern search algorithm (Wu *et al.*, 1995). In the second step, individual reads found to contain

a query DR sequence are searched for two or more copies of the query DRs. In the third step, the sequence fragments between the DRs detected in these sequence reads are extracted into a unique spacer set, which are then clustered using CD-HIT into a non-redundant spacer set stored in FASTA format (Li and Godzik, 2006). MetaCRAST is implemented in Perl as a command line tool to analyze metagenomes in FASTA format. Optionally, the user can specify the maximum spacer length, the distance metric used for comparing DRs to reads (Hamming or Levenshtein), whether to search for the reverse complement of the DR, the CD-HIT similarity threshold for clustering spacers, and the maximum number of threads to use to parallelize the search.

2.2 Evaluation with simulated and real metagenomes

To study the relationship between CRISPR spacer detection and read length or sequencing technology, simulated acid mine drainage (AMD) and enhanced biological phosphorus removal (EBPR) metagenomes were generated using Grinder (Angly *et al.*, 2012), which are described in the Supplementary Data. We generated simulated metagenomes over a range of read lengths using models of 454 (Balzer *et al.*, 2010) and Illumina (Korbel *et al.*, 2009) errors. We used highly simplified taxonomic profiles to model the AMD and EBPR metagenomes (Supplementary Tables S2 and S3). Simulated metagenomes were searched for CRISPR spacers using Crass (Skennerton *et al.*, 2013), Minced (Skennerton, 2013), and MetaCRAST. Detected spacers were clustered with CD-HIT with a similarity threshold of 0.9.

Similarly, CRISPR spacers were also detected by the aforementioned three tools in real AMD and EBPR metagenomes downloaded from iMicrobe (Hurwitz, 2014) and taxonomically profiled with MetaPhyler (Liu *et al.*, 2011). MetaCRAST analyses of the

real metagenomes were performed with taxonomy- or assembly-guided query DRs generated as follows. To make an assembly-guided query, CAP3-assembled contigs (Huang and Madan, 1999) were searched for CRISPR DRs using PILER-CR (Edgar, 2007), which finds CRISPRs in assembled genomes or contigs. These DRs formed an assembly-guided query, while DRs found in assembled *Leptospirillum* (AMD), *Ferroplasma* (AMD), and Candidatus *Accumulibacter phosphatis* (EBPR) genomes included in CRISPRdb (Grissa *et al.*, 2007) formed a taxonomy-guided query.

3. Results and Discussion

3.1 Effect of read length and sequencing technology on CRISPR detection

We first investigated the relationships between detected spacers and read length or sequencing technology. Performance, here determined by the number of spacers detected, consistently increased with read length over all 454 tests (Supplementary Figure S2). While the total number of spacers detected by Crass and MetaCRAST converged as read length increased, the total number of spacers detected by Minced steadily increased even beyond the true number of spacers found in the genomes used to generate the simulated metagenomes. We speculate that Minced inconsistently determined DR lengths amongst different CRISPR-containing reads, leading to the same spacers being inappropriately truncated or extended. Meanwhile, amongst metagenomes simulated with the Illumina model, MetaCRAST detected significantly more spacers than Crass and Minced for average read lengths of 200 bp or greater (Supplementary Figure S2). Crass detected more spacers than Minced and MetaCRAST for short Illumina reads (100 and 150 bp), however (Supplementary Figure S2).

3.2 Evaluation on real AMD and EBPR metagenomes

We also evaluated MetaCRAST against Crass and Minced using real AMD and EBPR metagenomes (Tyson *et al.*, 2004; Martín *et al.*, 2006). While taxonomy-guided queries consistently found fewer spacers than the other two methods, an assembly-guided MetaCRAST search identified more spacers than Crass did in the AMD metagenome (Supplementary Figure S3). In both AMD and EBPR metagenomes, many common spacers were detected with Crass, MetaCRAST (assembly-guided query), and Minced (7.1% of all detected spacers for AMD and 2.5% for EBPR - see Supplementary Figures S4 and S5). Despite this, there were also many spacers detected with Crass and Minced not identified with MetaCRAST searches (Supplementary Figures S4 and S5). Notably, however, none of the spacers detected with MetaCRAST using the taxonomy-guided query overlapped with the Crass-detected spacers (Supplementary Figures S3 and S4), suggesting the power of the reference-guided MetaCRAST to complement Crass and Minced as a CRISPR detection strategy.

Acknowledgements

Thanks to Michael Crowder and Gary Lorigan (Miami University) for feedback on the project and manuscript.

Authors' contributions

AM and CL developed the proposed tool, analyzed the data, and wrote the manuscript. CL edited the manuscript. AM developed and performed the bioinformatics analyses.

Funding

The project was funded partially by Committee on Faculty Research (CFR) program, the Office for the Advancement of Research & Scholarship (OARS), and by an Academic Challenge grant from the Department of Biology (Miami University).

References

Anderson,R.E. *et al.* (2011) Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage: CRISPR spacers reveal hosts of marine vent viral assemblage. *FEMS Microbiol. Ecol.*, **77**, 120–133.

Angly,F.E. *et al.* (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94–e94.

Balzer,S. *et al.* (2010) Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

Bland,C. *et al.* (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.

Edgar,R.C. (2007) PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.

Edwards,R.A. *et al.* (2015) Computational approaches to predict bacteriophage–host relationships. *FEMS Microbiol. Rev.*, fuv048.

Grissa,I. *et al.* (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.

Huang,X. and Madan,A. (1999) CAP3: A DNA Sequence Assembly Program. *Genome Res.*, **9**, 868–877.

185 Hurwitz,B. (2014) iMicrobe: Advancing Clinical and Environmental Microbial Research
186 using the iPlant Cyberinfrastructure. Plant and Animal Genome.
187 Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error
188 models for inferring genomic structural variants from massive paired-end
189 sequencing data. *Genome Biol.*, **10**, R23.
190 Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large
191 sets of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
192 Liu,B. *et al.* (2011) Accurate and fast estimation of taxonomic profiles from
193 metagenomic shotgun sequences. *BMC Genomics*, **12**, 1–10.
194 Makarova,K.S. *et al.* (2013) Comparative genomics of defense systems in archaea and
195 bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
196 Martín,H.G. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus
197 removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.
198 Rho,M. *et al.* (2012) Diverse CRISPRs Evolving in Human Microbiomes. *PLoS Genet.*,
199 **8**, e1002441.
200 Rousseau,C. *et al.* (2009) CRISPI: a CRISPR interactive database. *Bioinformatics*, **25**,
201 3317–3318.
202 Sanguino,L. *et al.* (2015) Linking environmental prokaryotic viruses and their host
203 through CRISPRs. *FEMS Microbiol. Ecol.*, **91**, fiv046.
204 Skennerton,C. minced - Mining CRISPRs in Environmental Datasets. *GitHub*.
205 Skennerton,C.T. *et al.* (2013) Crass: identification and reconstruction of CRISPR from
206 unassembled metagenomic data. *Nucleic Acids Res.*, **41**, e105–e105.

Sorek,R. *et al.* (2008) CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.

Tyson,G.W. *et al.* (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, **428**, 37–43.

Wu,S. *et al.* (1995) A Subquadratic Algorithm for Approximate Regular Expression Matching. *J. Algorithms*, **19**, 346–360.

Figures

Figure 1: An outline of the MetaCRAST workflow (A) and comparison of per-read CRISPR detection strategies (B) between MetaCRAST and existing *de novo* detection tools (e.g., Crass, Minced). DR represents direct repeat, while S represents spacer.