Published online 19 March 2013

doi:10.1093/nar/gkt183

Crass: identification and reconstruction of CRISPR from unassembled metagenomic data

Connor T. Skennerton[1,2], Michael Imelfort[1,2] and Gene W. Tyson[1,2,*]

[1]Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences and [2]Advanced Water Management Centre, The University of Queensland, St. Lucia, Brisbane, Queensland 4072, Australia

Received November 12, 2012; Revised January 23, 2013; Accepted February 26, 2013

ABSTRACT

Clustered regularly interspaced short palindromic repeats (CRISPR) constitute a bacterial and archaeal adaptive immune system that protect against bacteriophage (phage). Analysis of CRISPR loci reveals the history of phage infections and provides a direct link between phage and their hosts. All current tools for CRISPR identification have been developed to analyse completed genomes and are not well suited to the analysis of metagenomic data sets, where CRISPR loci are difficult to assemble owing to their repetitive structure and population heterogeneity. Here, we introduce a new algorithm, Crass, which is designed to identify and reconstruct CRISPR loci from raw metagenomic data without the need for assembly or prior knowledge of CRISPR in the data set. CRISPR in assembled data are often fragmented across many contigs/scaffolds and do not fully represent the population heterogeneity of CRISPR loci. Crass identified substantially more CRISPR in metagenomes previously analysed using assembly-based approaches. Using Crass, we were able to detect CRISPR that contained spacers with sequence homology to phage in the system, which would not have been identified using other approaches. The increased sensitivity, specificity and speed of Crass will facilitate comprehensive analysis of CRISPRs in metagenomic data sets, increasing our understanding of phage-host interactions and co-evolution within microbial communities.

INTRODUCTION

Clustered regularly interspaced short palindromic repeats (CRISPR) are an adaptive immune system found in half of the sequenced bacterial and almost all archaeal genomes (1). CRISPR are composed of an array of conserved direct repeats (DRs) separated by unique spacer sequences and are typically located adjacent to a leader sequence and CRISPR-associated (*cas*) genes (2). Previous research has shown that spacers often correspond to plasmid or phage DNA and act as a targeting mechanism for degradative enzymes encoded by the *cas* genes (1,3,4). Newly acquired spacers are inserted into a CRISPR locus at the end of the array closest to the leader sequence. This directionality preserves the history of phage infections and can be used to study evolution and epidemiology of bacterial strains with higher resolution than other phylogenetic markers (5, 6). Genome sequencing has revealed substantial diversity in the complement of the *cas* genes in different organisms and a wide variety of DR sequence types (1,7,8). As a result, the most common way of defining CRISPR loci and characterizing the spacer sequences has been to search for regularly spaced repeats in sequenced genomes (9–11).

During the past decade, there has been an exponential growth in the amount of metagenomic sequence data generated for natural microbial communities. Analysis of CRISPR in model organisms and metagenomic data sets has revealed remarkable diversity in spacer complement reflecting the rapid co-evolution of phage and their hosts (6,12–17). Recent studies have determined CRISPR content in metagenomic data sets based on analysis of the assembled fraction (14,16,17). However, this approach is limited as modern genome assembly algorithms filter out or collapse repetitive regions, and therefore CRISPR may not be properly assembled into contigs. Furthermore, populations of microorganisms can have highly diverse spacer arrangements between individuals, thus CRISPR loci found in metagenomic assemblies may only represent the most dominant strain in the community, and not be indicative of the true spacer diversity.

Here, we present a new algorithm called Crass, which has been designed specifically to identify and reconstruct

*To whom correspondence should be addressed. Tel: +617 336 53829; Fax: +617 336 54511; Email: g.tyson@awmc.uq.edu.au

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2013. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

CRISPR loci from unassembled metagenomic data sets. Crass is able to locate individual reads that contain DRs and cluster them together based on DR type. Entire CRISPR arrays are reconstructed using a novel graph approach to accurately describe the spacer arrangement and strain diversity for each locus. As a final step, Crass can output assembled contigs of individual strains using external assembly software such as Velvet (18). Investigation of CRISPR diversity using Crass will enable a deeper understanding of phage-host co-evolution in microbial communities.

MATERIALS AND METHODS

Algorithm details

The algorithm used in Crass can be broken into four parts: (i) initial search and refinement, (ii) exhaustive search, (iii) identification of the correct DR sequence and (iv) graph construction and refinement.

Initial search and refinement

The initial repeat detection is segregated into two separate algorithms based on the length of the read, henceforth referred to as the short-read and long-read algorithms. The short-read algorithm is optimized for scanning reads that can only contain a maximum of two copies of a DR (<176 bp; Illumina and Ion Torrent PGM based on current read lengths). The search algorithm scans a read for two copies of a kmer, equal to the minimum DR length (23 bp in the default settings), which are separated by a distance of at least the size of a spacer (S; 26 bp by default). The algorithm compares a subsequence beginning at position *P* to the subsequence of the read beginning at P + S. If no match is found, P is incremented one nucleotide, and the search is repeated until such point as the sub-sequence of the read beginning at position P + S is less than the minimum allowed size for a DR. However, if a match is found, the matching kmers are extended to the right for as long as the nucleotides at the extending positions continue to be identical.

The long-read algorithm (reads 177–2000 bp) uses a modification of the technique developed in CRT (10). This algorithm searches for two identical copies of a relatively short kmer (8 by default) that are separated by S. When a repeating kmer is found, it is used as a seed sequence for subsequent searches for the same kmer at the same interval across the remainder of the read. If the seed sequence is found in the read at least three times, then all of the matching regions are extended in both directions. Reads that are identified during the initial search stage are subject to quality control measures based on currently known CRISPR loci.

All currently known DRs and spacers fall within defined length ranges (Supplementary Figure S1A) and have no internal repeating motifs (Supplementary Figure S1B). These two parameters are used to filter out other repeat types, such as microsatellites that are composed of short (2–5 bp) repeating sequences. Finally, the identified repeating subsequence and the spacer region are compared with each other to distinguish CRISPR DRs from other repeating motifs. A comparison of all known CRISPR loci showed that the DR should not contain sequence homology to the spacer regions, and the individual spacers should not contain any sequence homology to each other (Supplementary Figure S1C).

Exhaustive search

The initial search identifies many potential DR types; however, many DR-containing reads will be missed owing to the small chance of there being two copies of a DR in short-read data (Supplementary Figure S2). To recruit reads that contain only one DR, Crass uses the Wu–Manber multipattern search algorithm (19). To improve the speed of this exhaustive search, Crass creates a non-redundant set of DR types by using single-linkage clustering and removing DR types that are perfect substrings of others. A DR type is added to an existing cluster only if it contains at least six, 7 bp kmers with one of the DR types already binned into that cluster. If there is not enough matching kmers to any of the DR types in any existing DR cluster, the sequence is used as the seed for a new cluster.

Identification of the correct DR sequence

The correct DR sequence for each DR cluster is identified by alignment of the reads and identification of highly conserved nucleotide positions. Performing a multiple sequence alignment of every read in a DR cluster is computationally demanding. Instead, Crass aligns all variants to the longest DR type in a cluster to determine an alignment offset, which is the number of nucleotides difference between the start of the longest DR type to the start of each of the other variants. The reads are then aligned using the position of the first DR in the read and the alignment offset calculated for that particular DR type. The boundary of the DR is determined by highly conserved positions (minimum 85%) in the consensus of the alignment. A caveat of this approach is that two or more highly similar, but distinct DR types, can be grouped together during the clustering step (Supplementary Figure S3). Crass detects these distinct variants by identifying nucleotide positions within the DR boundaries that fall between 30 and 85% conservation amongst the reads. These positions are examined to determine whether there is enough read depth to constitute a different DR type, as single SNPs in low coverage alignments have a greater effect of the conservation of that position. Crass then recursively separates the reads associated with each variant into different DR types.

Graph construction and refinement

Crass constructs a graph of the spacer arrangement, using the spacers as the nodes. Edges in the graph are created between two spacers if they lie sequentially in a read. The direction of the edge is calculated such that the sequence of the DR that lies between the two spacers is in its lowest lexicographical form. Short-read data complicates this process, as there is only a small probability of finding two full-length spacers within a read (Supplementary Figure S2). Crass solves this problem by building a preliminary graph (*p-graph*) that guides the construction of the final spacer graph. Nodes in the p-graph (*p-nodes*) are constructed from the first and last *k* bases of each spacer in each read. Thus, a read with only one full-length spacer and one partial spacer will contain enough information to be linked into the p-graph. Multiple p-nodes may originate from the same spacer, one from the first k bases and the other from the last k bases. P-nodes are linked together by two types of directed edges, 'inner edges' and 'jumping edges', depending on whether they originate from the same spacer (inner edges), or from adjacent spacers (jumping edges; Figure 1). A complete and accurate p-graph contains chains of p-nodes that are connected such that each p-node is typically joined in both the forward and reverse directions to at least one other p-node, and by a combination of both inner and jumping edges.

Sequencing errors typically result in p-nodes that lack either jumping or inner edges, and are therefore attached to the rest of the p-graph by only a single edge. Such nodes appear to dangle from the main body of the graph giving the appearance of *fur*. Alternatively, if these p-nodes lie adjacent to naturally occurring forks, they can become connected to the p-nodes lying after the fork, creating a 'bubble' in the graph (Figure 1). P-nodes that produce bubbles are recognized because they are only connected to the main body of the graph by edges of a single direction. Removing a bubble-creating p-node can create fur and vice versa and therefore removal of erroneous p-nodes uses an iterative algorithm, which alternates between removal of fur and bubbles until one complete iteration results in the removal of no p-nodes (Figure 1).

The cleaned p-graph is transformed into the final spacer by joining p-nodes that share an inner edge. Jumping edges between p-nodes define connections between their corresponding spacer nodes such that the directionality of the original edge is conserved. The resulting spacer graph typically contains a small number of erroneous nodes that could not be filtered during p-graph cleaning but can be removed from the spacer graph. Fur and bubbles are identified in the spacer graph in same way as the p-graph. Fur is identified as a node that has only a single edge to a node that contains at least three edges. Nodes that cause bubbles are identified, as they will contain multiple edges but only in a single direction. The same iterative approach used in the p-graph is used in the spacer graph to alternate the removal of fur and bubbles until no spacer nodes are removed.

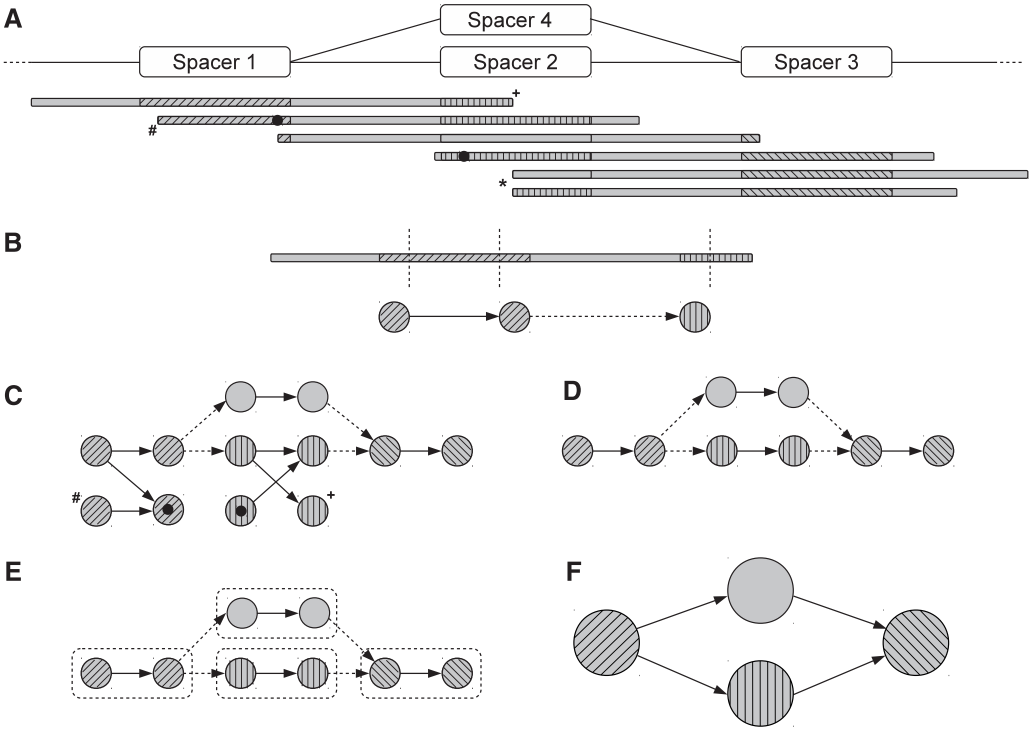

Figure 1. Construction and refinement of the preliminary and final spacer graph. A schematic illustrating graph construction and potential problems in determining the correct spacer order. (**A**) An arrangement of four spacers representing CRISPR spacer heterogeneity, where spacer 2 and spacer 4 are both connected to spacer 1 and spacer 3. Sequencing reads that contain these spacers are shown (grey bars), some of which contain spacer 2 and others that contain spacer 4 (without hatching). Sequencing errors are marked with black circles. Incomplete spacer sequences that are found in some reads are marked with hash, plus and asterisk symbols. (**B**) Each read is used to create a small portion of a preliminary spacer graph (p-graph). Nodes are created from k-mers, which are cut from the ends of each spacer region (delineated by dashed vertical lines). Edges are either 'inner' edges, connecting nodes from the same spacer (solid arrow), or 'jumping' edges between different spacers (dashed arrow). (**C**) The initial version of the p-graph is produced by combining nodes derived from all reads, including k-mers from incomplete spacer sequences. (**D**) The p-graph after removal of fur caused by sequencing errors or incomplete spacers. (**E**) Pairs of nodes joined by inner edges are concatenated together to form spacer-nodes in the spacer graph. Jumping edges remain in the spacer graph, as they represent a DR sequence. (**F**) Each node now represents a correctly ordered spacer in the final spacer graph.

Synthetic data set generation and comparison

Eight genomes containing between zero and seven CRISPR loci were chosen at random from CRISPRdb (20) (Table 1). Synthetic Illumina data sets (101 bp reads) representing $\sim$20× coverage were generated for each genome using Grinder 0.4.5 (21); command-line options: –cf 20 –rd 101 –md poly4. Crass 0.3.1 was run on each data set using a kmer length of 9 (all other parameters default). The spacers identified by Crass were mapped onto the reference genome using blastn 2.2.25+ (22) to determine whether they were correctly positioned. The spacer graphs for each data set were also analysed to determine whether the ordering of spacers generated by Crass accurately reflected the CRISPR loci found in the original genome assembly.

Acid mine drainage comparison

The UBA data set was obtained from the NCBI Trace Archive (Project 18537) and was analysed with Crass 0.3.1. The complete set of spacers identified by Andersson and Banfield (13) in the original analysis was retrieved from the Supplementary information (http://www.sciencemag.org/content/320/5879/1047/suppl/DC1). Spacers that matched to the UBA BS data set [also analysed by Andersson and Banfield (13), but not publically available] were removed by mapping the complete spacer set against the reads of the UBA data set using Usearch 4.2.66 (23) requiring a 100% match between a read and the spacer. The remaining spacers were then tested to see whether they were adjacent to their DR sequence; those that were not were deemed to be of phage/plasmid origin and removed. The resulting 1527 spacers were then compared with the complete set of spacers identified by Crass using blastn 2.2.22 (24).

Global ocean survey comparison

The full Global Ocean Survey (GOS) data set was downloaded from CAMERA (http://camera.calit2.net/) and analysed with Crass 0.3.1. The 'high quality' CRISPR cassettes identified by Sorokin *et al.* (14) were collapsed into unique DR types using GNU grep. The unique DR types and their corresponding spacers were compared against the DRs and spacers identified by Crass using blast 2.2.22 (24). Spacers that appeared to be missed by Crass were manually inspected by determining their positions in the raw metagenomic reads. Spacers that were not adjacent to a DR were removed as well as any that overlapped with the DR sequence. A number of unique spacers were actually found by both analyses; however, significant differences in the DR boundaries resulted in sequences that failed to be identified by blast. The two largest DR types were analysed for mis-assemblies in the scaffolds to determine whether they accurately reflected the spacer arrangement. DR containing reads that were identified from both of these DR types were reassembled with Geneious 5.6.3 *de novo* assembler (25), and the position of the spacers was identified by mapping with bowtie 2.0.0-beta5 (26).

Discovery of CRISPR loci in the Enhanced biological phosphorus removal metagenomic samples

The microbial enhanced biological phosphorus removal (EBPR) data set was assembled using Velvet 1.0.18 (18)

Table 1. Specificity and sensitivity analysis of Crass on synthetic short read data sets

	Total spacers	Detected spacers	Missing edges	Erroneous edges	Specificity	Sensitivity
Bacteroides fragilis YCH46						
CRISPR1	9	7	3	0	1	0.63
Acinetobacter sp. ADP1						
CRISPR1	6	6	0	0	1	0.83
CRISPR2	21	21	0	0	1	1.00
CRISPR3	90	88	2	0	1	0.98
Sulfolobus solfataricus P2						
CRISPR1	102	102	0	5	1	0.95
CRISPR2	94	94	0	0	1	0.96
CRISPR3	31	31	0	0	1	1.00
CRISPR4	95	95	0	3	1	0.97
CRISPR5	6	5	1	0	1	0.80
CRISPR6	22	22	0	0	1	1.00
CRISPR7	65	64	1	0	1	0.98
Natrialba magadii						
CRISPR1	27	18	11	0	0.89	0.58
Helicobacter pylori B8	0	0	N/A	N/A	1	N/A
Magnetospirillum magneticum AMB-1	0	0	N/A	N/A	1	N/A
Tsukamurella paurometabola	0	0	N/A	N/A	1	N/A
Oligotropha carboxidovorans OM5	0	0	N/A	N/A	1	N/A
Overall	568	553	18	8	0.99	0.89

Crass was used to examine synthetic data sets constructed from four genomes that contained between one and seven CRISPR loci, in addition to four genomes that did not contain CRISPRs. The specificity of Crass was calculated by determining the number of detected spacers that did not originate from CRISPRs; the sensitivity was determined by comparing the reconstructed spacer ordering to the ordering found in the genome.

with a kmer length of 37, insert size of 300 bp, a kmer cut-off of 2, and the expected kmer coverage of 100. The phage data set was assembled using Metavelvet 0.3 (27) and scaffolded with Bambus2 (28). CRISPRs were identified from the microbial assembly using PILER-CR 1.06 (9). Crass 0.3.1 was used on the same raw data using the kmer length (-K) set to 9; all other parameters were left as default. Spacers identified by Crass were compared with the phage contigs using blastn 2.2.25+ (22) with the following parameters changed from their defaults: -word_size 16 -evalue 1 e-6. Only hits that spanned the length of the spacer and contained a maximum of three mismatches were considered protospacers. The spacer graph of the most abundant DR type in the EBPR data set was manually curated through identification of the leader sequence and removal of erroneous nodes. To identify the leader sequence, all of the reads and their pairs from the DR type were mapped against the contigs to determine which spacers were adjacent to the flanking regions of the CRISPR. Four spacers mapped to a contig in the metagenomic assembly that contained *cas* genes, indicating that they were adjacent to the leader sequence. An extra node representing the leader sequence was added to the spacer graph (Supplementary Figure S4B; green circle). All low coverage nodes in the spacer graph (Supplementary Figure S4; blue circles) were manually examined at the read level to determine whether the edges (linkage between nodes) were supported, resulting in the removal of three nodes and an edge (Supplementary Figure S4B and C) from the final spacer arrangement (Supplementary Figure S4D). The raw sequence data for the EBPR data set can be found under NCBI bio-project PRJNA81811.

RESULTS

Overview of the Crass algorithm

Crass has been designed to process shotgun metagenomic data from Illumina, Ion Torrent PGM, Roche 454, and Sanger platforms using an iterative search approach that does not rely on preassembled contigs or prior knowledge of the CRISPRs in the metagenomic data set. Using the raw unassembled reads, Crass searches all sequences for possible DR. CRISPR DRs appear in reads as short subsequences that are perfectly repeated, separated by a unique subsequence. As the initial search identifies only repeated subsequences, a number of other repeat types, not originating from CRISPR, are also identified. These false positive matches are filtered out based on a number of criteria identified from analysing previously characterized CRISPR loci (see 'Materials and Methods' section). If reads pass these tests, they are considered to be a candidate CRISPR-containing read and their corresponding repeated region a putative CRISPR DR type.

The short reads produced by Illumina (100–150 bp) and Ion Torrent PGM (100–200 bp) introduce additional complications when identifying DRs. Given that spacers and DRs range in length from 23 to 50 bp (Supplementary Figure S1A), there is at best a 46% chance of two full-length copies of the DR being present in a read of 100–150 bp (Supplementary Figure S2). However, the initial DR search strategy will only identify reads with at least two copies of a DR. Crass re-examines all of the unused reads in the data set for the DR types found in the initial search phase to identify reads that contain a single copy of a DR. In many cases, the DRs identified in short read data sets are truncated, as they occur at the start or end of a read, or are artificially extended by incorporation of bases from adjacent spacers. The resulting variants of a DR type, and their reverse complement, are identified using single-linkage clustering, and an alignment of all reads for each DR cluster is constructed to identify a consensus DR sequence.

For each DR type, Crass attempts to reconstruct the CRISPR loci by ordering spacers based on their co-occurrence in individual reads. CRISPR loci reconstructions are represented as graphs, where the nodes (spacer nodes) represent the spacer sequences, and the directed edges represent the DR (Figure 2). Pairs of spacer nodes are joined by directed edges if their corresponding spacers lie sequentially in a read. Shared spacer arrangements appear as a linear chain of nodes that fork at nodes in the graph where a common spacer is adjacent to multiple unique spacers. However, short reads and sequencing errors can complicate the graph building process, as they introduce superfluous spacer nodes that confound the graph (Figure 1). For example, spacers that lie too close to the end of a read will be truncated, producing a unique spacer node. Spacer graph construction involves differentiating 'real' forks, resulting from strain variation, from those that are a byproduct of the sequencing process (see 'Materials and Methods' section).

Crass produces an XML file that contains information for each DR type, including the DR sequence, and spacer sequences, coverage and order within the CRISPR loci. The reconstructed spacer order can be saved as encapsulated postscript images that can be used to guide the assembly process or aid in CRISPR diversity analyses. In addition, user-defined pathways through each CRISPR loci can also be extracted and assembled using external assembly programs such as Velvet (18).

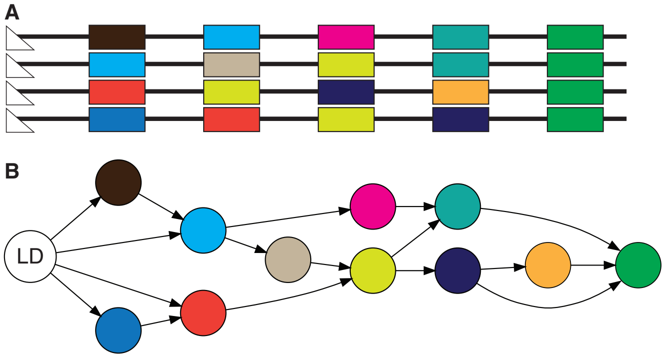

Figure 2. Comparison between different CRISPR loci visualization techniques. (**A**) Traditional approach to visualization where the spacers are shown as differently colored rectangles (the same colour refers to the same spacer) anchored to the leader sequence (white triangle). (**B**) The same CRISPR loci reconstructed by Crass into a spacer graph.

Assessing Crass performance

The specificity and sensitivity of Crass was analysed using synthetic data sets constructed from eight complete genomes (Table 1). Crass displayed high specificity, detecting all DR types in each genome. However, Crass erroneously detected one other repetitive element in the *Natrialba majadii* genome, resulting in the identification of three 'spacers' that were not from a *bona fide* CRISPR. Crass correctly identified the vast majority of spacers from each genome (553 of 568 total spacers), except in the case of *N. majadii* where there were significant variations in DR sequence composition. Individual DR units in this genome contained up to 4 bp changes from the DR consensus, and, as a result, Crass failed to identify nine spacers adjoining these variable DRs.

The sensitivity of the spacer reconstructions was determined by calculating the fraction of all spacers that were either erroneously linked or missing a correct edge in each CRISPR. Crass correctly ordered the majority of spacers from all CRISPR loci with the majority of errors being caused by missing spacers (Table 1). The only CRISPR where Crass incorrectly linked distal spacers together were found in the *Sulfolobus solfataricus* genome. There were five edges in CRISPR1 and three edges in CRISPR4 that were not supported in the genome. In both CRISPR, these errors were due to identical kmers being present at the beginning or end of non-adjacent spacers such that they became linked during spacer graph construction (Table 1; Supplementary Figure S5).

We next determined how Crass performed on real metagenomes that previously had their spacer complement analysed. An (AMD) metagenome generated using Sanger sequencing (114 mbp total size, 135 937 reads) that had been analysed at the individual read level (13) provided a robust data set for validating the sensitivity and specificity of Crass. The GOS metagenomic data sets (10 133 846 Sanger reads; 10.635Gbp and 2 538 672 Roche 454 reads; 963.763 mbp) were also examined to evaluate whether Crass' read level detection provided greater resolution over previous analyses performed using the assembled scaffolds (14).

Crass processed the AMD data set in 16 s and found all DR types and the majority of the spacers identified in the original analysis (Figure 3). Crass also discovered three novel DR types, likely missed owing to their low coverage in the metagenome (Supplementary Table S1). Alignment of reads representing all DR types showed that Crass was more sensitive at determining the DR boundaries, which were typically 1–3 nucleotides longer than those previously reported (Supplementary Figure S6A). This change in the DR boundaries also affected the spacer sequences, which incorrectly contained the start or end of the DR. There were also two DR types where Crass failed to call the boundary of the DR correctly. In one instance, Crass failed to split two DR types that differed by two nucleotides into separate CRISPR loci (Supplementary Figure S6B), and in the other instance, Crass extended the DR into the spacer sequence (Supplementary Figure S6C). The majority (99.62%) of the spacers were shared by both analyses; however, there were nine spacers that Crass failed to identify, which were deleted incorrectly during the graph cleaning process.

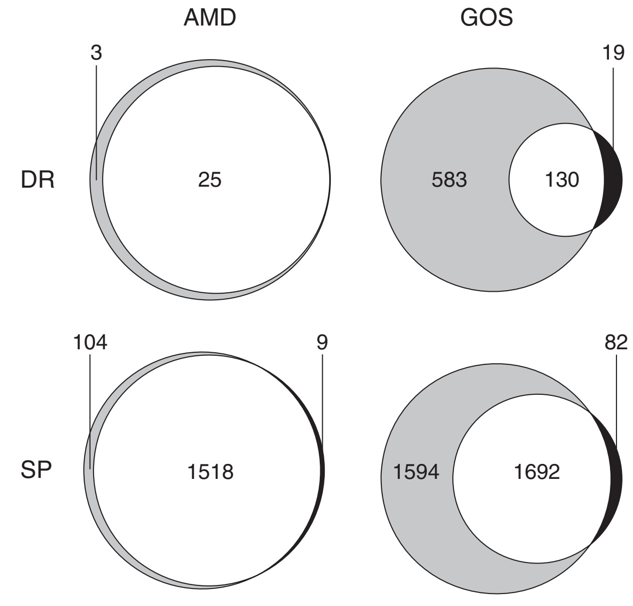

Figure 3. Summary of the number of repeats and spacers identified by Crass in comparison with the original analyses. The number of shared DRs and spacers for the AMD and GOS data sets are shown in the central white section of each Venn diagram. Sequences detected only by Crass are coloured grey, and those only found in the original analyses are coloured black.

The combined GOS metagenomic data sets were substantially larger than the AMD metagenome, and the CRISPR diversity was higher. Crass found 87% of the DR types, and 95% of the spacers originally identified in the GOS scaffolds (Figure 3). Crass processed the entire GOS data set in under 2 h and identified 713 DR types, 4.7× more than the original analysis. Crass identified 130 of the 194 'high-quality' CRISPR originally found in the GOS data set. However, the original analysis assumed that each scaffold containing a DR corresponded to a different CRISPR type, even if the DR had been previously found on another scaffold thereby over-inflating the number of CRISPR loci in the data. When collapsed to unique DR types, there were 149 DR types found in the original study. In most cases, Crass confirmed that the number of CRISPR loci were overestimated. For instance, one of the most abundant DR types identified by Crass was fragmented across 11 scaffolds, which resulted in 11 different CRISPR being reported. However, Crass identified a single DR type and arranged these reads into four discrete spacer graphs (Supplementary Figure S7).

From the 19 DR types that Crass did not find in the GOS data sets (Supplementary Table S1), nine were found in reads that contained two or less copies of the DR sequence, below the threshold for confident identification by Crass. Another six DR types were identified during the initial search but then removed during the subsequent filtering steps. Three DR types were removed, as the final

DR consensus sequence fell outside the acceptable size ranges. The other three were removed after the graph-building phase, as they contained less than three spacers.

Cross-validation of Crass results using coupled microbial-phage metagenomes

A 2 Gbp Illumina data set generated from an EBPR reactor community (see Supplementary Methods) was analysed using Crass. The raw data were assembled using Velvet, and the majority of contigs were putatively classified as belonging to *Candidatus* Accumulibacter phosphatis (29). Seventeen DR types containing 308 spacers were discovered from the assembly using PILER-CR; in comparison, Crass identified 72 DR types (6.7× more than with PILER-CR) containing 2304 spacers (Figure 4). The largest DR type identified in both analyses contained >400 spacers and showed evidence of strain heterogeneity near the leader sequence (Figure 5; green circle). The spacer graph forked into two main branches that were in equal coverage, likely representing micro-heterogeneity within the *C.* Accumulibacter Phosphatis (CAP) population in the community. One of the branches contained approximately twice the number of spacers as the other (Figure 5, label A), suggesting that some members of the population were infected by more phage in the past. There is also a third arrangement that could not be connected to the leader sequence (Figure 5, label C). The spacer graph ends in a conserved tail region (Figure 5, label D) that can be linked into a flanking contig from the assembly (Figure 5; grey circle). Half of the spacers in the main body of the graph were identified in 20 contigs from the assembly. These spacers did not fully correspond to any of the arrangements identified by Crass but contained fragmented arrangements of the CRISPR, including those near the leader sequence and the conserved tail region (Supplementary Figure S9).

A phage metagenome prepared at the same time as the microbial metagenome was sequenced (see Supplementary Methods) to find EBPR CRISPR spacers derived from phage in the bioreactor. The only spacers with matches to the phage metagenome were homologous to the most abundant phage in the bioreactor. There were eight spacers that had between zero and three mismatches to the phage genome and contained a CCN protospacer associated motif (Supplementary Figure S10). Although the CRISPR locus containing these spacers was not identified by PILER-CR in the contigs, Crass identified 131 spacers in the raw reads, making it the third largest CRISPR in the data set.

DISCUSSION

CRISPRs are an adaptive bacterial and archaeal immune system that directly target infecting phage types and as such, store a history of acquired phage resistance over time. Until recently our understanding of CRISPR evolution was largely based on isolate genomes (3,4,6, 30); however, the emergence of metagenomics now enables CRISPR diversity and dynamics to be explored in natural communities. Detection of CRISPRs in metagenomic data sets is confounded by their repetitive nature and strain heterogeneity, which complicate assembly. Most previous studies have only used CRISPR identified from assembled contigs to examine diversity and evolution (14, 16, 17); however, CRISPR loci are typically poorly assembled or not assembled at all (Figure 4; Supplementary Figures S7–S9). Crass implements an alternative approach that searches through raw metagenomic data for DR-containing reads and reconstructs the spacer ordering for each DR type.

The sensitivity and specificity of Crass was evaluated by comparing the detection of DRs and spacers in synthetic data sets and previous analysed metagenomes. In the synthetic data sets, Crass showed high specificity and

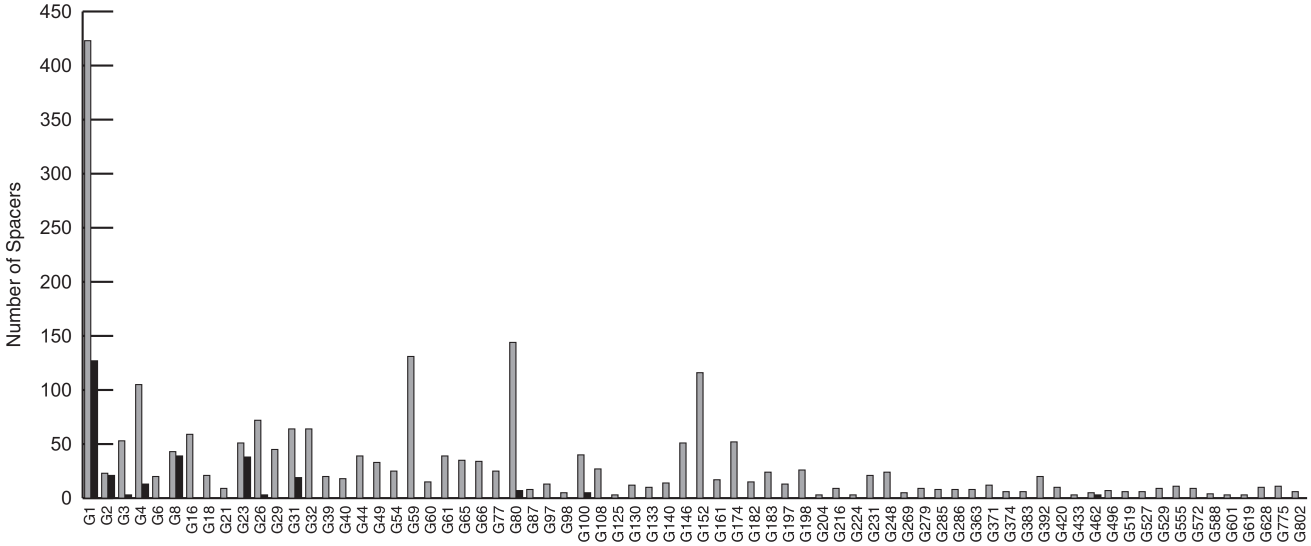

Figure 4. Identification of DR types and total spacer count from in the EBPR microbial metagenome. DR types identified by Crass are shown along the *x*-axis. Grey bars correspond to the number of spacers found for each DR type by Crass and black bars for PILER-CR.

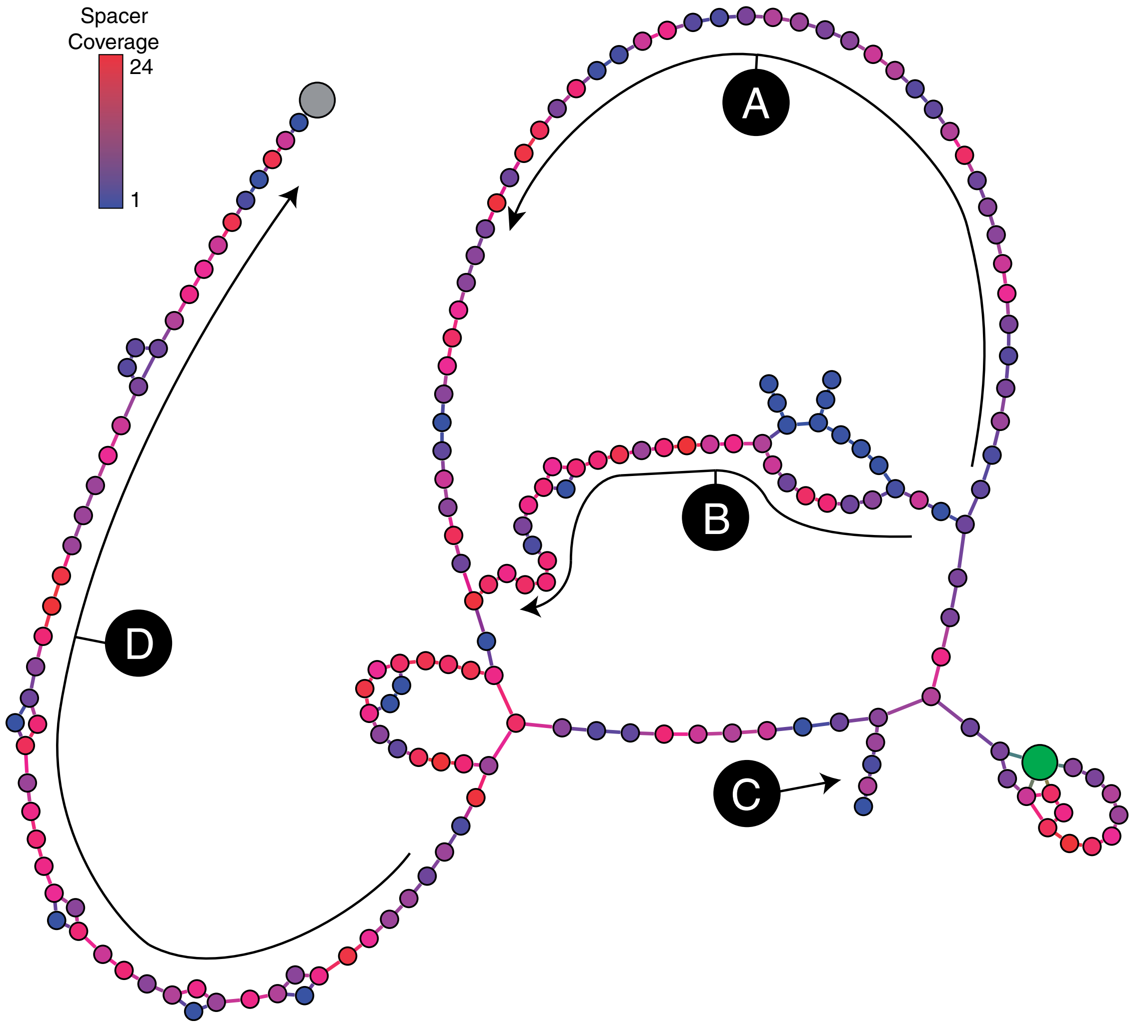

Figure 5. Reconstruction of the spacer arrangement of the most abundant CRISPR loci in the EBPR microbial metagenome. Each circle represents a spacer, and the lines connecting each spacer represent their positioning relative to other spacers. A spacer can be joined onto any number of other spacers (which indicates strain diversity in the population) and is coloured on a linear scale from blue to red, based on its coverage. The leader sequence (green circle) and distal end (grey circle) of the CRISPR are shown. There are two main spacer arrangements (**A** and **B**) from the leader to the tail region that merge into a conserved tail (**D**). A third arrangement contains unconnected spacers that may link into the leader sequence (**C**).

sensitivity when identifying CRISPR DRs and reconstructing the correct ordering of spacers (Table 1). However, owing to the strict filtering steps necessary to correctly group individual reads into DR types, spacers were missed in CRISPR loci where the DR sequence was not highly conserved (e.g. *N. magadii*). Crass performed comparably on the AMD data set identifying all of the DR types and the great majority of the spacers (Figure 3). Crass had higher specificity when calling the DR boundaries (Supplementary Figure S6) and identified novel spacers indicating that it had increased sensitivity when detecting DR containing reads. A clear advantage of Crass was the processing speed (~135 000 Sanger reads in 16 s) and detailed information about spacer ordering and diversity, which removes a major bottleneck when analysing CRISPRs in metagenomic data sets.

Crass-analysis of the GOS data, which had not previously been analysed at the read level, revealed substantially more DR types and spacers (Figure 3). The vast majority of unique Crass DR types were found on single reads that were not included in the assembly. However, some of these unique DR types were found in scaffolds but were missed as a result of the strict criteria used for DR identification in the original analysis, which required the DR to be detected by three separate programmes. The shared fraction of DR types from these programmes was relatively small resulting in many valid DR types being missed.

A detailed examination of the Crass spacers in the GOS data identified inconsistencies in the ordering for the most abundant DR types. One of the dominant DR types was reconstructed into four linear arrangements by Crass, but the same DR type was fragmented across 11 scaffolds, despite several of these scaffolds sharing identical spacer arrangements (Supplementary Figure S7). Conversely, there was evidence of spacers found by Crass that were missing from the assembled scaffold and likely the result of a mis-assembly (Supplementary Figure S8). Overall, fragmentation and mis-assembly of CRISPR loci created duplicated DR types and reduced the diversity of spacers, confounding the original analyses of the GOS data.

A comparison of DRs and spacers identified in the assembled EBPR metagenome by PILER-CR to the read level analysis of Crass revealed that a substantial number of DR types were not detected in the assembled data (Figure 4). The dominant DR type detected in the assembled contigs was substantially different to the reconstructed spacer ordering generated by Crass. This DR type was found on 20 contigs; however, Crass was able to create a spacer graph containing two main branches, likely representing population heterogeneity in this DR type (Figure 5). The assembly did not reconstruct either of these two strains, but instead produced contigs containing fragments from both branches. Using the assembly alone, it would be impossible to order the fragments correctly to reflect the strain variation that was identified with Crass.

Spacers detected in the EBPR metagenome with Crass were mapped to the phage metagenome sequences from the same time point. The only spacers with sequence homology to the phage metagenome corresponded to the genome of the most abundant phage in the bioreactor. These spacers belonged to the DR type containing the third largest number of spacers in the data set (Figure 4). Given that this DR type was low coverage and did not assemble, it likely originates from a low abundance member of the community. We posit that at the time of sampling, this host recently experienced a lytic event resulting in a high abundance of the attacking phage type and low relative abundance of the host population. Furthermore, mismatches to some of the spacers suggest that this phage is persistent in the system and rapidly evolving in response to the CRISPR conferred resistance of the host (Supplementary Figure S10). We speculate that the perfectly matched spacers were recently acquired and confer resistance to only a small number of individuals within this population.

The Crass algorithm balances trade-offs between processing speed and the accuracy of the spacer graph. Crass ran exceptionally fast on the data sets tested in this study, compared with other CRISPR finding tools (16 s AMD; 16 min EBPR; 90 min GOS), but its speed is primarily determined by the number of DR types found during the initial search phase and the size of the data set. To increase the processing speed, Crass avoids making pairwise sequence comparisons, instead using a kmer-based approach to link neighbouring spacers during graph building (see 'Materials and Methods' section). As a consequence, erroneous links can be made where many spacers begin or end in the same kmer (Supplementary Figure S7B). The kmer size is a user-defined parameter that should be maximized (ideally half the length of the spacer) to reduce the chances of erroneous links between spacers. Additionally, extra spacers can be created by sequencing errors that do not get resolved during graph cleaning (Supplementary Figure S7A). These spacers appear as bubbles in the final graph that must be resolved manually.

An important consideration when using Crass is that reads containing a specific DR type are analysed together; however, they may originate from multiple CRISPR loci belonging to different organisms, organisms within the same population or a duplicated locus within a single genome (13). Although Crass does not directly determine the number of discrete CRISPR loci, the spacer graph can be used to infer the number and diversity of loci. In the case of CRISPR loci derived from multiple unrelated organisms, Crass will likely create multiple unconnected graph arrangements indicating discrete CRISPR loci. Population heterogeneity within a CRISPR locus typically results in a graph that shares a common linear arrangement of spacers that splits into different pathways representing individuals within the population (Figure 5).

CONCLUSION

Crass provides a fast, accurate approach for exploring CRISPR diversity in metagenomic data sets without the need for assembly or prior knowledge of CRISPR in the data set. Examining CRISPR diversity in metagenomic data sets provides information important to understanding phage-host co-evolution. Substantially, more CRISPR loci and spacers could be identified in metagenomic data generated on all conventional sequencing platforms using Crass. The extra sensitivity and specificity of Crass revealed population heterogeneity and phage-host interactions that would not have been discovered in assembled data. Additionally, a fast and automated tool such as Crass is important for metagenomic investigation, as the size and complexity of these data sets is constantly increasing. The source code is licenced under the GNU public licence version 3 (GPLv3) and is freely available at http://ctskennerton.github.com/crass.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Table 1, Supplementary Figures 1–10, Supplementary Methods, Supplementary Results and Supplementary References [31,32].

ACKNOWLEDGEMENTS

The authors thank Florent Angly for providing the EBPR phage metagenome assembly for use in this work and Philip Hugenholtz for critically reviewing the manuscript.

FUNDING

Australian Research Council (ARC) [DP1093175] and strategic funds from the Australian Centre for Ecogenomics; G.W.T. is supported by an ARC Queen Elizabeth II fellowship [ARC-DP1093175]; C.T.S. is supported by an Australian Research Council Postgraduate Award (APA). Funding for open access charge: ARC grant [DP1093175].

Conflict of interest statement. None declared.

REFERENCES

1. Haft,D.H., Selengut,J., Mongodin,E.F. and Nelson,K.E. (2005) A guild of 45 CRISPR-associated (Cas) protein families and multiple CRISPR/Cas subtypes exist in prokaryotic genomes. *PLoS Comput. Biol.*, **1**, e60.
2. Sorek,R., Kunin,V. and Hugenholtz,P. (2008) CRISPR—a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat. Rev. Microbiol.*, **6**, 181–186.
3. Barrangou,R., Fremaux,C., Deveau,H., Richards,M., Boyaval,P., Moineau,S., Romero,D.A. and Horvath,P. (2007) CRISPR provides acquired resistance against viruses in prokaryotes. *Science*, **315**, 1709–1712.
4. Brouns,S.J., Jore,M.M., Lundgren,M., Westra,E.R., Slijkhuis,R.J., Snijders,A.P., Dickman,M.J., Makarova,K.S., Koonin,E.V. and van der Oost,J. (2008) Small CRISPR RNAs guide antiviral defense in prokaryotes. *Science*, **321**, 960–964.
5. Cui,Y., Li,Y., Gorge,O., Platonov,M.E., Yan,Y., Guo,Z., Pourcel,C., Dentovskaya,S.V., Balakhonov,S.V., Wang,X. *et al.* (2008) Insight into microevolution of *Yersinia pestis* by clustered regularly interspaced short palindromic repeats. *PloS One*, **3**, e2652.
6. Horvath,P., Romero,D.A., Coute-Monvoisin,A.C., Richards,M., Deveau,H., Moineau,S., Boyaval,P., Fremaux,C. and Barrangou,R. (2008) Diversity, activity, and evolution of CRISPR loci in Streptococcus thermophilus. *J. Bacteriol.*, **190**, 1401–1412.
7. Kunin,V., Sorek,R. and Hugenholtz,P. (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol.*, **8**, R61.
8. Makarova,K.S., Haft,D.H., Barrangou,R., Brouns,S.J., Charpentier,E., Horvath,P., Moineau,S., Mojica,F.J.M., Wolf,Y.I., Yakunin,A.F. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol.*, **9**, 467–477.
9. Edgar,R.C. (2007) PILER-CR: fast and accurate identification of CRISPR repeats. *BMC Bioinformatics*, **8**, 18.
10. Bland,C., Ramsey,T.L., Sabree,F., Lowe,M., Brown,K., Kyrpides,N.C. and Hugenholtz,P. (2007) CRISPR Recognition Tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. *BMC Bioinformatics*, **8**, 209.
11. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.*, **35**, W52–W57.
12. Tyson,G.W. and Banfield,J.F. (2008) Rapidly evolving CRISPRs implicated in acquired resistance of microorganisms to viruses. *Environ. Microbiol.*, **10**, 200–207.
13. Andersson,A.F. and Banfield,J.F. (2008) Virus population dynamics and acquired virus resistance in natural microbial communities. *Science*, **320**, 1047–1050.
14. Sorokin,V.A., Gelfand,M.S. and Artamonova,I.I. (2010) Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Appl. Environ. Microbiol.*, **76**, 2136–2144.
15. Pride,D.T., Sun,C.L., Salzman,J., Rao,N., Loomer,P., Armitage,G.C., Banfield,J.F. and Relman,D.A. (2011) Analysis of streptococcal CRISPRs from human saliva reveals substantial sequence diversity within and between subjects over time. *Genome Res.*, **21**, 126–136.
16. Rho,M., Wu,Y.W., Tang,H., Doak,T.G. and Ye,Y. (2012) Diverse CRISPRs evolving in human microbiomes. *PLoS Genet.*, **8**, e1002441.
17. Stern,A., Mick,E., Tirosh,I., Sagy,O. and Sorek,R. (2012) CRISPR targeting reveals a reservoir of common phages associated with the human gut microbiome. *Genome Res.*, **22**, 1985–1994.
18. Zerbino,D.R. and Birney,E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, **18**, 821–829.
19. Wu,S. and Manber,U. (1994). University of Arizona, Tuscon, USA, A fast algorithm for multi-pattern searching, Vol. TR 94–17.
20. Grissa,I., Vergnaud,G. and Pourcel,C. (2007) The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. *BMC Bioinformatics*, **8**, 172.
21. Angly,F.E., Willner,D., Rohwer,F., Hugenholtz,P. and Tyson,G.W. (2012) Grinder: a versatile amplicon and shotgun sequence simulator. *Nucleic Acids Res.*, **40**, e94.
22. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
23. Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.
24. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
25. Drummond,A.J., Ashton,B., Buxton,S., Cheung,M., Cooper,A., Duran,C., Field,M., Heled,J., Kearse,M., Markowitz,S. *et al.* (2012), Geneious 5.5 ed.
26. Langmead,B. and Salzberg,S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods.*, **9**, 357–359.
27. Namiki,T., Hachiya,T., Tanaka,H. and Sakakibara,Y. (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Res.*, **40**, e155.
28. Koren,S., Treangen,T.J. and Pop,M. (2011) Bambus 2: scaffolding metagenomes. *Bioinformatics*, **27**, 2964–2971.
29. Garcia Martin,H., Ivanova,N., Kunin,V., Warnecke,F., Barry,K.W., McHardy,A.C., Yeates,C., He,S., Salamov,A.A., Szeto,E. *et al.* (2006) Metagenomic analysis of two enhanced biological phosphorus removal (EBPR) sludge communities. *Nat. Biotechnol.*, **24**, 1263–1269.
30. Deveau,H., Barrangou,R., Garneau,J.E., Labonte,J., Fremaux,C., Boyaval,P., Romero,D.A., Horvath,P. and Moineau,S. (2008) Phage response to CRISPR-encoded resistance in streptococcus thermophilus. *J. Bacteriol.*, **190**, 1390–1400.
31. Lu,H., Oehmen,A., Virdis,B., Keller,J. and Yuan,Z. (2006) Obtaining highly enriched cultures of Candidatus Accumulibacter phosphates through alternating carbon sources. *Water Res.*, **40**, 3838–3848.
32. John,S.G., Mendez,C.B., Deng,L., Poulos,B., Kauffman,A.K., Kern,S., Brum,J., Polz,M.F., Boyle,E.A. and Sullivan,M.B. (2010) A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ. Microbiol. Rep.*, **3**, 195–202.