BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 13 2014, pages 1805–1813
doi:10.1093/bioinformatics/btu114

Genome analysis

Advance Access publication February 27, 2014

Accurate computational prediction of the transcribed strand of CRISPR non-coding RNAs

Ambarish Biswas[1], Peter C. Fineran[2,3] and Chris M. Brown[1,3,*]

[1]Department of Biochemistry, [2]Department of Microbiology and Immunology and [3]Genetics Otago, University of Otago, Dunedin 9054, New Zealand

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: CRISPR RNAs (crRNAs) are a type of small non-coding RNA that form a key part of an acquired immune system in prokaryotes. Specific prediction methods find crRNA-encoding loci in nearly half of sequenced bacterial, and three quarters of archaeal, species. These Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) arrays consist of repeat elements alternating with specific spacers. Generally one strand is transcribed, producing long pre-crRNAs, which are processed to short crRNAs that base pair with invading nucleic acids to facilitate their destruction. No current software for the discovery of CRISPR loci predicts the direction of crRNA transcription.

Results: We have developed an algorithm that accurately predicts the strand of the resulting crRNAs. The method uses as input CRISPR repeat predictions. CRISPRDirection uses parameters that are calculated from the CRISPR repeat predictions and flanking sequences, which are combined by weighted voting. The prediction may use prior coding sequence annotation but this is not required. CRISPRDirection correctly predicted the orientation of 94% of a reference set of arrays.

Availability and implementation: The Perl source code is freely available from http://bioanalysis.otago.ac.nz/CRISPRDirection.

Contact: chris.brown@otago.ac.nz

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 29, 2013; revised on January 30, 2014; accepted on February 20, 2014

1 INTRODUCTION

Many bacteria and archaea have defense systems that target incoming nucleic acids, termed CRISPR-Cas (Clustered Regularly Interspaced Short Palindromic Repeats–CRISPR-associated proteins) systems. Recent studies have shown that these RNA-mediated systems are a major antiviral mechanism in prokaryotes (Fineran and Charpentier, 2012; Makarova *et al.*, 2013; Sorek *et al.*, 2013; Wiedenheft *et al.*, 2012).

In bacterial genomic sequences, CRISPRs are characterized by repeated elements (repeats) interspersed with unique spacers. In most systems, RNA Polymerase transcribes the CRISPR array in one direction to produce the pre-CRISPR RNA (pre-crRNA), which is processed to provide separate small RNA effectors (crRNAs) that target foreign genetic elements (Lillestol *et al.*, 2009; Richter *et al.*, 2012a). The mechanism of targeting requires a number of Cas (CRISPR-associated proteins) that are typically encoded nearby to the CRISPR arrays. Most CRISPR-Cas systems target DNA from bacteriophages, plasmids and mobile elements in bacterial chromosomes (Brodt *et al.*, 2011; Makarova *et al.*, 2013; Sorek *et al.*, 2013); however, some CRISPR-Cas systems target RNA (Hale *et al.*, 2009).

The direction of transcription is unknown for most CRISPR arrays, and current algorithms do not predict direction. Normally, only one strand is transcribed and processed into small crRNAs, which determines base pairing target specificity. The crRNAs provide target specificity through base pairing for the CRISPR-Cas system. Therefore, target prediction algorithms permit both potential crRNAs to be analyzed, as the correct strand is often unknown (Biswas *et al.*, 2013).

There are several CRISPR discovery programs available to find these repeat arrays in genomic (Bland *et al.*, 2007; Edgar, 2007; Grissa *et al.*, 2007; Rousseau *et al.*, 2009) or metagenomic sequences (Rho *et al.*, 2012; Skennerton *et al.*, 2013). The most cited of these (CRISPRFinder) locates CRISPR arrays in 46% (1080 in 2355) of the complete genomes of bacteria and in 84% (126 in 150) of archaea. As CRT and PILER-CR are better suited to automation, they are used in bacterial genome annotation pipelines (Mavromatis *et al.*, 2009).

Recent CRISPR-Cas classifications include three types (I–III) and at least ten subtypes (Makarova *et al.*, 2011). These are based on multiple criteria, including the presence of specific Cas proteins. For those type III-B systems that target single-stranded RNA, appropriate assignment of the direction of CRISPR expression is essential to accurately identify RNA targets and avoid false target predictions. Whereas for type I and II, additional short specific sequences are associated 5′ or 3′ of true targets and function in different steps of the CRISPR-Cas process. These sequences are termed protospacer adjacent motifs (Mojica *et al.*, 2009), and recent work has suggested that these can be further discriminated into target interference motifs and spacer acquisition motifs that function at the different stages (Shah *et al.*, 2013). The identification of these additional determinants requires knowledge of the transcribed strand (Biswas *et al.*, 2013; Shah *et al.*, 2013). The ability to accurately identify the target strand would also be useful to determine if during conjugation the single strand initially transferred into the recipient is preferentially targeted (Westra *et al.*, 2013).

Repeats may be dissimilar between the same types in different species, and the relationship between sequence and type is complex, i.e. near identical repeats in the same species can be

*To whom correspondence should be addressed.

associated with different types, e.g. I-B and III-B (Nickel *et al.*, 2013), and multiple types commonly exist with a species (Makarova *et al.*, 2011). However, repeats can independently be clustered and grouped into sequence or structural classes (Kunin *et al.*, 2007). The most recent analysis puts 3527 repeats into 33 structural motif classes, 40 sequence family classes and 6 superclasses. Some of these classes are preferentially associated with specific CRISPR subtypes (Lange *et al.*, 2013).

In some types (I and II), the repeat RNA has a stable stem loop, typically with 4–6 bases in the stem (Nickel *et al.*, 2013; Scholz *et al.*, 2013). Pre-crRNA processing occurs within the repeat, resulting in 5′ and 3′ handles derived from the repeats and attached to the spacer (Fig. 1). This repeat structure within the pre-crRNA is required for processing in type I-F systems (Sternberg *et al.*, 2012). In some systems, the 5′ crRNA handle is UAAGAAA derived from the 3′ end of the repeat (Maier *et al.*, 2013; Wang *et al.*, 2011). In addition, it has recently been shown that for a few species using type II systems, crRNA transcription can occur from multiple promoter sequences (TANAAT −10 like) within the 3′ end of the repeat (Zhang *et al.*, 2013).

CRISPR arrays expand by acquiring specific sequences from invading nucleic acids. Although it is not completely understood, this adaptation typically involves addition of new spacers at the 5′ end of the array (Fineran and Charpentier, 2012; He and Deem, 2010; Lillestol *et al.*, 2009; Westra and Brouns, 2012). However, there is evidence of some systems acquiring new spacers at positions internal to the CRISPR array (Erdmann and Garrett, 2012). Spacers and repeats may subsequently degenerate, then be lost by successive point mutations, deletions or recombination (Gudbergsdottir *et al.*, 2011; He and Deem, 2010). However, as clear targets of most crRNAs are not present in sequence databases, it is difficult to determine if mutations have occurred in spacers. In addition, it is likely that both target nucleic acids and host crRNAs are accumulating mutations after the initial contact and spacer acquisition (Levin *et al.*, 2013; Sun *et al.*, 2013; Weinberger *et al.*, 2012). For example, phages can escape CRISPR-Cas interference via the acquisition of point mutations in their target sequences, and old spacers might accumulate mutations when there is no longer a selective advantage (i.e. in the absence of the target phage/plasmid) (Levin *et al.*, 2013; Vercoe *et al.*, 2013).

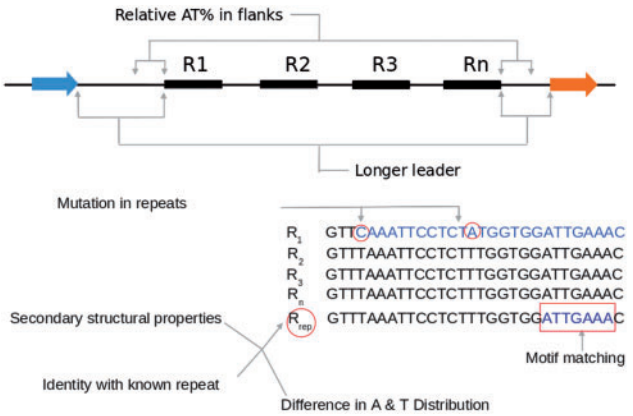

Fig. 1. Overview of the methods used in CRISPRDirection. Arrays were predicted using CRT and PILER-CR. An array with n repeats R_1–R_n is shown. Parameters were calculated from the repeats or using flanking regions, e.g. the distance to the next annotated protein CDS (arrows) (Section 2). Relative AT% in the flanks was determined in fixed windows. For each array, a representative repeat was determined and then assessed for motifs, composition and secondary structure (Section 2)

The 'leader' is the sequence that precedes the array and contains both the promoter and 5′ region before the first repeat (Pougach *et al.*, 2010; Przybilski *et al.*, 2011). In some characterized systems, there is clear accumulation of point mutations in the repeats (and spacers) at the 3′ end of the array (leader-distal end). Consistent with this observation, current data indicate that new spacers are acquired at the 5′ (leader) end of the CRISPR array, and older spacers and repeats decay at the 3′ end (trailer end spacers) (Swarts *et al.*, 2012; Weinberger *et al.*, 2012; Yosef *et al.*, 2012).

Previous studies have predicted the direction of repeats as parts of larger studies. It has been observed that repeats in archaea are A-rich and this has been used to determine direction in that domain (Chan *et al.*, 2012; Shah and Garrett, 2011). Recently, Lange *et al.* (2013) used A-richness, and the presence of ATTGAAA(C/G) at the 3′ end of repeats, to determine the direction of repeats (Lange *et al.*, 2013).

In this article, we aimed to use information within, and nearby, predicted CRISPR arrays to determine the direction of transcription. Prediction was performed as much as possible independently of the use of biological knowledge (e.g. repeat class), as this is not known for many sequences. The algorithm is described and available as a Perl module for a gene prediction and annotation pipeline. This CRISPRDirection algorithm predicts with >90% accuracy the direction of transcription and thus the correct sequences of the crRNAs that target foreign nucleic acids. CRISPRDirection will be useful for more accurate prediction of CRISPR arrays, and subsequent prediction of their types and targets, such as via CRISPRTarget (Biswas *et al.*, 2013).

2 METHODS

2.1 Preprocessing of arrays

2.1.1 Prediction of arrays Arrays were predicted with CRISPRFinder, PILER-CR and CRT on complete bacterial (2341 species with 4458 sequences) and archaeal (153 species with 236 sequences) genomes using the default parameters (Bland *et al.*, 2007; Edgar, 2007; Grissa *et al.*, 2007). Only CRISPR arrays with two or more spacers (i.e. three or more repeats) were used in the analyses. The three programs predicted 4053, 3471 and 3166 CRISPRs, respectively. There was partial or full overlap between all three predictions for 3115 arrays. Because arrays predicted with CRISPRFinder do not show variation in individual repeats in the final output, and its dependencies are not freely available, the CRISPRFinder-predicted arrays were not used. Because similar arrays were often predicted by CRT and PILER-CR, only the longer ones were used from the combined list of predictions. For overlapping predictions, the longest array was selected. The final list included 3571 CRISPR arrays. Those pipelines that include CRISPR annotation use a similar methodology.

2.1.2 Identifying representative repeat sequences As the PILER-CR output format provides most of the information required by CRISPRDirection, we converted the CRT-predicted arrays into PILER-CR format. However, as CRT does not provide a representative

repeat, we had to determine this for the arrays predicted by CRT. To find the representative/centroid repeat for each array, we used a match-scoring method, then the frequency of individual repeats. We first collected all the repeat sequences of an array and the number of individual occurrences in the array. At this stage if two repeats have single base difference, they were treated as two separate sequences. We initially selected the sequence with the least mismatches as the representative repeat. Where two or more sequences had the same degeneracy score, the most common repeat was selected. If these were equal, the repeat with fewer mismatches at the ends of the array was selected. The representative repeats were collected from the 3571 arrays, and after identical or reverse complemented repeats were removed, there were 2023 unique repeats.

2.1.3 Reference set of arrays and repeat sequences To develop a reference set of CRISPRs with known orientation, a set of 26 repeats with experimentally determined direction was used, 13 repeats were taken from Biswas *et al.* (2013) and supplemented with 13 repeats from the recent literature (Supplementary Table S1). These were chosen to cover each of the 10 subtypes and a range of structural and sequence classes (Section 4 and Supplementary Material S3) (Kunin *et al.*, 2007; Lange *et al.*, 2013). There were 208 arrays containing these exact 26 repeats.

To generate a larger reference set of arrays for our analysis, we searched the 2023 repeats for sequences similar to the 26 repeats of known orientation. In this comparison, repeats with a maximum of three mismatching bases were retained. This gave us a set of 120 unique repeats that were present in 575 distinct arrays (Supplementary Table S8).

We removed arrays with >95% sequence identity using CD-HIT-EST (Supplementary Table S2). This reduced the number of arrays to 460 arrays, 340 from 205 bacterial and 120 from 43 archaeal species. These 460 arrays were used as a reference set in our analyses.

2.2 Calculation of parameters

We derived and used a number of parameters and attributes from the repeats and regions immediately flanking the array. Each of these sets of parameters was used as a method to predict the direction of the array as either F (Forward) or R (Reverse). For some methods all 460 arrays could be used, and for others, not all arrays were suitable for the analysis.

If the array was not valid for the analysis, or the analysis results in a draw, it reports NP (No Prediction). For example, if there are no mutations in an array, the reported direction by the method dependent on these will be NP (Supplementary Fig. S4). A set of 100 CRISPR arrays were selected randomly in each run and the set is used for all the methods. This was repeated 100 times. The mean Positive Predictive Value (PPV) was calculated using resampling by re-substitution, as n is much larger than the number of predictors to be estimated. An alternative method of resampling by 10-fold cross-validation is provided in Supplementary File S2, which produced similar predictors and scores ± 0.03. As the random probability of getting a true prediction is 0.50, we used as a weighted score (score) for each method its PPV minus 0.50.

2.2.1 Screen for ATTGAAA-like sequences The sequence ATTGAAA and variants shown in Table 1 were searched for individually on both strands of the representative repeat (Fig. 1). Any single base, or no base, was allowed following the motif, to allow for inaccuracy in prediction of the repeat/spacer transition in tuning the algorithm.

2.2.2 Biased nucleotide composition in the 5′ and 3′ flanks relative AT richness In testing, we compared the percentage of nucleotides A and T (AT%) in different length windows ranging from 15 to 165 nucleotides with a step of 15 nucleotides flanking the 5′ and 3′ ends of the arrays (Fig. 1). The AT% in each of these regions were compared, in testing, to determine if they have an absolute minimum percentage difference ranging from 0.5 to 14.5%. When the AT differences for windows were greater than the minimum cutoff, a prediction was obtained. For example, if the 5′ end of the array showed a higher AT% for a window size (e.g. 135) and minimum cutoff (e.g. 2.5%) the prediction was F. Similarly, the predicted direction will be R when the 3′ end of the array shows a higher AT% than the 5′ end of the array. For all other cases, NPs are reported.

2.2.3 Degeneracy at the 3′ end of the array The array was divided into three sections: 5′, middle and 3′ sections, each with an equal number of repeats, where possible, or the nearest integer. For arrays with three or four repeats, both 5′ and 3′ sections had one repeat each. Of the 460 arrays, 164 had no mutations in the array and were excluded from this analysis. The remaining 296 had one or more mutations and a direction could be predicted. We calculated the total number of mutations present in these sections. In this analysis, a base was denoted as mutated only when it differed from an adjacent repeat (Supplementary Table S2). This rule of mutation was necessary as many long CRISPR arrays were shown to have a point mutation, which is propagated to the next repeats, resulting in two common repeats in a single array. Often, both the repeats were found to have more than one-third of repeats associated with them, making it difficult to select one as the representative repeat.

The score for the 5′ and 3′ sections were calculated in the following manner: 5′ section score = Total number of mutations + Observed mutation in the first repeat; 3′ section score = Total number of mutations + Observed mutation in the last repeat.

The 'observed mutation in the first/last repeat' can either be True (1) or False (0), and this was useful to correctly identify the direction when the total number of mutations in both 5′ and 3′ section was equal. The end with the lowest score was predicted to be the 5′ end.

2.2.4 Potential for RNA secondary structure in the repeat For this analysis, we first removed the redundant repeats from the sample (Supplementary Table S2) and used RNAfold to measure the Minimum Free Energy (MFE) of the repeats in both directions (Hofacker *et al.*, 1994). In addition, we masked bases from 0 to 11 bases from both ends and calculated the MFE of the center part of the repeats and their reverse complements. The direction with the lowest free energy for those with differences of >0.5 kcal/mol was designated as forward.

2.2.5 A relative to T in the repeat A/T ratio For this analysis, the relative number of A to T in the representative repeat was measured and the direction with the highest A/T ratio designated as forward.

2.2.6 Distance to the next coding gene at the 5′ and 3′ end length of leader analysis Using the start and stop positions of the 460 arrays, we obtained the positions of next CoDing Sequences (CDS) from their annotation files (NCBI Genbank format, gbk). All CDSs overlapping any CRISPR arrays were omitted (99 CDS). The closest CDS coordinate on the 5′ end of the array and the closest CDS coordinate on the 3′ end of the array were analyzed. The end with the longest distance was designated as forward.

2.2.7 Parameters used by CRISPRDirection.pl. method (sections 2.2.1–2.2.6): parameters if any, score (PPV-0.5) The high precision parameters and scores are in parentheses. A flow diagram of the methods is shown in Supplementary Figure S4.

(i) Specific ATTGAAA(N) motif: score, 0.50 (4.5)
(ii) Biased AT richness in flanks: window, 60 (120); AT% difference ≥ 10 (14); score, 0.32 (0.45)
(iii) Degeneracy at 3′ end: minimum mutations required, 1 (2), score, 0.41 (0.46)
(iv) RNA secondary structure: bases ignored at both ends, 5 (8), score, 0.37 (0.39)
(v) Longer distance to next CDS: ratio >200%; min length of longer sequence, 75; score, 0.18 (0.18)
(vi) A/T ratio in repeat: score, 0.37 (0.37).

Table 1. Presence of ATTGAAA-like motifs at the 3′ end of the array

Sequence	TP	FP	PPV
ATTGAAA(C/G)	130	0	1.00
ATTGAAA(N)	137	0	1.00
NTTGAAA(N)	140	0	1.00
ANTGAAA(N)	140	0	1.00
ATNGAAA(N)	148	0	1.00
ATTNAAA(N)	143	18	0.88
ATTGNAA(N)	111	0	1.00
ATTGANA(N)	137	0	1.00
ATTGAAN(N)	141	0	1.00
WWWGAAA(N)	240	9	0.96

Note: (N): any or no base, W: A or T, TP, true positive; FP, false positive; PPV, positive predictive value.

2.2.8 Final prediction The score for forward or reverse is the sum of the individual scores from each of the methods (e.g. Total_Forward_score is the sum of the forward scores, Supplementary Fig. S4). NP is not counted. The Confidence_score (CS) = absolute value of (Total_Forward_score − Total_Reverse_score). Predictions are made with 'high' confidence if the CS is ≥66% of the 'Sum of all Applied Scores' (SAS = Total_Forward_score + Total_Reverse_score) and also > 0.5; 'medium': CS <66% ≥33% SAS and 'low': <33% SAS. The required confidence score of >0.5 for 'High' confidence is required to exclude results solely from less accurate prediction methods, where the more accurate methods did not produce a prediction (NP).

2.3 Availability

An implementation in Perl with documentation and test data is available at http://bioanalysis.otago.ac.nz/CRISPRDirection. This will take as input a fasta or Genbank formatted, annotated or unannotated genomic sequences, or simply repeat sequences in fasta format. The algorithm is shown in Supplement S4, and output is shown in T5.

3 RESULTS

We aimed to develop an algorithm to detect the direction of CRISPR transcription. This would, as much as possible, be independent of specific biological knowledge for known types of arrays, as often this is neither known nor predicted. Such a method could be applied within a pipeline to new instances and potentially new classes of arrays. The predictive methods adopted were (i) searching for ATTGAAA-like sequence motifs in the repeats, (ii) finding biased nucleotide composition (relative AT richness) in the leader and trailer regions of the arrays, (iii) analysis of mutation(s) in the repeats (array degeneracy), (iv) analysis of RNA secondary structure in the repeats, (v) finding the longer distance to the next coding gene at the 5′ and 3′ ends and (vi) greater number A compared with T in the repeat. These methods are depicted in Figure 1.

A reference set of known direction was assembled by using a set of 26 repeats of experimentally determined orientation in 135 unique arrays. This was extend to 460 repeats by allowing up to three mismatches (see Section 2.1).

3.1 Specific sequences at the 3′ end of the repeat

A recent analysis of repeats noted that about a third of their set of repeats end in ATTGAAA(C/G), and they used that to predict repeat direction (Lange *et al*., 2013). To investigate whether this was a useful predictor on our independently created reference set, we searched for variants of ATTGAAA(N) at the 3′ end on both strands. An exact match was found in 30% (137) of the 460 reference arrays. In all these cases, this correctly predicted the direction of the array (PPV = 1.00, Table 1). Similar one-base variants of all the bases except the G slightly increased the number of predictions but retained PPVs of 1.00 (Table 1). Extensive substitution at the 5′ end, e.g. WWWGAAA(N), still retained a high PPV (0.96, Table 1). For those 137 arrays that had ATTGAAA(N), we used this as a certain predictor of the direction of these arrays. This parameter was given a score of 0.50 (PPV of 1.00–0.50 random chance).

3.2 Biased AT composition in the 5′ and 3′ flanks

The promoter/leader sequence 5′ of the CRISPR array would be expected to differ from the terminator sequence following it. These promoters have been described as being AT rich in several studies, and some AT-rich sequences (e.g. TATAAT-like sigma 70 promoter −10 elements) have been shown to be functional in CRISPR promoters (Lillestol *et al*., 2009; Pul *et al*., 2010). The AT% was calculated in windows at the two flanks and the percent difference compared. An AT% difference of 0.5–14% and windows up to 165 bp long were compared (see Section 2.2.2). The analysis was repeated 100 times for new randomly selected sets of 100 CRISPR arrays in each run, and the results were compared to determine the window length and minimum percentage difference in the AT that gave an optimal PPV (Supplementary Fig. S1). This sampling approach was used to calculate measures of variability within our set of 460 repeats, which is a sample of all possible CRISPR arrays.

As shown in Figure 2, a window size of 60 with a minimum AT% difference of 10% gave a high PPV (0.82) with a tolerable NP (TP 39, FP 9 and NP 53). Higher PPVs could be achieved, e.g. window size of 105, AT% difference >14%, PPV 0.95 (Supplementary Fig. S1). However, this combination could only be applied to one-third of the data. As this method is one of a combination to predict the probable direction, we used a window size of 60, >10% AT% as a balanced default for CRISPRDirection. Alternative combinations can be applied in the implementation, and results are shown in Supplementary Figure S1.

3.3 Degeneracy at the 3′ end of the array

Based on the idea that repeat sequences at the 3′ end may have accumulated mutations, we looked for differences at either end of the predicted arrays. Figure 3A shows that a PPV of >0.90 could be achieved (TP 89, FP 9 and NP 2). Some of the FPs may be caused by a random mutation or rare sequencing errors within the arrays. A higher accuracy (PPV 0.95) can be achieved by only including arrays with a minimum of two or three mutations (Fig. 3A). However, as that increases the number of NP (to ∼40%) we did not set any minimum mutation requirement

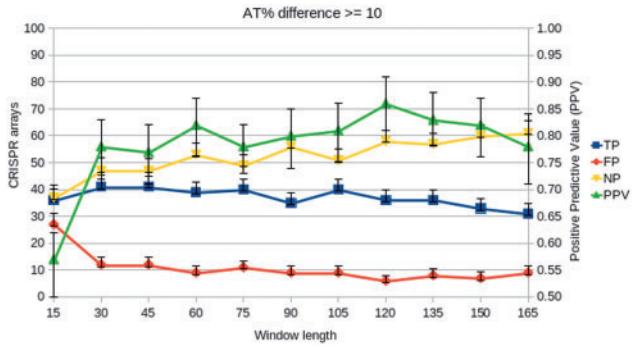

Fig. 2. Relative AT richness in the two flanks. AT% for windows of the specified lengths was calculated on either flank; the side with the higher AT% was used as a predictor of the leader. TP: True positives, FP: False positives, NP: No Prediction, PPV: Positive Predictive Value (green triangles). Data are the mean and SD (error bars) of 100 samples of 100 from the reference set of 460 arrays

by default. For 296 valid arrays (2.2.3) this method had 263 TPs, 28 FPs and 5 NPs.

3.4 Potential for RNA secondary structure in the repeat

Some CRISPR repeats are known to form small RNA secondary structures, which would be strand-specific. We used a simple method to test for the greater ability to form thermodynamically stable secondary structures on each strand. This method was designed to avoid requiring prior knowledge of classes or structures in different classes when predicting the direction.

The hypothesis was that theMFE of the forward strand would be lower compared with the MFE of the other strand. In an initial analysis, we observed that calculating the MFE of the whole repeats gives a poor prediction accuracy (Fig. 3B, PPV 0.55). The RNA secondary structural elements, where experimentally characterized, are typically present toward the center of the repeats rather than at the ends (Lange *et al.*, 2013). Therefore, we progressively masked the leading bases from the both ends of the repeats and calculated the MFE (Fig. 3B).

The accuracy of the prediction increased as the ends were masked by up to five bases (Fig. 3B and PPV 0.87). Further base neutralization showed that although the PPV was increased, the number of true predictions rapidly reduced (e.g. PPV 0.90 and NP 26% at 8). The average repeat length is 31 nucleotides, removing up to five bases from both sides (total 10 bases) still kept the secondary structural elements intact in the middle for most cases and gave a more accurate prediction. However, as more bases were masked, only longer repeats gave useful results, indicated by the rapid drop off after eight bases in Figure 3B.

3.5 Analysis of A relative to T in repeats—A/T ratio analysis

Several studies have also used A/T ratios in the repeat as a predictor of direction, as it has been observed that archaeal repeats are A rich (Chan *et al.*, 2012; Lange *et al.*, 2013). The notion is the forward strand will have more A than T. The PPV for this was 0.72 (68 TP, 27 FP, 5 NP of samples of 100). During this analysis, we observed that AT-rich repeats had a higher rate of

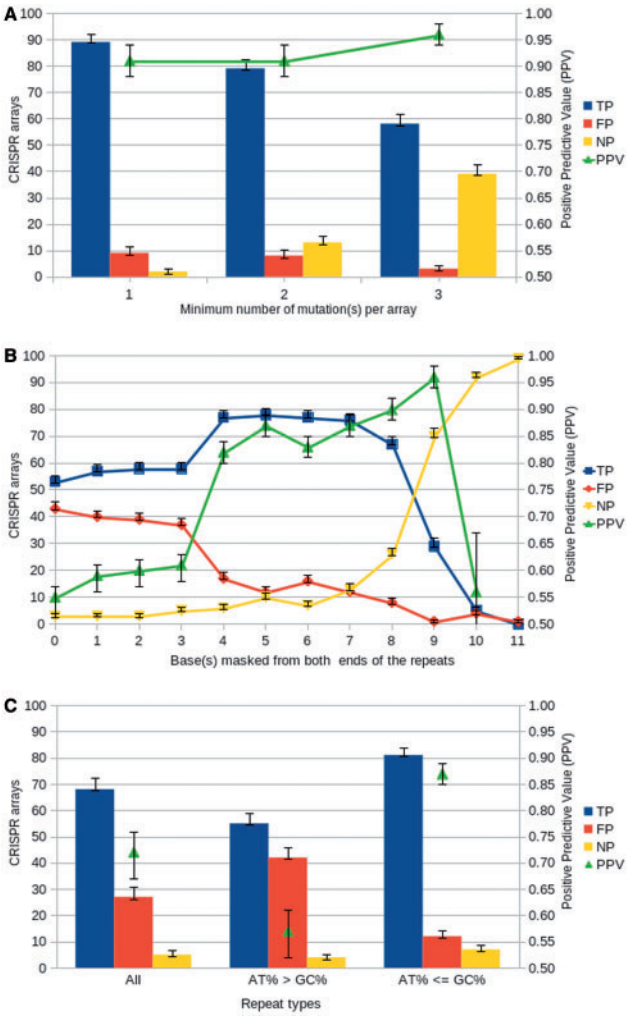

Fig. 3. Repeat parameters. (**A**) Mutations in the repeats at the 3′ (leader-distal) end of the array (Section 3.3). (**B**) Relative stability of predicted RNA secondary structure (Section 3.4). (**C**) Relative A/T ratio (Section 3.5). TP: True positives, FP: False positives, NP: No Prediction, PPV: Positive Predictive Value (green triangles). Data are the mean and SD (error bars) of 100 random samples of 100

false predictions. To investigate whether this has a clear pattern, which can be applied to improve the overall prediction, we separated the repeats into two groups. The first group contains the AT-rich repeats (i.e. AT% > GC%) and the second group contains AT-poor repeats and reapplied the method. The result (Fig. 3C) shows that this method is better applied to repeats that are AT poor (PPV 0.87).

3.6 Distance to the next coding gene at the 5′ and 3′ end length of leader analysis

The 'leader' region containing the promoter may be longer than the 'trailing' region containing the transcription terminator (Fig. 1). To test if this could be used to predict the direction, the positions of the nextCDS at the 5′ and 3′ end were obtained from their corresponding Genbank annotation. Only this

method, of the prediction methods implemented, relies on other types of gene annotation. The distances were compared to obtain a parameter that could be tested for use in prediction. As yet the NCBI Reference sequences do not include CRISPR predictions and leave them as long intergenic gaps. Hence, CDS annotations may not be accurate in these regions, and it is possible to have undetected CDS close to the CRISPR arrays, or falsely predicted CDS close or within the arrays. Indeed, 99 predicted probable arrays showed overlapping predicted CDS. These overlapping CDS were ignored for this analysis.

Distances ranged from 2 to 5252 bases to the next CDS for the set. However, often both distances were short (<75), making prediction unreliable. Hence, instead of only testing for the longer sequence, we also compared if the longer sequence was at least 120, 140, 160, 180 or >200% the size of the shorter, and required that at least one of the lengths be >75 nucleotides (Supplementary Fig. S4). NP was made if the ratio did not meet the criteria or the length was too short (Fig. 4). These predictions can be found in the Supplementary spreadsheet (Supplementary File S2). The maximum PPV obtained was 0.68 (>200%, Fig. 4), this was selected to be used by default. However, about half the arrays did not have such long leaders and resulted in NP calls (Fig. 4).

3.7 Combined prediction

Each of the six predictors was combined. To weight more accurate predictions higher (i.e. assigning higher scores to more accurate methods), each of the six predictions was weighted according to PPV (Section 2.2.8). Thus, the highest weighted prediction was ATTGAAA(N) (0.5) followed by degeneracy (0.41) and the lowest weighted prediction was longer leader (0.18). The combined method correctly predicted 424 arrays of the 460 CRISPRs used as reference with a PPV of 0.94, FP 27 and NP 4 (Bacteria: PPV 0.92, TP 317, FP 19, NP 4; Archaea: PPV 0.93, TP 112, FP 8, NP 0).

To further assess the robustness of the method, the algorithm was run without each of the methods in turn and without combinations of related methods. Leaving out the most accurate predictor (ATTGAAA(N) Table 1) had a small effect reducing the PPV by ~4% (Supplementary Table S4 and Supplementary Fig. S2). Leaving out the degeneracy had the greatest single impact on PPV, reducing it to 0.91. Leaving out the longer leader method, the only measure that depends on CDS annotation reduced had little effect. If only repeat sequences were used as input, an option in the implementation, PPV was 0.87 using three methods (i.e. Motif search, A/T ratio and RNA structure analysis) that only require the repeat.

An alternative prediction strategy was to predict using the highest practical PPV from each method (window size 120 and minimum AT% difference 14 in relative AT-richness analysis, 8 masked bases in the RNA secondary structure analysis and by requiring two or more mutation in the array degeneracy analysis). Using this alternative set of options stated as 'high precision parameters' in the section 2.2.7, we obtained a PPV of 0.95 against the whole reference set, with 413 TP, 21 FP and 26 NP (bacteria: PPV 0.95, NP 26; archaea: PPV 0.97, NP 4, Supplementary File S2).

Predictions differ in total scores; therefore, a prediction confidence could be determined. Predictions are flagged with 'high' confidence if they met defined criteria (Section 2.2.8 and Supplementary Fig. S4). Final predictions, and confidence scores, for the direction for the full dataset of 3571 arrays are available in the Supplementary Material on the companion Web site.

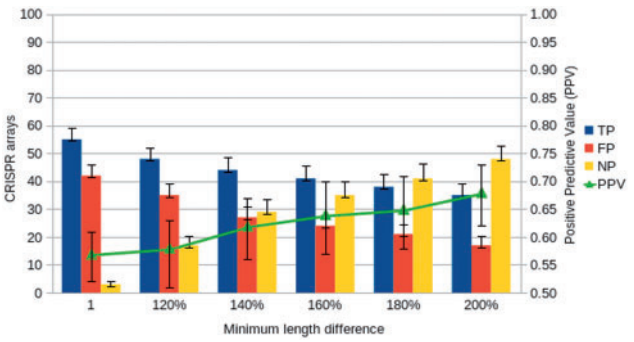

Fig. 4. The distance to the next CDS was determined for each flank. The side that was longer was predicted to be the leader. TP: True positives, FP: False positives, NP: No Prediction, PPV: Positive Predictive Value. Data are the mean and standard deviation of 100 random samples. PPV values are poor (near random) when the lengths are similar (~1). The highest PPV (0.68) is obtained when the ratio is >200%

4 DISCUSSION

The algorithm described works surprisingly well using a set of relatively easily calculable parameters. At the default level of accuracy, it correctly predicts 94% of reference arrays. Higher levels of accuracy (e.g. PPV 0.95 and NP 26) are possible by changing some of the parameters. This method can be built into a generic prokaryotic prediction pipeline, without a requirement of knowledge whether the sequence is from bacteria or archaea and without the need to define the type, subtype, class or superclass of the array or repeat. However, it is also possible for users analyzing specific genomes to incorporate biological knowledge, e.g. specific motifs or presence of RNA structure into the prediction.

4.1 Reference set

The method was tested on a set of 460 arrays, and these consisted of 208 arrays having exact matches to a set of 26 experimentally verified repeats from all 10 subtypes. This set was extended by allowing a small number of mismatches (up to 3 in repeats of mean length 31). Within genera and classes of arrays, there is often a large amount of substitution in arrays (Kunin *et al.*, 2007). Including this small tolerance might include some non-functional arrays, but is unlikely to reverse the arrays. We compared our array sets with the repeat sets used by Lange *et al.* (2013). Our set had members in 7 of the 10 largest structural classes and 9 of 10 largest sequence classes described (Supplementary Fig. S3 and Supplementary Tables S7 and S9) (Lange *et al.*, 2013). This set provides a new independent reference set for CRISPR analysis.

The similarity to known repeats could be extended to provide a predictive measure; however, its validation would require a

different reference set. We generated another reference set by using the 528 arrays that have ATTGAAA(N) at the 3′ end, and of these, 140 had <4 mismatches to the 26 known repeats (Supplementary Table S3). All 140 were correctly orientated. Up to 7 mismatches provided a PPV of 1.0 (250 arrays), and 43% of all the arrays had this amount of similarity (Supplement T3). This alternative approach is also implemented in CRISPRDirection. The use of this motif as a predictor (or any user-defined motif) can be included optionally in the CRISPRDirection program. The score can also be changed, e.g. increased to 4.5, to override the other predictions.

In both the previous sets, there are clearly some sequences with incorrect orientation, based on subsequent experimentally verified repeats (Kunin *et al*., 2007 and Lange *et al*., 2013). For example, the type I-F systems have a structured repeat that commonly ends in CUAAGAAA (Fig. 5 and Supplementary Table S7) (Haurwitz *et al*., 2010; Mojica *et al*., 2009; Przybilski *et al*., 2011; Richter *et al*., 2012b; Sternberg *et al*., 2012). These are represented in the classification of Kunin *et al*. (2007) as CRISPR-4. They comprise classes, M4 and F5, and incorrectly, its reverse complement, M6 and F8, in the classification of Lange *et al*. (2013). The work described here will allow further merging and refinement of these repeat classes, after prior prediction of the repeat direction.

4.2 Specific methods

We used one specific motif set in this analysis (AUUGAAA-like), as it was present in one-third of arrays. Cleavage immediately before this sequence generates the 5′ handle on the crRNA in several systems (Nickel *et al*., 2013; Scholz *et al*., 2013). Variations on this also provided strong predictors, as long as the G was retained (Table 1). These or other variants could also be used as predictors, particularly if the type was known. For example, to predict type I-F–associated CRISPRs, a variant of this in which there is reordering of the leading AU right region may be used (*UUA*GAAA) (Fig. 5). We searched for motifs at the end of the sequence but also allowed an additional base to be present in the repeat prediction. Current CRISPR array prediction algorithms cannot always accurately predict the termini of repeats, e.g. if the first base of the following spacer is also repeated (Fig. 5 and Supplementary Table S7).

Other sequence motifs can be defined by users in the repeats. For example, the presence of motif 'GUU' in the first bases of the repeats improved accuracy, and TATAAT-like sequences function in some repeats (Zhang *et al*., 2013). However, as some CRISPRs (e.g. I-E) do not obey these rules, and because our aim was to make the analysis unbiased to specific properties, we did not use more complex motif analysis by default here.

The relative AT% on either side of the repeat provided a useful predictor for 177 of the 460 arrays (Fig. 6). This is likely to represent the presence of AT-rich promoter sequences (e.g. TATAAT) and GC-rich structured terminator sequences. In addition, the leader region is important for acquisition of new spacers, which may require AT-rich binding sequences (Diez–Villasenor *et al*., 2013; Yosef *et al*., 2012). The automated determination of the direction of arrays will enable further computational analysis of these regulatory regions (Pul *et al*., 2010). The relative AT% could also possibly represent adjacent tracrRNA (Type II) or *cas* genes (Zhang *et al*., 2013).

For 197 of the reference arrays, the presence of a greater number of mutations at the 3′ trailer end provided a specific predictor (Fig. 3A). We believe this reflects the degeneration of older repeats at the trailer end. Therefore, CRISPRDirection will facilitate studies of the evolution of these arrays (Gudbergsdottir *et al*., 2011; He and Deem, 2010; Weinberger *et al*., 2012).

Calculation of a likelihood of forming an RNA structure worked poorly as a predictor on full-length repeats (Fig. 3B). This was surprising, as some types have structures (I, II) and these are well represented in our reference set. However, these structures are commonly small and located near the center of the repeat, with the termini involved in other processes (see above). Exclusion of the five bases at each end (e.g. UGAAA in more than one-third of the sequences) improved the PPV (Fig. 3B).

The relative number of A compared with T in repeats has previously been used to predict CRISPR direction. In particular, this is used as a predictor of the directions of arrays in the UCSC Archaeal Genome Browser (Chan *et al*., 2012). This gave a PPV of 0.72 overall, but was most effective (0.87) when applied to repeats with <50% AT (Fig. 3C).

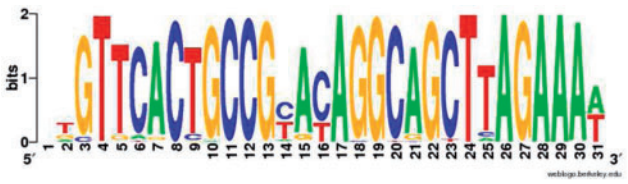

Fig. 5. Base conservation in repeats associated with type I-F systems. Classes F5 and the reverse complement of F8 (Lange *et al*., 2013) were merged to form one class, the orientation was determined experimentally from Przybilski, *et al*. (2011) and correctly predicted by CRISPRDirection. There are 109 sequences of this class in our reference dataset. The true length is 28 bases (3–30 on diagram). Extra base(s) at the left (numbered 1–2) and right (31) ends of some predicted repeats (6, or 22 of 223); in addition, some predictions are shorter. These are due to mispredictions of the ends of the repeats. The full alignment is in supplementary File S1 (T7)

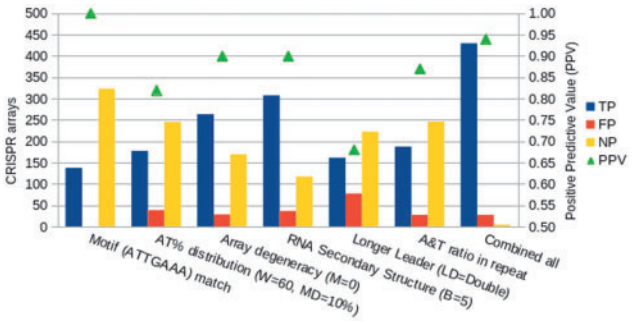

Fig. 6. Comparison of the six individual methods to the combined method. The number of true predictions (TP), false predictions (FP) and no predictions (NP) and PPV (green triangles) are shown for each. B = bases masked from either end, W = Window size, MD = Minimum AT difference, M = Minimum number of mutations per array

We examined the distance to the next annotated CDS as a predictor (Fig. 4). This was not an effective predictor unless the sequences were relatively long. This is the only measure used here that depends on prior annotation of the genome, and its exclusion (Supplementary Fig. S2) only reduced precision by 1%. Therefore, when CRISPRDirection is used as part of a pipeline it could be run before CDS prediction. This order would have the advantage of avoiding the many likely spurious CDS predictions within CRISPR arrays (e.g. 99 of 460 arrays in the RefSeq annotation). CRISPRDirection could be built into such pipelines, such as that used by the JGI-DOE (Mavromatis *et al*., 2009). A disadvantage of CRISPR prediction before CDS prediction is that repeats in some proteins may be falsely predicted as arrays by CRISPR prediction algorithms.

The combined analysis is robust to the removal of any of the predictors or combinations of predictors (Supplementary Fig. S2). Normally a prediction is made that depends on several methods. This confidence score normally ranges from 2.6 (where all parameters when added are concordant, Section 2.2.8) to 0.04 (when predictors are discordant or most are NP). For each prediction, the confidence score can be used to classify the prediction as high, medium or low confidence (Section 2.2.8). The 32 arrays, mainly bacterial, for which a false prediction was made had no obvious similarity in terms of species origin, repeat sequence or likely class (Supplementary Table S6). However, 30 of the 32 false predictions were classified as low confidence.

In the future, CRISPRDirection could be built into a more sophisticated CRISPR array finding algorithm. This could include additional automatically calculable parameters, or user inputted parameters. We observed that specific sequences are common near the ends of the repeats, such as GAA 3′, GTT 5′, and specific sequences are also found in the leader/promoter region. These could likely be used to refine the prediction algorithm, particularly in those cases where the direction is unable to be called with confidence. This could also be provided in a later interactive CRISPR prediction interface where expert users could input relevant predictors, or balance the parameters for their own species.

Funding: P.C.F. was supported by a Rutherford Discovery Fellowship from the Royal Society of NZ. C.M.B. was supported by a Human Frontier Science Program Grant (RGP0031/2009 to Ian Macara, Anne Spang and C.M.B). A.B. is a recipient of a University of Otago Postgraduate Scholarship.

Conflict of Interest: none declared.

REFERENCES

Biswas,A. *et al.* (2013) CRISPRTarget: bioinformatic prediction and analysis of crRNA targets. *RNA Biol*., **10**, 817–827.

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

Brodt,A. *et al.* (2011) CRISPR loci reveal networks of gene exchange in archaea. *Biol. Direct*, **6**, 65.

Chan,P.P. *et al.* (2012) The UCSC Archaeal Genome Browser: 2012 update. *Nucleic Acids Res*., **40**, D646–D652.

Diez-Villasenor,C. *et al.* (2013) CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of *Escherichia coli*. *RNA Biol*., **10**, 792–802.

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

Erdmann,S. and Garrett,R.A. (2012) Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. *Mol. Microbiol*., **85**, 1044–1056.

Fineran,P.C. and Charpentier,E. (2012) Memory of viral infections by CRISPR-Cas adaptive immune systems: acquisition of new information. *Virology*, **434**, 202–209.

Grissa,I. *et al.* (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*., **35**, W52–W57.

Gudbergsdottir,S. *et al.* (2011) Dynamic properties of the Sulfolobus CRISPR/Cas and CRISPR/Cmr systems when challenged with vector-borne viral and plasmid genes and protospacers. *Mol. Microbiol*., **79**, 35–49.

Hale,C.R. *et al.* (2009) RNA-guided RNA cleavage by a CRISPR RNA-Cas protein complex. *Cell*, **139**, 945–956.

Haurwitz,R.E. *et al.* (2010) Sequence- and structure-specific RNA processing by a CRISPR endonuclease. *Science*, **329**, 1355–1358.

He,J. and Deem,M.W. (2010) Heterogeneous diversity of spacers within CRISPR (clustered regularly interspaced short palindromic repeats). *Phys. Rev. Lett*., **105**, 128102.

Hofacker,I.L. *et al.* (1994) Fast folding and comparison of RNA secondary structures, *Monatsh. Chem*., **125**, 167–188.

Kunin,V. *et al.* (2007) Evolutionary conservation of sequence and secondary structures in CRISPR repeats. *Genome Biol*., **8**, R61.

Lange,S.J. *et al.* (2013) CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. *Nucleic Acids Res*., **41**, 8034–8044.

Levin,B.R. *et al.* (2013) The population and evolutionary dynamics of phage and bacteria with CRISPR-mediated immunity. *PLoS Genet*., **9**, e1003312.

Lillestol,R.K. *et al.* (2009) CRISPR families of the crenarchaeal genus Sulfolobus: bidirectional transcription and dynamic properties. *Mol. Microbiol*., **72**, 259–272.

Maier,L.K. *et al.* (2013) Essential requirements for the detection and degradation of invaders by the *Haloferax volcanii* CRISPR/Cas system I-B. *RNA Biol*., **10**, 865–874.

Makarova,K.S. *et al.* (2011) Evolution and classification of the CRISPR-Cas systems. *Nat. Rev. Microbiol*., **9**, 467–477.

Makarova,K.S. *et al.* (2013) Comparative genomics of defense systems in archaea and bacteria. *Nucleic Acids Res*., **41**, 4360–4377.

Mavromatis,K. *et al.* (2009) The DOE-JGI standard operating procedure for the annotations of microbial genomes. *Stand. Genomic Sci*., **1**, 63–67.

Mojica,F.J. *et al.* (2009) Short motif sequences determine the targets of the prokaryotic CRISPR defence system. *Microbiology*, **155**, 733–740.

Nickel,L. *et al.* (2013) Two CRISPR-Cas systems in Methanosarcina mazei strain Go1 display common processing features despite belonging to different types I and III. *RNA Biol*., **10**, 779–791.

Pougach,K. *et al.* (2010) Transcription, processing and function of CRISPR cassettes in *Escherichia coli*. *Mol. Microbiol*., **77**, 1367–1379.

Przybilski,R. *et al.* (2011) Csy4 is responsible for CRISPR RNA processing in *Pectobacterium atrosepticum*. *RNA Biol*., **8**, 517–528.

Pul,U. *et al.* (2010) Identification and characterization of *E. coli* CRISPR-cas promoters and their silencing by H-NS. *Mol. Microbiol*., **75**, 1495–1512.

Rho,M. *et al.* (2012) Diverse CRISPRs evolving in human microbiomes. *PLoS Genet*., **8**, e1002441.

Richter,C. *et al.* (2012a) Function and regulation of Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR)/CRISPR associated (Cas) systems. *Viruses*, **4**, 2291–2311.

Richter,C. *et al.* (2012b) *In vivo* protein interactions and complex formation in the *Pectobacterium atrosepticum* subtype I-F CRISPR/Cas system. *PLoS One*, **7**, e49549.

Rousseau,C. *et al.* (2009) CRISPI: a CRISPR interactive database. *Bioinformatics*, **25**, 3317–3318.

Scholz,I. *et al.* (2013) CRISPR-Cas Systems in the Cyanobacterium Synechocystis sp. PCC6803 exhibit distinct processing pathways involving at least two Cas6 and a Cmr2 protein. *PLoS One*, **8**, e56470.

Shah,S.A. and Garrett,R.A. (2011) CRISPR/Cas and Cmr modules, mobility and evolution of adaptive immune systems. *Res. Microbiol*., **162**, 27–38.

Shah,S.A. *et al.* (2013) Protospacer recognition motifs: mixed identities and functional diversity. *RNA Biol*., **10**, 891–899.

Skennerton,C.T. *et al.* (2013) Crass: identification and reconstruction of CRISPR from unassembled metagenomic data. *Nucleic Acids Res*., **41**, e105.

Sorek,R. *et al.* (2013) CRISPR-mediated adaptive immune systems in bacteria and archaea. *Annu. Rev. Biochem.*, **82**, 237–266.

Sternberg,S.H. *et al.* (2012) Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. *RNA*, **18**, 661–672.

Sun,C.L. *et al.* (2013) Phage mutations in response to CRISPR diversification in a bacterial population. *Environ. Microbiol.*, **15**, 463–470.

Swarts,D.C. *et al.* (2012) CRISPR interference directs strand specific spacer acquisition. *PLoS One*, **7**, e35888.

Vercoe,R.B. *et al.* (2013) Cytotoxic chromosomal targeting by CRISPR/Cas systems can reshape bacterial genomes and expel or remodel pathogenicity islands. *PLoS Genet.*, **9**, e1003454.

Wang,R. *et al.* (2011) Interaction of the Cas6 riboendonuclease with CRISPR RNAs: recognition and cleavage. *Structure*, **19**, 257–264.

Weinberger,A.D. *et al.* (2012) Persisting viral sequences shape microbial CRISPR-based immunity. *PLoS Comput. Biol.*, **8**, e1002475.

Westra,E.R. and Brouns,S.J. (2012) The rise and fall of CRISPRs—dynamics of spacer acquisition and loss. *Mol. Microbiol.*, **85**, 1021–1025.

Westra,E.R. *et al.* (2013) CRISPR-Cas systems preferentially target the leading regions of MOBF conjugative plasmids. *RNA Biol.*, **10**, 749–761.

Wiedenheft,B. *et al.* (2012) RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, **482**, 331–338.

Yosef,I. *et al.* (2012) Proteins and DNA elements essential for the CRISPR adaptation process in *Escherichia coli*. *Nucleic Acids Res.*, **40**, 5569–5576.

Zhang,Y. *et al.* (2013) Processing-independent CRISPR RNAs limit natural transformation in Neisseria meningitidis. *Mol. Cell*, **50**, 488–503.