

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

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

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

1 Classification: Microbiology, Ecology, Genomics, Bioinformatics, Computational Biology

2 Title: **MetaCRAST: Reference-guided extraction of CRISPR spacers from**

3 **unassembled metagenomes**

4 Abraham Moller^{1,†} and Chun Liang^{1,*}

5 ¹Department of Biology, Miami University, Oxford, Ohio 45056

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

7 †Current Address: Department of Microbiology and Immunology, Emory University,

8 Atlanta, Georgia 30322

9

10

11

12

13

14

15

16

17

18

19

20 **Abstract**

21 Clustered regularly interspaced short palindromic repeat (CRISPR) systems are
22 the adaptive immune systems of bacteria and archaea against viral infection. While
23 CRISPRs have been exploited as a tool for genetic engineering, their spacer sequences
24 can also provide valuable insights into microbial ecology by linking environmental viruses
25 to their microbial hosts. Despite this importance, metagenomic CRISPR detection
26 remains a major challenge. Here we present a reference-guided CRISPR spacer
27 detection tool (**Meta**genomic **CRISPR Reference-Aided Search Tool - Meta**CRAST) that
28 constrains searches based on user-specified direct repeats (DRs). These DRs could be
29 expected from assembly or taxonomic profiles of metagenomes. We compared the
30 performance of MetaCRAST to those of two existing metagenomic CRISPR detection
31 tools – Crass and MinCED – using both real and simulated acid mine drainage (AMD)
32 and enhanced biological phosphorus removal (EBPR) metagenomes. Our evaluation
33 shows MetaCRAST improves CRISPR spacer detection in real metagenomes compared
34 to the *de novo* CRISPR detection methods Crass and MinCED. Evaluation on simulated
35 metagenomes show it performs better than *de novo* tools for Illumina metagenomes and
36 comparably for 454 metagenomes. It also has comparable performance dependence on
37 read length and community composition, run time, and accuracy to these tools
38 MetaCRAST is implemented in Perl, parallelizable through the Many Core Engine (MCE),
39 and takes metagenomic sequence reads and direct repeat queries (FASTA or FASTQ)
40 as input. It is freely available for download at <https://github.com/molleraj/MetaCRAST>.

41

42

43 Introduction

44 The clustered regularly interspaced short palindromic repeat (CRISPR) arrays
45 found in prokaryotic genomes can help us better understand viral-microbial interactions
46 important in many ecosystems. Viruses can release cellular nutrients back into the
47 ecosystem through lytic infection, forming an ecological short-circuit called the viral shunt
48 (Weitz & Wilhelm, 2012). In this manner, viruses not only contribute to nutrient cycling in
49 individual ecosystems, but also to maintaining biogeochemical cycles on a broader scale.
50 The short spacers of viral DNA incorporated into CRISPR arrays form a historical record
51 of past infections, thus linking virus to host (Sorek, Kunin & Hugenholtz, 2008; Makarova,
52 Wolf & Koonin, 2013). This power of CRISPR spacers to determine viruses' host
53 specificity has recently been exploited using metagenomes from many ecosystems
54 (Anderson, Brazelton & Baross, 2011; Sanguino et al., 2015; Edwards et al., 2015). While
55 many tools exist for detecting CRISPRs in assembled genomes (Bland et al., 2007;
56 Edgar, 2007; Grissa, Vergnaud & Pourcel, 2007a; Rousseau et al., 2009), few exist for
57 CRISPR detection in metagenomic reads (Rho et al., 2012; Skennerton, Imelfort & Tyson,
58 2013; Skennerton).

59 The repetitive nature of CRISPRs makes them difficult to assemble from
60 metagenomes, necessitating special tools to detect them in unassembled reads. Several
61 tools have been developed to detect and assemble CRISPR arrays in unassembled reads
62 rather than assembled contigs. The tool MinCED (Mining CRISPRs in Environmental
63 Datasets), like metaCRT (Rho et al., 2012), is a modified version of CRT (Bland et al.,
64 2007) that detects CRISPR spacers (Skennerton), while the tool Crass (CRISPR
65 asssembler) detects and assembles CRISPR arrays (Skennerton, Imelfort & Tyson, 2013),

66 both from raw metagenomic reads. MinCED searches each read for CRISPRs using the
67 same strategy as CRT; it searches for appropriately spaced short k-mers from which it
68 extends longer repeats if appropriately frequent nucleotides are identified at the ends of
69 the growing repeats. Crass relies on a hybrid algorithm to detect spacers that blends
70 strategies of CRT (Bland et al., 2007) and CRISPRFinder (Grissa, Vergnaud & Pourcel,
71 2007b). In long reads (>177 bp), it searches for repeats using the CRT strategy previously
72 described. In short reads (<177 bp), on the other hand, it searches for appropriately
73 spaced full-length repeats (i.e., 20-50 bp) and extends these repeats only with identical
74 nucleotides, thus avoiding the potential errors caused by the CRT algorithm over- or
75 under-extending the few repeats found in a short sequence. Crass then searches further
76 for reads containing a single repeat, determines consensus direct repeats, uses the first
77 and last k-mers of detected spacers to build a graph of spacer arrangement, and
78 assembles CRISPR arrays using this graph. Both MinCED and Crass do not rely on prior
79 knowledge of direct repeat sequences, making them *de novo* detection methods. Instead,
80 they use heuristics to determine whether detected repeats are indeed CRISPRs. Such
81 heuristics include threshold array lengths to avoid short, spurious CRISPR arrays and
82 threshold repeat-spacer similarities to avoid arrays where spacers are too similar to
83 repeats (Bland et al., 2007; Grissa, Vergnaud & Pourcel, 2007a; Skennerton, Imelfort &
84 Tyson, 2013), which might indicate microsatellites rather than CRISPRs.

85 In this work, we present the Metagenomic CRISPR Reference-Aided Search Tool
86 (MetaCRAST), a novel reference-guided tool to improve CRISPR spacer detection in
87 unassembled metagenomic sequencing reads. While MetaCRAST, to our knowledge, is
88 the first reference-guided, read-dependent metagenomic CRISPR detection tool, prior

89 studies have used known direct repeats to improve CRISPR detection. The genomic
90 CRISPR identification algorithm CRISPRDetect matches newly identified direct repeats
91 to a reference library to refine repeat boundaries and validate arrays (Biswas et al., 2016).
92 Searching reference repeat libraries, together with annotating *cas* genes adjacent to
93 CRISPR arrays, has been used to exclude false positive “putative” CRISPRs from
94 CRISPR annotation (Zhang & Ye, 2017). Unlike MinCED and Crass, as a reference-
95 guided method, MetaCRAST constrains spacer detection by searching metagenomes for
96 direct repeats (DRs) that the user specifies. Relationships amongst these tools and such
97 differences in use are further illustrated in Figure 1. Such specified DRs may be selected
98 based on assembly or taxonomic profiling of metagenomic reads. MetaCRAST improves
99 CRISPR annotation by allowing users to control for the taxonomic composition of the
100 metagenome. It also avoids the rejection of true CRISPRs that can occur due to the
101 heuristics required for *de novo* detection methods. In addition, unlike Crass and MinCED,
102 MetaCRAST provides consistent performance over different read length Illumina
103 datasets.

104 **Materials and Methods**

105 **Algorithm and implementation**

106 MetaCRAST can constrain spacer detection by expected host species’ DRs or
107 DRs identified from assembly (Figure 2A). It searches each read for DR sequences
108 matching query DRs specified by the user. These DRs can be selected from CRISPR
109 arrays detected with genomic CRISPR detection tools such as PILER-CR (Edgar, 2007),
110 CRF (Wang & Liang, 2017), or CRISPRFinder (Grissa, Vergnaud & Pourcel, 2007b) in
111 fully assembled microbial genomes or assembled metagenomic contigs. The steps of the

112 MetaCRAST pipeline are outlined in Figure 2B. In the first step of the pipeline, reads
113 containing DRs within a certain Levenshtein edit distance (i.e., number of insertions,
114 deletions, or substitutions necessary to convert one sequence to another) of the query
115 DRs are quickly identified using the Wu-Manber multi-pattern search algorithm (Wu,
116 Manber & Myers, 1995). In the second step, individual reads found to contain a query DR
117 sequence are searched for two or more copies of the query DRs. In the third step, the
118 sequence fragments between the DRs detected in these sequence reads are extracted
119 into a comprehensive spacer set, which are then clustered using CD-HIT into a non-
120 redundant unique spacer set stored in FASTA format (Li & Godzik, 2006, p.).

121 MetaCRAST is implemented in Perl as a command line tool to analyze
122 metagenomes in FASTA or FASTQ formats. The tool has been implemented in several
123 versions that differ in metagenome loading method (using BioPerl or readfq, the latter of
124 which was paired either with the standard open routine to load a single file or mce_open
125 for parallel file loading). Optionally, the user can specify the maximum spacer length, the
126 distance metric used for comparing DRs to reads (Hamming or Levenshtein), whether to
127 search for the reverse complement of the DR, the CD-HIT similarity threshold for
128 clustering spacers, and the maximum number of threads to use to parallelize the search.
129 The reverse complement argument (-r) should be used when the CRISPR direction is
130 unknown. When the search is run in parallel, the FASTA (or FASTQ) file is split based on
131 the specified number of threads. All command line arguments are further described in
132 Table 1. Each split file is searched in parallel. An additional tool has been provided to
133 assist taxonomy-guided query selection. This tool searches a taxonomically-annotated

134 library of CRISPRdb DRs for those that belong to a particular taxon query (e.g.,
135 *Streptococcus*).

136 To analyze the distribution of taxonomic affiliations to direct repeats, we examined
137 all direct repeats found in microbial genomes using the CRISPRdb database. CRISPRdb
138 provides a library of direct repeats labeled with respective GenBank accessions in the
139 CRISPR utilities section of the database (Grissa, Vergnaud & Pourcel, 2007a). We
140 processed this library to assign taxonomy information based on GenBank accession.
141 Taxonomy information was extracted from GenBank records with the Perl module
142 Bio::DB::GenBank. Statistics describing the distribution of unique binomial names or
143 genres to which individual direct repeats affiliated was compiled with Microsoft Excel.
144 Binomial name (species-level) and genus statistics are presented in Table 2.

145 **Performance evaluation with simulated and real metagenomes**

146 To study the relationship between CRISPR spacer detection and read length or
147 sequencing technology, simulated acid mine drainage (AMD) and enhanced biological
148 phosphorus removal (EBPR) metagenomes were generated using Grinder (Angly et al.,
149 2012). We generated simulated metagenomes over a range of average read lengths (100
150 to 600 base pairs) using models of 454 (Balzer et al., 2010) and Illumina (Korbel et al.,
151 2009) errors. Following previous studies, we used a fourth-degree polynomial ($3e-3 +$
152 $3.3e-8 * i^4$, where i is the nucleotide position from the 5' end of the read, and the output
153 is percentage chance of an error at that position) to model the Illumina sequencing error
154 rate (Dohm et al., 2008; Korbel et al., 2009; Angly et al., 2012). This polynomial
155 determined the probability of substitution, insertion, or deletion at each base of a
156 simulated read (Korbel et al., 2009). For Illumina simulations, the ratio of substitutions to

157 insertions and deletions was set to 80:20 by default. For 454 metagenome simulations,
158 we modeled homopolymer errors as homopolymer length variation within simulated
159 reads. The distributions of homopolymer lengths were defined by the mean n and
160 standard deviation $0.03494 + n * 0.06856$, where n is the homopolymer length, based on
161 a prior study (Balzer et al., 2010; Angly et al., 2012).

162 We generated six simulated metagenomes per condition (average read length,
163 model, and microbial community). We used highly simplified taxonomic profiles to model
164 the AMD and EBPR metagenomes (Tables S1 and S2). To test the effects of community
165 composition on spacer detection, we simulated the AMD metagenome with a 454 error
166 model and 600 bp average read length, varying the relative proportions of *Leptospirillum*
167 and *Ferroplasma* genome used for the simulation (i.e., from 0 to 100% *Leptospirillum*).
168 All simulated metagenomes contained 100,000 reads. 454 metagenomes were
169 generated with this command: `grinder -reference_file AMDgenomes.fasta -`
170 `abundance_file AMDprofile.txt -total_reads 100000 -read_dist (one of 100, 150, 200, 250,`
171 `300, 400, or 600) normal 50 -homopolymer_dist balzer`. All 454 read length distributions
172 were normal with a standard deviation of 50 bp. Illumina metagenomes were generated
173 with this command: `grinder -reference_file AMDgenomes.fasta -abundance_file`
174 `AMDprofile.txt -total_reads 100000 -read_dist (one of 100, 150, 200, 250, or 300) -md`
175 `poly4 3e-3 3.3e-8`. All Illumina read length distributions were uniform with all reads having
176 exactly the average read length.

177 Simulated metagenomes were searched for CRISPR spacers using Crass
178 (Skenneron, Imelfort & Tyson, 2013), MinCED (Skenneron), and MetaCRIST. Crass
179 and MinCED were run with default parameters (`crass grinder-reads.fa; minced -spacers`

180 *grinder-reads.fa minced.crispr*). The default minimum and maximum DR lengths for both
181 Crass and MinCED were 23 and 47 bp. The default minimum and maximum spacer
182 lengths for both Crass and MinCED were 26 and 50 bp. MetaCRAST was run with a
183 taxonomy-guided query (Tables S3 and S4), a maximum spacer length of 60, a maximum
184 allowed edit distance (insertions, deletions, or substitutions) between query and target
185 direct repeats of 3, a CD-HIT clustering similarity threshold of 0.9, and a total of 16 parallel
186 threads (*MetaCRAST -p query.fa -i grinder-reads.fa -o MetaCRAST -d 3 -l 60 -c 0.90 -a*
187 *0.90 -n 16 -t tmp*). We selected a maximum allowed edit distance of 3 based on results
188 of our prior metagenomic CRISPR detection studies, which showed MetaCRAST
189 searches with a taxonomy-guided query found similar numbers of spacers to Crass when
190 we set this edit distance (Moller & Liang, 2017). For all analyses, detected spacers were
191 clustered with CD-HIT with a similarity threshold of 0.9 (*cdhit -i spacers.fa -o*
192 *spacersCD90.fa -c 0.9*) to reduce spacer redundancy. Performance on these simulated
193 metagenomes was evaluated based on total number of spacers detected, number of false
194 positive spacers detected, and run time for each average read length. For the mixed
195 composition simulated AMD metagenomes described above, spacers were aligned
196 against CRISPR spacers present in the source *Leptospirillum* and *Ferroplasma* genomes
197 and the number of matching true positive spacers for each organism reported.

198 The number of false positive spacers found in simulated metagenomes was
199 determined by comparing the total detected spacers with the expected CRISPRdb
200 spacers found in the source genomes used for the simulations (AMD and EBPR).
201 Alignments were made to the annotated CRISPRdb spacers using BLAST with an E-
202 value cutoff of 1e-6 (Altschul et al., 1990). This analysis was repeated with an E-value

203 cutoff of $1e-1$ to consider whether the original threshold was too stringent. The number
204 of detected spacers that were aligned to expected ones was subtracted from the total
205 number of spacers detected to determine the number of false positive spacers for a
206 particular method and condition. Cases where zero spacers were detected in a
207 metagenome were treated as zero false positive spacers and included in overall analysis.
208 Run times were determined for each metagenome and method using the built-in Linux
209 command *time*. Run time was calculated as the sum of the user and system time (together
210 the total CPU time).

211 Similarly, CRISPR spacers were also detected by the aforementioned three tools
212 in real AMD and EBPR metagenomes (Table S5) downloaded from iMicrobe (Hurwitz,
213 2014) and taxonomically profiled with MetaPhyler (Liu et al., 2011). MetaCRIST analyses
214 of the real metagenomes were performed with taxonomy- or assembly-guided query DRs
215 generated as follows. To make an assembly-guided query, CAP3-assembled contigs
216 (Huang & Madan, 1999) were searched for CRISPR DRs using PILER-CR (Edgar, 2007),
217 which finds CRISPRs in assembled genomes or contigs. These DRs formed an
218 assembly-guided query (Tables S6 and S7), while DRs found in assembled *Leptospirillum*
219 (AMD), *Ferroplasma* (AMD), and Candidatus *Accumulibacter phosphatis* (EBPR)
220 genomes included in CRISPRdb (Grissa, Vergnaud & Pourcel, 2007a) formed a
221 taxonomy-guided query (Tables S3 and S4). All of these aforementioned taxa were found
222 to be major components of the microbial community based on the AMD and EBPR
223 taxonomic profiles determined with MetaPhyler (Tables S8 and S9).

224

225

226 Results

227 Effects of read length, sequencing technology, and community composition on 228 CRISPR spacer detection

229 We first investigated the relationships between detected spacers and read length
230 or sequencing technology. Performance, here determined by the number of spacers
231 detected, increased with read length over all 454 tests (Figure 3). While the total number
232 of spacers detected by Crass and MetaCRASST converged as read length increased, the
233 total number of spacers detected by MinCED steadily increased even beyond the true
234 number of spacers found in the genomes used to generate the simulated metagenomes.
235 We speculate that MinCED inconsistently determined DR lengths amongst different
236 CRISPR-containing reads due to its CRT-based algorithm, leading to the same spacers
237 being inappropriately truncated or extended. Meanwhile, amongst metagenomes
238 simulated with the Illumina model, MetaCRASST detected significantly more spacers than
239 Crass and MinCED for average read lengths of 200 bp or greater (Figure 3; $p < 0.05$ for
240 both AMD and EBPR simulations using unpaired t-tests). Crass detected more spacers
241 than MinCED and MetaCRASST for short Illumina reads (100 and 150 bp), however
242 (Figure 3; $p < 0.05$ for both AMD and EBPR simulations using unpaired t-tests).

243 We also tested the effects of community composition on CRISPR detection for
244 each of the three methods using AMD metagenomes simulated with a 454 error model
245 and 600 bp average read length. We selected the 600 bp average read length for all
246 mixed metagenomes to minimize differences in detection between methods based on
247 read length (Figure 3). We varied the relative abundances of *Leptospirillum* and
248 *Ferroplasma* from 0 to 100 percent in our taxonomic profiles, thus varying the proportions

249 of CRISPR arrays specific to each included in the simulated metagenomes. For all
250 detection methods, detected spacers specific to a genome decreased as the relative
251 proportion of that taxon decreased, with roughly the same pattern for each method (Figure
252 4). As in the read length studies, MinCED consistently detected far more genome-specific
253 spacers in the metagenomes than were originally present in the source genomes (Figure
254 4). This may account for its steeper increase in detected genome-specific spacers as the
255 proportion of the corresponding genome in the simulated metagenomes increased.

256 **Evaluation of CRISPR spacer detection on real AMD and EBPR metagenomes**

257 We also evaluated MetaCRAST against Crass and MinCED using real AMD and
258 EBPR metagenomes (Tyson et al., 2004; Martín et al., 2006). While taxonomy-guided
259 queries consistently found fewer spacers than the other two methods (583 compared to
260 2486 for Crass and 4265 for MinCED in the AMD metagenome; 196 compared to 1014
261 for Crass and 1821 for MinCED in the EBPR metagenome), an assembly-guided
262 MetaCRAST search identified more spacers than Crass did in the AMD metagenome
263 (2813 compared to 2486 - Figure 5A). In both AMD and EBPR metagenomes, many
264 common spacers were detected amongst Crass, MetaCRAST (assembly-guided query),
265 and MinCED (7.1% of all detected spacers for AMD and 2.5% for EBPR – Figures 5B and
266 5C). Despite this, there were also many spacers detected with Crass and MinCED not
267 identified with MetaCRAST searches (Figures 5B and 5C). Notably, however, none of the
268 spacers detected with MetaCRAST using the taxonomy-guided query overlapped with the
269 Crass-detected spacers (Figures 5B and 5C), suggesting MetaCRAST can detect
270 spacers missed by Crass given an appropriate taxonomy-guided query.

271

272 **Evaluation of accuracy and runtime performance**

273 In addition to our studies comparing detected spacers over a variety of conditions,
274 we evaluated all three detection methods for spacer detection accuracy and run time
275 (Figures 6 and 7). We performed these evaluations on the simulated AMD and EBPR
276 metagenomes previously used to examine effects of read length and sequencing
277 technology on CRISPR detection (Figure 3). For AMD metagenomes simulated with the
278 454 model, MinCED detected significantly more false positive spacers than Crass or
279 MetaCRAST for average read lengths of 200 bp or more (Figure 6; $p < 0.05$ using
280 unpaired t-tests). Crass and MetaCRAST, on the other hand, did not have statistically
281 significant differences in detected false positive spacers over the entire range of average
282 read lengths ($p > 0.05$ using unpaired t-tests). For the AMD Illumina metagenomes, on
283 the other hand, MetaCRAST generated the largest number of false positive spacers for
284 average read lengths greater than 200 bp (Crass for average read lengths of 150 bp and
285 lower), but not by a statistically significant margin compared with MinCED ($p > 0.05$ using
286 unpaired t-tests). For the EBPR metagenomes simulated with the 454 model, there were
287 remarkably few false positive spacers detected with all methods over the full range of
288 average read lengths. For the EBPR Illumina metagenomes, MinCED generated the
289 largest number of false positive spacers for average read lengths greater than 200 bp
290 (Crass for average read lengths of 150 bp and lower), with MetaCRAST overlapping its
291 pattern closely (Figure 6). Because of this overlap, differences between MinCED and
292 MetaCRAST false positive spacers were not statistically significant ($p > 0.05$ using
293 unpaired t-tests), (EBPR Illumina metagenomes, Figure 6). MetaCRAST did detect more
294 false positives than MinCED for the 200 bp read length ($p < 0.05$ using unpaired t-tests,

295 EBPR Illumina metagenomes, Figure 6). We note that these false positive spacers are
296 only detected spacers that did not align to expected ones. The false positives do not
297 necessarily include improperly truncated or extended spacers, which we suspect MinCED
298 creates, leading to its artificially high spacer counts (Figure 3). We repeated this false
299 positive spacer analysis using a weaker E-value threshold of $1e-1$ (Figure S1). Using this
300 weaker threshold decreased the number of false positive spacers identified in all
301 conditions (Figure S1).

302 We also evaluated relative speed of the detection methods using the Linux function
303 time. We evaluated seven different combinations of algorithms, implementations, and
304 parameters. We evaluated both Crass and MinCED with default parameters. For
305 MetaCRAST, we evaluated five different conditions differing in parallelization and
306 metagenome loading method - BioPerl for loading and 16 threads, BioPerl and a single
307 thread, readfq with `mce_open` for loading and 16 threads, readfq with `mce_open` and a
308 single thread, and readfq with the standard open routine and a single thread (Figure 7).
309 We used CPU time (user and system time) rather than wall clock time (real time) as a
310 measure of speed performance.

311 We noticed steady increases in run time with increasing read length for all
312 detection methods, metagenomes, and sequencing technologies (Figure 7). MetaCRAST
313 showed a linear CPU time dependence on read length in all cases ($R^2 > 0.98$ in all cases;
314 p-values calculated from Pearson correlation were less than $1e-5$ in all cases), while
315 linear correlations for MinCED and crass were much weaker ($R^2 < 0.88$ in all cases; p-
316 values calculated from the Pearson correlations were more than 0.05 for Illumina datasets
317 but between $9e-4$ and $8e-3$ for 454 datasets). Among MetaCRAST implementations, the

318 readfq/open version used the least CPU time by statistically significant margins for all
319 conditions (Figure 7; $p < 0.05$ in all cases using unpaired t-tests). MetaCRAST was slower
320 than Crass for all read lengths by statistically significant margins (Figure 7; $p < 0.05$ in all
321 cases using unpaired t-tests). On the other hand, it was faster than MinCED for 454 read
322 lengths between 100 and 400 bp and Illumina read lengths between 100 and 250 bp
323 (Figure 7; $p < 0.05$ using unpaired t-tests).

324 **Taxonomic affiliations of CRISPR direct repeats annotated in CRISPRdb**

325 To analyze how direct repeats affiliated to taxa, we examined all direct repeats
326 annotated in microbial genomes using the CRISPRdb database. We used a Perl script to
327 assign taxonomy information based on GenBank accession using the module
328 Bio::DB::GenBank. The results of this analysis for species (binomial name) and genus-
329 level designations are presented in Table 2. The average number of unique taxon
330 designations per DR was greater at the species level than the genus level (1.308
331 compared to 1.063). Variation was also greater for species-level designations compared
332 to genus-level (standard deviation of 1.567 compared 0.521). Both species- and genus-
333 level analyses identified DRs that were affiliated to many taxa (a maximum of 20 genres
334 and 46 species). We acknowledge that our analysis does not examine the number of
335 unique DRs per taxon. It also only considers independent, unique DRs, ignoring the
336 possibility that many unique DRs may have closely related sequences.

337 **Discussion**

338 In this work, we present and evaluate a novel reference-guided method for
339 CRISPR detection in unassembled metagenomic reads. This method searches

340 metagenomic reads for user-specified direct repeats which could be provided through
341 taxonomy- or assembly-guided searches (Figures 1 and 2). We analyzed currently known
342 DRs with respect to their taxonomic designations to determine the robustness of
343 taxonomy-guided searches (Table 2). We found that most DRs in fact do affiliate to a
344 single species or genus, but that there are exceptions that may have arisen through
345 horizontal gene transfer (Table 2). This analysis does not consider small polymorphisms
346 between closely related DRs. Depending on the circumstance, it may be important to
347 consider whether one DR could be present in multiple taxa found in a sample.

348 Our studies of simulated metagenomes show distinct advantages for Crass and
349 MetaCRAST depending on average read length (Figure 3). While the modified assembly
350 procedure and exhaustive searches Crass provides make it well suited for short read 454
351 and Illumina metagenomes, MetaCRAST outperforms Crass for long read Illumina
352 metagenomes (Figure 3). We speculate that heuristics to avoid misassembly of CRISPR
353 arrays or improper repeat detection may hinder Crass in these long-read Illumina
354 metagenomes. We also noted that all three algorithms detected far more spacers in 454
355 compared to Illumina metagenomes (Figure 3). We have two possible explanations for
356 this phenomenon. First, our algorithms may have handled homopolymer error better than
357 the substitution error simulated in the Illumina metagenomes. Second, our Illumina model
358 may have introduced higher error rates than the 454 error model, making it more difficult
359 to find multiple similar DRs in the reads. The very high numbers of MinCED-detected
360 spacers are deceptive because this algorithm has the potential for substantial errors in
361 determining repeat and spacer lengths (Figures 3 and 4). Inconsistencies in defining
362 repeat length leads to false splitting of identical spacers into different groups.

363 Studies on real metagenomes suggest substantial advantages for Crass and
364 MinCED in terms of numbers of detected spacers (Figure 5). While in most cases
365 MetaCRAST detected fewer spacers than Crass or MinCED, it did identify spacers unique
366 to those from the two other methods. This suggests that it can complement these
367 methods, finding spacers missed due to the heuristics that Crass and MinCED use to
368 avoid false positives (Figure 5). We had expected that MetaCRAST would underperform
369 compared to Crass and MinCED in these real metagenomes, because the taxonomy-
370 guided queries we used did not fully account for all the taxa found with taxonomic profiling.
371 We only used one or two genomes to simulate the AMD and EBPR metagenomes,
372 making the simulated metagenomes much simpler in taxonomic diversity. This
373 simplification was what made MetaCRAST detection performance comparable to that of
374 Crass and MinCED for the simulations.

375 Accuracy was roughly similar amongst the three tools (Figure 6). Relaxing the error
376 threshold reduced false positive spacers detected by all tools, suggesting sequencing
377 error rather than algorithm issues could account for some of these false positive spacers
378 (Figure S1). MetaCRAST follows the same pattern of increasing run time with average
379 read length as the other two tools, and it is comparable in run time to MinCED (Figure 7).
380 MetaCRAST run time increases linearly with average read length (Figure 7). We
381 acknowledge that implementation of the algorithm in a compiled language or increasing
382 the number of threads used to parallelize the search could further improve MetaCRAST
383 speed. Nonetheless, while MetaCRAST is not as fast as the compiled algorithm Crass
384 under the conditions tested, it does identify spacers distinct from these methods in real

385 metagenomes and outperforms it in overall spacer detection for simulated Illumina
386 metagenomes.

387 Recent studies of computational methods for determining phage-host interactions
388 suggest CRISPR spacer alignment is a highly accurate signature of phage-host
389 interaction but that most identified CRISPR spacers do not align to known phage
390 genomes (Edwards et al., 2015). This suggests that it is critical to improve metagenomic
391 CRISPR spacer detection to increase the chances of matching spacers to viral genomes.
392 More broadly, increasing spacer matching would provide a fuller appreciation of a
393 microbial ecosystem's phage-host interaction space. We have recently used MetaCRIST
394 to improve our determination of virus-host interactions in solar salterns (Moller & Liang,
395 2017), complementing Crass with our spacer detection method. MetaCRIST
396 complements *de novo* methods like Crass because it avoids the heuristics they use to
397 reduce false positive spacers. Using a targeted direct repeat query, our tool can avoid the
398 false negative bias of these approaches. We anticipate that MetaCRIST will be of great
399 interest to microbial ecologists interested in phage-host interactions because it
400 complements existing *de novo* methods to improve metagenomic CRISPR detection.

401 **Acknowledgements**

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

404

405

406

407 **References**

- 408 Altschul SF., Gish W., Miller W., Myers EW., Lipman DJ. 1990. Basic local alignment
409 search tool. *Journal of Molecular Biology* 215:403–410. DOI: 10.1016/S0022-
410 2836(05)80360-2.
- 411 Anderson RE., Brazelton WJ., Baross JA. 2011. Using CRISPRs as a metagenomic tool
412 to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage:
413 CRISPR spacers reveal hosts of marine vent viral assemblage. *FEMS*
414 *Microbiology Ecology* 77:120–133. DOI: 10.1111/j.1574-6941.2011.01090.x.
- 415 Angly FE., Willner D., Rohwer F., Hugenholtz P., Tyson GW. 2012. Grinder: a versatile
416 amplicon and shotgun sequence simulator. *Nucleic Acids Research* 40:e94–e94.
417 DOI: 10.1093/nar/gks251.
- 418 Balzer S., Malde K., Lanzén A., Sharma A., Jonassen I. 2010. Characteristics of 454
419 pyrosequencing data—enabling realistic simulation with flowsim. *Bioinformatics*
420 26:i420–i425. DOI: 10.1093/bioinformatics/btq365.
- 421 Biswas A., Staals RHJ., Morales SE., Fineran PC., Brown CM. 2016. CRISPRDetect: A
422 flexible algorithm to define CRISPR arrays. *BMC Genomics* 17:356. DOI:
423 10.1186/s12864-016-2627-0.
- 424 Bland C., Ramsey TL., Sabree F., Lowe M., Brown K., Kyrpides NC., Hugenholtz P.
425 2007. CRISPR Recognition Tool (CRT): a tool for automatic detection of
426 clustered regularly interspaced palindromic repeats. *BMC Bioinformatics* 8:209.
427 DOI: 10.1186/1471-2105-8-209.

- 428 Dohm JC., Lottaz C., Borodina T., Himmelbauer H. 2008. Substantial biases in ultra-
429 short read data sets from high-throughput DNA sequencing. *Nucleic Acids*
430 *Research* 36:e105. DOI: 10.1093/nar/gkn425.
- 431 Edgar RC. 2007. PILER-CR: Fast and accurate identification of CRISPR repeats. *BMC*
432 *Bioinformatics* 8:18. DOI: 10.1186/1471-2105-8-18.
- 433 Edwards RA., McNair K., Faust K., Raes J., Dutilh BE. 2015. Computational
434 approaches to predict bacteriophage–host relationships. *FEMS Microbiology*
435 *Reviews*:fuv048. DOI: 10.1093/femsre/fuv048.
- 436 Grissa I., Vergnaud G., Pourcel C. 2007a. The CRISPRdb database and tools to display
437 CRISPRs and to generate dictionaries of spacers and repeats. *BMC*
438 *Bioinformatics* 8:172. DOI: 10.1186/1471-2105-8-172.
- 439 Grissa I., Vergnaud G., Pourcel C. 2007b. CRISPRFinder: a web tool to identify
440 clustered regularly interspaced short palindromic repeats. *Nucleic Acids*
441 *Research* 35:W52–W57. DOI: 10.1093/nar/gkm360.
- 442 Huang X., Madan A. 1999. CAP3: A DNA Sequence Assembly Program. *Genome*
443 *Research* 9:868–877. DOI: 10.1101/gr.9.9.868.
- 444 Hurwitz B. 2014. iMicrobe: Advancing Clinical and Environmental Microbial Research
445 using the iPlant Cyberinfrastructure. In: *Plant and Animal Genome*,.
- 446 Korb J., Abyzov A., Mu XJ., Carriero N., Cayting P., Zhang Z., Snyder M., Gerstein
447 MB. 2009. PEMer: a computational framework with simulation-based error
448 models for inferring genomic structural variants from massive paired-end
449 sequencing data. *Genome Biology* 10:R23. DOI: 10.1186/gb-2009-10-2-r23.

- 450 Li W., Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of
451 protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. DOI:
452 10.1093/bioinformatics/btl158.
- 453 Liu B., Gibbons T., Ghodsi M., Treangen T., Pop M. 2011. Accurate and fast estimation
454 of taxonomic profiles from metagenomic shotgun sequences. *BMC Genomics*
455 12:S4. DOI: 10.1186/1471-2164-12-S2-S4.
- 456 Makarova KS., Wolf YI., Koonin EV. 2013. Comparative genomics of defense systems
457 in archaea and bacteria. *Nucleic Acids Research* 41:4360–4377. DOI:
458 10.1093/nar/gkt157.
- 459 Martín HG., Ivanova N., Kunin V., Warnecke F., Barry KW., McHardy AC., Yeates C.,
460 He S., Salamov AA., Szeto E., Dalin E., Putnam NH., Shapiro HJ., Pangilinan
461 JL., Rigoutsos I., Kyrpides NC., Blackall LL., McMahon KD., Hugenholtz P. 2006.
462 Metagenomic analysis of two enhanced biological phosphorus removal (EBPR)
463 sludge communities. *Nature Biotechnology* 24:1263–1269. DOI:
464 10.1038/nbt1247.
- 465 Moller AG., Liang C. 2017. Determining virus-host interactions and glycerol metabolism
466 profiles in geographically diverse solar salterns with metagenomics. *PeerJ*
467 5:e2844. DOI: 10.7717/peerj.2844.
- 468 Rho M., Wu Y-W., Tang H., Doak TG., Ye Y. 2012. Diverse CRISPRs Evolving in
469 Human Microbiomes. *PLoS Genetics* 8:e1002441. DOI:
470 10.1371/journal.pgen.1002441.

- 471 Rousseau C., Gonnet M., Romancer ML., Nicolas J. 2009. CRISPI: a CRISPR
472 interactive database. *Bioinformatics* 25:3317–3318. DOI:
473 10.1093/bioinformatics/btp586.
- 474 Sanguino L., Franqueville L., Vogel TM., Larose C. 2015. Linking environmental
475 prokaryotic viruses and their host through CRISPRs. *FEMS Microbiology Ecology*
476 91:fiv046. DOI: 10.1093/femsec/fiv046.
- 477 Skennerton C. minced - Mining CRISPRs in Environmental Datasets. Available at
478 <https://github.com/ctSkennerton/minced> (accessed May 27, 2016).
- 479 Skennerton CT., Imelfort M., Tyson GW. 2013. Crass: identification and reconstruction
480 of CRISPR from unassembled metagenomic data. *Nucleic Acids Research*
481 41:e105–e105. DOI: 10.1093/nar/gkt183.
- 482 Sorek R., Kunin V., Hugenholtz P. 2008. CRISPR — a widespread system that provides
483 acquired resistance against phages in bacteria and archaea. *Nature Reviews*
484 *Microbiology* 6:181–186. DOI: 10.1038/nrmicro1793.
- 485 Tyson GW., Chapman J., Hugenholtz P., Allen EE., Ram RJ., Richardson PM.,
486 Solovyev VV., Rubin EM., Rokhsar DS., Banfield JF. 2004. Community structure
487 and metabolism through reconstruction of microbial genomes from the
488 environment. *Nature* 428:37–43. DOI: 10.1038/nature02340.
- 489 Wang K., Liang C. 2017. CRF: detection of CRISPR arrays using random forest. *PeerJ*
490 5:e3219. DOI: 10.7717/peerj.3219.
- 491 Weitz JS., Wilhelm SW. 2012. Ocean viruses and their effects on microbial communities
492 and biogeochemical cycles. *F1000 Biology Reports* 4. DOI: 10.3410/B4-17.

493 Wu S., Manber U., Myers E. 1995. A Subquadratic Algorithm for Approximate Regular
494 Expression Matching. *Journal of Algorithms* 19:346–360. DOI:
495 10.1006/jagm.1995.1041.

496 Zhang Q., Ye Y. 2017. Not all predicted CRISPR–Cas systems are equal: isolated cas
497 genes and classes of CRISPR like elements. *BMC Bioinformatics* 18:92. DOI:
498 10.1186/s12859-017-1512-4.

499

500

501

502

503

504

505

506

507

508

509

510

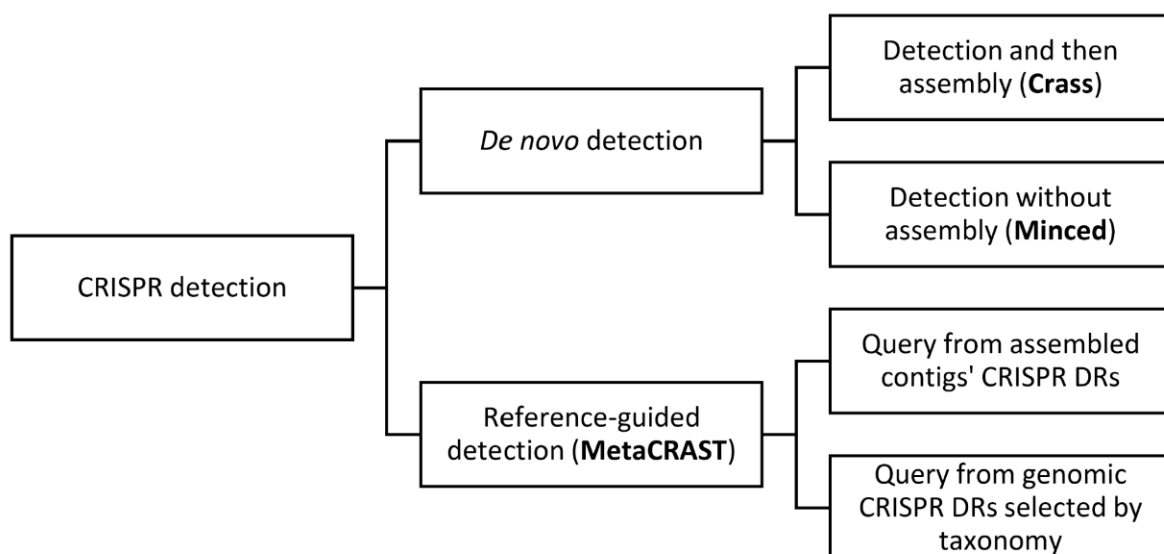
511

512

513 **Figures**

514 Figure 1: This diagram outlines relationships amongst different metagenomic CRISPR
 515 detection methods. CRISPR detection can be performed either using specified direct
 516 repeats (reference-guided detection) or without prior knowledge of direct repeat
 517 sequences (*de novo* detection). *De novo* detection searches raw metagenomic reads for
 518 direct repeat sequences of the appropriate length and spacing (i.e., 25-60 bp long repeats
 519 with 25-60 bp spacers between them). *De novo* detection techniques either detect
 520 spacers in reads only (MinCED) or assemble reads into arrays (Crass). Reference-guided
 521 CRISPR detection, on the other hand, searches reads for user-specified direct repeat
 522 sequences, and extracts spacers from between direct repeat sequences identified in
 523 reads containing direct repeats. While the query is user-specified, general strategies for
 524 generating a query include using direct repeats found in assembled metagenomic contigs
 525 with CRISPR array detection tools (e.g., PILER-CR) or direct repeats found in genomic
 526 CRISPR arrays (e.g., those found in microbial genomes included in CRISPRdb) that
 527 might be expected based on taxonomic profiles. An example of the latter strategy would
 528 be searching for known genomic *Streptococcus pyogenes* direct repeats if *Streptococcus*
 529 *pyogenes* is found in the metagenome's taxonomic profile.

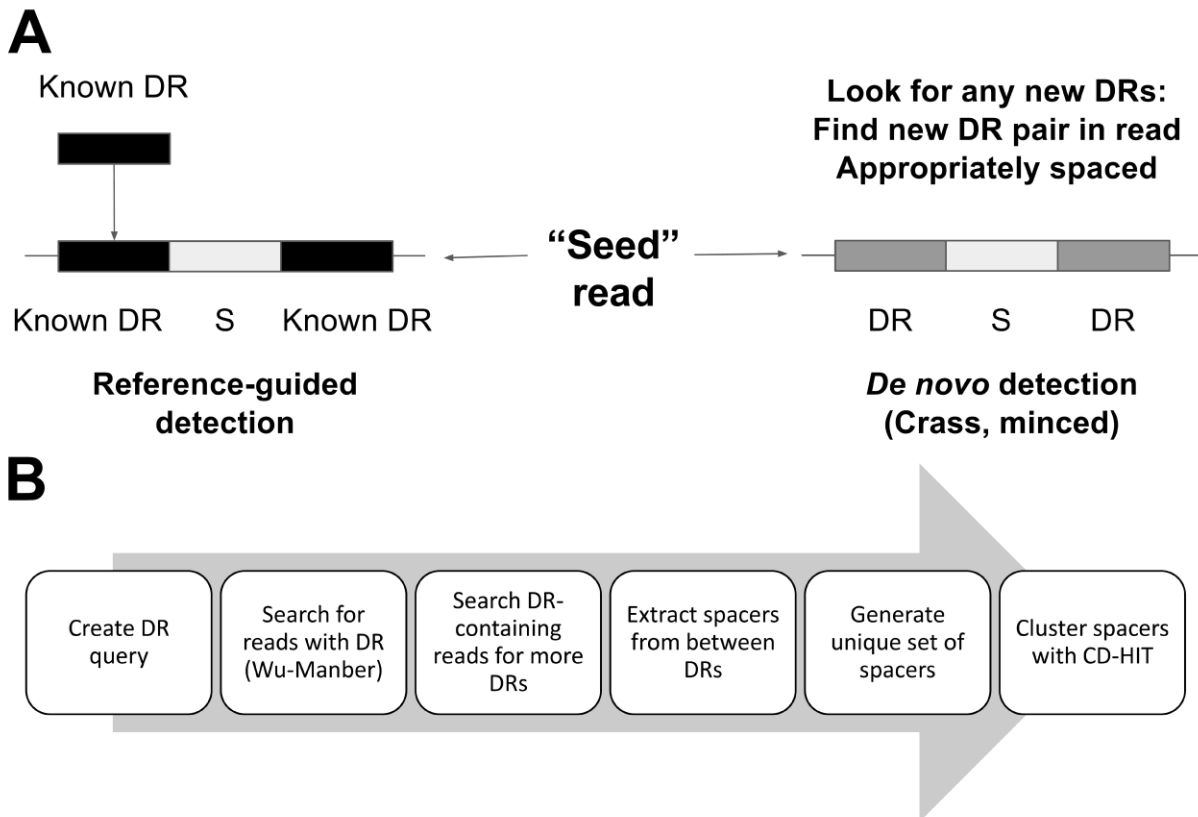
530



531

532 Figure 2: A comparison of per-read CRISPR detection strategies (A) between
 533 MetaCRAST and existing *de novo* detection tools (e.g., Crass, MinCED) and an outline
 534 of the MetaCRAST workflow (B). DR represents direct repeat, while S represents spacer.

535



536

537

538

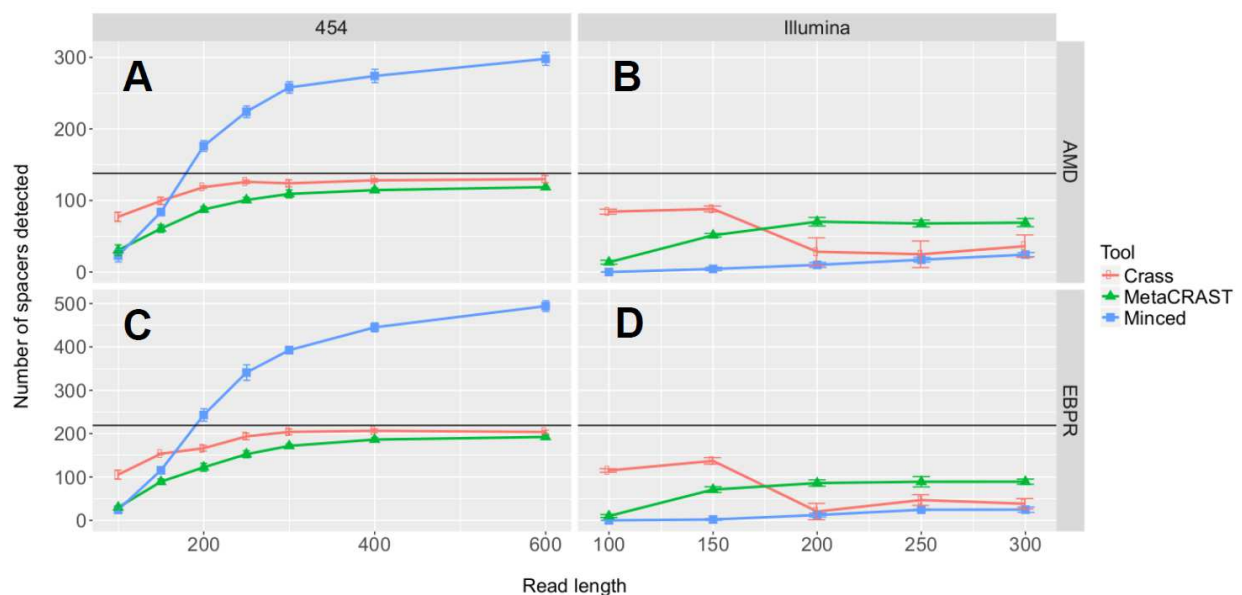
539

540

541

542 Figure 3: Evaluation of MetaCRAST, Crass, and MinCED performance on simulated AMD
543 (A and B) and EBPR (C and D) metagenomes. The procedure used to generate the
544 simulated metagenomes is described in Materials and Methods. All data points represent
545 the averages of six individual simulations and are presented with error bars representing
546 two times the standard error above and two below the average. The true number of
547 spacers expected in each simulated metagenome is marked with a black line (138
548 expected in the AMD metagenomes; 219 in the EBPR metagenomes).

549



550

551

552

553

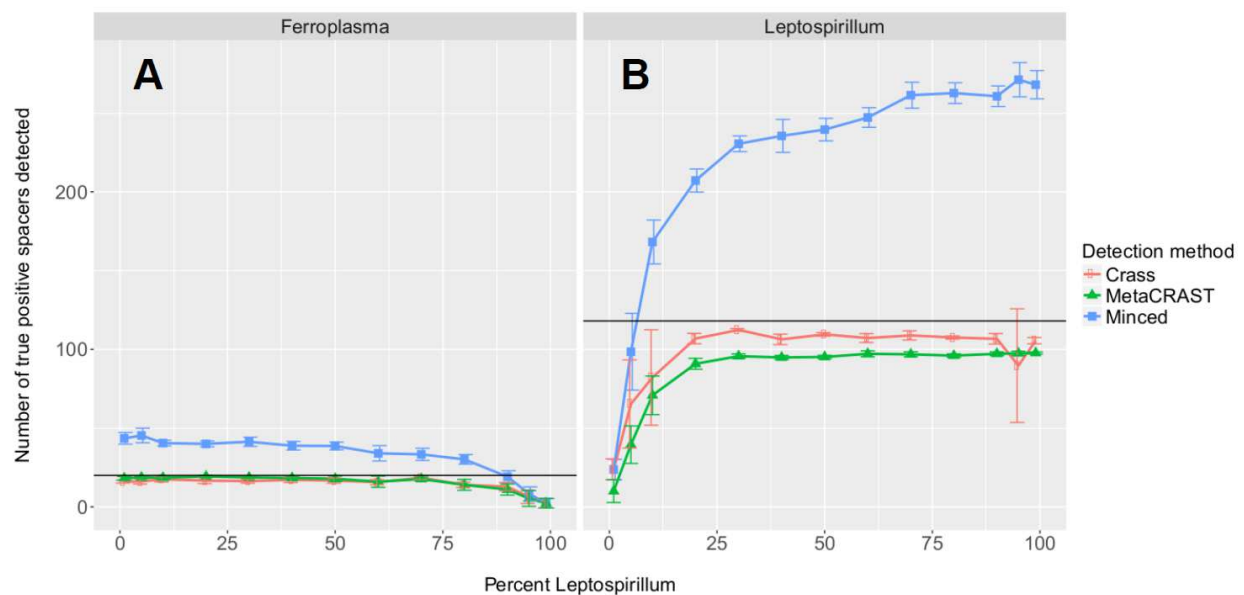
554

555

556

557 Figure 4: Evaluation of MetaCRAST, Crass, and MinCED performance on simulated
558 metagenomes with varying proportions of *Ferroplasma acidarmanus* fer1 and
559 *Leptospirillum* sp. Group II 'CF-1' genome sequences. Simulated metagenomes were
560 generated with Grinder. The data points shown represent the average number of “true
561 positive” spacers detected that matched spacers in corresponding *Ferroplasma* or
562 *Leptospirillum* CRISPR arrays (A and B, respectively). All data points represent the
563 averages of six individual simulations and are presented with error bars representing two
564 times the standard error above and two below the average. The true number of spacers
565 expected for each genome is marked with a black line (20 expected in the *Ferroplasma*
566 genome; 118 in the *Leptospirillum* genome).

567



568

569

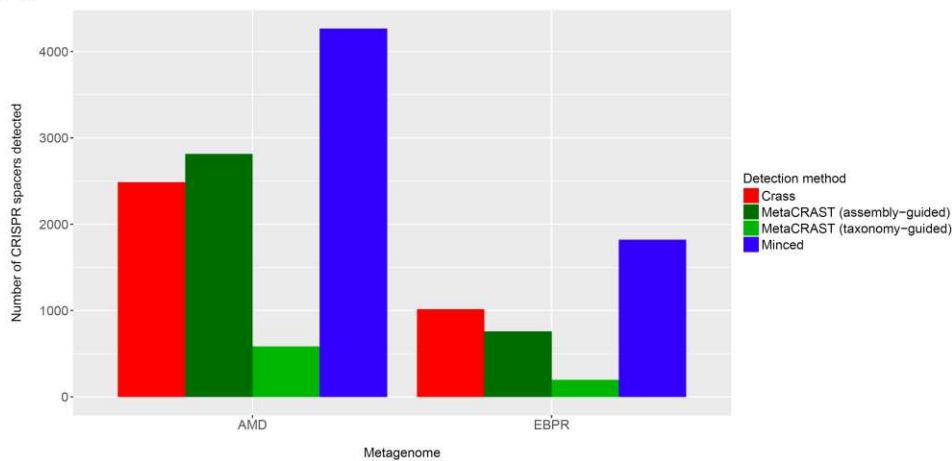
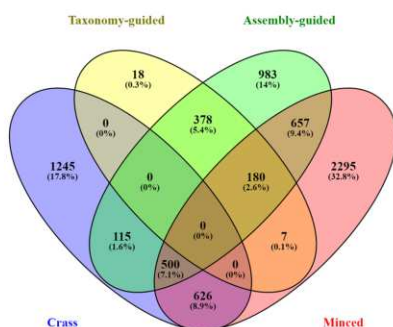
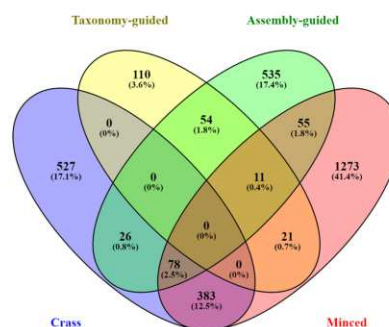
570

571

572

573 Figure 5: Evaluation of MetaCRAST, Crass, and MinCED on real AMD and EBPR
 574 metagenomes. A) Total number of CRISPR spacers detected in real AMD and EBPR
 575 metagenomes using four different detection methods – Crass (*de novo*), MetaCRAST
 576 (using assembly-guided queries), MetaCRAST (using taxonomy-guided queries), and
 577 MinCED (*de novo*). Taxonomy-guided and assembly-guided queries are provided as
 578 Tables S3-S4 and S6-S7. B) Comparison of spacers detected in the real AMD
 579 metagenome using Crass (*de novo*), MetaCRAST (using taxonomy-guided queries),
 580 MetaCRAST (using assembly-guided queries), and MinCED (*de novo*). Comparison was
 581 performed using Venny 2.1 (<http://bioinfogp.cnb.csic.es/tools/venny/>). C) Comparison of
 582 spacers detected in the real EBPR metagenome using the same methods as in B.
 583 Comparison was performed using Venny 2.1.

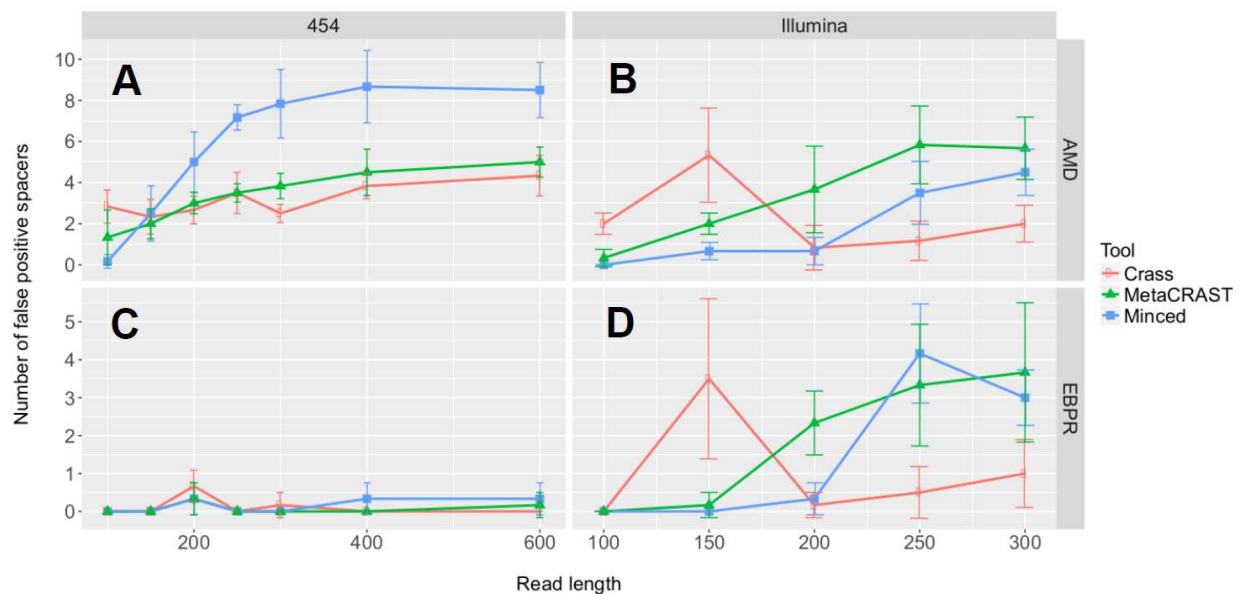
584

A**B****C**

585

586 Figure 6: Evaluation of MetaCRAST, Crass, and MinCED false positive detection on
587 simulated AMD (A and B) and EBPR (C and D) metagenomes. The procedure for
588 generating the simulated metagenomes is described in Materials and Methods. The
589 number of detected spacers matching expected ones was subtracted from the total
590 number of spacers detected to determine the number of false positive spacers for a
591 particular method and condition. All data points represent the averages of three
592 individual simulations and are presented with error bars representing two times the
593 standard error above and two below the average.

594



595

596

597

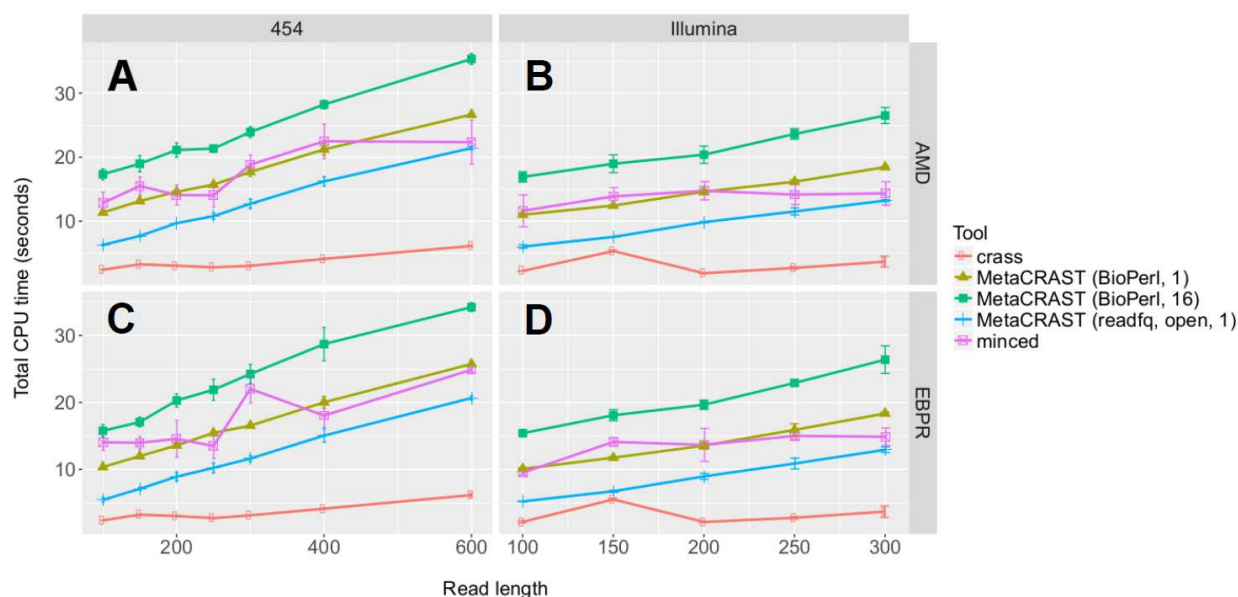
598

599

600

601

602 Figure 7: Evaluation of MetaCRAST, Crass, and MinCED run times on simulated AMD (A
 603 and B) and EBPR (C and D) metagenomes. We evaluated seven different combinations
 604 of algorithms, implementations, and parameters. We evaluated both Crass and MinCED
 605 with default parameters. For MetaCRAST, we evaluated five different conditions differing
 606 in parallelization and metagenome loading method - BioPerl loading and 16 threads,
 607 BioPerl and a single thread, readfq with mce_open for loading and 16 threads, readfq
 608 with mce_open and a single thread, and readfq with the standard open routine and a
 609 single thread. The procedure for generating the simulated metagenomes is described in
 610 Materials and Methods. Run time was calculated as the sum of the user and system time
 611 (together the total CPU time). All data points represent the averages of three individual
 612 simulations and are presented with error bars representing two times the standard error
 613 above and two below the average.



614

615

616

617

618

619 **Tables**

620 Table 1: Command line arguments for MetaCRAST. Required arguments are in bold.

Argument	Description
-p	Pattern file containing query DR sequences in FASTA or FASTQ format
-i	Input metagenome in FASTA or FASTQ format
-o	Output directory for detected reads and spacers
-d	Allowed edit distance (insertions, deletions, or substitutions) for initial read detection with the Wu-Manber algorithm and subsequent DR detection steps
-t	Temporary directory to put metagenome parts (use this if -n option also selected)
-q	Input metagenome is a FASTQ file (directs use of fastq-splitter.pl instead of fasta-splitter.pl)
-h	Use Hamming distance metric (substitutions only - no insertions or deletions) to find direct repeat locations in reads (default: use Levenshtein distance metric - look for sequences matching DR within insertion, deletion, and/or substitution edit distance)
-r	Search for reverse complement of CRISPR direct repeat sequences
-l	Maximum spacer length in bp
-c	CD-HIT similarity threshold for clustering spacers detected for each query direct repeat (value from 0 to 1)
-a	CD-HIT similarity threshold for clustering all detected spacers (value from 0 to 1)
-n	Number of processors to use for parallel processing (and number of temporary metagenome parts)

621

622

623

624

625

626 Table 2: Distribution statistics for binomial name and genus-level taxonomic affiliation of
627 CRISPRdb direct repeats. A library of direct repeats labeled with respective GenBank
628 accessions from CRISPRdb was processed to assign taxonomy information based on
629 GenBank accession. Taxonomy information was extracted from GenBank records with
630 the Perl module Bio::DB::GenBank. Statistics describing the distribution of binomial
631 names or genres to which individual direct repeats affiliated were compiled with
632 Microsoft Excel.

Statistic	Binomial names	Genres
Mean	1.308	1.063
Median	1	1
Mode	1	1
Minimum	1	1
Maximum	46	20
Standard deviation	1.567	0.521

633