RESEARCH

Evaluating tools for metagenomic analysis of CRISPR loci

Madeleine Bonsma-Fisher[??*], Ue-Yu Pen[??], Christian Farfan Centeno[??], Sidhartha Goyal[??] and Michael Brudno[??]

[*]Correspondence:
mbonsma@physics.utoronto.ca
[??]Department of Physics, University of Toronto, St. George St., Toronto, Canada
Full list of author information is available at the end of the article
[†]Equal contributor

Abstract

In this work we tested recently developed tools for CRISPR detection in metagenomic data. Crass, CRISPRfinder, CRISPRAlign, and MetaCRT are applied to both simulated CRISPR data and real metagenomic data from the Human Microbiome Project. The goal is to reveal what types of information are best provided by each tool, and how they might be combined to yield a better understanding of the data. Our results show that no tool is perfect. CRISPRAlign detected the most repeats and spacers in a Human Microbiome Project dataset, but was given the "correct" repeat sequences a priori. Crass detected the most spacers in two of the three simulated datasets tested. However, roughly half of the repeat groups detected by Crass were not present in the original data. The results given in this report are very helpful to our own research group when making choices for future studies and can benefit other researchers as well.

First part title: Text for this section.

Second part title: Text for this section.

Keywords: CRISPR; algorithm; CRISPRfinder; Crass; CRISPRAlign; MetaCRT

Introduction

The CRISPR-Cas system is a recently discovered adaptive immune system in bacteria and archaea [1–3]. Organisms possessing this mechanism can sample pieces of invasive viral or plasmid DNA, integrate it into their own genome, and then launch a targeted immune response using these fragments. The foreign fragments, called spacers, are inserted between short repeated segments called repeats (or direct repeats) into the CRISPR locus. Organisms may possess more than one CRISPR locus in their genome. Many spacers can be stored in a CRISPR locus, but most CRISPR loci contain under 50 spacers. New spacers are added always at the leader end of the locus, meaning the CRISPR locus can be studied as a genetic record of immunization [4].

Alongside ever-increasing CRISPR discoveries, DNA sequencing technology has also advanced dramatically in recent years [5–7]. The advent of next-generation sequencing technology opened the door to metagenomic analysis of all DNA present in an environmental sample [8]. The power of metagenomics combined with the window into microbial communities that CRISPR provides has prompted recent research to investigate CRISPR dynamics in microbiomes [9–12] as well as development of specialized tools to detect, assemble, and analyze metagenomic CRISPR data [13–15].

Analysis of CRISPRs in metagenomic data is challenging since assembly algorithms can be confused by the repetitive nature of CRISPR loci or by strain variation. This challenge is much less pronounced in a clonal population: assembling a clonal CRISPR locus is straightforward if read lengths are longer than the period of the repeat. (CRISPR repeats can range in length from 23 to 55 nucleotides and spacers from 21 to 72 nucleotides; a typical combined spacer-repeat segment ranges between 60 and 75 nucleotides [4]. Many common sequencing platforms yield reads much longer than this [16].) Metagenomes, however, are an entirely different story. In an actively evolving population of bacteria and phages, changes in spacer content can be observed in a matter of days [17]. A single species may have as many unique arrangements of spacers within the CRISPR locus as there are organisms in that species. As in [17], one approach to capture this spacer diversity is to use PCR to amplify CRISPR arrays if the organism of interest's genome is known [18]. In a diverse metagenomic population for which many organisms may be without reference genomes, however, it is unrealistic to specifically target certain CRISPRs as a means of learning about the entire population.

Detected CRISPR diversity is likely to decrease after metagenomic assembly because of the tendency of assembly algorithms to attempt to find consensus sequences. The problem is compounded by sequencing errors present in all real data. Most assemblers use certain techniques to reduce sensitivity to errors, which can also mask novel CRISPRS. Assembled CRISPRs may be chimeric (consisting of separate genomes artificially joined together), and simulations indicate chimeric sequences occur in high rates when assembling complex metagenomic data sets [19]. This means that in de novo assembly of environmental samples, chimeric CRISPRS, sequencing error CRISPRs and truly novel low frequency CRISPRs can be hard to separate.

An alternative approach is to identify CRISPR segments directly from whole-genome shotgun (WGS) metagenomic data. Reads are sampled with high coverage from the target genomes at random positions, ideally creating enough overlaps to deduce consensus sequences. CRISPR segments can be detected in raw reads before assembly, however, which has the potential to preserve much more of the natural strain diversity of the population than will be present after an assembler has determined consensus sequences.

There exist specific tools to analyze CRISPRs in metagenomic data. In this project we characterize and test recently developed tools for CRISPR detection and assembly in metagenomic data, comparing the same measured quantities for each tool. Our results show some of the strengths and weaknesses of each approach (detecting spacers and repeats from unassembled reads or from assembled metagenomes). We quantify the amount of data detected by each program as well as the accuracy of the detected results using simulated data, and we compare these results with real metagenomic data from the Human Microbiome Project. Previous studies have applied several of these tools to the Human Microbiome Project datasets and have combined several approaches [11,12]. Here we show a standardized comparison of the tools themselves, focused less on situation-specific results from the data.

Methods

Simulations

We simulated metagenomic data using MetaSim [20]. We simulated 101 bp metagenomic paired-end reads with an error profile modified from the MetaSim Illumina default to ensure that the resulting simulated data is realistically close to the HMP data set, which consists of 101 bp paired-end reads from the Illumina GAIIx platform [21]. The number of read pairs used was chosen to give approximate coverages of 2, 5, and 20. Coverage was estimated by multiplying the number of reads generated for one of the genomes by the average read length and dividing by the total length of that genome. This is a relatively stable estimate, since MetaSim weights the generated reads by the length of each genome in the database.

Previous work to detect CRISPRs has failed to accurately capture the complexity of true metagenomic data. In [11] and [12], a simulated dataset is created with six and eight complete genomes respectively. These simulations greatly underestimate the complexity of CRISPRs in a real metagenomic dataset. As outlined above, in an actively evolving population there will be not only many different species but also many unique CRISPR loci within species. In a simulated dataset with only one consensus sequence per species, assigning spacers to a repeat will be unnaturally simple.

We created three simulated datasets using whole genome sequences from several organisms that contain CRISPR loci according to CRISPRdb [22].

One of the datasets uses the same 6 organisms as [11] - Streptococcus mutans NN2025, Escherichia coli K-12 MG1655, Azospirillum sp. B510, Deferribacter desulfuricans SSM1, Dehalococcoides sp. GT, and Erwinia amylovora ATCC 49946. This group contains no exactly overlapping repeats and only two repeats that overlap with less than 2 nucleotide differences. This dataset is referred to as "6 organisms non-overlapping" in the analysis.

The second dataset is made with 6 organisms (E. coli K-12 MG1655, E. coli UMNK88, E. coli HS, Lactobacillus acidophilus La-14, L. amylovorus GRL 1118, and L. crispatus ST1) for which some of the repeats are identical between organisms and others are very similar to each other. This dataset, referred to as "6 organisms overlapping", is intended to probe whether or not chimeral or fictitious CRISPRs are more likely to occur when multiple different organisms possess the same repeat.

The last dataset is made with10 organisms for which all repeats are more than 2 nucleotides different from each other (E. coli K-12 MG1655, L. crispatus ST1, Streptococcus anginosus C1051, Bacteroides fragilis NCTC 9343, Clostridium difficile M120, Parabacteroides distasonis ATCC 8503, Campylobacter sp. 03-427, Bifidobacterium longum DJO10A, B. breve UCC2003, and Olsenella uli DSM 7084). This dataset is called "10 organisms non-overlapping".

All organisms in the "6 organisms overlapping" and "10 organisms non-overlapping" datasets were chosen from a list of species or genera that are present in the human body as reported by the Human Microbiome Project so that the resulting metagenomic data would be as realistic as possible. In all three datasets, the number of unique spacers is very close to the total number of spacers (6 organisms non-overlapping: 422 total, 415 unique; 6 organisms overlapping: 174 total, 163 unique; 10 organisms non-overlapping: 485 total, 476 unique), meaning that there are no CRISPR loci that are exactly duplicated even if they share the same repeat.

CRISPR detection

Crass [13]: Crass is made to identify CRISPRs from metagenomic data by detecting the DRs without prior knowledge of the CRISPR. Crass finds the repeats in following steps: first is to identify the sequences with repeated short substrings and create a database of potential repeats. Second is to find sequences with one copy of any repeat in the database and group similar repeat together base on single-linkage clustering and k-mer composition. After that, Crass builds a path graph from the spacer ends, and outputs a spacer graph. In this work, the length that considered acceptable for the size of a direct repeat is 23-47 dp for a spacer is 26-50 dp, and the minimum number of spacers allowed for a putative CRISPR to be considered real is 3, and the window length for finding seed sequences when using the long read search algorithm is set as 8bp.

CRISPRAlign [14] was used to identify repeats in unassembled reads that are similar to given reference CRISPR sequence. It first detects substrings in the reads that are similar to the given reference sequence, and was then set to accept partial repeats at the end of reads. Sequences were considered a match when within a set number of nucleotide differences, which we have set to 2.

MetaCRT [14] was used to identify novel CRISPRS in metagenomic assemblies by detecting repetitive subsequences that are separated by similar distances (i.e. likely spacers) and then filtering for CRISPR-specific motifs, such as non-repeating spacers and aforementioned similar size. It considers incomplete repeats at the end of assembly contigs. No run-time parameters available, any post data processing must be done by the user.

CRISPRFinder [22]: CRISPRFinder detect possible locations of CRISPRs (consisting of at least one motif) by finding maximal repeats. A maximal repeat (of length 23-55 bp) is defined as a repeat that are not able to be extended in either direction without experiencing a mismatch. A CRISPR pattern of two DRs and a spacer may be considered as a maximal repeat where the repeated sequences are separated by a sequence of approximately the same length [27].

MetaVelvet [23,24] was used to assemble metagenomic data, both from the simulations and the HMP. Single genome assemblers for metagenomic data are limited, in that sequences of recurring species are misidentified as repeats in a single genome. For CRISPR datasets this is especially problematic. MetaVelvet is a metagenomic extension to the single genome assembler known as Velvet. Velvet attempts to assemble the given raw reads by applying a series of heuristics to reduce graph complexity: (i) the length of the K-mers is odd to exclude nodes from palindromic repeats (ii) The minimum expected frequency of K-mers in reads controls which K-mers are eroded on a first pass (iii) The expected coverage of the genome controls processes involved in path reduction. For the simulations, Velvet was run with the expected insert length of 200, and expected coverage set to match the simulation. After the normal Velvet assembly is run, metaVelvet decomposes the graph into individual sub-graphs, by detecting peaks in the distribution of k-mer frequencies, and assumed these peaks correspond to individual species. Contigs and scaffolds are then constructed based on the decomposed subgraphs.

Analysis

The list of repeats and spacers found was analyzed for each program and each individual dataset. Several quantities were counted for the simulated data:

- the total number of repeats and spacers found;
- the total number of unique repeats and spacers found;
- the total number of repeats and spacers that match the known CRISPRs in the reference genomes used to make the simulated data (either exactly or within a cutoff);
- the total number of spacers that matched the reference genomes but had been paired with the wrong repeat.

For the HMP data, only the total unique repeats and spacers were counted since no comparison could be made to an exact known CRISPR complement.

Unless otherwise noted, only the total number of unique repeats and spacers was used in further analysis. This was because MetaCRT, CRISPRAlign, and CRISPRfinder did not group repeats and spacers by type, and so with short sequences this produced artificially high numbers of total repeats and spacers when many reads with the same repeat and spacer were detected.

We used two measures of 'sameness' for repeat and spacer sequences: the first considers two sequences the same only if they are an exact match, and the second considers two sequences the same if their global alignment score is no less than 2 lower than the length of the shorter sequence. (Match, mismatch, and gap scores were (1,0,0).) This metric is a measure of edit distance without penalizing start or end gaps. Additionally, since the direction of the CRISPR locus in a bacterial genome is ambiguous, we considered two repeats or spacers to match if either they or their reverse complements could be aligned within the selected cutoff.

An appropriate cutoff is one for which the total number of spacer types does not change significantly within a certain range - this indicates that the grouping is tolerant of small amounts of error (sequencing or biological) but not artificially grouping different types together. This effect was present in the detected spacer sequences; Figure 1 (left) shows the total number of spacer types for one of the simulated datasets and a set of spacers detected using CRISPRAlign (which doesn't group detected sequences based on similarity like Crass does) as the cutoff is increased. The number falls initially as inexact matches are included, but plateaus beginning at a cutoff of 2. For repeats, however, the CRISPRAlign-detected number of types did not show this pattern (Figure 1 right), meaning that there is really no reasonable choice of cutoff. To match the spacer results, we used a cutoff of 2 in all analysis.

When normalizing simulated data results, we used the total number of exactly unique repeats and spacers in the reference CRISPRs, also referred to as the original sequences. In every case, the genomes making up the simulated data contained at least once duplicate repeat sequence, usually two CRISPR regions in the same genome that had the same repeat sequence.

Results

All measured quantities for the HMP data can be found in Table 1 in the appendix. All measured quantities for the simulated data can be found in Tables 2 through 4 in the appendix. The following figures and accompanying text highlight a few of the most significant results.

Simulated data

To measure the accuracy of each tool, we calculated the fraction of repeats and spacers found by the tool that matched the reference CRISPRs within the chosen cutoff of 2. These we call "matched repeats" or "matched spacers". Repeats or spacers that don't match the reference CRISPRs at all we call "ghosts". Spacers that do match something in the reference CRISPR but haven't been paired with the correct repeat sequence by the tool are called "mismatched spacers." The mismatch quantity is a way to estimate the likelihood of chimeric CRISPRs if the data were to be assembled - these spacers are correctly identified, but they've been assigned to the wrong CRISPR sequence. The quantity matched spacers plus mismatched spacers is a measure of accuracy: this is the fraction of the original spacers present in the simulated data that were correctly identified by the tool.

We tested three different coverages in the simulated data. In almost every dataset with each program, as the coverage increases, the total number of detected repeats and spacers increases as well. This is reasonable, since higher coverage decreases the chances of missing things due to gaps. However, increasing coverage does not necessarily increase the accuracy of spacer and repeat detection. Figure 2 shows the unnormalized numbers of matched and mismatched repeats and matched, mismatched, and ghost spacers. The ratio between total, matched, and ghost spacers remains roughly the same as the coverage increases, and the ratio of total to mismatched repeats also remains roughly constant. From this we conclude that there is no extra advantage to high coverage when detecting CRISPRs in metagenomic data except that the raw number of "correct" sequences detected will likely increase as the coverage increases.

Figure 3 summarizes our results from running each tool on the simulated data. Shown are the fractional matched, mismatched, and ghost spacers (top row), and matched and ghost repeats (bottom row) for each of the three simulated datasets at 20x coverage. Each bar is normalized to the total number of detected sequences for that particular tool and dataset; the total number is shown above each bar. Note that the first two bars for each plot present the results of MetaCRT and CRISPRfinder on the unchopped simulated reference genomes, and they both perform very well - in essence this is a measure of how well those tools would do if you had a perfect assembler that returned the complete original genome sequences.

For the tools that run on the raw reads, as shown in the rest of bars on each plot, it can be seen that, generally, CRISPRAlign does the best job of finding spacers and repeats correctly (with lower fractions of ghost spacers and repeats) regardless of the presence of error or not. It even appears better than the results of MetaCRT and CRISPRfinder on the unchopped simulated reference genomes. For the middle and right-hand column, however, CRISPRAlign finds many fewer spacers and repeats. This few-but-accurate result may be caused by the working mechanism of CRISPRAlign, which is designed to identify repeats in unassembled reads that are similar to given reference repeat sequences. Working with the reference gives CRISPRAlign a good chance of finding the "right" repeats, and hence finding the "right" spacers. It is possible that the allowed tolerance for error in CRISPRAlign was set too small. On the other hand, Crass shows its rather balanced property of finding good amount of spacers with reasonable quality while MetaCRT seems to

be doing a better job of finding repeats. Crass and CRISPRAlign are affected by the organism overlapping of the dataset and get worse results compared with other datasets, while MetaCRT seems to be doing better in the organism overlapping dataset. Figure 3 also shows the effect of errors in the database, and in most cases the errors cause the fraction of correctly detected repeats and spacers to drop, as can be expected.

Figure 4 compares the results of MetaCRT run on simulated data, first on the already-assembled reference genomes, and second on the simulated reads after assembly with MetaVelvet. MetaCRT detected fewer spacers and repeats when it was run on the simulated reads after assembly with MetaVelvet, compared to when it was run on the original genome files. For two of the datasets (6 organisms overlapping and 10 organisms non-overlapping), the total number of detected spacers decreased only slightly, but the relative number of ghost spacers increased dramatically. This is an example of an accuracy issue that would not be detected in simulations that merely count the total number of detected spacers. The decrease in total number and increase in ghost spacers are good indications that a conventional metagenomic assembler may have difficulty assembling CRISPR regions. Interestingly, the number of mismatched spacers decreased after assembly in all three cases.

HMP data

Figure 5 shows the performance of the various CRISPR tools on two HMP data sets: SRS018394 (Supragingval plaque) and SRS019071 (Tongue dorsum). The number of unique repeats and spacers are displayed side by side for CRASS and CRISPRAlign (run on the raw reads from each sample), and MetaCRT and CRISPRFinder (run on the metagenomic assembly for each sample provided by the Human Microbiome Project) . CRASS clearly finds the greatest amount of unique spacers de novo, with no reference repeats or prior assembly provided by another program. Both metaCRT and CRISPRFinder take the HMP assembled data as input, and perform similarly, despite the differing algorithmic approaches. All three of these tools find a similar number of unique repeats, and are greatly outperformed by CRISPRAlign in this regard.

While analysis of the results shown in Figure 3 indicate that CRISPRAlign is not likely to find many Ghost repeats, this increase in repeats found is not strictly a triumph of the CRISPRAlign algorithm. Since CRISPRAlign works by looking for sequences similar to input repeats, it's success is based chiefly on being fed the "correct" repeats, and then looking for similarities. However it doesn't group them by type, and thus will report back a multiplicity of "unique" spacers that are actually the same spacer due to errors in the reads. A more trustworthy measure of performance is the number of unique spacers found. Here CRASS and CRISPRAlign both greatly outperform the other two tools in raw number of spacers identified. However as addressed seen in Figure 3, this is no guarantee of accuracy.

Discussion

Several features of our results suggest that the results of CRISPR detection in metagenomic data should be interpreted with caution. It is clear from Figure 3 that the accuracy of each tool is strongly dependent on the dataset, which suggests

that extreme care must be taken when searching for CRISPRs in metagenomic data. Also, in almost every case of the simulated data we detected ghost spacers and repeats, indicating that none of the programs can perfectly detect CRISPR sequences. The exception to this is CRISPRAlign, but it must be stressed that CRISPRAlign searches only for matches to repeat sequences it has been given and so its task is somewhat simplified from the others. CRISPRfinder and MetaCRT in the simulated data had the advantage of being given the assembled reference genomes as input instead of simulated short reads. Crass, the only true de novo detector in raw reads, still picks up a fairly large fraction of fictitious spacers and repeats.

The question of what "accurate" means can also be debated. The reference CRISPR repeat and spacer content, which was used in the simulated data was taken from the CRISPRdb website [Grissa2007], which uses CRISPRfinder to construct its CRISPRs. Hence, CRISPRfinder results matched the "correct" answer very closely. This is a potential bias in the results, since the actually correct CRISPR sequence is difficult to verify (except perhaps by manually inspecting the genome, which we did not do).

The number of matched, mismatched, and ghost repeats and spacers can in general depend very strongly on the cutoff. We found that the number of spacer types was relatively stable within a certain range of cutoff, but the number of repeat types did not plateau in the same way. This means that, had we chosen a different cutoff, the results for the repeats in particular could look quite different.

In our analysis, we did not attempt targeted assembly, the method of choice in [11]. Targeted assembly involves attempting metagenomic assembly on raw reads, running CRISPR identification tools, such as metaCRT, and then somehow filtering reads corresponding to the found CRISPRS through certain heuristic choices. The remaining reads are then assembled, theoretically providing more accurate sequences for the assembled CRISPR loci, as the information used to construct the final assembly of contigs has been pruned down to 'useful' reads only. We theorize that CRASS, which was not used for targeted assembly in [11], can prove useful in this approach. Firstly, referring to Figure 5, we see that CRASS finds a comparable amount of unique repeats as CRISPRFinder and metaCRT. Despite being prone to discovering a high fraction of ghost repeats in the simulation results of Figure 3, we see that CRASS does not report a number of repeats scaling up as suddenly as the other tools with more complicated data sets, suggesting that the performance in Figure 5 when working with real data indicates good performance. As such, CRASS could theoretically provide a better pool of reads to to filter, resulting in a better final assembly.

There are many more patterns and anomalies in our results than have been discussed here. It may be that the only safe conclusion is that no tool can work perfectly on metagenomic data.

We have not considered the final step that some may wish to take when analyzing CRISPR data: that of actually assembling the CRISPR locus from fragmentary repeat and spacer information. Tools that return a single consensus sequence for a given repeat type will often be inadequate for use in metagenomic data, since there will be many possible arrangements of spacers that are truly present in the sample

that won't be captured. The next best thing is a graph of spacer connectedness for a given repeat sequence (as in the output of Crass) which doesn't remove the option of multiple paths. Even in this case, it is clear that the number of possible paths grows exponentially with the number of nodes, and the true number of paths is most likely much less than the possible paths.

One possible way to deal with this problem is to use extra information: the coverage of each edge in the graph. In the simplest case where each edge was present only once in the data (coverage = 1), traversing that edge would then remove that edge from all other possible paths, since if another path went through that edge then the coverage of that edge would be 2. This could potentially be extended to real data - edges will be weighted by coverage, and any final set of paths should produce a graph with the same coverages if one worked backwards. Once a final set has been obtained (in general there may be many that satisfy the coverage criteria), any information about the expected distribution of locus sizes could be used to narrow down which set is realistic. Even this heuristic may be too computationally intensive, however, and it is possible that it won't narrow down the options enough.

Conclusions

In this work, we have compared several recently developed tools for CRISPR detection. Our results go several steps beyond merely counting the number of discovered repeats and spacers by also providing a measure of accuracy for each tool's performance on simulated data.

We find that CRISPRAlign generally performs well given the correct repeats to search for in the simulated data. In the HMP data it detects the most repeats and spacers of all the tools. CRISPRfinder and MetaCRT are both quite accurate when run on the original assembled genomes used in the simulated data. In most cases in the simulated raw reads, MetaCRT returned the most ghost spacers of all the tools, with Crass not far behind. Crass detected the highest number of spacers in two of the three simulated datasets; on average 77% of them were actually present in the original data. We found no significant increase in accuracy of spacer and repeat discovery as coverage is increased.

Our results serve as a reference for each of the tools and highlight several of the challenges present in detecting CRISPRs in metagenomic data.

Sub-sub heading for section
Text for this sub-sub-heading ...

Sub-sub-sub heading for section Text for this sub-sub-sub-heading ... (also see [**?**, **?**, **?**, **?**, **?**]).

Competing interests
The authors declare that they have no competing interests.

Author's contributions
Text for this section ...

Acknowledgements
We thank Dr. Sidhartha Goyal and Dr. Michael Brudno for helpful discussions and advice.

Figure 1 Sample figure title. A short description of the figure content should go here.

Figure 2 Sample figure title. Figure legend text.

Table 1 Sample table title. This is where the description of the table should go.

	B1	B2	B3
A1	0.1	0.2	0.3
A2	...	..	.
A3	..	.	.

Figures
Tables
Additional Files
Additional file 1 — Sample additional file title
Additional file descriptions text (including details of how to view the file, if it is in a non-standard format or the file extension). This might refer to a multi-page table or a figure.

Additional file 2 — Sample additional file title
Additional file descriptions text.