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 phage infection. CRISPR spacer sequences can provide valuable ecological insights by linking environmental viruses to their microbial hosts. Despite this importance, metagenomic CRISPR detection remains a major challenge. Here we present a reference-guided CRISPR detection tool (Metagenomic CRISPR Reference-Aided Search Tool - MetaCRAST) that constrains searches based on user-specified direct repeats. These DRs could be expected from assembly or taxonomic profiles of metagenomic sequence data. Our evaluation shows MetaCRAST improves CRISPR spacer detection in real metagenomes compared to *de novo* detection methods. Simulations show it performs better than comparable tools when analyzing Illumina metagenomes.

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

Contact: 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 numerous 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). The repetitive nature of CRISPRs makes them difficult to assemble from metagenomes, necessitating special tools to detect them in unassembled reads. The tool Minced detects CRISPR spacers (Skennerton, 2013), while the tool Crass detects and assembles CRISPR arrays (Skennerton *et al.*, 2013), both from raw reads. These two tools do not rely on prior knowledge of direct repeat sequences, making them *de novo* detection methods. In this work, we present MetaCRAST, a new reference-guided tool to improve CRISPR detection in unassembled metagenomic sequencing reads. Unlike Minced and Crass, MetaCRAST constrains spacer detection by searching metagenomes for direct repeats (DRs) the user specifies. Relationships amongst these tools are further illustrated in Supplementary Figure S1. Such specified DRs may be selected based on taxonomic profiling or assembly of metagenomic reads. MetaCRAST

improves CRISPR annotation by allowing the user to control for the taxonomic composition of the metagenome, avoiding unexpected DR sequences. 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 may be selected from genomic CRISPR arrays or assembled metagenomic contigs, for example. 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 of the pipeline, 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 as spacers. These spacers are then collected as a non-redundant set, further clustered using CD-HIT (Li and Godzik, 2006), and stored in FASTA format. 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 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). Further details about the simulated metagenomes are provided in the Supplementary Data. Simulated metagenomes were searched for CRISPRs 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.

Real AMD and EBPR metagenomes were downloaded from iMicrobe (Hurwitz, 2014) and taxonomically profiled with MetaPhyler (Liu *et al.*, 2011). Real metagenomes were searched for CRISPRs as described previously, while CAP3-assembled contigs (Huang and Madan, 1999) were searched for CRISPRs using PILER-CR (Edgar, 2007), which finds CRISPRs in assembled genomes or contigs. MetaCRAST analyses of the real metagenomes were performed with query DRs identified from taxonomy or assembly. DRs detected with PILER-CR in the CAP3 contigs 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

To investigate the relationships between detected spacers and read length or sequencing technology, 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). 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 attribute this to the possibility 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. Crass detected more spacers than Minced and MetaCRAST for short Illumina reads (100 and 150 bp), however.

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 far fewer spacers than the other methods, an assembly-guided MetaCRAST search identified many more spacers than Crass did in the AMD metagenome (Supplementary Figure S3). In both AMD and EBPR metagenomes, there was significant overlap amongst spacers detected with Crass, MetaCRAST (assembly-guided query), and Minced (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.

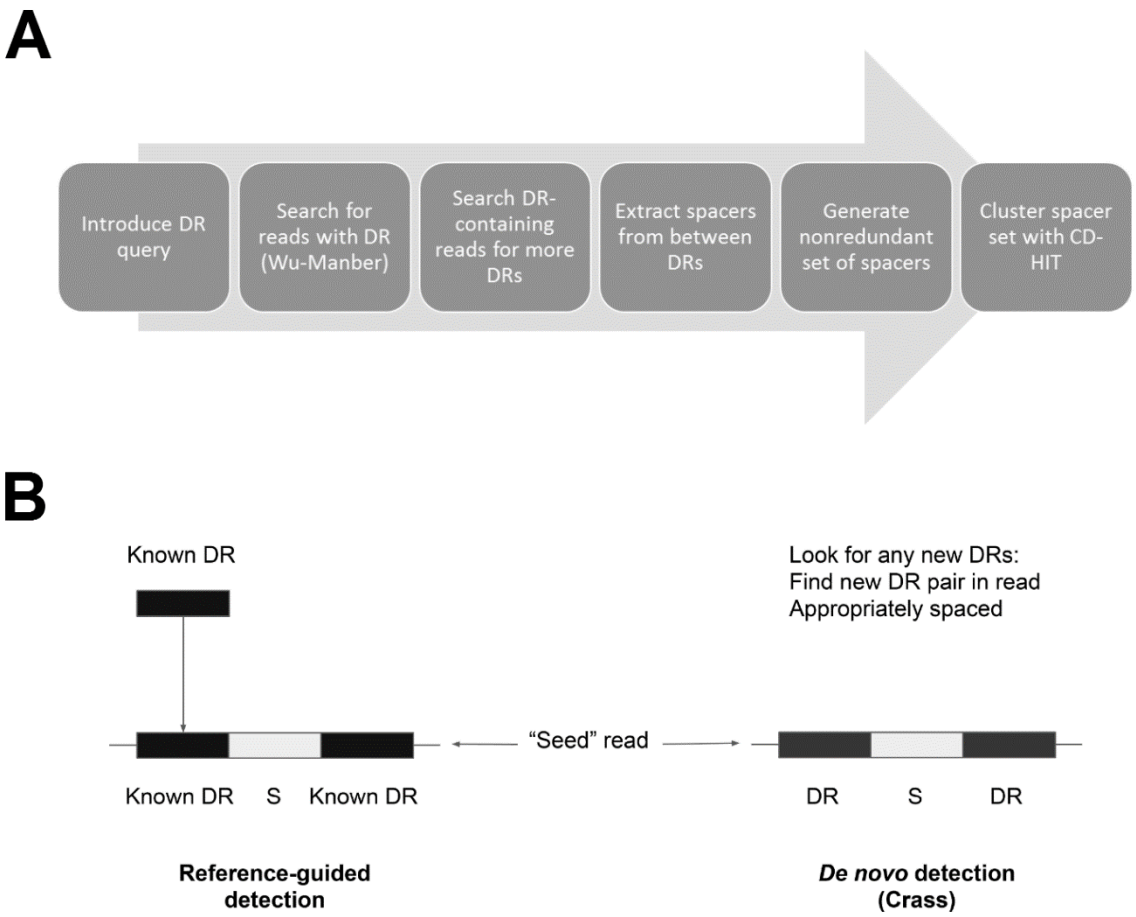
Balzer,S. *et al.* (2010) Characteristics of 454 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*, **26**, i420–i425.

161 Bland,C. *et al.* (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of
162 clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
163 Edgar,R.C. (2007) PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC*
164 *Bioinformatics*, **8**, 18.
165 Edwards,R.A. *et al.* (2015) Computational approaches to predict bacteriophage–host
166 relationships. *FEMS Microbiol. Rev.*, fuv048.
167 Grissa,I. *et al.* (2007) The CRISPRdb database and tools to display CRISPRs and to
168 generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
169 Huang,X. and Madan,A. (1999) CAP3: A DNA Sequence Assembly Program. *Genome*
170 *Res.*, **9**, 868–877.
171 Hurwitz,B. (2014) iMicrobe: Advancing Clinical and Environmental Microbial Research
172 using the iPlant Cyberinfrastructure. Plant and Animal Genome.
173 Korbel,J.O. *et al.* (2009) PEMer: a computational framework with simulation-based error
174 models for inferring genomic structural variants from massive paired-end
175 sequencing data. *Genome Biol.*, **10**, R23.
176 Liu,B. *et al.* (2011) Accurate and fast estimation of taxonomic profiles from metagenomic
177 shotgun sequences. *BMC Genomics*, **12**, 1–10.
178 Li,W. and Godzik,A. (2006) Cd-hit: a fast program for clustering and comparing large sets
179 of protein or nucleotide sequences. *Bioinformatics*, **22**, 1658–1659.
180 Makarova,K.S. *et al.* (2013) Comparative genomics of defense systems in archaea and
181 bacteria. *Nucleic Acids Res.*, **41**, 4360–4377.
182 Martín,H.G. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus
183 removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.

184 Rousseau,C. *et al.* (2009) CRISPI: a CRISPR interactive database. *Bioinformatics*, **25**,
185 3317–3318.
186 Sanguino,L. *et al.* (2015) Linking environmental prokaryotic viruses and their host through
187 CRISPRs. *FEMS Microbiol. Ecol.*, **91**, fiv046.
188 Skennerton,C. minced - Mining CRISPRs in Environmental Datasets. *GitHub*.
189 Skennerton,C.T. *et al.* (2013) Crass: identification and reconstruction of CRISPR from
190 unassembled metagenomic data. *Nucleic Acids Res.*, **41**, e105–e105.
191 Sorek,R. *et al.* (2008) CRISPR — a widespread system that provides acquired resistance
192 against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
193 Tyson,G.W. *et al.* (2004) Community structure and metabolism through reconstruction of
194 microbial genomes from the environment. *Nature*, **428**, 37–43.
195 Wu,S. *et al.* (1995) A Subquadratic Algorithm for Approximate Regular Expression
196 Matching. *J. Algorithms*, **19**, 346–360.
197
198
199
200
201
202
203
204
205
206

207 **Figures**

208 **Figure 1:** An outline of the MetaCRAST workflow (A) and comparison of per-read
209 CRISPR detection strategies (B) between MetaCRAST and existing *de novo* detection
210 tools (e.g., Crass, Minced).



211

212